International Journal of Approximate Reasoning 151 (2022) 64–84

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A geometric characterization of sensitivity analysis in monomial models

Manuele Leonelli [a,*,1], Eva Riccomagno [b]

[a] School of Science and Technology, IE University, Spain
[b] Dipartimento di Matematica, Universitá degli Studi di Genova, Italy

ARTICLE INFO

Article history:
Received 13 January 2021
Received in revised form 2 August 2022
Accepted 8 September 2022
Available online 14 September 2022

Keywords:
Bayesian network classifiers
Covariation
I-projections
Monomial models
Sensitivity analysis

ABSTRACT

Sensitivity analysis in probabilistic discrete graphical models is usually conducted by varying one probability at a time and observing how this affects output probabilities of interest. When one probability is varied, then others are proportionally covaried to respect the sum-to-one condition of probabilities. The choice of proportional covariation is justified by multiple optimality conditions, under which the original and the varied distributions are as close as possible under different measures. For variations of more than one parameter at a time and for the large class of discrete statistical models entertaining a regular monomial parametrisation, we demonstrate the optimality of newly defined proportional multi-way schemes with respect to an optimality criterion based on the I-divergence. We demonstrate that there are varying parameters' choices for which proportional covariation is not optimal and identify the sub-family of distributions where the distance between the original distribution and the one where probabilities are covaried proportionally is minimum. This is shown by adopting a new geometric characterization of sensitivity analysis in monomial models, which include most probabilistic graphical models. We also demonstrate the optimality of proportional covariation for multi-way analyses in Naive Bayes classifiers.

1. Introduction

The effect of changes of a quantitative model's inputs to outputs of interest is studied through sensitivity analysis. It is a critical step for both the construction and the validation of a model as well as the communication of its results. Because of the importance of the application of sensitivity analyses, their use has been constantly increasing in the past ten years [24]. The specific steps of a sensitivity analysis greatly vary depending on the quantitative model used [see 6,50, for an overview].

In the context of probabilistic graphical models, sensitivity analysis techniques are well-established [1,30,54,57] and, especially for discrete Bayesian networks (BNs), applied in many real-world problems [14,28,33,41,44]. A comprehensive review has been recently presented in [47]. For such models, the validation process can be broken down into two steps: the first concerns the auditing of the validity of the conditional independences implied by the underlying graphical structure; the second, assuming the graph is valid, checks the impact of the numerical elicited probabilities on outputs of interest. The focus of this paper lies in this second phase.

* Corresponding author.
E-mail address: manuele.leonelli@ie.edu (M. Leonelli).
[1] Full address: School of Science and Technology, Paseo de la Castellana 259, 28046 Madrid, Spain.

https://doi.org/10.1016/j.ijar.2022.09.006

The impact of the input probabilities in probabilistic graphical models is studied both at a local and a global level. At the local level, the effect of parameter variations on a specific output probability of interest is summarized by sensitivity functions [15,54]. These correspond to the probability of interest written as a function of a varying parameter. At the global level, the effect of parameter variations is quantified by divergences and distances between the original and the varied model's distribution. Common measures to quantify the dissimilarity between the two distributions are the Chan-Darwiche distance and the Kullback-Leibler divergence [12].

Because of both its simplicity and its proven theoretical justifications, the most common investigation is the so-called *one-way* sensitivity analysis, where the impacts of changes made to a single probability parameter are studied. One-way analyses are implemented in various pieces of software, e.g. Siamlam and the `bnmonitor` R package [40]. When one parameter is varied, then others are required to be adjusted, or *covaried*, to respect the sum-to-one condition of probabilities. Although there are various ways to covary probabilities, the most common covariation scheme is the *proportional* one, where, after a change to a parameter, the covarying parameters have the same proportion of the residual probability mass as they originally had [37,46]. Proportional covariation in one-way analyses is "optimal" since it minimizes a large array of divergences between the original and the varied probability distributions amongst any valid covariation scheme [12,39].

Multi-way methods, where two or more parameters are varied contemporaneously, have not been extensively studied in the literature [see 4,5,11,13,21,32,45, for some exceptions]. This is not only because they require a much more intensive computational power, but also, and more critically, because there are very little, if no, theoretical justifications in using a covariation scheme over another. Output probabilities have been shown to heavily depend on the covariation scheme used [39,46]. In [39] it has been recently proven that for specific multi-way analyses called *full single conditional probability table* (CPT) analyses, which highly restrict the parameters that can be varied, proportional covariation is optimal. However, to date there is no theoretical result guaranteeing the optimality of proportional covariation for multi-way variations.

This work provides a conclusive answer to the optimality of proportional covariation by characterizing the varying parameter choices in multi-way analyses for which proportional covariation minimizes the I-divergence, also known as Kullback-Leibler divergence [18]. This is achieved via a new formal, geometric characterization of sensitivity analysis in terms of I-projections of the original distribution into a well-specified subset of the probability simplex. New flexible ways for choosing parameters to vary are introduced in Definition 9 below and their optimality with respect to the I-divergence is demonstrated. Although the I-divergence was considered inappropriate for bounding belief changes in some situations by [12], we consider it in this work because of its geometric properties which are reviewed in Section 3.

We consider here the class of *monomial models* (MMs), where the probability of any element of the sample space is represented by a monomial, namely a power product of positive parameters [39]. We show that BNs are a specific instance of MMs, but many other well-known statistical models can also be cast as MMs [26,38,39]: for instance staged trees and chain event graphs [52], context-specific BNs [7], decomposable Markov networks and probabilistic chain graphs [36] (see Appendix B for details on the relationship between staged trees and MMs). Also, some relational models [34,35], which generalize standard log-linear models for the analysis of contingency tables, can be cast as MMs.

We chose to work with the more abstract MM model class for two reasons. On one hand, the monomial structure enables us to define new type of sensitivity analyses for which proportional covariation is shown to be optimal. On the other, the derived results do not simply apply to BNs but are also valid for a larger array of widely-used probabilistic graphical models. Although in this work the MM representation of probabilistic graphical models is used to have an abstract representation which is exploited to study properties of the models, it can also be useful in practice. The field of *algebraic statistics* [43,53] takes advantage of this abstract representation to devise efficient inferential and learning algorithms.

The optimality of proportional covariation in naive Bayes classifiers (e.g. [2]) for any combination of probabilities associated to feature variables is also proven here. Naive Bayes models are a specific type of BN classifiers often used to assign instances to a specific class in a classification problem. The tuning of the feature's probabilities is often critical to ensure that the classifier works reliably [5].

The paper is structured as follows. Section 2 includes the definition of MMs and some examples. Section 3 reviews the essential notions from information geometry. Section 4 summarizes relevant results of sensitivity analysis in BNs. Section 5 introduces new classes of multi-way sensitivity analyses. Section 6 gives a geometric characterization of sensitivity analysis and proves the optimality of our schemes. Our results apply to BNs, which are an important class of MMs, but more generally apply to a wider class of models as shown in Section 6.2. Section 7 discusses an applied sensitivity analysis in a medical application. The paper is concluded by a discussion. Longer proofs are collated in Appendix A.

2. Monomial discrete parametric models

Let P be a probability distribution over a finite set $\mathbb{Y}$. Write $\#\mathbb{Y} = q$, call $y \in \mathbb{Y}$ an atom and $\mathrm{P}(y)$ the atomic probability of y. The generic P can be seen as a point in the interior set of the q-dimensional simplex and we write $\mathrm{P} \in \Delta_{q-1}$. Let also $[k] = \{1, 2, \ldots, k\}$. Next, to $\mathbb{Y}$ we associate a particular class of parametric statistical models, called monomial models, in short MMs. Such models were introduced in [39] and here we provide a novel, more precise definition.

A MM is defined by three elements:

- a $q \times k$ matrix A with non-negative integer entries, i.e. $A \in \mathcal{M}_{q\times k}(\mathbb{Z}_{\geq 0})$;
- a k-dimensional parameter vector θ with positive real entries, i.e. $\theta = (\theta_i)_{i\in[k]} \in \mathbb{R}^k_{>0}$;

- a partition $S = \{S_1, \ldots, S_n\}$ of $[k]$ such that $\theta_{S_i} \in \Delta_{\#S_i - 1}$ for all $i \in [n]$, $n \leq k$.

There is a row of A for each atom y and A_y indicates the y-th row of A. The atomic probability of $y \in \mathbb{Y}$ given θ and A is defined as $\mathrm{P}(y) = \prod_{i \in [k]} \theta_i^{A_{y,i}} = \theta^{A_y}$. In an MM the θ parameters are grouped by the partition S in such a way that those in a group sum to one. For a subset $B \subset [k]$, the notation $\theta_B = (\theta_i)_{i \in B}$ indicates the sub-vector of elements of θ indexed by B and $\theta_B^{A_{y,B}} = \prod_{i \in B} \theta_i^{A_{y,i}}$ denotes the monomial associated to an event $y \in \mathbb{Y}$ where only parameters θ_i for $i \in B$ can have non-zero exponent.

The motivation behind the use of MMs is to give an abstract representation of many probabilistic graphical models. These usually factorize atomic probabilities via (conditional) probabilities associated to each vertex: such probabilities are represented by the vector θ. Since components of θ are conditional probabilities, sub-groups of θ must add up to one and these sub-groups are specified by the partition S. Furthermore, the structure of the graph specifies how (conditional) probabilities must be multiplied with each other to derive the atomic probabilities of the model: in a MM, for each $y \in \mathbb{Y}$, this is specified by the row of A corresponding to y. Therefore, the matrix A and the partition S are fixed and describe the structure of the model, whilst the vector of probabilities θ is allowed to vary. The class of multilinear MMs defined next includes these graphical models.

Definition 1. Let S be as above and $A \in \mathcal{M}_{q \times k}(\{0,1\})$. The class of multilinear monomial models (MMs) over $\mathbb{Y}$ associated to A and S is defined as

$$\mathrm{MM}(A, S) = \left\{ \mathrm{P} \in \Delta_{q-1} : \mathrm{P}(y) = \theta^{A_y} \text{ for } y \in \mathbb{Y} \text{ and } \theta \in \mathbb{R}^k_{>0} \right\},$$

where

$$\mathrm{P}(y) = \prod_{i \in [n]} \prod_{j \in S_i} \theta_j^{A_{y,j}} = \prod_{i \in [n]} \theta_{S_i}^{A_{y,S_i}}.$$

The assumption of strictly positive probabilities is often met in practice, for instance when probabilities are estimated from data using a Laplace smoothing approach [49]. The condition is imposed here to ensure the I-divergence exists and is finite. The word multilinear is motivated by the fact that all monomials defining the model are square-free, i.e. the exponents of the parameters are either zero or one. The class of (non-multilinear) MMs can be defined by dropping the assumptions that A has 0-1 entries [38].

Example 1. A simple example of a multilinear $\mathrm{MM}(A, S)$ over a finite set $\mathbb{Y}$ are the saturated models where $\mathrm{P}(y) = \theta_y$ for all $y \in \mathbb{Y}$, i.e. one parameter is associated to the probability of each atomic event. In this case $\theta \in \Delta_{q-1}$, A is the q-by-q identity matrix and S is the largest possible partition with $n = k$.

Not all combinations of A matrices, θ vectors and partition S give rise to MMs, as illustrated by the next example.

Example 2. Let $\mathbb{Y} = [4]$, A be the 4×3 matrix with rows $A_1 = (1,0,0)$, $A_2 = (0,1,0)$, $A_3 = (0,0,1)$ and $A_4 = (0,1,1)$, and $S = [4]$, i.e. the sum of all parameters must be one. Atomic probabilities are defined by $\mathrm{P}(y) = \theta_y$ for $y \in [3]$ and $\mathrm{P}(y) = \theta_2\theta_3$ for $y = 4$, entailing $\mathrm{P}(4) = \mathrm{P}(3)\,\mathrm{P}(2)$. These are not MMs since it is not possible to have $\sum_{i \in [4]} \mathrm{P}(i) = 1$.

Definition 2. Multilinear $\mathrm{MM}(A, S)$ are called *regular* if for all $y \in \mathbb{Y}$ and all $i \in [n]$, $A_{y,j} = 1$ for at most one $j \in S_i$ and zero otherwise.

Regular MMs are such that for every $y \in \mathbb{Y}$, $\mathrm{P}(y)$ has at most one parameter with a non-zero exponent from every sum-to-one group, that is for every $S_i \in S$. Henceforth we work with regular multilinear MMs. The authors have not been able to find multilinear MMs which are not regular nor to prove that any multilinear MMs are regular. All well-known probabilistic graphical models, such as BNs (shown in Section 2.1), staged trees (shown in Appendix B) and decomposable Markov networks can be cast as regular multilinear MMs, since they have an equivalent BN representation by choosing a directionality for the edges.

Example 3. Consider the simple scenario of two binary random variables Y_1 and Y_2 taking values in the set $\{1,2\}$. Define the probabilities

$$\begin{array}{lll} \mathrm{P}(Y_1 = 1) = \theta_1, & \mathrm{P}(Y_2 = 1 | Y_1 = 1) = \theta_3, & \mathrm{P}(Y_2 = 1 | Y_1 = 2) = \theta_5 \\ \mathrm{P}(Y_1 = 2) = \theta_2, & \mathrm{P}(Y_2 = 2 | Y_1 = 1) = \theta_4, & \mathrm{P}(Y_2 = 2 | Y_1 = 2) = \theta_6 \end{array}$$

The atomic probabilities are $\theta_1\theta_3$, $\theta_1\theta_4$, $\theta_2\theta_5$ and $\theta_2\theta_6$. This setting can be cast as a class of MMs with parameter $\theta = (\theta_i)_{i \in [6]}$, partition $S = \{\{1,2\}, \{3,4\}, \{5,6\}\}$ and the 4×6 matrix A

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	P(y)
1	0	1	0	0	0	$\theta_1\theta_3$
1	0	0	1	0	0	$\theta_1\theta_4$
0	1	0	0	1	0	$\theta_2\theta_5$
0	1	0	0	0	1	$\theta_2\theta_6$

where for clarity we labelled the columns and reported the atomic probabilities. Since no parameters belonging to a same element of the partition S appear in the same monomial, this class of MMs is regular.

2.1. Bayesian networks

Many discrete statistical problems in a variety of domains are often modelled using BNs and there are now thousands of practical applications of these models [3,8,22,42]. A BN expresses graphically a collection of conditional independences [20, 36]. For a random vector $Y = (Y_i)_{i\in[m]}$ taking values in the Cartesian product $\mathbb{Y} = \bigtimes_{i\in[m]} \mathbb{Y}_i$ and three disjoint subsets B, C, and D of $[m]$, the marginal vector Y_B is said to be conditionally independent of Y_C given Y_D, and written as $Y_B \perp\!\!\!\perp Y_C \mid Y_D$, if $\mathrm{P}(Y_B = b|Y_C = c, Y_D = d) = \mathrm{P}(Y_B = b|Y_D = d)$, for all $b \in \bigtimes_{i\in B} \mathbb{Y}_i$, $c \in \bigtimes_{i\in C} \mathbb{Y}_i$ and $d \in \bigtimes_{i\in D} \mathbb{Y}_i$.

