Identifiability and estimation of recursive max-linear models

Nadine Gissibl* Claudia Klüppelberg* Steffen Lauritzen†

January 10, 2019

Abstract

We address the identifiablity and estimation of recursive max-linear structural equation models represented by an edge weighted directed acyclic graph (DAG). Such models are generally unidentifiable and we identify the whole class of DAGs and edge weights corresponding to a given observational distribution. For estimation, standard likelihood theory cannot be applied because the corresponding families of distributions are not dominated. Given the underlying DAG, we present an estimator for the class of edge weights and show that it can be considered a generalized maximum likelihood estimator. In addition, we develop a simple method for identifying the structures of the DAGs. With probability tending to one at an exponential rate with the number of observations, this method correctly identifies the class of DAGs and, similarly, exactly identifies the possible edge weights.

MSC 2010 subject classifications: Primary 60E15, 62H12; secondary 62G05, 60G70, 62-09

Keywords and phrases: Causal inference, Bayesian network, directed acyclic graph, extreme value theory, generalized maximum likelihood estimation, graphical model, identifiability, max-linear model, structural equation model.

1 Introduction

Establishing and understanding cause-effect relations is an omnipresent desire in science and daily life. It is especially important when dealing with extreme events, because they are mostly dangerous and very costly; knowing and understanding the causes of such events and their causal relations could help us to deal better with them. Examples include incidents at airplane landings (Gissibl et al. [13]), flooding in river networks (Asadi et al. [1]), financial risk (Einmahl et al. [8]), and chemical pollution of rivers (Hoef et al. [15]). Such applications, where extreme risks may propagate through a network, have been the motivation behind the definition of *recursive max-linear* (ML) models in Gissibl and Klüppelberg [12]. Recursive ML models are *structural equation models* (SEMs) represented by a *directed acyclic graph* (DAG) and thereby obey the basic Markov properties associated with directed graphical models (Lauritzen [21], Lauritzen

*Center for Mathematical Sciences, Technische Universität München, 85748 Garching, Boltzmannstrasse 3, Germany; e-mail: n.gissibl@tum.de, cklu@tum.de

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

et al. [22]). Both SEMs (see for example Bollen [3], Pearl [23]) and directed graphical models (see for example Koller and Friedman [19], Lauritzen [20], Spirtes et al. [27]) are well-established concepts for the understanding and quantification of causal inference from observational data. We note that Hitz and Evans [14] and Engelke and Hitz [9] discuss graphical models for extremes that are based un undirected graphs.

Recursive ML models are defined by a DAG, a collection of edge weights, and a vector of independent innovations. Important research problems that are addressed for recursive SEMs are the question of *identifiability* of the coefficients and the associated DAG from the observational distribution. Although the true DAG and edge weights for a recursive ML model are not identifiable from the observational distribution, the so-called *max-linear coefficient matrix* is identifiable and determines the possible class of DAGs and edge weights uniquely.

We shall show that estimation and structure learning of recursive ML models can be done in a simple and efficient fashion by exploiting properties of the ratios between observable components of the model. For a sufficiently large number of observations, these ratios identify the true ML coefficient matrix with a probability that converges exponentially fast to 1. For the situation where the DAG is known, we show that our estimator can be considered a maximum likelihood estimator in an extended sense, originally introduced by Kiefer and Wolfowitz [18].

Our paper is organized as follows. In Section 2 we introduce the model class of recursive ML models and the notation used throughout. In Section 3 we discuss the identifiability of a recursive ML model from its observational distribution. Here we show distributional properties of the ratio between two components. Based on these properties, we suggest an identification method. Section 4 is then devoted to the estimation of recursive ML models where we assume the DAG to be known. Among other outstanding properties, we show that the proposed estimates are *generalized maximum likelihood estimates (GMLEs)* in the sense of Kiefer–Wolfowitz. The main part is here the derivation of a specific Radon-Nikodym derivative. In Section 5 we complement the theoretical findings on the identifiability of recursive ML models with an efficient procedure to learn recursive ML models from observations only, even when the DAG itself is also unknown. Section 6 concludes and suggests further directions of research.

2 Preliminaries — recursive max-linear models {ch4:s2}

In this section we introduce notation and summarize the most important properties of recursive ML models needed. A recursive ML model for a random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ is specified by an underlying structure in terms of a DAG $\mathcal{D}$ with nodes $V = \{1, \dots, d\}$, positive *edge weights* c_{ki} for $i \in V$ and $k \in \mathrm{pa}(i)$, and independent positive random variables $Z_1, \dots, Z_d$ with support $\mathbb{R}_+ := (0, \infty)$ and atom-free distributions:

$$X_i = \bigvee_{k \in \mathrm{pa}(i)} c_{ki} X_k \vee Z_i, \quad i = 1, \dots, d, \tag{2.1}$$ {ch4:ml-sem}

where $\mathrm{pa}(i)$ are the parents of node i in $\mathcal{D}$. To highlight the DAG $\mathcal{D}$, we say that $\boldsymbol{X}$ follows a *recursive ML model* on $\mathcal{D}$. Note that this is a slight variation of the original definition in [12]. We shall refer to $\boldsymbol{Z} = (Z_1, \dots, Z_d)$ as the vector of *innovations*.

In the context of risk analysis, natural candidates for distributions of the innovations are extreme value distributions or distributions in their domain of attraction, resulting in a corresponding multivariate distribution (for details and background on multivariate extreme value models, see for example Beirlant et al. [2], de Haan and Ferreira [7], Resnick [24, 25]).

Throughout the paper we use the following notation. The sets $\mathrm{an}(i)$, $\mathrm{pa}(i)$, and $\mathrm{de}(i)$ contain the ancestors, parents, and descendants of node i in $\mathcal{D}$. We set $\mathrm{An}(i) = \mathrm{an}(i) \cup \{i\}$ and $\mathrm{Pa}(i) = \mathrm{pa}(i) \cup \{i\}$. For $U \subsetneq V$ we write $\boldsymbol{X}_U = (X_\ell, \ell \in U)$ and accordingly for $\boldsymbol{x} \in \mathbb{R}_+^d$, $\boldsymbol{x}_U = (x_\ell, \ell \in U)$.

Instead of $k \in \mathrm{pa}(i)$ we also write $k \to i$. Assigning the weight $d_{ji}(p) = \prod_{\nu=0}^{n-1} c_{k_\nu k_{\nu+1}}$ to every path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$ and denoting the set of all paths from j to i by P_{ji}, the non-negative matrix $B = (b_{ij})_{d \times d}$ with entries

$$b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \quad \text{for } j \in \mathrm{an}(i), \quad b_{ii} = 1, \quad \text{and} \quad b_{ji} = 0 \quad \text{for } j \in V \backslash \mathrm{An}(i), \tag{2.2}$$

{ch4:coeff}

is said to be the *ML coefficient matrix of* $\boldsymbol{X}$. This means for distinct $i, j \in V$, b_{ji} is positive if and only if there is a path from j to i; in that case b_{ji} is the maximum weight of all paths from j to i, where the weight of a path is the product of all edge weights c_{ki} along this path. We say that a path from j to i whose weight equals b_{ji} is *max-weighted*.

The components of $\boldsymbol{X}$ can also be expressed as max-linear functions of their ancestral innovations and an independent one; the corresponding *ML coefficients* are the entries of B:

$$X_i = \bigvee_{j=1}^{d} b_{ji} Z_j = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j, \quad i = 1, \ldots, d; \tag{2.3}$$

{ch4:ml-noise}

see Theorem 2.2 of [12].

For two non-negative matrices F and G, where the number of columns in F is equal to the number of rows in G, we define the matrix product $\odot : \overline{\mathbb{R}}_+^{m \times n} \times \overline{\mathbb{R}}_+^{n \times p} \to \overline{\mathbb{R}}_+^{m \times p}$ by

$$(F = (f_{ij})_{m \times n}, G = (g_{ij})_{n \times p}) \mapsto F \odot G := \Big(\bigvee_{k=1}^{n} f_{ik} g_{kj} \Big)_{m \times p}, \tag{2.4}$$

{ch2:odot}

where $\overline{\mathbb{R}}_+ = [0, \infty)$. The triple $(\overline{\mathbb{R}}_+, \vee, \cdot)$, is an idempotent semiring with 0 as 0-element and 1 as 1-element and the operation $\odot$ is therefore a matrix product over this semiring; see for example Butkovič [4]. Denoting by $\mathcal{M}$ all $d \times d$ matrices with non-negative entries and by $\vee$ the componentwise maximum between two matrices, $(\mathcal{M}, \vee, \odot)$ is also a semiring with the null matrix as 0-element and the $d \times d$ identity matrix I_d as 1-element.

The matrix product $\odot$ allows us to represent the ML coefficient matrix B of $\boldsymbol{X}$ in terms of the weighted adjacency matrix $(c_{ij} \mathbb{1}_{\mathrm{pa}(j)}(i))_{d \times d}$ of $\mathcal{D}$ since (2.2) and (2.3) simply become

$$B = (I_d \vee C)^{\odot(d-1)} = \bigvee_{k=0}^{d-1} C^{\odot k}, \qquad \boldsymbol{X} = \boldsymbol{Z} \odot B, \tag{2.5}$$

{eq:dotrepr}

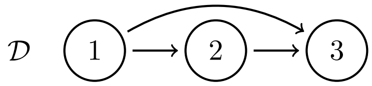
where we have let $A^{\odot 0} = I_d$ and $A^{\odot k} = A^{\odot(k-1)} \odot A$ for $A \in \overline{\mathbb{R}}_+^{d \times d}$ and $k \in \mathbb{N}$; see Proposition 1.6.15 of Butkovič [4] as well as Theorem 2.4 and Corollary 2.5 of [12].

3 Identifiability of a recursive max-linear model

{ch4:s4}

In this section we discuss the question of identifiability of the elements of a recursive ML model from the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$. We begin with an example.

Example 3.1. [The DAG and the edge weights are not necessarily identifiable] {ch4:exprob2}
Consider a recursive ML model on the DAG $\mathcal{D}$ depicted below with edge weights c_{12}, c_{23}, c_{13}.



According to (2.1), the components of $\boldsymbol{X}$ have the following representations

$$X_1 = Z_1, \quad X_2 = c_{12}X_1 \vee Z_2, \quad \text{and} \quad X_3 = c_{13}X_1 \vee c_{23}X_2 \vee Z_3.$$

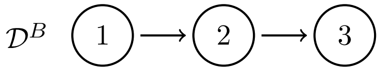
but also representations in terms of the innovations using (2.3) as

$$X_1 = Z_1, \quad X_2 = c_{12}Z_1 \vee Z_2, \quad \text{and} \quad X_3 = (c_{12}c_{23} \vee c_{13})Z_1 \vee c_{23}Z_2 \vee Z_3,$$

If $c_{13} \leqslant c_{12}c_{23}$ we have for any $c_{13}^* \in [0, c_{12}c_{23}]$ that $b_{13} = c_{12}c_{23} \vee c_{13}^* = c_{12}c_{23} \vee c_{13} = c_{12}c_{23}$; so we could also write

$$X_3 = c_{13}^* X_1 \vee c_{23}X_2 \vee Z_3$$

without changing the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$. This implies that if $c_{13} \leqslant c_{12}c_{23}$, $\boldsymbol{X}$ follows a recursive ML model on $\mathcal{D}$ with edge weights c_{12}, c_{23}, c_{13}^* but also on the DAG $\mathcal{D}^B$ depicted below with edge weights c_{12}, c_{23}. Consequently, we can neither identify $\mathcal{D}$ nor the value c_{13} from the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$. However, note that the ML coefficient $b_{13} = c_{12}c_{23} \vee c_{13}$ is uniquely determined.



