Bayesian Networks for Max-linear Models

Claudia Klüppelberg and Steffen Lauritzen

Abstract We study Bayesian networks based on max-linear structural equations as introduced in Gissibl and Klüppelberg [16] and provide a summary of their independence properties. In particular we emphasize that distributions for such networks are generally not faithful to the independence model determined by their associated directed acyclic graph. In addition, we consider some of the basic issues of estimation and discuss generalized maximum likelihood estimation of the coefficients, using the concept of a generalized likelihood ratio for non-dominated families as introduced by Kiefer and Wolfowitz [21]. Finally we argue that the structure of a minimal network asymptotically can be identified completely from observational data.

1 Introduction

The type of model we are studying has been motivated by applications to risk analysis, where extreme risks play an essential role and may propagate through a network. For example, say, if an extreme rainfall happens on a specific location near a river network, it may effect water levels at other parts of the network in an essentially deterministic fashion. Similar phenomena occur in the analysis of risk for other complex systems.

Specifically, the model presented in (1) below arose in the context of technical risk analysis, more precisely, in an investigation of the "runway overrun" event of airplane landing. Numerous variables contribute to this event and extraordinary val-

Claudia Klüppelberg
Center for Mathematical Sciences, Technical University of Munich, 85748 Garching, Boltzmannstrasse 3, Germany; e-mail: cklu@tum.de

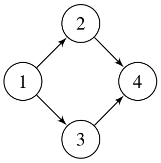
Steffen Lauritzen
Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark; e-mail: lauritzen@math.ku.dk

ues of some variables lead invariably to a runway overrun (see [18] for more details) naturally leading to questions about cause and effect of risky events. Other potential examples for risk-related cause and effect relations include chemical pollution of rivers ([35]), flooding in river networks ([1]), financial risk ([11]), and many others.

Statistical theory and applications of extreme value theory until the 1990s mainly focused on i.i.d. data as, for instance, yearly maximal water levels to predict future floodings or peaks over thresholds used to estimate the Value-at-Risk (e.g. [12]). From this, both theory and applications moved on to multivariate data, modelling risks like joint wind and wave extremes as well as extreme risks in financial portfolios [2]. The investigation of extremes in time series models have proved useful in financial and environmental risk analysis, and also in telecommunication (see e.g. the book [13]). More recently, extreme space-time models have been suggested and applied to environmental risk data [4, 6, 8, 19].

The paper focuses on first steps reporting on the methodological development associated with a specific class of network models. We begin with introducing our leading example of a recursive max-linear model which is Example 2.1 of [16]:

Example 1. Consider the network in the figure below:



Each node i in the network represents a random variable X_i and the joint distribution of $X=(X_1,X_2,X_3,X_4)$ is determined by a system of *max-linear structural equations*

$$X_1=Z_1,\ X_2=\max(c_{21}X_1,Z_2),\ X_3=\max(c_{31}X_1,Z_3),\ \max(c_{42}X_2,c_{43}X_3,Z_4),$$

where Z_1,Z_2,Z_3,Z_4 are independent positive random variables and the coefficients c_{ji} are all strictly positive.

The interpretation of a system like this is that each node in the network is subjected to a random shock Z_i and the effect from shocks of other nodes pointing to it, the latter being attenuated or amplified by the coefficients c_{ji}. To simplify notation here and later we write $a\vee b$ for $\max(a,b)$. We can alternatively represent $X=(X_1,X_2,X_3,X_4)$ directly in terms of the noise variables as

$$\begin{aligned}
X_1 &= Z_1\\
X_2 &= c_{21}X_1\vee Z_2 = c_{21}Z_1\vee Z_2\\
X_3 &= c_{31}X_1\vee Z_3 = c_{31}Z_1\vee Z_3\\
X_4 &= c_{42}X_2\vee c_{43}X_3\vee Z_4\\
&= c_{42}(c_{21}Z_1\vee Z_2)\vee c_{43}(c_{31}Z_1\vee Z_3)\vee Z_4\\
&= (c_{42}c_{21}\vee c_{43}c_{31})Z_1\vee c_{42}Z_2\vee c_{43}Z_3\vee Z_4.
\end{aligned}$$

We may then summarize the above coefficients to the noise variables $Z_1,\dots,Z_4$ in the matrix

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ c_{21} & 1 & 0 & 0 \\ c_{31} & 0 & 1 & 0 \\ c_{42}c_{21} \vee c_{43}c_{31} & c_{42} & c_{43} & 1 \end{pmatrix},$$

□

In greater generality we may write such a *recursive max-linear model* as

$$X_v = \bigvee_{u\in \mathrm{pa}(v)} c_{vu}X_k \vee c_{vv}Z_v, \quad v = 1,\dots,d, \tag{1}$$

where $\mathrm{pa}(v)$ denotes parents of v in a directed acyclic graph (DAG) and Z_v represent independent noise variables. The present article is concerned with such models and summarizes basic elements of Gissibl and Klüppelberg [16] and Gissibl et al. [17].

In this setting, natural candidates for the noise distributions are extreme value distributions or distributions in their domains of attraction resulting in a corresponding multivariate distribution with dependence structure given by the DAG (for details and background on multivariate extreme value models see e.g. [10, 27, 28]). The paper is structured as follows.

In Section 2 we establish the necessary terminology (Section 2.1), introduce Bayesian networks (Section 2.2), and basic properties of conditional independence (Section 2.3). In Section 2.4 we establish basic Markov properties of Bayesian networks. In Section 3 we study the specific Markov properties of Bayesian networks given by max-linear structural equations as in (1) and in Section 4 we study statistical properties of the models.

2 Preliminaries

2.1 Graph terminology

A *graph* as we use it here is determined by a finite vertex set V, an edge set E, and a map that to each edge e in E associates its endpoints $u,v \in V$. Our graphs are *simple* so that there are no self-loops (edges with identical endpoints) and no multiple edges. Therefore we can identify an edge e with its endpoints u,v so we can write $e = uv$. An edge uv of a *directed* graph points *from u to v* and we write $u \to v$. Then u is a *parent* of v and v is a *child* of u. The set of parents of v is denoted $\mathrm{pa}(v)$ and the set of children of u is $\mathrm{ch}(u)$. If uv is an edge we also say that u and v are adjacent and write $u \sim v$ whether or not the edge is directed.

