Technische Universität München
Zentrum Mathematik
Lehrstuhl für Mathematische Statistik

Graphical modeling of extremes: Max-linear models on directed acyclic graphs

Nadine Manuela Gissibl

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Prof. Dr. Noam Berger Steiger
Prüfer der Dissertation:	1.	Prof. Dr. Claudia Klüppelberg
	2.	Prof. Dr. John H.J. Einmahl (Universität Tilburg) (nur schriftliche Beurteilung)
	3.	Prof. Dr. Steffen L. Lauritzen (Universität Kopenhagen)

Die Dissertation wurde am 06.06.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 07.08.2018 angenommen.

Abstract

Graphical modeling has mainly been limited to discrete and Gaussian distributions, distributions that lead to severe underestimation of large risks and, therefore, to unsuitable models in the context of risk assessment. This thesis deals with the development and investigation of a class of graphical models that finds its application in situations where extreme risks play an essential role and may propagate through a network, for example, when modeling water levels or pollution concentrations in a river or when modeling risks in a large industrial structure.

We use the concept of structural equation modeling to introduce the class of recursive max-linear models. The causal structure of a recursive max-linear model is represented by a directed acyclic graph, and the node variables are max-linear functions of their parental node variables and independent noise variables. Natural candidates for the noise distributions are extreme value distributions or distributions in their domains of attraction resulting in a corresponding multivariate distribution.

First, we study structural properties of recursive max-linear models. Different directed acyclic graphs and weights in the max-linear structural equations may lead to the same recursive max-linear model; but all of them lead to the same max-linear representations of the model in terms of the noise variables. We characterize these graphs and weights and point out the minimum directed acyclic graph that represents the causal structure of the model. In particular, we address the relation between the weights in the structural equations and the coefficients in the max-linear representations in detail. Further, we give necessary and sufficient conditions on a max-linear model to be a recursive max-linear model. Throughout we exploit the natural orders between the node variables and between the max-linear coefficients, for example, to obtain reduced model representations.

In the second part of the thesis, we assume regularly varying noise variables leading to extremal dependence between the components of a recursive max-linear model. The focus is on the matrix of pairwise tail dependence coefficients, which measure the extremal dependence between two random variables. Motivated by the fact that a multivariate Gaussian distribution is completely determined by its mean and its covariance matrix, we investigate how far the coefficients of a recursive max-linear model and its underlying graph can be recovered from its tail dependence matrix. For example, the associated minimum graph is identifiable from the tail dependence matrix and a causal ordering of any associated graph. We present a procedure that, given a tail dependence matrix of a recursive max-linear model, finds all recursive max-linear models with this tail dependence matrix.

In the last part, we consider the identifiability and estimation of recursive max-linear models. We show that the max-linear coefficient matrix and the noise distributions can be identified from the distribution of a recursive max-linear model. To infer these quantities from observational

data, we cannot apply standard methods as the assumptions usually made are not satisfied. However, we can use the distributional properties of the ratio between two components to find, with probability 1, the true max-linear coefficient matrix exactly, provided the number of observations is sufficiently large. An estimate we suggest if the true underlying graph is known has the same property. We prove that this estimate can be considered a maximum likelihood estimate in an extended definition.

Zusammenfassung

Die graphische Modellierung beschränkt sich bisher hauptsächlich auf diskrete und Gaußsche Verteilungen, Verteilungen, die zu einer starken Unterschätzung großer Risiken und damit zu ungeeigneten Modellen im Rahmen der Risikobewertung führen. Diese Arbeit beschäftigt sich mit der Entwicklung und Untersuchung einer Klasse von graphischen Modellen, die ihre Anwendung in Situationen findet, in denen extreme Risiken eine wesentliche Rolle spielen und sich über ein Netzwerk ausbreiten können. Beispiele sind die Modellierung von Wasserständen oder Schadstoffkonzentrationen in einem Fluss oder von Risiken in einer großen industriellen Struktur.

Wir verwenden das Konzept der Strukturgleichungsmodellierung, um die Klasse rekursiver max-linearer Modelle einzuführen. Die kausale Struktur eines rekursiven max-linearen Modells wird durch einen gerichteten azyklischen Graphen repräsentiert. Die Knotenvariablen sind max-lineare Funktionen der elterlichen Knotenvariablen und unabhängiger Fehlerterme. Natürliche Kandidaten für die Fehlertermverteilungen sind Extremwertverteilungen oder Verteilungen in deren Anziehungsbereichen, die zu einer entsprechenden multivariaten Verteilung führen.

Zunächst untersuchen wir strukturelle Eigenschaften rekursiver max-linearer Modelle. Unterschiedliche gerichtete azyklische Graphen und Gewichte in den max-linearen Strukturgleichungen können zum selben rekursiven max-linearen Modell führen; aber alle von ihnen führen zu den gleichen max-linearen Darstellungen des Modells bezüglich der Fehlerterme. Wir charakterisieren diese Graphen und Gewichte und heben den minimalen gerichteten azyklischen Graphen, der die kausale Struktur des Modells repräsentiert, besonders hervor. Insbesondere beleuchten wir ausführlich den Zusammenhang zwischen den Gewichten in den Strukturgleichungen und den Koeffizienten in den max-linearen Darstellungen. Ferner geben wir notwendige und hinreichende Bedingungen für ein max-lineares Modell, ein rekursives max-lineares Modell zu sein, an. Durchgehend nutzen wir die natürlichen Ordnungen zwischen den Knotenvariablen und zwischen den max-linearen Koeffizienten aus, um beispielsweise reduzierte Modelldarstellungen zu erhalten.

Im zweiten Teil der Arbeit gehen wir von regelmäßig variierenden Fehlertermen aus, die zu einer extremalen Abhängigkeit zwischen den Komponenten eines rekursiven max-linearen Modells führen. Der Fokus liegt auf der Matrix der paarweisen Tail-Dependence-Koeffizienten, welche die extremale Abhängigkeit zwischen zwei Zufallsvariablen messen. Motiviert durch die Tatsache, dass multivariate Gauß-Verteilungen vollständig durch ihren Erwartungswert und ihre Kovarianzmatrix bestimmt sind, untersuchen wir, wie weit sich die Koeffizienten eines rekursiven max-linearen Modells und sein zugrunde liegender Graph aus seiner Tail-Dependence-Matrix bestimmen lassen. Beispielsweise ist der zugehörige Minimalgraph aus der Tail-Dependence-Matrix und einer kausalen Ordnung eines jeden zugehörigen Graphen identifizierbar. Wir stellen ein

Verfahren vor, welches für eine gegebene Tail-Dependence-Matrix eines rekursiven max-linearen Modells alle rekursiven max-linearen Modelle mit dieser Tail-Dependence-Matrix findet.

Im letzten Teil beschäftigen wir uns mit der Identifizierbarkeit und der Schätzung rekursiver max-linearer Modelle. Wir zeigen, dass die max-lineare Koeffizientenmatrix und die Fehlertermverteilungen anhand der Verteilung eines rekursiven max-linearen Modells identifiziert werden können. Um diese Größen aus Beobachtungsdaten zu gewinnen, können wir keine Standardmethoden anwenden, da die dabei üblicherweise getroffenen Annahmen nicht erfüllt sind. Wir können jedoch die Verteilungseigenschaften des Quotienten zweier Komponenten verwenden, um mit Wahrscheinlichkeit 1 die wahre max-lineare Koeffizientenmatrix zu finden, vorausgesetzt, die Anzahl der Beobachtungen ist hinreichend groß. Ein Schätzer, den wir empfehlen, wenn der wahre zugrunde liegende Graph bekannt ist, hat die gleiche Eigenschaft. Wir beweisen, dass dieser Schätzer in einer erweiterten Definition als Maximum-Likelihood-Schätzer angesehen werden kann.

Acknowledgements

First of all, I would like to thank my supervisor Claudia Klüppelberg for making this thesis, including all related experiences, possible and for introducing me to this interesting field of research. Working with her and conducting numerous fruitful discussions have been a pleasure. I have learned a lot from her, both scientifically and personally. She also gave me the chance to become a member of an interdisciplinary industrial project. I am extremely grateful for this opportunity. Moreover, I thank her for initiating a collaboration with Steffen Lauritzen. In this context, financial support from the Alexander von Humboldt Foundation is gratefully acknowledged.

Next, I want to express my gratefulness to Steffen Lauritzen, not only for new ideas for my work and for acting as a referee of this thesis. Our enlightening discussions and his constructive feedback helped me a lot.

Further, I thank all the current and former members of the Chair of Mathematical Statistics, for their company and many helpful discussions, but especially for the encouragement in the last months. Special thanks go to Andrea and Stephan for being so helpful; and to Sven for sharing the office with me and making my time in the office so enjoyable.

I also thank Justus Hartl, Moritz Otto, Zhongwei Zhang, and Mario Krali for their interest in my research topic and for working on it in their master's theses. Many interesting things came out of their work and resulted in publications.

The TUM International Graduate School of Science and Engineering (IGSSE) provided financial support to spend two months at the Seminar for Statistics at ETH Zurich, a time that has greatly benefited my research. I acknowledge further support from the International School of Applied Mathematics (ISAM) and the German LUFO IV/4 project 'SaMSys – Safety Management System in Order to Improve Flight Safety'.

Last but not least, I want to thank my family for their love, support, and encouragement, especially during my time as a doctoral candidate.

Contents

Chapter 1

Introduction

In this chapter we first introduce briefly the key concepts and methods we use subsequently in this thesis. We then outline the scope of the thesis and state the main results. Finally, we give an overview and summary of research topics and works on the model developed in this thesis.

1.1 General introduction and motivation

This thesis aims to develop a class of graphical models for modeling extreme events.

1.1.1 Graphical models

Probabilistic graphical models (*graphical models* for short) are a marriage between probability theory and graph theory and are a useful tool to reduce the complexity of multivariate statistical modeling. Profound introductions into graphical modeling can be found in Koller and Friedman [45] and Lauritzen [47]. Each node of the graph is identified with a random variable, and the edges in the graph are used to encode conditional independence relations between the random variables. So graphical models provide a simple way to visualize the structure of a probabilistic model, and model properties can be read off directly from the graph. It is therefore not astonishing that this rich class is widely used in various areas of application as, for example, in artificial intelligence, biology, decision support systems, engineering, finance, genetics, geology, and medicine (see e.g. Pourret et al. [59], the above textbooks and references therein).

Directed acyclic graphs

In this thesis we focus on *directed acyclic graphs (DAGs)* leading to *directed graphical models*, also called Bayesian networks. Figure 1.1.1 shows examples of graphs. The left graph is no DAG, since it has an undirected edge between 3 and 5; the middle graph is no DAG, since it has a (directed) cycle $1 \to 2 \to 4 \to 3 \to 1$; the right graph is a DAG.

Markov properties

A graphical model is directed if the distribution satisfies the *local Markov property* with respect to the DAG: each variable is conditionally independent of its non-descendants (excluding the parents) given its parents in the DAG (cf. Chapter 3.2 of [47]). Applying the local Markov to the DAG from Figure 1.1.1 yields the conditional independence relations $X_2 \perp\!\!\!\perp (X_3, X_5) \mid$

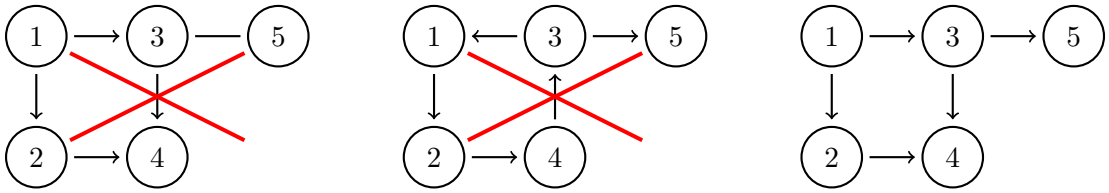


Figure 1.1.1: Example of graphs. We identify the nodes $1, \ldots, 5$ with the random variables $X_1, \ldots, X_5$. The right graph is a DAG: all edges are directed and there are no (directed) cycles.

$X_1,\ X_3 \perp\!\!\!\perp X_2 \mid X_1,\ X_4 \perp\!\!\!\perp (X_1, X_5) \mid (X_2, X_3)$, and $X_5 \perp\!\!\!\perp (X_1, X_2, X_4) \mid X_3$. All conditional independence relations that are implied by the local Markov property are encoded in the DAG via the *global Markov property*. The criterion of d-separation explains how these relations can be read off from the DAG (see Corollary 3.23 and Proposition 3.25 of [47] for precise definitions). If the distribution has a density with respect to a product measure, then the density factorizes according to the DAG (see second paragraph of Section 3.2.2 and Theorem 3.27 of [47]). This property is called *recursive factorization*. It is equivalent to the local and hence the global Markov property and is probably the most commonly used property to define directed graphical models.

Recursive structural equation models

Given a DAG, a *recursive structural equation model (recursive SEM)* is a multivariate statistical model where every random variable is associated with a node and can be written as a (measurable) function of its parents and an independent noise variable. The distributions of recursive SEMs satisfy, by construction, the local and hence the global Markov property with respect to the associated DAG (see Theorem 1.4.1 and the related discussion in Pearl [55]). Thus recursive SEMs offer a possibility to construct directed graphical models.

Causal models

Establishing and understanding cause-effect relations is an omnipresent desire in science and daily life. It is especially important when dealing with extreme risks, because knowing and understanding the causes of extreme events could help us to deal better with such events. Recursive SEMs play an important role in the field of causal inference; cf. Bollen [5], Pearl [55], and Spirtes et al. [69]. The causal structure of a recursive SEM is described by the associated DAG. Using DAGs has the advantage that parental variables can be considered to be direct causes of its children. As an example, the edge $1 \to 3$ in the DAG from Figure 1.1.1 reflects a (direct) causal influence of X_1 (*cause*) on X_3 (*effect*).

Structure learning

Assume we have observations of a graphical model or a recursive SEM and we want to adress the estimation of the DAG *(structure learning)*. We want to stress that in this thesis we assume that all variables are observed, that is, there are no hidden variables. Since, with respect to a given distribution, many different DAGs satisfy the Markov property (because of the equivalence, it

does not matter if we use the local or global Markov property), the Markov property is not enough to be able to recover the DAG from the observations. In addition to the Markov property, many common structure learning methods assume *faithfulness*, which means that the distribution is assumed to have no conditional independence relations except those represented in the DAG by the Markov property (cf. Sections 3.4.3, 3.5.2 of [69]). Two DAGs can be different but still entail the same conditional independence relations via the Markov property; such DAGs are called *Markov equivalent*. For a characterization of these DAGs, see e.g. Verma and Pearl [71]. Thus any method that assumes faithfulness and learns by observed conditional independence relations cannot distinguish between such DAGs and identifies Markov equivalence classes. A well-known example of such a method is the PC algorithm (Spirtes and Glymour [67]). So, without further assumptions, the graphs in the Markov equivalence classes cannot be distinguished.

Restricted recursive SEMs

However, if we put certain restrictions on the functions of recursive SEMs, their noise variables, or both, then for some of these classes the DAG is identifiable from the distribution. For the causal inference this would mean that, if the data follow a recursive SEM from such a restricted class of recursive SEMs and assuming that all variables are observed, the causal structure can be inferred from observational data only. Important research tasks for *restricted recursive SEMs* include the *identifiability* of the coefficients and the associated DAG from the distribution and the *structure learning* from a finite sample. The book by Peters et al. [58] provides a nice overview and introduction into this field of research and summarizes the current state of research. Recently, the identifiability theory was mainly elaborated for additive recursive SEMs with Gaussian noise (see e.g. Ernest et al. [21] and references therein).

1.1.2 Extreme value theory

Extreme value theory is concerned with the analysis and quantification of very rare and unusal events. Examples for such events include hurricanes, extreme wind gusts, floods, and heavy rainfall. They are from particular interest for society and industry as they are mostly dangerous and very costly.

Motivating example: extreme risks in networks

The development of extreme value theory and extreme value statistics has always been driven by applications. So the research presented in this thesis has been motivated by applications, more precisely by a technical risk analysis. Involved in an industrial project, we applied extreme value statistics to the safety of airplane landings and estimated the risk of serious incidents during an airplane landing (Gissibl et al. [29], Wang et al. [72]). One specifically risky event is the so-called runway overrun, which describes the fact that an airplane is unable to stop before the end of the runway. A case study can be found in Ayra [3]. As part of this projekt, the DAG shown in Figure 1.1.2 was developed. It shows the dependence structure, given by the DAG, between numerous physical quantities that may contribute to a runway overrun. Extraordinary values of

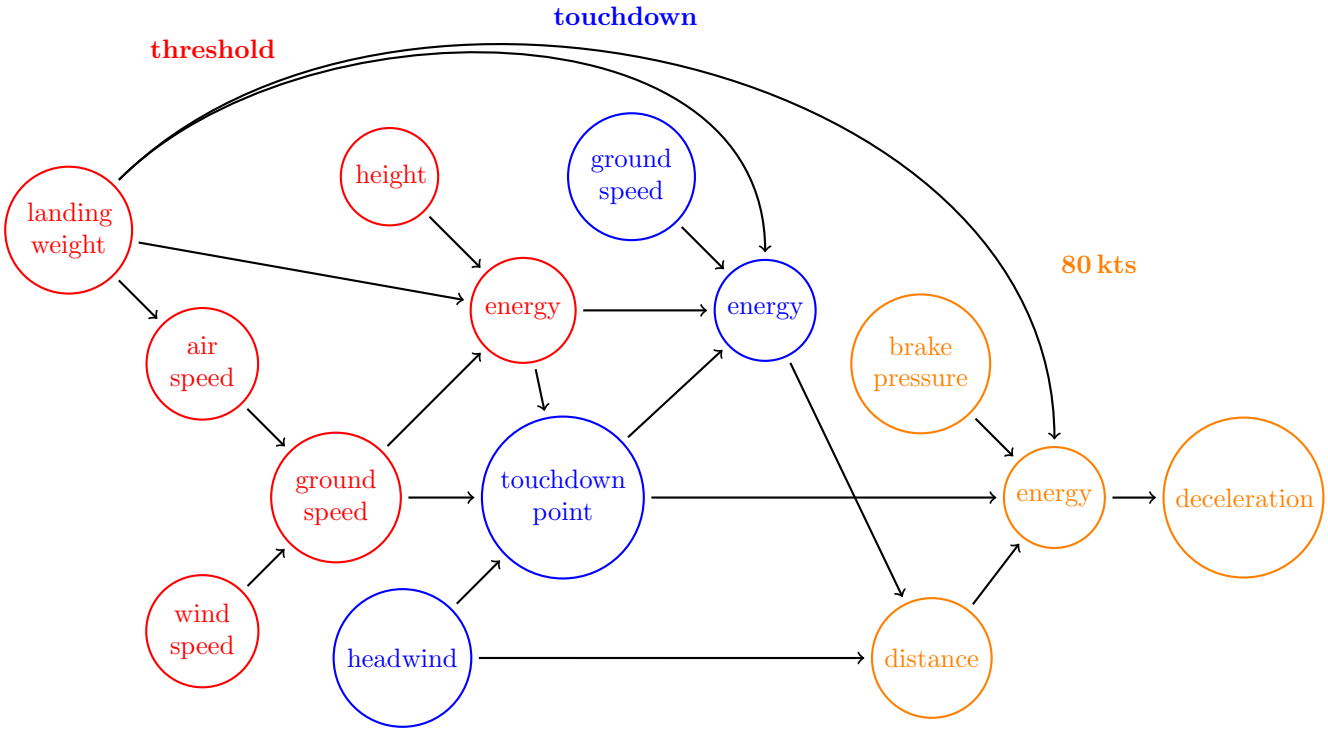


Figure 1.1.2: DAG describing the relationships between the physical quantities that may contribute to a runway overrun during an airplane landing.

some quantities may propagate through the DAG and lead invariably to a runway overrun. The thesis aims to develop models that are able to model such situations. To give another example, a river naturally forms a DAG. In the modeling of floods in rivers (Asadi et al. [2]), the extreme dependence structure, given by the DAG underlying the river, should be taken into account. Further examples include chemical pollution of rivers (Hoef et al. [35]), financial risk (Einmahl et al. [20]), and many others. Here graphical models appear as a natural class of models. So far, however, they are less suitable for modeling extreme events.

Graphical models mainly underestimate extreme risks

Despite the broad scope of applicability, graphical modeling of continuous random variables has mainly been limited to Gaussian distributions; see e.g. [45, 47]. In the context of risk assessment, risk variables are usually modeled by continuous variables; however, it has been known for a long time that Gaussian models in almost all cases underestimate extreme events severely. Thus there is a need for graphical models that do not underestimate extreme risks.

We now introduce the key concepts of extreme value theory needed in this thesis, but would like to mention that the focus of this thesis is on the graphical modeling with its associated concepts and only basic concepts of extreme value theory are used. Detailed introductions into extreme value theory are given in de Haan and Ferreira [14], Beirlant et al. [4], and Resnick [60, 61].

In extreme value theory *max-stable distributions* occur as limit distributions of normalized maxima. We deal with a conceptually simple but useful class of distributions whose max-stable distributions have a special property.

Max-linear (ML) models

A *max-linear (ML) model* is a multivariate probabilistic model where every component is a max-linear function of independent random variables. ML models are a natural extremal analogue of linear models. The random variables are usually assumed to be standard 1-Fréchet distributed. Similar to Wang and Stoev [73], we generally allow random variables with support $\mathbb{R}_+$. Within the class of max-stable distributions, the *spectral measure* of a max-linear model, which describes the dependence structure, is discrete. Conversely, every max-stable random vector with discrete spectral measure is max-linear. Another interesting result is that every max-stable distribution can be approximated arbitrarily well via a ML model (e.g. Yuen and Stoev [75], Section 2.2). ML models have been investigated, generalized, and applied to real world problems by many researchers; see e.g. Cui and Zhang [11], Einmahl et al. [19], Falk et al. [23], Kiriliouk [41], Schlather and Tawn [64], Strokorb and Schlather [70], and [73].

Tail dependence coefficients

The dependence structure of max-stable distributions is described by rather complex measures such as the *exponent measure*, the spectral measure, the *stable tail dependence function*, and the *Pickand's dependence function*. This complexity makes it difficult to estimate them, see e.g. [17, 19] and the references therein. Therefore, simpler extremal dependence measures are often considered. In this thesis we consider the *(upper) tail dependence coefficient* between two random variables, which goes back to Sibuya [66]. It is, roughly speaking, the probability of observing a large value in one variable provided that a large value has been observed for the other variable. The tail dependence coefficient is a special case of the *extremogram* (Davis and Mikosch [12]), which is a natural extremal analogue of the correlation function for stationary processes. It finds its application in many situations. One problem which is addressed for tail dependence coefficients is, for example, the construction of max-stable distributions with given tail dependence coefficients (see e.g. Falk [22] and [23, 64, 70]). Note that the tail dependence coefficient is not only defined for max-stable distributions, but is only meaningful if the distribution is heavy-tailed. Therefore, when working with tail dependence coefficients, we require corresponding distributions for the random variables of the ML models.

1.2 Scope and goals of this thesis

The main goal of this thesis is to develop a statistical model that can be applied in a variety of different areas and in situations where extreme risks play an essential role and may propagate through a network. To this end, we use the concept of structural equation modeling and define the class of *recursive max-linear (ML) models*.

Recursive ML models

A recursive ML model on a DAG $\mathcal{D}$ is a SEM where every random variable is a max-linear function of the parental node variabes in $\mathcal{D}$ and an independent noise variable with support $\mathbb{R}_+ = [0, \infty)$. The precise mathematical definition can be found in Section 2.1. We may think of the positive weights in the max-linear structural equations as relative quantities reflecting that a risk may originate with certain proportions in its different ancestors.

In this thesis we address different research problems for this model class.

Chapter 2: Max-linear models on directed acyclic graphs

In Chapter 2 we shed light on the structural properties of a recursive ML model. In the subsequent chapters, we use these properties extensively.

Almost all results we present in this chapter are based on the fact that there is no cancellation on $\mathbb{R}_+ = [0, \infty)$ with respect to the maximum $\vee$. This means that for some $a, b, c \in \mathbb{R}_+$, $a \vee c = b \vee c$ but $a \neq b$. In contrast, the addition, for example, has this property, i.e., for all $a, b, c \in \mathbb{R}_+$, if $a + c = b + c$, then $a = b$. That this property does not hold for the maximum leads to a complexity reduction of the model in many ways what we would like to discuss in the following.

The actual central property of a recursive ML model is the max-linearity in terms of its noise variables. The corresponding max-linear (ML) coefficients can be obtained by a path analysis of $\mathcal{D}$; more precisely, the computation of the ML coefficients corresponds to the algebraic path problem over the max-times semiring $(\mathbb{R}_+, \vee, \cdot)$ (see e.g. Mahr [51] and Rote [62]). Each path is assigned a weight and that is the product of the edge weights along the path. The problem consists then in finding a path between two nodes having maximum weight. The most well-known problem of this kind is probably the shortest path problem. What the product is in our problem, is there the sum, and the minimum corresponds to the maximum. So only the *max-weighted paths* are relevant for the ML model representation of a recursive ML model. As a consequence, a polytree represents the max-linear structural equation and the max-linear model representation of a component of a recursive ML model. With this, we often find further conditional independence relations that are not entailed by the (global) Markov property applied to its DAG. Hence, a recursive ML model is generally not faithful.

Again because of the above property of the maximum, different DAGs and edge weights may define the same recursive ML model. However, all of them lead to the same ML coefficient matrix and can be computed from it. We characterize these DAGs and edge weights. Here the minimum DAG representing the causal structure plays an important role. This result has direct implications for the identifiability and the estimation of a recursive ML model: the true DAG and the true weights of the max-linear structural equations cannot be recovered in general.

We also specify necessary and sufficient conditions on a matrix to be the ML coefficient matrix of a recursive ML model, both in the case where the associated DAG is given and in the case where it is arbitrary. Using the matrix product over $(\mathbb{R}_+, \vee, \cdot)$, the ML coefficient matrix is in both cases the solution of a fixed point equation.

In the last part of this chapter, given a set of node variables of a recursive ML model, we

investigate which insights can be gained about the others. We present a minimal subset of the given node variables that provides the same information.

Chapter 3: Tail dependence of recursive max-linear models with regularly varying noise variables

In Chapter 3 we assume regularly varying noise variables (cf. Section 3.1.1 below). This leads to models treated in classical multivariate extreme value theory, and we may have extremal dependence between two components of a recursive ML model. The latter would not be the case if we did not assume heavy-tailed noise variables, and the matrix of pairwise tail dependence coefficients, further referred to as *tail dependence matrix*, would be meaningless. The question of identifiability of restricted SEMs and the fact that a multivariate Gaussian distribution is completely defined by its mean and its covariance matrix motivated us to study the identifiability of a recursive max-linear model from its tail dependence matrix in this heavy-tailed setting.

We know from above that we cannot identify either its true DAG or its true weights in the max-linear structural equations. It is also not possible to recover the ML coefficient matrix. In fact, uncountably many recursive ML models with arbitrary index of regular variation have the same tail dependence matrix. However, we show that the DAGs representing the max-linear structural equations are identifiable from the tail dependence matrix and some additional information on the associated DAG such as the reachability matrix (i.e., the matrix whose ij-th entry is one if $i = j$ or i is an ancestor of j and zero else) or only a causal ordering (i.e., a permutation σ on $V = \{1, \dots, d\}$ such that $\sigma(j) < \sigma(i)$ for all i and their ancestors j).

We call a recursive ML model *max-weighted* if all paths are max-weighted. Because of its simple structure, this subclass of recursive ML models plays a special role, not only in this chapter. Here we can recover the DAGs from the tail dependence matrix and the initial nodes of the associated DAG. Further, we present necessary and sufficient conditions on a matrix to be the tail dependence matrix of a recursive max-weighted model on a given DAG.

We propose an algorithm that, given a tail dependence matrix of a recursive ML model, finds all recursive ML models with this tail dependence matrix. We also develop such a procedure especially for the subclass of recursive max-weighted models.

Another interesting problem we address is how DAGs of recursive ML models with the same tail dependence matrix relate to each other. For example, an initital node in a DAG of a recursive max-weighted model is again an initial node in a DAG of a recursive max-weighted model with the same tail dependence matrix or it is a terminal node.

Chapter 4: Identifiability and estimation of recursive max-linear models

In Chapter 4 we study the identifiability and estimation of recursive ML models. We relax again the assumptions on the noise variables and assume here independent and atomfree noise variables with support $\mathbb{R}_+$. In risk settings, which we have in mind when thinking of possible applications, it is natural to require the noise variables to have positive infinite support and atomfree distributions.

In contrast to the identifiability problem from Chapter 3, given the distribution of a recursive ML model, the ML coefficient matrix and hence the class of DAGs and edge weights defining the recursive ML model can be recovered; furthermore, the noise distributions are identifiable. To explain this easily and convincingly, we consider the ratio between two components and show that its support determines the relationship between the two corresponding nodes in the associated DAG uniquely. We use these theoretical findings to propose a simple procedure to learn the ML coefficient matrix from observational data. It has the nice property to identify, with probability 1, the true ML coefficient matrix for a sufficiently large number of observations.

The main part of this chapter deals with the parameter learning of recursive ML models with known DAG. The statistical theory of recursive ML models is generally challenging, here, for example, since no σ-finite measure exists on the space of observations that dominates the distributional family of recursive ML models on the given DAG. So standard likelihood theory does not apply. Kiefer and Wolfowitz [40] extended the standard definition of a maximum likelihood estimate (MLE) to the non-dominated case. Following this approach, our goal is to find the corresponding MLEs of the ML coefficient matrix. But, like the standard definition, the Kiefer-Wolfowitz approach suffers from the problem that the recommended densities are only uniquely defined up to almost sure equality. That the densities can be changed on null sets leads to different MLEs depending on the used density version. We illustrate this difficulty by examples. Among the potential MLEs, one stands out. It is based on the minimal observed ratio between two components and is, with probability 1, equal to the true ML coefficient matrix if the number of observations is sufficiently large. We prove that this estimate is a MLE in the sense of Kiefer-Wolfowitz but also discuss others. The most elaborate step is to derive the densities. Here we find interesting relationships with other classes of graphical models.

Each chapter of the thesis is based on a paper or a manuscript which is very close to be submitted:

Chapter 2: N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693–2720, 2018.

Chapter 3: N. Gissibl, C. Klüppelberg, and M. Otto. Tail dependence of recursive max-linear models with regularly varying noise variables. *Econometrics and Statistics*, 6:149–167, 2018.

Chapter 4: N. Gissibl, C. Klüppelberg, and S. L. Lauritzen. Identifiability and estimation of recursive max-linear models. In preparation, 2018.

The individual chapters are basically self-contained. They introduce the notation, methodology, and literature which is needed to understand the chapters in their respective introductory sections. Different notations, abbreviations, and model assumptions on a recursive ML model seem reasonable in different settings; therefore, they might differ from chapter to chapter.

Research on recursive ML models

We conclude the Introduction with the presentation of further research works on recursive ML models, which shows the variety of topics linked to this model class. Overall, we observe that extreme value theory and extreme value statistics are slowly starting to make their way into graphical modeling (see e.g. Hitz and Evans [34] and Papastathopoulos and Strokorb [54]).

Because of the accumulated knowledge on recursive ML models, we expect that a consequent use of algebraic theory based on properties of the max-times semiring $(\mathbb{R}_+, \vee, \cdot)$ (see e.g. Butkovič [7]) would simplify the theory of recursive ML models. Zhang [76] started to investigate this, for example, by finding the fixed point of the max-linear structural equations defining a recursive ML model. Further promising results in this direction are achieved in this master's thesis.

As already mentioned, a recursive ML model is generally not faithful to its DAG. Klüppelberg and Lauritzen [43] prove that a recursive ML model is faithful if and only if the DAG has at most one path between two nodes. This paper also provides a detailed overview and summary of the methodological concepts of this model; for example, the necessary graph terminology, basic properties of conditional independence, and the Markov properties of directed graphical models and their relation to SEMs are discussed.

Parts of the paper

N. Gissibl and C. Klüppelberg. Prediction of recursive max-linear models. In preparation, 2018.

have been contained in the very first version of Gissibl and Klüppelberg [27] available online using the link `https://arxiv.org/pdf/1512.07522v1.pdf`. In this paper we assume that some node variables of a recursive ML model are observed and we want to predict the values of the remaining. In a first step, we investigate representations of node variables in terms of a given subset of node variables and a minimal number of noise variables. In some situations, this leads to almost sure equality between two distinct appropriately scaled node variables. This result has direct implications for the prediction problem: some of the unknown node variables can, with probability 1, be predicted exactly. We use these results to provide an algorithm for the prediction problem. Given observations of parts of a recursive ML model, it predicts the other node variables. To prove its correctness, we also determine reduced forms of regular conditional distributions compared to previous representations (see e.g. [73]).

Chapter 3 of this thesis is the continuation of the work presented in Otto [53]. The focus of this master's thesis is on the so-called *homogeneous model* (see Example 3.3.1 below). The homogeneous model is a special case of a recursive ML model and is completely determined by the reachability matrix of any underlying DAG. Because of its simple structure, it was the starting point for the development of recursive ML models and is still an important model to try new ideas and approaches for recursive ML models. In [53] a consistent and asymptotically normal estimator for the tail dependence matrix is suggested. The performance of this estimator and the algorithms developed in this master's thesis to recover homogeneous models as far as possible from their tail dependence matrices is evaluated on simulated data sets.

As in Chapter 3, Krali [46] requires regularly varying noise variables in recursive ML models. The author proposes a scaling technique for causal order search and an estimation procedure

for the scaling parameters. Algorithms developed here are tested for their performance in a simulation study with the result that they perform nicely even in high dimensions.

Hartl [32] has done important preliminary work for Chapter 4 of this thesis. The estimate mentioned above that is based on the minimal observed ratio between two components is investigated and first properties are presented, mainly under the assumption of standard Fréchet distributed noise variables. In addition, the author illustrates the performance of this estimate on simulated data and makes first attempts to infer the minimum underlying causal structure of a recursive ML model from observational data. For the latter a method is suggested which also seems to work if the data do not follow a recursive ML model exactly. We come back to this in the conclusion and outlook of Chapter 4.

A recursive ML model has already been fitted to real data. In fact, Einmahl et al. [20] present an estimator of the tail dependence function in the context of extreme value theory and apply it to data from the EURO STOXX 50 Index assuming an underlying recursive ML model on a known DAG with standard 1-Fréchet distributed noise variables.

The last work related to recursive ML models we would like to mention is that of Klüppelberg and Sönmez [44]. They extend the definition of this model class to infinite graphs and investigate their relations to classical percolation theory, more precisely to nearest neighbor bond percolation.

Chapter 2

Max-linear models on directed acyclic graphs

Abstract

We consider a new recursive structural equation model where all variables can be written as max-linear function of their parental node variables and independent noise variables. The model is max-linear in terms of the noise variables, and its causal structure is represented by a directed acyclic graph. We detail the relation between the weights of the recursive structural equation model and the coefficients in its max-linear representation. In particular, we characterize all max-linear models which are generated by a recursive structural equation model and show that its max-linear coefficient matrix is the solution of a fixed point equation. We also find the minimum directed acyclic graph representing the recursive structural equations of the variables. The model structure introduces a natural order between the node variables and between the max-linear coefficients. This yields representations of the vector components, which are based on the minimum number of node and noise variables.

MSC 2010 subject classifications: Primary 60G70, 60E15, 05C20; secondary 05C75

Keywords and phrases: Directed acyclic graph, graphical model, max-linear model, minimal representation, path analysis, structural equation model

2.1 Introduction

Graphical models are a popular tool to analyze and visualize conditional independence relations between random variables (see e.g. Koller and Friedman [45] and Lauritzen [47]). Each node in a graph indicates a random variable, and the graph encodes conditional independence relations between the random variables. We focus on directed graphical models, also called Bayesian networks, where edge orientations come along with an intuitive causal interpretation. The conditional independence relations between the random variables, which are encoded by a directed acyclic graph (DAG), can be explored using the (directed) Markov property: each variable is conditionally independent of its non-descendants (excluding the parents) given its parents (cf. [47], Chapter 3.2).

Despite many areas of application for directed graphical models, ranging from artificial intelligence, decision support systems, and engineering to genetics, geology, medicine, and finance

(see e.g. Pourret et al. [59]), graphical modeling of random vectors has mainly been limited to discrete and Gaussian distributions; see e.g. [45, 47]. In the context of risk assessment, risk exposures are usually modeled by continuous variables, however, the assumption of Gaussianity leads invariably to severe underestimation of large risks and therefore to unsuitable models.

Recursive structural equation models (recursive SEMs) offer a possibility to construct directed graphical models; cf. Bollen [5], Pearl [55], and Spirtes et al. [69]. For a given DAG $\mathcal{D} = (V, E)$ with nodes $V = \{1, \dots, d\}$ and edges $E = \{(k, i) : i \in V \text{ and } k \in \mathrm{pa}(i)\}$ define

$$X_i = f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, Z_i), \quad i = 1, \dots, d, \tag{2.1.1}$$

where $\mathrm{pa}(i)$ denotes the parents of node i in $\mathcal{D}$ and f_i is a real-valued measurable function; $Z_1, \dots, Z_d$ are independent noise variables. Thus a recursive SEM is specified by an underlying causal structure given by a DAG $\mathcal{D}$, the functions f_i, and the distributions of Z_i for $i = 1, \dots, d$. In this setting the distribution of $\boldsymbol{X} = (X_1, \dots, X_d)$ is uniquely defined by the distributions of the noise variables and, denoting by $\mathrm{nd}(i)$ the non-descendants of node i,

$$X_i \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{nd}(i) \setminus \mathrm{pa}(i)} \mid \boldsymbol{X}_{\mathrm{pa}(i)}, \quad i = 1, \dots, d; \tag{2.1.2}$$

i.e., the distribution of $\boldsymbol{X}$ is Markov relative to $\mathcal{D}$ (see Theorem 1.4.1 and the related discussion in [55]). Recently, recursive linear SEMs and generalizations in a Gaussian setting have received particular attention; see Bühlmann et al. [6], Ernest et al. [21], and references therein.

Our focus is not on sums but on maxima, where natural candidates for the noise distributions are the extreme value distributions or distributions in their domains of attraction; see e.g. Resnick [60, 61]. We define a *recursive max-linear (ML) model* $\boldsymbol{X}$ *on a DAG* $\mathcal{D}$ by

$$X_i := \bigvee_{k \in \mathrm{pa}(i)} c_{ki} X_k \vee c_{ii} Z_i, \quad i = 1, \dots, d, \tag{2.1.3}$$

with independent random variables $Z_1, \dots, Z_d$ with support $\mathbb{R}_+ = [0, \infty)$ and positive weights c_{ki} for all $i \in V$ and $k \in \mathrm{pa}(i) \cup \{i\}$.

This new model is motivated by applications to risk analysis, where extreme risks play an essential role and may propagate through a network. In such a risk setting, it is natural to require the noise variables to have positive infinite support. Moreover, we may think of the edge weights in (2.1.3) as relative quantities so that a risk may originate with certain proportions in its different ancestors.

In this chapter we investigate structural properties as well as graph properties of a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D}$. We will show that $\boldsymbol{X}$ is a *max-linear (ML) model* (for background on ML models in the context of extreme value theory, see e.g. de Haan and Ferreira [14], Chapter 6) in the sense that

$$X_i = \bigvee_{j=1}^{d} b_{ji} Z_j, \quad i = 1, \dots, d, \tag{2.1.4}$$

with $Z_1, \dots, Z_d$ as in (2.1.3), and $B = (b_{ij})_{d \times d}$ is a matrix with nonnegative entries. We call B

max-linear (ML) coefficient matrix of $\boldsymbol{X}$ and its entries *max-linear (ML) coefficients.*

The ML coefficients of $\boldsymbol{X}$ can be determined by a path analysis of $\mathcal{D}$. Throughout we write $k \to i$ if there is an edge from k to i in $\mathcal{D}$. We assign a weight to every path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$, which is the product of the edge weights along p multiplied by the weight of the noise variable Z_j (a concept which goes back to Wright [74]):

$$d_{ji}(p) = c_{k_0,k_0} c_{k_0,k_1} \dots c_{k_{n-2},k_{n-1}} c_{k_{n-1},k_n} = c_{k_0,k_0} \prod_{\nu=0}^{n-1} c_{k_\nu,k_{\nu+1}}. \tag{2.1.5}$$

We will show that the ML coefficients are given for $i \in V$ by

$$b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \text{ for } j \in \mathrm{an}(i), \quad b_{ii} = c_{ii}, \quad \text{and} \quad b_{ji} = 0 \text{ for } j \in V \setminus (\mathrm{an}(i) \cup \{i\}), \tag{2.1.6}$$

where P_{ji} is the set of paths from j to i and $\mathrm{an}(i)$ the ancestors of i.

The computation in (2.1.6) corresponds to the algebraic path problem over the max-times semiring $(\mathbb{R}_+, \vee, \cdot)$ (see e.g. Mahr [51] and Rote [62]). We present it in matrix form using the matrix product over $(\mathbb{R}_+, \vee, \cdot)$. We apply this concept in the two different situations, where the DAG $\mathcal{D}$ is given, and we test if a given ML coefficient matrix is consistent with $\mathcal{D}$, but also later on, when we check if a given matrix defines a recursive SEM on some unspecified DAG.

From (2.1.6) it is clear that not all paths are needed for representing $\boldsymbol{X}$ as ML model (2.1.4). This perception leads to a complexity reduction of the model in different ways and in different situations. For every specific component X_i of $\boldsymbol{X}$ only those paths with terminal node i which carry the maximum weight are relevant for its representation (2.1.4), and we call them *max-weighted paths.* All other paths can be disposed of without changing this representation. It is even sufficient to consider only one in $\mathcal{D}$ max-weighted path from every ancestor of i to i. Consequently, X_i can be represented as component of a recursive ML model on a polytree with node set $\mathrm{an}(i) \cup \{i\}$ and with the same weights and noise variables as in the original representation (2.1.3).

However, in general none of these individual polytrees represents all components of $\boldsymbol{X}$ in the sense of (2.1.3) simultaneously. Still there may be subgraphs of $\mathcal{D}$ and weights such that all components of $\boldsymbol{X}$ have representation (2.1.3), and we present all such possible subgraphs and weights. In particular, we characterize the smallest subgraph of this kind, which we call *minimum max-linear (ML) DAG of* $\boldsymbol{X}$, and point out its prominent role.

We show that all DAGs and weights which represent $\boldsymbol{X}$ as in (2.1.3) can be identified from the ML coefficient matrix B of $\boldsymbol{X}$. In this context, we also give necessary and sufficient conditions on a matrix to be the ML coefficient matrix of some recursive ML model.

It is a simple but important observation that there is a natural order between the components of $\boldsymbol{X}$; from (2.1.3) we see immediately that $X_i \geq c_{ki} X_k$ holds for all $i \in V$ and $k \in \mathrm{pa}(i)$. For every component of $\boldsymbol{X}$ and some $U \subseteq V$, we find lower and upper bounds in terms of $\boldsymbol{X}_U := (X_\ell, \ell \in U)$. Often we do not need all components of $\boldsymbol{X}_U$ to compute the best bounds of X_i in terms of components of $\boldsymbol{X}_U$. If $i \in U$, then an upper and lower bound is given by X_i itself; otherwise, for a lower bound we only need to consider a component X_j of $\boldsymbol{X}_U$ if $j \in \mathrm{an}(i)$,

but no max-weighted path from j to i passes through some node in $U \setminus \{j\}$. A similar result and concept applies for the upper bound of X_i. Thus the max-weighted paths also lead in this context indirectly to a complexity reduction. We will also use the max-weighted ancestors of i in U to obtain a minimal representation of X_i in terms of $\boldsymbol{X}_U$ and noise variables.

This chapter is organized as follows. In Section 2.2 we discuss the max-linearity of a recursive ML model $\boldsymbol{X}$ and express its ML coefficient matrix in terms of a weighted adjacency matrix of a corresponding DAG. Section 2.3 introduces the important notion of a max-weighted path and studies its consequences for the ML coefficients. In Section 2.4 we give necessary and sufficient conditions for a ML model being a recursive ML model on a given DAG. Section 2.5 is devoted to the minimum ML DAG of $\boldsymbol{X}$ as the DAG with the minimum number of edges within the class of all DAGs representing $\boldsymbol{X}$ in the sense of (2.1.3). In Section 2.6, given a set of node variables, we investigate which information can be gained for the other components of $\boldsymbol{X}$. This results in lower and upper bounds for the components. Finally, we derive a minimal representation for the components of $\boldsymbol{X}$ as max-linear function of a subset of node variables and certain noise variables.

We use the following notation throughout this chapter. For a node $i \in V$, the sets $\mathrm{an}(i)$, $\mathrm{pa}(i)$, and $\mathrm{de}(i)$ contain the *ancestors, parents*, and *descendants of i in $\mathcal{D}$*. Furthermore, we use the notation $\mathrm{An}(i) = \mathrm{an}(i) \cup \{i\}$, $\mathrm{Pa}(i) = \mathrm{pa}(i) \cup \{i\}$, and $\mathrm{De}(i) = \mathrm{de}(i) \cup \{i\}$. We write $U \subseteq V$ for a non-empty subset U of nodes, $\boldsymbol{X}_U = (X_\ell, \ell \in U)$, and $U^c = V \setminus U$. All our vectors are row vectors. We also extend the previous notation in a natural way by writing $\mathrm{an}(U) = \bigcup_{i \in U} \mathrm{an}(i)$, $\mathrm{An}(U) = \mathrm{an}(U) \cup U$, and so on. For a matrix B with nonnegative entries, we write $\mathrm{sgn}(B)$ for the matrix with entries equal to 1 if the corresponding entry in B is positive and 0 else. We denote by $\mathbb{1}_U$ the indicator function of U and set $\mathbb{1}_\varnothing \equiv 0$. In general, we consider statements for $i \in \varnothing$ as invalid. For arbitrary (possibly random) $a_i \in \mathbb{R}_+$, we set $\bigvee_{i \in \varnothing} a_i = 0$ and $\bigwedge_{i \in \varnothing} a_i = \infty$.

2.2 Max-linearity of a recursive ML model

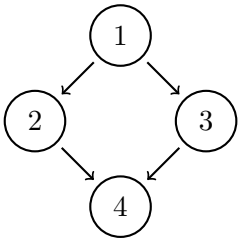
For a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D} = (V, E)$, given by (2.1.3), we derive its max-linear representation (2.1.4). We start with our leading example, a recursive ML model on the diamond-shaped DAG depicted below.

Example 2.2.1. [Max-linear representation of a recursive ML model]
Consider a recursive ML model $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ on the DAG

$$\mathcal{D} = (V, E) = (\{1, 2, 3, 4\}, \{(1, 2), (1, 3), (2, 4), (3, 4)\})$$

with weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$. We obtain for the components of $\boldsymbol{X}$:

$$\begin{aligned} X_1 &= c_{11} Z_1 \\ X_2 &= c_{12} X_1 \vee c_{22} Z_2 = c_{12} c_{11} Z_1 \vee c_{22} Z_2 \\ X_3 &= c_{13} X_1 \vee c_{33} Z_3 = c_{13} c_{11} Z_1 \vee c_{33} Z_3 \end{aligned}$$



$$\begin{aligned} X_4 &= c_{24}X_2 \vee c_{34}X_3 \vee c_{44}Z_4 \\ &= c_{24}(c_{12}c_{11}Z_1 \vee c_{22}Z_2) \vee c_{34}(c_{13}c_{11}Z_1 \vee c_{33}Z_3) \vee c_{44}Z_4 \\ &= (c_{24}c_{12}c_{11} \vee c_{34}c_{13}c_{11})Z_1 \vee c_{24}c_{22}Z_2 \vee c_{34}c_{33}Z_3 \vee c_{44}Z_4. \end{aligned}$$

Thus $\boldsymbol{X}$ satisfies (2.1.4) with ML coefficient matrix

$$B = \begin{bmatrix} c_{11} & c_{11}c_{12} & c_{11}c_{13} & c_{11}c_{12}c_{24} \vee c_{11}c_{13}c_{34} \\ 0 & c_{22} & 0 & c_{22}c_{24} \\ 0 & 0 & c_{33} & c_{33}c_{34} \\ 0 & 0 & 0 & c_{44} \end{bmatrix}.$$

We observe that the ML coefficients satisfy indeed (2.1.6). Moreover, B is an upper triangular matrix, since $\mathcal{D}$ is well-ordered (cf. Remark 2.2.3(ii)). □

The following result shows that such a representation can be obtained in general: every component of a recursive ML model has a max-linear representation in terms of its ancestral noise variables and an independent one. It provides a general method to calculate the ML coefficients by a path analysis as described in (2.1.5) and (2.1.6).

Theorem 2.2.2. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$, and let $B = (b_{ij})_{d\times d}$ be the matrix with entries as defined in* (2.1.6)*. Then*

$$X_i = \bigvee_{j\in \mathrm{An}(i)} b_{ji}Z_j, \quad i = 1, \dots, d; \tag{2.2.1}$$

i.e., B is the ML coefficient matrix of $\boldsymbol{X}$.

Proof. Without loss of generality we assume throughout this proof that $\mathcal{D}$ is well-ordered (cf. Remark 2.2.3(ii)). We prove the identity (2.2.1) by induction on the number of nodes of $\mathcal{D}$. For $d = 1$ we have by (2.1.3)

$$X_1 = c_{11}Z_1 = b_{11}Z_1,$$

where the last equality holds by (2.1.6). Suppose that (2.2.1) holds for a recursive ML model $\boldsymbol{X}$ of dimension d; i.e.,

$$X_k = \bigvee_{j\in \mathrm{An}(k)} b_{jk}Z_j = \bigvee_{j\in \mathrm{an}(k)} \bigvee_{p\in P_{jk}} d_{jk}(p)Z_j \vee c_{kk}Z_k, \quad k = 1, \dots, d.$$

Now consider a $(d+1)$-variate recursive ML model. We first investigate the nodes $i \in \{1, \dots, d\}$. Since $\mathcal{D}$ is well-ordered, we have $(d+1) \in V \setminus \mathrm{pa}(i)$. Hence, it suffices to show representation (2.2.1) with respect to the subgraph $\mathcal{D}[\{1, \dots, d\}] = (\{1, \dots, d\}, E \cap (\{1, \dots, d\} \times \{1, \dots, d\}))$. However, this holds by the induction hypothesis. So we can use this representation for $i \in \{1, \dots, d\}$ and

(2.A.1) to obtain

$$\begin{aligned} X_{d+1} &= \bigvee_{k\in\mathrm{pa}(d+1)} c_{k,d+1}X_k \vee c_{d+1,d+1}Z_{d+1} \\ &= \bigvee_{k\in\mathrm{pa}(d+1)} \bigvee_{j\in\mathrm{an}(k)} \bigvee_{p\in P_{jk}} c_{k,d+1}d_{jk}(p)Z_j \vee \bigvee_{k\in\mathrm{pa}(d+1)} c_{k,d+1}c_{kk}Z_k \vee c_{d+1,d+1}Z_{d+1} \\ &= \bigvee_{j\in\mathrm{an}(d+1)} \Big(\bigvee_{k\in\mathrm{de}(j)\cap\mathrm{pa}(d+1)} \bigvee_{p\in P_{jk}} c_{k,d+1}d_{jk}(p) \vee \bigvee_{k\in\mathrm{pa}(d+1)\cap\{j\}} c_{k,d+1}c_{kk}\Big) Z_j \vee c_{d+1,d+1}Z_{d+1}. \end{aligned}$$

Observe that every path from some j to $d+1$ is of the form $p = [j \to \cdots \to k \to d+1]$ for some $k \in \mathrm{de}(j) \cap \mathrm{pa}(d+1)$, or an edge $j \to d+1$ corresponding to $j \in \mathrm{pa}(d+1)$. By (2.1.5) the path p has weight $d_{j,d+1}(p) = d_{jk}(p)c_{k,d+1}$, and the edge $j \to d+1$ has weight $d_{j,d+1}([j \to d+1]) = c_{jj}c_{j,d+1}$. This yields

$$X_{d+1} = \bigvee_{j\in\mathrm{an}(d+1)} \bigvee_{p\in P_{j,d+1}} d_{j,d+1}(p)Z_j \vee c_{d+1,d+1}Z_{d+1} = \bigvee_{j\in\mathrm{An}(d+1)} b_{j,d+1}Z_j,$$

where we have used that $b_{j,d+1} = \bigvee_{p\in P_{j,d+1}} d_{j,d+1}(p)$ for $j \in \mathrm{an}(d+1)$ and $b_{d+1,d+1} = c_{d+1,d+1}$. □

By (2.1.6) the ML coefficient b_{ji} of $\boldsymbol{X}$ is different from zero if and only if $j \in \mathrm{An}(i)$. This information is contained in the *reachability matrix* $R = (r_{ij})_{d\times d}$ *of* $\mathcal{D}$, which has entries

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

If the ji-th entry of R is equal to one, then i *is reachable from* j.

Remark 2.2.3. Let R be the reachability matrix of $\mathcal{D}$.

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

(ii) The DAG $\mathcal{D}$ can be *well-ordered*, which means that the set $V = \{1, \dots, d\}$ of nodes can be linearly ordered in a way compatible with $\mathcal{D}$ such that $k \in \mathrm{pa}(i)$ implies $k < i$ (see e.g. Appendix A of Diestel [15]). If $\mathcal{D}$ is well-ordered, then B and R are upper triangular matrices. □

Finding the ML coefficient matrix B from $\mathcal{D}$ and the weights in (2.1.3) by a path analysis as described in (2.1.5) and (2.1.6) would be very inefficient. We may, however, compute B by means of a specific matrix multiplication.