If we however assume that $c_{13} > c_{12}c_{23}$, only $\mathcal{D}$ and the edge weights c_{12}, c_{23}, c_{13} represent $\boldsymbol{X}$ in the sense of (2.1). Thus in this case the DAG and the edge weights are identifiable from the distribution $\mathcal{L}(\boldsymbol{X})$. □

As conclusion of Example 3.1, it is generally not possible to identify the true DAG $\mathcal{D}$ and the edge weights c_{ki} underlying $\boldsymbol{X}$ in representation (2.1) from $\mathcal{L}(\boldsymbol{X})$, since several DAGs and edge weights may exist such that $\boldsymbol{X}$ has this representation. The smallest DAG of this kind is the DAG that has an edge $k \to i$ if and only if $k \to i$ is the only max-weighted path from k to i. We call this DAG $\mathcal{D}^B$ the *minimum ML DAG of* $\boldsymbol{X}$ and note that this is uniquely determined from the ML coefficient matrix B. All other DAGs representing $\boldsymbol{X}$ are those that include the edges of $\mathcal{D}^B$ and whose nodes have the same ancestors. The edge weights c_{ki} in the representation (2.1) of $\boldsymbol{X}$ are only uniquely determined for edges contained in $\mathcal{D}^B$; namely, by b_{ki}; otherwise, c_{ki} may be any number in $(0, b_{ki}]$. All this information can be found in Section 5 of [12] with its main results in Theorems 5.3, 5.4.

Based on the above observations, we investigate the identifiability of the whole class of DAGs and edge weights representing the max-linear structural equations (2.1) of $\boldsymbol{X}$ from $\mathcal{L}(\boldsymbol{X})$. Since this class can be recovered from B, it suffices to clarify whether B is identifiable from $\mathcal{L}(\boldsymbol{X})$. There are many ways to prove that this is indeed the case. The way we present in this section suggests a simple procedure to estimate B from independent realizations of $\boldsymbol{X}$ (see Algorithm 5.1 below). An alternative way can be found in Appendix 4.A.1 of [11].

The ratios $\boldsymbol{Y} = \{Y_{ij} = X_j/X_i,\ i, j = 1 \ldots, d\}$ between all pairs of components of $\mathbf{X}$ are the essential quantities used to identify B from $\mathcal{L}(\boldsymbol{X})$. We first present distributional properties of these ratios, where we let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space of $\boldsymbol{Z}$ and, hence, of $\boldsymbol{X}$. In what follows, we use the standard convention and write events such as $\{\omega \in \Omega : X_i(\omega) < X_j(\omega)\}$ as

$\{X_i < X_j\}$, etc. Unsurprisingly, because of the max-linear representation (2.3) of the components of $\boldsymbol{X}$, the ratios inherit their distributional properties from the innovations. It plays an important role that

$$\text{the event } \big\{Z_i = xZ_j\big\} \text{ for distinct } i, j \in V \text{ and } x \in \mathbb{R}_+ \text{ has probability zero,} \tag{3.1}$$ {ch4:noiseNull}

which follows from the independence of the innovations and the fact that their distributions are atom-free.

{lem:XioverXj}

Lemma 3.2. *Let $i, j \in V$ be distinct.*

(a) The ratio $Y_{ji} = X_i/X_j$ has an atom in $x \in \mathbb{R}_+$ if and only if $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \varnothing$ and $x = b_{\ell i}/b_{\ell j}$ for some $\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)$.

(b) We have

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

where $\mathrm{supp}(Y_{ji})$ denotes the support of Y_{ji}.

Proof. To establish (a) note that (2.3) and (3.1) imply that the sets $\{X_i = xX_j\} = \big\{\bigvee_{\ell\in\mathrm{An}(i)} b_{\ell i}Z_\ell = \bigvee_{\ell\in\mathrm{An}(j)} xb_{\ell j}Z_\ell\big\}$ and

$$\Big\{\bigvee_{\substack{\ell\in\mathrm{An}(i)\cap\mathrm{An}(j):\\ b_{\ell i}=b_{\ell j}x}} b_{\ell i}Z_\ell > \bigvee_{\substack{\ell\in\mathrm{An}(i)\cap\mathrm{An}(j):\\ b_{\ell i}\neq b_{\ell j}x}} (b_{\ell i} \vee xb_{\ell j})Z_\ell \vee \bigvee_{\ell\in\mathrm{An}(i)\setminus\mathrm{An}(j)} b_{\ell i}Z_\ell \vee \bigvee_{\ell\in\mathrm{An}(j)\setminus\mathrm{An}(i)} xb_{\ell j}Z_\ell\Big\}$$

differ only by a set of probability zero. Since the innovations are independent and have support $\mathbb{R}_+$ the conclusion follows.

To establish (b) note that the support $\mathbb{R}_+$ of the innovations and the representation (2.3) yield

$$\mathrm{supp}(Y_{ji}) = \left\{\frac{\bigvee_{\ell\in\mathrm{An}(i)} b_{\ell i}z_\ell}{\bigvee_{\ell\in\mathrm{An}(j)} b_{\ell j}z_\ell} : \boldsymbol{z}_{\mathrm{An}(i)\cup\mathrm{An}(j)} \in \mathbb{R}_+^{|\mathrm{An}(i)\cup\mathrm{An}(j)|}\right\}.$$

The continuity of the function

$$\mathbb{R}_+^{|\mathrm{An}(i)\cup\mathrm{An}(j)|} \to \mathbb{R}_+, \quad \boldsymbol{z}_{\mathrm{An}(i)\cup\mathrm{An}(j)} \mapsto \frac{\bigvee_{\ell\in\mathrm{An}(i)} b_{\ell i}z_\ell}{\bigvee_{\ell\in\mathrm{An}(j)} b_{\ell j}z_\ell}$$

implies that $\mathrm{supp}(Y_{ji})$ is an interval in $\mathbb{R}_+$. Since for $j \in \mathrm{an}(i)$ by Corollary 3.13 of [12] $b_{ji} \leqslant Y_{ji}$ and by (a) b_{ji} is an atom of Y_{ji}, it suffices to show that $j \in \mathrm{an}(i)$ if $\mathrm{supp}(Y_{ji})$ has a positive lower bound. For this assume that $j \notin \mathrm{an}(i)$. Because of the positive lower bound of $\mathrm{supp}(Y_{ji})$, there exists some $a \in \mathbb{R}_+$ such that

$$\bigvee_{\ell\in\mathrm{An}(i)\cap\mathrm{An}(j)} ab_{\ell j}z_\ell \vee \bigvee_{\ell\in\mathrm{An}(j)\setminus\mathrm{An}(i)} ab_{\ell j}z_\ell \leqslant \bigvee_{\ell\in\mathrm{An}(i)} b_{\ell i}z_\ell \tag{3.2}$$ {ch4:lowboundsu

for all $\boldsymbol{z}_{\mathrm{An}(i)\cup\mathrm{An}(j)} \in \mathbb{R}_+^{|\mathrm{An}(i)\cup\mathrm{An}(j)|}$. As $\mathrm{An}(j)\setminus\mathrm{An}(i) \neq \varnothing$, for fixed $\mathbf{z}_{\mathrm{An}(i)} \in \mathbb{R}_+^{|\mathrm{An}(i)|}$, we can choose z_ℓ for some $\ell \in \mathrm{An}(j)\setminus\mathrm{An}(i)$ so large that $ab_{\ell j}z_\ell$ is greater than the maximum on the right-hand side of (3.2). This contradicts (3.2). Hence, $j \in \mathrm{an}(i)$. □

Table 3.1: Distributional properties of Y_{ji} for distinct $i, j \in V$. ch4:table:rel

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

In Table 3.1 we summarize the results of Lemma 3.2: depending on the relationship between i and j in $\mathcal{D}$, the support and atoms of Y_{ji} are shown.

Table 3.1 and the fact that $b_{ji} = 0$ for $j \notin \operatorname{An}(i)$ (cf. (2.2)) suggest the following algorithm to find B from $\mathcal{L}(\boldsymbol{X})$ since we can identify the support of Y_{ji} from $\mathcal{L}(\boldsymbol{X})$. This proves the identifiability of B from $\mathcal{L}(\boldsymbol{X})$. In fact, it is sufficient to know $\operatorname{supp}(Y_{ji})$ for all $i, j \in V$ with $i \neq j$ rather than the whole distribution $\mathcal{L}(\boldsymbol{X})$.

Algorithm 3.3. [Find B from $\mathcal{L}(\boldsymbol{X})$] {ch4:alg1}

1. For all $i \in V = \{1, \ldots, d\}$, set $b_{ii} = 1$.

2. For all $i, j \in V$ with $i \neq j$, find $\operatorname{supp}(Y_{ji})$:

   if $\operatorname{supp}(Y_{ji}) = [a, \infty)$ for some $a \in \mathbb{R}_+$, then set $b_{ji} = a$;

   else, set $b_{ji} = 0$.

So far we have shown that the ML coefficient matrix B of $\boldsymbol{X}$ can be obtained from $\mathcal{L}(\boldsymbol{X})$. Since all DAGs and edge weights that represent $\boldsymbol{X}$ in the sense of (2.1) can be determined from B, the only quantities we do not know about yet but appear in the definition of $\boldsymbol{X}$ are the innovations. In what follows we show that the distribution of the innovation vector $\boldsymbol{Z}$ is also identifiable from $\mathcal{L}(\boldsymbol{X})$. For this, due to the identifiability of B from $\mathcal{L}(\boldsymbol{X})$ and the independence of the innovations, it suffices to provide an algorithm that determines the distributions of the innovations from $\mathcal{L}(\boldsymbol{X})$ and B. Note that B also determines the ancestral relationships between any pair of nodes in that $j \in \operatorname{An}(i)$ for any DAG representing $\boldsymbol{X}$ if and only if $b_{ji} > 0$.

We denote by F_{Z_i} the distribution function of the innovation Z_i. Even with this algorithm, we do not have to know the whole distribution $\mathcal{L}(\boldsymbol{X})$; it is enough to know the ML coefficient matrix B and the univariate marginal distribution functions of $\mathcal{L}(\boldsymbol{X})$.

Algorithm 3.4. [Find $F_{Z_1}(x), \ldots, F_{Z_d}(x)$ for $x \in \mathbb{R}_+$ from B and $\mathcal{L}(\boldsymbol{X})$] {ch4:alg3}
For $\nu = 0, \ldots, d-1$,
for $i \in V$ such that $|\operatorname{an}(i)| = |\{j \in V \backslash \{i\} : b_{ji} \neq 0\}| = \nu$, set

$$F_{Z_i}(x) = \frac{\mathbb{P}(X_i \leqslant x)}{\prod_{j \in \operatorname{an}(i)} F_{Z_j}(x/b_{ji})}.$$

Here we have used the convention that $\prod_{j \in \varnothing} a_j = 1$. The correctness of Algorithm 3.4 follows from the independence of the innovations and representation (2.3). Finally, we summarize the main result of this section.