A *walk* ω from u to v of *length* n is a sequence of vertices $\omega = [u = u_0, u_1, \dots, u_n = v]$ so that $u_{i-1} \sim u_i$ for all $i = 1,\dots,n$. A walk is a *cycle* if $u = v$. A *path* is a walk with no repeated vertices. The walk is *directed* from u to v if $u_{i-1} \to u_i$ for all i. If all edges in a graph $\mathscr{D} = (V,E)$ are directed, $\mathscr{D}$ is a *directed graph*. A directed graph is *acyclic* if it has no directed cycles. A *DAG* is a directed acyclic graph. A DAG is a *tree* if every vertex has at most one parent and a *polytree* if there is at most one path between two vertices u and v.

If there is a directed path from u to v in $\mathscr{D}$ we say that u is an *ancestor* of v and v a *descendant* of u and write $u \rightsquigarrow v$ or $v \leftsquigarrow u$. The set of ancestors of v is denoted $\mathrm{an}(v)$. A set $A \subseteq V$ is said to be *ancestral* if $\mathrm{an}(v) \subset A$ for all $v \in A$, or, alternatively, if $\mathrm{pa}(v) \subset A$ for all $v \in A$. For a subset A of V we let $\mathrm{An}(A)$ denote the smallest ancestral set containing A.

We say that the vertex set V of a DAG $\mathscr{D}$ is *well-ordered* if $V = \{1,\dots,d\}$ and all edges in $\mathscr{D}$ point from low to high, i.e. if $ij \in E \implies i < j$. Then the set of *predecessors* of a vertex i is $\mathrm{pr}(i) = \{1,\dots,i-1\}$.

For a DAG $\mathscr{D}$ we define its *moral graph* $\mathscr{D}^m$ as the simple, undirected graph with the same vertex set but with u and v adjacent in $\mathscr{D}^m$ if and only if either $u \sim v$ in $\mathscr{D}$ or if u and v have a common child. For further general graph terminology we refer the reader to West [38] but some of the concepts above are illustrated in Fig. 1.

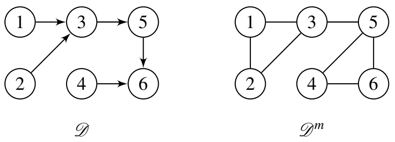


Fig. 1 A DAG $\mathscr{D}$ and its moral graph $\mathscr{D}^m$. In $\mathscr{D}$, 3 has parents $1,2$ and 5 is a child of 3. The DAG $\mathscr{D}$ is a polytree. The node 6 is a descendant of 1, and 2 is an ancestor of 4. The set $\{1,2,3,5\}$ is ancestral in $\mathscr{D}$. With the node numbering given, the DAG is well-ordered.

2.2 Bayesian networks

A real-valued Bayesian network associated to a given DAG $\mathscr{D} = (V,E)$ is determined by specifying random variables $X = (X_v, v \in V)$ and the conditional distribution of each of these, given values of their parent variables; for example as

$$P(X_v \leq x \,|\, X_{\mathrm{pa}(v)}) = F(x \,|\, x_{\mathrm{pa}}(v)).$$

Because there are no directed cycles in $\mathscr{D}$ there is a unique joint distribution corresponding to this specification.

Alternatively, as in Example 1, we can specify these conditional distributions through *structural equations* which describe the conditional distribution of X_v conditionally on $X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}$ in a functional form. More precisely a system of equations of the form

$$X_v = g_v(X_{\mathrm{pa}(v)}, Z_v), \quad v \in V, \tag{2}$$

where $(Z_v)_{v\in V}$ are independent noise variables and g_v suitable functions.

A system of structural equations as above is sometimes referred to as a *data generating mechanism*, interpreting each equation as a way of generating random variables with the desired conditional distribution.

An important instance of these models are *linear structural equation models* where the functions g_v are linear and hence

$$X_v = \sum_{u\in \mathrm{pa}(v)} c_{vu}X_u + c_{vv}Z_v, \quad v \in V, \tag{3}$$

where $c_{vu}, u \in \mathrm{pa}(v), c_{vv}$ are *structural coefficients*, see for example Bollen [3]. In general, a structural equation system need not be associated with a DAG, but if it is, the equation system is said to be *recursive*.

If the distributions of Z_v have heavy tails and all structural coefficients are nonnegative, the sum tends to be dominated by the largest term:

$$\sum_{u\in \mathrm{pa}(v)} c_{vu}X_u + c_{vv}Z_v \approx \bigvee_{u\in \mathrm{pa}(v)} c_{vu}X_u \vee c_{vv}Z_v$$

and hence for such cases, the max-linear variant in (4) as described in more detail in Section 3 below.

2.3 Conditional independence

The notion of conditional independence is at the heart of graphical models, including Bayesian networks. For three random variables (X,Y,Z) we say that X is conditionally independent of Y given Z if the conditional distribution of X given (Y,Z) does not depend on Y and we then write $X \perp\!\!\!\perp Y \,|\, Z$ or $X \perp\!\!\!\perp_P Y \,|\, Z$ if we wish to emphasize the dependence on the joint distribution P of (X,Y,Z).

The notion of conditional independence has a number of important properties, see e.g. Dawid [9] or Lauritzen [23].

