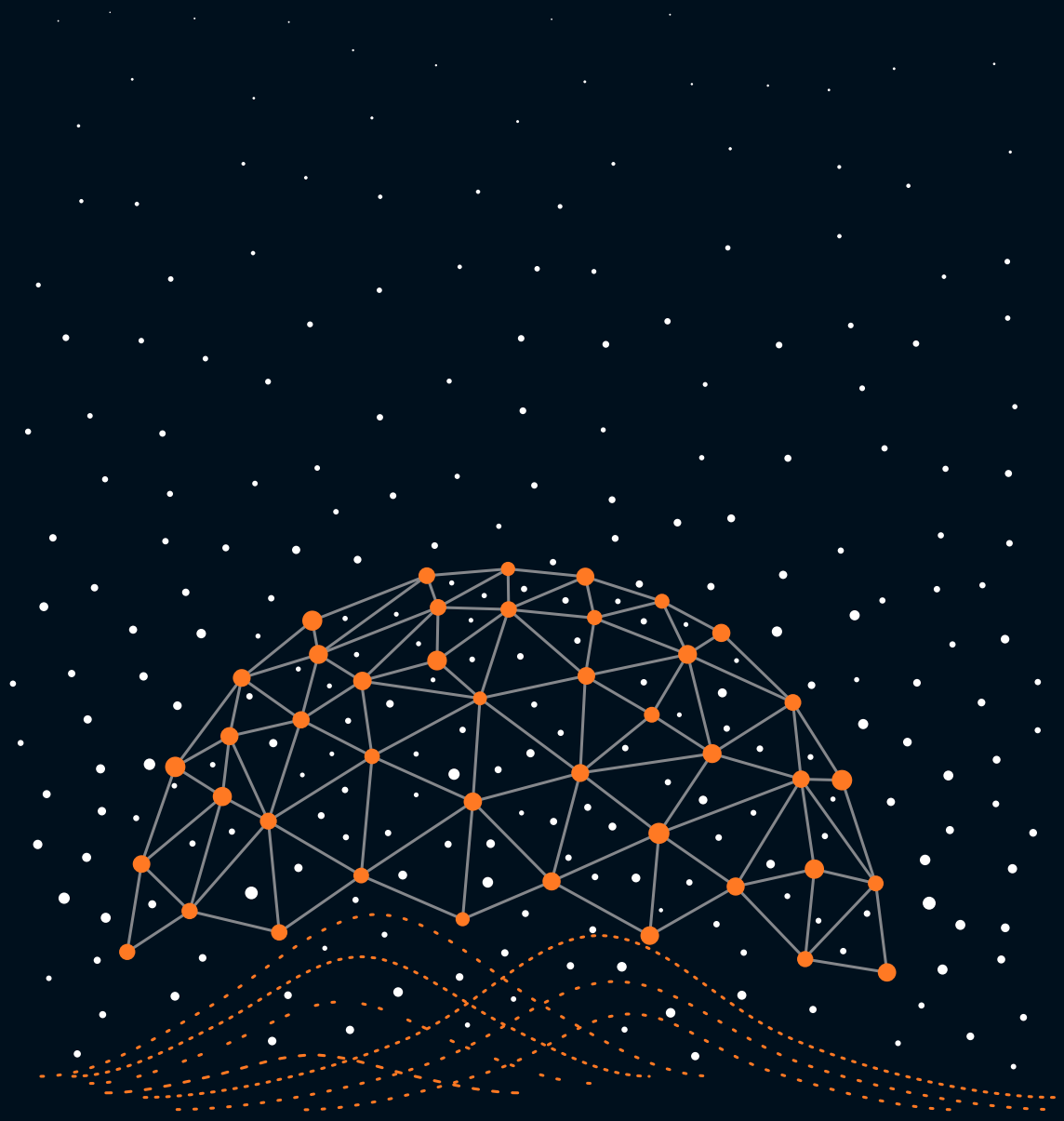
PROCEEDINGS



OF THE 13TH WORKSHOP ON UNCERTAINTY PROCESSING

Proceedings of the 13th Workshop on Uncertainty Processing *(WUPES'25)* Třešť, Czech Republic

Milan Studený, Nihat Ay, Andrea Capotorti, László Csirmaz,
Radim Jiroušek, Gernot D. Kleiter, Prakash P. Shenoy
(editors)

June 4-7, 2025

Published by:
MatfyzPress,
Publishing House of the Faculty of Mathematics and Physics Charles University
Sokolovská 83, 186 75 Praha 8, Czech Republic
as the 719. publication.
Printed by Reprostředisko MFF UK.

Printed in Prague — May 2025

Organized by:
Institute of Information Theory and Automation, Czech Academy of Sciences
Faculty of Management, Prague University of Economics and Business

Credits:
Editors: Milan Studený, Nihat Ay, Andrea Capotorti, László Csirmaz, Radim Jiroušek,
Gernot D. Kleiter, Prakash P. Shenoy
LaTeXeditor: Václav Kratochvíl
Cover design: Jiří Přibil
using LaTeX's 'confproc' package, version 0.8 by V. Verfaille

ISBN: 978-80-7378-525-3

13th WORKSHOP ON UNCERTAINTY PROCESSING

WUPES '25

TŘEŠŤ

Organized by:

Institute of Information Theory and Automation,
Czech Academy of Sciences

&

Faculty of Management,
Prague University of Economics and Business

Třešť

June 4-7, 2025

Programme and Conference Committee:
Milan Studený - chair, *Czech Academy of Sciences, Prague*
Nihat Ay, *Hamburg University of Technology, Germany*
Andrea Capotorti, *Università di Perugia, Italy*
László Csirmaz, *Alfréd Rényi Inst. of Mathematics, Hungary*
Radim Jiroušek, *Czech Academy of Sciences, Prague*
Gernot D. Kleiter, *Universität Salzburg, Austria*
Prakash P. Shenoy, *University of Kansas, USA*

Organizing Committee:
Jirka Vomlel - chair, *Czech Academy of Sciences, Prague*
Radim Jiroušek, *Czech Academy of Sciences, Prague*
Václav Kratochvíl, *Czech Academy of Sciences, Prague*
Milan Studený, *Czech Academy of Sciences, Prague*

Foreword

When I agreed to write a preface for these proceedings, I planned to incorporate good ideas from the previous twelve prefaces. Wanting to contribute something original, I considered what was new about the thirteenth meeting. It returns to Castle Třešť, but this return is not unique, as the first two meetings were held in the small village of Alšovice. Unable to come up with a better idea, I decided that this preface itself should be original. Thus, these are the first WUPES meeting proceedings to be opened with an AI-generated preface (I hope the rest of the proceedings are not written this way).

In Nučice, May 15, 2025

Radim Jiroušek

Preface

We are pleased to present the proceedings of the 13th Workshop on Uncertainty Processing (WUPES 2025), held at Castle Třešť in the Czech Republic. This historic venue, which also hosted WUPES in 1994, offers a fitting backdrop for the kind of focused and collegial atmosphere the workshop is known for. The return to Třešť evokes the long-standing tradition of WUPES as a meeting place for researchers interested in the many facets of uncertainty in artificial intelligence and related areas.

WUPES is a small and informal workshop that encourages the presentation of unfinished work, exploratory ideas, and early-stage results. Rather than emphasizing polished papers or rigid peer review, WUPES values open discussion, constructive feedback, and the exchange of perspectives across a variety of theoretical frameworks and applications. This years contributions reflect that diversity, with topics including probabilistic reasoning, fuzzy systems, belief functions, Bayesian networks, and more.

The spirit of WUPES lies in its openness and its scale — small enough for meaningful conversation, yet broad enough to bring together a range of approaches. We hope these proceedings offer a snapshot of current thinking in the field and help spark further dialogue and collaboration.

We are grateful to all the authors who contributed their work. Special thanks go to the local organizing team for making this return to Castle Třešť possible and for ensuring a welcoming setting for all participants.

We look forward to the continued growth of the WUPES community and to future meetings that carry on this unique and valuable tradition.

Somewhere on the internet, May 15, 2025

On behalf of the WUPES 2025 Organizing Committee

Language Model: ChatGPT (OpenAI)

List of Content

213 List of Authors

Inference for max-linear Bayesian networks with noise

Mark Adams*[1], Kamillo Ferry†[2], and Ruriko Yoshida‡[1]

[1]Department of Operations Research, Naval Postgraduate School
{*mark.p.adams,ryoshida*}*@nps.edu*
[2]Institute of Mathematics, Technische Universität Berlin
ferry@math.tu-berlin.de

Abstract

Max-Linear Bayesian networks provide a powerful framework for causal inference in extreme-value settings. We consider Max-Linear Bayesian networks with noise parameters with a given topology in terms of the max-plus algebra by taking its logarithm. Then, we show that an estimator of a parameter for each edge in a directed acyclic graph is distributed normally. We end this paper with computational experiments empirically studying the limiting conditions of the expectation and maximization algorithm and showing how quadratic optimization can be an alternative method to parameter estimation.

1 Introduction

Identifying and quantifying causal relationships is an objective in scientific inquiry and applied decision making processes. This objective becomes especially critical in the analysis of extreme events, which, despite their low frequency, can lead to disproportionately severe consequences in terms of cost and impact. Gaining insight into the underlying causal mechanisms is essential for informing risk management strategies, and guiding the development of effective mitigation policies. Gissibl and Klüppelberg (2018) contend that *max-linear Bayesian networks* (MLBNs) provide a powerful framework for studying causal relationships in extreme-value settings, where interactions between variables follow a max-linear structure. These max-linear models have found a broad range of applications, e. g. in environmental sciences for modeling flooding events as demonstrated by Engelke and Hitz (2020), and finances as explored by Einmahl, Kiriliouk, and Segers. (2018).

*MA is partially supported by NSF Statistics Program DMS 2409819.

†KF is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

‡RY is partially supported by NSF Statistics Program DMS 2409819.

Full version: `https://arxiv.org/abs/2505.00229`

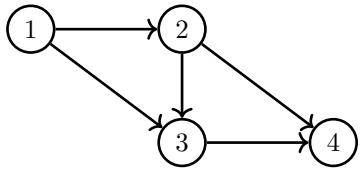


Figure 1: A DAG consisting of four vertices. Each vertex i in the network represents a random variable X_i in the joint distribution of a max-linear structure $X = (X_1, X_2, X_3, X_4)$.

A max-linear Bayesian network is a statistical model that is described by a weighted directed acyclic graph (DAG) in the following way. Let $G = (V, E)$ be a DAG with weight matrix $C = (c_{ij}) \in \mathbb{R}_{\geq 0}^{n \times n}$. Then, the MLBN on G for a random vector $X = (X_1, \ldots, X_n)$ is defined by the recursive structural equations

$$X_j = \bigvee_{i \in \mathrm{pa}(j)} c_{ij} X_i \vee c_{jj} Z_j, \quad i = 1, \ldots, n, \tag{1}$$

where $\vee$ denotes taking the maximum, $\mathrm{pa}(j)$ denotes the *parents* of vertex j, and $Z_1, \ldots, Z_n$ are independent non-negative random variables called *innovations*.

Example 1.1. Figure 1 represents a DAG on four vertices for a MLBN with a random variable $X = (X_1, X_2, X_3, X_4)$, a vector of innovations $Z = (Z_1, Z_2, Z_3, Z_4)$ and a weight matrix $C \in \mathbb{R}_{\geq 0}^{4 \times 4}$. The structural equations (1) for the model are given by:

$$\begin{aligned} X_1 &= Z_1, \\ X_2 &= c_{12} X_1 \vee Z_2, \\ X_3 &= c_{13} X_1 \vee c_{23} X_2 \vee Z_3, \\ X_4 &= c_{24} X_2 \vee c_{34} X_3 \vee Z_4, \end{aligned} \quad \text{where} \quad C = \begin{pmatrix} 1 & c_{12} & c_{13} & 0 \\ 0 & 1 & c_{23} & c_{24} \\ 0 & 0 & 1 & c_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Explicitly writing down the solution given by the matrix C allows us to express the random vector $X = (X_1, X_2, X_3, X_4)$ by

$$\begin{aligned} X_1 &= Z_1, \\ X_2 &= c_{12} Z_1 \vee Z_2, \\ X_3 &= (c_{13} \vee c_{12} c_{23}) Z_1 \vee c_{23} Z_2 \vee Z_3, \\ X_4 &= (c_{12} c_{24} \vee c_{13} c_{14} \vee c_{12} c_{23} c_{34}) Z_1 \vee (c_{24} \vee c_{23} c_{34}) Z_2 \vee c_{34} Z_3 \vee Z_4. \end{aligned} \tag{2}$$

In other words, equation (2) describes a matrix-vector equation $X = Z \cdot C^*$ where

$$C^* = \begin{pmatrix} 1 & c_{12} & c_{13} \vee c_{12} c_{23} & c_{12} c_{24} \vee c_{13} c_{14} \vee c_{12} c_{23} c_{34} \\ 0 & 1 & c_{23} & c_{24} \vee c_{23} c_{34} \\ 0 & 0 & 1 & c_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A central challenge in the analysis and application of MLBNs lies in the estimation of parameter matrix C. These parameters act as multiplicative weights along the directed edges of the network, while the vertices assume values through max-linear operations.

Klüppelberg and Lauritzen (2019) demonstrate that as a result of max-linear operations, standard likelihood based estimation techniques are not directly applicable. In particular, Gissibl, Klüppelberg, and Lauritzen (2021) identify possible edge weights with a sufficient sample size and without noise in the model. Buck and Kluppelberg (2020) derive estimators under the assumption of one sided noise.

Inspired by Tran (2022) and the Latent Tree problem, we model sensor collection error and develop a statistical framework for parameter estimation under more relaxed noise constraints.

1.1 Problem Statement

We develop a statistical estimation framework for the parameter matrix of a MLBN in the presence of multiplicative noise, assuming that the underlying DAG structure is known. We introduce a modification to the standard max-linear recursive equation by incorporating a strictly positive noise variable $E_j > 0$ into each structural equation. The resulting model takes the form:

$$X_j = \bigvee_{i \in \mathrm{pa}(j)} (c_{ij} X_i \vee c_{jj} Z_j) E_j, \quad i = 1, \ldots, n, \tag{3}$$

for a matrix $C = (c_{ij}) \in \mathbb{R}_{\geq 0}^{n \times n}$, and E_j represents multiplicative noise, capturing variability that distorts the observed values of X_j.

Our goal is to develop statistically sound and computationally efficient inference procedures that leverage the algebraic structure of the max-times semiring and the sparsity inherent in the DAG, enabling accurate estimation of C from observed data subject to noise.

1.2 Motivation

Gissibl, Klüppelberg, and Lauritzen (2021) established the identifiability of max-linear recursive equations in the noise-free setting by exploiting the structure of consistently occurring observations. A key challenge in our setting is the disentanglement of the multiplicative noise component from the underlying distribution, especially when noise obscures the max-linear dependencies. Developing accurate and robust estimators for the edge weights of the MLBN is crucial, as it enables more reliable inference of causal pathways, and improves our ability to reason about the causal impacts of extreme events in complex systems.

Understanding the causal relations between variables enables informed decisions about the effects of interventions, which is essential in informed policy design. In the context of extreme events and MLBNs, the model's structure implies that system behavior is governed by dominant risk pathways. Consequently, effective policies should focus on mitigating the most influential risk factors.

1.3 Contributions of this paper

This work contributes to the statistical foundations of MLBN by addressing parameter estimation in structured graphical models under noise. We demonstrate how logarithmic transformations can be leveraged to estimate parameters in the presence of noise and we provide a theoretical justification for using Gaussian Mixture Models (GMMs) as a statistical estimation tool within a MLBN framework. In parallel, we propose a quadratic optimization problem for estimating parameters for MLBN as an alternative. In the context of tropical geometry, and in the framework of our quadratic optimization problem, the feasible region forms a *polytrope*, which is a tropical polytope that is also clasically convex (see Joswig and Kulas (2010) for details on polytropes). Our problem becomes estimating the *Kleene star* C^* from the observations.

1.4 Organization of this paper

Our paper is organized as follows. In Section (2), we provide an overview of relevant key concepts, including graph terminology, tropical geometry, and max-linear models. In Section 3, we formally define the estimation problem, introduce log-space representations, establish the justification for using GMMs in parameter estimation, and provide a geometric estimation method. In Section 4, we present analytical results, and discuss the conditions under which GMM-based estimation fails. Section 5 concludes the paper and summarizes our key finding discussing their implications and providing the basis for future research directions.

2 Preliminaries

2.1 Graph Terminology

In the following, we study simple directed acyclic graphs (DAG) $G = (V, E)$ defined by a finite sets of vertices $V = [n] \coloneqq \{1, \dots, n\}$ and edges $E \subset V \times V$. In the terminology of Lauritzen (2004) we consider pure graphs only consisting of directed edges.

An edge $e \in E$ is defined by its *source* i and *target* j. This way, we keep to the same graph notation as Améndola et al. (2022) except i becomes the parent and j becomes the child given an edge $i \to j$. A path $i \rightsquigarrow j$ in G is defined as a sequence of distinct nodes $(i = d_0, d_1, \dots, d_\ell = j)$ such that $d_k \to d_{k+1}$ is an edge in G for all $0 \leq k < \ell$.

The set of *parents* of j is denoted $\mathrm{pa}(j)$ and the set of *children* of i is $\mathrm{ch}(i)$. These relationships can further be categorized into ancestors and descendants. Here j is a *descendant* of i and i is an *ancestor* of j if there exists a path from i to j, denoted by $i \rightsquigarrow j$. We denote the set of ancestors by $\mathrm{an}(i)$ and define the set of extended ancestors as $\overline{\mathrm{an}}(i) = \mathrm{an}(i) \cup \{i\}$.

A weighted directed graph is a directed graph together with a weight matrix $C \in \mathbb{R}_{\geq 0}^{n \times n}$ such that $c_{ii} = 1$ and $c_{ij} > 0$ whenever $i \to j \in E$.

2.2 Tropical semirings and polytropes

Max-linear Bayesian networks are inherently tropical objects. For this, we introduce the necessary preliminaries from tropical geometry to make our setting precise. There are two tropical semirings that are relevant for us, the *max-times semiring* $\mathbb{R}_{\geq 0}$ equipped with operations

$$a \vee b := \max(a, b), \quad a \cdot b := ab \quad \text{for } a, b \in \mathbb{R}_{\geq 0} := [0, \infty).$$

and the *max-plus semiring* $(\mathbf{T}, \oplus, \odot)$ where

$$a \oplus b := \max(a, b), \quad a \odot b := a + b \quad \text{for } a, b \in \mathbf{T} := \mathbb{R} \cup \{-\infty\}.$$

The semirings $\mathbb{R}_{\geq 0}$ and $\mathbf{T}$ are isomorphic by taking the logarithm resp. exponentiation. While max-linear Bayesian networks are defined over the max-times semiring, only geometry over the max-plus semiring allows for the necessary comparison to Euclidean geometry. Multiplication of matrices over these semirings is carried out analogously to the classical case using the corresponding addition and multiplication of the semiring.

It follows from equation (3) that a MLBN can be expressed as $X = Z \cdot C^*$ where C^* is the *Kleene star*, such that, $C^* = I_n \vee C \vee C^2 \vee \cdots \vee C^{n-1}$. Puente (2013) show that in this setting, $\log X$ lies in the set

$$\mathrm{Q}(A) = \{\, x \in \mathbb{R} \mid x_j - x_i \geq a_{ij} \text{ for } 1 \leq i, j \leq n \,\} \subseteq \mathbb{R}^n \tag{4}$$

which is the *polytrope* $\mathrm{Q}(A)$ associated to $A = \log C$.

Remark 2.1. Améndola and Ferry (2024) characterized for DAGs G the perturbations of weight matrices A preserving the associated Kleene star A^*. This happens in terms of the optimal transport problem on G. Theorem 4.9 of Améndola and Ferry (2024) proves that a hyperplane $\{\, x_j - x_i = a_{ij} \,\}$ defines a facet of the polytrope $\mathrm{Q}(A)$ if and only if the edge $i \to j$ is the unique optimal path connecting i to j. This means that edges in G might become irrelevant depending on the weights A.

2.3 Recursive structural equations and max-linear Bayesian networks

Gissibl and Klüppelberg (2018) introduce Max-linear Bayesian networks as a *recursive structural equation model* over the max-times semiring. We provide an example in equation (2).

By repeated substitution, this recursive equation system admits the solution $X = Z \cdot C^*$ where C^* is the Kleene star over the max-times semiring. By assumption, C is the weight matrix of a directed acyclic graph making C^* well-defined.

After applying a logarithmic transformation, the set of possible observations for $\log X$ forms a polytrope in $\mathbf{T}\mathbb{A}^{n-1}$. For this reason, we discuss the properties of $\log X$ from now on.

If $\omega := \log C^*$ denotes the logarithmic Kleene star of the weight matrix for the MLBN X, it follows from (4) that the observations of the difference

$$Y_{ij} := \log X_j - \log X_i$$

will be bounded from below. Gissibl, Klüppelberg, and Lauritzen (2021) gave an extensive characterisation of atoms occuring in the distribution of Y_{ij}.

Lemma 2.2 (Gissibl, Klüppelberg, and Lauritzen, Lemma 3.4)**.** *Let $i \neq j \in V(G)$ be distinct nodes of the underlying graph G. Then, the random variable Y_{ij} has an atom at $\omega_{kj} - \omega_{ki}$ for every common ancestor k and these are the only atoms. In particular, if i is an ancestor of j, then there is an atom at ω_{ij}.*

As a consequence of Lemma 2.2, for a sample $Y_{ij}^1, \dots, Y_{ij}^N$ of differences Y_{ij} without noise, the estimator

$$\hat{\omega}_{ij} = \min_{\nu=1}^{N} \left(Y_{ij}^{\nu}\right) \tag{5}$$

will be exactly equal to the true parameter with high probability as shown by Klüppelberg and Lauritzen (2019).

Remark 2.3. The phenomenon discussed in Remark 2.1 applies to max-linear Bayesian networks, particularly when an edge is either removed or rendered functionally insignificant within the network. We say that *structural inactivation* of an edge occurs when $\mathbb{P}_{c_{ij}}(X_j = X_i c_{ij}) = 0$.

Due to the statistical nature, structural inactivation can occur due to substantial reduction in the weight of parameter c_{ij}. In our setting, we define an edge to be approaching structurally inactivation when $\mathbb{P}_{c_{ij}}(X_j = X_i c_{ij}) < 0.05$. The threshold of 0.05 is adopted due to its conventional use in capturing tail dependence, as well as its empirical relevance as demonstrated by our observations in Table 1.

3 Parameter estimation under uncertainty

Assume that we are given observations of a MLBN $X = (X_1, \dots, X_n)$ with the presence of a noise E_j log-normally distributed. This means in particular that $\varepsilon_j \coloneqq \log E_j \sim N(0, \sigma_j^2)$ with $\sigma_j > 0$. We decide to work over the max-plus semiring. Thus, the problem we study is to estimate the parameters ω_{ij} given noisy data that satisfies the equations

$$\log X_j \odot \varepsilon_j = \left(\bigoplus_{i \in \mathrm{pa}(j)} \log c_{ij} \odot \log Z_i \oplus \log c_{jj} \odot \log Z_j \right) \odot \varepsilon_j. \tag{6}$$

3.1 Gaussian Mixture Models

Each random variable X_j arises as the maximum over several weighted random variables. We can see this as one specific observation for $\log X_j$ being selected at random from the expressions $\omega_{ij} + \log Z_i$ for each path from i to j in the underlying graph. In this section, we elaborate how in the noisy setting, the above observation leads to the application of Gaussian mixture models.

A *mixture* is a random variable X with density f given by the convex combination of probability densities f_k, that is

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

where K is the number of *mixture components* and $\pi_k \geq 0$ are the *mixing weights* satisfying $\sum_{k=1}^{K} \pi_k = 1$. If D_k are probability distributions with density f_k, we may denote X being a mixture by $X \sim \sum_{k=1}^{K} \pi_K D_k$. If for each $1 \leq k \leq K$, f_k is the density of a Gaussian distribution with mean μ_k and variance σ_k^2 we say that X is a *Gaussian mixture*. In this case, we write $X \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)$.

Under noisy conditions, the discrete atoms of Y_{ij} as described in Lemma 2.2 become normally distributed. This suggests that we may see the noisy differences $Y_{ij} + (\varepsilon_j - \varepsilon_i)$ as distorted Gaussian mixtures in the following way. The following is a generalization of Lemma 2.2 to the noisy setting.

Theorem 3.1. *Assume that $\varepsilon_j \sim N(0, \sigma_j^2)$ with $\sigma_j > 0$ for $j \in V(G)$. Then, there exists a distribution D and real numbers $0 \leq \pi_k, \pi \leq 1$ for every common ancestor k of i and j with $\pi + \sum_k \pi_k = 1$ such that Y_{ij} has as distribution the following finite mixture*

$$Y_{ij} + (\varepsilon_j - \varepsilon_i) \sim \sum_{k \in \overline{\text{an}}(i) \cap \overline{\text{an}}(j)} \pi_k N(\omega_{kj} - \omega_{ki}, \sigma_i^2 + \sigma_j^2) + \pi D.$$

Proof. For a proof, see Adams, Ferry, and Yoshida (2025). □

As a consequence of Lemma 2.2 and Theorem 3.1, when approximating $Y_{ij} + (\varepsilon_j - \varepsilon_i)$ by a Gaussian mixture, that the leftmost mixture component corresponds to the value of ω_{ij} we are interested in. That is, $\omega_{ij} = \min_k\{\mu_k\}$ where $\mu_k = \omega_{kj} - \omega_{ki}$ are the means of the Gaussian mixture in Theorem 3.1. This leads to the following estimator.

Corollary 3.2. *Assume that $\varepsilon_j \sim N(0, \sigma_j^2)$ with $\sigma_j > 0$ for $j \in V(G)$ and let $X^1, \ldots, X^N$ be an i.i.d. sample of the max-linear Bayesian network. If i is an ancestor of j, then*

$$\hat{\omega}_{ij} = \min_{\nu}(Y_{ij}^{\nu}) + \varepsilon_j - \varepsilon_i \sim N(\omega_{ij}, \sigma_i^2 + \sigma_j^2).$$

In particular, $N(\omega_{ij}, \sigma_i^2 + \sigma_j^2)$ occurs as a component of the mixture $Y_{ij} + (\varepsilon_j - \varepsilon_i)$.

Proof. By Lemma 2.2, the distribution of Y_{ij} contains an atom at ω_{ij} if i is an ancestor of j. By (4) this is in particular the minimum of the support of Y_{ij}. It follows from Theorem 3.1 that under the presence of noise this atom is replaced by the component $N(\omega_{ij}, \sigma_i^2 + \sigma_j^2)$. □

3.2 Geometric Estimation

In this section, we make use of the geometry of the polytrope associated to a MLBN. Since the facets of any polytrope are defined by Y_{ij}-hyperplanes, we can also estimate best-fit hyperplanes for the boundary of the support, turning the question of parameter estimation in Corollary 3.2 into an optimization problem. This point-of-view is advantageous when an edge is close to structural inactivation, or when the sample size N is small.

For a given sample $X^1, \dots, X^N$ with $X^\nu = (X_1^\nu, \dots, X_n^\nu)$ for $1 \leq \nu \leq N$, we need to solve the following optimization problem for $i < j$ and $i, j \in V(G)$ and $\nu = 1, \dots N$, where $\omega_{ij} \in \mathbb{R}$ and $\delta_{ij}^\nu \geq 0$ are decision variables:

Minimize	$K_1 \cdot \sum_{\nu=1}^{N} \sum_{i<j \in V(G)} \delta_{ij}^\nu + K_2 \cdot \sum_{i<j \in V(G)} \omega_{ij}^2$
with respect to	$\delta^\nu \in \mathbb{R}^{n \times n}, \nu \in [N]$ and $\omega \in \mathbb{R}^{n \times n}$
such that	$Y_{ij}^\nu \leq \omega_{ij} + \delta_{ij}^\nu$ and $\delta_{ij}^\nu \geq 0$

This is a dual optimization problem where the linear part of the constraints are known from (4) and the constants need to be found. The tuning parameters K_1 and K_2 allow us to put different emphasis on sharp boundaries in lieu of Lemma 2.2 compared to noisy, soft boundaries that are in line with Corollary 3.2.

4 Computational Experiments

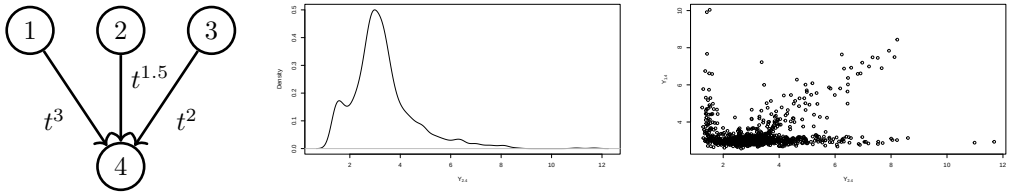
In this section, innovations Z_i are modeled as i. i. d. random variables following a Fréchet distribution with a common location parameter α, scale β, and shape ξ; that is, for each innovation we have

$$Z_i \sim \text{Fréchet}(\alpha, \beta = 1, \xi = 1), \text{ for some constant } \alpha \in \mathbb{R}.$$

Additionally, the standard deviation σ_i of the noise terms ε_i as defined in Section 3 are constrained to lie in the open interval $(0, .25]$ for all i. For a full explanation of the used software we refer to Adams, Ferry, and Yoshida (2025).

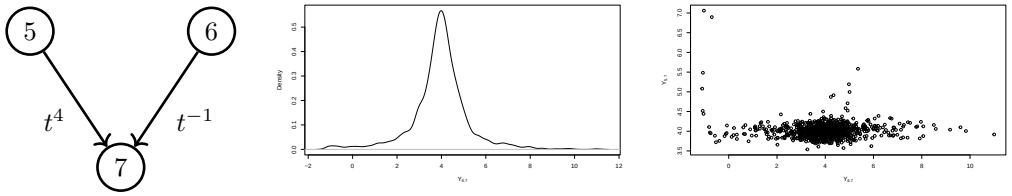
Example 4.1. Figure 2 shows an example of a random sample generated from the logarithms of the MLBN with Gaussian noise $N(0, 0.1)$ for all $j \in V(G)$. Knowing the structure of the network, we expect in accordance with Lemma 2.2 an atom in the distribution of Y_{14}, Y_{24} and Y_{34} each. For Y_{24}, this is shown in Figure 3b where there is a peak at $Y_{24} = 1.5$ corresponding to the value $\omega_{24} = 1.5$. In Figure 3c, we see a marginal picture of Y_{14} vs. Y_{24} with a horizontal boundary at $Y_{14} = 3$ and a vertical boundary at $Y_{24} = 1.5$.

The effectiveness of parameter estimation using GMMs is contingent upon the sample size and the extent to which each mixture component Y_{ij} is adequately represented. However, as the sample size decreases or if $i \to j$ is not adequately represented, the performance of GMM deteriorates, and the hyperplane method emerges as a more reliable alternative. This limitation is particularly notable in cases where a path in the network approaches structural inactivation.



(a) Graph structure (b) Density plot of Y_{24} (c) Marginal plot of Y_{14} vs. Y_{24}

Figure 2: A max-linear Bayesian network on 4 nodes with $N = 2000$ along with the density plot of Y_{24} and the marginal plot of Y_{14} vs. Y_{24}. These visualizations provide insights into the effectiveness of our methodology, naming the application of GMM and the geometric of the associated polytrope.



(a) Graph structure (b) Density plot of Y_{24} (c) Marginal plot of Y_{14} vs. Y_{24}

Figure 3: Illustration of structural inactivation in a MLBN, highlighting a case where the GMM-based method fails. The network structure, joint density plot, and marginal distribution plot of Y_{57} and Y_{67} are shown under the condition that the contribution of X_6 to X_7 is less than 1% of $N = 1000$.

Figure 3 illustrates the limiting behavior of an edge approaching structural inactivation. The empirical density plot of Y_{67} left exhibits heavy tailed behavior extending in both the positive and negative directions. This tail behavior is indicative of negligible dependency of a child from its parent vertex along an edge. Structurally, this behavior resembles that of an independent vertex, where observational data fail to provide definitive evidence of a causal relationship.

In Figure 3 we fixed the noise level σ and number N of observations. In doing this, we isolated a specific scenario that highlights the estimator's limitations. However, in a general setting, consistent estimation within the GMM framework is a function of noise levels σ_i and the number of atoms associated with edge $i \to j$.

Elevated noise levels σ_i obscure the separation between components, thereby increasing the minimum sample threshold required for accurate parameter estimation. Concurrently, the number of atoms associated with edge $i \to j$ must be sufficiently large, relative to the sample size and model complexity to guarantee reliable estimation of ω_{ij}.

To further assess the reliability of the GMM-based estimation procedure, we conducted an experiment fixing i. i. d. noise with standard deviation $\sigma = 0.1$ and systematically vary the sample size. We wish to determine the minimum number of observations along an

N	**Path %**	**Path obs.**	$\hat{\omega}_{ij}$
500	4.0%	20	0.097
1000	2.3%	23	-0.002
5000	1.48%	74	0.068
10000	1.32%	132	0.054
50000	1.16%	578	0.037

Table 1: Minimum number of edge-specific observations required for consistent estimation of $\omega_{ij} = 0$ under fixed noise $\sigma = 0.1$. For each size N, the corresponding percentage of observations along edge $i \to j$, the count of such observations, and the GMM estimate are reported. Estimates are presented on a $\log_2$ scale.

edge $i \to j$ required for the estimator to produce stable and accurate estimates for ω_{ij}.

In Table 1 we list for each sample size N the number of edge specific observations necessary for convergence, giving a quantitative lower bound on the signal strength required to ensure estimator consistency under fixed noise conditions. For example, in the largest of our experiments with $N = 50000$ the GMM estimator exhibits a significant upwards bias in estimating ω_{ij} when 1.16% of the total sample are associated with the edge $i \to j$. This behavior is indicative of a failure in statistical inference.

Remark 4.2. The reliability of the GMM-based estimation deteriorates as the noise level increases beyond $\sigma > 0.25$, due to the reduced separability of mixture components. However, the method also fails in the noiseless case, i. e. $\sigma = 0$, as the resulting Dirac measures are not identified within the GMM framework. In this case, the model violates fundamental assumptions of continuous mixture densities.

5 Conclusion

In this work, we presented two approaches for parameter estimation of MLBN under the presence of noise. The first, a GMM-based method, leverages the distributional regularization introduced by additive noise to infer latent edge weights through probabilistic mixture modeling.

The second method is based on the structure of associated polytropes. This approach formulates parameter estimation as a quadratic optimization problem and enables inference through facet alignment and directional projections, offering robustness when conventional statistical inference becomes unreliable.

References

Adams, Mark, Kamillo Ferry, and Ruriko Yoshida. *Inference for max-linear Bayesian networks with noise*, 2025. arXiv: `2505.00229 [stat.ML]`. (Cit. on pp. 7, 8).

Améndola, Carlos, and Kamillo Ferry. *Tropical combinatorics of max-linear Bayesian networks*, Nov. 15, 2024. arXiv: 2411.10394 [math.CO]. (Cit. on p. 5).

Améndola, Carlos, et al. "Conditional independence in max-linear Bayesian networks". *The Annals of Applied Probability* 32, no. 1 (Feb. 2022). ISSN: 1050-5164. https://doi.org/10.1214/21-aap1670. (Cit. on p. 4).

Buck, Johannes, and Claudia Kluppelberg. "Recursive max-linear models with propagating noise". *Electronic Journal of Statistics* (2020). https://api.semanticscholar.org/CorpusID:211678360. (Cit. on p. 3).

Einmahl, J. H. J., A. Kiriliouk, and J. Segers. "A continuous updating weighted least squares estimator of tail dependence in high dimensions." *Extremes* (2018). (Cit. on p. 1).

Engelke, Sebastian, and Adrien S. Hitz. "Graphical Models for Extremes". *Journal of the Royal Statistical Society* (2020). (Cit. on p. 1).

Gissibl, Nadine, and Claudia Klüppelberg. "Max-linear models on directed acyclic graphs". *Bernoulli* 24, no. 4A (2018): 2693–2720. https://doi.org/10.3150/17-BEJ941. (Cit. on p. 1).

— . "Max-linear models on directed acyclic graphs". *Bernoulli* 24, no. 4 (Nov. 1, 2018): 2693–2720. ISSN: 1350-7265. https://doi.org/10.3150/17-BEJ941. (Cit. on p. 5).

Gissibl, Nadine, Claudia Klüppelberg, and Steffen Lauritzen. *Identifiability and estimation of recursive max-linear models*, 1, Mar. 2021. https://doi.org/10.1111/sjos.12446. (Cit. on pp. 3, 6).

Joswig, Michael, and Katja Kulas. "Tropical and ordinary convexity combined". *Advances in Geometry* 10, no. 2 (Apr. 1, 2010): 333–352. ISSN: 1615-7168. https://doi.org/10.1515/advgeom.2010.012. (Cit. on p. 4).

Klüppelberg, Claudia, and Steffen Lauritzen. *Bayesian Networks for Max-linear Models*, 2019. arXiv: 1901.03948 [stat.ME]. (Cit. on pp. 3, 6).

Lauritzen, Steffen L. *Graphical models*. Reprinted 2004 with corrections. Oxford statistical science series. Oxford: Clarendon Press, 2004. ISBN: 9780198522195. (Cit. on p. 4).

Puente, María Jesús de la. "On tropical Kleene star matrices and alcoved polytopes". *Kybernetika* 49, no. 6 (2013): 897–910. ISSN: 0023-5954. https://doi.org/10338.dmlcz/143578. (Cit. on p. 5).

Tran, Ngoc M. *The tropical geometry of causal inference for extremes*, 2022. arXiv: 2207.10227 [math.ST]. (Cit. on p. 3).

Precomputing EMCC to Speed Up Causal Inference

Antonio Alves[1], Rafael Cabañas[1,2], and Antonio Salmerón[1,2]

[1]Department of Mathematics, University of Almería, Spain
[1]{*aga081,rcabanas,antonio.salmeron*}*@ual.es*
[2]Center for the Development and Transfer of Mathematical Research to Industry (CDTIME), University of Almería, Spain

Abstract

Structural causal models extend probabilistic graphical models to support causal and counterfactual reasoning; they distinguish between observed (endogenous) variables and latent (exogenous) variables. When exogenous distributions are unknown, some interventional and counterfactual queries cannot be identified. Recent methods attempt to bound unidentifiable queries by estimating the exogenous distributions, but their computational cost becomes prohibitive as model complexity increases. In this paper, we introduce precomputed expectation–maximization, a simple modification of the EMCC algorithm that identifies structures that remain constant across the iterations and evaluates them once in an upfront precomputation step. Experiments show that precomputed EM lowers runtime per iteration and yields progressively greater computational savings as the number of iterations and the structural complexity of the models grow.
Keywords: Structural causal models; expectation–maximization; causality; probabilistic graphical models.

1 Introduction

Structural causal models (SCMs) have emerged as the de facto framework for encoding and reasoning about cause-effect relationships (Pearl, 2009; Bareinboim et al., 2022). Formally, a SCM is a specialized probabilistic graphical model (PGM) that distinguished between endogenous (observable) and exogenous (latent) variables and incorporates structural equations in which the endogenous variables are functionally determined by the exogenous ones. This structure enables causal and counterfactual reasoning, allowing us to analyze how a system would behave under interventions of its variables or hypothetical scenarios. In many real-world domains, from epidemiology and economics to sociology and machine learning, exogenous variables must be inferred from purely observational data. This yields a partially specified SCM, where the unknown exogenous distributions must be estimated or bounded before any interventional or counterfactual queries can be addressed.

A variety of strategies have been proposed to bound non-identifiable causal queries in discrete SCMs. Kang and Tian (2012) first obtain bounds on causal queries through

systems of inequality constraints; however, the number of resulting constraints grows exponentially with model size, limiting its practical applicability. Bareinboim et al. (2022) propose bounding counterfactuals as a polynomial-programming problem and estimate credible intervals for the query using Monte-Carlo sampling. Similarly, Sachs et al. (2023) introduce a linear-programming framework for deriving bounds on these queries. Closely related to this work, Zaffalon et al. (2020) showed that any finite SCM can be transformed into an equivalent credal network (Cozman, 2000). This transformation requires solving a linear-programming problem for each exogenous variable; however, if the exogenous variables have a high cardinality, this approach may become infeasible. To mitigate this issue, Bjøru et al. (2024) proposed a divide-and-conquer strategy. Zaffalon et al. (2024) introduced EMCC for approximating bounds on non-identifiable queries; EMCC repeatedly applies the EM algorithm (Koller and Friedman, 2009) to obtain the specifications of the exogenous distributions. Unfortunately, each EM run becomes increasingly computationally demanding as the SCM's complexity grows.

This paper introduces the *precomputed expectation-maximization* method, which seeks to compute explicit specifications of the exogenous distributions, thereby enabling the evaluation of arbitrary causal and counterfactual queries. Precomputed EM exploits the structure of the SCM to identify and factor out repeated components in the EM updating rule, performing these calculations once upfront and eliminating redundant operations. The paper is structured as follows. Section 2 reviews structural causal models and EMCC. Section 3 introduces our precomputation statement and algorithm, illustrated with an example. Section 4 evaluates the method's performance on synthetic datasets across multiple model topologies. Section 5 concludes with directions for extending these ideas to more general causal frameworks.

2 Background

In this section, we introduce basic notation and provide background on fundamental concepts related to causal reasoning and the expectation-maximization algorithm within this framework.

2.1 Notation

In the context of general notation, we use uppercase letters (V) to denote random variables, lowercase letters (v) for specific values (or states), and Ω_V to represent the domain of V. Similarly, $\mathbf{V} = \{V_1, V_2, ..., V_n\}$ denotes a set of variables and $\mathbf{v}$ a specific joint realization of its domain, $\Omega_{\mathbf{V}} = \times_{V \in \mathbf{V}} \Omega_V$. We assume that $\Omega_{\mathbf{V}}$ is finite and discrete. The probability distribution over variables $\mathbf{V}$ is denoted by $P(\mathbf{V})$. For the sake of simplicity, we denote $P(v)$ as a shorthand of $P(V = v)$. In a directed graph, Pa_V represents the parents, that is, the immediate predecessors of V.

2.2 Structural causal models

Structural causal models (SCMs) are a type of probabilistic graphical models that serve as the foundational framework for causal reasoning. SCMs are formally defined as follows (Bareinboim et al., 2022):

Definition 2.1 (Structural causal model (SCM))**.** *A structural causal model is a 4-tuple* $\langle \boldsymbol{U}, \boldsymbol{X}, \mathscr{F}, \mathscr{P} \rangle$*, where*

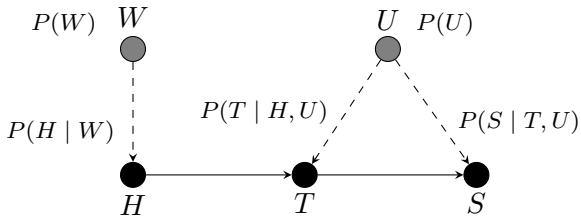
- $\mathbf{U}$ *is a set of exogenous variables that are determined by factors outside the model;*
- $\boldsymbol{X}$ *is a set of variables* $\{X_1, X_2, \ldots, X_n\}$*, called endogenous, that are determined by other (exogenous and endogenous) variables in the model, i.e. by variables in* $\boldsymbol{U} \cup \boldsymbol{X}$*.*
- $\mathscr{F}$ *is a set of functions* $\{f_{X_1}, f_{X_2}, \ldots, f_{X_n}\}$ *called structural equations (SEs), such that each of them is a function* $f_{X_i} : \Omega_{U_i} \cup \Omega_{\mathrm{Pa}_{X_i}} \to \Omega_{X_i}$*, where* $\mathrm{Pa}_{X_i} \subseteq \boldsymbol{X}$ *are the endogenous variables directly determining* X_i *and* $\boldsymbol{U}_i \subseteq \boldsymbol{U}$ *are the exogenous variables directly determining* X_i*.*
- $\mathscr{P}$ *is a set containing a probability distribution* $P(U)$ *for each* $U \in \boldsymbol{U}$*.*

Each SCM is associated with a directed acyclic graph (DAG) $\mathscr{G}$, known as the *causal graph*, where the nodes represent the variables in $\mathbf{U} \cup \mathbf{X}$. The edges connect a node in $\mathbf{U} \cup \mathrm{Pa}_{\mathbf{X}}$ to a node in $\mathbf{X}$, with each edge associated with a corresponding structural equation $f_{X_i} \in \mathscr{F}$.

To illustrate these concepts, consider the example depicted in Figure 1, which is an extension of the medical problem involving 700 patients introduced by Mueller et al. (2021). In the left part of the figure, the endogenous variables are represented as black nodes, specifically $\mathbf{X} = \{H, T, S\}$, where H denotes *hospital type*, T represents *treatment*, and S corresponds to *survival*. On the other hand, exogenous variables, $\mathbf{U} = \{U, W\}$, are represented as gray nodes. These variables act as root nodes in the graph and have endogenous variables as their children.

In the absence of expert knowledge, SEs can be inferred directly from the causal graph without loss of generality using *canonical specification* (Zhang et al., 2022). Consequently, the SEs f_W, f_T and f_S are deterministic degenerate conditional probability tables (CPTs), as shown in Figure 2, taking the form $P(H|W)$, $P(T|H, U)$ and $P(S|T, U)$. Meanwhile, the marginal distributions of the exogenous variables remain unknown, with their states encoding all possible deterministic mechanisms governing the relationships between these variables and their endogenous children.

Observe that in this example, each endogenous variable has exactly one exogenous parent; otherwise, the model is referred to as *non-Markovian*. Conversely, if every exogenous variable has exactly one endogenous child, the SCM is considered *Markovian*; if any exogenous variable has more than one endogenous child, it is known as *semi-Markovian*. This classification was introduced by Avin et al. (2005). In this paper, we consider both Markovian and semi-Markovian models. Note that the SCM in Figure 1 represents a



H	T	S	#
0	0	0	76
1	0	0	33
0	1	0	52
1	1	0	98
0	0	1	79
1	0	1	48
0	1	1	131
1	1	1	182

Figure 1: Example of a SCM (left) and associated data (right).

f_W

H	w_0	w_1
0	1	0
1	0	1

f_T

H	T	u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
0	0	1	1	1	0	0	0	0	0	0
0	1	0	0	0	1	1	1	1	1	1
1	0	1	1	1	1	1	1	0	0	0
1	1	0	0	0	0	0	0	1	1	1

f_S

T	S	u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
0	0	1	1	0	1	1	0	1	1	0
0	1	0	0	1	0	0	1	0	0	1
1	0	1	0	0	1	0	0	1	0	0
1	1	0	1	1	0	1	1	0	1	1

Figure 2: Structural equations.

semi-Markovian model, where T and S are *confounded components* (abbreviated as c-components) of U (Tian and Pearl, 2002).

Typically, exogenous variables are not directly observed, and only the endogenous variables are accessible through observational data. When the distributions of the exogenous variables are unknown, we denote these models as *partially specified* SCM (PSSCM). If the distributions of the exogenous variables are known, the model is referred to as a *fully specified* SCM. In the problem considered, we are given a PSSCM together with a dataset $\mathscr{D}$ containing observations on endogenous variables $\mathbf{X}$ (as shown in the right part of Figure 1), from which the empirical distribution $\widetilde{P}(\mathbf{X})$ can be computed. The main problem is how to define the distribution of the exogenous variables that is compatible with the observed data and, ultimately, derive a fully specified SCM from a partially specified SCM that aligns with the given data. Zaffalon et al. (2020) solve this issue by using the endogenous observations to impose linear constrains on the exogenous variables. This result in a non-deterministic system, that is, we have infinite compatible fully-specified SCMs for a PSSCM. However, a bound for the exogenous distributions can still be obtained.

2.3 EMCC

In practice, obtaining the exact bounds within which the exogenous distribution lies can be computationally intensive. To approximate the solution, Zaffalon et al. (2024) propose obtaining multiple distributions for the exogenous variables from the considered partially specified SCM and summarizing them through their lower and upper bounds. This approach is called *expectation-maximization for causal computation* (EMCC) and relies on the *expectation-maximization* (EM) algorithm to obtain these distributions.

The EM algorithm (Dempster et al., 1977) is a widely used method for parameter learning in PGMs. It optimizes the likelihood function by alternating between an expectation step (E-step), where expected sufficient statistics are computed, and a maximization step (M-step), where maximum likelihood estimation is applied to these statistics. A description of the EM approach for parameter learning in PGMs can be found in (Koller and Friedman, 2009).

In this context, the EMCC algorithm is applied to learn the latent variables $U \in \mathbf{U}$ within a Markovian or semi-Markovian model. Following Zaffalon et al. (2024), each iteration is formulated as follows:

$$\begin{aligned} \text{E-step:} \quad & M[u] = \sum_{(\mathbf{y},\mathbf{v}) \in \Omega_{\mathbf{Y},\mathbf{V}}} P_t\left(u|\mathbf{y},\mathbf{v}\right) \cdot M[\mathbf{y},\mathbf{v}] \\ \text{M-step:} \quad & P_{t+1}(u) = \frac{M[u]}{N}, \quad \forall u \in U, \end{aligned} \tag{1}$$

where $M[\cdot]$ denotes the count of occurrences of a realization and N the total count of observations. Note that the method in (1) improves computational efficiency compared to the basic EM approach. Rather than iterating over each individual observation in the dataset, we first aggregate the data by counting the occurrences of each configuration. The EM update is then performed over the distinct states, weighted by their frequencies. This allows the algorithm to process the data only once, significantly reducing the processing time.

3 Improvements

In line with the previous strategy, we leverage the structure of the causal network to minimize redundant computations at each iteration. As previously noted, the set of CPTs is deterministic, implying that any operation involving them remains unchanged across iterations. Consequently, precomputing these operations leads to a significant reduction in computational effort. This intuition is capture in the following theorem.

Theorem 3.1 (Precomputation of the EMCC)**.** *Let $\mathscr{M}$ be a Markovian or semi-Markovian model and let $\mathscr{D}$ be a dataset over the endogenous variables. The set of probabilities associated with the exogenous variables, $\{P(U)\}_{U \in \mathbf{U}}$, can be obtained by iteratively applying*

the following updating rule:

$$P_{t+1}(u) = \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} \frac{\phi_1(\mathbf{y},\mathbf{v},u)}{\sum_{u'\in\Omega_U}\phi_2(\mathbf{y},\mathbf{v},u')P_t(u')} P_t(u) \tag{2}$$

where $\mathbf{V}$ *denotes the set of endogenous children of* U*, and* $\mathbf{Y}$ *denotes the set of endogenous variables in* $\mathbf{X}\backslash\mathbf{V}$ *that are parents of* $\mathbf{V}$*. The quantities* ϕ_1 *and* ϕ_2 *are the precomputed factors defined as:*

$$\phi_1(\mathbf{Y},\mathbf{V},U) = \widetilde{P}\,(\mathbf{Y},\mathbf{V}) \cdot \prod_{V\in\mathbf{V}} P(V|Pa_V)$$
$$\phi_2(\mathbf{Y},\mathbf{V},U) = \prod_{V\in\mathbf{V}} P(V|Pa_V)$$

Proof. Consider the basic EM algorithm (Koller and Friedman, 2009) applied to the generalized semi-Markovian causal model:

$$\begin{aligned} &\text{E-step:} \quad M[u] = \sum_{(\mathbf{y},\mathbf{v})\in\mathcal{D}} P_t\,(u|\mathbf{y},\mathbf{v}) \\ &\text{M-step:} \quad P_{t+1} = \frac{M[u]}{N}. \end{aligned} \tag{3}$$

Before applying the method, we aggregate the data for each state of the endogenous variables, denoted as $M[\mathbf{y},\mathbf{v}]$. This aggregation involves calculating the frequency of each configuration of endogenous variables in the dataset. Then, we sum over these states, which leads to the following form of the EM algorithm, as shown in (1):

$$\begin{aligned} &\text{E-step:} \quad M[u] = \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} P_t\,(u|\mathbf{y},\mathbf{v}) \cdot M\,[\mathbf{y},\mathbf{v}]\,, \\ &\text{M-step:} \quad P_{t+1}(u) = \frac{M[u]}{N}. \end{aligned} \tag{4}$$

Combining both steps in (4) into a single process, and introducing the total number of observations:

$$\begin{aligned} P_{t+1}(u) &= \frac{1}{N} \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} P_t\,(u|\mathbf{y},\mathbf{v}) \cdot M\,[\mathbf{y},\mathbf{v}] \\ &= \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} P_t\,(u|\mathbf{y},\mathbf{v}) \cdot \widetilde{P}\,(\mathbf{y},\mathbf{v})\,, \end{aligned} \tag{5}$$

where $\widetilde{P}\,(\mathbf{y},\mathbf{v})$ represents the empirical probability for the endogenous observations. We now proceed by expanding the posterior probability as follows:

$$
\begin{aligned}
P_t\left(u|\mathbf{y},\mathbf{v}\right) &= \frac{P(\mathbf{y},\mathbf{v},u)}{P(\mathbf{y},\mathbf{v})} \\
&= \frac{P(\mathbf{y},\mathbf{v},u)}{\sum_{u'\in\Omega_U} P(\mathbf{y},\mathbf{v},u')} \\
&= \frac{P(u)\cdot\prod_{V\in\mathbf{V}} P(v|pa_V)}{\sum_{u'\in\Omega_U} P(u')\cdot\prod_{V\in\mathbf{V}} P,(v|pa_V)},
\end{aligned} \tag{6}
$$

and substituting (6) into (5) to obtain:

$$
P_{t+1}(u) = \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} \frac{\widetilde{P}\left(\mathbf{y},\mathbf{v}\right)\cdot\prod_{V\in\mathbf{V}} P(v|pa_V)\cdot P(u)}{\sum_{u\in\Omega_U}\prod_{V\in\mathbf{V}} P(v|pa_V)\cdot P(u)}. \tag{7}
$$

Note that each $P(v|pa_V)$ is a realization of a CPT. Instead, define the product of the CPTs as the factor product:

$$
\begin{aligned}
\phi_1(\mathbf{Y},\mathbf{V},U) &= \widetilde{P}\left(\mathbf{Y},\mathbf{V}\right)\cdot\prod_{V\in\mathbf{V}} P(V|Pa_V), \\
\phi_2(\mathbf{Y},\mathbf{V},U) &= \prod_{V\in\mathbf{V}} P(V|Pa_V).
\end{aligned} \tag{8}
$$

Observe that the empirical distribution remains unchanged across iterations and, as a result, can be included in the precomputation. By substituting the specific realization of the precomputed functions in (8) into (7), we obtain:

$$
P_{t+1}(u) = \sum_{(\mathbf{y},\mathbf{v})\in\Omega_{\mathbf{Y},\mathbf{V}}} \frac{\phi_1(\mathbf{y},\mathbf{v},u)}{\sum_{u'\in\Omega_U}\phi_2(\mathbf{y},\mathbf{v},u')P_t(u')} P_t(u) \tag{9}
$$

□

Example 3.2. *We illustrate the precomputation of EMCC (Theorem 3.1) using the semi-Markovian model and data from Section 2 (Figures 1, 2). Recall the endogenous variables (H,T,S) and exogenous variables U,W. For clarity, we denote the state of 0 of each endogenous variable by h_0,t_0,s_0 and the state of 1 by h_1,t_1,s_1, respectively. In this example, we show (i) the precomputation phase and (ii) a single iteration of the algorithm that updates the distribution of $P(U)$ using the precomputed factors.*

Initialization. *Draw U at random:*

$$
\begin{array}{rcccccccccc}
 & u_0 & & u_1 & & u_2 & & u_3 & & u_4 & \\
P(U) &= [0.0648 & & 0.2646 & & 0.0417 & & 0.0502 & & 0.0158 & \\
 & u_5 & & u_6 & & u_7 & & u_8 & & & \\
 & 0.2251 & & 0.0842 & & 0.0877 & & 0.1605] & & &
\end{array}
$$

Precompute the factors. *Compute*

$$
\begin{aligned}
\phi_1(H,T,S,U) &= \widetilde{P}(H,T,S)\cdot P(T|H,U)\cdot P(S|T,U), \\
\phi_2(H,T,S,U) &= P(T|H,U)\cdot P(S|T,U).
\end{aligned}
$$

We first compute ϕ_2, the product of deterministic CPTs:

$$\phi_2(H,T,S,U) = \begin{array}{cc} & \begin{matrix} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \end{matrix} \\ \begin{matrix} h_0 \\ \\ h_1 \\ \end{matrix} & \begin{matrix} t_0 \\ t_1 \\ t_0 \\ t_1 \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}.$$

$$\begin{array}{cc} & \begin{matrix} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \end{matrix} \\ \begin{matrix} t_0 \\ \\ t_1 \\ \end{matrix} & \begin{matrix} s_0 \\ s_1 \\ s_0 \\ s_1 \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} =$$

$$\begin{array}{cc} & \begin{matrix} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \end{matrix} \\ \begin{matrix} h_0 \\ \\ \\ \\ h_1 \\ \\ \\ \\ \end{matrix} & \begin{matrix} t_0 \\ \\ t_1 \\ \\ t_0 \\ \\ t_1 \\ \\ \end{matrix} \begin{matrix} s_0 \\ s_1 \\ s_0 \\ s_1 \\ s_0 \\ s_1 \\ s_0 \\ s_1 \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

The remaining factor, ϕ_1, is obtained by multiplying ϕ_2 with the empirical distribution $\widetilde{P}(H,T,S)$.

$$\phi_1(H,T,S,U) = \begin{array}{cc} & \begin{matrix} s_0 & & s_1 & \\ t_0 & t_1 & t_0 & t_1 \end{matrix} \\ \begin{matrix} h_0 \\ h_1 \end{matrix} & \begin{bmatrix} 0.109 & 0.075 & 0.113 & 0.187 \\ 0.050 & 0.140 & 0.066 & 0.260 \end{bmatrix} \end{array} \cdot \phi_2(H,T,S,U) =$$

$$\begin{array}{cc} & \begin{matrix} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \end{matrix} \\ \begin{matrix} h_0 \\ \\ \\ \\ h_1 \\ \\ \\ \\ \end{matrix} & \begin{matrix} t_0 \\ \\ t_1 \\ \\ t_0 \\ \\ t_1 \\ \\ \end{matrix} \begin{matrix} s_0 \\ s_1 \\ s_0 \\ s_1 \\ s_0 \\ s_1 \\ s_0 \\ s_1 \end{matrix} \begin{bmatrix} 0.109 & 0.109 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.113 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.075 & 0 & 0 & 0.075 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.187 & 0.187 & 0 & 0.187 & 0.187 \\ 0.050 & 0.050 & 0 & 0.050 & 0.050 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.066 & 0 & 0 & 0.066 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.140 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.260 & 0.260 \end{bmatrix} \end{array}$$

EM update for $P(U{=}u_2)$***.*** *The general EM-step is*

$$P_1(u_2) = \sum_{h,t,s \in \Omega_{H,T,S}} \frac{\phi_1(H=h, T=t, S=s, u_2)}{\sum_{u' \in \Omega_U} \phi_1(H=h, T=t, S=s, u')\, P(u')}\, P(u_2).$$

For the specific state $U = u_2$*, the factor* ϕ_1 *is non–zero only for the two configurations* $(H=0, T=0, S=1)$ *and* $(H=1, T=0, S=1)$*. The same rule can be applied for* ϕ_2*. Consequently, the update reduces to a sum over just these two configurations, greatly simplifying the computation.*

$$\begin{aligned} P_1(u_2) &= \left(\frac{\phi_1(0,0,1,u_2)}{\phi_2(0,0,1,u_2)\, P_0(u_2)} + \frac{\phi_1(1,0,1,u_2)}{\phi_2(1,0,1,u_2)\, P_0(u_2) + \phi_2(1,0,1,u_5)\, P_0(u_5)} \right) P_0(u_2) \\ &= \left(\frac{0.113}{1 \cdot 0.0417} + \frac{0.066}{1 \cdot 0.0427 + 1 \cdot 0.2251} \right) 0.0417 \\ &= 0.1233. \end{aligned}$$

Repeat for all u***.*** *We apply the same formula to each of the nine states of* U $u_0, \ldots, u_8$*.*

Resulting distribution. *After updating every entry, we obtain*

$$\begin{array}{rccccccccc} & u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \\ P_1(U) = [& 0.003 & 0.121 & 0.1233 & 0.0343 & 0.008 & 0.142 & 0.187 & 0.1254 & 0.229] \end{array}$$

4 Empirical validation

To quantify the computational behavior of our **precomputed EM** strategy in Theorem 3.1 and the **standard EM** formulation in (1), we performed experiments on simulated semi-Markovian causal models, focusing on the runtime performance and scalability under different structure complexities.

We considered four distinct DAG topologies derived from the initial configuration presented in Figure 1. Starting from this basic setup, we incrementally increased complexity by adding additional child nodes to the confounded components, resulting in the four configurations illustrated in Figure 3. Each topology was instantiated with ten independent models, yielding a total of 40 models. For each instance, conditional probability tables (CPTs) were set according to the canonical specification of the SEs. For each generated model, we randomly initialized the exogenous variables and sampled 10,000 observations of the endogenous variables $\mathbf{X}$ to form the dataset $\mathscr{D}$, from which the empirical distribution $\widetilde{P}(\mathbf{X})$ was computed.

The main results are summarized in Figure 4, where we report the average runtime for the precomputed EM and the standard EM algorithms continuously for each iteration up to a maximum of 20. Each subfigure corresponds to one of the four DAG topologies, with reported runtimes averaged over the ten independently generated models. From

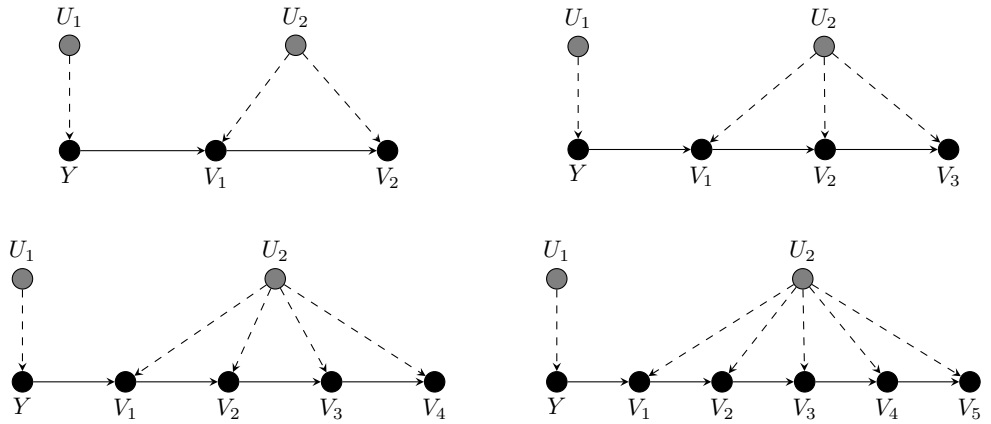


Figure 3: SCM topologies with an increasing number of children in a confounded component.

this empirical analysis, we observed clear computational trends demonstrating distinct advantages of our precomputed EM algorithm.

Initially, at the setup stage (iteration zero), the precomputed EM exhibited slower execution times relative to the standard EM algorithm. The preliminary cost aligns with theoretical expectations, reflecting the additional computations required to calculate and store these precomputed factors before the iterative phase. However, the initial delay is quickly outweighed by substantial efficiency gains in later iterations. During the initial iterations, the precomputed EM matches the runtime of the standard EM; as the number of iterations increases, it progressively outperforms the latter, underscoring the long-term computational advantage of the proposed approach.

Moreover, the experiments reveal a clear relationship between the size of a confounded component and the speed-up obtained by precomputed EM. As the number of child variables in a confounded component grows, the relative computational advantage of the precomputation strategy becomes increasingly pronounced. This trend is particularly relevant since larger confounded components involve more substantial computational overhead per iteration when using the standard EM approach, thereby amplifying the benefits provided by precomputing invariant factors.

5 Conclusion and future work

In this work, we have addressed the computational bottleneck of the EMCC algorithm to learn the distributions of exogenous variables in semi-Markovian SCM. By exploiting the fact that some CPTs remain unchanged throughout EM iterations, we showed (Theorem 3.1) how all deterministic operations can be precomputed at once, before the iterative phase begins. This yields a new *precomputed EM* update rule (2) whose per-iteration cost growth is less sensitive to increasing complexity in the DAG configuration.

Our empirical study demonstrated that, although precomputed EM incurs an up-

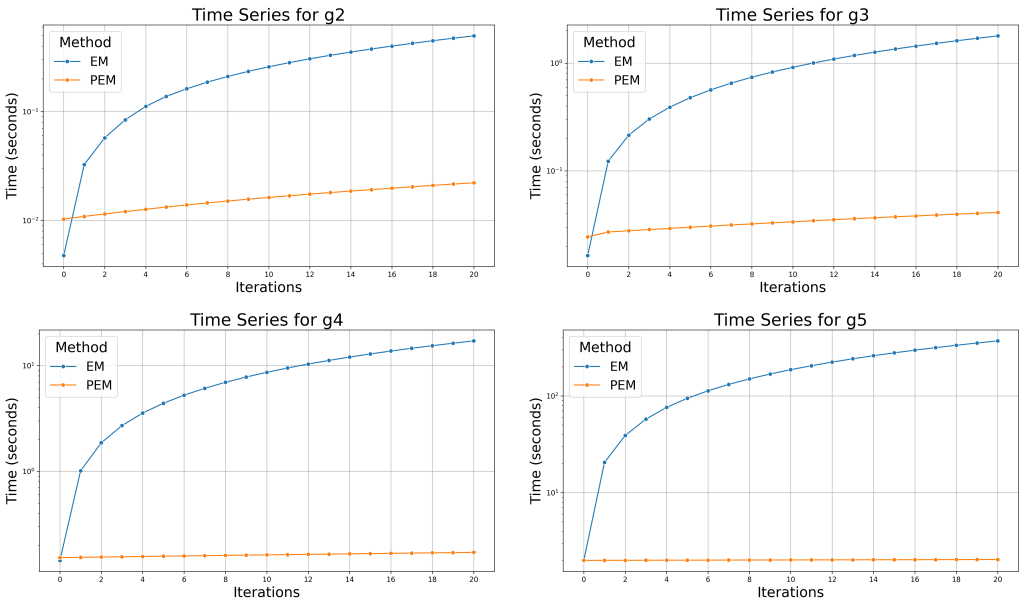


Figure 4: Computation time across the four DAG topologies

front overhead, it quickly surpasses the standard EM implementation. In particular, its advantage becomes more pronounced as (i) the number of EM iterations grows and (ii) the size of each confounded component increases. These results confirm that precomputed EM makes expectation-maximization feasible in settings where repeated CPT evaluations would otherwise make causal inference via EMCC prohibitive.

A current limitation of our approach is the memory consumed by storing the full set of precomputed factors, which scales with the number of endogenous configurations in a confounded component. In future work, we will explore tree-based representations that support pruning to reduce storage demands while preserving the bulk of the computational gains afforded by precomputation.

Acknowledgments

Grant PID2022-139293NB-C31 funded by MICIU/AEI/10.13039/501100011033 and by ERDF "A way of making Europe". Authors acknowledge the University of Almería Research and Transfer Programme funded by "Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía" through the European Regional Development Fund (ERDF), Operation Programme 2021-2027. Programme: Research and Innovation 54.A. In brief: PPIT-UAL, Junta de Andalucía- ERDF 2021-2027. Programme: 54.A. RC was also supported by "Plan Propio de Investigación y Transferencia 2024-2025" from University of Almería under the project P_LANZ_2024/003.

References

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 357–363, 2005.

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pages 507–556. 2022.

A. R. Bjøru, R. Cabañas, H. Langseth, A. Salmerón Cerdán, et al. A divide and conquer approach for solving structural causal models. 2024.

F. G. Cozman. Credal networks. *Artificial intelligence*, 120(2):199–233, 2000.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

C. Kang and J. Tian. Inequality constraints in causal models with hidden variables. *arXiv preprint arXiv:1206.6829*, 2012.

D. Koller and N. Friedman. Inference in graphical models. In *Probabilistic Graphical Models: Principles and Techniques*, chapter 19, pages 849–942. MIT Press, 2009.

S. Mueller, A. Li, and J. Pearl. Causes of effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730*, 2021.

J. Pearl. *Causality*. Cambridge university press, 2009.

M. C. Sachs, G. Jonzon, A. Sjölander, and E. E. Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In *International Conference on Probabilistic Graphical Models*, pages 581–592. PMLR, 2020.

M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, and D. Azzimonti. Efficient computation of counterfactual bounds. *International Journal of Approximate Reasoning*, 171: 109111, 2024.

J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pages 26548–26558. PMLR, 2022.

Common Cause Condition for Universal Approximation

Nihat Ay[1,2,3] and Leon Sierau[1]

[1]Hamburg University of Technology, Hamburg, Germany
{*nihat.ay,leon.sierau*}*@tuhh.de*
[2]Santa Fe Institute, Santa Fe, NM, USA
[3]Leipzig University, Leipzig, Germany

Abstract

In this article, we use the common cause principle to derive a necessary criterion for universal approximation. More precisely, a particular formalisation of the common cause principle within the context of Judea Pearl's causality theory allows us to infer the existence of a common cause based on the statistical dependence of the observed variables. Universal approximation has to incorporate all possible dependence structures of observed variables, which is only possible if the underlying network allows for common causes of higher order.

Keywords: Common cause principle, information theory, common ancestors, universal approximation.

1 Introduction

This article is based on the *common cause principle*, attributed to Reichenbach (1956). It postulates that the dependence of two variables X_2 and X_2 requires one of the following three cases to hold true:

1. X_1 is a cause of X_2,
2. X_2 is a cause of X_1, or
3. There exists a common cause of X_1 and X_2.

This principle can be easily derived within the context of Pearl's theory of causation (Pearl, 2000). In (Steudel and Ay, 2015), an extension of this principle to more than two variables has been formulated based on information-theoretic methods. Basically, it states that whenever the dependence of m observed variables $X_1, \ldots, X_m$ exceeds a particular

threshold that depends on c, $1 \leq c \leq m-1$, we can infer the existence of a variable from which there exists paths to at least $c+1$ of the observed variables. This provides a way to relate the dependence of observed variables to the underlying causal structure. In this article, this dependence will be used to connect the common cause principle to a necessary structural condition for universal approximation. More precisely, we consider a feed-forward network of binary stochastic units. The expressive power of such a network is maximal if it can approximate all stochastic maps, Markov kernels, which assign to an input state $x_1, \ldots, x_n$ a probability distribution over the output states $y_1, \ldots, y_m$.

A necessary condition for a feed-forward network to have maximal expressive power is that each input unit is connected to each output unit via at least one directed path. In particular, this implies that all output units must have at least one common cause, giving rise to the common cause condition for universal approximation. The necessity is intuitively clear. That is, because no mapping from a particular input unit to a particular output unit can be approximated in a network that does not allow for information transfer from the input to the output unit in question. A proof is given in Section 2. Further structural requirements on the depth and width of sigmoid belief networks for universal approximation have been studied by Merkh and Montúfar (2022).