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

$$(F = (f_{ij})_{m\times n}, G = (g_{ij})_{n\times p}) \mapsto F \odot G := \Big(\bigvee_{k=1}^{n} f_{ik}g_{kj} \Big)_{m\times p}. \tag{2.2.2}$$

The triple $(\mathbb{R}_+, \vee, \cdot)$, which is called max-times or subtropical algebra, is an idempotent semiring with 0 as 0-element and 1 as 1-element. The operation $\odot$ is therefore a matrix product over a

semiring. Such semirings are fundamental in tropical geometry; for an introduction see Butkovič [7] or Maclagan and Sturmfels [50]. The matrix product $\odot$ is associative: for $F \in \mathbb{R}_+^{m\times n}$, $G \in \mathbb{R}_+^{n\times p}$, and $H \in \mathbb{R}_+^{p\times q}$, $F\odot(G\odot H) = (F\odot G)\odot H$, and we have $(F\odot G)^\top = G^\top \odot F^\top$. Denoting by $\mathcal{B}$ all $d\times d$ matrices with nonnegative entries and by $\vee$ the componentwise maximum between two matrices, $(\mathcal{B}, \vee, \odot)$ is also a semiring with the null matrix as 0-element and the identity matrix $\mathrm{id}_{d\times d}$ as 1-element. This semiring is, however, not commutative, since $\odot$ is in general not. Consistent with a matrix product, we define powers recursively: $A^{\odot 0} := \mathrm{id}_{d\times d}$ and $A^{\odot n} := A^{\odot(n-1)} \odot A$ for $A \in \mathcal{B}$ and $n \in \mathbb{N}$.

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 $\mathcal{D}$.

Theorem 2.2.4. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$ as in (2.1.3). Define the matrices*

$$A := \mathrm{diag}(c_{11}, \dots, c_{dd}), \quad A_0 := \big(c_{ij}\mathbb{1}_{\mathrm{pa}(j)}(i)\big)_{d\times d}, \quad \text{and} \quad A_1 := \big(c_{ii}c_{ij}\mathbb{1}_{\mathrm{pa}(j)}(i)\big)_{d\times d}.$$

Then the ML coefficient matrix B of $\boldsymbol{X}$ from Theorem 2.2.2 has representation

$$B = A \quad \text{for } d = 1 \quad \text{and} \quad B = A \vee \bigvee_{k=0}^{d-2} \big(A_1 \odot A_0^{\odot k}\big) \quad \text{for } d \geq 2.$$

Proof. For $d = 1$ we know from (2.1.6) that $b_{11} = c_{11}$. Hence, $B = A$. Now assume that $d \geq 2$. First we show that if $\mathcal{D}$ has a path of length n (a path consisting of n edges) from node j to node i, then the ji-th entry of the matrix $A_1 \odot A_0^{\odot(n-1)}$ is equal to the maximum weight of all paths of lengths n from j to i, otherwise it is zero. The proof is by induction on n.

An edge $j \to i$, which is the only path of length $n = 1$, has the weight $d_{ji}([j \to i]) = c_{jj}c_{ji}$. Since the ji-th entry of the matrix $A_1 \odot A_0^{\odot 0} = A_1 \odot \mathrm{id}_{d\times d} = A_1$ is given by $c_{jj}c_{ji}\mathbb{1}_{\mathrm{pa}(i)}(j)$, the statement is true for $n = 1$.

Denote by $a_{0,ji}, a_{n,ji}$, and $a_{n+1,ji}$ the ji-th entry of $A_0, A_1\odot A_0^{\odot(n-1)}$, and $A_1\odot A_0^{\odot n}$, respectively. As $A_1\odot A_0^{\odot n} = (A_1\odot A_0^{\odot(n-1)})\odot A_0$, the ji-th entry of $A_1\odot A_0^{\odot n}$ is given by $a_{n+1,ji} = \bigvee_{k=1}^d a_{n,jk}a_{0,ki} = \bigvee_{k=1}^d a_{n,jk}c_{ki}\mathbb{1}_{\mathrm{pa}(i)}(k)$. We obtain from the induction hypothesis and (2.1.5) that $a_{n,jk}a_{0,ki}$ is zero if $\mathcal{D}$ does not contain a path of length n from j to k or the edge $k \to i$; otherwise it is equal to the maximum weight of all paths which consist of a path of length n from j to k and the edge $k \to i$. Since every path of length $n+1$ from j to i is of this form for some $k \in V$, the ji-th entry of $A_1 \odot A_0^{\odot n}$ is indeed equal to the maximum weight of all paths of length $n+1$ from j to i if there exists such a path, otherwise it is zero.

Finally, again by (2.1.6), for $i \in V$ and $j \in \mathrm{an}(i)$, the ML coefficient b_{ji} is equal to the maximum weight of all paths from j to i, and note that due to acyclicity, a path in $\mathcal{D}$ is at most of length $d-1$. Thus, if $j \in \mathrm{an}(i)$, then the ji-th entry of $\bigvee_{k=0}^{d-2} A_1 \odot A_0^{\odot k}$ is equal to b_{ji}, otherwise it is zero. Since by (2.1.6), $b_{ii} = c_{ii}$ and $b_{ji} = 0$ for $j \in V \setminus \mathrm{An}(i)$, the ML coefficient matrix B is given by

$$B = A \vee A_1 \vee \big(A_1 \odot A_0\big) \vee \big(A_1 \odot A_0^{\odot 2}\big) \vee \cdots \vee \big(A_1 \odot A_0^{\odot(d-2)}\big).$$

□

The following has been shown in the proof of Theorem 2.2.4.

Corollary 2.2.5. *If $\mathcal{D}$ has a path of length n from j to i, the ji-th entry of the matrix $A_1 \odot A_0^{\odot(n-1)}$ is equal to the maximum weight of all paths of length n from j to i, otherwise the entry is zero.*

Summarizing the noise variables of $\boldsymbol{X}$ into the vector $\boldsymbol{Z} = (Z_1, \dots, Z_d)$, the representation (2.2.1) of $\boldsymbol{X}$ can be written by means of the product $\odot$ as

$$\boldsymbol{X} = \boldsymbol{Z} \odot B = \Big(\bigvee_{j=1}^{d} b_{ji} Z_j, i = 1, \dots, d \Big) = \Big(\bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j, i = 1, \dots, d \Big).$$

Consequently, the definition of the matrix product $\odot$ modifies and extends the definition given in Wang and Stoev [73, Section 2.1, Eq. (2)].

2.3 Max-weighted paths

Given a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D} = (V, E)$ with weights c_{ki} for $i \in V$, $k \in \mathrm{Pa}(i)$ and ML coefficient matrix $B = (b_{ij})_{d \times d}$, we investigate the paths of $\mathcal{D}$ and their particular weights, the implications on the ML coefficients as well as induced subgraph structures leading to reduced representations of (2.1.3).

From (2.1.6) and (2.2.1) we know that a path p from j to i, whose weight $d_{ji}(p)$ is strictly smaller than b_{ji}, does not have any influence on the distribution of $\boldsymbol{X}$. This fact suggests the following definition.

Definition 2.3.1. Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with path weights as in (2.1.5) and ML coefficient matrix B. We call a path p from j to i a *max-weighted path* if $b_{ji} = d_{ji}(p)$. □

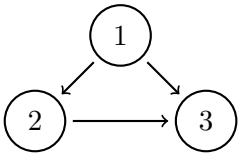
A prominent example, where all paths are max-weighted, is the following.

Example 2.3.2. [Polytree] A *polytree* is a DAG whose underlying undirected graph has no cycles; polytrees have at most one path between any pair of nodes. Thus, assuming that $\boldsymbol{X}$ is a recursive ML model on a polytree, all paths must be max-weighted. □

The following example indicates the importance and consequences of max-weighted paths.

Example 2.3.3. [Max-weighted paths, graph reduction]
Consider a recursive ML model $\boldsymbol{X} = (X_1, X_2, X_3)$ on the DAG

$$\mathcal{D} = (V, E) = (\{1, 2, 3\}, \{(1, 2), (1, 3), (2, 3)\})$$



with weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$ and ML coefficient matrix B. We distinguish between two situations:

(1) If $c_{13} > c_{12}c_{23}$, then the edge $1 \to 3$ is the unique max-weighted path from 1 to 3.

(2) If, however, $c_{13} \le c_{12}c_{23}$, then $b_{13} = c_{11}c_{12}c_{23} = \frac{b_{12}b_{23}}{b_{22}}$ and the path $[1 \to 2 \to 3]$ is max-weighted. We obtain in this case

$$X_3 = b_{13}Z_1 \vee b_{23}Z_2 \vee b_{33}Z_3 = \frac{b_{23}}{b_{22}}(b_{12}Z_1 \vee b_{22}Z_2) \vee b_{33}Z_3 = c_{23}X_2 \vee b_{33}Z_3.$$

Thus $\boldsymbol{X}$ is also a recursive ML model on the DAG

$$\mathcal{D}^B := (\{1,2,3\}, \{(1,2),(2,3)\}).$$

Here $\mathcal{D}^B$ is the DAG with the minimum number of edges such that $\operatorname{sgn}(B)$ is its reachability matrix. By Remark 2.2.3(i) there cannot be a smaller DAG representing $\boldsymbol{X}$ in the sense of (2.1.3). □

We present some immediate consequences of the path weights in (2.1.5) and the definition of max-weighted paths.

Remark 2.3.4. (i) If there is only one path between two nodes, it is max-weighted.

(ii) Every subpath of a max-weighted path is also max-weighted.

(iii) Every path which results from a max-weighted path by replacing a subpath with another max-weighted subpath is also max-weighted. □

To find for some $i \in V$ and $j \in \operatorname{an}(i)$ the ML coefficient b_{ji}, it suffices to know the weight c_{jj} of the noise variable Z_j and the edge weights along one arbitrary max-weighted path from j to i, since every max-weighted path from j to i has the same weight. This allows us to represent every component of $\boldsymbol{X}$ as component of a recursive ML model on a subgraph of $\mathcal{D}$. For this purpose we introduce the following definition.

Definition 2.3.5. Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$, and let $\overline{\mathcal{D}} = (\overline{V}, \overline{E})$ be a subgraph of $\mathcal{D}$. We denote by $\overline{\operatorname{pa}}(i)$ the parents of node i in $\overline{\mathcal{D}}$ and define

$$Y_i := \bigvee_{k \in \overline{\operatorname{pa}}(i)} c_{ki}Y_k \vee c_{ii}Z_i, \quad i \in \overline{V},$$

with the same weights and noise variables as in the representation of $\boldsymbol{X}$ in (2.1.3). We call the resulting recursive ML model $\boldsymbol{Y} = (Y_\ell, \ell \in \overline{V})$ *recursive ML submodel of* $\boldsymbol{X}$ *induced by* $\overline{\mathcal{D}}$. □

We summarize some immediate properties of $\boldsymbol{Y}$.

Remark 2.3.6. Let $i \in V$ with ancestors $\operatorname{an}(i)$ in $\mathcal{D}$. Denote by $\overline{B} = (\overline{b}_{ij})_{|\overline{V}| \times |\overline{V}|}$ the ML coefficient matrix of $\boldsymbol{Y}$.

(i) Every path in $\overline{\mathcal{D}}$ has the same weight (2.1.5) as in $\mathcal{D}$.

(ii) A path of $\overline{\mathcal{D}}$ which is in $\mathcal{D}$ a max-weighted path is also in $\overline{\mathcal{D}}$ max-weighted.

(iii) For $j \in \mathrm{an}(i)$, $\overline{\mathcal{D}}$ has one in $\mathcal{D}$ max-weighted path from j to i if and only if $\overline{b}_{ji} = b_{ji}$.

(iv) $\overline{\mathcal{D}}$ has one in $\mathcal{D}$ max-weighted path from every $j \in \mathrm{an}(i)$ to i if and only if $X_i = Y_i$. □

By Remark 2.3.4(ii) for every $i \in V$, there exists a polytree $\mathcal{D}_i$ of $\mathcal{D}$ with node set $\mathrm{An}(i)$ which has exactly one in $\mathcal{D}$ max-weighted path from every ancestor of i to i. There may even exist several such polytrees (cf. Example 2.3.8 below). We learn from the construction of $\mathcal{D}_i$ and Remark 2.3.4(ii) that indeed every path of $\mathcal{D}_i$ is in $\mathcal{D}$ max-weighted. Therefore, some component X_j of $\boldsymbol{X}$ coincides by Remark 2.3.6(iv) with the corresponding one of the recursive ML submodel of $\boldsymbol{X}$ induced by $\mathcal{D}_i$ if and only if $\mathcal{D}_i$ has at least one path from every ancestor of j in $\mathcal{D}$ to j. By construction of $\mathcal{D}_i$ this property holds obviously for X_i. We summarize this result as follows.

Proposition 2.3.7. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$. For some $i \in V$ and $\mathrm{An}(i)$ in $\mathcal{D}$ let $\mathcal{D}_i$ be a polytree with node set $\mathrm{An}(i)$ such that $\mathcal{D}_i$ has one in $\mathcal{D}$ max-weighted path from every $j \in \mathrm{an}(i)$ to i. Let $\boldsymbol{Y}_i = (Y_\ell, \ell \in \mathrm{An}(i))$ be the recursive ML submodel of $\boldsymbol{X}$ induced by $\mathcal{D}_i$. Then for all $j \in \mathrm{An}(i)$ which have the same ancestors in $\mathcal{D}_i$ and $\mathcal{D}$, we have $X_j = Y_j$.*

We discuss the recursive ML model from Example 2.2.1 in the context of Definition 2.3.1 and Proposition 2.3.7.

Example 2.3.8. [Continuation of Example 2.2.1: max-weighted paths, polytrees, conditional independence]
For the polytrees as in Proposition 2.3.7, we identify all max-weighted paths ending in node 4. By Remark 2.3.4(i), the paths $[2 \to 4]$ and $[3 \to 4]$ are max-weighted. For the weights of the paths $[1 \to 2 \to 4]$ and $[1 \to 3 \to 4]$, we have three situations:

$$c_{11}c_{12}c_{24} = c_{11}c_{13}c_{34}, \quad c_{11}c_{12}c_{24} > c_{11}c_{13}c_{34}, \quad \text{and} \quad c_{11}c_{12}c_{24} < c_{11}c_{13}c_{34}.$$

In the first situation, both paths from 1 to 4 are max-weighted. Thus there are two different polytrees having one in $\mathcal{D}$ max-weighted path from every ancestor of 4 to 4, namely,

$$\mathcal{D}_{4,1} = (\{1,2,3,4\}, \{(1,2),(2,4),(3,4)\}) \quad \text{and} \quad \mathcal{D}_{4,2} = (\{1,2,3,4\}, \{(1,3),(2,4),(3,4)\}).$$

In the second situation, the path $[1 \to 2 \to 4]$ is the unique max-weighted path from 1 to 4 and, hence, $\mathcal{D}_{4,1}$ is the unique polytree as in Proposition 2.3.7 for node 4. The third situation is symmetric to the second, such that $\mathcal{D}_{4,2}$ is also such a unique polytree.

Now let $\boldsymbol{Y}_1 = (Y_{1,1}, Y_{1,2}, Y_{1,3}, Y_{1,4})$ and $\boldsymbol{Y}_2 = (Y_{2,1}, Y_{2,2}, Y_{2,3}, Y_{2,4})$ be the recursive ML submodels of $\boldsymbol{X}$ induced by $\mathcal{D}_{4,1}$ and $\mathcal{D}_{4,2}$. The distributions of $\boldsymbol{X}$, $\boldsymbol{Y}_1$, and $\boldsymbol{Y}_2$ are Markov relative to $\mathcal{D}$, $\mathcal{D}_{4,1}$, and $\mathcal{D}_{4,2}$, respectively. For a DAG, the local Markov property as specified in (2.1.2), is by Proposition 4 of Lauritzen et al. [48] equivalent to the global Markov property (for a definition see Corollary 3.23 of [47]). Using this property we find

$$Y_{1,1} \perp\!\!\!\perp Y_{1,4} \mid Y_{1,2} \quad \text{and} \quad Y_{2,1} \perp\!\!\!\perp Y_{2,4} \mid Y_{2,3}.$$

If the path $[1 \to 2 \to 4]$ is max-weighted, we have by Proposition 2.3.7 that

$$Y_{1,1} = X_1, \quad Y_{1,2} = X_2, \quad \text{and} \quad Y_{1,4} = X_4,$$

hence, $X_1 \perp\!\!\!\perp X_4 \mid X_2$. Accordingly, if $[1 \to 3 \to 4]$ is max-weighted, then

$$Y_{2,1} = X_1, \quad Y_{2,3} = X_3, \quad \text{and} \quad Y_{2,4} = X_4,$$

and $X_1 \perp\!\!\!\perp X_4 \mid X_3$ holds. Since the only conditional independence property encoded in $\mathcal{D}$ by the (global) Markov property is $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$, we can identify additional conditional independence properties of $\boldsymbol{X}$ from the polytrees in Proposition 2.3.7. □

Remark 2.3.9. (i) Assume the situation of Proposition 2.3.7. Let V_i be the set of all nodes in $\mathrm{An}(i)$ which have the same ancestors in $\mathcal{D}$ and $\mathcal{D}_i$. Since the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Markov relative to $\mathcal{D}$ and $\mathcal{D}_i$, respectively, conditional independence properties of $\boldsymbol{X}$ are encoded in $\mathcal{D}$ and of $\boldsymbol{Y}$ in $\mathcal{D}_i$. By Proposition 2.3.7 the conditional independence relations between subvectors of $\boldsymbol{Y}_{V_i} = (Y_\ell, \ell \in V_i)$ which we can read off from $\mathcal{D}_i$ hold also between the corresponding subvectors of $\boldsymbol{X}$. Since missing edges correspond to conditional independence properties, and $\mathcal{D}_i$ is a subgraph of $\mathcal{D}$, we can often identify additional conditional independence properties of $\boldsymbol{X}$ from $\mathcal{D}_i$.

(ii) From (i) or Example 2.3.8 we learn that a recursive ML model on a DAG $\mathcal{D}$ is in general not *faithful*; i.e., not all its conditional independence properties are encoded in $\mathcal{D}$ by the (global) Markov property. □

As can be seen from Examples 2.3.3 and 2.3.8, any reduction of a recursive ML model depends on the existence of max-weighted paths that pass through specific nodes. The following result shows how we can obtain this information from its ML coefficient matrix.

Theorem 2.3.10. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with ML coefficient matrix B. Let further $U \subseteq V$, $i \in V$ and $j \in \mathrm{an}(i)$, and recall from Remark 2.2.3(i) that $b_{ji} > 0$.*

(a) There is a max-weighted path from j to i which passes through some node in U if and only if

$$b_{ji} = \bigvee_{k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)} \frac{b_{jk} b_{ki}}{b_{kk}}. \tag{2.3.1}$$

(b) No max-weighted path from j to i passes through some node in U if and only if

$$b_{ji} > \bigvee_{k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)} \frac{b_{jk} b_{ki}}{b_{kk}}. \tag{2.3.2}$$

This holds also for $U = \varnothing$.

Proof. First assume that $\mathrm{De}(j) \cap U \cap \mathrm{An}(i) = \varnothing$. Thus no path, hence also no max-weighted path, from j to i passes through some node in U, and it suffices to verify (b). Since the right-hand

side of (2.3.2) is zero if and only if $\mathrm{De}(j) \cap U \cap \mathrm{An}(i) = \varnothing$ and the ML coefficient b_{ji} is positive, (b) is proven for this case (including the case that $U = \varnothing$).

Now assume that $\mathrm{De}(j) \cap U \cap \mathrm{An}(i) = \{k\}$, which implies that there is a path from j to i passing through $k \in U$. If $k = i$ or $k = j$, there is obviously a max-weighted path from j to i passing through i or j and (2.3.1) is always valid.

Next assume that $k \in V \setminus \{i, j\}$, and let p_1 and p_2 be max-weighted paths from j to k and from k to i, respectively. Denote by p the path from j to i consisting of the subpaths p_1 and p_2. By (2.1.5) and the definition of a max-weighted path, we obtain

$$d_{ji}(p) = \frac{1}{c_{kk}} d_{jk}(p_1) d_{ki}(p_2) = \frac{b_{jk} b_{ki}}{b_{kk}}.$$

Since p is max-weighted if and only if $b_{ji} = d_{ji}(p)$ and this is not the case if and only if $b_{ji} > d_{ji}(p)$, we have shown (a) and (b) for the situation of $\mathrm{De}(j) \cap U \cap \mathrm{An}(i) = \{k\}$. In particular, it follows that $b_{ji} \geq \frac{b_{jk} b_{ki}}{b_{kk}}$ for all $k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)$.

Assume now that $\mathrm{De}(j) \cap U \cap \mathrm{An}(i)$ contains more than one element and that a max-weighted path from j to i passes through some node $k \in U$. We know from above that this is equivalent to

$$b_{ji} = \frac{b_{jk} b_{ki}}{b_{kk}} \quad \text{and} \quad b_{ji} \geq \frac{b_{jl} b_{\ell i}}{b_{\ell\ell}} \text{ for all } \ell \in \big(\mathrm{De}(j) \cap U \cap \mathrm{An}(i)\big) \setminus \{k\},$$

which is again equivalent to (2.3.1). Similarly, we obtain (b). □

Remark 2.3.11. Recall the matrix product $\odot$ from (2.2.2), and let R be the reachability matrix of $\mathcal{D}$. We obtain from $R = \mathrm{sgn}(B)$ (Remark 2.2.3(i)) that for $i, j \in V$

$$\bigvee_{k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} = \bigvee_{k=1}^{d} \frac{b_{jk} b_{ki}}{b_{kk}} \mathbb{1}_U(k) =: \bigvee_{k=1}^{d} b_{jk} b_{U,ki}$$

is the ji-th entry of the matrix $B \odot B_U$ with $B_U = (b_{U,ij})_{d \times d}$. Thus we may decide whether there is a max-weighted path between two nodes that passes through some node in U by comparing the entries of the matrices B and $B \odot B_U$. Such use of the matrix product $\odot$ can be made at various points throughout the thesis, for example, in Remark 2.5.2(ii), Theorem 2.5.3, and Lemma 2.6.3(b). □

The following corollary gives an important property of the ML coefficients. The first part has been shown in the proof of Theorem 2.3.10, the second part follows from Remark 2.2.3(i).

Corollary 2.3.12. *For all $i \in V$, $k \in \mathrm{An}(i)$, and $j \in \mathrm{An}(k)$, $b_{ji} \geq \frac{b_{jk} b_{ki}}{b_{kk}} > 0$. Indeed, $b_{ji} \geq \frac{b_{jk} b_{ki}}{b_{kk}}$ holds for all $i, j, k \in V$.*

We learn immediately from (2.1.3) that $c_{ki} X_k \leq X_i$ for all $i \in V$ and $k \in \mathrm{pa}(i)$. From Corollary 2.3.12 we find such inequalities also for components, whose nodes are not connected by an edge but by a path of arbitrary length.

Corollary 2.3.13. *For all $i \in V$ and $j \in \mathrm{An}(i)$ we have $\frac{b_{ji}}{b_{jj}} X_j \leq X_i$.*

Proof. Note that $\mathrm{An}(j) \subseteq \mathrm{An}(i)$. Using the max-linear representation (2.2.1) of X_i and X_j as well as Corollary 2.3.12, we obtain

$$X_i = \bigvee_{\ell \in \mathrm{An}(i)} b_{\ell i} Z_\ell \geq \bigvee_{\ell \in \mathrm{An}(j)} b_{\ell i} Z_\ell \geq \bigvee_{\ell \in \mathrm{An}(j)} \frac{b_{\ell j} b_{ji}}{b_{jj}} Z_\ell = \frac{b_{ji}}{b_{jj}} \bigvee_{\ell \in \mathrm{An}(j)} b_{\ell j} Z_\ell = \frac{b_{ji}}{b_{jj}} X_j.$$

□

2.4 ML coefficients leading to a recursive ML model on a given DAG

Recall the definition of a ML model given in (2.1.4). From Theorem 2.2.2 we know that every recursive ML model is max-linear. In this section we provide necessary and sufficient conditions on a ML model to be a recursive ML model on a given DAG $\mathcal{D}$.

It can be shown that every ML model which is a recursive SEM as given in (2.1.1) with unspecified functions $f_1, \ldots, f_d$ must be a recursive ML model. That a recursive ML model is also a recursive SEM follows immediately from its recursive definition. To summarize, a ML model can be represented as a recursive SEM (2.1.1) on a DAG $\mathcal{D}$ if and only if it has a recursive ML representation (2.1.3) relative to the same DAG $\mathcal{D}$.

Motivated by Remark 2.2.3(i), in what follows we assume that $\mathrm{sgn}(B)$ is the reachability matrix R of $\mathcal{D}$. In our investigation the DAG with the minimum number of edges, such that $R = \mathrm{sgn}(B)$, will play an important role. This has already been indicated in Example 2.3.3.

We give a general definition of the DAG with the minimum number of edges that represents the same reachability relation as a given DAG.

Definition 2.4.1. Let $\mathcal{D} = (V, E)$ be a DAG. The DAG $\mathcal{D}^{\mathrm{tr}} = (V, E^{\mathrm{tr}})$ is the *transitive reduction of* $\mathcal{D}$ if the following holds:

(a) $\mathcal{D}^{\mathrm{tr}}$ has a path from node j to node i if and only if $\mathcal{D}$ has a path from j to i, and

(b) there is no graph with less edges than $\mathcal{D}^{\mathrm{tr}}$ satisfying condition (a). □

Since we work with finite DAGs throughout, the transitive reduction is unique and is also a subgraph of the original DAG. The transitive reduction of a DAG can be obtained by successively examining its edges in any order and deleting an edge $k \to i$ if it contains a path from k to i which does not include this edge. For these properties and further details, see e.g. Aho et al. [1]. In what follows we need the notion of $\mathrm{pa}^{\mathrm{tr}}(i)$, the parents of i in $\mathcal{D}^{\mathrm{tr}}$.

We present necessary and sufficient conditions on B to be the ML coefficient matrix of a recursive ML model on $\mathcal{D}$.

Theorem 2.4.2. *Let $\mathcal{D} = (V, E)$ be a DAG with reachability matrix R and $\boldsymbol{X}$ a ML model as in* (2.1.4) *with ML coefficient matrix B such that* $\mathrm{sgn}(B) = R$. *Define*

$$A := \mathrm{diag}(b_{11}, \ldots, b_{dd}) \quad \textit{and} \quad A_0 := \Big(\frac{b_{ij}}{b_{ii}} \mathbb{1}_{\mathrm{pa}(j)}(i) \Big)_{d \times d}.$$

Then $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$ if and only if the following fixed point equation holds:

$$B = A \vee B \odot A_0, \tag{2.4.1}$$

where $\odot$ is the matrix product from (2.2.2). In this case,

$$X_i = \bigvee_{k \in \mathrm{pa}(i)} \frac{b_{ki}}{b_{kk}} X_k \vee b_{ii} Z_i, \quad i = 1, \dots, d.$$

Proof. First we investigate the fixed point equation (2.4.1) and compute the ji-th entry of $B \odot A_0$. By definition, together with $\mathrm{sgn}(B) = R$, it is equal to

$$\bigvee_{k=1}^{d} \frac{b_{jk} b_{ki}}{b_{kk}} \mathbb{1}_{\mathrm{pa}(i)}(k) = \bigvee_{k \in \mathrm{De}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}}.$$

We have $\mathrm{De}(j) \cap \mathrm{pa}(i) = \varnothing$ for $j \in V \setminus \mathrm{an}(i)$ and $\mathrm{De}(j) \cap \mathrm{pa}(i) = \mathrm{de}(j) \cap \mathrm{pa}(i)$ for $j \in \mathrm{an}(i) \setminus \mathrm{pa}(i)$. Moreover, for $j \in \mathrm{pa}^{\mathrm{tr}}(i)$ using that $\mathrm{de}(j) \cap \mathrm{pa}(i) = \varnothing$, we obtain $\mathrm{De}(j) \cap \mathrm{pa}(i) = \{j\}$. Thus, taking also the matrix A into account, (2.4.1) is equivalent to

$$b_{ji} = \begin{cases} 0, & \text{if } j \in V \setminus \mathrm{An}(i), \\ b_{ii}, & \text{if } j = i, \\ \displaystyle\bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}}, & \text{if } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i), \\ \displaystyle b_{ji} \vee \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}}, & \text{if } j \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i), \\ b_{ji}, & \text{if } j \in \mathrm{pa}^{\mathrm{tr}}(i) \end{cases}$$

for all $i, j \in V$. In this equation the first row is automatically satisfied, since $R = \mathrm{sgn}(B)$, also the second and the last one hold trivially. To summarize, the fixed point equation (2.4.1) is satisfied if and only if for all $i \in V$ the following identities hold:

$$b_{ji} = \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \qquad \text{for all } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i), \tag{2.4.2}$$

$$b_{ji} = b_{ji} \vee \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \quad \text{for all } j \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i). \tag{2.4.3}$$

Thus it suffices to show that $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$ if and only if (2.4.2) and (2.4.3) hold for all $i \in V$.

First assume that $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$, and let $i \in V$ and $j \in \mathrm{an}(i)$. Since every path from j to i passes through at least one parent node of i, there must be a max-weighted path from j to i passing through some node in $\mathrm{pa}(i)$. Using (2.3.1) with $U = \mathrm{pa}(i)$ and noting that $j \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i) = \mathrm{De}(j) \cap \mathrm{pa}(i)$, we find for $j \in \mathrm{an}(i) \setminus \mathrm{pa}(i)$ Eq. (2.4.2) and for $j \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i)$ Eq. (2.4.3).

For the converse statement, assume that (2.4.2) and (2.4.3) hold. For $j \in \mathrm{pa}^{\mathrm{tr}}(i)$ we have $\mathrm{de}(j) \cap \mathrm{pa}(i) = \varnothing$, such that the right-hand side of (2.4.3) is equal to b_{ji}. Thus (2.4.3) holds for

all $j \in \mathrm{pa}(i)$. Since $\mathrm{sgn}(B) = R$, we have $X_i = \bigvee_{j=1}^{d} b_{ji} Z_j = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j$. We split up the index set and use (2.4.2) in the first place and (2.4.3) for all $j \in \mathrm{pa}(i)$ in the second place to obtain

$$\begin{aligned} X_i &= \bigvee_{j \in \mathrm{an}(i) \setminus \mathrm{pa}(i)} b_{ji} Z_j \vee \bigvee_{j \in \mathrm{pa}(i)} b_{ji} Z_j \vee b_{ii} Z_i \\ &= \bigvee_{j \in \mathrm{an}(i) \setminus \mathrm{pa}(i)} \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j \vee \bigvee_{j \in \mathrm{pa}(i)} b_{ji} Z_j \vee \bigvee_{j \in \mathrm{pa}(i)} \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j \vee b_{ii} Z_i \\ &= \bigvee_{j \in \mathrm{an}(i)} \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j \vee \bigvee_{j \in \mathrm{pa}(i)} b_{ji} Z_j \vee b_{ii} Z_i. \end{aligned}$$

Interchanging the first two maximum operators by (2.A.1) yields

$$\begin{aligned} X_i &= \bigvee_{k \in \mathrm{pa}(i)} \bigvee_{j \in \mathrm{an}(k)} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j \vee \bigvee_{k \in \mathrm{pa}(i)} b_{ki} Z_k \vee b_{ii} Z_i \\ &= \bigvee_{k \in \mathrm{pa}(i)} \frac{b_{ki}}{b_{kk}} \Big(\bigvee_{j \in \mathrm{an}(k)} b_{jk} Z_j \vee b_{kk} Z_k \Big) \vee b_{ii} Z_i \\ &= \bigvee_{k \in \mathrm{pa}(i)} \frac{b_{ki}}{b_{kk}} X_k \vee b_{ii} Z_i. \end{aligned}$$

□

In the proof of Theorem 2.4.2 we have shown that, under the required conditions, the fixed point equation (2.4.1) holds if and only if (2.4.2) and (2.4.3) hold for all nodes. We summarize this in part (a) of the following corollary. Part (b) has also been verified in the proof of Theorem 2.4.2. The final statement is based on the fact that for $k \in \mathrm{pa}(i)$ we have $\mathrm{de}(k) \cap \mathrm{pa}(i) = \varnothing$ if and only if $k \in \mathrm{pa}^{\mathrm{tr}}(i)$.

Corollary 2.4.3. *(a) Assume the situation of Theorem 2.4.2. Then $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$ if and only if for every $i \in V$,*

$$b_{ji} = \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \quad \text{for all } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i), \tag{2.4.4}$$

$$b_{ji} \geq \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \quad \text{for all } j \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i). \tag{2.4.5}$$

(b) Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with ML coefficient matrix B. Then for every $i \in V$ and $k \in \mathrm{pa}(i)$,

$$b_{ki} \geq \bigvee_{\ell \in \mathrm{de}(k) \cap \mathrm{pa}(i)} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}}.$$

Moreover, the right-hand side is equal to zero if and only if $k \in \mathrm{pa}^{\mathrm{tr}}(i)$, and in this case the inequality is strict.

By (2.4.4) and (2.4.5) exactly those ML coefficients b_{ji} for $i \in V$ and $j \in \mathrm{an}(i)$, such that $j \to i$ is an edge in $\mathcal{D}^{\mathrm{tr}}$, do not have to meet any specific conditions apart from being positive.

In summary, given a DAG $\mathcal{D}$ with d nodes, both Theorem 2.4.2 and Corollary 2.4.3(a) characterize all ML coefficient matrices of any recursive ML model possible on $\mathcal{D}$ as all nonnegative $d\times d$ matrices that are weighted reachability matrices of $\mathcal{D}$ and satisfy (2.4.1), equivalently (2.4.4) and (2.4.5). If we can verify these two properties for a nonnegative $d\times d$ matrix B, then it is the ML coefficient matrix of a recursive ML model on $\mathcal{D}$, and weights in its representation (2.1.3) are given by $c_{ki} = \frac{b_{ki}}{b_{kk}}$ for $k \in \mathrm{pa}(i)$ and $c_{ii} = b_{ii}$.

2.5 Graph reduction for a recursive ML model

From Proposition 2.3.7 we know that every component of a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D} = (V, E)$ satisfies (2.1.3) on a subgraph of $\mathcal{D}$. These subgraphs, however, usually vary from one component to another. On the other hand, we know from Example 2.3.3 that the whole vector $\boldsymbol{X}$ may also be a recursive ML model on a subgraph of $\mathcal{D}$. This raises the question of finding the smallest subgraph of $\mathcal{D}$ such that $\boldsymbol{X}$ is a recursive ML model on this DAG. We define and characterize this minimum DAG before we point out its prominent role in the class of all DAGs representing $\boldsymbol{X}$ in the sense of (2.1.3).

Definition 2.5.1. Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with ML coefficient matrix B. We call the DAG

$$\mathcal{D}^B = (V, E^B) := \Big(V, \Big\{(k,i) \in E : b_{ki} > \bigvee_{\ell \in \mathrm{de}(k)\cap \mathrm{pa}(i)} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}}\Big\}\Big) \tag{2.5.1}$$

the *minimum max-linear (ML) DAG of* $\boldsymbol{X}$. □

We summarize some properties of $\mathcal{D}^B$ as follows.

Remark 2.5.2. (i) The minimum ML DAG $\mathcal{D}^B = (V, E^B)$ is a subgraph of the original DAG $\mathcal{D} = (V, E)$. Observe from Corollary 2.4.3(b) that the transitive reduction $\mathcal{D}^{\mathrm{tr}} = (V, E^{\mathrm{tr}})$ of $\mathcal{D}$ is also a subgraph of $\mathcal{D}$. In summary, we have $E^{\mathrm{tr}} \subseteq E^B \subseteq E$. This implies that the DAGs $\mathcal{D}^B$ and $\mathcal{D}$ have the same reachability matrix, which is $\mathrm{sgn}(B)$ by Remark 2.2.3(i).

(ii) By Theorem 2.3.10(b) the minimum ML DAG $\mathcal{D}^B$ contains exactly those edges $k \to i$ of $\mathcal{D}$, where no max-weighted path from k to i passes through some node in $\mathrm{pa}(i) \setminus \{k\}$. This means that $\mathcal{D}^B$ has an edge $k \to i$ if and only if it is the only max-weighted path from k to i in $\mathcal{D}$. The DAG $\mathcal{D}^B$ can be obtained from $\mathcal{D}$ by deleting an edge $k \to i$ if $\mathcal{D}$ contains a max-weighted path from k to i which does not include this edge. Note the analogy to finding the transitive reduction $\mathcal{D}^{\mathrm{tr}}$ of $\mathcal{D}$ described below Definition 2.4.1. An algorithm is by comparison of ML coefficients and motivated by Corollary 2.4.3(b): for all $i \in V$ and $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i)$ remove the edge $k \to i$ from $\mathcal{D}$ if

$$b_{ki} = \bigvee_{\ell \in \mathrm{de}(k)\cap \mathrm{pa}(i)} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}}.$$

□

The method described in Remark 2.5.2(ii) determines $\mathcal{D}^B$ from $\mathcal{D}$ and the ML coefficient matrix B. Indeed, we can also identify $\mathcal{D}^B$ directly from B without knowing $\mathcal{D}$.

Theorem 2.5.3. *Let $\boldsymbol{X}$ be a recursive ML model with ML coefficient matrix B. Then the minimum ML DAG of $\boldsymbol{X}$ can be represented as*

$$\mathcal{D}^B = \Big(V, \Big\{(k,i) \in V \times V : k \neq i \text{ and } b_{ki} > \bigvee_{\substack{\ell=1 \\ \ell \neq i,k}}^{d} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}}\Big\}\Big); \tag{2.5.2}$$

in particular, $\mathcal{D}^B$ is identifiable from B.

Proof. Let $\mathcal{D}$ be a DAG which represents $\boldsymbol{X}$ in the sense of (2.1.3). Such a DAG exists by the definition of a recursive ML model. We show that the edge set in (2.5.2) coincides with E^B as defined in (2.5.1). Assume first that (k,i) is contained in the edge set in (2.5.2). Since $\operatorname{sgn}(B)$ is the reachability matrix of $\mathcal{D}$ (cf. Remark 2.2.3(i)), we have

$$b_{ki} > \bigvee_{\substack{\ell=1 \\ \ell \neq i,k}}^{d} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}} = \bigvee_{\ell \in \operatorname{de}(k) \cap \operatorname{an}(i)} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}}. \tag{2.5.3}$$

Since the right-hand side of (2.5.3) is nonnegative, we must have $b_{ki} > 0$ and, hence, $k \in \operatorname{an}(i)$. By Theorem 2.3.10(b) no max-weighted path from k to i passes through some node in $V \setminus \{i,k\}$. Thus the edge $k \to i$ must be the only max-weighted path from k to i and, hence, by Remark 2.5.2(ii) it must be an edge in E^B as in (2.5.1).

For the converse, let $(k,i) \in E^B$. Since by Remark 2.5.2(ii) this edge is the only max-weighted path from k to i, no max-weighted path passes through some node in $V \setminus \{i,k\}$. This is by Theorem 2.3.10(b) equivalent to (2.5.3) and (k,i) belongs to the edge set in (2.5.2). □

We characterize all DAGs and specify all weights such that $\boldsymbol{X}$ satisfies (2.1.3). The minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$ is the smallest DAG of this kind and has unique weights in representation (2.1.3) in the sense that all irrelevant weights are set to zero. We can add any edge $k \to i$ into $\mathcal{D}^B$ with weight $c_{ki} \in (0, \frac{b_{ki}}{b_{kk}}]$ representing $\boldsymbol{X}$ again in the sense of (2.1.3) as long as the graph represents the same reachability relation as $\mathcal{D}^B$. As a consequence, to find B by a path analysis as described in (2.1.5) and (2.1.6), it suffices to know $\mathcal{D}^B$ and the weights in representation (2.1.3) relative to $\mathcal{D}^B$.

Theorem 2.5.4. *Let $\boldsymbol{X}$ be a recursive ML model with ML coefficient matrix B. Let further $\mathcal{D}^B = (V, E^B)$ be the minimum ML DAG of $\boldsymbol{X}$ and $\operatorname{pa}^B(i)$ the parents of node i in $\mathcal{D}^B$.*

(a) The minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$ is the DAG with the minimum number of edges such that $\boldsymbol{X}$ satisfies (2.1.3)*. The weights in* (2.1.3) *are uniquely given by $c_{ii} = b_{ii}$ and $c_{ki} = \frac{b_{ki}}{b_{kk}}$ for $i \in V$ and $k \in \operatorname{pa}^B(i)$.*

(b) Every DAG with node set V that has at least the edges of $\mathcal{D}^B$ and the same reachability matrix as $\mathcal{D}^B$ represents $\boldsymbol{X}$ in the sense of (2.1.3) *with weights given for all $i \in V$ by*

$$c_{ii} = b_{ii}, \quad c_{ki} = \frac{b_{ki}}{b_{kk}} \text{ for } k \in \operatorname{pa}^B(i), \quad \text{and} \quad c_{ki} \in \Big(0, \frac{b_{ki}}{b_{kk}}\Big] \text{ for } k \in \operatorname{pa}(i) \setminus \operatorname{pa}^B(i).$$

There are no further DAGs and weights such that $\boldsymbol{X}$ has representation (2.1.3).

Proof. (a) Let $\mathcal{D}$ be a DAG and c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$ weights such that $\boldsymbol{X}$ has representation (2.1.3). By Remark 2.5.2(i) $\mathcal{D}^B$ is a subgraph of $\mathcal{D}$.

First we prove that $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}^B$ with weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}^B(i)$ by showing that all components of $\boldsymbol{X}$ coincide with those of the recursive ML submodel of $\boldsymbol{X}$ induced by $\mathcal{D}^B$ (see Definition 2.3.5). By Remark 2.3.6(iv) it suffices to verify for all $i \in V$ and $j \in \mathrm{an}(i)$ that $\mathcal{D}^B$ has one in $\mathcal{D}$ max-weighted path from j to i. Among all max-weighted paths from j to i in $\mathcal{D}$, let p be one with maximum length, and assume that p includes an edge, say $k \to \ell$, which is not contained in $\mathcal{D}^B$. The DAG $\mathcal{D}$ has by Remark 2.5.2(ii), however, a max-weighted path p_1 from k to ℓ which does not include the edge $k \to \ell$. Note that p_1 consists of more edges than the path $[k \to \ell]$. Thus, by replacing in p the edge $k \to \ell$ by p_1, we obtain by Remark 2.3.4(iii) a max-weighted path from j to i consisting of more edges than p. Since this is a contradiction to the fact that p has maximum length among all max-weighted paths from j to i, p must be in $\mathcal{D}^B$.

Since every edge $k \to i$ in $\mathcal{D}^B$ is by Remark 2.5.2(ii) the only max-weighted path from k to i in $\mathcal{D}$, the weights in (2.1.3) are uniquely given, and we have by Definition 2.3.1 and (2.1.5) that $b_{ki} = c_{kk}c_{ki} = b_{kk}c_{ki}$, which implies $c_{ki} = \frac{b_{ki}}{b_{kk}}$. For the same reason there cannot be a DAG with less edges than $\mathcal{D}^B$ such that $\boldsymbol{X}$ has representation (2.1.3).

(b) First we show that $\boldsymbol{X}$ satisfies (2.1.3) relative to a DAG $\mathcal{D}$ with the properties and weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$ (the parents in $\mathcal{D}$). Note that the DAG $\mathcal{D}^B$ is a subgraph of $\mathcal{D}$ and both DAGs have the same reachability relation. Since $\boldsymbol{X}$ is by part (a) a recursive ML model on $\mathcal{D}^B$, we may use Corollary 2.3.13 with the ancestors in $\mathcal{D}^B$: for every $i \in V$ and $k \in \mathrm{pa}(i)$, since k is an ancestor of i in $\mathcal{D}^B$ and $\frac{b_{ki}}{b_{kk}} \geq c_{ki}$, we have

$$X_i \geq \frac{b_{ki}}{b_{kk}} X_k \geq c_{ki} X_k.$$

With this we obtain from representation (2.1.3) of X_i relative to $\mathcal{D}^B$ that

$$X_i = \bigvee_{k \in \mathrm{pa}^B(i)} c_{ki} X_k \vee c_{ii} Z_i = \bigvee_{k \in \mathrm{pa}^B(i)} c_{ki} X_k \vee \bigvee_{k \in \mathrm{pa}(i) \setminus \mathrm{pa}^B(i)} c_{ki} X_k \vee c_{ii} Z_i,$$

which is (2.1.3) relative to $\mathcal{D}$.

It remains to show that there are no further DAGs and weights such that $\boldsymbol{X}$ has representation (2.1.3). By Remark 2.5.2(i) every DAG that represents $\boldsymbol{X}$ in the sense of (2.1.3) must have the same reachability matrix as $\mathcal{D}^B$ and must contain at least the edges of $\mathcal{D}^B$. By (2.1.5) and (2.1.6) the weights in representation (2.1.3) of $\boldsymbol{X}$ have to satisfy $c_{ki} \leq \frac{b_{ki}}{b_{kk}}$ for all $i \in V$ and $k \in \mathrm{pa}(i)$. The statement follows, since the weights c_{ki} are by part (a) uniquely with respect to $\mathcal{D}^B$. □

As explained before Theorem 2.5.4, we can add edges into $\mathcal{D}^B$ while keeping the same reachability relation and still having representation (2.1.3) for $\boldsymbol{X}$. In what follows we will use the DAG with the maximum number of edges with this property.

Definition 2.5.5. Let $\mathcal{D} = (V, E)$ be a DAG. The *transitive closure* $\mathcal{D}^{\mathrm{tc}} = (V, E^{\mathrm{tc}})$ *of* $\mathcal{D}$ is the DAG that has an edge $j \to i$ if and only if $\mathcal{D}$ has a path from j to i. □

The transitive reduction is essentially the inverse operation of the transitive closure: for the transitive reduction one reduces the number of edges and for the transitive closure one adds edges while maintaining the identical reachability relation. The transitive reduction of a DAG $\mathcal{D}$ is a subgraph of $\mathcal{D}$, and $\mathcal{D}$ is again a subgraph of the transitive closure. Moreover, all DAGs with the same reachability matrix have the same transitive reduction and the same transitive closure, and every node has in all such DAGs the same ancestors and descendants.

The following is a consequence of Theorem 2.5.4(b) and Remark 2.2.3(i).

Corollary 2.5.6. *The recursive ML model* $\boldsymbol{X}$ *is also a recursive ML model on the transitive closure of every DAG with reachability matrix* $\operatorname{sgn}(B)$.

We use this corollary to obtain necessary and sufficient conditions on a ML coefficient matrix B as in (2.1.4) to be the ML coefficient matrix of a recursive ML model. In contrast to Theorem 2.4.2 and Corollary 2.4.3(a), we do not require that B belongs to a specific given DAG.

Theorem 2.5.7. *Let* $\boldsymbol{X}$ *be a ML model as in* (2.1.4) *with ML coefficient matrix* B *such that* $\operatorname{sgn}(B)$ *is the reachability matrix of some DAG. Define*

$$A := \operatorname{diag}(b_{11}, \dots, b_{dd}), \quad B_0 := \Big(\frac{b_{ij}}{b_{ii}}\Big)_{d\times d}, \quad \textit{and} \quad A_0^{\mathrm{tc}} := B_0 - \mathrm{id}_{d\times d},$$

where $\mathrm{id}_{d\times d}$ *denotes the identity matrix. Then* $\boldsymbol{X}$ *is a recursive ML model if and only if the following fixed point equation holds:*

$$B = B \odot B_0, \quad \textit{which is equivalent to} \quad B = A \vee B \odot A_0^{\mathrm{tc}}, \tag{2.5.4}$$

where $\odot$ *is the matrix product from* (2.2.2).

Proof. Let $\mathcal{D}^{\mathrm{tc}}$ be the transitive closure of a DAG with node set $V = \{1, \dots, d\}$ and reachability matrix $\operatorname{sgn}(B)$. For $i \in V$ we denote by $\mathrm{pa}(i)$ and $\mathrm{an}(i)$ the parents and ancestors of node i in $\mathcal{D}^{\mathrm{tc}}$, respectively, and observe from the definition of $\mathcal{D}^{\mathrm{tc}}$ that $\mathrm{an}(i) = \mathrm{pa}(i)$ for all $i \in V$.

First we show that $\boldsymbol{X}$ is a recursive ML model if and only if the fixed point equation $B = A \vee B \odot A_0^{\mathrm{tc}}$ holds. By Corollary 2.5.6 $\boldsymbol{X}$ is a recursive ML model if and only if it is a recursive ML model on $\mathcal{D}^{\mathrm{tc}}$. By Theorem 2.4.2 it suffices to show that A_0^{tc} is equal to the weighted adjacency matrix $A_0 = \Big(\frac{b_{ij}}{b_{ii}} \mathbb{1}_{\mathrm{pa}(j)}(i)\Big)_{d\times d}$. Since B_0 is a weighted reachability matrix of $\mathcal{D}^{\mathrm{tc}}$, we obtain

$$A_0^{\mathrm{tc}} = B_0 - \mathrm{id}_{d\times d} = \Big(\frac{b_{ij}}{b_{ii}} \mathbb{1}_{\mathrm{an}(j)}(i)\Big)_{d\times d} = \Big(\frac{b_{ij}}{b_{ii}} \mathbb{1}_{\mathrm{pa}(j)}(i)\Big)_{d\times d}.$$

It remains to show that $B \odot B_0 = A \vee B \odot A_0^{\mathrm{tc}}$. By the definition of the matrix product $\odot$ the

ji-th entry of $A \vee B \odot A_0^{\mathrm{tc}}$ is equal to

$$\begin{aligned} b_{ji}\mathbb{1}_{\{i\}}(j) \vee \bigvee_{k=1}^{d} b_{jk}\Big(\frac{b_{ki}}{b_{kk}} - \mathbb{1}_{\{i\}}(k)\Big) &= b_{ji}\mathbb{1}_{\{i\}}(j) \vee \bigvee_{\substack{k=1\\k\neq i}}^{d} \frac{b_{jk}b_{ki}}{b_{kk}} \\ &= b_{ji}\mathbb{1}_{\{i\}}(j) \vee \bigvee_{\substack{k=1\\k\neq i,j}}^{d} \frac{b_{jk}b_{ki}}{b_{kk}} \vee b_{ji}\mathbb{1}_{V\setminus\{i\}}(j) \\ &= \bigvee_{k=1}^{d} \frac{b_{jk}b_{ki}}{b_{kk}}, \end{aligned}$$

which is the $ji-$th entry of the matrix $B \odot B_0$. □

A nonnegative symmetric matrix is by Theorem 2.5.7 the ML coefficient matrix of a recursive ML model if and only if it is a weighted reachability matrix of a DAG and satisfies (2.5.4). Assume that we have verified these properties for a matrix B. In order to find now all recursive ML models which have ML coefficient matrix B, we can first use (2.5.2) to derive the minimum ML DAG $\mathcal{D}^B$ from B and then Theorem 2.5.4(b) to find all DAGs and weights as in (2.1.3) such that (2.1.6) holds.

2.6 Backward and forward information in a recursive ML model

We investigate relations between the components of a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D} = (V, E)$ with ML coefficient matrix B. More precisely, we apply our previous results to identify those components of $\boldsymbol{X}$ which provide maximal information on some other component.

We know already from Corollary 2.3.13 that $X_i \le \frac{b_{ii}}{b_{i\ell}} X_\ell$ for all $i \in V$ and $\ell \in \mathrm{De}(i)$ so that for some node set $U \subseteq V$ and all $i \in V$,

$$\bigvee_{j\in \mathrm{An}(i)\cap U} \frac{b_{ji}}{b_{jj}} X_j \le X_i \le \bigwedge_{\ell\in \mathrm{De}(i)\cap U} \frac{b_{ii}}{b_{i\ell}} X_\ell. \tag{2.6.1}$$

The values of the bounds in (2.6.1) can often be found as the maximum and minimum over a smaller number of nodes in U. We illustrate this by the following example.

Example 2.6.1. [Continuation of Examples 2.2.1 and 2.3.8: bounds]
For $U = \{1, 2\}$ and $i = 4$ we find by (2.6.1) the lower bound

$$\frac{b_{14}}{b_{11}} X_1 \vee \frac{b_{24}}{b_{22}} X_2 \le X_4. \tag{2.6.2}$$

We discuss the lower bound and distinguish between two cases.

First assume that the path $[1 \to 2 \to 4]$ is max-weighted, which is by Theorem 2.3.10(a) equivalent to $b_{14} = \frac{b_{12}b_{24}}{b_{22}}$. From Corollary 2.3.13 or (2.6.1) we obtain

$$\frac{b_{12}}{b_{11}} X_1 \le X_2, \quad \text{equivalently} \quad \frac{b_{14}}{b_{11}} X_1 \le \frac{b_{24}}{b_{22}} X_2.$$

Therefore, the lower bound of X_4 in (2.6.2) is always $\frac{b_{24}}{b_{22}}X_2$.

Now assume that the path $[1 \to 2 \to 4]$ is not max-weighted. Since this is the only path from 1 to 4 passing through node 2, this is by Theorem 2.3.10(b) equivalent to $b_{14} > \frac{b_{12}b_{24}}{b_{22}}$. From the max-linear representation (2.2.1) of X_1 and X_2 we have $\frac{b_{24}}{b_{22}}X_2 < \frac{b_{14}}{b_{11}}X_1$ if and only if

$$\frac{b_{12}b_{24}}{b_{22}}Z_1 \vee b_{24}Z_2 < b_{14}Z_1, \quad \text{equivalently} \quad b_{24}Z_2 < b_{14}Z_1.$$

The event $\{b_{24}Z_2 < b_{14}Z_1\}$ has positive probability, since Z_1 and Z_2 are independent with support $\mathbb{R}_+$, giving $\frac{b_{14}}{b_{11}}X_1$ as lower bound. But also the event $\{\frac{b_{14}}{b_{11}}X_1 \leq \frac{b_{24}}{b_{22}}X_2\}$ has positive probability, giving the lower bound $\frac{b_{24}}{b_{22}}X_2$. Thus only in the first case the number of nodes in the lower bound in (2.6.1) can be reduced. □

We will find that a node $j \in \mathrm{An}(i) \cap U$ is relevant for the lower bound in (2.6.1) if no max-weighted path from j to i passes through some other node in U. Observe that this includes the observation made in Example 2.6.1. The nodes in the upper bound of (2.6.1) have a similar characterization. We present a formal definition of these particular ancestors and descendants, characterize them below in Lemma 2.6.3, and give an example afterwards.

Definition 2.6.2. Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$, $U \subseteq V$ and $i \in V$.

(a) We call a node $j \in \mathrm{An}(i) \cap U$ *lowest max-weighted ancestor of i in U* if no max-weighted path from j to i passes through some node in $U \setminus \{j\}$. We denote the set of the lowest max-weighted ancestors of i in U by $\mathrm{An}^U_{\mathrm{low}}(i)$.