A BN over a vector of discrete variables $Y = (Y_i)_{i\in[m]}$ is a pair (G, θ), where: G is a DAG whose vertices are the variables in Y and, roughly speaking, edges represent direct probabilistic dependencies between the variables; θ is a vector of (conditional) probabilities $\mathrm{P}(Y_i = y_i|Y_{\Pi_i} = y_{\Pi_i})$ for each $y_i \in \mathbb{Y}_i$ given a, possibly empty, vector value $y_{\Pi_i} \in \mathbb{Y}_{\Pi_i}$ of the parents of Y_i in G. The BN represents the joint distribution over Y factorized according to the structure of G:

$$\mathrm{P}(Y = y) = \prod_{i\in[m]} \mathrm{P}(Y_i = y_i|Y_{\Pi_i} = y_{\Pi_i}).$$

A BN encodes conditional independences of the form $Y_i \perp\!\!\!\perp Y_{\Lambda_i}|Y_{\Pi_i}$, where Y_{Λ_i} are the non-descendants of Y_i in G.

For a variable Y_i with parents Y_{Π_i}, let $\theta_{y_i y_{\pi_i}} = \mathrm{P}(Y_i = y_i|Y_{\Pi_i} = y_{\pi_i})$. The probability of any atom $y = (y_1, \ldots, y_m) \in \mathbb{Y}$ can then be written as the monomial $\mathrm{P}(Y = y) = \prod_{i\in[m]} \theta_{y_i y_{\pi_i}}$ [10,19]. Notice that $(\theta_{y_i y_{\pi_i}})_{y_i\in\mathbb{Y}_i} \in \Delta_{\#\mathbb{Y}_i - 1}$ for all $i \in [m]$ and $y_{\pi_i} \in \mathbb{Y}_{\Pi_i}$, encoding the sum-to-one condition of a conditional probability mass function, i.e. $\sum_{y_i\in\mathbb{Y}_i} \mathrm{P}(Y_i = y_i|Y_{\Pi_i} = y_{\Pi_i}) = 1$. Thus a BN is a multilinear regular MM where the sets S_i, $i \in [n]$, denote the conditional probability mass functions of any vertex given a specific combination of parents, since each of the associated conditional distributions must respect the sum-to-one condition. Thus while the classical representation of a BN uses a graph and parameter CPTs local to every variable/node to compactly represent a joint probability distribution, its MM representation uses a matrix and a parameter vector partitioned into sub-vectors per variable to represent a joint distribution.

Example 4. We are interested in studying how a population's health (Y_3) is affected by both sports activity (Y_1) and alcoholic drinking habits (Y_2). These three variables are categorized into high, medium and low, coded with 3, 2 and 1 respectively. Health's levels are assumed to be a function of both sports activity and drinking habits and that people who work out a lot tend to drink less alcohol. This situation can be depicted by a BN with probabilities

$$\mathrm{P}(Y_1 = i) = \theta_i, \quad \mathrm{P}(Y_2 = j|Y_1 = i) = \theta_{ji}, \quad \mathrm{P}(Y_3 = l|Y_2 = j, Y_1 = i) = \theta_{lji}$$

where $\sum_{k\in[3]} \theta_k = 1$, $\sum_{k\in[3]} \theta_{ki} = 1$ and $\sum_{k\in[3]} \theta_{kji} = 1$ for all $i, j \in [3]$. The associated MM is given in Table 1 where the monomial representation of the 27 atomic probabilities is listed. Of course the parameters θ_k, θ_{ki} and θ_{kji} can be renamed to give some $(\theta_l)_{l\in[39]}$. The A matrix has dimension 27×39 and is very sparse: in each row there is a one in three positions and zero otherwise.

3. I-projections

As a measure of closeness of two distributions we consider the I-divergence. Below we follow [18, Chapter 3].

Definition 3. Let P and Q be two probability distributions over a finite space $\mathbb{Y}$. The I-divergence (or Kullback-Leibler divergence) from P to Q is defined as

$$\mathcal{D}(\mathrm{Q}\,||\,\mathrm{P}) = \sum_{y\in\mathbb{Y}} \mathrm{Q}(y) \ln \frac{\mathrm{Q}(y)}{\mathrm{P}(y)},$$

where ln is the natural logarithm. It is often of interest to find the distribution that, within a given set, is closest to a given P. *I-projections* formalize this idea. I-projections are used e.g. for maximum likelihood estimation in the context of exponential families.

Table 1
Monomial atomic probabilities for the BN of Example 4.

$\theta_1\theta_{11}\theta_{111}$	$\theta_1\theta_{11}\theta_{211}$	$\theta_1\theta_{11}\theta_{311}$	$\theta_1\theta_{21}\theta_{121}$	$\theta_1\theta_{21}\theta_{221}$	$\theta_1\theta_{21}\theta_{321}$	$\theta_1\theta_{31}\theta_{131}$	$\theta_1\theta_{31}\theta_{231}$	$\theta_1\theta_{31}\theta_{331}$
$\theta_2\theta_{12}\theta_{112}$	$\theta_2\theta_{12}\theta_{212}$	$\theta_2\theta_{12}\theta_{312}$	$\theta_2\theta_{22}\theta_{122}$	$\theta_2\theta_{22}\theta_{222}$	$\theta_2\theta_{22}\theta_{322}$	$\theta_2\theta_{32}\theta_{132}$	$\theta_2\theta_{32}\theta_{232}$	$\theta_2\theta_{32}\theta_{332}$
$\theta_3\theta_{13}\theta_{113}$	$\theta_3\theta_{13}\theta_{213}$	$\theta_3\theta_{13}\theta_{313}$	$\theta_3\theta_{23}\theta_{123}$	$\theta_3\theta_{23}\theta_{223}$	$\theta_3\theta_{23}\theta_{323}$	$\theta_3\theta_{33}\theta_{133}$	$\theta_3\theta_{33}\theta_{233}$	$\theta_3\theta_{33}\theta_{333}$

Definition 4. Let L be a closed, convex set in the pointwise topology of distributions over $\mathbb{Y}$. The I-projection of a distribution P over $\mathbb{Y}$ onto L is a distribution $\mathrm{P}^* \in L$ such that

$$\mathcal{D}(\mathrm{P}^* \,||\, \mathrm{P}) = \min_{\mathrm{Q}\in L} \mathcal{D}(\mathrm{Q}||\,\mathrm{P}).$$

If $\mathrm{P} \in L$ then $\mathrm{P}^* = \mathrm{P}$. The fact that L is closed and convex guarantees that P^* exists in L and for strictly positive probabilities P^* is unique.

Theorem 1. *For all* $\mathrm{Q} \in L$ *it holds*

$$\mathcal{D}(\mathrm{Q}||\,\mathrm{P}) \geq \mathcal{D}(\mathrm{Q}||\,\mathrm{P}^*) + \mathcal{D}(\mathrm{P}^*\,||\,\mathrm{P}), \tag{1}$$

where P^* *is the I-projection of* P *in* L*. If the inequality (1) holds for some* $\mathrm{P}^* \in L$ *and all* $\mathrm{P} \in L$*, then* P^* *is the I-projection of* Q *onto* L*.*

As a straightforward consequence of Theorem 1, if the Pythagorean identity

$$\mathcal{D}(\mathrm{Q}||\,\mathrm{P}) = \mathcal{D}(\mathrm{Q}||\,\mathrm{R}) + \mathcal{D}(\mathrm{R}||\,\mathrm{P}) \tag{2}$$

holds for all $\mathrm{Q} \in L$ and a specific $\mathrm{R} \in L$ then $\mathrm{R} = \mathrm{P}^*$. Theorem 1 is used extensively in Section 6 to prove the optimality of the new multi-way covariation schemes introduced in Section 5.

4. Sensitivity analysis

Sensitivity analysis of a probabilistic model amounts to varying the values for one or more model parameters simultaneously and investigating the effects on a probability of interest. Varying one parameter at a time is called a *one-way* sensitivity analysis. It serves to reveal the effect of the parameter under study on a probability of interest. Analyses where n parameters are varied simultaneously are called *n-way* sensitivity analyses. These reveal how the n parameters interact in their effect on a probability of interest. In this section we review relevant definitions and results about sensitivity analysis in BNs, but rephrased in the context of monomial models.

4.1. Covariation

When some parameters of a (conditional) probability distribution are varied to a new specific value, then the remaining parameters need to be adjusted (or to *covary*) to respect the sum-to-one condition of probability measures. In the binary case when one of the two parameters is varied this is straightforward, since the second parameter will be equal to one minus the other. But in generic discrete finite cases there are various considerations to be taken into account, as reviewed below.

We start by giving an alternative definition of a covariation scheme to [46]. The parameters to be modified are a subset of a simplex and only those in the same simplex are allowed to covary. The other model parameters are unchanged. Our definition of covariation allows more than one or no parameters to be varied and maps into a probability simplex, i.e. the scheme is valid [46]. Let k be the number of parameters in the model, $\emptyset$ the empty set and for a vector v let $|v|$ denote the sum of its components.

Definition 5. When $V = \emptyset$, the $\tilde{\theta}_V$-covariation scheme is the identity function. For $\emptyset \neq V \subset S \subseteq [k]$, let $\theta_S \in \Delta_{\#S-1}$ be partitioned as $\theta_S = (\theta_V, \theta_{S\setminus V})$ and let $\tilde{\theta}_V$ be such that $|\tilde{\theta}_V| \in (0,1)$. A *$\tilde{\theta}_V$-covariation scheme* is a function σ from $\Delta_{\#S-1}$ to $\Delta_{\#S-1}$ which fixes the subvector θ_V of θ_S to $\tilde{\theta}_V$, i.e.

$$\begin{aligned} \sigma: \quad & \Delta_{\#S-1} \longrightarrow \Delta_{\#S-1} \\ & (\theta_V, \theta_{S\setminus V}) \longmapsto (\tilde{\theta}_V, \cdot). \end{aligned}$$

Thus θ_S denotes a vector of parameters that need to respect the sum to one condition, $\tilde{\theta}_V$ denotes the new numerical specification of the parameters varied, i.e. those with index in a set V, and the values of the parameters with index in $[k]\setminus S$ do not vary. Below we define some frequently applied covariation schemes.

Definition 6. In the notation of Definition 5

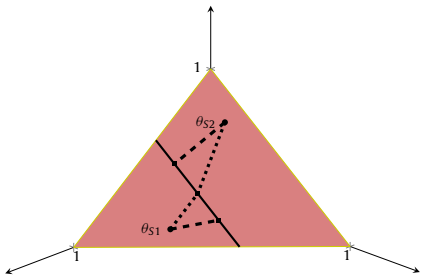


Fig. 1. Graphical representation of uniform and proportional covariation in Example 5 for two generic points $\theta_{S1}, \theta_{S2} \in \Delta_2$.

- the $\tilde{\theta}_V$*-proportional* covariation scheme $\sigma_{\text{pro}}(\theta_S) = (\tilde{\theta}_V, \tilde{\theta}_{S\setminus V})$ is defined by setting

$$\tilde{\theta}_j = \frac{1-|\tilde{\theta}_V|}{1-|\theta_V|}\theta_j \qquad \text{for all } j \in S \setminus V.$$

- The $\tilde{\theta}_V$*-uniform* covariation scheme, $\sigma_{\text{uni}}(\theta_S) = (\tilde{\theta}_V, \tilde{\theta}_{S\setminus V})$ is defined by setting

$$\tilde{\theta}_j = \frac{1-|\tilde{\theta}_V|}{\#S - \#V} \qquad \text{for all } j \in S \setminus V.$$

Different covariation schemes may entertain different properties which, depending on the domain of application, might be more or less desirable [see 39,46, for a list].

Example 5. Consider $\theta_S = (\theta_1, \theta_2, \theta_3) \in \Delta_2$, $V = \{1\}$ and $\tilde{\theta}_V = 0.4$. The simplex Δ_2 is given by the surface in Fig. 1, whilst the dark full line denotes the image of any $\tilde{\theta}_V$-covariation scheme σ which fixes $\tilde{\theta}_1 = 0.4$, i.e. the set defined by the intersection of the simplex with the line defined by $\theta_1 = 0.4$. That is, this line describes all possible ways θ_2 and θ_3 can be covaried. When σ is the uniform covariation scheme any $\theta_S \in \Delta_2$ is projected to the same point, as illustrated by the dotted lines in Fig. 1. Conversely, the dashed lines refer to the proportional covariation scheme which can project points $\theta_S \in \Delta_2$ to different elements.

The order-preserving covariation scheme in Definition 7 follows from changing one parameter and its definition is slightly convoluted. Let thus $V = \{v\}$ have just one element and consider $S \supset V$ such that $|\theta_S| = 1$. Assume that θ_v is not the largest component of θ_S and order the components of θ_S from the smallest to the largest. Without loss of generality by reorganising the indices in θ, we can assume that the ordered components of θ_S are $\theta_1 \le \cdots \le \theta_v < \cdots \le \theta_{\#S}$.

Definition 7. In the notation of Definition 5 the $\tilde{\theta}_V$*-order preserving* covariation scheme is defined, according to whether θ_v is increased or decreased, by setting

$$\tilde{\theta}_j = \begin{cases} \dfrac{\tilde{\theta}_v}{\theta_v}\theta_j, & \text{if } j < v \text{ and } \tilde{\theta}_v \le \theta_v, \\ -\theta_j \dfrac{1-\theta_{\text{suc}}}{\theta_{\text{suc}}}\dfrac{\tilde{\theta}_v}{\theta_v} + \dfrac{\theta_j}{\theta_{\text{suc}}}, & \text{if } j > v \text{ and } \tilde{\theta}_v \le \theta_v, \\ \theta_j \dfrac{\theta_{\max} - \tilde{\theta}_v}{\theta_{\max} - \theta_v}, & \text{if } j < v \text{ and } \tilde{\theta}_v > \theta_v, \\ (\theta_{\max} - \tilde{\theta}_v)\dfrac{\theta_j - \theta_{\max}}{\theta_{\max} - \theta_v} + \theta_{\max}, & \text{if } j > v \text{ and } \tilde{\theta}_v > \theta_v \end{cases}$$

where $\theta_{\max} = 1/(1 + \#S - v)$ is the upper bound for $\tilde{\theta}_v$ and $\theta_{\text{suc}} = \sum_{k=v+1}^{\#S} \theta_k$ is the original total mass of the parameters in θ_S larger than θ_v.

The order-preserving covariation scheme preserves the order (from smallest to largest) of the parameters, whilst proportional and uniform ones do not. This is useful in situations where there is qualitative information about one probability being larger than another: it then makes sense to preserve this order during sensitivity analysis.

Definitions 5 to 7 assume $\theta_S \in \Delta_{\#S-1}$. Next we specialise them to apply to the parameter vector θ of a MM.

Definition 8. In the notation of Definitions 2 and 5, let θ be the parameter vector of a MM. For $V \subset [k]$, let $V_i = S_i \cap V$ and σ_i a $\tilde{\theta}_{V_i}$-covariation scheme for each $i \in [n]$. Then:

Table 2
Probabilities associated to the BN in Example 4.