This article outlines an approach from causality theory which complements previous studies. Our main result, stated in Section 3, derives the necessary common cause condition outlined above from the information-theoretic generalization of the common cause principle presented in (Steudel and Ay, 2015). In Section 4, this simple criterion is compared with necessary requirements based on the parameter counting argument.

2 A simple graphical criterion for universal approximation

We will associate various kinds of maps with feed-forward networks. To incorporate all of them, consider two non-empty finite sets X and Y. The set of all stochastic mappings from X to Y is denoted by $\mathcal{K}$. (The letter "$\mathcal{K}$" stands for *kernel*, more precisely *Markov kernel*, a term that is also used to denote a stochastic map.) More formally,

$$\begin{aligned}\mathcal{K} &:= \left\{ K : \mathsf{X} \times \mathsf{Y} \to [0,1], \quad (x,y) \mapsto K(y|x) \ : \ \sum_y K(y|x) = 1 \ \text{ for all } x \in \mathsf{X} \right\} \\ &\subseteq \mathbb{R}^{\mathsf{X}\times\mathsf{Y}}.\end{aligned}$$

Note that this is a convex polytope in $\mathbb{R}^{\mathsf{X}\times\mathsf{Y}}$ with extreme points $\mathrm{Ext}\,(\mathcal{K})$ which can be identified with the deterministic maps $\mathsf{X} \to \mathsf{Y}$. To each deterministic map $f : \mathsf{X} \to \mathsf{Y}$, we assign the kernel

$$K^f(y|x) := \begin{cases} 1, & \text{if } y = f(x) \\ 0, & \text{otherwise} \end{cases} .$$

With this definition, we have

$$\mathrm{Ext}\,(\mathcal{K}) \ = \ \left\{ K^f \ : \ f : \mathsf{X} \to \mathsf{Y} \right\}.$$

Finally, $\mathcal{K}$ carries the natural topology induced by $\mathbb{R}^{\mathsf{X}\times\mathsf{Y}}$.

We now consider a finite set N of units consisting of *input units* I and *computational units* Λ. The set Λ of computational units is subdivided into *output units* O and *hidden units* H. For simplicity, all units are assumed to be binary, that is, the corresponding variables X_i take values in $\{\pm 1\} = \{-1, +1\}$. We are particularly interested in $\mathcal{K}_{O|I}$ defined for $\mathsf{X} = \{\pm 1\}^I$ and $\mathsf{Y} = \{\pm 1\}^O$. These are the kind of input-output maps that we ultimately want to represent or approximate by a family of input-output maps. Below, this family will be obtained in terms of a feed-forward network. As an intermediate step, we incorporate the hidden units and also consider $\mathcal{K}_{O,H|I} = \mathcal{K}_{\Lambda|I}$ defined for $\mathsf{X} = \{\pm 1\}^I$ and $\mathsf{Y} = \{\pm 1\}^O \times \{\pm 1\}^H = \{\pm 1\}^\Lambda$. This is the set of maps from the input to the hidden *and* output states. In order to obtain maps from the input to the output states only, we have to marginalise out the hidden states. This is obtained in terms of the marginalisation map

$$\pi_O : \ \mathcal{K}_{O,H|I} \ \to \ \mathcal{K}_{O|I}$$

defined by

$$K(x_O|x_I) \ = \ \sum_{x_H} K(x_O, x_H|x_I).$$

At this point, a *model* $\mathcal{M}$ is simply a subset of $\mathcal{K}_{O,H|I}$. The image of $\mathcal{M}$ under π_O, denoted by $\mathcal{M}_{O|I}$, consists of those maps that can be represented by the network.

Definition 1 (Universal approximator)**.** Let $\mathcal{M}$ be a model in $\mathcal{K}_{O,H|I}$. We call $\mathcal{M}$ a *universal approximator* (*in* $\mathcal{K}_{O|I}$), if every element of $\mathcal{K}_{O|I}$ can be approximated arbitrarily well by elements of $\mathcal{M}_{O|I} = \pi_O(\mathcal{M})$. We can reformulate this in terms of the closure of $\mathcal{M}_{O|I}$:

$$\mathcal{K}_{O|I} \ = \ \mathrm{cl}\left(\mathcal{M}_{O|I}\right).$$

The model $\mathcal{M}$ is a *deterministic universal approximator* (*in* $\mathcal{K}_{O|I}$), if every element of $\mathrm{Ext}\left(\mathcal{K}_{O|I}\right)$ can be approximated arbitrarily well by elements of $\mathcal{M}_{O|I}$, or, equivalently, $\mathrm{Ext}\left(\mathcal{K}_{O|I}\right) \subseteq \mathrm{cl}\left(\mathcal{M}_{O|I}\right)$. Clearly, every universal approximator is a deterministic universal approximator.

We now examine the properties of Definition 1 under further marginalisation of the output states. More precisely, consider

$$O \ \subseteq \ O' \ \subseteq \ \Lambda,$$

and the corresponding marginalisation

$$\pi_{O',O} : \ \mathcal{K}_{O'|I} \ \to \ \mathcal{K}_{O|I},$$

defined by

$$\pi_{O',O}(K(x_O \mid x_I)) \ := \ \sum_{x_{O'\setminus O}} K(x_O, x_{O'\setminus O} \mid x_I).$$

Clearly, we have $\pi_O = \pi_{O',O} \circ \pi_{O'}$, and $\mathcal{M}_{O'|I}$ is mapped onto $\mathcal{M}_{O|I}$. It is easy to see that

$$\pi_{O',O}\left(\mathrm{Ext}(\mathcal{K}_{O'|I})\right) \ = \ \mathrm{Ext}(\mathcal{K}_{O|I}). \tag{1}$$

This follows from the fact that $\pi_{O',O}$ is an affine map between convex sets, that is

$$\pi_{O',O}((1-t)K_1 + tK_2) \;=\; (1-t)\pi_{O',O}(K_1) + t\pi_{O',O}(K_2), \tag{2}$$

for all $K_1, K_2 \in \mathcal{K}_{O'|I}$ and $t \in [0,1]$.

Proposition 2 (Marginal universality)**.** *Let $\mathcal{M}$ be a model in $\mathcal{K}_{\Lambda|I}$, and let $O \subseteq O' \subseteq \Lambda$. Then the following holds:*

1. *If $\mathcal{M}$ is a universal approximator in $\mathcal{K}_{O'|I}$ then it is also a universal approximator in $\mathcal{K}_{O|I}$.*
2. *If $\mathcal{M}$ is a deterministic universal approximator in $\mathcal{K}_{O'|I}$ then it is also a deterministic universal approximator in $\mathcal{K}_{O|I}$.*

Proof. The proof is divided into three steps where we abbreviate $\pi_{O',O}$ by π.

Step 1: Consider the open set $\mathcal{U} := \mathcal{K}_{O|I} \setminus \mathrm{cl}(\mathcal{M}_{O|I})$. In what follows, we show that the open preimage $\mathcal{U}' := \pi^{-1}(\mathcal{U})$ and the closed set $\mathrm{cl}\left(\mathcal{M}_{O'|I}\right)$ are disjoint.

Obviously, we have

$$\pi^{-1}\left(\mathrm{cl}(\mathcal{M}_{O|I})\right) \;\supseteq\; \pi^{-1}\left(\mathcal{M}_{O|I}\right) \;\supseteq\; \mathcal{M}_{O'|I}. \tag{3}$$

Given that the set on the LHS of (3) is closed, due to the continuity of π, this implies

$$\pi^{-1}\left(\mathrm{cl}(\mathcal{M}_{O|I})\right) \;\supseteq\; \mathrm{cl}\left(\mathcal{M}_{O'|I}\right). \tag{4}$$

This finally yields

$$\begin{aligned} \mathcal{U}' \cap \mathrm{cl}(\mathcal{M}_{O'|I}) &\overset{(4)}{\subseteq} \pi^{-1}(\mathcal{U}) \cap \pi^{-1}\left(\mathrm{cl}(\mathcal{M}_{O|I})\right) \\ &= \pi^{-1}\left(\left(\mathcal{K}_{O|I} \setminus \mathrm{cl}(\mathcal{M}_{O|I})\right) \cap \mathrm{cl}(\mathcal{M}_{O|I})\right) \\ &= \pi^{-1}(\emptyset) \\ &= \emptyset. \end{aligned} \tag{5}$$

Step 2: We show

$$\mathcal{E}' \;\subseteq\; \mathrm{cl}(\mathcal{M}_{O'|I}) \quad\Rightarrow\quad \mathcal{E} \;:=\; \pi(\mathcal{E}') \;\subseteq\; \mathrm{cl}(\mathcal{M}_{O|I}). \tag{6}$$

We prove this by contradiction and assume that there exists $E \in \mathcal{E} \setminus \mathrm{cl}(\mathcal{M}_{O|I})$. Then the open set $\mathcal{U} = \mathcal{K}_{O|I} \setminus \mathrm{cl}(\mathcal{M}_{O|I})$ contains E and is therefore non-empty. Choose $E' \in \mathcal{E}'$ satisfying $\pi(E') = E$. Clearly, E' is an element of $\mathcal{U}' = \pi^{-1}(\mathcal{U})$. By (5), $\mathcal{U}'$ and $\mathrm{cl}(\mathcal{M}_{O'|I})$ are disjoint, so that $E' \in \mathcal{E}' \setminus \mathrm{cl}(\mathcal{M}_{O'|I})$. That means $\mathcal{E}' \;\not\subseteq\; \mathrm{cl}(\mathcal{M}_{O'|I})$.

Step 3: We can finally verify the two statements of the proposition by choosing $\mathcal{E}'$ and $\mathcal{E}$ in (6) appropriately. For universal approximation, we set

$$\mathcal{E}' \;:=\; \mathcal{K}_{O'|I}, \quad \text{and} \quad \mathcal{E} \;:=\; \pi(\mathcal{E}') \;=\; \mathcal{K}_{O|I}.$$

The statement about deterministic universal approximation is obtained by setting

$$\mathcal{E}' \;:=\; \mathrm{Ext}\left(\mathcal{K}_{O'|I}\right), \quad \text{and} \quad \mathcal{E} \;:=\; \pi(\mathcal{E}') \;\overset{(1)}{=}\; \mathrm{Ext}\left(\mathcal{K}_{O|I}\right).$$

□

In what follows, we specify the kind of models $\mathcal{M}$ we consider in this article. We interpret input-output maps as being generated by a feed-forward network consisting of layers $L_0, L_1, \ldots, L_D$ where $L_0 = I$ is the input layer and $L_D = O$ is the output layer. Thus, $N = L_0 \cup L_1 \cup \cdots \cup L_D$, and $\Lambda = N \setminus L_0$. Given that only L_0 and L_D are observable, we refer to the layers $L_1, \ldots, L_{D-1}$ as the hidden layers, and their union is denoted by H. To define the network, we now consider directed edges between the nodes. We assume that couplings are only allowed between neighbouring layers, that is $(i, i') \in E$ and $i \in L_k$ always implies $i' \in L_{k+1}$. In particular, units within a layer are not connected via edges, and edges cannot jump over a layer. We write n for the number of input units and m for the number of output units, that is $|L_0| = n$ and $|L_D| = m$.

We now define models in terms of parametrisations that are consistent with the given network structure. More precisely, for each unit $i \in \Lambda$, we consider a parametrisation

$$\theta_i \;\mapsto\; K_{\theta_i}.$$

With $\theta := (\theta_i)_{i\in\Lambda}$ we define a kernel $K_\theta \in \mathcal{K}_{O,H|I}$ by

$$K_\theta(x_O, x_H|x_I) \;:=\; \prod_{i\in\Lambda} K_{\theta_i}(x_i|x_{\mathrm{pa}(i)}). \tag{7}$$

Collecting all these kernels K_θ, we obtain the model

$$\mathcal{M} \;:=\; \{K_\theta \;:\; \theta \in \Theta\} \;\subseteq\; \mathcal{K}_{\Lambda|I},$$

which we refer to as a *feed-forward model*.

Proposition 2 implies a simple graphical requirement for a feed-forward model $\mathcal{M}$ to be a (deterministic) universal approximator in $\mathcal{K}_{O|I}$, which is stated in the following proposition.

Proposition 3 (Graphical condition for universal approximation)**.** *Let $\mathcal{M}$ be a feed-forward model which we assume to be a deterministic universal approximator in $\mathcal{K}_{O|I}$. Then:*

(1) For every input unit $i \in I$ and every output unit $o \in O$, there exists a directed path from i to o.

(2) In consequence, there exists a common cause $i \in N$ of all output units $o \in O$. We refer to this as the common cause condition (for determinsitic universal approximation).

Proof. Assume that there exist $i \in I$ and $o \in O$ that are not connected by a directed path from i to o. Then X_o is conditionally independent of X_i given $X_{I\setminus i}$ for all $K_\theta \in \mathcal{M}$. Therefore, $\mathcal{M}_{O|I}$ is contained in the closed set $\mathcal{C}$ of kernels $K \in \mathcal{K}_{O|I}$ for which X_o is conditionally independent of X_i given $X_{I\setminus i}$. This implies that also its closure, $\mathrm{cl}(\mathcal{M}_{O|I})$, is contained in $\mathcal{C}$. Clearly, the complement of $\mathcal{C}$ is an open and non-empty set. It consists of those kernels for which X_o is not conditionally independent of X_i given $X_{I\setminus i}$. Such a map is given, for instance, by the copy map which assigns as output of o simply the input of i. □

3 Information-theoretic derivation of the common cause condition

The main result of this article is concerned with deriving the second criterion of Proposition 3 by making use of Theorem 4 from (Steudel and Ay, 2015), stated below. It relates the stochastic dependencies of observed variables to the underlying causal structure given in terms of a directed acyclic graph $G = (N, E)$. A probability distribution p of random variables $X_N = (X_i)_{i\in N}$ is said to factorise according to G, if it satisfies

$$p(x_N) \;=\; \prod_{i\in N} p(x_i|x_{\mathrm{pa}(i)}).$$

Theorem 4. *Let $G = (N, E)$ be a directed acyclic graph, and let $X_N = (X_i)_{i\in N}$ be random variables with a distribution that factorises according to G. For a subset O of N, define the set A_{c+1} of all common ancestors of O of order $c+1$, that is*

$$A_{c+1} \;:=\; \{j \in N \;:\; j \text{ reaches more than } c \text{ units in } O\}\,.$$

Then, with

$$I_c(X_O) \;:=\; \sum_{k\in O} H(X_k) - c\cdot H(X_O), \tag{8}$$

the entropy of these common ancestors is lower bounded as

$$H(X_{A_{c+1}}) \;\geq\; \frac{1}{|O|-c}\, I_c(X_O). \tag{9}$$

In particular, if $I_c(X_O)$ on the RHS of (9) is positive, then there exists a unit $j \in N$ and corresponding units $k_1, \ldots, k_{c+1} \in O$ such that $j \rightsquigarrow k_i$, $i = 1, \ldots, c+1$.

Let us now relate the setting of this theorem to the setting of a feed-forward network, as introduced in Section 2. Clearly, the feed-forward network represents a directed acyclic graph G. Furthermore, if we multiply a kernel of the structure (7) with the uniform input distribution μ on $\{\pm 1\}^I$, then we obation a joint distribution p on $\{\pm 1\}^N$ that factorises according to G. Below, we will require a reformulation the function I_c, defined by (8), to a function on kernels $K \in \mathcal{K}_{O|I}$:

$$I_c :\; \mathcal{K}_{O|I} \;\to\; \mathbb{R}, \qquad K \;\mapsto\; I_c(K) := I_c(K_*(\mu)). \tag{10}$$

Stated in words, we simply take the image of the uniform distribution μ with respect to K, the K-push forward of μ, and then evaluate I_c of it, as defined by (8). Obviously, the map (10) is continuous (in K).

Theorem 5. *The common cause condition for deterministic universal approximation, second condition of Proposition 3, follows from the information-theoretic inequality (9) of Theorem 4.*

Proof. To prove this theorem, we define a deterministic map $f : \mathsf{X} \to \mathsf{Y}$ that cannot be approximated arbitrarily well if the condition is not satisfied. For that, we enumerate the units of the input and output layer as $L_0 = \{1, 2, \ldots, n\}$ and $L_D = \{1, 2, \ldots, m\}$, respectively. We then define f as the map that simply copies the state of the first input unit into the states of all output units, that is $f : \{\pm 1\}^n \to \{\pm 1\}^m$, $(x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_1)$.

We now evaluate $I_c(K^f)$, as defined by (10). It is easy to see that image of the uniform distribution with respect to K^f is given by

$$K^f_*(\mu)(x_O) \;=\; \begin{cases} 0.5 & \text{if } x_i = -1 \text{ for all } i \in O \\ 0.5 & \text{if } x_i = +1 \text{ for all } i \in O \\ 0 & \text{otherwise} \end{cases} . \tag{11}$$

That implies

$$I_{|O|-1}(K^f) \;=\; \sum_{i \in O} H(X_i) - (|O|-1)H(X_O) \;=\; |O|\,\ln 2 - (|O|-1)\ln 2 \;=\; \ln 2.$$

Now we choose $0 < \varepsilon < \ln 2$. Given that I_c is continuous, the preimage $\mathcal{U}$ of the open interval $]\ln 2 - \varepsilon, \ln 2 + \varepsilon[$ is an open neighbourhood of K^f in $\mathcal{K}_{O|I}$. If K^f can be approximated arbitrarily well by the model then *every* open neighbourhood of K^f, in particular $\mathcal{U}$, has a non-empty intersection with $\mathcal{M}_{O|I}$. Therefore, there exists $K \in \mathcal{M}$, such that $\pi_O(K) \in \mathcal{U}$, and therefore

$$\ln 2 - \varepsilon \;<\; I_{|O|-1}(\pi_O(K)) \;<\; \ln 2 + \varepsilon.$$

In particular, $I_{|O|-1}(\pi_O(K)) > 0$. Clearly, $\mu \otimes K$ factorises according to G and by Theorem 4, we have a common ancestor of order $|O| = m$. □

Figure 1 visualizes the common cause condition for deterministic universal approximation. Furthermore, it shows that the first condition of Proposition 3 is stronger than the second one. This is because the first condition would also exclude both networks shown in Figure 1, whereas the common cause condition only excludes the network on the RHS.

The common cause condition implies that if we upper bound the number of outgoing edges of each unit, universal approximation can only be achieved if the network is deep enough. In the following corollary of Theorem 5, we consider a simple instance of this intuition.

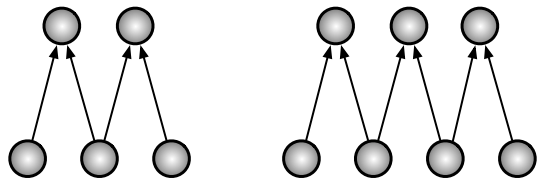


Figure 1: The network on the LHS is not excluded by the common cause condition because there is an input node that reaches all output nodes. This condition is not satisfied for the network on the RHS.

Corollary 6. *Assume that in the feed-forward network each node of a layer L_i can reach at most $\rho \geq 2$ nodes of the next layer L_{i+1}, and consider a corresponding feed-forward model $\mathcal{M}$. Then for $\mathcal{M}$ to be a deterministic universal approximator it is necessary that the depth of the network, D, satisfies*

$$D \;\geq\; \frac{\ln m}{\ln \rho}. \tag{12}$$

In particular, condition (12) is necessary for universal approximation.

Proof. By iteration, each node of layer L_i can reach at most ρ^k nodes of layer L_{i+k}. Now, for a depth D of the network with

$$\rho^D \;<\; m$$

there is no input node that reaches all output nodes so that the network cannot be a deterministic universal approximator according to Theorem 5. Thus, it is necessary that

$$\rho^D \;\geq\; m,$$

which is equivalent to (12). □

If we set $m = \rho^k$ in (12) then we obtain $D \geq k$. In the example shown in Figure 2, $\rho = 2$, and $m = 4 = \rho^2$, so that the depth D has to be at least 2. However, as one can see, this is not sufficient for the network to have a node that reaches all output nodes.

Note that while the first criterion of Proposition 3 is stronger than the second, no stronger condition on the necessary depth of the network follows from it. More precisely, if $\rho^D \geq m$ it is possible that every input unit reaches every output unit.

The following requirement on the depth of the feed-forward network is based on condition (1) of Proposition 3 and cannot be derived from condition (2), the common cause condition. Instead of bounding the number of outgoing edges, we now bound the number of incoming edges. In the neuroscience terminology, this corresponds to having narrow receptive field (i.e, each neuron in a layer L_i receives inputs from at most a bounded number of neurons in the previous layer L_{i-1}). A condition on the necessary depth similar to the condition of Corollary 6 follows directly.

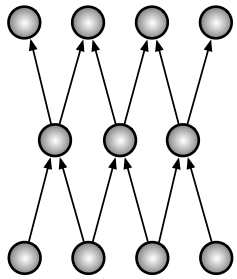


Figure 2: For this network, we have four output nodes, that is $m = 4$, and each node of one layer connects to at most two nodes of the following layer, that is $\rho = 2$. The necessary condition (12) then translates to $D \geq 2$. However, there is no node in the network that reaches all output nodes.

Corollary 7. *Assume that in the feed-forward network each node of a layer L_i can receive input from at most $\rho \geq 2$ nodes of the previous layer L_{i-1}, and consider a corresponding feed-forward model $\mathcal{M}$. Then for $\mathcal{M}$ to be a deterministic universal approximator it is necessary that the depth of the network, D, satisfies*

$$D \;\geq\; \frac{\ln n}{\ln \rho}. \tag{13}$$

Proof. By the same iteration argument as in the proof of Corollary 6, every unit of the output layer L_D can be reached by at most ρ^D units of the input layer. Hence, for all output units to be reached by all input units, a necessary condition immediately follows with

$$\rho^D \;\geq\; n \quad \Leftrightarrow \quad D \;\geq\; \frac{\ln n}{\ln \rho}.$$

□

4 Stochastic neurons and the parameter counting argument

The typical way to obtain necessary conditions for universal approximation is based on simple parameter counting. To be more precise, we consider a stochastic neuron $j \in L_1 \cup \cdots \cup L_D$ which is defined in terms of a weight vector $w_j = (w_{ij})_{i \in \mathrm{pa}(j)}$ and a threshold ϑ_j. The probability for the stochastic neuron j to generate the state $+1$ is then given by

$$K_{w_j,\vartheta_j}(+1|x_{\mathrm{pa}(j)}) \;:=\; \frac{1}{1 + e^{-2\left(\sum_{i \in \mathrm{pa}(j)} w_{ij} x_i - \vartheta_j\right)}}.$$

If each node of the network has at most ρ outgoing edges, then the total number of parameters is upper bounded by

$$\rho \cdot (|L_0| + \cdots + |L_{D-1}|) + |L_1| + \cdots + |L_D|. \tag{14}$$

For simplicity, let us consider the special case where all layers have width m. Then the bound (14) reduces to

$$D \cdot m \cdot (\rho + 1)\,. \tag{15}$$

We obtain the same bound on the number of parameters, if each computational node has at most ρ incoming edges. More precisely, the number of parameters is then upper bounded by

$$\rho \cdot (|L_1| + \cdots + |L_D|) + |L_1| + \cdots + |L_D|$$

which also reduces to (15) if all layers have the same width m. For a feed-forward model to be a universal approximator, the bound (15) has to exceed the dimension of $\mathcal{K}_{O|I}$, the set of the input-output kernels, that is

$$D \cdot m \cdot (\rho + 1) \;\geq\; 2^m(2^m - 1).$$

This is equivalent to

$$D \;\geq\; \frac{2^m(2^m - 1)}{m \cdot (\rho + 1)}. \tag{16}$$

Condition (16) of the parameter counting argument is typically stronger than the conditions (12) and (13), respectively. However, Proposition 3 equips us with very simple graphical conditions to rule out universal approximation and thereby complement parameter counting.

Consider for example an arbitrarily deep network with some bound ρ on the number of children (or parents) of each unit and $D \gg \frac{2^m(2^m-1)}{m\cdot(\rho+1)}$. If there exists a pair (i, o), $i \in I, o \in O$ such that there exists no directed path $(i \rightsquigarrow o)$, the corresponding feed-forward model cannot be a universal approximator. This situation is displayed in Figure 3.

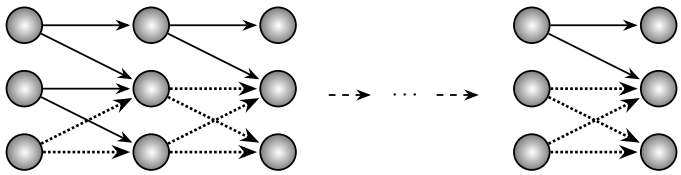


Figure 3: A deep network with $\rho = 2$, $m = 3$ and $D \gg \frac{2^m(2^m-1)}{m\cdot(\rho+1)}$. All edges that lie on directed paths originating at the bottom-most input neuron are drawn as dotted arrows. Because the same sparse connection pattern is repeated between every pair of successive layers, the top-most output neuron is unreachable from the bottom-most input neuron. Hence, although the network meets the depth condition suggested by the parameter-counting argument, it still falls short of being a universal approximator.

References

T. Merkh and G. Montúfar. *Stochastic Feedforward Neural Networks: Universal Approximation*, page 267–313. Cambridge University Press, 2022.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

H. Reichenbach. *The Direction of Time*. University of California Press, 1956.

B. Steudel and N. Ay. Information-theoretic inference of common ancestors. *Entropy*, 17 (4):2304–2327, 2015.

Transparency and Accuracy? Models for Bonus Allocation in the Age of Regulation

Vladislav Bína[1], Mojmír Sabolovič[2], and Stanislav Tripes[2]

[1]Department of Data Analytics, Faculty of Management,
Prague University of Economics and Business
vladislav.bina@vse.cz
[2]Department of Management, Faculty of Management,
Prague University of Economics and Business
{*mojmir.sabolovic,stanislav.tripes*}*@vse.cz*

Abstract

Artificial Intelligence and Machine Learning systems are increasingly being deployed in Human Resources (HR), including sensitive tasks such as bonus allocation. This paper benchmarks modeling techniques – ranging from transparent white-box models (stepwise regression, Bayesian network, fuzzy rules) to high-performing black-box algorithms (random forest, gradient boosting, neural networks) – on small sample HR data. In addition, a large language model (LLM) was tasked with deriving human-understandable rules directly from the training set, offering a novel alternative aligned with explainability requirements. The results reveal that while black-box models deliver superior predictive accuracy, they struggle to meet AI Act transparency obligations. The study concludes with recommendations for balancing performance, interpretability, and regulatory compliance and outlines the possibilities for the integration of decision support systems using interpretable methods.

1 Introduction

The deployment of Artificial Intelligence (AI) in Human Resources Management (HR) has moved from experimental proofs of concept to routine practice in recruitment, performance evaluation, and bonus allocation. McKinsey's 2024 global AI survey (McKinsey Global Institute, 2024) reports that AI has moved into the mainstream: more than three-quarters of organizations already apply it in at least one business function, and almost a third of large companies use algorithms for critical HR decisions such as hiring, performance evaluation, and employee development. Although only 12 percent of respondents

report routine use of generative AI in HR, adoption is accelerating. Large enterprises both spearhead uptake and redesign internal processes to embed AI, and HR emerges as one of the functions with the highest intensity of AI deployment and perceived impact.

At the same time, prominent failures have revealed the legal and ethical risks of opaque "black-box" models. The (Larsson et al., 2024) argues that bias is an intrinsic feature of HR AI systems because they inherit training data patterns and their algorithms are hard to audit. Citing documented cases in which automated hiring tools disadvantaged candidates by gender or ethnicity, the study calls for tighter transparency requirements and stronger regulatory oversight.

1.1 Regulatory background

In March 2024 the Council of the European Union adopted *Artificial Intelligence Act* (Regulation (EU) 2024/1689, (European Union, 2024), hereafter 'AI Act'), the first horizontal, risk-based regulation of AI systems worldwide[1]. Applications that evaluate or classify individuals for employment, promotion, or termination and monitor or evaluate performance and behavior are explicitly listed as *high-risk* (Annex III 4). Providers of such high-risk systems must therefore (i) establish a risk management system, (ii) register the AI system in the public EU database, (iii) generate comprehensive technical documentation, (iv) guarantee human oversight, (v) ensure traceability and appropriate accuracy, robustness, and cybersecurity, (vi) conduct data governance, establish a quality management system(ISACA, 2024).

From a modeling perspective, these requirements shift attention to **transparency** and **interpretability**. Whereas large language models, deep neural networks, or gradient-boosting ensembles often deliver state-of-the-art predictive accuracy, their inner logic is nontrivial to convey to nontechnical stakeholders. The emerging literature on eXplainable AI (XAI) documents this tension between accuracy and comprehensibility (Molnar and Freiesleben, 2024; Shwartz-Ziv and Tishby, 2017). In HR contexts, the risk is amplified: models operate on small, heterogeneous data sets that combine behavioral, transactional, and demographic attributes, making them vulnerable to both sampling variance and indirect discrimination(Bujold et al., 2024; Capasso et al., 2024).

1.2 Research gap

Recent empirical studies have benchmarked white-box and black-box algorithms in standardized machine learning repositories (Kruschel et al., 2025; Fischer et al., 2023). However, analogous comparisons of *real HR data sets* - which typically contain *many* correlated predictors but observations of *few* - remain scarce. Moreover, just few works evaluate models simultaneously on (i) predictive performance, (ii) formal interpretability, and (iii) explicit compliance with the AI Act transparency clauses. Addressing this gap is timely and relevant: legal scholars, advisory companies, and international organizations warn that business leaders 'risk sleepwalking toward AI misuse' in HR if they continue

[1]The regulation was published in the EU Official Journal on July 12th 2024 and enters into force on August 1st 2024 with an implementation period of 24 months for most provisions.

to prioritize performance over auditability(Herrera-Poyatos et al., 2025; Marwala, 2025; Mökander et al., 2021; Oxborough et al., 2018).

1.3 Objectives and contributions

The authors aim to make the following contributions.

1. We created eight prediction models that span the white-box/black-box spectrum, i.e. stepwise linear regression, a hybrid Bayesian network, a Mamdani-type fuzzy rule system, random forest, gradient-boosting regression, and a single-hidden-layer neural network.

2. Use the training and testing data set in the ChatGPT large language model and compare the prediction results with the other models.

3. Using an anonymized data set of employee performance, evaluation and annual bonuses from the Faculty of Management (38 employees, 59 potential predictors), we benchmark these models on a hold-out test set using several common error metrics (RMSE, MAE, RMSLE, *etc.*).

4. We map every model to the transparency obligations of the AI Act and discuss trade-offs between numeric accuracy and regulatory fitness.

2 Methods and their properties

2.1 Data set

The empirical study uses an anonymized personnel table used by the management of faculty for periodic evaluation, planning, and decision-making. The table contains *38* employees (rows) and *59* candidate predictors (columns), including

- simplified organizational role, study-administration engagement, department,
- quantitative teaching load (guaranteed and taught courses, English-taught share, etc.),
- research indicators (publications, citations, research grants),
- service indicators (expert panels, state-exam committees, outreach activities),
- the continuous target variable concerning employee annual bonuses.

The feature-to-observation ratio ($p/n \approx 1.6$) typifies the 'wide but shallow' setting (Ahn, 2006) that can be encountered in HR analytics of smaller organizations. The predictors mix numerical, ordinal, and nominal scales and can contain sporadic missing values.

2.2 Pre-processing

(a) **Cleaning.** Related columns with vast majority of zeros were added together (like counts in categories of impact factors journals or different types of mobilities and outreach activities).

(b) **Cleaning.** Non-informative columns were dropped (variants of citation counts correlated with WOS citations). Character variables were converted to factors.

(c) **Missing values.** There were no missing values.

(d) **Feature reduction.** We were searching for near-zero-variance predictors. However, these were already removed in earlier phases. Similarly, similar to rare factor levels ($< 5\%$ frequency), which were already removed in the categorization phase of the faculty function variable.

(e) **Train–test split.** An 80:20 stratified split (stratified on the binary flag bonus>0) was fixed with the setup of a pseudorandom number generator seed.

(f) **Scaling.** For algorithms sensitive to magnitude (artificial neural network), the numeric input was centered and scaled; the target was transformed by $\log(1 + y)$ when required.

2.3 Model catalogue

Following the transparency imperative of the AI Act, we explicitly include *white-box* techniques whose internal logic can be inspected directly and add problematic (but powerful) state-of-the-art *black-box* learners complemented with partial post hoc explanations. The comparison of both categories can highlight the possible loss caused by regulatory environment in EU. Table 1 summarizes key characteristics.

Table 1: Overview of candidate models. "Interpretable in principle" means that coefficients, rules or CPTs can be inspected without auxiliary tools.

Family	Concrete algorithm	Transparency level
White-box	Stepwise linear regression (AIC)	coefficients visible
	Conditional Gaussian BN	graph + CPTs, causal queries
	Fuzzy rule system	linguistically readable rules
Dubious	Estimates by LLM (ChatGPT)	explanation by system of rules?
Black-box	Random forest	variable importance + SHAP
	Extreme gradient boosting	tree-SHAP explanations
	Multilayer perceptron (nnet)	post-hoc SHAP, Integ. gradients
	Gradient b. regr. via ChatGPT	tree-SHAP explanations

2.3.1 White-box models

Stepwise linear regression (LM–AIC). We fit an ordinary least squares model (Chambers, 1992) and apply bidirectional AIC search to control over-parameterization (Venables and Ripley, 2002). The significant number of zero values in the response variable led us to the idea of a more sophisticated model. The variant of Tweedie GLM (Dunn and Smyth, 2005) was rejected due to harder interpretability, zero inflated Poisson, or negative binomial regression (Zeileis et al., 2008), which are usable only for count data (bonuses in general cannot be regarded as count data). The last possibility considered was the combination of logistic regression model (zero or nonzero bonus) and GLM Gamma with log link (for nonzero bonuses only) (Venables and Ripley, 2002). The last possibility is interpretable but is infeasible for a small sample of data with a higher number of features.

Hybrid Bayesian network (BN–CG). Discrete predictors remain categorical, continuous predictors (including the logarithmic transformation of the bonus) are treated as Gaussian nodes. A hill-climbing structure search with BIC–CG score learns the directed acyclic graph under the constraint that continuous nodes cannot be parents of discrete nodes (Scutari, 2010). The model parameters are then learned using MLE-CG, which is a combination of maximum likelihood estimators for conditional probability tables (discrete nodes) and least squares regression models (continuous nodes) (Azzimonti et al., 2019). The network gives predictions (insertion of evidence), contains directly interpretable conditional distributions (Lauritzen, 1996), and provides a possibility of causal interpretation (Pearl, 2009).

Fuzzy rule system (FRS-WM). We employ the fuzzy rule approach introduced in (Mamdani and Assilian, 1975) and implemented in the FRBS R package (Riza et al., 2015). Each input variable domain is partitioned into five Gaussian membership functions. The Wang–Mendel algorithm (Wang and Mendel, 1992) is generally used to handle regression tasks and induces interpretable IF–THEN rules; defuzzification employs a weighted average. The output membership sets are defined on the $\log(1 + y)$ scale to avoid dominance of the zero cluster, and we used a finer partition into nine membership functions.

2.3.2 Black-box models

Random forest (RF). We used Breiman's random forest algorithm (based on the Breiman and Cutler procedure) used for classification and regression (Breiman, 2001). We grow 500 trees with one third of candidate splits for the regression task. Out-of-bag error guides hyper-parameter tuning.

Partial interpretability is provided by feature importance scores (Gini importance) which give only global explanations and SHapley Additive Explanations (SHAP) values which provide more informative contributions of single features to particular predictions (Lundberg and Lee, 2017).

Extreme gradient boosting (XGB). XGBoost is a scalable end-to-end tree boosting system able to work with sparse data and weighted quantile sketch for approximate tree learning (Chen and Guestrin, 2016). Gradient-boosted trees with lower learning rate 0.1 which is more robust to overfitting but slower to compute. Therefore, it we incresed number of iterations to 300 rounds and limited the maximal tree depth to 6.

Although we ranked this approach among black-box models, it has a partial interpretability. SHAP interaction values can be used for local explanations (Jabeur et al., 2024).

Single-hidden-layer multilayer perceptron. Using `nnet`, we train a basic example of the artificial neural network approach, a single hidden layer multilayer perceptron (Ripley, 1996) with a size of 20 neurons in the hidden layer with skipped connection of input and output. The L2-decay was set to 0.01 and a maximum of 1000 iterations were performed on z-scored predictors and log-scaled target (Venables and Ripley, 2002). The output clipping prevents numeric overflow.

Gradient Boosting Regression (GBR) in ChatGPT. We asked ChatGPT with the GPT-4.5 model to train on the same example of training set used in the above methods and then to provide predictions for cases in the test set (done only with a single partition used in the white-box approaches above). The large language model generated Python code and used a Gradient Boosting Regressor algorithm implemented in the scikit-learn Python package (Pedregosa et al., 2011), which is closely related to the XGB approach mentioned above and based on Friedman's gradient boosting algorithm (Friedman, 2001).

Gradient boosting iteratively builds an additive model by sequentially fitting weak decision trees to the residual errors of previous ensemble members. The hyperparameters were tuned using cross-validation and the model handled categorical variables using one-hot encoding. Interpretability is only partial, can be realized by analyzing feature importance derived from the trained model and providing insights into the global contributions of individual predictors.

Large Language Model-based Heuristic (LLM). We employed a topical heuristic approach using a large language model (LLM), specifically GPT-4.5 (OpenAI, 2025), to derive a prediction directly from training data. Rather than using traditional numerical optimization methods, the LLM analyzed the training data set, identified key features impacting bonuses (such as administrative roles, academic positions, publication activities, and teaching workload), and formulated explicit IF–THEN decision rules. The predictions were manually derived by applying these rules to new data.

The following prompt was used with the anonymized training and testing CSV files:

> Simply study the "train" file as a large language model to understand how bonuses are allocated. Then, without training any machine learning models, attempt to assign bonuses to the "test" file based on the same principles.

ChatGPT itself declares this method to be interpretable, transparent, and suitable for scenarios that require high explainability and compliance with regulatory requirements

such as the AI Act (European Union, 2024). Since this is just a declaration appearing as an output of LLM and we were not able to control the real procedure. Also, we performed a single run, since we cannot be sure whether the LLM under same account does not share previous results. Therefore, we labeled the method as 'dubious' in Table 1.

2.4 Hyper-parameter tuning

As we briefly mentioned above, the white-box models used internal criteria (AIC, BIC, rule strength) while the black-box learners were tuned by five-fold cross-validation in the training split, minimizing the root mean square error (RMSE). The kind reader can imagine that this aspect of model optimization can be advanced based on user's more advanced experience with particular models, and models could be fine-tuned even further. However, we did not focus on this aspect; we aimed to present a comparison of traditional approaches with the capabilities of emerging large language model tools.

2.5 Evaluation metrics

To assess accuracy, we report seven scalar metrics commonly used in regression benchmarking (Hyndman and Koehler, 2006):

RMSE	root-mean-squared error	$\sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$
MAE	mean absolute error	$\frac{1}{n}\sum_i \lvert y_i - \hat{y}_i \rvert$
MedAE	median absolute error	$\operatorname{median} \lvert y_i - \hat{y}_i \rvert$
RMSLE	RMSE in log space	$\sqrt{\frac{1}{n}\sum_i (\log(1+y_i) - \log(1+\hat{y}_i))^2}$
MAPE	mean absolute percentage error	$\frac{1}{n}\sum_i \frac{\lvert y_i - \hat{y}_i \rvert}{y_i + \varepsilon}$
sMAPE	symmetric MAPE	$\frac{1}{n}\sum_i \frac{\lvert y_i - \hat{y}_i \rvert}{(\lvert y_i \rvert + \lvert \hat{y}_i \rvert + 2\varepsilon)/2}$
R^2	coefficient of determination	$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$

Where $\varepsilon = 10^{-6}$ prevents division by zero for zero bonuses. RMSLE is used to predict the correct order of magnitude of a variable, since it uses a logarithm, it produces NaN if negative prediction appears. All metrics were evaluated on the previously unseen test set to emulate real deployment.

3 Results

Predictive accuracy. Gradient boosting (XGBoost) delivers the second lowest *RMSE* (5 400 CZK) and a strong R^2 (0.92), marginally ahead of the LLM-generated gradient boosting (5 431 CZK). The large language model (LLM) rule set produced by ChatGPT attains the best *RMSE* (4 899 CZK), *MAE* (2 000 CZK) and perfect median error, indicating few large outliers but an acceptable central tendency. Among *white-box* methods,

Table 2: Test–set accuracy of all candidate models (values rounded to two decimals).

Model	RMSE	MAE	MedAE	MAPE	sMAPE	R^2	RMSLE
LM–step AIC	31 703.2	26 920.7	27 224.4	1.02×10^{10}	1.60	0.32	–
BN–hybrid	6 918.6	4 125.5	970.5	2.57×10^{7}	1.66	0.66	3.82
FRS–WM	12 331.3	7 993.4	3 997.2	1.90×10^{6}	1.33	0.55	4.85
Random Forest	10 883.8	8 957.1	7 759.8	6.56×10^{9}	1.15	0.13	6.64
XGBoost	5 400.3	5 065.2	5 343.0	3.18×10^{9}	1.12	0.92	6.18
MLP (nnet)	24 692.4	16 270.1	7 526.9	2.98×10^{8}	1.69	0.41	4.78
ChatGPT – GBR	5 431.2	4 209.0	3 129.4	1.29×10^{9}	1.18	0.69	4.90
ChatGPT – LLM	4 899.0	2 000.0	0.0	0.3	0.14	0.37	0.37

the hybrid Bayesian network clearly outperforms stepwise linear regression and the fuzzy rule system, reducing the RMSE by roughly 80%.

Interpretability and compliance. Stepwise LM, the Bayesian network and the fuzzy rule base provide intrinsic explanations (coefficients, conditional-probability tables, linguistic rules) and therefore satisfy the AI Act's transparency clause without auxiliary tools. Random forest and XGBoost require post-hoc SHAP analysis, while MLP is essentially opaque; all three are consequently classified as high-risk systems that are not transparent and sufficiently interpretable. Therefore, it cannot be used as an automated decision system even with human oversight.

The ChatGPT LLM solution produces amazing accuracy of prediction. It exactly guessed five out of six bonuses, where the last one in reality was set lower for reasons not contained in the HR evaluation data set. However, this unbelievable model strength in this specific task on a small sample needs to be tested in a correct scientific manner, not by anecdotic evidence as in this case. The LLM formulates compact natural-language rules, but its provenance and training data remain opaque, raising traceability concerns.

Let us present example of a LLM rule-based decision (prediction of bonus) in case of two emloyees accompanied by a thorough justification and argumentation.

Employee 3
Assigned bonus: 0 CZK
Criteria:

- **Position:** Without administrative involvement.
- **Study administration involvement:** Not involved.
- **Academic activity:** Low to average, 35 citations, 4 new, without significant publishing activity.
- **Teaching activity:** Low workload (0.6 FTE), average student-hours, and only 2 courses taught.
- **Other activities:** Higher participation in faculty-organized events (24), but low mobility and very limited training.

Conclusion: The employee does not meet the threshold of activity that would justify awarding an extraordinary bonus.

Employee 6
Assigned bonus: 25,000 CZK
Criteria:

- **Position:** Member of the Academic Senate, implying a higher level of responsibility.
- **Study administration involvement:** Actively engaged in study administration.
- **Academic activity:** Slightly lower publication activity (8 citations in total, only 1 new), but this lower level is compensated by other factors.
- **Teaching activity:** Very high teaching load (2,282 student-hours) and a large number of taught courses.
- **Other activities:** Significant participation in training sessions (29), solid involvement in faculty-organized events (14).

Conclusion: The employee demonstrates a high level of commitment in administrative roles, teaching, and education, fully justifying a higher bonus of 25,000 CZK.

Overall assessment. On this "wide-but-shallow" HR data set the performance gap between the interpretable and black-box models narrows: the Bayesian network achieves an *RMSE* only 28% higher than XGBoost while maintaining full explainability.

Given the preference of the EU AI Act for transparent systems, the hybrid BN offers the best trade-off between precision and regulatory fitness, while pure black-box models will not be usable as a support for decision-making in high-risk fields like HR.

4 Conclusion and future directions

This study demonstrates the applicability and comparative strengths of eight modeling approaches for the allocation of bonuses in a small academic organization. Although black-box models, particularly ensemble methods such as gradient boosting, excel in predictive performance, they fall short of the stringent transparency and interpretability requirements imposed by the AI Act. In contrast, white-box models, like stepwise regression, Bayesian networks, and fuzzy rules—offer interpretable structures but struggle with predictive accuracy in sparse data contexts. The LLM approach showed good prediction accuracy accompanied by rule-based argumentation, but failed to deliver complete transparency and auditability.

Despite these insights, several limitations should be addressed in future research. More extensive data sets and longitudinal evaluation are necessary to validate the stability of model performance over time. In addition, refining the LLM heuristic approach

through reinforcement learning from human feedback (RLHF), integration with formal rule-learning algorithms could enhance its consistency and traceability and thorough experimental design.

Our findings underscore that the trade-off between predictive power and regulatory compliance will remain a key challenge for AI applications in high-risk decision support system resources, especially under the evolving European AI regulatory framework.

References

J. Ahn. *High dimension, low sample size data analysis.* PhD thesis, The University of North Carolina at Chapel Hill, 2006.

L. Azzimonti, G. Corani, and M. Zaffalon. Hierarchical estimation of parameters in bayesian networks. *Computational Statistics & Data Analysis*, 137:67–91, 2019. doi: 10.1016/j.csda.2019.02.004.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

A. Bujold, I. Roberge-Maltais, X. Parent-Rocheleau, J. Boasen, S. Sénécal, and P.-M. Léger. Responsible artificial intelligence in human resources management: a review of the empirical literature. *AI and Ethics*, 4(4):1185–1200, 2024.

M. Capasso, P. Arora, D. Sharma, and C. Tacconi. On the right to work in the age of artificial intelligence: Ethical safeguards in algorithmic human resource management. *Business and Human Rights Journal*, pages 1–15, 2024.

J. M. Chambers. *Linear Models.* Wadsworth & Brooks/Cole, Pacific Grove, CA, 1992.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

P. K. Dunn and G. K. Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15:267–280, 2005.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 168, 12 July 2024, pp. 1–189, 2024. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689`. Entered into force 1 August 2024.

S. F. Fischer, M. Feurer, and B. Bischl. Openml-ctr23 – a curated tabular regression benchmarking suite. In *Proceedings of the AutoML 2023 Workshop*, 2023. URL `https://openreview.net/forum?id=HebAOoMm94`.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

A. Herrera-Poyatos, J. D. Ser, M. L. de Prado, F.-Y. Wang, E. Herrera-Viedma, and F. Herrera. Responsible artificial intelligence systems: A roadmap to society's trust through trustworthy ai, auditability, accountability, and governance, 2025. URL `https://arxiv.org/abs/2503.04739`.

R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001. Discusses MAE, RMSE, MAPE, sMAPE, ME and related metrics, and contrasts their statistical properties.

ISACA. Understanding the eu ai act: Requirements and next steps. White paper, ISACA, 2024. URL `https://www.isaca.org/resources/white-papers/2024/understanding-the-eu-ai-act`.

S. B. Jabeur, S. Mefteh-Wali, and J.-L. Viviani. Forecasting gold price with the xgboost algorithm and shap interaction values. *Annals of Operations Research*, 334(1):679–699, 2024.

S. Kruschel, N. Hambauer, S. Weinzierl, S. Zilker, M. Kraus, and P. Zschech. Challenging the performance–interpretability trade-off: An evaluation of interpretable machine learning models. *Business & Information Systems Engineering*, 2025. doi: 10.1007/s12599-024-00922-2.

S. Larsson, J. M. White, and C. Ingram Bogusz. The artificial recruiter: Risks of discrimination in employers' use of ai and automated decision-making. *Social Inclusion*, 12, 2024.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man–Machine Studies*, 7(1):1–13, 1975. doi: 10.1016/S0020-7373(75)80002-2.

T. Marwala. The algorithmic problem in artificial intelligence governance. Technical report, United Nations University, UNU Centre, January 2025. URL `https://unu.edu/article/algorithmic-problem-artificial-intelligence-governance`.

McKinsey Global Institute. The state of AI in 2024—and a half-decade in review. Technical report, McKinsey & Company, 2024. URL `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2024`. Global AI Survey.

J. Mökander, J. Morley, M. Taddeo, and L. Floridi. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27(4):44, 2021.

C. Molnar and T. Freiesleben. *Supervised Machine Learning for Science: How to stop worrying and love your black box*. Christoph Molnar, 2024.

OpenAI. Gpt-4.5 system card, 2025. URL `https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf`.

C. Oxborough, E. Cameron, A. Rao, A. Birchall, A. Townsend, and C. Westermann. Explainable AI: Driving business value through greater understanding. White paper, PricewaterhouseCoopers LLP (PwC UK), London, 2018. URL `https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf`.

J. Pearl. *Causality*. Cambridge university press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python, 2011. URL `https://scikit-learn.org/`.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996. ISBN 978-0521460865.

L. S. Riza, C. Bergmeir, F. Herrera, and J. M. Benítez. frbs: Fuzzy rule-based systems for classification and regression in R. *Journal of Statistical Software*, 65(6):1–30, 2015. doi: 10.18637/jss.v065.i06.

M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03. URL `https://www.jstatsoft.org/index.php/jss/article/view/v035i03`.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information, 2017. URL `https://arxiv.org/abs/1703.00810`.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

L.-X. Wang and J. M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(6):1414–1427, 1992. doi: 10.1109/21.199452.

A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25, 2008. doi: 10.18637/jss.v027.i08. URL `https://www.jstatsoft.org/index.php/jss/article/view/v027i08`.

Polyhedral aspects of maxoids

Tobias Boege[1], Kamillo Ferry[2], Benjamin Hollering[3], and Francesco Nowell[2]

[1]UiT The Arctic University of Norway, Tromsø, Norway
post@taboege.de
[2]Technical University of Berlin, Germany
{*ferry,nowell*}*@math.tu-berlin.de*
[3]Technical University of Munich, Germany
benjamin.hollering@tum.de

Abstract

The conditional independence (CI) relation of a distribution in a max-linear Bayesian network depends on its weight matrix through the C^*-separation criterion. These CI models, which we call *maxoids*, are compositional graphoids which are in general not representable by Gaussian or discrete random variables. We prove that every maxoid can be obtained from a transitively closed weighted DAG and show that the stratification of generic weight matrices by their maxoids yields a polyhedral fan.

1 Introduction

Linear structural equation models, sometimes called Bayesian networks, are of critical importance in modern data science and statistics through their applications to causality Pearl (2009) and probabilistic inference Koller and Friedman (2009). These statistical models use directed acyclic graphs (DAGs) to represent causal relationships and conditional independencies between random variables. Recently, there has been a focus on developing graphical models which are able to capture causal relations between extreme events. The two main approaches employ *Hüssler–Reiss distributions* Engelke et al. (2024, 2025) and max-linear Bayesian networks, the latter of which are the main subject of this paper.

Max-linear Bayesian networks (MLBNs), were introduced in Gissibl and Klüppelberg (2018) to model *cascading failures*. They are used in areas where these failures lead to catastrophic events, such as financial risk and water contamination Leigh et al. (2019); Rochet and Tirole (1996). A random vector $X = (X_1, \ldots, X_n)$ is distributed according to the max-linear model on a DAG $\mathcal{G}$ if it satisfies the system of *recursive structural equations*

$$X_i = \bigvee_{j \in \mathrm{pa}(i)} c_{ij} X_j \vee Z_i, \qquad c_{ij}, Z_i \geq 0, \tag{1.1}$$

where $\vee = \max$, the c_{ij} are edge weights, $\mathrm{pa}(i)$ is the set of parents of i in $\mathcal{G}$, and the Z_i are independent, atom-free, continuous random variables.

The structural equations mimic Bayesian networks in the extreme value setting. Despite this similarity, the conditional independence (CI) theory of MLBNs turns out to be more subtle in certain aspects than that of classical Bayesian networks which are governed by the well-known d-separation criterion. In addition to the d-separations of the DAG, a max-linear model may satisfy other CI statements which depend on the weight matrix C appearing in (1.1). Améndola et al. (2022) observed that multiple distinct CI structures can arise for the same DAG, each for a set of C-matrices with positive Lebesgue measure. They introduced the graphical $*$-separation criterion which is complete but not strongly complete for CI implication in MLBNs, and the C^*-separation criterion which takes C into account and completely characterizes the CI structure of an MLBN. Moreover, the following chain of implications from d- over $*$- to C^*-separation is valid for all MLBNs:

$$[i \perp_d j \mid L] \implies [i \perp_* j \mid L] \implies [i \perp_{C^*} j \mid L].$$

In this paper, we focus on the CI structures which arise from C^*-separation since they are the most refined according to these implications and MLBNs are generically faithful to them. We call $\mathcal{M}_*(\mathcal{G}, C) \coloneqq \{\, [I \perp\!\!\!\perp J \mid L] : [I \perp_{C^*} J \mid L] \text{ in } (\mathcal{G}, C) \,\}$ the *maxoid* associated to the DAG $\mathcal{G}$ with coefficient matrix C and note that this is essentially the *global Markov property* of $\mathcal{G}$ with respect to C^*-separation with given coefficient matrix C. We show that $\mathcal{M}_*(\mathcal{G}, C)$ is a *compositional graphoid* and that the set of distinct maxoids associated to a fixed DAG $\mathcal{G}$ are in correspondence with the cones of a complete fan for which we provide an explicit representation of the inequalities. The following is our main result.

Theorem. *For any DAG $\mathcal{G}$ there is a hyperplane arrangement $\mathcal{H}_\mathcal{G} \subseteq \mathbb{R}^E$ such that for every $C \in \mathbb{R}^E \setminus \mathcal{H}_\mathcal{G}$ the set*

$$\operatorname{cone}_\mathcal{G}(C) := \left\{ C' \in \mathbb{R}^E \setminus \mathcal{H}_\mathcal{G} : \mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_*(\mathcal{G}, C') \right\}$$

is a full-dimensional open polyhedral cone. The collection of all closures of such cones for a fixed $\mathcal{G}$ forms a complete polyhedral fan $\mathcal{F}_\mathcal{G}$ in $\mathbb{R}^E$. Moreover the map which sends a cone of $\mathcal{F}_\mathcal{G}$ to its maxoid is an inclusion-reversing surjection.

One immediate consequence of the above theorem together with the results of Améndola et al. (2022) is that the maximal cones of $\mathcal{F}_\mathcal{G}$ correspond to the distinct CI structures which can arise from a max-linear Bayesian network with positive Lebesgue measure for the choice of C. In this sense, the maximal cones correspond to the *generic* CI structures of an MLBN supported on $\mathcal{G}$. Similarly, we call a weight matrix C *generic* if it does not lie on $\mathcal{H}_\mathcal{G}$. As we will show in Section 2, if there exist two nodes $i, j \in V(\mathcal{G})$ such that there are at least two distinct paths between i and j, then $\mathcal{F}_\mathcal{G}$ has at least two distinct full-dimensional cones. This provides a strong contrast to classical linear structural equation models which are generically faithful to d-separation, and thus almost every distribution in the model exhibits the same CI structure; cf. Lauritzen (1996).

Our results also elucidate which CI structures may arise from a given graph and when two graphs exhibit the same generic CI structure. This is critical for determining if the graph structure may be recovered using only conditional independencies as is typically done in constraint-based causal discovery algorithms, e.g.,Spirtes et al. (2000).

The remainder of this paper is organized as follows. In Section 2 we recall the details of the $*$- and C^*-separation criteria and use this to provide an explicit description of the linear inequalities which define $\mathrm{cone}_{\mathcal{G}}(C)$ via the *critical paths* of $\mathcal{G}$. We then relate the set of maxoids arising from a DAG $\mathcal{G}$ to the cones of the associated polyhedral fan. In Section 3 we show that every maxoid is a compositional graphoid but provide counterexamples which demonstrate that maxoids need not be representable by either regular Gaussian or discrete distributions. This again provides a contrast to Bayesian networks for which Gaussian and discrete distributions are the primary parametric families which are studied.

2 The Polyhedral Geometry of C^*-separation

Let $\mathcal{G} = (V, E)$ be a DAG on $|V| = n$ vertices and denote the set of coefficient matrices supported on $\mathcal{G}$ by $\mathbb{R}^E_{>0}$, i.e., all $n \times n$ matrices C with $c_{ij} = 0$ if $i \to j \notin E$ and $c_{ij} > 0$ otherwise. We recall that a random vector X is distributed according to the max-linear model on $\mathcal{G}$ if it satisfies eq. (1.1). This system of equations has solution $X = C^*Z$ where the matrix-vector product is done in max-times arithmetic. C^* is the *Kleene star* matrix of C whose entries are given by

$$(C^*)_{ij} = \max_{\pi \in P(i,j)} \prod_{e \in \pi} c_{ij},$$

where $P(i, j)$ denotes the set of all directed paths from i to j in $\mathcal{G}$ and $\prod_{e\in\pi} c_{ij}$ is the weight of the path π. The conditional independence structure of max-linear models depends on inequalities between the weights of paths, which in the above form would not be polyhedral. To solve this, we note that the coordinate-wise logarithm is an isomorphism which maps $\mathbb{R}^E_{>0} \to \mathbb{R}^E$. This transformation takes us from max-times to max-plus arithmetic and in this new coordinate system the Kleene star is given by

$$(C^*)_{ij} = \max_{\pi \in P(i,j)} \sum_{e \in \pi} c_{ij} = \max_{\pi \in P(i,j)} \omega_C(\pi).$$

It is natural to extend the logarithm to send 0 to $-\infty$ and thus when embedding $C \in \mathbb{R}^E$ into $\mathbb{R}^{n\times n}$ we use the convention that $c_{ij} = -\infty$ if $i \to j \notin E(\mathcal{G})$. Those familiar with tropical geometry will notice that C is actually a matrix over the tropical semiring with max-plus arithmetic, however this will not be relevant for the results which we present here. For the remainder of this paper, we will exclusively utilize the max-plus convention.

A path π' is *critical* if $\omega_C(\pi') = \max_{\pi\in P(i,j)} \omega_C(\pi)$. If there is a unique critical path between every pair of nodes i, j then we say that C is *generic* and denote this unique path by $\pi^{ij}_{\mathrm{crit}}(\mathcal{G}, C)$, omitting the pair $(\mathcal{G}, C)$ when it is clear from context. Note that if C is not generic, then its entries satisfy some non-trivial linear equation of the form $\omega_C(\pi) = \omega_C(\pi')$ for $\pi, \pi' \in P(i, j)$. Hence, the set of generic matrices is the complement of a hyperplane arrangement $\mathcal{H}_{\mathcal{G}}$ whose defining equations depend on $\mathcal{G}$. We are now ready to introduce C^*-separation.

Definition 2.1. Let $(\mathcal{G}, C)$ be a weighted DAG with vertex set V and $L \subseteq V$. The *critical DAG* $\mathcal{G}^*_{C,L}$ is the DAG on V such that $i \to j \in E(\mathcal{G}^*_{C,L})$ whenever i and j are connected via a directed path, and no critical path from i to j in $\mathcal{G}$ intersects L.

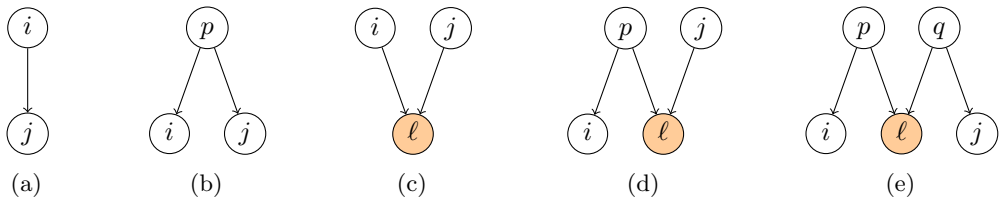


Figure 1: The types of $*$-connecting paths between i and j given L in a critical DAG $\mathcal{G}^*_{C,L}$. The colored colliders ℓ must belong to L; the non-colliders p, q must not belong to L.

Two nodes i and j are *C^*-connected given L* if there exists a path from i to j in $\mathcal{G}^*_{C,L}$ of the form pictured in Figure 1. If no such path exists, then i and j are *C^*-separated given L* which is denoted $[i \perp_{C^*} j \mid L]$.

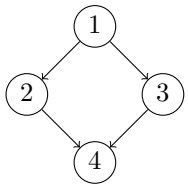
Theorem 2.2 ((Améndola et al., 2022, Theorem 6.18))**.** *Let $(\mathcal{G}, C)$ be a weighted DAG and X be a random vector distributed according to the max-linear model on $(\mathcal{G}, C)$. Then*

$$[i \perp_{C^*} j \mid L] \implies [i \perp\!\!\!\perp j \mid L].$$

Moreover, the converse holds for all but a Lebesgue null set of weight matrices C.

C^*-separation generally entails more CI statements than d-separation. In particular note that a $*$-connecting path can have at most one collider in its conditioning set whereas d-separation allows any number of colliders in a connecting path. Moreover, it suffices to block only a single critical path from i and j in order to separate them.

Example 2.3. Consider the diamond graph $\mathcal{G}$ with weight matrix C:



$$C = \begin{pmatrix} -\infty & c_{12} & c_{13} & -\infty \\ -\infty & -\infty & c_{24} & -\infty \\ -\infty & -\infty & -\infty & c_{34} \\ -\infty & -\infty & -\infty & -\infty \end{pmatrix}$$

Observe that $P(1,4) = \{\pi_2, \pi_3\}$ where $\pi_i = 1 \to i \to 4$. If C satisfies $\omega_C(\pi_2) > \omega_C(\pi_3)$, then $\mathcal{G}^*_{C,\{2\}}$ is exactly the diamond above because π_2 is a critical path from 1 to 4 which intersects the conditioning set $\{2\}$. Since neither π_2 nor π_3 are of the forms displayed in Figure 1, this MLBN satisfies $[1 \perp_{C^*} 4 \mid 2]$. On the other hand, if $\omega_C(\pi_3) > \omega_C(\pi_2)$ then a similar argument yields that $[1 \perp_{C^*} 4 \mid 3]$. Thus we get two distinct maxoids which correspond to whether π_2 or π_3 is the critical path. Moreover, the maxoid $\mathcal{M}_*(\mathcal{G}, C)$ is completely determined by which side of the hyperplane

$$c_{12} + c_{24} = \omega_C(\pi_2) = \omega_C(\pi_3) = c_{13} + c_{34}$$

the matrix C lies on.

Our goal in the remainder of this section is to develop the observations from the previous example into a general result which connects weighted DAGs and their maxoids using polyhedral geometry. We begin with a sequence of lemmas which further elucidate the connection between the critical paths in $(\mathcal{G}, C)$ and the CI structure $\mathcal{M}_*(G, C)$.

Lemma 2.4. *Let $(\mathcal{G}, C)$ and $(\mathcal{G}', C')$ be two weighted DAGs on the same node set and generic weights C and C'. Then*

$$\mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_*(\mathcal{G}', C') \iff \pi^{ij}_{crit}(\mathcal{G}, C) = \pi^{ij}_{crit}(\mathcal{G}', C') \text{ for all } i \neq j,$$

i.e., two weighted DAGs have the same critical paths if and only if their maxoids coincide.

Proof. " $\Longleftarrow$ ": If $(\mathcal{G}, C)$ and $(\mathcal{G}', C')$ have the same critical paths then they give rise to the same critical DAG for any L, implying equal maxoids.

" $\Longrightarrow$ " by contraposition: Suppose that $\pi = \pi^{ij}_{\mathrm{crit}}(\mathcal{G}, C) \neq \pi^{ij}_{\mathrm{crit}}(\mathcal{G}', C') = \pi'$ and denote the nodes on π' as follows:

$$\pi' : i = \ell'_0 \to \ell'_1 \cdots \to \ell'_{m-1} \to \ell'_m = j. \tag{2.1}$$

We may assume that π does not contain any of the nodes $\ell'_1, \ldots, \ell'_{m-1}$ (if this is not the case, then we may replace j with an internal node common to both π and π' and have shorter but still differing critical paths). Then clearly $[i \perp_{C^*} j \mid \ell'_{m-1}]$ holds in $(\mathcal{G}', C')$ but not in $(\mathcal{G}, C)$, implying inequality of the respective maxoids. □

Lemma 2.5. *Let $(\mathcal{G}, C)$ be a weighted DAG and $\overline{\mathcal{G}}$ the transitive closure of $\mathcal{G}$. There exists a matrix $\overline{C}$ supported on $\overline{\mathcal{G}}$ such that $\mathcal{M}_*(\overline{\mathcal{G}}, \overline{C}) = \mathcal{M}_*(\mathcal{G}, C)$.*

Proof. For a fixed $(\mathcal{G}, C)$ with edge set E , let $\overline{E}$ be the edge set of its transitive closure $\overline{\mathcal{G}}$. For any two path-connected nodes $i, j \in V$, let ε_{ij} be the weight of the (not necessarily unique) critical $i - j$ path in $(\mathcal{G}, C)$, and fix a $-\infty < \delta < \min_{i,j} \varepsilon_{ij}$. One possible choice of $\overline{C}$ is given by

$$\overline{C} = (\overline{c}_{ij})_{i,j \in V} = \begin{cases} c_{ij} & \text{if} \quad (i,j) \in E, \\ \delta & \text{if} \quad (i,j) \in \overline{E} \setminus E, \\ -\infty & \text{otherwise.} \end{cases}$$

By construction, no edge in $\overline{E} \setminus E$ is contained in any critical path of $(\overline{\mathcal{G}}, \overline{C})$. Thus, the statement follows from Lemma 2.4. □

A related notion is the *weighted transitive reduction*.

Definition 2.6. The *weighted transitive reduction* $\mathcal{G}^{\mathrm{tr}}_C$ of a weighted DAG $(\mathcal{G}, C)$ is the subgraph of $\mathcal{G}$ with edges determined as follows:

$$i \to j \in E(\mathcal{G}^{\mathrm{tr}}_C) \iff i \to j \text{ is the unique critical } i - j \text{ path in } (\mathcal{G}, C).$$

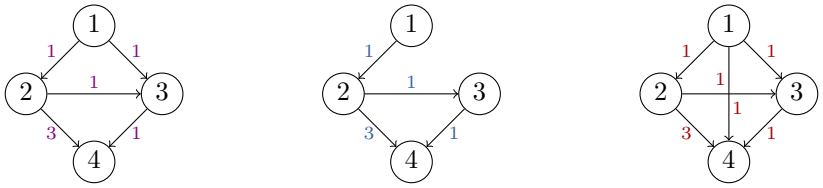


Figure 2: For appropriate C^{tr} and $\overline{C}$, $\mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_*(\mathcal{G}_C^{\mathrm{tr}}, C^{\mathrm{tr}}) = \mathcal{M}_*(\overline{\mathcal{G}}, \overline{C})$ holds.

Remark 2.7. Another consequence of Lemma 2.4 is that for any weighted DAG $(\mathcal{G}, C)$ with generic C we have

$$\mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_*(\mathcal{G}_C^{\mathrm{tr}}, C^{\mathrm{tr}}), \tag{2.2}$$

where C^{tr} is any matrix supported on $\mathcal{G}_C^{\mathrm{tr}}$ which gives rise to the same critical paths as $(\mathcal{G}, C)$. Combined with Lemma 2.5, this means that the maxoid of *any* weighted DAG arises as a maxoid of its transitive closure for an appropriately chosen weight matrix.

Example 2.8. The maxoid corresponding to the weighted DAG on the left in Figure 2 is

$$\mathcal{M}_*(\mathcal{G}, C) = \{[1 \perp\!\!\!\perp 3 \mid 2]\ ,\ [1 \perp\!\!\!\perp 3 \mid 2, 4]\ ,\ [1 \perp\!\!\!\perp 4 \mid 2]\ ,\ [1 \perp\!\!\!\perp 4 \mid 2, 3]\}.$$

This maxoid is also realized by the weighted transitive reduction $\mathcal{G}_C^{\mathrm{tr}}$ and transitive closure $\overline{\mathcal{G}}$ when C^{tr} and $\overline{C}$ are chosen according to Lemma 2.5 and Remark 2.7. In this example, $c_{24}^{\mathrm{tr}} > c_{23}^{\mathrm{tr}} + c_{34}^{\mathrm{tr}}$ and $\overline{c}_{14} < \min\{\overline{c}_{12} + \overline{c}_{24}\ ,\ \overline{c}_{13} + \overline{c}_{34}\}$ must hold.

Theorem 2.9. *Let $(\mathcal{G}, C)$ be a weighted DAG with generic $C \in \mathbb{R}^E \setminus \mathcal{H}_{\mathcal{G}}$. The set*

$$\mathrm{cone}_{\mathcal{G}}(C) := \left\{C' \in \mathbb{R}^E \setminus \mathcal{H}_{\mathcal{G}} : \mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_*(\mathcal{G}, C')\right\} \tag{2.3}$$

is a full-dimensional open polyhedral cone defined by linear inequalities of the form

$$\omega_{C'}(\pi_{crit}^{ij}(\mathcal{G}, C)) > \omega_{C'}(\pi), \quad \text{for each } \pi \in P(i,j) \setminus \{\pi_{crit}^{ij}(\mathcal{G}, C)\}, \tag{2.4}$$

for all distinct $i, j \in V$.

Proof. By Lemma 2.4, the set $\mathrm{cone}_{\mathcal{G}}(C)$ consists of all generic weight matrices C' supported on $\mathcal{G}$ and giving rise to the same critical paths as C. This is precisely what is encoded in the inequalities (2.4) for all $i, j \in V$. These strict linear inequalities in the entries of C' define an open polyhedral cone in $\mathbb{R}^E$ disjoint from $\mathcal{H}_{\mathcal{G}}$. The cone is non-empty as $C \in \mathrm{cone}_{\mathcal{G}}(C)$ is given, and full-dimensional because ε-perturbations of C in the direction of any c_{ij} preserve its critical paths. □

Remark 2.10. A minimal description of the cone defined in (2.4) can be obtained by considering only pairs i, j which are connected by multiple *disjoint* paths, in the sense that any two of them form a simple cycle in the skeleton of $\mathcal{G}$. Indeed, if two $i - j$ paths π_1 and π_2 contain a common intermediate node k, then the linear inequality corresponding to the comparison of $\omega_C(\pi_1)$ and $\omega_C(\pi_2)$ is already implied by the linear inequalities which arise from comparing their respective $i - k$ and $k - j$ portions.

We now study the case where the weight matrix lies on the boundary of a cone. For generic C, let $\tilde{C}$ be a matrix lying on a *facet* of the euclidean closure of $\mathrm{cone}_{\mathcal{G}}(C)$. This means that for some pair $i, j \in V$, equality holds in (2.4) and thus there are two critical $i-j$ paths in $(\mathcal{G}, \tilde{C})$: one is the unique critical $i-j$ path $\pi^{ij}_{\mathrm{crit}}(\mathcal{G}, C)$, and the other we denote by π'. We assume that the paths are disjoint in the sense of Remark 2.10 and that all matrices on the facet of $\mathrm{cone}_{\mathcal{G}}(C)$ on which $\tilde{C}$ lies give rise to the same critical paths as C outside of those which factor through the directed $i-j$ portion of the DAG.

