Sensitivity analysis beyond linearity

Manuele Leonelli

School of Mathematics and Statistics, University of Glasgow, UK.

Abstract

A wide array of graphical models can be parametrised to have atomic probabilities represented by monomial functions. Such a monomial structure has proven very useful when studying robustness under the assumption of a multilinear model where all monomials have either zero or one exponents. Robustness in probabilistic graphical models is usually investigated by varying some of the input probabilities and observing the effects of these on output probabilities of interest. Here the assumption of multilinearity is relaxed and a general approach for one-way sensitivity analysis in non-multilinear models is presented. It is shown that in non-multilinear models sensitivity functions have a polynomial form, conversely to multilinear models where these are simply linear. The form of various divergences and distances under different covariation schemes is also formally derived. Proportional covariation is proven to be optimal in non-multilinear models under some specific choices of varied parameters. The methodology is illustrated throughout by an educational application.

Keywords: Covariation, Monomial models, Probabilistic graphical models, Sensitivity analysis, Staged trees

1. Introduction

Sensitivity methods have received great attention in the literature of probabilistic graphical models in the past twenty years. Sensitivity analysis is a fundamental part of any applied analysis, carried out to validate the construction of a probabilistic graphical model and investigate its robustness to misspecification of its probabilities. Such methods have been successfully used in a variety of applications (e.g. Nur et al., 2009; Oberguggenberger et al., 2009; Pollino et al., 2007; Uusitalo, 2007).

Research has mostly focused on Bayesian network (BN) models (Koller et al., 2009; Smith, 2010), although sensitivity results also exist for Markov networks (Chan and Darwiche, 2005b) and chain event graphs (Leonelli et al., 2017a). Sensitivity analysis in BNs usually consists of two phases: first, some parameters of the model are varied and the effect of these variations on output probabilities of interest are investigated; second, once parameter variations are identified, the effect of these are summarized by a distance or divergence measure between the original and the varied distributions underlying the BN. Although sensitivity methods exist for continuous random variables under the assumption of Gaussianity (e.g Castillo and Kjærulff, 2003; Gómez-Villegas et al., 2013; Görgen and Leonelli, 2018), henceforth we focus on the most common case of discrete random variables only.

For the first phase of a sensitivity analysis, a simple mathematical function, usually termed *sensitivity function*, describes an output probability of interest as a function of the BN parameters. This is a (multi-) linear function of the varied parameters for marginal output probabilities (Castillo et al., 1997; Coupé and Van Der Gaag, 2002). Conversely, if the probability of interest is a conditional probability, then the sensitivity function is a ratio of (multi-) linear functions.

For the second phase, the Chan-Darwiche distance (Chan and Darwiche, 2005a), Kullback-Leibler divergence (Kullback and Leibler, 1951) and ϕ-divergences (Ali and Silvey, 1966) are often used to measure the overall effect of parameter variations. One important line of research has focused on identifying parameter *covariations*, i.e. ways to adjust parameters so to respect the sum to one condition after a parameter variation, that minimize such distances. Proportional covariation (Laskey, 1995; Renooij, 2014), which assigns the same proportion of residual probability mass to covarying paramaters after a variation, is the gold-standard method since this has been shown to minimize the above-mentioned divergences in a variety of settings (Chan and Darwiche, 2002; Leonelli et al., 2017a), although not all (Leonelli and Riccomagno, 2018).

Most of the above-mentioned results, although specifically derived for BNs, hold for a variety of models whose atomic probabilities can be written as a multilinear polynomial (Leonelli et al., 2017a). The multilinear structure of atomic probabilities in BNs has been known for quite some time (Castillo et al., 1995; Darwiche, 2003), but other models entertain the same property under specific parametrisations, for instance stratified staged trees (Görgen et al., 2015), context-specific BNs (Boutilier et al., 1996) and influence diagrams (Leonelli et al., 2017b).

The development of sensitivity methods for models whose atomic probabilities cannot be written as multilinear polynomials have been limited. Results have been derived for dynamic Bayesian networks (DBNs) (Charitos and van der Gaag, 2006a,b), Markov chains (de Cooman et al., 2008) and hidden Markov models (Amsalu et al., 2017; Renooij, 2012). The atomic probabilities of all these model classes have a non-square-free polynomial representation, as demonstrated in Brandherm and Jameson (2004) since they all have a DBN characterisation. Non-multilinear atomic probabilities are often associated to models whose probabilities are recursively updated through time in a dynamic fashion, although this does not necessarily have to be the case as demonstrated by the examples below.

This work presents a general framework for one-way sensitivity analysis in models whose atomic probabilities have a non-multilinear structure and therefore can be applied to the already mentioned model classes of DBNs and hidden Markov models. The monomial representation of a statistical model introduced in Leonelli and Riccomagno (2018) is used here to encompass all classes of discrete models with non-multilinear atomic probabilities. For such models, the form of the sensitivity functions and their properties are derived. Furthermore, results about the computation of the CD distance and ϕ-divergences under various covariation schemes are derived. In particular, it is proven that, for specific choices of parameters to be varied, proportional covariation is optimal, in the sense that it minimizes the CD distance between the original and varied distributions amongst all possible ways to covary parameters. Therefore, this work extends the results of Leonelli et al. (2017a) for multilinear models to non-multilinear ones, as well as proposing sensitivity methods similar to those of Renooij (2012) and Charitos and van der Gaag (2006a) but which apply to a much more general class of models.

The paper is structured as follows. Section 2 reviews monomial models and shows that staged trees have in general a non-multilinear polynomial representation. This section further introduces a running example from an educational application. Section 3 reviews covariation methods for probabilities. Section 4 reports the derivations of the sensitivity functions for non-multilinear models, whilst Section 5 deals with divergences and their computation. The paper is concluded with a discussion.