$\theta_1 = 0.2$	$\theta_2 = 0.3$	$\theta_3 = 0.5$	$\theta_{11} = 0.2$	$\theta_{21} = 0.3$	$\theta_{31} = 0.5$
$\theta_{12} = 0.3$	$\theta_{22} = 0.3$	$\theta_{32} = 0.4$	$\theta_{13} = 0.7$	$\theta_{23} = 0.2$	$\theta_{33} = 0.1$
$\theta_{111} = 0.1$	$\theta_{211} = 0.2$	$\theta_{311} = 0.7$	$\theta_{112} = 0.1$	$\theta_{212} = 0.3$	$\theta_{312} = 0.6$
$\theta_{113} = 0.2$	$\theta_{213} = 0.3$	$\theta_{313} = 0.5$	$\theta_{121} = 0.1$	$\theta_{221} = 0.4$	$\theta_{321} = 0.5$
$\theta_{122} = 0.3$	$\theta_{222} = 0.6$	$\theta_{322} = 0.1$	$\theta_{123} = 0.3$	$\theta_{223} = 0.5$	$\theta_{323} = 0.2$
$\theta_{131} = 0.8$	$\theta_{231} = 0.1$	$\theta_{331} = 0.1$	$\theta_{132} = 0.7$	$\theta_{232} = 0.2$	$\theta_{332} = 0.1$
$\theta_{133} = 0.4$	$\theta_{233} = 0.5$	$\theta_{333} = 0.1$			

- a $\tilde{\theta}_V$*-covariation scheme* for θ is a function $\sigma : \bigtimes_{i\in[n]} \Delta_{\#S_i-1} \to \bigtimes_{i\in[n]} \Delta_{\#S_i-1}$ such that $\sigma_{|S_i} = \sigma_i$, where $\sigma_{|S_i}$ denotes the restriction of σ over $\Delta_{\#S_i-1}$.
- a $\tilde{\theta}_V$-covariation scheme for θ is called *proportional* if σ_i is a $\tilde{\theta}_{V_i}$-proportional covariation scheme whenever $V_i \neq \emptyset$.

Definition 8 formalizes how parameters in a MM need to covary for any choice of varied parameters $\tilde{\theta}_V$. For instance, in a BN model the sets S_i, $i \in [n]$, denote the conditional probability distributions of any vertex given a specific combination of parents. If a full single CPT analysis is performed then $\tilde{\theta}_V$ includes one parameter from each conditional distribution associated to a given vertex. For such distributions, since V_i is non-empty, a standard σ_i covariation scheme is applied. In all other cases V_i is empty and the $\tilde{\theta}_V$-covariation scheme for θ returns the original value of the parameters since σ_i is defined as the identity function.

Our definition of a covariation scheme for a MM model encompasses all types of sensitivity analyses usually considered in BN models: one-way sensitivity analysis if V consists of one element only, full single CPT analyses as illustrated in the previous paragraph, multi-way analyses where multiple parameters from the same conditional distribution are varied, or generic multi-way analyses where any combination of parameters can be varied.

Our definition of a $\tilde{\theta}_{V_i}$-covariation scheme σ guarantees that if $\mathrm{P} \in \mathrm{MM}(A, S)$ then $\sigma(\mathrm{P}) \in \mathrm{MM}(A, S)$ for all covariation schemes σ, i.e. $\sigma(\mathrm{P})$ belongs to the same monomial model.

4.2. Global dissimilarity

In a sensitivity analysis, once a parameter is varied and those that must respect the sum-to-one condition are covaried, it is of interest to assess how close the original and varied BN distributions are. This closeness can be quantified using different distances or divergences. The most commonly used distance in sensitivity studies is the so called Chan-Darwiche (CD) distance [12]. The CD distance is equal to the DeRobertis distance, which has been used for quite some time in the Bayesian inference literature [27]. For two probability distributions P and Q over a finite space $\mathbb{Y}$ this is

$$\mathcal{D}_{\mathrm{CD}}(\mathrm{P}, \mathrm{Q}) = \ln \max_{y\in\mathbb{Y}} \left(\frac{\mathrm{P}(y)}{\mathrm{Q}(y)}\right) - \ln \min_{y\in\mathbb{Y}} \left(\frac{\mathrm{P}(y)}{\mathrm{Q}(y)}\right) = \max_{y,y'\in\mathbb{Y}} \ln \left(\frac{\mathrm{P}(y)\,\mathrm{Q}(y')}{\mathrm{P}(y')\,\mathrm{Q}(y)}\right),$$

where ln is the usual natural logarithm. Until recently, proportional covariation had a theoretical justification only for one-way analyses in BN models, since this scheme minimizes the CD distance between the original and the varied distributions [12]. In [39] it is proven that this is also true for full single CPT analyses in any multilinear MM.

Proportional covariation also minimizes the ϕ-divergence from the original to the varied distribution in full single CPT analyses [39]. The ϕ-divergence from P to Q is defined as

$$\mathcal{D}_\phi(\mathrm{Q}||\mathrm{P}) = \sum_{y\in\mathbb{Y}} \mathrm{P}(y)\phi\left(\frac{\mathrm{Q}(y)}{\mathrm{P}(y)}\right), \qquad \phi \in \Phi,$$

where Φ is the class of convex functions $\phi(x)$, $x \geq 0$, such that $\phi(1) = 0$, $0\,\phi(0/0) = 0$ and $0\phi(x/0) = \lim_{x\to\infty} \phi(x)/x$. The I-divergence in Definition 3 can be seen as a special instance of ϕ-divergences for $\phi(x) = x\ln(x)$.

Example 6. The BN of Example 4 is refined with the numerical specification of its parameters given in Table 2. Fig. 2 reports the CD distance and I-divergence for variations of θ_2 under different three covariation schemes. These plots show the optimality of proportional covariation (short-dashed line) which for both metrics takes smaller values than the other schemes.

5. A novel categorization of multi-way sensitivity analyses

Full single CPT analyses, for which the optimality of proportional covariation has been already proven, highly restrict the parameters that can be varied. To our knowledge, the only attempt in defining other multi-way sensitivity analyses is given in [5], where *balanced* analyses are introduced. In a nutshell, these reduce a multi-way problem into a one-way analysis by restricting the possible parameter variations.

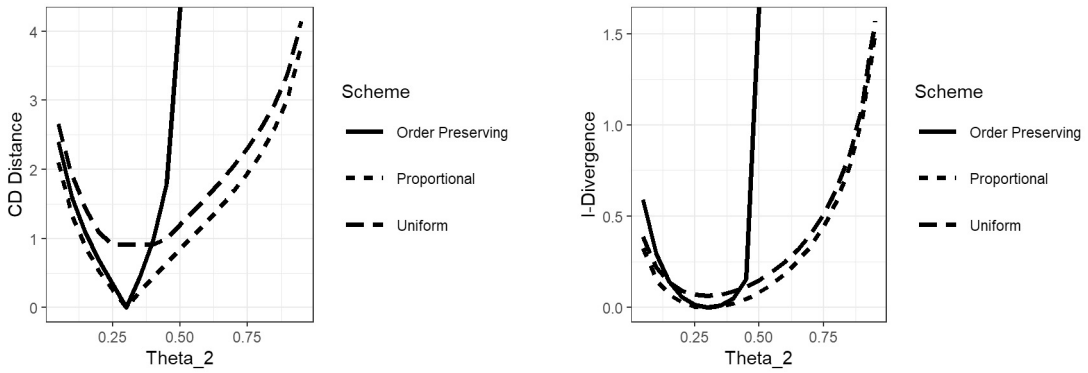


Fig. 2. Measures of dissimilarity for the BN in Example 4 under different covariation schemes: CD distance (left) and I-divergence (right).

Because of the monomial structure of their atomic probabilities, more general ways to choose parameters to be varied can be defined for MMs. These new multi-way analyses depend on the partition $\{S_1, \ldots, S_n\}$ of the k parameters of a MM and on the sets $V_i = V \cap S_i$, where $V \subset [k]$ and $i \in [n]$. Let $C = \bigcup_{i \in [n]: V_i \neq \emptyset} S_i$ be the union of all S_i for which V_i is not empty and $F = [k] \setminus C$. The set F includes the indices of the parameters that do not need to be covaried, whilst C is the index set of the (co)varied parameters. Definition 9 gives special ways to choose V which depend on the model structure. Below by division we mean standard division between monomials [17].

Definition 9. A sensitivity analysis is said to be

- *simple* if, for all $i, j \in C$, $\theta_i \theta_j$ does not divide θ^{A_y} for any $y \in \mathbb{Y}$;
- *complete* if, for all $H \in \times_{i \in [n]: V_i \neq \emptyset} S_i$, there exists at least one $y \in \mathbb{Y}$ such that θ_H divides θ^{A_y};
- *ordered* if, given an ordered sequence of sets $S_{k_1}, \ldots, S_{k_l}$ such that $V_{k_i} = V \cap S_{k_i} \neq \emptyset$, $i \in [l]$, and $\{S_{k_1}, \ldots, S_{k_l}\} \subseteq \{S_1, \ldots, S_n\}$, for all H in the set

$$\left\{S_{k_1} \setminus V\right\} \cup_{i \in [l-1]} \left\{ \times_{j \in [i]} V_{k_j} \times \{S_{k_{i+1}} \setminus V\} \right\} \cup \{\times_{i \in [l]} V_{k_i}\}, \tag{3}$$

there exists at least one $y \in \mathbb{Y}$ such that θ_H divides θ^{A_y}.

Recall that the set C in a simple sensitivity analysis includes the indices of the varied or covaried parameters. In a complete analysis all sets $H \in \times_{i \in [n]: V_i \neq \emptyset} S_i$ include the same number of indices, equal to $\sum_{i \in [n]} \mathbb{1}_{\{S_i \cap V \neq \emptyset\}}$. Conversely in an ordered analysis different sets H include a different number of indices. This implies that all monomials θ_H have the same degree in complete analyses, whilst they have different degrees in ordered ones.[2]

Example 7. For an ordered analysis suppose $\{S_{k_1}, \ldots, S_{k_l}\} = \{S_1, S_2, S_3\}$. Then (3) becomes

$$\{S_1 \setminus V\} \cup \{V_1 \times \{S_2 \setminus V\}\} \cup \{V_1 \times V_2 \times \{S_3 \setminus V\}\} \cup \{V_1 \times V_2 \times V_3\}.$$

Although the definition of such new analyses may appear obscure, these have a very simple graphical interpretation and include some well-known sensitivity analyses. In a simple sensitivity analysis no (co)varied parameters appear in the same monomial. It thus includes the following analyses in BN models:

- one-way sensitivity analyses: one parameter of a CPT of a vertex is varied;
- full single CPT analyses: one parameter from each conditional distribution of a vertex is varied;
- multi-way analyses where two or more parameters from one conditional distribution are varied;
- multi-way analyses where parameters from CPTs associated to incompatible parent configurations are varied. To see this consider the BN of Example 4 and suppose the parameters θ_{22} and θ_{311} are varied, implying respectively that $Y_1 = 2$ and $Y_1 = 1$. From Table 1 we can see that no two parameters from the associated conditional probability mass functions appear in the same monomial.

It is straightforward to see that no two (co)varied parameters appear in the same monomial representing the probabilities of a BN for all choices listed above. For instance, in a one-way sensitivity analysis, the (co)varied parameters belong to

[2] The degree of a monomial is the sum of the exponents of all its unknowns.

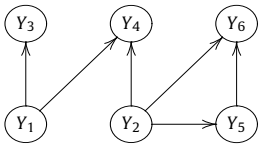


Fig. 3. A BN to illustrate complete and ordered sensitivity analyses.

the same conditional distribution and cannot appear jointly in the same atomic probability. Similarly, in a full single CPT analysis all (co)varied parameters belong to one CPT of a vertex of a BN and again these probabilities cannot appear in the factorization of the same atomic probability. The same reasoning applies for the other cases.

Complete sensitivity analyses are such that all possible combinations of (co)varied parameters in different sets S_i, for $i \in [n]$ such that $V_i \neq \emptyset$, appear in at least one monomial. The simplest possible example of such analyses is in the case of a BN consisting of two independent random variables where one parameter from each distribution is varied. But more generally such analyses are associated to varied parameters in CPTs implying disjoint parent sets. To illustrate this consider the BN in Fig. 3. The variation of one parameter from the distribution of $Y_3|Y_1$ and another from the distribution of $Y_5|Y_2$ would give a complete sensitivity analysis since the conditioning variables are different. As an additional example, consider the BN of Example 4 and suppose θ_1 and θ_{122} are varied. Notice that $(\theta_1, \theta_2, \theta_3) \in \Delta_2$ and $(\theta_{122}, \theta_{222}, \theta_{322}) \in \Delta_2$. This choice of parameters would be a complete analysis if every element of

$$(\theta_1\theta_{122}, \theta_1\theta_{222}, \theta_1\theta_{322}, \theta_2\theta_{122}, \theta_2\theta_{222}, \theta_2\theta_{322}, \theta_3\theta_{122}, \theta_3\theta_{222}, \theta_3\theta_{322})$$

appeared in at least one monomial in Table 1. Since this is not the case, this choice of parameters does not correspond to a complete sensitivity analysis. More generally, it can be seen that for any choice of two parameters each from a different conditional distribution (possibly in the same CPT) there is no pair of θ's such that all combinations of covaried parameters appear in the same monomial.

Ordered analyses imply an order over the varied parameters. A varying parameter needs to be a probability which is conditional on the events associated to preceding varying parameters in this order. An example from Fig. 3 illustrates this for BNs. Suppose the parameter associated to $\mathrm{P}(Y_2 = y_2)$ is varied. Then in an ordered analysis any parameter from $\mathrm{P}(Y_5 = y_5|Y_2 = y_2)$ can be varied and, if so, also any parameter from $\mathrm{P}(Y_6 = y_6|Y_5 = y_5, Y_2 = y_2)$. As an additional example, consider again the BN in Example 4 and suppose the parameters θ_1 and θ_{311} are varied, defining the set V of varied indexes. Let $\theta_{S_1} = (\theta_1, \theta_2, \theta_3)$ and $\theta_{S_2} = (\theta_{111}, \theta_{211}, \theta_{311})$. For this choice of parameters (3) can be written as $S_1 \setminus V \cup \{V_1 \times \{S_2 \setminus V\}\} \cup \{V_1 \times V_2\}$, where $V_i = S_i \cap V$, for $i = 1, 2$. The monomials indexed by $S_1 \setminus V$ are (θ_2, θ_3), those indexed by $V_1 \times \{V_1 \times \{S_2 \setminus V\}\}$ are $(\theta_1\theta_{111}, \theta_1\theta_{211})$ and those indexed by $V_1 \times V_2$ are $(\theta_1\theta_{311})$. This choice of parameters corresponds to an ordered analysis since any of these monomials divides at least one monomial atomic probability in Table 1. Conversely, the choice of varied parameters θ_1 and θ_{122}, for instance, would not correspond to an ordered sensitivity analysis.