Theorem 2.11. *In the setting described above, the following holds:*

$$\mathcal{M}_*(\mathcal{G}, \tilde{C}) = \mathcal{M}_*(\mathcal{G}, C) \cup \mathcal{M}_*(\mathcal{G}, C'), \tag{2.5}$$

where C' is a matrix supported on $\mathcal{G}$ giving rise to the same critical paths as C except for in the directed $i-j$ portion, where the unique critical path is π'.

Proof. We first consider the simplified case where $\mathcal{G}$ consists solely of the two directed $i-j$ paths. For readability, we set $\pi := \pi^{ij}_{\mathrm{crit}}(\mathcal{G}, C)$ and refer to the intermediate nodes of π and π' using the notation in (2.1). In this setting, i and j are the only two nodes which are connected by more than one path. Because of this, it suffices to prove both inclusions in (2.5) only for separation statements of the form $[i \perp_{C^*} j \mid L]$.

"$\subseteq$": Let $L \subset V \setminus ij$. Note that if $[i \perp\!\!\!\perp j \mid L] \in \mathcal{M}_*(\mathcal{G}, \tilde{C})$ holds, then L intersects $\pi \cup \pi'$ non-trivially. Indeed, if $L \cap (\pi \cup \pi') = \emptyset$, then the critical DAG $\mathcal{G}^*_{\tilde{C},L}$ contains the edge $i \to j$, implying $*$-connectedness. Thus, this choice of L also separates i and j in $(\mathcal{G}, C)$ or $(\mathcal{G}, C')$, implying the first inclusion.

"$\supseteq$": In $(\mathcal{G}, C)$ any $L \subseteq V \setminus ij$ which intersects π non-trivially gives rise to the statement $[i \perp_{C^*} j \mid L]$. This choice of L also separates i and j in $(\mathcal{G}, \tilde{C})$. (Recall that the condition for the edge $i \to j$ to be present in $\mathcal{G}^*_{\tilde{C},L}$ is that *no* critical $i-j$ path in $(\mathcal{G}, \tilde{C})$ factors through L.) Analogously, any statement of the form $[i \perp_{C^*} j \mid L]$ in $\mathcal{M}_*(\mathcal{G}, C')$ also holds in $(\mathcal{G}, \tilde{C})$.

In the more general setting where $\mathcal{G}$ does not consist solely of directed $i-j$ paths, additional $*$-connecting $i-j$ paths may exist. Thus, additional nodes which are not contained in π and π' may be needed to separate i and j. However, these nodes will be required to separate i and j in all three weighted DAGs, since, by our starting assumption, these three matrices give rise to the same critical paths outside of the directed $i-j$ portion of $\mathcal{G}$. Furthermore, if i' and j' are nodes such that a path between them factors through π (and thus also π'), then a similar argument immediately shows that any L which separates them in $(\mathcal{G}, \tilde{C})$ must also separate them in either $(\mathcal{G}, C)$ or $(\mathcal{G}, C')$. Lastly, any separation which does not involve π and π' will be present in all three maxoids by assumption and thus the remaining CI statements will be the same as well. □

Remark 2.12. In the setting of Theorem 2.11, given $\tilde{C}$ one can obtain a matrix with the properties of C' by replacing $\tilde{c}_{\ell,\ell'}$ with $\tilde{c}_{\ell,\ell'} + \varepsilon$, where $(\ell, \ell') \in \pi'$ and $\varepsilon > 0$ fulfills

$$\varepsilon < \min_{i',j' \in V} \left\{ \min_{\pi_1, \pi_2 \in P(i',j')} |\omega_{\tilde{C}}(\pi_1) - \omega_{\tilde{C}}(\pi_2)| \right\}. \tag{2.6}$$

This makes π' the unique critical $i-j$ path while preserving all other critical paths.

Theorem 2.11 implies that facets (and by extension, lower-dimensional faces) of the euclidean closure of $\text{cone}_{\mathcal{G}}(C)$ correspond to non-generic maxoids which arise as unions of generic maxoids.

Corollary 2.13. *The euclidean closures of the open cones corresponding to the generic maxoids of $\mathcal{G}$ form a complete polyhedral fan, $\mathcal{F}_{\mathcal{G}}$, in $\mathbb{R}^E$. The maximal cones of $\mathcal{F}_{\mathcal{G}}$ are in bijection with the generic maxoids of $\mathcal{G}$. Moreover, the function Φ which sends a cone of $\mathcal{F}_{\mathcal{G}}$ to its maxoid is an inclusion-reversing surjection:*

$$F_1 \text{ is a face of } F_2 \implies \Phi(F_1) \supseteq \Phi(F_2) \quad \text{for all } F_1, F_2 \in \mathcal{F}_{\mathcal{G}}.$$

Remark 2.14. It is not hard to show that $\mathcal{F}_{\mathcal{G}}$ is the *Gröbner fan* of the ideal $I_{\mathcal{G}} = \langle \sum_{\pi \in P(i,j)} \prod_{e \in \pi} x_e : i, j \in V, |P(i,j)| > 1 \rangle$; see Sturmfels (1996). Indeed, any weight matrix $C \in \mathbb{R}^E$ defines a term order which picks out the critical $i - j$ paths of $(\mathcal{G}, C)$ as the initial term of the generator $f_{ij} = \sum_{\pi \in P(i,j)} \prod_{e \in \pi} x_e$ in $I_{\mathcal{G}}$.

Example 2.15. The fan associated to the diamond graph from Example 2.3 consists of two maximal cones in $\mathbb{R}^4$ separated by the hyperplane $c_{12} + c_{24} = c_{13} + c_{34}$. The corresponding maxoids are

$$\begin{aligned}
\mathcal{M}_1 &= \{[1 \perp\!\!\!\perp 3 \mid 2]\,, [1 \perp\!\!\!\perp 4 \mid 2,3]\,,\ [1 \perp\!\!\!\perp 4 \mid 2]\} && \text{for } c_{12} + c_{24} > c_{13} + c_{34} \\
\mathcal{M}_2 &= \{[1 \perp\!\!\!\perp 3 \mid 2]\,, [1 \perp\!\!\!\perp 4 \mid 2,3]\,,\ [1 \perp\!\!\!\perp 4 \mid 3]\} && \text{for } c_{12} + c_{24} < c_{13} + c_{34} \\
\mathcal{M}_3 &= \{[1 \perp\!\!\!\perp 3 \mid 2]\,, [1 \perp\!\!\!\perp 4 \mid 2,3]\,,\ [1 \perp\!\!\!\perp 4 \mid 2]\,,\ [1 \perp\!\!\!\perp 4 \mid 3]\} && \text{for } c_{12} + c_{24} = c_{13} + c_{34}.
\end{aligned}$$

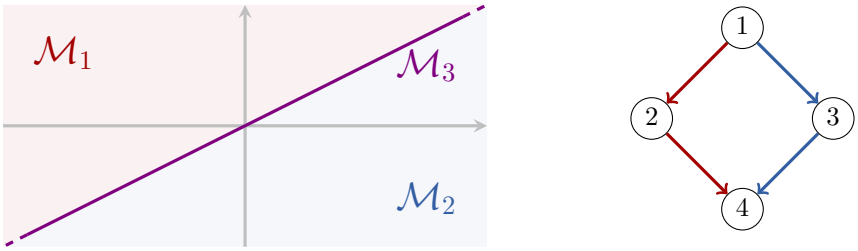


Figure 3: A projection of the fan $\mathcal{F}_{\mathcal{G}}$ of the diamond which is the Gröbner fan of the ideal $I_{\mathcal{G}} = \langle x_{12}x_{24} + x_{13}x_{34} \rangle$.

3 Representability of maxoids

The polyhedral fan $\mathcal{F}_{\mathcal{G}}$ provides the maxoids associated to a given DAG with a geometric structure which is both interesting and practically useful: it gives an efficient algorithm for solving the CI implication problem for maxoids on a given DAG. A similar connection has been previously exploited in the framework of *structural imsets* by Bouckaert et al. (2010).

However, the polyhedral fan in our case is specific to the graph and the map from its cones to maxoids does not in general induce a Galois connection. As a result, the extraction of conditional independence features from the polyhedral geometry is not straightforward.

In this section we focus on logical properties of all maxoids, independent of the underlying DAG, in the context of conditional independence implication. Like many other types of graphical models (cf. Lauritzen and Sadeghi (2018)), C^*-separation satisfies the *compositional graphoid properties*, i.e., every maxoid is closed under the following equivalence and implications for all disjoint $I, J, K, L \subseteq N$:

$$\begin{aligned}
\textbf{Semigraphoid:}\quad & [I \perp\!\!\!\perp J \mid L] \wedge [I \perp\!\!\!\perp K \mid JL] && \iff && [I \perp\!\!\!\perp JK \mid L],\\
\textbf{Intersection:}\quad & [I \perp\!\!\!\perp J \mid KL] \wedge [I \perp\!\!\!\perp K \mid JL] && \implies && [I \perp\!\!\!\perp JK \mid L], \text{ and}\\
\textbf{Composition:}\quad & [I \perp\!\!\!\perp J \mid L] \wedge [I \perp\!\!\!\perp K \mid L] && \implies && [I \perp\!\!\!\perp JK \mid L].
\end{aligned}$$

Whereas the Semigraphoid property holds for the CI statements satisfied by any random vector, Intersection and Composition provide non-trivial additional structure. Améndola et al. (2022) mention without proof that C^*-separation satisfies the compositional graphoid properties. We supply the routine proof below and then delve into the question of what distinguishes maxoids from other types of compositional graphoids.

Proposition 3.1. *Maxoids are compositional graphoids.*

Proof. Consider any maxoid $\mathcal{M} = \mathcal{M}_*(\mathcal{G}, C)$ for a given DAG $\mathcal{G}$ and weight matrix C supported on $\mathcal{G}$. All separation statements below are with respect to $(\mathcal{G}, C)$. By the definition of C^*-separation, the assumption $[I \not\perp_{C^*} J \mid L]$ implies the existence of a $*$-connecting path π between I and J in the critical DAG $\mathcal{G}^*_{C,L}$. A fortiori, π also connects I and KL in $\mathcal{G}^*_{C,L}$, hence $[I \not\perp_{C^*} JK \mid L]$. Now consider a $*$-connecting path π between I and K in $\mathcal{G}^*_{C,JL}$. If it contains a collider $j \in J$, then the portion of π from I to j is a $*$-connecting path between I and J in $\mathcal{G}^*_{C,L}$. Otherwise the collider (if any) is in L and π yields a $*$-connecting between I and K in $\mathcal{G}^*_{C,L}$. In both cases, we obtain a $*$-connecting path between I and JK in $\mathcal{G}^*_{C,L}$. By contraposition, these two arguments prove the "only if" part of the Semigraphoid property.

The "if" direction is proved by contraposition as well. Assume that $[I \not\perp_{C^*} JK \mid L]$ and $[I \perp_{C^*} J \mid L]$, i.e., there exists a $*$-connecting path π from I to JK but not one from I to J in $\mathcal{G}^*_{C,L}$. Hence, π must connect I and K and cannot contain any node from J. But then π also $*$-connects I and K in $\mathcal{G}^*_{C,JL}$, thus $[I \not\perp_{C^*} K \mid JL]$ holds.

For Intersection, use again contraposition. Assume $[I \not\perp_{C^*} JK \mid L]$ and $[I \perp_{C^*} J \mid KL]$. By the Semigraphoid property and the symmetry with respect to exchanging J and K, we can split $[I \not\perp_{C^*} JK \mid L]$ into two cases: $[I \not\perp_{C^*} K \mid L]$ or $[I \not\perp_{C^*} J \mid KL]$. The second case contradicts our other assumption. In the former case, let π denote an $*$-connecting path between I and K in $\mathcal{G}^*_{C,L}$. We may assume that this path is as short as possible, i.e., does not contain any other node of K. If it contains a node $j \in J$, then the portion from I to j $*$-connects I and J in $\mathcal{G}^*_{C,KL}$ which is impossible. Hence π is free of nodes from J and thus $*$-connects I and K also in $\mathcal{G}^*_{C,JL}$ which is the required conclusion of Intersection.

The Composition property holds almost by definition. Any $*$-connecting path from I to JK in $\mathcal{G}^*_{C,L}$ connects either I to J or I to K, which is the contrapositive of the assertion of Composition. □

Algebraic statistics today, by and large, deals with CI models on discrete and regular Gaussian random variables. Note that the parametrization of MLBNs in (1.1) does not produce jointly Gaussian distributions as the maximum of Gaussians does not follow a Gaussian distribution. On the other hand, discrete distributions are not atom-free and are thus incompatible with this parametrization. Nevertheless, it is reasonable to ask whether maxoids, as abstract conditional independence models, can be represented using one of these two distribution classes. We answer this question negatively by highlighting features of maxoids which serve as obstructions to Gaussian and discrete representability. This means that maxoids are a new and rather exotic class of compositional graphoids.

In Drton and Xiao (2010) the term *semigaussoid* is used to refer to compositional graphoids. What is missing from a semigaussoid to a *gaussoid* is the closedness under the following implication:

$$\textbf{Weak Transitivity:} \quad [i \perp\!\!\!\perp j \mid L] \wedge [i \perp\!\!\!\perp j \mid kL] \iff [i \perp\!\!\!\perp k \mid L] \vee [j \perp\!\!\!\perp k \mid L],$$

for all distinct i, j, k and $L \subseteq N \setminus ijk$. The following example shows that maxoids need not satisfy Weak Transitivity. By results of Lněnička and Matúš (2007), this provides examples of maxoids which — in contrast to classical d-separation graphoids — cannot be faithfully represented by a regular Gaussian random vector.

Example 3.2. Consider the diamond graph $\mathcal{G}$ as described in Example 2.3 with weight matrix C satisfying $\omega_C(\pi_3) > \omega_C(\pi_2)$. The maxoid of $(\mathcal{G}, C)$ consists precisely of the d-separations in $\mathcal{G}$ plus $[1 \perp\!\!\!\perp 4 \mid 3]$. As $[1 \perp\!\!\!\perp 4 \mid 3]$ and $[1 \perp\!\!\!\perp 4 \mid 2, 3]$ hold without $[1 \perp\!\!\!\perp 2 \mid 3]$ or $[2 \perp\!\!\!\perp 4 \mid 3]$, this CI structure violates Weak Transitivity and cannot be faithfully represented by a regular Gaussian distribution.

This violation of Weak Transitivity has the following geometric consequence. The space of regular Gaussian distributions which are Markov to this CI structure is the union of two standard Bayesian networks on subgraphs of the diamond $\mathcal{G}$: one has the edge $1 \to 2$ removed (so that it satisfies $[1 \perp\!\!\!\perp 2 \mid 3]$) and the other has the edge $2 \to 4$ removed (and hence satisfies $[2 \perp\!\!\!\perp 4 \mid 3]$). For an algebraic explanation of this phenomenon we refer to Drton et al. (2024).

Example 3.3. The Cassiopeia graph $\mathcal{G}$ (see also Améndola et al. (2022)) arises from Figure 1e by reversing the direction of all arrows. Its associated fan $\mathcal{F}_\mathcal{G}$ has only one maximal cone and thus all generic C give rise to the same CI structure, which has a peculiarity: since the path from i to j has two colliders, it is not $*$-connecting given p and q. We have $\mathcal{M}_*(\mathcal{G}, C) = \mathcal{M}_d(\mathcal{G}) \cup \{[i \perp\!\!\!\perp j \mid p, q]\}$. One easily computes that the CI structure of the conditional distribution given q equals

$$\Big\{[i \perp\!\!\!\perp \ell], [i \perp\!\!\!\perp \ell \mid j], [i \perp\!\!\!\perp \ell \mid j, p], [i \perp\!\!\!\perp j],$$
$$[i \perp\!\!\!\perp j \mid \ell], [i \perp\!\!\!\perp j \mid p], [\ell \perp\!\!\!\perp p \mid j], [\ell \perp\!\!\!\perp p \mid i, j]\Big\}.$$

It follows from implication (I:13) of Studený (2021) (with $X = i, Y = j, Z = \ell, U = p$) that every discrete distribution satisfying these CI statements must also satisfy $[i \perp\!\!\!\perp j \mid \ell, p]$. This shows that the Cassiopeia maxoid cannot be represented by discrete random variables.

Acknowledgements

T.B. was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101110545. K.F. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). B.H. was supported by the Alexander von Humboldt Foundation. F.N. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Priority Programme "Combinatorial Synergies" (SPP 2458, project ID: 539875257).

References

C. Améndola, C. Klüppelberg, S. Lauritzen, and N. M. Tran. Conditional independence in max-linear Bayesian networks. *Ann. Appl. Probab.*, 32(1):1–45, 2022. doi: 10.1214/21-AAP1670.

R. Bouckaert, R. Hemmecke, S. Lindner, and M. Studený. Efficient algorithms for conditional independence inference. *J. Mach. Learn. Res.*, 11:3453–3479, 2010. URL `www.jmlr.org/papers/v11/bouckaert10b.html`.

M. Drton and H. Xiao. Smoothness of Gaussian conditional independence models. In *Algebraic methods in statistics and probability II*, volume 516 of *Contemporary Mathematics*, pages 155–177. American Mathematical Society (AMS), 2010. doi: 10.1090/conm/516/10173.

M. Drton, L. Henckel, B. Hollering, and P. Misra. Faithlessness in Gaussian graphical models, 2024. URL `https://arxiv.org/abs/2404.05306`.

S. Engelke, M. Hentschel, M. Lalancette, and F. Röttger. Graphical models for multivariate extremes, 2024. URL `https://arxiv.org/abs/2402.02187`.

S. Engelke, N. Gnecco, and F. Röttger. Extremes of structural causal models, 2025. URL `https://arxiv.org/abs/2503.06536`.

N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693–2720, 2018. doi: 10.3150/17-BEJ941.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

S. Lauritzen and K. Sadeghi. Unifying Markov properties for graphical models. *Ann. Statist.*, 46(5):2251–2278, 2018. doi: 10.1214/17-AOS1618.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

C. Leigh, O. Alsibai, R. J. Hyndman, S. Kandanaarachchi, O. C. King, J. M. McGree, C. Neelamraju, J. Strauss, P. D. Talagala, R. D. Turner, et al. A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of the Total Environment*, 664:885–898, 2019.

R. Lněnička and F. Matúš. On Gaussian conditional independence structures. *Kybernetika*, 43(3):327–342, 2007. URL `https://www.kybernetika.cz/content/2007/3/327`.

J. Pearl. *Causality.* Cambridge University Press, 2009.

J.-C. Rochet and J. Tirole. Interbank lending and systemic risk. *Journal of Money, Credit and Banking*, 28(4):733–762, 1996.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

M. Studený. Conditional independence structures over four discrete random variables revisited: Conditional Ingleton inequalities. *IEEE Trans. Inf. Theory*, 67(11):7030–7049, 2021. doi: 10.1109/TIT.2021.3104250.

B. Sturmfels. *Gröbner bases and convex polytopes*, volume 8 of *Univ. Lect. Ser.* American Mathematical Society (AMS), 1996.

On the Intersection and Composition properties of conditional independence

Tobias Boege

UiT The Arctic University of Norway, Tromsø, Norway
post@taboege.de

Abstract

Compositional graphoids are fundamental discrete structures which appear in probabilistic reasoning, particularly in the area of graphical models. They are semigraphoids which satisfy the Intersection and Composition properties. These important properties, however, are not enjoyed by general probability distributions. We survey what is known in terms of sufficient conditions for Intersection and Composition and derive a set of new sufficient conditions in the context of discrete random variables based on conditional information inequalities for Shannon entropies.

1 Introduction

Dawid (1980) found fundamental relations among the valid conditional independence (CI) statements for any finite system N of jointly distributed random variables which became known later as the *semigraphoid properties*. They consist of the following assertions and implications, for any four disjoint subsets $I, J, K, L \subseteq N$; see (Studený, 2005, Section 2.2.2):

Triviality $[I \perp\!\!\!\perp \emptyset \mid L]$,

Symmetry $[I \perp\!\!\!\perp J \mid L] \iff [J \perp\!\!\!\perp I \mid L]$,

Decomposition $[I \perp\!\!\!\perp JK \mid L] \implies [I \perp\!\!\!\perp J \mid L]$,

Weak union $[I \perp\!\!\!\perp JK \mid L] \implies [I \perp\!\!\!\perp K \mid JL]$,

Contraction $[I \perp\!\!\!\perp J \mid L] \wedge [I \perp\!\!\!\perp K \mid JL] \implies [I \perp\!\!\!\perp JK \mid L]$.

The Triviality axiom is inconsequential as it does not interact with the other axioms in a way that produces other, non-trivial statements. Throughout this paper, we accept the Symmetry axiom and formally identify any CI symbol $[I \perp\!\!\!\perp J \mid K]$ with its symmetric version $[J \perp\!\!\!\perp I \mid K]$. This leaves Decomposition, Weak union and Contraction as the defining traits of a semigraphoid. They can be restated more succinctly as an equivalence:

$$[I \perp\!\!\!\perp JK \mid L] \iff [I \perp\!\!\!\perp J \mid L] \wedge [I \perp\!\!\!\perp K \mid JL].$$

Since the roles of J and K are interchangeable in the left-hand side, we may consider a symmetrized version which is the starting point for our investigation:

$$[I \perp\!\!\!\perp JK \mid L] \iff \begin{cases} ①\,[I \perp\!\!\!\perp J \mid L] \wedge ②\,[I \perp\!\!\!\perp K \mid JL] \wedge \\ ③\,[I \perp\!\!\!\perp K \mid L] \wedge ④\,[I \perp\!\!\!\perp J \mid KL]. \end{cases}$$

The present paper addresses the question under which circumstances are subsets of the statements on the right-hand side sufficient to imply $[I \perp\!\!\!\perp JK \mid L]$ on the left-hand side, provided that all random variables are discrete.

By Contraction, ① ∧ ② as well as ③ ∧ ④ are always sufficient; hence, any 3-subset of $\{①, ②, ③, ④\}$ is sufficient. Up to interchanging J and K, this leaves only three configurations of the 2-element subsets to consider:

- The implication ① ∧ ③ $\implies$ $[I \perp\!\!\!\perp JK \mid L]$ is the converse of (the symmetrized version of) Decomposition, called *Composition*.
- Similarly, ② ∧ ④ $\implies$ $[I \perp\!\!\!\perp JK \mid L]$ is the converse of (the symmetrized version of) Weak union and is called *Intersection*.
- Finally, the two symmetric implications ① ∧ ④ $\implies$ $[I \perp\!\!\!\perp JK \mid L]$ and ② ∧ ③ $\implies$ $[I \perp\!\!\!\perp JK \mid L]$ seem to be almost entirely disregarded in the literature, to the point where we could not find an established name for these implications.

The focus of this paper is on sufficient conditions for Intersection and Composition; the nameless third implication is only briefly discussed in Section 5. Unlike the semigraphoid properties, Intersection and Composition are not universally valid: there exist discrete probability distributions which satisfy the premises but not the conclusion $[I \perp\!\!\!\perp JK \mid L]$. Nevertheless, they can be verified for several families of *graphical models* (see Lauritzen and Sadeghi (2018)) which play a prominent role in applications. It is also worth mentioning that the geometric notion of *partial orthogonality* which has uses in machine learning as a measure of semantic independence satisfies Composition; see Jiang et al. (2023).

Intersection classically appears as a technical condition which ensures the equivalence of different Markov properties of graphical models (see (Lauritzen, 1996, Theorem 3.7)). It also guarantees the uniqueness of Markov boundaries by Pearl and Paz (1985) and drives certain identifiability results described in Peters (2015). The Composition property is needed in the correctness proof of the IAMB algorithm to find Markov boundaries; cf. Peña et al. (2007). Continuing this line of work, more recent research of Amini et al. (2022) seeks to decouple structure learning algorithms from the graphical representation and faithfulness assumptions to generalize them to situations in which only formal properties of the independence model, such as Intersection and Composition, are assumed. This has renewed interest in sufficient conditions under which these properties hold.

The remainder of this paper is organized as follows. Section 2 performs routine manipulations to reduce Intersection and Composition to a standard form in which they turn out to be logical converses. Sections 3 and 4 survey known sufficient conditions for Intersection and Composition, respectively, discuss some interesting example classes, and derive a set of new sufficient conditions. Further remarks are collected in Section 5.

Notational conventions

Our notation for conditional independence statements largely follows the standard reference Studený (2005). In particular, N is a finite set indexing a system of jointly distributed random variables. Subsets of N are usually called $I, J, K, L, \ldots$ and elements $i, j, k, l, \ldots$. An element $i \in N$ also denotes the singleton subset $\{ i \} \subseteq N$. Union of subsets of N is abbreviated to $IJ = I \cup J$. A CI statement $[I \perp\!\!\!\perp J \mid K]$ is read as "I is independent of J given K". In Sections 3 and 4 we work concretely with four discrete random variables denoted $\mathsf{X}, \mathsf{Y}, \mathsf{Z}, \mathsf{G}$. Throughout we employ concepts such as entropy and conditional mutual information from Shannon theory for which Yeung (2005) is an accessible reference.

2 Preliminary reductions

It is well-known that any CI statement $[I \perp\!\!\!\perp J \mid K]$ with pairwise disjoint sets $I, J, K \subseteq N$ is equivalent modulo the semigraphoid axioms to a conjunction of *elementary* CI statements:

$$[I \perp\!\!\!\perp J \mid K] \iff \bigwedge_{i \in I} \bigwedge_{j \in J} \bigwedge_{K \subseteq L \subseteq IJK \setminus ij} [i \perp\!\!\!\perp j \mid L]. \tag{1}$$

The proof of this fact merely combines Decomposition and Weak union (with Symmetry) in one direction and Contraction in the other. Since the semigraphoid axioms hold for any system of discrete random variables, we may reformulate Intersection and Composition in terms of elementary CI using (1) and arrive at the following equivalent formulations:

$$\begin{aligned} \textbf{Intersection} \quad & [i \perp\!\!\!\perp j \mid kL] \wedge [i \perp\!\!\!\perp k \mid jL] \implies [i \perp\!\!\!\perp j \mid L] \wedge [i \perp\!\!\!\perp k \mid L], \\ \textbf{Composition} \quad & [i \perp\!\!\!\perp j \mid L] \wedge [i \perp\!\!\!\perp k \mid L] \implies [i \perp\!\!\!\perp j \mid kL] \wedge [i \perp\!\!\!\perp k \mid jL]. \end{aligned}$$

This is the form in which these properties are often presented in the literature on gaussoids, such as Lněnička and Matúš (2007). This also shows that Intersection and Composition are logical converses of each other modulo the semigraphoid properties.

The final reduction concerns the conditioning set L which is common to all statements in the above CI implication formulas. The "full" Intersection and Composition properties demand the above CI implications to hold for each choice of distinct $i, j, k \in N$ and $L \subseteq N \setminus ijk$. Each quadruple (i, j, k, L) encodes an *instance* of the property. In a given instance (i, j, k, L), we may marginalize the distribution to $ijkL$ and condition on L. Thus, we arrive at the following problem formulation which is addressed in this paper.

Problem. For jointly distributed discrete random variables $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$, find sufficient conditions such that

$$\begin{aligned} \textbf{Intersection} \quad & [\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}] \wedge [\mathsf{X} \perp\!\!\!\perp \mathsf{Z} \mid \mathsf{Y}] \implies [\mathsf{X} \perp\!\!\!\perp \mathsf{Y}] \wedge [\mathsf{X} \perp\!\!\!\perp \mathsf{Z}], \text{ respectively,} \\ \textbf{Composition} \quad & [\mathsf{X} \perp\!\!\!\perp \mathsf{Y}] \wedge [\mathsf{X} \perp\!\!\!\perp \mathsf{Z}] \implies [\mathsf{X} \perp\!\!\!\perp \mathsf{Y} \mid \mathsf{Z}] \wedge [\mathsf{X} \perp\!\!\!\perp \mathsf{Z} \mid \mathsf{Y}]. \end{aligned}$$

If $T(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ is a sufficient condition for Intersection or, respectively, Composition to hold in a trivariate discrete distribution, then a sufficient condition for the full Intersection or Composition property is obtained as a conjunction of $T(i, j, k \mid L = \omega)$ over all quadruples (i, j, k, L) and all events ω of L.

3 The Intersection property

The problem of finding sufficient conditions for the Intersection property has received considerable attention from a variety of research communities. The most widely known and the simplest general condition on a distribution which ensures Intersection is that the probability density is strictly positive. This is sufficient but not necessary and, depending on the application, may be too restrictive. We begin with two examples of the failure of Intersection which illustrate this condition.

Example 3.1 (Three binary random variables)**.** The joint distribution of three binary random variables is given by eight non-negative real numbers $p_{000}, p_{001}, \ldots, p_{111}$ which are indexed by triples over the set $\{0, 1\}$ and sum to one. The set of all such distributions is known as the probability simplex $\Delta(2, 2, 2)$. A generic choice of these values leads to a distribution which does not satisfy any CI statement and therefore satisfies Intersection vacuously. To violate Intersection, at least its premises must be fulfilled. The set of such distributions is the intersection of $\Delta(2, 2, 2)$ with an algebraic variety V and its structure can be examined using primary decomposition in `Macaulay2` (Grayson and Stillman) as described in Kahle et al. (2019).

Macaulay2

```
needsPackage "GraphicalModels";
R = markovRing(3:2);
I = conditionalIndependenceIdeal(R, {{{1},{2},{3}}, {{1},{3},{2}}});
J = conditionalIndependenceIdeal(R, {{{1},{2,3},{}}});
decompose(I:J)
```

The above decomposition describes the two irreducible components of V in $\Delta(2, 2, 2)$ on which there are distributions which violate Intersection. They are given by the conditions

$$p_{000} = p_{011} = p_{100} = p_{111} = 0, \text{ or} \tag{2}$$

$$p_{001} = p_{010} = p_{101} = p_{110} = 0. \tag{3}$$

As expected, violations of Intersection can only occur on the boundary of $\Delta(2, 2, 2)$ where the probability mass function has zeros and not all of the eight joint events are possible. Choosing one of the two sets of zero constraints and using generic values for the remaining four probabilities (which must sum to one) yields two 3-parameter families of distributions which satisfy the premises but not the conclusion of Intersection.

Example 3.2 (Functional dependencies)**.** The random variable X depends functionally on Y if the conditional entropy $H(X \mid Y)$ vanishes. This is equivalent to the existence of a deterministic function f such that $\Pr[X = f(Y)] = 1$, i.e., the value of Y determines the outcome of X almost surely. In this case (and if X is non-constant overall), the joint distribution cannot be strictly positive. Functional dependencies occur frequently in the context of relational databases and may present themselves in measurements of physical quantities because of the laws of nature. If X functionally depends on Z and on Y, then $[X \perp\!\!\!\perp Y \mid Z]$ and $[X \perp\!\!\!\perp Z \mid Y]$ hold and the mutual information $I(X : Y, Z)$ simplifies to $H(X)$. Thus, if X is a function of both Y and Z but non-constant (hence has positive Shannon entropy $H(X)$), then the conclusion of Intersection is not satisfied.

Remark 3.3. Note that the conditions (2) and (3) in Example 3.1 enforce in both cases that Y is a function of Z and vice versa. It is possible to violate Intersection without any functional dependencies in the distribution, but this requires larger state spaces.

The positivity of the entire distribution is unnecessarily restrictive. A more refined support condition has been developed independently by groups of statisticians, information theorists and algebraists. It is based on the following concept.

Definition 3.4. Let Y and Z be jointly distributed discrete random variables with state spaces Q_Y and Q_Z, respectively. Their *characteristic bipartite graph* $G(Y, Z)$ is the bipartite graph on $Q_Y \sqcup Q_Z$ with an edge between events y and z if and only if $\Pr[Y = y, Z = z] > 0$.

This graph appears in the work of Gács and Körner (1973) on common information where it is used to construct a random variable $GK(Y, Z)$ which solves the following optimization problem aimed at extracting the maximum entropy of a random variable which is simultaneously a function of Y and of Z:

$$\begin{aligned} &\max H(G) \\ &\text{s.t. } H(G \mid Y) = H(G \mid Z) = 0. \end{aligned} \tag{4}$$

The optimal value is known as the *Gács–Körner common information*. The solution $GK(Y, Z)$ has as its events the connected components of $G(Y, Z)$ and is specified as a function of (Y, Z) to evaluate to the connected component in which the outcomes of Y and Z both lie. Since by construction $\Pr[Y = y, Z = z] > 0$ if and only if y and z lie in the same connected component, $G(Y, Z)$ is well-defined and satisfies the functional dependence constraints in (4). In our context, its significance lies in the following fact:

Theorem 3.5. If $[X \perp\!\!\!\perp Y \mid Z]$ and $[X \perp\!\!\!\perp Z \mid Y]$, then $[X \perp\!\!\!\perp Y, Z \mid GK(Y, Z)]$.

In information theory, this is sometimes called the *double Markov property* after Exercise 16.25 in the book of Csiszár and Körner (2011).

Corollary 3.6 (Gács–Körner criterion)**.** If $G(Y, Z)$ is connected, then $[X \perp\!\!\!\perp Y \mid Z] \wedge [X \perp\!\!\!\perp Z \mid Y] \implies [X \perp\!\!\!\perp Y, Z]$.

Proof. If $G(Y, Z)$ is connected, then $GK(Y, Z)$ is a constant random variable and thus $[X \perp\!\!\!\perp Y, Z \mid GK(Y, Z)]$ simplifies to $[X \perp\!\!\!\perp Y, Z]$, the desired conclusion of Intersection. □

This sufficient condition for one instance of Intersection indirectly also targets the support of the distribution but instead of requiring positivity everywhere, it only requires enough positivity on the marginal distribution of (Y, Z) to make their characteristic bipartite graph connected. One can show that under the premises of Intersection, indeed every connected component of $G(Y, Z)$ is a complete bipartite graph. An equivalent condition in terms of σ-algebras is already present in Dawid (1980) and features in other works under the name *measurable separability*. San Martín et al. (2005) provide an overview of the history of this idea on the statistics side.

In algebraic statistics, a similar result is known as the Cartwright–Engström conjecture which was recorded in Drton et al. (2009) and resolved by Fink (2011). It asserts that the binomial ideal corresponding to the premises of Intersection has one associated prime for each possible shape of $G(Y, Z)$. Intersection holds for all distributions in the unique component for which the graph is connected and all other components contain distributions violating Intersection. This explains the computational results observed in Example 3.1.

Example 3.7 (Incompleteness of the Gács–Körner criterion)**.** The following table defines a joint distribution of four binary random variables in which G is the Gács–Körner common information of Y and Z. Since G is non-constant, the criterion of Corollary 3.6 does not apply. Nevertheless, the distribution satisfies $[X \perp\!\!\!\perp Y, Z]$ and therefore Intersection.

X	Y	Z	G	Pr
0	0	1	1	$1/4$
0	1	0	0	$1/4$
1	0	1	1	$1/4$
1	1	0	0	$1/4$

The first contribution of this paper is a set of new sufficient conditions for Intersection. Like the Gács–Körner criterion above, they are formulated *synthetically*, i.e., in terms of an auxiliary random variable G which satisfies additional CI constraints with respect to X, Y, Z. In this situation, the random variables are subject to powerful information-theoretic inequalities. We take advantage of recent work of Studený (2021) which elucidates the connections between CI implications on four discrete random variables and special information-theoretic constraints known as *conditional Ingleton inequalities*.

Theorem 3.8 (Conditional Ingleton criterion)**.** Let X, Y, Z be jointly distributed discrete random variables. If there exists a discrete G jointly distributed with X, Y, Z satisfying any of the following conditions: (i) $[X \perp\!\!\!\perp G]$ and $[Y \perp\!\!\!\perp Z \mid G]$, (ii) $[Y \perp\!\!\!\perp G]$ and $[X \perp\!\!\!\perp Z \mid G]$, or (iii) $[Z \perp\!\!\!\perp G]$ and $[X \perp\!\!\!\perp Y \mid G]$; then $[X \perp\!\!\!\perp Y \mid Z] \wedge [X \perp\!\!\!\perp Z \mid Y] \implies [X \perp\!\!\!\perp Y, Z]$ holds.

Proof. Given $[X \perp\!\!\!\perp Y \mid Z]$ and $[X \perp\!\!\!\perp Z \mid Y]$, the conditions (ii) and (iii) are symmetric with respect to exchanging Y and Z and both follow from rule (I:2) in Studený (2021). Condition (i) is covered by rule (I:4). □

In order to compare Theorems 3.5 and 3.8, assume that G is a function of Y and of Z. In condition (ii), the independence assumption $[Y \perp\!\!\!\perp G]$ then implies that G is constant and hence the further assumption $[X \perp\!\!\!\perp Z \mid G]$ simplifies to the desired conclusion already; a similar argument applies to condition (iii). Regarding condition (i), the assumption $[Y \perp\!\!\!\perp Z \mid G]$ is equivalent to $H(G) = I(Y : Z)$. This is highly unusual when G is a function of Y and of Z. Indeed, the Gács–Körner theorem (see, e.g., Csirmaz (2023)) asserts that this can only happen if the probability table of (Y, Z) can be brought into block-diagonal form by permutations of its rows and columns and each block has rank one. It appears that the two criteria in Theorems 3.5 and 3.8 are complementary and neither implies the other. Note that the distribution given in Example 3.7 does not satisfy the Gács–Körner criterion but does satisfy the conditional Ingleton criterion Theorem 3.8 (i) since $G = Z = 1 - Y$ are functionally equivalent and the marginal (X, G) is uniform.

4 The Composition property

The previous section showed that the Intersection property is well-studied. By comparison, not much is known about the failure modes of Composition. Studený (2005), Corollary 2.4, shows that Gaussians (even with singular covariance matrices) satisfy Composition. Since many types of graphical models can be faithfully represented by Gaussians, they inherit the Composition property from the Gaussian. In the discrete setting, a known sufficient condition is *multivariate total positivity of order 2* (MTP_2) which is a type of log-supermodularity condition on the density function. Fallat et al. (2017) show that MTP_2 implies upward stability (i.e., $[I \perp\!\!\!\perp J \mid K] \implies [I \perp\!\!\!\perp J \mid L]$ for any $L \supseteq K$), which is far stronger than Composition. We again begin the investigation with two example classes in which Composition is violated.

Example 4.1 (Matroids)**.** Matroids provide a class of functional dependence structures which are incompatible with the Composition property. For background information on matroids and their probabilistic representations, we refer to Matúš (1994) and its references. In a matroid M on ground set N, any two elements which are not loops are either functionally equivalent or independent. Under the benign assumption that the matroid is *simple*, i.e., contains no loops and no two functionally equivalent elements, the independence $[i \perp\!\!\!\perp j]$ holds for any $i \neq j$. If the full Composition property were to hold as well, then it follows inductively that $[I \perp\!\!\!\perp J]$ for all disjoint $I, J \subseteq N$. This implies that the matroid is free and hence there is only one simple matroid satisfying Composition.

Example 4.2 (Three binary random variables)**.** The assumptions $[X \perp\!\!\!\perp Y]$ and $[X \perp\!\!\!\perp Z]$ define a *marginal independence model* which can be easily parametrized using the results of Kirkup (2007). For binary states, this parametrization is as follows:

$$\begin{aligned} p_{000} &= \alpha\beta\gamma - \delta, & p_{100} &= \overline{\alpha}\beta\gamma - \varepsilon, \\ p_{001} &= \alpha\beta\overline{\gamma} + \delta, & p_{101} &= \overline{\alpha}\beta\overline{\gamma} + \varepsilon, \\ p_{010} &= \alpha\overline{\beta}\gamma + \delta, & p_{110} &= \overline{\alpha}\overline{\beta}\gamma + \varepsilon, \\ p_{011} &= \alpha\overline{\beta}\overline{\gamma} - \delta, & p_{111} &= \overline{\alpha}\overline{\beta}\overline{\gamma} - \varepsilon, \end{aligned} \tag{5}$$

where $\alpha, \beta, \gamma \in (0,1)$ and $\overline{x} = 1 - x$; the values of δ and ε are subject to the conditions that all these probabilities must be non-negative. If $\alpha = \beta = \gamma = 1/2$, $\delta = 0$ and $\varepsilon > 0$ is small, then the parametrization defines a probability distribution which satisfies $[X \perp\!\!\!\perp Y]$ and $[X \perp\!\!\!\perp Z]$ but $I(X : Y, Z) = 8\varepsilon^2 + \mathcal{O}(\varepsilon^3) > 0$. Hence, this distribution violates Composition.

On the other hand, the parametrization technique from Boege et al. (2022) can also be used to describe the distributions on which $[X \perp\!\!\!\perp Y, Z]$ holds true. It is the submodel parametrized by (5) where $\varepsilon = \delta \cdot \overline{\alpha}/\alpha$.

Example 4.2 shows that there are strictly positive distributions which do not satisfy the Composition property. Thus, Composition does not admit sufficient conditions which require a "richness of support" like Corollary 3.6 in the case of Intersection. The moral of the Cartwright–Engström story is that a sufficient condition may still be encoded in the primary decomposition of the Composition ideal, even if it does not take the form of

support constraints. Recall that the Intersection ideal has one minimal prime for each possible characteristic bipartite graph of Y and Z. This rich structure invites further investigation which leads to Corollary 3.6. However, Kirkup (2007) proved that the Composition ideal has only one minimal prime whose variety contains any probability distribution at all. Thus, there is no relevant structure in the primary decomposition and this approach is also a dead end. The only strategy which we found applicable to Composition is the one using conditional information inequalities.

Theorem 4.3 (Dual conditional Ingleton criterion)**.** Let X, Y, Z, G be jointly distributed discrete random variables satisfying any of the following conditions: (i) $[X \perp\!\!\!\perp G \mid Y, Z]$ and $[Y \perp\!\!\!\perp Z \mid X]$, (ii) $[Y \perp\!\!\!\perp G \mid X, Z]$ and $[X \perp\!\!\!\perp Z \mid Y]$, or (iii) $[Z \perp\!\!\!\perp G \mid X, Y]$ and $[X \perp\!\!\!\perp Y \mid Z]$. Then $[X \perp\!\!\!\perp Y \mid G] \wedge [X \perp\!\!\!\perp Z \mid G] \implies [X \perp\!\!\!\perp Y, Z \mid G]$ holds.

Proof. Analogously to the proof of Theorem 3.8, the conditions (ii) and (iii) are symmetric and they follow from (I:14) in Studený (2021). Condition (i) follows from (I:19). □

Note that in Theorem 4.3, the Composition property is obtained simultaneously for all conditional distributions given the auxiliary G. This makes the criterion appear to be somewhat harder to work with as it requires a suitable coupling of the conditional distributions through G. The following Examples 4.4 and 4.5 show possible applications and provide a glimpse at the complexity of the underlying real algebraic geometry. Concerning comparisons with pre-existing criteria for Composition, we remark that every Bayesian network which satisfies the four CI statements required in Theorem 4.3 (i) also satisfies $[X \perp\!\!\!\perp Y, Z, G]$ which is much stronger than required for the instance of Composition.

Example 4.4. Start with any distribution satisfying $[Y \perp\!\!\!\perp Z \mid X]$ and take a function G of (Y, Z). The resulting joint distribution satisfies $[X \perp\!\!\!\perp G \mid Y, Z]$ and thus Theorem 4.3 (i). If none of X, Y, Z, in turn, functionally depend on G, the conditional distributions are non-trivial and satisfy Composition (perhaps vacuously).

Example 4.5. Consider the two distributions p and q of three jointly distributed binary random variables X, Y, Z parametrized as follows:

$$\begin{array}{llll} p_{000} = 0, & p_{100} = 0, & q_{000} = {}^1\!/_4\,\xi(4\eta - 1), & q_{100} = {}^1\!/_4\,\overline{\xi}(4\eta - 1) \\ p_{001} = {}^1\!/_2\,\alpha, & p_{101} = {}^1\!/_2\,\overline{\alpha}, & q_{001} = {}^1\!/_4\,\xi, & q_{101} = {}^1\!/_4\,\overline{\xi} \\ p_{010} = {}^1\!/_4\,\alpha, & p_{110} = {}^1\!/_4\,\overline{\alpha}, & q_{010} = \xi\overline{\eta}, & q_{110} = \overline{\xi}\overline{\eta}, \\ p_{011} = {}^1\!/_4\,\alpha, & p_{111} = {}^1\!/_4\,\overline{\alpha}, & q_{011} = 0, & q_{111} = 0, \end{array} \tag{6}$$

for $\alpha, \xi \in (0, 1)$ and $\eta \in [{}^{17}\!/_{18}, 1)$. Both of these families of distributions satisfy an instance of Composition. Indeed, the mixture of the two distributions defined via

$$\Pr[X = x, Y = y, Z = z, G = g] = \overline{g}\lambda p_{xyz} + g\overline{\lambda} q_{xyz}, \text{ where } g \in \{0, 1\},$$

satisfies the conditions of Theorem 4.3 (i) whenever $\xi = \alpha$ and $4\lambda = 1 \pm \sqrt{\frac{18\eta - 17}{2\eta - 1}}$.

5 Remarks

Duality. Intersection and Composition are not only converses modulo the semigraphoid axioms but also *dual*. For an elementary CI statement $[i \perp\!\!\!\perp j \mid L]$ over ground set N, the *dual statement* is $[i \perp\!\!\!\perp j \mid L]^* \coloneqq [i \perp\!\!\!\perp j \mid N \setminus ijL]$. Applying duality statement-wise transforms

$$\begin{aligned}
&\textbf{Intersection} && [i \perp\!\!\!\perp j \mid kL] \wedge [i \perp\!\!\!\perp k \mid jL] \implies [i \perp\!\!\!\perp j \mid L] \wedge [i \perp\!\!\!\perp k \mid L] \text{ to} \\
&\textbf{Intersection}^* && [i \perp\!\!\!\perp j \mid Ł] \wedge [i \perp\!\!\!\perp k \mid Ł] \implies [i \perp\!\!\!\perp j \mid kŁ] \wedge [i \perp\!\!\!\perp k \mid jŁ],
\end{aligned}$$

where $Ł = N \setminus ijkL$. But this is the Composition property with L replaced by Ł. Hence, the dual of a CI structure satisfying Intersection is a CI structure satisfying Composition and vice versa. Remarkably, the sets of sufficient conditions obtained in Theorems 3.8 and 4.3 are also formally dual to each other. This is a feature of the conditional information inequalities used in their proofs, although in general it is not true that any valid conditional information inequality can be dualized and remain valid.

Denote by $\mathcal{I}_4$ the set of CI structures which are representable by four discrete random variables and satisfy all instances of Intersection; analogously $\mathcal{C}_4$ for the Composition property. It can be verified that $|\mathcal{I}_4| = |\mathcal{C}_4| = 5\,736$ and they both have the same number of elements modulo the action of the symmetric group S_4 on the random variables. These orbit representatives are usually called *permutational types* and $\mathcal{I}_4$ and $\mathcal{C}_4$ both have 369 of them. However, this coincidence of numbers is *not* explained by duality. For example, the CI structure of the distribution in Studený (2021), Example 4, is in $\mathcal{I}_4 \cap \mathcal{C}_4$ but its dual is not probabilistically representable as it violates Studený's rule (I:1).

Moreover, the sets $\mathcal{I}_4$ and $\mathcal{C}_4$ have a natural lattice structure. Using code adapted from Boege et al. (2025), we have computed that $\mathcal{I}_4$ has 23 permutational types of irreducible elements and that $\mathcal{C}_4$ has 24 such permutational types. Hence, the lattices are not isomorphic. In view of this incompatibility, we believe that the coincidence of the cardinalities of $\mathcal{I}_4$ and $\mathcal{C}_4$ is an artifact of the small ground set size rather than a reflection of a deeper connection between the two properties.

Relation to Gaussianity. Regular Gaussian distributions satisfy both, Intersection and Composition. For the third implication $[X \perp\!\!\!\perp Y] \wedge [X \perp\!\!\!\perp Y \mid Z] \implies [X \perp\!\!\!\perp Y, Z]$ briefly discussed in Section 1, note that the premises are symmetric under exchanging X and Y, but the consequence is not. This means that $[X \perp\!\!\!\perp Y]$ and $[X \perp\!\!\!\perp Y \mid Z]$ may be derived from $[X \perp\!\!\!\perp Y, Z]$ as well as $[Y \perp\!\!\!\perp X, Z]$ using the semigraphoid axioms. A more symmetric formulation of the converse implication

$$[X \perp\!\!\!\perp Y] \wedge [X \perp\!\!\!\perp Y \mid Z] \implies [X \perp\!\!\!\perp Y, Z] \vee [Y \perp\!\!\!\perp X, Z]$$

is sometimes referred to as *weak transitivity* and is known to hold for Gaussians as well. This analysis suggests that the realm of Gaussian random variables is usually more pleasant to work in as far as elementary properties of conditional independence, such as the semigraphoid properties and their converses, are concerned.

The third implication. The proper (unsymmetrized) form of the third converse implication has been considered in the work of Dawid (1980) which features a sufficient condition derived from a generalization of Basu's theorem. For discrete random variables, our approach of using conditional information inequalities also applies.

Theorem 5.1. Let X, Y, Z be jointly distributed discrete random variables. If there exists a discrete G jointly distributed with X, Y, Z satisfying $[X \perp\!\!\!\perp Z \mid G]$ and $[Z \perp\!\!\!\perp G \mid Y]$, then $[X \perp\!\!\!\perp Y] \wedge [X \perp\!\!\!\perp Y \mid Z] \implies [X \perp\!\!\!\perp Y, Z]$ holds.

Proof. This follows from (I:7) in Studený (2021). □

Conditional information inequalities for tight Composition. Matúš (2006) gives a complete characterization of the entropy profiles of discrete X, Y, Z which satisfy $[X \perp\!\!\!\perp Y]$ and $[X \perp\!\!\!\perp Z]$ and such that every variable is a function of the other two. The latter condition is often referred to as *tightness* of the entropy profile. Such linear information inequalities can be used to give sufficient conditions for Composition in terms of the joint entropies of the random variables. Unfortunately, the tightness assumption which drives Matúš's proof is very restrictive in this case. It is easy to check that a tight distribution which satisfies $[X \perp\!\!\!\perp Y, Z]$ must have X constant and Y and Z functions of each other. It would certainly be interesting to investigate conditional information inequalities for entropy vectors satisfying $[X \perp\!\!\!\perp Y]$ and $[X \perp\!\!\!\perp Z]$ but which are not tight.

Operational characterizations of Theorems 3.8 and 4.3. One of the merits of the Gács–Körner criterion for Intersection is that the auxiliary variable GK can be directly constructed and has an operational interpretation as the common information of Y and Z. Both of these aspects have to be left unexplored in this article for the auxiliary variables appearing in Theorems 3.8 and 4.3. It would be interesting to attach an operational meaning to these random variables or to provide direct constructions, even in special cases.

Limits of discrete distributions. Some information-theoretic constructions define a sequence of random variables Y_n as functions of n i.i.d. copies of a given variable X. A desired information-theoretic effect only occurs "in the limit" of this construction which is possibly realized by a random variable Y^* with infinite support. For example, consider again the setup in Example 4.4. Using Matúš (2007), Theorem 3 and Corollary 2, it is possible to construct discrete random variables Y^*, Z^*, G_1^*, G_2^* with possibly infinite support such that:

- $H(G_1^* \mid Y^*) = I(X : Y^* \mid G_1^*) = 0$,
- $H(G_2^* \mid Z^*) = I(X : Z^* \mid G_2^*) = 0$, and
- the entropy profiles of (X, Y, Z) and (X, Y^*, Z^*) agree.

Then $G^* = (G_1^*, G_2^*)$ is a function of (Y^*, Z^*) and the distribution (X, Y^*, Z^*, G^*) satisfies the CI constraints of Theorem 4.3 (i) as well as the premises of Composition given G^*. However, it is not known whether the conditional Ingleton inequalities on which Theorem 4.3 relies are valid for distributions with infinite supports. See also Open Question 2 in Studený (2021) about the validity on these inequalities on the almost-entropic region.

Acknowledgements

I would like to thank Bryon Aragam for discussions on the Composition property. This research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101110545.

References

A. A. Amini, B. Aragam, and Q. Zhou. A non-graphical representation of conditional independence via the neighbourhood lattice, 2022.

T. Boege, S. Petrović, and B. Sturmfels. Marginal independence models. In *Proceedings of the 2022 International Symposium on Symbolic and Algebraic Computation*, ISSAC '22, pages 263–271. Association for Computing Machinery (ACM), 2022. doi: 10.1145/3476446.3536193.

T. Boege, J. H. Bolt, and M. Studený. Self-adhesivity in lattices of abstract conditional independence models. *Discrete Applied Mathematics*, 361:196–225, 2025. doi: 10.1016/j.dam.2024.10.006.

L. Csirmaz. A short proof of the Gács–Körner theorem, 2023.

I. Csiszár and J. Körner. *Information theory. Coding theorems for discrete memoryless systems*. Cambridge University Press, 2nd ed. edition, 2011. doi: 10.1017/CBO9780511921889.

A. P. Dawid. Conditional independence for statistical operations. *Ann. Stat.*, 8:598–617, 1980. doi: 10.1214/aos/1176345011.

M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Semin.* Birkhäuser, 2009. doi: 10.1007/978-3-7643-8905-5.

S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in Markov structures. *Ann. Stat.*, 45(3):1152–1184, 2017. doi: 10.1214/16-AOS1478.

A. Fink. The binomial ideal of the intersection axiom for conditional probabilities. *J. Algebr. Comb.*, 33(3):455–463, 2011. doi: 10.1007/s10801-010-0253-5.

P. Gács and J. Körner. Common information is far less than mutual information. *Probl. Control Inf. Theory*, 2:149–162, 1973.

D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at `http://www2.macaulay2.com`. Version 1.22.

Y. Jiang, B. Aragam, and V. Veitch. Uncovering meanings of embeddings via partial orthogonality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23. Curran Associates Inc., 2023.

T. Kahle, J. Rauh, and S. Sullivant. Algebraic aspects of conditional independence and graphical models. In M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors, *Handbook of graphical models*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pages 61–80. CRC Press, 2019.

G. A. Kirkup. Random variables with completely independent subcollections. *J. Algebra*, 309(2):427–454, 2007. doi: 10.1016/j.jalgebra.2006.06.023.

S. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press, 1996.

S. Lauritzen and K. Sadeghi. Unifying Markov properties for graphical models. *Ann. Statist.*, 46(5):2251–2278, 2018. doi: 10.1214/17-AOS1618.

R. Lněnička and F. Matúš. On Gaussian conditional independence structures. *Kybernetika*, 43(3):327–342, 2007.

F. Matúš. Probabilistic conditional independence structures and matroid theory: Background. *Int. J. Gen. Syst.*, 22(2):185–196, 1994. doi: 10.1080/03081079308935205.

F. Matúš. Piecewise linear conditional information inequality. *IEEE Trans. Inf. Theory*, 52(1):236–238, 2006. doi: 10.1109/TIT.2005.860438.

F. Matúš. Two constructions on limits of entropy functions. *IEEE Trans. Inf. Theory*, 53 (1):320–330, 2007. doi: 10.1109/TIT.2006.887090.

J. Pearl and A. Paz. GRAPHOIDS: A graph-based logic for reasoning about relevance relations, or When would x tell you more about y if you already know z. Technical Report CSD-850038, UCLA Computer Science Department, 1985.

J. M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. Towards scalable and data efficient learning of Markov boundaries. *Int. J. Approx. Reasoning*, 45(2):211–232, 2007. doi: 10.1016/j.ijar.2006.06.008.

J. Peters. On the intersection property of conditional independence and its application to causal discovery. *J. Causal Inference*, 3(1):97–108, 2015. doi: 10.1515/jci-2014-0015.

E. San Martín, M. Mouchart, and J.-M. Rolin. Ignorable common information, null sets and Basu's first theorem. *Sankhyā*, 67(4):674–698, 2005.

M. Studený. *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer, 2005.

M. Studený. Conditional independence structures over four discrete random variables revisited: conditional Ingleton inequalities. *IEEE Trans. Inf. Theory*, 67(11):7030–7049, 2021. doi: 10.1109/TIT.2021.3104250.

R. W. Yeung. *A first course in information theory*. Information Technology: Transmission, Processing and Storage. Springer, 2005. doi: 10.1007/978-1-4419-8608-5.

Information Fusion in Sentiment Fuzzy Rule-Based systems: How to Improve Readability and Robustness through SMART Operators

Andrea Capotorti[1], Davide Petturiti[1], and Barbara Vantaggi[2]

[1]University of Perugia, Italy
{*andrea.capotorti, davide.petturiti*}*@unipg.it*
[2]University of Rome "la Sapienza", Italy
barbara.vantaggi@uniroma1.it

Abstract

We propose the adoption of the newly introduced SMART operators to improve the readability of a Fuzzy rule-based system developed for sentiment analysis. Specifically, the S-or operator will be utilized to combine the results of various rules generated from a designated lexicon. Conversely, the S-and operator will merge the diverse inferences derived from multiple lexicons.

1 Introduction

Fuzzy rule-based systems, such as those discussed in Alonso Moral et al. (2021)), offer outputs that are inherently explainable. This is particularly evident when these systems are utilized in sentiment analysis, as demonstrated in Liu and Cocea (2017). However, traditional Mamdani-type systems (Mamdani (1974); Mamdani and Assilian (1975)), employing standard min-max aggregation operators, often yield outputs that are challenging to interpret prior to the final defuzzification step . To address this issue, we propose the adoption of the SMART disjunctive operator (referred to as S-or hereafter), which was recently introduced in Capotorti and Figà-Talamanca (2020). By utilizing the S-or operator instead of the conventional max operator for aggregation, we aim to produce more easily interpretable outputs for the consequents of the rules. This approach enhances the overall interpretability of the fuzzy rule-based system, ultimately improving the clarity and understanding of the system's outputs.

In sentiment analysis, Fuzzy rule-based systems typically rely on scores derived from lexicons to provide crisp inputs (see, e.g., Nadali et al. (2010)). However, it is important to note that different lexicons can yield varying results, as demonstrated in previous studies, such as Chauhan et al. (2023); Vashishtha and Susan (2019). Usually, only the

final defuzzified crisp values are utilized to evaluate classifier performance, overlooking the richness and diversity inherent in the various Fuzzy outputs. To address this limitation, we apply a novel approach of merging multiple Fuzzy outputs using the SMART conjunctive operator (S-and hereafter) before the defuzzification process, as suggested in Capotorti and Figà-Talamanca (2020). This method allows for the creation of an ensemble of classifications, offering a unique perspective on sentiment analysis. Our approach differs significantly from existing methods, such as Shapiro and Moritz Sudhof (2020)), and provides a fresh and innovative solution to the challenges in sentiment analysis.

The paper is organized in the following way: In Section 2 we briefly illustrate a Mamdami-type Fuzzy Rule-Base apt to deal with sentiment analysis with two different scores obtained through a lexicon and how the S-or operator can be employed to obtain a more interpretable output membership. Moreover, we propose a new dufuzzification operator that is more accurate of the usual center of area (COA). In Section 3.1 we propose the adoption of the other aggregation operator S-and to join the information stemming from different sources, in particular different lexicons, obtaining and ensemble of classifiers. Finally, a concluding section summarizes the contribution and hints about future developments.

2 FUZZY RULE-BASED SYSTEM

Our proposal is based on the Fuzzy Rule-Based system outlined in the study by Vashishtha and Susan (2019). This system was chosen for its clear and easily understandable design, as well as its strong classification performance. However, it is important to note that our method is versatile and can be applied to any Fuzzy Rule-Based system of the Mamdani type.

In the aforementioned paper by Vashishtha and Susan (2019), the authors utilized a system consisting of 9 well-defined rules, as detailed in Tab. 1; Each rule is composed of two antecedent variables, each with three Fuzzy subsets (Low, Medium, High), and one consequent variable with three Fuzzy subsets (Negative, Neutral, Positive).

Table 1: The nine Mamdani Fuzzy rules proposed in Vashishtha and Susan (2019).

RULE	Pos. score	Neg. score	Sentiment
R1	Low	Low	Neutral
R2	Medium	Low	Positive
R3	High	Low	Positive
R4	Low	Medium	Negative
R5	Medium	Medium	Neutral
R6	High	Medium	Positive
R7	Low	High	Negative
R8	Medium	High	Negative
R9	High	High	Neutral

The two input variables, $X1$ and $X2$, respectively represent the negative and the

positive sentiment score of a specific text obtained by adopting one specific lexicon (e.g. SentiWordNet Baccianella et al. (2010), AFINN Nielsen (2011) or VADER Hutto and Gilbert (2015)). For the three linguistic terms Low-Medium-High, authors use a standard Fuzzy partition of the domain of each input score with three triangular memberships $\mu_i(x)$, $i \in \{\text{Low, Medium, High}\}$, as those depicted in Fig.1 Usually, but not necessarily, input

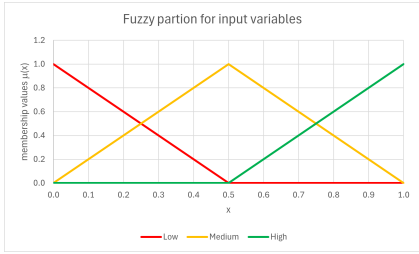


Figure 1: Fuzzy partition for input variables.

domains are normalized into the real unit interval $[0, 1]$.

For the output variable Y, representing the overall sentiment judgment of the selected text (in the aforementioned paper authors analyze Twitter posts), authors use a similar Fuzzy partition with three linguistic terms, but on the standard scale of numbers between 0 an 10, with memberships $\mu_i(y)$, $i \in \{neg, neu, pos\}$, illustrated in Fig.2

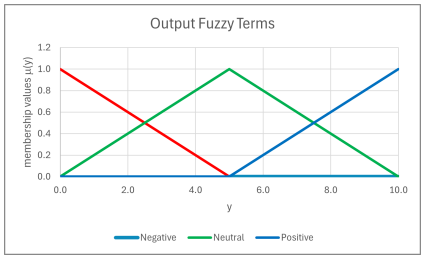


Figure 2: Fuzzy linguistic terms for the output variable.

The sentiment classification for each text is performed through the following steps.

- The two input scores $x1$ and $x2$ are fuzzified into the membership values $\mu_{j1}(x1)$ and $\mu_{j2}(x2)$, respectively, of each of the 9 rules in Tab.1.

- Through the usual minimum t-norm, the firing strength of the j-th rule is obtained as:

$$w_{Rj} = \mu_{j1}(x1) \wedge \mu_{j2}(x2). \tag{1}$$

- The usual maximum t-conorm $\vee$ is applied for combining the firing strengths derived in the previous step in order to obtain the fulfillment of each output term:

$$w_{neg} = w_{R4} \vee w_{r7} \vee w_{R8} \quad (2)$$
$$w_{neu} = w_{R1} \vee w_{R5} \vee w_{R9} \quad (3)$$
$$w_{pos} = w_{R2} \vee w_{R3} \vee w_{R6} \quad (4)$$

- The resultant output memberships for all terms are obtained through the usual minimum t-norm $\wedge$:

$$\tilde{\mu}_{neg}(y) = w_{neg} \wedge \mu_{neg}(y); \quad (5)$$
$$\tilde{\mu}_{neu}(y) = w_{neu} \wedge \mu_{neu}(y); \quad (6)$$
$$\tilde{\mu}_{pos}(y) = w_{pos} \wedge \mu_{pos}(y). \quad (7)$$

- The final output memberships $\tilde{\mu}_j$, $j \in \{neg, neu, pos\}$, are again aggregated through the usual maximum t-conorm $\vee$ before being defuzzified:

$$\mu_{\vee-\max}(y) = \tilde{\mu}_{neg}(y) \vee \tilde{\mu}_{neu}(y) \vee \tilde{\mu}_{pos}(y). \quad (8)$$

The usual aggregation produces memberships $\mu_{\vee-\max}$ that are not easy to interpret due to their potential extreme vagueness, e.g. those depicted in Fig.3.

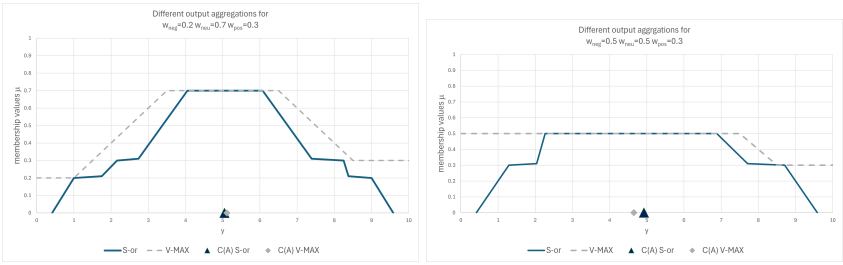


Figure 3: Comparison of aggregated output memberships obtained by $\vee-$ max or S-or operators.

We perform here the same exercise by applying the disjunctive aggregation operator S-or in the last step; the results highlight that this alternative merge produces less vague and concave memberships. Indeed, $\mu_{S-or} \subseteq \mu_{\vee-\max}$ using usual Fuzzy inclusion relation, and μ_{S-or} displays nested α-cuts. Such membership can be easy interpreted as a truncated Fuzzy number. Fig.3 provides a visual comparison of μ_{S-or} w.r.t. $\mu_{\vee-\max}$).

2.1 The S-or aggregation operator

The S-or function operates by calculating a weighted average of the extreme values of the different α-cuts of the m memberships to be aggregated. The weights are carefully adjusted to achieve a specific behavior in the aggregation process, focusing on the outermost values. This approach aligns with the canonical max t-conorm $\vee$ if applied vertically. The weighting emphasizes the disagreement among the different α-cuts in terms of weak overlapping. To achieve this, when fixing an α-cut, the weights of the outermost values, identified by indexes in O_l for the left extrema and O_r for the right extrema, are calculated as $\frac{1}{m}(1+\epsilon_j)$, with

$$\epsilon_j = \begin{cases} \frac{\sum_{f=1}^{M} \frac{1}{f}\pi_f^j}{\Delta} & \text{if } \Delta \neq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{9}$$

In (9) the π_f^j's are the lengths of the various overlappings while Δ is the range of the α-cuts, as depicted in Fig.4.

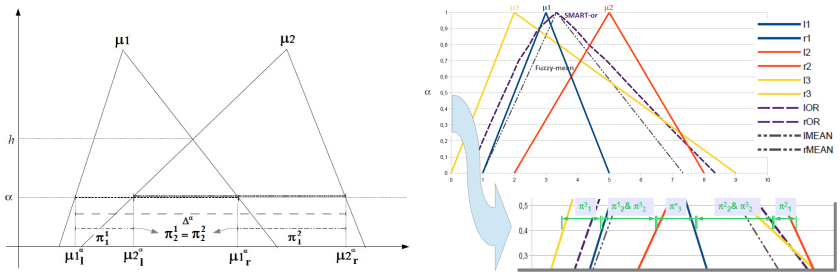


Figure 4: Overlapping lengths π_f^j and range Δ of α-cuts involved in the weights ϵ_j for the S-or aggregation: above between two memberships; below among three.

For the two inner extremes, i.e. the largest left extreme and the lowest right extreme, the weights are simply given by $\frac{1}{m}(1-\sum_{j\in O_*}\epsilon_j)$.

The memberships m that need to be aggregated are truncated by the firing strengths in equations (5, 6, 7). As a result, these memberships may not be normal, meaning there is no value $\tilde{y}$ for which $\tilde{\mu}_j(\tilde{y}) = 1$. This is different from the original formulation in Capotorti and Figà-Talamanca (2020), where the memberships were normal fuzzy numbers. Our goal is to ensure that the output of this merging process is a fuzzy number, possibly truncated. When the number of α-cuts to aggregate changes, the resulting extremes must align with those of the lower levels. This alignment can be achieved through a proper translation and deformation of the extremes.

Due to the necessary discretization of the α levels to be considered in practical applications, we can formulate the transformation by referring to two consecutive values α_m and α_{m+1}. The extremes of the transformed α_{m+1}-cut are computed recursively as follows:

$$\widehat{\mu}_l^{\alpha_{m+1}} = \widehat{\mu}_l^{\alpha_m} + \varrho^{\alpha_m} |\overline{\mu}_l^{\alpha_{m+1}} - \overline{\mu}_l^{\alpha_m}| \tag{10}$$

$$\widehat{\mu}_r^{\alpha_{m+1}} = \widehat{\mu}_r^{\alpha_m} - \varrho^{\alpha_m} |\overline{\mu}_r^{\alpha_{m+1}} - \overline{\mu}_r^{\alpha_m}| \tag{11}$$

with

$$\varrho^{\alpha_m} = \frac{\widehat{\mu}_r^{\alpha_m} - \widehat{\mu}_l^{\alpha_m}}{\overline{\mu}_r^{\alpha_m} - \overline{\mu}_l^{\alpha_m}} \tag{12}$$

and where the "overlined" extremes are those obtained by the previously described S-or and the "hatted" ones are those finally obtained by the transformation at the specified levels.

It is important to note that the typical method of discretizing the membership functions $\tilde{\mu_j}$, where $j \in \{neg, neu, pos\}$, results in step functions with a finite number of distinct α-cuts. This characteristic makes the S-or operator easily applicable and effectively implementable for our purposes. It is worth mentioning that the original formulation in Capotorti and Figà-Talamanca (2020) was intended for Fuzzy numbers, whereas we are working with more general Fuzzy quantities.

2.2 Defuzzification

The final step in text sentiment classification involves synthesizing the final output membership, denoted as $\mu_{\vee-\max}$ or μ_{S-or}, into a single real value that represents the overall sentiment of the text. This value is then used to assign a final sentiment label based on a predetermined decision rule. For example, the following decision rule adopted in Vashishtha and Susan (2019):

$$\text{Label} = \begin{cases} \text{Negative} & \text{if } COA \in [0, 3.3[\\ \text{Neutral} & \text{if } COA \in [3.3, 6.7[\\ \text{Positive} & \text{if } COA \in [6.7, 10] \end{cases} \tag{13}$$

where COA is the usual defuzzified center of the area value of their aggregated membership $\mu_{\vee-\max}$ usually approximated by

$$COA = \frac{\sum_{y_i \in \mathcal{Y}} y_i \, \mu_{\vee-\max}(y_i)}{\sum_{y_i \in \mathcal{Y}} \mu_{\vee-\max}(y_i)}, \tag{14}$$

summations being over the partition $\mathcal{Y} = \{y_0, y_1, \ldots, y_n\}$ chosen for the discrete operational representation of the memberships (in the aforementioned paper authors chose, e.g., $\mathcal{Y} = \{0, 1, \ldots, 10\}$).

Since the polygonal shape of the output membership μ_{S-or}, we suggest instead to approximate the center of the area through the more accurate value proposed in Naimi and Tahayori (2020) and defined as:

$$C(A) = \frac{\sum_{j=1}^{n} \left((y_j - y_{j-1}) \left[\mu_{j-1} \frac{(2y_{j-1} + y_j)}{3} + \mu_j \frac{(y_{j-1} + 2y_j)}{3} \right] \right)}{\sum_{j=1}^{n} ((y_j - y_{j-1})(\mu_{j-1} + \mu_j))}. \tag{15}$$

where μ_{j-1} and μ_j stay for $\mu_{S-or}(y_{j-1})$ and $\mu_{S-or}(y_j)$, respectively, and $\mathcal{Y} = \{0, 0.1, \ldots, 9.9, 10\}$.