2. Monomial models

A review of monomial models, in short MMs, as introduced in Leonelli and Riccomagno (2018) is given first. Let $\mathbb{Y}$ be a finite set with q elements and P a strictly positive probability density function for $\mathbb{Y}$. Let $\#\mathbb{Y} = q$, call $y \in \mathbb{Y}$ an atom and $\mathrm{P}(y)$ the atomic probability of y. The generic probability P can be seen as a point in the interior set of the q-dimensional simplex, i.e. $\mathrm{P} \in \Delta_{q-1}$. Next, a particular class of parametric statistical models, called MMs, is associated to $\mathbb{Y}$.

Let $[k] = \{1, 2, \ldots, k\}$. A MM is defined by three elements: a $q \times k$ matrix A with non-negative integer entries, $A \in \mathcal{M}_{q\times k}(\mathbb{Z}_{\geq 0})$; a k-dimensional parameter vector θ with positive real entries, $\theta = (\theta_i)_{i\in[k]} \in \mathbb{R}^k_{>0}$; and a partition $S = \{S_1, \ldots, S_n\}$ of $[k]$. 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 $\mathrm{P}(y) = \prod_{i\in[k]} \theta_i^{A_{y,i}} = \theta^{A_y}$. The partition S of $[k]$ is such that $\theta_{S_i} = (\theta_j)_{j\in S_i} \in \Delta_{\#S_i-1}$. The atomic probability of $y \in \mathbb{Y}$ can then be written as

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

where $\theta_S^{A_{y,S}} = \prod_{i\in S} \theta_i^{A_{y,i}}$ denotes the monomial associated to an event $y \in \mathbb{Y}$ where only parameters θ_i for $i \in S$ can have non-zero exponent. For $A \in \mathcal{M}_{q\times k}(\mathbb{Z}_{\geq 0})$, $B \subseteq [q]$ and $C \subseteq [k]$, $A_{B,C}$ denotes the submatrix of A with B rows and C columns.

Definition 1. The MM over $\mathbb{Y}$ associated to A, θ and S, where S is such that $\theta_{S_i} \in \Delta_{\#S_i-1}$, is defined as

$$\mathrm{MM}(A, \theta, S) = \left\{ \mathrm{P} \in \Delta_{q-1} : \mathrm{P}(y) = \prod_{i\in[n]} \theta_{S_i}^{A_{y,S_i}} \text{ for } y \in \mathbb{Y} \text{ and } \theta \in \mathbb{R}^k_{>0} \right\}$$

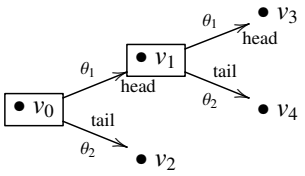


Figure 1: The staged tree of a repeated coin toss from Example 1.

A MM(A, θ, S) is said to be multilinear if $A \in \mathcal{M}_{q\times k}(\{0, 1\})$.

A MM is multilinear if all its monomials are square-free, i.e. the exponents of the parameters are either zero or one. Leonelli et al. (2017a) and Leonelli and Riccomagno (2018) give a thorough investigation of sensitivity analysis in multilinear MMs. Here conversely the focus is on models which are not necessarily multilinear.

Example 1. Consider a simple coin toss game. The probability of head (H) is θ_1, whilst tail (T) has probability θ_2, where $\theta_1 + \theta_2 = 1$. If the result of the first toss is head, then the coin is tossed a second time. This situation can be represented by a MM with parameter vector $\theta = (\theta_1, \theta_2)$, degenerate partition of [2] including one element only, and matrix A defined as

$$A = \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

where the first column of A relates to θ_1 and the second to θ_2. The model is such that $\mathrm{P}(HH) = \theta_1^2$, $\mathrm{P}(HT) = \theta_1\theta_2$ and $\mathrm{P}(T) = \theta_2$. This MM is non-multilinear since the matrix A includes an entry equal to 2.

Since DBNs have already been shown to have a non-multilinear monomial structure in Brandherm and Jameson (2004), here the focus is on staged trees, which are introduced next.

2.1. Staged trees

Graphical models represented by *event trees* $\mathcal{T} = (V, E)$ are considered here, which are directed rooted trees where each inner vertex $v \in V$ has at least two children. In this context, the sample space of the model corresponds to the set of root-to-leaf paths in the graph and each directed path, which is a sequence of edges $r = (e \mid e \in E(r))$, for $E(r) \subset E$ has a meaning in the modelling context. Each edge $e \in E$ is associated to a primitive probability $\theta_e \in (0, 1)$ such that on each *floret* $\mathcal{F}(v) = (v, E(v))$, where $E(v) \subseteq E$ is the set of edges emanating from $v \in V$, the primitive probabilities sum to unity. The probability of an atom is then simply the product of the primitive probabilities along the edges of its path: $\mathrm{P}(r) = \prod_{e\in E(r)} \theta_e$.

Definition 2. Let $\theta_v = (\theta_e \mid e \in E(v))$ be the vector of primitive probabilities associated to the floret $\mathcal{F}(v)$, $v \in V$, in an event tree $\mathcal{T} = (V, E)$. A *staged tree* is an event tree as above where, for some $v, w \in V$, the floret probabilities are identified $\theta_v = \theta_w$. Then, $v, w \in V$ are in the same *stage*.

Two vertices are thus in the same stage if they have the same (conditional) distribution over their edges. When drawing a tree, vertices in the same stage are either framed using the same shape or equally colored in order to have a visual counterpart of that information. Setting floret probabilities equal can be thought of as representing conditional independence information. Staged trees are capable of representing all conditional independence hypotheses within discrete BNs, whilst at the same time being more flexible in expressing modifications of these (Collazo et al., 2018; Smith and Anderson, 2008).

Staged trees are MMs whose atomic probabilities can either be multilinear or not (Görgen et al., 2015). The following example gives a simple illustration of a non-multilinear staged tree.