Proposition 1. *Let $(\Omega,\mathbb{F},P)$ be a probability space and X, Y, Z, W random variables on Ω. Then the following properties hold.*

(C1) *If $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$ (symmetry);*
(C2) *If $X \perp\!\!\!\perp Y \,|\, Z$ and $W = \phi(Y)$ then $X \perp\!\!\!\perp W \,|\, Z$ (reduction);*
(C3) *If $X \perp\!\!\!\perp (Y,Z) \,|\, W$ then $X \perp\!\!\!\perp Y \,|\, (Z,W)$ (weak union);*
(C4) *If $X \perp\!\!\!\perp Z \,|\, Y$ and $X \perp\!\!\!\perp W \,|\, (Y,Z)$ then $X \perp\!\!\!\perp (Z,W) \,|\, Y$ (contraction);*

It is occasionally important to abstract the notion of conditional independence away from necessarily being concerned with probability measures. An (abstract) *independence model* $\perp_\sigma$ over V is a ternary relation over subsets of a finite set V. The independence model is a *semi-graphoid* if the following holds for mutually disjoint subsets A, B, C, D:

(S1) *If $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$ (symmetry);*
(S2) *If $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ (decomposition);*
(S3) *If $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, (C \cup D)$ (weak union);*
(S4) *If $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, (B \cup C)$, then $A \perp_\sigma (B \cup D) \,|\, C$ (contraction);*

Further, the independence model is a *graphoid* if it also satisfies

(S5) *If $A \perp_\sigma B \,|\, (C \cup D)$ and $A \perp_\sigma C \,|\, (B \cup D)$ then $A \perp_\sigma (B \cup C) \,|\, D$ (intersection).*

We shall in particular be interested in distributions on product spaces $\mathscr{X} = \times_{v \in V} \mathscr{X}_v$ where V is a finite set. For $A \subseteq V$ we write $x_A = (x_v, v \in A)$ to denote a generic element in$\mathscr{X}_A = \times_{v \in A} \mathscr{X}_v$, and similarly $X_A = (X_v)_{v \in A}$.

If P is a probability distribution on $\mathscr{X}$, we can now define an independence model $\perp\!\!\!\perp$ by the relation

$$A \perp\!\!\!\perp B \,|\, C \iff X_A \perp\!\!\!\perp_P X_B \,|\, X_C$$

and it follows from Proposition 1 that $\perp\!\!\!\perp$ *is a semi-graphoid*; in general $\perp\!\!\!\perp$ is not a graphoid without further assumptions on P.

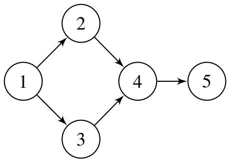
Another important independence model is determined by *separation* in an undirected graph. More precisely, if $\mathscr{G} = (V, E)$ is an undirected graph we can define an independence model $\perp_\mathscr{G}$ by letting $A \perp_\mathscr{G} B \,|\, S$ mean that all paths in $\mathscr{G}$ from A to B intersect S. Then it is easy to see that $\perp_\mathscr{G}$ is always a graphoid; indeed the term graphoid refers to this fact.

For a directed graph, the relevant notion of separation is more subtle. A vertex u is a *collider* on a path π if two arrowheads meet on the walk at u, i.e. if the following situation occurs $\pi = [\cdots \to u \leftarrow \cdots]$.

We say that a path π from u to v in a DAG $\mathscr{D}$ is *connecting* relative to S, if all colliders on π are in the ancestral set $\mathrm{An}(S)$, and all non-colliders are outside S. A path that is not connecting relative to S is said to be *blocked* by S. We then define an independence model $\perp_\mathscr{D}$ relative to a directed graph $\mathscr{D}$ as follows:

Definition 1. For three disjoint subsets A, B, and S of the vertex set V of a graph $\mathscr{G} = (V, E)$ we say that A and B are *$\mathscr{D}$-separated* by S if all paths from A to B are blocked by S and we then write $A \perp_\mathscr{D} B \,|\, S$. □

Example 2. Consider the network in the figure below, only slightly more complex than in Example 1:



We have $2 \perp_\mathscr{D} 3 \,|\, 1$ since the path $2 \leftarrow 1 \rightarrow 3$ is blocked as the non-collider 1 is in $S = \{1\}$ whereas the path $2 \rightarrow 4 \rightarrow 3$ is blocked because the collider 4 is not an ancestor of $S = \{1\}$; on the other hand it holds that $\neg(2 \perp_\mathscr{D} 3 \,|\, \{1,5\})$ since now the second path is rendered active as the collider 4 is in $\mathrm{An}(\{1,5\})$. □

Note that this definition in a natural way extends that of $\perp_{\mathscr{G}}$ for an undirected graph, as an undirected graph does not have colliders. The independence model $\perp_{\mathscr{D}}$ also satisfies the graphoid axioms, see e.g. Lauritzen and Sadeghi [22].

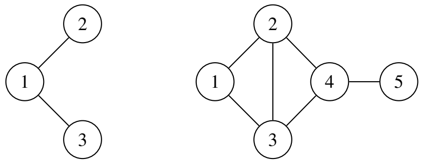
There is an alternative method for checking $\mathscr{D}$-separation in terms of standard separation in a suitable undirected graph, associated with the query. More precisely we say that A is *m-separated* from B by S and we write $A \perp_m B \,|\, S$ if S separates A from B in the moral graph $(\mathscr{D}_{\mathrm{An}(A\cup B\cup S)})^m$. We then have:

Proposition 2. *Let A, B and S be disjoint subsets of the nodes of a directed acyclic graph $\mathscr{G}$. Then $A \perp_{\mathscr{D}} B \,|\, S \iff A \perp_m B \,|\, S$.*

For a proof, see Richardson [29], amending an inaccuracy in Lauritzen et al. [24].

Example 3. To illustrate the alternative procedure, we again consider the network in Example 2.

If we wish to check whether $2 \perp_{\mathscr{D}} 3 \,|\, 1$ we consider the subgraph induced by the ancestral set of $\{1,2,3\}$ and moralize to obtain the graph to the left in the figure below. Since 1 is a separator in this graph, we conclude that $2 \perp_{\mathscr{D}} 3 \,|\, 1$.



On the other hand, if the query is whether $2 \perp_{\mathscr{D}} 3 \,|\, \{1,5\}$ we have $\mathrm{An}(\{1,5\}) = V$ and thus the relevant moral graph is given to the right in the figure above; in this graph, 2 and 3 are not separated by $\{1,5\}$ so we conclude $\neg(2 \perp_{\mathscr{D}} 3 \,|\, \{1,5\})$. □

2.4 Markov properties of Bayesian networks

It follows directly from the construction of a Bayesian network, that the joint distribution P satisfies the *well-ordered Markov property* (O) w.r.t. $\mathscr{D}$ if for some well-ordering of V, every variable is conditionally independent of its predecessors given its parents

$$v \perp\!\!\!\perp \mathrm{pr}(v) \,|\, \mathrm{pa}(v)$$

for all $v \in V = \{1,\dots,d\}$.