{ch4:identify}

Theorem 3.5. *Let $\mathcal{L}(\boldsymbol{X})$ be the distribution of $\boldsymbol{X}$ following a recursive ML model. Then its ML coefficient matrix B and the distribution of its innovation vector $\boldsymbol{Z}$ are identifiable from $\mathcal{L}(\boldsymbol{X})$. Furthermore, the class of all DAGs and edge weights that could have generated $\boldsymbol{X}$ by* (2.1) *can be obtained.*

4 Estimation with known directed acyclic graph {ch4:s5}

In this section we consider independent realizations $\boldsymbol{x}^{(t)} = \big(x_1^{(t)}, \ldots, x_d^{(t)}\big)$, $t = 1, \ldots, n$, of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ following a recursive ML model with its DAG $\mathcal{D}$ given. Further, we consider the distribution of the innovation vector to be fixed, but the only information we eventually use is that it has independent, atom-free margins with support $\mathbb{R}_+$, hence our estimation results also cover the situation where this distribution is unknown. Our aim is the estimation of the edge weights c_{ki} and the ML coefficient matrix B. Since $\boldsymbol{X}$ may satisfy (2.1) with respect to $\mathcal{D}$ for different systems of edge weights c_{ki} (see Example 3.1) and thus these are generally not identifiable from $\mathcal{L}(\boldsymbol{X})$, we usually have no chance to estimate the true edge weights from $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ consistently, but only the ML coefficient matrix B. Using the fact that the class of possible edge weights can be determined from B, we obtain automatically an estimate of this class; see Corollary 4.12 below.

The ML coefficient matrix $\boldsymbol{B}$

In the following we let $\mathcal{B}(\mathcal{D})$ denote the class of possible ML coefficient matrices of all recursive ML models on $\mathcal{D}$. For B being a matrix with non-negative entries and diagonal elements $b_{ii} = 1$ we define $B_0 := (b_{ij}\mathbb{1}_{\mathrm{pa}(j)}(i))_{d\times d}$. Then it holds that $B \in \mathcal{B}(\mathcal{D})$ if and only if B satisfies the following

$$[b_{ji} > 0 \iff j \in \mathrm{An}(i)] \text{ and } B = I_d \vee (B \odot B_0); \tag{4.1}$$ {eq:fixpoint}

see Theorem 4.2 or Corollary 4.3(a) of [12].

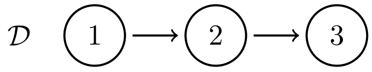
A simple estimate of $\boldsymbol{B}$ {ch4:s52}

Next we discuss a sensible estimate of B. Table 3.1 shows that for $j \in \mathrm{an}(i)$ the minimal value that can be observed for the ratio $Y_{ji} = X_i/X_j$ is b_{ji}, which is an atom of Y_{ji}. This suggests the following estimate $\breve{B}$ of the ML coefficient matrix:

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

Davis and Resnick [6] suggested such minimal observed ratios as estimates for parameters in max-ARMA processes. For n sufficiently large, we can expect to observe the atoms b_{ji} for $j \in \mathrm{an}(i)$ in the sample $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and, hence, to estimate the ML coefficients exactly. However, if n is not large we may with positive probability have that $\breve{B}$ is not an ML coefficient matrix of any recursive ML model on $\mathcal{D}$ as the following simple example shows:

Example 4.1. [$\breve{B}$ is not necessarily in $\mathcal{B}(\mathcal{D})$]
Consider the DAG



and assume we observe $\breve{b}_{31} > \breve{b}_{32}\breve{b}_{21}$. Then the matrix $\breve{B}$ fails to satisfy (4.1) and hence is not an element of $\mathcal{B}(\mathcal{D})$. □

However, if we only estimate the ML coefficients corresponding to edges in $\mathcal{D}$ and then compute an estimate based on Lemma 4.2 below this phenomenon cannot occur.

{lem:solving}

Lemma 4.2. *Let $B_0 \in \mathbb{R}_+^{d\times d}$ be a matrix with $b_{ij} > 0 \iff i \to j$. A matrix $A \in \mathbb{R}_+^{d\times d}$ satisfies*

$$[a_{ji} > 0 \iff j \in \mathrm{An}(i)] \textit{ and } A = I_d \vee (A \odot B_0);$$

if and only if $A = (I_d \vee B_0)^{\odot(d-1)}$.

Proof. We first show that $A = (I_d \vee B_0)^{\odot(d-1)}$ is a solution. We have ([4], Proposition 1.6.10)

$$(I_d \vee B_0)^{\odot(d-1)} = \bigvee_{k=0}^{d-1} B_0^{\odot k} = \bigvee_{k=0}^{\infty} B_0^{\odot k}$$

and hence

$$I_d \vee (A \odot B_0) = I_d \vee \{(I_d \vee B_0)^{\odot(d-1)} \odot B_0\} = I_d \vee \{\bigvee_{k=1}^{\infty} B_0^{\odot k}\} = \bigvee_{k=0}^{\infty} B_0^{\odot k} = A.$$

It is easy to see directly that $B_0^{\odot k} = 0$ for $k \geqslant d$ and hence if $\breve{A}$ is a solution to (4.1) we get by iteration, using that $(M \vee N) \odot K = (M \odot K) \vee (N \odot K)$,

$$\begin{aligned}
\breve{A} &= I_d \vee (\breve{A} \odot B_0) \\
&= I_d \vee [\{I_d \vee (\breve{A} \odot B_0)\} \odot B_0] \\
&= I_d \vee B_0 \vee (\breve{A} \odot B_0^{\odot 2}) \\
&= \cdots \\
&= (I_d \vee B_0)^{\odot(d-1)} \vee (\breve{A} \odot B_0^{\odot d}) = (I_d \vee B_0)^{\odot(d-1)} = A
\end{aligned}$$

and hence the solution to the equation is unique. □

For a path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$ and a realization $\boldsymbol{x}^{(t)}$ such that $\breve{b}_{ji}x_j^{(t)} = x_i^{(t)}$ we have

$$\Big(\bigwedge_{s=1}^{n} y_{01}^{(s)}\Big)\Big(\bigwedge_{s=1}^{n} y_{12}^{(s)}\Big)\cdots\Big(\bigwedge_{s=1}^{n} y_{n-1,n}^{(s)}\Big) \leqslant y_{01}^{(t)}y_{12}^{(t)}\cdots y_{n-1,n}^{(t)} = y_{ji}^{(t)} = \breve{b}_{ji} = \bigwedge_{s=1}^{n} y_{ji}^{(s)}. \qquad (4.2)$$

{ch4:eq:wtBwhB

Thus we may define the estimate $\widehat{B}$ by first calculating the matrix $\breve{B}_0 = (\breve{b}_{ij}\mathbb{1}_{\mathrm{pa}(j)}(i))_{d\times d}$ and then iterating the $\odot$-matrix product as:

$$\widehat{B} = (I_d \vee \breve{B}_0)^{\odot(d-1)}. \qquad (4.3)$$

{ch4:eq:Bhat}

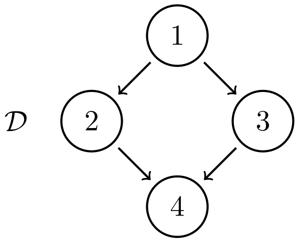
It then follows from (4.2) that $\widehat{B}_0 = \breve{B}_0$ and from Lemma 4.2 that $\widehat{B}$ is the unique element of $\mathcal{B}(\mathcal{D})$ with this property. By Lemma 3.2(b), we also have

$$b_{ji} \leqslant \widehat{b}_{ji} \leqslant \breve{b}_{ji} \text{ for } j \in \mathrm{an}(i).$$

Consequently, when using $\widehat{B}$ or $\breve{B}$ as an estimate of B, we never underestimate a ML coefficient; furthermore, the matrix $\widehat{B}$ always estimates B more precisely than $\breve{B}$ and since we always have $\widehat{B} \in \mathcal{B}(\mathcal{D})$, $\widehat{B}$ seems to be clearly preferable as an estimate of B.

The following two examples show how effective the estimate $\widehat{B}$ can be; in particular, n does not necessarily need to be large.

Example 4.3. [One observation may be enough to estimate B exactly] {ch4:examBhat}
Consider the DAG

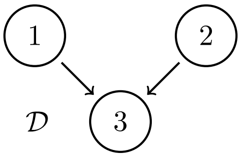


and assume that the paths $[1 \to 2 \to 4]$ and $[1 \to 3 \to 4]$ are both max-weighted, which is equivalent to $b_{12}b_{24} = b_{13}b_{34}$. If we observe the event

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

then $\widehat{B} = B$ so we estimate all ML coefficients exactly. Note that this event has positive probability and occurs $\mathbb{P}$-almost surely if and only if Z_1 realizes all node variables; i.e., if $X_2 = b_{12}Z_1$, $X_3 = b_{13}Z_1$, and $X_4 = b_{14}Z_1$. □

Example 4.4. Consider the DAG



We have four events that may occur, namely,

$$F_1 = \{X_3 = b_{13}X_1\} \cap \{X_3 = b_{23}X_2\}, \quad F_2 = \{X_3 = b_{13}X_1\} \cap \{X_3 > b_{23}X_2\},$$
$$F_3 = \{X_3 > b_{13}X_1\} \cap \{X_3 = b_{23}X_2\}, \quad F_4 = \{X_3 > b_{13}X_1\} \cap \{X_3 > b_{23}X_2\},$$

When excluding the null event F_1, we estimate B by $\widehat{B}$ exactly if and only if the observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ include both of the events F_2 and F_3; so two observations may be enough to estimate B exactly. □

Since by Table 3.1 $\mathbb{P}(X_i = b_{ki}X_k) > 0$ for $k \in \mathrm{pa}(i)$, it follows from the Borel-Cantelli lemma that $\widehat{b}_{ki}$ $\mathbb{P}$-almost surely equals the true value for n sufficiently large. Thus, if n is large, $\widehat{B}$ finds, with probability 1, the true B. In [6] this is discussed for the time-series framework used there and in Davis and McCormick [5] they show that under suitable assumptions, this estimator is asymptotically Fréchet distributed . Assuming the probability of $\{X_i = b_{ki}X_k\}$ is known, we show next how one has to choose n to observe this event with probability greater than $1 - p$ for some $p \in (0, 1)$. We also prove that the probability for estimating the true b_{ki} converges exponentially fast to 1.

{ch4:PropBhat}

Proposition 4.5. *Let $\boldsymbol{X}^{(t)} = \big(X_1^{(t)}, \dots, X_n^{(t)}\big)$ for $t = 1, \dots, n$ be a sample from a recursive ML model on a DAG $\mathcal{D}$ with ML coefficient matrix B. Let $i \in V$ and $k \in \mathrm{pa}(i)$. It then holds that*

$$\mathbb{P}\left(\bigwedge_{t=1}^{n} Y_{ki}^{(t)} = b_{ki}\right) \geqslant 1 - p \textit{ for some } p \in (0,1)$$

if and only if

$$n \geqslant \frac{\ln(p)}{\ln(\mathbb{P}(Y_{ki} > b_{ki}))}.$$

Furthermore, the convergence $\mathbb{P}\big(\bigwedge_{t=1}^{n} Y_{ki}^{(t)} = b_{ki}\big) \to 1$ as $n \to \infty$ is exponentially fast.

Proof. First note that the events $\{X_i = b_{ki}X_k\}$ and $\{X_i > b_{ki}X_k\}$ are complementary and both have positive probability. Further, using that $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)}$ are independent and identically distributed yields

$$\mathbb{P}\Big(\bigwedge_{t=1}^{n} Y_{ki}^{(t)} = b_{ki}\Big) = 1 - \mathbb{P}\Big(\bigwedge_{t=1}^{n} Y_{ki}^{(t)} > b_{ki}\Big) = 1 - \prod_{t=1}^{n} \mathbb{P}(Y_{ki}^{(t)} > b_{ki}) = 1 - \mathbb{P}(Y_{ki} > b_{ki})^n.$$

Altogether, the statements follow. □

In conclusion, $\widehat{B}$ has the nice property to be 'geometrically consistent' in the sense that the probability of $\{\widehat{B} = B\}$ converges exponentially fast to one.