Example 2. The MM of Example 1 can be depicted as the staged tree in Figure 1, which has two inner-vertices, v_0 and v_1, in the same stage. The tree has three root-to-leaf paths ending in the leaves v_3 (head and head), v_4 (head and tail) and v_2 (tail). The edges emanating from the inner-vertices v_0 and v_1 are associated to the primitive probabilities θ_1 and θ_2 representing the probability of head and tail respectively.

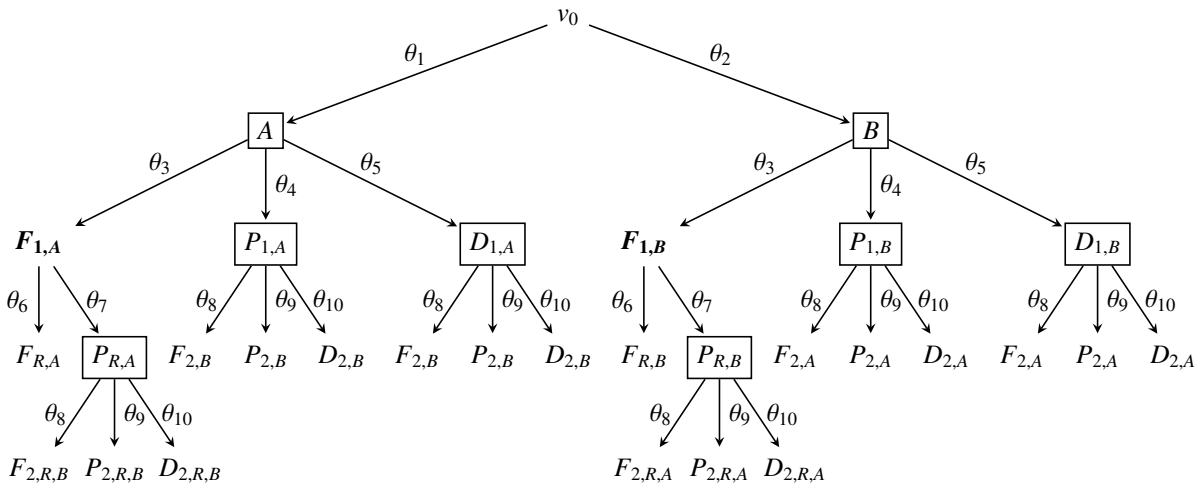


Figure 2: The staged tree of the educational application of Section 2.2 under the first set of hypotheses.

2.2. An example

To illustrate the construction of a staged tree and its monomial representation, an example from an educational application is considered. This example was first introduced in Freeman and Smith (2011).

In a one-year program students take components A and B, but not everyone in the same order: students are first allocated to study either module A or B for the first six months and then the other for the final six months. After the first six months students are examined on their allocated component and can be awarded a distinction (D), a pass (P) or a fail (F). If failed, they can resit the exam with the possibility of passing and thus be allowed to the second component. Students who fail the resit are withdrawn from the program. For the second module students can again either fail, pass or be awarded a distinction, but with no possibility of resitting. With an obvious extension of the labeling, the process can be depicted by the tree in Figure 2

Various hypotheses of conditional independence, corresponding to equal primitive probabilities of multiple florets, can be embedded in the above educational scenario. One set of such hypotheses was given in Freeman and Smith (2011) as:

- The components A and B are equally hard: this is depicted by framing the vertices A and B by a square in Figure 2.
- The chances of passing the first module after a fail do not depend on the module taken: this is depicted by the bold font of $F_{1,A}$ and $F_{1,B}$ in Figure 2.
- The distribution of grades for the last six months does not depend on the module taken nor on the results of the first part: this is depicted by framing $P_{R,A}$, $P_{1,A}$, $D_{1,A}$, $P_{R,B}$, $P_{1,B}$ and $D_{1,B}$ by a rectangle in Figure 2.

These hypotheses give the staged tree of Figure 2, which can be equally represented by a MM with parameter vector $(\theta_1, \dots, \theta_{10})$, matrix $A = (A_{11}, A_{12})^{\mathrm{T}}$, with

$$A_{11} = \begin{pmatrix} 1&0&1&0&0&1&0&0&0&0\\ 1&0&1&0&0&0&1&1&0&0\\ 1&0&1&0&0&0&1&0&1&0\\ 1&0&1&0&0&0&1&0&0&1\\ 1&0&0&1&0&0&0&1&0&0\\ 1&0&0&1&0&0&0&0&1&0\\ 1&0&0&1&0&0&0&0&0&1\\ 1&0&0&0&1&0&0&1&0&0\\ 1&0&0&0&1&0&0&0&1&0\\ 1&0&0&0&1&0&0&0&0&1 \end{pmatrix} \qquad A_{12} = \begin{pmatrix} 0&1&1&0&0&1&0&0&0&0\\ 0&1&1&0&0&0&1&1&0&0\\ 0&1&1&0&0&0&1&0&1&0\\ 0&1&1&0&0&0&1&0&0&1\\ 0&1&0&1&0&0&0&1&0&0\\ 0&1&0&1&0&0&0&0&1&0\\ 0&1&0&1&0&0&0&0&0&1\\ 0&1&0&0&1&0&0&1&0&0\\ 0&1&0&0&1&0&0&0&1&0\\ 0&1&0&0&1&0&0&0&0&1 \end{pmatrix}$$