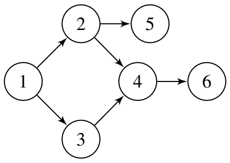
We further say that P obeys the *local Markov property* (L) w.r.t. $\mathscr{D}$ if every variable is conditionally independent of its non-descendants, given its parents:

$$v \perp\!\!\!\perp (\mathrm{nd}(v) \setminus \mathrm{pa}(v)) \,|\, \mathrm{pa}(v).$$

And, finally, P satisfies the *global Markov property* (G) w.r.t. $\mathscr{D}$ if

$$A \perp_{\mathscr{D}} B \,|\, C \implies A \perp\!\!\!\perp B \,|\, C.$$

Example 4. Consider the network in the figure below:



The numbering of the nodes here constitute a well-ordering so, for example, (O) implies $5 \perp\!\!\!\perp \{1,3,4\} \,|\, 2$, whereas the local Markov property (L) implies $5 \perp\!\!\!\perp \{1,3,4,6\} \,|\, 2$; the global Markov property implies, for example, $5 \perp\!\!\!\perp \{1,6\} \,|\, 4$. □

In the case of undirected graphs, the local and global Markov properties are different [23, Section 3.2], but here we have

Theorem 1. *Let $\mathscr{D}$ be a directed acyclic graph with $V = \{1,\dots,d\}$ well-ordered and P a probability distribution on $\mathscr{X} = \times_{v \in V} \mathscr{X}_v$. Then we have*

$$\text{(O)} \iff \text{(L)} \iff \text{(G)}.$$

In words, if P satisfies any of these Markov properties, it satisfies all of them.

Proof. This fact is established in [24, Corollary 2] for any semi-graphoid independence model $\perp_\sigma$. □

Note that in particular it is true that *if P satisfies (O) w.r.t. one well-ordering, it satisfies (O) w.r.t. all well-orderings.*

The global Markov property gives a sufficient condition for conditional independence in terms of $\mathscr{D}$-separation. Another central concept is that of faithfulness, formally defined below

Definition 2. A probability distribution P on $\mathscr{X} = \times_{v \in V} \mathscr{X}_v$ is said to be *faithful* to a DAG $\mathscr{D}$ if

$$A \perp_{\mathscr{D}} B \,|\, C \iff A \perp\!\!\!\perp_P B \,|\, C.$$

In other words, if $\mathscr{D}$-separation is also necessary for conditional independence. □

Generally, most probability distributions are faithful [25], but we shall later see that this is not the case for the special Bayesian networks we study here.

Finally, we need to emphasize that two different DAGs can define exactly the same independence model. Consider two graphs $\mathscr{D}_1$ and $\mathscr{D}_2$ as well as their associated independence models $\perp_{\mathscr{D}_1}$ and $\perp_{\mathscr{D}_2}$. It may well happen that even though the graphs are different, their independence models might be identical, see for example Figure 2 below.

Here all independence models are the same although the graphs are different. This also means that any probability distribution P which satisfies the global Markov property for any of them, automatically satisfies the global Markov property for all of them. We formally define

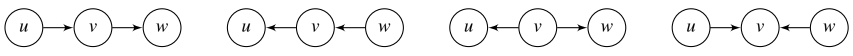


Fig. 2 The DAGs to the left of the figure are Markov equivalent; the only non-trivial element of their independence models is $u \perp_{\mathscr{D}} w \,|\, v$. The DAG to the right in the figure has a different independence model, since there $u \perp_{\mathscr{D}} w$.

Definition 3. Two DAGs $\mathscr{D}_1$ and $\mathscr{D}_2$ are *Markov equivalent* if and only if their independence models coincide, i.e. if $A \perp_{\mathscr{D}_1} B \,|\, C \iff A \perp_{\mathscr{D}_2} B \,|\, C$. □

The following result was shown by Frydenberg [14] and Verma and Pearl [36] and gives a necessary and sufficient condition for two DAGs to be Markov equivalent.

Theorem 2. *Two directed acyclic graphs $\mathscr{D}_1 = (V, E_1)$ and $\mathscr{D}_2 = (V, E_2)$ are Markov equivalent if and only if they have the same skeleton* $\mathrm{ske}(\mathscr{D}_1) = \mathrm{ske}(\mathscr{D}_2)$ *and the same unshielded colliders.*

Here the *skeleton* $\mathrm{ske}(\mathscr{D})$ of a DAG $\mathscr{D}$ is the undirected graph with $u \sim v$ in $\mathrm{ske}(\mathscr{D})$ if $u \sim v$ in $\mathscr{D}$, and an *unshielded collider* is a triple $u \to w \leftarrow v$ with $u \not\sim v$.

3 Recursive max-linear structural equation models

We shall be interested in Bayesian networks defined through structural equation systems (2) where the functions g_v are *max-linear,* i.e. the additions in (3) are replaced with the operation of forming the maximum.

Henceforth we assume that the vertex set of our DAG $\mathscr{D} = (V, E)$ is well-ordered so $V = \{1, \dots, d\}$ and assume a data generating mechanism specified via a *recursive max-linear structural equation model*, which has representation

$$X_v = \bigvee_{u \in \mathrm{pa}(v)} c_{vu} X_u \vee c_{vv} Z_v, \quad v = 1, \dots, d, \tag{4}$$

where $Z_1, \dots, Z_d$ are independent and identically distributed with a continuous distribution having support $\mathbb{R}_+ = (0, \infty)$, and $c_{vu} > 0$, $u \in \mathrm{pa}(v)$, c_{vv} are *structural coefficients* in the equations or *edge weights* for the associated DAG $\mathscr{D}$.

Following Gissibl and Klüppelberg [16] we say this is a *recursive max-linear model*. Note that our use of indices for edge weights here is the opposite of that used in [16].