(b) We call a node $\ell \in \mathrm{De}(i) \cap U$ *highest max-weighted descendant of i in U* if no max-weighted path from i to ℓ passes through some node in $U \setminus \{\ell\}$. We denote the set of the highest max-weighted descendants of i in U by $\mathrm{De}^U_{\mathrm{high}}(i)$. □

For $i \in U$ we find that the only lowest max-weighted ancestor and the only highest max-weighted descendant of i in U is the node i itself. For $i \in U^c = V \setminus U$ a simple characterization of $\mathrm{An}^U_{\mathrm{low}}(i)$ and $\mathrm{De}^U_{\mathrm{high}}(i)$ is given next; this allows us to identify these nodes via the ML coefficient matrix of $\boldsymbol{X}$.

Lemma 2.6.3. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$, $U \subseteq V$ and $i \in V$.*

(a) If $i \in U$, then $\mathrm{An}^U_{\mathrm{low}}(i) = \mathrm{De}^U_{\mathrm{high}}(i) = \{i\}$.

(b) If $i \in U^c$, then

$$\mathrm{An}^U_{\mathrm{low}}(i) = \Big\{ j \in \mathrm{an}(i) \cap U : b_{ji} > \bigvee_{k \in \mathrm{de}(j) \cap U \cap \mathrm{an}(i)} \frac{b_{jk}b_{ki}}{b_{kk}} \Big\}, \tag{2.6.3}$$

$$\mathrm{De}^U_{\mathrm{high}}(i) = \Big\{ \ell \in \mathrm{de}(i) \cap U : b_{i\ell} > \bigvee_{k \in \mathrm{de}(i) \cap U \cap \mathrm{an}(\ell)} \frac{b_{ik}b_{k\ell}}{b_{kk}} \Big\}. \tag{2.6.4}$$

Proof. (a) follows immediately from the definition.

(b) Since $i \in U^c$, we have by Definition 2.6.2(a) that $\mathrm{An}^U_{\mathrm{low}}(i) \subseteq \mathrm{an}(i) \cap U$. For $j \in \mathrm{an}(i) \cap U$ we know from Theorem 2.3.10(b) that no max-weighted path from j to i passes through some node in $U \setminus \{j\}$ if and only if

$$b_{ji} > \bigvee_{k \in \mathrm{de}(j) \cap U \cap \mathrm{an}(i)} \frac{b_{jk} b_{ki}}{b_{kk}},$$

where we have used that $i \in U^c$. Similarly, we obtain (2.6.4). □

Example 2.6.4. [Continuation of Examples 2.2.1, 2.3.8, and 2.6.1: $\mathrm{An}^U_{\mathrm{low}}(4)$]
In order to find the lowest max-weighted ancestors of node 4 in $U = \{1, 2\}$, first observe that the only max-weighted path $[2 \to 4]$ from 2 to 4 does not pass through any node in $U \setminus \{2\}$. Therefore, we have by Definition 2.6.2(a) that $2 \in \mathrm{An}^U_{\mathrm{low}}(4)$. For node 1 we consider – as in Example 2.6.1 – two cases and use (2.6.3):

(1) If $b_{14} = \frac{b_{12} b_{24}}{b_{22}}$, then $\mathrm{An}^U_{\mathrm{low}}(4) = \{2\}$.

(2) If $b_{14} > \frac{b_{12} b_{24}}{b_{22}}$, then $\mathrm{An}^U_{\mathrm{low}}(4) = \{1, 2\}$.

Comparing this with Example 2.6.1 shows that the lower bound of X_4 in (2.6.2) is always realized by some lowest max-weighted ancestor of node 4 in U. □

We prove that the lower and upper bounds in (2.6.1) are always realized by some lowest max-weighted ancestor and highest max-weighted descendant in U, respectively. For the lower bound this is based on the fact that between all nodes and their ancestors in U there is always a max-weighted path which contains a lowest max-weighted ancestor in U. For the upper bound we use the existence of a max-weighted path between all nodes and their descendants in U that passes through some highest max-weighted descendant in U. Before we state the modified lower and upper bounds in Proposition 2.6.6, we provide a useful characterization for a path analysis, which includes these statements.

Lemma 2.6.5. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$. Furthermore, let $U \subseteq V$, $i \in V$, $j \in \mathrm{an}(i)$, and $\ell \in \mathrm{de}(i)$.*

(a) $\mathcal{D}$ has a max-weighted path from j to i passing through some node in U if and only if it has a max-weighted path from j to i passing through some node in $\mathrm{An}^U_{\mathrm{low}}(i)$.

(b) $\mathcal{D}$ has a max-weighted path from i to ℓ passing through some node in U if and only if it has a max-weighted path from i to ℓ passing through some node in $\mathrm{De}^U_{\mathrm{high}}(i)$.

Proof. We only show (a); part (b) can be proved analogously. Assume that a max-weighted path from j to i passes through some node in $\mathrm{An}^U_{\mathrm{low}}(i)$. Since $\mathrm{An}^U_{\mathrm{low}}(i) \subseteq U$, there is obviously also a max-weighted path from j to i that passes through some node in U.

For the converse, we may assume that $i \in U^c$, since by Lemma 2.6.3(a) $\mathrm{An}^U_{\mathrm{low}}(i) = \{i\}$ for $i \in U$ and hence every max-weighted path contains a node in $\mathrm{An}^U_{\mathrm{low}}(i)$. Among all max-weighted paths from j to i let p be one with maximum number of nodes in U. Denote by k_1 the lowest node on p contained in U; i.e., the subpath of p from k_1 to i contains no other node of U. Assume

that $k_1 \notin \mathrm{An}^U_{\mathrm{low}}(i)$. Since $k_1 \in U$ and $i \in U^c$, there is by Definition 2.6.2(a) a max-weighted path p_1 from k_1 to i that passes through some node $k_2 \in U$ with $k_2 \neq k_1$. Thus, by replacing in p the subpath from k_1 to i by p_1, we obtain by Remark 2.3.4(iii) a max-weighted path from j to i containing more nodes in U than p. This is however a contradiction. Hence, $k_1 \in \mathrm{An}^U_{\mathrm{low}}(i)$, and p is a max-weighted path from j to i that passes through some node in $\mathrm{An}^U_{\mathrm{low}}(i)$. □

Proposition 2.6.6. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D} = (V, E)$ with ML coefficient matrix B. Let $U \subseteq V$ and $i \in V$. Then*

$$\bigvee_{j \in \mathrm{An}(i) \cap U} \frac{b_{ji}}{b_{jj}} X_j = \bigvee_{j \in \mathrm{An}^U_{\mathrm{low}}(i)} \frac{b_{ji}}{b_{jj}} X_j \quad \text{and} \quad \bigwedge_{\ell \in \mathrm{De}(i) \cap U} \frac{b_{ii}}{b_{i\ell}} X_\ell = \bigwedge_{\ell \in \mathrm{De}^U_{\mathrm{high}}(i)} \frac{b_{ii}}{b_{i\ell}} X_\ell. \tag{2.6.5}$$

Proof. Note from Definition 2.6.2(a) that $\mathrm{An}^U_{\mathrm{low}}(i) \subseteq \mathrm{An}(i) \cap U$. To show the first equality, take some $k \in (\mathrm{An}(i) \cap U) \setminus \mathrm{An}^U_{\mathrm{low}}(i)$. Observe from Lemma 2.6.3(a) that $k \neq i$ and, hence, $k \in \mathrm{an}(i) \cap U$. By Lemma 2.6.5(a) there must be a max-weighted path from k to i which passes through some node $j \in \mathrm{An}^U_{\mathrm{low}}(i)$. By (2.3.1) and Corollary 2.3.13, we obtain

$$\frac{b_{ki}}{b_{kk}} X_k = \frac{b_{kj} b_{ji}}{b_{kk} b_{jj}} X_k \leq \frac{b_{ji}}{b_{jj}} X_j. \tag{2.6.6}$$

Since for all $k \in (\mathrm{An}(i) \cap U) \setminus \mathrm{An}^U_{\mathrm{low}}(i)$ there exists some $j \in \mathrm{An}^U_{\mathrm{low}}(i)$ such that (2.6.6) holds, the first equality of (2.6.5) follows. The second equality may be verified analogously. □

So far, for every component of $\boldsymbol{X}$, we have identified a lower and upper bound in terms of the components of $\boldsymbol{X}_U = (X_\ell, \ell \in U)$. However, we cannot say anything about the quality of the bounds. For example, we do not know in which situation a component attains one of the bounds. We clarify this by writing all components of $\boldsymbol{X}$ as max-linear functions of the components of $\boldsymbol{X}_U$ and certain noise variables. There are many such representations, since we can always include non-relevant ancestral components with appropriate ML coefficients as we know from Corollary 2.3.13. To find the relevant components of $\boldsymbol{X}_U$ and noise variables, we focus on those with the minimum number of components of $\boldsymbol{X}_U$ and the minimum number of noise variables. For $i \in V$ we denote by $\mathrm{an}^U_{\mathrm{nmw}}(i)$ the set of all $j \in \mathrm{an}(i)$ such that no max-weighted path from j to i passes through some node in U. By Theorem 2.3.10(b) we have

$$\mathrm{an}^U_{\mathrm{nmw}}(i) = \Big\{ j \in \mathrm{an}(i) : b_{ji} > \bigvee_{k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \Big\}. \tag{2.6.7}$$

Since $j \in \mathrm{an}(i) \setminus \mathrm{an}^U_{\mathrm{nmw}}(i)$ if and only if there is a max-weighted path from j to i passing through some node in U, we obtain from Theorem 2.3.10(a)

$$\mathrm{an}(i) \setminus \mathrm{an}^U_{\mathrm{nmw}}(i) = \Big\{ j \in \mathrm{an}(i) : b_{ji} = \bigvee_{k \in \mathrm{De}(j) \cap U \cap \mathrm{An}(i)} \frac{b_{jk} b_{ki}}{b_{kk}} \Big\}. \tag{2.6.8}$$

Theorem 2.6.7. *Let $\boldsymbol{X}$ be a recursive ML model on a DAG $\mathcal{D}$ with ML coefficient matrix B, and let $U \subseteq V$. Furthermore, let $\mathrm{An}^U_{\mathrm{low}}(i)$ be the lowest max-weighted ancestors of node i in U*

as in Definition 2.6.2(a), and define $\mathrm{An}^U_{\mathrm{nmw}}(i) := (\mathrm{an}^U_{\mathrm{nmw}}(i) \cup \{i\}) \cap U^c$. *Then for every* $i \in V$,

$$X_i = \bigvee_{k \in \mathrm{An}^U_{\mathrm{low}}(i)} \frac{b_{ki}}{b_{kk}} X_k \vee \bigvee_{j \in \mathrm{An}^U_{\mathrm{nmw}}(i)} b_{ji} Z_j. \tag{2.6.9}$$

This representation of X_i *as a max-linear function of the components of* $\boldsymbol{X}_U$ *and noise variables involves the minimum number of components of* $\boldsymbol{X}_U$ *and the minimum number of noise variables.*

Proof. We distinguish between nodes $i \in U$ and $i \in U^c$. For $i \in U$ we know from Lemma 2.6.3(a) that $\mathrm{An}^U_{\mathrm{low}}(i) = \{i\}$. Furthermore, we have $\mathrm{An}^U_{\mathrm{nmw}}(i) = \varnothing$, since $i \in U$ and every path, hence every max-weighted path, from some $j \in \mathrm{an}(i)$ to i passes through some node in U, namely i itself. Thus we obtain (2.6.9). The second statement is obvious.

Now assume that $i \in U^c$, and note that in this case $\mathrm{An}^U_{\mathrm{nmw}}(i) = \mathrm{an}^U_{\mathrm{nmw}}(i) \cup \{i\}$. Applying the first equality in (2.6.5) and (2.2.1) as well as (2.A.2) in a second step to interchange the first two maximum operators, we have

$$\bigvee_{k \in \mathrm{An}^U_{\mathrm{low}}(i)} \frac{b_{ki}}{b_{kk}} X_k = \bigvee_{k \in \mathrm{an}(i) \cap U} \frac{b_{ki}}{b_{kk}} \Big(\bigvee_{j \in \mathrm{An}(k)} b_{jk} Z_j \Big) = \bigvee_{j \in \mathrm{an}(i)} \bigvee_{k \in \mathrm{De}(j) \cap \mathrm{an}(i) \cap U} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j. \tag{2.6.10}$$

We split up the set $\mathrm{an}(i)$ into $\mathrm{an}^U_{\mathrm{nmw}}(i)$ and $\mathrm{an}(i) \setminus \mathrm{an}^U_{\mathrm{nmw}}(i)$ as well as the set $\mathrm{An}^U_{\mathrm{nmw}}(i)$ into $\mathrm{an}^U_{\mathrm{nmw}}(i)$ and $\{i\}$ to obtain that the right-hand side of (2.6.9) is equal to

$$\bigvee_{j \in \mathrm{an}(i) \setminus \mathrm{an}^U_{\mathrm{nmw}}(i)} \bigvee_{k \in \mathrm{De}(j) \cap \mathrm{an}(i) \cap U} \frac{b_{jk} b_{ki}}{b_{kk}} Z_j \vee \bigvee_{j \in \mathrm{an}^U_{\mathrm{nmw}}(i)} \Big(\bigvee_{k \in \mathrm{De}(j) \cap \mathrm{an}(i) \cap U} \frac{b_{jk} b_{ki}}{b_{kk}} \vee b_{ji} \Big) Z_j \vee b_{ii} Z_i.$$

Noting that $i \in U^c$ when using (2.6.8) and (2.6.7) yields

$$\bigvee_{j \in \mathrm{an}(i) \setminus \mathrm{an}^U_{\mathrm{nmw}}(i)} b_{ji} Z_j \vee \bigvee_{j \in \mathrm{an}^U_{\mathrm{nmw}}(i)} b_{ji} Z_j \vee b_{ii} Z_i = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j = X_i.$$

In order to verify that for $i \in U^c$ (2.6.9) is the representation of X_i with the minimum number of components of $\boldsymbol{X}_U$ and the minimum number of noise variables, we prove that each term on the right-hand side of (2.6.9) has to appear, since otherwise some noise variable Z_j in representation (2.2.1) of X_i would have a weight strictly less than b_{ji}. We compare the noise variables on the right-hand sides of (2.6.9) and (2.6.10). Since $b_{ii} Z_i$ does not appear in (2.6.10), it has to appear in (2.6.9). For $j \in \mathrm{an}^U_{\mathrm{nmw}}(i)$ it follows from (2.6.7) that if Z_j appears in (2.6.10), then with a coefficient strictly less than b_{ji}. The maximum over $\mathrm{An}^U_{\mathrm{nmw}}(i)$ must therefore appear in (2.6.9). Definition 2.6.2(a) implies that no max-weighted path from $j \in \mathrm{An}^U_{\mathrm{low}}(i)$ to i passes through some node in $\mathrm{de}(j) \cap \mathrm{an}(i) \cap U$. Thus observe from (2.6.10) and (2.3.2) that only the term $\frac{b_{ji}}{b_{jj}} X_j$ provides Z_j with the weight b_{ji} in (2.6.9) and the term $\frac{b_{ji}}{b_{jj}} X_j$ has to appear in (2.6.9). □

We use Theorem 2.6.7 to obtain for every component X_i of $\boldsymbol{X}$ a minimal representation in terms of the components of $\boldsymbol{X}_{\mathrm{pa}(i)}$ and noise variables.

Corollary 2.6.8. *Let* $\mathcal{D}^B$ *be the minimum ML DAG of* $\boldsymbol{X}$ *as in Definition 2.5.1 with parents*

$\mathrm{pa}^B(i)$ *of node* i *in* $\mathcal{D}^B$*. Then for all* $i \in V$ *we have* $\mathrm{An}_{\mathrm{low}}^{\mathrm{pa}(i)}(i) = \mathrm{pa}^B(i)$ *and*

$$X_i = \bigvee_{k \in \mathrm{pa}^B(i)} \frac{b_{ki}}{b_{kk}} X_k \vee b_{ii} Z_i = \bigvee_{k \in \mathrm{pa}^B(i)} c_{ki} X_k \vee c_{ii} Z_i. \tag{2.6.11}$$

Proof. Recall from (2.5.1) that

$$\mathrm{pa}^B(i) = \Big\{ k \in \mathrm{pa}(i) : b_{ki} > \bigvee_{\ell \in \mathrm{de}(k) \cap \mathrm{pa}(i)} \frac{b_{k\ell} b_{\ell i}}{b_{\ell\ell}} \Big\},$$

and observe from this and (2.6.3) that $\mathrm{An}_{\mathrm{low}}^{\mathrm{pa}(i)}(i) = \mathrm{pa}^B(i)$. Since every path from $j \in \mathrm{an}(i)$ to i passes through some node in $\mathrm{pa}(i)$, there is always a max-weighted path from j to i containing some node of $\mathrm{pa}(i)$. Hence, $\mathrm{An}_{\mathrm{nmw}}^{\mathrm{pa}(i)}(i) = (\mathrm{an}_{\mathrm{nmw}}^{\mathrm{pa}(i)}(i) \cup \{i\}) \cap (\mathrm{pa}(i))^c = \{i\}$. Thus we obtain from (2.6.9) the first equality in (2.6.11). For the second, recall from Theorem 2.5.4(a) that $b_{ii} = c_{ii}$ and $\frac{b_{ki}}{b_{kk}} = c_{ki}$ for $k \in \mathrm{pa}^B(i)$. □

Remark 2.6.9. Representation (2.6.11) complements Theorem 2.5.4(a); we find again that the minimum ML DAG $\mathcal{D}^B$ yields the minimal representation of $\boldsymbol{X}$ as a recursive ML model. □

The following example illustrates and discusses representation (2.6.9).

Example 2.6.10. [Continuation of Examples 2.2.1, 2.3.8, 2.6.1, and 2.6.4: minimal representation of X_4 by X_1, X_2 and X_2]
We consider again $U = \{1, 2\}$ and $i = 4$. Obviously, there are max-weighted paths from 1 and 2 to 4 passing through some node in $U = \{1, 2\}$. Hence, $1, 2 \in \mathrm{an}(4) \setminus \mathrm{an}_{\mathrm{nmw}}^U(4)$. Since no max-weighted path from 3 to 4 passes through 1 or 2, we have $\mathrm{An}_{\mathrm{nmw}}^U(4) = (\mathrm{an}_{\mathrm{nmw}}^U(4) \cup \{4\}) \cap U^c = \{3, 4\}$. In Example 2.6.4 we have already determined the set $\mathrm{An}_{\mathrm{low}}^U(4)$ depending on the ML coefficients. Thus we distinguish again between two cases:

(1) If $b_{14} = \frac{b_{12} b_{24}}{b_{22}}$, then $X_4 = \frac{b_{24}}{b_{22}} X_2 \vee b_{34} Z_3 \vee b_{44} Z_4$.

We want to remark that the conditional independence properties of $\boldsymbol{X}$ are reflected in this representation: from Example 2.3.8 we know that $X_1 \perp\!\!\!\perp X_4 \mid X_2$ if the path $[1 \to 2 \to 4]$ is max-weighted, which is the case here. So it is obvious that X_1 does not need to appear in the minimal representation of X_4 as max-linear function of X_1 and X_2.

(2) If $b_{14} > \frac{b_{12} b_{24}}{b_{22}}$, then $X_4 = \frac{b_{14}}{b_{11}} X_1 \vee \frac{b_{24}}{b_{22}} X_2 \vee b_{34} Z_3 \vee b_{44} Z_4$.

In particular, $\frac{b_{14}}{b_{11}} X_1 > \frac{b_{24}}{b_{22}} X_2$ is possible with positive probability; in (1) this is not possible (see Example 2.6.1).

For $U = \{2\}$ and $i = 4$ we have $\mathrm{An}_{\mathrm{low}}^U(4) = \{2\}$. Similarly as above we obtain that $2 \in \mathrm{an}(4) \setminus \mathrm{an}_{\mathrm{nmw}}^U(4)$ and $3, 4 \in \mathrm{An}_{\mathrm{nmw}}^U(4)$. It remains to discuss node 1, which gives rise to the same two cases as above:

(1) If the path $[1 \to 2 \to 4]$ is max-weighted, then $X_4 = \frac{b_{24}}{b_{22}} X_2 \vee b_{34} Z_3 \vee b_{44} Z_4$.

(2) If the path $[1 \to 2 \to 4]$ is not max-weighted, then $X_4 = \frac{b_{24}}{b_{22}} X_2 \vee b_{14} Z_1 \vee b_{34} Z_3 \vee b_{44} Z_4$.

Such minimal representations become relevant, when $\boldsymbol{X}$ is partially observed. If, for example, X_2 is observed and B is known, then the prediction problem of X_4 can be solved by (conditional) simulation of the relevant noise variables and direct computation of X_4. In case (1) we need to simulate independent Z_3, Z_4, whereas in case (2) additionally Z_1 has to be simulated conditioned on X_2. We discuss such prediction problems in Gissibl and Klüppelberg [28]. □

Appendix 2.A An auxiliary lemma

Lemma 2.A.1. *Let $\mathcal{D} = (V, E)$ be a DAG and $U \subseteq V$. For nonnegative functions $a(i, j, k)$, $i, j, k \in V$, we have for all $i \in V$,*

$$\bigvee_{k \in \mathrm{pa}(i)} \bigvee_{j \in \mathrm{an}(k)} a(i,j,k) = \bigvee_{j \in \mathrm{an}(i)} \bigvee_{k \in \mathrm{de}(j) \cap \mathrm{pa}(i)} a(i,j,k), \tag{2.A.1}$$

$$\bigvee_{k \in \mathrm{an}(i) \cap U} \bigvee_{j \in \mathrm{An}(k)} a(i,j,k) = \bigvee_{j \in \mathrm{an}(i)} \bigvee_{k \in \mathrm{De}(j) \cap \mathrm{an}(i) \cap U} a(i,j,k). \tag{2.A.2}$$

Proof. Since we take maxima, we only have to prove that each combination of nodes (k, j) on the left-hand side appears also on the right-hand side and vice versa. In order to prove (2.A.1), it suffices to show that

$$k \in \mathrm{pa}(i) \text{ and } j \in \mathrm{an}(k) \quad \text{if and only if} \quad j \in \mathrm{an}(i) \text{ and } k \in \mathrm{de}(j) \cap \mathrm{pa}(i).$$

By observing that $\mathrm{an}(\mathrm{pa}(i)) \subseteq \mathrm{an}(i)$ and $j \in \mathrm{an}(k)$ if and only if $k \in \mathrm{de}(j)$, this equivalence is obvious. Eq. (2.A.2) is proved in the same way. □

Chapter 3

Tail dependence of recursive max-linear models with regularly varying noise variables

Abstract

Recursive max-linear structural equation models with regularly varying noise variables are considered. Their causal structure is represented by a directed acyclic graph (DAG). The problem of identifying a recursive max-linear model and its associated DAG from its matrix of pairwise tail dependence coefficients is discussed. For example, it is shown that if a causal ordering of the associated DAG is additionally known, then the minimum DAG representing the recursive structural equations can be recovered from the tail dependence matrix. For a relevant subclass of recursive max-linear models, identifiability of the associated minimum DAG from the tail dependence matrix and the initial nodes is shown. Algorithms find the associated minimum DAG for the different situations. Furthermore, given a tail dependence matrix, an algorithm outputs all compatible recursive max-linear models and their associated minimum DAGs.

MSC 2010 subject classifications: Primary 60G70, 05C20; secondary 05C75, 62-09, 65S05

Keywords and phrases: Causal inference, directed acyclic graph, extreme value theory, graphical model, max-linear model, max-stable model, regular variation, structural equation model, tail dependence coefficient

3.1 Introduction

Causal inference is fundamental in virtually all areas of science. Examples for concepts established over the last years to understand causal inference include structural equation modeling (see e.g. Bollen [5], Pearl [55]) and graphical modeling (see e.g. Koller and Friedman [45], Lauritzen [47], Spirtes et al. [69]).

In extreme risk analysis, it is especially important to understand causal dependencies. We consider *recursive max-linear models (RMLMs)*, which are max-linear structural equation models whose causal structure is represented by a *directed acyclic graph (DAG)*. Such models are directed graphical models (see [55], Theorem 1.4.1); i.e., the DAG encodes conditional independence

relations in the distribution via the (directed global) Markov property. RMLMs were introduced and studied in Chapter 2. They may find their application in situations when extreme risks play an essential role and may propagate through a network, for example, when modeling water levels or pollution concentrations in a river or when modeling risks in a large industrial structure. In Einmahl et al. [20] a RMLM was fitted to data from the EURO STOXX 50 Index, where the DAG structure was assumed to be known.

In this chapter we assume *regularly varying noise variables*. This leads to models treated in classical multivariate extreme value theory. The books by Beirlant et al. [4], de Haan and Ferreira [14], and Resnick [60, 61] provide a detailed introduction into this field. A RMLM with regularly varying noise variables is in the *maximum domain of attraction of an extreme value (max-stable) distribution*. The spectral measure of the limit distribution, which describes the dependence structure given by the DAG, is discrete. Every max-stable random vector with discrete spectral measure is *max-linear (ML)*, and every multivariate max-stable distribution can be approximated arbitrarily well via a ML model (e.g. Yuen and Stoev [75], Section 2.2). This demonstrates the important role of ML models in extreme value theory. They have been investigated, generalized, and applied to real world problems by many researchers; see e.g. Cui and Zhang [11], Einmahl et al. [19], Falk et al. [23], Kiriliouk [41], Schlather and Tawn [64], Strokorb and Schlather [70], and Wang and Stoev [73].

One main research problem that is addressed for restricted recursive structural equation models, where the functions are required to belong to a specified function class, is the *identifiability* of the coefficients and the DAG from the observational distribution. Recently, particular attention in this context has been given to recursive structural equation models with additive Gaussian noise; see e.g. Ernest et al. [21], Peters et al. [57], and references therein. For RMLMs this problem is investigated in Chapter 4. In the present chapter we discuss the identifiability of RMLMs from their *(upper) tail dependence coefficients (TDCs)*.

The TDC, which goes back to Sibuya [66], measures the extremal dependence between two random variables and is a simple and popular dependence measure in extreme value theory. Methods to construct multivariate max-stable distributions with given TDCs have been proposed, for example, by Falk [22], [23], [64], and [70]. Somehow related we identify all RMLMs with the same given TDCs.

3.1.1 Problem description and important concepts

First we briefly review RMLMs and introduce the TDC formally. We then describe the idea of this work in more detail and state the main results.

Max-linear models on DAGs

Consider a *RMLM* $\boldsymbol{X} = (X_1, \dots, X_d)$ *on a DAG* $\mathcal{D} = (V, E)$ with nodes $V = \{1, \dots, d\}$ and edges $E = \{(k, i) : i \in V \text{ and } k \in \mathrm{pa}(i)\}$:

$$X_i = \bigvee_{k \in \mathrm{pa}(i)} c_{ki} X_k \vee c_{ii} Z_i, \quad i = 1, \dots, d, \tag{3.1.1}$$

where $\mathrm{pa}(i)$ denotes the parents of node i in $\mathcal{D}$ and $c_{ki} > 0$ for $k \in \mathrm{pa}(i) \cup \{i\}$; the noise variables $Z_1, \dots, Z_d$, represented by a generic random variable Z, are assumed to be independent and identically distributed with support $\mathbb{R}_+ := (0, \infty)$ and regularly varying with index $\alpha \in \mathbb{R}_+$, abbreviated by $Z \in \mathrm{RV}(\alpha)$. Denoting the distribution function of Z by F_Z, the latter means that

$$\lim_{t\to\infty} \frac{1 - F_Z(xt)}{1 - F_Z(t)} = x^{-\alpha}$$

for every $x \in \mathbb{R}_+$. Examples for F_Z include Cauchy, Pareto, and log-gamma distributions. For details and background on regular variation, see e.g. [60, 61].

The properties of the noise variables imply the existence of a normalizing sequence $a_n \in \mathbb{R}_+$ such that for independent copies $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)}$ of $\boldsymbol{X}$,

$$a_n^{-1} \bigvee_{\nu=1}^{n} \boldsymbol{X}^{(\nu)} \xrightarrow{d} \boldsymbol{M}, \quad n \to \infty, \tag{3.1.2}$$

where $\boldsymbol{M}$ is a non-degenerate random vector with distribution function denoted by G and all operations are taken componentwise. Thus $\boldsymbol{X}$ is in the maximum domain of attraction of G; we write $\boldsymbol{X} \in \mathrm{MDA}(G)$. The limit vector $\boldsymbol{M}$ (its distribution function G) is necessarily max-stable: in the present situation we have for all $n \in \mathbb{N}$ and independent copies $\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(n)}$ of $\boldsymbol{M}$, the distributional equality $n^{1/\alpha}\boldsymbol{M} \overset{d}{=} \bigvee_{\nu=1}^{n} \boldsymbol{M}^{(\nu)}$. Furthermore, $\boldsymbol{M}$ is again a RMLM on $\mathcal{D}$, with the same weights in (3.1.1) as $\boldsymbol{X}$ and standard α-Fréchet distributed noise variables, i.e.,

$$F_Z(x) = \Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, \quad x \in \mathbb{R}_+.$$

A proof of (3.1.2) as well as an explicit formula for G and its univariate and bivariate marginal distributions can be found in Appendix 3.A.2, Proposition 3.A.2.

In what follows we summarize the most important properties of $\boldsymbol{X}$ presented in Chapter 2 which are needed throughout this chapter. Every component of $\boldsymbol{X}$ can be written as a max-linear function of its ancestral noise variables:

$$X_i = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j, \quad i = 1, \dots, d, \tag{3.1.3}$$

where $\mathrm{An}(i) = \mathrm{an}(i) \cup \{i\}$ and $\mathrm{an}(i)$ are the ancestors of i in $\mathcal{D}$ (see Theorem 2.2.2). For $i \in V$, $b_{ii} = c_{ii}$. For $j \in \mathrm{an}(i)$, b_{ji} can be determined by a path analysis of $\mathcal{D}$ as explained in the following. Throughout we write $k \to i$ whenever $\mathcal{D}$ has an edge from k to i. With every path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$ we associate a weight, which we define to be the product of the edge weights along p multiplied by c_{jj}. The coefficient b_{ji} is then the maximum weight of all paths from j to i. In summary, we have for $i \in V$ and $j \in \mathrm{an}(i)$,

$$b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \quad \text{with} \quad d_{ji}(p) := c_{k_0 k_0} \prod_{\nu=0}^{n-1} c_{k_\nu k_{\nu+1}}, \tag{3.1.4}$$

where P_{ji} is the set of all paths from j to i. For all $i \in V$ and $j \in V \setminus \mathrm{An}(i)$ we set $b_{ji} = 0$. We call

the coefficients b_{ji} *ML coefficients (MLCs)* and summarize them in the *ML coefficient matrix (MLCM)* $B = (b_{ij})_{d\times d}$. For the *reachability matrix* R *of* $\mathcal{D}$, whose ji-th entry is one if $j \in \mathrm{An}(i)$ and zero else, we find

$$R = \mathrm{sgn}(B), \tag{3.1.5}$$

where sgn denotes the signum function and is taken componentwise. As a consequence, the ancestors and descendants of every node in $\mathcal{D}$ can be obtained from B.

Not all paths are needed for computing b_{ji} in (3.1.4). We call a path p from j to i *max-weighted path from* j *to* i if it realizes the maximum in (3.1.4), i.e., if $b_{ji} = d_{ji}(p)$. The concept of max-weighted paths is essential. This has been worked out in Chapter 2. For example, max-weighted paths may lead to more conditional independence relations in the distribution of $\boldsymbol{X}$ than those encoded by $\mathcal{D}$ via the Markov property (see Remark 2.3.9). RMLMs where all paths are max-weighted play a central role in this chapter; we call them *recursive max-weighted models (RMWMs)*.

Further DAGs and weights may exist such that $\boldsymbol{X}$ satisfies (3.1.1); for a detailed characterization of these DAGs and weights, see Theorem 2.5.4. The smallest DAG of this kind is the one that has an edge $k \to i$ if and only if this is the only max-weighted path from k to i in $\mathcal{D}$ (see Remark 2.5.2(ii) and Theorem 2.5.4(a)). We call this DAG $\mathcal{D}^B$, the *minimum ML DAG of* $\boldsymbol{X}$. It can be determined from B (see Theorem 2.5.3). The other DAGs representing $\boldsymbol{X}$ in the sense of (3.1.1) are those that have at least the edges of $\mathcal{D}^B$ and the same reachability matrix. For edges contained in $\mathcal{D}^B$, the weights from (3.1.1) are uniquely defined by B. From these weights the weights for the other edges can be derived.

Remark 3.1.1. The random vector $\boldsymbol{X}$ and its distribution are characterized by the distribution F_Z of the noise variables and the max-linear dependence structure induced by $\mathcal{D}$. So computing the max-stable limit distribution G concerns only the marginal limits, whereas the max-linear dependence structure remains always the same (cf. also the proof of Proposition 3.A.2). This restrictive dependence structure of $\boldsymbol{X}$ can be generalized naturally within the framework of multivariate regular variation. See [60, 61] for background on multivariate regular variation.

In the literature various equivalent formulations of regular variation for random vectors can be found. The extent of a possible generalization can be probably best understood when considering an equivalent representation of the dependence in a regularly varying vector. A random vector $\boldsymbol{X} \in \mathbb{R}_+^d$ is regularly varying with index $\alpha \in \mathbb{R}_+$ if and only if there exists a random vector Θ with values in $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\}$, where $\|\cdot\|$ is any norm in $\mathbb{R}_+^d$, such that for every $x \in \mathbb{R}_+$,

$$\frac{\mathbb{P}(\|\boldsymbol{X}\| > tx, \boldsymbol{X}/\|\boldsymbol{X}\| \in \cdot)}{\mathbb{P}(\|\boldsymbol{X}\| > t)} \xrightarrow{v} x^{-\alpha}\mathbb{P}(\Theta \in \cdot), \quad t \to \infty. \tag{3.1.6}$$

The notation $\xrightarrow{v}$ stands for vague convergence on the Borel σ-algebra of $\mathbb{S}^{d-1}$. We immediately find from (3.1.6) that the dependence structure of $\boldsymbol{X}$ is for moderate values of $\|\boldsymbol{X}\|$ arbitrary; only when $\|\boldsymbol{X}\|$ becomes large, the dependence structure becomes that of Θ. When assuming that the dependence structure in the limit is max-linear given by $\mathcal{D}$ and the marginal limits are α-Fréchet

(with an appropriate scale parameter), then $\boldsymbol{X} \in \mathrm{MDA}(G)$ with G still as in Proposition 3.A.2; hence, $\boldsymbol{X}$ would have the same TDCs as in the present less general framework. So similarly to the flexibility of the margins, expressed by $Z \in \mathrm{RV}(\alpha)$, there would also be flexibility in the dependence structure.

In this chapter the restriction to the limiting max-linear dependence provides a sufficient model as the focus lies on the causal structure in terms of the DAGs. This allows for a more concise notation and makes the focus of the chapter more transparent. □

The tail dependence matrix of $\boldsymbol{X}$

For $i \in V$ we denote the distribution function of component X_i of the RMLM $\boldsymbol{X}$ by F_i and its generalized inverse by $F_i^{\leftarrow}(u) = \inf\{x \in \mathbb{R}_+ : F(x) \geq u\}$ for $0 < u < 1$. The TDC between X_i and X_j is then given by the limit

$$\chi(i,j) = \lim_{u \uparrow 1} \mathbb{P}(X_i > F_i^{\leftarrow}(u) \mid X_j > F_j^{\leftarrow}(u)).$$

We summarize all TDCs in the *tail dependence matrix (TDM)* $\chi = (\chi(i,j))_{d\times d}$.

Recall the fact that $Z \in \mathrm{RV}(\alpha)$ and the definition of the MLCM B of $\boldsymbol{X}$. Defining then the *standardized MLCM of* $\boldsymbol{X}$ by

$$\overline{B} = \big(\overline{b}_{ij}\big)_{d\times d} := \Big(\frac{b_{ij}^{\alpha}}{\sum_{k\in \mathrm{An}(j)} b_{kj}^{\alpha}}\Big)_{d\times d}, \tag{3.1.7}$$

the TDC between X_i and X_j can be computed as

$$\chi(i,j) = \chi(j,i) = \sum_{k\in \mathrm{An}(i)\cap \mathrm{An}(j)} \overline{b}_{ki} \wedge \overline{b}_{kj}. \tag{3.1.8}$$

By (3.1.5) and (3.1.7) it is the sum of the pairwise minima of the i-th and j-th column of $\overline{B}$. A proof of (3.1.8) is given in Appendix 3.A.2. There we implicitly show that $\boldsymbol{X}$ and the limit vector $\boldsymbol{M}$ from (3.1.2) have the same TDM χ.

The TDC $\chi(i,j)$ is zero if and only if i and j do not have common ancestors. Therefore, the *initial nodes of* $\mathcal{D}$ (i.e., the nodes without parents) constitute a set V_0 of maximum cardinality such that $\chi(i,j)$ is zero for all distinct $i,j \in V_0$. This property turns out to be helpful when identifying from χ. We also show that $\chi(i,j)$ is zero if and only if X_i and X_j are independent, which is reminiscent of the multivariate Gaussian distribution with its equivalence between independence and zero correlation.

Obviously, when investigating χ, understanding the structure of $\overline{B}$ is essential. Not surprisingly, $\overline{B}$ inherits structural properties from B. For example, $\overline{B}$ is again a MLCM of a RMLM on the same DAG $\mathcal{D}$, and its columns add up to one. Properties of $\overline{B}$, which we use throughout this chapter, are summarized in Appendix 3.A.2, Lemma 3.A.1.

Identifiability from χ

The main goal of this chapter is to investigate how far the dependence structure of $\boldsymbol{X}$ and the DAG $\mathcal{D}$ can be recovered from the TDM χ. We call two RMLMs that have the same TDM χ-*equivalent*. For example, $\boldsymbol{X}$ and the limit vector $\boldsymbol{M}$ from (3.1.2) are χ-equivalent. The set

$$\left\{(\widetilde{b}_{ij})_{d\times d} \in \mathbb{R}_+^{d\times d} : \widetilde{b}_{ij} = \beta_j \overline{b}_{ij}^{1/\widetilde{\alpha}} \text{ for all } i,j \in V \text{ and } \beta_j \in \mathbb{R}_+\right\} \tag{3.1.9}$$

contains the MLCMs of all RMLMs that have the same standardized MLCM $\overline{B}$ as $\boldsymbol{X}$ and regularly varying noise variables with index $\widetilde{\alpha} \in \mathbb{R}_+$; this can be verified by using Theorem 2.5.7. Obviously, all the corresponding RMLMs are also χ-equivalent to $\boldsymbol{X}$. Therefore, given χ only, we can never identify the true representations (3.1.1) and (3.1.3) of $\boldsymbol{X}$ and the DAG $\mathcal{D}$.

The RMLM $\boldsymbol{X}$ has the same minimum ML DAG $\mathcal{D}^B$ as every RMLM with MLCM $\overline{B}$ (Lemma 3.A.1(e)). As a consequence, $\mathcal{D}^B$ can be determined from $\overline{B}$ (cf. Theorem 2.5.3). This raises the question of whether $\overline{B}$ and, hence, the minimum ML DAG of $\boldsymbol{X}$ are identifiable from χ. The answer is generally no, quite simply due to the symmetry of χ.

Example 3.1.2. [$\overline{B}$ is not identifiable from χ]
Consider two RMLMs on the DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ with standardized MLCMs

$$\mathcal{D}_1 \quad \begin{array}{ccc} 1 & \longrightarrow & 2 \end{array} \quad \overline{B}_1 = \begin{bmatrix} 1 & b \\ 0 & 1-b \end{bmatrix} \quad \text{and} \quad \overline{B}_2 = \begin{bmatrix} 1-b & 0 \\ b & 1 \end{bmatrix} \quad \begin{array}{ccc} 1 & \longleftarrow & 2 \end{array} \quad \mathcal{D}_2$$

for some $b \in (0,1)$. For both we find the same TDCs: $\chi(1,1) = \chi(2,2) = 1$, $\chi(1,2) = \chi(2,1) = b$. □

We show, however, that $\overline{B}$ can be computed recursively from χ and some additional information on the DAG $\mathcal{D}$. This may be its reachability matrix R but also only a *causal ordering* σ; i.e., σ is a permutation on $V = \{1,\dots,d\}$ such that $\sigma(j) < \sigma(i)$ for all $i \in V$ and $j \in \mathrm{an}(i)$. If $\boldsymbol{X}$ is max-weighted, then $\overline{B}$ is identifiable from χ and the initial nodes V_0 of $\mathcal{D}$.

The question also arises which RMLMs are all χ-equivalent to $\boldsymbol{X}$ and what their minimum ML DAGs are. Since by (3.1.9) every MLCM of a RMLM with TDM χ can be obtained from its particular standardized version, it suffices to clarify which the standardized MLCMs of all RMLMs with TDM χ are. To this end we use the identifiability results mentioned above to develop an algorithm that computes these matrices from χ. The proposed procedure can be considerably simplified for RMWMs.

Another interesting point is how DAGs of χ-equivalent RMLMs relate to each other. Here we also investigate the RMWMs as a relevant subclass of RMLMs separately. For example, an initial node in a DAG of a RMWM is again an initial node in a DAG of a χ-equivalent RMWM or it must be a terminal node (i.e., a node without descendants).

This chapter is organized as follows. We provide some basic results in Section 3.2. For a RMLM $\boldsymbol{X}$ we investigate its TDM χ and link it to its standardized MLCM $\overline{B}$ and its associated DAG $\mathcal{D}$. Here we discuss the situations when two components of $\boldsymbol{X}$ have zero tail dependence. We also

introduce the important concept of χ-cliques, which allows us to identify potential initial node sets in $\mathcal{D}$ from χ. Section 3.3 is devoted to RMWMs. We point out the specific properties of χ which lead to the identifiability of $\overline{B}$ from χ and the initial nodes. We also present necessary and sufficient conditions on a matrix to be the TDM of a RMWM. In Section 3.4 we then study different identifiability problems based on χ. We propose algorithms to compute $\overline{B}$ from χ and some further information on $\mathcal{D}$ such as a causal ordering. We also explain how the standardized MLCMs of all RMLMs that have TDM χ can be determined. In Section 3.5 we consider χ-equivalent RMLMs and analyze relationships between them and their DAGs. We use these results to investigate whether RMWMs on different DAGs can be χ-equivalent at all and if so under which conditions. Section 3.6 concludes and suggests further directions of research.

Note that all recursion formulas presented in the chapter are well-defined, since we work with DAGs. Throughout we illustrate our findings with examples for the (standardized) MLCM of a RMLM on a given DAG. It can be verified by Theorem 2.4.2 or Corollary 2.4.3(a) that the presented matrices are indeed MLCMs of RMLMs on the particular DAGs. Moreover, we use the following notation throughout the chapter. We denote the ancestors, parents, and descendants of node i in $\mathcal{D}$ by $\mathrm{an}(i)$, $\mathrm{pa}(i)$, and $\mathrm{de}(i)$, respectively. We define $\mathrm{An}(i) := \mathrm{an}(i) \cup \{i\}$, $\mathrm{Pa}(i) := \mathrm{pa}(i) \cup \{i\}$, and $\mathrm{De}(i) := \mathrm{de}(i) \cup \{i\}$. For (possibly random) $a_i \in \mathbb{R}$ we set $\bigvee_{i\in\varnothing} a_i = 0$ and $\sum_{i\in\varnothing} a_i = 0$. We generally consider statements for $i \in \varnothing$ as invalid.

3.2 A recursive ML model and its tail dependence matrix

In this section for a RMLM $\boldsymbol{X}$ on a DAG $\mathcal{D}$, we highlight some relations between its TDM χ, its standardized MLCM $\overline{B}$, and the DAG $\mathcal{D}$. They prove particularly useful when we identify the RMLMs that are χ-equivalent to $\boldsymbol{X}$ in Section 3.4.4 or investigate DAGs of χ-equivalent RMLMs in Section 3.5.

3.2.1 The tail dependence coefficients and max-weighted paths

We start with lower and upper bounds for the TDC between two components of $\boldsymbol{X}$ such that in $\mathcal{D}$ the two corresponding nodes are connected by a path. We also show that max-weighted paths lead to simple expressions for the TDCs and to nice relationships between them. Exactly these properties motivate us to consider RMWMs in detail later on.

Lemma 3.2.1. *Let $i \in V$ and $j \in \mathrm{an}(i)$.*

(a) We have $0 < \frac{\overline{b}_{ji}}{\overline{b}_{jj}} \le \chi(j,i)$ with equality if and only if there is a max-weighted path from every $k \in \mathrm{An}(j)$ to i passing through j. In that case, $\chi(j,i) = \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki}$.

(b) We have $\chi(j,i) \le \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki} < 1$.

(c) Let $k \in \mathrm{de}(j) \cap \mathrm{an}(i)$. If there is a max-weighted path from every $\ell \in \mathrm{An}(j)$ to k and from every $\ell \in \mathrm{An}(j)$ to i passing through j as well as from every $\ell \in \mathrm{An}(k)$ to i passing through

k, then

$$\chi(j,i) = \chi(j,k)\chi(k,i) < \chi(j,k) \wedge \chi(k,i). \tag{3.2.1}$$

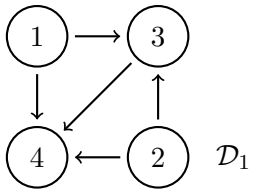
Proof. As $\mathrm{An}(j) \subseteq \mathrm{An}(i)$, we have by (3.1.8), $\chi(j,i) = \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki} \wedge \overline{b}_{kj}$.
(a) For $k \in \mathrm{An}(j)$, by Lemma 3.A.1(d), (f), $\frac{\overline{b}_{kj}\overline{b}_{ji}}{\overline{b}_{jj}} \le \overline{b}_{ki} \wedge \overline{b}_{kj}$ with equality if and only if there is a max-weighted path from k to i passing through j. With this, using also Lemma 3.A.1(b), (a), we obtain $\chi(j,i) \ge \frac{\overline{b}_{ji}}{\overline{b}_{jj}} \sum_{k\in\mathrm{An}(j)} \overline{b}_{kj} = \frac{\overline{b}_{ji}}{\overline{b}_{jj}} > 0$ with equality if and only if there is a max-weighted path from every $k \in \mathrm{An}(j)$ to i passing through j. In that case Lemma 3.A.1(d) yields $\chi(j,i) = \sum_{k\in\mathrm{An}(j)} \frac{\overline{b}_{kj}\overline{b}_{ji}}{\overline{b}_{jj}} = \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki}$.
(b) As $\mathrm{An}(j) \subsetneq \mathrm{An}(i)$, by Lemma 3.A.1(a), (b) we find $\chi(j,i) \le \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki} < \sum_{k\in\mathrm{An}(i)} \overline{b}_{ki} = 1$.
(c) The equality in (3.2.1) follows from (a) and Lemma 3.A.1(d), the inequality then from the strict inequalities in (a) and (b). □

In the proof of Lemma 3.2.1 we have used that for $i \in V$, $k \in \underline{\mathrm{an}}(i)$, and $j \in \mathrm{an}(k)$, $\mathcal{D}$ has a max-weighted path from j to i passing through k if and only if $\overline{b}_{ji} = \frac{\overline{b}_{jk}\overline{b}_{ki}}{\overline{b}_{kk}}$ (Lemma 3.A.1(d)). As to the equality in (3.2.1), one could expect that the MLCs can be replaced by the corresponding TDCs. The following example disproves this. In particular, it proves that the converse of Lemma 3.2.1(c) is not true in general and also that we may have the equality in (3.2.1) although $k \notin \mathrm{de}(j) \cap \mathrm{an}(i)$.

Example 3.2.2. [$\chi(j,i) = \chi(j,k)\chi(k,i)$ is neither necessary nor sufficient for $\overline{b}_{ji} = \frac{\overline{b}_{jk}\overline{b}_{ki}}{\overline{b}_{kk}}$]

(1) Consider a RMLM on $\mathcal{D}_1$ with standardized MLCM

$$\overline{B} = \begin{bmatrix} 1 & 0 & 0.4 & 0.3 \\ 0 & 1 & 0.4 & 0.25 \\ 0 & 0 & 0.2 & 0.125 \\ 0 & 0 & 0 & 0.325 \end{bmatrix}.$$



$\mathcal{D}_1$

As $\overline{b}_{24} = \frac{\overline{b}_{23}\overline{b}_{34}}{\overline{b}_{33}}$, the path $[2 \to 3 \to 4]$ is max-weighted. Computing χ we find $\chi(2,4) < \chi(2,3)\chi(3,4)$. That is, $\chi(2,4) \ne \chi(2,3)\chi(3,4)$ although there is a max-weighted path from 2 to 4 passing through 3.

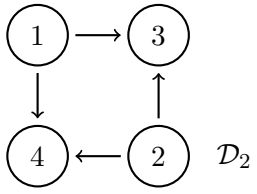
(2) Now consider a RMLM on $\mathcal{D}_1$ with standardized MLCM

$$\overline{B} = \begin{bmatrix} 1 & 0 & 0.1 & 0.085 \\ 0 & 1 & 0.8 & 0.5 \\ 0 & 0 & 0.1 & 0.04 \\ 0 & 0 & 0 & 0.375 \end{bmatrix}.$$

The path $[2 \to 3 \to 4]$ is not max-weighted, since $\frac{\overline{b}_{23}\overline{b}_{34}}{\overline{b}_{33}} \ne \overline{b}_{24}$. However, we have $\chi(2,3)\chi(3,4) = \chi(2,4)$. In summary, $\chi(2,3)\chi(3,4) = \chi(2,4)$ although there is no max-weighted path from 2 to 4 passing through 3.

(3) Finally, consider a RMLM on $\mathcal{D}_2$ with standardized MLCM

$$\overline{B} = \begin{bmatrix} 1 & 0 & 1/3 & 1/6 \\ 0 & 1 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/2 \end{bmatrix}.$$



$\mathcal{D}_2$

Here we find $\chi(1,3)\chi(3,4) = \chi(1,4)$; but 3 is not an ancestor of 4. According to this the equality in (3.2.1) may hold although $k \notin \mathrm{de}(j) \cap \mathrm{an}(i)$. □

3.2.2 The tail dependence coefficients and the initial nodes

In this section we mainly investigate how χ and $\mathcal{D}$ relate to each other.

Two components of $\boldsymbol{X}$ are independent if and only if the TDC between them is zero. We link these two properties with the relationship between the two corresponding nodes in $\mathcal{D}$.

Theorem 3.2.3. *Let $\boldsymbol{X}$ be a RMLM on a DAG $\mathcal{D} = (V, E)$ with TDM χ and $i, j \in V$. Then the following statements are equivalent:*

(a) *X_i and X_j are independent.*

(b) $\mathrm{An}(i) \cap \mathrm{An}(j) = \varnothing$.

(c) $\chi(i,j) = 0$.

Proof. The equivalence between (a) and (b) follows from representation (3.1.3) for X_i and X_j and the distributional properties of the noise variables. The one between (b) and (c) is immediate by (3.1.8) and Lemma 3.A.1(a). □

Remark 3.2.4. (i) Let R be the reachability matrix of $\mathcal{D}$. The ij-th (ji-th) entry of $R^T R$ equals the cardinality of $\mathrm{An}(i) \cap \mathrm{An}(j)$. Thus by Theorem 3.2.3, $\mathrm{sgn}(\chi) = \mathrm{sgn}(R^T R)$. That is, we learn from $\chi(i,j) > 0$ only that $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \varnothing$ but not whether i and j are connected by a path as is the case for the (standardized) MLCs (Lemma 3.A.1(a) and (3.1.5), respectively).

(ii) In the more general framework of Remark 3.1.1, parts (a) and (b) of Theorem 3.2.3 would have to be replaced by

(a') X_i and X_j are asymptotically independent; i.e., the corresponding components of the limit vector are independent.

(b') The dependence structure in the limit is given by a DAG, in which $\mathrm{An}(i) \cap \mathrm{An}(j) = \varnothing$.

The equivalence between (a') and (c) is a well-known result in extreme value theory; see e.g. Theorem 6.2.3 and the subsequent remark in [14]. □

In what follows we investigate the relationship between χ and the initial nodes V_0 of $\mathcal{D}$. This is motivated by the fact that a RMLM is recursively defined by the structure of $\mathcal{D}$. For example, to obtain representation (3.1.3) of $\boldsymbol{X}$ from its representation (3.1.1) recursively, we would start with representation (3.1.3) of the components X_i with $i \in V_0$. Then by proceeding iteratively we would substitute the parental variables in (3.1.1) by their representation (3.1.3). Such an iterative procedure starting with the initial nodes could also identify all RMLMs which have (the given) TDM χ.

The TDC between two components of $\boldsymbol{X}$ simplifies considerably when in $\mathcal{D}$ one of the corresponding nodes is an initial node. If both nodes are initial nodes, then the TDC between them is zero. We provide these and further related results.

Lemma 3.2.5. *(a) For distinct $i, j \in V_0$, $\chi(i,j) = 0$.*

(b) Let $W \subseteq V$ such that $\chi(i,j) = 0$ for all distinct $i, j \in W$. Then $|W| \leq |V_0|$.

(c) For $i \in V$ and $j \in V_0$, $\mathrm{An}(i) \cap V_0 = \{k \in V_0 : \chi(k,i) > 0\}$ and $\mathrm{De}(j) = \{k \in V : \chi(j,k) > 0\}$.

(d) For $i \in V$ and $j \in V_0$, $\chi(j,i) = \overline{b}_{ji}$.

Proof. (a) and (c) follow from the fact that initial nodes have no ancestors and Theorem 3.2.3. (b) Assume that $|W| > |V_0|$. Since for every $i \in V$ there is some $j \in \mathrm{An}(i) \cap V_0$, we have $j \in \mathrm{An}(i_1) \cap \mathrm{An}(i_2)$ for some $j \in V_0$ and distinct $i_1, i_2 \in W$. As $\mathrm{An}(i_1) \cap \mathrm{An}(i_2) \neq \varnothing$, again by Theorem 3.2.3, $\chi(i_1, i_2) \neq 0$. This is, however, a contradiction to the fact that $\chi(i_1, i_2) = 0$ as $i_1, i_2 \in W$. Hence, $|W| \leq |V_0|$.
(d) As $\mathrm{An}(j) = \{j\}$, we obtain from (3.1.8) by Lemma 3.A.1(a), (f), $\chi(j,i) = \sum_{k=1}^{d} \overline{b}_{ki} \wedge \overline{b}_{kj} = \overline{b}_{ji} \wedge \overline{b}_{jj} = \overline{b}_{ji}$. □

From Lemma 3.2.5(a), (b) we learn that V_0 is one of the node sets of maximum cardinality such that for every two distinct nodes, the TDC between their corresponding components of $\boldsymbol{X}$ is zero. We introduce a concept which allows us to determine these sets from χ by a graph. For an illustration of these notions, we refer to Example 3.4.12 below.