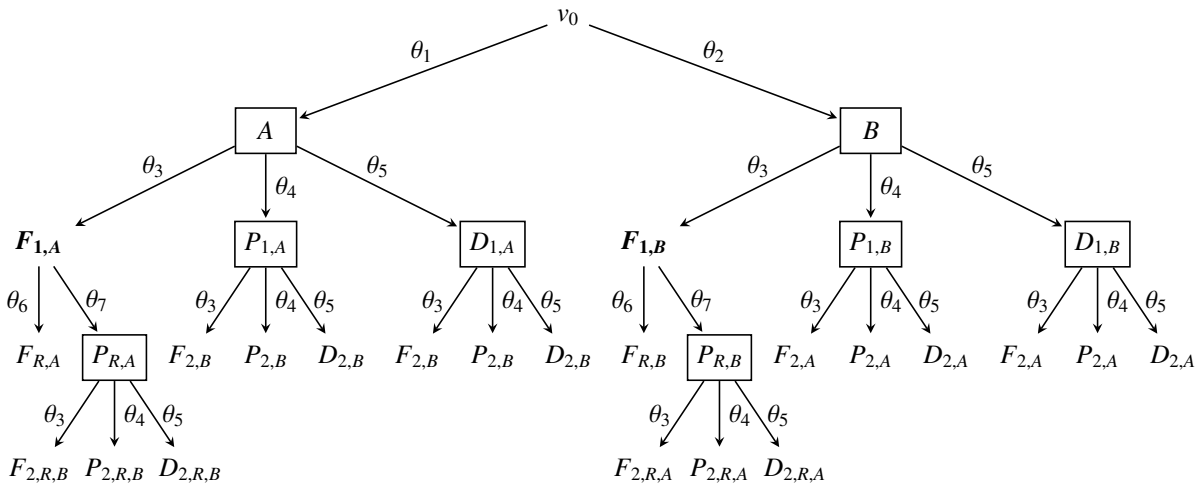


Figure 3: The staged tree of the educational application of Section 2.2 under the second set of hypotheses.

and partition $S = \{S_1, S_2, S_3, S_4\}$ where $S_1 = \{1, 2\}$, $S_2 = \{3, 4, 5\}$, $S_3 = \{6, 7\}$ and $S_4 = \{8, 9, 10\}$. This model is multilinear since all entries of A are either zero or one. Graphically this could have also been deduced by noticing that no vertices along a root-to-leaf path are in the same stage.

A second set of hypotheses may embellish the first one by assuming that the distribution of grades of students not experiencing fails are the same in all components. This additional hypothesis gives the staged tree in Figure 3 where vertices A, B, $P_{R,A}$, $P_{1,A}$, $D_{1,A}$, $P_{R,B}$, $P_{1,B}$ and $D_{1,B}$ are now all in the same stage. This staged tree can be written as a MM with parameter $(\theta_1, \ldots, \theta_7)$, matrix $A = (A_{21}, A_{22})^{\mathrm{T}}$ with

$$A_{21} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 & 0 \end{pmatrix} \qquad A_{22} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 & 0 \end{pmatrix}$$

and partition $S = \{S_1, S_2, S_3\}$ where $S_1 = \{1, 2\}$, $S_2 = \{3, 4, 5\}$ and $S_3 = \{6, 7\}$. Under this additional hypothesis the staged tree does not entertain a multilinear monomial parametrization, but only a non-multilinear one. For such models there is currently no established sensitivity theory to investigate their robustness.

3. Covariation

The basic underlying idea of sensitivity analysis is to vary some of the model's parameters and observe how such variations affect outputs of interest. However, when variations are performed, then some of 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 multiple ways to covary parameters.

The theory of covariation from Renooij (2014) is reviewed next, with a particular focus on its specific characterization for MMs given in Leonelli and Riccomagno (2018). For a set S and $i \in S$, let S^{-i} denote $S \setminus \{i\}$, $-S_j$ denote the set $[k] \setminus S_j$ and let $|v|$ denote the sum of the elements of a vector v.

Definition 3. Let θ be the parameter vector of a MM and θ_i be the parameter varied where $i \in S_j$. Let θ be partitioned as $\theta = (\theta_i, \theta_{S_j^{-i}}, \theta_{-S_j})$ and let $\tilde{\theta}_i \in (0,1)$. A $\tilde{\theta}_i$-*covariation scheme* is a function $\sigma : \bigtimes_{l\in[n]} \Delta_{\#S_l-1} \to \bigtimes_{l\in[n]} \Delta_{\#S_l-1}$ which fixes θ_i to $\tilde{\theta}_i$ and does not change θ_{-S_j}, i.e.

$$\begin{aligned}\sigma : \quad \bigtimes_{l\in[n]} \Delta_{\#S_l-1} &\longmapsto \bigtimes_{l\in[n]} \Delta_{\#S_l-1}\\ (\theta_i, \theta_{S_j^{-i}}, \theta_{-S_j}) &\longmapsto (\tilde{\theta}_i, \cdot, \theta_{-S_j}).\end{aligned}$$

Thus θ_{S_j} denotes a vector of parameters that need to respect the sum to one condition, $\tilde{\theta}_i$ denotes the new numerical specification of the parameter varied and θ_{-S_j} the parameter vector which is not affected by the variation. Consider as an example a staged tree model. In a staged tree the sets S_l, $l \in [n]$, denote the conditional probability distributions of florets in different stages. Suppose one parameter from one stage is varied. Then the parameters associated to that same stage are covaried, whilst all others are held fixed.

Definition 4. In the notation of Definition 3

- the $\tilde{\theta}_i$-*proportional* covariation scheme $\sigma_{\text{pro}}(\theta) = (\tilde{\theta}_i, \tilde{\theta}_{S_j^{-i}}, \theta_{-S_j})$ is defined by setting

$$\tilde{\theta}_l = \frac{1-\tilde{\theta}_i}{1-\theta_i}\theta_l \qquad \text{for all } l \in S_j^{-i}.$$

- The $\tilde{\theta}_i$-*uniform* covariation scheme, $\sigma_{\text{uni}}(\theta) = (\tilde{\theta}_i, \tilde{\theta}_{S_j^{-i}}, \theta_{-S_j})$ is defined by setting

$$\tilde{\theta}_l = \frac{1-\tilde{\theta}_i}{\#S_j - 1} \qquad \text{for all } l \in S_j^{-i}.$$