For simplicity we assume throughout the rest of the paper that $c_{vv} = 1$ for all $v \in V$. For a path $\pi = [u = k_0 \to k_1 \to \cdots \to k_n = v]$ of length n from u to v, we define the quantities

$$d_{vu}(\pi) := \prod_{l=0}^{n-1} c_{k_{l+1} k_l} \quad \text{and} \quad b_{vu} := \bigvee_{\pi \in \Pi_{uv}} d_{vu}(\pi), \tag{5}$$

where Π_{uv} denotes all paths from u to v. In summary, we define

$$b_{vu} = \bigvee_{\pi \in \Pi_{uv}} d_{vu}(\pi) \text{ for } u \in \mathrm{an}(v);\ b_{vv} = c_{vv} = 1;\ b_{vu} = 0 \text{ for } u \in V \setminus \mathrm{An}(v), \tag{6}$$

where $\mathrm{An}(v) = \mathrm{an}(v) \cup \{v\}$ is the smallest ancestral set containing vertex v. We then arrange these coefficients in the *max-linear coefficient matrix* $B = (b_{vu})_{d\times d}$ and find

$$X_v = \bigvee_{u \in \mathrm{An}(v)} b_{vu} Z_u, \quad v = 1,\dots,d. \tag{7}$$

This equation represents X as a *max-linear model* as defined for instance in Wang and Stoev [37].

For two non-negative matrices F and G, where the number n of columns in F is equal to the number of rows in G we introduce the product $\odot$ as

$$(F \odot G)_{vu} = \Big(\bigvee_{k=1}^{n} f_{vk} g_{ku}\Big). \tag{8}$$

If we collect the noise variables into the column vector $Z = (Z_1,\dots,Z_d)'$, the representation (7) of X can then be written as

$$X = B \odot Z = \big(\bigvee_{u=1}^{d} b_{vu} Z_j, i = 1,\dots,d\big) = \big(\bigvee_{u \in \mathrm{An}(v)} b_{vu} Z_j, i = 1,\dots,d\big).$$

Given the DAG $\mathscr{D}$ and the edge weights c_{ik} with $c_{ii} = 1$ for all $i = 1,\dots,d$, the max-linear coefficient matrix B can be found by iterating the weighted adjacency matrix $C = (c_{vu} \mathbf{1}_{\mathrm{Pa}(v)}(u))_{d\times d}$ of $\mathscr{D}$ using this matrix multiplication; here $\mathbf{1}_{\mathrm{Pa}(v)}$ denotes the indicator function of $\mathrm{Pa}(v) = \mathrm{pa}(v) \cup \{v\})$:

$$B = \bigvee_{k=0}^{d-1} C^{\odot k} = C^{\odot (d-1)}, \tag{9}$$

cf. Butkovič [5], Lemma 1.4.1. For more details see [16], Theorem 2.4.

By (6) the max-linear coefficient b_{vu} of X is different from zero if and only if $u \in \mathrm{An}(v)$. This information is contained in the *reachability matrix* $R = (r_{vu})_{d\times d}$ of $\mathscr{D}$, which has entries

$$r_{vu} := \begin{cases} 1, & \text{if there is a path from } u \text{ to } v \text{, or if } u = v, \\ 0, & \text{otherwise.} \end{cases}$$

If the vu-th entry of R is equal to one, then *v is reachable from u*. In the context of a DAG $\mathscr{D}$ with its reachability matrix R and a recursive max-linear model X on $\mathscr{D}$ with max-linear coefficient matrix B it will be useful to keep the following in mind.

Remark 1. Let $\mathscr{D}$ be a DAG with reachability matrix R.

(i) The max-linear coefficient matrix B is a weighted reachability matrix of $\mathscr{D}$; i.e., $R = \mathrm{sgn}(B)$.

(ii) Since V is assumed well-ordered, B and R are lower triangular matrices. □

From (6) and (7) we conclude that a path π from u to v, whose weight $d_{vu}(\pi)$ is strictly less than b_{vu} does not have any influence on X_i. For $v \in V$ and $u \in \operatorname{an}(v)$ we call a path π from u to v *max-weighted*, if $b_{vu} = d_{vu}(\pi)$, and investigate its relevance for the recursive max-linear model in further detail.

Firstly we note that we can remove an edge from $\mathscr{D}$ which is not part of a max-weighted path without changing the distribution of X. The DAG obtained in this way is termed the *minimum max-linear* DAG $\mathscr{D}^B$. In the special case where $\mathscr{D}$ is a polytree, all paths are necessarily max-weighted and we clearly have

Proposition 3. *If $\mathscr{D}$ is a polytree, it holds that $\mathscr{D}^B = \mathscr{D}$.*

The following result describes exactly all DAGs and edge weights possible for a given max-linear coefficient matrix. Recall that we have set $c_{vv} = 1$.

Theorem 3. *[Gissibl and Klüppelberg [16], Theorem 5.4]*
Let X be given by a recursive max-linear structural equation system with coefficient matrix B. Let further $\mathscr{D}^B = (V, E^B)$ be the minimum max-linear DAG *of X and* $\operatorname{pa}^B(v)$ *the parents of v in $\mathscr{D}^B$.*

(a) $\mathscr{D}^B$ is the DAG *with the minimum number of edges such that X satisfies* (4)*. The weights in* (4) *are uniquely given by $c_{vv} = b_{vv}$ and $c_{vs} = b_{vs}$ for $v \in V$ and $s \in \operatorname{pa}^B(v)$.*

(b) Every DAG *with vertex set V that has at least the edges of $\mathscr{D}^B$ and the same reachability matrix as $\mathscr{D}^B$ represents X in the sense of* (4) *with weights satisfying*

$$c_{vv} = b_{vv},\, c_{vs} = b_{vs} \textit{ for } s \in \operatorname{pa}^B(v), \textit{ and } c_{vs} \in (0, b_{vs}) \textit{ for } s \in \operatorname{pa}(v) \setminus \operatorname{pa}^B(v).$$

There are no further DAG*s and weights such that X has representation* (4)*.*

In general, recursive max-linear models are not faithful to their DAG, not even if $\mathscr{D} = \mathscr{D}^B$, see Remark 3.9 (ii) in [16]. This is illustrated in Example 5 below.