6. The optimality of proportional covariation

After the variation of some parameters of a MM with indices in V to a value $\tilde{\theta}_V$, a $\tilde{\theta}_V$-covariation scheme needs to be applied to respect all sum-to-one conditions of the model. In this section, given $\mathrm{P} \in \mathrm{MM}(A, S)$ and its $\tilde{\theta}_V$-proportional covariation $\tilde{\mathrm{P}}$ we first determine a family L of Q densities for which the Pythagorean equality $\mathcal{D}(\mathrm{Q}||\mathrm{P}) = \mathcal{D}(\mathrm{Q}||\tilde{\mathrm{P}}) + \mathcal{D}(\tilde{\mathrm{P}}||\mathrm{Q})$ holds. This will later allow us to demonstrate for which cases a $\tilde{\theta}_V$-proportional covariation scheme is optimal.

6.1. A geometric characterization of sensitivity analysis

In the notation of Definition 8, let $\emptyset \neq V \subset [k]$, $C = \bigcup_{i \in [n]: V_i \neq \emptyset} S_i$ and $F = [k] \setminus C$. Let $\Delta^F = \bigtimes_{i \in [n]: V_i = \emptyset} \Delta_{\#S_i - 1}$ and $\Delta^C = \bigtimes_{i \in [n]: V_i \neq \emptyset} \Delta_{\#S_i - 1}$. The set $[k]$ is so partitioned into V, $C \setminus V$ and F, namely the index set of the varied, covaried and fixed parameters. A generic parameter vector can be written as $\theta = (\theta_F, \theta_V, \theta_{C\setminus V})$ and for $\mathrm{P} \in \mathrm{MM}(A, S)$ the atomic probability of $y \in \mathbb{Y}$ can be written as $\mathrm{P}(y) = \theta_F^{A_{y,F}} \theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}$. For any given $\theta_F \in \Delta^F$, $\mathrm{Slice}(\theta_F)$ is the subset of densities in $\mathrm{MM}(A, S)$ for which the parameters indexed by F take value θ_F, namely

$$\mathrm{Slice}(\theta_F) = \left\{ \mathrm{P} \in \mathrm{MM}(A, S) : \mathrm{P}(y) = \theta_F^{A_{y,F}} \theta_C^{A_{y,C}} \text{ for some } \theta_C \in \Delta^C \text{ and all } y \in \mathbb{Y} \right\},$$

where $\theta_C^{A_{y,C}} = \theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}$. It holds $\mathrm{MM}(A, S) = \bigcup_{\theta_F \in \Delta^F} \mathrm{Slice}(\theta_F)$.

Example 8. Consider the BN in Example 4 with monomial atomic probabilities given in Table 1. Suppose the parameters θ_1, θ_{11} and θ_{111} are varied. Then $\theta_V = (\theta_1, \theta_{11}, \theta_{111})$, $\theta_{C\setminus V} = (\theta_2, \theta_3, \theta_{21}, \theta_{31}, \theta_{211}, \theta_{311})$ and

$$\theta_F = (\theta_{i2}, \theta_{i3}, \theta_{i21}, \theta_{i31}, \theta_{i12}, \theta_{i22}, \theta_{i23}, \theta_{i13}, \theta_{i23}, \theta_{i33})_{i\in[3]}.$$

In $\text{Slice}(\theta_F)$ the entries in θ_F are fixed to the values given in Table 2 and are not allowed to vary, whilst those in θ_V and $\theta_{C\setminus V}$ can vary. Atomic probabilities in $\text{Slice}(\theta_F)$ are still written as $\text{P}(y) = \theta_i\theta_{ji}\theta_{lji}$, for appropriate choices of the indexes i, j and l.

Theorem 2 shows that $\text{P} \in \text{MM}(A, S)$ and its $\tilde{\theta}_V$-proportional covariation mass function belong to the same slice.

Theorem 2. *Let $\theta = (\theta_F, \theta_V, \theta_{C\setminus V})$ be the parameter vector of $\text{P} \in \text{MM}(A, S)$ and let $\tilde{\theta} = (\tilde{\theta}_F, \tilde{\theta}_V, \tilde{\theta}_{C\setminus V})$ be the parameter vector of the $\tilde{\theta}_V$-proportional covariation of P called $\tilde{\text{P}}$. Then $\tilde{\text{P}} \in \text{Slice}(\theta_F)$, that is $\tilde{\theta}_F = \theta_F$.*

Proof. The proof follows straightforward from Definition 8. Indeed $\tilde{\theta}_F = \theta_F$, $\tilde{\theta}_V$ is given and $\tilde{\theta}_{C\setminus V} = \left(\left(\frac{1-|\tilde{\theta}_{V_i}|}{1-|\theta_{V_i}|}\theta_j\right)_{j\in S_i}\right)_{i\in[n]}$. □

Theorem 2 guarantees that proportional covariation schemes in MMs do not affect probabilities that customarily are not changed in sensitivity analysis, i.e. those in θ_F. Next, let

$$L_{\text{sensi}}(\theta_F, \tilde{\theta}_V) = \text{Slice}(\theta_F) \cap \{P \in \text{MM}(A, S) : \theta_V = \tilde{\theta}_V\},$$

denote the family of distributions where only the parameters $\theta_{C\setminus V}$ can vary. This is the set of distributions which are considered in sensitivity analysis, where only parameters that must be covaried after a parameter variation are allowed to change. In other words, $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ includes all distributions that do *not* assume a specific covariation scheme. It follows from Theorem 2 that $\tilde{\text{P}} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$.

Example 9. Consider the setup of Example 5 where $\theta = (\theta_1, \theta_2, \theta_3) \in \Delta_2$, $V = 1$ and $\tilde{\theta}_V = 0.4$. Since this situation corresponds to a simple discrete probability distribution, $\text{Slice}(\theta_F) = \Delta_2$, i.e. there are no parameters in θ_F. So $\text{Slice}(\theta_F)$ is the simplex represented in Fig. 1. The space $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ is represented by the full line traversing the simplex, corresponding to the intersection of $\tilde{\theta}_1 = 0.4$ with Δ_2.

The above example highlighted that $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ includes the model's distributions where only the parameters in $C \setminus V$, those that need to be covaried, can vary. Henceforth, we thus consider distributions in $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$.

Now the objective is to identify if, for a distribution $\text{P} \in MM(A, S)$, proportional covariation is its I-projection into $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$, meaning that out of all possible covariation schemes the proportional scheme is the one minimizing the I-divergence. We achieve this by checking if the Pythagorean equality in (2) holds for the proportional covariation scheme. The following subsection illustrates our approach in a simple example.

6.2. A comprehensive example

Consider the following setup: a student can either fail (coded as 1), pass (coded as 2) or obtain a distinction (coded as 3) in an exam. If the student fails, she is given an additional chance where again she can either fail, pass or get a distinction. The probabilities for the two tries are assumed to be different.

The above scenario can be modelled by a $\text{MM}(A, S)$ with parameters $(\theta_1, \theta_2, \theta_3, \psi_1, \psi_2, \psi_3)$ such that $|(\theta_1, \theta_2, \theta_3)| = 1$ and $|(\psi_1, \psi_2, \psi_3)| = 1$, and matrix A

	θ_1	θ_2	θ_3	ψ_1	ψ_2	ψ_3	P(y)
y_1	1	0	0	1	0	0	$\theta_1\psi_1$
y_2	1	0	0	0	1	0	$\theta_1\psi_2$
y_3	1	0	0	0	0	1	$\theta_1\psi_3$
y_4	0	1	0	0	0	0	θ_2
y_5	0	0	1	0	0	0	θ_3

(4)

where for clarity we labelled the columns and the rows with the associated parameters and events, respectively, and reported the atomic probabilities. The parameters θ's represent the outcome of the first try, whilst the parameters ψ's of the second one. For illustration purposes, we consider the following five choices of varied parameters θ_V: (θ_1), (θ_2), (ψ_1), (θ_1, ψ_1) and (θ_2, ψ_1), where in the last two cases the two parameters are varied simultaneously. For $\text{P}, \tilde{\text{P}}, \text{Q} \in \text{MM}(A, S)$, we denote by $(\theta_i, \psi_i)_{i\in[3]}$, $(\tilde{\theta}_i, \tilde{\psi}_i)_{i\in[3]}$ and $(\bar{\theta}_i, \bar{\psi}_i)_{i\in[3]}$ the parameter vectors of P, $\tilde{\text{P}}$ and Q respectively. The I-divergence from P to Q takes the form

$$\mathcal{D}(\text{Q}\,||\,\text{P}) = \sum_{i=2,3} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\theta_i}\right) + \bar{\theta}_1 \sum_{i\in[3]} \bar{\psi}_i \ln\left(\frac{\bar{\theta}_1\bar{\psi}_i}{\theta_1\psi_i}\right).$$

The Pythagorean equality in equation (2) can then be written as

$$\sum_{i=2,3} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\theta_i}\right) + \bar{\theta}_1 \sum_{i\in[3]} \bar{\psi}_i \ln\left(\frac{\bar{\theta}_1\bar{\psi}_i}{\theta_1\psi_i}\right) - \sum_{i=2,3} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\tilde{\theta}_i}\right) - \bar{\theta}_1 \sum_{i\in[3]} \bar{\psi}_i \ln\left(\frac{\bar{\theta}_1\bar{\psi}_i}{\tilde{\theta}_1\tilde{\psi}_i}\right)$$
$$- \sum_{i=2,3} \tilde{\theta}_i \ln\left(\frac{\tilde{\theta}_i}{\theta_i}\right) - \tilde{\theta}_1 \sum_{i\in[3]} \tilde{\psi}_i \ln\left(\frac{\tilde{\theta}_1\tilde{\psi}_i}{\theta_1\psi_i}\right) = 0. \tag{5}$$

Next we look at the form of the above equality for each of the possible varied parameter choices. For each case, we consider only densities $\mathrm{Q} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$, that is all possible ways to covary relevant parameters.

1. For $\tilde{\theta}_V = \tilde{\theta}_1$, we consider Q such that $\mathrm{Q} \in \text{Slice}(\psi_1, \psi_2, \psi_3)$ and $\bar{\theta}_1 = \tilde{\theta}_1$. Then $\tilde{\mathrm{P}}$ has parameter vector $(\tilde{\theta}_i, \psi_i)_{i\in[3]}$, whilst Q has parameters $(\tilde{\theta}_1, \bar{\theta}_2, \bar{\theta}_3, \psi_1, \psi_2, \psi_3)$. Under these conditions and noticing that $\tilde{\psi} = \bar{\psi} = \psi$, (5) can be simplified to
$$\sum_{i\in[3]} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\theta_i}\right) - \sum_{i\in[3]} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\tilde{\theta}_i}\right) - \sum_{i\in[3]} \tilde{\theta}_i \ln\left(\frac{\tilde{\theta}_i}{\theta_i}\right) = 0. \tag{6}$$
By substituting $\tilde{\theta}_i = \theta_i(1-\tilde{\theta}_1)/(1-\theta_1)$ into the logarithms, (6) reduces to
$$\ln\left(\frac{1-\theta_1}{1-\tilde{\theta}_1}\right) \sum_{i=2,3} (\tilde{\theta}_i - \bar{\theta}_i) = 0,$$
which holds for all $\mathrm{Q} \in L_{\text{sensi}}((\psi_1, \psi_2, \psi_3), \tilde{\theta}_1)$ since $\sum_{i=2,3} \bar{\theta}_i = \sum_{i=2,3} \tilde{\theta}_i = 1 - \tilde{\theta}_1$.
2. For $\tilde{\theta}_V = \tilde{\theta}_2$, we consider Q such that $\mathrm{Q} \in \text{Slice}(\psi_1, \psi_2, \psi_3)$ and $\bar{\theta}_2 = \tilde{\theta}_2$. Then $\tilde{\mathrm{P}}$ has parameter vector $(\tilde{\theta}_i, \psi_i)_{i\in[3]}$, whilst Q has parameters $(\bar{\theta}_1, \tilde{\theta}_2, \bar{\theta}_3, \psi_1, \psi_2, \psi_3)$. Under these conditions and noticing that $\tilde{\psi} = \bar{\psi} = \psi$, (5) can be written as equation (6) which can be simplified as in the previous case to show that the equality holds for all $\mathrm{Q} \in L_{\text{sensi}}((\psi_1, \psi_2, \psi_3), \tilde{\theta}_2)$.
3. For $\tilde{\theta}_V = \tilde{\psi}_1$, we consider Q such that $\mathrm{Q} \in \text{Slice}(\theta_1, \theta_2, \theta_3)$ and $\bar{\psi}_1 = \tilde{\psi}_1$. Then $\tilde{\mathrm{P}}$ has parameter vector $(\theta_i, \tilde{\psi}_i)_{i\in[3]}$, whilst Q has parameters $(\theta_1, \theta_2, \theta_3, \tilde{\psi}_1, \bar{\psi}_2, \bar{\psi}_3)$. Under these conditions and noticing that $\tilde{\theta} = \theta$, (5) can be written as
$$\theta_1 \left(\sum_{i=2,3} \bar{\psi}_i \ln\left(\frac{\bar{\psi}_i}{\psi_i}\right) - \sum_{i=2,3} \bar{\psi}_i \ln\left(\frac{\bar{\psi}_i}{\tilde{\psi}_i}\right) - \sum_{i=2,3} \tilde{\psi}_i \ln\left(\frac{\tilde{\psi}_i}{\psi_i}\right) \right) = 0,$$
which can be simplified as in the two previous cases to show that the equality holds for all $\mathrm{Q} \in L_{\text{sensi}}((\theta_1, \theta_2, \theta_3), \tilde{\psi}_1)$.
4. For $\tilde{\theta}_V = (\tilde{\theta}_1, \tilde{\psi}_1)$, we consider Q such that $\bar{\theta}_1 = \tilde{\theta}_1$ and $\bar{\psi}_1 = \tilde{\psi}_1$, since there are no parameters with index in F. Then $\tilde{\mathrm{P}}$ has parameter vector $(\tilde{\theta}_i, \tilde{\psi}_i)_{i\in[3]}$, whilst Q has parameters $(\tilde{\theta}_1, \bar{\theta}_2, \bar{\theta}_3, \tilde{\psi}_1, \bar{\psi}_2, \bar{\psi}_3)$. Under these conditions, (5) can be written as
$$\left(\sum_{i=2,3} \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\theta_i}\right) - \bar{\theta}_i \ln\left(\frac{\bar{\theta}_i}{\tilde{\theta}_i}\right) - \tilde{\theta}_i \ln\left(\frac{\tilde{\theta}_i}{\theta_i}\right) \right)$$
$$+ \theta_1 \left(\sum_{i=2,3} \bar{\psi}_i \ln\left(\frac{\bar{\theta}_1\bar{\psi}_i}{\theta_1\psi_i}\right) - \sum_{i=2,3} \bar{\psi}_i \ln\left(\frac{\bar{\theta}_1\bar{\psi}_i}{\tilde{\theta}_1\tilde{\psi}_i}\right) - \sum_{i=2,3} \tilde{\psi}_i \ln\left(\frac{\tilde{\theta}_1\tilde{\psi}_i}{\theta_1\psi_i}\right) \right) = 0.$$
The above equation can be simplified by combining the steps used in the previous cases to show that the equality holds for all $\mathrm{Q} \in L_{\text{sensi}}(\cdot, (\tilde{\theta}_1, \tilde{\psi}_1))$.
5. For $\tilde{\theta}_V = (\tilde{\theta}_2, \tilde{\psi}_1)$, we consider Q such that $\bar{\theta}_2 = \tilde{\theta}_2$ and $\bar{\psi}_1 = \tilde{\psi}_1$, since there are no parameters with index in F. Then $\tilde{\mathrm{P}}$ has parameter vector $(\tilde{\theta}_i, \tilde{\psi}_i)_{i\in[3]}$, whilst Q has parameters $(\bar{\theta}_1, \tilde{\theta}_2, \bar{\theta}_3, \tilde{\psi}_1, \bar{\psi}_2, \bar{\psi}_3)$. Under these conditions and using similar steps to the previous points, (5) can be written as
$$\ln\left(\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)(\bar{\theta}_3 - \tilde{\theta}_3) + \tilde{\psi}_1 \ln\left(\frac{\tilde{\psi}_1}{\psi_1}\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)(\bar{\theta}_1 - \tilde{\theta}_1) + (1-\tilde{\psi}_1)\ln\left(\frac{1-\tilde{\psi}_1}{1-\psi_1}\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)(\bar{\theta}_1 - \tilde{\theta}_1) = 0. \tag{7}$$
Although the terms in (5) can be re-arranged differently, the equality cannot be derived and consequently for the choice $\tilde{\theta}_V = (\tilde{\theta}_2, \tilde{\psi}_1)$ the family of distributions for which the Pythagorean identity holds is restricted.