Definition 3.2.6. Let χ be the TDM of a RMLM on a DAG $\mathcal{D}$ and $W \subseteq V$.

(a) We call the undirected graph that has nodes V and an edge between k and i if and only if $\chi(k,i) > 0$, *χ-graph*.

Let now $\mathcal{D}^\chi$ be the complement of the χ-graph; i.e., $\mathcal{D}^\chi$ is the graph with the same node set V but the edge set consists of the edges that are not present in the χ-graph.

(b) We call W a *χ-clique* if it is a clique in $\mathcal{D}^\chi$; i.e., every two (distinct) nodes in W are connected by an edge in $\mathcal{D}^\chi$.

(c) We call W a *maximum χ-clique* if it is a maximum clique in $\mathcal{D}^\chi$; i.e., W is a clique in $\mathcal{D}^\chi$ such that no clique in $\mathcal{D}^\chi$ with higher cardinality exists. □

The χ-graph associated with the TDM χ of $\boldsymbol{X}$ corresponds to the covariance graph of the random vector $\boldsymbol{X}$ introduced in Cox and Wermuth [10], in which two (distinct) nodes are connected by an edge if and only if their corresponding components are dependent (cf. Theorem 3.2.3). In the non-Gaussian case, however, the name covariance graph is misleading.

The following theorem is an immediate consequence of Definition 3.2.6 and Lemma 3.2.5(a), (b).

Theorem 3.2.7. *Let $\boldsymbol{X}$ be a RMLM on a DAG $\mathcal{D}$ with TDM χ. Then the set V_0 is a maximum χ-clique.*

Theorem 3.2.7 raises the question of how V_0 is related to possible other maximum χ-cliques.

Lemma 3.2.8. *Let W be a maximum χ-clique.*

(a) There is only one bijection $\varphi : V_0 \to W$ such that for every $j \in V_0$, $\chi(j, \varphi(j)) > 0$ and $\chi(j, i) = 0$ for all $i \in W \setminus \{\varphi(j)\}$.

(b) Let φ be the bijection from (a). Then for $j \in V_0$, $\mathrm{An}(\varphi(j)) \cap V_0 = \{j\}$ and $\mathrm{De}(j) \cap W = \{\varphi(j)\}$. In particular, if $j \neq \varphi(j)$, then $\mathcal{D}$ has a path from j to $\varphi(j)$.

(c) Let $i, j \in V \setminus W$. If $\chi(i, j) < \sum_{k \in W} \chi(k, i) \wedge \chi(k, j)$, then $V_0 \neq W$.

Proof. (a) Since maximum χ-cliques have the same cardinality, we know from Theorem 3.2.7 that $|V_0| = |W|$. As for every $i \in W$, $\mathrm{An}(i) \cap V_0 \neq \varnothing$, it suffices by Lemma 3.2.5(c) to show that $|\mathrm{De}(j) \cap W| = 1$ for $j \in V_0$. We first assume that $|\mathrm{De}(j) \cap W| > 1$. Using Theorem 3.2.3 similarly as in the proof of Lemma 3.2.5(b) yields a contradiction. Hence, $|\mathrm{De}(j) \cap W| \leq 1$. As $|V_0| = |W|$, $|\mathrm{De}(j) \cap W| = 1$ must hold.
(b) follows from (a) and Lemma 3.2.5(c).
(c) Assume that $V_0 = W$. Using Lemma 3.A.1(a) and Lemma 3.2.5(d) we obtain from (3.1.8)

$$\chi(i, j) = \sum_{k=1}^{d} \overline{b}_{ki} \wedge \overline{b}_{kj} \geq \sum_{k \in W} \chi(k, i) \wedge \chi(k, j).$$

Since this contradicts the conditions of (c), V_0 and W must be different. □

3.3 A recursive max-weighted model and its tail dependence matrix

In this section we focus on RMWMs, i.e., RMLMs where all paths are max-weighted. We first present some structural properties of a RMWM $\boldsymbol{X}$ on a DAG $\mathcal{D}$ with standardized MLCM $\overline{B}$. We then investigate its TDM χ and show that the assumption of all paths in $\mathcal{D}$ being max-weighted involves simple relations between the TDCs and the (standardized) MLCs. Finally, we give necessary and sufficient conditions on a matrix to be the TDM of a RMWM on a given DAG.

3.3.1 Some structural properties of a recursive max-weighted model

All RMLMs on polytrees are RMWMs simply because in a polytree there is at most one path between every two (distinct) nodes (see also Example 2.3.2). Furthermore, a RMWM can be constructed on every DAG, as the following example shows. Note the particularly simple structure of the introduced class of RMLMs.

Example 3.3.1. [The homogeneous model]
Let $\mathcal{D} = (V, E)$ be a DAG with $V = \{1, \dots, d\}$ and $Z_1, \dots, Z_d$ as in (3.1.1). Consider the RMLM defined by

$$X_i := \frac{1}{|\mathrm{An}(i)|^{1/\alpha}} \Big(\bigvee_{k \in \mathrm{pa}(i)} |\mathrm{An}(k)|^{1/\alpha} X_k \vee Z_i \Big), \quad i = 1, \dots, d.$$

We find that every path p from j to i has the same weight $d_{ji}(p) = |\mathrm{An}(i)|^{-1/\alpha}$. As a consequence, every path is max-weighted and $\boldsymbol{X}$ is a RMWM. Its representation (3.1.3) is given by

$$X_i = \frac{1}{|\mathrm{An}(i)|^{1/\alpha}} \bigvee_{j \in \mathrm{An}(i)} Z_j, \quad i = 1, \dots, d.$$

For the TDC from (3.1.8) between X_i and X_j, we have

$$\chi(i,j) = \sum_{k \in \mathrm{An}(i) \cap \mathrm{An}(j)} \frac{1}{|\mathrm{An}(i)|} \wedge \frac{1}{|\mathrm{An}(j)|} = \frac{|\mathrm{An}(i) \cap \mathrm{An}(j)|}{|\mathrm{An}(i)| \vee |\mathrm{An}(j)|}.$$

If $j \in \mathrm{an}(i)$, then this reduces to $\chi(j,i) = |\mathrm{An}(j)|/|\mathrm{An}(i)|$. Finally, by Proposition 3.A.2 the components of the limit vector $\boldsymbol{M}$ introduced in (3.1.2) are standard α-Fréchet distributed. □

Recall from the Introduction the prominent role of the minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$, which equals the minimum ML DAG $\mathcal{D}^{\overline{B}}$ of a RMLM with MLCM $\overline{B}$ (Lemma 3.A.1(e)). The fact that $\boldsymbol{X}$ is max-weighted ensures that $\mathcal{D}^{\overline{B}}$ only depends on $\mathrm{sgn}(\overline{B})$ but not on the precise values of the standardized MLCs. Since $\mathrm{sgn}(\overline{B})$ is the reachability matrix of $\mathcal{D}$ ((3.1.5) and Lemma 3.A.1(a)), $\mathcal{D}^{\overline{B}}$ can be determined from pure graph theoretical properties. To clarify this we introduce a basic concept in graph theory, which goes back to Aho et al. [1].

Definition 3.3.2. Let $\mathcal{D}$ be a DAG.

(a) An edge $k \to i$ is redundant if $\mathcal{D}$ has another path from k to i.

(b) The DAG $\mathcal{D}^{\mathrm{tr}}$ obtained from $\mathcal{D}$ by deleting its redundant edges is called transitive reduction of $\mathcal{D}$. □

Since $\mathcal{D}^{\overline{B}}$ has an edge $k \to i$ if and only if this is the only max-weighted path from k to i in $\mathcal{D}$, the fact that $\mathcal{D}$ has only max-weighted paths yields part (i) of the following remark. By Definition 3.3.2 and Lemma 3.A.1(a) (ii) is a consequence of (i).

Remark 3.3.3. Let $\mathcal{D}^{\mathrm{tr}}$ be the transitive reduction of $\mathcal{D}$.

(i) The DAGs $\mathcal{D}^{\overline{B}}$ and $\mathcal{D}^{\mathrm{tr}}$ coincide.

(ii) $\mathcal{D}^{\overline{B}}$ is the DAG with the minimum number of edges that has reachability matrix $\operatorname{sgn}(\overline{B})$.

(iii) Even if $\boldsymbol{X}$ is a RMLM but not max-weighted, it may happen that $\mathcal{D}^{\overline{B}} = \mathcal{D}^{\mathrm{tr}}$ with all paths max-weighted in $\mathcal{D}^{\overline{B}}$. In that case all results presented in this section hold with respect to $\mathcal{D}^{\mathrm{tr}}$. □

3.3.2 Properties of the tail dependence coefficients of a recursive max-weighted model

The following result points out the simple structure of χ. It follows from Lemma 3.2.1(a), (c), since in $\mathcal{D}$ all paths are max-weighted.

Lemma 3.3.4. *Let $i \in V$.*

(a) For $j \in \operatorname{An}(i)$, $\chi(j,i) = \frac{\overline{b}_{ji}}{\overline{b}_{jj}} = \sum_{k \in \operatorname{An}(j)} \overline{b}_{ki} = \sum_{k \in \operatorname{An}(j)} \overline{b}_{kk} \chi(k,i)$.

(b) For $k \in \operatorname{an}(i)$ and $j \in \operatorname{an}(k)$, $\chi(j,i) = \chi(j,k)\chi(k,i) < \chi(j,k) \wedge \chi(k,i)$.

(c) For $j \in \operatorname{an}(i)$ and some path $[j = k_0 \to k_1 \to \cdots \to k_n = i]$, $\chi(j,i) = \prod_{\nu=0}^{n-1} \chi(k_\nu, k_{\nu+1})$.

The equality $\chi(j,i) = \chi(j,k)\chi(k,i)$ for some $j \in \operatorname{An}(i) \cap \operatorname{An}(k)$ does not necessarily imply that $k \in \operatorname{An}(i)$ (cf. part (3) of Example 3.2.2). For RMWMs, however, whenever these products hold for all $j \in \operatorname{An}(i) \cap \operatorname{An}(k) \cap V_0$, where V_0 are again the initial nodes in $\mathcal{D}$, we can conclude that $k \in \operatorname{An}(i)$.

Proposition 3.3.5. *For $i, k \in V$, $k \in \operatorname{An}(i)$ if and only if $\chi(j,i) = \chi(j,k)\chi(k,i)$ for all $j \in \operatorname{An}(i) \cap \operatorname{An}(k) \cap V_0$.*

Proof. Assume that $\chi(j,i) = \chi(j,k)\chi(k,i)$ for all $j \in \operatorname{An}(i) \cap \operatorname{An}(k) \cap V_0$. We first show that $\chi(\ell,i) \le \chi(\ell,k)$ for every $\ell \in \operatorname{An}(i) \cap \operatorname{An}(k)$. We obtain for $j \in \operatorname{An}(\ell) \cap V_0$, using the assumptions and Lemma 3.3.4(b),

$$\chi(k,i) = \frac{\chi(j,i)}{\chi(j,k)} = \frac{\chi(j,\ell)\chi(\ell,i)}{\chi(j,\ell)\chi(\ell,k)} = \frac{\chi(\ell,i)}{\chi(\ell,k)}.$$

Hence, $\chi(\ell,i) = \chi(\ell,k)\chi(k,i)$ and $\chi(\ell,i) \le \chi(\ell,k)$. Together with Lemma 3.3.4(a) we then find from (3.1.8)

$$\begin{aligned}\chi(k,i) &= \sum_{\ell \in \operatorname{An}(k) \cap \operatorname{An}(i)} \overline{b}_{\ell\ell}(\chi(\ell,k) \wedge \chi(\ell,i)) = \sum_{\ell \in \operatorname{An}(k) \cap \operatorname{An}(i)} \overline{b}_{\ell\ell}\chi(\ell,i) \\ &= \chi(k,i) \sum_{\ell \in \operatorname{An}(k) \cap \operatorname{An}(i)} \overline{b}_{\ell\ell}\chi(\ell,k).\end{aligned}$$

By the assumptions and Theorem 3.2.3 $\chi(k,i) > 0$ so that $\sum_{\ell \in \operatorname{An}(k) \cap \operatorname{An}(i)} \overline{b}_{\ell\ell}\chi(\ell,k) = 1$. As $1 = \sum_{\ell \in \operatorname{An}(k)} \overline{b}_{\ell\ell}\chi(\ell,k)$ (Lemma 3.3.4(a)) and $\overline{b}_{\ell\ell}\chi(\ell,k) > 0$ for all $\ell \in \operatorname{An}(k)$ (Lemma 3.A.1(a) and Theorem 3.2.3), we have $\operatorname{An}(i) \cap \operatorname{An}(k) = \operatorname{An}(k)$. This finally implies that $\operatorname{An}(k) \subseteq \operatorname{An}(i)$, equivalently $k \in \operatorname{An}(i)$.

The converse statement holds due to Lemma 3.3.4(b). □

In Lemma 3.3.4(a) we have written the positive standardized MLCs as functions of themselves and TDCs. We now present expressions for them only in terms of TDCs.

Proposition 3.3.6. *For $i \in V$ and $j \in \mathrm{An}(i)$,*

$$\overline{b}_{ji} = \chi(j,i) - \sum_{k \in \mathrm{an}(j)} \lambda_{jk}\chi(k,i) \quad \textit{with} \quad \lambda_{jk} = 1 - \sum_{\ell \in \mathrm{de}(k) \cap \mathrm{an}(j)} \lambda_{j\ell}. \tag{3.3.1}$$

Proof. As by Lemma 3.3.4(a) $\overline{b}_{ji} = \chi(j,i) - \sum_{k \in \mathrm{an}(j)} \overline{b}_{ki}$, it suffices to show that $\sum_{k \in \mathrm{an}(j)} \lambda_{jk}\chi(k,i) = \sum_{k \in \mathrm{an}(j)} \overline{b}_{ki}$. Using again Lemma 3.3.4(a) yields

$$\sum_{k \in \mathrm{an}(j)} \lambda_{jk}\chi(k,i) = \sum_{k \in \mathrm{an}(j)} \lambda_{jk} \sum_{\ell \in \mathrm{An}(k)} \overline{b}_{\ell i}.$$

Noting that $k \in \mathrm{an}(j)$ and $\ell \in \mathrm{An}(k)$ if and only if $\ell \in \mathrm{an}(j)$ and $k \in \mathrm{De}(\ell) \cap \mathrm{an}(j)$, we can interchange the two summation operators to obtain

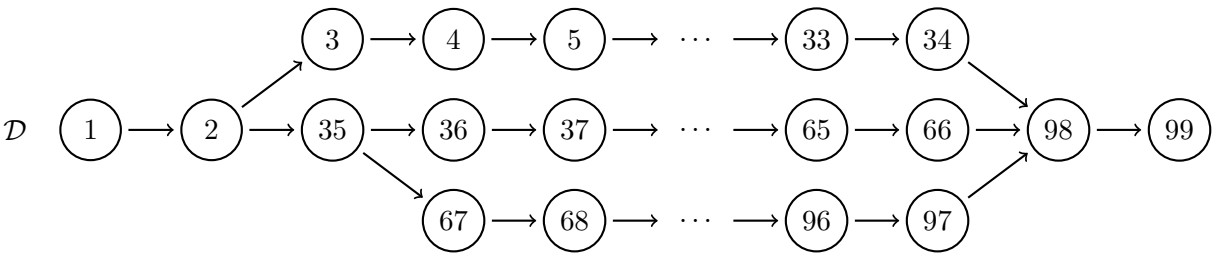
$$\sum_{k \in \mathrm{an}(j)} \lambda_{jk} \sum_{\ell \in \mathrm{An}(k)} \overline{b}_{\ell i} = \sum_{\ell \in \mathrm{an}(j)} \overline{b}_{\ell i} \sum_{k \in \mathrm{De}(\ell) \cap \mathrm{an}(j)} \lambda_{jk} = \sum_{\ell \in \mathrm{an}(j)} \overline{b}_{\ell i} \Big(\lambda_{j\ell} + \sum_{k \in \mathrm{de}(\ell) \cap \mathrm{an}(j)} \lambda_{jk} \Big) = \sum_{\ell \in \mathrm{an}(j)} \overline{b}_{\ell i},$$

where we have used the definition of $\lambda_{j\ell}$ for the last equality. □

Before we give an example of representation (3.3.1), we summarize some characteristics of the coefficients λ_{jk}. Denoting by $\mathrm{pa}^{\mathrm{tr}}(j)$ the parents of j in the transitive reduction $\mathcal{D}^{\mathrm{tr}}$ of $\mathcal{D}$, we have $\lambda_{jk} = 1$ for $k \in \mathrm{pa}^{\mathrm{tr}}(j)$ as $\mathrm{de}(k) \cap \mathrm{an}(j) = \varnothing$. For $k \in \mathrm{an}(j) \setminus \mathrm{pa}^{\mathrm{tr}}(j)$ it can be verified that $\lambda_{jk} \neq 0$ if and only if there exists no $\widetilde{k} \in \mathrm{de}(k) \cap \mathrm{an}(j)$ such that $|\mathrm{De}(\widetilde{k}) \cap \mathrm{pa}^{\mathrm{tr}}(j)| = |\mathrm{De}(k) \cap \mathrm{pa}^{\mathrm{tr}}(j)|$.

Example 3.3.7. [On representation (3.3.1)]
Consider a RMWM $\boldsymbol{X}$ on the DAG $\mathcal{D}$ depicted below, and note that here $\mathcal{D} = \mathcal{D}^{\mathrm{tr}}$. We determine, as an example, representation (3.3.1) for the MLCs $\overline{b}_{36,66}$ and $\overline{b}_{98,99}$:

$$\overline{b}_{36,66} = \chi(36,66) - \chi(35,66),$$
$$\overline{b}_{98,99} = \chi(98,99) - \chi(34,99) - \chi(66,99) - \chi(97,99) + \chi(2,99) + \chi(35,99).$$



□

We address again the interrelations between the TDCs and prove that every TDC can be written as linear combination of minima of two TDCs.

Proposition 3.3.8. *For $i, j \in V$,*

$$\chi(i,j) = \sum_{k \in \mathrm{An}(i) \cap \mathrm{An}(j)} \mu_{ij,k} (\chi(k,i) \wedge \chi(k,j)) \quad \textit{with} \quad \mu_{ij,k} = 1 - \sum_{\ell \in \mathrm{de}(k) \cap \mathrm{An}(i) \cap \mathrm{An}(j)} \mu_{ij,\ell}. \tag{3.3.2}$$

Proof. Applying Lemma 3.3.4(a) and Lemma 3.A.1(b), (d) we obtain for $k \in \mathrm{An}(i) \cap \mathrm{An}(j)$,

$$\chi(k,i) \wedge \chi(k,j) = \frac{\overline{b}_{ki}}{\overline{b}_{kk}} \wedge \frac{\overline{b}_{kj}}{\overline{b}_{kk}} = \Big(\frac{\overline{b}_{ki}}{\overline{b}_{kk}} \wedge \frac{\overline{b}_{kj}}{\overline{b}_{kk}} \Big) \Big(\sum_{\ell \in \mathrm{An}(k)} \overline{b}_{\ell k} \Big) = \sum_{\ell \in \mathrm{An}(k)} \frac{\overline{b}_{\ell k} \overline{b}_{ki}}{\overline{b}_{kk}} \wedge \frac{\overline{b}_{\ell k} \overline{b}_{kj}}{\overline{b}_{kk}} = \sum_{\ell \in \mathrm{An}(k)} \overline{b}_{\ell i} \wedge \overline{b}_{\ell j}.$$

With this we then have

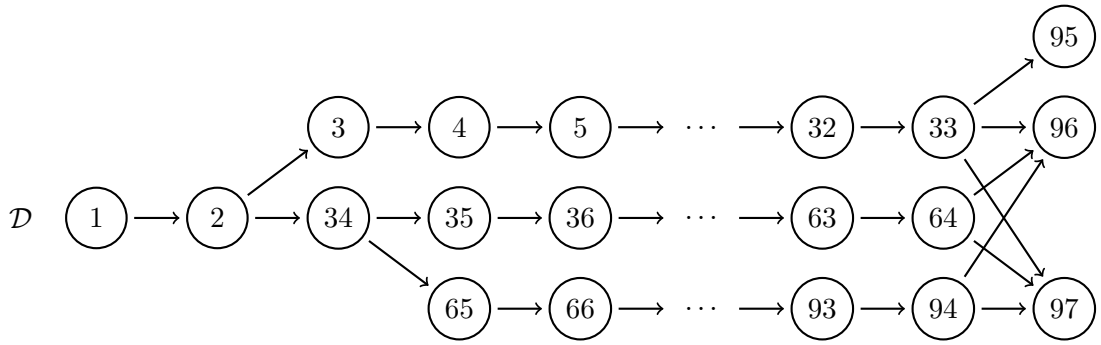
$$\sum_{k \in \mathrm{An}(i) \cap \mathrm{An}(j)} \mu_{ij,k} (\chi(k,i) \wedge \chi(k,j)) = \sum_{k \in \mathrm{An}(i) \cap \mathrm{An}(j)} \mu_{ij,k} \sum_{\ell \in \mathrm{An}(k)} \overline{b}_{\ell i} \wedge \overline{b}_{\ell j}.$$

Using that $k \in \mathrm{An}(i) \cap \mathrm{An}(j)$ and $\ell \in \mathrm{An}(k)$ if and only if $\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)$ and $k \in \mathrm{De}(\ell) \cap \mathrm{An}(i) \cap \mathrm{An}(j)$ to interchange the summation operators similarly as in the proof of Proposition 3.3.6 and the definition of $\mu_{ij,\ell}$ similarly as the one of $\lambda_{j\ell}$ there, we finally find (3.3.2). □

For $i, j \in V$ denote by $\mathrm{lca}(i,j)$ the lowest common ancestors of i and j; i.e., $k \in \mathrm{lca}(i,j)$ if and only if $k \in \mathrm{An}(i) \cap \mathrm{An}(j)$ and $\mathcal{D}$ has no path from k to another node in $\mathrm{An}(i) \cap \mathrm{An}(j)$. For $\mu_{ij,k}$ from (3.3.2) we have $\mu_{ij,k} = 1$ for $k \in \mathrm{lca}(i,j)$ as in that case $\mathrm{de}(k) \cap \mathrm{An}(i) \cap \mathrm{An}(j) = \varnothing$. It can be verified that $\mu_{ij,k} = 0$ for $k \in (\mathrm{An}(i) \cap \mathrm{An}(j)) \setminus \mathrm{lca}(i,j)$ if and only if there exists some $\widetilde{k} \in \mathrm{de}(k) \cap \mathrm{An}(i) \cap \mathrm{An}(j)$ such that $|\mathrm{De}(\widetilde{k}) \cap \mathrm{lca}(i,j)| = |\mathrm{De}(k) \cap \mathrm{lca}(i,j)|$. With this, if $j \in \mathrm{An}(i)$, then $\mu_{ij,j} = 1$ and $\mu_{ij,k} = 0$ for $k \in \mathrm{an}(j)$. Thus in that case the right-hand side of the first equality in (3.3.2) is equal to $\chi(j,i) \wedge \chi(j,j) = \chi(j,i)$, and representation (3.3.2) is trivial. Note the analogy of the coefficients $\mu_{ij,k}$ to the coefficients λ_{jk} in (3.3.1).

Example 3.3.9. [On representation (3.3.2)]
Consider a RMWM on the DAG $\mathcal{D}$ depicted below. We present, as an example, representation (3.3.2) for the TDCs $\chi(95,96)$ and $\chi(96,97)$:

$$\begin{aligned}
\chi(95,96) =& \chi(33,95) \wedge \chi(33,96), \\
\chi(96,97) =& \chi(33,96) \wedge \chi(33,97) + \chi(64,96) \wedge \chi(64,97) + \chi(94,96) \wedge \chi(94,97) \\
& - \chi(34,96) \wedge \chi(34,97) - \chi(2,96) \wedge \chi(2,97).
\end{aligned}$$



□

We conclude this section with necessary and sufficient conditions on a matrix to be the TDM of a RMWM on a given DAG $\mathcal{D}$. To be such a matrix, the ij-th (ji-th) entry of the matrix must satisfy a property depending on the relationship between i and j in $\mathcal{D}$. For example, based on Theorem 3.2.3, it must be zero if and only if $\mathrm{An}(i) \cap \mathrm{An}(j) = \varnothing$. By Lemma 3.A.1(e), Remark 3.3.3(i), and Theorem 2.5.4, a RMWM on $\mathcal{D}$ is a RMWM on every DAG that has reachability matrix R of $\mathcal{D}$. Consequently, it would be sufficient to specify R and to require the four conditions below for any DAG with reachability matrix R such as the transitive reduction $\mathcal{D}^{\mathrm{tr}}$ of $\mathcal{D}$.

Theorem 3.3.10. *Let $\mathcal{D} = (V, E)$ be a DAG with nodes $V = \{1, \dots, d\}$ and reachability matrix R. Let $\chi = (\chi(i,j))_{d\times d}$ be a symmetric matrix with ones on the diagonal. For $i \in V$ define $\overline{b}_{ii} := 1 - \sum_{k\in\mathrm{an}(i)} \overline{b}_{kk}\chi(k,i)$ recursively. Then χ is the TDM of a RMWM $\boldsymbol{X}$ on $\mathcal{D}$ if and only if the following conditions hold:*

(a) $\mathrm{sgn}(\chi) = \mathrm{sgn}(R^T R)$.

(b) For all $i \in V$, $\overline{b}_{ii} > 0$.

(c) For all $i \in V$, $j \in \mathrm{an}(i)$, and $k \in \mathrm{de}(j) \cap \mathrm{pa}(i)$, $\chi(j,i) = \chi(j,k)\chi(k,i)$.

(d) For all $i, j \in V$ such that $i \notin \mathrm{An}(j)$ and $j \notin \mathrm{An}(i)$ but $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \varnothing$,

$$\chi(i,j) = \sum_{k\in\mathrm{An}(i)\cap\mathrm{An}(j)} \overline{b}_{kk}(\chi(k,i) \wedge \chi(k,j)).$$

In that case $\overline{b}_{ii}$ is the i-th diagonal entry of the standardized MLCM $\overline{B}$ of $\boldsymbol{X}$. Furthermore, for $i, j \in V$, $\overline{b}_{ji} = 0$ if $j \in V \setminus \mathrm{An}(i)$, and $\overline{b}_{ji} = \overline{b}_{jj}\chi(j,i)$ if $j \in \mathrm{an}(i)$.

Proof. Assume that χ is the TDM of a RMWM $\boldsymbol{X}$ on $\mathcal{D}$. The statements (a) and (c) follow from Remark 3.2.4(i) and Lemma 3.3.4(b). By Lemma 3.3.4(a) $\overline{b}_{ii}$ is the i-th diagonal entry of the standardized MLCM $\overline{B}$ of $\boldsymbol{X}$. Since all $\overline{b}_{ii}$ are positive according to Lemma 3.A.1(a), assertion (b) holds. The representation of $\chi(i,j)$ in (d) is again a consequence of Lemma 3.3.4(a).

Assume now that (a)–(d) hold. For every $i \in V$ define $\overline{b}_{ji} := \overline{b}_{jj}\chi(j,i)$ for all $j \in \mathrm{an}(i)$ and for all $j \in V \setminus \mathrm{An}(i)$, $\overline{b}_{ji} := 0$. We first show that $\overline{B} = (\overline{b}_{ij})_{d\times d}$ is the MLCM of a RMWM on $\mathcal{D}$, where weights from its representation (3.1.1) are given by $c_{ii} := \overline{b}_{ii}$ and $c_{ki} := \frac{\overline{b}_{ki}}{\overline{b}_{kk}} = \chi(k,i)$ for $i \in V$ and $k \in \mathrm{pa}(i)$. As $\mathrm{sgn}(\chi) = \mathrm{sgn}(R^T R)$ and $\overline{b}_{ii} > 0$, the weights c_{ki} for $i \in V$ and $k \in \mathrm{Pa}(i)$ are positive, which is a necessary condition for them by the definition of a RMLM in (3.1.1). Let $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$ be a path in $\mathcal{D}$. Using (c) iteratively yields

$$d_{ji}(p) = c_{jj}\prod_{\nu=0}^{n-1} c_{k_\nu,k_{\nu+1}} = \overline{b}_{jj}\prod_{\nu=0}^{n-1}\chi(k_\nu,k_{\nu+1}) = \overline{b}_{jj}\chi(j,k_2)\prod_{\nu=2}^{n-1}\chi(k_\nu,k_{\nu+1}) = \ldots = \overline{b}_{jj}\chi(i,j) = \overline{b}_{ji}.$$

This implies that $\overline{B} = (\overline{b}_{ij})_{d\times d}$ is the MLCM of a RMWM $\boldsymbol{X}$. Since it suffices to specify one RMLM that has TDM χ, we may assume that $Z \in \mathrm{RV}(1)$. Denoting the TDM of $\boldsymbol{X}$ by $\overline{\chi} = (\overline{\chi}(i,j))_{d\times d}$, it remains to show that $\overline{\chi} = \chi$. Since the diagonal entries of χ equal one, the equality of the diagonal entries is obvious. For $i, j \in V$ such that $\mathrm{An}(i) \cap \mathrm{An}(j) = \varnothing$, the ij-th

(ji-th) entries of χ and $\overline{\chi}$ are zero and, hence, equal due to condition (a) and Theorem 3.2.3. The matrix $\overline{B}$ is the standardized MLCM of $\boldsymbol{X}$ as $\alpha = 1$ and $\bar{b}_{ii} = 1 - \sum_{k\in\mathrm{an}(i)} \bar{b}_{ki}$ for every $i \in V$. Thus for $i \in V$ and $j \in \mathrm{an}(i)$ we have by Lemma 3.3.4(a) and the definition of $\overline{B}$ that $\overline{\chi}(j,i) = \frac{\bar{b}_{ji}}{\bar{b}_{jj}} = \chi(i,j)$. Finally, for $i, j \in V$ such that $j \notin \mathrm{An}(i)$ and $i \notin \mathrm{An}(j)$ but $\mathrm{An}(i)\cap\mathrm{An}(j) \neq \varnothing$, using Lemma 3.3.4(a), the result shown before, and condition (d), we obtain

$$\overline{\chi}(i,j) = \sum_{k\in\mathrm{An}(i)\cap\mathrm{An}(j)} \bar{b}_{kk}(\overline{\chi}(k,i) \wedge \overline{\chi}(k,j)) = \sum_{k\in\mathrm{An}(i)\cap\mathrm{An}(j)} \bar{b}_{kk}(\chi(k,i) \wedge \chi(k,j)) = \chi(i,j).$$

□

In Example 3.5.5 below we present a possible application of Theorem 3.3.10.

Remark 3.3.11. In Theorem 3.3.10 the coefficients $\bar{b}_{ii}$ can also be defined by $1-\sum_{k\in\mathrm{an}(i)} \lambda_{ik}\chi(k,i)$ with λ_{ik} as in (3.3.1). We give a sketch of a proof of this assertion: we show that $\lambda_{ik} = 1 - \sum_{\ell\in\mathrm{de}(k)\cap\mathrm{an}(i)} \lambda_{\ell k}$ and use this to verify that if (c) holds, then the assertion is valid as well. Moreover, condition (d) can be replaced by

(d') For all $i, j \in V$ such that $i \notin \mathrm{An}(j)$ and $j \notin \mathrm{An}(i)$ but $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \varnothing$,

$$\chi(i,j) = \sum_{k\in\mathrm{An}(i)\cap\mathrm{An}(j)} \mu_{ij,k}(\chi(k,i) \wedge \chi(k,j)) \quad \text{with} \quad \mu_{ij,k} \text{ as in (3.3.2).}$$

By going through the proof of Theorem 3.3.10, we observe that this can be done due to the representation of $\chi(i,j)$ in (3.3.2). □

3.4 Identifiability problems based on the tail dependence matrix of a recursive ML model

Throughout this section we assume that the TDM χ of a RMLM $\boldsymbol{X}$ on a DAG $\mathcal{D}$ with standardized MLCM $\overline{B}$ is given. We first show the identifiability of $\overline{B}$ from χ and the reachability matrix R of $\mathcal{D}$. We then assume that the reachability relation of $\mathcal{D}$ is not fully known but only a causal ordering σ. This still leads to identifiability of $\overline{B}$ from χ. We also investigate whether $\overline{B}$ can be recovered from χ and the initial nodes V_0 of $\mathcal{D}$. It turns out that this is generally not possible, but we verify it for RMWMs. We prove the different identifiability results by providing algorithms which compute $\overline{B}$ from χ and the additionally known information on $\mathcal{D}$. Finally, based on these results we present an approach, which finds the standardized MLCMs of all RMLMs with TDM χ. Since this method simplifies for RMWMs considerably, we give an adapted and modified version for this subclass of RMLMs.

3.4.1 Identifiability from the tail dependence matrix and the reachability matrix

The following algorithm computes $\overline{B}$ from χ and R recursively. The rows of $\overline{B}$ are filled up successively until $\overline{B}$ is obtained, where the number of ancestors determines the order in which

the rows are treated. The existence of such an algorithm proves the identifiability of $\overline{B}$ from χ and R.

Algorithm 3.4.1. [Find $\overline{B}$ from χ and R]
For $\nu = 0, \dots, d-1$,
for $j \in V$ such that $|\mathrm{an}(j)| = \nu$, set

$$\overline{b}_{ji} = 0 \text{ for all } i \in V \setminus \mathrm{De}(j) \quad \text{and} \quad \overline{b}_{ji} = \chi(j,i) - \sum_{k \in \mathrm{an}(j)} \overline{b}_{ki} \wedge \overline{b}_{kj} \text{ for all } i \in \mathrm{De}(j). \tag{3.4.1}$$

Eq. (3.4.1) follows from Lemma 3.A.1(a), (3.1.8), and Lemma 3.A.1(f). If $\boldsymbol{X}$ is max-weighted, then by Lemma 3.3.4(a) (3.4.1) can be replaced by

$$\overline{b}_{ji} = 0 \text{ for all } i \in V \setminus \mathrm{De}(j), \quad \overline{b}_{jj} = 1 - \sum_{k \in \mathrm{an}(j)} \overline{b}_{kj}, \quad \text{and} \quad \overline{b}_{ji} = \overline{b}_{jj}\chi(j,i) \text{ for all } i \in \mathrm{de}(j). \tag{3.4.2}$$

To avoid the iterative loop, we can also use (3.3.1) for computing the diagonal entries of $\overline{B}$. Note, however, that this requires to calculate the coefficients λ_{jk} appearing in (3.3.1) recursively as well.

3.4.2 Identifiability from the tail dependence matrix and a causal ordering

So far we have dealt with the identifiability from χ and the reachability matrix R of $\mathcal{D}$. Here we investigate the identifiability from χ and a causal ordering σ of $\mathcal{D}$. If R is given, then we know for every two (distinct) $i, j \in V$ whether there is a path from j to i; but from σ we only learn that there is no path from j to i if $\sigma(j) > \sigma(i)$.

There exists a causal ordering for every DAG due to the acyclicity (see also Diestel [15], Appendix A). However, it is not necessarily unique. For example, the DAG $\mathcal{D}_1$ from Example 3.2.2 has the identity function on $V = \{1,2,3,4\}$ and the permutation $\widetilde{\sigma}$ on V given by $\widetilde{\sigma}(2) = 1$, $\widetilde{\sigma}(1) = 2$, $\widetilde{\sigma}(3) = 3$, $\widetilde{\sigma}(4) = 4$ as causal orderings.

The DAG $\mathcal{D}$ has a causal ordering which can be completely described by its initial nodes V_0 and χ as follows.

Lemma 3.4.2. *We denote the initial nodes by $V_0 = \{i_1, \dots, i_{|V_0|}\}$ and define $V_0^i := \{k \in V_0 : \chi(k,i) > 0\}$ for $i \in V$. Then $\mathcal{D}$ has a causal ordering σ such that*

$$\sigma(i_\nu) = \nu \text{ for } \nu = 1, \dots, |V_0| \quad \text{and} \quad \text{for all } i,j \in V,\ \sigma(j) < \sigma(i) \text{ whenever } |V_0^j| < |V_0^i|. \tag{3.4.3}$$

Proof. Recall from Lemma 3.2.5(c) that $V_0^j = \mathrm{An}(j) \cap V_0$ and $V_0^i = \mathrm{An}(i) \cap V_0$. With this it is not difficult to see that $\mathcal{D}$ has such a causal ordering. □

Now we give an iterative procedure which computes $\overline{B}$ from χ and σ. Obviously, this proves the identifiability of $\overline{B}$ from χ and σ. Here the rows of $\overline{B}$ are also filled up successively, where the order of the nodes given by σ defines the order in which the rows are treated.

Algorithm 3.4.3. [Find $\overline{B}$ from χ and σ]
For $\nu = 1, \ldots, d$,
for $j \in V$ such that $\sigma(j) = \nu$, set

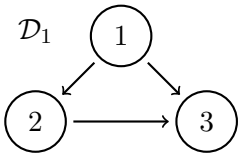
$$\begin{aligned} &\overline{b}_{ji} = 0 \text{ for all } i \in V \text{ such that } \sigma(j) > \sigma(i), \\ &\overline{b}_{ji} = \chi(j,i) - \bigvee_{k:\sigma(k)<\sigma(j)} \overline{b}_{ki} \wedge \overline{b}_{kj} \text{ for all } i \in V \text{ such that } \sigma(j) \leq \sigma(i). \end{aligned} \tag{3.4.4}$$

Eq. (3.4.4) can be obtained from (3.1.8) by using Lemma 3.A.1(a), the definition of a causal ordering, and Lemma 3.A.1(f).

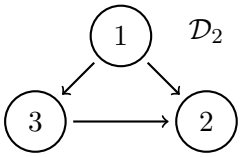
3.4.3 Identifiability of recursive max-weighted models from the tail dependence matrix and the initial nodes

In what follows we assume $\boldsymbol{X}$ to be max-weighted. Then recalling Lemma 3.2.5(c), Proposition 3.3.5 involves a procedure to determine R from χ and V_0. Since Algorithm 3.4.1 computes $\overline{B}$ from χ and R, we can identify $\overline{B}$ from χ and V_0. This is usually not possible outside the class of RMWMs.

Example 3.4.4. [$\overline{B}$ is generally not identifiable from χ and V_0]
Consider two RMLMs on $\mathcal{D}_1$ and $\mathcal{D}_2$ with standardized MLCMs $\overline{B}_1$ and $\overline{B}_2$ given by



$$\overline{B}_1 = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0 & 0.8 & 0.4 \\ 0 & 0 & 0.3 \end{bmatrix} \quad \text{and} \quad \overline{B}_2 = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0 & 0.4 & 0 \\ 0 & 0.4 & 0.7 \end{bmatrix}.$$



We find by Lemma 3.A.1(d) that none of the two models is max-weighted. Since both have the same χ and $\mathcal{D}_1$ and $\mathcal{D}_2$ share the same initial node $V_0 = \{1\}$, we cannot distinguish between $\overline{B}_1$ and $\overline{B}_2$ based on χ and V_0. □

Proceeding as suggested by Proposition 3.3.5 to recover R from χ and V_0 is very tedious, since many conditions may need to be verified. Therefore, we introduce an alternative method which computes $\overline{B}$ from χ and V_0: we first determine a causal ordering σ of $\mathcal{D}$ and apply then Algorithm 3.4.3 to obtain $\overline{B}$. From the next proposition we learn how a causal ordering σ of $\mathcal{D}$ can be computed from χ and V_0; note that we encountered property (i) in (3.4.3).

Proposition 3.4.5. *Let V_0^i for $i \in V$ be as in Lemma 3.4.2. Every permutation σ on V such that for all $i, j \in V$,*

(i) $\sigma(j) < \sigma(i)$ whenever $|V_0^j| < |V_0^i|$ and

(ii) $\sigma(j) < \sigma(i)$ whenever $|V_0^j| = |V_0^i|$ and $\max_{k \in V_0^i} \chi(k,i) < \max_{k \in V_0^j} \chi(k,j)$

is a causal ordering of $\mathcal{D}$.

Proof. Assume that σ is no causal ordering of $\mathcal{D}$, i.e., $\sigma(j) > \sigma(i)$ for some $i \in V$ and $j \in \mathrm{an}(i)$. Recall from Lemma 3.2.5(c) that $V_0^j = \mathrm{An}(j) \cap V_0$ and $V_0^i = \mathrm{An}(i) \cap V_0$. As $j \in \mathrm{an}(i)$, $V_0^j \subseteq V_0^i$. But then because of the properties of σ, $V_0^j = V_0^i$ and $\max_{k \in V_0^j} \chi(k,j) \leq \max_{k \in V_0^j} \chi(k,i)$. Assume now that $j \in V_0^j$, and note that $i \notin V_0^j$ as $j \in \mathrm{an}(i)$. Then, since for $i_1, i_2 \in V$ the TDC $\chi(i_1, i_2) = 1$ if and only if $i_1 = i_2$ (cf. (3.1.8) and Lemma 3.A.1(a)), we find $1 = \max_{k \in V_0^j} \chi(k,j) \leq \max_{k \in V_0^j} \chi(k,i) < 1$. This contradiction proves that $j \notin V_0^j$, which implies again that $V_0^j = \mathrm{an}(j) \cap V_0$. As $\max_{k \in V_0^j} \chi(k,j) \leq \max_{k \in V_0^j} \chi(k,i)$, $\chi(k,j) \leq \chi(k,i)$ for some $k \in \mathrm{an}(j) \cap V_0$. Observe from Lemma 3.3.4(b) that $j \notin \mathrm{an}(i)$, since otherwise $\chi(k,i) < \chi(k,j)$. This, however, contradicts our original assumption, and σ must be a causal ordering of $\mathcal{D}$. □

Finally, we clarify the precise steps of our approach to determine $\overline{B}$ from χ and V_0.

Algorithm 3.4.6. [Modification of Algorithm 3.4.3 for RMWMs: find $\overline{B}$ from χ and V_0]

1. Find a causal ordering σ of $\mathcal{D}$ from χ and V_0:
   for $\nu = 1, \ldots, |V_0|$,

   find all $j \in V$ such that $|V_0^j| = |\{k \in V_0 : \chi(k,j) > 0\}| = \nu$ and summarize them in the set A_ν;

   sort the nodes $k_1, \ldots, k_{|A_\nu|}$ from A_ν such that

   $$\max_{\ell \in V_0} \chi(\ell, k_1) \geq \max_{\ell \in V_0} \chi(\ell, k_2) \geq \ldots \geq \max_{\ell \in V_0} \chi(\ell, k_{|A_\nu|});$$

   for $\mu = 1, \ldots, |A_\nu|$,

   set $\sigma(k_\mu) = \sum_{\lambda=1}^{\nu-1} |A_\lambda| + \mu$, where $\sum_{\lambda=1}^{0} := 0$.

2. Apply Algorithm 3.4.3 to obtain $\overline{B}$ from χ and σ.

Observe from Proposition 3.4.5 that every permutation σ on V which can be chosen in step 1. is indeed a causal ordering of $\mathcal{D}$.

3.4.4 Identifiability from the tail dependence matrix

We now combine the previous results to find the standardized MLCMs of all RMLMs that have TDM χ. In the first part we deal with general RMLMs. Because of the identifiability properties derived in Section 3.4.3, we assume in the second part that χ is the TDM of a RMWM. We provide an algorithm, which outputs the standardized MLCMs of all RMWMs that have TDM χ.

(General) recursive max-linear models

Every permutation $\widetilde{\sigma}$ on $V = \{1, \ldots, d\}$ is a causal ordering of a DAG with nodes V but not necessarily of a DAG that corresponds to a RMLM with TDM χ. But if this is the case, then applying Algorithm 3.4.3 with $\sigma = \widetilde{\sigma}$ yields the corresponding standardized MLCM $\overline{B}$. This suggests the following procedure to prove the existence of a RMLM which has TDM χ and whose associated DAG has causal ordering $\widetilde{\sigma}$: first apply Algorithm 3.4.3 with $\sigma = \widetilde{\sigma}$, and check

then whether the obtained matrix $\overline{B}$ is the standardized MLCM of a RMLM which has TDM χ and whose associated DAG has causal ordering $\widetilde{\sigma}$. In the second step it is enough to verify that $\overline{B}$ is the MLCM of a RMLM, which can be done by Theorem 2.5.7.

Lemma 3.4.7. *Let $\widetilde{\sigma}$ be a permutation on V and $\overline{B}$ the matrix obtained by applying Algorithm 3.4.3 with $\sigma = \widetilde{\sigma}$. If $\overline{B}$ is the MLCM of a RMLM (RMWM), then $\overline{B}$ is the standardized MLCM of a RMLM (RMWM) which has TDM χ and whose associated DAG has causal ordering $\widetilde{\sigma}$.*

Proof. Let $\boldsymbol{X}$ be the RMLM (RMWM) with MLCM $\overline{B}$ and $Z \in \mathrm{RV}(1)$. Its existence is guaranteed as $\overline{B}$ is the MLCM of a RMLM (RMWM). We show that $\boldsymbol{X}$ has standardized MLCM $\overline{B}$ and TDM χ as well as that its associated DAG $\mathcal{D}$ has causal ordering $\widetilde{\sigma}$. Recall from (3.1.5) that $\mathrm{sgn}(\overline{B})$ is the reachability matrix of $\mathcal{D}$. Thus by (3.4.4) $\widetilde{\sigma}$ is a causal ordering of $\mathcal{D}$ and $\overline{b}_{ii} = 1 - \sum_{k \in \mathrm{an}(i)} \overline{b}_{ki}$ for every $i \in V$. As the latter holds and $\alpha = 1$, $\overline{B}$ is the standardized MLCM of $\boldsymbol{X}$. The fact that $\boldsymbol{X}$ has TDM χ also follows from (3.4.4). □

So the following "naive" method finds the standardized MLCMs of all RMLMs that have TDM χ: for every permutation on V compute the matrix $\overline{B}$ from Algorithm 3.4.3, and check whether it is the MLCM of a RMLM; if so, then $\overline{B}$ is the standardized MLCM of a RMLM with TDM χ. However, the number of permutations on V to be investigated can often be significantly reduced. By Theorem 3.2.7 and Lemma 3.2.8(c) the set of all maximum χ-cliques W (see Definition 3.2.6) such that $\chi(i,j) \geq \sum_{k \in W} \chi(k,i) \wedge \chi(k,j)$ for all $i, j \in V \setminus W$ contains the initial node sets of all DAGs underlying RMLMs with TDM χ. Hence, it suffices to investigate the causal orderings of the DAGs that have such initial nodes W. But also the number of causal orderings to be investigated for every such set W can be reduced further by Lemma 3.4.2: it is enough to consider those permutations on V that satisfy the properties σ has in (3.4.3) with $V_0 = W$. The following algorithm describes the precise steps of the proposed method to find the standardized MLCMs of all RMLMs with TDM χ.

Algorithm 3.4.8. [Find all $\overline{B}$ from χ]

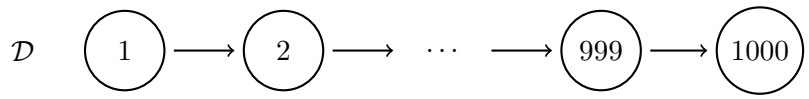
1. Find all maximum χ-cliques:
   (a) find the complement $\mathcal{D}^{\chi}$ of the χ-graph;
   (b) find all maximum cliques of $\mathcal{D}^{\chi}$.
2. For every maximum χ-clique $W = \{i_1, \ldots, i_{|W|}\}$,
   (a) check $\chi(i,j) \geq \sum_{k \in W} \chi(k,i) \wedge \chi(k,j)$ for all $i, j \in V \setminus W$;
   if not, then there is no RMLM with TDM χ on a DAG with initial nodes W;
   else,
   (b) for every permutation $\widetilde{\sigma}$ on $V = \{1, \ldots, d\}$ such that
   $\widetilde{\sigma}(i_\nu) = \nu$ for $\nu = 1, \ldots, |W|$ and
   $\widetilde{\sigma}(j) < \widetilde{\sigma}(i)$ whenever $|\{k \in W : \chi(k,j) > 0\}| < |\{k \in W : \chi(k,i) > 0\}|$,
   i. apply Algorithm 3.4.3 with $\sigma = \widetilde{\sigma}$;

ii. check whether $\overline{B}$ obtained in i. is the MLCM of a RMLM, for example, using Theorem 2.5.7;
if not, then there is no RMLM with TDM χ on a DAG with causal ordering $\widetilde{\sigma}$;
else, $\overline{B}$ is the standardized MLCM of a RMLM with TDM χ.

When the algorithm returns a standardized MLCM $\overline{B}$ of a RMLM with TDM χ in step ii., then it is not necessary to perform steps i., ii. for further permutations on V which are causal orderings of DAGs with reachability matrix $\operatorname{sgn}(\overline{B})$, since all of them would lead to the same $\overline{B}$. For the application of Algorithm 3.4.8, we have assumed so far that χ is the TDM of a RMLM. If this is not the case, Algorithm 3.4.8 would not produce any output. The same applies to Algorithm 3.4.11 below if χ is not the TDM of a RMWM.

One could drop step 2.(a) and perform step 2.(b) for all maximum χ-cliques. However, the performance of step 2.(a) can be very effective.

Example 3.4.9. [Not all maximum χ-cliques are initial node sets]
Consider the TDM χ of a RMLM on the DAG $\mathcal{D}$ depicted below. Note that such a RMLM is max-weighted, since $\mathcal{D}$ is a polytree (cf. Section 3.3.1). Theorem 3.2.3 yields that the sets $\{1\},\dots,\{1000\}$ are the maximum χ-cliques. For $k \in \{2,\dots,999\}$ we know from Lemma 3.3.4(b) that $\chi(1,1000) < \chi(1,k) \wedge \chi(k,1000)$. The property tested in step 2.(a) is therefore not fulfilled for the maximum χ-cliques $W \in \{\{2\},\dots,\{999\}\}$. However, we can verify by Lemma 3.3.4(b) that it is fullfilled for $W \in \{\{1\},\{1000\}\}$. Consequently, step 2.(b) needs only be performed for $W \in \{\{1\},\{1000\}\}$ and not for the other 998 maximum χ-cliques.



□

It is indeed necessary to perform step ii., i.e., to verify that a matrix $\overline{B}$ obtained in i. is a MLCM of a RMLM.

Example 3.4.10. [Not every $\overline{B}$ obtained in ii. belongs to a RMLM]
Consider the TDM

$$\chi = \begin{bmatrix} 1 & 1/10 & 1/3 \\ 1/10 & 1 & 13/30 \\ 1/3 & 13/30 & 1 \end{bmatrix}.$$

Performing steps i. and ii. of Algorithm 3.4.8 with $\widetilde{\sigma}$ being the identity function on $V = \{1,2,3\}$ and also with $\widetilde{\sigma}$ given by $\widetilde{\sigma}(1) = 1$, $\widetilde{\sigma}(3) = 2$, $\widetilde{\sigma}(2) = 3$ (note that these permutations are really tested in step 2.(b)), we find

$$\overline{B}_1 = \begin{bmatrix} 1 & 1/10 & 1/3 \\ 0 & 9/10 & 1/3 \\ 0 & 0 & 1/3 \end{bmatrix} \quad \text{and} \quad \overline{B}_2 = \begin{bmatrix} 1 & 1/10 & 1/3 \\ 0 & 17/30 & 0 \\ 0 & 1/3 & 2/3 \end{bmatrix}.$$

As can be verified by Theorem 2.4.2, the matrix $\overline{B}_1$ is the MLCM of a RMLM on the DAG $\mathcal{D}_1$ depicted in Example 3.4.4. Although $\operatorname{sgn}(B_2)$ is the reachability matrix of a DAG, namely of the DAG $\mathcal{D}_2$ from Example 3.4.4, which is a necessary property of a matrix to be the MLCM of a RMLM according to (3.1.5), it is no MLCM of a RMLM. □

Recursive max-weighted models

Assume now that χ is the TDM of a RMWM. We modify and adapt Algorithm 3.4.8 to obtain a procedure which outputs the standardized MLCMs of all RMWMs with TDM χ. Among the maximum χ-cliques which we find in step 2.(a) of Algorithm 3.4.8 are the initial node sets of the DAGs underlying the RMWMs that have TDM χ. We learn from Proposition 3.4.5 and Lemma 3.4.7 that a maximum χ-clique is such an initial node set if and only if the matrix $\overline{B}$ obtained by Algorithm 3.4.6 is the MLCM of a RMWM. In that case, $\overline{B}$ is obviously the standardized MLCM of a RMWM with TDM χ. These observations lead to the following procedure.

Algorithm 3.4.11. [Modification of Algorithm 3.4.8 for RMWMs: find all $\overline{B}$ from χ]

1. Find all maximum χ-cliques (cf. step 1. of Algorithm 3.4.8).
2. For every maximum χ-clique W,
   (a) check $\chi(i,j) \geq \sum_{k \in W} \chi(k,i) \wedge \chi(k,j)$ for all $i, j \in V \setminus W$;
   if not, then there is no RMWM with TDM χ on a DAG with initial nodes W;
   else,
      i. apply Algorithm 3.4.6 with $V_0 = W$;
      ii. check the following properties for the matrix $\overline{B}$ obtained in i.:
         - $\operatorname{sgn}(\overline{B})$ is the reachability matrix of a DAG
         - for all $i \in V$, $j \in \operatorname{an}(i)$, and $k \in \operatorname{de}(j) \cap \operatorname{pa}(i)$, $\overline{b}_{ji} = \frac{\overline{b}_{jk}\overline{b}_{ki}}{\overline{b}_{kk}}$

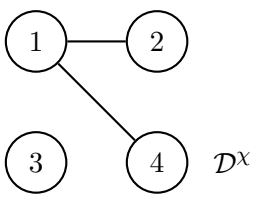
      if not, then there is no RMWM with TDM χ on a DAG with initial nodes W;
      else, $\overline{B}$ is the standardized MLCM of a RMWM with TDM χ.

That the properties we verify for the matrix $\overline{B}$ in step ii. are sufficient for $\overline{B}$ to be the MLCM of a RMWM can be verified by Corollary 2.4.3(a).