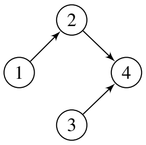
Example 5. [Example 3.8 of [16] and continuation of Example 1:] We note that the paths $[1 \to 2]$, $[1 \to 3]$, $[2 \to 4]$, and $[3 \to 4]$ are max-weighted as they are the only directed paths between their endpoints. It therefore holds that $\mathscr{D}^B = \mathscr{D}$ since they are the unique max-weighted paths. Still, the distribution determined by this recursive system is never faithful to $\mathscr{D}$, as we shall see below.

Concerning the paths from node 1 to 4 we have three situations:

$$c_{42}c_{21} = c_{43}c_{31}, \quad c_{42}c_{21} > c_{43}c_{31}, \quad \text{or} \quad c_{42}c_{21} < c_{43}c_{31}.$$

In the first situation, both paths from 1 to 4, $[1 \to 2 \to 4]$ and $[1 \to 3 \to 4]$, are max-weighted whereas in the other situations only one of them is.

If the path $[1 \to 2 \to 4]$ is max-weighted, we can consider the subdag $\tilde{\mathscr{D}}$ obtained from $\mathscr{D}$ by removing the edge $1 \to 3$:



In other words, we are changing the edge weights by letting $\tilde{c}_{31} = 0$, keeping the other edge weights unchanged. The new max-linear coefficient matrix becomes

$$\tilde{B} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ c_{21} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ c_{42}c_{21} & c_{42} & c_{43} & 1 \end{pmatrix}$$

where we have exploited that $c_{ii} = 1$. The max-linear coefficient matrix for the marginal distribution of (X_1, X_2, X_4) is obtained by ignoring the third row and since only entries in the third row have changed, we see that (X_1, X_2, X_4) has the same joint distribution in the model determined by $\mathscr{D}$ as it has in the model determined by $\tilde{\mathscr{D}}$.

But as we clearly have $1 \perp_{\tilde{\mathscr{D}}} 4 \mid 2$, we conclude that $X_1 \perp\!\!\!\perp X_4 \mid X_2$ in the model determined by $\tilde{\mathscr{D}}$ and hence also by $\mathscr{D}$. But since $\neg(1 \perp_{\mathscr{D}} 4 \mid 2)$, the distribution is not faithful to $\mathscr{D}$.

If $[1 \to 3 \to 4]$ is also max-weighted, the similar argument yields $X_1 \perp\!\!\!\perp X_4 \mid X_3$, so the distribution is *not faithful to* $\mathscr{D}$ for any allocation of edge weights. □

We note that [16] suggest in their Remark 3.9(i) that additional conditional independence relations that are valid for a given DAG can be revealed by considering a system of submodels determined by appropriate subgraphs, but here we refrain from giving a complete description of all valid conditional independence relations.

4 Statistical properties

The statistical theory of recursive max-linear models is challenging because standard assumptions for smooth statistical models are not satisfied. For example, if we for a given DAG $\mathscr{D}$ consider the family $\mathscr{P}$ of distributions with coefficients adapted to $\mathscr{D}$, this family is not dominated by any measure on the space of observations, so standard likelihood theory does not apply. On the other hand, as we shall see, estimation of coefficients and identification of the network structure for recursive max-linear models can be made in a simple fashion and procedures are more efficient than usual in that estimates of coefficients and structures converge at exponential rates to the true values. Here we shall give a summary of the most important findings in Gissibl et al. [17].

Throughout the following we consider a sample $\mathbf{x} = (X^1 = x^1, \ldots, X^n = x^n)$ from a distribution P given by the recursive max-linear model (4).

4.1 Estimation of coefficients

We first consider the situation where the DAG $\mathscr{D} = (V, E)$ and for the sake of simplicity we assume the distribution of noise variables $Z_v, v \in V$, is completely known, the coefficients c_{vv} are all equal to one, whereas the edge weights $C = \{c_{vu}, u \in \text{pa}(v), v \in V\}$ are all strictly positive, but otherwise unknown. We let $\mathscr{C}$ denote the set of all possible coefficients and P_C denote the distribution of X determined by the corresponding recursive model (4).

The family $\mathscr{P} = P_C, C \in \mathscr{C}$, is not dominated by any fixed σ-finite measure μ on $\mathscr{X}$, as the support of P_C varies strongly with the coefficients; more precisely, the distributions have disjoint atomic components. This is a disadvantage in the sense that we cannot define a standard likelihood function; but, as we shall see, an advantage since these atomic components help identifying P_C from a given sample. We illustrate this by a simple example.

Example 6. [Estimation from the atoms]
Consider the simple DAG $1 \to 2$ with just two nodes and a single directed edge, and let $c = c_{21}$ be the corresponding coefficient.

Then P_c has support on the cone given as $x_2 \geq cx_1 \geq 0$ and the line $A_c = \{x_2 = cx_1\}$ is an atom for P_c because $P_c(A_c) = P(Z_2 \leq cX_1) = P(Z_2 \leq cZ_1) > 0$ (cf. Remark **??**(ii)).

Still, since then $\{c\}$ is the only atom in P_c for $Y = X_2/X_1$, the sample will for large n with high probability have repeated values of Y and c will be the only value that is repeated. In other words, $\hat{c} = \min\{y^\nu = x_2^\nu/x_1^\nu, \nu = 1, \dots, n\}$ will be exactly equal to the true parameter with high probability. A similar estimator has been considered by Davis and Resnick [7] in a time-series framework. □

Although most likelihood theory is concerned with dominated families, Kiefer and Wolfowitz [21] considered the non-dominated case. Their formulation has been used rarely — an exception being Johansen [20]; see also Scholz [30] and Gill et al. [15], for example. This formulation turns out to be exactly what we need to discuss estimation of C in a formal way.

For two probability measures P and Q on a measurable space $(\mathscr{X}, \mathbb{E})$, we define the *generalized likelihood ratio* $\rho_x(P, Q)$ at the observation x as

$$\rho_x(P,Q) = \frac{\mathrm{d}P}{\mathrm{d}(P+Q)}(x) \tag{10}$$

where $\mathrm{d}P/\mathrm{d}(P+Q)$ is the density of P w.r.t. $P+Q$; the density always exists as, clearly, $P(A)+Q(A) = 0 \implies P(A) = 0$ so P is absolutely continuous w.r.t. $P+Q$.