Notice that the first four choices of parameters corresponded to the analyses introduced in Definition 9, namely simple in the first three cases and ordered in the fourth, whilst the last choice of parameters does not correspond to any of the sensitivity analyses of Definition 9.

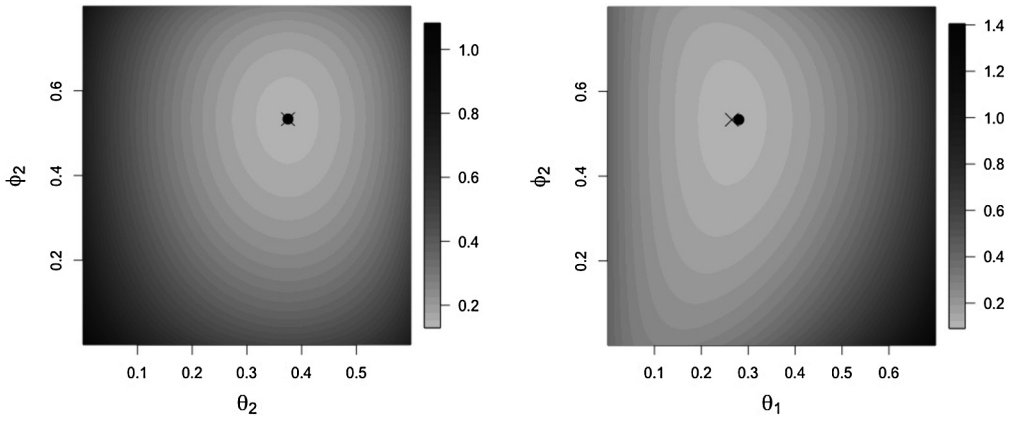


Fig. 4. I-divergences for the examples of Section 6.2 (case 4: left; case 5: right). The star represents the minimum I-divergence and the dot represents proportional covariation.

Suppose we now refine the model definition with the following probability specifications: $\theta_1 = 0.2$, $\theta_2 = 0.5$, $\theta_3 = 0.3$, $\psi_1 = 0.4$, $\psi_2 = 0.4$ and $\psi_3 = 0.2$. Next consider the choices of parameters varied in points 4 and 5 above. First suppose that $\tilde{\theta}_1 = 0.4$ and $\tilde{\psi}_1 = 0.2$. For this choice of varied parameters we showed that the Pythagorean identity holds for all $\mathrm{Q} \in L_{\text{sensi}}(\cdot, (\tilde{\theta}_1, \tilde{\psi}_1))$. In this case proportional covariation minimizes the I-divergence between the original and the varied distribution, as reported on the left of Fig. 4. Consider now case 5 and suppose $\tilde{\theta}_2 = 0.3$ and $\tilde{\psi}_1 = 0.2$. We showed that the Pythagorean identity does not hold for all $\mathrm{Q} \in L_{\text{sensi}}(\cdot, (\tilde{\theta}_2, \tilde{\psi}_1))$. The identity holds in the restricted family of distributions characterized by (7). In this case the I-divergence is not minimized by proportional covariation as reported on the right of Fig. 4. It can be further showed that the CD distance is minimized by proportional covariation for the choice of varied parameters in point 4, whilst it is not for the varied parameters of point 5.

6.3. Proportional covariation and I-projections

Given a $\mathrm{P} \in \mathrm{MM}(A, S)$ and its $\tilde{\theta}_V$-proportional covariation $\tilde{\mathrm{P}}$, Theorem 3 identifies the set of distributions $\mathrm{Q} \in L_{\text{sensi}}$ that satisfy the Pythagorean identity in (2). This is the first step towards the proof of optimality of proportional covariation. For $H \subseteq C$ define

$$\mathbb{Y}_H = \left\{ y \in \mathbb{Y} : A_{y,i} = 1 \text{ for all } i \in H \text{ and } A_{y,i} = 0 \text{ for all } i \in C \setminus H \right\}.$$

Given a parameter vector θ_H including some of the covaried parameters, $H \subseteq C$, $\mathbb{Y}_H$ denotes the subset of the sample space including atomic probabilities where all elements of θ_H have a non-zero exponent.

For $y \in \mathbb{Y}$, with a slight abuse of notation, if $A_{y,i} = 1$ for all $i \in B \subseteq [k]$ we write $\theta_B^{A_{y,B}} = \prod_{i \in B} \theta_i^{A_{y,i}} = \theta_B$. Thus the symbol θ_B might indicate the vector $(\theta_i)_{i \in B}$ and the square-free monomial $\prod_{i \in B} \theta_i$. The context clarifies which interpretation applies. In particular for all $y \in \mathbb{Y}_H$ and $B \subseteq H$ we write $\theta_B^{A_{y,B}} = \theta_B$ and $\mathrm{P}(y) = \theta_F^{A_{y,F}} \theta_H$ for any $\mathrm{P} \in \mathrm{MM}(A, S)$.

Theorem 3. *Let $\emptyset \neq V \subset [k]$, $\mathrm{P} \in \mathrm{MM}(A, S)$, with parameter vector $(\theta_F, \theta_V, \theta_{C \setminus V})$, and $\tilde{\mathrm{P}}$ the $\tilde{\theta}_V$-proportional covariation of P with parameter $\tilde{\theta} = (\theta_F, \tilde{\theta}_V, \tilde{\theta}_{C \setminus V})$. For any $H \subseteq C$, define $\tilde{\theta}_{V \cap H} = 1$ if $V \cap H = \emptyset$ and $\sum_{y \in \mathbb{Y}_H} \theta_F^{A_{y,F}} = 0$ if $\mathbb{Y}_H = \emptyset$. A probability mass function $\mathrm{Q} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ with parameter $\bar{\theta} = (\theta_F, \bar{\theta}_V, \bar{\theta}_{C \setminus V})$ satisfies*

$$\sum_{H \subseteq C, H \neq \emptyset} \tilde{\theta}_{V \cap H} (\bar{\theta}_{\{C \setminus V\} \cap H} - \tilde{\theta}_{\{C \setminus V\} \cap H}) \ln \left(\alpha \frac{\tilde{\theta}_{V \cap H}}{\theta_{V \cap H}} \right) \sum_{y \in \mathbb{Y}_H} \theta_F^{A_{y,F}} = 0, \tag{8}$$

where $\alpha = \prod_{i \in [n] : S_i \cap H \neq \emptyset} \prod_{j \in \{S_i \setminus V\} \cap H} (1 - |\tilde{\theta}_{V_i}|)/(1 - |\theta_{V_i}|)$, if and only if $\mathcal{D}(\mathrm{Q} || \mathrm{P}) = \mathcal{D}(\mathrm{Q} || \tilde{\mathrm{P}}) + \mathcal{D}(\tilde{\mathrm{P}} || \mathrm{Q})$.

The proof is in Appendix A.1. The first condition that $\mathrm{Q} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ corresponds to consider every possible way to covary parameters. The second condition given in (8) has one term only depending on the probability mass function Q, namely $\bar{\theta}_{\{C \setminus V\} \cap H}$, whilst all others can be derived straightforwardly once $\tilde{\theta}_V$ and $\mathrm{P} \in \mathrm{MM}(A, S)$ are given.

Example 10. Consider the setup of Section 6.2, where it was showed that for the variation of the parameter θ_1 the Pythagorean identity holds. Conversely, for the simultaneous variation of parameters θ_2 and ψ_1 it did not hold. Therefore, the equality in (8) must hold in the first case, but not in the second. We next check this is indeed the case.

Table 3
Elements of the sample space (left column) and contribution to (8) (right column) for the setup in Example 10 when θ_2 and ψ_1 are varied.

θ_2	$\tilde{\theta}_2(1-1)\ln\left(1\cdot\frac{\tilde{\theta}_2}{\theta_2}\right)\cdot 1$
θ_3	$1\cdot(\bar{\theta}_3-\tilde{\theta}_3)\ln\left(\frac{1-\tilde{\theta}_2}{1-\theta_2}\cdot\frac{1}{1}\right)\cdot 1$
$\theta_1\psi_1$	$\tilde{\psi}_1(\bar{\theta}_1-\tilde{\theta}_1)\ln\left(\frac{\tilde{\psi}_1}{\psi_1}\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)\cdot 1$
$\theta_1\psi_2$	$1\cdot(\bar{\theta}_1\bar{\psi}_2-\tilde{\theta}_1\tilde{\psi}_2)\ln\left(\frac{1-\tilde{\psi}_1}{1-\psi_1}\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)\cdot 1$
$\theta_1\psi_3$	$1\cdot(\bar{\theta}_1\bar{\psi}_3-\tilde{\theta}_1\tilde{\psi}_3)\ln\left(\frac{1-\tilde{\psi}_1}{1-\psi_1}\frac{1-\tilde{\theta}_2}{1-\theta_2}\right)\cdot 1$

When θ_1 is varied, the set C of (co)varied parameters is such that $\theta_C=(\theta_1,\theta_2,\theta_3)$. The summation in (8) considers all subsets H of C such that $\mathbb{Y}_H$ is non-empty: this is the case only for the indexes of θ_2 and θ_3. So (8) can be written as

$$1\cdot(\bar{\theta}_2-\tilde{\theta}_2)\ln\left(1\cdot\frac{1-\tilde{\theta}_1}{1-\theta_1}\right)\cdot 1+1\cdot(\bar{\theta}_3-\tilde{\theta}_3)\ln\left(1\cdot\frac{1-\tilde{\theta}_1}{1-\theta_1}\right)\cdot 1. \tag{9}$$

Since $\bar{\theta}_2+\bar{\theta}_3=\tilde{\theta}_2+\tilde{\theta}_3=1-\tilde{\theta}_1$, (9) is equal to zero, as expected.

When θ_2 and ψ_1 are varied, the set C of (co)varied parameters includes all model's parameters and the summation in (8) considers all subsets H of C coinciding with the sample space given in (4). So (8) can be written as the sum of the expressions on the rhs of Table 3. It can be noticed that this sum is equal to (7): the first row of Table 3 is zero, the second row is equal to the first term in (7), the third row is equal to the second term in (7) and the sum of the fourth and fifth rows is equal to the third term in (7) since $\bar{\tilde{\psi}}_2+\bar{\tilde{\psi}}_3=\tilde{\psi}_2+\tilde{\psi}_3=1-\tilde{\psi}_1$. As already noticed, for this choice of parameters the Pythagorean identity does not hold for all distributions in $L_{\text{sensi}}(\theta_F,\tilde{\theta}_V)$.

Corollary 1. *In the notation of Theorem 3, the probability mass function* $\mathrm{Q}\in L_{\text{sensi}}(\theta_F,\tilde{\theta}_V)$ *satisfies*

$$\sum_{H\subseteq C,H\neq\emptyset}\tilde{\theta}_{V\cap H}(\bar{\theta}_{\{C\setminus V\}\cap H}-\tilde{\theta}_{\{C\setminus V\}\cap H})\ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right)\sum_{y\in\mathbb{Y}_H}\theta_F^{A_{y,F}}\geq 0, \tag{10}$$

if and only if $\mathcal{D}(\mathrm{Q}\,||\,\mathrm{P})\geq\mathcal{D}(\mathrm{Q}\,||\tilde{\mathrm{P}})+\mathcal{D}(\tilde{\mathrm{P}}||\,\mathrm{Q})$.

This result follows by substituting the equalities in the proof of Theorem 3 with inequalities.

Since for $\tilde{\mathrm{P}}$, P and all the distributions Q characterized by (10) the Pythagorean inequality holds, then it can be proven that $\tilde{\mathrm{P}}$ is the I-projection of P into this well-specified family of distributions. Let

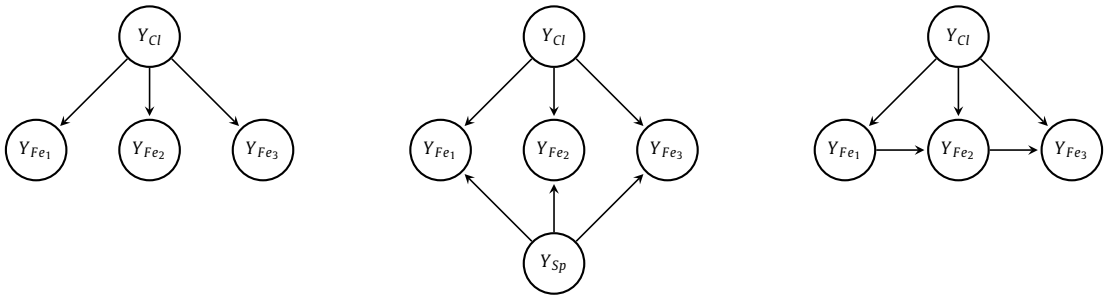
$$L_{\text{constr}}(\theta_F,\tilde{\theta}_V)=L_{\text{sensi}}(\theta_F,\tilde{\theta}_V)\cap\left\{\mathrm{Q}\in\mathrm{MM}(A,S):\sum_{H\subseteq C,H\neq\emptyset}\tilde{\theta}_{V\cap H}(\bar{\theta}_{\{C\setminus V\}\cap H}-\tilde{\theta}_{\{C\setminus V\}\cap H})\ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right)\sum_{y\in\mathbb{Y}_H}\theta_F^{A_{y,F}}\geq 0\right\}.$$

Corollary 2. *In the notation of Theorem 3,* $\tilde{\mathrm{P}}$ *is the I-projection of* P *in the set* $L_{\text{constr}}(\theta_F,\tilde{\theta}_V)$.