The center of the area is quite insensitive to the specific shape of the membership functions, consequently, we obtain almost the same values with $\mu_{\vee-\max}$ or μ_{S-or}. Hence, concerning the classification aim, our proposal reaches the same performances of Vashishtha and Susan (2019). In fact, our goal is to improve the interpretability rather than the performance and we expect that this characteristic of our approach will be valuable when applied to real and extensive datasets.

3 An Ensemble of Sentiment Classifier

As mentioned in the introduction, sentiment classifications are typically conducted using different lexicons, with each lexicon providing a classification for a given instance. However, there are instances where different lexicons may produce varying classifications for the same data point. For example, in the Sanders dataset referenced in Vashishtha and Susan (2019), tweet # 3420 is labeled as Neutral using the AFINN lexicon with a COA of 5.4, but is classified as Positive when processed through the VADER lexicon with a COA of 7.67. To address this issue, the results from different lexicons can be aggregated to provide a more comprehensive classification. One approach is to calculate the arithmetic mean of the centers of the areas $C(A)$'s in (15), but this method may not account for the unique shapes of the membership functions. A weighted average could offer a more nuanced solution, although determining the appropriate weights can be challenging.

An alternative method is to utilize the conjunctive SMART operator S-and, as introduced in Capotorti and Figà-Talamanca (2020). This operator considers the agreements among the α-cuts of the initial membership functions, resulting in a more robust aggregation of output memberships. By combining the output memberships obtained from different lexicons, a single aggregated membership can be obtained using the S-and operator. The center of the area of this aggregated membership can then be used to determine the classification label through a decision rule.

Precisely, by aggregating the different output memberships μ_{s-or}'s obtained for different lexicons, a single aggregated joined membership μ_{S-and} is obtained; by computing the center of the area (15) of μ_{S-and} we can eventually decide the classification label through some decision rule, like, e.g., (13). As a result, this ensemble approach provides a weighted mean of the different classifications, with the final center of the area of the aggregated membership representing a combination of the centers of the individual memberships, implicitly weighted based on their agreement or disagreement.

3.1 The S-and aggregation operator

In order to gain a simple understanding of the S-and operator, let us briefly outline the main steps involved. For more detailed information, please refer to the source cited as Capotorti and Figà-Talamanca (2020).

The S-and operator is a variation of the Fuzzy mean that considers the full or partial overlap among the α-cuts of the Fuzzy memberships being combined.

Let m represent the number of membership functions to be aggregated, $I_l \subset \{1, \ldots, m\}$ denote the set of indices of the last $m-1$ left extremes, and $I_r \subset \{1, \ldots, m\}$ represent the set of indices of the first $m-1$ right extremes (in ascending order) of the m α-cuts in input. The construction of the S-and operator follows a similar basic rule to the S-or operator described in Subsection 2.1, but requires a more complex formulation. The computation of the α-cuts of the result varies depending on whether α is below or above the value $h \in (0, 1)$, which represents the highest level of non-empty intersection among all the m α-cuts of the original Fuzzy numbers (see Fig.4 for the case $m = 2$).

If $\alpha \leq h$, the extremes of the α-cuts are determined as convex combinations of the original extremes using coefficients given by:

$$\frac{1}{m}(1 + \gamma_j) \qquad j \in I_l \text{ or } I_r \ , \tag{16}$$

where the quantities

$$\gamma_j = \frac{\sum_{f=1}^{m} \frac{1}{m+1-f} \pi_f^j}{\sum_{k=1}^{m} \sum_{f=1}^{m} \frac{1}{m+1-f} \pi_f^k} \tag{17}$$

represent the weighted normalized contribution of the $m-1$ most relevant extremes, specifically those in I_l for the left extremes and those in I_r for the right extremes.

It is worth to remark that the factor $\frac{1}{m+1-f}$ for each term in the numerator of γ_j is directly proportional to the number of overlaps, which represent the level of agreement. Additionally, in cases where all m α-cuts align, the values of γ_j can be consistently set to zero.

Furthermore, the coefficients associated with the m-th position, specifically those linked to the outermost left extreme in $\{1, \ldots, m\} \setminus I_l$ and the outermost right extreme in $\{1, \ldots, m\} \setminus I_r$, can be expressed as follows:

$$\frac{1}{m}(1 - \sum_{j=I_*} \gamma_j) \qquad I_* = I_l,\ I_r \text{ respectively.} \tag{18}$$

In situations where α exceeds the threshold h, the formal definition becomes more intricate as various subgroups of intersections involving two or more memberships need to be identified.

The logic underlying the method is to compute the S-and operator within each subgroup obtaining different intermediate α-levels, and merge them in a second step, by applying the S-or operator. For the technical details, we refer the reader once again to Capotorti and Figà-Talamanca (2020).

Furthermore, when utilizing the S-and operator, it is necessary that any changes in the number of memberships to be merged at a certain α-level result in the nesting of the resulting extremes within those of the lower levels. This nesting process should be achieved through translation or deformation, employing the same technique described for the S-or operator in equations (10-11).

In Figure 5, the results of the S-and and of the two S-or outputs for tweet #3420 from the Sanders dataset are plotted, as documented in Vashishtha and Susan (2019). The AFINN lexicon assigned a positive score of $x_1 = 3.0$ and a negative score of $x_2 = 0.0$ to the

tweet. By applying equations (2-4), we calculated a negative firing strength of $w_{neg} = 0.0$, a neutral firing strength of $w_{neu} = 0.625$, and a positive firing strength of $w_{pos} = 0.375$.

When processed through the VADER lexicon, the tweet received a positive score of $x_1 = 0.5$ and a negative score of $x_2 = 0.0$, resulting in a negative firing strength of $w_{neg} = 0.0$, a neutral firing strength of $w_{neu} = 0.0$, and a positive firing strength of $w_{pos} = 1.0$.

The membership value μ_{S-and} is compared with the original $\mu_{\vee-\max}$ values. It is important to note that the center of the area $C(A)$ is calculated to be 6.19, which differs from the arithmetic mean of 6.99 for the two original $C(A)$ values, which were 5.65 and 8.33, respectively.

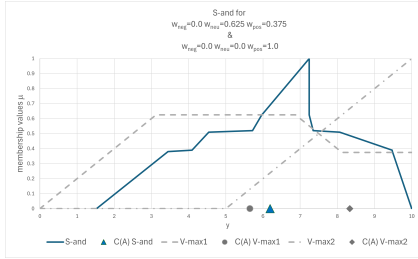


Figure 5: Comparison of the S-and membership and $C(A)$ w.r.t. two $\vee-\max$ of Vashishtha and Susan (2019) for tweet #3420 of Sanders dataset.

4 Conclusion

We have shown how the implementation of the recently introduced SMART Fuzzy aggregation operators, S-or and S-and, can improve the interpretability of a rule-based system proposed in Vashishtha and Susan (2019) for sentiment classification. Our innovative approach can be seamlessly integrated into any other Fuzzy rule-based system of Mandami type, regardless of the lexicon used.

Specifically, the S-or aggregator generates a truncated Fuzzy number as the output of the inference process, which is more precise and easier to interpret compared to the traditional V-max aggregation method. The S-or membership represents a Fuzzy number centered around its $C(A)$ value, with a weight determined by the maximum height of the membership.

On the other hand, the S-and operator enables the creation of an ensemble of sentiment classifiers utilizing various lexicons, resulting in a final crisp sentiment score that is a weighted average of scores obtained from individual lexicons. These weights implicitly capture the level of agreement or disagreement among the original Fuzzy outputs.

In conclusion, the SMART Fuzzy aggregation operators offer a versatile and effective solution for enhancing the interpretability and performance of rule-based sentiment classification systems. Further research will be devoted to apply these operators to classify the

sentiment of news articles retrieved from financial US newspaper and to finally assess the impact of the fuzzy-based sentiment score in the return and volatility dynamics of major US stocks.

As a distinguished target problem, we envisage the application of the present method to find associations between the sentiment of central banks (FED and ECB) communications (via policy statements, post-meeting press conferences, economic forecasts, monetary policy reports, speeches, interviews, and testimony to parliament) and the monetary policy stance, similarly to what has been done in Hilscher et al. (2024).

Acknowledgments:

We acknowledge the support of the PRIN 2022 project "Models for dynamic reasoning under partial knowledge to make interpretable decisions" (Project number: 2022AP3B3B, CUP Master: J53D23004340006, CUP: B53D23009860006) funded by the European Union – Next Generation EU.

References

J. M. Alonso Moral, C. Castiello, L. Magdalena, and C. Mencar. *An Overview of Fuzzy Systems*, pages 25–47. Springer International Publishing, Cham, 2021. ISBN 978-3-030-71098-9. doi: 10.1007/978-3-030-71098-9_2. URL https://doi.org/10.1007/978-3-030-71098-9_2.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 2200 —- 2204, 2010. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034054192&partnerID=40&md5=42db614bb847384aba6a349b4e0c98f9. Cited by: 2502.

A. Capotorti and G. Figà-Talamanca. Smart-or and smart-and fuzzy average operators: A generalized proposal. *Fuzzy Sets and Systems*, 395:1–20, 2020. ISSN 0165-0114. doi: https://doi.org/10.1016/j.fss.2019.04.027. URL https://www.sciencedirect.com/science/article/pii/S0165011419302349. Aggregation Operations.

S. Chauhan, J. P. Shet, S. M. Beram, V. Jagota, M. Dighriri, M. W. Ahmad, M. S. Hossain, and A. Rizwan. Rule based fuzzy computing approach on self-supervised sentiment polarity classification with word sense disambiguation in machine translation for hindi language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5), may 2023. ISSN 2375-4699. doi: 10.1145/3574130. URL https://doi.org/10.1145/3574130.

J. Hilscher, K. Nabors, and A. Raviv. Information in central bank sentiment: An analysis of fed and ecb communication. *Available at SSRN*, 2024. URL http://dx.doi.org/10.2139/ssrn.4797935.

C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 01 2015.

H. Liu and M. Cocea. Fuzzy rule based systems for interpretable sentiment analysis. In *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, pages 129–136, 2017. doi: 10.1109/ICACI.2017.7974497.

E. Mamdani. Application of fuzzy algorithms for control of simple dynamic plant. *Proceedings of the Institution of Electrical Engineers*, 121:1585–1588(3), December 1974. ISSN 0020-3270. URL https://digital-library.theiet.org/content/journals/10.1049/piee.1974.0328.

E. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1):1–13, 1975. ISSN 0020-7373. doi: https://doi.org/10.1016/S0020-7373(75)80002-2. URL https://www.sciencedirect.com/science/article/pii/S0020737375800022.

S. Nadali, M. A. A. Murad, and R. A. Kadir. Sentiment classification of customer reviews based on fuzzy logic. In *2010 International Symposium on Information Technology*, volume 2, pages 1037–1044, 2010. doi: 10.1109/ITSIM.2010.5561583.

M. Naimi and H. Tahayori. Centroid of polygonal fuzzy sets. *Applied Soft Computing*, 95: 106519, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106519. URL https://www.sciencedirect.com/science/article/pii/S1568494620304580.

F. Å. Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011. URL http://ceur-ws.org/Vol-718/paper_16.pdf.

A. H. Shapiro and D. W. Moritz Sudhof. Measuring news sentiment. Working Paper 2017-01, Federal Reserve Bank of San Francisco, 2020. URL https://doi.org/10.24148/wp2017-01.

S. Vashishtha and S. Susan. Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138:112834, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.112834. URL https://www.sciencedirect.com/science/article/pii/S0957417419305366.

Conditional Independence Constraints in Score-Based Learning of Bayesian Networks

James Cussens[1]

[1]School of Computer Science, University of Bristol
james.cussens@bristol.ac.uk

Abstract

A method for adding conditional independence (CI) constraints to score-based Bayesian Network Structure Learning is presented. As well as showing how to check that a DAG meets a CI constraint, we consider methods for propagation, conflict analysis and separation. Throughout, forward-chaining in propositional logic is the main algorithm. Some preliminary experiments are presented which indicate that the chosen propagation method is slow, but that implementing separation is beneficial.

1 Introduction

There are two main approaches to *Bayesian Network Structure Learning (BNSL)*, i.e. learning the structure of a Bayesian network from data. In the *constraint-based* approach a series of carefully chosen conditional independence tests are performed on the data and a DAG that satisfies the results of those tests is constructed. In the *score-based* approach each candidate DAG has a score (determined by the data). This is typically a penalised likelihood, such as BIC, or some Bayesian measure such as marginal likelihood or posterior probability. The job of a score-based algorithm for BNSL is simply to find a DAG with maximal score.

In this paper we address the issue of incorporating conditional independence (CI) constraints into a score-based approach. Where these CI constraints come from is not our focus: they might either come from a domain expert who knows that certain CI relations hold or they might be inferred from data, leading in effect, to a hybrid constraint-based/score-based approach to BNSL.

Throughout we focus on a particular score-based algorithm: GOBNILP (Cussens, 2011)—which encodes BNSL as an integer program and uses a cutting plane approach to ensure acyclicity of the graph. GOBNILP has two implementations: one in Python and one in C. Here we focus on the C implementation which uses the SCIP (Solving Constraint

Integer Programs) library (Bolusani et al., 2024). SCIP has a 'plug-in' architecture which allows a user to write *constraint handlers* for particular sorts of constraints. In this paper the design of a constraint handler for CI constraints is described and evaluated.

2 Implementing conditional independence constraints

2.1 Constraint representation

GOBNILP assumes that the score for a given DAG is a sum of *local scores*, one for each BN variable. The local score is determined by the choice of parent set for each BN variable. As a result GOBNILP creates binary integer program (IP) variables called *family variables*. The family variable $x_{i \leftarrow J}$ takes the value 1 if J is the parent set for BN variable i and 0 otherwise.

GOBNILP imposes constraints to ensure that any assignment of values to all family variables represents a DAG. One could effect CI constraints on family variables but we choose not to do that. One problem with family variables is that there are exponentially-many of them (so that for big problems GOBNILP either artificially fixes most of them to zero or attempts to add them to the IP during the course of solving—a method known as *pricing*). To evade this problem we define CI constraints on *arrow variables*, where the arrow variable $x_{i \leftarrow j}$ takes the value 1 if the DAG has an arrow from j to i and 0 otherwise. Using p to denote the number of BN variables (= vertices in the DAG) there are only $p(p-1)$ arrow variables. Arrow and family variables are connected by the following linear equation (where P denotes the set of DAG vertices):

$$x_{i \leftarrow j} = \sum_{J \subseteq P \setminus \{i\}, j \in J} x_{i \leftarrow J} \tag{1}$$

2.2 Constraint checking

The very minimum that any SCIP constraint handler must do is to decide whether an assignment of values to a constraint's variables satisfy the constraint or not. Given a DAG $\mathcal{G}$ (defined by arrow variables) and a particular CI constraint $A \perp B | S$ we can say that the constraint is satisfied if A and B are separated by S in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$, the moral graph of the DAG restricted to the smallest ancestral set containing $A \cup B \cup S$. This test, of course, gives the same result as testing for d-separation in $\mathcal{G}$ (Lauritzen, 1996, Proposition 3.25).

To determine whether A and B are separated by S in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$ the set of propositional logic definite clauses shown in Fig 1 are constructed and it is then determined whether the proposition λ can be deduced. (We will call these definite clauses *rules*.) The standard forward-chaining algorithm for propositional logic is used; see Russell and Norvig (2010, Fig 7.15) for more information on this simple algorithm. In Fig 1 α_i indicates that i is in $\mathrm{An}(A \cup B \cup S)$, $y_{i \leftarrow j}$ indicates that there is an arrow from j to i in $\mathcal{G}_{\mathrm{An}(A \cup B \cup S)}$, z_{i-j} indicates that i and j are connected (not necessarily directly) in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$ and λ indicates that A and B are **not** separated by S in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$.

$$x_{i \leftarrow j} \qquad i \leftarrow j \in \mathcal{G} \tag{2}$$

$$\alpha_i \qquad i \in (A \cup B \cup S) \tag{3}$$

$$\alpha_i \wedge x_{i \leftarrow j} \rightarrow \alpha_j \tag{4}$$

$$\alpha_i \wedge x_{i \leftarrow j} \rightarrow y_{i \leftarrow j} \tag{5}$$

$$\alpha_i \wedge x_{i \leftarrow j} \rightarrow z_{i-j} \qquad \{i, j\} \cap S = \emptyset \tag{6}$$

$$y_{i \leftarrow j} \wedge y_{i \leftarrow k} \rightarrow z_{j-k} \qquad \{j, k\} \cap S = \emptyset \tag{7}$$

$$z_{i-j} \wedge z_{i-k} \rightarrow z_{j-k} \qquad \{i, j, k\} \cap S = \emptyset \tag{8}$$

$$z_{i-j} \rightarrow \lambda \qquad i \in A, j \in B \tag{9}$$

Figure 1: Rules for checking a CI constraint

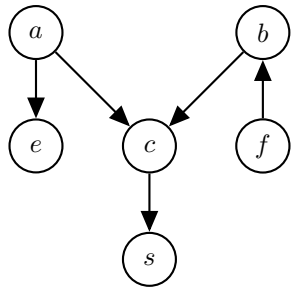


Figure 2: $\mathcal{G}_1$: A graph which violates the constraint $\{a\} \perp \{b\} | \{s\}$

To see how the constraint checking process works, consider the DAG $\mathcal{G}_1$ in Fig 2 which violates the constraint $\{a\} \perp \{b\} | \{s\}$. The proof of λ in this case is illustrated by the proof graph in Fig 3. The proof graph is constructed by adding a vertex for each fact newly derived by forward-chaining and adding arrows to it from any earlier facts which allowed its inference.

2.3 Constraint propagation

GOBNILP uses SCIP to perform a *branch-and-bound* search for an optimal DAG. When (as in GOBNILP) we have only binary IP variables, branching is the process of choosing a variable and creating two sub-problems: one where that variable has the value 0 and one where it has the value 1. Successive branching leads to sub-problems where a number of variables have fixed values: a partial assignment. If we can determine that a given constraint is violated by a partial assignment then the branch in question can be cut off since it cannot lead to a feasible solution (let alone an optimal one). In addition, constraints can be used to fix as yet unfixed variables. For example if we have a CI constraint that i and j must be independent and we have a sub-problem where $x_{i \leftarrow k} = 1$

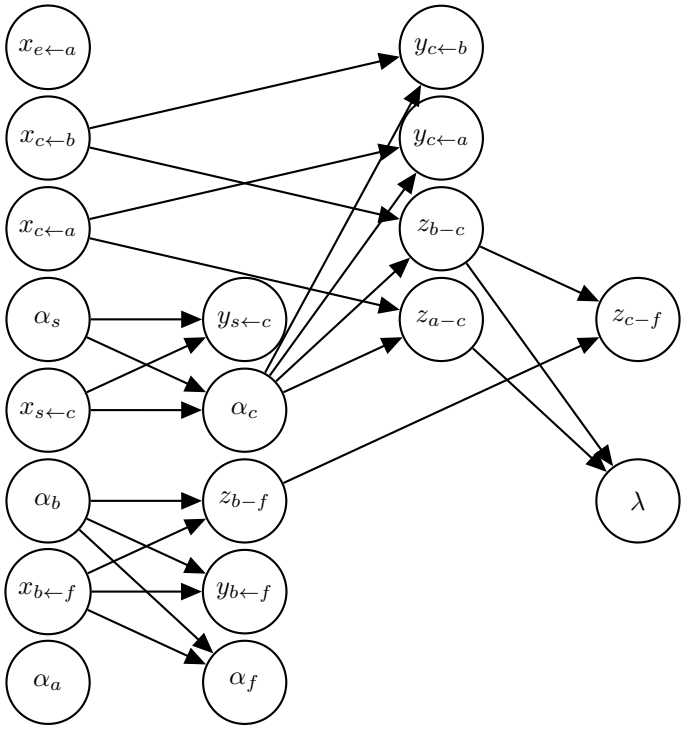


Figure 3: Illustration of the proof that $\mathcal{G}_1$ in Fig 2 violates the constraint $\{a\}\perp\{b\}|\{s\}$

then we can fix $x_{k\leftarrow j}$ to 0. This is called *propagation*.

In GOBNILP's CI constraint handler, propagation is effected as follows. The rules in Fig 1 are constructed with the modification that only arrows fixed to 1 lead to the corresponding proposition $x_{i\leftarrow j}$ being included in (2). If λ can be deduced we have established *infeasibility* and simply cut off the current branch. If not, we see whether propagation is possible.

Our desire to perform propagation is why forward-chaining inference (as opposed to, say, resolution) is used. Once forward-chaining has terminated without inferring λ, then for each unfixed arrow variable $x_{i\leftarrow j}$ we add $x_{i\leftarrow j}$ as a new fact and see whether, with this addition, λ can be deduced, again using forward-chaining. If so, we fix $x_{i\leftarrow j}$ to 0. This approach to propagation where we see whether a particular fixing leads to infeasibility, is called *probing*. In general probing can be expensive. However, since we construct and keep hold of the set of facts deducible prior to any fixings we typically have few additional iterations of forward-chaining to do when probing. On the other hand since *each* unfixed arrow variable is considered many probings are done which is expensive.

2.4 Conflict analysis

Cutting off infeasible branches of the branch-and-bound tree is important for efficient solving. However, if we can also supply the 'reason' for the infeasibility to the solver (in this case SCIP) this can also improve performance. In the case of a CI constraint, the reason for infeasibility is the existence of a set of arrow variables which entail λ using the rules in Fig 1. If the proof graph is constructed then the set of variables leading to infeasibility is readily recovered: it is just the set of $x_{i\leftarrow j}$ nodes which are ancestors of λ. So, for example, in Fig 3, the set of ancestors of λ are $\{x_{s\leftarrow c}, x_{c\leftarrow b}, x_{c\leftarrow a}\}$.

Once an infeasibility-entailing set of variables has been identified the solver has the option of adding an associated *conflict clause* to the problem, which in our example would be the constraint:

$$\neg x_{s\leftarrow c} \vee \neg x_{c\leftarrow b} \vee \neg x_{c\leftarrow a} \tag{10}$$

To see the benefit of such clauses imagine that the 5 arrow variables in the first layer of the proof graph (Fig 3) had got set to 1 as a result of this sequence of branching decisions: $(x_{b\leftarrow f} = 1, x_{e\leftarrow a} = 1, x_{s\leftarrow c} = 1, x_{c\leftarrow b} = 1, x_{c\leftarrow a} = 1)$. So we have reached a depth 5 node in the branch-and-bound tree that we now know to be infeasible due to the $\{a\}\perp\{b\}|\{s\}$ constraint. It follows that

$$\neg x_{b\leftarrow f} \vee \neg x_{e\leftarrow a} \vee \neg x_{s\leftarrow c} \vee \neg x_{c\leftarrow b} \vee \neg x_{c\leftarrow a} \tag{11}$$

is a valid clause. However adding such a clause to the problem is pointless since the search will never revisit this depth 5 node we know already to be infeasible. Clause (10) however is useful. If we add it then any node reached by a sequence of branching decisions of the form $(\ldots, x_{s\leftarrow c} = 1, \ldots, x_{c\leftarrow b} = 1, \ldots x_{c\leftarrow a} = 1)$ (in fact, any sequence fixing these 3 arrow variables to 1 in any order) can immediately be flagged as infeasible.

Also if some of the fixings of the variables $x_{s\leftarrow c}$, $x_{c\leftarrow b}$ and $x_{c\leftarrow a}$ to 1 were the result of propagations using other constraints, then we might be able to infer additional conflict

clauses. This is done by essentially extending the proof graph Fig 3 'leftwards' to represent these propagations. In practice, the work of doing all this is left to SCIP. We just inform SCIP of the (minimal) sets of arrow variables whose locally fixed values led to infeasibility and leave it to create what it judges are useful conflict clauses.

2.5 Separation

Since we are always dealing with finite DAGs it follows that any constraint on DAG structure can be expressed as a finite set of linear constraints. In this section we derive the linear representation of CI constraints. Returning to the running example we have that any DAG containing the arrows $\{x_{s\leftarrow c}, x_{c\leftarrow b}, x_{c\leftarrow a}\}$ will violate the constraint $\{a\}\perp\{b\}|\{s\}$. It follows that the linear constraint:

$$x_{s\leftarrow c} + x_{c\leftarrow b} + x_{c\leftarrow a} \leq 2 \tag{12}$$

is implied by the constraint $\{a\}\perp\{b\}|\{s\}$.

More generally, it is useful to consider *active chains* used in the definition of d-separation. An active chain for a CI constraint $A\perp B|S$ is a chain connecting an element of A to an element of B which is not blocked by S. For each active chain there is a minimal set of arrows which establish that it is an active chain. This set will include all the arrows in the chain together with a (possibly empty) set of arrows which establish that any colliders in the chain which are not in S have an element of S as a descendant. In our running example, the active chain is (a, c, b), where c is a collider with s as a descendant.

Let Active($A\perp B|S$) be the set of all minimal sets of arrows which entail an active chain, or equivalently which allow λ to be deduced. It follows that the constraint $A\perp B|S$ is equivalent to the following set of linear constraints.

$$\sum_{x_{i\leftarrow j}\in\mathcal{C}} x_{i\leftarrow j} \leq |\mathcal{C}| - 1 \qquad \mathcal{C} \in \text{Active}(A\perp B|S) \tag{13}$$

The linear inequalities in (13) can be used to perform *separation*. As part of the process of solving an integer program (IP) an IP solver will solve the *linear relaxation* of the problem which is a linear program (LP) where the integrality constraint on integer variables is removed. Solving this LP provides a useful bound and is typically quick to do. The solution to the LP typically contains variables with fractional values. In the case of GOBNILP's binary arrow variables we will typically get values in the interval $[0, 1]$ other than 0 or 1.

If we can find linear inequalities from (13) that are violated by the solution to the linear relaxation, then we can add them as *cutting planes* (which separate the linear relaxation solution from the set of feasible solutions) and resolve the (now more constrained) linear relaxation. Adding linear inequalities from CI constraints at this point in the solving process will increase efficiency.

So, given an assignment of (possibly fractional) values to the arrow variables $x_{i\leftarrow j}$ how can we find a violated linear inequality $\sum_{x_{i\leftarrow j}\in\mathcal{C}} x_{i\leftarrow j} \leq |\mathcal{C}| - 1$ for some $\mathcal{C} \in$ Active($A\perp B|S$)? We do this by adapting the forward-chaining algorithm to infer positive

lower bounds associated with facts. Initially, each α_i for $i \in A \cup B \cup S$ is given a lower bound $\ell(\alpha_i)$ of 1 and each $x_{i \leftarrow j}$ with a positive value $x^*_{i \leftarrow j} > 0$ in the solution to the linear relaxation is given $\ell(x_{i \leftarrow j}) = x^*_{i \leftarrow j}$ as its lower bound. All other facts h have an associated lower bound of $\ell(h) = 0$. Forward-chaining uses the rules in Fig 1 to increase lower bounds as follows. If $b_1 \rightarrow h$ is a rule then $\ell(h)$ is updated to $\ell(b_1)$ if this increases $\ell(h)$. If $b_1 \wedge b_2 \rightarrow h$ is a rule then $\ell(h)$ is updated to $\ell(b_1) + \ell(b_2) - 1$ if this increases $\ell(h)$. This update rule is an example of a Boole-Fréchet inequality (Boole, 1854) and the lower bounds can be viewed as lower bounds on the probability of the facts, although this interpretation is not exploited here.

It is easy to adapt forward-chaining to effect these updates and also to record a proof graph. Indeed if all initial positive lower bounds were 1, then the procedure is exactly the same as normal forward-chaining. Only one alteration is required to the construction of the proof graph: if h is a vertex already in the proof graph and its lower bound gets increased using a rule $b_1 \rightarrow h$ (resp. $b_1 \wedge b_2 \rightarrow h$) then any arrows to it in the proof graph are removed and one from b_1 is added (resp. one from both b_1 and b_2 are added).

If this procedure produces a lower bound for λ which is positive then the $x_{i \leftarrow j}$ ancestors of λ in the proof graph are found. Call the set of such ancestors $\mathcal{C}$. We will show below that if λ has a positive lower bound then $\sum_{x_{i \leftarrow j} \in \mathcal{C}} x^*_{i \leftarrow j} > |\mathcal{C}| - 1$ so the cut $\sum_{x_{i \leftarrow j} \in \mathcal{C}} x_{i \leftarrow j} \leq |\mathcal{C}| - 1$ is added. It is easy to see that this is a valid cut since the proof graph constructed during the updates of lower bounds also serves as a proof that $\mathcal{C} \in \text{Active}(A \perp B | S)$. Fig 4 shows a proof that $\lambda \geq 0.2$ from $x_{c \leftarrow a} \geq 0.6$, $x_{c \leftarrow b} \geq 0.8$ and $x_{s \leftarrow c} \geq 0.9$. (Fig 4 only shows the lower bounds needed for this proof to avoid clutter.) Note that $0.6+0.8+0.9 = 2.3 > 2$ so $x_{c \leftarrow a} + x_{c \leftarrow b} + x_{s \leftarrow c} \leq 2$ separates any linear relaxation solution where $x^*_{c \leftarrow a} = 0.6$, $x^*_{c \leftarrow b} = 0.8$ and $x^*_{s \leftarrow c} = 0.9$. Note also that, although our cut-finding algorithm is correct, it is not complete. If we had $x^*_{s \leftarrow c} = 0.7$ rather than $x^*_{s \leftarrow c} = 0.9$ then forward-chaining would infer $\alpha_c \geq 0.7$, $z_{b-c} \geq 0.5$ and $z_{a-c} \geq 0.3$ and no positive lower bound for λ could be inferred, even though $x_{c \leftarrow a} + x_{c \leftarrow b} + x_{s \leftarrow c} \leq 2$ is still a cut since $0.6 + 0.8 + 0.7 = 2.1 > 2$.

If a positive lower bound $\ell(\lambda)$ is inferred for λ then we have a set $H = \{\lambda\}$ such that $\sum_{h \in H} \ell(h) > |H| - 1$. The following theorem shows that for any set H obeying this strict inequality if we replace a fact in H by its parent(s) to get a new set H' then we also have $\sum_{h \in H'} \ell(h) > |H'| - 1$. If we start with $H = \{\lambda\}$ and repeatedly apply this result until we have arrive at a set H' where no fact has parents then we $\sum_{h \in H'} \ell(h) > |H'| - 1$. H' will contain α_i facts and $x_{i \leftarrow j}$ facts. All α_i have $\ell(\alpha_i) = 1$ so if we remove all of them form H' we will still have $\sum_{h \in H'} \ell(h) > |H'| - 1$ and so the remaining $x_{i \leftarrow j}$ facts provide a cut.

Theorem 1. *Let H be a set of vertices in a proof graph for lower bounds produced using forward-chaining such that $\sum_{h \in H} \ell(h) > |H| - 1$ where $\ell(h)$ is the lower bound for fact h. Let $h' \in H$ have a non-empty parent set $\text{Pa}(h')$ in the proof graph and let $H' = H \setminus \{h'\} \cup \text{Pa}(h')$. Then $\sum_{h \in H'} \ell(h) > |H'| - 1$.*

Proof. We simply check the 5 possible situations. Suppose $\text{Pa}(h') = \{b_1\}$ so $\ell(b_1) = \ell(h')$. If (i) $b_1 \in H$ then $\sum_{h \in H'} \ell(h) = \left[\sum_{h \in H} \ell(h)\right] - \ell(h') > |H| - 1 - \ell(h') \geq |H| - 2 = |H'| - 1$. If (ii) $b_1 \notin H$ then $\sum_{h \in H'} \ell(h) = \sum_{h \in H} \ell(h) > |H| - 1 = |H'| - 1$. Now suppose

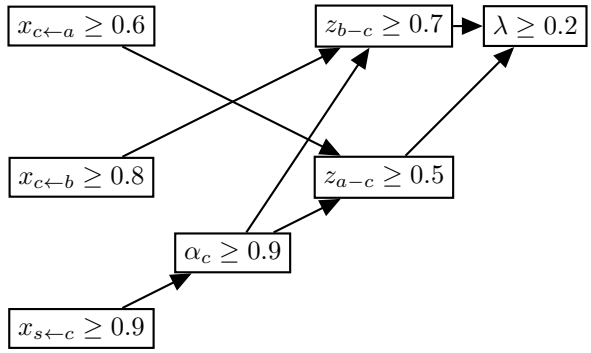


Figure 4: Simplified proof tree for $\lambda \geq 0.2$

$\mathrm{Pa}(h') = \{b_1, b_2\}$ so $\ell(b_1) + \ell(b_2) - 1 = \ell(h')$. If (iii) $\{b_1, b_2\} \subseteq H$ then $\sum_{h\in H'} \ell(h) = \left[\sum_{h\in H} \ell(h)\right] - \ell(h') > |H| - 1 - \ell(h') \geq |H| - 2 = |H'| - 1$. If (iv) $b_1 \in H, b_2 \notin H$, then $\sum_{h\in H'} \ell(h) = \left[\sum_{h\in H} \ell(h)\right] - \ell(h') + \ell(b_2) > |H| - 1 + 1 - \ell(b_1) \geq |H'| - 1$. Similarly for $b_1 \notin H, b_2 \in H$. If (v) $b_1 \notin H, b_2 \notin H$ then $\sum_{h\in H'} \ell(h) = \left[\sum_{h\in H} \ell(h)\right] - \ell(h') + \ell(b_1) + \ell(b_2) > |H| - 1 + 1 = |H'| - 1$. □

2.6 Presolving

CI constraints can play a part in *presolving*, a process which simplifies the problem before solving proper begins. Most obviously, given a CI constraint $A\perp B|S$, we can fix any $x_{i\leftarrow j}$ to 0 in presolving if $i \in A$, $j \in B$ or vice-versa. Also if $i \in A$, $j \in B$, $k \notin S$ then we can add the following *set packing* constraints during presolving:

1. $x_{i\leftarrow k} + x_{k\leftarrow i} + x_{j\leftarrow k} \leq 1$
2. $x_{j\leftarrow k} + x_{k\leftarrow j} + x_{i\leftarrow k} \leq 1$

More complicated presolving is also possible. If we have not only arrow variables $x_{i\leftarrow j}$ but also (binary) variables representing ancestor relations then some of those can be fixed to 0 in presolving: if i is required to be independent of j then neither can be the ancestor of the other and they may not have a common ancestor. In addition, although not currently implemented, one could use CI constraints to prevent some family variables from being created (rather than being pointlessly created and then fixed to 0 in presolving). For example, if we have a CI constraint $A\perp B|S$ then no parent set for any $s \in S$ can intersect with both A and B.

3 Preliminary Experiments

The CI constraint handling method described in Section 2 has been implemented as SCIP constraint handler and integrated into the GOBNILP algorithm. To check the imple-

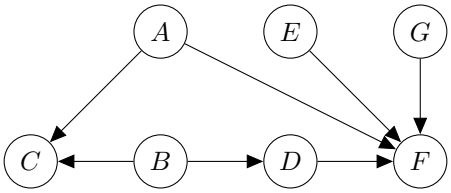


Figure 5: 'True' DAG used to simulated data.

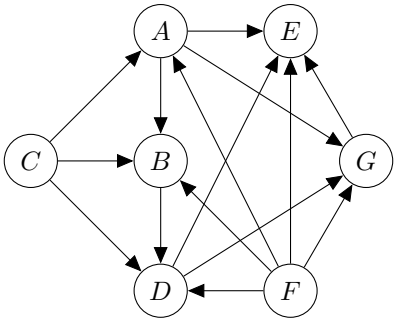


Figure 6: Optimal learned DAG with $C \perp F$ constraint. BIC score is -12009. Takes 1.75 seconds to learn

mentation and assess the effectiveness of propagating and separating some preliminary experiments have been conducted.

5000 datapoints generated from the 7 node 7 arc Gaussian network shown in Fig 5 (from the `bnlearn` (Scutari, 2010) R package) were provided as input to GOBNILP and the globally optimal DAG for this data, according to the BIC score, was found in under 1 second. This learned DAG was, unsurprisingly, Markov equivalent to the true DAG in Fig 5 and had BIC score -6997.753.

Figs 6 and 7 show the optimal DAGs when the constraints $C \perp F$ and $C \perp F|D$, respectively, are added. For this very small example at least, dealing with a CI constraint did not slow down learning much: it took 1.75 seconds and 1.69 seconds, respectively. It is interesting that in both cases to ensure that (i) the constraint and (ii) the dependencies suggested by the data are respected, quite a dense network is required.

To examine CI constraint handling for a bigger problem, a dataset with 100 datapoints on 20 continuous variables $X_0 \dots X_{19}$ was used for learning. With a parent set cardinality upper bound of 4 the optimal DAG, according to the BIC score, was learned in 19 seconds seconds. With a single CI constraint $X_7 \perp X_{19}$ an optimal DAG was learned in 44 seconds and with a second CI constraint $X_{11} \perp X_{14}|X_9$ an optimal DAG was learned in 59 seconds. However, if constraint propagation was turned off finding an optimal DAG took only 24 and 33 seconds, respectively. This indicates that the *probing* approach to propagation is too slow and so an alternative faster and perhaps less complete approach to propagation is worth exploring. When in addition to propagation being turned off, separation was

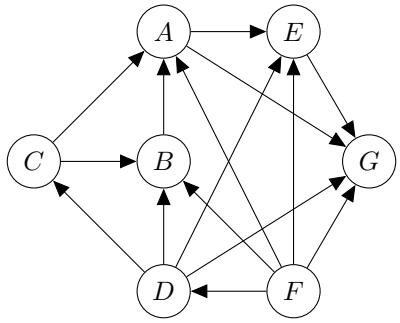


Figure 7: Optimal learned DAG with $C \perp F | D$ constraint. BIC score is -8187. Takes 1.69 seconds to learn

also turned off, finding an optimal DAG took 25 and 60 seconds, so it appears that implementing separation is beneficial.

References

S. Bolusani, M. Besançon, K. Bestuzheva, A. Chmiela, J. Dionísio, T. Donkiewicz, J. van Doornmalen, L. Eifler, M. Ghannam, A. Gleixner, C. Graczyk, K. Halbig, I. Hedtke, A. Hoen, C. Hojny, R. van der Hulst, D. Kamp, T. Koch, K. Kofler, J. Lentz, J. Manns, G. Mexi, E. Mühmer, M. E. Pfetsch, F. Schlösser, F. Serrano, Y. Shinano, M. Turner, S. Vigerske, D. Weninger, and L. Xu. The SCIP Optimization Suite 9.0. Technical report, Optimization Online, February 2024. URL `https://optimization-online.org/2024/02/the-scip-optimization-suite-9-0/`.

G. Boole. *An Investigation of the Laws of Thought, On Which Are Founded the Mathematical Theories of Logic and Probability.* Walton and Maberly, London, 1854.

J. Cussens. Bayesian network learning with cutting planes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 153–160. AUAI Press, 2011.

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

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Pearson, 3rd edition, 2010.

M. Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

How Sir Harold Jeffreys would create a belief function based on data

Milan Daniel[1], Radim Jiroušek[2], and Václav Kratochvíl[2]

[1]Institute of Computer Sciences, Czech Academy of Sciences
[1]*milan.daniel@cs.cas.cz*
[2]Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic
[2] {*radim;velorex*}*@utia.cas.cz*

Abstract

Not all normalized nonnegative monotone set functions are belief functions. This paper investigates ways to modify them to obtain a belief function that preserves some of their properties. The problem is motivated by an approach to data-based learning of belief function models. The approach is based on the idea that classical methods of mathematical statistics can provide estimates of lower bounds for unknown probabilities. Thus, methods of mathematical statistics can be used to obtain a reasonable rough estimate, which is further elaborated to obtain a desired belief function model.

1 Introduction

When learning a probabilistic model from data, you need to determine a large number of parameters, i.e., find estimates for many probabilities. But you are not completely sure about any of them especially if you have a limited amount of data. Using Bayesian statistics, you get a posterior distribution of the considered probability (usually suggesting to accept the most probable value), and if you want to be careful when applying the frequentist approach, you should consider some confidence intervals of the needed parameters. Thus, when learning models from data, it may seem more natural to do so in the framework of belief function theory (Shafer, 1976) than in the framework of probability theory. However, even this approach is not trivial. Although it has some similar properties, the function whose values are estimates of the lower bounds of the confidence intervals for each of the probabilities considered does not form a belief function. This happens only in small examples. Otherwise, we get more general monotonic set functions (usually called capacities), which need to be processed in order to be transformed into belief functions. And the possible ways to do this are the subject of this paper.

Thus, in this paper we explore the possibilities of searching for a belief function that would be obtained as a modification of the values of the statistical estimates of confidence intervals. After introducing the necessary terminology and notation, we begin to study the properties of functions whose values correspond to the statistical estimates of the lower bounds of confidence intervals. In this paper, we study only two types of their approximation by belief functions. In Section 3 we propose a procedure for finding approximations that preserve all the information extracted from the data by the statistical estimates, and in Sections 4 and 5 we study the lower approximations that do not add to the model any information that is not encoded in the considered statistical estimates.

2 Necessary notions from theory of belief functions

In the whole paper, Ω denotes a finite *frame of discernment*. A non-negative set function $f : 2^\Omega \to \mathbb{R}^+$ is called *pseudo-belief function* (PBF) if $f(\emptyset) = 0$, and it is

monotonic for $a \subset b \subseteq \Omega : f(a) \leq f(b)$, and

normalized $f(\Omega) = 1$.

Each PBF f is connected with a set of probability distributions defined on Ω. A *credal set* of PBF f is the following set of probability distributions.

$$\mathcal{P}(f) = \left\{\pi \text{ defined on } \Omega : \left(\forall a \subseteq \Omega : \pi(a) \geq f(a)\right)\right\}. \tag{1}$$

There is a natural partial ordering for PBFs. $f \leq g$ means that $f(a) \leq g(a)$ for all $a \subseteq \Omega$. $f < g$ denote that $f \leq g$ and $f \neq g$. Notice also that $f \leq g$ is equivalent with $\mathcal{P}(f) \supseteq \mathcal{P}(g)$.

We say a function $f : 2^\Omega \to \mathbb{R}^+$ is a *belief function* (BF) if it is a PBF and for all non-empty $a \subseteq \Omega$

$$\sum_{b \subseteq a} -1^{|a \setminus b|} f(b) \geq 0.$$

The set of all belief functions defined on given Ω is denoted by $\mathcal{BF}$.

Thus, for each BF f one can define non-negative function m_f on 2^Ω called *basic probability assignment* (BPA) m_f by the following expression

$$m_f(a) = \sum_{b \subseteq a} -1^{|a \setminus b|} f(b), \tag{2}$$

for which

$$f(a) = \sum_{b \subseteq a} m_f(b). \tag{3}$$

The sets $a \subseteq \Omega$ for which $m_f(a) > 0$ are called *focal elements* of f. Notice that if all focal elements of BF f are singletons ($|a| = 1$), then also $\mathcal{P}$ is a singleton. These BFs are called *Bayesian*. The representation of BF f using its BPA m_f is often preferred. It makes

the introduction of some notions, such as the following notion of pignistic transform and simple specification, more intuitive.

Pignistic transform of BF f is a specific element of its credal set $\mathcal{P}(f)$ (Dubois and Prade, 1982). It is a probability distribution defined for all $\omega \in \Omega$

$$\pi_f(\omega) = \sum_{a \subseteq \Omega : \omega \in a} \frac{m_f(a)}{|a|}.$$

Using Equation (3), it is trivial to show that $\pi_f(a) \geq f(a)$. It means that $\mathcal{P}(f) \neq \emptyset$ for all BFs f. Note that it does not hold for all PBFs. As a trivial example consider $\Omega = \{\omega_1, \omega_2\}$, and PBF g defined $g(\{\omega_1\} = g(\{\omega_2\} = 0.6$, $g(\Omega) = 1$, which obviously complies with the definition of PBFs. As a little bit more sophisticated example consider an arbitrary Ω, and PBF g defined

$$g(a) = \begin{cases} 0 & \text{if } |a| < |\Omega| - 1, \\ 1 - \frac{1}{|\Omega|+1} & \text{if } |a| = |\Omega| - 1, \\ 1 & \text{if } a = \Omega. \end{cases}$$

It means that for $\pi \in \mathcal{P}(g)$, $\pi(\omega) \leq \frac{1}{|\Omega|+1}$ for each $\omega \in \Omega$, which cannot hold for any probability distribution π.

The following trivial assertion holds.

Lemma 1 *Let g be a PBF on Ω. $\mathcal{P}(g) \neq \emptyset$ if and only if there exists BF f, for which $f \geq g$.*

Proof. If $\mathcal{P}(g) \neq \emptyset$ denote $\pi \in \mathcal{P}(g)$, and define Bayesian BF f through its BPA $m_f(\{\omega\}) = \pi(\omega)$. Since $\pi(a) \geq g(a)$ for all $a \subseteq \Omega$, the also $f \geq g$.

The opposite part of the equivalence is even simpler. The credal set $\mathcal{P}(f)$ of BF f is always nonempty, and therefore $\mathcal{P}(g) \supseteq \mathcal{P}(f)$ must also be nonempty. □

In (Jiroušek and Kratochvíl, 2025), the following notion was defined for BFs. We say that BF f is a *simple specification* of BF g if m_f is created from m_g by shifting some of its mass from some focal element to its subset; more precisely, there exist subsets $a \subset b \subseteq \Omega$ such that $m_f(a) = m_g(a) + \varepsilon$, and $m_f(b) = m_g(b) - \varepsilon$, and all the remaining focal elements of m_f are the copies of the focal elements of m_g, i.e., for all $c \in \Omega \setminus \{a, b\}$, $m_f(c) = m_g(c)$. Thus, this operation means that $f(c) = g(c) + \varepsilon$ for all $c \subset \Omega$, for which $(a \subseteq c) \setminus (b \subseteq c)$, and for all remaining c, $f(c) = g(c)$. So we see that $f > g$. In this paper we will apply this notion to all PBFs with the same effect. We will also generalize this notion in the sense that we will consider $\varepsilon < 0$. The reader immediately sees that for negative values of ε, $f < g$, and thus we will call this modification *simple generalization.*

Recall that in (Jiroušek and Kratochvíl, 2025) we proved the following assertion stating that if $f > g$, then f can be obtained from g by a sequence of simple specifications.

Lemma 2 *Let BFs f and g are defined on Ω. If $g < f$, then there exists a finite sequence of BFs $g = h_1, h_2, \ldots, h_k = f$ such that each h_{i+1} is a simple specification of h_i.*

In the computational procedures introduced below we will use *mass redistribution*, which consists of several simple specifications (generalizations) performed simultaneously. By this new term we understand the process, when masses assigned to several subsets are changed. We change PBF g to PBF f by *redistributing* ε from $b \subseteq \Omega$ to r its proper subsets $a_1, \ldots, a_r$ if

(i) for all $\ell = 1, \ldots, r$, $a_\ell \subsetneq b$;

(ii) $\varepsilon_1, \ldots, \varepsilon_r$ are such that $\sum_{\ell=1}^{r} \varepsilon_\ell = \varepsilon$, and for all ℓ, $\frac{\varepsilon_\ell}{\varepsilon} > 0$;

(iii) $m_f(b) = m_g(b) - \varepsilon$;

(iv) for all $\ell = 1, \ldots, r$, $m_f(a_\ell) = m_g(a_\ell) + \varepsilon_\ell$;

(v) for all the remaining $c \subseteq \Omega \setminus \{b, a_1, \ldots, a_r\}$, $m_f(c) = m_g(c)$.

Notice that if $\varepsilon > 0$, f is a specification of g (i.e., $f > g$), if $\varepsilon < 0$, f is a generalization of g (i.e., $f < g$) because the condition (ii) guarantees that all ε_ℓ are of the same sign; they are all positive or negative.

Each PBF g splits the whole set of BFs $\mathcal{BF}$ into three disjoint parts: *inner* (*upper*) *envelop* of g

$$\overline{\mathcal{B}}(g) = \{f \in \mathcal{BF} : f \geq g\},$$

outer (*lower*) *envelop* of g

$$\underline{\mathcal{B}}(g) = \{f \in \mathcal{BF} : f \leq g\},$$

and the set of BFs, which are *incomparable* with g, i.e., $\mathcal{BF} \setminus \left(\underline{\mathcal{B}}(g) \cup \overline{\mathcal{B}}(g)\right)$.

3 Upper Approximations of Pseudo-Belief Functions

Consider a general PBF $g : 2^\Omega \to [0, 1]$. Let us explore ways to find a suitable approximation of g with some BF. There are several possible ways to do this. In this paper, we will only consider approximations by BFs either from $\underline{\mathcal{B}}(g)$ or $\overline{\mathcal{B}}(g)$. Approximations from $\underline{\mathcal{B}}(g)$ are supported by the fact that they do not add any information that is not contained in g. But the only task we can solve optimally is to look for the solution in $\overline{\mathcal{B}}(g)$. In this case the optimal solution is obtained by the following *Upper approximation procedure* presented here in the form that produces both the BF f and the corresponding BPA m_f (it is a trivial application of Formulas (2) and (3) to show that the procedure produces a consistent pair of functions).

Upper Approximation Procedure
For $k = 1, ..., |\Omega|$
For $a \subseteq \Omega : |a| = k$
$f(a) := \max\left[g(a),\ \sum_{b \subsetneq a} m_f(b)\right]$;
$m_f(a) := \max\left[0,\ g(a) - \sum_{b \subsetneq a} m_f(b)\right]$;
If $f(\Omega) > 1$ **Then Fail;**

If the procedure does not fail, then

- f is a BF;
- $f \geq g$;
- f is the lowest element of $\overline{\mathcal{B}}(g)$.

The first two properties are obvious from the procedure. The last one is proved in the following assertions.

Lemma 3 *Let BF f be an output of the Upper approximation procedure applied to PBF g (it means the procedure does not fail). Then there does not exist $\hat{f} \in \overline{\mathcal{B}}(g)$ for which $f > \hat{f}$.*

Proof. Assume the opposite. Let $\hat{f} \in \overline{\mathcal{B}}(g)$, and $f > \hat{f}$. By Lemma 2, there exists in $\overline{\mathcal{B}}(g)$ a BF for which f is its simple specification. Thus, without loss of generality, we can assume that f is a simple specification of $\hat{f}$. It means that there exists $a \subseteq \Omega$, for which $f(a) > \hat{f}(a)$. It follows from the definition of simple specification that for all $b \subseteq \Omega$, $|b| < |a|$, BFs f and $\hat{f}$ coincide, i.e., for all $b \subsetneq a$, $f(b) = \hat{f}(b)$ and $m_f(b) = m_{\hat{f}}(b)$. The latter equality yields that

$$\hat{f}(a) \geq \sum_{b \subsetneq a} m_f(b).$$

Since $\hat{f} \in \overline{\mathcal{B}}(g)$, we also know that $\hat{f}(a) \geq g(a)$. It means that

$$\hat{f}(a) \geq \max\left[g(a); \textstyle\sum_{b \subsetneq a} m_f(b)\right] = f(a),$$

which is in contradiction with our assumption that $f(a) > \hat{f}(a)$. □

In other words, Lemma 3 states that Upper approximation procedure finds a solution that is not dominated by another BF. The following theorem states the solution is unique.

Theorem 4 *Let BF f be an output of the Upper approximation procedure applied to PBF g (it means the procedure does not fail). Then $\overline{\mathcal{B}}(f) = \overline{\mathcal{B}}(g)$.*

Proof. Since $f \in \overline{\mathcal{B}}(g)$, it is obvious that $\overline{\mathcal{B}}(f) \subseteq \overline{\mathcal{B}}(g)$. Therefore, assuming the opposite, there must be $\overline{f} \in \overline{\mathcal{B}}(g) \setminus \overline{\mathcal{B}}(f)$ for which there exists at least one $a \subseteq \Omega$ such that $\overline{f}(a) < f(a)$. Consider such a with the smallest cardinality, which means that for all $b \subsetneq a$, $f(b) \geq \overline{f}(b)$. This choice guarantees that $\hat{f} : 2^\Omega \to [0,1]$, $\hat{f}(a) = \overline{f}(a)$, and $\hat{f}(c) = f(c)$ for all remaining $c \in 2^\Omega \setminus \{a\}$ is correctly defined BF (it is monotonic). However, $\hat{f} < f$, which contradicts Lemma 3. □

Thus, the set of optimal upper BF-approximations of g

$$\overline{\mathcal{A}}(g) = \left\{f \in \overline{\mathcal{B}} : \overline{\mathcal{B}}(g) \cap \underline{\mathcal{B}}(f) = \{f\}\right\}$$

is a one-point set. It means that if the procedure fails, then $\overline{\mathcal{B}}(g) = \emptyset$. There cannot exist BF $\hat{f}$, for which $\sum_{b \subsetneq a} m_{\hat{f}}(b) > g(\Omega) = 1$.

Cybersecurity data example - part I

Table 1: Features (variables) characterizing organizations

Notation	Variable	# of values
L	Legislation (regulated/unregulated by law)	2
S	Security knowledge status of decision makers	5
T	Total security score of organization	5
U	User experience	4
V	Volume of resources invested in cybersecurity	5
W	Willingness of CIOs to educate themselves	5

In the current section, we illustrate the above-presented ideas concerning the data-based learning of BF models using the *cybersecurity* data collected by Švadlenka (2025) (for a more detailed description, see his cited PhD thesis). The data describe six characteristics (variables – see Table 1) of fifty-two organizations (records).

For the sake of simplicity, we start with the simplest possible case, considering only two dichotomous variables. We consider variables L and W, the latter binarized as shown in Table 2. Thus, we consider $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, and the available data, which are summarized in the 2×2 contingency table (Table 2).

Table 2: 2×2 contingency table

	$W =\leq 3$	$W > 3$
$L = 1$	$\{\omega_1\}$	$\{\omega_2\}$
$L = 2$	$\{\omega_3\}$	$\{\omega_4\}$

	$W =\leq 3$	$W > 3$
$L = 1$	14	17
$L = 2$	15	6

The learning process is based on the idea that a belief function of an event a is a lower bound for the possible probability of that event. The lower bound of a binomial confidence interval has a similar property. Namely, there is only a small chance that an actual probability of event a is less than the lower bound of a confidence interval computed from the given data. As the title of this article suggests, we are considering Jeffreys confidence intervals. We started with $\alpha = 0.05$ (Lee, 1989), although other levels and other estimates of confidence intervals can (and will) be used. The lower bounds of these intervals for $n = 52$ are tabulated in Table 3. The values are calculated using the R command (Crawley, 2012)

```
for(x in 0:52){print(c(x,qbeta(0.05/2, x+0.5, 52-x+0.5)))}.
```

Using this approach, we get a function $g : 2^\Omega \to [0, 1]$ that is monotone. If we also set its value for Ω to one, we get a PBF that fits the given data. This function is tabulated

Table 3: Lower limits of Jeffreys confidence intervals with $\alpha = 0.05$, and $n = 52$

	$0*$	$1*$	$2*$	$3*$	$4*$	$5*$
$*0$	0	0.1033921	0.2615272	0.4415067	0.6422402	0.9136317
$*1$	0.0020829	0.1179209	0.2786434	0.460575	0.6637672	0.9136317
$*2$	0.0080773	0.1327931	0.2959618	0.4798459	0.6856412	0.9530633
$*3$	0.0165147	0.1479794	0.3134786	0.4993247	0.7079003	
$*4$	0.0265221	0.1634559	0.3311906	0.5190176	0.7305931	
$*5$	0.0376467	0.1792032	0.3490958	0.5389325	0.7537837	
$*6$	0.0496248	0.1952054	0.367193	0.5590787	0.7775591	
$*7$	0.06228799	0.2114491	0.3854819	0.5794675	0.8020423	
$*8$	0.0755216	0.2279231	0.4039627	0.6001125	0.8274177	
$*9$	0.08924337	0.2446184	0.422637	0.6210301	0.8539855	

in Table 4. Since it is not a BF, we let the Upper approximation process modify it. The whole process is recorded in the table. Looking at the resulting function f, we see that it is not a BF, it is not normalized. This means that, due to Theorem 4, no BF dominates g. However, this does not mean that there is no way to find a BF corresponding to the given data with the properties of the upper approximation. Two ways to modify the function g are presented here, both related to the reliability of the process by which g was constructed. The alternative way — using a generalization of Shafer's discounting — is analyzed in our next contribution in these proceedings Daniel et al. (2025).

Table 4: Application of the Upper approximation process with $\alpha = 0.05$

a	# occurrences	$g(a)$	$\sum_{b \subsetneq a} m_f(b)$	$f(a)$	$m_f(a)$
$\{\omega_1\}$	14	0.163	0	0.163	0.163
$\{\omega_2\}$	17	0.211	0	0.211	0.211
$\{\omega_3\}$	15	0.179	0	0.179	0.179
$\{\omega_4\}$	6	0.050	0	0.050	0.050
$\{\omega_1, \omega_2\}$	31	0.461	0.375	0.461	0.086
$\{\omega_1, \omega_3\}$	29	0.423	0.343	0.423	0.080
$\{\omega_1, \omega_4\}$	20	0.262	0.213	0.262	0.048
$\{\omega_2, \omega_3\}$	32	0.480	0.391	0.480	0.089
$\{\omega_2, \omega_4\}$	23	0.313	0.261	0.313	0.052
$\{\omega_3, \omega_4\}$	21	0.279	0.229	0.279	0.050
$\{\omega_1, \omega_2, \omega_3\}$	46	0.778	0.809	0.809	0
$\{\omega_1, \omega_2, \omega_4\}$	37	0.579	0.611	0.611	0
$\{\omega_1, \omega_3, \omega_4\}$	35	0.539	0.571	0.571	0
$\{\omega_2, \omega_3, \omega_4\}$	38	0.600	0.632	0.632	0
Ω	52	1.000	1.009	1.009	0

We know that with $\alpha = 0.05$ it can happen that $\pi(a) < g(a)$ with probability 0.025.

Therefore, the probability that all 15 probabilities considered are greater than the respective values of the function g is only 0.975^{15}, which is less than 0.7. To deal with such unreliability, we can either decrease α, or apply the idea of Shafer (1976) to *discount* PBF g. The latter approach is to follow the idea of Shafer (1976) proposed for inaccurate sources: take a coefficient of *discount rate* δ and recompute values of g for all proper subsets $a \subsetneq \Omega$ as $g(a) := (1-\delta)g(a)$.

In the following, we consider the first case and lower the level of the confidence intervals used. First, we computed the function g based on the estimates of the Jeffreys confidence intervals with $\alpha = 0.03$. It turned out that this PBF did not have a non-empty credal set either. We succeeded in obtaining a suitable PBF by considering the estimates of Jeffreys confidence intervals with $\alpha = 0.02$. The corresponding PBF g and a log of the corresponding calculation are presented in Table 5. It is worth noting that a very similar result can be obtained by applying the same algorithm to the original g presented in Table 4, discounted at the discount rate of $\delta = 0.9$.

Table 5: Process of computation of the Upper approximation; $\alpha = 0.02$

a	# occurrences	$g(a)$	$\sum_{b \subsetneq a} m_f(a)$	$f(a)$	$m_f(a)$
$\{\omega_1\}$	14	0.147	0	0.147	0.147
$\{\omega_2\}$	17	0.192	0	0.192	0.192
$\{\omega_3\}$	15	0.162	0	0.162	0.162
$\{\omega_4\}$	6	0.041	0	0.041	0.041
$\{\omega_1, \omega_2\}$	31	0.436	0.339	0.436	0.097
$\{\omega_1, \omega_3\}$	29	0.398	0.308	0.398	0.090
$\{\omega_1, \omega_4\}$	20	0.241	0.187	0.241	0.053
$\{\omega_2, \omega_3\}$	32	0.455	0.354	0.455	0.101
$\{\omega_2, \omega_4\}$	23	0.291	0.233	0.291	0.058
$\{\omega_3, \omega_4\}$	21	0.257	0.202	0.257	0.055
$\{\omega_1, \omega_2, \omega_3\}$	46	0.755	0.788	0.788	0
$\{\omega_1, \omega_2, \omega_4\}$	37	0.554	0.587	0.587	0
$\{\omega_1, \omega_3, \omega_4\}$	35	0.513	0.547	0.547	0
$\{\omega_2, \omega_3, \omega_4\}$	38	0.575	0.608	0.608	0
Ω	52	1.000	0.994	1.000	0.006

It is perhaps unnecessary to say that decreasing the level of confidence α (or discount rate δ) finally always results in finding a PBF g with a nonempty credal set.

In this example, we were interested in finding a BF on Ω, whose cardinality was 4. To have a simple example, we had to binarize variable W. When considering non-simplified two-dimensional contingency tables of the considered cybersecurity example, the cardinality of the considered space of discernment increases up to 25. Not to speak of considering three-dimensional contingency tables, where the cardinality of Ω can grow up to 125. Thus, we easily go beyond the capacities of current computers, and the open question is whether there are similar approaches that take into account the need to keep the complexity of the resulting BPs reasonable.

4 Lower Approximations of Pseudo-Belief Functions

As mentioned above, the lower approximations should be preferred if one does not want the approximation to contain any information not contained in PBF g. Thus, in this case, we consider the set of optimal outer approximations

$$\underline{\mathcal{A}}(g) = \left\{ f \in \underline{\mathcal{B}} : \underline{\mathcal{B}}(g) \cap \overline{\mathcal{B}}(f) = \{f\} \right\}.$$

It is a set of Pareto optimal outer approximations of g that are not dominated by any other outer approximation of g. The problem remains which one to choose and how to compute it. As in many other areas of research where one has to choose one solution from a set of Pareto optimal solutions, it depends on whether there is some additional information or some supporting criterion to take into account. This is all subject to further research.

A relatively simple way to find a lower approximation is to apply the following Easy lower approximation procedure. In contrast to the Upper approximation procedure introduced in Section 3, where the procedure was unambiguously described, the following pseudo-code uses a step that can be implemented in several different ways.

Easy Lower Approximation Procedure
For $k = 1, ..., |\Omega|$
While $A := \left\{ a \subseteq \Omega : |a| = k \;\; \& \;\; g(a) < \sum_{b \subsetneq a} m_f(b) \right\} \neq \emptyset$
Choose any $a \in A$; $\varepsilon := g(a) - \sum_{b \subsetneq a} m_f(b)$;
Redistribute ε to $m_f(b_\ell), b_\ell \subsetneq a, \quad \ell = 1, \ldots, r$;
For $a \subseteq \Omega : |a| = k$
$m_f(a) := g(a) - \sum_{b \subsetneq a} m_f(b)$;
For $a \subseteq \Omega$
$f(a) := \sum_{b \subseteq a} m_f(b)$;

Recall that the step "*Redistribute* ε to $m_f(b_\ell), b_\ell \subsetneq a, \ell = 1, \ldots, r$;" means that you have to choose the system $b_\ell \subsetneq a, \ell = 1, \ldots, r$, and split ε into corresponding ε_ℓ so that all ε_ℓ are negative and $\varepsilon = \sum_{\ell=1}^{r} \varepsilon_\ell$. No matter how it is implemented, the process of redistributing a negative value ε to m_f always realizes several simple generalizations we discussed in Section 2. Thus, when the redistribution is finished, $g(a) = \sum_{b \subsetneq a} m_f(b)$. It can always be realized in such a way that all $m_f(b_\ell)$ are non-negative. In the example below, we will implement this step so that a third of ε is added to all $b \subset a : |b| = |a| - 1$. We take a third of ε because, in the case of the following example, it is applied when $|\{b \subset a : |b| = |a| - 1\}| = 3$.

As said, the discussed process of redistribution can always be done, because $\sum_{b \subsetneq a} m_f(b) > |\varepsilon|$. This can be implemented in many ways. It is also a topic for further research to study which of them is preferable. Note that if possible, one should redistribute ε to $m_f(b)$, for $b \subset a : |b| = |a| - 1$, and not to $m_f(b)$, for $b \subsetneq a$ with $|b| < |a| - 1$. This is because the simple generalization when ε_i is subtracted from $m_f(a)$ and added to $m_f(b)$ with $|b| = |a| - 2$ can be realized as two successive simple generalizations, first

from a to c (for $b \subset c \subset a$), and second from c to b, which would give a hint that the resulting BF does not belong to the Pareto optimal outer approximations.

There are even more open questions regarding the redistribution step. Although it is quite likely, we are not sure whether one can always redistribute ε only to sets b, for $b \subset a : |b| = |a| - 1$. However, the most important open question is whether there is an implementation that guarantees that the resulting approximation is Pareto optimal. Greater chances of producing Pareto optimal solutions have Advanced lower approximation procedure, which is described after an example in the following section. We will see that in the advanced procedure, the redistribution process is considered simultaneously for all subsets of the same cardinality, rather than separately in a cycle. However, this increases the computational complexity of the whole process.

Cybersecurity data example - part II

Let us get back to considering the cybersecurity data and the function g defined in Table 4, where its values correspond to the Jeffreys estimates of lower bounds of confidence intervals with $\alpha = 0.05$.

Table 6: Application of the Easy lower approximation process with $\alpha = 0.05$

		$k=1,2$	$k=3$									
a	$g(a)$	$m_f(a)$	ε	$m_f(a)$	ε	$m_f(a)$	ε	$m_f(a)$	ε	$m_f(a)$	$f(a)$	$f^\star(a)$
$\{\omega_1\}$	0.163	0.163		0.163		0.163		0.163		0.163	0.163	0.163
$\{\omega_2\}$	0.211	0.211		0.211		0.211		0.211		0.211	0.211	0.211
$\{\omega_3\}$	0.179	0.179		0.179		0.179		0.179		0.179	0.179	0.179
$\{\omega_4\}$	0.050	0.050		0.050		0.050		0.050		0.050	0.050	0.050
$\{\omega_1,\omega_2\}$	0.461	0.086		0.075		0.068		0.068		0.068	0.443	0.461
$\{\omega_1,\omega_3\}$	0.423	0.080		0.070		0.070		0.065		0.065	0.407	0.407
$\{\omega_1,\omega_4\}$	0.262	0.048		0.048		0.041		0.037		0.037	0.250	0.246
$\{\omega_2,\omega_3\}$	0.480	0.089		0.079		0.079		0.079		0.076	0.466	0.464
$\{\omega_2,\omega_4\}$	0.313	0.052		0.052		0.045		0.045		0.042	0.303	0.298
$\{\omega_3,\omega_4\}$	0.279	0.050		0.050		0.050		0.045		0.042	0.271	0.279
$\{\omega_1,\omega_2,\omega_3\}$	0.778		-0.031	0		0.007		0.012		0.015	0.778	0.778
$\{\omega_1,\omega_2,\omega_4\}$	0.579				-0.021	0		0.005		0.008	0.579	0.579
$\{\omega_1,\omega_3,\omega_4\}$	0.539						-0.014	0		0.003	0.539	0.539
$\{\omega_2,\omega_3,\omega_4\}$	0.600								-0.009	0	0.600	0.600
Ω	1									0.041	1	1

When applied to this function g, Easy lower approximation procedure skips the While cycle for $k = 1, 2$, because for these k, set $A = \left\{a \subseteq \Omega : |a| = k \ \ \& \ \ g(a) < \sum_{b \subsetneq a} m_f(b)\right\}$

is empty. A is nonempty only for $k = 3$. The whole calculation is shown in Table 6. The values of the resulting BF are in the column headed by f. We do not know if f is Pareto optimal or not. We do not know it even for the solution $f^\star$ in the last column, which was computed by the procedure described in the following section. The reader certainly noticed that two solutions f and $f^\star$ are incomparable.

5 Advanced Lower Approximations

The idea of this procedure, here called Advanced lower approximation, is based on the behavior of the Easy lower approximation procedure. The reader can see it in the example presented above. After applying the redistribution step to $a = \{\omega_1, \omega_2, \omega_3\} \in A$ (for $k = 3$) we get $g(a) = \sum_{b \subsetneq a} m_f(b)$, and therefore $m_f(a) = 0$. But in the next step of this cycle, the redistribution is applied to $\bar{a} = \{\omega_1, \omega_2, \omega_4\} \in A$. When this redistribution process is finished, we get, analogously, $g(\bar{a}) = \sum_{b \subsetneq \bar{a}} m_f(b)$, but for the preceding a we get $g(a) > \sum_{b \subsetneq a} m_f(b)$, and therefore $m_f(a) > 0$. Therefore, in the procedure we are going to describe, we leave the idea of redistributing negative ε's in a cycle, we want to redistribute all values simultaneously to get $m_f(a) = 0$ for all $a \in A$.

Perhaps the first idea could be to set up a system to be solved by linear programming methods. Unknown variables are all potential shifts of mass functions. Assume a fixed k, and that we only want to move masses to subsets whose cardinality is one less than the considered k. Denote $\ell_k = |\{b \subset a : |b|+1 = |a| = k\}|$. Then we get $\ell_k.|A_k|$ unknown variables. The constraints are given by the equality that for all $a \in A_k$, $g(a) = \sum_{b \subsetneq a} m_f(b)$, and the sum of all shifts from the set a must equal ε_a, and that all new values of $m_f(b)$ for all b are nonnegative. Such an approach is possible, but we believe it is unnecessarily computationally expensive. Much simpler is the following iterative process.

Advanced Lower Approximation Procedure
For $k = 1, ..., |\Omega|$
Set $A_k = \left\{ a \subseteq \Omega : |a| = k \;\; \& \;\; g(a) < \sum_{b \subsetneq a} m_f(b) \right\} \neq \emptyset$;
If $A_k \neq \emptyset$ **Then**
Untill $\max[|\varepsilon_a|] < 10^{-8}$
For $a \in A_k \;\; \varepsilon_a := g(a) - \sum_{b \subsetneq a} m_f(b)$;
Choose $a \in A_k : \max[|\varepsilon_a|]$;
Redistribute ε_a to $m_f(b_\ell), b_\ell \subsetneq a, \ell = 1, \ldots, r$;
For $a \subseteq \Omega : |a| = k$
$m_f(a) := g(a) - \sum_{b \subsetneq a} m_f(b)$;
For $a \subseteq \Omega$
$f(a) := \sum_{b \subseteq a} m_f(b)$;

Let us add two comments to the iterative process introduced above (realized in the Untill cycle). First, the criterion that ends the cycle means that you are willing to accept that all differences are so small that they can be considered zero. Second, realize that

the values of ε_a reach both negative and positive values. When splitting this value for Redistribution, all of its parts must be of the same sign. This is also why we choose a according to the absolute value of ε_a.

This procedure yielded the BF $f^\star$ from Table 6 Let us also remark that, up to now, we have always got along with the simplest possible realization of the Redistribution step by one simple cycle:

For $b \subset a : |b| + 1 = |a|$

$$m_f(b) := m_f(b) + \frac{\varepsilon_a}{|\{b \subset a : |b| + 1 = |a|\}|}.$$

6 Conclusion

The paper introduces two types of approximations of PBFs by BFs. The research is motivated by the idea that data-based learning of belief function models can start with a normalized monotone set function whose values are defined by lower bounds of the corresponding confidence intervals. Although the paper presents simple algorithms for obtaining both lower and upper bounds of PBFs, it raises more questions than it answers. The main questions concern the optimality of the proposed lower approximations. There is also more room for computational experiments and the design of heuristic algorithms, because, as it is quite natural in the framework of belief functions, all computational processes are of very high computational complexity. Thus, the paper provides a good basis for future research in several directions.

References

M. J. Crawley. *The R book*. John Wiley & Sons, 2012.

M. Daniel, R. Jiroušek, and V. Kratochvíl. Discounting or optimizing? different approaches to pseudo-belief function correction. In *Proceedings of the 13th Workshop on Uncertainty Processing (WUPES'25)*, pages 104–115, 2025.