The idea here is that if $\rho_x(P,Q) > \rho_x(Q,P)$, then P is a more likely explanation of x than Q. We note in particular that if P and Q have densities f and g w.r.t. a σ-finite measure μ, we have $\rho_x(P,Q) = f(x)/\{f(x)+g(x)\}$ so then $\rho_x(P,Q) > \rho_x(Q,P)$ if and only if $f(x) > g(x)$. Hence ρ_x extends the standard likelihood ratio in a natural way.

Clearly, the generalized likelihood ratio suffers from the same problem as the usual likelihood ratio: the densities are only defined almost surely, so can be changed on $P+Q$-null sets; therefore, a version of $\mathrm{d}P/\mathrm{d}(P+Q)$ must be chosen independently of the observation x.

Next we say that if $\mathscr{P}$ is a family of probability distributions, $\hat{P}$ is a *generalized maximum likelihood estimate* (GMLE) of P based on $x \in \operatorname{supp}(\hat{P})$ if

$$\rho_x(\hat{P},Q) \geq \rho_x(Q,\hat{P}) \text{ for all } Q \in \mathscr{P},$$

i.e. if $\hat{P}$ explains x at least as well as any other member of $\mathscr{P}$.

Example 7. [Continuation of Example 6: GMLE]
We illustrate use of the generalized maximum likelihood ratio for the model described in Example 6. To identify the density, we consider two values $c > c^*$ where we have

$$\rho_x(c,c^*) = \frac{\mathrm{d}P_c}{\mathrm{d}(P_c+P_{c^*})}(x_1,x_2) = \begin{cases} 1/2 & \text{for } x_2 > cx_1 \\ 1 & \text{for } x_2 = cx_1 \\ 0 & \text{for } x_2 < cx_1 \end{cases}$$

and

$$\rho_x(c^*,c) = \frac{\mathrm{d}P_{c^*}}{\mathrm{d}(P_c+P_{c^*})}(x_1,x_2) = \begin{cases} 1/2 & \text{for } x_2 > cx_1 \\ 0 & \text{for } x_2 = cx_1 \\ 1 & \text{for } cx_1 > x_2 \geq c^*x_1 \\ 0 & \text{for } x_2 < c^*x_1. \end{cases}$$

If $c = c^*$ we may let

$$\rho_x(c,c) = \rho_x(c,c^*) = \rho_x(c^*,c) = \frac{1}{2}\mathbf{1}_{\{x_2 \geq cx_1\}}.$$

Thus, if we consider a full sample, let $\hat{c} = \min\{y^\nu = x_2^\nu/x_1^\nu, \nu = 1,\ldots,n\}$ and $n_+(c,\mathbf{x}) = \#\{\nu : y^\nu > c\}$, we get:

$$\rho_{\mathbf{x}}(\hat{c},c) = \prod_{\nu=1}^n \rho_{x^\nu}(\hat{c},c) = \begin{cases} 0 & \text{if } c > \hat{c} \text{ and } c \in \{y^\nu, \nu = 1,\ldots,n\} \\ 2^{-n_+(c,\mathbf{x})} & \text{if } c > \hat{c} \text{ and } c \notin \{y^\nu, \nu = 1,\ldots,n\} \\ 2^{-n} & \text{if } c = \hat{c} \\ 2^{-n_+(\hat{c},\mathbf{x})} & \text{if } c < \hat{c}, \end{cases}$$

whereas

$$\rho_{\mathbf{x}}(c,\hat{c}) = \prod_{\nu=1}^n \rho_{x^\nu}(c,\hat{c}) = \begin{cases} 0 & \text{if } c > \hat{c} \\ 2^{-n} & \text{if } c = \hat{c} \\ 0 & \text{if } c < \hat{c}. \end{cases}$$

Clearly, $\rho_{\mathbf{x}}(\hat{c},c) \geq \rho_{\mathbf{x}}(c,\hat{c})$ showing that $\hat{c}$ is the unique GMLE of c. □

Indeed, *it holds in general for a recursive max-linear model that*

$$\hat{c}_{ij} = \bigwedge_{\nu=1}^{n} \frac{x_i^\nu}{x_j^\nu}, \quad i \in V, j \in \mathrm{pa}(i)$$

is a GMLE of the edge weights. We refer to [17] for further details but should point out that in the general case, the GMLE is not unique. Since the distribution of X only depends on the edge weights through the max-linear coefficient matrix B, only B is uniquely estimable from a sample. We clearly have by (9) for the GMLE that

$$\hat{B} = B(\hat{C}) = \bigvee_{k=0}^{d-1} \hat{C}^{\odot k} = \hat{C}^{\odot(d-1)}.$$

An alternative estimate of the max-linear coefficient matrix is given as

$$\tilde{b}_{ij} = \bigwedge_{\nu=1}^{n} \frac{x_i^\nu}{x_j^\nu}, \quad i \in V, j \in \mathrm{an}(i).$$

Although this estimate is also sensible and asymptotically consistent, it is less efficient than the GMLE as X_i^ν/X_j^ν only attends its minimum value when all noise variables on the path from j to i are smaller than $b_{ij}X_j^\nu$ for the same ν, whereas the minima for the X_v^ν/X_u^ν on the path from j to i can be attained for different νs.

4.2 Identification of structure

General methods for identifying the structure of DAG $\mathscr{D}$ from a sample are often based on an assumption of faithfulness, so that observed conditional independence relations can be translated back to the structure of the DAG since then any observed conditional independence must correspond to a separation in $\mathscr{D}$, see for example Spirtes et al. [33]. Also, as noted in Theorem 2, two DAGs can be different but still Markov equivalent and thus any method based on observed direct conditional independence relations cannot distinguish between DAGs that are Markov equivalent.

As shown in Example 5, faithfulness is violated for max-linear Bayesian networks whenever $\mathscr{D}$ is not a polytree. However, as we shall see below, the minimal DAG $\mathscr{D}^B$ of a max-linear Bayesian network can still be completely recovered from observations.