- The $\tilde{\theta}_i$-*linear* covariation scheme $\sigma_{\text{lin}}(\theta) = (\tilde{\theta}_i, \tilde{\theta}_{S_j^{-i}}, \theta_{-S_j})$ is defined by setting

$$\tilde{\theta}_l = \gamma_l\tilde{\theta}_i + \delta_l \qquad \text{for all } l \in S_j^{-i},$$

  where γ_l and δ_l need to be chosen so that $\tilde{\theta}_i + |\tilde{\theta}_{S_j^{-i}}| = 1$

Different covariation schemes may entertain different properties which, depending on the domain of application, might be more or less desirable (see Leonelli et al., 2017a; Renooij, 2014, for a list). Applying a linear covariation scheme is very natural: if for instance $\delta_l = -\gamma_l$, then $\tilde{\theta}_l = \delta_l(1-\tilde{\theta}_i)$ and the scheme assigns a proportion δ_l of the remaining probability mass to $\tilde{\theta}_l$. Notice that uniform and proportional schemes are specific instances of linear covariations. Another used covariation scheme is the order-preserving one (see Renooij, 2014, for details).

4. Sensitivity functions

Sensitivity functions represent the functional relationship between a parameter being varied and the output probability of an event of interest. These are often used in practice since, for instance, the parameter specifications of a MM may imply event probabilities which appear to be unreasonable to a user, although being a coherent consequence of his/her beliefs. Sensitivity functions depict the required change of a parameter that would give a reasonable event probability.

Consider a $MM(A, \theta, S)$ and an event $E \subset \mathbb{Y}$ of interest. Definition 5 gives the probability of an event E as a function of a covariation scheme.

Definition 5. Let σ be a $\tilde{\theta}_i$-covariation scheme. For $\mathrm{P} \in MM(A, \theta, S)$, the probability $\sigma(\mathrm{P})(E)$ read as a function of $\tilde{\theta}_i$ is called the sensitivity function associated to σ.

The following theorem derives the general form of sensitivity functions in non-multilinear MMs as well as their form for specific covariation schemes.

Theorem 1. *Let* $\mathrm{P} \in MM(A, \theta, S)$*,* $E \subset \mathbb{Y}$ *and suppose the parameter* θ_i *is varied, where* $i \in S_j$*. Then*

- *for a generic* θ_i*-covariation scheme* σ

$$\sigma(\mathrm{P})(E) = \sum_{y \in E} \tilde{\theta}_{S_j}^{A_{y,S_j}} \theta_{-S_j}^{A_{y,-S_j}} \tag{1}$$

- *for proportional covariation* σ_{pro}

$$\sigma_{\text{pro}}(\mathrm{P})(E) = \sum_{y \in E} \tilde{\theta}_i^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|A_{y,S_j^{-i}}|} \theta_{S_j^{-i}}^{A_{y,S_j^{-i}}} \theta_{-S_j}^{A_{y,-S_j}} \tag{2}$$

- *for uniform covariation* σ_{uni}

$$\sigma_{\text{uni}}(\mathrm{P})(E) = \sum_{y \in E} \tilde{\theta}_i^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{\#S_j - 1} \right)^{|A_{y,S_j^{-i}}|} \theta_{-S_j}^{A_{y,-S_j}} \tag{3}$$

- *for linear covariation* σ_{lin}

$$\sigma_{\text{lin}}(\mathrm{P})(E) = \sum_{y \in E} \tilde{\theta}_i^{A_{y,i}} \prod_{k \in S_j^{-i}} (\gamma_k \tilde{\theta}_i + \delta_k)^{A_{y,k}} \theta_{-S_j}^{A_{y,-S_j}} \tag{4}$$

Proof. For equation (1) notice that

$$\sigma(\mathrm{P})(E) = \sum_{y \in E} \tilde{\theta}^{A_y} = \sum_{y \in E} \tilde{\theta}_{S_j}^{A_{y,S_j}} \tilde{\theta}_{-S_j}^{A_{y,-S_j}} = \sum_{y \in E} \tilde{\theta}_{S_j}^{A_{y,S_j}} \theta_{-S_j}^{A_{y,-S_j}}.$$

The form of the sensitivity function under different covariation schemes follows from equation (1) by plugging-in their definition given in Definition 4. □

From Theorem 1 is then easy to deduce the polynomial properties of the sensitivity function in general MMs.

Corollary 1. *For proportional, uniform and linear* $\tilde{\theta}_i$*-covariation schemes, the sensitivity function* $\sigma(\mathrm{P})(E)$ *is a polynomial in* $\tilde{\theta}_i$ *of degree* $\max_{y \in E} |A_{y,S_j}|$*.*

This follows from the form of the sensitivity functions given in equation (2)-(4).

Notice that differently to multilinear MMs, where the sensitivity function is linear for any linear covariation scheme, the sensitivity function is more generally polynomial in non-multilinear MMs. However, there are cases where sensitivity functions are simply linear, as formalized by the following corollary.

Corollary 2. *In the notation of Theorem 1, if* $0 \leq |A_{y,S_j}| \leq 1$ *for all* $y \in E$*, then* $\sigma(\mathrm{P})(E)$ *is a linear function of* $\tilde{\theta}_i$ *for any linear* $\tilde{\theta}_i$*-covariation scheme.*

This follows from Corollary 1 since if $0 \leq |A_{y,S_j}| \leq 1$ then the sensitivity function is a polynomial of degree 1.

The previous results formalize the form of sensitivity functions for marginal probabilities. Conditional sensitivity functions represent the functional relationship between conditional probabilities and a parameter varied.

Corollary 3. *The conditional sensitivity function* $\sigma(P)(E \mid C)$ *is the ratio of sensitivity functions* $\sigma(P)(E \cap C)/\sigma(P)(C)$*, where each of these have the properties formalized in Theorem 1, Corollary 1 and Corollary 2.*

This result easily follows from the definition of conditional probability.