D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In *IFAC Symposium on Theory and Application of Digital Control (IFAC 1982)*, pages 167–181. North Holland, 1982.

R. Jiroušek and V. Kratochvíl. About twenty-five naughty entropies in belief function theory: do they measure informativeness? *International Journal of Approximate Reasoning*, page 109454, 2025. doi: 10.1016/j.ijar.2025.109454.

P. M. Lee. *Bayesian statistics*. Oxford University Press London:, 1989.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976. ISBN 9780691100425.

R. Švadlenka. *Analýza implementace bezpečnostních opatření pod tíhou legislativních změn*. PhD thesis, Prague University of Economics and Business, 2025.

Discounting or Optimizing? Different Approaches to Pseudo-Belief Function Correction

Milan Daniel[1], Radim Jiroušek[2], and Václav Kratochvíl[2]

[1]Institute of Computer Sciences, Czech Academy of Sciences
[1]*milan.daniel@cs.cas.cz*
[2]Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic
[2] {*radim;velorex*}*@utia.cas.cz*

Abstract

We present and compare several approaches for transforming pseudo-belief functions, constructed from Jeffreys confidence intervals on observational data, into proper belief functions. Two main classes of methods are examined: one based on polyhedral geometry using various optimization strategies, and the other employing generalized belief discounting. Finally, the proposed methods are evaluated on real cybersecurity data and compared with standard upper and lower approximations of pseudo-belief.

1 Introduction

In a companion paper in these proceedings Daniel et al. (2025), belief functions were estimated from data using lower bounds based on Jeffrey's binomial confidence intervals. These bounds may not directly correspond to any valid belief function, leading to the notion of *pseudo-belief functions* — representations that preserve the intended epistemic meaning but violate some mathematical constraints of belief calculus.

This paper presents two complementary groups of methods for correcting such pseudo-belief functions and obtaining valid belief functions from them. Each method has its own motivation and interpretation:

- The first approach is based on *polyhedral geometry*. It considers the lower bounds (e.g., Jeffreys-type) as defining a polyhedron of admissible belief functions and selects one or more representative elements from this set using geometric or optimization-based criteria. This approach is constructive and data-driven.

- The second approach, introduced in this paper, generalizes the classical method of *belief discounting* as defined by Shafer (1976). It assumes a partial reliability of the original pseudo-belief function and proportionally reduces its support, originally transferring the remainder to total ignorance. We utilize the remainder for negative belief mass correction here. The result is a valid belief function that retains the internal structure of the original function.

While these two approaches differ in spirit — geometric reconstruction versus numerical correction — they share the same objective: to obtain valid belief functions that are compatible with uncertain or incomplete evidence. Moreover, belief discounting is a linear transformation and can be interpreted as a special case of movement within the credal set (i.e., the set of all belief functions consistent with given information), hence admitting a geometric interpretation.

By combining these perspectives, the paper contributes to the broader effort of belief function learning: deriving reasonable epistemic representations from empirical data, even in the presence of imprecision or ambiguity.

2 Preliminaries

This paper builds on a preceding contribution in these proceedings (Daniel et al., 2025), where belief and pseudo-belief functions were introduced and motivated by data-driven lower bounds such as Jeffreys intervals. These bounds may not always define a valid belief function, leading to pseudo-belief structures that require correction.

We explore two complementary correction strategies, each developed in a separate section. The first is based on polyhedral geometry and is presented next. The second relies on belief discounting and follows afterward. Here we briefly recall the key concepts common to both.

A belief function over a finite frame Ω is defined via a basic probability assignment (BPA) $m : 2^\Omega \to [0, 1]$ satisfying $m(\emptyset) = 0$, $\sum m(A) = 1$. Belief and plausibility functions are given by:

$$\mathrm{bel}_m(A) = \sum_{B \subseteq A} m(B), \quad \mathrm{pl}_m(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{1}$$

Pseudo-belief functions generalize this by allowing some negative masses while preserving the belief–plausibility duality.

An important correction tool is *discounting* (Shafer, 1976), used when evidence is not fully reliable. Given trust $1 - \delta$, the discounted mass function is defined as:

$$m^\delta(A) = (1 - \delta)m(A) \text{ for } A \neq \Omega, \quad m^\delta(\Omega) = (1 - \delta)m(\Omega) + \delta,$$

yielding a belief function bel^δ with $\mathrm{bel}^\delta(A) = (1 - \delta)\,\mathrm{bel}(A)$ for $A \neq \Omega$, $\mathrm{bel}^\delta(\Omega) = 1$. Discounting increases uncertainty while preserving belief ratios.

Finally, when belief or plausibility bounds are defined by inequalities, they form a polyhedron

$$\mathcal{P} = \{x \in \mathbb{R}^n : Mx \leq b\},$$

which may or may not correspond to any belief function. We use geometric methods to modify such sets into valid belief structures — a topic of the next section.

3 Geometric Correction via Polyhedral Optimization

In the preceding chapters, we introduced pseudo-belief functions derived from empirical data using Jeffreys confidence intervals, as motivated in Daniel et al. (2025). These define lower and upper bounds for the belief and plausibility of each subset $A \subseteq \Omega$, resulting in a system of linear inequalities that constrains possible belief functions.

Let $\mathbf{bel}_J \in \mathbb{R}^{2^n}$ denote the vector of lower bounds (Jeffreys intervals) for each subset A. The inequality

$$\mathrm{bel}_J(A) \leq \sum_{B \subseteq A} m(B)$$

can be rewritten in matrix form as

$$Mm \geq \mathbf{bel}_J,$$

where $m \in \mathbb{R}^{2^n}$ is a vector of bpa values and $M \in \{0,1\}^{2^n \times 2^n}$ is the inclusion matrix with entries $M_{[A,B]} = 1$ iff $B \subseteq A$.

If we add constraints for normalization and non-negativity,

$$\sum_{A \subseteq \Omega} m(A) = 1, \qquad m(A) \geq 0,$$

we obtain a polytope $\mathcal{P}_J^* \subset \mathbb{R}^{2^n}$ containing all belief functions consistent with the empirical bounds. The polytope may have many vertices, corresponding to different consistent BPAs.

Polyhedral geometry offers strong tools for selecting one specific point $m \in \mathcal{P}_J^*$ by optimizing a suitable objective function. This approach draws on the convex geometry of belief spaces as explored in Cuzzolin (2010, 2020) and relies on standard polyhedral optimization methods (Ziegler, 1995; Bagnara et al., 2008).

We consider four optimization criteria:

Zero Objective (ZO). Selects any feasible point:

$$\text{Minimize } f_{\mathrm{ZO}}(m) = 0.$$

Sparsity (SP). Minimizes the number of focal elements:

$$\text{Minimize } f_{\mathrm{SP}}(m) = \sum_{A \subseteq \Omega, A \neq \emptyset} \delta[m(A) > 0],$$

where $\delta[\cdot]$ is the indicator function. This is approximated in LP by introducing binary variables z_A and constraints $m(A) \leq M \cdot z_A$, for large M.

Cardinality-Weighted (CW). Penalizes small subsets:

$$\text{Minimize } f_{\mathrm{CW}}(m) = \sum_{A \subseteq \Omega, A \neq \emptyset} \frac{1}{|A|} m(A).$$

Dubois–Prade Entropy (HD). Maximizes entropy:

$$\text{Maximize } H_D(m) = \sum_{A \subseteq \Omega, A \neq \emptyset} m(A) \log |A|,$$

as proposed in Dubois and Prade (1987).

The resulting belief mass assignments under each objective are shown below for the example with $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$:

Table 1: Comparison of belief mass assignments under different objective functions. Column m_J shows the pseudo-belief mass vector constructed directly from Jeffreys intervals (Daniel et al., 2025).

A	$\text{bel}_J(A)$	$m_J(A)$	$m_{ZO}(A)$	$m_{SP}(A)$	$m_{CW}(A)$	$m_{HD}(A)$
$\{\omega_1\}$	0.164	0.163	0.400	0.163	0.164	0.174
$\{\omega_2\}$	0.211	0.211	0.321	0.297	0.211	0.222
$\{\omega_3\}$	0.179	0.179	0.229	0.317	0.179	0.190
$\{\omega_4\}$	0.050	0.050	0.050	0.222	0.050	0.060
$\{\omega_1, \omega_2\}$	0.461	0.086	0	0	0.081	0.065
$\{\omega_1, \omega_3\}$	0.423	0.080	0	0	0.075	0.059
$\{\omega_1, \omega_4\}$	0.262	0.048	0	0	0.044	0.027
$\{\omega_2, \omega_3\}$	0.480	0.089	0	0	0.089	0.068
$\{\omega_2, \omega_4\}$	0.314	0.052	0	0	0.052	0.031
$\{\omega_3, \omega_4\}$	0.279	0.050	0	0	0.050	0.029
$\{\omega_1, \omega_2, \omega_3\}$	0.778	-0.031	0	0	0	0
$\{\omega_1, \omega_2, \omega_4\}$	0.580	-0.032	0	0	0	0
$\{\omega_1, \omega_3, \omega_4\}$	0.539	-0.032	0	0	0	0
$\{\omega_2, \omega_3, \omega_4\}$	0.600	-0.032	0	0	0	0
Ω	1.000	0.117	0	0	0	0.075
H_D			0	0	0.271	0.297

Discussion. Each optimization criterion leads to a different belief assignment. The **Sparsity (SP)** solution concentrates belief on a minimal number of focal elements, which may be beneficial for interpretability. The **Cardinality-Weighted (CW)** solution favors larger sets, thereby reflecting cautious reasoning. The **Dubois–Prade entropy (HD)** solution maximizes epistemic uncertainty and spreads mass over larger focal elements, constrained by empirical bounds. Interestingly, the **CW** solution closely resembles the approximation obtained in (Daniel et al., 2025) using the Upper Approximation Procedure, which supports the validity of our optimization-based approach.

These results illustrate that geometric correction methods not only ensure consistency with empirical estimates but also allow tailoring belief functions according to different modeling principles or user preferences.

4 A Generalization of Belief Discounting

Let us assume a PBF bel_J constructed by application of the method of Jeffreys confidence intervals (Daniel et al., 2025) or any general PBF, which does not satisfy the classic Shafer's definition of BF - there are some focal elements in respective BPA with negative mass, cf. negative ε in lower approximation in Daniel et al. (2025). Our aim is simply to find an acceptable way how to eliminate these negative belief masses. From the definition of a PBF, all pseudo-beliefs are non-negative, thus also all belief masses of singletons are obviously non-negative. Hence, an issue can appear only for $X \subseteq \Omega$, $|X| \geq 2$.

We can consider the negative mass $m(X)$ as a consequence of over information about the case which is the subject of the belief; excess information on subsets of X, which can be corrected by removal of at least belief mass corresponding to the sum of negative belief masses.

Example 1 In the simplest case of such a PBF, i.e., bel_S on $|\Omega_2| = 2$, $m_S(\Omega_2) = \varepsilon < 0$, we can simply solve the problem by discounting with discounting rate $\delta \geq -\varepsilon/(1-\varepsilon)$: $m_S^\delta(\Omega_2) = (1-\delta)m_S(\Omega_2) + \delta \geq (1-(-\varepsilon/(1-\varepsilon))\varepsilon - \varepsilon/(1-\varepsilon) = \varepsilon(1+\varepsilon/(1-\varepsilon) - 1/(1-\varepsilon) = \varepsilon(1-1) = 0$, $m_S^\delta(\omega_i) = (1-\delta)m_S(\omega_i)$, which is also non-negative, as $1-\delta = 1+\varepsilon/(1-\varepsilon) = (1-\varepsilon)/(1-\varepsilon) + \varepsilon/(1-\varepsilon) = 1/(1-\varepsilon) > 0$. Just the same procedure we can use for correction of any PBF on $|\Omega_n| = n$, with the only negative belief mass $m(\Omega_n)$.

In general, we have to correct all negative belief masses, of any focal element $|X| > 1$. We can do it in three different ways:

(i) *local correction* of $m(X)$, related to the only focal element $X \subset \Omega$ and its subsets,
(ii) *layered correction* of all $m(X)$, s.t. $|X| = k$, related only to focal elements $|Y| \leq k$,
(iii) *global correction* which correct all the focal elements with negative masses together.

Motivated by the above solution of the simplest PBF case we will try to generalize discounting as it follows.

Local discounting on $X \subset \Omega$: $m^{\ulcorner X \urcorner \delta}(A) = (1-\delta)m(A)$ for any $A \subset X$, $m^{\ulcorner X \urcorner \delta}(X) = m(X) + \delta \sum_{A \subset X} m(A)$ and $m^{\ulcorner X \urcorner \delta}(A) = m(A)$ for other subsets $A \subseteq \Omega$, i.e., for $A \not\subseteq X$.
Property 1: If X is disjoint with all focal elements of the same and less cardinality which are not its subsets ($X \cap Y = \emptyset$ for all $|Y| \leq |X|$ s.t. $Y \not\subseteq X$) we can describe reasonable property of this definition of local discounting: ($bel^{\ulcorner X \urcorner \delta}(A) = (1-\delta)bel(A)$ for any $A \subset X$, $bel^{\ulcorner X \urcorner \delta}(A) = bel(A)$ if A $\not\subset X$).
Property 2: If X is disjoint with all focal elements of the same cardinality, we can describe the following property of this definition of local discounting: ($bel^{\ulcorner X \urcorner \delta}(A) = (1-\delta)bel(A)$ for any $A \subset X$, $(1-\delta)bel(A) \leq bel^{\ulcorner X \urcorner \delta}(A) \leq bel(A)$ for any $|A| < |X| : A \not\subset X \& A \cap X \neq \emptyset$, $m^{\ulcorner X \urcorner \delta}(A) = m(A)$ otherwise.

Example 2 Let suppose $|\Omega_7| = 7$, with only two focal elements of cardinality 3, $|A_i| = 3$: $A_1 = \{\omega_1, \omega_2, \omega_3\}$, $A_2 = \{\omega_4, \omega_5, \omega_6\}$. Property 2 holds for both $bel^{\ulcorner A_i \urcorner \delta}$. If further $m(X) = 0$ for any $X \subseteq A_1 \cap A_2$, property 1 also holds for both $bel^{\ulcorner A_i \urcorner \delta}$. If there are added f.e.s $A_3 = \{\omega_4, \omega_5, \omega_7\}$, $A_4 = \{\omega_5, \omega_6, \omega_7\}$, the properties does hold only for $bel^{\ulcorner A_2 \urcorner \delta}$.

Hence, the above useful properties does not hold for general PBFs. Jeffreys bel_J has often focal elements intersecting with the others of the same cardinality. Thus for our reason the definition of local discounting should be improved in the future.

Cardinality or k-discounting for $1 < k \leq n$: $m^{k\delta}(A) = (1-\delta)m(A)$ for any $A : |A| < k$, $m^{k\delta}(A) = m(A) + \delta \sum_{B \subset A} \frac{m(A)}{\sum_{B \subset C : |C| = k} m(C)} m(B)$ for any $|A| = k$, and $m^{k\delta}(A) = m(A)$ for any $|A| > k$.

The formula - more precisely the coefficient of $m(B)$ seems to be rather complicated here, nevertheless we need to distribute any $m(B)$ among subsets of cardinality k, resp. just among all C, such that $B \subset C$ & $|C| = k$.

Observation 1 We can observe that $bel^{k\delta}(A) = (1-\delta)bel(A)$ for $|X| < k$ and $(1-\delta)bel(A) \leq bel^{k\delta}(A) \leq bel(A)$ for $|A| \geq k$: the first equality holds for $m(A) = 0$ and the second if there is the only one focal element of cardinality k, specially for $k = n$. [To be proved]

Observation 2 We can observe $m^{\lceil \Omega_n \rceil \delta}(A) = bel^{n\delta}(A) = bel^{\delta}(A)$ for $|\Omega_n| = n$. Thus both local and cardinality discounting are generalization of the original Shafer's discounting.

Proof. $m^{\lceil \Omega_n \rceil \delta}(\Omega_n) = m(\Omega_n) + \delta \cdot \sum_{A \subset \Omega_n} m(A) = (1-\delta)m(\Omega_n) + \delta m(\Omega_n) + \delta \cdot \sum_{A \subset \Omega_n} m(A) = (1-\delta)m(\Omega_n) + \delta(m(\Omega_n) + \sum_{A \subset \Omega_n} m(A)) = (1-\delta)m(\Omega_n) + \delta = m^{\delta}(\Omega_n)$. $m^{n\delta}(\Omega_n) = m(\Omega_n) + \delta \cdot \sum_{B \subset \Omega_n} m(\Omega_n)/m(\Omega_n) \cdot m(B) = (1-\delta)m(\Omega_n) + \delta m(\Omega_n) + \delta \cdot \sum_{B \subset \Omega_n} m(B)) = (1-\delta)m(\Omega_n) + \delta = m^{\delta}(\Omega_n)$.

5 Reduction of Over-Belief by Generalized Discounting

5.1 General Remarks on PBF Correction

Motivated by the successful Example 1, we have generalized belief discounting to allow analogous correction of general PBFs in the previous section.

As we have not yet been fully successful with the generalization of local discounting for general PBFs while preserving the ratios of belief masses, we resort to using only cardinality-based discounting in our corrections here. It affects all focal elements of a given cardinality (it distributes the discounted belief mass among all of them) and, similarly to classical Shafer discounting, adds the discounted mass to only one cardinality. Thus, our local and global corrections are in fact mixtures of local/global and layered corrections.

We must correct all belief masses < 0, i.e., such that $\sum_{B \subset A} m(B) > bel(A)$, i.e., with ratio $R(A) = bel(A) / \sum_{B \subset A} < 1$. If the lowest ratio among focal elements of cardinality k is used, then the strongest correction is performed, and the entire cardinality level is corrected. If the highest ratio is used, the smallest correction is performed, and only the focal element with that ratio is corrected. Note that a more complex formula for distributing the discounted belief mass is applied in cardinality discounting compared to classical discounting, and thus determining δ is also more complex, even though the underlying idea is analogous to that in Example 1.

5.2 Local, Layered, and Global Corrections of PBFs

1 Local correction should be the most precise, nevertheless it appears more complicated both from the theoretical point of view and also due to its computational complexity. As the theoretical part is not yet fully investigated, we adopt a mixture of local and layered approaches, correcting entire cardinality levels or individual focal elements one by one.

2 Layered correction is a compromise approach that corrects entire cardinality levels.

3 Global correction should correct all negative belief masses of the entire PBF together, if possible. Nevertheless, we again use only a mixture with layered correction.

5.3 Local Correction Algorithms

5.3.1 Algorithm 1-ugr

For each cardinality with negative pseudo-belief mass(es), we repeatedly utilize cardinality discounting with minimal correction discount rates (i.e., upward from minimal correction), correcting focal elements of cardinality k one by one using discount rates $\delta_A = -\sum_{|B|=k} m(B)/\sum_{C\subset A} m(C)$ for $|A| = k$, ordered from minimal to maximal, correcting pseudo-belief mass $m(A)$ if it is negative.

Algorithm 1

Compute pseudo m_J by Möbius transformation from bel_J, i.e., from the lower bounds of estimated confidence intervals;
For all $X \subseteq \Omega_n$:
$m_1(X) := m_J(X)$,
$R(X) := \min(1, bel_J(X)/\sum_{Y\subset X} m_J(X))$ for $|X| > 1$, $bel_J(X) > 0$,
$R(X) := 1$ for $|X| = 1$ or $bel_J(X) = 0$ or $\sum_{Y\subset X} m_J(X) = 0$.
$n := |\Omega_n|$;
For $k = 2, ..., n$:
$r_k := \max_{|X|=k} R(X)$,
If $r_k < 1$ **Then** $RFE := \{X \subseteq \Omega_n \mid |X| = k\}$
While $RFE \neq \emptyset$:
$AFE := \{X \in RFE \mid \text{with max } \sum_{Y\in RFE} m_{k-1}(Y)\}$
$sum_A := \sum_{B\subset A} m_{k-1}(B)$ for some $A \in AFE$,
$\delta_A := -\sum_{|B|=k} m_{k-1}(B)/sum_A$,
$m_k(X) := (1 - \delta_A) \cdot m_{k-1}(X)$ for $|X| < k$,
$m_k(X) := m_{k-1}(X) + \left|m_{k-1}(X) \cdot \sum_{Y\subset X} m(Y)/sum_A\right|$ for $|X| = k$,
$m_k(\Omega_n) := m(\Omega_n) + \sum_{|X|=k} \left(m_{k-1}(X)/sum_A \cdot \sum_{Y\not\subset X} m_{k-1}(Y)\right)$
Else For all $X \subseteq \Omega_n$: $m_k(X) := m_{k-1}(X)$;
% Now, $m(X) \geq 0$ for all X s.t. $|X| \leq k$
For all $X \subseteq \Omega_n$: $m(X) := m_n(X)$.

In the case of our cybersecurity data example from Daniel et al. (2025), there are four negative pseudo-belief masses of 3-element focal elements: $m(\{\omega_1, \omega_2, \omega_3\}) = -0.03139$, $m(\{\omega_1, \omega_2, \omega_4\}) = -0.03158$, $m(\{\omega_1, \omega_3, \omega_4\}) = -0.03159$, and $m(\{\omega_2, \omega_3, \omega_4\}) = -0.03157$ (see Table 1 and also the red values in Table 2). The corresponding discount rates are: $\delta_{(123)} = 0.155935$, $\delta_{(234)} = 0.044932$, $\delta_{(124)} = 0.002343$, $\delta_{(134)} = 0.000130$. For the final mass assignment further denoted as m_1 and the corresponding BF bel_1, see Tables 2 and 3. The belief function bel_1 is in fact a composition of four 3-discountings:

$$bel_1 = (((bel_J^{3\delta_{(123)}})^{3\delta_{(234)}})^{3\delta_{(124)}})^{3\delta_{(134)}}.$$

5.3.2 Algorithm 1-dgr

We utilize cardinality discounting with the maximal correction rate (i.e., the highest rate necessary to correct all negative belief masses of focal elements of cardinality k). Since this correction always affects all focal elements of cardinality k simultaneously, it essentially corresponds to Algorithm 2 from the next subsection.

5.3.3 Algorithms 1-ulr and 1-dlr

These algorithms aim to be closer to truly local corrections, either upward from the minimal or downward from the maximal correction. However, it is still under investigation whether it is possible to define an improved version of local discounting that preserves belief mass proportions as much as possible.

5.4 Layered Correction Algorithm

We apply cardinality discounting with the maximal discount rate, i.e., the minimal rate that corrects all negative belief masses of focal elements of a given cardinality k. This correction thus always affects all focal elements of that cardinality together.

In the case of our cybersecurity data example from Daniel et al. (2025), the maximal discount rate is $\delta_{(134)} = 0.2211$. Since negative pseudo-belief masses only appear at cardinality 3, the resulting BF further denoted bel_2 is a simple application of 3-discounting: $bel_2 = bel_J^{3\delta_{(134)}}$, see Tables 2 and 3. Note that this discount rate corresponds to $\{\omega_1, \omega_3, \omega_4\}$, which was the last focal element corrected in Algorithm 1-ugr. Nevertheless, the discount rate differs because here it is applied directly to the original bel_J, whereas in Algorithm 1-ugr it was applied after three previous corrections.

Algorithm 2. Layered Correction

Compute pseudo m_J by Möbius transformation from bel_J;
For all $X \subseteq \Omega_n$:
$m_1(X) := m_J(X)$,
$R(X) := \min(1, bel_J(X)/\sum_{Y \subset X} m_J(X))$ for $|X| > 1$, $bel_J(X) > 0$,
$R(X) := 1$ for $|X| = 1$ or $bel_J(X) = 0$ or $\sum_{Y \subset X} m_J(X) = 0$,
$n := |\Omega_n|$;
For $k = 2, \ldots, n$:
$r_k := \min_{|X|=k} R(X)$,
If $r_k < 1$ **Then**
$sum_\delta := \min_{|X|=k} \sum_{B \subset X} m_{k-1}(B)$,
$\delta_k := -\sum_{|B|=k} m_{k-1}(B)/sum_\delta$,
$m_k(X) := (1 - \delta_k) \cdot m_{k-1}(X)$ for $|X| < k$,
$m_k(X) := m_{k-1}(X) + |m_{k-1}(X) \cdot \sum_{Y \subset X} m_{k-1}(Y)/sum_\delta|$ for $|X| = k$,
$m_k(\Omega) := m_{k-1}(\Omega_n) + \sum_{|X|=k} \Big(m_{k-1}(X)/sum_\delta \cdot \sum_{Y \not\subset X} m_{k-1}(Y) \Big)$,
Else For all $X \subseteq \Omega_n$: $m_k(X) := m_{k-1}(X)$;
% $m_{k-1}(X) \geq 0$ for all $|X| \leq k$
For all $X \subseteq \Omega_n$: $m(X) := m_n(X)$.

5.5 Global Correction Algorithm(s)

Unfortunately, we do not yet have a truly global correction. Since cardinality discounting only corrects one cardinality level at a time, this method corresponds to an upside-down layered discounting — starting from cardinality n downward.

In the cybersecurity data example, there is only one cardinality (3) with negative pseudo-belief masses. Thus, the result of this approach is again $bel_3 = bel_J^{3\delta_{(134)}}$.

Algorithm 3. Global Correction

Compute pseudo m_J by Möbius transformation from bel_J; $n := |\Omega_n|$;
For all $X \subseteq \Omega_n$:
$m_1(X) := m_J(X)$,
$R(X) := \min(1, bel_J(X)/\sum_{Y \subset X} m_J(X))$ for $|X| > 1$, $bel_J(X) > 0$,
$R(X) := 1$ for $|X| = 1$ or $bel_J(X) = 0$ or $\sum_{Y \subset X} m_J(X) = 0$;
For $k = n, n-1, \ldots, 2$:
$r_k := \min_{|X|=k} R(X)$,
If $r_k < 1$ **Then**
$sum_\delta := \min_{|X|=k} \sum_{B \subset X} m_{k-1}(B)$,
$\delta_k := -\sum_{|B|=k} m_{k-1}(B)/sum_\delta$,
$m_k(X) := (1 - \delta_k) \cdot m_{k-1}(X)$ for $|X| < k$,
$m_k(X) := m_{k-1}(X) + |m_{k-1}(X) \cdot \sum_{Y \subset X} m_{k-1}(Y)/sum_\delta|$ for $|X| = k$,
$m_k(\Omega) := m_{k-1}(\Omega_n) + \sum_{|X|=k} \Big(m_{k-1}(X)/sum_\delta \cdot \sum_{Y \not\subset X} m_{k-1}(Y) \Big)$,
Else For all $X \subseteq \Omega_n$: $m_k(X) := m_{k-1}(X)$;
% $m_{k-1}(X) \geq 0$ for all $|X| \leq k$
For all $X \subseteq \Omega_n$: $m(X) := m_n(X)$.

5.5.1 Alternative Algorithms

There are two ideas for alternative algorithms. The first is a layered approach that mirrors Algorithm 1 in reverse: stepwise correction from the smallest to the largest $R(X)$ within each cardinality, moving downward from n to 2. The second is an open question: whether it is possible to define some generalized discounting operation that can correct all "negative" cardinalities at once.

6 Comparison on Cybersecurity Data Example

Let us compare lower and advanced lower approximations described in Daniel et al. (2025) with the approaches studied here. Specially, with geometric cardinality-weighted minimization (CW) and Dubois-Prade entropy (HD) and also with local and layered correction based on generalized discounting.

As all studied approaches are just in the process of their development, also our implementations are still in progress. Thus, we currently have correct comparable results only on 4-element frames of discernment now. General procedures are still in the middle of their tuning. Hence we will compare our approaches on the simplest case defined on cybersecurity data by Table 2 in Daniel et al. (2025): having 52 data records on the

4-element frame of discernment. For approximated/corrected belief mass assignments see Table 2, for approximated/corrected BFs see Table 3. Note that indices of m_1, m_2, bel_1, bel_2 refer results of Algorithms 1 and 2 here, not intermediate steps of their processing.

We skip here upper approximations from Daniel et al. (2025) as their results are not BFs in general, as it was presented there. In the case of Table 4 (there) $f(a)$ is even more general than pseudo-belief function as the sum of corresponding belief masses over all subsets of Ω is greater than 1. We also skip zero objective (ZO) and Sparsity (SP) geometric approximation, as ZO with its zero objective function returns an ad-hoc BF from the corresponding polytope and SP assigns all belief masses to singletons, thus a large information is added there and the belief structure of pseudo-belief bel_J is completely lost there.

Finally, we have to recall that both correction Algorithms 2 and 3 produce the same results on our simple data example, having negative pseudo-belief masses only on focal elements of cardinality 3. Having our still limited experience with pseudo-belief functions based on Jeffreys confidence interval, we have a hypothesis, the Jeffreys PBFs on $|\Omega| = 4$ have either no negative pseudo-belief masses or have negative belief masses only on focal elements of cardinality 3, hence Algorithms 2 and 3 both produce the same results on any data on any 4-element frame.

Table 2: Comparion pseudo-belief masses derived from Jeffreys intervals $m_J = m_g$ with degree of confidence $\alpha = 0.05$ on cybersecurity data on $|\Omega| = 4$ with its corrections: m_f and m_{f^*} from Daniel et al. (2025),m_{CW} and m_{HD} obtained by geometric Cardinality-Weighted and Dubois-Prade entropy, m_1 and m_2 by correction Algorithms 1 and 2.

A	$bel_J(A)$	$\sum_{B\subsetneq A} m_J(B)$	$m_J(A)$	$m_f(A)$	$m_{f^*}(A)$	$m_{CW}(A)$	$m_{HD}(A)$	$m_1(A)$	$m_2(A)$
$\{\omega_1\}$	0.1635	0	0.16346	0.1635	0.1635	0.1635	0.1739	0.1314	0.1273
$\{\omega_2\}$	0.2114	0	0.21145	0.2114	0.2114	0.2114	0.2219	0.1700	0.1647
$\{\omega_3\}$	0.1792	0	0.17920	0.1792	0.1792	0.1792	0.1897	0.1441	0.1396
$\{\omega_4\}$	0.0496	0	0.04962	0.0496	0.0496	0.0496	0.0603	0.0399	0.0387
$\{\omega_1,\omega_2\}$	0.4606	0.3749	0.08567	0.0682	0.0857	0.0810	0.0647	0.0689	0.0667
$\{\omega_1,\omega_3\}$	0.4226	0.3427	0.07998	0.0648	0.0643	0.0754	0.0590	0.0643	0.0623
$\{\omega_1,\omega_4\}$	0.2615	0.2131	0.04845	0.0367	0.0327	0.0438	0.0273	0.0390	0.0377
$\{\omega_2,\omega_3\}$	0.4798	0.3901	0.08919	0.0756	0.0735	0.0892	0.0683	0.0717	0.0695
$\{\omega_2,\omega_4\}$	0.3135	0.2611	0.05240	0.0422	0.0366	0.0524	0.0313	0.0421	0.0408
$\{\omega_3,\omega_4\}$	0.2786	0.2288	0.04982	0.0420	0.0497	0.0498	0.0287	0.0409	0.0388
$\{\omega_1,\omega_2,\omega_3\}$	0.7776	0.8089	-0.03139	0.0149	0	0	0	0	0.0131
$\{\omega_1,\omega_2,\omega_4\}$	0.5795	0.6111	-0.03158	0.0078	0	0	0	0	0.0022
$\{\omega_1,\omega_3,\omega_4\}$	0.5389	0.5705	-0.03159	0.0031	0	0	0	0	0
$\{\omega_2,\omega_3,\omega_4\}$	0.6001	0.6317	-0.03157	0	0	0	0	0	0.0034
Ω	1.0000	0.8830	0.11690	0.0409	0.0538	0	0.0749	0.1884	0.1952

What can we see in the tables?

The important is that both f and f^* and also both m_1 and m_2 not increase or even decrease their value comparing with $g = bel_J$, i.e., all four are $\leq g$; this corresponds to the fact that f and f^* are lower approximation and bel_i's are constructed using generalized

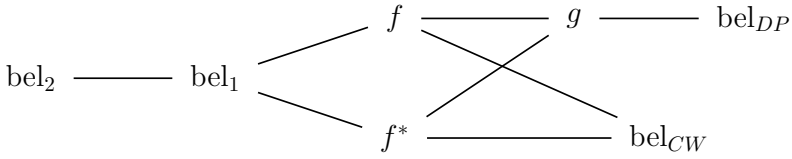
Table 3: Comparison of belief functions — corrected pseudo-beliefs derived from Jeffreys intervals with degree of confidence $\alpha = 0.05$ on cybersecurity data

A	$bel_J(A)$	$\sum_{b \subsetneq a} m_J(A)$	$m_J(A)$	$f(A)$	$f^*(A)$	$bel_{CW}(A)$	$bel_{HD}(A)$	$bel_1(A)$	$bel_2(A)$
$\{\omega_1\}$	0.1635	0.0000	0.16346	0.1635	0.1635	0.1635	0.1739	0.1314	0.1273
$\{\omega_2\}$	0.2114	0.0000	0.21145	0.2114	0.2114	0.2114	0.2219	0.1700	0.1647
$\{\omega_3\}$	0.1792	0.0000	0.17920	0.1792	0.1792	0.1792	0.1897	0.1441	0.1396
$\{\omega_4\}$	0.0496	0.0000	0.04962	0.0496	0.0496	0.0496	0.0603	0.0399	0.0387
$\{\omega_1, \omega_2\}$	0.4606	0.3749	0.08567	0.4431	0.4606	0.4560	0.4606	0.3704	0.3587
$\{\omega_1, \omega_3\}$	0.4226	0.3427	0.07998	0.4075	0.4069	0.4180	0.4226	0.3399	0.3292
$\{\omega_1, \omega_4\}$	0.2615	0.2131	0.04845	0.2498	0.2457	0.2569	0.2615	0.2103	0.2037
$\{\omega_2, \omega_3\}$	0.4798	0.3901	0.08919	0.4663	0.4642	0.4798	0.4798	0.3859	0.3738
$\{\omega_2, \omega_4\}$	0.3135	0.2611	0.05240	0.3033	0.2977	0.3135	0.3135	0.2521	0.2442
$\{\omega_3, \omega_4\}$	0.2786	0.2288	0.04982	0.2708	0.2785	0.2786	0.2786	0.2241	0.2170
$\{\omega_1, \omega_2, \omega_3\}$	0.7776	0.8089	-0.03139	0.7776	0.7776	0.7997	0.7776	0.6505	0.6432
$\{\omega_1, \omega_2, \omega_4\}$	0.5795	0.6111	-0.03158	0.5795	0.5795	0.6018	0.5795	0.4914	0.4782
$\{\omega_1, \omega_3, \omega_4\}$	0.5389	0.5705	-0.03159	0.5389	0.5389	0.5613	0.5389	0.4588	0.4444
$\{\omega_2, \omega_3, \omega_4\}$	0.6001	0.6317	-0.03157	0.6001	0.6001	0.6317	0.6001	0.5080	0.4954
Ω	1.0000	0.8831	0.11690	1.0000	1.0000	0.9954	1.0000	1.0000	1.0000

discounting. Thus all these corrections decreases information of the original PBF g. Moreover, it holds $bel_2 \leq bel_1 \leq f \leq g$ and also $bel_2 \leq bel_1 \leq f^* \leq g$, while f and f^* are mutually $\leq$-incomparable. $bel_2 \leq bel_1$, thus bel_1 is closer to g, nevertheless it has higher computational complexity, which does not play any role on our small example on $|\Omega| = 4$. Both f and f^* are even closer to original g, nevertheless they use ad-hoc negative belief mass redistribution, which may despite closeness to g to add a piece of ad-hoc information, while both bel_1 and bel_2 satisfy all belief proportions at any cardinality of $A \subset \Omega$, hence they better keeps the belief structure of the original PBF $g = bel_J$.

bel_{CW} is $\leq$-incomparable both with bel_{HD} and g while $bel_{HD} \geq g$, thus also $\geq$ all other which are $\leq g$. $bel_{CW} \geq bel_1$, bel_2, f and f^*. Thus both these geometric corrections add some extra information. bel_{CW} has a strange ad-hoc feature: that belief of some couples are increased ($\{\omega_1, \omega_2\}$,$\{\omega_1, \omega_3\}$,$\{\omega_1, \omega_4\}$), while the other keep the same belief mass as the original PBF g has ($\{\omega_2, \omega_3\}$,$\{\omega_2, \omega_4\}$,$\{\omega_3, \omega_4\}$). Hence, there is a challenging open problem: finding more convenient optimization criteria for pseudo-belief correction.

We can summarize our $\leq$-comparison by the following schema.



Unfortunately, negative pseudo-belief masses are relatively quite small and their values

are almost the same (differ on 5th decimal place) in the compared example, thus also the differences of their corrections are rather similar. Hence we have to compare our approaches not only on greater frame of discernment but also on various cases on Ω_4.

Finally, we have to note that this comparison is related only to the one simple case or real data on very small frame of discernment. It is rather a presentation how we can compare our approaches in near future having processed more examples.

One of the interesting open questions is which relations from the $\leq$-comparison schema are general, which of them are frequent, and which of them are rare or even exceptional.

7 Conclusion

Following our preceding contribution Daniel et al. (2025), we have proposed and presented several methods for transforming pseudo-belief functions into classical belief functions. The investigated procedures are based on fundamentally different approaches to correcting pseudo-beliefs. All the presented methods have been compared with those from Daniel et al. (2025) using a simple example based on real cybersecurity data. The implementation of our algorithms is currently under development. This will allow us to perform more comprehensive comparisons on larger frames of discernment and to address several open questions that have emerged in this interesting area of research.

References

R. Bagnara, P. M. Hill, and E. Zaffanella. The parma polyhedra library: Toward a complete set of numerical abstractions for the analysis and verification of hardware and software systems. *Science of Computer Programming*, 72(1-2):3–21, 2008.

F. Cuzzolin. The geometry of consonant belief functions: simplicial complexes of necessity measures. *Fuzzy Sets and Systems*, 161(10):1459–1479, 2010.

F. Cuzzolin. *The geometry of uncertainty: the geometry of imprecise probabilities*. Springer Nature, 2020.

M. Daniel, R. Jiroušek, and V. Kratochvíl. How Sir Harold Jeffreys Would Create a Belief Function Based on Data. In *Proceedings of the 13th Workshop on Uncertainty Processing (WUPES'25)*, pages 92–103, 2025.

D. Dubois and H. Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy sets and systems*, 24(2):161–182, 1987.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976. ISBN 9780691100425.

G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. ISBN 9780387943657. doi: 10.1007/978-1-4613-8431-1.

Surjective Independence of Causal Influences for Local Bayesian Network Structures

Kieran Drury[1], Martine J. Barons[1], and Jim Q. Smith[1]

[1]Department of Statistics, University of Warwick
{*kieran.drury, martine.barons, j.q.smith*}*@warwick.ac.uk*

Abstract

Bayesian network modelling is an established, powerful, descriptive tool for the representation of large, uncertain systems. However, the very expressiveness of Bayesian networks can introduce fresh challenges due to the large number of relationships they often model. It is thus essential to supplement any available data with elicited expert judgements. This in turn leads to two key challenges: the cognitive burden of these judgement is often very high, and there are a very large number of judgements required to obtain a full probability model.

We can mitigate both issues by introducing assumptions such as independence of causal influences (ICI) on the local structures throughout the network, restricting the parameter space of each conditional probability table. However, the use of ICI is often unjustified and overly strong. In this paper, we relax this assumption by partitioning the parents into blocks which themselves independently influence the child, producing the surjective independence of causal influences (SICI) model. We demonstrate that this modification can dramatically ease the burden of any necessary expert judgement elicitation while ensuring faithful belief representations. This further reduces the client resources required to fully construct a model.

1 Introduction

Bayesian network (BN) modelling (see e.g. Korb and Nicholson (2011); Pearl (1988)) has now been a highly established, reliable and intuitive tool among statisticians, computer scientists and AI practitioners for a number of decades. One powerful use of BNs is as a decision support tool - modelling a complex real-world system to test potential policies before implementation by a decision centre (DC) (see e.g. Jensen and Nielsen (2007); Korb and Nicholson (2011); Smith (2010)). This use of BNs has become widespread in the 21st century due to the increased complexity and interconnectedness of the environments in which many decision problems are based.

One fundamental challenge - which we have experienced in many applications - of BN modelling, especially for decision support, is the lack of sufficient data for fully calibrating a model (French et al., 2021; Werner et al., 2017). BNs piece together sub-systems from a variety of distinct domains, leading to a large number of complex relationships with high-order interactions needing to be modelled (see examples in Korb and Nicholson (2011); Barons et al. (2022a)). We therefore often need to rely on expert judgement to parameterise each of these relationships (see e.g. Burgman (2015); French et al. (2021); O'Hagan et al. (2006); Werner et al. (2017)).

Expert judgement elicitation does not, however, come without its own problems. If the system is too complex to be modelled through the available data, an insufficiently managed elicitation process can easily become intractable (Werner et al., 2017). This can happen in several ways - a lack of available, sufficiently knowledgeable experts; a lack of time and money available for eliciting this knowledge; the inability to suppress cognitive biases during the elicitation; and the difficulty of translating experience and knowledge into the required probabilistic assessments (Korb and Nicholson, 2011; Woudenberg et al., 2015; Burgman, 2015; O'Hagan, 2019; O'Hagan et al., 2006). These issues form the so-called "knowledge bottleneck" (Korb and Nicholson, 2011).

Even if the above issues are addressed through careful structuring of the elicitation process, two significant problems persist. The first is simply the number of probabilistic judgements that are required to embellish the whole network (Smith, 2010). The second is that the judgements required from experts are often highly complex due to high-order interactions that are present in BNs, and because the judgements required from experts are usually probabilistic (see Woudenberg et al. (2015) for an example of this in practice). These issues lead to a high elicitation burden for the experts (Werner et al., 2017). They often become fatigued and more susceptible to a number of cognitive biases when this burden is high, threatening to corrupt the judgements they provide even further (Burgman, 2015; Barons et al., 2022b). One way to reduce this elicitation burden is to restrict the model space through applying local structure assumptions across the network, often by assuming particular local causal interaction models (Zagorecki and Druzdzel, 2006).

A popular class of causal interaction models relies on the assumption of *independence of causal influence* (ICI - see e.g. Heckerman (1993); Zhang and Poole (1996)) which assumes that each parent node influences the child node independently. ICI has been utilised within many sub-classes of local BN structure models such as noisy-OR (Pearl, 1988) and its extensions (Díez, 1993; Henrion, 1989; Srinivas, 1993), as well as in CPT interpolation methods such as those reviewed by Mkrtchyan et al. (2016). Despite this, we have found that, while it does simplify the elicitation process, its underlying assumptions are usually too strong and rigid to faithfully represent experts' beliefs in complex systems.

In this paper, we present a simple, practicable methodology for modifying a network structure into a form that better incorporates the ICI assumption, while allowing more freedom for interactions between parents for whom the original ICI model would be too rigid. Like the ICI model, our new modified local network structure model - named the *surjective independence of causal influences* (SICI) model - introduces latent causal mechanisms acting as mediators between the parent set and the child node. Whereas ICI sets a bijection between the parents and these mechanisms, SICI uses a more general

surjective mapping between these sets, allowing the parents to be partitioned into blocks that themselves exhibit ICI. We can therefore modify existing approximate CPT population methods for use in the SICI framework. The SICI model thereby allows quantitative embellishment of a BN to be performed with a significantly reduced burden in a way that is flexible enough to more faithfully model expert beliefs about the real-world system.

The paper is laid out as follows. In Section 2, we review Bayesian networks and explore how they can be elicited through expert judgement. In Section 3, we explore the assumption of independence of causal influences and detail some of its uses. In Section 4, we introduce the surjective independence of causal influences model, detailing its mathematical foundations and giving some practical examples of such models. We explore how this modified network structure more flexibly accommodates the assumption of ICI, enabling efficient yet faithful elicitation of BN models. We finish in Section 5 with a discussion about this new methodology and future research directions.

2 Bayesian Networks and their Elicitation

In this paper, we assume that we require at least some expert judgement to be embedded into the model. Such a BN may be referred to as a Bayesian belief network (BBN). There are two main stages to building such a network. The first is the construction of the network structure - i.e. which variables to include, how to define them, the possible values they have and which variables to draw edges between. This is referred to as the *qualitative stage* of the process, utilising *qualitative* or *soft elicitation* of expert judgements (see Cain (2001); Korb and Nicholson (2011); and Wilkerson and Smith (2021)).

The second stage concerns quantifying the relationships within the network. Modelling discrete BNs, as we assume in this paper, and as is common in practice, involves populating each child node's conditional probability table (CPT). This is the *quantitative stage* of the process, utilising *quantitative elicitation* of expert judgements (see Cain (2001); Korb and Nicholson (2011); and O'Hagan et al. (2006)).

The qualitative elicitation process typically consists of natural language discussions with domain experts in order to understand how *they* picture the structure of the real-world system. In contrast, the quantitative elicitation process can involve a high number of probabilistic judgements about high-order interactions that those not trained in probability struggle to instinctively comprehend. This renders the quantitative elicitation process the most burdensome stage for experts. The number of probabilistic judgements required to populate each CPT is one part of this problem. Consider the child node Y with parent nodes $\mathbf{X} = \{X_1, \ldots, X_n\}$. Let $l_1, \ldots, l_n$ denote the number of possible states for each of the parent nodes, and l_c that for the child. It can easily be checked that there are $(\prod_{i=1}^{n} l_i) \cdot (l_c - 1)$ probabilities to be determined to fully populate the CPT of $Y|\mathbf{X}$. The number of probabilistic judgements required across the network can quickly become enormous. This, combined with the high cognitive burden that probabilistic judgements and high-order interactions bring, makes direct quantitative elicitation often intractable.

The question of how we can reduce the elicitation burden faced by experts is therefore of high importance. Methods for structuring the quantitative elicitation process

such as the Delphi method (Rowe and Wright, 1999), the Sheffield Elicitation Framework (SHELF; Gosling, 2018) and the IDEA protocol (Hanea et al., 2017) mitigate the effects of cognitive biases when providing probabilistic judgements, somewhat reducing the cognitive burden faced by experts. However, these methods do not reduce the number of probabilistic judgements required, nor do they remove the high-order interactions or the probabilistic nature of the required judgements. There is therefore scope for reducing the elicitation burden further than what these methods provide. The main question is how to do this while maintaining faithfulness of the model.

Consequently, methods have been developed for simplifying the structures found within BNs to reduce the number of quantitative assessments required to fully embellish the model. Variables modelled in a BN are often influenced by just a small subset of the other variables in the network through mechanisms that are invariant to variables outside this local structure (Pearl, 2009). The local structure we refer to in this paper simply considers a node and its set of parent nodes. This local structure can be modified without impacting other local structures in the network due to the highly compartmentalised structure a BN exhibits. In this way, these local structures can be simplified through some assumption about how the causal mechanisms operate between a child and its set of parents.

A number of local structure models have been developed that ease the quantitative elicitation process by embedding some such assumption. Many of these models require all nodes to be binary, including noisy-OR (Pearl, 1988) and its extensions (Henrion, 1989; Lemmer and Gossink, 2004; Quintanar-Gago and Nelson, 2021), and the intercausal cancellation model (Woudenberg et al., 2015). A notable example which allows for n-ary nodes is the noisy-MAX model (Díez, 1993; Srinivas, 1993). Further details and extensions of these models can be seen in Díez and Druzdzel (2006). Other methods have been developed to interpolate or otherwise approximate missing CPT values, often just from assessments of the influence of each parent, also reducing the quantity and complexity of the required quantitative judgements. Some such methods are analysed in Mkrtchyan et al. (2016), and some more recent approximate CPT population methods can be seen in Hassall et al. (2019); Phillipson et al. (2021) and Mascaro and Woodberry (2022). Both these CPT approximation methods and the above local structure models restrict the parameter space for each CPT, enabling each CPT to be approximated by a much smaller number of expert judgements than direct elicitation of the original CPT.

These methods each construct CPTs through the influence of each individual parent, or through linear interpolation between elicited rows of the CPT considering one parent change at a time. The modelled information therefore often solely concerns the influence of each individual parent, ignoring any interactions between parents. Some approximation to the conditional probability mass function of the child is used that combines these marginal contributions without embedding any interaction terms. Such an approximation would be valid if the influence of each parent on the child node was independent of the values taken by the other parents - an assumption known as *independence of causal influences*. In the next section, we explore this key assumption so that we can develop methodology addressing the representation of local network structures that justifies the use of models that utilise this property.

3 Independence of Causal Influences

Though models utilising independence of causal influence (ICI) had been used implicitly before (such as in Pearl, 1988), the concept was first formally introduced under the name 'causal independence' by Heckerman (1993), later renamed 'independence of causal influences' (ICI) by Zhang and Yan (1998). ICI is a local structure assumption that simplifies the embellishment of a BN, tackling the challenges of BN calibration and quantitative elicitation discussed in Section 2. However, we have found that the ICI model is too strong and rigid to faithfully model expert beliefs about many real-world systems.

Consider a BN whose structure has already been elicited, and denote the child node of a local structure by Y. Let its parents be written as $\text{Pa}(Y) = \mathbf{X} = \{X_1, \ldots, X_n\}$. We will assume that no two parents are adjacent for simplicity, though this need not be the case. The initial, unmodified local structure for the child node Y is given in Figure 1.

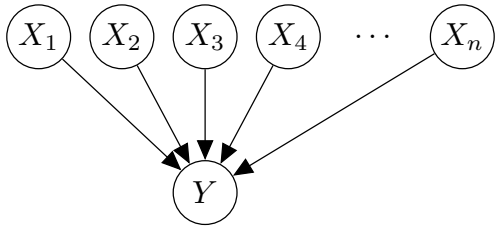


Figure 1: Initial local structure consisting of child node Y and its parent set $\mathbf{X}$

Now suppose we know, through expert judgement or otherwise, that each parent independently influences the value taken by the child, and thus this local structure satisfies ICI. This can be thought of as each parent influencing the child node through its own independent causal mechanism. We can therefore modify our representation of the local structure by introducing a set of mechanisms, one for each parent (see e.g. Heckerman, 1993; Heckerman and Breese, 1996; van Gerven et al., 2008). We denote these mechanisms by $\mathbf{M} = \{M_1, \ldots, M_n\}$ where each mechanism typically has the same set of potential values as Y. The mechanisms explicitly quantify the effect of each parent individually on the child, converting the parent value into a probability mass function over the same states as the child. The mechanisms are combined through the deterministic function f, mapping the outputs of the mechanisms to a value taken by the child. The CPT of $Y|\mathbf{X}$ is calculated through the following definition of an ICI model (van Gerven et al., 2008):

$$p(y|\mathbf{x}) = \sum_{f(\mathbf{m})=y} \prod_{i=1}^{n} p(m_i|x_i). \tag{1}$$

The ICI model structure is shown in Figure 2. ICI has been extended by Zagorecki and Druzdzel (2006) to allow the combination function f to be stochastic, and other extensions and properties of ICI have been explored by Heckerman and Breese (1996).

A key benefit of the ICI model is in the number of parameters required to populate the CPT of the child. The number of parameters in the CPT of $Y|\mathbf{X}$ without the ICI assump-

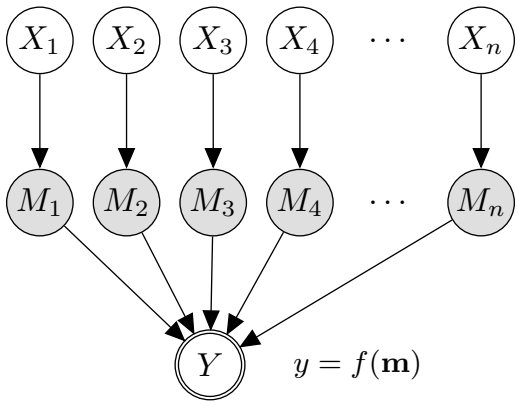


Figure 2: Local ICI model structure

tion grows exponentially in the number of parents, n. When assuming the ICI model, this often reduces to a linear growth (van Gerven et al., 2008), significantly reducing the parameter space and thus the resource requirements for the quantitative elicitation. Further, this structure helps elicit more accurate judgements due to the reduced cognitive burden it brings, and the model becomes more accessible for clients.

Some of the models described in Section 2 can easily be written as ICI models. For example, consider the noisy-OR model (Pearl, 1988) over a set of binary nodes. This model introduces inhibitor nodes that then define each of the mechanisms as presented here; a mechanism M_j takes the value 1 (or 'true' etc.) if the cause is present (i.e. $X_j = 1$) and the inhibitor node is false (taking value 0), else it takes value 0. Then the function f is simply the deterministic OR function; if any of the mechanisms M_j take the value 1, the effect will be present (i.e. $Y = 1$) (Heckerman and Breese, 1996). A similar construction is simple for the noisy-MAX model (Díez, 1993; Srinivas, 1993) which models n-ary variables with the combination function f being the deterministic MAX function (Heckerman and Breese, 1996). Many other models and methods given in Section 2 implicitly assume this structure, assuming that the probability mass function of the child can be approximated through the combination of contributions made individually by each parent.

In the ICI model, an important underlying assumption is that there exists a bijection between the parents and the latent mechanisms. In practice, this would require that each parent in the real-world system affects its child through a mechanism unique to that variable, and that these mechanisms operate independently of each other. We argue that this is an excessively strong assumption, rendering the ICI model too restrictive as it does not allow any interactions between parents' causal mechanisms. In the next section, we present the SICI model which generalises the ICI model, allowing for some low-order interactions between parents.

4 The Surjective ICI Model

The surjective independence of causal influences (SICI) model generalises the ICI model by allowing the mapping between the parent set $\mathbf{X}$ and the mechanism set $\mathbf{M}$ to be a surjection rather than a bijection. This surjective mapping allows the modeller to introduce typically basic, low-level interactions between some parents, usually through a composition of deterministic logical operators that can easily be elicited from experts. A consequence of this is that the number of mechanisms, m, introduced in the SICI model is bounded above by the number of mechanisms introduced in the ICI model - i.e. $m \leq n$. We denote the surjective mapping by $\phi : \mathbf{X} \to \mathbf{M}$. This surjective relationship allows multiple parents to feed into the same mechanism where they combine through the parent-mechanism combination function $f_{(\cdot)}$. The mechanisms then combine in much the same way as they do in the ICI model through the mechanism-child combination function f, though we assume this function to be stochastic to allow for greater flexibility in the CPT parameters. This does not introduce much complexity as CPT approximation methods utilising the ICI assumption (such as those in Section 2) can be applied to approximate the CPT of $Y|\mathbf{M}$ *in a justifiable way*, given that the surjection is constructed to exhibit ICI among the mechanism nodes. The stochasticity of f also allows the functions $f_{(\cdot)}$ to be deterministic, enabling basic interactions to be modelled with ease - though these can also be modelled stochastically within the SICI model.

Of course, the partition that best exhibits ICI across the mechanism nodes may be the partition formed of the singleton parents - in which case $m = n$ and ϕ becomes the bijection seen in the ICI model. Thus it is clear to see that the SICI model is a generalisation of the ICI model. Hence the SICI model can accommodate the ICI assumption through ϕ at least as well as, and often better than, the ICI model.

Through the elicitation of the functions f and $f_{(\cdot)}$, we can simply determine the probability mass functions $p(m_i|\mathbf{x}_{(i)})$, where $\mathbf{x}_{(i)} = \phi^{-1}(m_i)$ denotes the parents of mechanism m_i, as well as the probability mass function $p(y|\mathbf{m})$ for the child node - which may be implicit within the CPT approximation for $Y|\mathbf{X}$. We can then use the following definition to obtain the original CPT of $Y|\mathbf{X}$, analogous to that for the ICI model:

$$p(y|\mathbf{x}) = \sum_{\mathbf{m}} p(y|\mathbf{m})p(\mathbf{m}|\mathbf{x}) = \sum_{\mathbf{m}} \left(p(y|\mathbf{m}) \prod_{i=1}^{m} p(m_i|\mathbf{x}_{(i)}) \right). \tag{2}$$

A graphical representation of the SICI model for a choice of surjection ϕ is shown in Figure 3. Note that we can always ensure the SICI model is planar (i.e. a tree) by ordering the parents and mechanisms such that $\phi(X_1) = M_1$, $\phi(X_n) = M_m$ and:

$$\begin{aligned} &\forall\ i \in \{1, \ldots, n-1\},\ k \in \{1, \ldots, n-i\}, \\ &\qquad \phi(X_i) = \phi(X_{i+k}) \implies \phi(X_i) = \phi(X_{i+1}) = \ldots = \phi(X_{i+k-1}) = \phi(X_{i+k}). \end{aligned} \tag{3}$$

Below we introduce a causal interaction model that is a member of the SICI model family. This is a generalisation of the noisy-OR model (Pearl, 1988) which we name the surjective noisy-OR model. In this model, the parent set is partitioned into blocks that share a common inhibitor variable; parents in the same block feed into a shared

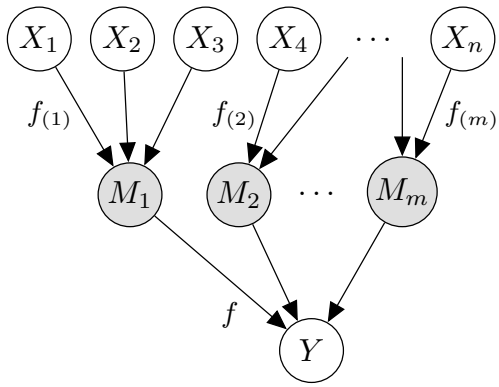


Figure 3: SICI model with $\phi^{-1}(M_1) = \{X_1, X_2, X_3\}$, $\phi(X_4) = M_2$ and $\phi(X_n) = M_m$

mechanism node, say, M_i, just as described above for the general SICI model, but with an additional inhibitor node I_i also feeding into the mechanism M_i for each $i = 1, \ldots, m$. The parent-mechanism combination function for mechanism M_i is now denoted $f_{(i)} \wedge \neg I_i$, where $f_{(i)}$, as before, describes how the parents in the block collectively influence the child through their common causal mechanism. The mechanism node now also depends on this causal mechanism not being inhibited by I_i. The mechanism-child combination function is simply the deterministic OR function over the mechanisms. This model satisfies the definition of the SICI model, and is shown in Figure 4 for a given mapping ϕ on 6 parents. Note that, while the functions $f_{(i)}$ need to be determined, the number of quantitative parameters to be determined has fallen from 6 for the standard noisy-OR model to 3 for the surjective noisy-OR model - in general giving a saving of $n - m$ parameters.

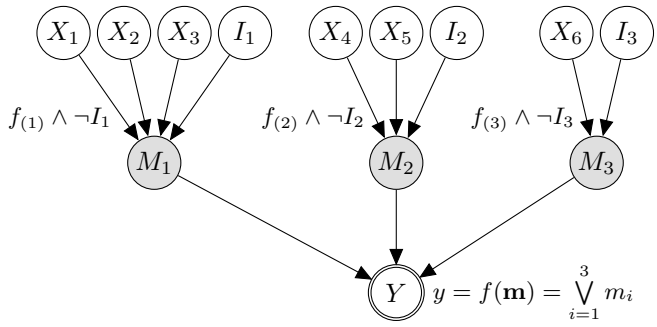


Figure 4: The surjective noisy-OR Model - a member of the SICI model family

Finally, we briefly mention how the SICI model better accommodates many existing CPT approximation methods compared to the ICI model. While many such methods exist, we demonstrate this with Hassall's algorithm (Hassall et al., 2019) due to its simplicity. Hassall's algorithm uses linear interpolation for each parent over its possible states onto

the set $[0, 1]$ in combination with an elicited influence score w_i for the parent to quantify the effect of the parent's state on the child. This requires only n quantitative judgements, a significant reduction compared to direct elicitation of the full CPT.

We can consider Hassall's algorithm as a form of ICI model which allows the mechanism nodes to be defined deterministically and the child node to be defined stochastically. This is, of course, a rather simple form of ICI model, but shares the same essence as the originally defined ICI model. Considering Hassall's algorithm for the case where each node is binary, it can easily be checked that we obtain the following:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}. \tag{4}$$

By reversing the stochasticity in the ICI model, the probability mass function of $Y|\mathbf{X}$ simply becomes that of $Y|\mathbf{M}$. This is what happens for Hassall's algorithm by defining $M_i = w_i X_i$. This is a very restrictive probability mass function in just the same way as the ICI model is restrictive; both embed the same underlying assumption of ICI.

Hassall's algorithm can be made less restrictive and more applicable to many domains by embedding it within the SICI framework. To do this, the parent set is partitioned and the parent-mechanism combination functions are elicited from experts. Then, a weight can be elicited per mechanism, requiring $m \leq n$ such judgements, to reflect the influence score of each mechanism through which subsets of parents combine to impact the child node. Equation 4 would then be amended to sum over the m mechanisms, with the numerator substituting M_i in place of each X_i term. This allows a more faithful approximation of the CPT to be constructed while maintaining a heavily reduced elicitation burden through significant parameter savings compared to direct elicitation.

5 Discussion

In this paper, we introduce the surjective independence of causal influences (SICI) model as a generalisation of the ICI model. The SICI model is easily elicited through natural language conversations with experts and naturally combines with existing CPT approximation methods that facilitate efficient quantitative elicitation of ICI structures - here applied to subsets of the parent set. The SICI model is less restrictive than the ICI model, allowing interactions between parents and easily permitting both types of combination function to be stochastic. The interactions can be embedded through simple compositions of deterministic logical operates, or more complex, possibly stochastic, relationships if necessary. This can be achieved with few, if any, quantitative expert judgements. The CPT for $Y|\mathbf{M}$ in the SICI model can be *justifiably* approximated through existing CPT approximation methods alluded to in Section 2 through the embedding of the ICI assumption in the partition of the parent set - as demonstrated in Section 4 with the surjective noisy-OR model and Hassall's algorithm. This reduces the quantity and complexity of the required expert judgements by introducing significant parameter savings to the local structure. The SICI model does introduce complexity to the qualitative stage of the

elicitation, though this is far less burdensome for experts than a complex quantitative elicitation process. We therefore argue that this shift in complexity comes with a net benefit. As a result, the SICI methodology facilitates quicker achievement of the modelling objectives while significantly reducing client resource requirements. In addition to the ability to model interactions, the SICI model can handle large parent sets more efficiently than the ICI model, eliminating the practical need to compromise on small parent sets. These factors ensure the elicited model is more faithful than what is possible through the use of the ICI model, giving the client crucial faith in the model outputs.

The underlying assumption of ICI is not one that we have found to be appropriate in many practical modelling domains. The SICI methodology satisfies the clear need to generalise the ICI model to weaken this assumption, enabling faithful BN modelling with minimal client resource requirements. Our next steps in the development of the SICI methodology are to create a complete BN model for an existing research application of ours fully utilising the SICI methodology; to compare this model to an existing BN model for the same application that does not utilise SICI; to explore the performance of different subclasses of SICI model (such as one that has no deterministic relationships); and to develop a complete framework for the construction of elicited BN models that fully incorporates this SICI methodology with adapted existing quantitative elicitation methodologies. However, SICI is, of course, not the only method for efficient CPT approximation, and we will continue to explore and report on these methods going forward.

References

M. J. Barons, T. C. O. Fonseca, A. Davis, and J. Q. Smith. A decision support system for addressing food security in the united kingdom. *J. R. Stat. Soc. A*, 185(2):447–470, 2022a.

M. J. Barons, S. Mascaro, and A. M. Hanea. Balancing the elicitation burden and the richness of expert input when quantifying discrete Bayesian networks. *Risk Anal.*, 42 (6):1196–1234, 2022b.

M. A. Burgman. *Trusting Judgements: How to Get the Best out of Experts.* Cambridge University Press, 2015.

J. Cain. *Planning improvements in natural resource management.* UK Centre for Ecology and Hydrology, 2001.

F. J. Díez. Parameter adjustment in Bayes networks. the generalized noisy OR–gate. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 99–105. Morgan Kaufmann, San Matteo, CA, 1993.

F. J. Díez and M. J. Druzdzel. Canonical probabilistic models for knowledge engineering. *UNED, Madrid, Spain, Technical Report CISIAD-06*, 1, 2006.

S. French, A. M. Hanea, T. Bedford, and G. F. Nane. Introduction and overview of structured expert judgement. In *Expert Judgement in Risk and Decision Analysis*, page 1–16. Springer International Publishing, 2021.

J. P. Gosling. SHELF: The Sheffield Elicitation Framework. In L. Dias, A. Morton, and J. Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, page 61–93. Springer International Publishing, 2018.

A. M. Hanea, M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. Investigate Discuss Estimate Aggregate for structured expert judgement. *Int. J. Forecast.*, 33(1):267–279, 2017.

K. L. Hassall, G. Dailey, J. Zawadzka, A. E. Milne, J. A. Harris, R. Corstanje, and A. P. Whitmore. Facilitating the elicitation of beliefs for use in Bayesian belief modelling. *Environ. Model. Softw.*, 122:104539, 2019.

D. Heckerman. Causal independence for knowledge acquisition and inference. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 122—-127. Elsevier, 1993.

D. Heckerman and J. Breese. Causal independence for probability assessment and inference using bayesian networks. *IEEE Trans. Syst. Man. Cybern. A Syst. Hum.*, 26(6): 826–831, 1996.

M. Henrion. Some practical issues in constructing belief networks. In *Uncertainty in Artificial Intelligence 3*, pages 161–173, 1989.

F. V. Jensen and T. Nielsen. *Bayesian networks and decision graphs (2nd edition).* Information science and statistics. Springer, New York, NY, 2007.

K. B. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence.* CRC Press, Boca Raton, FL, 2011.

J. F. Lemmer and D. E. Gossink. Recursive noisy OR–a rule for estimating complex probabilistic interactions. *IEEE Trans. Syst. Man Cybern. B Cybern.*, 34(6):2252–2261, 2004.

S. Mascaro and O. Woodberry. A flexible method for parameterizing ranked nodes in Bayesian networks using Beta distributions. *Risk Anal.*, 42(6):1179–1195, 2022.

L. Mkrtchyan, L. Podofillini, and V. N. Dang. Methods for building conditional probability tables of bayesian belief networks from limited judgment: An evaluation for human reliability application. *Reliab. Eng. Syst. Saf.*, 151:93–112, 2016.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: Eliciting experts' probabilities.* John Wiley & Sons, Chichester, UK, 2006.

A. O'Hagan. Expert knowledge elicitation: Subjective but scientific. *Am. Stat.*, 73:69–81, 2019.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference (2nd Edition)*. Cambridge University Press, New York, NY, Sept. 2009.

F. Phillipson, P. Langenkamp, and R. Wolthuis. Alternative initial probability tables for elicitation of Bayesian belief networks. *Math. Comput. Appl.*, 26(3):54, 2021.

D. A. Quintanar-Gago and P. F. Nelson. The extended recursive noisy OR model: Static and dynamic considerations. *Int. J. Approx. Reason.*, 139:185–200, 2021.

G. Rowe and G. Wright. The Delphi technique as a forecasting tool: issues and analysis. *Int. J. Forecast.*, 15(4):353–375, 1999.

J. Q. Smith. *Bayesian decision analysis: Principles and practice*. Cambridge University Press, Cambridge, UK, 2010.

S. Srinivas. A generalization of the noisy-OR model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 208–215. Morgan Kaufmann, San Matteo, CA, 1993.

M. A. J. van Gerven, P. J. F. Lucas, and T. P. van der Weide. A generic qualitative characterization of independence of causal influence. *Int. J. Approx. Reason.*, 48(1): 214–236, 2008.

C. Werner, T. Bedford, R. M. Cooke, A. M. Hanea, and O. Morales-Nápoles. Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *Eur. J. Oper. Res.*, 258(3):801–819, 2017.

R. L. Wilkerson and J. Q. Smith. Customize structural elicitation. In A. M. Hanea, G. F. Nane, T. Bedford, and S. French, editors, *Expert Judgement in Risk and Decision Analysis*, pages 83—-113. Springer International Publishing, 2021.

S. P. Woudenberg, L. C. van der Gaag, and C. M. Rademaker. An intercausal cancellation model for Bayesian-network engineering. *Int. J. Approx. Reason.*, 63:32–47, 2015.

A. Zagorecki and M. J. Druzdzel. Probabilistic independence of causal influences. In M. Studený and J. Vomlel, editors, *European Workshop on Probabilistic Graphical Models*, 2006.

N. L. Zhang and D. Poole. Exploiting causal independence in bayesian network inference. *J. Artif. Intell. Res.*, 5:301–328, 1996.

N. L. Zhang and L. Yan. Independence of causal influence and clique tree propagation. *Int. J. Approx. Reason.*, 19(3–4):335–349, 1998.

Joint Additive Gaussian Processes for Microbial Species Distribution Modeling

Thomas Heede[1], Abdulkadir Çelikkanat[1], Francesco Delogu[2], Andres R. Masegosa[1], Mads Albertsen[2], and Thomas Dyhre Nielsen[1]

[1]Department of Computer Science, Aalborg University, Denmark
{*thhe,abce,arma,tdn*}*@cs.aau.dk*
[2]Department of Chemistry and Bioscience, Aalborg University, Denmark
{*frde,ma*}*@bio.aau.dk*

Abstract

Microorganisms are fundamental to the functioning of every ecosystem on Earth. Yet, the majority of microbial species remain uncultured and uncharacterized. Expanding our understanding or generating hypotheses about how different factors affect species can help accelerate the discovery of new insights. Species Distribution Modeling (SDM) has traditionally been the primary approach for gaining such insights. However, previous models have often struggled with capturing non-linear responses and have largely focused on environmental predictors. In this paper, we instead explore an additive Gaussian Process (GP) framework to jointly predict species responses to environmental features and spatial effects, while also leveraging model interpretability to enable domain analysis. The model is compared to existing baseline models across several real-world datasets, showing promising results. We demonstrate how the interpretable nature of the model can provide insight into the relationship between environmental features and species community compositions as well as support uncertainty estimation for species response curves.

1 Introduction

The distribution of species over geographical scales is governed by a complex set of factors, and understanding these factors is crucial for the conservation of biodiversity, ecosystem management, and climate change research (Timmis et al., 2017). While Species Distribution Modeling (SDM) (Miller, 2010) has traditionally focused on larger organisms, microbial life also plays a key role in the functioning of ecosystems. However, microbial SDM is challenging due to differences in the observation of occurrences and the physiology of microbes compared to eukaryotes (e.g., plants and animals).

Microbial distributions can be examined as functions of several distinct classes of explanatory features. Environmental features, such as pH and temperature, can be counted

as characteristics that determine the suitable environmental space for a species. The relationship between species and the environment is mediated by the physiology of the species, which can be represented (often not completely) in the model using species-specific traits (Tremlová and Münzbergová, 2007). The coordinates of which a microbe is found can also be used as an explanatory feature, as the spatial distance between sites can influence how likely two sites should be similar. Species can migrate and be dispersed, increasing the chance of detection at sites not necessarily suitable for them (Malard and Guisan, 2023). Other species are also a relevant set of features because species-species interactions (e.g., mutualism or competition) can further constrain the chance of observing a given species. Ultimately, the combination of these feature sets can shed light on the holistic explanation of the biogeographical patterns of microbes (Malard and Guisan, 2023).