The matrix $\widehat{\boldsymbol{B}}$ is a generalized maximum likelihood estimate

{ch4:s51}

For $B \in \mathcal{B}(\mathcal{D})$ and a fixed distribution of the innovation vector we let P_B denote the probability measure induced by a recursive ML model on $\mathcal{D}$ with ML coefficient matrix B, i.e. the distribution of $\boldsymbol{X}$ where $\boldsymbol{X} = \boldsymbol{Z} \odot B$. We shall denote the family of these probability measures by $\mathcal{P}(\mathcal{D})$.

We cannot use standard maximum likelihood methods to estimate B, since the family $\mathcal{P}(\mathcal{D})$ is not dominated (cf. Example 4.4.1 of [11]) and hence the standard likelihood function is not well defined. However, there exist generalizations of maximum likelihood estimation (GMLE) that cover the undominated case as well; Kalbfleisch and Prentice [17], Kiefer and Wolfowitz [18], and Scholz [26] suggested such extensions. We essentially follow the Kiefer–Wolfowitz definition of a GMLE; as also done, for example, by Gill et al. [10] and Johansen [16] and in the following we shall show that $\widehat{B}$ can be seen as a maximum likelihood estimate of B in the extended sense introduced by Kiefer and Wolfowitz in [18].

Let $\mathcal{P}$ be a family of probability measures on $(\mathbb{R}_+^d, \mathbb{B}(\mathbb{R}_+^d))$ where $\mathbb{B}(\mathbb{R}_+^d)$ denotes the Borel σ-algebra on $\mathbb{R}_+^d$, and $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ a random sample from some $P_0 \in \mathcal{P}$. For $P, Q \in \mathcal{P}$ and $\boldsymbol{x} \in \mathbb{R}_+^d$ we define

$$\rho(\boldsymbol{x}, P, Q) := \frac{dP}{d(P+Q)}(\boldsymbol{x}),$$

where $dP/d(P+Q)$ denotes a density of P with respect to $P+Q$. Then we call $\widehat{P}$ a *generalized maximum likelihood estimate* of P_0 if

$$\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, \widehat{P}, \widehat{P}) \neq 0 \quad \text{and} \quad \prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, Q, \widehat{P}) \leqslant \prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, \widehat{P}, Q) \text{ for all } Q \in \mathcal{P}. \tag{4.4}$$

{ch4:def:gmle}

Since P is absolutely continuous with respect to $P+Q$, the density $dP/d(P+Q)$ always exists according to the Radon-Nikodym theorem. This means that the GMLE is well-defined, save for the usual ambiguity in the method of maximum likelihood that densities are only defined up to null sets and therefore a specific choice of densities must be made. The Kiefer–Wolfowitz definition extends the definition of a MLE in a very natural way as it simply says that $\widehat{P}$ is the MLE in the smaller family $\{\hat{P}, Q\}$ for any $Q \in \mathcal{P}$. In [18] only the second condition in (4.4) is required, but the first condition is implicit. The first step in verfying that $\widehat{B}$ is a GMLE of B is to specify densities of P_B with respect to $P_B + P_{B*}$ for any two $B, B^* \in \mathcal{B}(\mathcal{D})$. For this purpose we determine a partition $\big\{A_0(B,B^*), A_{1/2}(B,B^*), A_1(B,B^*)\big\}$ of $\mathbb{R}^d_+$ that satisfies the following three properties,

$$\begin{aligned} &\text{(A):} \quad P_B(A_0(B,B^*)) = 0,\\ &\text{(B):} \quad P_B(A \cap A_{1/2}(B,B^*)) = P_{B*}(A \cap A_{1/2}(B,B^*)) \text{ for every } A \in \mathbb{B}(\mathbb{R}^d_+),\\ &\text{(C):} \quad P_{B*}(A_1(B,B^*)) = 0. \end{aligned} \tag{4.5}$$ {eq:partition}

Then we choose as density the measurable function from $\mathbb{R}^d_+$ to $\{0, 1/2, 1\}$ defined as

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{x}, B, B^*) := \frac{1}{2} \cdot \mathbb{1}_{A_{1/2}(B,B^*)}(\boldsymbol{x}) + \mathbb{1}_{A_1(B,B^*)}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in A_0(B,B^*),\\ \frac{1}{2}, & \text{if } \boldsymbol{x} \in A_{1/2}(B,B^*),\\ 1, & \text{if } \boldsymbol{x} \in A_1(B,B^*). \end{cases} \tag{4.6}$$ {ch4:dens:gen}

This is a valid density because, using the properties (A), (B), (C), we obtain for every $A \in \mathbb{B}(\mathbb{R}^d_+)$,

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

We begin with an example that shall help to get an idea and provide insights into the concepts and arguments we shall use in the general case. It is deliberately very detailed.

Example 4.6. [How to find a density and the associated GMLEs] {ch4:ex1}
For $B, B^* \in \mathcal{B}(\mathcal{D})$ where $\mathcal{D} = (\{1,2\}, 1 \to 2)$, we show that the partition

$$\begin{aligned} \Big\{ A_0(B,B^*) &:= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 < b_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b^*_{12}x_1 > b_{12}x_1\big\},\\ A_{1/2}(B,B^*) &:= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b_{12}x_1 = b^*_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 > (b_{12} \vee b^*_{12})x_1\big\},\\ A_1(B,B^*) &:= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : b^*_{12}x_1 > x_2 \geqslant b_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b_{12}x_1 > b^*_{12}x_1\big\}\Big\} \end{aligned}$$

of $\mathbb{R}^2_+$ satisfies properties (A), (B), (C) of (4.5). Figure 4.1 shows the corresponding density $\rho(\cdot, B, B^*)$ from (4.6) for the three possible order relations between b_{12} and b^*_{12}.

Since by Table 3.1, $\operatorname{supp}(X_2/X_1) = [b_{12}, \infty)$ and b_{12} is the only atom of X_2/X_1, property (A) is true. By reversing the roles of B and B^*, (C) follows from (A). The condition (B) is obvious if $b_{12} = b^*_{12}$. Assume that $b_{12} \neq b^*_{12}$. We then have by definition of $\boldsymbol{X}$ that $\{\boldsymbol{X} \in A_{1/2}(B,B^*)\} = \{X_2 > (b_{12} \vee b^*_{12})X_1\} = \{Z_2 > (b_{12} \vee b^*_{12})Z_1\}$ and $X_2 = Z_2$ on $\{Z_2 > (b_{12} \vee b^*_{12})Z_1\}$. With this, using that $A_{1/2}(B^*,B) = A_{1/2}(B,B^*)$. We finally obtain for $A \in \mathbb{B}(\mathbb{R}^2_+)$,

$$\begin{aligned} P_B(A \cap A_{1/2}(B,B^*)) &= \mathbb{P}(\{\boldsymbol{X} \in A\} \cap \{Z_2 > (b_{12} \vee b^*_{12})Z_1\})\\ &= \mathbb{P}(\{(Z_1,Z_2) \in A\} \cap \{Z_2 > (b_{12} \vee b^*_{12})Z_1\})\\ &= P_{B*}(A \cap A_{1/2}(B^*,B)) = P_{B*}(A \cap A_{1/2}(B,B^*)). \end{aligned}$$

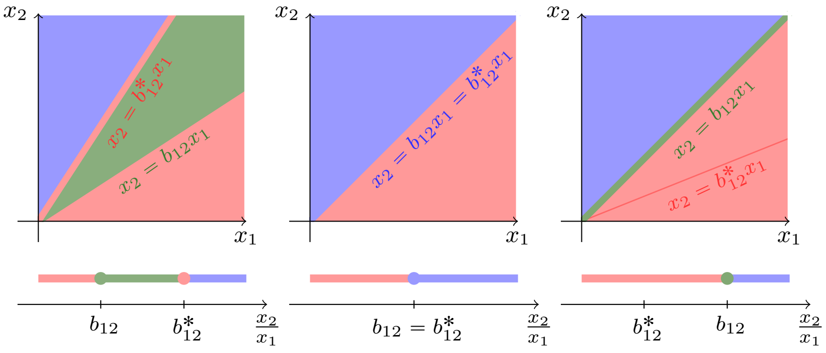


Figure 4.1: The density $\rho(\cdot, B, B^*)$ from Example 4.6 shown as a contour plot (top line) and as a function of $y_{12} = x_2/x_1$ (bottom line) for the three situations $b_{12} < b_{12}^*$ (left-hand side), $b_{12} = b_{12}^*$ (middle), and $b_{12} > b_{12}^*$ (right-hand side). The area where it is $0/\frac{1}{2}/1$ is coloured in red/blue/green. ch4:fig1:DAG12

We now use the density found to determine the GMLE of B. The only ML coefficient we have to estimate is b_{12}. As before we let $\widehat{b}_{12} = \breve{b}_{12}$ be the minimal observed ratio of X_2/X_1 and let $\widehat{B}$ be the corresponding ML coefficient matrix from (4.3). Defining $n(B, B^*) = |\{t : \boldsymbol{x}^{(t)} \in A_{1/2}(B, B^*)\}|$ and using that $n(B, B^*) = n(B^*, B)$, we obtain

$$\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, B, B^*) = 2^{-n(B,B^*)} \prod_{t=1}^{n} \mathbb{1}_{\mathbb{R}_+^d \setminus A_0(B,B^*)}\big(\boldsymbol{x}^{(t)}\big),$$
$$\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, B^*, B) = 2^{-n(B,B^*)} \prod_{t=1}^{n} \mathbb{1}_{\mathbb{R}_+^d \setminus A_0(B^*,B)}\big(\boldsymbol{x}^{(t)}\big).$$

Let now $\widetilde{B}$ be an arbitrary potential GMLE of B. Then $P_{\widetilde{B}} \in \mathcal{P}(\mathcal{D})$ satisfies the first condition in (4.4) if and only if

$$\widetilde{b}_{12} x_1^{(t)} \leqslant x_2^{(t)} \text{ for all } t, \text{ equivalently } \widetilde{b}_{12} \leqslant \widehat{b}_{12} \tag{4.7}$$ {ch4:cond1:DAG1