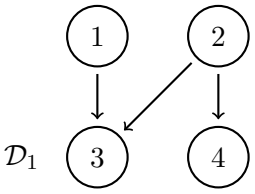
To conclude this section, we highlight the essential steps of Algorithm 3.4.11 with an example.

Example 3.4.12. [The class of RMWMs is not closed under χ-equivalence]
Consider the TDM

$$\chi = \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 0 & 1 & 0.6 & 0.5 \\ 0.2 & 0.6 & 1 & 0.5 \\ 0 & 0.5 & 0.5 & 1 \end{bmatrix}.$$

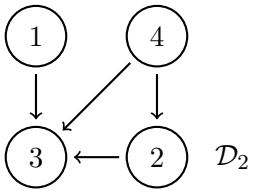


We read from the complement $\mathcal{D}^\chi$ of the χ-graph that the sets $W_1 = \{1,2\}$ and $W_2 = \{1,4\}$ are the maximum χ-cliques. Applying Algorithm 3.4.6 with $V_0 = W_1$ and $V_0 = W_2$, we get the matrices



$\mathcal{D}_1$

$$\overline{B}_1 = \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 0 & 1 & 0.6 & 0.5 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix} \text{ and } \overline{B}_2 = \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 0 & 0.5 & 0.1 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0.5 & 0.5 & 1 \end{bmatrix}.$$



$\mathcal{D}_2$

The matrix $\overline{B}_1$ is the MLCM of a RMWM on $\mathcal{D}_1$, whereas $\overline{B}_2$ is not the MLCM of a RMWM, but it is the MLCM of a RMLM on $\mathcal{D}_2$. Therefore, all RMWMs with TDM χ have the same standardized MLCM $\overline{B}_1$, and $\mathcal{D}_1$ is their associated DAG. Furthermore, all these models are χ-equivalent to the RMLMs with standardized MLCM $\overline{B}_2$. □

3.5 χ-equivalent recursive ML models and their DAGs

In this section we mainly present interrelations between DAGs of χ-equivalent RMLMs.

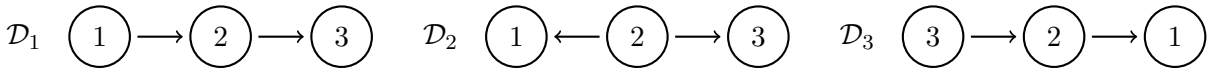
One of the best known equivalence relations on the set of DAGs is certainly the Markov equivalence: two DAGs are Markov equivalent if they entail the same conditional independence relations via the Markov property; for a characterization of such DAGs, see e.g. Verma and Pearl [71]. The associated DAG of a recursive linear Gaussian structural equation model can be identified from the distribution only up to a Markov equivalence class (under the assumption of faithfulness; see e.g. Spirtes and Zhang [68]). In the following example we discuss the relation between χ-equivalence of RMLMs and Markov equivalence of their associated DAGs.

Example 3.5.1. [The difference between χ-equivalence of RMLMs and Markov equivalence of their DAGs]

(1) Undirected graphs underlying Markov equivalent DAGs coincide. Example 3.4.12 clarifies that this does not hold for DAGs of χ-equivalent RMLMs. Such DAGs are therefore not necessarily Markov equivalent.

(2) For the TDCs of a RMLM $\boldsymbol{X}$ on $\mathcal{D}_1$, which is always a RMWM, we have by Lemma 3.3.4(b) that $\chi(1,3) < \chi(1,2) \wedge \chi(2,3)$. Since $\mathcal{D}_2$ has initial node 2, by Lemma 3.2.8(c) there cannot be a RMLM that is χ-equivalent to $\boldsymbol{X}$ on $\mathcal{D}_2$. Thus although the DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent, there exist no χ-equivalent RMLMs on $\mathcal{D}_1$ and $\mathcal{D}_2$.

(3) As can be verified by Theorem 3.3.10, RMLMs on the Markov equivalent DAGs $\mathcal{D}_1$ and $\mathcal{D}_3$ are always χ-equivalent. This shows that there can be χ-equivalent RMLMs on Markov equivalent DAGs.



□

DAGs of χ-equivalent RMLMs have the same number of initial nodes, since the initial node sets of such DAGs are maximum χ-cliques, which have the same cardinality by definition. We learn from Algorithm 3.4.3 that if the standardized MLCMs of two χ-equivalent RMLMs differ, then the causal orderings of their associated DAGs must also differ. So for these two DAGs there exist nodes $i, j \in V$ such that one DAG has a path from j to i and the other has one from i to j. We provide further properties of two DAGs underlying χ-equivalent RMLMs.

Proposition 3.5.2. *Let $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ be χ-equivalent RMLMs on DAGs $\mathcal{D}$ and $\widetilde{\mathcal{D}}$, respectively. We denote the initial nodes in $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ by V_0 and $\widetilde{V}_0$, the ancestors of i by $\mathrm{an}(i)$ and $\widetilde{\mathrm{an}}(i)$, and the descendants of i by $\mathrm{de}(i)$ and $\widetilde{\mathrm{de}}(i)$.*

(a) There is only one bijection $\varphi : V_0 \to \widetilde{V}_0$ such that for every $j \in V_0$, $\chi(j, \varphi(j)) > 0$ and $\chi(j, \widetilde{j}) = 0$ for all $\widetilde{j} \in \widetilde{V}_0 \setminus \{\varphi(j)\}$.

Let φ be the bijection from (a) and $j \in V_0$.

(b) We have $\mathrm{An}(\varphi(j)) \cap V_0 = \widetilde{\mathrm{De}}(\varphi(j)) \cap V_0 = \{j\}$. In particular, if $j \neq \varphi(j)$, then $\mathcal{D}$ has a path from j to $\varphi(j)$, and $\widetilde{\mathcal{D}}$ has one from $\varphi(j)$ to j.

(c) We have $\mathrm{De}(j) = \widetilde{\mathrm{De}}(\varphi(j))$.

(d) For $i \in V$, $\widetilde{\mathrm{An}}(i) \cap \widetilde{V}_0 = \{\varphi(j) : j \in \mathrm{An}(i) \cap V_0\}$.

Proof. (a) is immediate by Lemma 3.2.8(a), since $\widetilde{V}_0$ is a maximum χ-clique.
(b) Since $\widetilde{V}_0$ is a maximum χ-clique, according to Lemma 3.2.8(b), $\mathrm{An}(\varphi(j)) \cap V_0 = \{j\}$. Note that for every $\widetilde{j} \in \widetilde{V}_0$, $\chi(\widetilde{j}, \varphi^{-1}(\widetilde{j})) > 0$ and $\chi(\widetilde{j}, j) > 0$ for all $j \in V_0 \setminus \{\varphi^{-1}(\widetilde{j})\}$, where $\varphi^{-1} : \widetilde{V}_0 \to V_0$ denotes the inverse of φ. As V_0 is a maximum χ-clique, we therefore have again by Lemma 3.2.8(b) that $\widetilde{\mathrm{De}}(i) \cap V_0 = \{\varphi^{-1}(i)\}$ with $i = \varphi(j)$, which is obviously equivalent to $\widetilde{\mathrm{De}}(\varphi(j)) \cap V_0 = \{j\}$.
(c) Let $i \in \mathrm{De}(j)$. By (b) $j \in \mathrm{An}(\varphi(j)) \cap \mathrm{An}(i)$ and, consequently, by Theorem 3.2.3 $\chi(\varphi(j), i) > 0$. Lemma 3.2.5(c) then yields that $i \in \widetilde{\mathrm{De}}(\varphi(j))$. Hence, $\mathrm{De}(j) \subseteq \widetilde{\mathrm{De}}(\varphi(j))$. From this, by reversing the roles of $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ and noting that $\chi(\widetilde{j}, \varphi^{-1}(\widetilde{j})) > 0$ for all $\widetilde{j} \in \widetilde{V}_0$, we observe that $\widetilde{\mathrm{De}}(\varphi(j)) \subseteq \mathrm{De}(j)$.
(d) can be verified by (c). □

Recursive max-weighted models

Now we consider χ-equivalent RMWMs and investigate their DAGs. Because of Theorem 3.2.7, Algorithm 3.4.6, and Lemma 3.A.1(e), if a TDM χ of a RMWM has one maximum χ-clique W, all RMWMs with TDM χ (the models are then χ-equivalent by definition) have the same standardized MLCM and, hence, the same minimum ML DAG, which again has initial nodes W. By Algorithm 3.4.6 the initial nodes of DAGs of χ-equivalent RMLMs with different standardized

MLCMs must also differ. We present further interrelationships between DAGs of χ-equivalent RMWMs with regard to their initial nodes.

Theorem 3.5.3. *Let $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ be χ-equivalent RMWMs on DAGs $\mathcal{D}$ and $\widetilde{\mathcal{D}}$, respectively. We denote by V_0 and $\widetilde{V}_0$ the initial nodes in $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ and by V_∞ and $\widetilde{V}_\infty$ their terminal nodes. Let $\varphi : V_0 \to \widetilde{V}_0$ be the bijection from Proposition 3.5.2(a) and $j \in V_0$ such that $j \neq \varphi(j)$.*

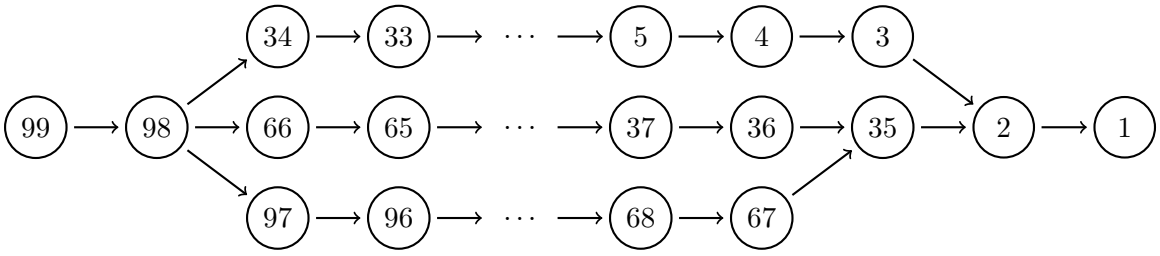
(a) We have $\varphi(j) \in V_\infty$. In particular, $\widetilde{V}_0 \subseteq (V_0 \cap \widetilde{V}_0) \cup V_\infty$.

(b) If $p = [j = k_0 \to k_1 \to \cdots \to k_{n-1} \to k_n = \varphi(j)]$ is a path in the transitive reduction $\mathcal{D}^{\mathrm{tr}}$ of $\mathcal{D}$, then $\widetilde{p} = [\varphi(j) = k_n \to k_{n-1} \to \cdots \to k_1 \to k_0 = j]$ is a path in the transitive reduction $\widetilde{\mathcal{D}}^{\mathrm{tr}}$ of $\widetilde{\mathcal{D}}$.

Proof. We denote by $\mathrm{an}(i)$ and $\widetilde{\mathrm{an}}(i)$ the ancestors of i in $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ and by $\mathrm{de}(i)$ and $\widetilde{\mathrm{de}}(i)$ its descendants.
(a) Assume that $\varphi(j) \notin V_\infty$. Consequently, by Proposition 3.5.2(b) $\mathcal{D}$ has a path from j to some $i \neq \varphi(j)$ passing through $\varphi(j)$. Replacing V_0 by $\widetilde{V}_0$, we learn from the the proof of Lemma 3.2.8(c) that $\chi(j,i) \geq \chi(j,\varphi(j)) \wedge \chi(\varphi(j),i)$. But this contradicts Lemma 3.3.4(b). Hence, $\varphi(j) \in V_\infty$.
(b) To prove that $\widetilde{p}$ is a path in $\widetilde{\mathcal{D}}^{\mathrm{tr}}$, because of the properties of $\widetilde{\mathcal{D}}^{\mathrm{tr}}$, it suffices to show that for $\nu = 0, \ldots, n-1$, $k_{\nu+1} \in \widetilde{\mathrm{an}}(k_\nu)$ and $\widetilde{\mathrm{de}}(k_{\nu+1}) \cap \widetilde{\mathrm{an}}(k_\nu) = \varnothing$. Recalling from Proposition 3.5.2(b) that $\mathrm{An}(\varphi(j)) \cap V_0 = \{j\}$, we observe that $\mathrm{An}(k_\nu) \cap \mathrm{An}(k_{\nu+1}) \cap V_0 = \{j\}$. We then obtain from Proposition 3.5.2(d) that $\widetilde{\mathrm{An}}(k_\nu) \cap \widetilde{\mathrm{An}}(k_{\nu+1}) \cap \widetilde{V}_0 = \{\varphi(j)\}$. By Lemma 3.3.4(b) we have $\chi(k_\nu, \varphi(j)) = \chi(k_\nu, k_{\nu+1})\chi(k_{\nu+1}, \varphi(j))$. As $\widetilde{\mathrm{An}}(k_\nu) \cap \widetilde{\mathrm{An}}(k_{\nu+1}) \cap \widetilde{V}_0 = \{\varphi(j)\}$, using Proposition 3.3.5 then proves that $k_{\nu+1} \in \widetilde{\mathrm{an}}(k_\nu)$. To show that $\widetilde{\mathrm{de}}(k_{\nu+1}) \cap \widetilde{\mathrm{an}}(k_\nu) = \varnothing$, assume the converse. Let $\ell \in \widetilde{\mathrm{de}}(k_{\nu+1}) \cap \widetilde{\mathrm{an}}(k_\nu)$. By reversing the roles of $\mathcal{D}^{\mathrm{tr}}$ and $\widetilde{\mathcal{D}}^{\mathrm{tr}}$ and noting that for every $\widetilde{j} \in \widetilde{V}_0$, $\chi(\widetilde{j}, \varphi^{-1}(\widetilde{j})) > 0$ and $\chi(\widetilde{j}, j) > 0$ for all $j \in V_0 \setminus \{\varphi^{-1}(\widetilde{j})\}$, where $\varphi^{-1} : \widetilde{V}_0 \to V_0$ denotes the inverse of φ, we know from above that then $k_\nu \in \mathrm{an}(\ell)$ and $\ell \in \mathrm{an}(k_{\nu+1})$, i.e., $\mathrm{de}(k_\nu) \cap \mathrm{an}(k_{\nu+1}) \neq \varnothing$. But this is in contradiction to the fact that p is a path in $\mathcal{D}^{\mathrm{tr}}$. Hence, $\widetilde{\mathcal{D}}^{\mathrm{tr}}$ must contain $\widetilde{p}$. □

In the next example we use Theorem 3.5.3 to find RMWMs that are χ-equivalent to a given one.

Example 3.5.4. [Continuation of Example 3.3.7: find χ-equivalent RMWMs]
By Theorem 3.2.3 the sets $\{1\}, \ldots, \{99\}$ are the maximum χ-cliques. Since 99 is the only terminal node in $\mathcal{D}$, it may be the only initial node of a DAG that underlies a potential RMWM with the same TDM χ as $\boldsymbol{X}$ and differs from $\mathcal{D}$. Thus the DAG



is the transitive reduction $\widetilde{\mathcal{D}}^{\mathrm{tr}}$ of such a DAG. To verify the existence of a RMWM with TDM χ on a DAG whose transitive reduction is $\widetilde{\mathcal{D}}^{\mathrm{tr}}$, we may compute the matrix $\overline{B}$ from (3.4.2) and check then whether it is the MLCM of a RMWM. □

We conclude this section with an example investigating whether a RMWM on a known DAG is χ-equivalent to a RMWM on another given DAG.

Example 3.5.5. [The existence of χ-equivalent RMWMs on given DAGs]
We consider a RMWM $\boldsymbol{X}$ with TDM χ on $\mathcal{D}_1$ and clarify when $\boldsymbol{X}$ is χ-equivalent to a RMWM on $\mathcal{D}_2$. Note that all RMLMs on $\mathcal{D}_1$ and on $\mathcal{D}_2$ are max-weighted. By Theorem 3.3.10 we find

$$\chi(1,2) = 0, \quad \chi(1,4) = 0, \quad \chi(1,3) > 0, \quad 1 - \chi(1,3) - \chi(2,3) > 0,$$
$$1 - \chi(2,4) > 0, \quad \chi(3,4) = \chi(2,3) \wedge \chi(2,4) > 0$$

and also that χ is the TDM of a RMWM on $\mathcal{D}_2$ if and only if

$$\chi(1,2) = 0, \quad \chi(1,4) = 0, \quad \chi(1,3) > 0, \quad 1 - \chi(1,3) - \chi(3,4) > 0,$$
$$1 - \chi(2,4) > 0, \quad \chi(2,3) = \chi(2,4) \wedge \chi(3,4) > 0.$$

This implies that $\boldsymbol{X}$ is χ-equivalent to a RMWM on $\mathcal{D}_2$ if and only if $\chi(2,3) = \chi(3,4)$.

As shown in Example 3.4.12 the matrix χ given therein is the TDM of a RMWM on $\mathcal{D}_1$. As $\chi(2,3) = 0.6 \neq \chi(3,4) = 0.5$, such a model cannot be χ-equivalent to a RMWM on $\mathcal{D}_2$. Of course, we already know this from Example 3.4.12.

□

3.6 Conclusion and outlook

A RMLM is not restricted to heavy-tailed noise variables, but is defined in (2.1.3) for independent noise variables with support $\mathbb{R}_+$. Only, if the noise variables are heavy-tailed, the TDM is meaningful (not identical to the zero matrix) for modeling the dependence structure in a RMLM.

In this heavy-tailed setting, we considered the problem of identifying a RMLM $\boldsymbol{X}$ on a DAG $\mathcal{D}$ from its TDM χ. Simply because of the symmetry of χ, the identifiability of $\boldsymbol{X}$ is not possible in general. RMLMs with arbitrary index of regular variation and MLCM whose column sums are also arbitrary have TDM χ. As our focus was on the causal structure of $\boldsymbol{X}$ represented by $\mathcal{D}$, we concentrated on the standardized model, where the index of regular variation is one and the columns of its MLCM $\overline{B}$ add up to one. We showed that $\overline{B}$ can be recovered from χ and some additional information on $\mathcal{D}$ such as the full reachability relation or only a causal ordering. In these situations we can also determine the minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$, the smallest

DAG which represents the recursive max-linear dependence structure of $\boldsymbol{X}$. We developed an algorithm which outputs the standardized MLCMs of all RMLMs having TDM χ. Moreover, we found the RMWMs as a relevant subclass of RMLMs. The simple structure of their TDMs allows for identifiability of $\overline{B}$ and $\mathcal{D}^B$ from χ and the initial nodes of $\mathcal{D}$. This led to a simpler approach to find the standardized MLCMs of all RMWMs with TDM χ.

Finally, we would like to say a few words about how the results of this chapter can be applied statistically. The first step would be the estimation of χ. Of course, this is usually based on observations, from which we could learn more than only the extreme dependence between every two components of $\boldsymbol{X}$. Extremal data are, however, sufficient for estimating χ. Estimators can be derived from estimators of the tail dependence function (Huang [36]). An estimator for the latter that is suitable for our situation is, for example, the empirical one introduced in [36] and studied further in Drees and Huang [17]. Many modifications of this estimator can be found (see e.g. the textbooks [4, 14]). A parametric estimator has been suggested in [20].

Since zero tail dependence is essential for the causal dependence between two components (see Theorem 3.2.3), we would also want to test zero tail dependence between every two components of $\boldsymbol{X}$. This is equivalent to testing asymptotic independence (cf. Remark 3.2.4(ii)). Corresponding tests, which can be consulted in this context, were introduced in Coles et al. [9], Draisma et al. [16], Ledford and Tawn [49], and Peng [56]. A similar problem occurs with Gaussian graphical model selection. It can be performed by testing conditional independence relations, which is equivalent to testing zero entries in the inverse of the covariance matrix (cf. [47], Proposition 5.2). We plan to investigate variants of methods developed in this context, for example, in Drton and Perlman [18], Friedman et al. [24], Kalisch and Bühlmann [39], Meinshausen and Bühlmann [52], and Rothman et al. [63]. A further goal will be to derive relations between (conditional) independence in a regularly varying graphical model and its TDM.

Appendix 3.A

3.A.1 Properties of the standardized ML coefficient matrix of a recursive ML model

We summarize some properties of the standardized MLCM $\overline{B}$ defined in (3.1.7), which are used throughout the chapter.

Lemma 3.A.1. *Let $\boldsymbol{X}$ be a RMLM on a DAG $\mathcal{D}$ with MLCM B and standardized MLCM $\overline{B}$.*

(a) We have $\operatorname{sgn}(\overline{B}) = \operatorname{sgn}(B)$.

(b) For $i \in V$, $\sum_{k\in\operatorname{An}(i)} \overline{b}_{ki} = \sum_{k=1}^{d} \overline{b}_{ki} = 1$.

(c) The matrix $\overline{B}$ is the MLCM of a RMLM on $\mathcal{D}$.

(d) For $i \in V$, $k \in \operatorname{an}(i)$, *and* $j \in \operatorname{an}(k)$, $\overline{b}_{ji} \geq \frac{\overline{b}_{jk}\overline{b}_{ki}}{\overline{b}_{kk}}$ *with equality if and only if there is a max-weighted path from j to i passing through k.*

(e) The minimum ML DAGs $\mathcal{D}^B$ and $\mathcal{D}^{\overline{B}}$ coincide.

(f) For distinct $i, j \in V$, $\overline{b}_{jj} > \overline{b}_{ji}$.

Proof. (a) and (b) are immediate consequences of the definition of $\overline{B}$ and (3.1.5).
(c) can be verified by Theorem 2.4.2.
(d) The inequality follows from (c) and Corollary 2.3.12 and the rest of the statement from Theorem 2.3.10(a) and by observing that $\overline{b}_{ji} = \frac{\overline{b}_{jk}\overline{b}_{ki}}{\overline{b}_{kk}}$ if and only if $b_{ji} = \frac{b_{jk}b_{ki}}{b_{kk}}$.
(e) is a consequence of Theorem 2.5.3 and the definition of $\overline{B}$.
(f) For $j \in V \setminus \mathrm{An}(i)$ we have immediately by (a) that $\overline{b}_{ji} = 0 < \overline{b}_{jj}$. For $j \in \mathrm{An}(i)$ we obtain by parts (b) and (d),

$$1 = \sum_{k\in\mathrm{An}(j)} \overline{b}_{ki} + \sum_{k\in\mathrm{An}(i)\setminus\mathrm{An}(j)} \overline{b}_{ki} \geq \frac{\overline{b}_{ji}}{\overline{b}_{jj}} \sum_{k\in\mathrm{An}(j)} \overline{b}_{kj} + \sum_{k\in\mathrm{An}(i)\setminus\mathrm{An}(j)} \overline{b}_{ki} = \frac{\overline{b}_{ji}}{\overline{b}_{jj}} + \sum_{k\in\mathrm{An}(i)\setminus\mathrm{An}(j)} \overline{b}_{ki}.$$

Since $\mathrm{An}(i) \setminus \mathrm{An}(j) \neq \varnothing$ and $\overline{b}_{ki} > 0$ for all $k \in \mathrm{An}(i) \setminus \mathrm{An}(j)$, we find $1 > \frac{\overline{b}_{ji}}{\overline{b}_{jj}}$, equivalently $\overline{b}_{jj} > \overline{b}_{ji}$. □

3.A.2 Derivation of the tail dependence matrix of a recursive ML model

We first prove (3.1.2) and specify G and its univariate and bivariate marginal distributions.

Proposition 3.A.2. *Let $\boldsymbol{X}$ be a RMLM on a DAG $\mathcal{D}$ with MLCM B. Then $\boldsymbol{X} \in \mathrm{MDA}(G)$ with*

$$G(\boldsymbol{x}) = \exp\Big\{ -\sum_{j=1}^{d} \bigvee_{i\in\mathrm{De}(j)} \Big(\frac{b_{ji}}{x_i}\Big)^{\alpha} \Big\}, \quad \boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}_+^d.$$

Let $\boldsymbol{M} = (M_1, \dots, M_d)$ be a random vector with distribution function G. Then for $i, j \in V$ the distribution functions of M_i and (M_i, M_j) are given by

$$G_i(x_i) = \exp\Big\{ -x_i^{-\alpha} \sum_{j\in\mathrm{An}(i)} b_{ji}^{\alpha} \Big\} \quad \textit{and} \quad G_{ij}(x_i, x_j) = \exp\Big\{ -\sum_{k\in\mathrm{An}(i)\cup\mathrm{An}(j)} \Big(\frac{b_{ki}}{x_i}\Big)^{\alpha} \vee \Big(\frac{b_{kj}}{x_j}\Big)^{\alpha} \Big\}.$$

Proof. As $Z \in \mathrm{RV}(\alpha)$, there exists a normalizing sequence $a_n \in \mathbb{R}_+$ such that for every $x \in \mathbb{R}_+$,

$$\lim_{n\to\infty} F_Z^n(a_n x) = \Phi_\alpha(x) \tag{3.A.1}$$

(e.g. [60], Proposition 1.11). Using (3.1.3), the independence of the noise variables, and (3.A.1), we obtain for $\boldsymbol{x} \in \mathbb{R}_+^d$,

$$\begin{aligned}
\big[\mathbb{P}\big(\boldsymbol{X} \leq a_n \boldsymbol{x}\big)\big]^n &= \Big[\mathbb{P}\Big(\bigvee_{j\in\mathrm{An}(i)} b_{ji} Z_j \leq a_n x_i,\ i \in V \Big)\Big]^n \\
&= \Big[\mathbb{P}\Big(Z_j \leq a_n \bigwedge_{i\in\mathrm{De}(j)} \frac{x_i}{b_{ji}},\ j \in V\Big)\Big]^n \\
&= \prod_{j=1}^{d} F_Z^n\Big(a_n \bigwedge_{i\in\mathrm{De}(j)} \frac{x_i}{b_{ji}}\Big) \\
&\xrightarrow[n\to\infty]{} \prod_{j=1}^{d} \Phi_\alpha\Big(\bigwedge_{i\in\mathrm{De}(j)} \frac{x_i}{b_{ji}}\Big) = G(\boldsymbol{x}).
\end{aligned}$$

This proves that $\boldsymbol{X} \in \mathrm{MDA}(G)$ (cf. Eq. (3.1.2)). Finally, the distribution functions of M_i and (M_i, M_j) are obtained by letting all other components of $\boldsymbol{x}$ in G tend to ∞ and recalling (3.1.5). □

Proof of (3.1.8). For every $k \in V$ we have $n(1-F_k(a_{k,n})) \to 1$ as $n \to \infty$ with $a_{k,n} := F_k^{\leftarrow}\big(1-\frac{1}{n}\big) = \big(\frac{1}{1-F_k}\big)^{\leftarrow}(n)$. Thus,

$$\begin{aligned}\chi(i,j) &= \lim_{n\to\infty} \frac{\mathbb{P}(X_i > a_{i,n}, X_j > a_{j,n})}{1-F_j(a_{j,n})} \\ &= \lim_{n\to\infty} n[1-F_i(a_{i,n}) + 1 - F_j(a_{j,n}) - 1 + \mathbb{P}(X_i \le a_{i,n}, X_j \le a_{j,n})] \\ &= 2 - \lim_{n\to\infty} n[1-\mathbb{P}(X_i \le a_{i,n}, X_j \le a_{j,n})].\end{aligned}$$

By Proposition 5.10(b), whose conditions are satisfied according to Proposition 3.A.2, and Eq. (5.38) of [60], we find

$$\chi(i,j) = 2 + \log G_{ij}((-1/\log G_i)^{\leftarrow}(1), (-1/\log G_j)^{\leftarrow}(1)),$$

where $(-1/\log G_i)^{\leftarrow}$ and $(-1/\log G_j)^{\leftarrow}$ denote the generalized inverses of the functions $-1/\log G_i$ and $-1/\log G_j$. With the representations for G_i, G_j, and G_{ij} from Proposition 3.A.2, we then obtain by a simple calculation

$$\chi(i,j) = 2 - \sum_{k\in \mathrm{An}(i)\cup\mathrm{An}(j)} \overline{b}_{ki} \vee \overline{b}_{kj}.$$

Finally, using Lemma 3.A.1(b), (a) yields

$$\begin{aligned}\chi(i,j) &= \sum_{k\in \mathrm{An}(i)\cup\mathrm{An}(j)} \overline{b}_{ki} + \sum_{k\in \mathrm{An}(i)\cup\mathrm{An}(j)} \overline{b}_{kj} - \sum_{k\in \mathrm{An}(i)\cup\mathrm{An}(j)} \overline{b}_{ki} \vee \overline{b}_{kj} \\ &= \sum_{k\in \mathrm{An}(i)\cup\mathrm{An}(j)} \overline{b}_{ki} \wedge \overline{b}_{kj} = \sum_{k\in \mathrm{An}(i)\cap\mathrm{An}(j)} \overline{b}_{ki} \wedge \overline{b}_{kj}.\end{aligned}$$

We learn from this proof that $\boldsymbol{X}$ and the limit vector $\boldsymbol{M}$ from (3.1.2) have the same TDM, since $\boldsymbol{M} \in \mathrm{MDA}(G)$. □

Chapter 4

Identifiability and estimation of recursive max-linear models

Abstract

We address the identifiablity and estimation of recursive max-linear structural equation models. Such models are generally unidentifiable: several DAGs and edge weights representing the max-linear structural equations may exist. We show that the whole class of DAGs and edge weights is identifiable. For estimation, standard likelihood theory and classical methods cannot be applied because assumptions usually made are not satisfied. We develop a simple learning method which, with probability 1, identifies the true class for a sufficiently large number of observations. Given the true underlying DAG, we present an estimator for the class of edge weights that can be considered a maximum likelihood estimator in a generalized setting. Given many observations, this estimator has also the nice property to estimate, almost surely, the true class of edge weights exactly.

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

Keywords and phrases: Causal inference, directed acyclic graph, generalized maximum likelihood estimation, graphical model, identifiability, max-linear model, nonparametric maximum likelihood estimation, structural equation model

4.1 Introduction

Establishing and understanding cause-effect relations is an omnipresent desire in science and daily life. It is especially important when dealing with extreme risks. Examples for such situations include incidents at airplane landings (Gissibl et al. [29]; cf. Figure 1.1.2), flooding in river networks (Asadi et al. [2]), financial risk (Einmahl et al. [20]), and chemical pollution of rivers (Hoef et al. [35]). Such applications to risk analysis, where extreme risks may propagate through a network, were the motivation behind the *recursive max-linear (ML) models* defined in Chapter 2. Recursive ML models are by definition *structural equation models (SEMs)* whose causal structure is represented by a *directed acyclic graph (DAG)* and, hence, by Theorem 1.4.1 of Pearl [55] *directed graphical models*. SEMs (see e.g. Bollen [5], [55]) and graphical models (see e.g. Koller and Friedman [45], Lauritzen [47], Spirtes et al. [69]) are well-established concepts to the understanding and quantification of causal inference from observational data.

Important research problems that are addressed for classes of recursive SEMs, as is the class of recursive ML models, are the *identifiability* of the coefficients and the associated DAG from the observational distribution as well as the estimation of the DAG (*structure learning*) from a finite sample. The book by Peters et al. [58] provides a profound introduction into this field of research and summarizes the current state of research.

We study these problems for recursive ML models. Throughout we assume that all variables are observed, that is, there are no hidden variables.

Recursive ML models are defined by a DAG, edge weights, and independent noise variables. Different DAGs and edge weights can define the same model (cf. Theorem 2.5.4). The so-called *max-linear (ML) coefficient matrix* determines this class of DAGs and edge weights uniquely. So the true DAG and edge weights are not identifiable; but, as we shall see, the ML coefficent matrix and, hence, the class of DAGs and edge weights defining the underlying model.

This identifiability result has direct implications for structure learning: if the data follow a recursive ML model, the associated class of DAGs and edge weights can be inferred from observational data only.

Several approaches for structure learning, which can be split mainly into score-based and constraint-based methods, have been proposed. Constraint-based methods, such as the PC algorithm (Spirtes and Glymour [67]), assume faithfulness to the underlying DAG (see e.g. Remark 2.3.9(ii) for the definition of faithfulness). However, recursive ML models are never faithful unless the underlying DAG has at most one path between two nodes (see Remark 2.3.9(i) and Theorem 4 of Klüppelberg and Lauritzen [43]). On the other hand, score-based methods (see Chickering [8], Geiger and Heckerman [25], Heckerman et al. [33], and references therein) require distributional properties that are not valid for recursive ML models or would at least restrict the model class. So we cannot use standard methods for structure learning without further ado.

Of course, we meet the same challenge in parameter learning where the DAG is assumed to be known. There exists no σ-finite measure on the space of observations that dominates the distributional family of recursive ML models on a given DAG. As a consequence, we cannot use standard maximum likelihood estimation methods. We suggest an estimator that can be considered a maximum likelihood estimator in an extended definition originally introduced by Kiefer and Wolfowitz [40] for covering the nonparametric case.

But for all that, estimation and structure learning of recursive ML models can be done in a simple and efficient fashion. Exploiting the distributional properties of the ratios between two components of a recursive ML model, we present appropriate procedures. For a sufficiently large number of observations, they identify, with probability 1, the true ML coefficient matrix and, hence, the true associated class of DAGs and edge weights; the convergence is geometrically fast.

This chapter is organized as follows. In Section 4.2 we present the model class of recursive ML models and introduce the notation used throughout this chapter. In Section 4.3 we discuss the identifiability of recursive ML models. Here we show distributional properties of the ratio between two components of a recursive ML model. Based on these properties, we propose an identification method. Section 4.4 is devoted to the estimation of recursive ML models in the situation where the DAG is known. We follow the Kiefer-Wolfowitz approach to determine *generalized maximum*

likelihood estimates (GMLEs). The main part is the derivation of a specific Radon-Nikodym derivative. Here we make comparisons with the case where we can define a standard likelihood function. To conclude this section, we point out a GMLE and its outstanding properties. In Section 4.5 we complement the theoretical findings with an efficient procedure to learn recursive ML models from observations only. Section 4.6 concludes and suggests further directions of research. In Appendix 4.A we give an alternative identification algorithm. In addition, we prove further distributional properties of the ratio between two components, which are not needed in the main part of the chapter, but are useful for a deeper understanding.

4.2 Preliminaries – Recursive ML models

We consider recursive ML models, which have been introduced in Chapter 2. In this section we introduce some notations and summarize the most important properties needed throughout this chapter.

A recursive ML model $\boldsymbol{X} = (X_1, \dots, X_d)$ is specified by an underlying (causal) structure in terms of a DAG $\mathcal{D} = (V, E)$ with nodes $V = \{1, \dots, d\}$, positive *edge weights* c_{ki} for $i \in V$ and $k \in \mathrm{pa}(i)$, and independent random variables $Z_1, \dots, Z_d$ with support $\mathbb{R}_+ := (0, \infty)$ and atomfree distributions:

$$X_i = \bigvee_{k \in \mathrm{pa}(i)} c_{ki} X_k \vee Z_i, \quad i = 1, \dots, d, \tag{4.2.1}$$

where $\mathrm{pa}(i)$ are the parents of node i in $\mathcal{D}$. To highlight the DAG $\mathcal{D}$, we say that $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$. In the original definition of a recursive ML model in (2.1.3), the weights of the *noise variables* Z_i in (4.2.1) do not necessarily have to be equal to one but can be any positive real number. Such a recursive ML model has then representation (4.2.1) with appropriately scaled noise variables. In addition, the distributional properties of the noise variables are in (2.1.3) slightly different. In the context of risk analysis, natural candidates for the noise distributions 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 e.g. Beirlant et al. [4], de Haan and Ferreira [14], Resnick [60, 61]).

Occasionally, we write $k \to i$ instead of $k \in \mathrm{pa}(i)$. 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}, we call the nonnegative matrix B with entries

$$b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \text{ for all } j \in \mathrm{an}(i), \quad b_{ii} = 1, \quad \text{and} \quad b_{ji} = 0 \text{ for all } j \in V \setminus \mathrm{An}(i), \tag{4.2.2}$$

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 call a path from j to i whose weight equals the maximum weight b_{ji} *max-weighted*.

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

$$X_i = \bigvee_{j=1}^{d} b_{ji} Z_j = \bigvee_{j \in \mathrm{An}(i)} b_{ji} Z_j, \quad i = 1, \dots, d, \tag{4.2.3}$$

where $\mathrm{An}(i) = \mathrm{an}(i) \cup \{i\}$ and $\mathrm{an}(i)$ are the ancestors of i in $\mathcal{D}$ (cf. Theorem 2.2.2).

We have presented those properties of $\boldsymbol{X}$ we need throughout the whole chapter. Further properties of $\boldsymbol{X}$ from Chapter 2 are introduced where they are needed.

Throughout this chapter we use the following notation. The sets $\mathrm{an}(i)$, $\mathrm{pa}(i)$, $\mathrm{de}(i)$, and $\mathrm{nd}(i)$ contain the ancestors, parents, descendants, and non-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 $\boldsymbol{x} \in \mathbb{R}_+^d$, $\boldsymbol{x}_U = (x_\ell, \ell \in U)$. Generally, we consider statements for $i \in \varnothing$ as invalid. Furthermore, we set $\bigvee_{i \in \varnothing} a_i = 0$, $\bigwedge_{i \in \varnothing} a_i = \infty$, $\prod_{i \in \varnothing} a_i = 1$, and $\frac{a_i}{0} = \infty$ for (possibly random) $a_i \in \mathbb{R}_+$ as well as $\bigcup_{i \in \varnothing} A_i = \varnothing$ and $\bigcap_{i \in \varnothing} A_i = \mathbb{R}_+^d$ for $A_i \subseteq \mathbb{R}_+^d$.

4.3 Identifiability of a recursive ML model

We discuss the identifiability of a recursive ML model $\boldsymbol{X}$ from its distribution $\mathcal{L}(\boldsymbol{X})$. We start with an example.

Example 4.3.1. [The DAG and the edge weights are not necessarily identifiable]
Consider a recursive ML model $\boldsymbol{X} = (X_1, X_2, X_3)$ on the DAG $\mathcal{D}$ depicted below with edge weights c_{12}, c_{23}, c_{13} such that $c_{13} \le c_{12}c_{23}$. By (4.2.1) the components of $\boldsymbol{X}$ have the following representations,

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

From these and the order between the edge weights, we observe that

$$X_3 = c_{13}^* X_1 \vee c_{23}X_2 \vee Z_3 \quad \text{for every } c_{13}^* \in [0, c_{12}c_{23}].$$

This implies that $\boldsymbol{X}$ is a recursive ML model on $\mathcal{D}$ with edge weights $c_{12}, c_{23}, c_{13}^* \in (0, c_{12}c_{23}]$ as well as on the DAG $\mathcal{D}^B$ depicted below with edge weights c_{12}, c_{23}. Consequently, we cannot identify $\mathcal{D}$ as well as c_{13} from the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$. However, note that the ML coefficient matrix B is unique.

If $c_{13} > c_{12}c_{23}$, then $\mathcal{D}$ and the edge weights c_{12}, c_{23}, c_{13} are the only DAG and edge weights, respectively, that represent $\boldsymbol{X}$ in the sense of (4.2.1). Thus they are identifiable from $\mathcal{L}(\boldsymbol{X})$.

□

As conclusion of Example 4.3.1, it is generally not possible to identify the true DAG $\mathcal{D}$ and the edge weights c_{ki} underlying $\boldsymbol{X}$ in representation (4.2.1) from $\mathcal{L}(\boldsymbol{X})$. The reason for this is that 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* $\boldsymbol{X}$. It is the smallest DAG representing the causal structure of $\boldsymbol{X}$. The further DAGs are the DAGs that have at least the edges of $\mathcal{D}^B$ and a path from j to i if and only if $\mathcal{D}^B$ has a path from j to i. The edge weights c_{ki} in representation (4.2.1) of $\boldsymbol{X}$ are only uniquely given for edges contained in $\mathcal{D}^B$. In that case, $c_{ki} = b_{ki}$; otherwise, we may have $c_{ki} \in (0, b_{ki}]$. All these DAGs and edge weights lead via (4.2.2) to the same ML coefficient matrix B and can be determined from B. All this can be found in Section 2.5 with its main results in Theorems 2.5.3, 2.5.4.

Based on the above observations, we investigate the identifiability of this class of DAGs and edge weights from $\mathcal{L}(\boldsymbol{X})$. Since it 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 used in this section suggests a simple procedure, which is presented in Algorithm 4.5.1 below, to estimate B from independent realizations of $\boldsymbol{X}$. We demonstrate an alternative way in Appendix 4.A.1.

We know from (4.2.2) that $b_{ii} = 1$ and $b_{ji} \neq 0$ if and only if $j \in \mathrm{An}(i)$. Hence, to show the identifiability of B from $\mathrm{supp}(\boldsymbol{X})$, it suffices to find a quantity that can be determined from $\mathcal{L}(\boldsymbol{X})$ and that specifies for distinct $i, j \in V$ whether $j \in \mathrm{an}(i)$ and if so, defines b_{ji}. It turns out that the support of $\frac{X_i}{X_j}$, denoted by $\mathrm{supp}\big(\frac{X_i}{X_j}\big)$, is such a quantity. Because of the max-linear representation (4.2.3) of the components of $\boldsymbol{X}$, it is clear that it depends on the distributional properties of the noise variables. We first discuss some consequences of these properties. Recall that the noise variables are assumed to be independent with support $\mathbb{R}_+$ and atomfree distributions. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space of $(Z_1, \dots, Z_d)$ and, hence, of $\boldsymbol{X}$. We write events such as $\{\omega \in \Omega : X_i(\omega) < X_j(\omega)\}$ or $\{\omega \in \Omega : Z_i(\omega) < Z_j(\omega)\}$ more conveniently as $\{X_i < X_j\}$ or $\{Z_i < Z_j\}$, respectively.

The independence of the noise variables and their atomfree distributions imply that

$$\text{the event } \big\{Z_i = xZ_j\big\} \text{ for distinct } i, j \in V \text{ and } x \in \mathbb{R}_+ \text{ has probability zero.} \tag{4.3.1}$$

This plays an important role in determing the atoms of $\frac{X_i}{X_j}$. It shows, together with (4.2.3), that the sets $\{X_i = xX_j\} = \big\{\bigvee_{\ell \in \mathrm{An}(i)} b_{\ell i} Z_\ell = \bigvee_{\ell \in \mathrm{An}(j)} x b_{\ell j} Z_\ell\big\}$ and

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

differ only by a set of probability zero. Therefore, $\frac{X_i}{X_j}$ has an atom in x if and only if $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \varnothing$ and $x = \frac{b_{\ell i}}{b_{\ell j}}$ for some $\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)$ as the noise variables are independent and have support $\mathbb{R}_+$.

Table 4.1: Distributional properties of $\frac{X_i}{X_j}$.

Relationship between i and j	$\operatorname{supp}\big(\frac{X_i}{X_j}\big)$	Atoms
$j \in \operatorname{an}(i)$	$[b_{ji}, \infty)$	$\frac{b_{\ell i}}{b_{\ell j}}$ for $\ell \in \operatorname{An}(j)$
$i \in \operatorname{an}(j)$	$\big(0, \frac{1}{b_{ij}}\big]$	$\frac{b_{\ell i}}{b_{\ell j}}$ for $\ell \in \operatorname{An}(i)$
$j \in \operatorname{nd}(i)$ and $i \in \operatorname{nd}(j)$:		
$\operatorname{an}(i) \cap \operatorname{an}(j) \neq \varnothing$	$\mathbb{R}_+$	$\frac{b_{\ell i}}{b_{\ell j}}$ for $\ell \in \operatorname{an}(i) \cap \operatorname{an}(j)$
$\operatorname{an}(i) \cap \operatorname{an}(j) = \varnothing$	$\mathbb{R}_+$	-

The support of the noise variables of $\mathbb{R}_+$ and representation (4.2.3) are the reason why

$$\operatorname{supp}\Big(\frac{X_i}{X_j}\Big) = \Big\{ \frac{\bigvee_{\ell \in \operatorname{An}(i)} b_{\ell i} z_\ell}{\bigvee_{\ell \in \operatorname{An}(j)} b_{\ell j} z_\ell} : \boldsymbol{z}_{\operatorname{An}(i) \cup \operatorname{An}(j)} \in \mathbb{R}_+^{|\operatorname{An}(i) \cup \operatorname{An}(j)|} \Big\}.$$

Since the function

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

is continuous, $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ is an interval in $\mathbb{R}_+$. Assume that $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ has a positive lower bound; i.e., there exists some $a \in \mathbb{R}_+$ such that

$$\bigvee_{\ell \in \operatorname{An}(i) \cap \operatorname{An}(j)} a b_{\ell j} z_\ell \vee \bigvee_{\ell \in \operatorname{An}(j) \setminus \operatorname{An}(i)} a b_{\ell j} z_\ell \leq \bigvee_{\ell \in \operatorname{An}(i) \cap \operatorname{An}(j)} b_{\ell i} z_\ell \vee \bigvee_{\ell \in \operatorname{An}(i) \setminus \operatorname{An}(j)} b_{\ell i} z_\ell \tag{4.3.2}$$

for all $\boldsymbol{z}_{\operatorname{An}(i) \cup \operatorname{An}(j)} \in \mathbb{R}_+^{|\operatorname{An}(i) \cup \operatorname{An}(j)|}$. If $\operatorname{An}(j) \setminus \operatorname{An}(i) \neq \varnothing$, we can choose z_ℓ for some $\ell \in \operatorname{An}(j) \setminus \operatorname{An}(i)$ so large that $a b_{\ell j} z_\ell$ is greater than the maximum on the right-hand side of (4.3.2). This contradicts (4.3.2), and we necessarily have that $\operatorname{An}(j) \setminus \operatorname{An}(i) = \varnothing$, equivalently, $j \in \operatorname{An}(i)$. We then find that (4.3.2) holds if and only if $a \leq \bigwedge_{\ell \in \operatorname{An}(i)} \frac{b_{\ell i}}{b_{\ell j}}$; otherwise, there are $\boldsymbol{z}_{\operatorname{An}(i) \cup \operatorname{An}(j)}$ such that the maximum on the left-hand side of (4.3.2) is greater than the one on the right-hand side. Hence, $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ has a positive lower bound if and only if $j \in \operatorname{An}(i)$. It remains to clarify whether $\bigwedge_{\ell \in \operatorname{An}(i)} \frac{b_{\ell i}}{b_{\ell j}}$ is then contained in the interval $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$. This is indeed the case, since by Corollary 2.3.12 $\bigwedge_{\ell \in \operatorname{An}(i)} \frac{b_{\ell i}}{b_{\ell j}} = b_{ji}$ and b_{ji} is an atom of $\frac{X_i}{X_j}$. Conversely, we obtain that $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ is bounded from above if and only if $i \in \operatorname{An}(j)$. In that case, the upper bound is $\frac{1}{b_{ij}}$, which is an atom of $\frac{X_i}{X_j}$ and contained in $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$. In Table 4.1 we give $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ depending on the relationship between i and j in $\mathcal{D}$; the atoms of $\frac{X_i}{X_j}$ are shown as well.

Table 4.1 and (4.2.2) suggest the following algorithm to compute B from $\mathcal{L}(\boldsymbol{X})$. This proves the identifiability of B from $\mathcal{L}(\boldsymbol{X})$. Instead of the whole distribution $\mathcal{L}(\boldsymbol{X})$, it suffices to know $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$ for all $i, j \in V$ with $i < j$.

Algorithm 4.3.2. [Find B from $\mathcal{L}(\boldsymbol{X})$]

1. For all $i \in V = \{1, \ldots, d\}$, set $b_{ii} = 1$.
2. For all $i, j \in V$ with $i < j$, find $\operatorname{supp}\big(\frac{X_i}{X_j}\big)$:

if $\operatorname{supp}\left(\frac{X_i}{X_j}\right) = [a, \infty)$ for some $a \in \mathbb{R}_+$, then set $b_{ji} = a$ and $b_{ij} = 0$;

else, if $\operatorname{supp}\left(\frac{X_i}{X_j}\right) = (0, a]$ for some $a \in \mathbb{R}_+$, then set $b_{ij} = \frac{1}{a}$ and $b_{ji} = 0$;

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

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 weights that represent $\boldsymbol{X}$ in the sense of (4.2.1) can be determined from B, the only quantities we do not know about yet but are needed to define $\boldsymbol{X}$ are the noise variables. In fact, the distribution of the noise vector $(Z_1, \ldots, Z_d)$ is identifiable from $\mathcal{L}(\boldsymbol{X})$. Because of the identifiability of B from $\mathcal{L}(\boldsymbol{X})$, we can prove this by providing an algorithm that determines the distributions of the noise variables from $\mathcal{L}(\boldsymbol{X})$ and B. Its correctness follows from the independence of the noise variables. We denote by F_{Z_i} the distribution function of the noise variable Z_i. It is enough to know the univariate marginal distribution functions of $\mathcal{L}(\boldsymbol{X})$ instead of the whole distribution $\mathcal{L}(\boldsymbol{X})$.

Algorithm 4.3.3. [Find $F_{Z_1}(x), \ldots, F_{Z_d}(x)$ for $x \in \mathbb{R}_+$ from B and $\mathcal{L}(\boldsymbol{X})$]
For $\nu = 0, \ldots, d-1$,

for $i \in V$ such that $|\operatorname{an}(i)| = \nu$, set

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

Finally, we summarize the main result of this section again.

Theorem 4.3.4. *Let $\mathcal{L}(\boldsymbol{X})$ be the distribution of a recursive ML model $\boldsymbol{X}$. Then its ML coefficient matrix B and the distribution of its noise vector $(Z_1, \ldots, Z_d)$ are identifiable from $\mathcal{L}(\boldsymbol{X})$. Furthermore, all edge weights and DAGs that could have generated $\boldsymbol{X}$ by* (4.2.1) *can be obtained.*

Figure 4.3.1 gives an overview of how all these quantities can be determined from $\mathcal{L}(\boldsymbol{X})$. To conclude, recursive ML models with different ML coefficient matrices or different distributions of the noise vectors can never have the same distribution. Conversely, all recursive ML models with the same distribution must have the same ML coefficient matrix and identically distributed noise vectors. However, such recursive ML models can have different underlying DAGs and edge weights with the result that the true underlying DAG and the corresponding edge weights cannot be identified in general.

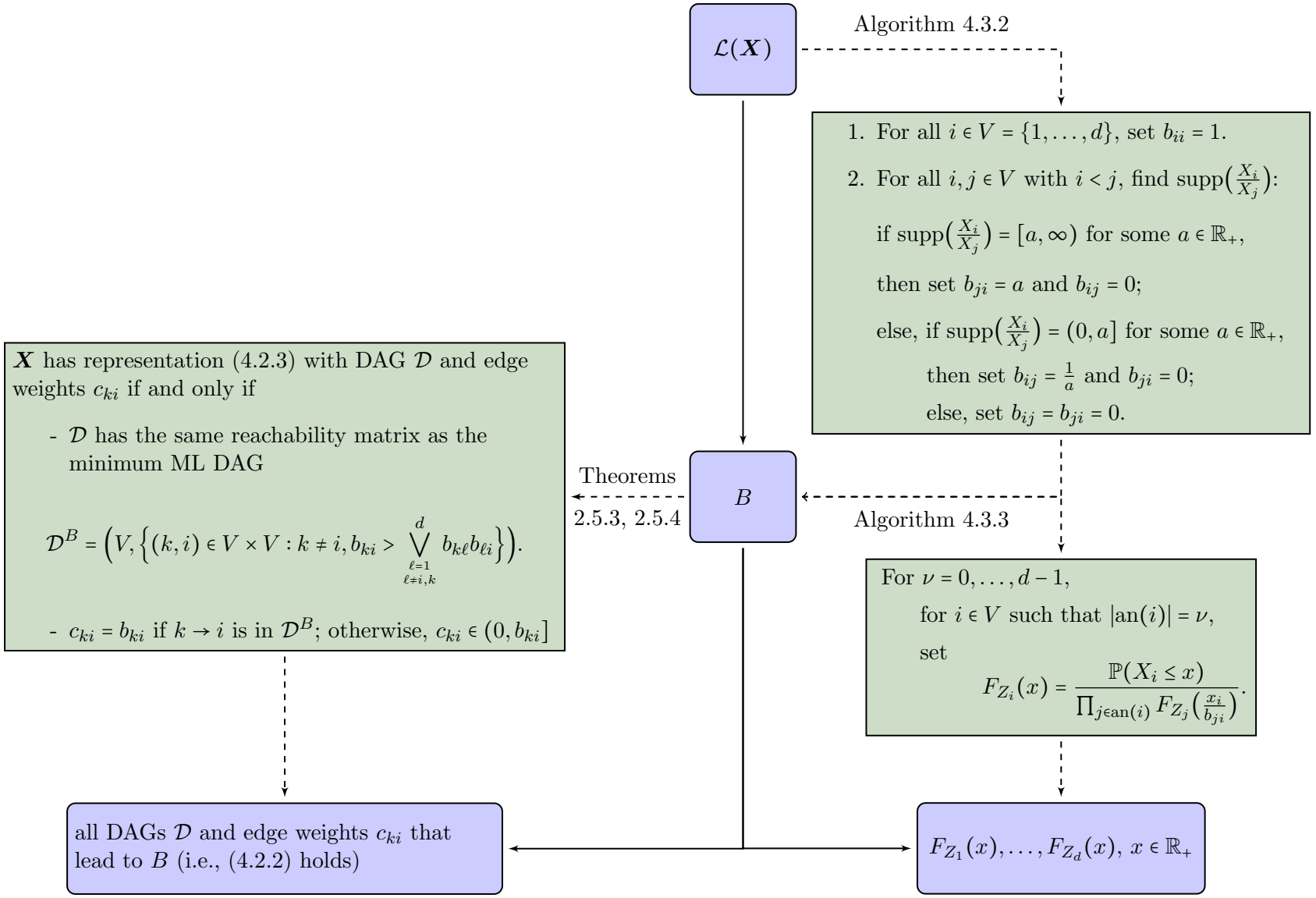


Figure 4.3.1: How the ML coefficient matrix B, the distributions of the noise variables $F_{Z_1}, \dots, F_{Z_d}$, as well as all potential DAGs $\mathcal{D}$ and edge weights c_{ki} of a recursive ML model $\boldsymbol{X}$ can be identified from its distribution $\mathcal{L}(\boldsymbol{X})$ (cf. Theorem 4.3.4). In the paragraph above Remark 2.2.3, the definition of the reachability matrix of a DAG is given.