A primary goal of SDM is to identify the drivers of species distribution. A continuation of this is to obtain insight into the environmental niche of the species, such as through species response curves, charting the probability of species occurrence across environmental gradients. This requires a model that can capture and reason about the complex and non-linear effects of environmental features, as well as spatial effects. Established SDM frameworks (Tikhonov et al., 2020; Phillips et al., 2006) cannot natively capture such potentially non-linear relationships between the species and the environmental features. For example, were species able to only live in a specific range of a feature, a linear relationship would be unable to capture such a relation. Accurate uncertainty quantification is equally critical in this domain: field observations are often sparse or unevenly distributed, so practitioners must know where predictions are trustworthy and where limited evidence could render management decisions risky. Gaussian processes (GPs) (Williams and Rasmussen, 2006) naturally fulfill this need of functional flexibility and principled uncertainty estimates, making them a natural choice for the present study.

In this paper, we propose an additive GP-based framework for the analysis and interpretation of microbial SDM, integrating environmental and spatial features. We explore the proposed model based on a large real-world dataset (Singleton et al., 2024), comprising more than 15000 species and 2300 geo-located sample sites across Denmark. For evaluating the predictive performance of the model, we compare with three baseline methods over two additional real-world datasets, showing promising results. The implementation of the framework is publicly available on the project's GitHub repository.[1]

2 Related Work

Species Distribution Modeling (SDM) provides a principled way to analyze and predict how species are arranged across space and time, by relating their occurrences to environmental features (Elith and Leathwick, 2009; Araújo et al., 2019). SDM started with the use of percentile-based envelope methods (Nix et al., 1986) and then Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) (James et al., 2013).Modern SDM is based on machine-learning strategies such as maximum-entropy modeling (Phillips et al., 2006) and ensemble models (Araújo and New, 2007). These approaches

[1] `github.com/MicrobialDarkMatter/Biogeography`

established a solid baseline, yet they typically ignored two ecological realities: (i) organisms do not stay in the same place – they disperse and migrate – so purely environmental niches seldom tell the full story, and (ii) many relationships between species and their environment are non-linear (Valavi et al., 2022).

Recent work, therefore, augments environmental predictors with an explicit spatial term. Adding coordinates, distance-based kernels, or latent spatial effects helps account for detections in sites that appear environmentally unsuitable but are reachable through dispersal (Malard and Guisan, 2023). A relevant recent example is the Spatial Implicit Neural Representation of Cole et al. (2023), which learns a continuous field over geography and can fill large data gaps seamlessly.

A widely adopted probabilistic framework that unifies environmental, spatial, and species-specific trait information is HMSC (Tikhonov et al., 2020). HMSC embeds linear environmental effects, spatial information, and species-specific traits into one Bayesian hierarchical model, giving ecologists a coherent toolbox for joint species analysis. Unfortunately, its linear environmental assumption is restrictive: true ecological responses often peak, plateau, or threshold.

Gaussian Processes naturally overcome this limitation. They offer a non-parametric, uncertainty-aware way to capture smooth yet flexible responses (Williams and Rasmussen, 2006). Additive GP models can tie spatial proximity and environmental similarity together so that two sites co-vary only when both their environment and locations are alike (Vanhatalo et al., 2020). Multi-output GPs (MOGPs) generalize to whole communities, modeling several species jointly and thus borrowing strength across taxa while still allowing non-linear effects (Alvarez and Lawrence, 2011). In head-to-head evaluations with HMSC and other baselines, MOGPs yield superior predictive accuracy whenever environmental responses deviate from linearity, though current implementations usually omit traits or spatial components (Ingram et al., 2020).

Interpretability is another essential feature: practitioners want to translate complex models into ecological insight. Species-specific response curves – the probability of presence as each covariate varies – are a popular diagnostic (Hurford et al., 2019; Bazzichetto et al., 2023). Early Bayesian curve models that imposed Gaussian shapes (Schurr et al., 2012) or hierarchical logistic regressions (Jansen and Oksanen, 2013) improved flexibility, but either enforced symmetry or failed to quantify posterior uncertainty fully. Modern GP-based curves inherit both flexibility and calibrated uncertainties, partly resolving these issues.

Finally, while SDM methods were designed for macro-organisms, they are increasingly applied to microbial communities (Delgado-Baquerizo et al., 2018; Mod et al., 2021). Microbial SDMs must cope with compositional sequencing data, extreme sparsity, and blurred taxonomic resolution; nonetheless, linking environmental features to microbial biogeography promises fresh ecological insights (Barberán et al., 2014).

Together, these developments motivate the use of integrative, non-linear, and interpretable models – such as the one we explore in this work – that marry environmental and spatial information in a single coherent open-box framework.

3 Methodology

To model and understand the factors shaping species distributions, we require a structured representation of ecological data that captures both environmental conditions and spatial characteristics. In microbial ecology, this includes not only the presence or absence of species across locations but also detailed descriptions of site-specific environmental features and spatial coordinates that may influence ecological responses. Below we formalize this setup and introduce the notation used throughout the paper.

We work with presence-absence data for J species surveyed at I sampling sites. Each site i provides three kinds of information: (i) **Occurrence** – a binary matrix $\mathbf{Y} \in \{0,1\}^{I\times J}$, $Y_{ij} = 1$ if species j is observed at site i, 0 otherwise; (ii) **Environment** – an environmental feature matrix $\mathbf{X} \in \mathbb{R}^{I\times E}$, whose E columns record variables such as pH, temperature, or soil moisture for each site; (iii) **Spatial** – a coordinate matrix $\mathbf{S} \in \mathbb{R}^{I\times 2}$, giving longitude and latitude (in radians) for every site.

These three matrices – $\mathbf{Y}$ (occurrence), $\mathbf{X}$ (environment), and $\mathbf{S}$ (spatial) – constitute the full input to the species-distribution model analyzed in the remainder of the paper.

3.1 Multi-output Gaussian Process

We assume that at each site i, the occurrence of species j follows a Bernoulli distribution:

$$Y_{ij} \sim Bern(\sigma(\eta_{ij})), \tag{1}$$

where $\sigma(\cdot)$ is the logistic function. Inspired by Ingram et al. (2020), we employ a multi-output Gaussian process (GP) model, which assumes that the response function of each species i (or, more precisely, η_{ij}) is given by a linear combination of n_l non-linear functions that each follow a GP model. These functions are defined over the environmental features and can thus be interpreted as latent function representations of these features. The latent functions are combined with species-specific weights to capture the environmental responses of the individual species.

$$\eta_{(i,j)} := \sum_{l=1}^{n_l} f_{(i,l)} w_{(l,j)} + b_j, \tag{2}$$

where $\mathbf{f}_{(:,l)} \sim \mathcal{GP}\left(0, \kappa_{\theta_l}(x, x')\right)$ for each latent feature $l \in \{1, \ldots, n_l\}$. The latent functions are thus shared across species, and species with similar presence/absence patterns will therefore have similar weights. The prior distribution for the weights $w_{(l,j)}$ are defined so that $w_{(l,j)} \sim \mathcal{N}(0,1)$ for each $(l,j) \in [n_l] \times [J]$. Lastly, $b_j \sim N(0,1)$ is a species-specific bias term, independent of the sampling site i.

For the experimental results in Section 4 we employ an RBF kernel defined over environmental feature pairs $\left(\mathbf{X}_{(i,:)}, \mathbf{X}_{(i',:)}\right)$. We use $\theta_l = \{\ell_l, c_l\}$ to denote length scale and coefficient parameters, and we place an Automatic Relevance Determination (ARD) prior (Wipf and Nagarajan, 2007) on each of the length scales to capture feature importance.

For the spatial correlation among the sampling sites, we follow Tikhonov et al. (2020) and augment the model with an additive spatial component:

$$\eta_{(i,j)} := \underbrace{\sum_{l=1}^{n_l} f_{(i,l)} w_{(l,j)}}_{Environment} + \underbrace{\sum_{m=1}^{n_m} g_{(i,m)} v_{(m,j)}}_{Spatial} + b_j. \tag{3}$$

We place a GP prior on the columns of $\mathbf{g} \in \mathbb{R}^{I \times n_m}$: $\mathbf{g}_{(:,m)} \sim \mathcal{GP}\left(0, \kappa_{\alpha_m}(x, x')\right)$ for each latent dimension $m \in [n_m]$ and $v_{(m,j)} \sim \mathcal{N}(0,1)$. In our experiments, we again use a standard RBF kernel $\kappa_{\alpha_m}(\cdot,\cdot)$ with input pairs $\left(\mathbf{S}_{(i,:)}, \mathbf{S}_{(i',:)}\right)$. Due to numerical instability, we transform the coordinates through standard normalization. As each dataset only contains smaller regions of Earth, the discrepancy from calculating the distances in Euclidean space is deemed tolerable.

For scalability and efficient learning and inference, we employ a variational sparse Gaussian process approximation (Hensman et al., 2015). Specifically, we introduce inducing points for both the environmental ($\mathbf{f}_{(:,l)}$) and spatial ($\mathbf{g}_{(:,m)}$) latent processes and adopt a generalized mean-field variational distribution:

$$q(\mathbf{f}, \mathbf{w}, \mathbf{g}, \mathbf{v}, \mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \prod_{l=1}^{n_l} q(\mathbf{f}_{(:,l)}) \prod_{m=1}^{n_m} q(\mathbf{g}_{(:,m)}) \prod_{l,j} q(w_{l,j}) \prod_{m,j} q(v_{m,j}) \prod_{j=1}^{J} q(b_j) q(\boldsymbol{\alpha}) q(\boldsymbol{\beta}).$$

The variational distributions for $\mathbf{f}$ and $\mathbf{g}$ are constructed using sparse GP approximations via learnable inducing points Hensman et al. (2015), and the weights $(\mathbf{w}, \mathbf{v})$ and bias term $(\mathbf{b})$ are assumed to follow Gaussian variational posteriors. We have used a Gamma distribution for the hyperparameters of the kernels, $(\boldsymbol{\theta}, \boldsymbol{\alpha})$, to ensure positivity.

For model learning we optimize the Evidence Lower Bound using stochastic gradient descent, with mini-batching over sites to handle large datasets. For each latent dimension and environmental feature, we learn a separate length scale value ($\ell_{(l,e)}$), which allows us to determine the input relevance similar to Williams and Rasmussen (2006, Section 5.1). The ARD priors on the kernel length scales are jointly optimized to enable data-driven selection of relevant environmental features. This strategy allows us to infer the latent structure while preserving the interpretability of species-environment relationships.

Compared with the HMSC framework Tikhonov et al. (2020), the model presented here do not assume linear environmental effects instead learns fully non-linear response surfaces for each environmental feature while modeling spatial autocorrelation through a separate GP component. Furthermore, whereas the multi-output GP (MOGP) Ingram et al. (2020) jointly shares latent environmental functions across species, it does not include any explicit spatial process, so it integrates geographic and environmental factors. Our formulation retains the cross-species sharing of environmental structure but augments it with an additive spatial GP so that environmental and spatial features can be interrogated independently. Finally, the use of ARD priors and a scalable sparse variational inference scheme allows our model to yield species-specific response curves, feature importance, and calibrated uncertainties for large species communities.

4 Experiments

In this section, we present both quantitative and qualitative experimental results to assess the effectiveness of our proposed additive multi-output Gaussian Process (GP) model for species distribution modeling presented in Section 3.1. Using three real-world datasets with distinct characteristics (described in Section 4.1), we benchmark our model against three established baseline methods— Logistic Regression (LR), MOGP (Ingram et al., 2020), and HMSC (Tikhonov et al., 2020) — across several standard evaluation metrics (Section 4.2). Our experiments are designed to evaluate not only predictive accuracy but also the interpretability of the model. In particular, in Section 4.3, we explore how the inferred latent functions and species-specific weights has the potential to provide valuable insights into domain properties, including the interplay between environmental conditions and species community composition.

4.1 Data Sources

For model evaluation, we consider three different datasets. The Microflora Danica (*MfD*) dataset (Singleton et al., 2024), which motivated this work, originally contains approximately 10,000 sample sites. After preprocessing to retain only those sites with reliable spatial coordinates and no missing data, we are left with 2,337 sample sites for our experiments. The dataset includes nearly 20,000 species, but is filtered down to 15,013 species, as around 5,000 species are not observed in the selected sites. The *Butterfly* dataset (Ovaskainen et al., 2016) is one of the commonly used benchmark datasets in ecology. It was sampled across a grid of 10×10 kilometers in all of Great Britain. The last dataset, *NY*, consists of microbial species in Central Park, New York (NY) (Ramirez et al., 2014)[2]. All environmental variables are standard normalized. Summary details of the datasets used in the experiments can be found in Table 1.

Dataset	#Sites (I)	#Species (J)	#Env (E)	Location
Microbes, MfD	2,337	15,013	105	Denmark
Butterfly	2,609	55	4	Great Britain
Microbes, NY	579	12599	4	Central Park

Table 1: Characteristics of the datasets used in the experiments. Each sample site across all three datasets includes latitude and longitude coordinates.

The three datasets exhibit significantly different characteristics. For instance, the two microbial datasets (that is, *MfD* and *NY*) are considerably more sparse than the *Butterfly* dataset, as illustrated in Figure 1: 16% of the species in the *MfD* dataset occurs less than 3 times and 41% of the species occur between 3 and 23 times out of the 2,337 samples. A similar trend is also visible for the *NY* dataset, except that no species are observed in

[2]The NY dataset is available from `microbeatlas.org`

less than 0.1% of the samples. In contrast to microbial datasets, the *Butterfly* dataset is less sparse, with 40% of the species occurring in more than half of the samples.

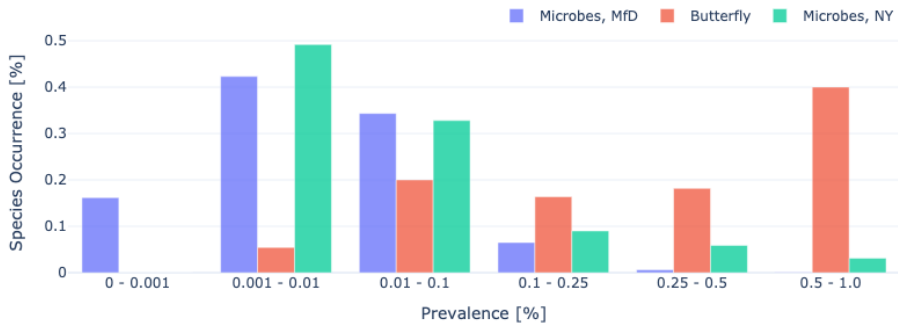


Figure 1: Percentage of species for different prevalence frequencies overall sample sites. For example, for the MfD dataset, approximately 16% of the species occur in at most 0.1% of all sample sites.

4.2 Quantitative results

To evaluate the performance of the proposed model, we compare with three baseline models: Logistic Regression (LR) (Hosmer Jr et al., 2013), MOGP (Ingram et al., 2020), and HMSC (Tikhonov et al., 2020) across four metrics: Receiver Operating Characteristic Area Under the Curve (ROC AUC), Negative Log Likelihood (NLL), Mean Absolute Error (MAE), and Precision/Recall Area Under the Curve (PR AUC). For the analysis, the datasets are split such that 80% of the data is used for training and 20% is used for testing. Species occurring in less than 10 samples are removed. The results are averaged over five runs, but the standard deviations were found to be negligible.

The models are trained for 200 epochs, until convergence with a learning rate of 0.01 optimizing the ELBO. For the results reported in this section, both of the proposed models have $n_l = 10$ latent functions, f, and use 200 inducing points for the environmental GP. The spatial model variant includes $n_m = 5$ latent functions, g, and uses 200 inducing points for the spatial GP. Where applicable, MOGP and HMSC utilize the same set of parameters.

The hyperparameters were chosen based on preliminary experiments into model sensitivity w.r.t. changes in latent features and inducing points. We observed that the number of latent features (n_l) for the environmental GP is the most significant hyperparameter. For its low values, the model is unable to capture the complexities of environmental features.

The results of the evaluation is shown in Table 2. It is worth noting that for LR, a separate model is trained for each individual species, whereas the other methods learn across all species simultaneously. All models are able to be run with less than 16 GB RAM, with execution times in minutes.

Our proposed model consistently outperforms both MOGP and HMSC across all evaluation metrics and datasets, with the version incorporating spatial features achieving the strongest overall performance. These results underscore the benefits of modeling nonlinear relationships between species and environmental variables—something HMSC does not support—as well as the value of jointly incorporating environmental and spatial information -in opposite to MOGP.

Interestingly, the relatively simple model logistic regression (LR) occasionally matches or even slightly outperforms more sophisticated approaches, particularly in the NY dataset. This is a well-known phenomenon in machine learning, where simpler models can perform well in low-signal or low-complexity settings. Nonetheless, in such cases, the performance gap between LR and our method remains small, further emphasizing the robustness and adaptability of our proposed framework.

		Baselines			**Ours**	
		LR	MOGP	HMSC	E	E + S
MfD	*ROC AUC* ↑	0.847	0.820	0.802	<u>0.877</u>	**0.881**
	PR AUC ↑	0.354	0.310	0.299	<u>0.395</u>	**0.408**
	NLL ↓	**0.197**	0.260	0.243	0.235	<u>0.231</u>
	MAE ↓	0.101	0.109	0.121	**0.094**	**0.094**
Butterfly	*ROC AUC* ↑	<u>0.866</u>	0.854	0.844	0.865	**0.905**
	PR AUC ↑	<u>0.722</u>	0.709	0.686	0.721	**0.781**
	NLL ↓	**0.296**	0.341	0.315	0.347	<u>0.312</u>
	MAE ↓	0.186	0.185	0.205	<u>0.177</u>	**0.149**
NY	*ROC AUC* ↑	**0.654**	0.568	0.619	<u>0.643</u>	0.642
	PR AUC ↑	**0.403**	0.348	0.386	<u>0.402</u>	<u>0.402</u>
	NLL ↓	**0.484**	0.512	0.513	0.523	<u>0.522</u>
	MAE ↓	0.310	0.340	0.361	**0.305**	<u>0.306</u>

Table 2: Model performance across datasets, averaged over five runs. The best results are shown in bold, second-best results are underlined.

4.3 Qualitative results

This section offers a qualitative analysis of the learned model, using the MfD dataset to showcase how this model has the potential to support deeper ecological insight. Rather than providing an exhaustive exploration, our goal is to illustrate the model's potential for uncovering meaningful patterns and guiding further investigation in this domain.

Response curves with uncertainty

Response curves translate a fitted SDM model into an intuitive ecological picture: they show how the predicted probability of encountering a species changes as a single environ-

mental feature is swept across its observed range, while all other features remain at the values actually recorded. Such curves let an ecologist read off optima, tolerance limits, or threshold effects at a glance and provide concrete guidance for management interventions (e.g., "keep soil pH below 6.2 to deter *S. aureus*").

Because our model is fully Bayesian, every response curve comes with a posterior distribution, not just a single line. We draw functions from the posterior predictive to compute a point wise mean curve together with credibility bands that widen where data are sparse or where the species shows highly variable behavior. These quantified uncertainties are crucial: they tell the analyst when an apparent preference is well supported and when it may be an artifact of limited data. Figure 2 illustrates this output, plotting the mean response (solid line) and its 95% credibility interval (shaded) for two species under different pH levels.

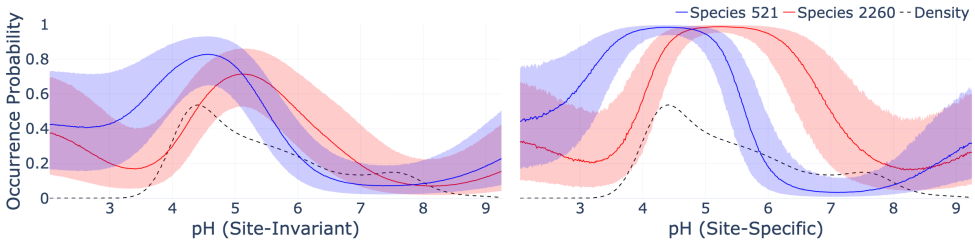


Figure 2: Response curves for pH for two species: the left-hand panel present the marginal curves obtained by averaging across all sites, while the right-hand panel show the site-specific curves for one selected location. Dotted lines trace a kernel-density estimate of the empirical pH distribution in the full dataset, and the shaded ribbons mark 95% credible intervals.

In Figure 2, we observe model prediction of how species 521 and 2260 are responding to changes in pH. On the left, we show the site-invariant response $P(Y|\text{pH})$ and on the right, we show the site-specific response $P(Y|\text{pH}, \mathcal{D}_{(i)})$. Each approach has its own merits. The site-invariant version can shed light on how a species might respond under controlled conditions, such as when cultivating it in a laboratory, by abstracting away site-specific factors. In contrast, the site-specific version captures how changes to a variable, like pH, affect species at a particular location. For example, imagine a farmer who knows the environmental conditions of a field and has access to biochemicals to adjust soil pH. Using the site-specific model, one could predict how those changes would impact the species.

Predicting the occurrence probabilities with the site-invariant approach, species 521 and 2260 are expected to be occurring with their highest probability for pH in ranges 4 to 5 and 4.5 to 5.5, respectively. For the site-specific approach, we can be much more certain of the species occurrence for larger intervals, with near certainty of being present for pH ranges from 4 to 5 and 5 to 6 for species 521 and 2260, respectively. Both approaches show an increase in uncertainty where the kernel density estimate is low. They also converge towards the logits, which prior mean is influenced by the bias term, making them converge towards the marginal probability of observing the individual species.

Clustering of Species

The proposed model in Section 3.1 includes species-specific weights (i.e., $w_{(l,j)}$ in Equation 3) that model the contributions of the shared latent environmental functions to the species occurrence. The weights associated with a particular species can thus be interpreted as a latent representation of the species, which can be used as a basis for species clustering. For illustration, we have performed a k-means clustering (with $k = 5$) of the species based on the mean values of their latent weight representations. The relative number of species belonging to each of the five clusters can be interpreted as an (abstract) representation of a species community. Figure 3 shows the conditional distribution of species for four different types of land use. Encouragingly, we see a strong relationship between the land use type and the species community composition, which supports that idea that the model has the potential to capture meaning biological insights.

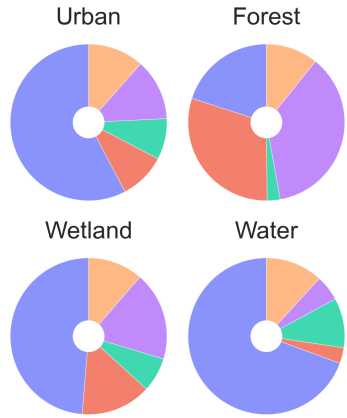


Figure 3: Proportion of species occurrence in different land usages for five species clusters.

5 Conclusion

In this paper, we have explored the use of a modeling framework that relies on additive Gaussian processes to model the distribution of species that, in contrast to previous proposals, combines (non-linear) environmental and spatial features. Beyond predicting species occurrence with higher quality, the open-box nature of the presented approach has the potential to facilitate the analysis of the impact of environmental features (e.g., pH) on the presence of given species – by the use of response curves – while accounting for the inherent uncertainty due to limited and noisy data. Our framework demonstrates superior or comparable performance to state-of-the-art models across multiple datasets. Importantly, we prioritize interpretability over black-box designs, enabling biologists to derive meaningful insights from the model's outputs.

Despite its strengths, the proposed architecture has various limitations. The current framework assumes presence-absence observations, while microbial species datasets often consist of relative abundances (i.e., counts). Additionally, our model does not yet account for observation noise, species traits, or phylogenetic relationships—factors that could further enhance predictive accuracy and ecological interpretability. Since collecting microbial species data is time-consuming and costly, data, sparsity also poses a challenge.

Our work serves as a foundation for several promising extensions. As a future work, we will incorporate species traits and phylogenetic tree information to better capture genomic and evolutionary influences. Extending the model to handle relative abundance data, and explicitly modeling observation noise and missing data mechanisms to improve robustness are other promising research directions. We will also leverage hyperbolic em-

beddings to efficiently represent phylogenetic structure in lower-dimensional latent spaces, avoiding the need for additional dimensionality reduction operations for interpretability. Furthermore, exploring species interactions and alternative likelihood formulations could broaden the model's applicability.

Acknowledgements

This work was supported by a research grant (VIL50093) from VILLUM FONDEN.

References

M. A. Alvarez and N. D. Lawrence. Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research*, 12:1459–1500, 2011.

M. Araújo et al. Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1):eaat4858, 2019.

M. B. Araújo and M. New. Ensemble forecasting of species distributions. *Trends in ecology & evolution*, 22(1):42–47, 2007.

A. Barberán et al. Why are some microbes more ubiquitous than others? predicting the habitat breadth of soil bacteria. *Ecology letters*, 17(7):794–802, 2014.

M. Bazzichetto et al. Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. *GEB*, 32(10):1717–1729, 2023.

E. Cole et al. Spatial implicit neural representations for global-scale species mapping. In *International conference on machine learning*, pages 6320–6342. PMLR, 2023.

M. Delgado-Baquerizo et al. A global atlas of the dominant bacteria found in soil. *Science*, 359(6373):320–325, 2018.

J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.*, 40(1):677–697, 2009.

J. Hensman et al. Scalable variational gaussian process classification. In *Artificial intelligence and statistics*, pages 351–360. PMLR, 2015.

D. W. Hosmer Jr et al. *Applied logistic regression*. John Wiley & Sons, 2013.

A. Hurford et al. Skewed temperature dependence affects range and abundance in a warming world. *Proceedings of the Royal Society B*, 286(1908):20191157, 2019.

M. Ingram et al. Multi-output gaussian processes for species distribution modelling. *Methods in ecology and evolution*, 11(12):1587–1598, 2020.

G. James et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

F. Jansen and J. Oksanen. How to model species responses along ecological gradients–h uisman–o lff–f resco models revisited. *J. Veg. Sci*, 24(6):1108–1117, 2013.

L. A. Malard and A. Guisan. Into the microbial niche. *TREE*, 38(10):936–945, 2023.

J. Miller. Species distribution modeling. *Geography Compass*, 4(6):490–509, 2010.

H. K. Mod et al. Predicting spatial patterns of soil bacteria under current and future environmental conditions. *The ISME journal*, 15(9):2547–2560, 2021.

H. A. Nix et al. A biogeographic analysis of australian elapid snakes. *Atlas of elapid snakes of Australia*, 7:4–15, 1986.

O. Ovaskainen et al. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, 7(4):428–436, 2016.

S. J. Phillips et al. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

K. S. Ramirez et al. Biogeographic patterns in below-ground diversity in new york city's central park are similar to those observed globally. *Proc. R. Soc. B*, 281(1795), 2014.

F. M. Schurr et al. How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, 39(12):2146–2162, 2012.

C. Singleton et al. Microflora danica: the atlas of danish environmental microbiomes. *bioRxiv*, 2024. doi: 10.1101/2024.06.27.600767.

G. Tikhonov et al. Joint species distribution modelling with the r-package hmsc. *Methods in ecology and evolution*, 11(3):442–447, 2020.

K. Timmis et al. The contribution of microbial biotechnology to sustainable development goals. *Microbial biotechnology*, 10(5):984–987, 2017.

K. Tremlová and Z. Münzbergová. Importance of species traits for species distribution in fragmented landscapes. *Ecology*, 88(4):965–977, 2007.

R. Valavi et al. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological monographs*, 92(1):e01486, 2022.

J. Vanhatalo et al. Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data. *Bayesian Analysis*, 15(2):415–447, 2020.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

D. Wipf and S. Nagarajan. A new view of automatic relevance determination. *NeurIPS*, 20, 2007.

Decision Analysis with a Set of Interval Priority Weight Vectors

Masahiro Inuiguchi[1] and Shigeaki Innan[1]

[1]Graduate School of Engineering Science, The University of Osaka
{*inuiguti,innan*}*@sys.es.osaka-u.ac.jp*

Abstract

The advantage of the interval priority weight estimation from a crisp pairwise comparison matrix has been shown over the crisp priority weight estimation when the DM's evaluation is assumed to be vague. However, the usefulness of the estimated interval priority weights in the decision analysis has not yet been shown. In this paper, we show the decision analysis based on the maximin, maximax, and minimax regret criteria under interval priority weights. A deeper decision analysis using interval priority weights is demonstrated in three examples.

1 Introduction

Analytic Hierarchy Process (AHP) (Saaty, 1980; Saaty and Vargas, 2012) is a structural and analytic approach to multiple criteria decision problems. The pairwise comparison matrix (PCM) given by the decision maker (DM) is usually inconsistent as human evaluation is not precise. The consistency index is defined, and when it is in the acceptable range, the priority weights are estimated by minimizing the errors, where the inconsistency is regarded as an error. On the other hand, considering that the inconsistency of the PCM comes from the vague evaluation, the interval AHP (Sugihara and Tanaka, 2001) was proposed. In this approach, the interval priority weights instead of crisp ones are estimated. Because the original method estimates too narrow interval priority weights, various estimation methods of interval priority weights are proposed (Innan and Inuiguchi, 2024). It is shown that the accuracy scores in the problems of ordering alternatives using the estimated several interval priority weights are better than those by the crisp priority weights estimated by eigenvalue and geometric mean methods of the classical AHP (see Inuiguchi et al. (2022, 2025)). However, the usefulness of the interval priority weights in the decision analysis has not yet been studied considerably. The estimated interval priority weights reflect the degree of inconsistency of the given PCM in their widths, although the crisp priority weights of the classical AHP have no inconsistency. Therefore, a more detailed analysis is expected by considering the vagueness of the priority weights.

In this paper, using the best-performed estimation method (Inuiguchi et al., 2025) of interval priority weights in accuracy scores in the previous numerical experiments, the usefulness of interval priority weights is demonstrated by didactic numerical examples. We consider examples where sufficiently consistent PCMs are given. Among many conceivable decision analyses using interval priority weights, we apply a few famous decision rules under uncertainty. In the analysis, the nonuniqueness (Inuiguchi, 2016) of the solution to the estimation problem of interval priority weights is fully introduced. More concretely, the proposed decision analysis uses a set of solutions obtained from the standard solution satisfying center normalization.

This paper is organized as follows. Next section describes the estimation method of interval priority weights used in this paper. It is the best performed method. Decision analyses under interval priority weights are described in Section 3. In Section 4, the decision problem treated in this paper is explained. Then three examples are given to demonstrate the differences of the analyses between the classical and Interval AHP.

2 Estimation Method of Interval Priority Weights

The estimation method of interval priority weights used in this paper is explained. It performs the best in numerical experiments so far (Inuiguchi et al., 2025). Given a PCM $A = (a_{ij})$, calculating n preliminary interval priority weight vectors composed of $W_i(k) = [w_i^{\mathrm{L}}(k), w_i^{\mathrm{U}}(k)]$, $i \in N = \{1, 2, ..., n\}$ by the following procedure for each $k \in N$, we obtain an estimated interval priority weight vector composed of $\hat{W}_i = \sum_{k \in N} W_(k)/n$, $i \in N$, where the PCM $A = (a_{ij})$ satisfies $a_{ij} = 1/a_{ji} > 0$, $i < j$, $i, j \in N$ and $a_{ii} = 1$, $i \in N$ and we impose the constraints, $a_{ij} \in W_i(k)/W_j(k)$ (the reproducibility of a_{ij}), $\sum_{i \in N \setminus j} w_i^{\mathrm{L}} + w_j^{\mathrm{U}} \leq 1$, $\sum_{i \in N \setminus j} w_i^{\mathrm{U}} + w_j^{\mathrm{L}} \geq 1$, $j \in N$ (the normality condition of $W_i(k)$, $i \in N$), and $\sum_{i \in N}(w_i^{\mathrm{L}} + w_i^{\mathrm{U}}) = 2$ (center normalization of $W_i(k)$, $i \in N$).

⟨1⟩ Estimate $w_i(k)$, $i \in N \setminus k$ by the EV method of the classical AHP, where $N \setminus k = N \setminus \{k\}$. The ratios of the centers of $W_i(k)$, $i \in N \setminus k$ are fixed by the raios between $w_i(k)$, $i \in N \setminus k$.

⟨2⟩ Minimize the sum of widths of $W_i(k)$, $i \in N \setminus k$ and the width of $W_k(k)$, lexicographically, subject to the constraints of the reproducibility of a_{ij}, the normality condition of $W_i(k)$, $i \in N$, the center normalization of $W_i(k)$, $i \in N$, and the equality of the ratios of the centers of $W_i(k)$, $i \in N \setminus k$ to the ratios between $w_i(k)$, $i \in N \setminus k$ obtained in ⟨1⟩.

Given $\hat{W}_i$, $i \in N$, we obtain the minimum and maximum solutions by t^{L} and t^{U}

defined by

$$t^{\mathrm{L}} = \frac{1}{\min\limits_{i\in N}\left(w_i^{\mathrm{L}} + \sum\limits_{i\in N\setminus j} w_j^{\mathrm{U}}\right)}, \tag{1}$$

$$t^{\mathrm{U}} = \frac{1}{\max\limits_{i\in N}\left(w_i^{\mathrm{U}} + \sum\limits_{i\in N\setminus j} w_j^{\mathrm{L}}\right)}. \tag{2}$$

Then, the following set of interval priority weight vectors is obtained as the solution set of the estimation problem:

$$\mathcal{W} = \left\{ t\hat{W}_i,\ i \in N \;\middle|\; t \in [t^{\mathrm{L}}, t^{\mathrm{U}}] \right\}. \tag{3}$$

We note that the consistency of the PCM $A = (a_{ij})$ is evaluated by a consistency index C.I. and a consistency ratio C.R. (Saaty and Vargas, 2012) defined by

$$\text{C.I.} = \frac{\lambda_{\max} - n}{n-1}, \quad \text{C.R.} = \frac{\text{C.I.}}{\text{R.I.}}, \tag{4}$$

where $\lambda_{\max}$ is the largest eigenvalue of A and R.I. (Saaty and Vargas, 2012) is the randomness index defined by the average of the resulting consistency index C.I. depending on the order of the matrix. It is known that we may accept the priority weights obtained from PCM A when its C.I. or C.R. is not greater than 0.1.

3 Decision Analysis under Interval Priority Weights

We consider a multiple criteria decision making (MCDM) problem with m alternatives $o_k,\ k \in M = \{1, 2, ..., m\}$ whose marginal utility score in the i-th criteria is given as $u_i(o_k)$ $(i \in N)$. Because the criteria weight vector is obtained as a set of interval priority weight vectors, one of the decision rules under uncertainty (French, 1986) is applied to order the alternatives. We assume that the DM evaluates the criteria weights vaguely by a consistent interval PCM but we including the DM cannot know which interval priority weight vector is used for the evaluation of alternatives, or the interval priority weight vector can change by the DM's mood, the situation, and the time. Namely, the multiplier $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$ is not known but assumed to take a value in the range. Therefore, $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$ is treated as a parameter and the various decision rules under uncertainty (French, 1986) are applied under interval priority weights $t\hat{W}_i,\ i \in N$ for the decision analysis. In this paper, we consider the maximin rule, the maximax rule, and the minimax regret rule as the decision rules under uncertainty Inuiguchi et al. (2022).

For each $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, the maximin rule evaluates an alternative o_k by the worst total utility score $\check{U}(o_k|t)$ obtained as the optimal value of the linear programming (LP)

problem,

$$\begin{aligned} \min \ & \check{U}(o_k|t) = \sum_{i\in N} u_i(o_k)w_i \\ \text{sub. to} \ & \sum_{i\in N} w_i = 1,\ tw_i^{\mathrm{L}} \le w_i \le tw_i^{\mathrm{U}},\ i \in N, \end{aligned} \tag{5}$$

and orders alternatives o_k, $k \in N$ in the decreasing order of $\check{U}(o_k|t)$. Namely, the larger $\check{U}(o_k|t)$, the better.

Similarly, for each $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, the maximax rule evaluates an alternative o_k by the best total utility score $\hat{U}(o_k|t)$ obtained as the optimal value of the LP problem,

$$\begin{aligned} \max \ & \hat{U}(o_k|t) = \sum_{i\in N} u_i(o_k)w_i \\ \text{sub. to} \ & \sum_{i\in N} w_i = 1,\ tw_i^{\mathrm{L}} \le w_i \le tw_i^{\mathrm{U}},\ i \in N, \end{aligned} \tag{6}$$

and orders alternatives o_k, $k \in N$ in the decreasing order of $\hat{U}(o_k|t)$. Namely, the larger $\hat{U}(o_k|t)$, the better.

Finally, for each $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, the minimax regret rule evaluates first the worst disadvantage of an alternative o_k over another alternative o_l by the maximum utility difference $d\check{U}(o_k, o_l|t)$ obtained as the optimal value of the LP problem,

$$\begin{aligned} \max \ & d\check{U}(o_k, o_l|t) = \sum_{i\in N} (u_i(o_l) - u_i(o_k))w_i \\ \text{sub. to} \ & \sum_{i\in N} w_i = 1,\ tw_i^{\mathrm{L}} \le w_i \le tw_i^{\mathrm{U}},\ i \in N. \end{aligned} \tag{7}$$

Then the maximum regret $R(o_k|t)$ of an alternative o_k is defined by

$$R(o_k|t) = \max_{l\in N\setminus k} d\check{U}(o_k, o_l|t). \tag{8}$$

Accordingly, this rule orders alternatives o_k, $k \in N$ in the increasing order of $R(o_k|t)$. Namely, the smaller $R(o_k|t)$, the better.

We note that LP problems (5), (6) and (7) are solved simlply by a greedy method. From this fact, we know that $\check{U}(o_k|t)$, $\hat{U}(o_k|t)$ and $R(o_k|t)$ are relatively easily obtained as a piecewise linear function of t.

Consider a utility score obtained by multiplying the marginal utility score by the lower bound of interval priority weight in each criterion. We understand this utility score shows the fundamental score because w_i, $i \in N$ obtained by solving LP problems (5), (6) and (7) are not less than the lower bounds tw_i^{L}, $i \in N$. As the sum of lower bounds of interval priority weights is not greater than 1, the total score is not composed only of fundamental scores but bonus scores. The bonus score is obtained by assigning the bonus weight to criteria and the assignment of the bonus weight is different by the decision rule. In the maximin rule, bonus weights are assigned to the criteria having the bad marginal scores to see the harsh evaluation. In the maximax rule, bonus weights are assigned to

the criteria having good marginal scores to see the lenient evaluation. Finally, in each maximum utility difference evaluation, bonus weights are assigned to criteria having big exceeding marginal scores to see the terrible losses. We note that the bonus weights are bounded by the widths of interval priority weights, i.e., the differences between the upper and lower bounds.

Parameter t controls the ratio of the fundamental score to the total utility score. Let L be the sum of lower bounds of the interval priority weights, i.e., $L = \sum_{i \in N} w_i^{\mathrm{L}}$. For $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, we have $tL < 1$ as far as $\sum_{i \in N} w_i^{\mathrm{L}} < \sum_{i \in N} w_i^{\mathrm{U}}$. For $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, $100tL\%$ is assigned to the fundamental score, and $100(1-tL)\%$ is assigned to the bonus score in the total utility score. Therefore, the fundamental score ratio becomes the minimum when $t = t^{\mathrm{L}}$ and the maximum when $t = t^{\mathrm{U}}$.

4 The Setting of MCDM Problem

Any MCDM problem can be treated by the interval AHP. In this paper, we consider an MCDM problem for ordering alternatives in the setting of the rookie draft in professional baseball (Kinoshita, 2006). This MCDM problem considers five criteria, C_1: 'Physical Fitness', C_2: 'Good Taste', C_3: 'Personality', C_4: 'Circumstance', and C_5: 'Talent'. Using those criteria, three players A, B, and C, which are alternatives o_1, o_2, and o_3, respectively should be ordered. The DM gives the PCM A (Kinoshita, 2006) defined by

$$A = \begin{pmatrix} 1 & 3 & 3 & 4 & 5 \\ \frac{1}{3} & 1 & 1 & 2 & 3 \\ \frac{1}{3} & 1 & 1 & 2 & 3 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 & 3 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix}. \tag{9}$$

The consistency index and ratio are obtained as $C.I. = 0.0279553$ and $C.R. = 0.02496$. Both of them are sufficiently less than 0.1, and thus the preference information expressed by A is considered consistent.

In the classical AHP, the crisp priority weights w_i, $i \in N$ are estimated by the EV method and the GM method (Saaty, 1980; Saaty and Vargas, 2012). In the EV method, the normalized eigenvector $\boldsymbol{w}$ satisfying

$$A\boldsymbol{w} = \lambda_{\max}\boldsymbol{w}, \tag{10}$$

is calculated as the estimated crisp priority weight vector. On the other hand, in the GM method, the i-th component of the crisp priority weight vector $\boldsymbol{w}$ is calculated by

$$w_i = \frac{\left(\prod_{j=1}^{n} a_{ij}\right)^{\frac{1}{n}}}{\sum_{k=1}^{n}\left(\prod_{j=1}^{n} a_{kj}\right)^{\frac{1}{n}}}, \ i \in N. \tag{11}$$

By the EV method, the crisp priority weights w_i, $i \in N$ of the criteria C_i, $i \in N$ are obtained as

$$w_1 = 0.455525,\ w_2 = 0.183092,\ w_3 = 0.183092,\ w_4 = 0.116967,\ w_5 = 0.0613244. \tag{12}$$

Similarly, by the GM method, the crisp priority weights ω_i, $i \in N$ of the criteria C_i, $i \in N$ are obtained as

$$\omega_1 = 0.454727,\ \omega_2 = 0.184885,\ \omega_3 = 0.184885,\ \omega_4 = 0.115159,\ \omega_5 = 0.0603434. \tag{13}$$

The obtained crisp priority weights are similar between the EV and GM methods. We use the crisp priority weights w_i, $i \in N$ obtained by the EV method.

On the other hand, applying the estimation method of interval priority weights described in Section 2, we obtain the following interval priority weights satisfying the center normalization:

$$W_1 = [0.436191, 0.457725],\ W_2 = [0.152574, 0.214312],\ W_3 = [0.152574, 0.214312],\\ W_4 = [0.093944, 0.147591],\ W_5 = [0.033225, 0.097551]. \tag{14}$$

Then, from (1) and (2), we obtain

$$t^{\mathrm{L}} = 0.937062 \quad \text{and} \quad t^{\mathrm{U}} = 1.072002. \tag{15}$$

In what follows, using three MCDM problems with three alternatives ($m = 3$), which are different in sets of three alternatives, we demonstrate the usefulness of the estimated interval priority weights in the decision analysis.

4.1 MCDM Problem I

We consider the MCDM problem where the marginal utility scores $u(o_j)$ of alternatives o_j, $j = 1, 2, 3$ under each criterion C_i, $i \in N$ are given in Table 1. Those marginal utility scores are given in the book (Kinoshita, 2006). The total utility scores $V(o_j)$, $j = 1, 2, 3$ of three alternatives using the crisp priority weights w_i, $i \in N$ are obtained as

$$V(o_1) = 0.223324, \quad V(o_2) = 0.376315, \quad V(o_3) = 0.400361. \tag{16}$$

Then, we obtain the preference order, $o_3 \succ o_2 \succ o_1$ for the DM, where $o_j \succ o_k$ means that the DM prefers o_j to o_k.

On the other hand, using the set of interval priority weight vectors, $\{(tW_1, ..., tW_n)^{\mathrm{T}} \mid t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]\}$, the minimum total utility score $\check{U}(o_k|t)$ and the maximum total utility score

Table 1: The marginal utility scores of three alternatives of MCDM Problem I

	C_1	C_2	C_3	C_4	C_5
o_1	0.105	0.649	0.098	0.2	0.25
o_2	0.637	0.072	0.187	0.2	0.25
o_3	0.258	0.279	0.715	0.6	0.5

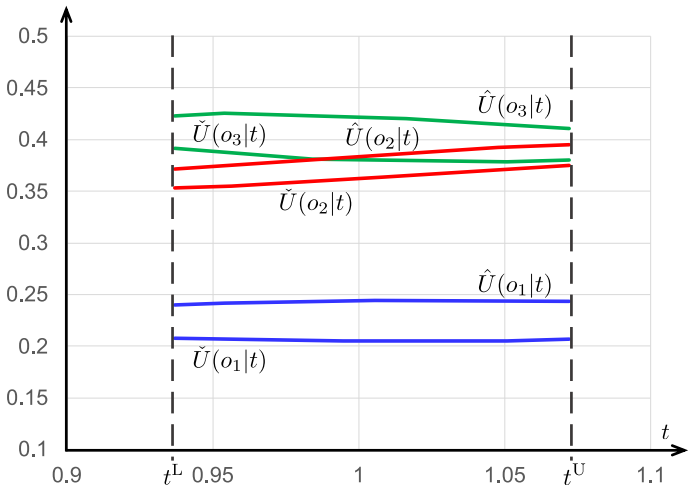


Figure 1: $\check{U}(o_k|t)$ and $\hat{U}(o_k|t)$ in the MCDM problem I

$\hat{U}(o_k|t)$ are obtained as shown in Figure 1. From Figure 1, we confirm $\check{U}(o_3|t) \geq \check{U}(o_2|t) \geq \check{U}(o_1|t)$ and $\hat{U}(o_3|t) \geq \hat{U}(o_2|t) \geq \hat{U}(o_1|t)$. Therefore, we think that the preference order $o_3 \succ o_2 \succ o_1$ obtained by the EV method is reasonable. In this example, without calculating the minimax regret $R(o_k|t)$, we may conclude the preference order $o_3 \succ o_2 \succ o_1$ is adequate. As shown in this example, the decision analysis using the interval priority weights may corroborate the result obtained in the classical AHP.

4.2 MCDM Problem II

We consider the MCDM problem where the marginal utility scores $u(o_j)$ of alternatives o_j, $j = 1, 2, 3$ under each criterion C_i, $i \in N$ are given in Table 2. The total utility scores $V(o_j)$, $j = 1, 2, 3$ of three alternatives of Table 2 using the crisp priority weights w_i, $i \in N$ are obtained as

$$V(o_1) = 0.381441, \quad V(o_2) = 0.384809, \quad V(o_3) = 0.380260. \tag{17}$$

Although the differences are small, the preference order of the three alternatives given in Table 2 is estimated as $o_2 \succ o_1 \succ o_3$.

Table 2: The marginal utility scores of three alternatives of MCDM Problem II

	C_1	C_2	C_3	C_4	C_5
o_1	0.637	0.1	0.187	0.2	0.25
o_2	0.26	0.265	0.78	0.5	0.27
o_3	0.38	0.378	0.378	0.378	0.4

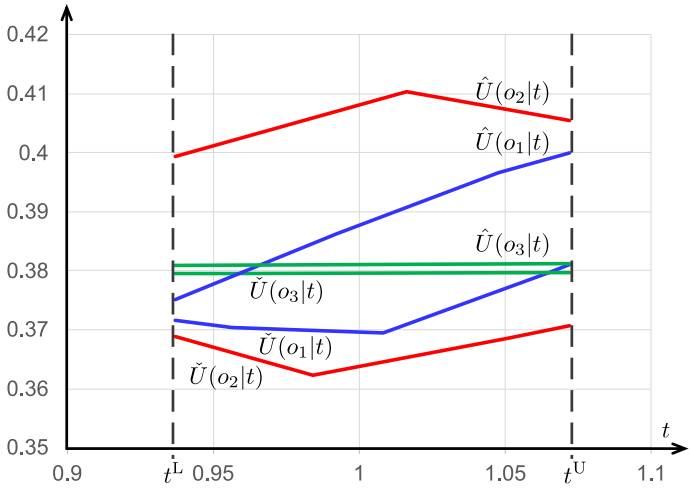


Figure 2: $\check{U}(o_k|t)$ and $\hat{U}(o_k|t)$ in the MCDM problem II

Using the set of interval priority weight vectors, $\{(tW_1, ..., tW_n)^{\mathrm{T}} \mid t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]\}$, the minimum total utility score $\check{U}(o_k|t)$ and the maximum total utility score $\hat{U}(o_k|t)$ are obtained as shown in Figure 2. The scale of the vertical axis in Figure 2 is finer than that in Figure 1. Although the variation ranges of the minimum total utility scores $\check{U}(o_k|t)$, $k = 1, 2, 3$ and the maximum total utility scores $\hat{U}(o_k|t)$, $k = 1, 2, 3$ are small, the variation patterns are different to a certain extent. There are big overlaps among the interval total utility scores of those three alternatives. The alternative o_2 takes the smallest minimum total utility score among three alternatives for every $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$. Nevertheless, the largest difference in minimum total utility scores from the other alternatives is less than 0.02. On the contrary, the maximum total utility score of the alternative o_2 is significantly larger than the others. Therefore, o_2 can be the best alternative. However, if the DM wants to avert strongly from the worst result and accept rather small total utility scores, the alternative o_3 can be recommended. o_1 is completely worse than o_3 when t is small. This fact comes from the fact that o_1 takes the smallest marginal scores in all criteria except C_1.

For $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, we obtain maximum regret $R(o_k|t)$ as shown in Figure 3. From Figure 3, although the alternative o_1 takes the smallest maximum regret for large $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$, it takes the largest maximum regret in a wide range of $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$. Moreover, its largest maximum regret is larger than the other two alternatives and the difference is more than 0.01. Considering those facts, the alternative o_1 is not a good solution. From Figure 3, the alternative o_2 is a good solution from the viewpoint of the maximum regret.

In this example, the results are the same as in the classical AHP in the sense that o_2 is the best. On the other hand, the results are different as the preference order between o_1 and o_2 reverses.

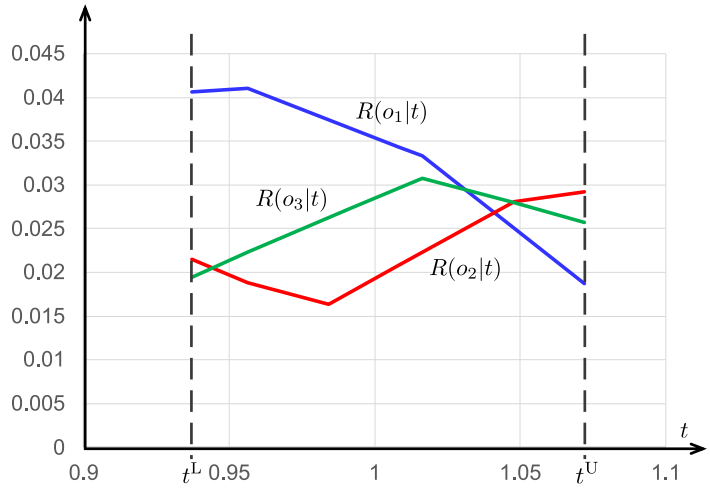


Figure 3: $R(o_k|t)$ in the MCDM problem II

4.3 MCDM Problem III

We consider the MCDM problem where the marginal utility scores $u(o_j)$ of alternatives o_j, $j = 1, 2, 3$ under each criterion C_i, $i \in N$ are given in Table 3. The total utility scores $V(o_k)$, $k = 1, 2, 3$ using crisp priority weights obtained by the EV method are

$$V(o_1) = 0.399709, \quad V(o_2) = 0.399509, \quad V(o_3) = 0.399075. \tag{18}$$

Again the differences among those total utility scores are very small and we obtain the preference order $o_1 \succ o_2 \succ o_3$. As the differences are very small, the DM wonders whether this preference order should be accepted as the final decision.

The minimum total utility score $\check{U}(o_k|t)$ and the maximum total utility score $\hat{U}(o_k|t)$ are obtained as shown in Figure 4 using the set of interval priority weight vectors, $\{(tW_1, ..., tW_n)^{\mathrm{T}} \mid t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]\}$. The scale of the vertical axis in Figure 4 is finer than that in Figure 2, i.e., much finer than that in Figure 1.

As shown in the MCDM problem II, although the variation ranges of the minimum total utility scores $\check{U}(o_k|t)$, $k = 1, 2, 3$ and the maximum total utility scores $\hat{U}(o_k|t)$, $k = 1, 2, 3$ are small, their variation patterns are different. There are big overlaps among the interval total utility scores of those three alternatives, and the minimum and maximum

Table 3: The marginal utility scores of three alternatives of MCDM Problem III

	C_1	C_2	C_3	C_4	C_5
o_1	0.648	0.12	0.17	0.23	0.4
o_2	0.28	0.3	0.69	0.55	0.43
o_3	0.322	0.414	0.414	0.414	0.854

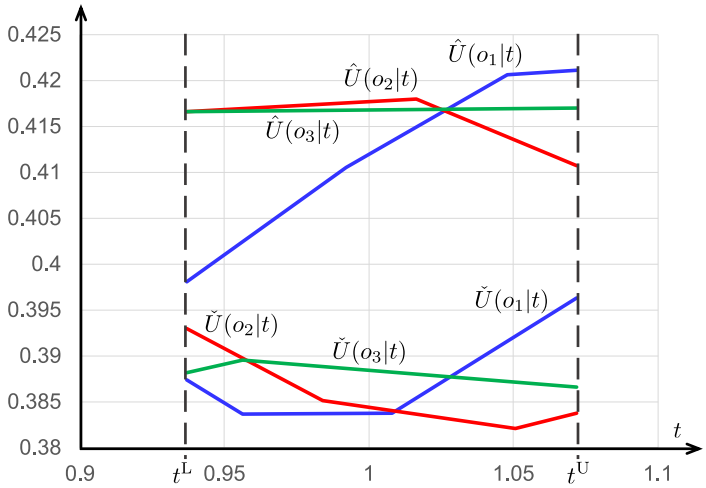


Figure 4: $\check{U}(o_k|t)$ and $\hat{U}(o_k|t)$ in the MCDM problem III

total utilities are close to one another. Generally speaking, it is difficult to tell which is the best from Figure 4. However, if we had to select one alternative, we would take the alternative o_3 because the variations of the minimum and maximum total utility scores $\check{U}(o_3|t)$ and $\hat{U}(o_3|t)$ are comparatively small.

Then we calculate the maximum regret $R(o_k|t)$ for $t \in [t^{\mathrm{L}}, t^{\mathrm{U}}]$ and $k = 1, 2, 3$. We obtain Figure 5. From Figure 5, the alternative o_1 takes the smallest maximum regret when t is large, i.e., when the ratio of the fundamental score is large. When the DM does not decide how large the ratio of the fundamental score is, the alternative o_1 is not always a good solution because the variation of its maximum regret is too large. However, if the DM accepts the difference of maximum regrets at most 0.02, the alternative o_1 would be the best because of its good performance when t is large. When t is small, i.e., when the ratio of the fundamental score is small, the alternative o_3 would be good as its maximum regret is smallest or near smallest. If the decision maker prefers the stable maximum regret, i.e., the maximum regret not greater than 0.03, the alternative o_3 is the best.

In this example, the result of the analysis using interval priority weights is very different from that using priority weights of the classical AHP because o_2 is never recommended while it is the best in the classical AHP.

In the proposed analysis using interval priority weights, the recommended alternative can depend on the DM, more precisely, her/his evaluation attitude and policy, i.e., harsh or lenient, and the importance of the fundamental score.

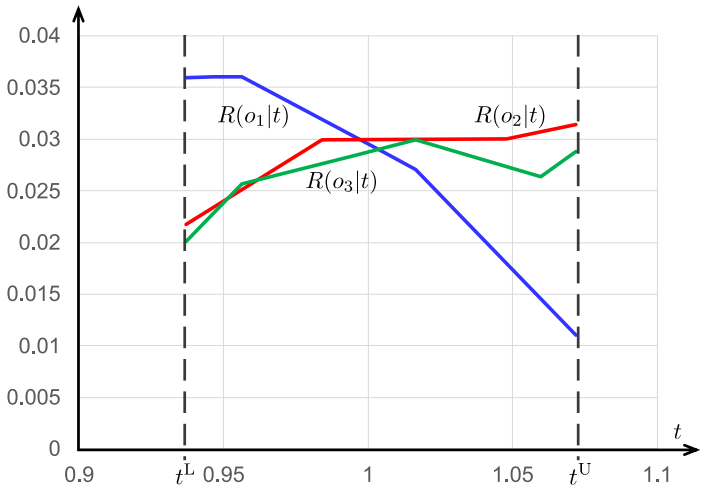


Figure 5: $R(o_k|t)$ in the MCDM problem III

5 Concluding Remarks

We have demonstrated a few possible analyses using interval priority weights under a crisp PCM that is considered consistent. When the crisp total utility scores are sufficiently different in the classical AHP, the result in the interval AHP will not be very different. However, as shown in subsections 4.2 and 4.3, when the crisp total utility scores are close, the result in the interval AHP can be different. In our analysis, the recommended alternative depends on the DM's evaluation attitude and policy, i.e., harsh or lenient, and the importance of the fundamental score.

The vagueness of evaluation estimated from a given crisp PCM is preserved in the interval priority weights. This vagueness makes the analyses richer than the classical AHP. The proposed analysis by the interval AHP does not require the additional preference data from the DM. We obtain the result of the analysis from the same crisp PCM. For the final recommendation, we may ask the DM about her/his evaluation attitude and style.

We do not intend to replace the classical AHP with the interval AHP. We want to recommend the analysts to use the interval AHP together with the classical AHP for getting a second opinion.

In the analysis of the interval AHP, we can know the possible orders of alternatives obtained from the given PCM. Then we may apply the interval AHP to the PCM whose consistency is not sufficient for the primary decision analysis. By this primary analysis, we may find which part of the data in the given PCM would be questionable. The application to the interval AHP to the inconsistent PCM would be one of the future topics.

Acknowledgement

This work is supported by JSPS KAKENHI Grant Number JP23K04272.

References

S. French. *Decision Theory: An Introduction to the Mathematics of Rationality*. Number v. 1-2 in Ellis Horwood series in mathematics and its applications. Ellis Horwood, 1986. URL `https://books.google.co.jp/books?id=swc7xgEACAAJ`.

S. Innan and M. Inuiguchi. Parameter-free interval priority weight estimation methods based on minimum conceivable ranges under a crisp pairwise comparison matrix. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 28(2):333–351, 2024. doi: 10.20965/jaciii.2024.p0333.

M. Inuiguchi. Non-uniqueness of interval weight vector to consistent interval pairwise comparison matrix and logarithmic estimation methods. In V.-N. Huynh, M. Inuiguchi, B. Le, B. N. Le, and T. Denoeux, editors, *Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 39–50, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49046-5. doi: 10.1007/978-3-319-49046-5_4.

M. Inuiguchi, A. Hayashi, and S. Innan. Comparing the ranking accuracies among interval weight estimation methods at the standard, minimum and maximum solutions under crisp pairwise comparison matrices. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–4, 2022. doi: 10.1109/SCISISIS55246.2022.10002032.

M. Inuiguchi, Y. Hong, and S. Innan. Modifying submodels in estimation methods using minimum possible ranges for interval priority weights under a crisp pairwise comparison matrix. In V.-N. Huynh, K. Honda, B. Le, M. Inuiguchi, and H. T. Huynh, editors, *Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 16–28, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-4603-6. doi: 10.1007/978-981-96-4603-6_2.

E. Kinoshita. *Yokuwakaru AHP (in Japanese)*. Ohmsha, 2006.

T. L. Saaty. *The Analytic Hierarchy Process*. McGraw-Hill, New York, 1980.

T. L. Saaty and L. G. Vargas. *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process, Revised 2nd Edition*. Springer, New York, 2012. doi: 10.1007/978-1-4614-3597-6.

K. Sugihara and H. Tanaka. Interval evaluations in the analytic hierarchy process by possibility analysis. *Computational Intelligence*, 17(3):567–579, 2001. doi: 10.1111/0824-7935.00163.

A Genetic Algorithm-Based Heuristic for Large Tram Network Scheduling

Markéta Jirmanová[1] and Martin Plajner[2]

[1]Czech Technical University in Prague, Konviktska 20, Prague, Czechia
jirmamar@fd.cvut.cz
[2]Institute of Information Theory and Automation, Czech Academy of Sciences,
Pod vodárenskou věží 4, Prague 8, 182 08, Czechia
plajner@utia.cas.cz

Abstract

Planning efficient tram schedules for municipal transportation presents significant challenges, often relying on time-consuming manual methods that struggle with network changes and complexity. This paper introduces a genetic algorithm-based heuristic approach to automate and optimize tram timetabling. The heuristic was evaluated on three selected subsets of the Prague tram network with different size. The largest example is a problem potentially computationally infeasible for non-heuristic approaches due to its vast combinatorial space. Results demonstrate that the proposed heuristic significantly accelerates the scheduling process, improves service levels by optimizing vehicle distribution and minimizing wait times, and reduces manual effort, offering a more efficient and adaptable tool for municipal transport planning.

Keywords: Tram Scheduling, Timetabling, Genetic Algorithm, Heuristic Optimization, Public Transport, Urban Mobility, Prague Tram Network

1 Introduction

Efficiently planning tram schedules, specifically departure timetables, for municipal transportation systems represents a complex challenge with far-reaching consequences for commuters, city inhabitants, and overall traffic flow. Current practices frequently involve manual adjustments across multiple spreadsheets, demanding significant expertise from planners to identify viable schedule combinations (Guihaire and Hao (2008)). This methodology is not only time-consuming but also struggles to adapt readily to network alterations, such as those necessitated by construction diversions, placing considerable

strain on planning personnel. Furthermore, suboptimal scheduling can result in diminished service quality – manifested as irregular headways, poor connections, and excessive passenger wait times – and inflated operational expenditures due to inefficient resource utilization (Pietrzak and Pietrzak (2022), Shelat et al. (2022)).

The structure of a timetable profoundly impacts the operational quality of public transport. From the passenger's viewpoint, key decision factors for choosing a transport mode hinge on reliability and travel time. Consequently, the waiting time experienced by passengers is a critical component of mode choice (Shelat et al., 2021a). This encompasses waiting not just for vehicles on a specific route, but for all public transport services operating within the same inter-stop segment. An optimally designed schedule should aim to maximize the even distribution of trams along shared track sections, thereby minimizing passenger wait times and enhancing the overall travel experience. The scheduling process is further complicated by additional considerations, including interconnections between different tram routes, requirements for modal transfers (e.g., with metro or bus lines), constraints imposed by physical infrastructure like tram stop capacity (especially at termini or busy junctions) and street congestion affecting travel time reliability, and the broader demands of urban mobility patterns which fluctuate throughout the day and week (Shi et al., 2017).

Addressing these challenges requires moving beyond traditional manual methods towards more automated and optimized approaches. This paper proposes and evaluates a heuristic optimization technique based on Genetic Algorithms (GAs) specifically tailored for the large-scale tram network scheduling problem. We aim to develop a tool that can generate high-quality, robust, and adaptable tram timetables significantly faster than manual methods, while explicitly optimizing for passenger service quality and operational efficiency.

In the context of Prague, the tram network is one of the densest and most intensively used in Europe, with more than 150 kilometers of lines and almost 30 regular daily lines serving hundreds of thousands of passengers daily. The operational complexity of such a system is significant, especially in the city centre where multiple lines share the same route sections, resulting in track capacity at the limit of operational capacity (pid (2025)). This makes creating the optimal timetable a really challenging task. In this article we use results and observation provided in Chmelová (2025), which describes the Prague tram network in detail, work on basic concepts, and provides exact solutions for simple scheduling cases.

The remainder of this paper is structured as follows: Section 2 provides a review of public transport scheduling current state and optimization techniques. Section 3 formally defines the tram scheduling problem and its constraints as well as details of the proposed Genetic Algorithm heuristic, including its components and objective function formulation. Section 4 describes the experiment with the Prague tram network. In this section the comparison with exact solution is performed first and then a larger case is solved. Finally, Section 5 concludes the paper and suggests avenues for future research.

2 Timetable Scheduling

Public transport scheduling has been an active area of research for decades, falling under the broader category of vehicle routing and scheduling problems. The specific task of timetable generation aims to determine the departure times for all trips of all routes within a given planning horizon.

2.1 Common Approach in Modern City Infrastructure

Public transport timetable scheduling varies across cities, reflecting diverse operational strategies, technological advancements, and policy priorities. While some cities rely on traditional methods, others have adopted innovative approaches to enhance efficiency and passenger satisfaction. For example, in Beijing, the integration of big data analytics into bus scheduling has marked a significant shift from manual methods. By leveraging vast amounts of GPS and smart card data, a data-driven timetable optimization model was developed. This model considers time-dependent running times and passenger demand, aiming to maximize passenger volume while adhering to operational constraints (Zhao et al. (2020)).

In Kaunas, Lithuania, the VIVALDI project introduced the PIKAS software to optimize public transport schedules. This system allowed for the adjustment of timetables based on fluctuating passenger flows throughout the day, improving service quality and coordination among buses, trolleybuses, and microbuses (CIVITAS Initiative (2011)).

Zurich, Switzerland, employs the "Zurich model," emphasizing a dense network with short headways and high priority at intersections. This approach has maintained a high public transport modal share by ensuring reliability and efficiency without extensive underground infrastructure (Moglestue (2005)).

In Copenhagen, the public transport network employs a hybrid approach combining both schedule-based and frequency-based services. The A-bus lines and the Metro operate on high-frequency intervals, allowing passengers to use these services without consulting specific timetables during peak hours. This system enhances flexibility and reduces waiting times. Conversely, services like the S-train and regional buses adhere to fixed schedules, catering to areas with lower demand. This dual system is designed to optimize efficiency and meet diverse passenger needs across the city (Eltved et al. (2018)).

These examples illustrate the spectrum of timetable scheduling practices, from data-driven optimization in Beijing to frequency-based models in Copenhagen, each tailored to local needs and technological capabilities.

2.2 Traditional and Analytical Methods

Historically, timetabling relied heavily on manual methods based on planner experience and simple rules of thumb. While effective in smaller systems, these methods become intractable for large, complex networks. Early analytical approaches often involved network flow models or queuing theory, but these typically required simplifying assumptions

that limited their applicability to real-world scenarios (e.g., in trams scheduling nicely over-viewed in Törnquist (2006)).

Integer programming (IP) formulations have been proposed (e.g., very early article by Ryan and Foster (1981)), capable of finding optimal solutions for smaller problems. However, the combinatorial complexity of timetabling, especially with synchronization constraints and large networks, renders exact IP methods computationally infeasible for practical, large-scale applications like metropolitan tram networks.

To address these challenges, some cities have adopted advanced software solutions. For instance, Transportes Sul do Tejo (TST), a public transport operator in the Lisbon metropolitan area, implemented Optibus' timetable optimization software. This allowed TST to significantly reduce planning time and improve operational efficiency, achieving a 10% reduction in peak vehicle requirements (Optibus (2023)).

2.3 Heuristic and Metaheuristic Approaches

Given the computational difficulty, research has increasingly focused on heuristic and metaheuristic techniques. These methods aim to find near-optimal solutions within reasonable computational time. For example, simulated annealing (Fan and Machemehl, 2006) and tabu search (Krajewska and Kopfer, 2009) have been applied to various public transport scheduling problems. Genetic Algorithms (GAs), inspired by natural evolution, have proven particularly effective for complex combinatorial optimization problems due to their ability to explore large solution spaces effectively. It is used, for example, in Naumov (2020).

Genetic Algorithms (GAs) have been effectively applied to various public transport scheduling problems due to their ability to explore large solution spaces. In railway systems, Arenas et al. Arenas et al. (2014) and Yao et al. Yao et al. (2022) demonstrated their use in periodic timetabling and dense traffic corridors. For bus networks, GAs have been used to optimize schedules while minimizing environmental impacts Zhao et al. (2020) and improving electric bus operations based on demand data Wang et al. (2024). Applications to tram scheduling are less common, but Popescu and Dumitrescu Popescu and Dumitrescu (2021) applied a GA to optimize tram operations in congested areas using AVL data. Our study builds on this foundation, targeting the specific challenges of large-scale tram networks with shared tracks and multi-line coordination.

However, applications specifically focused on the nuances of large-scale *tram* network timetabling, considering factors like shared track sections and detailed passenger waiting time minimization across multiple interacting lines, are less common. This study aims to fill this gap by developing a tailored GA heuristic.

2.4 Factors and Parameters Affecting Scheduling

When constructing a timetable using heuristic methods, such as Genetic Algorithms (GAs), a key component is the design of the objective function. This function guides the optimization process and must reflect the real-world goals of both passengers and operators. However, not all parameters are equally important. At this stage, we distinguish

between core (crucial) factors that must be included from the beginning and others that can be added later to refine the solution.

This prioritization reflects three main considerations. First, we emphasize parameters with a direct impact on perceived service quality, such as regularity and waiting time, which are among the most influential factors in mode choice and user satisfaction van Oort and van Nes (2011); Shelat et al. (2021b). Second, we consider operational feasibility—certain infrastructure constraints (e.g., platform capacity or turnaround times) are indispensable for schedule validity. Third, we account for the increasing modeling complexity: while energy efficiency or workload balancing are desirable, they require richer data and complex formulations, making them more appropriate for later stages Cats and Jenelius (2014).

Crucial parameters:

- *Waiting time for passengers:* Especially on shared track segments, overall waiting time is minimized when vehicle departures are evenly distributed (Shelat et al., 2021b).
- *Headway regularity:* Avoiding bunching and ensuring uniform vehicle spacing improves both service reliability and perceived frequency (Ceder, 2007).
- *Transfer coordination:* Aligning departure times at interchange points can significantly reduce total travel time and improve network accessibility (van Oort and van Nes, 2011).
- *Infrastructure constraints:* These include stop and depot capacities, available turnback facilities, and signal timing restrictions. If violated, a schedule is physically infeasible.

Secondary parameters:

- *Energy consumption:* Scheduling that avoids simultaneous acceleration of multiple vehicles can reduce power demand spikes, as discussed by Naumov and Dmitriev (2020).
- *Vehicle and driver workload balance:* Even distribution of vehicle usage and crew shifts supports long-term operational sustainability.
- *Robustness to disturbances:* Timetables can be designed to better absorb delays by incorporating slack or buffer times (Cats and Jenelius, 2014).

At this stage, we focus only on the crucial parameters that directly affect user experience and operational feasibility. These serve as the foundation for the first version of our objective function.

3 Problem Formulation

Let the tram network be represented by a set of routes R and a set of stops S. Each route $r \in R$ consists of an ordered sequence of stops. Let each stop in the set S be indexed

using natural indexing. Let the subset S^r be the set of stops visited by a single route. The goal is to determine the departure time $d_{r,k}$ for the k-th trip of each route $r \in R$ from its starting terminal in the planning horizon (e.g., the morning peak period).

3.1 Constraints

Several constraints must be satisfied:

- **Headway Constraints:** Minimum and maximum time intervals between consecutive trips on the same route must be respected to maintain service frequency and prevent vehicle bunching.

$$H_{\min} \leq d_{r,k+1} - d_{r,k} \leq H_{\max}$$

- **Travel Time Constraints:** The time taken to travel between stops is assumed based on operational data. This determines arrival times at downstream stops based on departure times. Individual times of the route r to travel between stops are described by the matrix

$$T = \{t^r_{i,j}; i, j \in \{1, 2, \ldots, |S|\}\},$$

  where $t_{i,j} = \inf$ for non-existent connections.
- **Synchronization Constraints:** Desired time intervals for transfers between specific connecting routes at transfer stops must be maintained.

$$\text{Sync}_{\min} \leq (\text{Arr}_{r',j'} - \text{Dep}_{r,j}) \leq \text{Sync}_{\max}$$

  (where Arr and Dep are arrival/departure times at the transfer stop j for routes r' and r).
- **Stop Capacity Constraints:** The number of trams simultaneously occupying a stop (especially termini or platforms) cannot exceed its physical capacity $c_i, i \in \{1, 2, \ldots, |S|\}$.