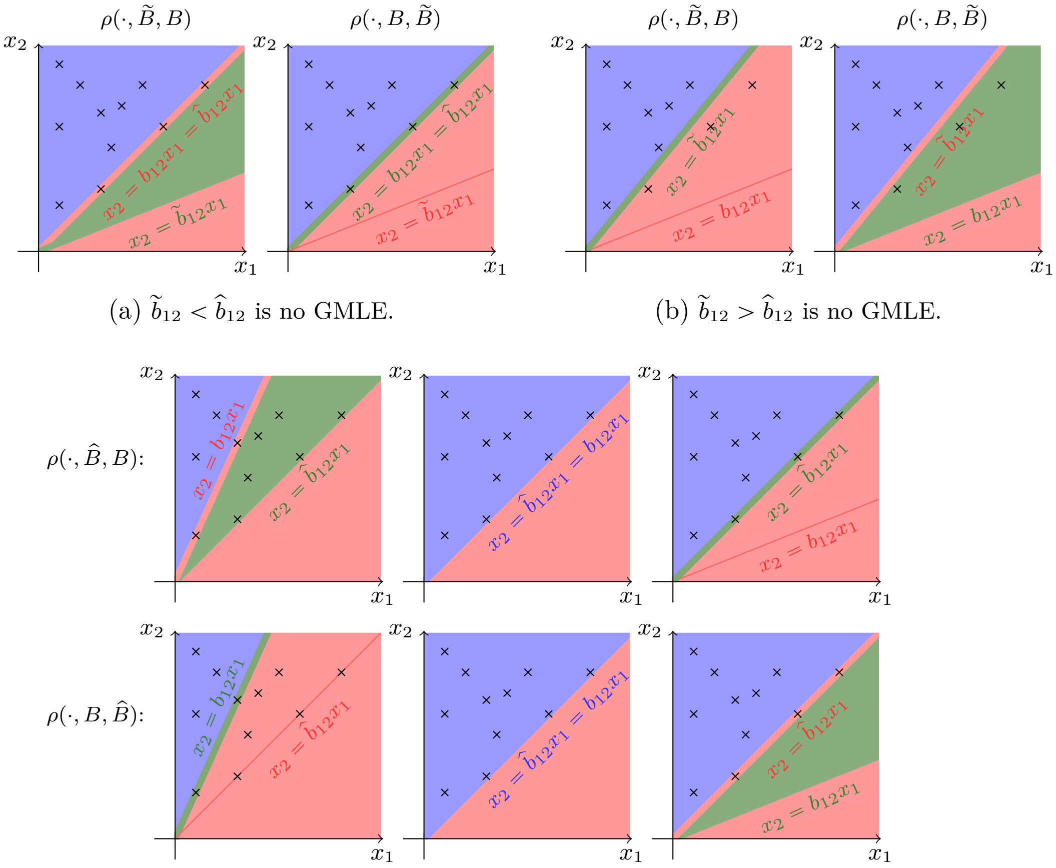
and the second if and only if

$$\text{for all } B \in \mathcal{B}(\mathcal{D}), \text{ if some } \boldsymbol{x}^{(t)} \in A_0(\widetilde{B}, B), \text{ then some } \boldsymbol{x}^{(s)} \in A_0(B, \widetilde{B}). \tag{4.8}$$ {ch4:cond2:DAG1

In summary, some $\widetilde{B} \in \mathcal{B}(\mathcal{D})$ is a GMLE of B if and only if (4.7) and (4.8) are satisfied. We discuss the possible GMLEs of b_{12} in detail.

(a) $\widetilde{b}_{12} < \widehat{b}_{12}$ is no GMLE:
Set $b_{12} = \widehat{b}_{12}$, and let $\boldsymbol{x}^{(t)}$ be such that $\widehat{b}_{12} x_1^{(t)} = x_2^{(t)}$. Then $\boldsymbol{x}^{(t)} \in \big\{\boldsymbol{x} \in \mathbb{R}_+^2 : x_2 = b_{12} x_1 > \widetilde{b}_{12} x_2\big\} \subseteq A_0(\widetilde{B}, B)$ but no $\boldsymbol{x}^{(s)} \in A_0(B, \widetilde{B}) = \big\{\boldsymbol{x} \in \mathbb{R}_+^2 : x_2 < b_{12} x_1\big\}$. This contradicts (4.8); consequently, $\widetilde{b}_{12}$ cannot be a GMLE of b_{12}. In Figure 4.2(a) we illustrate this situation. On the left-hand side a contour plot of the density $\rho(\cdot, \widetilde{B}, B)$ is shown, on the right-hand side of $\rho(\cdot, B, \widetilde{B})$. The crosses represent the realizations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$. In the left plot crosses are in the 0-area coloured in red, namely, those that realize $\widehat{b}_{12}$, but in the right plot not. So $\widetilde{B}$ cannot be a GMLE of B.

(b) $\widetilde{b}_{12} > \widehat{b}_{12}$ is no GMLE:
This follows directly from (4.7). Figure 4.2(b) shows a situation that contradicts (4.8), similarly to Figure 4.2(a) in (1).



(a) $\widetilde{b}_{12} < \widehat{b}_{12}$ is no GMLE.

(b) $\widetilde{b}_{12} > \widehat{b}_{12}$ is no GMLE.

(c) $\widetilde{b}_{12} = \widehat{b}_{12}$ is a GMLE.

Figure 4.2: Discussion of the GMLEs of b_{12} with respect to the density from Figure 4.1.; see further explanation in (a), (b), and (c) of Example 4.6. ch4:fig2:DAG12

(c) $\widetilde{b}_{12} = \widehat{b}_{12}$ is a GMLE:
Condition (4.7) holds obviously. To prove (4.8), assume for some $B \in \mathcal{B}(\mathcal{D})$ that some $\boldsymbol{x}^{(t)} \in A_0(\widehat{B},B)$. By definition of $A_0(\widehat{B},B)$, $x_2^{(t)} = b_{12}x_1^{(t)} > \widehat{b}_{12}x_1^{(t)}$, which implies that $b_{12} > \widehat{b}_{12}$. For $\boldsymbol{x}^{(s)}$ such that $\widehat{b}_{12}x_1^{(s)} = x_2^{(s)}$, we then find that $x_2^{(s)} < b_{12}x_1^{(s)}$. Hence, $\boldsymbol{x}^{(s)} \in A_0(B,\widehat{B})$, and $\widehat{b}_{12}$ is a GMLE of b_{12}. We learn this informally from Figure 4.2(c). The top line shows contour plots of $\rho(\cdot,\widehat{B},B)$ for the three different orders between b_{12} and $\widehat{b}_{12}$, and the bottom line shows the corresponding contour plots of $\rho(\cdot,B,\widehat{B})$. The two plots on the left-hand side correspond to the situation from above: in the upper plot there are realizations in the 0-area, namely those that are on the line $\{\boldsymbol{x} \in \mathbb{R}_+^2 : x_2 = b_{12}x_1\}$, but then there are also realizations in the 0-area of the lower plot (those that lie below this line). Hence, (4.8) holds. Since there is no realization in the 0-area of the middle and right plot in the top line, (4.8) is automatically satisfied if $b_{12} \leqslant \widehat{b}_{12}$. □

In what follows we specify, for the general case, one density of P_B with respect to $P_B + P_{B*}$ that has a representation as in (4.6) and leads to $\widehat{B}$ as a GMLE of B.

Our partition $\{A_0(B,B^*), A_{1/2}(B,B^*), A_1(B,B^*)\}$ of $\mathbb{R}_+^d$ is based on the following representation for the components of $\boldsymbol{X}$:

$$X_i = \bigvee_{k \in \mathrm{pa}(i)} b_{ki}X_k \vee Z_i; \quad \text{in particular, } X_i \geqslant \bigvee_{k \in \mathrm{pa}(i)} b_{ki}X_k, \quad i \in V. \tag{4.9}$$ {ch4:proof:dens

We begin with the specification of $A_{1/2}(B,B^*)$ and prove a property needed subsequently to verify property (B). Have in mind that if $b_{ki} > b^*_{ki}$ for all $k \in \mathrm{pa}(i)$ or $b_{ki} < b^*_{ki}$ for all $k \in \mathrm{pa}(i)$ then $\big\{\boldsymbol{x} \in \mathbb{R}^d_+ : x_i = \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\big\} = \varnothing$.

{ch4:lem}

Lemma 4.7. *Let $B, B^* \in \mathcal{B}(\mathcal{D})$ and define*

$$\Omega(B,B^*) := \bigcap_{i=1}^{d} \Big\{ \bigvee_{j\in\mathrm{An}(i):b_{ji}=b^*_{ji}} b_{ji}Z_j > \bigvee_{j\in\mathrm{an}(i):b_{ji}\neq b^*_{ji}} (b_{ji} \vee b^*_{ji})Z_j \Big\},$$

$$A_{1/2}(B,B^*) := \bigcap_{i=1}^{d} \Big[\Big\{\boldsymbol{x} \in \mathbb{R}^d_+ : x_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k = \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k\Big\} \cup \Big\{\boldsymbol{x} \in \mathbb{R}^d_+ : x_i > \bigvee_{k\in\mathrm{pa}(i)} (b_{ki} \vee b^*_{ki})x_k\Big\}\Big].$$

Then for every $F \in \mathcal{F}$,

$$\mathbb{P}(F \cap \{\boldsymbol{X} \in A_{1/2}(B,B^*)\}) = \mathbb{P}(F \cap \Omega(B,B^*)). \tag{4.10}$$ {ch4:lem:eq3}

Proof. First, define for $i \in V$

$$\Omega^{1,i}_{1/2} := \Big\{X_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}X_k = \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}X_k\Big\}, \quad \Omega^{2,i}_{1/2} := \Big\{X_i > \bigvee_{k\in\mathrm{pa}(i)} (b_{ki} \vee b^*_{ki})X_k\Big\},$$
$$\Omega_i := \Big\{ \bigvee_{j\in\mathrm{An}(i):b_{ji}=b^*_{ji}} b_{ji}Z_j > \bigvee_{j\in\mathrm{an}(i):b_{ji}\neq b^*_{ji}} (b_{ji} \vee b^*_{ji})Z_j \Big\}.$$

The proof is by induction on the number of nodes of $\mathcal{D}$. For $d = 1$ the statement is clear. Assume now that $\mathcal{D} = (V,E)$ has $d+1$ nodes and that the assertion holds with respect to DAGs with at most d nodes. Furthermore, assume without loss of generality that $d+1$ is a terminal node (i.e., $\mathrm{de}(d+1) = \varnothing$). Since $(X_1,\dots,X_d)$ follows a recursive ML model on the DAG $(\{1,\dots,d\}, E \cap (\{1,\dots,d\} \times \{1,\dots,d\}))$ with ML coefficient matrix $B = (b_{ij})_{d\times d}$ and $B^* = (b^*_{ij})_{d\times d}$ is the ML coefficient matrix of a recursive ML model on this DAG as well, the induction hypothesis yields that

$$\mathbb{P}(F \cap \{\boldsymbol{X} \in A_{1/2}(B,B^*)\}) = \mathbb{P}\Big(F \cap \bigcap_{i=1}^{d+1} \big(\Omega^{1,i}_{1/2} \cup \Omega^{2,i}_{1/2}\big)\Big) = \mathbb{P}\Big(F \cap \bigcap_{i=1}^{d} \Omega_i \cap \big(\Omega^{1,d+1}_{1/2} \cup \Omega^{2,d+1}_{1/2}\big)\Big). \tag{4.11}$$ {ch4:lem:eq2}

For every $i \in V$ we have by (2.3) on Ω_i that

$$X_i = \bigvee_{j\in\mathrm{An}(i)} b_{ji}Z_j = \bigvee_{j\in\mathrm{An}(i)} b^*_{ji}Z_j. \tag{4.12}$$ {ch4:lem:eq1}

Noting from the proof of Theorem 4.2 of [12] that

$$\bigvee_{k\in\mathrm{pa}(d+1)} b_{k,d+1}X_k = \bigvee_{k\in\mathrm{pa}(d+1)} b_{k,d+1} \bigvee_{j\in\mathrm{An}(k)} b_{jk}Z_j = \bigvee_{j\in\mathrm{an}(d+1)} b_{j,d+1}Z_j,$$

we obtain from (4.12) on $\bigcap_{i=1}^{d} \Omega_i$,

$$\bigvee_{k\in\mathrm{pa}(d+1)} b^*_{k,d+1}X_k = \bigvee_{k\in\mathrm{pa}(d+1)} b^*_{k,d+1} \bigvee_{j\in\mathrm{An}(k)} b^*_{jk}Z_j = \bigvee_{j\in\mathrm{an}(i)} b^*_{j,d+1}Z_j.$$