4.4 Estimation of a recursive ML model with known DAG

In this section we assume that independent realizations $\boldsymbol{x}^{(t)} = \left(x_1^{(t)}, \ldots, x_d^{(t)}\right)$, $t = 1, \ldots, n$, of a recursive ML model $\boldsymbol{X} = (X_1, \ldots, X_d)$ and its DAG $\mathcal{D}$ are given. Our goal is the estimation of its edge weights c_{ki}, its ML coefficient matrix B, and the distribution of its noise vector $(Z_1, \ldots, Z_d)$. In the first part we suggest GMLEs; in the second we discuss the preferred GMLE of B in detail.

4.4.1 Generalized maximum likelihood estimation

The recursive ML model $\boldsymbol{X}$ may satisfy (4.2.1) with respect to $\mathcal{D}$ for various edge weights c_{ki} (see Example 4.3.1 and Theorem 2.5.4(b)). As a consequence, we usually have no chance to estimate the true edge weights of $\boldsymbol{X}$ from $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ exactly, although $\mathcal{D}$ is known. But this is theoretically possible for its ML coefficient matrix B, since B is identifiable from the distribution $\mathcal{L}(\boldsymbol{X})$ of $\boldsymbol{X}$ (see Algorithm 4.3.2). That is why we start with the estimation of B. To obtain then estimates of these various edge weights, we use the fact that this class of edge weights can be determined from B (see Figure 4.3.1). We present the corresponding result in Corollary 4.4.30 below.

ML coefficient matrix B

Before we start estimating B, we introduce some notation. We denote by $\mathcal{B}$ the class of the ML coefficient matrices of all recursive ML models on $\mathcal{D}$. This means that our estimate of B should be an element of $\mathcal{B}$. For a characterization of the class $\mathcal{B}$, see Theorem 2.4.2 or Corollary 2.4.3(a). In Remark 4.4.5 below we present further necessary and sufficient conditions for a matrix to be contained in $\mathcal{B}$. For $B \in \mathcal{B}$ we denote by P_B the probability measure induced by a recursive ML model on $\mathcal{D}$ with ML coefficient matrix B. To estimate B, we consider the family of these probability measures, denoted by $\mathcal{P}(\mathcal{D})$ in the following. Theorem 4.3.4 allows us to assume that the distributions of the underlying noise vectors are identical. So throughout this section the noise vectors of all recursive ML models on $\mathcal{D}$ are assumed to have the same distribution. Further, we assume that the only information we have about this distribution is that it has support $\mathbb{R}_+$ and independent, atomfree margins.

We cannot use standard maximum likelihood methods to estimate B, since $\mathcal{P}(\mathcal{D})$ is not dominated; i.e., there exists no σ-finite measure μ such that every probability measure in $\mathcal{P}(\mathcal{D})$ is absolutely continuous with respect to μ. We illustrate this by a simple example.

Example 4.4.1. [$\mathcal{P}(\mathcal{D})$ is not dominated]
Consider the DAG $\mathcal{D}$ with nodes $\{1, 2\}$ and an edge from 1 to 2. We assume that $\mathcal{P}(\mathcal{D})$ is dominated: there exists a $\sigma-$finite measure μ such that for every $B \in \mathcal{B}$,

$$P_B(A) = 0 \text{ whenever } \mu(A) = 0 \text{ for } A \in \mathcal{B}(\mathbb{R}_+^2), \tag{4.4.1}$$

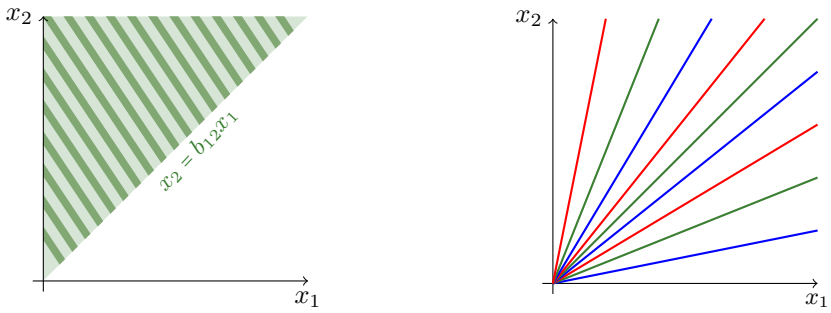


Figure 4.4.1: On the left-hand side $\text{supp}(P_B)$ from Example 4.4.1 is shown, on the right-hand side the set $A(b_{12})$ for different values of $b_{12} \in \mathbb{R}_+$.

where $\mathcal{B}(\mathbb{R}_+^2)$ denotes the Borel $\sigma-$algebra on $\mathbb{R}_+^2$. From Table 4.1 we know that

$$P_B\big(\{(b_{12}x, x) \in \mathbb{R}_+^2 : x \in \mathbb{R}_+\}\big) =: P_B(A(b_{12})) > 0.$$

Hence, by (4.4.1), $\mu(A(b_{12})) > 0$ for every $b_{12} \in \mathbb{R}_+^2$. Thus there are uncountably many disjoint sets with positive μ-measure. This contradicts the σ-finiteness of μ, and $\mathcal{P}(\mathcal{D})$ cannot be dominated. In Figure 4.4.1 the support of P_B, $\text{supp}(P_B)$, and the set $A(b_{12})$ for different values $b_{12} \in \mathbb{R}_+^2$ are depicted; these sets play an important role in the further course of this section.

By the same argumentation as in this example, $\mathcal{P}(\mathcal{D})$ is for all $\mathcal{D}$ not dominated. □

We cannot use densities as likelihoods as with the classical maximum likelihood estimation. However, there exist definitions of generalized MLEs that cover the undominated case as well; Kalbfleisch and Prentice [38], [40], and Scholz [65] suggested such extensions. Their goal was to investigate how a nonparametric MLE should be defined, a problem where typically no common σ-finite dominating measure exists; think, for example, of the problem of estimating an arbitrary unknown distribution. We follow the Kiefer-Wolfowitz definition of a GMLE; this was also done, for example, by Gill [26] and Johansen [37].

Let $\mathcal{P}$ be a family of probability measures on $(\mathbb{R}_+^d, \mathcal{B}(\mathbb{R}_+^d))$, where $\mathcal{B}(\mathbb{R}_+^d)$ denotes the Borel σ-algebra on $\mathbb{R}_+^d$, and $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ 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 $\frac{dP}{d(P+Q)}$ denotes a density of P with respect to $P+Q$. Then we call $\widehat{P}_0$ a *generalized maximum likelihood estimate (GMLE)* of P_0 if

$$\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, \widehat{P}_0, \widehat{P}_0) \neq 0 \quad \text{and} \quad \prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, Q, \widehat{P}_0) \leq \prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)}, \widehat{P}_0, Q) \text{ for all } Q \in \mathcal{P}. \tag{4.4.2}$$

Since P is absolutely continuous with respect to $P+Q$, the density $\frac{dP}{d(P+Q)}$ always exists according to the Radon-Nikodym theorem. This means that the GMLE is well-defined. The idea of the Kiefer-Wolfowitz definition is very logical and their definition extends the definition of a MLE in a very natural way. The approach is to consider pairwise comparisons of possible distributions P and Q for the observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ only: strict inequality in the second condition of (4.4.2) means that $\widehat{P}_0$ is a more likely explanation of the sample $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ than Q. When using the Kiefer-Wolfowitz approach, only the second condition in (4.4.2) is usually required. But the first condition is implicitly in the Kiefer-Wolfowitz definition and requiring it leads in

our case to other GMLEs, as we show in Example 4.4.2 below. This condition excludes – with a suitable choice of ρ – GMLEs $\widetilde{B}$ where the observations $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ could not have been generated by $P_{\widetilde{B}}$. We go into it and clarify it in Examples 4.4.2, 4.4.3.

If $\mathcal{P}$ is dominated by a σ-finite measure, then the Kiefer-Wolfowitz definition is equivalent to the usual definition of a MLE; if P_0 is completely unknown, then the empirical distribution function is a GMLE.

In order to be able to compute the GMLEs of B, we need to find for any two $B, B^* \in \mathcal{B}$, a density of P_B with respect to $P_B + P_{B^*}$. For this 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 three properties,

(A) $P_B(A_0(B,B^*)) = 0$,

(B) $P_B(A \cap A_{1/2}(B,B^*)) = P_{B^*}(A \cap A_{1/2}(B,B^*))$ for every $A \in \mathcal{B}(\mathbb{R}^d_+)$, and

(C) $P_{B^*}(A_1(B,B^*)) = 0$.

Then the measurable function from $\mathbb{R}^d_+$ to $\{0, 1/2, 1\}$ such that

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{x}, B, B^*) = \frac{1}{2} \cdot \mathbb{1}_{A_{1/2}(B,B^*)}(\boldsymbol{x}) + \mathbb{1}_{A_1(B,B^*)}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in A_0(B,B^*), \\ \frac{1}{2}, & \text{if } \boldsymbol{x} \in A_{1/2}(B,B^*), \\ 1, & \text{if } \boldsymbol{x} \in A_1(B,B^*), \end{cases} \tag{4.4.3}$$

is a density as desired. That is because, using the properties (A), (B), (C), we obtain for every $A \in \mathcal{B}(\mathbb{R}^d_+)$,

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

We begin with two examples that shall help to get an idea and provide insights into the concepts and arguments we use in the general case. They are deliberately very detailed.

Example 4.4.2. [Continuation of Example 4.4.1: how to find a density as in (4.4.3) and the GMLEs]
Recall the sets depicted in Figure 4.4.1. For $B, B^* \in \mathcal{B}$ we show that the partition

$$\begin{aligned} \big\{A_0(B,B^*) &:= \mathbb{R}^2_+ \setminus \operatorname{supp}(P_B) \cup \big[\big(\operatorname{supp}(P_B) \setminus A(b_{12})\big) \cap A(b^*_{12})\big] \\ &= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 < b_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b^*_{12}x_1 > b_{12}x_1\big\}, \\ A^{B+B^*}_{1/2} &:= \big[A(b_{12}) \cap A(b^*_{12})\big] \cup \big[\big(\operatorname{supp}(P_B) \setminus A(b_{12})\big) \cap \big(\operatorname{supp}(P_{B^*}) \setminus A(b^*_{12})\big)\big] \\ &= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b_{12}x_1 = b^*_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 > (b_{12} \vee b^*_{12})x_1\big\}, \\ A_1(B,B^*) &:= \big[\operatorname{supp}(P_B) \cap \big(\mathbb{R}^2_+ \setminus \operatorname{supp}(P_{B^*})\big)\big] \cup \big[A(b_{12}) \cap \big(\operatorname{supp}(P_{B^*}) \setminus A(b^*_{12})\big)\big] \\ &= \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : b^*_{12}x_1 > x_2 \geq b_{12}x_1\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b_{12}x_1 > b^*_{12}x_1\big\}\big\} \end{aligned}$$

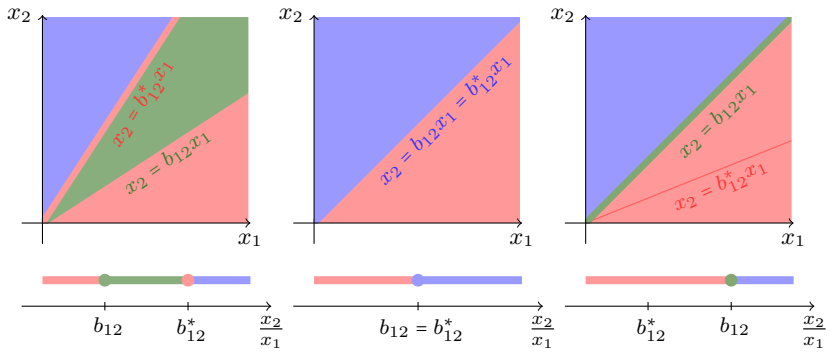


Figure 4.4.2: The density $\rho(\cdot, B, B^*)$ given in (4.4.4) as a contour plot (top line) and as a function of $\frac{x_2}{x_1}$ (bottom line) for the three situations $b_{12} < b_{12}^*$ (left-hand side), $b_{12} = b_{12}^*$ (middle), and $b_{12} > b_{12}^*$ (right-hand side). The area where it is $0/\frac{1}{2}/1$ is coloured in red/blue/green.

of $\mathbb{R}_+^2$ satisfies properties (A), (B), (C), leading to the function

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{x}, B, B^*) = \frac{1}{2} \cdot \mathbb{1}_{\{b_{12}x_1\} \cap \{b_{12}^* x_1\}}(x_2) + \frac{1}{2} \cdot \mathbb{1}_{(b_{12} \vee b_{12}^*, \infty)}(x_2) + \mathbb{1}_{[b_{12}, b_{12}^*)}(x_2) + \mathbb{1}_{\{b_{12}\} \cap (b_{12}^*, \infty)}(x_2) \tag{4.4.4}$$

from $\mathbb{R}_+^2$ to $\{0, 1/2, 1\}$ as a density of P_B with respect to $P_B + P_{B^*}$. Note that, obviously, $\{\boldsymbol{x} \in \mathbb{R}_+^2 : x_2 = b_{ki}^* x_1 > b_{ki} x_1\} = \varnothing$ if $b_{12} \geq b_{12}^*$, and corresponding for the other sets. Figure 4.4.2 shows the density $\rho(\cdot, B, B^*)$ for the three possible orders between b_{12} and b_{12}^*.

Let $\boldsymbol{X}$ be a recursive ML model on $\mathcal{D}$ with ML coefficient matrix B and Z_1, Z_2 its noise variables. Since by Table 4.1 b_{12} is the only atom of $\frac{X_2}{X_1}$ and $\operatorname{supp}\left(\frac{X_2}{X_1}\right) = [b_{12}, \infty)$, property (A) is true. By reversing the roles of B and B^*, (C) follows from (A). Recalling that we assume identically distributed noise vectors, (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 $A_{1/2}(B^*, B) = A_{1/2}(B, B^*)$ and that the noise vectors are identically distributed, we finally obtain for $A \in \mathcal{B}(\mathbb{R}_+^2)$,

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

So far we know that the function in (4.4.4) is a density of P_B with respect to $P_B + P_{B^*}$. We use this density to determine the GMLEs of B. The only ML coefficient we have to estimate is b_{12}. Define $\widehat{b}_{12}$ as the minimal observed ratio of $\frac{X_2}{X_1}$, i.e., $\widehat{b}_{12} = \bigwedge_{t=1}^n \frac{x_2^{(t)}}{x_1^{(t)}}$, and let $\widehat{B}$ denote the corresponding ML coefficient matrix. Then, by the definition of ρ and its defining sets, some $P_{\widetilde{B}} \in \mathcal{P}(\mathcal{D})$ satsifies the first condition in (4.4.2) if and only if

$$\widetilde{b}_{12} x_1^{(t)} \leq x_2^{(t)} \text{ for all } t, \text{ equivalently } \widetilde{b}_{12} \leq \widehat{b}_{12}. \tag{4.4.5}$$

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

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

Hence, some $P_{\widetilde{B}} \in \mathcal{P}(\mathcal{D})$ satsifies the second condition in (4.4.2) if and only if

$$\text{for all } B \in \mathcal{B}, \text{ if some } \boldsymbol{x}^{(t)} \in A_0(\widetilde{B}, B), \text{ then some } \boldsymbol{x}^{(s)} \in A_0(B, \widetilde{B}). \tag{4.4.6}$$

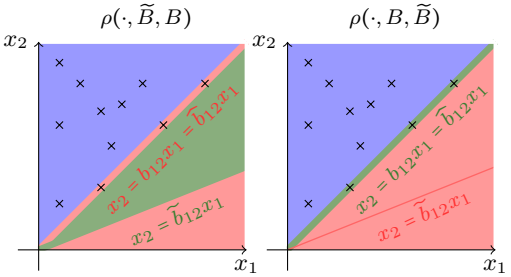
In summary, some $\widetilde{B}$ is a GMLE of B if and only if (4.4.5) and (4.4.6) are satisfied. We discuss the possible GMLEs of b_{12} in detail.

(1) $\widetilde{b}_{12} \in (0, \widehat{b}_{12})$ is no GMLE:
Set $b_{12} = \widehat{b}_{12}$, and let $\boldsymbol{x}^{(t)}$ 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 (4.4.6); consequently, $\widetilde{b}_{12}$ cannot be a GMLE of b_{12}. In Figure 4.4.3(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 $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$. 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.

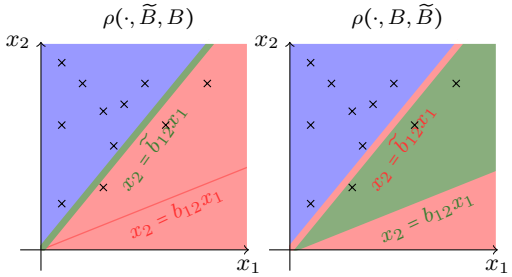
(2) $\widetilde{b}_{12} > \widehat{b}_{12}$ is no GMLE if $\widetilde{b}_{12} x_1^{(t)} \neq x_2^{(t)}$ for all t:
Figure 4.4.3(b) shows a situation that contradicts (4.4.6), similarly to Figure 4.4.3(a) in (1). So $\widetilde{b}_{12}$ is no GMLE. That could not be the case because of the necessary condition in (4.4.5), but more on this in (4).

(3) $\widehat{b}_{12}$ is a GMLE:
Condition (4.4.5) holds obviously. To prove (4.4.6), assume for some $B \in \mathcal{B}$ 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 4.4.3(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 $\big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 = b_{12}x_1\big\}$, but then there are also realizations in the 0-area of the lower plot (those that lie below this line). Hence, (4.4.6) holds. Since there is no realization in the 0-area of the middle and right plot in the top line, (4.4.6) is automatically satisfied if $b_{12} \leq \widehat{b}_{12}$.

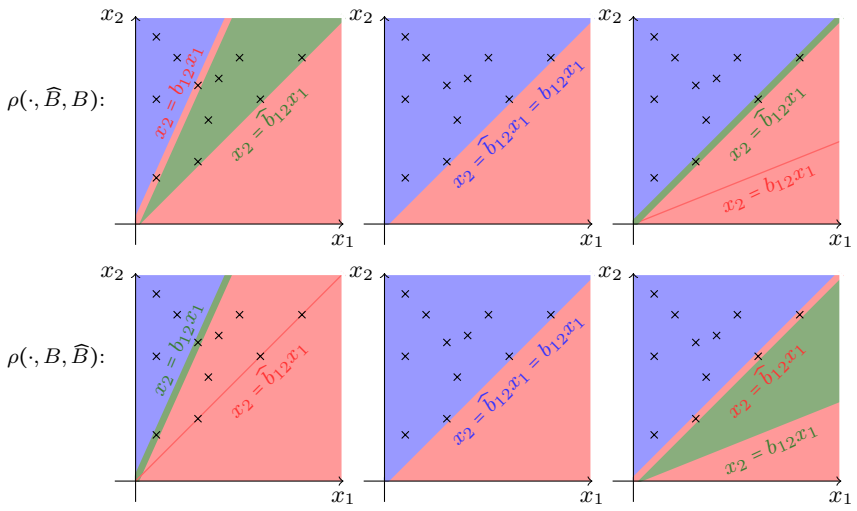
(4) $\widetilde{b}_{12} = \frac{x_2^{(t)}}{x_1^{(t)}} > \widehat{b}_{12}$ would be a GMLE if we do not require (4.4.5):
Similarly to Figure 4.4.3(c) in (3), Figure 4.4.3(d) indicates that (4.4.6) holds. Since $\operatorname{supp}(P_B) = \big\{\boldsymbol{x} \in \mathbb{R}^2_+ : x_2 \geq b_{12}x_1\big\}$ and all $\boldsymbol{x}^{(t)} \in \operatorname{supp}(P_B)$, it makes sense to require



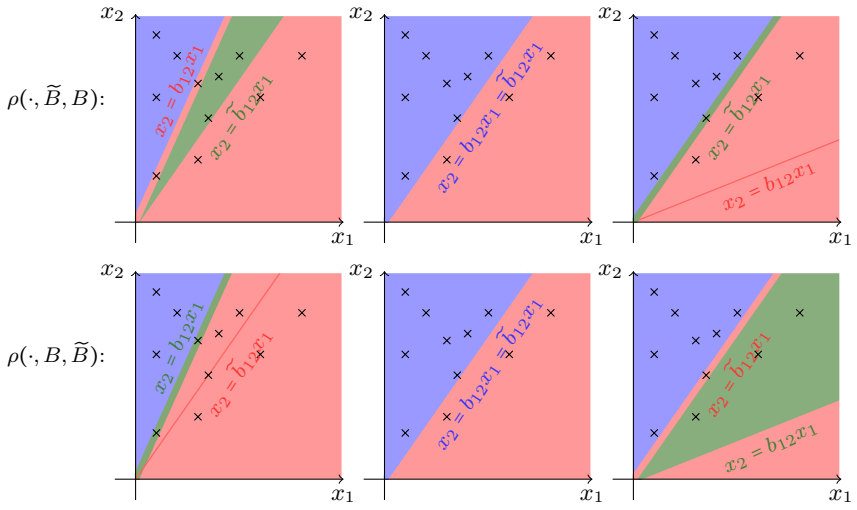
(a) $\widetilde{b}_{12} \in (0, \widehat{b}_{12})$ is no GMLE.



(b) $\widetilde{b}_{12} > \widehat{b}_{12}$ is no GMLE if $\widetilde{b}_{12} \neq \frac{x_2^{(t)}}{x_1^{(t)}}$ for all t.



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



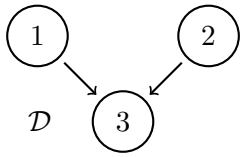
(d) $\widetilde{b}_{12} = \frac{x_2^{(t)}}{x_1^{(t)}} > \widehat{b}_{12}$ would be a GMLE if we do not require the first condition in (4.4.2).

Figure 4.4.3: Discussion of the GMLEs of b_{12} with respect to the density from Figure 4.4.2.

(4.4.5) and, therefore, to exclude here $\widetilde{b}_{12}$ as a GMLE of b_{12}. In summary, when setting the density of P_B with respect to $P_B + P_{B^*}$ outside P_B to zero, what is allowed due to property (A), then the first condition in (4.4.2) guarantees that for a GMLE $\widetilde{B}$ of B all $\boldsymbol{x}^{(t)} \in \operatorname{supp}(P_{\widetilde{B}})$. □

A density of P_B with respect to $P_B + P_{B^*}$ is only uniquely defined up to almost sure equality. This is exactly what may cause problems in the Kiefer-Wolfowitz definition of a GMLE. So various versions of the Radon-Nikodym derivatives can lead to many different GMLEs. Since the author of [65] missed the specification of the exact version in the Kiefer-Wolfowitz definition, he developed another definition of a nonparametric MLE. However, the same problem can arise with the classical definition of the MLE what [65] illustrates by the Gaussian distribution, and the classical MLE became accepted anyhow. In the following example we investigate how the GMLEs of B differ depending on the choice of the density.

Example 4.4.3. [The GMLEs depend on the choice of the density ρ]
Consider the DAG



and define $\widehat{b}_{13} := \bigwedge_{t=1}^{n} \frac{x_3^{(t)}}{x_1^{(t)}}$ and $\widehat{b}_{23} := \bigwedge_{t=1}^{n} \frac{x_3^{(t)}}{x_2^{(t)}}$. Let $\widehat{B}$ be the corresponding ML coefficient matrix. Figure 4.4.4 shows twelve different densities of P_B with respect to $P_B + P_{B^*}$ for the nine combinations of the three orders between b_{13} and b_{13}^* as well as the three orders between b_{23} and b_{23}^*. That these are really densities can be derived from Theorem 4.4.7 below or at least by the same arguments used in its proof. In Table 4.2 the corresponding GMLEs of B are presented. Depending on the observations, we obtain different GMLEs. For the densities ρ_1 and ρ_2, we discuss the potential GMLEs in Figures 4.4.5, 4.4.6 similarly as in Figure 4.4.3 for Example 4.4.2.

We have four events that may occur, namely,

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

where F_1 is by (4.2.3) and (4.3.1) the only null event. Assuming we exclude such null events in the observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and we observe $x_3^{(t)} = \widehat{b}_{13}x_1^{(t)} = \widehat{b}_{23}x_2^{(t)}$ for some t, then $(\widehat{b}_{13}, \widehat{b}_{23})$ is no exact estimate of (b_{13}, b_{23}). The densities $\rho_2, \rho_4, \rho_6, \rho_8, \rho_{10}, \rho_{11}$ recognize this and suggest $(\widehat{b}_{13}, \widehat{b}_{23})$ as a GMLE of (b_{13}, b_{23}) only if we do not observe $x_3^{(t)} = \widehat{b}_{13}x_1^{(t)} = \widehat{b}_{23}x_2^{(t)}$ for some t. This is because these densities are zero on the null set $\{\boldsymbol{x} \in \mathbb{R}_+^3 : x_3 = b_{13}x_1 = b_{23}x_2\}$.

Generally, all densities assume that at least one ML coefficient is estimated exactly. All of them consider $\widehat{B}$ as the unique GMLE if the minimal observed ratios of $\frac{X_3}{X_1}$ and $\frac{X_3}{X_2}$ have not occured in the same observation. Concerning the possible events, this seems to be reasonable. Apart from that, each density leads to more meaningful and less useful GMLEs. Consider, for example, the third situation from Table 4.2. The minimal observed ratio $\widehat{b}_{13}$ of $\frac{X_3}{X_1}$ has then been observed at least twice. Excluding null events would mean that $\widehat{b}_{13}$ is an atom of $\frac{X_3}{X_1}$. Since the

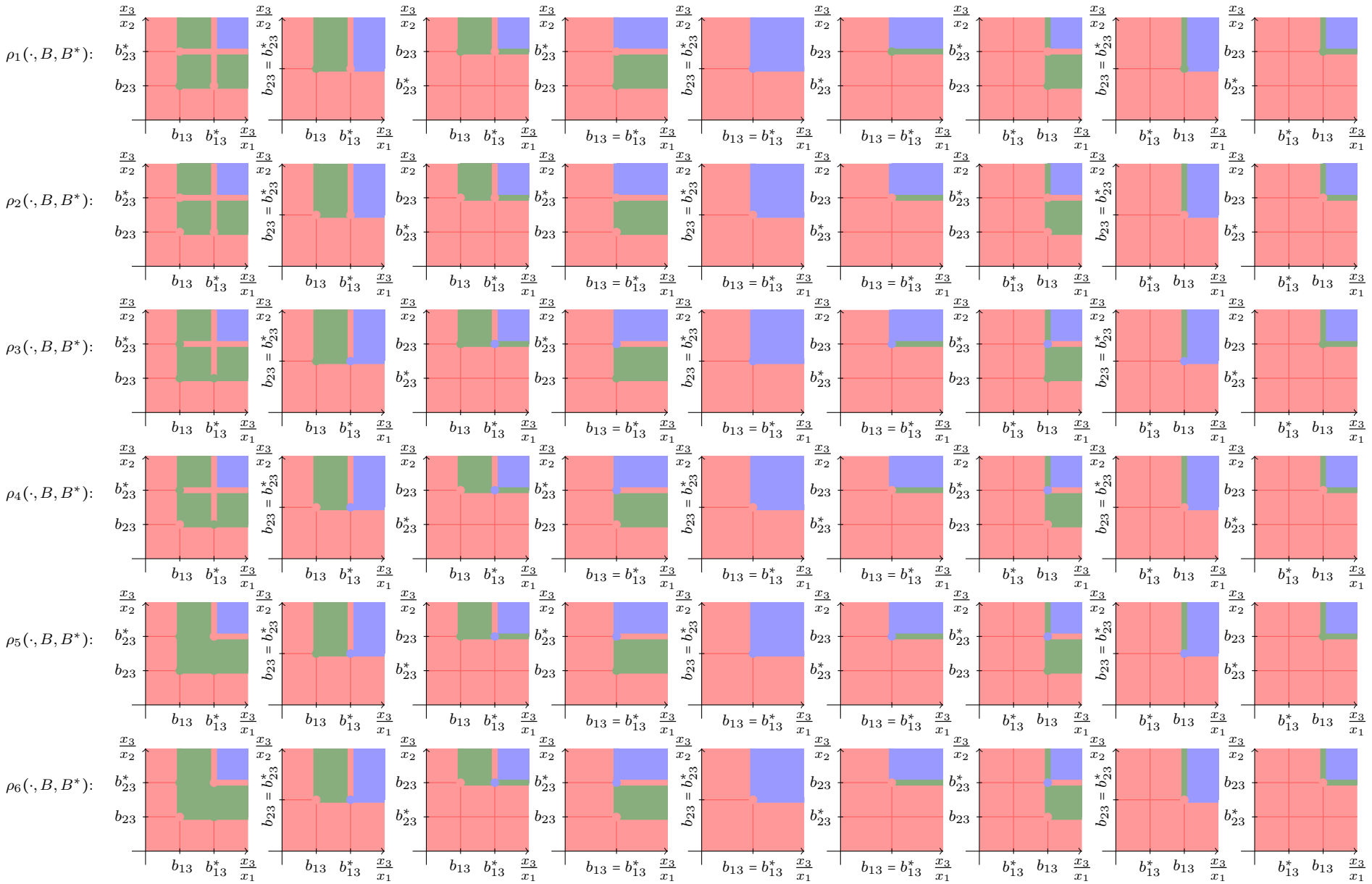


Figure 4.4.4: Different densities $\frac{dP_B}{d(P_B+P_{B^*})}$ for $B, B^* \in \mathcal{B}$ with the DAG $\mathcal{D}$ from Example 4.4.3 as a function of $\frac{x_3}{x_1}$ and $\frac{x_3}{x_2}$. The area where the respective density is $0/\frac{1}{2}/1$ is coloured in red/blue/green. The corresponding GMLEs are given in Table 4.2.

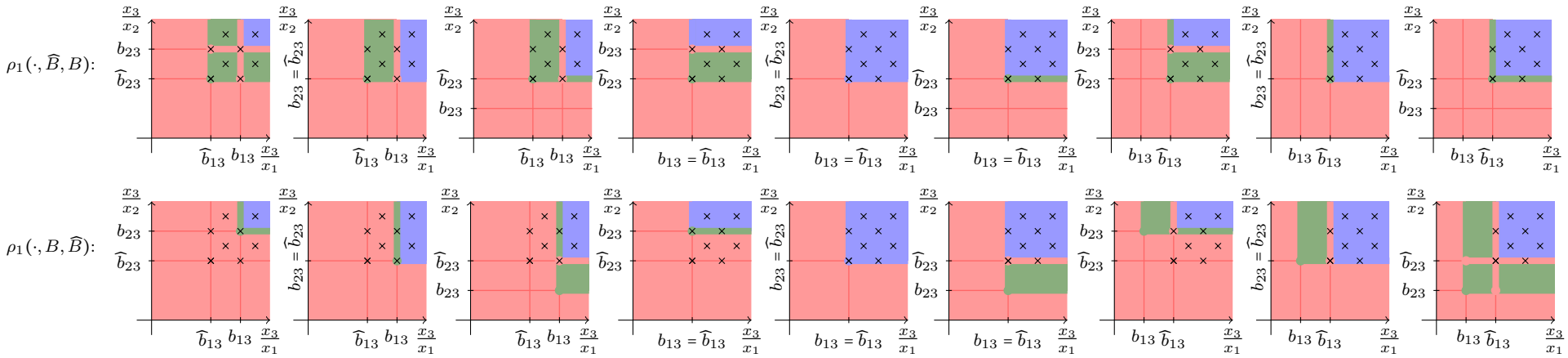
Figure 4.4.4: continued.

Table 4.2: The GMLEs corresponding to the densities depicted in Figure 4.4.4.

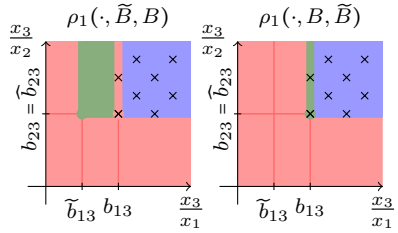
Situation	GMLE	Density
$\nexists s : \frac{x_3^{(s)}}{x_1^{(s)}} = \widehat{b}_{13}, \frac{x_3^{(s)}}{x_2^{(s)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widehat{b}_{23})$	ρ_1–ρ_{12}
$\exists s : \frac{x_3^{(s)}}{x_1^{(s)}} = \widehat{b}_{13}, \frac{x_3^{(s)}}{x_2^{(s)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widehat{b}_{23})$	$\rho_1, \rho_3, \rho_5, \rho_7, \rho_9, \rho_{12}$
$\exists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} > \widehat{b}_{23}$	$(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$	$\rho_2, \rho_4, \rho_6, \rho_8, \rho_{10}, \rho_{11}$
$\exists u : \frac{x_3^{(u)}}{x_1^{(u)}} > \widehat{b}_{13}, \frac{x_3^{(u)}}{x_2^{(u)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23})$	$\rho_2, \rho_4, \rho_6, \rho_8, \rho_{10}, \rho_{11}$
$\exists s : \frac{x_3^{(s)}}{x_1^{(s)}} = \widehat{b}_{13}, \frac{x_3^{(s)}}{x_2^{(s)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widehat{b}_{23})$	$\rho_1, \rho_3, \rho_5, \rho_7, \rho_9, \rho_{12}$
$\exists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} > \widehat{b}_{23}$	$(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$	ρ_2, ρ_{11}
$\nexists u : \frac{x_3^{(u)}}{x_1^{(u)}} > \widehat{b}_{13}, \frac{x_3^{(u)}}{x_2^{(u)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23})$	$\rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_8, \rho_9, \rho_{10}, \rho_{11}, \rho_{12}$
$\exists s : \frac{x_3^{(s)}}{x_1^{(s)}} = \widehat{b}_{13}, \frac{x_3^{(s)}}{x_2^{(s)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widehat{b}_{23})$	$\rho_1, \rho_3, \rho_5, \rho_7, \rho_9, \rho_{12}$
$\nexists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} > \widehat{b}_{23}$	$(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$	$\rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_8, \rho_9, \rho_{10}, \rho_{11}, \rho_{12}$
$\exists u : \frac{x_3^{(u)}}{x_1^{(u)}} > \widehat{b}_{13}, \frac{x_3^{(u)}}{x_2^{(u)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23})$	ρ_2, ρ_{11}
$\exists s : \frac{x_3^{(s)}}{x_1^{(s)}} = \widehat{b}_{13}, \frac{x_3^{(s)}}{x_2^{(s)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widehat{b}_{23})$	$\rho_1, \rho_3, \rho_5, \rho_7, \rho_9, \rho_{12}$
$\nexists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} > \widehat{b}_{23}$	$(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$	$\rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_8, \rho_9, \rho_{10}, \rho_{11}, \rho_{12}$
$\nexists u : \frac{x_3^{(u)}}{x_1^{(u)}} > \widehat{b}_{13}, \frac{x_3^{(u)}}{x_2^{(u)}} = \widehat{b}_{23}$	$(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23})$	$\rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_8, \rho_9, \rho_{10}, \rho_{11}, \rho_{12}$

only atom of $\frac{X_3}{X_1}$ is b_{13} (cf. Table 4.1), b_{13} is estimated by $\widehat{b}_{13}$ exactly and $(\widetilde{b}_{13}, \widehat{b}_{23})$ for $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$ is rather an inappropriate estimate. All densities except ρ_2 and ρ_{11} recognize this and do not consider this estimate a GMLE of b_{13}.

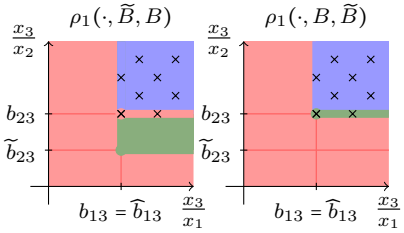
To summarize, setting a density used to determine GMLEs of B to zero on all null events as described above, then the first condition in (4.4.2) guarantees that $\widetilde{B} \in \mathcal{B}$ can only be a GMLE of B if no observation belongs to a $P_{\widetilde{B}}$-null event. Furthermore, we notice that the distributional properties of $\boldsymbol{X}$ allow us to decide whether a GMLE is a sensible estimate or not. So the uncertainty about how to select unproblematic density versions does not matter too much in our situation. □



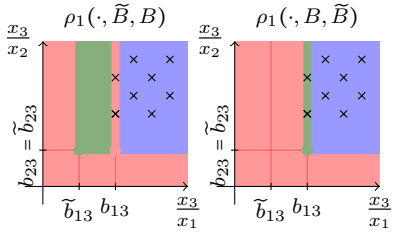
(a) $(\widehat{b}_{13}, \widehat{b}_{23})$ is a GMLE.



(b) $(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} < \widehat{b}_{13}$ is no GMLE.

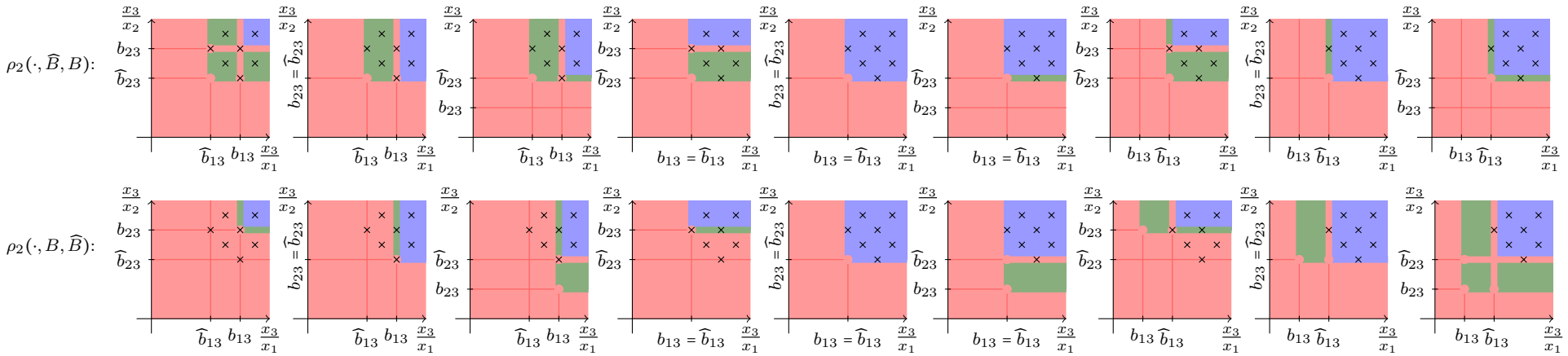


(c) $(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} < \widehat{b}_{23}$ is no GMLE.

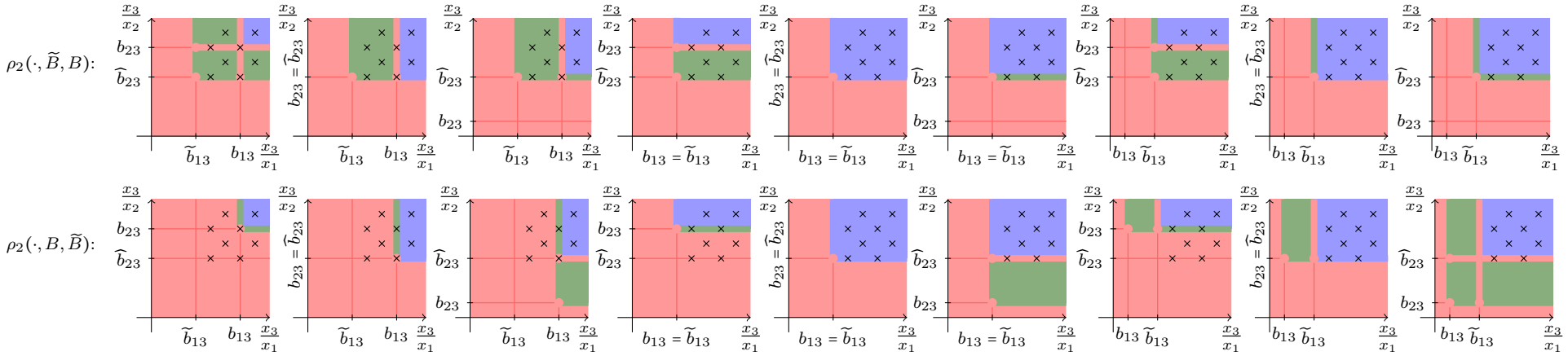


(d) $(\widetilde{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{13} < \widehat{b}_{13}$ and $\widetilde{b}_{23} < \widehat{b}_{23}$ is no GMLE.

Figure 4.4.5: Discussion of the possible GMLEs of (b_{13}, b_{23}) with respect to the density ρ_1 from Figure 4.4.4.

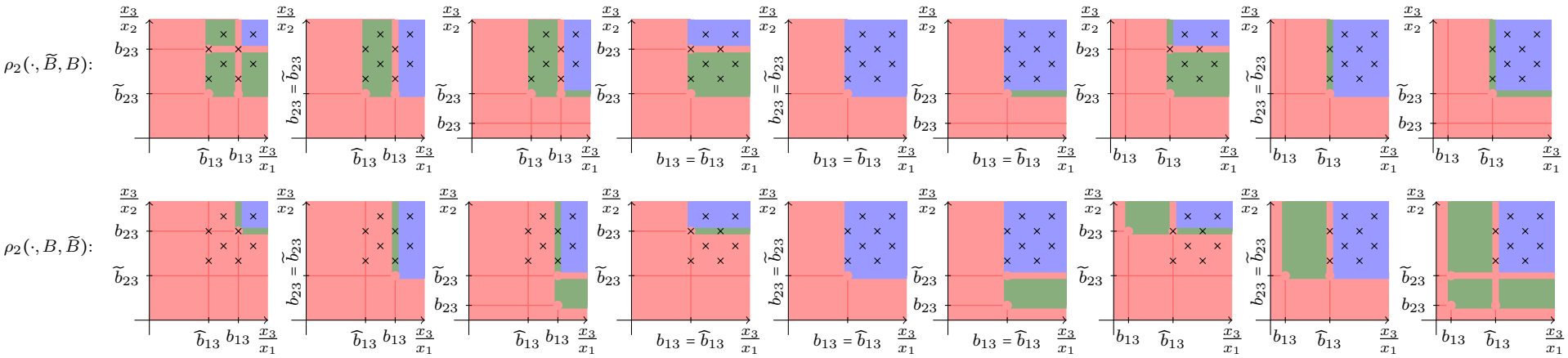


(a) $(\widehat{b}_{13}, \widehat{b}_{23})$ is a GMLE if $\nexists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}$.

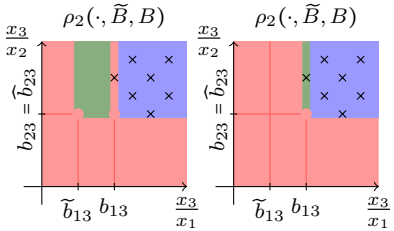


(b) $(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$ is a GMLE if $\exists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}$.

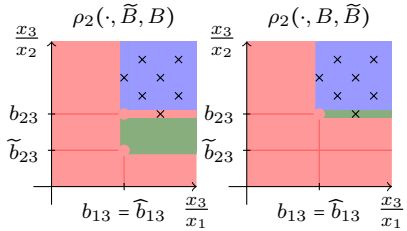
Figure 4.4.6: Discussion of the possible GMLEs of (b_{13}, b_{23}) with respect to the density ρ_2 from Figure 4.4.4.



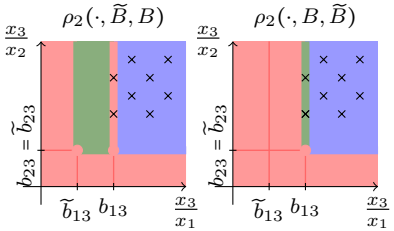
(c) $(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23})$ is a GMLE if $\exists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}$.



(d) $(\widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} < \widehat{b}_{13}$ is no GMLE if $\nexists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}$.



(e) $(\widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} < \widehat{b}_{23}$ is no GMLE if $\nexists t : \frac{x_3^{(t)}}{x_1^{(t)}} = \widehat{b}_{13}, \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}$.



(f) $(\widetilde{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{13} < \widehat{b}_{13}$ and $\widetilde{b}_{23} < \widehat{b}_{23}$ is no GMLE.

Figure 4.4.6: continued.

Example 4.4.3 raises the question what a sensible estimator of B is in the general case. Recall that $\mathcal{D}$ is assumed to be known. Table 4.1 shows that for $j \in \mathrm{an}(i)$ the minimal value that can be observed for $\frac{X_i}{X_j}$ is b_{ji}, which is an atom of $\frac{X_i}{X_j}$. This suggests the following estimate of the ML coefficients:

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

Davis and Resnick [13] suggested such minimal observed ratios for Max-ARMA processes. For $j \in V \setminus \mathrm{an}(i)$ we estimate b_{ji} obviously exactly. 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 these ML coefficients exactly as well. However, if n is not large, $\breve{B}$ is not necessarily a ML coefficient matrix of a recursive ML model on $\mathcal{D}$, as the following simple example shows.

Example 4.4.4. [$\breve{B}$ is not necessarily in $\mathcal{B}$]
Consider the DAG

$\mathcal{D}$ 

and assume we observe $\breve{b}_{13} > \breve{b}_{12}\breve{b}_{23}$; this can happen, for example, if the observations only belong to the events $\{X_2 > b_{12}X_1\} \cap \{X_3 > b_{23}X_2\}$ and $\{X_2 = b_{12}X_1\} \cap \{X_3 > b_{23}X_2\}$. The matrix $\breve{B}$ is, however, not contained in $\mathcal{B}$ (see (2.4.4)) and, therefore, no suitable estimate of B. □

The ML coefficients $(b_{ki}, i \in V, k \in \mathrm{pa}(i))$ define B uniquely, as the following remark makes clear. This means that it suffices to find appropriate estimates of the ML coefficients $(b_{ki}, i \in V, k \in \mathrm{pa}(i))$. If we estimate these ML coefficients exactly, then the remaining ML coefficients as well. The remark follows from the definition of a recursive ML model on $\mathcal{D}$ and Theorem 2.5.4(b).

Remark 4.4.5. Let $\mathcal{D}^{\mathrm{tr}}$ be the transitive reduction of $\mathcal{D}$. For the definition of $\mathcal{D}^{\mathrm{tr}}$, see e.g. Definition 2.4.1. We denote by $\mathrm{pa}^{\mathrm{tr}}(i)$ the parents of i in $\mathcal{D}^{\mathrm{tr}}$.

(i) $B = (b_{ij})_{d\times d} \in \mathcal{B}$ if and only if for every $i \in V$,

$$b_{ii} = 1, \ b_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i), \ b_{ki} > 0 \text{ for } k \in \mathrm{pa}^{\mathrm{tr}}(i), \ \text{and } b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \text{ for } j \in \mathrm{an}(i),$$

where $d_{ji}(p) = \prod_{\nu=0}^{n-1} b_{k_\nu k_{\nu+1}}$ for a path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$.

(ii) Coefficients $(b_{ki}, i \in V, k \in \mathrm{pa}(i))$ are entries of a matrix $B \in \mathcal{B}$ if and only if for every $i \in V$, $b_{ki} > 0$ for $k \in \mathrm{pa}^{\mathrm{tr}}(i)$ and $b_{ki} \geq \bigvee_{p \in P_{ki} \setminus \{[k \to i]\}} d_{ki}(p)$ for $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i)$. In this case, the remaining ML coefficients are uniquely given for $i \in V$ by

$$b_{ii} = 1, \ b_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i), \ \text{and } b_{ji} = \bigvee_{p \in P_{ji}} d_{ji}(p) \text{ for } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i).$$

□

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

$$\Big(\bigwedge_{s=1}^{n}\frac{x_{k_1}^{(s)}}{x_{k_0}^{(s)}}\Big)\Big(\bigwedge_{s=1}^{n}\frac{x_{k_2}^{(s)}}{x_{k_1}^{(s)}}\Big)\cdots\Big(\bigwedge_{s=1}^{n}\frac{x_{k_n}^{(s)}}{x_{k_{n-1}}^{(s)}}\Big) \le \frac{x_{k_1}^{(t)}}{x_{k_0}^{(t)}}\frac{x_{k_2}^{(t)}}{x_{k_1}^{(t)}}\cdots\frac{x_{k_n}^{(t)}}{x_{k_{n-1}}^{(t)}} = \frac{x_i^{(t)}}{x_j^{(t)}} = \breve{b}_{ji} = \bigwedge_{s=1}^{n}\frac{x_i^{(s)}}{x_j^{(s)}}. \tag{4.4.7}$$

By Remark 4.4.5(ii) this proves that the coefficients $(\breve{b}_{ki}, i \in V, k \in \mathrm{pa}(i))$ are entries of a unique matrix $\widehat{B} \in \mathcal{B}$. The entries of $\widehat{B}$ can be computed by

$$\begin{aligned}&\widehat{b}_{ii} = 1,\ \widehat{b}_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i),\ \widehat{b}_{ki} = \bigwedge_{t=1}^{n}\frac{x_i^{(t)}}{x_k^{(t)}} \text{ for } k \in \mathrm{pa}(i), \text{ and}\\ &\widehat{b}_{ji} = \bigvee_{p\in P_{ji}} \widehat{d}_{ji}(p) \text{ for } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i),\end{aligned} \tag{4.4.8}$$

where $\widehat{d}_{ji}(p) = \prod_{\nu=0}^{n-1}\widehat{b}_{k_\nu k_{\nu+1}}$ for a path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$. We can use Theorem 2.2.4 with $c_{ki} = b_{ki}$ and, hence, the matrix product $\odot$ defined in (2.2.2) to compute $\widehat{B}$ more efficiently than by the path analysis described in (4.4.8). An even more efficient computation of $\widehat{B}$ by $\odot$ is shown in Theorem 4.2 of Zhang [76].

We learn from (4.4.8) and Remark 4.4.5(i) that $\breve{B} \in \mathcal{B}$ if and only if $\breve{B} = \widehat{B}$. By (4.4.7) and Remark 4.4.5(ii) we have $b_{ji} \le \widehat{b}_{ji} \le \breve{b}_{ji}$ 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 $\widehat{B}$ estimates B more precisely than $\breve{B}$. So there is no reason why we should prefer $\breve{B}$ to $\widehat{B}$. As explained above when we have introduced $\breve{B}$, if n is large, we estimate $(b_{ki}, i \in V, k \in \mathrm{pa}(i))$ with high probability by $(\widehat{b}_{ki}, i \in V, k \in \mathrm{pa}(i))$ exactly and, hence, by $\widehat{B}$ the whole matrix B. In that case, we also have $\widehat{B} = \breve{B}$. We explain this 'exact estimation' more precisely in Section 4.4.2. In summary, $\widehat{B}$ seems to be a reasonable estimate of B and the only reasonable for n sufficiently large. As discussed and observed in Example 4.4.3, we can say relatively well whether $\widehat{B}$ is a reasonable estimate of B or whether some ML coefficients are overestimated. The distributional properties of the ratios between two components of $\boldsymbol{X}$ help us with this. The same applies to all GMLEs we obtain.

Because of the mentioned properties of $\widehat{B}$, we would expect that $\widehat{B}$ is a GMLE of B and for n sufficiently large even that it is the unique GMLE. That is exactly what we show 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 (4.4.3) and leads to $\widehat{B}$ as a GMLE of B. This density has a particularly nice structure and representation. However, one should have in mind from Example 4.4.3 and Table 4.2 that several such densities may exist, even where $\widehat{B}$ is no 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 followig representation for the components of the recursive ML model $\boldsymbol{X}$:

$$X_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}X_k \vee Z_i; \quad \text{in particular, } X_i \ge \bigvee_{k\in\mathrm{pa}(i)} b_{ki}X_k, \tag{4.4.9}$$

which has been shown in Theorem 2.4.2. We start with the specification of $A_{1/2}(B,B^*)$ and prove a property needed subsequently to verify property (B). Have in mind that, obviously,

for $i \in V$, $\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}^* x_k = \bigvee_{k\in\mathrm{pa}(i)} b_{ki} x_k\} = \varnothing$ if there is no $\boldsymbol{x} \in \mathbb{R}_+^d$ such that $x_i = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}^* x_k = \bigvee_{k\in\mathrm{pa}(i)} b_{ki} x_k$. This is, for example, the case if $b_{ki} > b_{ki}^*$ for all $k \in \mathrm{pa}(i)$ or $b_{ki} < b_{ki}^*$ for all $k \in \mathrm{pa}(i)$.

Lemma 4.4.6. *Let $B, B^* \in \mathcal{B}$. Let $\boldsymbol{X}$ be a recursive ML model on $\mathcal{D}$ with ML coefficient matrix B and $Z_1, \dots, Z_d$ its noise variables. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space of $(Z_1, \dots, Z_d)$ and, hence, of $\boldsymbol{X}$. Furthermore, we 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 $F \in \mathcal{F}$,

$$\mathbb{P}(F \cap \{\boldsymbol{X} \in A_{1/2}(B, B^*)\}) = \mathbb{P}(F \cap \Omega(B, B^*)). \tag{4.4.10}$$

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

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

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 less than $d+1$ nodes. Furthermore, assume without loss of generality that $d+1$ is a terminal node (i.e., $\mathrm{de}(d+1) = \varnothing$). Since $(X_1, \dots, X_d)$ is 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

$$\mathbb{P}(F \cap \{\boldsymbol{X} \in A_{1/2}(B, B^*)\}) = \mathbb{P}\Big(F \cap \bigcap_{i=1}^{d+1} \big(\Omega_{1/2}^{1,i} \cup \Omega_{1/2}^{2,i}\big)\Big) = \mathbb{P}\Big(F \cap \bigcap_{i=1}^{d} \Omega_i \cap \big(\Omega_{1/2}^{1,d+1} \cup \Omega_{1/2}^{2,d+1}\big)\Big). \tag{4.4.11}$$

For every $i \in V$ we have by (4.2.3) on Ω_i by definition,

$$X_i = \bigvee_{j\in\mathrm{An}(i)} b_{ji} Z_j = \bigvee_{j\in\mathrm{An}(i)} b_{ji}^* Z_j. \tag{4.4.12}$$

Noting from the proof of Theorem 2.4.2 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 (4.4.12) 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 (4.2.3),

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

From (4.3.1) we then finally observe that $\bigcap_{i=1}^d \Omega_i \cap \big(\Omega^{1,d+1}_{1/2} \cup \Omega^{2,d+1}_{1/2}\big)$ and $\bigcap_{i=1}^d \Omega_i \cap \Omega_{d+1}$ only differ by a set of probability zero, and, hence, (4.4.10) follows from (4.4.11). □

The complete partition $\big\{A_0(B,B^*), A_{1/2}(B,B^*), A_1(B,B^*)\big\}$ of $\mathbb{R}^d_+$ we suggest is as follows,

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

Every vector $\boldsymbol{x}\in\mathbb{R}^d_+$ belongs to exactly one of these sets. This has to be understood that intersections correspond to all components have to satisfy something, and unions correspond to at least one component satisfies something. Then by definition all sets are disjoint, and by definition of $A_1(B,B^*)$ they really partition $\mathbb{R}^d_+$. We now show that this partition leads to a density as in (4.4.3) and, therefore, verify properties (A)–(C). In Example 4.4.2 we have already done this for the simple DAG $1\to 2$. For an arbitrary DAG $\mathcal{D}$, we can proceed likewise.