3.2 Genetic Algorithm Heuristic

The core decision variables are the discretized departure times ($d_{r,k}$) for each trip (k) of every tram route (r) within the planning period. Each individual in the genetic algorithm is the departures vector, each chromosome is a single tram departure for a single trip. In the GA process we use following mechanisms

- *Mutation* is simply altering a single chromosome to a different departure time. The driving parameter sets probability of a single chromosome mutation.
- *Crossover* randomly mixes departure times of two parents. In case crossover happens, we keep only offspring. The driving parameter sets the probability of crossover occurrence.
- *Elitism* is a driving parameter which sets the number of top scoring individuals which go to the next iteration unchanged.

Objective Function & Prague Specific Settings

The objective function, or fitness function F, evaluates the quality of a proposed timetable. It combines several key performance indicators into a single weighted score allowing planners to prioritize specific goals. The main components considered are:

- **Passenger Wait Time Minimization:** Aims to reduce passenger waiting by evenly distributing trams, especially on shared tracks.
- **Synchronization Quality:** Penalizes poor connections at transfer points to improve network accessibility.
- **Headway Regularity:** Promotes uniform spacing between trams on the same route to enhance service reliability.
- **Constraint Violation Penalties:** Discourages solutions that violate operational constraints like stop capacity.

In the experimental part we work with the tram lines of Prague. In Prague tram lines planning, times are rounded to the nearest minute. It is always assumed that a tram stays at a stop for the minute it arrives. Also, during the individual time horizon, each tram operates with the same interval T^* during the whole horizon. This reduces the problem to find only the first departure time $d_{r,1}$.

In this article we aim only at minimization of passenger wait time while satisfying constraints. The fitness function optimizes intervals between individual trams visiting stops where each stop $s_i \in S$ has its own weight w_i influencing its importance in the network. With $a^i_{q,r}$ as the interval of two routes q, r at the stop i

$$\hat{a}^i_{q,r} = \begin{cases} |((d_{q,1} + t_{s^q_1,i}) \bmod T^*) - ((d_{r,1} + t_{s^r_1,i}) \bmod T^*)|, & i \in S^q \cap S^r \\ 0, & otherwise \end{cases}$$

$$a^i_{q,r} = min(\hat{a}^i_{q,r} \quad , \quad T^* - \hat{a}^i_{q,r})$$

the fitness function is defined as

$$F = - \sum_{s_i \in S} w_i \sum_{\forall q,r \in R} a^i_{q,r}$$

4 Prague Tram Network Case Study

This section details the application and evaluation of the proposed genetic algorithm-based heuristic using the Prague tram network as a real-world case study. Prague tram network is one of Europe's densest and most utilized tram systems, featuring over 150 kilometers of lines and nearly 30 regular daily routes that serve hundreds of thousands of passengers each day. The network's complexity is particularly high in the city center, where multiple lines share track sections, often operating at the limits of track capacity. This shared infrastructure makes optimal timetable coordination exceptionally challenging. The network comprises 26 main tram lines, which translates to 52 directional routes

ID	No. of stops	No. of lines	Solver time [s]
1	12	7	3.51
2	65	7	19.81
3	12	20	75600 (unfinished)

Table 1: Example cases settings

when considering both directions of travel. Practical implication for GA algorithm are discussed in 3.2.

Having this settings, for a single time period like the peak hour in Prague with 52 directional routes and an 8-minute departure window $T^* = 8$, the number of potential combinations is enormous 8^{52}. Considering multiple trips per route and finer time discretization would make the search space astronomically larger.

4.1 Scheduling Task

The goal is to generate a timetable that optimizes the objective function, focusing on minimizing passenger waiting time while respecting operational constraints.

To evaluate the proposed genetic algorithm heuristic, we applied it to three problems for comparison against solutions obtained via exact, non-heuristic optimization methods, providing a benchmark to assess the heuristic's accuracy and efficiency in controlled scenarios. We use three examples described and optimized in Chmelová (2025). These origin by sub-selecting specific lines and stops from the Prague's network and performing computations only for this sub-selection. Settings of three cases is detail in Table 1.

4.2 Simulation & Results

The GA test was performed on a desktop computer running Intel core i5 chip with 48GB RAM. We used different settings of parameters for GA, with each set of parameters 10 runs were performed. As results we show the ratio against the optimal value (exact solution) and the percentage of runs which were able to obtain optimum. We measure the time for a single run. The simulation runs for a defined number of iteration and we observe the iteration with the last improvement of the target value.

For the case of the first example consult Figure 1. This plot shows behaviour based on chosen parameters. It is clearly visible that high levels of mutation degrades solution. Crossover parameter does not provide significant stable changes across different population sizes. We can also observe that higher level of elitism yields better results. Based on this observation we show results for crossover and mutation parameters set to 0.25, elitism to 20 in the remaining experiments.

Figure 4.2 displays scoring and time results for the first case. For each population size the algorithm was able to find an optimal solution in at least one case out of 10. The lowest averaged ratio against optimal value is 0.91. Nevertheless, only the smallest population with few generations did not found optimal solution in any of 10 repetition. For all others we were able to obtain optimum at least once. We can observe linear growth

of time with population size and number of iterations. Observing the last change of the objective value shows that lower number of iterations is possible, although the increase for 200 iterations clearly shows that in some cases even a late iteration provides changes of fitness. The time of exact solution is 3.5s, GA with the smallest settings take 8.4s (10 runs in the set) and 15.1s for the second smallest case. For larger population it takes significantly longer, although for practical application it is still reasonable timing.

Figure 3 shows results for remaining two cases in the same way as described in the paragraph above. In the third case, the exact solution was not reached after 75600 seconds and we compare GA results to the best integer solution found by the optimization method in that time. As a result, we are able to find better solution in some runs. The shortest time for case 2 and 3 is 40s and 11s respectively. In the second case it is still larger than the exact solution. In the third case it is significantly lower and we can say that capabilities of exact solution hit its threshold. The increase in times for GA method is linear in terms of number of lines and stops which makes it usable for even large networks.

Figure 1: default

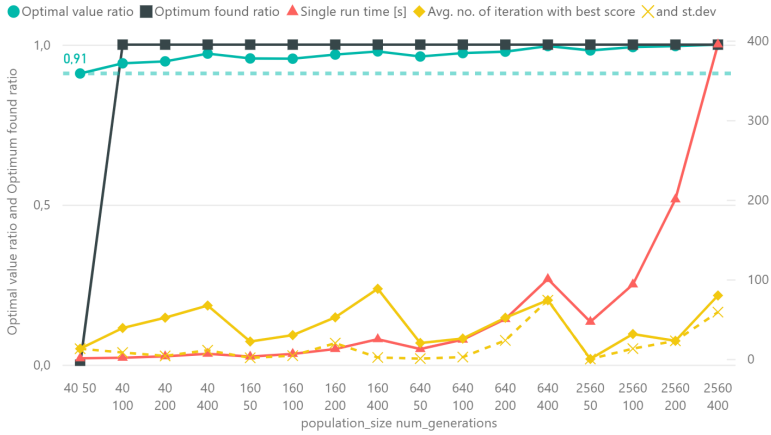


Figure 2: Results of the GA optimizing the first case. Averaged values out of instances with the same settings (10 runs). mutation_prob = 0.5, crossover_prob = 0.5, elitism = 20

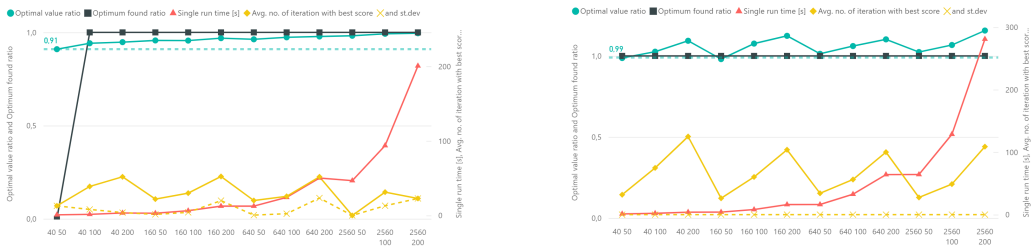


Figure 3: Results of the GA optimizing the second and third case. Averaged values out of instances with the same settings (10 runs). mutation_prob = 0.5, crossover_prob = 0.5, elitism = 20

5 Conclusion and Future Work

The proposed genetic algorithm efficiently provides high-quality, feasible solutions for tram scheduling, as demonstrated on benchmark cases. It effectively distributes tram departures on shared sections to minimize average passenger waiting times, with resulting timetables adhering to crucial operational constraints. Experiments confirm the algorithm achieves near-optimal results in reasonable times.

This work offers a method for an automated tool for transportation planners, speeding up timetable generation and reducing manual work. It improves adaptability to network changes and optimizes schedules based on defined objectives, enhancing passenger service and resource efficiency. Unlike current methods, this heuristic allows planners to weight criteria and optimize for specific goals, providing a practical solution for complex urban tram networks like Prague's.

This paper presents a promising method to provide tram line planners with an effective tool for creating new, more efficient timetables faster. Follow-up research is needed to:

- evaluate different objective functions and different interval options,
- include model changes and interconnections,
- improve the usage of passengers headcount,
- compute more complex test cases, and
- compare theoretical timetables with the real one.

References

Tramvaje, 2025. URL `https://pid.cz/tramvaje/`.

D. Arenas, R. Chevirer, S. Hanafi, and J. Rodriguez. Solving the periodic timetabling problem using a genetic algorithm. *arXiv preprint arXiv:1411.6998*, 2014.

O. Cats and E. Jenelius. Robustness in railway timetable planning: Analytical and simulation methods. *Transportation Research Part B: Methodological*, 68:1–15, 2014.

A. Ceder. *Public transit planning and operation: theory, modeling and practice*. CRC press, 2007.

T. Chmelová. Model for optimizing the planning of prague tram timetables. Master's thesis, Prague University of Economics and Business, Prague, May 2025. Awaiting defense.

CIVITAS Initiative. Optimising public transport timetables. `https://civitas.eu/mobility-solutions/optimising-public-transport-timetables`, 2011. Accessed: 2025-05-04.

M. Eltved, O. A. Nielsen, and T. K. Rasmussen. The influence of frequency on route choice in mixed schedule- and frequency-based public transport systems – the case of the greater copenhagen area. *Case Studies on Transport Policy*, 6(4):507–515, 2018. doi: 10.1016/j.cstp.2018.07.001.

W. Fan and R. Machemehl. Using a simulated annealing algorithm to solve the transit route network design problem. *Journal of Transportation Engineering-asce - J TRANSP ENG-ASCE*, 132, 02 2006. doi: 10.1061/(ASCE)0733-947X(2006)132:2(122).

V. Guihaire and J.-K. Hao. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10):1251–1273, 2008. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2008.03.011.

M. A. Krajewska and H. Kopfer. Transportation planning in freight forwarding companies: Tabu search algorithm for the integrated operational transportation planning problem. *European Journal of Operational Research*, 197(2):741–751, 2009. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2008.06.042.

A. Moglestue. Zürich: A city and its trams. *Tramways & Urban Transit*, 68(812):380–385, 2005.

A. Naumov and A. Dmitriev. Integrated optimization of energy-efficient timetables and vehicle schedules in electric public transport. *Sustainable Cities and Society*, 52:101859, 2020.

V. Naumov. Genetic-based algorithm of the public transport lines synchronization in a transfer node. *Transportation Research Procedia*, 47:315–322, 2020. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2020.03.104.

Optibus. Transportes sul do tejo achieves operational excellence using optibus' timetable optimization, 2023. URL `https://optibus.com/case/transportes-sul-do-tejo-achieves-operational-excellence-using-optibus-timetable-optimization/`. Accessed: 2025-05-04.

K. Pietrzak and O. Pietrzak. Tram system as a challenge for smart and sustainable urban public transport: Effects of applying bi-directional trams. *Energies*, 15(15), 2022. ISSN 1996-1073.

C. Popescu and D. Dumitrescu. The use of the genetic algorithms for optimizing public transport schedules in congested urban areas. *IOP Conf. Ser.: Materials Science and Engineering*, 1037(1):012062, 2021.

D. M. Ryan and B. A. Foster. An integer programming approach to scheduling. *Computer scheduling of public transport urban passenger vehicle and crew scheduling*, pages 269–280, 1981.

S. Shelat, O. Cats, and J. van Lint. Quantifying travellers' evaluation of waiting time uncertainty in public transport networks. *Travel Behaviour and Society*, 25:209–222, 2021a. ISSN 2214-367X. doi: https://doi.org/10.1016/j.tbs.2021.07.009.

S. Shelat, O. Cats, and N. van Oort. Understanding passenger route choice behavior using network-wide smart card data. *Transportation Research Part C: Emerging Technologies*, 124:102977, 2021b.

S. Shelat, O. Cats, N. Oort, and J. Lint. Evaluating the impact of waiting time reliability on route choice using smart card data. *Transportmetrica A: Transport Science*, pages 1–19, 02 2022. doi: 10.1080/23249935.2022.2028929.

J. Shi, Y. Sun, P. Schonfeld, and J. Qi. Joint optimization of tram timetables and signal timing adjustments at intersections. *Transportation Research Part C: Emerging Technologies*, 83:104–119, 2017. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2017.07.014.

J. Törnquist. Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms. In L. G. Kroon and R. H. Möhring, editors, *5th Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS'05)*, volume 2 of *Open Access Series in Informatics (OASIcs)*, pages 1–23, Dagstuhl, Germany, 2006. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-939897-00-2. doi: 10.4230/OASIcs.ATMOS.2005.659.

N. van Oort and R. P. van Nes. Service reliability in transit networks: Impact of synchronizing schedules on waiting times and passenger satisfaction. *Transportation Research Record*, 2216(1):124–132, 2011.

L. Wang, H. Zhang, and Q. Liu. Optimization model for electric bus scheduling based on od data and improved genetic algorithm. In *Proc. of 3rd ICAUS 2023*, pages 162–175. Springer, 2024.

Z. Yao, L. Nie, and Z. He. A genetic algorithm for heterogeneous high-speed railway timetabling with dense traffic. *Journal of Rail Transport Planning & Management*, 23: 100334, 2022.

X. Zhao, Y. Li, Y. Wang, and Y. Zhang. Single bus line timetable optimization with big data: A case study in beijing. *Information Sciences*, 512:282–297, 2020. doi: 10.1016/j.ins.2019.10.042.

Fuzzy Bayesian Networks with Likert Scales

Jan Mrógala[1], Irina Perfilieva[1], and Jiří Vomlel*[1,2]

[1]Institute for Research and Applications of Fuzzy Modeling
University of Ostrava
30. dubna 22, 701 03 Ostrava 1, Czechia

[2]Institute of Information Theory and Automation
Czech Academy of Sciences
Pod Vodárenskou věží 4, 182 00, Prague 8, Czechia
vomlel@utia.cas.cz

Abstract

Our work is motivated by the applications of probabilistic models in the social sciences, in which surveys and questionnaires are typically used to collect respondents' opinions via a Likert scale. The dividing lines between the states on the Likert scale are vague, so it is natural to interpret them using fuzzy numbers instead of integers. We treat the true model variables as hidden continuous variables, the values of which are observed only through their fuzzified counterparts. This approach seems more conceptually appropriate in the context of surveys and questionnaires, since the modeled variables are continuous by nature but are only observed on a fuzzy, discrete scale. Probabilistic inference with continuous variables is challenging when the assumption of normality of the variables' distribution is violated, which is particularly true for variables modeling polarizing issues. We approximate continuous, multidimensional probability distributions using an F-transform composed of basic functions with central points, called nodes, at a multidimensional grid. We illustrate the suggested approach using a small Bayesian network model of data from the survey "Dividing Lines in Czech Society."

1 Introduction

Uncertainty can manifest in multiple ways. One aspect is the likelihood of a variable's state, typically expressed as a probability. Another aspect is the vagueness of a variable's

*Corresponding author

state, which can be represented by a fuzzy set.

Our work is motivated by the application of probabilistic models in the social sciences, where surveys and questionnaires are commonly used to collect respondents' opinions via a Likert scale. Likert scales consist of several states, which are usually represented by a set of consecutive integers. A seven-point Likert scale, for example, has the following states: (1) strongly disagree, (2) disagree, (3) somewhat disagree, (4) neutral, (5) somewhat agree, (6) agree, and (7) strongly agree. The dividing lines between these states are vague, so it is natural to interpret them using fuzzy sets.

Bayesian networks (Pearl, 1988; Jensen and Nielsen, 2007; Koller and Friedman, 2009)) are probabilistic graphical models in which directed acyclic graphs are used to model relations between random variables. Pan and Liu introduced fuzzy Bayesian networks in (Pan and Liu, 2000). Their formalism fuzzifies a Bayesian network by replacing continuous variables with discrete ones. The mapping between the two is approximated by a conditional Gaussian distribution. The second form replaces continuous variables with discrete partners only when necessary and uses conditional Gaussian regression for continuous dependencies.

We treat continuous variables as hidden variables, observing their values only through their fuzzified counterparts. We extend Bayesian networks with additional variables. Each additional variable represents a fuzzy observation of the corresponding unobserved continuous variable. In the context of surveys and questionnaires, this approach seems more conceptually appropriate, since the modeled variables are continuous by nature but are observed on a fuzzy, discrete scale. Unfortunately, this proposal makes inference challenging because it must be performed with continuous variables. Furthermore, the assumption of normality of the variables' distribution is often significantly violated. Quite often, probability distributions have several modes. This is particularly true of polarizing issues. Therefore, we approximate the continuous, multidimensional probability distributions of our fuzzy Bayesian network using the F-transform (**?**)perfilieva-2006, perfilieva-2008), which is composed of basic functions with central points (called nodes) at a multidimensional grid. We illustrate the suggested approach using data from the survey "Dividing Lines in Czech Society", discussed in (Buchtík, 2023).

The paper is organized as follows: In Section 2 we introduce the necessary terminology from the fuzzy set theory and present three types of membership functions used in this paper. These membership functions form the basis of F-transform and its inverse and they are described in Section 3. Finally, in Section 4 the F-transform using Gaussian basic functions is applied to inference in a small Bayesian network. We conclude the paper with a discussion of future research directions.

2 Fuzzy sets and fuzzy partitions

In this paper we will treat the studied variables as continuous random variables that are hidden and their values are observed only through their fuzzified counterparts with their values from a Likert scale[1]. This idea is not entirely new since some authors have

[1]In this paper we consider the Likert scale with seven values, i.e., $n = 7$.

already suggested to identify each Likert response category with a fuzzy set (Gil and González-Rodríguez, 2012).

A fuzzy set is a pair (U, A) where U is a set (often required to be non-empty) and A is a mapping from universe U to $[0, 1]$ referred as the membership function. The fuzzy set is often identified with its membership function. An observation $X = k$ of a random variable X, where k is a value from the Likert scale $\{1, \ldots, n\}$ will be interpreted as a fuzzy set $A_k : [1, n] \to [0, 1]$.

A family of fuzzy sets $\{A_k, k = 1, \ldots, n\}$ where all A_k are defined on the same universe U is called fuzzy partition. In this paper the universe U will be an interval of real numbers $[a, b]$. In Figure 1 we present examples of two fuzzy partitions that we discuss in more detail later in this section. In Perfilieva (2006) a fuzzy partition $A_1, \ldots, A_n$ of $[a, b]$ satisfying so called Ruspini conditions was introduced. Let $x_1 < \ldots < x_n$ be fixed nodes from the interval of real numbers $[a, b]$ such that $x_1 = a$, $x_n = b$ and $n \geq 2$.

Definition 1 (Ruspini partition)**.** We say that the fuzzy sets $A_1, \ldots, A_n$, identified with their membership functions defined on $[a, b]$, establish a Ruspini partition of $[a, b]$ if they fulfill the following conditions for $k = 1, \ldots, n$:

1. $A_k : [a, b] \to [0, 1]$, $A_k(x_k) = 1$;
2. $A_k(x) = 0$ if $x \notin (x_{k-1}, x_{k+1})$, where for uniformity of notation, we set $x_0 = a$ and $x_{n+1} = b$;
3. $A_k(x)$ is continuous;
4. $A_k(x)$, for $k = 2, \ldots, n$, strictly increases on $[x_{k-1}, x_k]$ and $A_k(x)$, for $k = 1, \ldots, n-1$, strictly decreases on $[x_k, x_{k+1}]$;
5. for all $x \in [a, b]$,
$$\sum_{k=1}^{n} A_k(x) = 1.$$

The membership functions $A_1, \ldots, A_n$ are called *basic functions*. A point $x \in [a, b]$ is *covered* by basic function A_k if $A_k(x) > 0$. The shape of the basic functions is not predetermined, so it can be selected based on additional requirements (such as smoothness).

The first basic function considered in this paper is the Gaussian function defined for $x \in [a, b]$ as:

$$A_k(x) \quad = \quad \frac{1}{\sigma_X^k \cdot \sqrt{2\pi}} \exp\left(-\frac{(x - c_X^k)^2}{2 \cdot (\sigma_X^k)^2}\right)$$

where parameters are the values σ_X^k of the standard deviation and the means c_X^k specify centers of the fuzzy sets[2]. In the following experiments, we always positioned the centers

[2]In all our experiments we assume that standard deviations are equal, i.e. $\sigma_X^1 = \ldots = \sigma_X^7$.

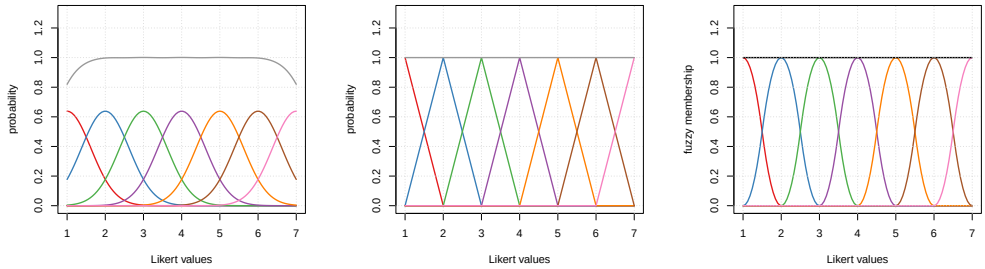


Figure 1: Fuzzy partitions using Gaussian (left), triangular (middle), and Π-curves (right) basic functions.

at certain points of the considered grid[3]. This fuzzy partition does not satisfy the conditions of the Ruspini partition. See the right hand side of Figure 1 where the top black curve corresponds to the sum of values of all seven basic functions.

The second option we consider are triangular functions. Each basic function T_k is specified by its center c_X^k and its width w_X^k for $x \in [a, b]$ as:

$$T_k(x) \quad = \quad \max\left\{0, 1 - \frac{2|x - c_X^k|}{w_X^k}\right\}$$

Another option, commonly used in the fuzzy sets applications are Π-curves. They are named after their shape that for certain parametrizations resemble the letter Π. They are composed from the so called S-functions defined for $x, u, v \in [a, b]$, $u < v$ as:

$$S(x, u, v) \quad = \quad \begin{cases} 0 & \text{if } x \leq u \\ 2\left(\frac{x-u}{v-u}\right)^2 & \text{if } u < x \leq \frac{u+v}{2} \\ 1 - 2\left(\frac{x-v}{v-u}\right)^2 & \text{if } \frac{u+v}{2} < x \leq v \\ 1 & \text{if } v < x \end{cases}$$

In this paper we will use Π-curves defined for $x \in [1, 7]$ as:

$$\Pi_k(x) \quad = \quad S(x, k - s, k) - S(x, k, k + s) \ ,$$

where k are coordinates of the centers of individual basic functions and s is their span.[4]. The advantage of this fuzzy transformation is that for the above parametrization satisfies the Ruspini condition.

[3]The position of centers could be considered as another parameter to be specified with the help of a domain expert or by an optimization algorithm.

[4]In this paper we use $k \in \{1, \ldots, 7\}$ with $s = 1$ or $k \in \{1, 4, 7\}$ with $s = 2$.

Since the fuzzy transformations of two variables X and Y are independent then an observation $X = i, Y = j$, where i, j are the Likert values in data can be interpreted as a two-dimensional fuzzy set with the basic function

$$\begin{aligned} f_{i,j}(x, y|c_X, c_Y, \sigma_X, \sigma_Y) &= f_i(x|c_X, \sigma_X) \cdot f_i(y|c_Y, \sigma_Y) \\ &= \frac{1}{2 \cdot \pi \cdot \sigma_X \cdot \sigma_Y} \exp -\frac{1}{2} \cdot \left(\frac{(x - c_X)^2}{\sigma_X^2} + \frac{(y - c_Y)^2}{\sigma_Y^2} \right) \end{aligned}$$

where the parameters are the standard deviations σ_X and σ_Y of X and Y, respectively and the mean values c_X, c_Y that specifies the center of the fuzzy set[5]. Similarly, for the Π-curves we get

$$\Pi_{i,j}(x, y) = \Pi_i(x) \cdot \Pi_j(y) \ .$$

3 F-transform

In this section, we briefly introduce F-transform - a technique that we will use to approximate multidimensional probability distributions. In this paper, F-transform is used to represent a discrete probability distribution by a finite set of its F-transform components. We present the definition of F-transform (Perfilieva, 2006) for a discrete probability distribution of two variables[6] defined on the grid

$$\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2 = \{1, \ldots, N\} \times \{1, \ldots, M\} \ .$$

Let $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$ be basic functions. Further, let $A_1, \ldots, A_n : [1, N] \to [0, 1]$ and $B_1, \ldots, B_m : [1, M] \to [0, 1]$ be fuzzy partitions of $[1, N]$ and $[1, M]$, respectively. Assume that the grid $\mathbb{X}$ is *sufficiently dense with respect to the chosen partitions*, which means that $(\forall k)(\exists x \in \mathbb{X}_1) : A_k(x) > 0$, and $(\forall l)(\exists y \in \mathbb{X}_2 : B_l(y) > 0$.

Definition 2. A $n \times m$ matrix U of real numbers is called *F-transform* of P with respect to $\{A_1, \ldots, A_n\}$ and $\{B_1, \ldots, B_m\}$ if for all $k = 1, \ldots, n,\ l = 1, \ldots, m$,

$$U_{kl} = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} P(X = x_i, Y = y_j) \cdot A_k(x_i) \cdot B_l(y_j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} A_k(x_i) \cdot B_l(y_j)} \ . \tag{1}$$

The elements U_{kl} are called *components of F-transform*.

In Figure 2 we illustrate F-transform computation for a single variable X (Referendum about EU) resulting in the vector

$$(U_1, U_2, U_3, U_4, U_5, U_6, U_7) = (\ 0.168,\ 0.101,\ 0.101,\ 0.144,\ 0.087,\ 0.107,\ 0.197\ ) \ .$$

We can try to reconstruct the original probability distribution P and in this way get an approximation $\hat{P}$ of P by the application of the inverse F-transform to its F-transform represented by matrix U.

[5] In all our experiments we assume that standard deviations are equal, i.e. $\sigma_X^1 = \sigma_Y^1 = \ldots = \sigma_X^7 = \sigma_Y^7$ and also the correlation $\rho_{X,Y}$ between X and Y in the basic function $f_{i,j}$ is zero.

[6] The generalization to more dimensional probability distributions is straightforward.

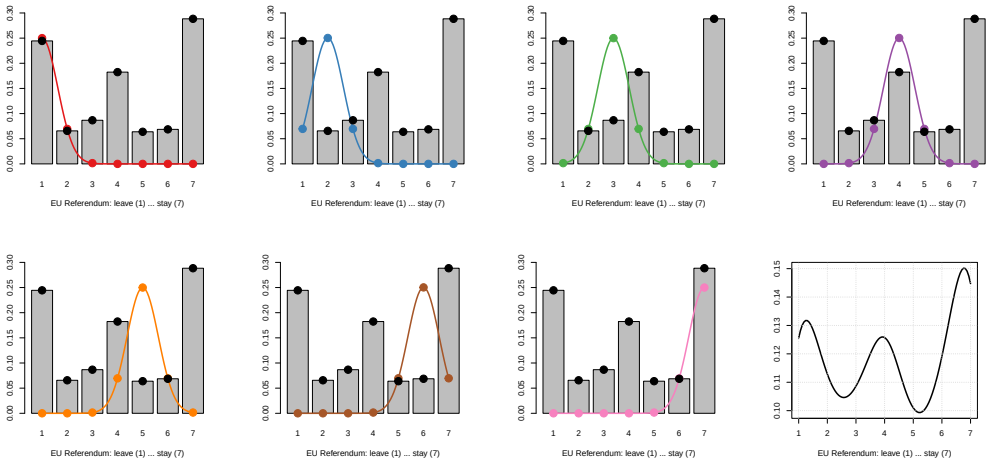


Figure 2: Illustration of the F-transform computations using Gaussian basic functions. The values denoted by dots from the Gaussian curve and from the corresponding columns are multiplied and all products are summed up. The last plot on the bottom right is the result of the inverse F-transform for the full interval $[1, 7]$.

Definition 3. The *inverse F-transform* maps U to $\hat{P} : \mathbb{X} \to [0, 1]$ with respect to $\{A_1, \ldots, A_n\}$ and $\{B_1, \ldots, B_m\}$ and it is defined as follows:

$$\hat{P}(X = x_i, Y = y_j) \quad = \quad \frac{\sum_{k=1}^{n} \sum_{l=1}^{m} U_{kl} \cdot A_k(i) \cdot B_l(j)}{\sum_{k=1}^{n} \sum_{l=1}^{m} A_k(i) \cdot B_l(j)} \quad . \tag{2}$$

In Figure 3 we compare the square root error of the F-transform approximations on a regular 7x7 and 3x3 grids with respect to the original distribution for two variables (Geopolitcs and Referendum about EU), both reported using the seven-value Likert scale. It is not surprising that the F-transform on the 7x7 grid is able to fit perfectly the original table if the fuzziness is very low. More interesting and useful are approximations on the sparser grid 3x3. In this way we achieve a compression of the original table and since the application of F-transform corresponds to smoothing we can expect that these compressed probability tables represent well the actual values. In our experiment we use triangular and Gaussian fuzzy membership functions.

In Figure 4 we can compare the original probability distribution having 7x7 values with its approximations using F-transform on a regular 3x3 grid. The Gaussian membership function has both sigma values equal to 0.8. The triangular fuzzy membership function has both width values equal to 3. The Π-curves membership function has the parameter $s = 1.8$. The Gaussian and Π-curves membership functions seem preferable for the consequent computations since their better correspond to the original table.

With the increasing value of the standard deviation σ the number of modes of the probability distribution is decreasing. We note that rather than an optimization task to

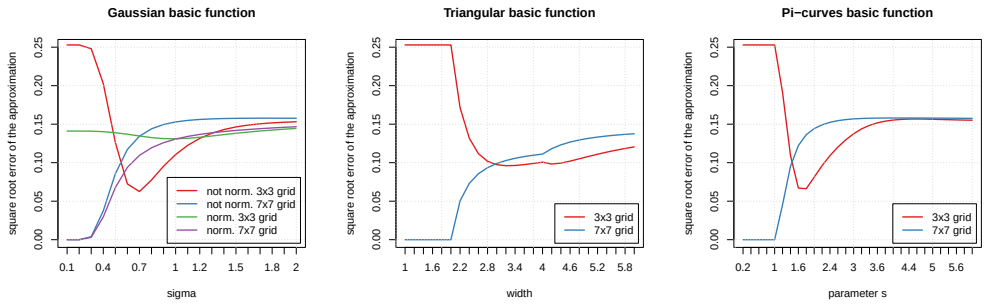


Figure 3: Square root error of the F-transform approximations on a regular 7x7 and 3x3 grids using Gaussian (left), triangular (middle), and Π-curves (right) basic functions.

be solved it is a task for a domain expert to decide how the variables relate, e.g., how many modes the probability distribution is expected to have.

4 BN inference using F-transform

To illustrate the application of F-transform for probabilistic inference in Bayesian networks we use the dataset from the survey "Dividing Lines in Czech Society" (Buchtík, 2023). This survey contained 30 pairs of contradictory statements related to Czech society, denoted $A_1, \ldots, A_{30}$, 14 pairs of contradictory statements that relate to the respondents personally, denoted $D_1, \ldots, D_{14}$, and answers to several demographic questions. The dataset contains answers of 1661 respondents. We performed a projection on the original dataset, retaining only four variables:

A_{15} Geopolitics: integral part of Western Europe (1) vs. neutral bridge (7),

A_{18} Referendum about EU: leave (1) vs. stay (7),

A_{19} Czech Society after 1989: the right direction (1) vs. the wrong direction (7), and

D_{11} Personal benefits from the EU membership: beneficial (1) vs. non-beneficial (7).

In Figure 5, we present the Bayesian network structure representing the relationships between these four variables learned by optimizing the BIC score.

For the illustration of the suggested method we will use a simple but still interesting scenario. Assume two of three child variables were observed and we want to compute the conditional probability of the parent variable and the remaining child. Without loss of generality assume that $A_i = a$ and $A_j = b$ was observed and we want to compute conditional probability distributions $P(D|A_i = a, A_j = b)$ and $P(A_k|A_i = a, A_j = b)$. We

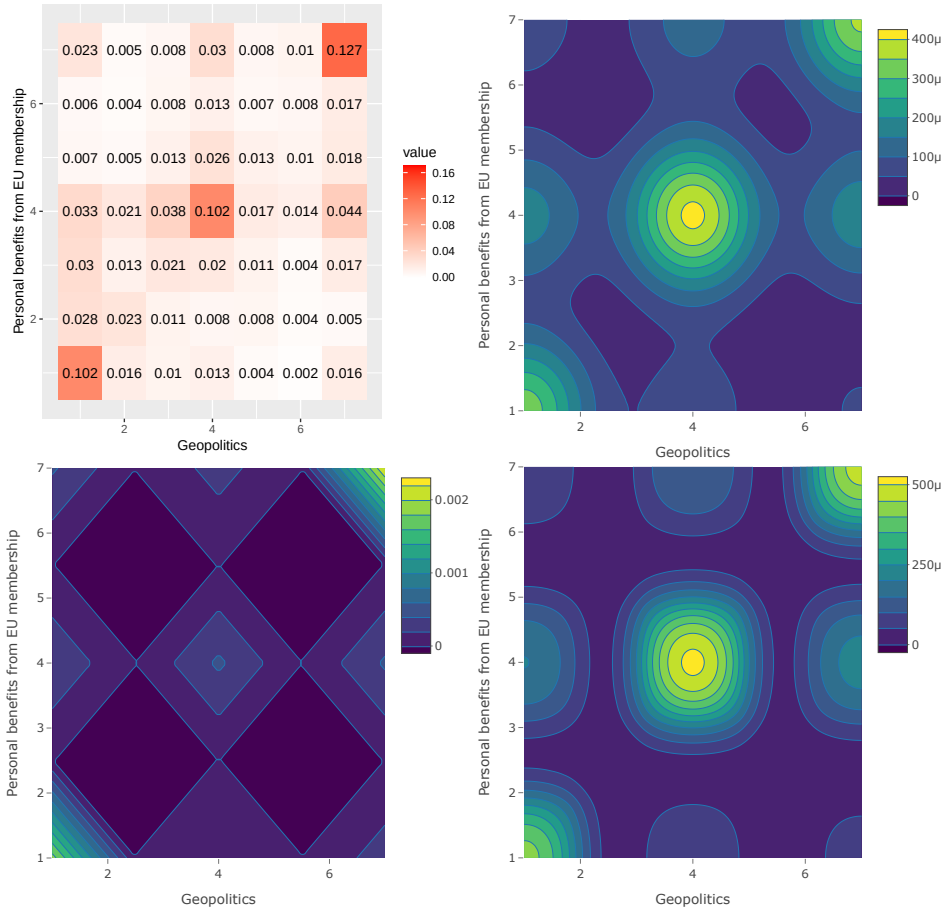


Figure 4: Comparison of the original probability table (top) and its F-transform approximations on a regular 3x3 grid using Gaussian (topright), triangular (bottomleft) and Π-curves (bottomright) membership functions.

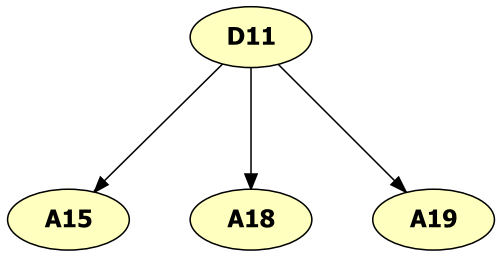


Figure 5: Bayesian network representing relations between four variables from the "Dividing Lines in Czech Society" survey.

perform the following computations for all combinations of values c of A_k and d of D:

$$\psi(D=d) = P(A_i=a|D=d)\cdot P(A_j=b|D=d)\cdot P(D=d) \quad (3)$$

$$p = \sum_d \psi(D=d) \quad (4)$$

$$P(D=d|A_i=a,A_j=b) = \frac{1}{p}\cdot\psi(D=d) \quad (5)$$

$$P(A_k=c|A_i=a,A_j=b) = \sum_d P(A_k=c|D=d)\cdot P(D=d|A_i=a,A_j=b) \ , \quad (6)$$

where p denotes the probability of evidence $P(A_i=a,A_j=b)$ and serves as the normalization constant. In the context of our example, assume $A_i=A_{15}$, $A_j=A_{18}$, $A_k=A_{19}$, and $D=D_{11}$. Then the formulas (3), (4), and (5) correspond to the multiplication of one row of probability table $P(A_{15},D_{11})$ corresponding to $A_{15}=a$ with one row of $P(A_{18},D_{11})$ corresponding to $A_{18}=b$ and normalizing the resulting vector. The formula (6) is multiplication of each column of probability table $P(A_{19},D_{11})$ with corresponding element of the resulting vector and computing the marginal sums over all columns.

All two-dimensional distributions used in the above computations are parametrized on the $3x3$ grid. Note that instead of $3\cdot 7\cdot 7=147$ parameters required for the original discrete model only $3\cdot 3\cdot 3=27$ parameters are needed. Generally, for complex models the savings using F-transform can be substantially larger and may correspond to a shift from an intractable to a tractable model. in Figure 7 (left) the mean square error of the approximation is presented as a function of σ of the Gaussian basic function for the two versions of F-transform (not-normalized and normalized). We can see that the lowest error is achieved by the not-normalized version with $\sigma=0.8$. On the right hand side of this figure an example of the comparison of a conditional probability distribution in the original model and its approximation is presented. In the plot $P(A_{19}|A_{15}=1,A_{18}=3)$ is presented. Note that the observation $A_{18}=3$ is not on the 3x3 grid but it does not pose any problem in the inference algorithm since using the inverse F-transform the probability distribution given any value of A_{18} from interval $[1,7]$ can be estimated.

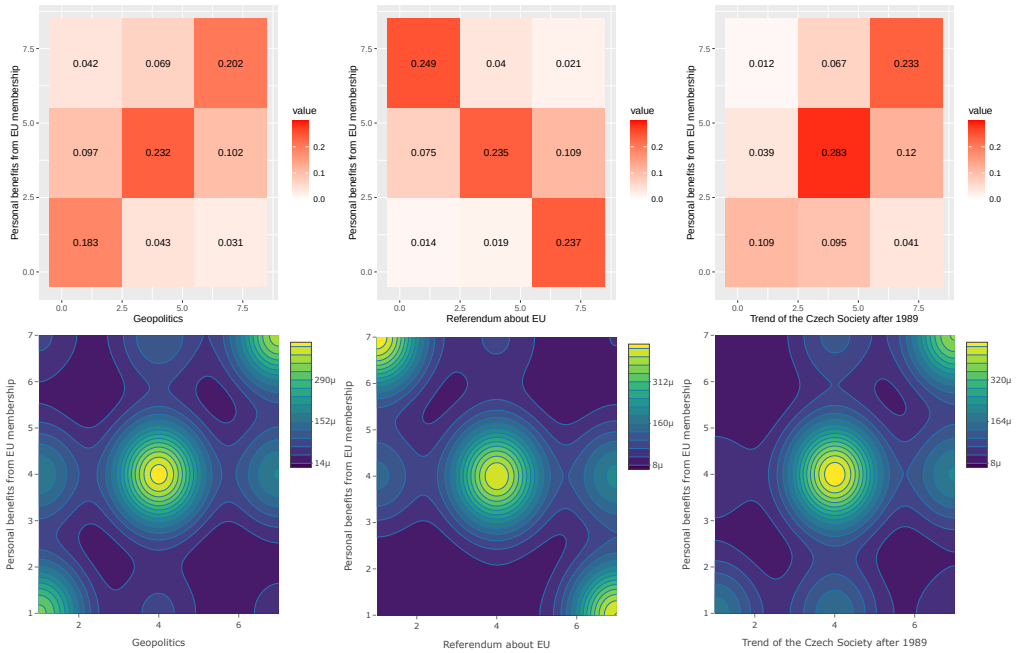


Figure 6: Matrices of the F-transform on a 3x3 grid (top) and their corresponding approximations after the inverse F-transform (bottom).

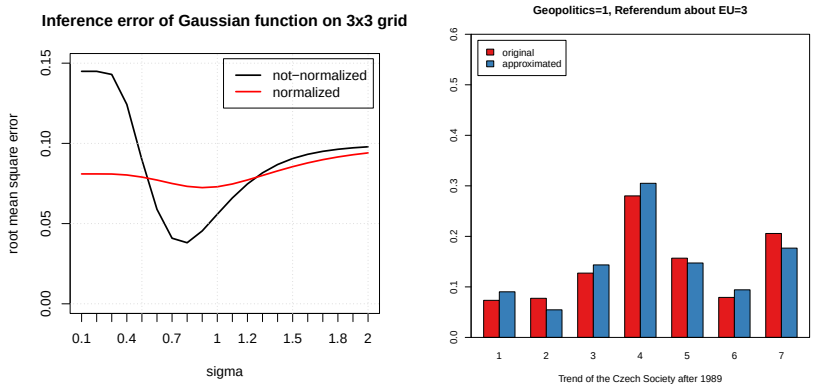


Figure 7: The root mean square error of the approximation as a function of σ of the Gaussian basic function (left) and the comparison of a conditional probability distribution in the original model and its approximation (right).

5 Conclusions

We have argued that Likert scale can be naturally interpreted using fuzzy sets. F-transform can be used to approximate probability tables in Bayesian networks, which can lead to computationally more efficient probabilistic inference. However, this paper represents a starting point and future developments in several directions are necessary.

Efficient inference procedure is still to be described and implemented. We would like to use keypoints instead of regular grids selected so that they represent well the modelled probability distribution with least memory requirements. A procedure for the combination of two probability tables with different keypoints is to be devised. Another dimension to be explored are more general shapes of membership functions in more dimensional transformations, e.g. using general covariance matrix instead of just a diagonal one.

The F-transform representation of the probability distribution resembles Kernel Density Estimates (Rosenblatt, 1956; Silverman, 1986). They differ in that the Kernel Density Estimates place basic functions at every datapoint in the dataset. However, some of the procedures used there may be useful also for the F-transform representation.

Acknowledgment

Developed within the project of the University of Ostrava: Social Dimension of New Technologies in the Energy Sector in the Ostrava Metropolitan Area (reg. number CZ.02.01.01/00/23_021/000859), with financial support from the European Union through the Jan Amos Komenský Operational Programme.

References

M. Buchtík. *Různá vyprávění o jedné společnosti*. Friedrich-Ebert-Stiftung and Masarykova demokratická akademie Praha, 2023.

M. A. Gil and G. González-Rodríguez. Fuzzy vs. likert scale in statistics. In E. Trillas, P. P. Bonissone, L. Magdalena, and J. Kacprzyk, editors, *Combining Experimentation and Theory: A Hommage to Abe Mamdani*, pages 407–420. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-24666-1. URL `https://doi.org/10.1007/978-3-642-24666-1_27`.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer New York, NY, 2 edition, 2007. doi: 10.1007/978-0-387-68282-2.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

H. Pan and L. Liu. Fuzzy Bayesian networks—a general formalism for representation, inference and learning with hybrid Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(07):941–962, 2000.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

I. Perfilieva. Fuzzy transforms: Theory and applications. *Fuzzy Sets and Systems*, 157: 993–1023, 2006. URL https://doi.org/10.1016/j.fss.2005.11.012.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. URL https://doi.org/10.1214/aoms/1177728190.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, London, 1986. ISBN 978-0412246203. URL https://doi.org/10.1201/9781315140919.

Structural Learning of BN2A models

Iván Pérez[1,2,3] and Jiří Vomlel[1]

[1]Institute of Information Theory and Automation, Czech Academy of Sciences
{*cabrera, vomlel*}*@utia.cas.cz*
[2]Institute of Computer Science, Czech Academy of Sciences
[3]Faculty of Mathematics and Physics, Charles University

Abstract

Probabilistic graphical models, particularly Bayesian networks, provide a flexible framework for representing dependencies among random variables and have been widely applied in domains such as medicine, biology, and educational testing. Our work focuses on BN2A networks - a specialized bipartite Bayesian network architecture, where the first layer consists of hidden variables and the second layer consists of observed variables. In BN2A models all variables are assumed to be binary. The variables in the second layer depend on the variables in the first layer and this dependence is characterized by conditional probability tables representing Noisy-AND models. In this work, we propose an Expectation-Maximization (EM) algorithm for learning the structure of BN2A models, that is, for learning the relationship between hidden variables and observed variables. To validate our structural learning algorithm, we designed two experiments using educational assessment data. For the first experiment, we used synthetic data generated from a BN2A model that we previously defined, while for the second experiment we used a well-known real-world dataset in the field of Cognitive Diagnostic Models, the Fraction Subtraction dataset. Our proposed algorithm has interesting potential use cases. One key application is to generate a reasonably accurate BN2A structure model for educational diagnosis, particularly in scenarios where no prior model exists. Depending on the required level of accuracy, the estimated model can be used directly to analyze skill profiles or serve as an initial framework for test designers, who can further refine it before implementation.

1 Introduction

Bayesian networks are a popular framework for modeling probabilistic relationships between random variables and have been used successfully in educational tests (Almond et al. (2015), Vomlel (2004)). There is interest in a particular type of Bayesian networks we have called BN2A, which are characterized as bipartite networks, where the first layer

consists of hidden variables (which commonly represent skills) and the second layer consists of observed variables (which represent questions in a test). In BN2A models all variables are assumed to be binary. The variables in the second layer depend on the variables in the first layer and this dependence is characterized by conditional probability tables (CPTs) representing a Noisy-AND model. In Fig. 1 we give an example of a directed bipartite graph that can define the structure of a BN2A model.

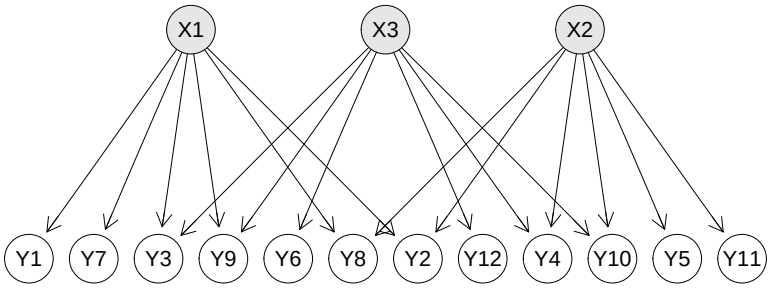


Figure 1: BN2A model with three hidden variables and 12 observed variables.

This paper is structured as follows. In Section 2 we formally introduce the BN2A model and present their corresponding CPTs, leaky Noisy-AND. In Section 3 we present an algorithm for learning the structure of the BN2A models while in Section 4, we illustrate how this algorithm works conducting two experiments. For the first experiment, we used synthetic data generated from a BN2A model that we previously defined, while for the second experiment we used a well-known real-world dataset in the field of Cognitive Diagnostic Models, the Fraction Subtraction dataset (Tatsuoka, 2002). Finally, we summarise the contribution of this paper and discuss our future work in Section 5.

2 BN2A model

Let $\mathbf{X}$ denote the vector $(X_1, \ldots, X_K)$ of K hidden variables, and similarly let $\mathbf{Y}$ denote the vector $(Y_1, \ldots, Y_L)$ of L observed dependent variables. The hidden variables are also called attributes or skills in the context of cognitive diagnostic models (CDMs), and observed dependent variables are usually called items or questions in the same context. All variables are assumed to be binary, taking states from $\{0, 1\}$. The state space of the multidimensional variable $\mathbf{X}$ is denoted $\mathbb{X}$ and is equal to the Cartesian product of the state spaces of $X_k, k = 1, \ldots, K$:

$$\mathbb{X} \quad = \quad \times_{k=1}^{K} \mathbb{X}_k = \{0, 1\}^K \ . \tag{1}$$

Similarly, the state space of multidimensional variable $\mathbf{Y}$ is denoted $\mathbb{Y}$ and is equal to the Cartesian product of state spaces of $Y_\ell, \ell = 1, \ldots, L$:

$$\mathbb{Y} \quad = \quad \times_{\ell=1}^{L} \mathbb{Y}_\ell = \{0,1\}^L \ . \tag{2}$$

The basic building blocks of a BN2A model are conditional probability tables (CPTs) specified in the form of a Noisy-AND model. Let Y_ℓ be an observed dependent variable and $pa(Y_\ell)$ be the subset of indexes of related variables from $\mathbf{X}$. They are referred to as the parents of Y_ℓ.

Definition 1 (Noisy-AND model)**.**
A conditional probability table $P(Y_\ell|\mathbf{X}_{pa(Y_\ell)})$ represents a Noisy-AND model if

$$P(Y_\ell = y_\ell|\mathbf{X}_{pa(Y_\ell)} = \mathbf{x}_{pa(Y_\ell)}) \quad = \quad \begin{cases} q_{\ell,0} \cdot \prod\limits_{i \in pa(Y_\ell)} (q_{\ell,i})^{(1-x_i)} & \text{if } y_\ell = 1 \\ 1 - q_{\ell,0} \cdot \prod\limits_{i \in pa(Y_\ell)} (q_{\ell,i})^{(1-x_i)} & \text{if } y_\ell = 0. \end{cases} \tag{3}$$

Note that if $x_i = 1$ then $(q_{\ell,i})^{(1-x_i)} = 1$ and if $x_i = 0$ then $(q_{\ell,i})^{(1-x_i)} = q_{\ell,i}$. The interpretation is that if $X_i = 1$, then this variable definitely enters the AND relation with the value 1. If $X_i = 0$, then there is still a probability $q_{\ell,i}$ that it enters the AND relation with value 1. The model also contains an auxiliary parent X_0 which is always 0 and thus enters the AND relation with probability $q_{\ell,0}$ for the value 1. This probability is traditionally called *leak* probability and allows non-zero probability of $Y_\ell = 0$ even if all parents of Y_ℓ have value 1. In CDM, this model is known as the Reduced Reparametrized Unified Model (RRUM) (Hartz and Roussos, 2008) and it is a special case of the Generalized Noisy Inputs, Deterministic AND (GNIDA) gate model (de la Torre, 2011).

The prior probability of the hidden skill for $k = 1, \ldots, K$ is defined as

$$P(X_k = x_k) \quad = \quad (p_k)^{x_k}(1-p_k)^{(1-x_k)} \ , \tag{4}$$

which means that if $x_k = 1$ then it is p_k and if $x_k = 0$ then it equals $1 - p_k$.

Now we are ready to define a special class of Bayesian network models with hidden variables, called BN2A model.

Definition 2 (BN2A model)**.**
A BN2A model is a pair (G, P), where G is a directed bipartite graph with its nodes divided into two layers. The nodes of the first layer correspond to the hidden variables $X_1, \ldots, X_K$ and the nodes of the second layer correspond to the observed variables $Y_1, \ldots, Y_L$. All edges are directed from a node of the first layer to a node of the second layer. The symbol P refers to the joint probability distribution over the variables corresponding to the nodes of the graph G. The probability distribution is parameterized by a vector of model parameters $(\mathbf{p}, \mathbf{q})$:

$$(\mathbf{p}, \mathbf{q}) \quad = \quad \big((p_k)_{k \in \{1,\ldots,K\}}, (q_{\ell,k})_{\ell \in \{1,\ldots,L\}, k \in \{0\} \cup pa(Y_\ell)}\big) \ . \tag{5}$$

The bipartite graph G of a BN2A model can also be specified by an incidence matrix, in the context of CDM, traditionally denoted by $\mathbf{Q}$. A $\mathbf{Q}$-matrix is an $L \times K$ binary matrix, with entries $\mathbf{Q}_{\ell,k} \in \{0, 1\}$ that indicate whether or not the ℓ^{th} observed dependent variable is linked to the k^{th} hidden variable:

$$\mathbf{Q}[\ell, k] \quad = \quad \begin{cases} 1 & \text{if } X_k \in pa(Y_\ell) \\ 0 & \text{otherwise.} \end{cases}$$

3 Structural learning

As previously mentioned, our primary focus lies in applying BN2A models to educational testing. In practice, education experts often possess test outcome data but lack a predefined cognitive model structure. This creates a fundamental need for computational tools that can automatically learn the underlying model structure from available data — particularly important since BN2A models can not only effectively represent the relationship between latent skills and observed responses, but are also inherently interpretable for educational applications.

In this section we propose a method for learning the structure of a BN2A model from data $\mathbf{D}$ where states of variables $\mathbf{Y}$ are observed but variables $\mathbf{X}$ are hidden, i.e., their states are unobserved. The algorithm is basically a version of the structural EM algorithm proposed by Friedman (1998).

We further assume that the number K of hidden variables is known.[1] Clearly, a naive approach of evaluating all possible BN2A structures would quickly become intractable for larger values of K and L.

Recall that L denotes the dimension of the vector of observed variables $\mathbf{Y} = (Y_1, \ldots, Y_L)$ and K denotes the dimension of the vector of hidden variables $\mathbf{X} = (X_1, \ldots, X_K)$. Let for all $\mathbf{x} \in \mathbb{X}$ the function $N(\mathbf{x})$ denote the number of occurrences of vector $\mathbf{x}$ in data $\mathbf{D}$. Similarly, let for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{X} \times \mathbb{Y}$ the function $N(\mathbf{x}, \mathbf{y})$ denote the estimated number[2] of occurrences of vector $(\mathbf{x}, \mathbf{y})$ in completed data $\mathbf{C}$ consisting of values of $(\mathbf{X}, \mathbf{Y})$.

3.1 Algorithm

The structure learning algorithm is presented in Algorithm 1. This algorithm requires a dataset $\mathbf{D}$ consisting of n data vectors with values of $\mathbf{Y}$ and the dimension K of $\mathbf{X}$ as its input. In general, the algorithm alternates between two major steps - the E-step and the M-step as described below.

[1] Generally, one can consider K also as a free parameter and from best models of different K values select one that maximizes an evaluation criteria, which in our study is BIC.

[2] The number of occurrences is computed as their expected number. Therefore, the numerical values are non-negative real numbers, i.e., they need not (and typically they are not) natural numbers. This does not cause any problems in subsequent computations.

Input : $\mathbf{D}$ – dataset consisting of n data vectors with values of $\mathbf{Y}$
K – the dimension of $\mathbf{X}$
$\mathbf{Q}'$ – initial Q matrix, e.g. the unit $L \times K$ matrix
$(\mathbf{p}, \mathbf{q})$ – initial parameter values, e.g. GDINA($\mathbf{D}, \mathbf{Q}', K$, "RRUM")
Output: $\mathbf{Q}$ – the Q-matrix of BN2A
$\mathbf{p}$ – prior probabilities of the hidden variables
$\mathbf{q}$ – parameters of the Noisy-AND models

Set $\mathbf{Q}$ as the zero $L \times K$ matrix;
while $\mathbf{Q} \neq \mathbf{Q}'$ **do**
  if $\mathbf{Q}$ *is not the zero* $L \times K$ *matrix* **then**
    $\mathbf{Q}' \leftarrow \mathbf{Q}$;
  end
  E-step
  for $\mathbf{x} \in \{0,1\}^K$ **do**
    $P(\mathbf{x}) = \prod_{k=1}^{K} (p_i)^{x_i} \cdot (1-p_i)^{1-x_i}$;
    for $\ell \in \{1, \ldots, L\}$ **do**
      $pa(Y_\ell) = \{k \in \{1, \ldots, K\} : \mathbf{Q}'[\ell, k] = 1\}$;
      $R = \{0\} \cup pa(Y_\ell)$;
      $P(Y_\ell = 1|\mathbf{x}) = \prod_{i \in R} (q_{\ell,i})^{1-x_i}$;
      $P(Y_\ell = 0|\mathbf{x}) = 1 - \prod_{i \in R} (q_{\ell,i})^{1-x_i}$;
    end
  end
  M-step
  $\mathbf{p} = (\frac{N(x_1)}{n}, \ldots, \frac{N(x_k)}{n})$;
  for $\ell \in \{1, \ldots, L\}$ **do**
    $M = \{1, \ldots, L\} \setminus \{\ell\}$;
    $N(\mathbf{x}, y_\ell) = \sum_{\mathbf{y}_M} N(\mathbf{x}, \mathbf{y})$;
    $BIC^* = -\infty$;
    for $R \subseteq \{1, \ldots, K\}$ **do**
      for $(\mathbf{x}_R, y_\ell) \in \{0,1\}^{|R|+1}$ **do**
        $S = \{1, \ldots, K\} \setminus R$;
        $N(\mathbf{x}_R, y_\ell) = \sum_{\mathbf{x}_S} N(\mathbf{x}, y_\ell)$;
      end
      $\mathbf{q}_{\ell,R} = \arg\max_{\mathbf{q}'} LL(\mathbf{q}')$;
      $BIC = LL(\mathbf{q}_{\ell,R}) - \frac{\log n}{2} \cdot (|R| + 1)$;
      if $BIC > BIC^*$ **then**
        $BIC^* = BIC$;
        $\mathbf{q}_\ell = \mathbf{q}_{\ell,R}$;
        for $k \in \{1, \ldots, K\}$ **do**
          $\mathbf{Q}[\ell, k] \leftarrow \mathbb{I}(k \in R)$;
        end
      end
    end
  end
end

Algorithm 1: Learning the structure of BN2A.

E-step

In the E-step it is assumed a model structure specified by a matrix $\mathbf{Q}$ is known. In the first iteration we assume the structure is a complete graph represented by an all-ones matrix $\mathbf{Q}$. For the given $\mathbf{Q}$ the parameters of Noisy-AND models of all CPTs are learned using an R library called GDINA (Ma and de la Torre, 2020) for the particular RRUM model, which is equivalent to the BN2A model. These parameters are then used to estimate values of hidden variables. In this way a complete data set $\mathbf{C}$ is created.

M-step

The complete data $\mathbf{C}$ represent the input of the M-step. In this step a BN2A model that best describes the data is learned. As the model quality criteria the well-known BIC criterion is used. It is the log-likelihood penalized by a penalty proportional to the number of model parameters. The important observation is that the learning algorithm finds the best parent set and optimal parameters independently for each variable Y_ℓ, $\ell = 1, \ldots, L$, which still guarantees the global optimality. This does not hold for general Bayesian networks since the independent learning may result in directed cycles that are forbidden. But for models with their structure defined by a bipartite graph the global optimality is guaranteed since bipartite graphs cannot contain directed cycles. For each considered parent set the parameters of the Noisy-AND model maximizing the log-likelihood for the given data are learned using a gradient method and its BIC is computed. Then, the parent set with the maximum BIC is selected, and based on it, the corresponding row of the matrix $\mathbf{Q}$ is formed, which serves as input for the E-step. An important observation is that the log-likelihood of Noisy-AND is a nicely shaped function with a unique maximum. This follows from the concavity of the log-likelihood of the Noisy-AND function:

$$\begin{aligned} LL(\mathbf{q}_\ell) &= \sum_{\mathbf{x}_R} N(\mathbf{x}_R, Y_\ell = 1) \cdot \log \prod_{i \in T} (q_{\ell,i})^{1-x_i} \\ &\quad + \sum_{\mathbf{x}_R} N(\mathbf{x}_R, Y_\ell = 0) \cdot \log \left(1 - \prod_{i \in T} (q_{\ell,i})^{1-x_i} \right) \quad . \end{aligned} \tag{6}$$

The concavity can be proven similarly as in (Vomlel et al., 2023, Lemma 1).

4 Experiments

To show how Algorithm 1 works, we conducted two experiments. For the first experiment, we used synthetic data generated from a BN2A model representing a test consisting of 12 items measuring three skills, while for the second experiment we used a well-known real-world dataset in the field of Cognitive Diagnostic Models, the Fraction Subtraction dataset (Tatsuoka, 2002).

4.1 A BN2A model with $K = 3$ and $L = 12$

For the first experiment, we used the structure shown in Figure 1 which presents dependencies of 12 questions on three proposed skills. In this model, note that that six questions require only one skill, while the others questions require two skills to be answered correctly.

As in any knowledge domain, some skills are easy to master and others are difficult; thus, tests typically require skills of varying mastery prevalence in the population. In our experiment, the simulated proportion of skill mastery (prior probabilities of mastering the skills) was set as $p_1 = 0.8$, $p_2 = 0.6$, and $p_3 = 0.4$. The leak parameters $q_{\ell,0}$ represent the probability of answering question ℓ correctly when the student masters all required skills. These parameters were selected to range from 0.6 to 0.9. The failure parameters $q_{\ell,k}$ act as penalty factors for lacking the k-th skill when answering question ℓ. These parameters were randomly chosen in the range of 0.1 to 0.4 to have a realistic model. With the structure in Figure 1 and the above parameters, we randomly generated a dataset $\mathbf{D}$ of size $n = 10^4$ —a feasible sample size for large-scale standardized tests.

Before running the algorithm, we randomly created ten initial vectors in the parameter space. For each, we computed the log-likelihood of the complete model (unit Q-matrix) using the GDINA library (Ma and de la Torre, 2020), and initialized the algorithm using the parameter vector corresponding to the highest log-likelihood value. This helps avoid getting stuck in a local maximum. Once we execute Algorithm 1 using the dataset $\mathbf{D}$ and the proposed number of latent variables ($K = 3$), the model structure was correctly learned in the first iteration, and was completed in the second iteration, confirming that the structure was the same in both iterations.

The algorithm learned the correct structure but with a different labeling of the hidden variables. To facilitate the comparison of the results we have permuted the columns of the learned model. Table 1 compares the prior probabilities for each skill, while Table 2 compares the leak and failure parameters for each test question. In Table 2, a dash (-) indicates that there is no relationship between the question and the corresponding skill. In general, it can be seen that, for the learned model, the prior probabilities and the leak and failure parameters are close to the original model.

p_1	p_2	p_3
0.800	0.600	0.400

p_1	p_2	p_3
0.7943	0.5954	0.4028

Table 1: Comparison of prior probabilities for skill mastery: original values (left) and learned values (right).

4.2 Tatsuoka's fraction subtraction dataset

The fraction subtraction test (Tatsuoka and Tatsuoka, 1987) was designed as a diagnostic tool to detect common error patterns and maladaptive solution strategies in fraction arithmetic. A detailed presentation of the test questions appears in Table 3.

ℓ	$q_{\ell,0}$	$q_{\ell,1}$	$q_{\ell,2}$	$q_{\ell,3}$
1	0.900	0.400	-	-
2	0.800	0.200	0.100	-
3	0.700	0.100	-	0.300
4	0.600	-	0.200	0.100
5	0.700	-	0.200	-
6	0.800	-	-	0.300
7	0.950	0.300	-	-
8	0.850	0.400	0.200	-
9	0.750	0.300	-	0.200
10	0.650	-	0.200	0.400
11	0.750	-	0.100	-
12	0.850	-	-	0.200

ℓ	$q_{\ell,0}$	$q_{\ell,1}$	$q_{\ell,2}$	$q_{\ell,3}$
1	0.8936	0.3707	-	-
2	0.7885	0.1942	0.1079	-
3	0.6836	0.0725	-	0.3059
4	0.5990	-	0.1897	0.0895
5	0.6916	-	0.2097	-
6	0.7867	-	-	0.2965
7	0.9473	0.2889	-	-
8	0.8522	0.4057	0.2092	-
9	0.7363	0.2969	-	0.1917
10	0.6378	-	0.1905	0.3929
11	0.7438	-	0.1070	-
12	0.8457	-	-	0.2016

Table 2: Comparison of leak and failure parameters: original values (left) and learned values (right).

The fraction subtraction dataset has become a benchmark in cognitive diagnostic modeling research, serving as a foundational testbed for over three decades. This seminal dataset has been extensively used to develop, validate, and compare diagnostic classification models due to its well-documented cognitive structure and pedagogical relevance. Its popularity stems from the clear mapping between mathematical skills and item responses, making it an ideal case study for educational testing research. The dataset's widespread adoption across numerous studies (e.g., de la Torre and Douglas (2004), DeCarlo (2011), Culpepper (2019)) establishes it as a gold standard for evaluating new methodological approaches in cognitive diagnosis.

The dataset contains binary response patterns from N = 536 middle school students on J = 20 test questions. The data matrix (536 × 20) represents each student's performance, where rows correspond to individual students and columns represent test questions. A value of 1 indicates a correct response, while 0 denotes an incorrect response.

Fraction subtraction problems were constructed to include the basic skills required for solving problems correctly, such as borrowing, converting a whole number to a simple fraction, and getting the common denominator. Each question assesses combinations of eight core cognitive skills. Table 4 displays the operational definitions of the eight measured skills while Table 5 (left) shows the corresponding skill-question mapping (Q-matrix) for the Fraction subtraction test.

Several noteworthy observations emerge from examining Tatsuoka's proposed Q-matrix. First, the importance of the skill X_7 (Subtract numerators) becomes evident, as it is required for all test items except Y_9. Second, half of the measured skills (specifically X_1, X_3, X_6, and X_8) appear in no more than three questions each. This distribution aligns with both the skill definitions and question requirements: for instance, while many problems involve mixed numbers (whole numbers with fractions), only specific items like Y_7, Y_{15} and Y_{19} necessitate skill X_1 (Convert a whole number to a fraction) for their solution.

No.	Question	No.	Question
1	$\frac{5}{3} - \frac{3}{4}$	11	$4\frac{1}{3} - 2\frac{4}{3}$
2	$\frac{3}{4} - \frac{3}{8}$	12	$\frac{11}{8} - \frac{1}{8}$
3	$\frac{5}{6} - \frac{1}{9}$	13	$3\frac{3}{8} - 2\frac{5}{6}$
4	$3\frac{1}{2} - 2\frac{3}{2}$	14	$3\frac{4}{5} - 3\frac{2}{5}$
5	$4\frac{3}{5} - 3\frac{4}{10}$	15	$2 - \frac{1}{3}$
6	$\frac{6}{7} - \frac{4}{7}$	16	$4\frac{5}{7} - 1\frac{4}{7}$
7	$3 - 2\frac{1}{5}$	17	$7\frac{3}{5} - 2\frac{4}{5}$
8	$\frac{2}{3} - \frac{2}{3}$	18	$4\frac{1}{10} - 2\frac{8}{10}$
9	$3\frac{7}{8} - 2$	19	$4 - 1\frac{4}{3}$
10	$4\frac{4}{12} - 2\frac{7}{12}$	20	$4\frac{1}{3} - 1\frac{5}{3}$

Table 3: Tatsuoka's Fraction subtraction test

In our experiment, we tested different values of K ranging from 2 to 5. Table 5 (right) presents the learned Q-matrix and represents the optimal question-skill relationships for $K = 5$. To compare the two Q-matrices in Table 5, we computed their BIC values: the Tatsuoka expert-defined Q-matrix yields -5,303.9, while our data-learned Q-matrix achieves -4,635.5. This result suggests that our model, despite its lower dimensionality (5 vs. 8 skills), captures the data's essential structure more parsimoniously.

When learning a model's structure solely from response data, the latent variables lack explicit definitions. However, by analyzing the test items and their associations with learned skills (hidden variables), we can interpret these relationships. Our analysis of the learned Q-matrix reveals two significant patterns:

Skill X_4 appears in all items except Y_{18}, aligning with Tatsuoka's definition of "Subtract numerators" as a fundamental operation in fraction subtraction. Skill X_2 is associated with 15 of the 20 test items, with 12 of these matching Tatsuoka's original classification as "Separate a whole number from a fraction" - indicating the learned structure successfully recovers this semantically meaningful dimension from response patterns alone.

Examining the learned model parameters, the prior probabilities are: $p_1 = 0.582, p_2 = 0.525, p_3 = 0.764, p_4 = 0.805$, and $p_5 = 0.707$. Notably, the highest prior probability corresponds to X_4 (that we align with "Subtract numerators"), suggesting this is a foundational skill typically mastered even before formal fraction operation instruction.

Table 6 presents the corresponding BN2A model parameters (leak and failure parameters) of the learned structure. The leak parameters ($q_{\ell,0}$) generally exceed 0.8, consistent with theoretical expectations - these represent the probability of correct response given

Skill	Description
X_1	Convert a whole number to a fraction.
X_2	Separate a whole number from a fraction.
X_3	Simplify before subtracting.
X_4	Find a common denominator.
X_5	Borrow from whole number part.
X_6	Column borrow to subtract the second numerator from the first.
X_7	Subtract numerators.
X_8	Reduce answers to simplest form.

Table 4: Skills definition for the Fraction subtraction test.

mastery of required skills. The single exception ($q_{13,0} = 0.684$) aligns with failure parameters that strongly penalize lack of skill mastery.

Finally, the BN2A model's failure parameters ($q_{\ell,k}$) quantify the performance penalty for lacking specific skills per question. These values could help educators to: (a) identify particularly skill-sensitive items, (b) develop targeted instructional interventions, and (c) create adaptive assessment variants.

5 Discussion

In this paper, we focused on BN2A models—Bayesian networks where conditional probability tables (CPTs) are represented by Noisy-AND models with a bipartite graph structure, and all nodes in the first layer are hidden. These models are of particular interest due to: a) their interpretability in educational testing contexts, and b) their parameter efficiency: the number of parameters scales proportionally to $K \cdot L$, significantly fewer than in general bipartite Bayesian networks where CPTs can grow exponentially.

We proposed a Structural EM algorithm for learning the structure of BN2A models. A key advantage of this approach is that it independently identifies the optimal parent set and parameters for each observed variable while still guaranteeing global optimality, a property not shared by general Bayesian network learning methods.

To validate the algorithm, we conducted two experiments. On synthetic data the algorithm successfully recovered the ground-truth BN2A structure and estimated parameters with high accuracy. On Real-world data – the Fraction Subtraction dataset (a benchmark in Cognitive Diagnostic Modeling) – the algorithm learned a simpler model (with fewer hidden variables) without sacrificing interpretability.

Our algorithm enables practical applications, such as generating data-driven BN2A models for educational diagnostics—especially in settings where no prior expert-defined model exists. Depending on accuracy requirements, the learned model can directly analyze skill profiles, or serve as an initial framework for test designers to refine before deployment. As a next step, we plan to apply this method to a large-scale dataset of mathematics test results provided by the Czech Republic's Ministry of Education, Youth, and Sports.