Thus, again by (2.3),

$$\begin{aligned}
\bigcap_{i=1}^{d} \Omega_i \cap \Omega_{1/2}^{1,d+1} &= \bigcap_{i=1}^{d} \Omega_i \cap \{ \bigvee_{j\in \mathrm{An}(d+1)} b_{j,d+1} Z_j = \bigvee_{j\in \mathrm{an}(d+1)} b_{j,d+1} Z_j = \bigvee_{j\in \mathrm{an}(d+1)} b^*_{j,d+1} Z_j \}, \\
\bigcap_{i=1}^{d} \Omega_i \cap \Omega_{1/2}^{2,d+1} &= \bigcap_{i=1}^{d} \Omega_i \cap \{ \bigvee_{j\in \mathrm{An}(d+1)} b_{j,d+1} Z_j > \bigvee_{j\in \mathrm{an}(d+1)} (b_{j,d+1} \vee b^*_{j,d+1}) Z_j \} \\
&= \bigcap_{i=1}^{d} \Omega_i \cap \big\{ b_{j,d+1} Z_j > \bigvee_{j\in \mathrm{an}(d+1)} (b_{j,d+1} \vee b^*_{j,d+1}) Z_j \big\}.
\end{aligned}$$

From (3.1) we then finally observe that $\bigcap_{i=1}^d \Omega_i \cap \big(\Omega_{1/2}^{1,d+1} \cup \Omega_{1/2}^{2,d+1}\big)$ and $\bigcap_{i=1}^d \Omega_i \cap \Omega_{d+1}$ only differ by a set of probability zero, and, hence, (4.10) follows from (4.11). □

As partition $\big\{A_0(B,B^*), A_{1/2}(B,B^*), A_1(B,B^*)\big\}$ of $\mathbb{R}_+^d$ we suggest

$$\begin{aligned}
\Big\{ A_0(B,B^*) &= \bigcup_{i\in V} \big[\{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i < \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k \} \cup \{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} x_k > \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k \} \big], \\
A_{1/2}(B,B^*) &= \bigcap_{i\in V} \big[\{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i = \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} x_k \} \cup \{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i > \bigvee_{k\in \mathrm{pa}(i)} (b_{ki} \vee b^*_{ki}) x_k \} \big], \\
A_1(B,B^*) &= \mathbb{R}_+^d \backslash \big(A_0(B,B^*) \cup A_{1/2}(B,B^*) \big) \Big\}.
\end{aligned}$$

With this partition we then have:

{ch4:dens:gener

Theorem 4.8. *Let $B, B^* \in \mathcal{B}(\mathcal{D})$. Then the function $\rho : \mathbb{R}_+^d \to \{0, 1/2, 1\}$*

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{x}, B, B^*) = \frac{1}{2} \cdot \mathbb{1}_{A_{1/2}(B,B^*)}(\boldsymbol{x}) + \mathbb{1}_{A_1(B,B^*)}(\boldsymbol{x}) = \begin{cases} 0, & \textit{if } \boldsymbol{x} \in A_0(B,B^*), \\ \frac{1}{2}, & \textit{if } \boldsymbol{x} \in A_{1/2}(B,B^*), \\ 1, & \textit{if } \boldsymbol{x} \in A_1(B,B^*), \end{cases} \tag{4.13}$$

{ch4:dens:gen1}

is a density of P_B with respect to $P_B + P_{B^*}$.

Proof. We must verify properties (A)–(C) of (4.5).
(A) Since V is finite, it suffices to show for every $i \in V$,

$$P_B\big(\{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i < \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k \}\big) = \mathbb{P}\big(X_i < \bigvee_{k\in \mathrm{pa}(i)} b_{ki} X_k\big) = 0, \tag{4.14}$$

{ch4:gmle:eq5}

$$P_B\big(\{ \boldsymbol{x} \in \mathbb{R}_+^d : x_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} x_k > \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k \}\big) = \mathbb{P}\big(X_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} X_k > \bigvee_{k\in \mathrm{pa}(i)} b_{ki} X_k\big) = 0.$$

The former is immediate by (4.9). By the same argument we have for the latter,

$$\begin{aligned}
0 &\leqslant \mathbb{P}\big(\bigvee_{k\in \mathrm{pa}(i)} b_{ki} X_k \vee Z_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} X_k > \bigvee_{k\in \mathrm{pa}(i)} b_{ki} X_k \big) = \mathbb{P}\big(Z_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} X_k > \bigvee_{k\in \mathrm{pa}(i)} b_{ki} X_k \big) \\
&\leqslant \mathbb{P}\big(Z_i = \bigvee_{k\in \mathrm{pa}(i)} b^*_{ki} \bigvee_{j\in \mathrm{An}(k)} b_{jk} Z_j \big) = 0,
\end{aligned}$$

where we have used (2.3) and (3.1) for the last inequality and equality, respectively. Thus we have verified (A).

(B) Recall that P_B and P_{B*} share the same innovation vector when represented by a recursive ML model. Furthermore, note that the set $\Omega(B, B^*)$ from Lemma 4.7 is a subset of $\bigcap_{i\in V}\big\{X_i = \bigvee_{j\in\mathrm{An}(i):b_{ji}=b^*_{ji}} b_{ji}Z_j\big\}$. We have $\Omega(B, B^*) = \Omega(B^*, B)$ and hence we obtain from (4.10) for $A \in \mathbb{B}(\mathbb{R}^d_+)$,

$$\begin{aligned} P_B(A \cap A_{1/2}(B, B^*)) &= \mathbb{P}(\{\boldsymbol{X} \in A\} \cap \Omega(B, B^*)) = \mathbb{P}\big(\big\{\big(\bigvee_{j\in\mathrm{An}(i):b_{ji}=b^*_{ji}} b_{ji}Z_j, i \in V\big) \in A\big\} \cap \Omega(B, B^*)\big) \\ &= \mathbb{P}\big(\big\{\big(\bigvee_{j\in\mathrm{An}(i):b_{ji}=b^*_{ji}} b^*_{ji}Z_j, i \in V\big) \in A\big\} \cap \Omega(B^*, B)\big) = P_{B*}(A \cap A_{1/2}(B, B^*)). \end{aligned}$$

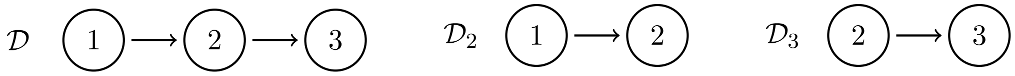
(C) We observe from the definition of $A_0(B, B^*)$ and $A_{1/2}(B, B^*)$ that

$$\begin{aligned} &A_1(B, B^*) = \mathbb{R}^d_+ \backslash \big(A_0(B, B^*) \cup A_{1/2}(B, B^*)\big) \\ &\subseteq \bigcup_{i\in V} \big[\big\{\boldsymbol{x} \in \mathbb{R}^d_+ : \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k > x_i \geqslant \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^d_+ : x_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k > \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k\big\}\big] \\ &\subseteq A_0(B^*, B). \end{aligned}$$

Since $A_0(B^*, B)$ is a P_{B*}-null set by (A), this holds for the subset $A_1(B, B^*)$ as well. □

We observe an interesting relation between the density (4.13) for $\mathcal{D}$ and corresponding densities for subgraphs of $\mathcal{D}$.

Example 4.9. [Local densities ρ_i] {ch4:exam:margi
Consider the DAGs



Let ρ, ρ_2, and ρ_3 be the corresponding densities from (4.13). For the ML coefficient matrix B of a recursive ML model on $\mathcal{D}$, let B_2 and B_3 be the ML coefficient matrices of recursive ML models on $\mathcal{D}_2$ and $\mathcal{D}_3$ with edge weight $c_{12} = b_{12}$ and $c_{23} = b_{23}$, and let starred quantities denote the same for B^*. We then find for $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3_+$,

$$\begin{aligned} &\rho(\boldsymbol{x}, B, B^*) \\ &= \big(\rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)}, B_2, B^*_2) \vee \rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)}, B_3, B^*_3)\big)\mathbb{1}_{(0,\infty)}\big(\rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)}, B_2, B^*_2) \wedge \rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)}, B_3, B^*_3)\big). \end{aligned}$$

This can be observed from Figure 4.3, where the densities are depicted as functions of x_2/x_1 and/or x_3/x_2 for all nine different orders between the ML coefficients in B and B^*.

Conversely, ρ_2 and ρ_3 can be derived from ρ as follows:

$$\begin{aligned} \rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)}, B_{12}, B^*_{12}) &= \min_{\{y\in\mathbb{R}_+:\rho((\boldsymbol{x}_{\mathrm{Pa}(2)},y),B,B^*)>0\}} \rho((\boldsymbol{x}_{\mathrm{Pa}(2)}, y), B, B^*), \\ \rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)}, B_{23}, B^*_{23}) &= \min_{\{y\in\mathbb{R}_+:\rho((y,\boldsymbol{x}_{\mathrm{Pa}(3)}),B,B^*)>0\}} \rho((y, \boldsymbol{x}_{\mathrm{Pa}(3)}), B, B^*), \end{aligned}$$

which we learn from Figure 4.3 again. □

We extend the findings from Example 4.9 to the general case. Furthermore, we show that the densities ρ_i are densities of regular conditional distributions.

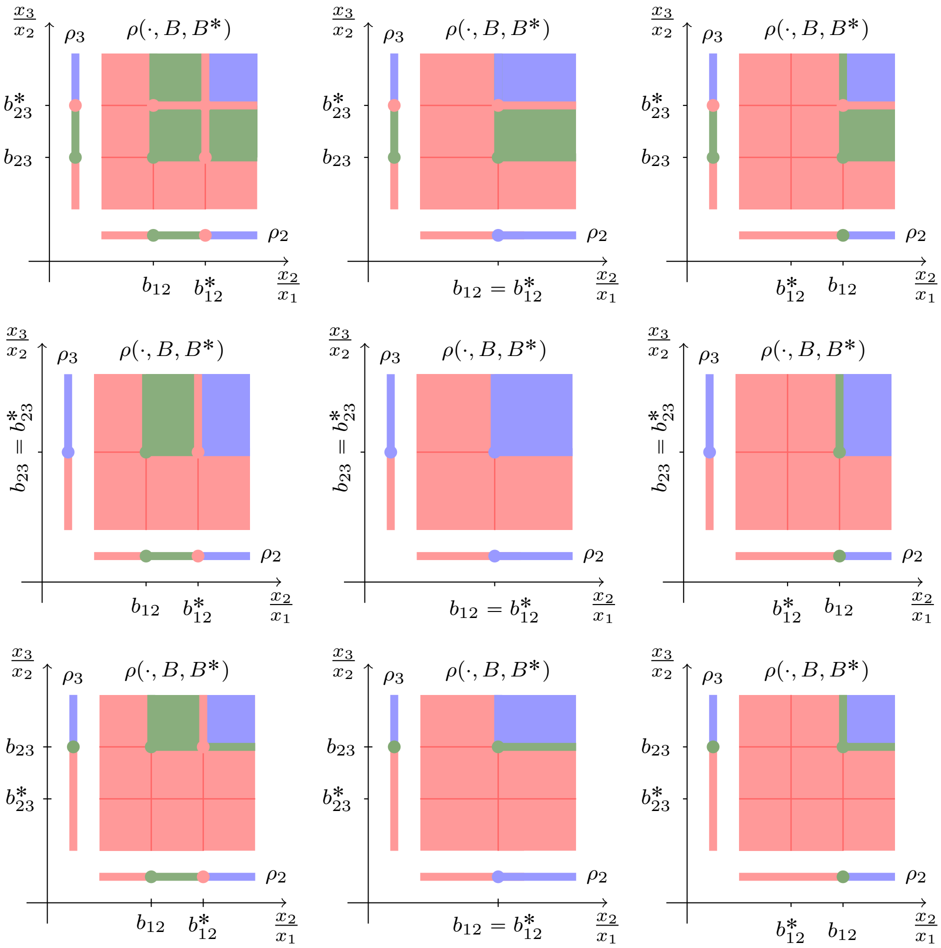


Figure 4.3: The densities $\rho(\boldsymbol{x}, B, B^*)$, $\rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)}, B_2, B_2^*)$, $\rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)}, B_3, B_3^*)$ from Example 4.9 as functions of x_2/x_1 and/or x_3/x_2. The area where the density is $0/\frac{1}{2}/1$ is coloured in red/blue/green. ch4:ex:fig:mar