This fact conforms with recent developments where the recursive linear structural equation systems have been shown to be completely identifiable if the errors follow a non-Gaussian distribution (Shimizu et al. [31]) and it has been shown that the faithfulness assumption can be considerably weakened also in other situations (Spirtes and Zhang [32], Peters and Bühlmann [26]).

To explain why the structure $\mathscr{D}^B$ is identifiable, we consider the statistics

$$Y_{ij} = X_i/X_j, \quad i, j \in V$$

and note that Y_{ij} has support $[b_{ij},\infty)$ and an atom in b_{ij} if and only if $j \in \operatorname{an}(i)$. Using this property one can show that the following estimate $\check{B}$ eventually identifies the max-linear coefficient matrix B.

$$\check{b}_{ij} = \begin{cases} \bigwedge_{\nu=1}^{n} y_{ij}^{\nu} & \text{if minimum value is attained at least twice in the sample,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathscr{D}^B$ is identifiable from B; we refer the reader to [17] for further details.

5 Conclusion

We have reviewed basic elements of Bayesian networks based on recursive max-linear structural equations and some of their statistical properties. We conclude this article by pointing out some natural extensions of this work that we hope to address in the future.

Firstly, it would be of interest to have a simple and complete description of all independence properties which hold for a distribution determined by a recursive max-linear equation system, i.e. a global Markov property for max-linear Bayesian networks.

Secondly, it appears that a consequent use of algebraic theory; see e.g. Butkovič [5], based on properties of the max-times semiring $\mathbb{S} = ([0,\infty], \vee, \cdot)$ would be able to simplify the theory of these models.

Finally, we should emphasize that the models heuristically can be seen as limiting cases of standard linear recursive models where error distributions have heavy tails and therefore the maximal element of any sum will almost completely dominate the sum; a rigorous study of this limiting process will enhance the understanding of this class of models.

Acknowledgements

The authors have benefited from discussions with Nadine Gissibl and financial support from the Alexander von Humboldt Stiftung.

References

[1] P. Asadi, A. C. Davison, and S. Engelke. Extremes on river networks. *Ann. Appl. Stat.*, 9(4):2023–2050, 12 2015.

[2] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications.* Wiley, Chichester, 2004.

[3] K. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

[4] S. Buhl, R. Davis, C. Klüppelberg, and C. Steinkohl. Semiparametric estimation for isotropic max-stable space-time processes. arXiv 1609.04967, 2016.

[5] P. Butkovič. *Max-linear Systems: Theory and Algorithms*. Springer, London, 2010.

[6] R. Davis, C. Klüppelberg, and C. Steinkohl. Statistical inference for max-stable processes in space and time. *J. Roy. Statist. Soc. Ser. B*, 75(5):791–819, 2013.

[7] R. A. Davis and S. I. Resnick. Basic properties and prediction of max-ARMA processes. *Advances in Applied Probability*, 21(4):781–803, 1989.

[8] A. Davison, S. Padoan, and M. Ribatet. Statistical modelling of spatial extremes. *Statistical Science*, 27(2):161–186, 2012.

[9] A. P. Dawid. Conditional independence for statistical operations. *Ann. Statist.*, 8:598–617, 1980.

[10] L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

[11] J. Einmahl, A. Kiriliouk, and J. Segers. A continuous updating weighted least squares estimator of tail dependence in high dimensions. To appear, 2017.

[12] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

[13] B. Finkenstädt and H. Rootzén. *Extreme Values in Finance, Telecommunication and the Environment*. Chapman and Hall/CRC, Boca Raton, 2004.

[14] M. Frydenberg. The chain graph Markov property. *Scand. J. Statist.*, 17:333–353, 1990.

[15] R. D. Gill, J. A. Wellner, and J. Præstgaard. Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scand. J. Statist.*, 16:97–128, 1989.

[16] N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. arXiv 1512.07522. *Bernoulli*, to appear., 2015.

[17] N. Gissibl, C. Klüppelberg, and S. Lauritzen. Estimation of recursive max-linear models. In preparation.

[18] N. Gissibl, C. Klüppelberg, and J. Mager. Big data: progress in automating extreme risk analysis. In W. Pietsch, J. Wernecke, and M. Ott, editors, *Berechenbarkeit der Welt?* Springer, Wiesbaden, 2017.

[19] R. Huser and A. Davison. Space-time modelling of extreme events. *J. Roy. Statist. Soc. Ser. B*, 76(2):439–461, 2014.

[20] S. Johansen. The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, 5(4):195–199, 1978.

[21] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:887–906, 1956.

[22] S. Lauritzen and K. Sadeghi. Unifying Markov properties for graphical models. arXiv:1608.05810, 2017. To appear in *The Annals of Statistics*.

[23] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

[24] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[25] C. Meek. Strong completeness and faithfulness in Bayesian networks. In P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–418. Morgan Kaufman Publishers, San Francisco, CA, August 1995.

[26] J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.

[27] S. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer, New York, 1987.

[28] S. Resnick. *Heavy-Tail Phenomena, Probabilistic and Statistical Modeling*. Springer, New York, 2007.

[29] T. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 30(1):145–157, 2003.

[30] F. W. Scholz. Towards a unified definition of maximum likelihood. *The Canadian Journal of Statistics*, 8:193–203, 1980.

[31] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Machine Learning Research*, 7:2003–2030, 2006.

[32] P. Spirtes and J. Zhang. A uniformly consistent estimator of causal effects under the *k*-triangle faithfulness assumption. *Statistical Science*, 29:662–678, 2014.

[33] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, New York, 2 edition, 2000.

[34] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

[35] J. Ver Hoef, E. Peterson, and D. Theobald. Spatial statistical methods that use flow and stream distance. *Environmental and Ecological Statistics*, 13(4): 449–464, 2006.

[36] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 255–270, Amsterdam, 1990.

[37] Y. Wang and S. A. Stoev. Conditional sampling for spectrally discrete max-stable random fields. *Advances in Applied Probability*, 43(2):461–483, 2011.

[38] D. B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ, USA, 2001.