	Skill							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y_1	0	0	0	1	0	1	1	0
Y_2	0	0	0	1	0	0	1	0
Y_3	0	0	0	1	0	0	1	0
Y_4	0	1	1	0	1	0	1	0
Y_5	0	1	0	1	0	0	1	1
Y_6	0	0	0	0	0	0	1	0
Y_7	1	1	0	0	0	0	1	0
Y_8	0	0	0	0	0	0	1	0
Y_9	0	1	0	0	0	0	0	0
Y_{10}	0	1	0	0	1	0	1	1
Y_{11}	0	1	0	0	1	0	1	0
Y_{12}	0	0	0	0	0	0	1	1
Y_{13}	0	1	0	1	1	0	1	0
Y_{14}	0	1	0	0	0	0	1	0
Y_{15}	1	0	0	0	0	0	1	0
Y_{16}	0	1	0	0	0	0	1	0
Y_{17}	0	1	0	0	1	0	1	0
Y_{18}	0	1	0	0	1	1	1	0
Y_{19}	1	1	1	0	1	0	1	0
Y_{20}	0	1	1	0	1	0	1	0

	Skill				
	X_1	X_2	X_3	X_4	X_5
Y_1	0	0	1	1	1
Y_2	0	0	0	1	1
Y_3	0	0	0	1	1
Y_4	0	1	0	1	1
Y_5	1	1	1	1	1
Y_6	0	0	0	1	0
Y_7	1	1	0	1	0
Y_8	0	1	0	1	1
Y_9	1	0	1	1	0
Y_{10}	0	1	0	1	1
Y_{11}	0	1	0	1	1
Y_{12}	0	1	0	1	0
Y_{13}	0	1	0	1	1
Y_{14}	0	1	0	1	0
Y_{15}	1	1	0	1	1
Y_{16}	0	1	1	1	1
Y_{17}	1	1	0	1	0
Y_{18}	0	1	0	0	1
Y_{19}	1	1	0	1	1
Y_{20}	0	1	0	1	1

Table 5: Q-matrix Structures: Original Expert-defined (Tatsuoka, left) and Data-driven Learned ($K = 5$, right).

ℓ	$q_{\ell,0}$	$q_{\ell,1}$	$q_{\ell,2}$	$q_{\ell,3}$	$q_{\ell,4}$	$q_{\ell,5}$
1	0.919	-	-	0.819	0.020	0.076
2	0.969	-	-	-	0.031	0.031
3	0.885	-	-	-	0.023	0.005
4	0.903	-	0.254	-	0.735	0.699
5	1.000	0.799	0.838	0.555	0.418	0.499
6	0.953	-	-	-	0.149	-
7	0.974	0.316	0.310	-	0.001	-
8	0.952	-	0.758	-	0.666	0.795
9	0.878	0.674	-	0.843	0.411	-
10	0.813	-	0.091	-	0.043	0.532
11	0.941	-	0.108	-	0.144	0.874
12	0.952	-	0.808	-	0.154	-
13	0.684	-	0.100	-	0.011	0.128
14	0.963	-	0.823	-	0.029	-
15	0.977	0.572	0.427	-	0.053	0.453
16	0.992	-	0.818	0.767	0.084	0.921
17	0.948	0.812	0.067	-	0.052	-
18	0.813	-	0.214	-	-	0.598
19	0.928	0.363	0.088	-	0.005	0.635
20	0.837	-	0.014	-	0.033	0.880

Table 6: Learned parameters of the BN2A model with $K = 5$.

References

R. G. Almond, R. J. Mislevy, L. Steinberg, D. Yan, and D. Williamson. *Bayesian Networks in Educational Assessment*. Springer Publishing Company, Incorporated, 2015. ISBN 149392124X. doi: https://doi.org/10.1007/978-1-4939-2125-6.

S. A. Culpepper. Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, 84(2):333–357, June 2019. doi: https://doi.org/10.1007/s11336-018-9643-8.

J. de la Torre. The generalized DINA model framework. *Psychometrika*, 76:179–199, 2011. doi: https://doi.org/10.1007/s11336-011-9207-7.

J. de la Torre and J. A. Douglas. Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353, 2004. doi: https://doi.org/10.1007/BF02295640.

L. T. DeCarlo. On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. *Applied Psychological Measurement*, 35(1): 8–26, 2011. doi: https://doi.org/10.1177/0146621610377081.

N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 129—138, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X. URL `https://dl.acm.org/doi/10.5555/2074094.2074110`.

S. Hartz and L. Roussos. The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series*, 2008(2):i–57, 2008. URL `https://doi.org/10.1002/j.2333-8504.2008.tb02157.x`.

W. Ma and J. de la Torre. GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14):1–26, 2020. doi: https://doi.org/10.18637/jss.v093.i14.

C. Tatsuoka. Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(3):337–350, 07 2002. ISSN 0035-9254. doi: https://doi.org/10.1111/1467-9876.00272.

K. K. Tatsuoka and M. M. Tatsuoka. Bug distribution and statistical pattern classification. *Psychometrika*, 52(2):193–206, 1987. doi: https://doi.org/10.1007/BF02294234.

J. Vomlel. Building adaptive tests using Bayesian networks. *Kybernetika*, 40(3):333–348, 2004. URL `https://www.kybernetika.cz/content/2004/3/333`.

J. Vomlel, V. Kratochvíl, and F. Kratochvíl. Structural learning of mixed noisy-OR Bayesian networks. *International Journal of Approximate Reasoning*, 161:108990, 2023. doi: https://doi.org/10.1016/j.ijar.2023.108990.

An experimental comparison of Bayesian network classifiers for duplicability detection

Angel T. Sáez-Ruiz[1], Ana D. Maldonado[1], Lorenzo Carretero-Paulet[2], Aaron Gálvez-Salido[2], and Rafael Rumí[1]

[1]Department of Mathematics, University of Almería, Almería, Spain
{*asr042,ana.d.maldonado,rrumi*}*@ual.es*
[2]Department of Biology and Geology, University of Almería, Almería, Spain
{*lpaulet,ags408*}*@ual.es*

Abstract

Gene duplications constitute the main source of raw genetic material upon which selection and other evolutionary forces act to generate new genes and gene functions. Gene duplication explains the existence of multigene families and much of the variation in gene number between organisms, populations, and species. Thus, understanding the mechanisms that drive the long-term retention of duplicated genes (i.e., their duplicability) is essential to understand how genomic changes ultimately contribute to phenotypic diversity and speciation. To classify genes according to their duplicability, we evaluated several Bayesian network classifiers. In particular, we compared models based on discretized variables, conditional Gaussian distributions and mixtures of polynomials to asses their performance in predicting gene duplicability.

Key words: gene duplication; hybrid Bayesian networks; conditional Gaussian; mixtures of polynomials.

1 Introduction

Among all the mechanisms involved in the origin of new genes and gene regions, gene duplication provides the main source of raw genetic material upon which mutation, selection and other evolutionary forces may act to evolve new or novel gene functions. Gene duplicates arising from whole-genome duplications (WGDs or polyploidization) (Carretero-Paulet and Van de Peer, 2020) involve the duplication of every gene in the genome, whereas those arising from small-scale duplications (SSDs) involve only one to

a few genes. Both WGDs and SSDs are widespread in plants and explain the existence of multigene families and of most of the variation in gene number between organisms, populations and species.

Although most gene duplicates are expected to neutrally accumulate deleterious mutations and eventually become a non-functional pseudogene or be eliminated from the genome, some of them are retained for longer evolutionary periods through the acquisition of specialized and/or novel biological functions. Indeed, many instances of them have been found at the basis of key morphological and metabolic adaptations of plants to their sessile lifestyle and of desirable agronomic traits (Vélez-Bermúdez et al., 2015; Quesada-Traver et al., 2022; Salojärvi et al., 2024). Therefore, understanding the mechanisms driving the long-term retention of genes after duplication, i.e., their duplicability, is crucial to understanding how changes at the genome level ultimately translate into phenotypic diversity and speciation.

Machine learning (Bishop and Nasrabadi, 2006; Murphy, 2012) techniques are now fundamental tools in many research areas. This field comprises a large amount of mathematical and statistical methods that have proven to be useful in a vast number of applications. In particular, Libbrecht and Noble (2015) showed several applications of machine learning techniques in genetics. Another relevant area is probabilistic machine learning (Bishop and Nasrabadi, 2006; Murphy, 2012; Ghahramani, 2015; Koller and Friedman, 2009), in which Bayesian Networks (BN), and more generally, probabilistic graphic models (PGMs), play a relevant role.

The purpose of this work is to make a comparative analysis of various BN classifiers for gene duplicability classification. In particular, we compare the performance of the Naive Bayes (NB), Tree Augmented Naive Bayes (TAN), and a non-restricted structure learned using the Hill-Climbing (HC) algorithm. These models are evaluated under different approaches to represent the conditional probability distributions (CPDs): discrete, conditional Gaussian (CG), and mixtures of polynomials (MoP). Moreover, we learned the aforementioned models with and without variable selection. The remainder of this paper is organized as follows. Section 2 describes the Bayesian Networks, the structures and CPD utilized, the variable selection method, the model validation approach, and the data description and preprocessing steps. The performance of the models is analyzed and discussed in Section 3. Finally, the main conclusions are presented in Section 4.

2 Methodology

From this point forward, uppercase letters will denote random variables, while lowercase letters will represent specific values of a random variable. Boldfaced characters will indicate random vectors (i.e., multidimensional random variables). The set of all possible values of a random vector $\boldsymbol{X}$ (i.e., its *support*) will be denoted by $\Omega_{\boldsymbol{X}}$.

2.1 Probabilistic graphical models: Bayesian networks

A *Bayesian Network* (BN) (Pearl, 1988) is a statistical multivariate model for a set of variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, in which independence relations are encoded by the

structure of an underlying *Directed Acyclic Graph* (DAG). More specifically, the DAG consists of nodes that represent random variables and directed edges between pairs of nodes, which indicate statistical dependencies. Each variable X_i, for $i = 1, \ldots, n$, is associated with a conditional probability distribution $p(x_i|Pa(x_i))$, given its parents in the DAG, denoted as $Pa(X_i)$. Consequently, the joint distribution of the random vector $\boldsymbol{X}$ factorizes as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|Pa(x_i)), \quad \forall x_1, \ldots, x_n \in \Omega_{X_1,\ldots,X_n}. \tag{1}$$

A simple example of a BN representing the joint distribution of five variables, X_1, X_2, X_3, X_4, and X_5, is illustrated in Figure 1. This network encodes the following factorization:

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3).$$

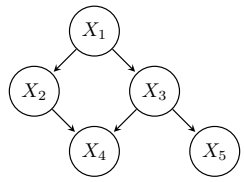


Figure 1: Example of a BN structure with five variables.

Originally proposed for discrete variables, BNs can also be used to model problems in continuous domains or hybrid domains, where both continuous and discrete variables coexist. There are several approaches to modelling hybrid domains with BNs. A commonly used method is to discretize the continuous variables, converting them into discrete ones and treating them as if they were inherently discrete. However, this approach may lead to a loss of information. Alternatively, to avoid discretization, the *Conditional Gaussian* (CG) model (Lauritzen, 1996) is widely used, despite imposing two major restrictions (i) continuous variables must follow a multivariate Gaussian distribution; (ii) discrete variables can not have continuous parents in the graph. These limitations have motivated the development of alternative approaches, such as *Mixtures of Truncated Basis Functions* (MoTBFs) (Langseth et al., 2012).

Formally, let $\boldsymbol{X}$ be a mixed n-dimensional random vector. Let $\boldsymbol{Y} = \{Y_1, \ldots, Y_d\}$ and $\boldsymbol{Z} = \{Z_c, \ldots, Z_c\}$ be the discrete and continuous parts of $\boldsymbol{X}$, respectively, with $d + c = n$. Let $\boldsymbol{\psi} = \{\psi_i(\cdot)\}_{i=0}^{\infty}$ with $\psi_i : \mathbb{R} \to \mathbb{R}$ define a collection of real basis functions. We say that a function $\hat{f} : \Omega_{\boldsymbol{X}} \to \mathbb{R}_0^+$ is a *mixture of truncated basis functions* potential to level k wrt. $\boldsymbol{\psi}$ if one of the following two conditions holds:

(i) $\hat{f}$ can be written as

$$\hat{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{y}, \boldsymbol{z}) = \sum_{i=0}^{k} \prod_{j=1}^{c} a_{i,\boldsymbol{y}}^{(j)} \psi_i(z_j), \tag{2}$$

where $a_{i,\boldsymbol{y}}^{(j)}$ are real numbers.

(ii) There is a partition $\mathcal{I}_1, \ldots, \mathcal{I}_m$ of $\Omega_{\boldsymbol{X}}$ for which the domain of the continuous variables, $\Omega_{\boldsymbol{Z}}$, is divided into hyper-cubes and such that f is defined as

$$f(\boldsymbol{x}) = f_l(\boldsymbol{x}) \quad \text{if} \quad \boldsymbol{x} \in \mathcal{I}_l,$$

where each f_l, $l = 1, \ldots, m$ can be written in the form of (2).

An MoTBF potential is said to be a *density* if $\sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{Y}}} \int_{\Omega_{\boldsymbol{Z}}} \hat{f}(\boldsymbol{y}, \boldsymbol{z}) d\boldsymbol{z} = 1$.

By defining $\boldsymbol{\psi} = \{\psi_i(z) = z^i\}_{i=0}^{\infty}$, the MoTBF model reduces to a Mixture of Polynomial (MoP) model (Langseth et al., 2012). For the purpose of comparing hybrid models, both CG and MoP distributions were used to represent the conditional probability distributions (CPDs) associated with the continuous variables in the model.

One of the most successful applications of Bayesian networks is *classification*, a prediction task in which a discrete target variable, C, referred to as the *class*, is forecasted based on the values of a set of *feature* variables, $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. The predicted value, c^*, of C is determined as the one that maximizes the posterior distribution of C given the observed values of the features.

$$c^* = \arg\max_{c \in \Omega_C} p(c|x_1, \ldots, x_n).$$

Note that, by Bayes' theorem

$$p(c|x_1, \ldots, x_n) = \frac{p(c)p(x_1, \ldots, x_n|c)}{p(x_1, \ldots, x_n)} \propto p(c)p(x_1, \ldots, x_n)|c),$$

which implies that solving the classification problem requires computing an n-dimensional distribution for $X_1, \ldots, X_n$ given C. This problem can be simplified by representing the joint distribution using a Bayesian network and leveraging the factorization encoded by its structure, as given in Equation (1).

Bayesian networks classifiers: NB and TAN

BNs are known for their high-quality performance in classification tasks and their ability to model complex probability distributions, including both discrete and continuous variables. In order to emphasize the importance of the class variable while reducing the number of parameters to be estimated from data, certain restricted structures are typically used, such as the *k-dependence Bayesian network classifiers* (kDB) (Sahami, 1996). In this type of model, the class variable is the parent of all other nodes, and each feature can have up to k additional parents besides the class. The *Naive Bayes* (NB) model and the *Tree Augmented Naive Bayes* (TAN) model are special cases of kDB, where $k = 0$ and $k = 1$, respectively.

Supposing that the feature variables of $\boldsymbol{X}$ are statistically independent of each other given the class variable, NB is the simplest classifier. Therefore, the joint distribution of

the NB model factorizes as

$$p(c, x_1, \ldots, x_n) = p(c) \prod_{i=1}^{n} p(x_i|c), \quad \forall c, x_1, \ldots x_n \in \Omega_{C,X_1,\ldots,X_n}.$$

Despite its strong independence assumption, NB often achieves remarkable classification performance while keeping a small number of parameters that need to be estimated. Figure 2a shows an example of the NB structure with four feature variables, where joint distribution is factorized as follows:

$$p(c, x_1, x_2, x_3, x_4) = p(c)p(x_1|c)p(x_2|c)p(x_3|c)p(x_4|c).$$

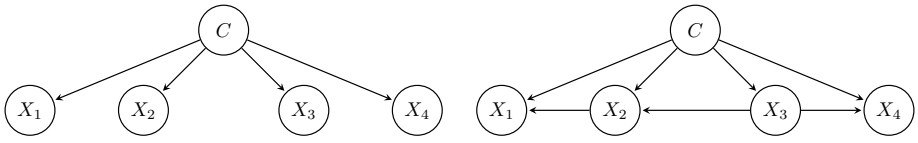


Figure 2: Examples of (a) NB with four feature variables, and (b) TAN with four feature variables and X_3 as root.

In order to improve the performance of classification problems, Friedman et al. (1997) introduced the *Tree Augmented Naive Bayes* (TAN) classifier, where each feature variable is allowed to have another feature variable as a parent, besides the class, as long as the resulting subgraph of feature variables forms a tree (i.e., it contains no directed cycles). The feature that has the class as its only parent is called the *root*. Therefore, the TAN model relaxes the independence assumption of NB. An example of a TAN structure with four feature variables is shown in Figure 2b, and its probability distribution factorizes as:

$$p(c, x_1, x_2, x_3, x_4) = p(c)p(x_1|c, x_2)p(x_2|c, x_3)p(x_3|c)p(x_4|c, x_3).$$

In general, kDB structures impose certain restrictions on the DAG from the outset. In particular, NB and TAN are among those with the most constraints. To relax these limitations, we also learned a model according to a non-restricted structure.

Non-restricted structure learning Bayesian Networks

Despite the performance of kDB, the structure of the DAG is restricted for each k. A non-restricted structure can be learned using the *Hill-Climbing algorithm* (HC) (Russell and Norvig, 2016). The objective of structure learning is to find the best DAG, $\mathcal{G}^*$, such that

$$\mathcal{G}^* = \arg\max_{\mathcal{G}} f(\mathcal{G}|D), \tag{3}$$

where $f(\mathcal{G}|D)$ is a scoring metric that evaluates the merit of any candidate DAG $\mathcal{G}$.

Being an iterative algorithm, HC usually starts with an empty graph (or a specific structure from an expert), and in each iteration, three possible actions (addition, deletion, or reversal of an edge between $X_i \rightarrow X_j$) are applied to improve the score of the structure. The *Bayesian Information Criterion* (BIC) was applied to maximize the scoring metric f in Equation (3). An example of DAG learned by the HC algorithm is shown in Figure 3.

2.2 Variable selection

We used two different strategies for variable selection based on the structure of the BN classifier. Firstly, in the case of HC, the Markov blanket (Koller and Friedman, 2009) of the class variable C was used as the method for variable selection. The Markov blanket of a node, consisting of its parents, its children, and the other parents of its children, d-separates C from the remaining nodes in the network. Consequently, once the Markov blanket is observed, the remaining feature variables provide no additional information for computing the probability of C. Figure 3 shows an example of the Markov blanket of the variable C.

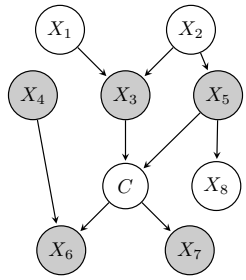


Figure 3: Example of a DAG learned using the *Hill-Climbing* algorithm. Variables in the Markov blanket of C are shaded in grey.

Secondly, a filter-wrapper approach was used for variable selection in both the NB and TAN models. Unlike DAGs learned using the HC algorithm, in the case of NB and TAN, all feature variables belong to the Markov blanket of the class C. Therefore, the Markov blanket cannot be applied for feature selection in this case. The filter-wrapper approach involves ranking the feature variables according to their conditional entropy with respect to the class, resulting in an ordered set of feature variables. The first variable in this set is used to train the initial model, and the remaining feature variables are sequentially inserted, one by one, in the specified order. Whenever the inclusion of a variable increases the accuracy of the model, it is retained; otherwise, it is excluded.

2.3 Validation of the model

In order to evaluate the predictive capability of the various models, the k-fold cross-validation technique (Stone, 1974) was used. This technique randomly splits the dataset into k subsets, and the method is repeated k times. In each iteration, one subset is used

to test the model trained on the remaining $k-1$ subsets. In this paper, a $k-$value of 10 was applied, and the global accuracy of the classification was taken as the performance metric.

Then, the ten accuracy measures of the models were compared in two ways: first, across the different DAG learning algorithms (i.e., NB, TAN, or HC), and second, across the different CPD representations (discrete, CG and MoP). In other words, we compared the performance of models with the same CPD representation but learned using different DAG learning algorithms (and vice-versa), to see how the choice of DAG algorithm (or CPD representation) affects the accuracy of the model. These comparisons were performed using the Friedman test, followed by the Wilcoxon signed-rank test for pairwise comparisons. For each test, a significance level of $\alpha = 0.05$ was considered.

2.4 Data description and pre-processing

Data were mined from different repositories of biological sequence data, including National Center for Biotechnology Information (NCBI) [1], UNIPROT [2] or INTERPRO [3]; comparative genomics platforms, including CoGe [4], PLAZA [5], or Phytozome [6]; species-specific genomics resources, such as TAIR (The Arabidopsis Information Resource [7]); biological network resources, including AraNet [8], STRING [9], or GEO (Gene Expression Omnibus, [10]); genome functional annotation databases, e.g. AraCyc [11]; and proteomic databases, e.g., qPTMplants [12].

The dataset built consists of 743 variables (continuous, binary and nominal) and each variable describes a different characteristic of each gene, including:

- *sequence-based features*: exon/intron number and length, nucleotide or amino acid composition, GC content, codon usage, sequence divergence, 5-methylcytosine hypermutations, selection rates, etc.
- *biological-based features*: phylogenetic distribution, AraCyc metabolic pathways, Plant Slim GO terms, expression breadth and level, KEGG biochemical pathways, essentiality, phenotype, etc.
- *molecular-based features*: epigenetic signatures and motifs notably methylation, INTERPRO protein functional domains, protein post-translational modifications, protein transmembrane domains, protein solvent accessibility, protein 2ary structure, protein molecular weight and isoelectric point, protein subcellular localization, protein-protein interactors, etc.

This dataset contains over 27,000 observations, and the class variable, *Duplicability*, is distributed as *Yes* (83%) and *No* (17%). In addition, some of the continuous variables had missing values, which were replaced by their respective mean values, and the only binary

[1] https://www.ncbi.nlm.nih.gov/ [2] https://www.uniprot.org/
[3] https://www.ebi.ac.uk/interpro/ [4] https://genomevolution.org/coge/
[5] https://bioinformatics.psb.ugent.be/plaza/ [6] https://phytozome.jgi.doe.gov/pz/portal.html
[7] https://www.arabidopsis.org/ [8] https://www.inetbio.org/aranet/ [9] https://string-db.org/
[10] https://www.ncbi.nlm.nih.gov/geo/ [11] https://www.plantcyc.org/databases/aracyc/15.0
[12] http://qptmplants.omicsbio.info/

variable with missing values was transformed into a nominal variable with three values: *Yes*, *No*, and *Unknown*. To fulfil the objective of this work – comparing BN models in hybrid domains using different strategies for CPD representation, namely discrete, CG and MoP distributions – a discretization method was chosen based on gain ratio maximization, with at most five intervals (Catlett, 1991).

3 Results and discussion

In this work, we have built 15 different classification models, one for each combination of CPD representation and DAG learning algorithm (including both selective and non-selective versions). For parameter learning, models based on discrete or MoP distributions have been learned using the MoTBFs package (Pérez Bernabé et al., 2020), whereas models based on CG distributions have been learned using the bnlearn package (Scutari, 2010).

Regarding structural learning, the topology for both discrete and MoP models was learned from the discretized dataset, while for the CG model, it was learned from the original hybrid dataset. The implementation of the HC algorithm from the R package bnlearn was used in all cases. The TAN implementation of bnlearn was used for the discrete and MoP models, while a modified TAN implementation, following the Chow-Liu scheme, was employed for the CG model to account for its topological restrictions.

A 10-fold cross-validation was carried out to test the predictive performance of the classifiers. The mean accuracy of the models, computed from the k-fold cross validation, along with the standard deviation, are shown in Table 1. The highest mean accuracy is observed for the selective-NB with a CG representation (0.8675), closely followed by the selective-TAN with MoP representation (0.8673). Only discrete NB, MoP TAN, and CG NB models yielded a global accuracy lower than the baseline frequency of *Duplicability Yes* class (83%). Regarding variability, the standard deviations are relatively low for most models, with TAN model using MoPs exhibiting the highest variability.

DAG learning algorithm	CPD representation		
	Discrete	CG	MoP
HC	0.8574 (0.0054)	0.8464 (0.0092)	0.865 (0.0043)
NB	0.7884 (0.0065)	0.8191 (0.0055)	0.8458 (0.0048)
TAN	0.8344 (0.0064)	0.8331 (0.0079)	0.8013 (0.0493)
Sel. NB	0.8347 (0.0056)	**0.8675** (**0.0030**)	0.8646 (0.0053)
Sel. TAN	0.8425 (0.0078)	0.8538 (0.0063)	0.8673 (0.0051)

Table 1: Mean classification accuracy and standard deviation (in parenthesis) for each trained model. The highest overall performance is highlighted in bold. MoP: Mixtures of Polynomials; CG: Conditional Gaussian; Sel. NB: Selective-NB; Sel. TAN: Selective-TAN.

In order to compare the accuracy among the CPD representations (Discrete, MoP, and CG) within each DAG learning algorithm (HC, NB, Selective-NB, TAN, Selective-TAN), the Friedman test was applied. Whenever significant differences were found, a posthoc Wilcoxon sign-rank test was conducted to compare pairs of CPD representations. Figure 4 illustrates these results for the selective models only (HC, Selective-NB, and Selective-TAN).

For each DAG learning algorithm, the discrete CDP representation is consistently outperformed by the hybrid ones (p-value < 0.05). In the case of hybrid representations, the MoP CPD is never outperformed by CG and shows better performance in HC and Selective-TAN (p-value ≥ 0.05). As for the non-selective TAN model, the Friedman test did not show any significant difference between the CPD representations (p-value ≥ 0.05).

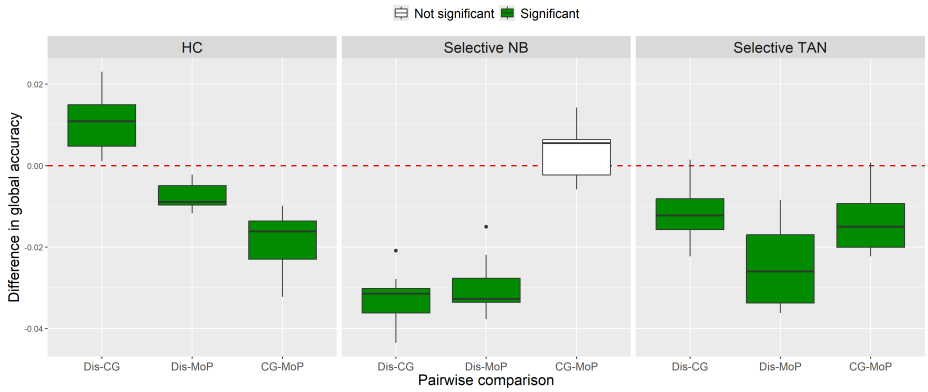


Figure 4: Box plots comparing model performance across CPD representations for each selective DAG learning algorithm considered. Green-filled boxes indicate significant differences (post hoc analysis after Friedman test). Dis: Discrete; CG: Conditional Gaussian; MoP: Mixture of Polynomials.

On the other hand, the Friedman test was applied to compare the classification accuracy among the DAG learning algorithms within each CPD representation. In general, DAG learning algorithms with feature selection are never outperformed by their counterparts without variable selection, at the 5% significance level. Figure 5 illustrates these results for the selective models only (HC, Selective-NB, and Selective-TAN).

Based on these results, the following observations can be made:

- In the case of discrete CPD, the HC structure is never outperformed and shows no significant difference only when compared to the Selective-TAN structure. Furthermore, the Selective-NB structure is outperformed by both the HC and Selective-TAN structures.

- In the case of MoP distributions, the post hoc analysis did not reveal significant differences between structures with feature selection (p-values ≥ 0.05). In particular,

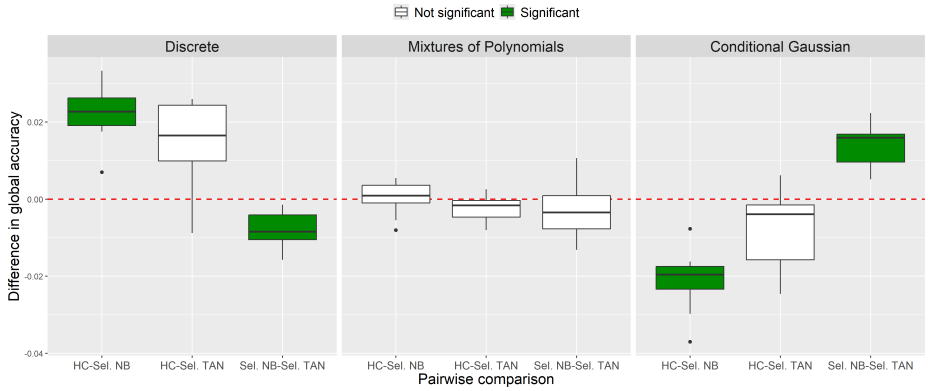


Figure 5: Box plots comparing model performance among selective DAG learning algorithms for each CPD representation considered. Green-filled boxes indicate significant differences (post hoc analysis after Friedman test). Sel. NB: Selective-NB; Sel. TAN: Selective-TAN.

HC outperforms models without variable selection, but fails to show significant differences when compared to Selective-NB o Selective-TAN. Similarly, Selective-NB performs better than the non-selective models (NB and TAN) but shows no significant difference with respect to the other 2 selective models. The same behaviour is observed for the Selective-TAN model.

- Regarding models using CG distributions, Selective-NB consistently outperformed the other structures (p-value < 0.05). Furthermore, Selective-TAN was not outperformed by any of the remaining topologies. Similarly, HC performed better than the non-selective models, worse than the Selective-NB, and comparably to the Selective TAN.

4 Conclusion

This study enabled a comparative analysis of Bayesian network classifiers with different topologies and CPD representations. Overall, DAG learning algorithms with feature selection were never outperformed by their counterparts without variable selection. In addition, hybrid models generally outperform discrete ones when continuous variables are discretised. Concerning hybrid models, BNs based on the MoP distribution are never outperformed by those using the CG model, and performed better in three out of the five DAG learning algorithms considered. Furthermore, the MoP representation provides some advantages over the CG model, as it does not impose structural restrictions and is flexible enough to fit a wide range of distributions.

One of the main expected impacts of the approach described in this paper was to obtain models that allow us to predict the duplicability of de novo annotated genes in

a newly sequenced genome, including those of wild or orphan species or varieties—i.e., those traditionally understudied and underutilised but with great nutritional, agronomic, and adaptive potential to local and changing environmental conditions (Carretero-Paulet et al., 2025). As future work, the model will be tested in distant taxonomic groups, such as animals or yeast. It is interesting to recall that repeated gene duplications in the human genome result in so-called copy number variations (CNVs), which are commonly associated with various neurological diseases and forms of cancer (Rice and McLysaght, 2017).

Acknowledgments

The work of R. Rumí and A. Maldonado is supported by grant PID2022-139293NB-C31 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU "A way of Making Europe", by the research group FQM-244, grant PPIT-UAL, Junta de Andalucía-ERDF 2021-2027, Objective RSO1.1. Programme: 54.A , and by the Center for Development and Transfer of Mathematical Research to Industry CDTIME (University of Almería). The work of L. Carretero-Paulet and A. Gálvez-Salido is supported by a "Proyectos I+D Generación de Conocimiento" grant from the Spanish Ministry of Science and Innovation (grant code: PID2020-113277GB-I00) and by the University of Almería Research and Transfer Programme 54.A funded by "Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía" through the European Regional Development Fund (ERDF), 2021-2027 to LCP.

References

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

L. Carretero-Paulet and Y. Van de Peer. The evolutionary conundrum of whole-genome duplication. *American journal of botany*, 107(8):1101, 2020.

L. Carretero-Paulet, A. J. Mendoza-Fernández, F. J. Alcalá, and A. J. Castro. Leveraging agrobiodiversity for sustainable transition in greenhouse-based intensive agriculture across mediterranean drylands. *Journal of Arid Environments*, 228:105354, 2025.

J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal, March 6–8, 1991 Proceedings 5*, pages 164–178. Springer, 1991.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29:131–163, 1997.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521 (7553):452–459, 2015. doi: 10.1038/nature14541.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

H. Langseth, T. D. Nielsen, R. Rumı, and A. Salmerón. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning*, 53(2):212–227, 2012.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015. doi: 10.1038/nrg3920.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

J. Pearl. Probabilist reasoning in intelligent systems. *Morgan Faufman*, 1988.

I. Pérez Bernabé, A. D. Maldonado González, A. Salmerón Cerdán, T. D. Nielsen, et al. Motbfs: An r package for learning hybrid bayesian networks using mixtures of truncated basis functions. 2020.

C. Quesada-Traver, A. Lloret, L. Carretero-Paulet, M. L. Badenes, and G. Ríos. Evolutionary origin and functional specialization of dormancy-associated mads box (dam) proteins in perennial crops. *BMC Plant Biology*, 22(1):473, 2022. doi: 10.1186/s12870-022-03856-7.

A. M. Rice and A. McLysaght. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nature communications*, 8(1):14366, 2017.

S. J. Russell and P. Norvig. *Artificial intelligence : a modern approach / Stuart J. Russell, Peter Norvig*. Prentice Hall series in Artificial Intelligence. Pearson, Boston, 3rd ed. [global ed.] edition, 2016. ISBN 9781292153964.

M. Sahami. Learning limited dependence bayesian classifiers. In *KDD*, volume 96, pages 335–338, 1996.

J. Salojärvi, A. Rambani, Z. Yu, R. Guyot, S. Strickler, M. Lepelley, C. Wang, S. Rajaraman, P. Rastas, C. Zheng, et al. The genome and population genomics of allopolyploid coffea arabica reveal the diversification history of modern coffee cultivars. *Nature genetics*, 56(4):721–731, 2024. doi: 10.1038/s41588-024-01695-w.

M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of statistical software*, 35:1–22, 2010.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

I. Vélez-Bermúdez, L. Carretero-Paulet, T. Legnaioli, D. Ludevid, M. Pagès, and M. Riera. Novel ck2α and ck2β subunits in maize reveal functional diversification in subcellular localization and interaction capacity. *Plant Science*, 235:58–69, 2015. ISSN 0168-9452. doi: https://doi.org/10.1016/j.plantsci.2015.03.005. URL `https://www.sciencedirect.com/science/article/pii/S0168945215000667`.

Block-Coordinate Descent Algorithm for Interventional Data in Directed Graphical Models

Jun Wu[1]

[1]Institute of Computer Science, Czech Academy of Science
wu@cs.cas.cz

Abstract

Computing maximum likelihood estimates in linear structural equation models is generally a difficult problem. The critical equations are usually non-linear and have numerous solutions, even for purely observational data. The block-coordinate descent (BCD) algorithm proposed by Drton et al. (2019)[1] is an efficient way to solve the optimization problem by decomposing it into a series of sub-problems with closed-form solutions, and which works with observational data. In this work, we describe the general problem of a BCD-type scheme for computing maximum likelihood estimates in linear structural equation models without hidden variables, integrating multiple observational and interventional environments. With interventional data, the degrees of both the original likelihood equations and the block-coordinate update equations could increase greatly. We study special setups in which the block optimization subproblems have a degree of at most 2 and provide closed-form solutions in these cases. Additionally, we discuss the potential applications of the model and algorithm to health and well-being data.

1 Introduction

Structural equation models (SEMs) encode the cause-effect relationships between random variables and error terms , and are widely used across various fields. Naturally, a structural equation model is associated with a directed (or mixed) graph, where the edges represent relations between variables. Research on structural equation models dates back to Wright's path diagrams (Wright, 1921, 1934), and Haavelmo's simultaneous equations (Haavelmo, 1943), with more recent integration into a general framework for causal modeling (Spirtes et al., 2000; Pearl, 2009).

In this work, we address the problem of computing maximum likelihood estimates (MLEs) in linear SEMs with Gaussian errors, using both observational and interventional data. For recursive SEMs with independent errors (represented by a directed

acyclic graph, DAG), the computation of MLEs with interventional data follows a similar approach as with observational data, involving regressing each node on its parental nodes using data from environments where the variable Y_i is not intervened (Hauser and Bühlmann, 2012; Hauser and Bühlmann, 2015). However, more general models may include bidirected edges and cycles, corresponding to hidden variables and self-regulatory feedback loops, respectively. For these models, quasi-Newton optimization and blockwise partial optimization methods have been proposed for purely observational scenarios. The R packages "sem" (Fox, 2006) and "lavaan" (Rosseel, 2012) use quasi-Newton methods. The residual iterative conditional fitting (RICF) algorithm gives a closed-form block update for acyclic directed mixed graph models (Drton et al., 2009). Recently, the block-coordinate descent (BCD) algorithm is proposed in Drton et al. (2019), extending the RICF algorithm to graphs with cycles.

Interventions across different environments induce variations in both the graph and equation structures, significantly increasing the complexity of the log-likelihood function and its optimization. Even in the purely observational case, the likelihood equations are typically high-degree algebraic functions of the data (Drton et al., 2019). The BCD algorithm attempts to perform low degree partial optimizations at each step for observational data. Extending the algorithm to accommodate both observational and interventional data is of great interest. However, for this extension, we focus on directed cyclic graphs without bidirected edges.

We consider linear SEMs associated with a directed graph and address the general question of BCD-type optimization. Several concrete examples of intervention targets, linear SEMs, and directed graphs are provided, along with the joint log-likelihood, critical point equations, and maximum likelihood degrees. For certain kinds of graphs and special intervention targets on arbitrary directed graphs, we show that the block update problem is a quadratic equation with degree 2. This indicates that the BCD-type method can still be applied to interventional data under certain conditions.

2 Linear structural equation models

2.1 Background

A structural equation model (SEM) is a equation system involving variables $Y_i : i \in V$ and stochastic errors $\{\epsilon_i : i \in V\}$, where V is the set of variables and $|V| = p$. It describes the quantitative mechanism by which a variable Y_i depends on other variables and their associated error. In this work, we adopt the notations introduced in Drton et al. (2019) and focus on linear SEMs with independent errors:

$$Y_i = \sum_{j \in V \setminus \{i\}} \beta_{ij} Y_j + \epsilon_i, \quad i \in V. \tag{1}$$

The coefficients can be summarized in an edge weight matrix $B = (\beta)_{ij}$ and we can express Y and ϵ in vector form

$$Y = BY + \epsilon, \quad \epsilon \sim N(0, \Omega),$$

where Ω is a (positive definite) diagonal matrix, and $\omega = \text{diag}(\Omega)$ denotes it diagonal part.

An SEM can be represented by a directed graph $G = (V, E)$, where the node set V corresponds to the random variables and the edge set E consists of ordered node pairs. A pair $(i, j) \in E$ defines a directed edge $i \to j \in G$, implying that Y_i has a causal effect on Y_j. In this context, we refer to i as a parent of j and j as a child of i: $i \in \text{pa}(j)$, $j \in \text{ch}(i)$. The weights of the parental edges of i are denoted by $B_{i,\text{pa}(i)}$. SEMs are not always recursive, meaning the associated graph may contain feedback loops. The strongly connected component (SCC) in a directed graph is a maximal subgraph in which there is a directed path between any pair of nodes. The strongly connected component that includes i in a directed graph G is denoted by $\mathcal{C}(i, G)$. When the graph is clear from context, we simply use $\mathcal{C}(i)$.

2.2 Interventional distributions

We are interested in data collected under different interventions. The type of interventions we consider are "hard" interventions that fix the values of the intervened variables in a randomized fashion so that they follow a controlled probability distribution (Pearl, 2009; Spirtes et al., 2000).

Throughout this paper, we use $I \subseteq V$ to denote the intervention target in one interventional environment, i.e., the set I indexes the intervened variables. Given an intervention target I, the manipulated graph $G_{\overline{I}}$ is obtained by removing all edges pointing to nodes in I from G, representing the structure of the SEM after the intervention. The collection $\mathcal{I} \subset 2^V$ denotes the family of intervention targets I_k's across all environments for which data is available. To distinguish from the entry indices in data matrices, we use (k)-superscripts to specify quantities from different interventional environments: $Y^{(k)} \in \mathbb{R}^{p \times n^{(k)}}$ is the data matrix for intervention target I_k, where each column is one sample and $n^{(k)}$ is the sample size; the parameters $(\Omega^{(k)}, B^{(k)})$ differs because of interventions.

Example 1. *In Figure 1, the linear SEM in the observational environment is*

$$\begin{cases} Y_1 &= \epsilon_1, \\ Y_2 &= \beta_{21} Y_1 + \beta_{24} Y_4 + \epsilon_2, \\ Y_3 &= \beta_{32} Y_2 + \epsilon_3, \\ Y_4 &= \beta_{41} Y_1 + \beta_{43} Y_3 + \epsilon_4, \end{cases}, \quad (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)^T \sim N(\mathbf{0}, \text{diag}(\omega_1, \omega_2, \omega_3, \omega_4)).$$

With intervention target $I = \{2\}$, the manipulated linear SEM becomes

$$\begin{cases} Y_1 &= \epsilon_1, \\ Y_2 &= \epsilon_2', \\ Y_3 &= \beta_{32} Y_2 + \epsilon_3, \\ Y_4 &= \beta_{41} Y_1 + \beta_{43} Y_3 + \epsilon_4, \end{cases}, \quad (\epsilon_1, \epsilon_2', \epsilon_3, \epsilon_4)^T \sim N(\mathbf{0}, \text{diag}(\omega_1, \omega_2', \omega_3, \omega_4)).$$

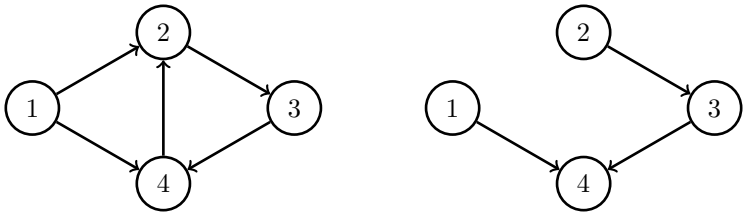


Figure 1: Original graph, and manipulated graph with intervention target $I = \{2\}$.

2.3 Likelihood inference

The log-likelihood of a single interventional dataset k is a function of $(\Omega^{(k)}, B^{(k)})$:

$$\ell_{G,Y}(\Omega^{(k)}, B^{(k)}) = -\log\det(\Omega^{(k)}) - \log\det(I - B^{(k)})^2 - \operatorname{tr}\left\{(I - B^{(k)})^T(\Omega^{(k)})^{-1}(I - B^{(k)})S^{(k)}\right\},$$

where $S^{(k)} = Y^{(k)}(Y^{(k)})^T/n^{(k)}$ is the sample covariance matrix of the k'th environment. And the total log-likelihood is

$$\ell_{G,Y^{(1)},\ldots,Y^{(K)}}(\Omega, B) = \sum_{k}^{K} n^{(k)} \cdot \ell_{G,Y^{(k)}}(\Omega^{(k)}, B^{(k)}). \tag{2}$$

To find the critical point(s), we take derivatives to derive the likelihood equations, as demonstrated in Proposition 1 of Drton et al. (2019). Typically, the likelihood equations are of high degree. In example 1, the equations for $\mathcal{I} = \{\emptyset\}$ has degree 6 and the equations for $\mathcal{I} = \{\emptyset, \{2\}\}$ has degree 9.[1] Following the procedures outlined in Drton and Richardson (2004); Drton et al. (2019), we apply the block-coordinate descent method to break the original problem into partial subproblems of lower degrees.

3 Block-coordinate descent

3.1 Block update problem

For each node i, the total log-likelihood (2) can be written as a function of $(\omega_{ii}, B_{i,\mathrm{pa}(i)})$

$$\ell_{G,Y^1,\ldots,Y^K}(\omega_{ii}, B_{i,\mathrm{pa}(i)}) = \sum_{k:i\notin I_k} \left(-n^{(k)}\log\omega_{ii}^{(k)} - \frac{1}{\omega_{ii}^{(k)}}\|Y_i^{(k)} - B_{i,\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^{(k)}\|^2 + n^{(k)}\log[(c_{i,0}^{(k)} + B_{i,\mathrm{pa}(i)}c_{i,\mathrm{pa}(i)}^{(k)})^2] \right). \tag{3}$$

This expression follows the likelihood formula in Drton et al. (2019) and is based on the result that $\det(I - B^{(k)}) = c_{i,0}^{(k)} + B_{i,\mathrm{pa}(i)}c_{i,\mathrm{pa}(i)}^{(k)}$. [2]

[1] See the accompanying Mathematica script.
[2] See Lemma 2 in Drton et al. (2019).

If $Y_i^{(k)} - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^{(k)} \neq 0$ for each k such that $i \notin I_k$, then

$$(\omega_{ii}^{(k)})^* = \frac{1}{\sum_{k:i\notin I_k} n^{(k)}} \sum_{k:i\notin I_k} n^{(k)} \|Y_i^{(k)} - B_{i,\mathrm{pa}(i)}^{(k)} Y_{\mathrm{pa}(i)}^{(k)}\|^2$$

maximizes the total log-likelihood with respect to $\omega_{ii}^{(k)}$. This leads to the following profile log-likelihood function for the parameter vector $B_{i,\mathrm{pa}(i)}$:

$$\ell(B_{i,\mathrm{pa}(i)}) = -\sum_{k:i\notin I_k} n^{(k)} \log \frac{\sum_{k:i\notin I_k} \|Y_i^{(k)} - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^{(k)}\|^2}{(c_{i,0}^{(k)} + B_{i,\mathrm{pa}(i)} c_{i,\mathrm{pa}(i)}^{(k)})^2}. \tag{4}$$

3.2 Degree 2 update, condition and formula

In general, the block update problem is challenging to solve. Unless the graph structure is restricted, a closed-form update is only achievable under specific conditions on the interventions. For node i, such a closed-form update is possible if the following sufficient condition is met:

$$\exists\, G' \subseteq G, \;\; \text{s.t. } \forall\, I \in \mathcal{I} \text{ and } i \notin I,\; \mathcal{C}(i, G_{\overline{I}}) = G' \text{ or } V[\mathcal{C}(i, G_{\overline{I}})] = \{i\}, \tag{5}$$

where the operator $V[\cdot]$ returns the set of the nodes in a (sub)graph. In other words, the condition requires that there could be at most 2 different structures of the strongly connected component containing i, and one of which is the singleton set $\{i\}$. For directed graphs where each strongly connected component contains **at most one cycle** (meaning any two cycles in the graph are disjoint), condition (5) is automatically satisfied for all nodes and any intervention family $\mathcal{I}$.

All the valid samples ($i \notin I_k$) are divided into these two groups, with total sample sizes n_1 and n_2. To estimate $(\omega_{ii}, B_{i,\mathrm{pa}(i)})$, we stack these data matrices $Y^{(k)}$ column-wise into a single matrix Y. If the $\mathcal{C}(i, G_{\overline{I}})$'s take values in $\{i\}$ and some G' across all I's, the profile log-likelihood function after optimizing over ω_{ii}, is given by

$$g_i(B_{i,\mathrm{pa}(i)}) = -n_1 \log\left(\frac{\|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}\|^2}{(c_{i,0} + B_{i,\mathrm{pa}(i)} c_{i,\mathrm{pa}(i)})^2}\right) - n_2 \log\left(\|Y_i - B_{i,\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}\|^2\right) + C.$$

Maximizing the profile log-likelihood function is equivalent to this minimization problem

$$\min_{\alpha\in\mathbb{R}^{|\,\mathrm{pa}(i)|}} n_1 \log \frac{\|Y_i^T - Y_{\mathrm{pa}(i)}^T \alpha\|^2}{(c_{i,0} + c_{i,\mathrm{pa}(i)}^T \alpha)^2} + n_2 \log\left(\|Y_i^T - Y_{\mathrm{pa}(i)}^T \alpha\|^2\right). \tag{6}$$

Next, we demonstrate that by applying the reparameterization techniques outlined in Drton et al. (2019), the minimization problem in (6) admits closed-form solutions with algebraic degree 2.

Theorem 1. *Given a node i, let $n_1, n_2 > 0$ be the total numbers of data corresponding to strongly connected components G' and $\{i\}$, respectively, and let $r = n_1/n_2$. Suppose*

that the stacked partial data matrix $Y_{\mathrm{pa}(i)\cup\{i\}}$ *has full rank* $|\,\mathrm{pa}(i)|+1 \le n_1+n_2$. *Let* $\hat{\alpha} = (Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}Y_{\mathrm{pa}(i)}Y_i^T$ *be the minimizer of* $\|Y_i^T - Y_{\mathrm{pa}(i)}^T\alpha\|^2$. *We define* $n := n_1+n_2, m := |\,\mathrm{pa}(i)|, c_0 := c_{i,0}$ *and* $c_1 := c_{i,\mathrm{pa}(i)} \neq 0$, $y_0^2 := \|Y_i^T - Y_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2$ *and* $l^2 := c_1^T(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1$. *Then the solution of the optimization problem in* (6) *satisfies*

$$\alpha^* = \hat{\alpha} + \delta\cdot(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1, \tag{7}$$

where δ *is a solution to the quadratic equation*

$$l^2\delta^2 + (c_1^T\hat{\alpha}+c_0)(r+1)\delta - ry_0^2 = 0. \tag{8}$$

Proof. By adopting the orthogonal transformation method in Drton et al. (2019), we further derive auxiliary properties relevant to our problem.

First, we find an orthogonal $m\times m$ matrix Q_1 such that $Q_1c_1 = (0,\ldots,0,\|c_1\|)^T$. Then, we compute a QR decomposition $Y_{\mathrm{pa}(i)}^TQ_1^T = Q_2^TR$, with $Q_2\in\mathbb{R}^{N\times m}$ orthogonal and $R\in\mathbb{R}^{N\times m}$ upper triangular. Since $Y_{\mathrm{pa}(i)}$ and $Y_{\mathrm{pa}(i)}^TQ_1^T$ have full ranks, we can assume that all diagonal entries of R are positive, making the matrix R unique for any given Q_1. After reparameterizing $\alpha' = Q_1\alpha$, the common L_2-norm term is transformed to

$$\begin{aligned} y_0^2 &= \|Y_i^T - Y_{\mathrm{pa}(i)}^T\alpha\|^2 = \|Q_2Y_i^T - R\alpha'\|^2 \\ &= \sum_{j=1}^{m}[(Q_2Y_i^T)_j - (R\alpha')_j]^2 + \sum_{j=m+1}^{N}(Q_2Y_i^T)_j^2, \end{aligned}$$

and the denominator is transformed to $(c_0+\|c_1\|\alpha'_m)^2$.

Since $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ with $R_1\in\mathbb{R}^{m\times m}$, we reparameterize again with $\alpha'' = R_1\alpha'$ and the original minimization problem becomes

$$\min_{\alpha''\in\mathbb{R}^m} n_1\log\frac{\sum_{j=1}^m[(Q_2Y_i^T)_j-\alpha''_j]^2+\sum_{j=m+1}^n(Q_2Y_i^T)_j^2}{(c_0+\|c_1\|R_{mm}^{-1}\alpha''_m)^2} + n_2\log\left(\sum_{j=1}^m[(Q_2Y_i^T)_j-\alpha''_j]^2+\sum_{j=m+1}^n(Q_2Y_i^T)_j^2\right). \tag{9}$$

Any solution must satisfies that $\alpha''_j = (Q_2Y_i^T)_j$ for $j\in[m-1]$. The optimal value of α''_m is given by

$$\operatorname*{argmin}_{\alpha''_m\in\mathbb{R}}\left\{n_1\log\frac{[(Q_2Y_i^T)_m-\alpha''_m]^2+\sum_{j=m+1}^n(Q_2Y_i^T)_j^2}{(c_0+\|c_1\|R_{mm}^{-1}\alpha''_m)^2} + n_2\log\left([(Q_2Y_i^T)_m-\alpha''_m]^2+\sum_{j=m+1}^n(Q_2Y_i^T)_j^2\right)\right\}, \tag{10}$$

i.e., maximizing

$$g_i(x) := n_1\log\left(x+\frac{c_0R_{mm}}{\|c_1\|}\right)^2 - (n_1+n_2)\log\left(x^2-2(Q_2Y_i^T)_mx+\sum_{j=m}^n(Q_2Y_i^T)_j^2\right)+C.$$

The univariate function g_i has derivative

$$g_i'(x) = \frac{2n_1}{x + c_0 R_{mm}/\|c_1\|} - 2(n_1+n_2)\frac{x - (Q_2 Y_i^T)_m}{x^2 - 2(Q_2 Y_i^T)_m x + \sum_{j=m}^N (Q_2 Y_i^T)_j^2}.$$

Let $a = 1, b = -(Q_2 Y_i^T)_m, c = \sum_{j=m}^n (Q_2 Y_i^T)_j^2$ and $\lambda = \|c_1\|/R_{mm} \neq 0$. The equation $g_i'(x) = 0$ has the form $ax^2 + 2bx + c = 0$. Using $r = n_1/n_2$, the two solutions are

$$\begin{aligned}\alpha_m'' &= \frac{b\lambda(r-1) - ac_0(r+1) \pm \sqrt{(b\lambda(r-1) - ac_0(r+1))^2 + 4a\lambda(c\lambda r - bc_0(r+1))}}{2a\lambda} \\ &= -\frac{b}{a} + \frac{(b\lambda - ac_0)(r+1) \pm \sqrt{(b\lambda - ac_0)^2(r+1)^2 + 4(ac - b^2)\lambda^2 r}}{2a\lambda}.\end{aligned}$$

The optimal solution in original coordinates is $\alpha = Q_1^T R_1^{-1} \alpha''$. Since $R_1^{-1}(Q_2 Y_i^T)$ is the linear regression coefficient vector of $Y_{\mathrm{pa}(i)}^T Q_1^T$ on Y_i^T, we have

$$Q_1^T R_1^{-1}(Q_2 Y_i^T)_{[m]} = (Y_{\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^T)^{-1} Y_{\mathrm{pa}(i)} Y_i^T := \hat{\alpha}.$$

Let $e_{m,m} = (0, \ldots, 0, 1)$ be the m-th canonical basis vector of $\mathbb{R}^m$ and let $e_{m,N}$ be the m-th canonical basis vector of $\mathbb{R}^N$ (i.e., the vector with a 1 in the m-th position and zeros elsewhere). Noticing that $R_{mm}^{-1}(Q_2 Y_i^T)_m$ is the m-th entry of $R_1^{-1}(Q_2 Y_i^T)_{[m]}$, and the last column of R_1^{-T} is $R_{mm}^{-1} e_{m,m}$, we can derive that

$$\lambda \cdot b = -\|c_1\| R_{mm}^{-1}(Q_2 Y_i^T)_m = -\langle Q_1 c_1, R_1^{-1}(Q_2 Y_i^T)_{[m]} \rangle = -\langle c_1, Q_1^T R^{-1}(Q_2 Y_i^T)_{[m]} \rangle = -c_1^T \hat{\alpha},$$

and

$$\begin{aligned}Q_1^T R_1^{-1} \|c_1\| R_{mm}^{-1} e_{m,m} &= Q_1^T R_1^{-1} R_1^{-T} Q_1 c_1 = (Q_1^T R^T R Q_1)^{-1} c_1 \\ &= (Q_1^T R^T Q_2 Q_2^T R Q_1)^{-1} c_1 = (Y_{\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^T)^{-1} c_1.\end{aligned}$$

The matrices Q_1, Q_2, and R may vary, but the value of R_{mm} (or equivalently, λ) is uniquely determined by $Y_{\mathrm{pa}(i)}$ and c_1. To see this, note that

$$\begin{aligned}Y_{\mathrm{pa}(i)}^T (Y_{\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^T)^{-1} c_1 &= Y_{\mathrm{pa}(i)}^T Q_1^T R_1^{-1} \|c_1\| R_{mm}^{-1} e_{m,m} = Q_2^T \begin{pmatrix} R_1 \\ 0 \end{pmatrix} R_1^{-1} \|c_1\| R_{mm}^{-1} e_{m,m} \\ &= Q_2^T \begin{pmatrix} I_m \\ 0 \end{pmatrix} \|c_1\| R_{mm}^{-1} e_{m,m} = Q_2^T \|c_1\| R_{mm}^{-1} e_{m,N}.\end{aligned}$$

Since Q_2 is orthogonal, the Euclidean norms of both sides must be equal. That is,

$$l = \sqrt{c_1^T (Y_{\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^T)^{-1} c_1} = \|Y_{\mathrm{pa}(i)}^T (Y_{\mathrm{pa}(i)} Y_{\mathrm{pa}(i)}^T)^{-1} c_1\| = \|c_1\| R_{mm}^{-1} = \lambda.$$

Then we can compute

$$c = b^2 + \sum_{j=m+1}^N (Q_2 Y_i^T)_j^2 = b^2 + \|Y_i^T - Y_{\mathrm{pa}(i)}^T \hat{\alpha}\|^2 = b^2 + y_0^2,$$

and

$$\alpha''_m = -b + \frac{(-c_1^T\hat{\alpha} - c_0)(r+1) \pm \sqrt{(c_1^T\hat{\alpha} + c_0)^2(r+1)^2 + 4rl^2y_0^2}}{2l}$$
$$:= (Q_2 Y_i^T)_{[m]} + \frac{(-c_1^T\hat{\alpha} - c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l}$$

Therefore, the two possible optimal vectors are

$$\alpha = Q_1^T R_1^{-1}(Q_2 Y_i^T)_{[m]} + \frac{-(c_1^T\hat{\alpha} + c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l^2}\|c_1\|R_{mm}^{-1} \cdot Q_1^T R_1^{-1} e_{m,m}$$
$$= \hat{\alpha} + \frac{-(c_1^T\hat{\alpha} + c_0)(r+1) \pm \sqrt{\Delta_{r,\hat{\alpha}}(l)}}{2l^2}(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1.$$

Each possible solution is the simple linear regression coefficient vector $\hat{\alpha}$ adding a multiple of $(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1$. The coefficient of the second term is a solution to the quadratic equation

$$l^2t^2 + (c_1^T\hat{\alpha} + c_0)(r+1)t - ry_0^2 = 0,$$

where $\hat{\alpha} = Y_{\mathrm{pa}(i)}^T(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}Y_i^T$, $y_0^2 = \|Y_i^T - Y_{\mathrm{pa}(i)}^T\hat{\alpha}\|^2$ and $l^2 = \|Y_{\mathrm{pa}(i)}^T(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1\|^2$. □

The update of ω_{ii} is given by

$$\omega_{ii}^* = \frac{1}{n_1+n_2}\|Y_i - B_{i,\mathrm{pa}(i)}^* Y_{\mathrm{pa}(i)}\|^2 \tag{11}$$

for the two possible $B_{i,\mathrm{pa}(i)}^*$'s. Then the profile log-likelihood has value

$$n_1 \log((c_{i,0} + B_{i,\mathrm{pa}(i)}^* c_{i,\mathrm{pa}(i)})^2) - (n_1+n_2)\log(\omega_{ii}^*) + C. \tag{12}$$

The choice of $B_{i,\mathrm{pa}(i)}^*$ corresponds to the larger log-likelihood value of the two candidates.

Remark 1.1. *The ratio $r = n_1/n_2$ influences the possible weights in the direction of $(Y_{\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}^T)^{-1}c_1$. When $n_2 = 0$, or equivalently $r = \infty$, the problem reduces to the purely observational case. In this scenario, equation* (8) *simplifies to a linear form: $(c_1^T\hat{\alpha} + c_0)t - y_0^2 = 0$, which coincides with the result established in Drton et al. (2019).*

At each update, the block-coordinate descent algorithm finds a local maximum of log-likelihood function with respect to a subset of all variables. Overall, the value of the log-likelihood function is non-decreasing throughout the iterations. To ensure that the algorithm is well-defined, each block update must have an optimal solution in which every ω_{ii} positive. This condition is equivalent to requiring $\|Y_i - B_{i,\mathrm{pa}(i)}Y_{\mathrm{pa}(i)}\| > 0$ during the update for every i, which in turn implies that Y_i is not in the row span $Y_{\mathrm{pa}(i)}$, i.e., the matrix $Y_{\mathrm{pa}(i)\cup\{i\}}$ has linearly independent rows. This requirement is consistent with condition $(\mathrm{A1})_i$ in Drton et al. (2019) for directed graphs. Using similar arguments, we can conclude that for generic triples $(Y, \omega_0, B_0) \in \mathbb{R}^{p\times N} \times \omega(V) \times \mathbf{B}(G)$, any finite number of iterations of Algorithm 1 have unique and feasible block updates when (ω_0, B_0) is used as the starting value.

Input: ω^0, B^0; $Y^{(1)}, \ldots, Y^{(K)}$; $I_1, \ldots, I_K$ and $n_1, \ldots, n_K$
repeat
  foreach $i \in V$ **do**
    if *the condition* (5) *does not hold* **then**
      stop: The block update cannot be solved in closed-form;
    end
    Find $S_2 = \{k : I_k \in \mathcal{I}_{-i},\ \mathcal{C}(i, G_{\overline{I_k}}) = \{i\}\}$;
    Find $S_1 = \{k : I_k \in \mathcal{I}_{-i}\} \setminus S_2$;
    Set $Y = [Y^{(k_1)}, \ldots, Y^{(k_l)}]$ with $k_1, \ldots, k_l \in S_1 \cup S_2$;
    Compute $n_1 = \sum_{k \in S_1} n_k$ and $n_2 = \sum_{k \in S_2} n_k$;
    if $n_1 = 0$ **then**
      Compute $\hat{B}_{i,\mathrm{pa}(i)}$ by solving least squares: $\arg\min_\beta \|Y_i^T - Y_{\mathrm{pa}(i)}^T \beta\|^2$;
    else
      if $n_2 = 0$ **then**
        Compute $\hat{B}_{i,\mathrm{pa}(i)}$ as the block-coordinate update for observational data;
      else
        Compute the two possible $\hat{B}_{i,\mathrm{pa}(i)}$ using (7) and (8);
        Compute corresponding $\hat{\omega}_{ii}$ and log-likelihood values using (11) and (12);
        Choose the larger log-likelihood value and the corresponding $(\hat{\omega}_{ii}, \hat{B}_{i,\mathrm{pa}(i)})$;
      end
    end
    Update ω and B_i using $\hat{\omega}_{ii}$ and $\hat{B}_{i,\mathrm{pa}(i)}$;
  end
until *convergence criterion is met*;

Algorithm 1: Block-coordinate descent, for directed graph and special intervention targets

4 Numerical experiments

Suppose there are K dataset with different interventions $\mathcal{I} = \{I_1, \ldots, I_K\}$, the simple aggregation method involves performing original BCD algorithm on each dataset, aggregating the K MLEs and compute the weighted average as the final estimates. A natural choice of weighting scheme is to use weights proportional to the sample sizes. We compare the performance of the simple aggregation method and our BCD-type algorithms on synthetic data.

In the simulations, we consider the special directed graphs that contain one unique cycle. First, we add the l-cycle $1 \to 2 \to \cdots \to l \to 1$ to the empty graph. Due to the component constraint, there remain $p(p-1)/2 - l(l-1)/2$ ordered pairs (i, j) with $i < j$ that can be assigned with nonzero edge weights. We simulate independent

uniform random variables $U_{ij} \sim U(0,1)$. If $U_{ij} < d$, the edge $i \to j$ is introduced. The sparsity parameter $d \in (0,1)$ controls the average number of edges in the graph. The edge generation is performed under a fixed topological ordering of the nodes. After adding the edges, we randomly permute the node labels. This construction ensures that the graph has a unique cycle of length l.

				RMSE		diff-llh		Running time	
p	m	k	d	**Agg**	**MLE**	**Agg**	**MLE**	**Agg**	**MLE**
5	25	0	0.2	0.0277	0.0267	0.2426	0.2274	1.04	0.63
5	25	0	0.3	0.0280	0.0266	0.2861	0.2629	1.07	0.86
5	25	3	0.2	0.2281	0.1732	0.3731	0.3020	10.71	3.92
5	25	3	0.3	0.2141	0.1498	0.4104	0.3282	11.16	4.00
5	25	4	0.2	0.0319	0.0901	0.3673	0.3170	9.08	4.19
5	25	4	0.3	0.0323	0.0989	0.3913	0.3367	9.02	4.45
5	50	0	0.2	0.0134	0.0130	0.1046	0.1010	0.97	0.63
5	50	0	0.3	0.0135	0.0130	0.1204	0.1144	1.08	0.70
5	50	3	0.2	0.0353	1.1436*	0.1686	0.1350	11.89	4.04
5	50	3	0.3	0.0339	1.0413*	0.1810	0.1440	11.93	4.21
5	50	4	0.2	0.0154	0.0124	0.1600	0.1417	8.70	4.09
5	50	4	0.3	0.0156	0.0125	0.1676	0.1477	8.86	4.06
10	50	0	0.2	0.0106	0.0100	0.2358	0.2147	2.24	1.48
10	50	0	0.3	0.0113	0.0104	0.3093	0.2714	2.60	1.72
10	50	3	0.2	0.0330	0.0112	0.2894	0.2401	16.99	5.84
10	50	3	0.3	0.0300	0.0113	0.3745	0.2932	18.72	6.20
10	50	4	0.2	0.0109	0.0098	0.2952	0.2533	13.23	5.35
10	50	4	0.3	0.0116	0.0101	0.3699	0.3045	14.03	5.72
10	100	0	0.2	0.0053	0.0051	0.1098	0.1053	2.21	1.58
10	100	0	0.3	0.0054	0.0051	0.1400	0.1314	2.70	2.02
10	100	3	0.2	0.0074	0.0057	0.1296	0.1126	18.24	6.20
10	100	3	0.3	0.0072	0.0057	0.1599	0.1356	19.76	6.64
10	100	4	0.2	0.0053	0.0048	0.1293	0.1169	13.44	5.51
10	100	4	0.3	0.0055	0.0049	0.1564	0.1380	14.49	5.97

Table 1: Statistics for randomly generated directed graphs with at most one unique cycle. Each row summarizes 1000 simulations. The columns "Agg" and "MLE" correspond to the aggregation method and Algorithm 1. "RMSE" represents the average root mean square error of the estimate for a single parameter among the total 1000 simulations. "diff-llh" is the average difference between the log-likelihood of the true parameters and that of the estimated parameters (the smaller the better). Running time is the average CPU time (in milliseconds). For the aggregation method, the reported running time includes both the BCD algorithm applied to each observational or interventional dataset and the subsequent aggregation steps.

We use 24 different configurations of (p, m, l, d), where m is the sample size of observational data. For each graph, we randomly select the number of interventional environments, ensuring that $|\mathcal{I}| \in \{1, 2, 3\}$. Each random intervention target is of size 2 or 3. We then compute the intervened model and simulate data of sample size $\max(m_k, p+1)$, $n_k \sim U[\lfloor (m+1)/2 \rfloor, m]$, for each intervention target. Consequently, the data for one graph is from both observation and interventions, with the total size ranging between $3m/2$ and $4m$. We consider $p \in \{5, 10\}$, $m \in \{5p, 10p\}$, and $l \in \{0, 3, 4\}$, with $d \in \{0.2, 0.3\}$.

We avoid using 2-cycles because they are not identifiable from observational data (Drton et al., 2019), and estimation results with interventional data are also unstable.

For each of the configurations (p, m, l, d), we simulate 1000 graphs using the procedure described above. In each simulation, all free entries of B are drawn independently from a uniform distribution on $[-2, -0.5] \cup [0.5, 2]$. The diagonal entries of Ω are randomly drawn from a uniform distribution on $[0.3, 1]$. For an intervention target set I, the corresponding columns in B are masked by zero: $B_{I,\cdot} = 0$. The interventional errors ϵ_I are sampled from $|I|$ independent standard normal distributions.

Simulations were performed on a desktop equipped with an AMD Ryzen 9 7950X3D processor (4.2 GHz), using R 4.4.2 on Windows 11. For each run, the maximum number of iterations was set to 5000. In all simulations, our algorithm converged, and the BCD algorithm also converged across all environments. Table 1 summarizes the simulation results.

Our MLE algorithm consistently achieves higher log-likelihood values and outperforms the aggregation method in terms of RMSE across most of the configurations. There are 2 exceptions where our method does not achieve lower RMSE. The RMSE may be influenced by extreme values in cycle parameter estimation, whereas the likelihood-based metric tends to be more stable. This behavior reflects the nature of cyclic models, where better likelihood does not necessarily imply a smaller RMSE due to the additional likelihood contribution from the cycle structure. In addition to achieving higher accuracy, our MLE algorithm is also faster than the aggregation method, as expected.

For real data example, one potential application is to model the relations between daily sleeping time, activity time and mental health measurements using data collected in the Healthy Aging in Industrial Environment study - Program 4 (4HAIE) (Elavsky et al., 2021). Beginning in 2019, this study intensively monitored air pollution and behavioral parameters. The timing of the study coincided with the COVID-19 and the pandemic restrictive measures implemented since March 2020 can be regarded as interventions affecting activity patterns (e.g., steps per day). After transforming the data to standard normal scale, variables corresponding to average sleeping time, average daily steps, and psychological scores fit naturally within a linear SEM framework that includes a feedback loop. Data collected before and after the implementation of the restrictive measures serve as observational and interventional data, respectively.

Acknowledgements

The study is supported by the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union.

References

M. Drton and T. S. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. In *Uncertainty in Artificial Intelligence: Proceedings of the* 20th *Conference*,

2004.

M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348, 2009. URL http://jmlr.org/papers/v10/drton09a.html.

M. Drton, C. Fox, and Y. S. Wang. Computation of maximum likelihood estimates in cyclic structural equation models. *Ann. Statist.*, 47(2):663–690, 04 2019. doi: 10.1214/17-AOS1602. URL https://doi.org/10.1214/17-AOS1602.

S. Elavsky, V. Jandačková, L. Knapová, V. Vašendová, M. Sebera, B. Kaštovská, D. Blaschová, J. Kühnová, R. Cimler, D. Vilímek, et al. Physical activity in an air-polluted environment: Behavioral, psychological and neuroimaging protocol for a prospective cohort study (Healthy Aging in Industrial Environment study – Program 4). *BMC Public Health*, 21:1–14, 2021.

J. Fox. Teacher's corner: Structural equation modeling with the sem package in R. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3):465–486, 2006. doi: 10.1207/s15328007sem1303_7. URL https://doi.org/10.1207/s15328007sem1303_7.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(79):2409–2464, 2012. URL http://jmlr.org/papers/v13/hauser12a.html.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society Series B*, 77(1):291–318, 2015. URL https://EconPapers.repec.org/RePEc:bla:jorssb:v:77:y:2015:i:1:p:291-318.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-89560-6; 0-521-77362-8. Models, reasoning, and inference.

Y. Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012. doi: 10.18637/jss.v048.i02. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i02.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, Cambridge, MA, second edition, 2000. ISBN 0-262-19440-6.

S. Wright. Correlation and causation. *J. Agricultural Research*, 20:557–585, 1921.

S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5(3):161–215, 1934.

List of Authors

The 13th Workshop on Uncertainty Processing

June 4-7, 2025

Třešť, Czech Republic

Editors

Milan Studený, Nihat Ay, Andrea Capotorti, László Csirmaz,
Radim Jiroušek, Gernot D. Kleiter, Prakash P. Shenoy

Organized by

Institute of Information Theory and Automation, Czech Academy of Sciences
Faculty of Management, Prague University of Economics and Business, Jindřichův Hradec

http://wupes.utia.cas.cz/

Nakladatelství MFF UK
MatfyzPress

PUBLISHING HOUSE
OF THE FACULTY OF MATHEMATICS AND PHYSICS
CHARLES UNIVERSITY IN PRAGUE

ISBN 978-80-7378-525-3