{ch4:prop:rhoi}

Proposition 4.10. *Let $B, B^* \in \mathcal{B}(\mathcal{D})$ and let $\boldsymbol{X} = \boldsymbol{Z} \odot B, \boldsymbol{X}^* = \boldsymbol{Z} \odot B^*$ follow corresponding recursive ML models on $\mathcal{D}$. For $i \in V$, let ρ_i be the density given in (4.13) with respect to the DAG $\mathcal{D}_i = (\mathrm{Pa}(i), \{(k,i) : k \in \mathrm{pa}(i)\})$ as well as B_i and B_i^* the ML coefficient matrices of recursive ML models on $\mathcal{D}_i$ with edge weights $c_{ki} = b_{ki}$ and $c_{ki}^* = b_{ki}^*$, respectively.*

(a) *We have for $\rho(\boldsymbol{x}, B, B^*)$ given in (4.13)*

$$\rho(\boldsymbol{x}, B, B^*) = \Big(\bigvee_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*)\Big) \mathbb{1}_{(0,\infty)}\Big(\bigwedge_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*)\Big). \qquad (4.15)$$

{ch4:rhorhoi}

(b) *The function ρ_i can be computed from ρ by*

$$\rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) = \min_{\{\boldsymbol{y} \in \mathbb{R}_+^d : \boldsymbol{y}_{\mathrm{Pa}(i)} = \boldsymbol{x}_{\mathrm{Pa}(i)}, \rho(\boldsymbol{y}, B, B^*) > 0\}} \rho(\boldsymbol{y}, B, B^*),$$

where we set $\min_{\boldsymbol{y} \in \varnothing} \rho(\boldsymbol{y}, B, B^*) = 0$.

(c) *The function $\rho_i : \mathbb{R}_+^d \to \{0, 1/2, 1\}$ such that $\boldsymbol{x}_{\mathrm{Pa}(i)} \mapsto \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*)$ is a density of $P_B^{i|\mathrm{pa}(i)}$ with respect to $P_B^{i|\mathrm{pa}(i)} + P_{B^*}^{i|\mathrm{pa}(i)}$, where $P_B^{i|\mathrm{pa}(i)}$ is a regular conditional distribution of X_i given $\boldsymbol{X}_{\mathrm{pa}(i)}$ and $P_{B^*}^{i|\mathrm{pa}(i)}$ one of X_i^* given $\boldsymbol{X}_{\mathrm{pa}(i)}^*$.*

Proof. Denoting by $A_0^i(B_i, B_i^*)$, $A_{1/2}^i(B_i, B_i^*)$, $A_1^i(B_i, B_i^*)$ the sets defining $\rho_i(\cdot, B_i, B_i^*)$, we have for the corresponding sets of ρ,

$$\begin{aligned}
A_0(B, B^*) &= \bigcup_{i\in V} \big\{\boldsymbol{x} \in \mathbb{R}_+^d : \boldsymbol{x}_{\mathrm{Pa}(i)} \in A_0^i(B_i, B_i^*)\big\},\\
A_{1/2}(B, B^*) &= \bigcap_{i\in V} \big\{\boldsymbol{x} \in \mathbb{R}_+^d : \boldsymbol{x}_{\mathrm{Pa}(i)} \in A_{1/2}^i(B_i, B_i^*)\big\},\\
A_1(B, B^*) &= \bigcap_{i\in V} \big\{\boldsymbol{x} \in \mathbb{R}_+^d : \boldsymbol{x}_{\mathrm{Pa}(i)} \in A_{1/2}^i(B_i, B_i^*) \cup A_1^i(B_i, B_i^*)\big\} \cap \big[\mathbb{R}_+^d \backslash A_{1/2}(B_i, B_i^*)\big].
\end{aligned}$$

From this we obtain (a) and (b). Now, to see (c) we reason as follows:

$$P_B^{i|\mathrm{pa}(i)}\big((0, x_i] \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = F_{Z_i}(x_i)\mathbb{1}_{[\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k, \infty)}(x_i), \quad \boldsymbol{x}_{\mathrm{Pa}(i)} \in \mathbb{R}_+^{|\mathrm{Pa}(i)|},$$

is a regular conditional distribution function of X_i given $\boldsymbol{X}_{\mathrm{pa}(i)}$. To see this, use (4.9) and the independence of the innovations to obtain

$$\begin{aligned}
P_B^{i|\mathrm{pa}(i)}\big((0, x_i] \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) &= \mathbb{P}(X_i \leqslant x_i \mid \boldsymbol{X}_{\mathrm{pa}(i)} = \boldsymbol{x}_{\mathrm{pa}(i)})\\
&= \mathbb{P}\big(\bigvee_{k\in\mathrm{pa}(i)} b_{ki}X_k \vee Z_i \leqslant x_i \mid \boldsymbol{X}_{\mathrm{pa}(i)} = \boldsymbol{x}_{\mathrm{pa}(i)}\big)\\
&= F_{Z_i}(x_i)\mathbb{1}_{[\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k, \infty)}(x_i).
\end{aligned}$$

Since $\boldsymbol{X}$ and $\boldsymbol{X}^*$ share the same innovation vector, we have

$$P_{B^*}^{i|\mathrm{pa}(i)}\big((0, x_i] \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = F_{Z_i}(x_i)\mathbb{1}_{[\bigvee_{k\in\mathrm{pa}(i)} b_{ki}^*x_k, \infty)}(x_i), \quad \boldsymbol{x}_{\mathrm{Pa}(i)} \in \mathbb{R}_+^{|\mathrm{Pa}(i)|},$$

is a regular conditional distribution function of X_i^* given $\boldsymbol{X}_{\mathrm{pa}(i)}^*$. Figure 4.4 depicts the two conditional distribution functions for the three possible orders between $\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k$ and $\bigvee_{k\in\mathrm{pa}(i)} b_{ki}^*x_k$. It then suffices to show for all $\boldsymbol{x}_{\mathrm{pa}(i)} \in \mathbb{R}_+^{|\mathrm{pa}(i)|}$ and $y \in \mathbb{R}_+$,

$$P_B^{i|\mathrm{pa}(i)}\big((0, y] \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = \int_{(0,y]} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*)\big(P_B^{i|\mathrm{pa}(i)} + P_{B^*}^{i|\mathrm{pa}(i)}\big)(dx_i \mid \boldsymbol{x}_{\mathrm{pa}(i)}),$$

and for this again by definition of ρ_i (cf. (4.6) and the related discussion) that

$$\begin{aligned}
&P_B^{i|\mathrm{pa}(i)}\big((0, y] \cap \big(0, \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\big) \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = 0,\\
&P_B^{i|\mathrm{pa}(i)}\big((0, y] \cap \{\bigvee_{k\in\mathrm{pa}(i)} b_{ki}^*x_k\} \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = 0 \quad \text{if} \bigvee_{k\in\mathrm{pa}(i)} b_{ki}^*x_k > \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k,\\
&P_B^{i|\mathrm{pa}(i)}\big((0, y] \cap \{\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\} \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = P_{B^*}^{i|\mathrm{pa}(i)}\big((0, y] \cap \{\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\} \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \bigvee_{k\in\mathrm{pa}(i)} b_{ki}^*x_k = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k,\\
&P_B^{i|\mathrm{pa}(i)}\big((0, y] \cap \big(\bigvee_{k\in\mathrm{pa}(i)} (b_{ki} \vee b_{ki}^*)x_k, \infty\big) \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big) = P_{B^*}^{i|\mathrm{pa}(i)}\big((0, y] \cap \big(\bigvee_{k\in\mathrm{pa}(i)} (b_{ki} \vee b_{ki}^*)x_k, \infty\big) \mid \boldsymbol{x}_{\mathrm{pa}(i)}\big).
\end{aligned}$$

Since F_{Z_i} is atom-free, this can be read directly from Figure 4.4. □

We now show that $\widehat{B}$ is indeed a GMLE in the sense of [18]. Note also that the GMLE is obtained by piecing together individual GMLEs corresponding to conditional distributions

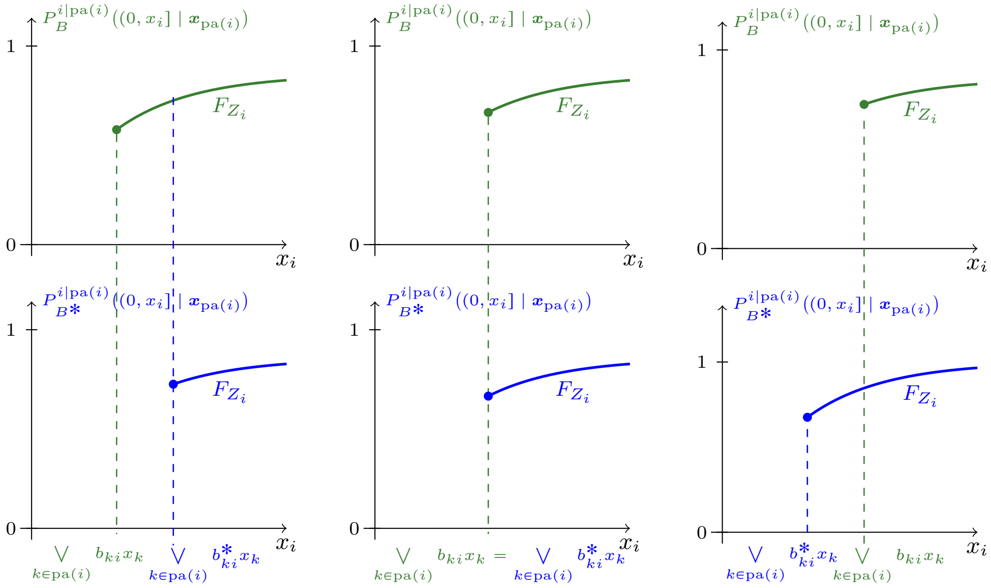


Figure 4.4: The conditional distribution functions from the proof of Proposition 4.10(c). ch4:fig:lemA4

of any variable given its parents. Thus this is similar to what is obtained in cases where the distributions have densities with respect to a product measure, as the maximum of the likelihood function is then obtained by maximizing each conditional likelihood function for the density of a node given its parents.

{ch4:the:gmle}

Theorem 4.11. *Let $\boldsymbol{x}^{(t)} = \big(x_1^{(t)}, \ldots, x_n^{(t)}\big)$ for $t = 1, \ldots, n$ be a sample from a recursive ML model on a DAG $\mathcal{D}$ with ML coefficient matrix $B \in \mathcal{B}(\mathcal{D})$ unknown.*

(a) The matrix $\widehat{B}$ from (4.3) *is a GMLE of B.*

(b) For every $i \in V$, $(\widehat{b}_{ki}, k \in \mathrm{pa}(i))$ is a GMLE of the ML coefficients $(b_{ki}, k \in \mathrm{pa}(i))$ of a random vector following a recursive ML model on $\mathcal{D}_i = (\mathrm{Pa}(i), \{(k,i) : k \in \mathrm{pa}(i)\})$ with edge weights $c_{ki} = b_{ki}$.

(c) For every $i \in V$ and $k \in \mathrm{pa}(i)$, $\widehat{b}_{ki}$ is the only GMLE of the ML coefficient b_{ki} of a random vector following a recursive ML model on $\mathcal{D}_{ki} = (\{k,i\}, \{(k,i)\})$ with edge weight $c_{ki} = b_{ki}$.