Proof. Let $\bar{L}$ be the smallest convex and closed subset of Δ_{q-1} which includes $L_{\text{constr}}(\theta_F,\tilde{\theta}_V)$. From Section 3, there exists a unique $\mathrm{P}^*\in\bar{L}$ such that $\mathcal{D}(\mathrm{Q}\,||\,\mathrm{P})\geq\mathcal{D}(\mathrm{Q}\,||\,\mathrm{P}^*)+\mathcal{D}(\mathrm{P}^*\,||\,\mathrm{P})$. But since $\tilde{\mathrm{P}}\in L_{\text{constr}}(\theta_F,\tilde{\theta}_V)\subseteq\bar{L}$ satisfies the Pythagorean identity then $\tilde{\mathrm{P}}=\mathrm{P}^*$. □

Csiszár and Shields [18] proved that the I-projection satisfies the Pythagorean identity using the fact that L is closed and convex. Here we took a different approach by taking advantage of the specific monomial form of the statistical models we study. By characterizing the class of distributions for which the Pythagorean identity holds, we have then been able to prove that proportional covariation is the I-projection within this family.

Although Corollary 2 demonstrates that proportional covariation minimizes the I-divergence between the original distribution and those in the set L_{constr}, it does not provide information on whether $L_{\text{constr}}(\theta_F,\tilde{\theta}_V)$ includes all distributions



(a) A Naive Bayes classifier. (b) A SPODE classifier with super parent Y_{Sp}. (c) A generic BN classifier.

Fig. 5. Examples of BN classifiers.

of interest in sensitivity analysis or not. More explicitly, Corollary 2 does not specify whether, given $\mathrm{P} \in \mathrm{MM}(A, S)$ and $\tilde{\theta}_V$, $\tilde{\mathrm{P}}$ is the I-projection of P in $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$. This is the case if and only if $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V) = L_{\text{constr}}(\theta_F, \tilde{\theta}_V)$, i.e. if for all $\mathrm{Q} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ the condition in (10) holds. Theorem 4 below states that for the multi-way analyses in Definition 9, proportional covariation is indeed the I-projection of the original distribution in $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$. Namely for such analyses $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V) = L_{\text{constr}}(\theta_F, \tilde{\theta}_V)$.

Theorem 4. *In the notation of Theorem 3, if $\tilde{\theta}_V$ is chosen according to a simple, complete or ordered sensitivity analysis, then $\tilde{\mathrm{P}}$ is the I-projection of* P *in $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$.*

The proof is given in Appendix A.2. The result holds for regular MMs and consequently for all graphical models entertaining a regular monomial parametrization. Illustrations of this result were given in Section 6.2: in the first four cases, corresponding to simple or ordered analyses, the Pythagorean identity holds for all $\mathrm{Q} \in L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$ and thus the $\tilde{\theta}_V$-proportional covariation scheme is the I-projection over the set of all distributions of interest. Conversely, in the fifth case of Section 6.2, which does not correspond to any of the new multi-way analyses of Definition 9, the Pythagorean identity holds in a restricted set of distributions, namely $L_{\text{constr}}(\theta_F, \tilde{\theta}_V)$. Thus, as specified by Corollary 2, proportional covariation is the I-projection over this restricted space only.

6.4. BN classifiers

BN classifiers are BNs whose graph entertains some specific properties designed for classification problems. BN classifiers have been successfully used in a wide array of real-world applications, with a competitive predictive performance against other classification techniques, because of their intuitiveness and computational efficiency [see e.g. 2, for a review]. A BN classifier is defined by partitioning the BN vertex set into the set of features and the classes. It is customary that the edge set of a BN classifier is such that feature variables are not allowed to have class children. For simplicity here we focus on univariate classification problems where there is a single class variable. However our results apply to multidimensional classes since in a BN classifier the class variables can be collapsed into a unique vertex.

BN classifiers range from the simplest naive Bayes classifier where the features are conditionally independent given the class variable (given in Fig. 5a), to generic dependence structures between the features (as for example in Fig. 5c). A BN classifier of interest is the super-parent-one-dependence estimator (SPODE) [31] where all features depend on one specific feature called super-parent (see Fig. 5b).

Since BN classifiers are BN models, they can be represented as MMs as shown in Example 4 [although other monomial representations exist 55,56]. Therefore, we can apply our methodology and deduce the following result.

Theorem 5. *Consider a naive Bayes classifier with features $Y_{Fe_1}, \ldots, Y_{Fe_m}$. In the notation of Theorem 3, if $\tilde{\theta}_V$ is chosen so that $V \subset \times_{i \in [m]} \mathbb{Y}_{Fe_i}$, then $\tilde{\mathrm{P}}$ is the I-projection of* P *in $L_{\text{sensi}}(\theta_F, \tilde{\theta}_V)$.*

Proof. This result follows from Theorem 4 by noticing that (co)varied parameters conditionally on different values of Y_{Cl} never appear in the same monomial, thus giving a simple sensitivity analysis. For each instantiation of Y_{Cl}, the feature variables are independent, thus giving a complete sensitivity analysis. Since for these two analyses the $\tilde{\theta}_V$-proportional covariation scheme is optimal the result then follows. □

Thus in a naive Bayes classifier for any choice of conditional probabilities from the feature variables to be varied, proportional covariation is optimal. This result can be extended straightforwardly to SPODE classifiers by excluding the super-parent node from the feature variables set. Then for any variation of probabilities of the other features, proportional covariation is optimal. For generic BN classifiers the optimality of proportional covariation holds for the cases formalized in Theorem 3 and Theorem 4.

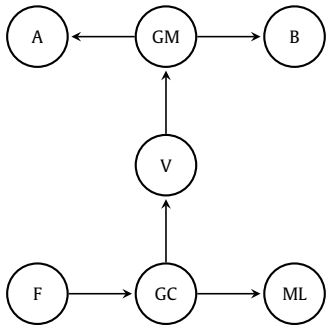


Fig. 6. Learnt BN for the cachexia dataset using hill-climbing.

7. A real-world application

In this section we study a subset of the dataset of Eisner et al. [23] including metabolomic information of 77 individuals: 47 of them suffering of cachexia, whilst the remaining do not. Cachexia is a metabolic syndrome characterized by loss of muscle with or without loss of fat mass. Although the study of Eisner et al. [23] included 71 different metabolics which could possibly distinguish individuals who suffer of Cachexia from those who do not, for our illustrative purposes we focus on only six of them: Adipate (A), Betaine (B), Fumarate (F), Glucose (GC), Glutamine (GM) and Valine (V). These metabolics are measured in a continuous scale and have been recently investigated in the context of Gaussian BNs [25]. The variables are discretized into three levels using the equal frequency method [see e.g. 48]. A BN over these variables together with the variable muscle loss (ML, taking values "yes" for individuals who suffered of cachexia and "no" otherwise) is learnt using the Bayesian hill-climbing method implemented in the `bnlearn` R package [51] and reported in Fig. 6. Since there are six ternary and one binary variables, the model has 1458 atomic probabilities. Furthermore, because of the network structure there are 54 parameters, and consequently the BN in Fig. 6 could be represented by a MM with associated matrix A of dimension 1458×54.

Eisner et al. [23] reports that the variables Adipate (A) and muscle loss (ML) have the largest mutual information. The estimated conditional probability of ML given A and computed using the `gRain` package [29] is

ML | A	low	average	high
yes	0.49	0.64	0.70
no	0.51	0.36	0.30

As the level of Adipate increases, the probability of having the disease increases and the probability of having the disease given a high level of Adipate is estimated as 0.70. Since Adipate is marginally a powerful predictor for muscle loss, due to the largest value of mutual information, it is of particular interest to estimate correctly the probability of ML = yes given A = high and to investigate the effect of misspecifications of this probability. Using the `sensquery` function of `bnmonitor` R package [40] the required individual parameter changes to obtain a conditional probability of either 0.68 or 0.72 are computed together with the associated CD distances and I-divergences. These are reported in Table 4. The software reports only one parameter change per CPT and uses proportional covariation, since the scheme is optimal for one-way analyses. There are multiple parameter changes that independently meet the constraint of the probability of ML = yes given A = high being equal to either 0.72 or 0.68. Although the interval of new probabilities of interest ([0.68,0.72]) is narrow and chosen so to best illustrate our methodology, the required changes in the CPTs defining the BN are on average equal to 0.25 and all require a change larger or equal to 0.1, with the exception of F = low for an increase to 0.72. The CD distances are comparable for both changes of ± 0.02, but the I-divergence is overall slightly larger when the conditional probability is decreased.

For ease of exposition we next consider only 2-way sensitivity analyses considering the parameters that `bnmonitor` suggested changing in one-way sensitivity analysis. Table 5 reports the results for every pair of parameters varied and using proportional covariation. The following conclusions can be made:

- the required changes to the parameters are smaller in absolute value compared to one-way analyses;
- the CD distance and the I-divergence takes a value in between those of the one-way analysis of the corresponding parameters;
- the vast majority of pairs of varied parameters correspond to the novel sensitivity analyses introduced in Section 5. For all these choices of parameters, Theorem 4 guarantees that proportional covariation minimizes the I-divergence between the original and the varied BNs;

Table 4
Required parameter changes in one-way sensitivity analysis to entertain constraints on the conditional probability of ML = yes given A = high. Columns: Parameter - varied probability; Old Value - original probability value; New Value - required probability change; Abs. Diff. - absolute value of the difference between Old Value and New Value; CD Dist. - CD distance between the original and the varied BN; I-Diver. - I-divergence between the original and the varied BN.

P(ML = yes | A = high) = 0.72					
Parameter	Old Value	New Value	Abs. Diff.	CD Dist.	I-Diver.
P(F = low)	0.34	0.25	0.09	0.42	0.0018
P(ML = yes | GC = low)	0.39	0.51	0.12	0.50	0.011
P(GC = high | F = low)	0.04	0.14	0.10	1.31	0.025
P(V = low | GC = average)	0.08	0.35	0.27	1.79	0.091
P(GM = low | V = low)	0.80	0.97	0.17	2.08	0.043
P(A = low | GM = average)	0.19	0.67	0.48	2.13	0.181

P(ML = ys | A = high) = 0.68					
Parameter	Old Value	New Value	Abs. Diff.	CD Dist.	I-Diverg.
P(F = low)	0.34	0.44	0.10	0.43	0.022
P(ML = yes | GC = low)	0.39	0.23	0.16	0.74	0.019
P(GM = low | V = low)	0.80	0.55	0.25	1.19	0.055
P(V = low | GC = low)	0.88	0.58	0.30	1.65	0.094
P(A = low | GM = low)	0.79	0.34	0.45	2.01	0.156
P(GC = average | F = low)	0.19	0.72	0.53	2.37	0.219

Table 5
All possible 2-way sensitivity analyses for the varied parameters in Table 4. Columns: Parameter1 - first varied probability; Parameter2 - second varied probability; Type - type of sensitivity analysis (either simple, complete, ordered or none); Avg. Change - mean of the absolute difference between the old and the new probabilities; CD Dist. - CD distance between the original and the varied BN; I-Diver. - I-divergence between the original and the varied BN.

P(ML = cachexic | A = high) = 0.72					
Parameter1	Parameter2	Type	Avg. Change	CD Dist.	I-Diver.
P(F = low)	P(ML = yes | GC = low)	complete	0.06	0.50	0.008
P(F = low)	P(GC = high | F = low)	ordered	0.05	1.03	0.015
P(F = low)	P(V = low | GC = average)	complete	0.06	0.78	0.017
P(F = low)	P(GM = low | V = low)	complete	0.05	0.48	0.018
P(F = low)	P(A = low | GM = average)	complete	0.19	1.71	0.102
P(ML = yes | GC = low)	P(GC = high | F = low)	**none**	0.06	0.47	0.009
P(ML = yes | GC = low)	P(V = low | GC = average)	simple	0.09	1.07	0.028
P(ML = yes | GC = low)	P(GM = low | V = low)	complete	0.08	1.50	0.027
P(ML = yes | GC = low)	P(A = low | GM = average)	complete	0.10	0.95	0.016
P(GC = high | F = low)	P(V = low | GC = average)	**none**	0.12	1.57	0.064
P(GC = high | F = low)	P(GM = low | V = low)	complete	0.08	1.73	0.024
P(GC = high | F = low)	P(A = low | GM = average)	complete	0.12	1.92	0.042
P(V = low | GC = average)	P(GM = low | V = low)	ordered	0.13	1.70	0.079
P(V = low | GC = average)	P(A = low | GM = average)	complete	0.19	2.33	0.079
P(GM = low | V = low)	P(A = low | GM = average)	**none**	0.15	1.93	0.047

P(ML = cachexic | A = high) = 0.68					
Parameter1	Parameter2	Type	Avg. Change	CD Dist.	I-Diver.
P(F = low)	P(ML = yes | GC = low)	complete	0.07	0.69	0.015
P(F = low)	P(GM = low | V = low)	complete	0.10	1.05	0.035
P(F = low)	P(V = low | GC = low)	complete	0.06	0.63	0.020
P(F = low)	P(A = low | GM = low)	complete	0.25	0.64	0.020
P(F = low)	P(GC = average | F = low)	ordered	0.23	2.00	0.166
P(ML = yes | GC = low)	P(GM = low | V = low)	complete	0.08	0.83	0.012
P(ML = yes | GC = low)	P(V = low | GC = low)	complete	0.09	1.08	0.017
P(ML = yes | GC = low)	P(A = low | GM = low)	complete	0.19	1.70	0.100
P(ML = yes | GC = low)	P(GC = average | F = low)	**none**	0.23	1.93	0.124
P(GM = low | V = low)	P(V = low | GC = low)	ordered	0.15	1.15	0.044
P(GM = low | V = low)	P(A = low | GM = low)	ordered	0.16	1.15	0.047
P(GM = low | V = low)	P(GC = average | F = low)	complete	0.21	2.05	0.074
P(V = low | GC = low)	P(A = low | GM = low)	complete	0.19	2.12	0.070
P(V = low | GC = low)	P(GC = average | F = low)	**none**	0.21	1.64	0.081
P(A = low | GM = low)	P(GC = average | F = low)	complete	0.27	2.55	0.157

- five out of the 30 pairs of varied parameters do not correspond to any of the novel sensitivity methods of Section 5. For such pairs, it can be shown that indeed proportional covariation does not minimize neither the CD distance nor the I-divergence.

8. Discussion

The representation of a wide array of statistical models in terms of the defining atomic monomial probabilities has proven useful for a variety of applications, including sensitivity analysis. In this paper, we took advantage of this representation to develop a formal geometric approach for sensitivity analysis which uses elements of information geometry. This approach has enabled us to demonstrate the optimality of proportional covariation for novel multi-way choices of varied parameters defined by the characteristics of the monomial atomic probabilities. Attention was devoted to BN classifiers where the tuning of the feature probabilities is often critical to ensure the classifier works reliably.