Example 3. To illustrate the different form of sensitivity functions in multilinear and non-multilinear models, consider the staged trees from the educational example of Section 2.2. The two staged tree structures are embellished by the probability specifications given in Table 1. For ease of comparison the probability distributions from the stages $\{v_0\}$ and $\{F_{1,A}, F_{1,B}\}$ are equally defined in the two trees. The distribution of the stage $\{A, B, P_{1,A}, D_{1,A}, P_{1,B}, D_{1,B}, P_{R,A}, P_{R,B}\}$ in the non-multilinear staged tree of Figure 3 is such that the parameters θ_3, θ_4 and θ_5 are chosen from the probabilities

Multilinear staged tree
$\theta_1 = 0.5, \theta_2 = 0.5, \theta_3 = 0.2, \theta_4 = 0.7, \theta_5 = 0.1, \theta_6 = 0.35, \theta_7 = 0.65, \theta_8 = 0.1, \theta_9 = 0.5, \theta_{10} = 0.4$
Non-multilinear staged tree
$\theta_1 = 0.5, \theta_2 = 0.5, \theta_3 = 0.15, \theta_4 = 0.6, \theta_5 = 0.25, \theta_6 = 0.35, \theta_7 = 0.65$

Table 1: Probability specifications for the staged trees in Section 2.2.

underlying the tree in Figure 2 as $(\theta_3 + \theta_8)/2$, $(\theta_4 + \theta_9)/2$ and $(\theta_5 + \theta_{10})/2$, respectively. Suppose the parameter θ_4 is varied in both cases: notice that for the first tree this is the probability of passing the exam in the second semester, whilst for the three in Figure 3 this is the probability of passing an exam at any point.

The probabilities of four events are considered here. First, the sensitivity function for a θ_4 variation of not being admitted to the second semester is for both trees $\theta_1\theta_3\tilde{\theta}_6 + \theta_2\theta_3\tilde{\theta}_6$, where $\tilde{\theta}_6$ depends on the covariation scheme used. Thus in both models this function is simply linear whenever the covariation scheme is linear, even though the second tree is a non-multilinear model. These sensitivity functions are reported in Figure 4a. Under uniform covariation, the sensitivity function is the same for the two trees, whilst under proportional covariation they differ.

The second event considered is passing both exams with distinction. For the multilinear tree the associated sensitivity function can be written as $(\theta_1 + \theta_2)\tilde{\theta}_5\theta_{10}$, whilst for the non-multilinear tree this is $(\theta_1 + \theta_2)\tilde{\theta}_5^2$. Thus in this case the sensitivity function is a non-linear function of the varied parameter, as reported in Figure 4b, but for both trees the sensitivity function is decreasing.

For the event of failing the exam in the second semester the sensitivity functions for the two trees are highly different, as reported in Figure 4c. For the multilinear tree, the sensitivity function is slightly increasing and almost identical for uniform and proportional covariation. Conversely, for the non-multilinear tree this is decreasing non-linearly. Formally, for the multilinear tree the sensitivity function is $\theta_8(\theta_1 + \theta_2)(\tilde{\theta}_3\theta_7 + \tilde{\theta}_4 + \tilde{\theta}_5)$, whilst for the non-multilinear tree this is $(\theta_1 + \theta_2)(\tilde{\theta}_3^2\theta_7 + \tilde{\theta}_4\tilde{\theta}_3 + \tilde{\theta}_5\tilde{\theta}_3)$.

Lastly, the conditional probability of obtaining a distinction in the first semester given that a distinction was given in the second one is computed. In this case, the sensitivity function is a ratio of polynomials and as such is not linear even for multilinear models. This is shown in Figure 4d. As for the first event considered, the sensitivity functions under uniform covariation are equal for the two trees.

5. Divergence quantification

Once viable parameter variations have been identified via the study of sensitivity functions as illustrated in Section 4, the overall effect that these would have on the model's distribution is studied. This is carried out by computing various distances and divergences between the original and the varied distributions.

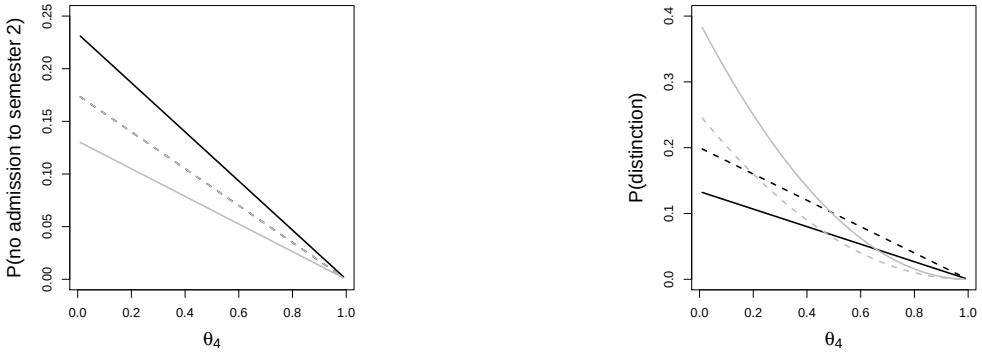
5.1. The CD distance in non-multilinear models

The measure of dissimilarity which is most commonly used in sensitivity analysis in graphical models is the so-called CD distance (Chan and Darwiche, 2005a).