Proof. (a) First, recall that $\widehat{B}$ is indeed a ML coefficient matrix of a recursive ML model on $\mathcal{D}$. The first condition in the definition of a GMLE in (4.4) is satisfied due to the definition of $\rho(\cdot, \widehat{B}, \widehat{B})$ since $A_{1/2}(\widehat{B}, \widehat{B}) = \mathbb{R}_+^d$. Since the densities $\rho(\cdot, \widehat{B}, B)$ and $\rho(\cdot, B, \widehat{B})$ have the values 0, 1, 1/2, and $A_{1/2}(\widehat{B}, B) = A_{1/2}(B, \widehat{B})$, to verify the second condition in (4.4), it suffices to show that there is some realization $\boldsymbol{x}^{(t_1)} \in A_0(B, \widehat{B})$ whenever there is some realization $\boldsymbol{x}^{(t_2)} \in A_0(\widehat{B}, B)$; cf. Example 4.6, in particular (4.8). So let $\boldsymbol{x}^{(t_2)} \in A_0(\widehat{B}, B)$ for some $t_2 \in \{1, \ldots, n\}$. We find, for some $i \in V$, from the definition of $A_0(\widehat{B}, B)$ and the fact that $x_i^{(t)} \geqslant \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki} x_k^{(t)}$,

$$\boldsymbol{x}^{(t_2)} \in \Big\{\boldsymbol{x} \in \mathbb{R}_+^d : \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki} x_k < x_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki} x_k\Big\}.$$

Hence, $x_i^{(t_2)} = b_{ki} x_k^{(t_2)}$ for some $k \in \mathrm{pa}(i)$ with $\widehat{b}_{ki} < b_{ki}$. Let now $t_1 \in \{1, \ldots, n\}$ such that $\bigwedge_{s=1}^n y_{ki}^{(s)} = y_{ki}^{(t_1)}$. As $\widehat{b}_{ki} = \bigwedge_{s=1}^n y_{ki}^{(s)}$, we have $x_i^{(t_1)} < b_{ki} x_k^{(t_1)}$ implying that $\boldsymbol{x}^{(t_1)} \in A_0(B, \widehat{B})$. The statement in (b) is a consequence of (a), and (c) has already been shown in Example 4.6. □

Figure 4.5 illustrates the DAGs $\mathcal{D}_i$ in Theorem 4.11(b) or Proposition 4.10.

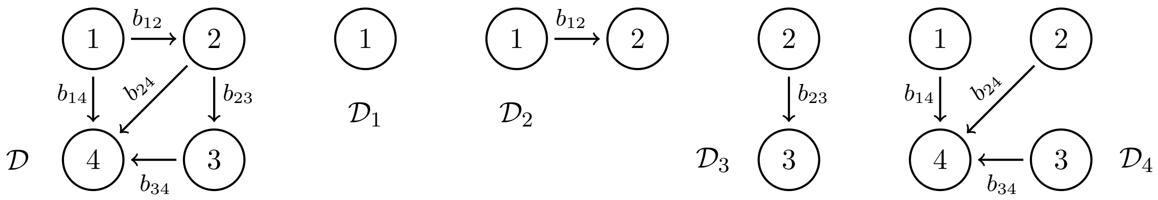


Figure 4.5: The DAGs $\mathcal{D}_i$ from Theorem 4.11(b) for a recursive ML model on the DAG $\mathcal{D}$ depicted on the left-hand side with ML coefficient matrix B. The edges are marked with the corresponding ML coefficients. Note that $b_{12}, b_{14}, b_{34}, b_{24}$ can be arbitary positive numbers but $b_{24} \geqslant b_{23}b_{34}$. ch4:exam:Di

Edge weights c_{ki}

We have started with the estimation of B as it is not possible to recover the true edge weights c_{ki} underlying representation (2.1) of $\boldsymbol{X}$ from $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, since different edge weights may lead to B. But we know what edge weights that are and, obviously, the probability measure P_B induced by $\boldsymbol{X}$ is the same for different edge weights that all result in B. As a consequence, all edge weights that lead, together with $\mathcal{D}$, to the GMLE $\widehat{B}$ of B are GMLEs of the true edge weights of $\boldsymbol{X}$ and its properties are inherited.

{ch4:co:edges}

Corollary 4.12. *Let c_{ki} for $i \in V$ and $k \in \mathrm{pa}(i)$ be the edge weights of representation* (2.1) *of $\boldsymbol{X}$ and $\mathcal{D}^{\widehat{B}}$ the minimum ML DAG based on $\widehat{B}$. We denote by $\mathrm{pa}^{\widehat{B}}(i)$ the parents of i in $\mathcal{D}^{\widehat{B}}$. Then every $(\widehat{c}_{ki}, i \in V, k \in \mathrm{pa}(i))$ such that*

$$\widehat{c}_{ki} = \widehat{b}_{ki} \;\; \textit{if } k \in \mathrm{pa}^{\widehat{B}}(i) \quad \textit{and} \quad \widehat{c}_{ki} \in (0, \widehat{b}_{ki}] \;\; \textit{if } k \in \mathrm{pa}(i) \backslash \mathrm{pa}^{\widehat{B}}(i)$$

is a GMLE of $(c_{ki}, i \in V, k \in \mathrm{pa}(i))$.

5 Learning the structure of a recursive max-linear model

{ch4:s6}

In contrast to the assumptions in the previous section, we now assume independent realizations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ of $\boldsymbol{X}$ following a recursive ML model but the underlying DAG $\mathcal{D}$ is unknown. We know from previous discussions that it is not possible to recover $\mathcal{D}$ and the true edge weights c_{ki} but, based on Corollary 4.12, it is possible to identify the ML coefficient matrix B and the class of all DAGs and edge weights that could have generated $\boldsymbol{X}$ via (2.1). We therefore again focus on the estimation of B.

Following Algorithm 3.3, it suffices for any pair of distinct $i, j \in V$ to decide whether $\mathrm{supp}(Y_{ji}) = \mathrm{supp}(X_j/X_i)$ has a positive lower bound, alternatively a finite upper bound, and if so, to estimate the bound. Recall from Table 3.1 that, if there is such a bound, then it is an atom of Y_{ji}. Since we can expect to observe atoms more than twice for n sufficiently large, we propose the following estimation method.

Algorithm 5.1. [Find an estimate $\breve{B}$ of B from $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$] {ch4:alg4}

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

2. For all $i, j \in V$ with $i \neq j$,

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

$$\text{else, set } \breve{b}_{ji} = 0.$$

In the second step rather two steps are summarized. The first step is concerned with estimating the ancestors of the nodes, the second with estimating the ML coefficients.

Note that the estimate $\breve{B}$ from Algorithm 5.1 is not necessarily a ML coefficient matrix of a recursive ML model. For example, the property that $b_{ji} > 0$ if $b_{jk}b_{ki} > 0$ (see, for example, Corollary 3.12 of [12]) is not guaranteed. Many modifications of $\breve{B}$ are possible, and here we shall not discuss this in detail. Rather we notice that Algorithm 5.1 outputs, $\mathbb{P}$-almost surely, the true ML coefficient matrix B if n is sufficiently large. As in the case where the DAG is known — see Proposition 4.5 — the probability that $\breve{b}_{ji}$ is equal to the true value b_{ji} converges to one at an exponential rate.

6 Conclusion and outlook {ch4:s7}

We studied the identifiability of the elements of a recursive ML model from the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$. The associated DAG and the edge weights are not identifiable, however, the ML coefficient matrix B. In other words, we can identify representation (2.3) but not (2.1). The class of all DAGs and edge weights that could have generated $\boldsymbol{X}$ via (2.1) and the distribution of the innovation vector are identifiable from $\mathcal{L}(\boldsymbol{X})$. As a consequence, we can recover B, the class of the DAGs and edge weights, and the innovation distributions from realizations of $\boldsymbol{X}$.

Parameter estimation and structure learning for recursive ML models seem to be challenging tasks because assumptions usually made in standard methods are not met. However, in both cases, B can be estimated by a simple procedure. The key idea of our approach is to consider the observed ratios between any pair of components, i.e. to perform a transformation on the realizations. The transformed realizations or rather the distributional properties of the corresponding random variables make it possible to identify, with probability 1, the true B whenever the number of observations n is sufficiently large. It would be interesting to investigate the relationship between the performance of our procedures and the number n of observations. Here, one possible question is how many observations are at least necessary to estimate B exactly; see, Example 4.3. In addition it would be interesting to study estimation of the DAG structure for moderate sample sizes, where exact estimation is not guaranteed.

We emphasize again that, although our estimates are derived under the assumption that the distribution of the innovation vector $\boldsymbol{Z}$ is fixed, the estimates do not depend on what this distribution is and would therefore also be valid in the situation where the innovations are independent with unkown distributions that are atom-free and have support equal to $\mathbb{R}_+$. Algorithm 3.4 provides a recursive procedure to obtain the distribution functions F_{Z_i} from B and the marginal distribution functions F_{X_i} of X_i. Estimating B by $\widehat{B}$ and the distributions F_{X_i}, for example, by their empirical versions, we can apply this procedure to find estimators of the distributions F_{Z_i} although it will formally violate the assumption of atom-freeness and thus it is both more efficient and formally correct to estimate these parametrically, or under suitable monotonicity restrictions.

An important goal for future work is to apply the procedures to real-world data. However, it is unreasonable to expect any non-simulated data to follow a recursive ML model exactly and the model should then be modified by adding appropriate noise terms. In particular we should not expect that we observe a minimal observed ratio more than twice, as we exploit in Algorithm 5.1. It seems to be more reasonable to expect values close to each other. We therefore want to develop methods based on accumulation points.

Acknowledgements

We thank Justus Hartl for providing a first discussion about the different estimators suggested in this paper in his master's thesis. NG acknowledges support by Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE). All authors benefited from financial support from the Alexander von Humboldt Stiftung.

References

[1] P. Asadi, A. C. Davison, and S. Engelke. Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050, 2015.

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

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

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

[5] R. A. Davis and W. P. McCormick. Estimation for first-order autoregressive processes with positve or bounded innovations. *Stoch. Proc. Appl.*, 31:237–250, 1989.

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

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

[8] J. H. J. Einmahl, A. Kiriliouk, and J. Segers. A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 21(2):205–233, 2018.

[9] S. Engelke and A. Hitz. Graphical models for extremes. arXiv:1812.01734, 2018.

[10] R. D. Gill, J. A. Wellner, and J. Præstgaard. Non-and semi-parametric maximum likelihood estimators and the von-Mises method (part 1). *Scandinavian Journal of Statistics*, 16(2):97–128, 1989.

[11] N. Gissibl. *Graphical Modeling of Extremes: Max-linear Models on Directed Acyclic Graphs.* PhD thesis, Technical University of Munich, 2018.

[12] N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A): 2693–2720, 2018.

[13] 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?*, pages 171–189. Springer VS, Wiesbaden, 2017.

[14] A. Hitz and R. Evans. One-component regular variation and graphical modeling of extremes. *J. Appl. Prob.*, 53:733–746, 2016.

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

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

[17] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.

[18] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956.

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

[20] S. L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, United Kingdom, 1996.

[21] S. L. Lauritzen. Causal inference from graphical models. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg, editors, *Complex Stochastic Systems*, pages 63–107. Chapman and Hall/CRC Press, London/Boca Raton, 2001.

[22] 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.

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

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

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

[26] F. W. Scholz. Towards a unified definition of maximum likelihood. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 8(2):193–203, 1980.

[27] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2000.