Although in this work we focused on models having multilinear atomic probabilities, our geometric approach could be used to investigate more general classes of models, for instance dynamic BNs whose atomic probabilities are not necessarily multilinear. Preliminary results suggest that the I-divergence exhibit different properties than in the multilinear case, with the potential of even more informative sensitivity investigations.

We concentrated on I-divergences but other measures of closeness between distributions could have been considered, for instance the already mentioned ϕ-divergences and CD distances. It is yet unknown whether our newly introduced covariation schemes would be optimal under these other measures. The example in Section 6.2 shows that in general proportional covariation does not minimize the CD distance in multi-way analyses. Furthermore, the choice of varied parameters not minimizing the CD distance is the same as for the I-divergence and therefore it may be expected that our newly introduced sensitivity analyses are such that proportional covariation is also optimal when considering the minimization of the CD distance as optimality criterion.

The `bnmonitor` R package was used in a real-world application to understand the relationship between various metabolics and muscle loss. However, its current implementation only provides a user-friendly implementation of one-way sensitivity methods. In a future version of `bnmonitor` we plan to also implement multi-way methods to provide a more comprehensive toolbox for sensitivity analysis in BNs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proofs

A.1. Proof of Theorem 3

Substituting $\bar{\theta}_V = \tilde{\theta}_V$ and $\mathrm{Q} \in \mathrm{Slice}(\theta_F)$, we can write $\mathcal{D}(\mathrm{Q}||\mathrm{P})$ as

$$\mathcal{D}(\mathrm{Q}||\mathrm{P}) = \sum_{y\in\mathbb{Y}} \theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}} \ln \frac{\theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}}}{\theta_F^{A_{y,F}} \theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}} = \sum_{y\in\mathbb{Y}} \theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}} \ln \frac{\tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}}}{\theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}}. \tag{A.1}$$

For all $\emptyset \neq H \subset [k]$, define $\mathbb{Y}_H^{=} = \{y \in \mathbb{Y} : A_{y,i} = 0, \text{ for all } i \in H\}$. Now (A.1) can be split as

$$\mathcal{D}(\mathrm{Q}||\mathrm{P}) = \sum_{y\in\mathbb{Y}\setminus\mathbb{Y}_C^{=}} \theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}} \ln \frac{\tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}}}{\theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}} + \sum_{y\in\mathbb{Y}_C^{=}} \theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}} \ln \frac{\tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}}}{\theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}}, \tag{A.2}$$

but since for all $y \in \mathbb{Y}_C^{=}$ and $i \in C$ $A_{y,i} = 0$, the second term on the rhs of (A.2) is equal to zero. The set $\mathbb{Y} \setminus \mathbb{Y}_C^{=}$ includes all events y for which $A_{y,i} = 1$ for at least one $i \in C$. Thus $\mathbb{Y} \setminus \mathbb{Y}_C^{=} = \bigcup_{H\subseteq C, H\neq\emptyset} \mathbb{Y}_H$, recalling that $\mathbb{Y}_H$ is the set of events y for which $A_{y,i} = 1$ for $i \in H$ and $A_{y,i} = 0$ for $i \in C \setminus H$. Furthermore since these sets $\mathbb{Y}_H$, for $H \subseteq C$, are disjoint we have that

$$\mathcal{D}(\mathrm{Q}||\mathrm{P}) = \sum_{H\subseteq C, H\neq\emptyset} \sum_{y\in\mathbb{Y}_H} \theta_F^{A_{y,F}} \tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}} \ln \frac{\tilde{\theta}_V^{A_{y,V}} \bar{\theta}_{C\setminus V}^{A_{y,C\setminus V}}}{\theta_V^{A_{y,V}} \theta_{C\setminus V}^{A_{y,C\setminus V}}}, \tag{A.3}$$

where terms in the internal sum are only for $\mathbb{Y}_H \neq 0$. For any $H \subseteq C$, $\mathrm{P} \in \mathrm{MM}(A, S)$ and $y \in \mathbb{Y}_H$, by multilinearity it holds

$$\theta_V^{A_{y,V}} = \prod_{i\in V\cap H} \theta_i = \theta_{V\cap H}, \qquad \theta_{C\setminus V}^{A_{y,C\setminus V}} = \prod_{i\in\{C\setminus V\}\cap H} \theta_i = \theta_{\{C\setminus V\}\cap H}. \tag{A.4}$$

Substituting (A.4) and using properties of the logarithm, (A.3) simplifies to

$$\mathcal{D}(\mathrm{Q}||\mathrm{P}) = \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\bar{\theta}_{\{C\setminus V\}\cap H}}{\theta_{\{C\setminus V\}\cap H}}. \tag{A.5}$$

Analogously

$$\mathcal{D}(\mathrm{Q}||\tilde{\mathrm{P}}) = \sum_{H\subseteq C, H\neq\emptyset} \sum_{y\in\mathbb{Y}_H} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\bar{\theta}_{\{C\setminus V\}\cap H}}{\tilde{\theta}_{\{C\setminus V\}\cap H}}, \tag{A.6}$$

$$\mathcal{D}(\tilde{\mathrm{P}}||\mathrm{P}) = \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{\{C\setminus V\}\cap H}}{\theta_{\{C\setminus V\}\cap H}}. \tag{A.7}$$

In (A.6) we used the assumption that $\bar{\theta}_V = \tilde{\theta}_V$ and in (A.6) and (A.7) we used Theorem 2. Next we use the fact that $\tilde{\theta}_{\{C\setminus V\}\cap H}$ is computed via proportional covariation. For $H \subseteq C$ it holds that

$$\begin{aligned}\tilde{\theta}_{\{C\setminus V\}\cap H} &= \prod_{\substack{i\in[n]: \\ S_i\cap H\neq\emptyset}} \prod_{j\in\{S_i\setminus V\}\cap H} \frac{1-|\tilde{\theta}_{V_i}|}{1-|\theta_{V_i}|} \theta_{\{S_i\setminus V\}\cap H} = \left(\prod_{\substack{i\in[n]: \\ S_i\cap H\neq\emptyset}} \prod_{j\in\{S_i\setminus V\}\cap H} \frac{1-|\tilde{\theta}_{V_i}|}{1-|\theta_{V_i}|} \right) \theta_{\{C\setminus V\}\cap H} \\ &= \alpha\theta_{\{C\setminus V\}\cap H} \end{aligned} \tag{A.8}$$

Substituting (A.8) into the logarithms in (A.6) and (A.7) and re-arranging the factors yields

$$\mathcal{D}(\mathrm{Q}||\tilde{\mathrm{P}}) = \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\bar{\theta}_{\{C\setminus V\}\cap H}}{\theta_{\{C\setminus V\}\cap H}} - \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln\alpha \tag{A.9}$$

$$\mathcal{D}(\tilde{\mathrm{P}}||\mathrm{P}) = \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln\alpha \tag{A.10}$$

At this stage the result is proven if under the condition in (8) the rhs of (A.5) is equal to the sum of the rhs of (A.9) and (A.10). Since the second term on the rhs of (A.5) is equal to the first term on the rhs of (A.9), we can write $\mathcal{D}(\mathrm{Q}\,||\,\mathrm{P}) = \mathcal{D}(\mathrm{Q}||\tilde{\mathrm{P}}) + \mathcal{D}(\tilde{\mathrm{P}}||\,\mathrm{P})$ as

$$\begin{aligned}&\sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \bar{\theta}_{\{C\setminus V\}\cap H} \ln\alpha \\ &\qquad - \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} - \sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \tilde{\theta}_{\{C\setminus V\}\cap H} \ln\alpha = 0. \end{aligned} \tag{A.11}$$

By rearranging the terms in (A.11) we have that

$$\sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} \left(\bar{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \bar{\theta}_{\{C\setminus V\}\cap H} \ln\alpha + \tilde{\theta}_{\{C\setminus V\}\cap H} \ln \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} + \tilde{\theta}_{\{C\setminus V\}\cap H} \ln\alpha \right) = 0, \tag{A.12}$$

which yields

$$\sum_{\substack{H\subseteq C, H\neq\emptyset \\ y\in\mathbb{Y}_H}} \theta_F^{A_{y,F}} \tilde{\theta}_{V\cap H} (\bar{\theta}_{\{C\setminus V\}\cap H} - \tilde{\theta}_{\{C\setminus V\}\cap H}) \ln \left(\alpha \frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}} \right) = 0. \tag{A.13}$$

Noticing that (A.13) equals (8), since only $\theta_F^{A_{y,F}}$ depends on the event $y \in \mathbb{Y}_H$, proves the result.

A.2. Proof of Theorem 4

The result is proven if the condition in (8) holds for all $\mathrm{Q} \in L$. For a simple analysis, for all $i, j \in C$, the monomial $\theta_i\theta_j$ does not divide θ^{A_y} for any $y \in \mathbb{Y}$. Thus all sets H in condition (8) that need to be considered, i.e. those such that $\mathbb{Y}_H$ is non-empty, have one element only because of regularity. If H is an element of V then $\bar{\theta}_{\{C\setminus V\}\cap H} - \tilde{\theta}_{\{C\setminus V\}\cap H} = 0$ by construction and the result thus follows. Conversely, if H is an element of $C \setminus V$, condition (8) holds if and only if

$$\sum_{j\in C\setminus V} \bar{\theta}_j - \tilde{\theta}_j = 0. \tag{A.14}$$

Next, (A.14) can be rewritten as

$$\sum_{i\in[n]:V_i\neq\emptyset} 1 - |\tilde{\theta}_{V_i}| - 1 + |\tilde{\theta}_{V_i}| = 0,$$

which is always true. This proves Theorem 4 for simple analyses.

In a complete sensitivity analysis all sets H in condition (8) that need to be considered, i.e. those such that $\mathbb{Y}_H$ is non-empty, are in $\times_{i\in[n]:V_i\neq\emptyset} S_i$, by regularity. Thus, (8) can be written as

$$\sum_{\substack{H\in\times_{i\in[n]:V_i\neq\emptyset}S_i,\\ H\neq\emptyset}} \tilde{\theta}_{V\cap H}(\bar{\theta}_{\{C\setminus V\}\cap H} - \tilde{\theta}_{\{C\setminus V\}\cap H}) \ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right) \sum_{y\in\mathbb{Y}_H} \theta_F^{A_{y,F}} = 0. \tag{A.15}$$

Suppose with no loss of generality that the sets S_i such that $V_i\neq\emptyset$ are those with index in the set $[r]$, $r\leq n$. Notice that

$$\bigtimes_{i\in[r]} S_i = \bigcup_{R\subseteq[r]} \left\{ \bigtimes_{i\in R} V_i \bigtimes \bigtimes_{i\in[r]\setminus R} \{C_i\setminus V_i\} \right\}. \tag{A.16}$$

Thus the result is proven if (A.15) holds for each $R\subseteq[r]$, i.e. if

$$\sum_{H\in R}\sum_{J\in[r]\setminus R} \tilde{\theta}_{V\cap H}(\bar{\theta}_{\{C\setminus V\}\cap J} - \tilde{\theta}_{\{C\setminus V\}\cap J}) \ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right) \sum_{y\in\mathbb{Y}_{\{H\cup J\}}} \theta_F^{A_{y,F}} = 0, \tag{A.17}$$

for H and J such that $\mathbb{Y}_{\{H\cup J\}}\neq\emptyset$. First notice that if $R=[r]$, then $\bar{\theta}_{\{C\setminus V\}\cap J} - \tilde{\theta}_{\{C\setminus V\}\cap J} = 0$ by construction and the result follows. Now fix an $R\subset[r]$ and suppose $k\in[r]\setminus R$. So, (A.17) can be written as

$$\sum_{H\in R}\sum_{J\in[r]\setminus R\setminus\{k\}}\sum_{j\in C_k\setminus V_k} \tilde{\theta}_{V\cap H}(\bar{\theta}_{\{C\setminus V\}\cap J}\bar{\theta}_j - \tilde{\theta}_{\{C\setminus V\}\cap J}\tilde{\theta}_j) \ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right) \sum_{y\in\mathbb{Y}_{\{H\cup J\cup\{j\}\}}} \theta_F^{A_{y,F}} = 0. \tag{A.18}$$

Noticing that $\sum_{j\in C_k\setminus V_k}\bar{\theta}_j = \sum_{j\in C_k\setminus V_k}\tilde{\theta}_j = 1-|\tilde{\theta}_{V_k}|$, (A.18) can be rearranged as

$$\sum_{H\in R}\sum_{J\in[r]\setminus R\setminus\{k\}} \tilde{\theta}_{V\cap H}\left(\bar{\theta}_{\{C\setminus V\}\cap J}(1-|\tilde{\theta}_{V_k}|) - \tilde{\theta}_{\{C\setminus V\}\cap J}(1-|\tilde{\theta}_{V_k}|)\right) \ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right) \sum_{y\in\mathbb{Y}_{\{H\cup J\cup\{j\}\}}} \theta_F^{A_{y,F}} = 0. \tag{A.19}$$

By applying the same steps as in (A.18)-(A.19) for all $k\in[r]\setminus R$, we have that

$$\sum_{H\in R} \tilde{\theta}_{V\cap H}\left(\prod_{k\in[r]\setminus R}(1-|\tilde{\theta}_{V_k}|) - \prod_{k\in[r]\setminus R}(1-|\tilde{\theta}_{V_k}|)\right) \ln\left(\alpha\frac{\tilde{\theta}_{V\cap H}}{\theta_{V\cap H}}\right) \sum_{y\in\mathbb{Y}_{\{H\cup J\cup\{j\}\}}} \theta_F^{A_{y,F}} = 0, \tag{A.20}$$

which always holds, thus proving Theorem 4 for complete analyses.

The proof for ordered analyses follows from the one of complete sensitivity analyses by noticing that the sets $R\subseteq[r]$ for which condition (A.17) needs to hold is a subset of those already demonstrated in the complete case.

Appendix B. Staged trees as monomial models

This appendix illustrates how staged trees can be cast as monomial models. To this end, we consider another plausible statistical model for the public health application in Example 4 is supported on the tree in Fig. B.7 as follows. The edges emanating from vertex v_0 denote the different levels of sports activity, whilst the edges emanating from v_1, v_2 and v_3 denote the levels of alcohol consumption conditional on the level of activity. Edges emanating from $v_4,\ldots,v_{12}$ are associated to the population's health conditional on both preceding variables. A (conditional/transition) probability is associated to each edge and probabilities from edges emanating from the same non-leaf vertex must sum to one. The atomic probabilities are then simply given by the product of the edge probabilities along a root-to-leaf path.

Staged trees [9,52] are a particular class of trees where conditional probability distributions emanating from different vertices are identified. This is denoted by framing vertices whose distributions are identified by the same shape. In Fig. B.7 the transition probabilities from vertex v_7 to v_{22} and from v_8 to v_{25} are equal because v_7 and v_8 are in the same stage, i.e. vertices whose distributions are identified. Setting transition probabilities equal can be thought of as representing

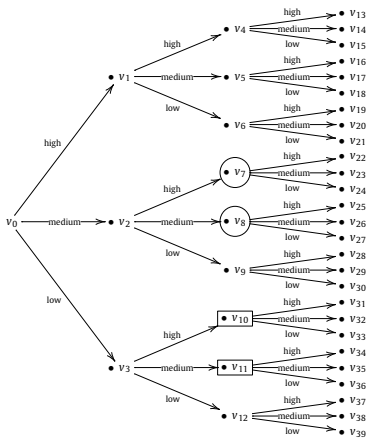


Fig. B.7. Staged tree representation of the application in Example 4. Vertices in the same stage are framed in the same shape.