Definition 6. The CD distance between two probability distributions $\tilde{\mathrm{P}}$ and P over a discrete sample space $\mathbb{Y}$ is

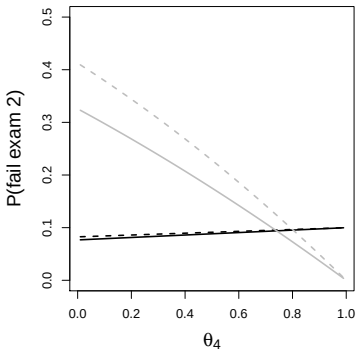
$$\mathcal{D}_{\mathrm{CD}}(\tilde{\mathrm{P}}, \mathrm{P}) = \log \max_{y \in \mathbb{Y}} \frac{\tilde{\mathrm{P}}(y)}{\mathrm{P}(y)} - \log \min_{y \in \mathbb{Y}} \frac{\tilde{\mathrm{P}}(y)}{\mathrm{P}(y)}.$$

For single and specific multi-way parameter variations, proportional covariation minimizes the CD distance in BN models, as well as in any multilinear MM (Chan and Darwiche, 2002; Leonelli et al., 2017a). However, in non-multilinear models even for single parameter variations proportional covariation does not minimize the CD distance in general as shown by the following example.

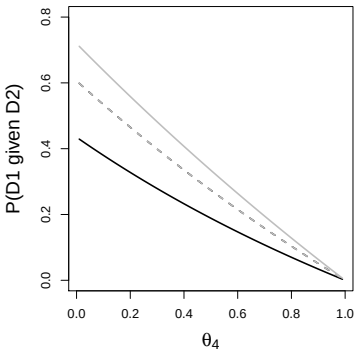


(a)

(b)



(c)



(d)

Figure 4: Sensitivity functions of four events for the staged trees of Section 2.2. Black lines: multilinear staged tree; Gray lines: non-multilinear staged tree; Full lines: proportional covariation; Dashed lines: uniform covariation.

Example 4. Consider two random variables Y_1 and Y_2 and suppose $\mathbb{Y}_1 = \mathbb{Y}_2 = [3]$. Suppose also

$$\theta_i = \mathrm{P}(Y_1 = i) = \mathrm{P}(Y_2 = i \,|\, Y_1 = j), \quad i \in [3], j \in [2].$$

and $\theta_{i+3} = \mathrm{P}(Y_2 = i \,|\, Y_1 = 3)$. The atomic probabilities of this model are clearly non-multilinear. Suppose θ_i is varied and θ_2 and θ_3 are covaried. Suppose $\theta_1 = 0.33$, $\theta_2 = 0.33$, $\theta_3 = 0.34$ and let θ_1 be varied to 0.4 (the value of θ_4, θ_5 and θ_6 does not affect the CD distance). In this situation the CD distance under a proportional scheme is larger than under a uniform scheme, which would therefore be preferred to the proportional one. Conversely, if θ_1 is set to 0.2 the distance is smaller under the proportional scheme than under the uniform one. Notice that if conversely the initial probabilities were $\theta_1 = 0.35$, $\theta_2 = 0.15$ and $\theta_3 = 0.5$, then proportional covariation would have smaller CD distance than uniform covariation for both variations of θ_1 to 0.2 and 0.4.

Therefore the optimality of proportional covariation may not only depend on the model, for instance multilinear or non-multilinear, but also on the numerical specification of its probabilities.

Next the form of the CD distance in MMs is derived in general and for specific covariation schemes. For all $\emptyset \neq H \subset [k]$ define $\mathbb{Y}_H^{=} = \{y \in \mathbb{Y} : A_{y,i} = 0 \text{ for all } i \in H\}$ and let $\mathbb{Y}_H^{\neq} = \mathbb{Y} \setminus \mathbb{Y}_H^{=}$. The set $\mathbb{Y}_H^{\neq}$ includes the events for which at least one parameter with index in H has a non-zero exponent.

Theorem 2. *Let* $\mathrm{P} \in MM(A, \theta, S)$ *and suppose the parameter* θ_i *is varied, where* $i \in S_j$*. Then*

- *for a generic* θ_i*-covariation scheme* σ

$$\mathcal{D}_{\mathrm{CD}}(\sigma(\mathrm{P}), \mathrm{P}) = \log \max_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}} - \log \min_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}} \tag{5}$$

- *for proportional covariation* σ_{pro}

$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{pro}}(\mathrm{P}), \mathrm{P}) = \log \max_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_i}{\theta_i} \right)^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|A_{y,S_j^{-i}}|} - \log \min_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_i}{\theta_i} \right)^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|A_{y,S_j^{-i}}|}$$

- *for uniform covariation* σ_{uni}

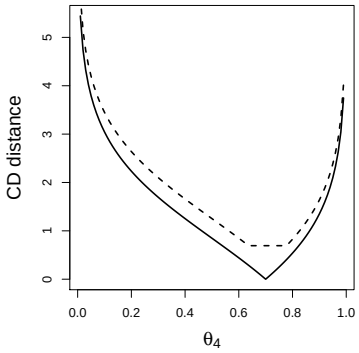
$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{uni}}(\mathrm{P}), \mathrm{P}) = \log \max_{y \in \mathbb{Y}_{S_j}^{\neq}} \frac{\theta_i^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{\#S_j - 1} \right)^{|A_{y,S_j^{-1}}|}}{\theta_{S_j}^{A_{y,S_j}}} - \log \min_{y \in \mathbb{Y}_{S_j}^{\neq}} \frac{\theta_i^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{\#S_j - 1} \right)^{|A_{y,S_j^{-1}}|}}{\theta_{S_j}^{A_{y,S_j}}}$$

- *for linear covariation* σ_{lin}

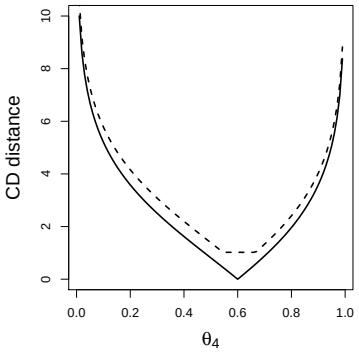
$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{lin}}(\mathrm{P}), \mathrm{P}) = \log \max_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_i}{\theta_i} \right)^{A_{y,i}} \prod_{k \in S_j} \left(\frac{\gamma_k \tilde{\theta}_i + \delta_k}{\theta_k} \right)^{A_{y,k}} - \log \min_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_i}{\theta_i} \right)^{A_{y,i}} \prod_{k \in S_j} \left(\frac{\gamma_k \tilde{\theta}_i + \delta_k}{\theta_k} \right)^{A_{y,k}}$$