Theorem 4.4.7. *Let $B, B^*\in\mathcal{B}$, where $\mathcal{B}$ contains the ML coefficient matrices of all recursive ML models on the DAG $\mathcal{D}$. Then the function from $\mathbb{R}^d_+$ to $\{0,1/2,1\}$ such that*

$$\boldsymbol{x}\mapsto \rho(\boldsymbol{x},B,B^*) = \frac{1}{2}\cdot \mathbb{1}_{A_{1/2}(B,B^*)}(\boldsymbol{x}) + \mathbb{1}_{A_1(B,B^*)}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x}\in A_0(B,B^*),\\ \frac{1}{2}, & \text{if } \boldsymbol{x}\in A_{1/2}(B,B^*),\\ 1, & \text{if } \boldsymbol{x}\in A_1(B,B^*)\end{cases} \tag{4.4.13}$$

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

Proof. Let $\boldsymbol{X}$ be a recursive ML model on $\mathcal{D}$ with ML coefficient matrix B and $Z_1,\dots,Z_d$ its noise variables.

(A) Since V is finite, it suffices to show for every i,

$$P_B\big(\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i < \bigvee_{k \in \mathrm{pa}(i)} b_{ki} x_k\}\big) = \mathbb{P}\big(X_i < \bigvee_{k \in \mathrm{pa}(i)} b_{ki} X_k\big) = 0, \tag{4.4.14}$$
$$P_B\big(\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = \bigvee_{k \in \mathrm{pa}(i)} b^*_{ki} x_k > \bigvee_{k \in \mathrm{pa}(i)} b_{ki} x_k\}\big) = \mathbb{P}\big(X_i = \bigvee_{k \in \mathrm{pa}(i)} b^*_{ki} X_k > \bigvee_{k \in \mathrm{pa}(i)} b_{ki} X_k\big) = 0.$$

The former is immediate by (4.4.9). By the same argument we have for the latter,

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

where we have used (4.2.3) and (4.3.1) for the last inequality and equality, respectively. Thus we have verified (A).
(B) Recall that the noise vectors of the recursive ML models are assumed to be identically distributed. Furthermore, note that the set $\Omega(B, B^*)$ from Lemma 4.4.6 is a subset of $\bigcap_{i \in V} \big\{X_i = \bigvee_{j \in \mathrm{An}(i): b_{ji} = b^*_{ji}} b_{ji} Z_j\big\}$. Thus, using that $\Omega(B, B^*) = \Omega(B^*, B)$, we then obtain from (4.4.10) for $A \in \mathcal{B}(\mathbb{R}_+^d)$,

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

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

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

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

As observed in Example 4.4.3, there exist several densities of P_B with respect to $P_B + P_{B^*}$. For example, the value on

$$A_{1/2}(B, B^*) \cap \bigcup_{i \in V} \Big[\bigcup_{k \in \mathrm{pa}(i): b^*_{ki} \neq \frac{b_{ji}}{b_{jk}} \forall j \in \mathrm{An}(k)} \Big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = b^*_{ki} x_k\Big\} \cup \bigcup_{k \in \mathrm{pa}(i): b_{ki} \neq \frac{b^*_{ji}}{b^*_{jk}} \forall j \in \mathrm{An}(k)} \Big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = b_{ki} x_k\Big\}\Big]$$

can be set arbitrarily, since this set is by property (C) from the proof of Theorem 4.4.7 and Table 4.1 a P_B- and a P_{B^*}-null set. The same applies to the sets

$$A_0(B, B^*) \cap \Big[\bigcup_{i \in V} \Big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i < \bigvee_{k \in \mathrm{pa}(i)} b^*_{ki} x_k\Big\} \cup \bigcup_{k \in \mathrm{pa}(i): b_{ki} \neq \frac{b^*_{ji}}{b^*_{jk}} \forall j \in \mathrm{An}(k)} \Big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = b_{ki} x_k\Big\}\Big],$$

$$A_1(B,B^*)\cap\Big[\bigcup_{i\in V}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i<\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\big\}\cup\bigcup_{k\in\mathrm{pa}(i):b^*_{ki}\neq\frac{b_{ji}}{b_{jk}}\,\forall j\in\mathrm{An}(k)}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i=b^*_{ki}x_k\big\}\Big].$$

Furthermore, we may set the density equal to one on

$$\bigcup_{i\in V}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i<\bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k\big\}\cup\bigcup_{k\in\mathrm{pa}(i):b_{ki}\neq\frac{b^*_{ji}}{b^*_{jk}}\,\forall j\in\mathrm{An}(k)}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i=b_{ki}x_k\big\}.$$

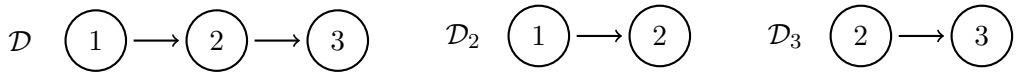
and equal to zero on

$$\bigcup_{i\in V}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i<\bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k\big\}\cup\bigcup_{k\in\mathrm{pa}(i):b^*_{ki}\neq\frac{b_{ji}}{b_{jk}}\,\forall j\in\mathrm{An}(k)}\big\{\boldsymbol{x}\in\mathbb{R}_+^d: x_i=b^*_{ki}x_k\big\}.$$

It is not necessary to use (4.4.9) but, according to (2.6.11), we could use the representation $X_i=\bigvee_{k\in\mathrm{pa}^B(i)} b_{ki}X_k\vee Z_i$, where $\mathrm{pa}^B(i)$ are the parents of i in the minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$. From these facts we can derive several densities of P_B with respect to $P_B+P_{B^*}$.

We observe an interesting relation between the density (4.4.13) for $\mathcal{D}$ and corresponding densities for subgraphs of $\mathcal{D}$.

Example 4.4.8. ["Marginal" densities ρ_i]
Consider the DAGs



Let ρ, ρ_2, and ρ_3 be the corresponding densities from (4.4.13). For the ML coefficient matrix B of a recursive ML model on $\mathcal{D}$, let B_{12} and B_{23} 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 $c_{23}=b_{23}$. Here B_{12} and B_{23} are the submatrices of B formed by the first two or the last two rows and columns, respectively. We then find for $\boldsymbol{x}=(x_1,x_2,x_3)\in\mathbb{R}_+^3$,

$$\begin{aligned}&\rho(\boldsymbol{x},B,B^*)\\&=\big(\rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)},B_{12},B^*_{12})\vee\rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)},B_{23},B^*_{23})\big)\mathbb{1}\big\{\rho_2(\boldsymbol{x}_{\mathrm{Pa}(2)},B_{12},B^*_{12})\wedge\rho_3(\boldsymbol{x}_{\mathrm{Pa}(3)},B_{23},B^*_{23})>0\big\}.\end{aligned}$$

This can be observed from Figure 4.4.7, where these densities are depicted as functions of $\frac{x_2}{x_1}$ and/or $\frac{x_3}{x_2}$ for all nine different orders between the ML coefficients. Conversely, ρ_2 and ρ_3 can be derived from ρ as follows:

$$\begin{aligned}\rho_2((x_1,x_2),B_{12},B^*_{12})&=\min_{y\in\{y\in\mathbb{R}_+:\rho((x_1,x_2,y),B,B^*)>0\}}\rho((x_1,x_2,y),B,B^*),\\\rho_3((x_2,x_3),B_{23},B^*_{23})&=\min_{y\in\{y\in\mathbb{R}_+:\rho((y,x_2,x_3),B,B^*)>0\}}\rho((y,x_2,x_3),B,B^*),\end{aligned}$$

which we learn from Figure 4.4.7 again. □

We extend the findings from Example 4.4.8 to the general case. Furthermore, we show that the densities ρ_i are (regular) conditional densities. For the definition of a regular conditional

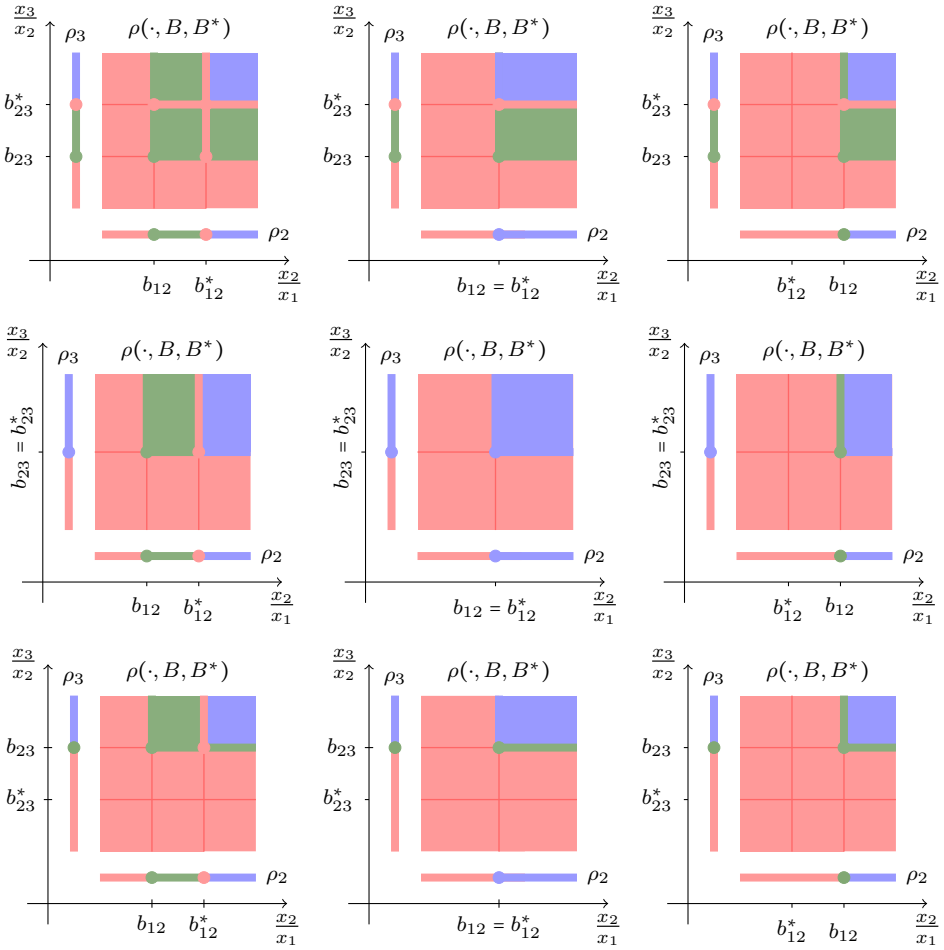


Figure 4.4.7: The densities $\rho(\boldsymbol{x} = (x_1, x_2, x_3), B, B^*)$, $\rho_2((x_1, x_2), B_{12}, B_{12}^*)$, $\rho_3((x_2, x_3), B_{23}, B_{23}^*)$ from Example 4.4.8 as functions of $\frac{x_2}{x_1}$ and/or $\frac{x_3}{x_2}$. The area where the respective density is $0/\frac{1}{2}/1$ is coloured in red/blue/green.

distribution, see e.g. Chapter 8.3 of Klenke [42].

Proposition 4.4.9. *Let $B, B^* \in \mathcal{B}$ and $\boldsymbol{X}, \boldsymbol{X}^*$ corresponding recursive ML models on $\mathcal{D}$. For $i \in V$, let ρ_i be the density given in (4.4.13) 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 (4.4.13),

$$\begin{aligned}\rho(\boldsymbol{x}, B, B^*) &= \Big(\bigvee_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*)\Big) \mathbb{1}\Big\{\bigwedge_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) > 0\Big\} \qquad (4.4.15)\\ &= \begin{cases} 0, & \text{if } \bigwedge_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) = 0, \\ \bigvee_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*), & \text{if } \bigwedge_{i \in V} \rho_i(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) > 0. \end{cases}\end{aligned}$$

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

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

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

(c) The function from $\mathbb{R}_+^d$ *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,

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

From this we can observe (a) and (b).
(c) We denote the noise variables of $\boldsymbol{X}$ again by $Z_1,\dots,Z_d$. It is not difficult to verify that

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

is a regular conditional distribution function of X_i given $\boldsymbol{X}_{\mathrm{pa}(i)}$. To get an idea for this, use (4.4.9) and the independence of the noise variables to obtain

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

Since we assume that the noise vectors of $\boldsymbol{X}$ and $\boldsymbol{X}^*$ are identically distributed,

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

is a regular conditional distribution function of X_i^* given $\boldsymbol{X}^*_{\mathrm{pa}(i)}$. Figure 4.4.8 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. (4.4.3) and the related discussion) that

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

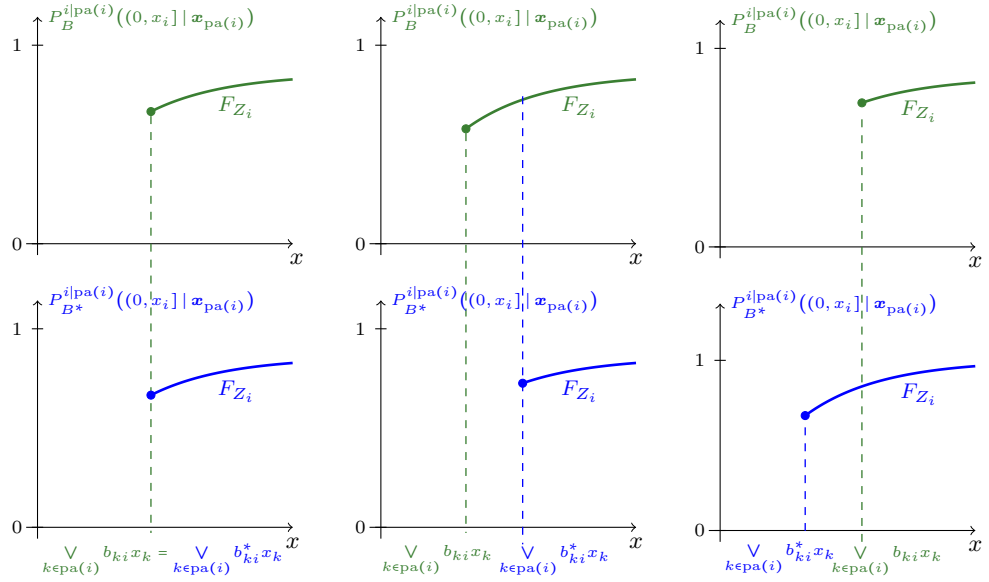


Figure 4.4.8: The conditional distribution functions from the proof of Proposition 4.4.9(c).

$$\text{if } \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki}x_k = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k,$$

$$P_B^{i|\mathrm{pa}(i)}\big((0,y]\cap\big(\bigvee_{k\in\mathrm{pa}(i)}(b_{ki}\vee b^*_{ki})x_k,\infty\big)\,|\,\boldsymbol{x}_{\mathrm{pa}(i)}\big) = P_{B^*}^{i|\mathrm{pa}(i)}\big((0,y]\cap\big(\bigvee_{k\in\mathrm{pa}(i)}(b_{ki}\vee b^*_{ki})x_k,\infty\big)\,|\,\boldsymbol{x}_{\mathrm{pa}(i)}\big).$$

Since F_{Z_i} is continuous, this can be read directly from Figure 4.4.8. □

Figure 4.4.9 below shows another example for the DAGs $\mathcal{D}_i$ from Proposition 4.4.9.

According to (4.4.15), the density ρ is the maximum of the conditional densities ρ_i. This is an interesting result in consideration of the following remark.

Remark 4.4.10. The distribution of $\boldsymbol{X}$ is Markov relative to $\mathcal{D}$ (see e.g. (2.1.2) and the related discussion). If the distribution $\mathcal{L}(\boldsymbol{Y})$ of a positive real-valued random vector $\boldsymbol{Y}$ satisfies this property and has density f with respect to a product measure, then $\mathcal{L}(\boldsymbol{Y})$ admits a *recursive factorization* according to $\mathcal{D}$; i.e.,

$$f(\boldsymbol{y}) = \prod_{i\in V} f_{i|\mathrm{pa}(i)}\big(y_i\,|\,\boldsymbol{y}_{\mathrm{pa}(i)}\big), \quad \boldsymbol{y}\in\mathbb{R}^d_+,$$

where the functions $f_{i|\mathrm{pa}(i)}(\cdot\,|\,\boldsymbol{y}_{\mathrm{pa}(i)})$ are densities for the conditional distribution of Y_i given $\boldsymbol{Y}_{\mathrm{pa}(i)} = \boldsymbol{y}_{\mathrm{pa}(i)}$. For more details, see Section 3.2.2 of Lauritzen [47]. □

In what follows we now determine all GMLEs of B with respect to the density ρ given in (4.4.13). The value of $\rho(\boldsymbol{x}^{(t)}, B, B^*)$ indicates the membership of $\boldsymbol{x}^{(t)}$, having the value 0 if $\boldsymbol{x}^{(t)}\in A_0(B,B^*)$, 1/2 if $\boldsymbol{x}^{(t)}\in A_{1/2}(B,B^*)$, and 1 if $\boldsymbol{x}^{(t)}\in A_1(B,B^*)$. So, to find the GMLEs of B, we have to understand when an observation $\boldsymbol{x}^{(t)}$ is in which of these sets. This is investigated in the next lemma, which is simply by definition of $A_0(B,B^*)$, $A_{1/2}(B,B^*)$, and $A_1(B,B^*)$.

Lemma 4.4.11. *Let $\widehat{B}$ be the matrix from* (4.4.8).

(a) *No* $\boldsymbol{x}^{(t)} \in A_0(B,B)$ *if and only if all* $\boldsymbol{x}^{(t)} \in A_{1/2}(B,B)$.

(b) $\boldsymbol{x}^{(t)} \in A_{1/2}(B,B^*)$ *if and only if* $\boldsymbol{x}^{(t)} \in A_{1/2}(B^*,B)$.

(c) *If* $\boldsymbol{x}^{(t)} \in A_1(B,B^*)$, *then* $\boldsymbol{x}^{(t)} \in A_0(B^*,B)$.

(d) *All* $\boldsymbol{x}^{(t)} \in A_{1/2}(B,B)$ *if and only if for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$, $b_{ki} \le \widehat{b}_{ki}$.

(e) *If all* $\boldsymbol{x}^{(t)} \in A_{1/2}(B,B)$, *then* $\boldsymbol{x}^{(t)} \in A_0(B,B^*)$ *if and only if* $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} b^*_{ki} x_k^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} b_{ki} x_k^{(t)}$ *for some* $i \in V$.

(f) *If all* $\boldsymbol{x}^{(t)} \in A_{1/2}(B,B)$, *then no* $\boldsymbol{x}^{(t)} \in A_0(\widehat{B},B)$.

By this lemma we find nice characterizations of the GMLEs of B. These characterizations only depend on the estimates of the ML coefficients that belong to an edge of $\mathcal{D}$. With the choice of the partition $\{A_0(B,B^*),\ A_{1/2}(B,B^*), A_1(B,B^*)\}$ defining ρ and Remark 4.4.5 in mind, this is exactly what was expected and makes sense. Since $\rho(\cdot,B,B^*)$ equals zero outside $\mathrm{supp}(P_B) = A_{1/2}(B,B)$, the first condition in (4.4.2) ensures that for a GMLE $\widetilde{B} \in \mathcal{B}$ all $\boldsymbol{x}^{(t)} \in \mathrm{supp}(P_{\widetilde{B}})$, which is a reasonable property. Throughout the following discussion about the GMLEs of B, we use implicitly that the matrix $\widehat{B}$ from (4.4.8) is an element of $\mathcal{B}$ as well as (4.4.8), and Remark 4.4.5. All GMLEs relate to the density ρ from (4.4.13) without mentioning this explicitly again.

Theorem 4.4.12. *Let* $\boldsymbol{x}^{(t)} = \left(x_1^{(t)},\dots,x_d^{(t)}\right)$ *for* $t = 1,\dots,n$ *be independent realizations of a recursive ML model* $\boldsymbol{X}$ *on a given DAG* $\mathcal{D}$ *with ML coefficient matrix* B. *Then* $\widetilde{B} \in \mathcal{B}$ *is a GMLE of* B *if and only if one of the following conditions is satisfied.*

(a) $\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)},\widetilde{B},\widetilde{B}) \neq 0$ *and* $\prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)},B,\widetilde{B}) \le \prod_{t=1}^{n} \rho(\boldsymbol{x}^{(t)},\widetilde{B},B)$ *for all* $B \in \mathcal{B}$.

(b) *All* $\boldsymbol{x}^{(t)} \in A_{1/2}(\widetilde{B},\widetilde{B})$ *and for all* $B \in \mathcal{B}$, *if some* $\boldsymbol{x}^{(t)} \in A_0(\widetilde{B},B)$, *then some* $\boldsymbol{x}^{(s)} \in A_0(B,\widetilde{B})$.

(c) $\widetilde{b}_{ki} \le \widehat{b}_{ki}$ *for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$, *and no* $\boldsymbol{x}^{(t)} \in A_0(\widetilde{B},\widehat{B})$.

(d) *All* $\boldsymbol{x}^{(t)} \in A_{1/2}(\widetilde{B},\widehat{B}) = A_{1/2}(\widehat{B},\widetilde{B})$.

(e) *For every* $i \in V$, $\widetilde{b}_{ki} \le \widehat{b}_{ki}$ *for all* $k \in \mathrm{pa}(i)$ *with strict inequality only if for every* $\boldsymbol{x}^{(t)}_{\mathrm{Pa}(i)}$ *such that* $\frac{x_i^{(t)}}{x_k^{(t)}} = \widehat{b}_{ki}$, $\widetilde{b}_{\widetilde{k}i} = \widehat{b}_{\widetilde{k}i} = \frac{x_i^{(t)}}{x_{\widetilde{k}}^{(t)}}$ *for some* $\widetilde{k} \in \mathrm{pa}(i)$.

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

Proof. (a) corresponds to the definition of a GMLE of B.
(b) is by definition of ρ and Lemma 4.4.11(a)–(c).
(c) We show the equivalence to (b). Assume (b) holds but not (c). With regard to Lemma 4.4.11(d), there must be some $\boldsymbol{x}^{(t)} \in A_0(\widetilde{B},\widehat{B})$. As $\widehat{B} \in \mathcal{B}$, we obtain from (b) that $\boldsymbol{x}^{(t)} \in A_0(\widehat{B},\widetilde{B})$. This contradicts, however, Lemma 4.4.11(f). By Lemma 4.4.11(d) it then remains to show that (c) implies

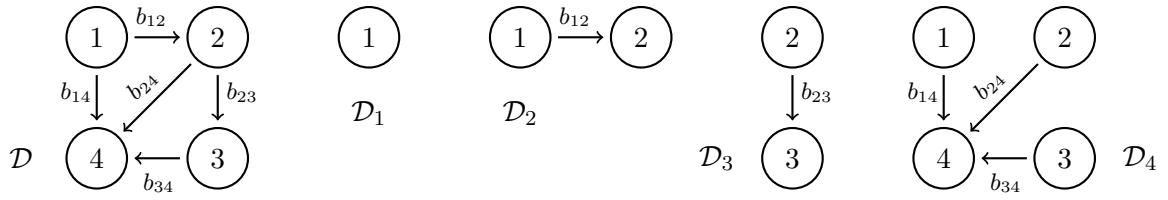


Figure 4.4.9: The DAGs $\mathcal{D}_i$ of the recursive ML models $\boldsymbol{Y}_i$ from Proposition 4.4.9 and Theorem 4.4.12(f) for a recursive ML model $\boldsymbol{X}$ on the DAG $\mathcal{D}$ depicted on the left-hand side with ML coefficient matrix B. The edges are marked with the corresponding ML coefficients. Note that $b_{12}, b_{14}, b_{34}, b_{24}$ can be arbitary positive numbers but $b_{24} \geq b_{23}b_{34}$.

the second property of (b). For this assume some $\boldsymbol{x}^{(t)} \in A_0(\widetilde{B}, B)$, where $B \in \mathcal{B}$. Lemma 4.4.11(e) yields for some i, $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} \widetilde{b}_{ki}x_k^{(t)}$. As $x_i^{(t)} \geq \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki}x_k^{(t)}$, we necessarily have that $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki}x_k^{(t)}$; otherwise, it would be a contradiction to (c) because of Lemma 4.4.11(e). Hence, $b_{ki} > \widehat{b}_{ki}$ for some $k \in \mathrm{pa}(i)$. For $\boldsymbol{x}^{(s)}$ such that $\frac{x_i^{(s)}}{x_k^{(s)}} = \widehat{b}_{ki}$, we then find $x_i^{(s)} < b_{ki}x_k^{(s)}$, which finally proves that $\boldsymbol{x}^{(s)} \in A_0(B, \widetilde{B})$, as $x_i^{(s)} < \bigvee_{k\in\mathrm{pa}(i)} b_{ki}x_k^{(s)}$.
(d) follows from (c) as we may observe from Lemma 4.4.11(c), (d), (f). By definition, if all $\boldsymbol{x}^{(t)} \in A_{1/2}(\widetilde{B}, \widehat{B})$, then all $\boldsymbol{x}^{(t)} \in A_{1/2}(\widetilde{B}, \widetilde{B})$. Thus by Lemma 4.4.11(d), (d) implies (c).
(e) It is not difficult to see that (c), (d), and (e) are equivalent.
(f) For $i \in V$ let $\big(y_\ell^{(t)}, \ell \in \mathrm{Pa}(i)\big)$ for $t = 1, \ldots, n$ be independent realizations of $\boldsymbol{Y}_i$. We know from (e) that $(\widetilde{b}_{ki}, k \in \mathrm{pa}(i))$ is a GMLE of $(b_{ki}, k \in \mathrm{pa}(i))$ if and only if for every $k \in \mathrm{pa}(i)$, $\widetilde{b}_{ki} \leq \bigwedge_{t=1}^n \frac{y_i^{(t)}}{y_k^{(t)}} =: \widehat{c}_{ki}$ with strict inequality only if for every t such that $y_i^{(t)} = \widehat{c}_{ki}y_k^{(t)}$, $\widetilde{b}_{\widetilde{k}i}y_{\widetilde{k}}^{(t)} = \widehat{c}_{\widetilde{k}i}y_{\widetilde{k}}^{(t)} = y_i^{(t)}$ for some $\widetilde{k} \in \mathrm{pa}(i)$. This shows the equivalence between (e) and (f). □

Theorem 4.4.12(f) seems to reduce the problem of finding GMLEs of B to the same problem for recursive ML models on DAGs that consist only of initial nodes (i.e., nodes without ancestors) and one terminal node (i.e., a node without descendants). These DAGs have already appeared in Proposition 4.4.9. Figure 4.4.9 shows them for an example. For some matrix $\widetilde{B}$ to be a GMLE of B, it is, however, not enough for property (f) of Theorem 4.4.12 to hold. $\widetilde{B} \in \mathcal{B}$ is necessary as we show by an example.

Example 4.4.13. [$\widetilde{B} \in \mathcal{B}$ in Theorem 4.4.12 is necessary but property (f) provides edge weights leading to a GMLE $\widetilde{B}$ of B]
Consider the DAG $\mathcal{D}$ from Example 4.3.1, and assume we observe that $\widehat{b}_{13} = \widehat{b}_{12}\widehat{b}_{23}$ and no $\boldsymbol{x}^{(t)}$ with $x_3^{(t)} = \widehat{b}_{13}x_1^{(t)}$, $x_3^{(t)} > \widehat{b}_{23}x_2^{(t)}$. This can happen, for example, when $n = 1$ or all observations belong to the event $\{X_2 > b_{12}X_1\} \cap \{X_3 > b_{23}X_2\} \cap \{X_3 > b_{13}X_1\}$. Then $(\widehat{b}_{12}, \widetilde{b}_{13}, \widehat{b}_{23})$ with $\widetilde{b}_{13} \in (0, \widehat{b}_{13})$ cannot be a GMLE of (b_{12}, b_{13}, b_{23}), since $\widetilde{b}_{13} \geq \widehat{b}_{12}\widehat{b}_{23}$ is a necessary property (see e.g. Corollary 2.4.3(a)). However, property (f) of Theorem 4.4.12 holds: that (e) holds is clear, and (f) is always equivalent to (e) (cf. the proof of Theorem 4.4.12(f)). Note that the edge weights $\widetilde{c}_{12} = \widehat{b}_{12}$, $\widetilde{c}_{23} = \widehat{b}_{23}$, $\widetilde{c}_{13} \in (0, \widehat{b}_{13})$ lead to the ML coefficient matrix $\widehat{B}$ and thus belong by Theorem 4.4.12(e) to a recursive ML model on $\mathcal{D}$ whose ML coefficient matrix is a GMLE of B (cf. Corollary 4.4.17 below). □

So far, to find all GMLEs of B, we would determine all vectors with property (c), (d), or (e) of Theorem 4.4.12 and then test by using Remark 4.4.5(ii) which of them lead to a matrix $\widetilde{B} \in \mathcal{B}$. However, the second step is not necessary, since the observation from Example 4.4.13 is valid for the general case: the vectors from Theorem 4.4.12(c), (d), (e) are the edge weights leading to a GMLE $\widetilde{B}$ of B.

Corollary 4.4.14. *Assume the situation of Theorem 4.4.12. Recall from the definition of $\widehat{B}$ that $\widehat{b}_{ki} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_k^{(t)}}$ for $k \in \mathrm{pa}(i)$. Let c_{ki} be the edge weights of $\boldsymbol{X}$ and $\boldsymbol{Y}$ a recursive ML model on $\mathcal{D}$ with edge weights $\widetilde{c}_{ki}$. Then the ML coefficient matrix $\widetilde{B}$ of $\boldsymbol{Y}$ is a GMLE of B if and only if one of the following conditions is satisfied.*

(a) For every $i \in V$, $\widetilde{c}_{ki} \le \widehat{b}_{ki}$ for all $k \in \mathrm{pa}(i)$ and there is no $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$ such that $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki} x_k^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} \widetilde{c}_{ki} x_k^{(t)}$.

(b) For every $i \in V$, $x_i^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} (\widetilde{c}_{ki} \vee \widehat{b}_{ki}) x_k^{(t)}$ or $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} \widetilde{c}_{ki} x_k^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki} x_k^{(t)}$ for all $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$.

(c) For every $i \in V$, $\widetilde{c}_{ki} \le \widehat{b}_{ki}$ for all $k \in \mathrm{pa}(i)$ with strict inequality only if for every $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$ such that $\frac{x_i^{(t)}}{x_k^{(t)}} = \widehat{b}_{ki}$, $\widetilde{c}_{\widetilde{k}i} = \widehat{b}_{\widetilde{k}i} = \frac{x_i^{(t)}}{x_{\widetilde{k}}^{(t)}}$ for some $\widetilde{k} \in \mathrm{pa}(i)$.

(d) For every $i \in V$, the vector $(\widetilde{c}_{ki}, k \in \mathrm{pa}(i))$ is a GMLE of the edge weights $(c_{ki}, k \in \mathrm{pa}(i))$ of a recursive ML model on $\mathcal{D}_i = (\mathrm{Pa}(i), \{(k,i) : k \in \mathrm{pa}(i)\})$.

Proof. Observe from the proof that the properties (c)–(f) of Theorem 4.4.12 are equivalent even if $\widetilde{B} \notin \mathcal{B}$. Therefore, by definition of the sets $A_0(\widetilde{B}, \widehat{B}), A_{1/2}(\widetilde{B}, \widehat{B})$, we obtain the equvivalence between (a)–(d).

It suffices to show the equivalence between (a) and property (c) of Theorem 4.4.12. By Lemma 4.4.11(e), Theorem 4.4.12(c) holds if and only if for every $i \in V$, $\widetilde{b}_{ki} \le \widehat{b}_{ki}$ for all $k \in \mathrm{pa}(i)$ and there is no $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$ such that $x_i^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} \widehat{b}_{ki} x_k^{(t)} > \bigvee_{k\in\mathrm{pa}(i)} \widetilde{b}_{ki} x_k^{(t)}$. We know from Theorem 2.5.4(b) that $\widetilde{c}_{ki} = \widetilde{b}_{ki}$ for $k \in \mathrm{pa}^{\widetilde{B}}(i)$ and $\widetilde{c}_{ki} \in (0, \widetilde{b}_{ki}]$ for $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\widetilde{B}}(i)$, where $\mathrm{pa}^{\widetilde{B}}(i)$ are the parents of i in the minimum ML DAG $\mathcal{D}^{\widetilde{B}}$. With this, using Corollary 2.6.8 and (2.6.5), we obtain

$$\bigvee_{k\in\mathrm{pa}(i)} \widetilde{b}_{ki} X_k = \bigvee_{k\in\mathrm{pa}^{\widetilde{B}}(i)} \widetilde{b}_{ki} X_k = \bigvee_{k\in\mathrm{pa}^{\widetilde{B}}(i)} \widetilde{c}_{ki} X_k = \bigvee_{k\in\mathrm{pa}(i)} \widetilde{c}_{ki} X_k.$$

Consequently, for every i and $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$, $\bigvee_{k\in\mathrm{pa}(i)} \widetilde{b}_{ki} x_k^{(t)} = \bigvee_{k\in\mathrm{pa}(i)} \widetilde{c}_{ki} x_k^{(t)}$. So it remains to show that $\widetilde{b}_{ki} \le \widehat{b}_{ki}$ for every i and $k \in \mathrm{pa}(i)$ if $\widetilde{c}_{ki} \le \widehat{c}_{ki}$ for every $i \in V$ and $k \in \mathrm{pa}(i)$. We find in that case for the weight of a path $p = [k_0 = k \to k_1 \to \cdots \to k_{n-1} \to k_n = i]$,

$$\widetilde{d}_{ki}(p) = \widetilde{c}_{k_0k_1} \widetilde{c}_{k_1k_2} \ldots \widetilde{b}_{k_{n-1},k_n} \le \widehat{b}_{k_0k_1} \widehat{b}_{k_1k_2} \ldots \widehat{b}_{k_{n-1},k_n} \le \widehat{b}_{ki},$$

where we have used (4.4.7) for the last inequality. Thus, by definition of $\widetilde{b}_{ki}$ in (4.2.2), $\widetilde{b}_{ki} \le \widehat{b}_{ki}$ for $k \in \mathrm{pa}(i)$. □

In the following algorithm we use Corollary 4.4.14(a) to identify all GMLEs of B. Of course, we could alternatively use Corollary 4.4.14(b),(c) in step 2. To avoid to compute the same ML coefficient matrix several times, since different edge weights may lead to the same ML coefficient matrix, we use Theorem 2.5.4(b) in step 3.(b), (c).

Algorithm 4.4.15. [Find all GMLEs $\widetilde{B}$ of B from $\mathcal{D}$ and $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$]

1. For every $i \in V$ and $k \in \mathrm{pa}(i)$, compute $\widehat{b}_{ki} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_k^{(t)}}$.
2. Find all vectors $(\widetilde{c}_{ki}, i \in V, k \in \mathrm{pa}(i))$ such that $\widetilde{c}_{ki} \leq \widehat{b}_{ki}$ and for every $i \in V$, there is no $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$ with $x_i^{(t)} = \bigvee_{k \in \mathrm{pa}(i)} \widehat{b}_{ki} x_k^{(t)} > \bigvee_{k \in \mathrm{pa}(i)} \widetilde{c}_{ki} x_k^{(t)}$. Summarize them in the set $\mathcal{C}$.
3. For $\widetilde{\boldsymbol{c}} \in \mathcal{C}$,
   (a) compute via (4.2.2) the corresponding ML coefficient matrix $\widetilde{B}$;
   (b) find $\mathcal{D}^{\widetilde{B}}$ via (2.5.2);
   (c) remove those vectors $\widetilde{\boldsymbol{c}} = (\widetilde{c}_{ki}, i \in V, k \in \mathrm{pa}(i))$ from $\mathcal{C}$ that lead to $\widetilde{B}$; i.e., $\widetilde{c}_{ki} = \widetilde{b}_{ki}$ for $k \in \mathrm{pa}^{\widetilde{B}}(i)$ and $\widetilde{c}_{ki} \in (0, \widetilde{b}_{ki}]$ for $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\widetilde{B}}(i)$;
   (d) perform step 3. for the next vector $\widetilde{\boldsymbol{c}} \in \mathcal{C}$.

We would like to recall once again the matrix product $\odot$ from (2.2.2) with which $\widetilde{B}$ and the adjacency matrix of $\mathcal{D}^{\widetilde{B}}$ can be computed more efficiently (see Theorem 2.2.4 and Remark 2.3.11 as well as Theorem 4.2 of [76]).

We know the following from Example 4.4.2 but, of course, it is an immediate consequence of Theorem 4.4.12(e) as well.

Corollary 4.4.16. *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 recursive ML model on $\mathcal{D}_{ki} = (\{k, i\}, \{(k, i)\})$ with edge weight $c_{ki} = b_{ki}$.*

As $\widehat{B} \in \mathcal{B}$, a further immediate consequence of Theorem 4.4.12(e) is that $\widehat{B}$ is a GMLE of B. We use part (d) to provide a necessary and sufficient condition for its uniqueness.

Corollary 4.4.17. *The matrix $\widehat{B}$ is a GMLE of B. It is the only GMLE if and only if there is no $B \in \mathcal{B} \setminus \{\widehat{B}\}$ such that all $\boldsymbol{x}^{(t)} \in A_{1/2}(B, \widehat{B})$.*

In what follows we give an explanation why $\widehat{B}$ is the unique GMLE if n is sufficiently large. For this we first present a situation where $\widehat{B}$ is the unique GMLE.

Corollary 4.4.18. *We denote by $\mathrm{pa}^{\widehat{B}}(i)$ the parents of i in the minimum ML DAG $\mathcal{D}^{\widehat{B}}$ of a recursive ML model with ML coefficient matrix $\widehat{B}$. The matrix $\widehat{B}$ is the unique GMLE of B if for every $i \in V$ and $k \in \mathrm{pa}^{\widehat{B}}(i)$, $x_i^{(t)} = \widehat{b}_{ki} x_k^{(t)} > \bigvee_{\widetilde{k} \in \mathrm{pa}(i) \setminus \{k\}} \widehat{b}_{\widetilde{k}i} x_{\widetilde{k}}^{(t)}$ for some $\boldsymbol{x}_{\mathrm{Pa}(i)}^{(t)}$.*

Proof. Assume a further GMLE $\widetilde{B} \in \mathcal{B}$ of B. We observe from Theorem 4.4.12(e) that $\widetilde{b}_{ki} = \widehat{b}_{ki}$ for $k \in \mathrm{pa}^{\widehat{B}}(i)$. Since $\widehat{B}$ and $\widetilde{B}$ differ, $\widetilde{b}_{ki} < \widehat{b}_{ki}$ for some $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{B}(i)$ (see Remark 4.4.5). By definition of $\mathcal{D}^{\widehat{B}}$ there must exist a max-weighted path $p = [k_0 = k \to k_1 \to \cdots \to k_n = i]$ that is contained in $\mathcal{D}^{\widehat{B}}$ (cf. Remark 2.5.2(ii)). With this we find that $\widehat{b}_{ki} = \widehat{b}_{k_0 k_1} \widehat{b}_{k_1 k_2} \dots \widehat{b}_{k_{n-1}, k_n} = \widetilde{b}_{k_0 k_1} \widetilde{b}_{k_1 k_2} \dots \widetilde{b}_{k_{n-1}, k_n} > \widetilde{b}_{ki}$. The inequality is a contradiction to $\widetilde{B} \in \mathcal{B}$. Hence, $\widetilde{B} = \widehat{B}$, and $\widehat{B}$ is the only GMLE. □

The events corresponding to Corollary 4.4.18 have positive probability.

Lemma 4.4.19. *Let $i \in V$ and $k \in \mathrm{pa}^B(i)$, where $\mathrm{pa}^B(i)$ are the parents of i in $\mathcal{D}^B$. Then the event $A_{ki} = \{X_i = b_{ki}X_k > \bigvee_{\widetilde{k}\in \mathrm{pa}(i)\setminus\{k\}} b_{\widetilde{k}i}X_{\widetilde{k}}\}$ has positive probability.*

Proof. We first show that

$$\widetilde{A}_{ki} = \Big\{ \bigvee_{j\in\mathrm{an}(i): b_{ji} > \bigvee_{\widetilde{k}\in\mathrm{pa}(i)\setminus\{k\}} b_{j\widetilde{k}}b_{\widetilde{k}i}} b_{ji}Z_j > \bigvee_{j\in\mathrm{an}(i): b_{ji} = \bigvee_{\widetilde{k}\in\mathrm{pa}(i)\setminus\{k\}} b_{j\widetilde{k}}b_{\widetilde{k}i}} b_{ji}Z_j \vee Z_i \Big\} \subseteq A_{ki}.$$

Using that for $j \in \mathrm{an}(i)$, $b_{ji} = \bigvee_{k\in\mathrm{pa}(i)} b_{jk}b_{ki}$, which follows, for example, from Corollary 2.4.3(a) and Remark 2.2.3(i), it is not difficult to verify that

$$\widetilde{A}_{ki} \subseteq \Big\{ \bigvee_{j\in\mathrm{An}(i)} b_{ji}Z_j = b_{ki} \bigvee_{j\in\mathrm{An}(k)} b_{jk}Z_j \Big\} \cap \bigcap_{\widetilde{k}\in\mathrm{pa}(i)\setminus\{k\}} \Big\{ \bigvee_{j\in\mathrm{An}(i)} b_{ji}Z_j > b_{\widetilde{k}i} \bigvee_{j\in\mathrm{An}(\widetilde{k})} b_{j\widetilde{k}}Z_j \Big\},$$

where we set $\bigcap_{\widetilde{k}\in\emptyset} F = \Omega$ for $F \in \mathcal{F}$. This is by (4.2.3) a subevent of A_{ki} and, hence, $\widetilde{A}_{ki} \subseteq A_{ki}$. It remains to show that $\mathbb{P}(\widetilde{A}_{ki}) > 0$. Since the noise variables are independent and have support $\mathbb{R}_+$, it suffices to show that $b_{ji} > \bigvee_{\widetilde{k}\in\mathrm{pa}(i)\setminus\{k\}} b_{j\widetilde{k}}b_{\widetilde{k}i}$ for some $j \in \mathrm{an}(i)$. Observing from (2.5.2) that k is such a node j finishes the proof. □

For n sufficiently large, we observe A_{ki} for every $i \in V$ and $k \in \mathrm{pa}^B(i)$. Then the estimate $\widehat{b}_{ki}$ for $k \in \mathrm{pa}^B(i)$ is equal to the true parameter b_{ki}. However, we do not know which edges are in $\mathcal{D}^B$ so that it may be not enough to observe the events A_{ki} from Lemma 4.4.19 only to estimate B exactly. We illustrate this by an example.

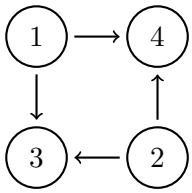
Example 4.4.20. [Observing the events from Lemma 4.4.19 is not enough]
Consider the DAG $\mathcal{D}$ from Example 4.3.1, and assume for the true ML coefficients that $b_{13} = b_{12}b_{23}$; the minimum ML DAG $\mathcal{D}^B$ of $\boldsymbol{X}$ is then the DAG $\mathcal{D}$ without the edge $1 \to 3$ (see Example 4.3.1). Furthermore, we assume that we have two observations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ only: $\boldsymbol{x}^{(1)}$ belongs to the event $\{X_2 = b_{12}X_1\} \cap \{X_3 > b_{23}X_2\} \cap \{X_3 > b_{13}X_1\}$ and $\boldsymbol{x}^{(2)}$ to $\{X_2 > b_{12}X_1\} \cap \{X_3 = b_{23}X_2\} \cap \{X_3 > b_{13}X_1\}$. Both events occur with positive probability. Then we estimate b_{12} and b_{23} by $\widehat{b}_{12}$ and $\widehat{b}_{23}$ exactly but would overstimate b_{13} with $\widehat{b}_{13}$. If we know $\mathcal{D}^B$, which is usually not the case, then we would estimate b_{13} by $\widehat{b}_{12}\widehat{b}_{23}$ and, hence, exactly. Assuming that $\frac{x_3^{(1)}}{x_1^{(1)}} > \frac{x_3^{(2)}}{x_1^{(2)}} = \widehat{b}_{13}$, we learn from Theorem 4.4.12(e) that the vectors $(\widehat{b}_{12}, \widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23}]$ and $(\widehat{b}_{12}, \widetilde{b}_{13}, \widehat{b}_{23})$ such that $\widehat{b}_{13} \geq \widetilde{b}_{13} \geq \widehat{b}_{12}\widehat{b}_{23}$ are the GMLEs of (b_{12}, b_{13}, b_{23}) as $\widehat{b}_{23} = \frac{x_3^{(2)}}{x_2^{(2)}} < \frac{x_3^{(1)}}{x_2^{(1)}}$. If $\widehat{b}_{13} = \frac{x_3^{(1)}}{x_1^{(1)}} < \frac{x_3^{(2)}}{x_1^{(2)}}$, then $(\widehat{b}_{12}, \widehat{b}_{13}, \widehat{b}_{23})$ is the only GMLE (cf. Corollary 4.4.18), and if $\widehat{b}_{13} = \frac{x_3^{(1)}}{x_1^{(1)}} = \frac{x_3^{(2)}}{x_1^{(2)}}$, then the vectors $(\widehat{b}_{12}, \widehat{b}_{13}, \widetilde{b}_{23})$ with $\widetilde{b}_{23} \in (0, \widehat{b}_{23}]$ are the GMLEs. Note, however, that we observe $\widehat{b}_{13} = \frac{x_3^{(1)}}{x_1^{(1)}} = \frac{x_3^{(2)}}{x_1^{(2)}}$ with probability zero, since $\frac{X_3}{X_1}$ has no atom in $\widehat{b}_{13}$ (see Table 4.1). □

If every $\widehat{b}_{ki}$ for $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^B(i)$ equals the true ML coefficient b_{ki} and the events A_{ki} from Lemma 4.4.19 are observed implicitly, then Corollary 4.4.18 yields that $\widehat{B}$ is the unique GMLE; furthermore, B is estimated exactly by $\widehat{B}$. These conditions apply in particular if n is sufficiently large. Thus we get the following corollary.

Corollary 4.4.21. *For n sufficiently large, $\widehat{B}$ is equal to the true B and is the unique GMLE of B.*

Recursive max-weighted models

Suppose we have the information that the model $\boldsymbol{X}$ underlying $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ is max-weighted; i.e., all paths are max-weighted. This subclass of recursive ML mdels was introduced and discussed in Chapter 3. For example, if $\mathcal{D}$ is a polytree (i.e., the underlying undirected graph has no cycles) or has at most one path between two nodes, then a recursive ML model on $\mathcal{D}$ is automatically max-weighted; the latter request to $\mathcal{D}$ is weaker than the first as the following DAG shows.



We observe from Example 4.4.20 that, when determing GMLEs of B, it makes sense to use the additional knowledge that $\boldsymbol{X}$ is max-weighted. This can be achieved by the use of a density specific to recursive max-weighted models. In what follows we discuss such a density and the corresponding GMLEs of B.

Analogous to Remark 4.4.5, we first present necessary and sufficient conditions for a matrix to be the ML coefficient matrix of a recursive max-weighted model on the given DAG $\mathcal{D}$.

Remark 4.4.22. Let $\mathcal{D}^{\mathrm{tr}}$ be the transitive reduction of $\mathcal{D}$. We denote by $\mathcal{B}_{\mathrm{mw}}$ the class of the ML coefficient matrices of all recursive max-weighted models on $\mathcal{D}$.

(i) $B = (b_{ij})_{d\times d} \in \mathcal{B}_{\mathrm{mw}}$ if and only if for every $i \in V$,

$$b_{ii} = 1,\ b_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i),\ b_{ki} > 0 \text{ for } k \in \mathrm{pa}^{\mathrm{tr}}(i),\ \text{and}$$
$$b_{ji} = d_{ji}(p) \text{ for } j \in \mathrm{an}(i) \text{ and every } p \in P_{ji},$$

where $d_{ji}(p) = \prod_{\nu=0}^{n-1} b_{k_\nu k_{\nu+1}}$ for a path $p = [j = k_0 \to k_1 \to \cdots \to k_n = i]$.

(ii) $B \in \mathcal{B}_{\mathrm{mw}}$ if and only if for every $i \in V$,

$$b_{ii} = 1,\ b_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i),\ b_{ki} > 0 \text{ for } k \in \mathrm{pa}^{\mathrm{tr}}(i),\ \text{and}$$
$$b_{ji} = b_{jk} b_{ki} \text{ for } k \in \mathrm{pa}(i) \text{ and } j \in \mathrm{an}(k).$$

(iii) Coefficients $(b_{ki}, i \in V, k \in \mathrm{pa}(i))$ are entries of a matrix $B \in \mathcal{B}_{\mathrm{mw}}$ if and only if for every $i \in V$, $b_{ki} > 0$ for $k \in \mathrm{pa}^{\mathrm{tr}}(i)$ and $b_{ki} = d_{ki}(p)$ for $k \in \mathrm{pa}(i) \setminus \mathrm{pa}^{\mathrm{tr}}(i)$ and every path $p \in P_{ki}$. In this case, the remaining ML coefficients are uniquely given for $i \in V$ by

$$b_{ii} = 1,\ b_{ji} = 0 \text{ for } j \in V \setminus \mathrm{An}(i),\ \text{and } b_{ji} = d_{ji}(p) \text{ for } j \in \mathrm{an}(i) \setminus \mathrm{pa}(i) \text{ and some } p \in P_{ji}.$$

□

According to Remark 4.4.22(ii), we have for every $i \in V$, $k \in \mathrm{pa}(i)$, and $j \in \mathrm{An}(k)$, $b_{ji} = b_{jk}b_{ki}$. Thus we find from Table 4.1

$$\mathbb{P}(X_i = b_{ki}X_k) > 0 \quad \text{and} \quad \mathbb{P}(X_i = xX_k) = 0 \text{ for } x \in \mathbb{R}_+ \setminus \{b_{ki}\}; \tag{4.4.16}$$

i.e., the distribution of $\frac{X_i}{X_k}$ has only one atom in b_{ki}.

Let now $B, B^* \in \mathcal{B}_{\mathrm{mw}}$. We define

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

$\big\{A_0^{\mathrm{mw}}(B,B^*), A_{1/2}^{\mathrm{mw}}(B,B^*), A_1^{\mathrm{mw}}(B,B^*)\big\}$ is a partition of $\mathbb{R}_+^d$. These sets define a density of P_B with respect to $P_B + P_{B^*}$, for which we obtain similar representations as in Proposition 4.4.9(a) for ρ.

Corollary 4.4.23. *Let $B, B^* \in \mathcal{B}_{mw}$.*

(a) The function from $\mathbb{R}_+^d$ to $\{0, 1/2, 1\}$ such that

$$\boldsymbol{x} \mapsto \rho_{mw}(\boldsymbol{x}, B, B^*) = \frac{1}{2}\cdot \mathbb{1}_{A_{1/2}^{mw}(B,B^*)}(\boldsymbol{x}) + \mathbb{1}_{A_1^{mw}(B,B^*)}(\boldsymbol{x}) \tag{4.4.17}$$

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

For $i \in V$, let $\rho_{mw,i}$ be the density given in (4.4.17) with respect to the DAG $\mathcal{D}_i$ from Proposition 4.4.9; we also use the notation B_i and B_i^* introduced there. Furthermore, for $i \in V$ and $k \in \mathrm{pa}(i)$, let $\rho_{k\to i}$ the density from (4.4.17) with respect to the DAG $\mathcal{D}_{ki} = (\{k,i\}, \{(k,i)\})$ as well as B_{ki} and B_{ki}^* the ML coefficient matrices of recursive ML models on $\mathcal{D}_{ki}$ with edge weight $c_{ki} = b_{ki}$ and $c_{ki}^* = b_{ki}^*$, respectively.