Table B.6
Monomial atomic probabilities of the staged tree in Fig. B.7.

$\theta_{v_0v_1}\theta_{v_1v_4}\theta_{v_4v_{13}}$	$\theta_{v_0v_1}\theta_{v_1v_4}\theta_{v_4v_{14}}$	$\theta_{v_0v_1}\theta_{v_1v_4}\theta_{v_4v_{15}}$
$\theta_{v_0v_1}\theta_{v_1v_5}\theta_{v_5v_{16}}$	$\theta_{v_0v_1}\theta_{v_1v_5}\theta_{v_5v_{17}}$	$\theta_{v_0v_1}\theta_{v_1v_5}\theta_{v_5v_{18}}$
$\theta_{v_0v_1}\theta_{v_1v_6}\theta_{v_6v_{19}}$	$\theta_{v_0v_1}\theta_{v_1v_6}\theta_{v_6v_{20}}$	$\theta_{v_0v_1}\theta_{v_1v_6}\theta_{v_6v_{21}}$
$\theta_{v_0v_2}\theta_{v_2v_7}\theta_{v_7v_{22}}$	$\theta_{v_0v_2}\theta_{v_2v_7}\theta_{v_7v_{23}}$	$\theta_{v_0v_2}\theta_{v_2v_7}\theta_{v_7v_{24}}$
$\theta_{v_0v_2}\theta_{v_2v_8}\theta_{v_8v_{25}}$	$\theta_{v_0v_2}\theta_{v_2v_8}\theta_{v_8v_{26}}$	$\theta_{v_0v_2}\theta_{v_2v_8}\theta_{v_8v_{27}}$
$\theta_{v_0v_2}\theta_{v_2v_9}\theta_{v_9v_{28}}$	$\theta_{v_0v_2}\theta_{v_2v_9}\theta_{v_9v_{29}}$	$\theta_{v_0v_2}\theta_{v_2v_9}\theta_{v_9v_{30}}$
$\theta_{v_0v_3}\theta_{v_3v_{10}}\theta_{v_{10}v_{31}}$	$\theta_{v_0v_3}\theta_{v_3v_{10}}\theta_{v_{10}v_{32}}$	$\theta_{v_0v_3}\theta_{v_3v_{10}}\theta_{v_{10}v_{33}}$
$\theta_{v_0v_3}\theta_{v_3v_{11}}\theta_{v_{11}v_{25}}$	$\theta_{v_0v_3}\theta_{v_3v_{11}}\theta_{v_{11}v_{26}}$	$\theta_{v_0v_3}\theta_{v_3v_{11}}\theta_{v_{11}v_{27}}$
$\theta_{v_0v_3}\theta_{v_3v_{12}}\theta_{v_{12}v_{28}}$	$\theta_{v_0v_3}\theta_{v_3v_{12}}\theta_{v_{12}v_{29}}$	$\theta_{v_0v_3}\theta_{v_3v_{12}}\theta_{v_{12}v_{30}}$

context-specific conditional independence information. Staged trees are capable of representing all conditional independence hypotheses within discrete BNs [52]. At the same time they are a larger class of statistical models, as illustrated next.

In Example 4 suppose the following equalities are believed to hold

$$\mathrm{P}(Y_3 = y_3 | Y_2 = 3, Y_1 = 2) = \mathrm{P}(Y_3 = y_3 | Y_2 = 2, Y_1 = 2), \tag{B.1}$$

$$\mathrm{P}(Y_3 = y_3 | Y_2 = 3, Y_1 = 1) = \mathrm{P}(Y_3 = y_3 | Y_2 = 2, Y_1 = 1), \tag{B.2}$$

for all $y_3 \in [3]$. For instance, equation (B.2) states that the probability distribution of health for individuals with high alcohol consumption and low physical activity is equal to that of individuals with medium alcohol consumption and low physical activity. Such context-specific independence constraints cannot be explicitly represented in a BN model. Conversely they have a straightforward representation in the staged tree reported in Fig. B.7 [which is stratified according to the definition of 16].

The atomic probabilities of this staged tree are multilinear monomials in the parameters associated to edges, where pairs of parameters from two vertices in the same stage are identified. Letting $\theta_{v_iv_j}$ denote the transition probability from v_i to v_j, the staged tree in Fig. B.7 has the following probabilities identified:

$$\begin{aligned} &\theta_{v_7v_{22}} = \theta_{v_8v_{25}}, && \theta_{v_7v_{23}} = \theta_{v_8v_{26}}, && \theta_{v_7v_{24}} = \theta_{v_8v_{27}}, \\ &\theta_{v_{10}v_{31}} = \theta_{v_{11}v_{34}}, && \theta_{v_{10}v_{32}} = \theta_{v_{11}v_{35}}, && \theta_{v_{10}v_{33}} = \theta_{v_{11}v_{36}}. \end{aligned} \tag{B.3}$$

The parameter equalities in the first row of equation (B.3) derive from equation (B.1). Similarly, the bottom row of (B.3) is associated to (B.2). Given these constraints, the atomic probabilities of the staged tree in Fig. B.7 are those reported in Table B.6. For a formal derivation see [26]. The A matrix has dimensions 27×33, is very sparse (again in each row there are 3 ones and zeros otherwise).

In general staged trees may not be multilinear MMs since two vertices in the same stage can possibly be along a same root-to-leaf path. However, staged trees that are multilinear are also regular since no parameters associated to emanating edges from vertices in the same stage appear in one atomic probability monomial.

References

[1] S.B. Amsalu, A. Homaifar, A. Esterline, A simplified matrix formulation for sensitivity analysis of hidden Markov models, Algorithms 10 (3) (2017) 97.
[2] C. Bielza, P. Larrañaga, Discrete Bayesian network classifiers: a survey, ACM Comput. Surv. 47 (1) (2014) 5.
[3] C. Bielza, P. Larrañaga, Bayesian networks in neuroscience: a survey, Front. Comput. Neurosci. 8 (2014) 131.

[4] J.H. Bolt, S. Renooij, Local sensitivity of Bayesian networks to multiple simultaneous parameter shifts, in: European Workshop on Probabilistic Graphical Models, Springer, 2014, pp. 65–80.
[5] J.H. Bolt, L.C. van der Gaag, Balanced sensitivity functions for tuning multi-dimensional Bayesian network classifiers, Int. J. Approx. Reason. 80 (2017) 361–376.
[6] E. Borgonovo, E. Plischke, Sensitivity analysis: a review of recent advances, Eur. J. Oper. Res. 248 (3) (2016) 869–887.
[7] C. Boutilier, N. Friedman, M. Goldszmidt, D. Koller, Context-specific independence in Bayesian networks, in: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, 1996, pp. 115–123.
[8] B. Cai, X. Kong, Y. Liu, J. Lin, X. Yuan, H. Xu, R. Ji, Application of Bayesian networks in reliability evaluation, IEEE Trans. Ind. Inform. 15 (4) (2018) 2146–2157.
[9] Federico Carli, Manuele Leonelli, Eva Riccomagno, Gherardo Varando, The R package stagedtrees for structural learning of stratified staged trees, J. Stat. Softw. 102 (6) (2022) 1–30.
[10] E. Castillo, J.M. Gutiérrez, A.S. Hadi, Sensitivity analysis in discrete Bayesian networks, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 27 (4) (1997) 412–423.
[11] H. Chan, A. Darwiche, Sensitivity analysis in Bayesian networks: from single to multiple parameters, in: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 317–325.
[12] H. Chan, A. Darwiche, A distance measure for bounding probabilistic belief change, Int. J. Approx. Reason. 38 (2005) 149–174.
[13] T. Charitos, L.C. Van Der Gaag, Sensitivity analysis of Markovian models, in: FLAIRS Conference, 2006, pp. 806–811.
[14] S.H. Chen, C.A. Pollino, Good practice in Bayesian network modelling, Environ. Model. Softw. 37 (2012) 134–145.
[15] V.M.H. Coupé, L.C. Van Der Gaag, Properties of sensitivity analysis of Bayesian belief networks, Ann. Math. Artif. Intell. 36 (4) (2002) 323–356.
[16] R.G. Cowell, J.Q. Smith, Causal discovery through MAP selection of stratified chain event graphs, Electron. J. Stat. 8 (1) (2014) 965–997.
[17] David Cox, John Little, Donal OShea, Ideals, Varieties, and Algorithms: an Introduction to Computational Algebraic Geometry and Commutative Algebra, Springer Science & Business Media, 2013.
[18] I. Csiszár, P.C. Shields, Information theory and statistics: a tutorial, Found. Trends Commun. Inf. Theory 1 (2004) 417–528.
[19] A. Darwiche, A differential approach to inference in Bayesian networks, J. ACM 50 (3) (2003) 280–305.
[20] A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.
[21] J. De Bock, C.P. De Campos, A. Antonucci, Global sensitivity analysis for MAP inference in graphical models, in: Advances in Neural Information Processing Systems, 2014, pp. 2690–2698.
[22] B. Drury, J. Valverde-Rebaza, M.F. Moura, A. de Andrade Lopes, A survey of the applications of Bayesian networks in agriculture, Eng. Appl. Artif. Intell. 65 (2017) 29–42.
[23] R. Eisner, C. Stretch, T. Eastman, J. Xia, D. Hau, S. Damaraju, R. Greiner, D.S. Wishart, V.E. Baracos, Learning to predict cancer-associated skeletal muscle wasting from 1h-nmr profiles of urinary metabolites, Metabolomics 7 (1) (2011) 25–34.
[24] F. Ferretti, A. Saltelli, S. Tarantola, Trends in sensitivity analysis practice in the last decade, Sci. Total Environ. 568 (2016) 666–670.
[25] C. Görgen, M. Leonelli, Model-preserving sensitivity analysis for families of Gaussian distributions, J. Mach. Learn. Res. 21 (84) (2020) 1–32.
[26] C. Görgen, M. Leonelli, J.Q. Smith, A differential approach for staged trees, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, 2015, pp. 346–355.
[27] P. Gustafson, L. Wasserman, Local sensitivity diagnostics for Bayesian inference, Ann. Stat. 23 (1995) 2153–2167.
[28] M. Hänninen, P. Kujala, Influences of variables on ship collision probability in a Bayesian belief network model, Reliab. Eng. Syst. Saf. 102 (2012) 27–40.
[29] S. Højsgaard, Graphical independence networks with the gRain package for R, J. Stat. Softw. 46 (10) (2012).
[30] B. Kamiński, M. Jakubczyk, P. Szufel, A framework for sensitivity analysis of decision trees, Cent. Eur. J. Oper. Res. 26 (1) (2018) 135–159.
[31] E.J. Keogh, M.J. Pazzani, Learning the structure of augmented Bayesian classifiers, Int. J. Artif. Intell. Tools 11 (4) (2002) 587–601.
[32] U. Kjærulff, L.C. van der Gaag, Making sensitivity analysis computationally efficient, in: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, 2000, pp. 317–325.
[33] J. Kleemann, E. Celio, C. Fürst, Validation approaches of an expert-based Bayesian belief network in northern Ghana, West Africa, Ecol. Model. 365 (2017) 10–29.
[34] A. Klimova, T. Rudas, Testing the fit of relational models, Commun. Stat., Theory Methods (2021) 1–20.
[35] A. Klimova, T. Rudas, A. Dobra, Relational models for contingency tables, J. Multivar. Anal. 104 (1) (2012) 159–173.
[36] D. Koller, N. Friedman, F. Bach, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
[37] K.B. Laskey, Sensitivity analysis for probability assessments in Bayesian networks, IEEE Trans. Syst. Man Cybern. 25 (6) (1995) 901–909.
[38] M. Leonelli, Sensitivity analysis beyond linearity, Int. J. Approx. Reason. 113 (2019) 106–118.
[39] M. Leonelli, C. Görgen, J.Q. Smith, Sensitivity analysis in multilinear probabilistic models, Inf. Sci. 411 (2017) 84–97.
[40] M. Leonelli, R. Ramanathan, R.L. Wilkerson, Sensitivity and robustness analysis in Bayesian networks with the bnmonitor R package, arXiv:2107.11785, 2021.
[41] T. Makaba, W. Doorsamy, B.S. Paul, Bayesian network-based framework for cost-implication assessment of road traffic collisions, Int. J. Intell. Transp. Syst. Res. (2020) 1–14.
[42] S. McLachlan, K. Dube, G.A. Hitman, N. Fenton, E. Kyrimi, Bayesian networks in healthcare: distribution by medical condition, Artif. Intell. Med. (2020) 101912.
[43] Giovanni Pistone, Eva Riccomagno, Henry P. Wynn, Algebraic Statistics: Computational Commutative Algebra in Statistics, Chapman and Hall/CRC, 2000.
[44] J. Pitchforth, K. Mengersen, A proposed validation framework for expert elicited Bayesian networks, Expert Syst. Appl. 40 (2013) 162–167.
[45] S. Renooij, Efficient sensitivity analysis in hidden Markov models, Int. J. Approx. Reason. 53 (9) (2012) 1397–1414.
[46] S. Renooij, Co-variation for sensitivity analysis in Bayesian networks: properties, consequences and alternatives, Int. J. Approx. Reason. 55 (2014) 1022–1042.
[47] J. Rohmer, Uncertainties in conditional probability tables of discrete Bayesian belief networks: a comprehensive review, Eng. Appl. Artif. Intell. 88 (2020) 103384.
[48] R.F. Ropero, S. Renooij, L.C. Van der Gaag, Discretizing environmental data for learning Bayesian-network classifiers, Ecol. Model. 368 (2018) 391–403.
[49] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2016.
[50] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, Wiley Online Library, 2004.
[51] M. Scutari, Learning Bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (3) (2010) 1–22.
[52] J.Q. Smith, P.E. Anderson, Conditional independence and graphs, Artif. Intell. 172 (2008) 42–68.
[53] Seth Sullivant, Algebraic Statistics, vol. 194, American Mathematical Soc., 2018.
[54] L.C. Van der Gaag, S. Renooij, V.M.H. Coupé, Sensitivity analysis of probabilistic networks, in: Advances in Probabilistic Graphical Models, Springer, 2007, pp. 103–124.
[55] G. Varando, C. Bielza, P. Larrañaga, Decision boundary for discrete Bayesian network classifiers, J. Mach. Learn. Res. 16 (2015) 2725–2749.
[56] G. Varando, C. Bielza, P. Larrañaga, Decision functions for chain classifiers based on Bayesian networks for multi-label classification, Int. J. Approx. Reason. 68 (2016) 164–178.
[57] C. Yakaboski, E. Santos, Bayesian knowledge base distance-based tuning, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 64–72.