(b) Writing $\boldsymbol{x}_{ki}$ for (x_k, x_i), we have for $\boldsymbol{x} \in \mathbb{R}_+^d$,

$$\begin{aligned}
\rho_{k\to i}(\boldsymbol{x}_{ki}, B_{ki}, B_{ki}^*) &= \min_{\boldsymbol{y}\in\{\boldsymbol{y}\in\mathbb{R}_+^d : \boldsymbol{y}_{ki}=\boldsymbol{x}_{ki}, \rho_{mw}(\boldsymbol{y},B,B^*)>0\}} \rho_{mw}(\boldsymbol{y}, B, B^*)\\
&= \min_{\boldsymbol{y}_{\mathrm{Pa}(i)}\in\{\boldsymbol{y}_{\mathrm{Pa}(i)}\in\mathbb{R}_+^{|\mathrm{Pa}(i)|} : \boldsymbol{y}_{ki}=\boldsymbol{x}_{ki}, \rho_{mw,i}(\boldsymbol{y}_{\mathrm{Pa}(i)},B_i,B_i^*)>0\}} \rho_{mw,i}(\boldsymbol{y}_{\mathrm{Pa}(i)}, B_i, B_i^*),\\
\rho_{mw,i}(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) &= \min_{\boldsymbol{y}\in\{\boldsymbol{y}\in\mathbb{R}_+^d : \boldsymbol{y}_{\mathrm{Pa}(i)}=\boldsymbol{x}_{\mathrm{Pa}(i)}, \rho_{mw}(\boldsymbol{y},B,B^*)>0\}} \rho_{mw}(\boldsymbol{y}, B, B^*)\\
&= \Big(\bigvee_{k\in\mathrm{pa}(i)} \rho_{k\to i}(\boldsymbol{x}_{ki}, B_{ki}, B_{ki}^*) \vee \frac{1}{2}\Big)\mathbb{1}\Big\{\bigwedge_{k\in\mathrm{pa}(i)} \rho_{k\to i}(\boldsymbol{x}_{ki}, B_{ki}, B_{ki}^*) > 0\Big\},\\
\rho_{mw}(\boldsymbol{x}, B, B^*) &= \Big(\bigvee_{i\in V}\bigvee_{k\in\mathrm{pa}(i)} \rho_{k\to i}(\boldsymbol{x}_{ki}, B_{ki}, B_{ki}^*) \vee \frac{1}{2}\Big)\mathbb{1}\Big\{\bigwedge_{i\in V}\bigwedge_{k\in\mathrm{pa}(i)} \rho_{k\to i}(\boldsymbol{x}_{ki}, B_{ki}, B_{ki}^*) > 0\Big\}
\end{aligned}$$

$$= \Big(\bigvee_{i\in V} \rho_{mw,i}(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) \Big) \mathbb{1}\Big\{ \bigwedge_{i\in V} \rho_{mw,i}(\boldsymbol{x}_{\mathrm{Pa}(i)}, B_i, B_i^*) > 0 \Big\}.$$

Proof. (a) We verify

(A) $P_B(A_0^{\mathrm{mw}}(B, B^*)) = 0$,

(B) $P_B(A \cap A_{1/2}^{\mathrm{mw}}(B, B^*)) = P_{B^*}(A \cap A_{1/2}^{\mathrm{mw}}(B, B^*))$ for every $A \in \mathcal{B}(\mathbb{R}_+^d)$, and

(C) $P_{B^*}(A_1^{\mathrm{mw}}(B, B^*)) = 0$

(cf. the discussion related to (4.4.3)).
(A) As $\bigcup_{i\in V} \big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i < \bigvee_{k\in \mathrm{pa}(i)} b_{ki} x_k\big\} = \bigcup_{i\in V} \bigcup_{k\in \mathrm{pa}(i)} \big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i < b_{ki} x_k\big\}$, it suffices by (4.4.14) and the definition of $A_0^{\mathrm{mw}}(B, B^*)$ to show that $\big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = b_{ki}^* x_k > b_{ki} x_k\big\}$ is a P_B-null set for $i \in V$ and $k \in \mathrm{pa}(i)$. This is immediate by (4.4.16).
(B) In Theorem 4.4.7 we have proved that (B) holds with $A_{1/2}^{\mathrm{mw}}(B, B^*)$ replaced by $A_{1/2}(B, B^*)$. As $A_{1/2}^{\mathrm{mw}}(B, B^*) \subseteq A_{1/2}(B, B^*)$, (B) follows.
(C) is a consequence of (A) as

$$\begin{aligned} A_1^{\mathrm{mw}}(B, B^*) &\subseteq \bigcup_{i\in V} \bigcup_{k\in \mathrm{pa}(i)} \Big[\big\{\boldsymbol{x} \in \mathbb{R}_+^d : b_{ki}^* x_k > x_i \geq b_{ki} x_k\big\} \cup \big\{\boldsymbol{x} \in \mathbb{R}_+^d : x_i = b_{ki} x_k > b_{ki}^* x_k\big\}\Big] \\ &\subseteq A_0^{\mathrm{mw}}(B^*, B). \end{aligned}$$

(b) can be observed similarly as Proposition 4.4.9(a), (b). □

Since the sets defining ρ_{mw} are simpler as the ones defining ρ, we would prefer ρ_{mw}. However, it is no density for general recursive ML models as we demonstrate by an example.

Example 4.4.24. [ρ_{mw} is no density for non-max-weighted models]
Consider the DAG $\mathcal{D}$ from Example 4.3.1, and let $B, B^* \in \mathcal{B}$ such that $b_{13} > b_{12} b_{23}$, $b_{12} = b_{12}^*$, $b_{13} = b_{13}^*$, and $b_{23}^* = \frac{b_{13}}{b_{12}}$. Furthermore, define $A_0^{23} = \big\{\boldsymbol{x} \in \mathbb{R}_+^3 : x_3 = b_{23}^* x_2\big\}$. We learn from Table 4.1 that $P_B(A_0^{23}) > 0$. As $A_0^{23} \subseteq A_0^{\mathrm{mw}}(B, B^*)$, we then obtain

$$\int_{A_0^{23}} \rho_{\mathrm{mw}}(\boldsymbol{x}, B, B^*)(P_B + P_{B^*})(d\boldsymbol{x}) = 0 < P_B\big(A_0^{23}\big).$$

This shows that $\boldsymbol{x} \mapsto \rho_{\mathrm{mw}}(\boldsymbol{x}, B, B^*)$ is no density of P_B with respect to $P_B + P_{B^*}$. □

Before we characterize the GMLEs of B with respect to ρ_{mw} given in (4.4.17), we summarize the most important properties needed for this. Similarly as in Lemma 4.4.11, they follow directly from the definition of the sets $A_0^{\mathrm{mw}}(B, B^*)$, $A_{1/2}^{\mathrm{mw}}(B, B^*)$, $A_1^{\mathrm{mw}}(B, B^*)$.

Lemma 4.4.25. *For $i \in V$ and $k \in \mathrm{pa}(i)$, set $\widehat{b}_{ki} = \bigwedge_{t=1}^n \frac{x_i^{(t)}}{x_k^{(t)}}$.*

(a) No $\boldsymbol{x}^{(t)} \in A_0^{mw}(B, B)$ if and only if all $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B, B)$.

(b) $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B, B^*)$ if and only if $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B^*, B)$.

(c) If $\boldsymbol{x}^{(t)} \in A_1^{mw}(B, B^*)$, then $\boldsymbol{x}^{(t)} \in A_0^{mw}(B^*, B)$.

(d) All $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B,B)$ *if and only if for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$, $b_{ki} \leq \widehat{b}_{ki}$.

(e) If all $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B,B)$, *then no* $\boldsymbol{x}^{(t)} \in A_0^{mw}(B^*,B)$ *if and only if for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$, $b_{ki}^* \leq \widehat{b}_{ki}$ *with equality if* $b_{ki} = \widehat{b}_{ki}$.

(f) If all $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(B,B)$, *then* $\boldsymbol{x}^{(t)} \in A_0^{mw}(B,B^*)$ *if and only if* $x_i^{(t)} = b_{ki}^* x_k^{(t)}$ *and* $b_{ki}^* > b_{ki}$ *for some* $i \in V$ *and* $k \in \mathrm{pa}(i)$.

Proposition 4.4.26. *Let* $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ *be independent realizations of a recursive max-weighted model* $\boldsymbol{X}$ *on a given DAG* $\mathcal{D}$ *with ML coefficient matrix* B. *Let* $\widehat{B}$ *be the matrix from* (4.4.8). *Then* $\widetilde{B} \in \mathcal{B}_{mw}$ *is a GMLE of* B *if and only if one of the following conditions is satisfied.*

(a) All $\boldsymbol{x}^{(t)} \in A_{1/2}^{mw}(\widetilde{B},\widetilde{B})$ *and for all* $B \in \mathcal{B}_{mw}$, *if some* $\boldsymbol{x}^{(t)} \in A_0^{mw}(\widetilde{B},B)$, *then some* $\boldsymbol{x}^{(s)} \in A_0^{mw}(B,\widetilde{B})$.

(b) $\widetilde{b}_{ki} \leq \widehat{b}_{ki}$ *for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$, *and if* $\widetilde{b}_{ki} < \widehat{b}_{ki}$ *for some* $i \in V$ *and* $k \in \mathrm{pa}(i)$, *then there is no* $B \in \mathcal{B}_{mw}$ *such that* $b_{ki} \leq \widehat{b}_{ki}$ *for every* $i \in V$ *and* $k \in \mathrm{pa}(i)$ *with equality if* $\widetilde{b}_{ki} = \widehat{b}_{ki}$ *and for at least one edge* $k \to i$ *with* $\widetilde{b}_{ki} < \widehat{b}_{ki}$.

Proof. (a) follows from the definition of ρ_{mw} and Lemma 4.4.25(a)–(c).
(b) By Lemma 4.4.25(d) and (a), $\widetilde{B} \in \mathcal{B}_{\mathrm{mw}}$ is a GMLE of B if and only if $\widetilde{b}_{ki} \leq \widehat{b}_{ki}$ for every i and $k \in \mathrm{pa}(i)$ and there is no $B \in \mathcal{B}_{\mathrm{mw}}$ such that some $\boldsymbol{x}^{(t)} \in A_0^{\mathrm{mw}}(\widetilde{B},B)$ but no $\boldsymbol{x}^{(t)} \in A_0^{\mathrm{mw}}(B,\widetilde{B})$. We finally observe from Lemma 4.4.25(e), (f) that this is equivalent to (b), which finishes the proof. □

The matrix $\widehat{B}$ is not necessarily in $\mathcal{B}_{\mathrm{mw}}$. Consider, for example, DAG $\mathcal{D}$ from Example 4.3.1: $\widehat{B} \notin \mathcal{B}_{\mathrm{mw}}$ if $\widehat{b}_{13} > \widehat{b}_{12}\widehat{b}_{23}$. We have observed this situation in Example 4.4.20. A further example can be found in Example 4.4.28 below. Of course, $\widehat{B}$ is no GMLE and cannot be the true B if $B \in \mathcal{B}_{\mathrm{mw}}$ but $\widehat{B} \notin \mathcal{B}_{\mathrm{mw}}$; at least one ML coefficient b_{ki} with $k \in \mathrm{pa}(i)$ must then have been overestimated. Next, we investigate the case where $\widehat{B} \in \mathcal{B}_{\mathrm{mw}}$. To guarantee this we could, for example, assume that $\mathcal{D}$ is a polytree or has at most one path between two nodes.

Corollary 4.4.27. *Assume the situation of Proposition 4.4.26.*

(a) If $\widehat{B} \in \mathcal{B}_{mw}$, *it is the only GMLE of* B.

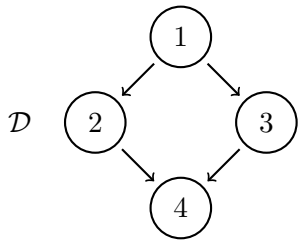
(b) For every $i \in V$, *the vector* $(\widehat{b}_{ki}, k \in \mathrm{pa}(i))$ *is the only GMLE of the ML coefficients* $(b_{ki}, k \in \mathrm{pa}(i))$ *of a recursive ML model* $\boldsymbol{Y}_i$ *on* $\mathcal{D}_i$ *with edge weights* $c_{ki} = b_{ki}$.

(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 recursive ML model on* $\mathcal{D}_{ki} = (\{k,i\}, \{(k,i)\})$ *with edge weight* $c_{ki} = b_{ki}$.

Proof. (a) By Proposition 4.4.26(b) $\widehat{B} \in \mathcal{B}_{\mathrm{mw}}$ is a GMLE of B. It is unique, since, otherwise, $\widehat{B}$ is such a matrix B that cannot exist according to Proposition 4.4.26(b).
(b) Let $\big(y_\ell^{(t)}, \ell \in \mathrm{Pa}(i)\big)$, $t = 1, \ldots, n$, be independent realizations of $\boldsymbol{Y}_i$. Since $\boldsymbol{Y}_i$ is max-weighted, we know from (a) that the vector $\big(\bigwedge_{t=1}^n \frac{y_i^{(t)}}{y_k^{(t)}}, k \in \mathrm{pa}(i)\big)$ is the only GMLE of $(b_{ki}, k \in \mathrm{pa}(i))$.
(c) follows from Example 4.4.2 or analogoulsy to (b) from (a). □

The densities $\rho_{k\to i}$ and $\rho_{\mathrm{mw},i}$ have a similar meaning for ρ_{mw} as ρ_i for ρ (cf. Proposition 4.4.9(a), (b) and Corollary 4.4.23(b)). Therefore, in the situation of Proposition 4.4.26, one could expect a similar result as Theorem 4.4.12(f). That this is not the case can be observed from Corollary 4.4.27(b), (c) and the next example.

Example 4.4.28. [If $\widehat{B} \notin \mathcal{B}_{\mathrm{mw}}$, then several GMLEs may exist]
Consider the DAG



and assume that $\boldsymbol{X}$ is max-weighted, which is equivalent to $b_{12}b_{24} = b_{13}b_{34}$. Furthermore, assume that $\widehat{b}_{12}\widehat{b}_{24} < \widehat{b}_{13}\widehat{b}_{34}$. Of course, we would rather not oberve this if n is sufficiently large. But this can happen if, for example, only the events $F_1 = \{X_2 = b_{12}X_1\} \cap \{X_3 > b_{13}X_1\} \cap \{X_4 > b_{24}X_2\} \cap \{X_4 > b_{34}X_3\}$ and $F_2 = \{X_2 > b_{12}X_1\} \cap \{X_3 > b_{13}X_1\} \cap \{X_4 = b_{24}X_2\} \cap \{X_4 > b_{34}X_3\}$ have occured. It is known that the model is max-weighted. So it is obvious to use ρ_{mw} for finding GMLEs. Then by Proposition 4.4.26(b) both the ML coefficient matrix corresponding to the edge weights $(\widehat{b}_{12}, \widetilde{b}_{13}, \widehat{b}_{24}, \widehat{b}_{34})$ such that $\widehat{b}_{12}\widehat{b}_{24} = \widetilde{b}_{13}\widehat{b}_{34}$ and the one corresponding to the edge weights $(\widehat{b}_{12}, \widehat{b}_{13}, \widehat{b}_{24}, \widetilde{b}_{34})$ such that $\widehat{b}_{12}\widehat{b}_{24} = \widehat{b}_{13}\widetilde{b}_{34}$ is a GMLE of B. This is because $\widehat{B} \notin \mathcal{B}_{\mathrm{mw}}$; otherwise, these matrices would not be GMLEs of B (cf. Proposition 4.4.26(b) and Corollary 4.4.26(a)). The ML coefficent matrix of a recursive max-weighted model on $\mathcal{D}$ with edge weights $(\widehat{b}_{12}, \widetilde{b}_{13}, \widehat{b}_{24}, \widetilde{b}_{34})$ such that $\widetilde{b}_{13} \le \widehat{b}_{13}$, $\widetilde{b}_{34} \le \widehat{b}_{34}$, and $\widehat{b}_{12}\widehat{b}_{24} = \widetilde{b}_{13}\widetilde{b}_{34}$ is no GMLE of B, since setting $b_{12} = \widehat{b}_{12}$, $b_{13} = \widehat{b}_{13}$, $b_{24} = \widehat{b}_{24}$, and $b_{34} = \frac{b_{12}b_{24}}{b_{34}}$ leads to a matrix $B \in \mathcal{B}_{\mathrm{mw}}$ that, according to Proposition 4.4.26(b), cannot exist if this would be a GMLE. In this way we can show that the two matrices listed above are the only GMLEs of B. One possible explanation why we find exactly these GMLES with ρ_{mw} is as follows (cf. Example 4.4.3). Since we have assumed observations underlying a recursive max-weighted model but observed that $\widehat{b}_{12}\widehat{b}_{24} < \widehat{b}_{13}\widehat{b}_{34}$, $\widehat{b}_{13}$ or $\widehat{b}_{23}$ cannot be the true value. Assuming that b_{12}, b_{24}, b_{34} are estimated exactly by $\widehat{b}_{12}, \widehat{b}_{24}, \widehat{b}_{34}$, $(\widehat{b}_{12}, \widetilde{b}_{13}, \widehat{b}_{24}, \widehat{b}_{34})$ with $\widehat{b}_{12}\widehat{b}_{24} = \widetilde{b}_{13}\widehat{b}_{34}$ is the only reasonable estimate of $(b_{12}, b_{24}, b_{13}, b_{34})$; similarly, assuming that $\widehat{b}_{12}, \widehat{b}_{13}, \widehat{b}_{24}$ are the true values, $(\widehat{b}_{12}, \widehat{b}_{13}, \widehat{b}_{24}, \widetilde{b}_{34})$ with $\widehat{b}_{12}\widehat{b}_{24} = \widehat{b}_{13}\widetilde{b}_{34}$ is the only sensible estimate. When b_{13}, b_{24}, b_{34} are estimated exactly by $\widehat{b}_{13}, \widehat{b}_{24}, \widehat{b}_{34}$, it makes no sense to estimate b_{12} by a value smaller than $\widehat{b}_{12}$, since these estimates do not belong to a recursive max-weighted model; the same applies to b_{24}. If we observe the events F_1 and F_2 both more than once, then we can expect that b_{12} and b_{24} are estimated exactly, since $\frac{X_2}{X_1}$ has only an atom in b_{12} and $\frac{X_4}{X_2}$ in b_{24} (cf. Table 4.1). □

If it is known that the realizations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ are generated by a recurisve max-weighted model, we can use both ρ and ρ_{mw} to find GMLEs of B. We show that parts (a), (b) of Corollary 4.4.27 are not true when replacing ρ_{mw} by ρ. Part (c) holds (see Corollary 4.4.16).

Example 4.4.29. [Continuation of Example 4.4.3: ρ versus ρ_{mw}]
Note that $\boldsymbol{X}$ is max-weighted. So we can consider GMLEs with respect to ρ and ρ_{mw}. By Theorem 4.4.12(e) the GMLEs of (b_{13}, b_{23}) with respect to ρ from (4.4.13) are

$$(\widehat{b}_{13}, \widehat{b}_{23}),$$

$$(\widetilde{b}_{13}, \widehat{b}_{23}) \text{ with } \widetilde{b}_{13} \in (0, \widehat{b}_{13}) \text{ if there is no } \boldsymbol{x}^{(t)} \text{ such that } \widehat{b}_{13} = \frac{x_3^{(t)}}{x_1^{(t)}} \text{ and } \frac{x_3^{(t)}}{x_2^{(t)}} > \widehat{b}_{23}, \text{ and}$$

$$(\widehat{b}_{13}, \widetilde{b}_{23}) \text{ with } \widetilde{b}_{23} \in (0, \widehat{b}_{23}) \text{ if there is no } \boldsymbol{x}^{(t)} \text{ such that } \frac{x_3^{(t)}}{x_1^{(t)}} > \widehat{b}_{13} \text{ and } \frac{x_3^{(t)}}{x_2^{(t)}} = \widehat{b}_{23}.$$

As $\widehat{B} \in \mathcal{B}_{\mathrm{mw}}$, $\widehat{B}$ is the only GMLE of B with respect to ρ_{mw} (see Corollary 4.4.27(a)).

The density ρ_{mw} equals ρ_1 from Figure 4.4.4 and ρ equals ρ_5. Thus we have verified the results presented in Table 4.2 for these two densities. □

Edge weights c_{ki}

In what follows we do not assume anymore that the observations $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ explicitly underlie a recursive max-weighted model but an arbitrary recursive ML model. We have started with the estimation of B as it is not possible to recover the true edge weights c_{ki} underlying representation (4.2.1) 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 (see e.g. Figure 4.3.1), and, obviously, the probability measure 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 a GMLE $\widetilde{B}$ of B are GMLEs of the true edge weights of $\boldsymbol{X}$. Since we consider $\widehat{B}$ to be the best possible estimate of B (see also Section 4.4.2 below) and it is the unique GMLE of B if n is sufficiently large (see Corollary 4.4.21), we formulate the following corollary. However, it remains valid if we replace $\widehat{B}$ by some $\widetilde{B} \in \mathcal{B}$ that satisfies one of the conditions of Theorem 4.4.12.

Corollary 4.4.30. *Assume the situation of Theorem 4.4.12. Let c_{ki} for $i \in V$ and $k \in \mathrm{pa}(i)$ be the edge weights of representation (4.2.1) 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*

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

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

If $\mathcal{D}$ is a polytree or has at most one path between two nodes, then $\mathcal{D}$, $\mathcal{D}^B$, and $\mathcal{D}^{\widehat{B}}$ are the same, since every edge $k \to i$ is the only and, hence, max-weighted path from k to i (cf. Remark 2.5.2(ii)). In that case, $\boldsymbol{X}$ is max-weighted, $\widehat{B} \in \mathcal{B}_{\mathrm{mw}}$, and the edge weights of $\boldsymbol{X}$ are unique. Therefore, by Corollary 4.4.27(a), with respect to ρ_{mw}, the vector $\left(\bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_k^{(t)}}, i \in V, k \in \mathrm{pa}(i) \right)$ is the unique GMLE of the edge weights $(c_{ki}, i \in V, k \in \mathrm{pa}(i))$.

To find all GMLEs of the edge weights and not only those that correspond to $\widehat{B}$, it is not necessary to determine all GMLEs of B before. In fact, we have characterized the GMLEs

of the edge weights in Corollary 4.4.14: a vector $(\widetilde{c}_{ki}, i \in V, k \in \mathrm{pa}(i)) \in \mathbb{R}_+^{|E|}$ is a GMLE of $(c_{ki}, i \in V, k \in \mathrm{pa}(i))$ if and only if it satisfies one of the properties of Corollary 4.4.14. Performing steps 1. and 2. of Algorithm 4.4.15, we obtain all these vectors from $\mathcal{D}$ and $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$; more precisely, the set $\mathcal{C}$ from step 2. contains all GMLEs of $(c_{ki}, i \in V, k \in \mathrm{pa}(i))$.

Distribution functions F_{Z_i} of the noise variables

Algorithm 4.3.3 provides an iterative procedure to obtain the distribution functions F_{Z_i} from B and the marginal distribution functions G_i of the variables X_i. Estimating B by $\widehat{B}$ and the distributions G_i, for example, by their empirical versions, we can apply this procedure to find an estimator of the distributions F_{Z_i}. Often, it is more efficient to estimate G_i parametrically. Under the assumption that the noise variables Z_i are regularly varying with the same index, we have computed the distributions G_i explicitly in Proposition 3.A.2. Besides B, we then would also have to estimate the index of regular variation.

4.4.2 An almost perfect estimate of B

As already indicated in the previous section, we would usually choose the matrix $\widehat{B}$ from (4.4.8) as an estimate of B. To clarify again why, we summarize and add properties of this estimate.

In the extended definition of a MLE introduced by Kiefer-Wolfowitz, $\widehat{B}$ can be considered a MLE (Corollary 4.4.17). By its definition it never underestimates a ML coefficient and identifies by Remark 4.4.5(ii) the true ML coefficient matrix exactly if and only if it identifies all b_{ki} for $k \in \mathrm{pa}(i)$ exactly. Since by Table 4.1 $\mathbb{P}(X_i = b_{ki}X_k) > 0$ 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 [13] this is discussed for the time-series framework used there.

The following two examples show how effective the estimate $\widehat{B}$ can be; in particular, n does not necessarily need to be large. The second example is a conclusion of Example 4.4.3.

Example 4.4.31. [Continuation of Example 4.4.28: one observation may be enough to estimate B exactly]
If we observe the event

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

then 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 4.4.32. [Continuation of Example 4.4.3: $\widehat{B}$ is the perfect estimate of B]
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 the events F_2 and F_3; otherwise, it would be enough to observe F_1. □

Obviously, the larger n the higher the probability that the event $\{X_i = b_{ki}X_k\}$ for $k \in \mathrm{pa}(i)$ is included in $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and, hence, the minimal observed ratio $\widehat{b}_{ki} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_k^{(t)}}$ of $\frac{X_i}{X_k}$ equals

the true b_{ki}. 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 $1-p$. We also prove that the probability for estimating the true b_{ki} converges geometrically fast to 1.

Proposition 4.4.33. *Let $\boldsymbol{X}^{(t)} = \big(X_1^{(t)}, \dots, X_n^{(t)}\big)$ for $t = 1, \dots, n$ be independent copies of a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D}$ with ML coefficient matrix B. Let $i \in V$ and $k \in \mathrm{pa}(i)$.*

(a) We have $\mathbb{P}\big(\frac{X_i}{X_k} = b_{ki}\big), \mathbb{P}\big(\frac{X_i}{X_k} > b_{ki}\big) \in (0,1)$.

(b) $\mathbb{P}\big(\bigwedge_{t=1}^n \frac{X_i^{(t)}}{X_k^{(t)}} = b_{ki}\big) \geq 1-p$ for some $p \in (0,1)$ if and only if $n \geq \frac{\ln(p)}{\ln(1-\mathbb{P}(\frac{X_i}{X_k}=b_{ki}))} = \frac{\ln(p)}{\ln(\mathbb{P}(\frac{X_i}{X_k}>b_{ki}))}$.

(c) The convergence $\mathbb{P}\big(\bigwedge_{t=1}^n \frac{X_i^{(t)}}{X_k^{(t)}} = b_{ki}\big) \to 1$ as $n \to \infty$ is geometrically fast.

Proof. First, recall, for example, from (4.4.9) or Corollary 2.3.13 that the events $\{X_i = b_{ki}X_k\}$ and $\{X_i > b_{ki}X_k\}$ are complementary. With this and the same argumentation we have used to verify that $\mathbb{P}\big(\frac{X_i}{X_k} = b_{ki}\big) > 0$ below of (4.3.1), we obtain that $\mathbb{P}\big(X_i > b_{ki}X_k\big) > 0$; cf. also Corollary 4.A.5(b) in Appendix 4.A.2, where we examine such events in more detail. Hence, (a) holds. Using that $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)}$ are independent and identically distributed yields

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

From this and the above properties, we then observe (b) and (c). □

In conclusion, $\widehat{B}$ has the nice property to be 'geometrically' consistent.

4.5 Structure learning of a recursive ML model

Contrary to the assumptions in the previous section, we now assume that independent realizations $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ of a recursive ML model $\boldsymbol{X}$ are given 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 Theorem 4.3.4, the ML coefficient matrix B and from this the distribution of the noise vector as well as all DAGs and edge weights that could have generated $\boldsymbol{X}$ via (4.2.1). Our first goal is, therefore, the estimation of B.

Algorithm 4.3.2 suggests a very simple procedure: it suffices for any pair of distinct $i, j \in V$ to decide whether $\mathrm{supp}\big(\frac{X_i}{X_j}\big)$ has a positive lower bound, alternatively a positive upper bound, and if so, to estimate the bound. By Table 4.1, if there is such a bound, then it is an atom of $\frac{X_i}{X_j}$. Since we can expect to observe atoms more than twice for n sufficiently large, we propose the following estimation method.

Algorithm 4.5.1. [Find an estimate $\breve{B}$ of B from $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$]

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

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

if $\#\Big\{t : \bigwedge_{s=1}^{n} \frac{x_i^{(s)}}{x_j^{(s)}} = \frac{x_i^{(t)}}{x_j^{(t)}}\Big\} \geq 2$, then set $\breve{b}_{ji} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_j^{(t)}}$ and $\breve{b}_{ij} = 0$;

else, if $\#\Big\{t : \bigvee_{s=1}^{n} \frac{x_i^{(s)}}{x_j^{(s)}} = \frac{x_i^{(t)}}{x_j^{(t)}}\Big\} \geq 2$, then set $\breve{b}_{ij} = \bigwedge_{t=1}^{n} \frac{x_j^{(t)}}{x_i^{(t)}}$ and $\breve{b}_{ji} = 0$;

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

In step 2. rather two steps are summarized. The first step is concerned with estimating the reachability matrix of $\mathcal{D}$, the second with learning the ML coefficients. As explained above, we can use $\breve{B}$ to derive estimates of the DAGs and edge weights that represent $\boldsymbol{X}$ by (4.2.1) as well as an estimate of the distribution of the noise vector (cf. the last two paragraphs of Section 4.4.1 and Figure 4.3.1).

Since by (4.2.2) $b_{ji} \neq 0$ if and only if $j \in \mathrm{An}(i)$, b_{ji} and b_{ij} are never both positive for distinct i, j and, if $b_{jk}b_{ki} > 0$, then $b_{ji} > 0$. Algorithm 4.5.1 is constructed in the way that $\breve{B}$ satisfies the former property automatically but not the second. To guarantee this, we could update the estimator $\breve{B}$ as follows:
for all distinct $i, j \in V$ with $i < j$,

if $\breve{b}_{jk}\breve{b}_{ki} > 0$ for some $k \in V \setminus \{i, j\}$, then set $\breve{b}_{ji} = \bigwedge_{t=1}^{n} \frac{x_i^{(t)}}{x_j^{(t)}}$;

if $\breve{b}_{ik}\breve{b}_{kj} > 0$ for some $k \in V \setminus \{i, j\}$, then set $\breve{b}_{ij} = \bigwedge_{t=1}^{n} \frac{x_j^{(t)}}{x_i^{(t)}}$.

But now the first property is not necessarily satisfied anymore. As a consequence, the estimate $\mathcal{D}^{\breve{B}}$ of the minimum ML DAG $\mathcal{D}^B$ obtained from $\breve{B}$ is not necessarily acyclic. There are certainly many ways to avoid this and to obtain better estimates of B. Decisive is, however, the following: because of the distributional properties of the ratios between two components of $\boldsymbol{X}$ summarized in Table 4.1, Algorithm 4.5.1 outputs, $\mathbb{P}$-almost surely, the true ML coefficient matrix B if n is sufficiently large. Similar statements as in Proposition 4.4.33 can be made about the convergence of $\breve{b}_{ji}$ to the true value b_{ji} and about the number of observations needed to estimate B by $\breve{B}$ with a certain probability exactly.

4.6 Conclusion and Outlook

We first studied the identifiability of a recursive ML model $\boldsymbol{X}$ from its distribution. Its true DAG and edge weights are not identifiable; however, its ML coefficient matrix B. This repesents the class of all DAGs and edge weights that could have generated $\boldsymbol{X}$ via (4.2.1). In other words, we can identify representation (4.2.3) but not (4.2.1). Beside B, the distribution of the noise vector is identifiable. As a consequence of these results, we can recover B and the noise distributions from realizations of $\boldsymbol{X}$.

Parameter and structure learning for recursive ML models seems to be a challenging task because assumptions usually made for the models in standard methods are not met. However, in both cases, B can be estimated very efficiently by a simple procedure. The key idea of our approach is to consider the observed ratios between any two node variables, that is, 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 n is sufficiently large. It would be interesting to investigate the relationship between the performance of our procedures and n. Here, one possible question is how many observations are at least necessary to estimate B exactly. This requires understanding which of the events such as F_1, F_2, F_3, F_4 from Example 4.4.3 have positive probability (cf. the discussion below Corollary 4.A.5 in Appendix 4.A.2).

We plan to evaluate the performance of our proposed methods on simulated data sets. In Hartl [32] first simulations were performed. They confirm the theoretical findings and our epectations on the quality of our estimates. A comparison with other methods makes only limited sense. On the one hand because of the discussed assumptions of the other methods, on the other hand because of the outstanding properties of our estimates. Compared with other methods, we also do not make concrete distributional assumptions; we only assume independent and atomfree noise variables with support $\mathbb{R}_+$. In risk settings, which we have in mind when thinking of possible applications, it is natural to require the noise variables to have positive infinite support and atomfree distributions. Another advantage of our procedures is that they can deal with arbitrary high dimensions and, as long as n is sufficiently large, they have the same performance as for smaller dimensions.

A further goal 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; especially to expect that we observe a minimal observed ratio more than twice, what we do in Algorithm 4.5.1. It seems to be more reasonable to expect values close to each other. We therefore want to develop methods based on accumulation points. First attempts have already been made in [32].

Appendix 4.A

4.A.1 Alternative procedure to identify the ML coefficient matrix of a recursive ML model from its distribution (cf. Algorithm 4.3.2)

There are many ways to show that the ML coefficient matrix B of a recursive ML model $\boldsymbol{X}$ on a DAG $\mathcal{D}$ is identifiable from its distribution $\mathcal{L}(\boldsymbol{X})$. We present an alternative to Algorithm 4.3.2.

Proposition 4.A.1. *Let $i, j \in V$ be distinct. Then $j \in \mathrm{an}(i)$ if and only if there exists some $a \in \mathbb{R}_+$ such that for all $x_i, x_j \in \mathbb{R}_+$ with $ax_j \geq x_i$,*

$$\mathbb{P}(X_i \leq x_i, X_j \leq x_j) = \mathbb{P}(X_i \leq x_i). \tag{4.A.1}$$

In that case, $a = b_{ji}$.

Proof. For the bivariate distribution function of (X_i, X_j), we obtain for $x_i, x_j \in \mathbb{R}_+$, using representation (4.2.3) and the independence of the noise variables,

$$\mathbb{P}(X_i \leq x_i, X_j \leq x_j) = \mathbb{P}\Big(\bigvee_{\ell \in \mathrm{An}(i)} b_{\ell i} Z_\ell \leq x_i, \bigvee_{\ell \in \mathrm{An}(j)} b_{\ell j} Z_\ell \leq x_j\Big)$$

$$= \prod_{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}} \wedge \frac{x_j}{b_{\ell j}}\Big) \prod_{\ell \in \mathrm{An}(i) \setminus \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) \prod_{\ell \in \mathrm{An}(j) \setminus \mathrm{An}(i)} F_{Z_\ell}\Big(\frac{x_j}{b_{\ell j}}\Big). \tag{4.A.2}$$

Assume that $j \in \mathrm{an}(i)$. We find from Corollary 2.3.12 that $\bigwedge_{\ell \in \mathrm{An}(j)} \frac{b_{\ell i}}{b_{\ell j}} = b_{ji}$ and, hence, if $b_{ji} x_j \geq x_i$, then $\frac{x_j}{b_{\ell j}} \geq \frac{x_i}{b_{\ell i}}$ for all $\ell \in \mathrm{An}(j)$. With this we obtain from (4.A.2) for $b_{ji} x_j \geq x_i$,

$$\mathbb{P}(X_i \leq x_i, X_j \leq x_j) = \prod_{\ell \in \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) \prod_{\ell \in \mathrm{An}(i) \setminus \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) = \prod_{\ell \in \mathrm{An}(i)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) = \mathbb{P}(X_i \leq x_i).$$

This proves (4.A.1).

Assume now that (4.A.1) holds and that $j \notin \mathrm{an}(i)$. Furthermore, note that the latter holds if and only if $\mathrm{An}(j) \setminus \mathrm{An}(i) \neq \varnothing$. Since the noise variables have support $\mathbb{R}_+$, we know that $\prod_{\ell \in \mathrm{An}(j) \setminus \mathrm{An}(i)} F_{Z_\ell}\Big(\frac{x_j}{b_{\ell j}}\Big) < 1$. Thus, using (4.A.1), (4.A.2), and the monotony of a distribution function yields for $a x_j \geq x_i$,

$$\begin{aligned}
\mathbb{P}(X_i \leq x_i) &= \prod_{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}} \wedge \frac{x_j}{b_{\ell j}}\Big) \prod_{\ell \in \mathrm{An}(i) \setminus \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) \prod_{\ell \in \mathrm{An}(j) \setminus \mathrm{An}(i)} F_{Z_\ell}\Big(\frac{x_j}{b_{\ell j}}\Big) \\
&< \prod_{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}} \wedge \frac{x_j}{b_{\ell j}}\Big) \prod_{\ell \in \mathrm{An}(i) \setminus \mathrm{An}(j)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big) \\
&\leq \prod_{\ell \in \mathrm{An}(i)} F_{Z_\ell}\Big(\frac{x_i}{b_{\ell i}}\Big),
\end{aligned}$$

which is obviously a contradiction. Hence, $j \in \mathrm{an}(i)$. □

Proposition 4.A.1 and (4.2.2) allow us by the following algorithm to obtain B from $\mathcal{L}(\boldsymbol{X})$. This again proves the identifiability of B from $\mathcal{L}(\boldsymbol{X})$. Instead of the whole distribution $\mathcal{L}(\boldsymbol{X})$, it suffices to know the bivariate marginal distribution functions of $\mathcal{L}(\boldsymbol{X})$. However, this is a stronger information on $\mathcal{L}(\boldsymbol{X})$ than we need in Algorithm 4.3.2.

Algorithm 4.A.2. [Find B from $\mathcal{L}(\boldsymbol{X})$]

1. For all $i \in V = \{1, \ldots, d\}$, set $b_{ii} = 1$.
2. For all $i, j \in V$ with $i < j$, find $\mathbb{P}(X_i \leq x_i, X_j \leq x_j)$:

   if $\mathbb{P}(X_i \leq x_i, X_j \leq x_j) = \mathbb{P}(X_i \leq x_i)$ for some $a \in \mathbb{R}_+$ and all $x_i, x_j \in \mathbb{R}_+$ with $a x_j \geq x_i$, then set $b_{ji} = a$ and $b_{ij} = 0$;

   else, if $\mathbb{P}(X_i \leq x_i, X_j \leq x_j) = \mathbb{P}(X_j \leq x_j)$ for some $a \in \mathbb{R}_+$ and all $x_i, x_j \in \mathbb{R}_+$ with $a x_i \geq x_j$, then set $b_{ij} = a$ and $b_{ji} = 0$;

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

4.A.2 Ratios between two components of a recursive ML model

When estimating, identifying, or learning the structure of a recursive ML model $\boldsymbol{X}$, the ratios between two components of $\boldsymbol{X}$ are crucial. Because of their importance, we show further distributional properties of these ratios. In particular, we relate different events that can be described

by such ratios to events that depend on noise variables only. Already presented results such as the properties shown in Table 4.1 occur again as a consequence, but are shown for the sake of completeness.

Denote the probability space of the noise vector $(Z_1, \dots, Z_d)$ by $(\Omega, \mathcal{F}, \mathbb{P})$, and define for every $i \in V$,

$$m_i : \Omega \to \mathrm{An}(i), \quad \omega \mapsto j \text{ for some } j \in \Big\{ j \in \mathrm{An}(i) : b_{ji} Z_j(\omega) \geq \bigvee_{\ell \in \mathrm{An}(i) \setminus \{j\}} b_{\ell i} Z_\ell(\omega) \Big\}.$$

The max-linear representation (4.2.3) yields that $X_i(\omega) = b_{m_i(\omega),i} Z_{m_i(\omega)}(\omega)$, in words, m_i indicates a noise variable which realizes X_i. By (4.3.1), with probability 1, the maximum value of $\{b_{ji} Z_j : j \in \mathrm{An}(i)\}$ is achieved for unique $j \in \mathrm{An}(i)$; consequently, m_i is $\mathbb{P}$-almost surely uniquely defined. Since the noise variables are independent with support $\mathbb{R}_+$, m_i is, with positive probability, equal to each node in $\mathrm{An}(i)$. Unsurprisingly, the ratios between two components of $\boldsymbol{X}$ inherit the distributional properties from those of the noise variables.

Theorem 4.A.3. *Let $i, j \in V$ and $x \in \mathbb{R}_+$.*

(a) For every $F \in \mathcal{F}$,

$$\mathbb{P}(F \cap \{X_i \leq x X_j\}) = \mathbb{P}\Big(F \cap \Big\{ \bigvee_{\ell \in \mathrm{An}(i) \setminus M^{\leq}} b_{\ell i} Z_\ell < \bigvee_{\ell \in M^{\leq}} x b_{\ell j} Z_\ell \Big\}\Big) =: \mathbb{P}\big(F \cap \Omega_{ij}^{\leq}(x)\big),$$

where $M_{ij}^{\leq} = \{\ell \in \mathrm{An}(j) : b_{\ell i} \leq b_{\ell j} x\}$.

(b) The event $\{X_i \leq x X_j\}$ has positive probability if and only if $b_{ji} \leq x$.

(c) The event $\{X_i \leq x X_j\}$ has positive probability for every $x \in \mathbb{R}_+$ if and only if $j \notin \mathrm{An}(i)$.

(d) On $\{X_i \leq x X_j\}$, $m_j \in M_{ij}^{\leq}$.

Proof. By (4.2.3) $\{X_i \leq x X_j\}$ equals

$$\Big\{ \bigvee_{\ell \in \mathrm{An}(i) \setminus \mathrm{An}(j)} b_{\ell i} Z_\ell \vee \bigvee_{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)} b_{\ell i} Z_\ell \leq \bigvee_{\ell \in \mathrm{An}(j) \setminus \mathrm{An}(i)} x b_{\ell j} Z_\ell \vee \bigvee_{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)} x b_{\ell j} Z_\ell \Big\}. \tag{4.A.3}$$

The maximum on the right-hand side cannot be attained in $x b_{\ell j} Z_\ell$ for $\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)$ with $x b_{\ell j} < b_{\ell i}$; otherwise, we would have a contradiction as $x b_{\ell j} Z_\ell$ is strictly smaller than the maximum on the left-hand side. We find by (4.2.2) that $\mathrm{An}(j) \setminus \mathrm{An}(i) \subseteq M_{ij}^{\leq}$. These two observations yield (d). For $k \in \mathrm{An}(i) \cap \mathrm{An}(j) \cap M_{ij}^{\leq}$, obviously, $b_{ki} Z_k \leq \bigvee_{\ell \in M_{ij}^{\leq}} x b_{\ell j} Z_\ell$. Hence, we may also remove the nodes of $M_{ij}^{\leq}$ appearing on the left-hand side of (4.A.3). All in all, the events $\{X_i \leq x X_j\}$ and $\big\{ \bigvee_{\ell \in \mathrm{An}(i) \setminus M_{ij}^{\leq}} b_{\ell i} Z_\ell \leq \bigvee_{\ell \in M_{ij}^{\leq}} x b_{\ell j} Z_\ell \big\}$ coincide. Since, according to (4.3.1), $\Omega_{ij}^{\leq}$ differs from the second set only by a null set, we have verified (a). From (a) and the fact that the noise variables are independent and have support $\mathbb{R}_+$, we learn that $\mathbb{P}(X_i \leq x X_j) > 0$ if and only if $M_{ij}^{\leq} \neq \varnothing$. As by Corollary 2.3.12 $\bigwedge_{\ell \in \mathrm{An}(j)} \frac{b_{\ell i}}{b_{\ell j}} = b_{ji}$, $M_{ij}^{\leq} \neq \varnothing$ if and only if $x \geq b_{ji}$. This shows (b). Assertion (c) is a consequence of (b), since by (4.2.2) $b_{ji} = 0$ if and only if $j \notin \mathrm{An}(i)$. □

Theorem 4.A.3(d) implies that the max-linear representation (4.2.3) of X_j can be reduced on $\{X_i \le xX_j\}$ to $\bigvee_{\ell \in M_{ij}^{\le}} b_{\ell j} Z_\ell$. Since the maximum value of $\{b_{\ell i} Z_\ell : \ell \in \mathrm{An}(j)\}$ is achieved on $\{X_i \le xX_j\}$ with positive probability for every $\ell \in M_{ij}^{\le}$, a further reduction is not possible.

By Theorem 4.A.3 we can draw conclusions about potential atoms of $\frac{X_i}{X_j}$: consider the intersection of the events $\{X_i \le xX_j\}$ and $\{X_i \ge xX_j\} = \{X_j \le \frac{1}{x} X_i\}$, which equals $\{X_i = xX_j\}$. In the following corollary we summarize some results regarding the event $\{X_i = xX_j\}$.

Corollary 4.A.4. *Let $i, j \in V$ and $x \in \mathbb{R}_+$.*

(a) For all $F \in \mathcal{F}$,

$$\begin{aligned}\mathbb{P}\big(F \cap \{X_i = xX_j\}\big) &= \mathbb{P}\Big(F \cap \Big\{ \bigvee_{\ell \in M_{ij}^{=}} b_{\ell i} Z_\ell > \bigvee_{\ell \in \mathrm{An}(i) \setminus M_{ij}^{=}} b_{\ell i} Z_\ell \vee \bigvee_{\ell \in \mathrm{An}(j) \setminus M_{ij}^{=}} x b_{\ell j} Z_\ell \Big\}\Big) \\ &=: \mathbb{P}\big(F \cap \Omega_{ij}^{=}(x)\big),\end{aligned}$$

where $M_{ij}^{=} = \{\ell \in \mathrm{An}(i) \cap \mathrm{An}(j) : b_{\ell i} = x b_{\ell j}\}$.

(b) The event $\{X_i = xX_j\}$ has positive probability if and only if $x = \frac{b_{\ell i}}{b_{\ell j}}$ for some $\ell \in \mathrm{An}(i) \cap \mathrm{An}(j)$.

(c) $\frac{X_i}{X_j}$ has atoms if and only if $\mathrm{An}(i) \cap \mathrm{An}(j) \neq \emptyset$.

(d) We have $\mathbb{P}$-almost surely on $\{X_i = xX_j\}$, $m_i = m_j \in M_{ij}^{=}$.

Proof. (a) It can be shown that $\Omega_{ij}^{=}(x) = \Omega_{ij}^{\le}(x) \cap \Omega_{ij}^{\le}(1/x)$. Hence, (a) follows from part (a) of Theorem 4.A.3.
(b) is a consequence of (a), since the noise variables are independent with support $\mathbb{R}_+$.
(c) is immediate by (b).
(d) The max-linear representation (4.2.3) shows that on $\Omega_{ij}^{=}(x)$, $m_i, m_j \in M_{ij}^{=}$. Since by (a) the set $\{X_i = xX_j\} \setminus \Omega_{ij}^{=}(x)$ is a null set, (d) holds. □

For the sake of completeness, we consider the events that are complementary to them in Theorem 4.A.3.

Corollary 4.A.5. *Let $i, j \in V$ and $x \in \mathbb{R}_+$. Let further $M_{ij}^{\le}$ be as defined in Theorem 4.A.3.*

(a) We have

$$\{X_i > xX_j\} = \Big\{ \bigvee_{\ell \in \mathrm{An}(i) \setminus M_{ij}^{\le}} b_{\ell i} Z_\ell > \bigvee_{\ell \in M_{ij}^{\le}} x b_{\ell j} Z_\ell \Big\} = \Big\{ \bigvee_{\ell \in M_{ij}^{>}} b_{\ell i} Z_\ell > \bigvee_{\ell \in \mathrm{An}(j) \setminus M_{ij}^{>}} x b_{\ell j} Z_\ell \Big\} = \Omega_{ij}^{>}(x),$$

where $M_{ij}^{>} = \mathrm{An}(i) \setminus M_{ij}^{\le} = \{\ell \in \mathrm{An}(i) : b_{\ell i} > b_{\ell j} x\}$.

(b) The event $\{X_i > xX_j\}$ has positive probability if and only if $x < \frac{1}{b_{ij}}$.

(c) The event $\{X_i > xX_j\}$ has positive probability for every $x \in \mathbb{R}_+$ if and only if $i \notin \mathrm{An}(j)$.

(d) On $\{X_i > xX_j\}$, $m_i \in M_{ij}^{>}$.

Proof. (a) In the proof of Theorem 4.A.3, we have shown that the events $\{X_i \le xX_j\}$ and $\big\{\bigvee_{\ell\in \mathrm{An}(i)\setminus M_{ij}^{\le}} b_{\ell i}Z_\ell \le \bigvee_{\ell\in M_{ij}^{\le}} xb_{\ell j}Z_\ell\big\}$ are equal. Since $\{X_i > xX_j\}$ is the complementary event of $\{X_i \le xX_j\}$, the assertion is clear.
(b), (c) can be obtained from (a) analogously as Theorem 4.A.3(b), (c) from part (a) there.
(d) follows from (a) and (4.2.3). □

We can use the results presented in this section to figure out whether events such as F_1, F_2, F_3, F_4 from Example 4.4.3 have positive probability. For example, we obtain from parts (a) of Corollaries 4.A.4, 4.A.5 that

$$\begin{aligned}
\mathbb{P}(F_1) &= \mathbb{P}(\{b_{13}X_1 = b_{23}X_2\} \cap \{X_3 = b_{13}X_1\} \cap \{X_3 = b_{23}X_2\}) \\
&= \mathbb{P}(\Omega_{21}^{=}(b_{13}/b_{23}) \cap \Omega_{31}^{=}(b_{13}) \cap \Omega_{32}^{=}(b_{23})) = \mathbb{P}(\varnothing \cap \{b_{13}Z_1 > Z_3\} \cap \{b_{23}Z_2 > Z_3\}) = 0, \\
\mathbb{P}(F_4) &= \mathbb{P}(\Omega_{31}^{>}(b_{13}) \cap \Omega_{32}^{>}(b_{23})) = \mathbb{P}(\{Z_3 > b_{13}Z_1\} \cap \{Z_3 > b_{23}Z_2\}) = \mathbb{P}(Z_3 > b_{13}Z_1 \vee b_{23}Z_2) > 0,
\end{aligned}$$

since the noise variables are independent and have support $\mathbb{R}_+$. Depending on which events may occur with positive probability simultaneously, more or less observations are needed to obtain good estimates of B using (4.4.8) when $\mathcal{D}$ is known or Algorithm 4.5.1 when $\mathcal{D}$ is unknown.

The distributional properties of $\frac{X_i}{X_j}$ presented in Table 4.1 also follow from this section; the atoms are given in Corollary 4.A.4(b), $\mathrm{supp}\big(\frac{X_i}{X_j}\big)$ can be determined, for example, from Theorem 4.A.3(b) and Corollary 4.A.5(b).

Bibliography

[1] A. V. Aho, M. R. Garey, and J. D. Ullman. The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137, 1972.

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

[3] E. Ayra. Risk analysis of runway overrun excursions at landing: a case study. `http://www.agifors.org/award/submissions2013/EduardoAyra.pdf`, 2013.

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

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

[6] P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

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

[8] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[9] S. Coles, J. Heffernan, and J. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365, 1999.

[10] D. R. Cox and N. Wermuth. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218, 1993.

[11] Q. Cui and Z. Zhang. Max-linear competing factor models. *Journal of Business & Economic Statistics*, 36(1):62–74, 2018.

[12] R. A. Davis and T. Mikosch. The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009, 2009.

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

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

[15] R. Diestel. *Graph Theory.* Graduate Texts in Mathematics, Vol. 173. Springer, Heidelberg, 4th edition, 2010.

[16] G. Draisma, H. Drees, A. Ferreira, and L. de Haan. Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, 10(2):251–280, 2004.

[17] H. Drees and X. Huang. Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis*, 64(1):25–47, 1998.

[18] M. Drton and M. D. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

[19] J. H. J. Einmahl, A. Krajina, and J. Segers. An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793, 2012.

[20] 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.

[21] J. Ernest, D. Rothenhäusler, and P. Bühlmann. Causal inference in partially linear structural equation models: identifiability and estimation. arXiv:1607.05980, *the Annals of Statistics*, to appear, 2018.

[22] M. Falk. On the generation of a multivariate extreme value distribution with prescribed tail dependence parameter matrix. *Statistics & Probability Letters*, 75(4):307–314, 2005.

[23] M. Falk, M. Hofmann, and M. Zott. On generalized max-linear models and their statistical interpolation. *Journal of Applied Probability*, 52(3):736–751, 2015.

[24] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[25] D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 235–243. Morgan Kaufmann, Seattle, WA, 1994.

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

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

[28] N. Gissibl and C. Klüppelberg. Prediction of recursive max-linear models. In preparation, 2018.

[29] 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.

[30] N. Gissibl, C. Klüppelberg, and S. L. Lauritzen. Identifiability and estimation of recursive max-linear models. In preparation, 2018.

[31] N. Gissibl, C. Klüppelberg, and M. Otto. Tail dependence of recursive max-linear models with regularly varying noise variables. *Econometrics and Statistics*, 6:149 – 167, 2018.

[32] J. Hartl. *Estimating the Coefficients of Max-Linear Structural Equation Models.* Master's thesis, Technical University of Munich, 2015.

[33] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–165. MIT Press, Cambridge, MA, 1999.

[34] A. Hitz and R. Evans. One-component regular variation and graphical modeling of extremes. *Journal of Applied Probability*, 53(3):733746, 2016.

[35] 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.

[36] X. Huang. *Statistics of Bivariate Extremes.* Ph.D. thesis, Erasmus University Rotterdam, Tinbergen Institute Research series No. 22, 1992.

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

[38] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data.* Wiley, New York, 1980.

[39] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[40] 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.

[41] A. Kiriliouk. Hypothesis testing for tail dependence parameters on the boundary of the parameter space with application to generalized max-linear models. arXiv:1708.07019, 2018.

[42] A. Klenke. *Probability Theory: A Comprehensive Course.* Springer, London, 2007.

[43] C. Klüppelberg and S. Lauritzen. Bayesian networks for max-linear models. Submitted, 2018.

[44] C. Klüppelberg and E. Sönmez. Max-linear models on infinte graphs generated by bernoulli bond percolation. arXiv:1804.06102, 2018.

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

[46] M. Krali. *Causality and Estimation of Multivariate Extremes on Directed Acyclic Graphs.* Master's thesis, Technical University of Munich, 2018.

[47] S. L. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

[48] 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.

[49] A. W. Ledford and J. A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

[50] D. Maclagan and B. Sturmfels. *Introduction to Tropical Geometry*. Graduate Studies in Mathematics, Vol. 161. American Mathematical Society, Providence, Rhode Island, 2015.

[51] B. Mahr. A birds eye view to path problems. In H. Noltemeier, editor, *Graphtheoretic Concepts in Computer Science: Proceedings of the International Workshop WG 80 Bad Honnef, June 15–18, 1980*, pages 335–353. Springer, Berlin, 1981.

[52] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[53] M. Otto. *Extremes on Directed Acyclic Graphs*. Master's thesis, Technical University of Munich, 2016.

[54] I. Papastathopoulos and K. Strokorb. Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15, 2016.

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

[56] L. Peng. Estimation of the coefficient of tail dependence in bivariate extremes. *Statistics & Probability Letters*, 43(4):399–409, 1999.

[57] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[58] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

[59] O. Pourret, P. Naïm, and B. E. Marcot, editors. *Bayesian Networks: A Practical Guide to Applications*. Wiley, Chichester, 2008.

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

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

[62] G. Rote. A systolic array algorithm for the algebraic path problem (shortest paths; matrix inversion). *Computing*, 34(3):191–219, 1985.

[63] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[64] M. Schlather and J. Tawn. Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes*, 5(1):87–102, 2002.

[65] 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.

[66] M. Sibuya. Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210, 1960.

[67] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[68] P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):1–28, 2016.

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

[70] K. Strokorb and M. Schlather. An exceptional max-stable process fully parameterized by its extremal coefficients. *Bernoulli*, 21(1):276–302, 2015.

[71] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artifical Intelligence*, pages 220–227, Cambridge, MA, 1990.

[72] C. Wang, L. Drees, N. Gissibl, L. Hoehndorf, J. Sembiring, and F. Holzapfel. Quantification of incident probabilities using physical and statistical approaches. In *6th International Conference on Research in Air Transportation. Istanbul, Turkey*, 2014.

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

[74] S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1934.

[75] R. Yuen and S. A. Stoev. CRPS M-estimation for max-stable models. *Extremes*, 17(3): 387–410, 2014.

[76] Z. Zhang. *An Algebraic Approach to Understanding Generalized Recursive Max-linear Model*. Master's thesis, Technical University of Munich, 2018. URL `https://mediatum.ub.tum.de/doc/1439499/1439499.pdf`.