Proof. For equation (1) notice that

$$\begin{aligned}
\mathcal{D}_{\mathrm{CD}}(\sigma(\mathrm{P}), \mathrm{P}) &= \log \max_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}^{A_y}}{\theta^{A_y}} \right) - \log \min_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}^{A_y}}{\theta^{A_y}} \right) \\
&= \log \max_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}_{S_j}^{A_{y,S_j}} \tilde{\theta}_{-S_j}^{A_{y,-S_j}}}{\theta_{S_j}^{A_{y,S_j}} \theta_{-S_j}^{A_{y,-S_j}}} \right) - \log \min_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}_{S_j}^{A_{y,S_j}} \tilde{\theta}_{-S_j}^{A_{y,-S_j}}}{\theta_{S_j}^{A_{y,S_j}} \theta_{-S_j}^{A_{y,-S_j}}} \right) \\
&= \log \max_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}} - \log \min_{y \in \mathbb{Y}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}} \\
&= \log \max_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}} - \log \min_{y \in \mathbb{Y}_{S_j}^{\neq}} \left(\frac{\tilde{\theta}_{S_j}}{\theta_{S_j}} \right)^{A_{y,S_j}},
\end{aligned}$$



(a) Multilinear tree.



(b) Non-multilinear tree.

Figure 5: CD distance for the staged trees of Section 2.2 for variations of θ_4. Full lines: proportional covariation; Dashed lines: uniform covariation.

where the last equality holds since, for all $y \in \mathbb{Y}_{S_j}^{=}$, $(\tilde{\theta}_{S_j}/\theta_{S_j})^{A_{y,S_j}} = 1$ and there are always both larger and smaller ratios between varied and original parameters.

The form of the CD distance under different covariation schemes follows from equation (5) by plugging-in their definition given in Definition 4. □

One of the reasons why the CD distance is commonly used for sensitivity analysis in BNs is that, for a single parameter variation, the distance between the BN distributions equals the distance between the single conditional probability distributions associated to the varied parameters (Chan and Darwiche, 2002). Theorem 2 demonstrates that this is true in general for non-multilinear models since the distance only depends on the parameter θ_{S_j}.

Example 5. As in Example 3, suppose the parameter θ_4 is varied in the two staged trees from the educational example of Section 2.2. From the results of Leonelli et al. (2017a), it can be deduced that for the multilinear tree, the CD distance between the original and varied distributions is simply

$$\log \max_{i=3,4,5} \frac{\tilde{\theta}_i}{\theta_i} - \log \min_{i=3,4,5} \frac{\tilde{\theta}_i}{\theta_i}. \tag{6}$$

Conversely, using Theorem 2, for the non-multilinear staged tree this equals

$$\log \max \left\{ \frac{\tilde{\theta}_3}{\theta_3}, \frac{\tilde{\theta}_3^2}{\theta_3^2}, \frac{\tilde{\theta}_4^2}{\theta_4^2}, \frac{\tilde{\theta}_5^2}{\theta_5^2}, \frac{\tilde{\theta}_3\tilde{\theta}_4}{\theta_3\theta_4}, \frac{\tilde{\theta}_3\tilde{\theta}_5}{\theta_3\theta_5}, \frac{\tilde{\theta}_4\tilde{\theta}_5}{\theta_4\theta_5} \right\} - \log \min \left\{ \frac{\tilde{\theta}_3}{\theta_3}, \frac{\tilde{\theta}_3^2}{\theta_3^2}, \frac{\tilde{\theta}_4^2}{\theta_4^2}, \frac{\tilde{\theta}_5^2}{\theta_5^2}, \frac{\tilde{\theta}_3\tilde{\theta}_4}{\theta_3\theta_4}, \frac{\tilde{\theta}_3\tilde{\theta}_5}{\theta_3\theta_5}, \frac{\tilde{\theta}_4\tilde{\theta}_5}{\theta_4\theta_5} \right\}. \tag{7}$$

The specific form of the CD distance for uniform covariation can be deduced from equation (7) by simply substituting $\tilde{\theta}_3$ and $\tilde{\theta}_5$ with $(1 - \tilde{\theta}_4)/2$. For proportional covariation the CD distance greatly simplifies and can be written as

$$\log \max \left\{ \frac{\tilde{\theta}_4^2}{\tilde{\theta}_4^2}, \frac{1-\tilde{\theta}_4}{1-\theta_4}, \frac{(1-\tilde{\theta}_4)^2}{(1-\theta_4)^2}, \frac{\tilde{\theta}_4(1-\tilde{\theta}_4)}{\theta_4(1-\theta_4)} \right\} - \log \min \left\{ \frac{\tilde{\theta}_4^2}{\tilde{\theta}_4^2}, \frac{1-\tilde{\theta}_4}{1-\theta_4}, \frac{(1-\tilde{\theta}_4)^2}{(1-\theta_4)^2}, \frac{\tilde{\theta}_4(1-\tilde{\theta}_4)}{\theta_4(1-\theta_4)} \right\},$$

which, as formalized by Theorem 2, only depends on the original and varied values of θ_4. The CD distances for proportional and uniform covariation and any possible varied value of θ_4 are reported in Figure 5. Notice that although for this application the CD distance for proportional covariation is always smaller than for uniform covariation, Example 4 above gives an illustration where this is not the case.

Theorem 2 and Example 5 show that for single parameter variations the CD distance in non-multilinear models does not simply correspond to the distance between distributions defined over one element of the partition S (as in equation (6) for the multilinear staged tree). However, there are parameter variations in non-multilinear models where this is the case as formalized by Corollary 4

Corollary 4. *In the notation of Theorem 2, suppose* $0 \leq |A_{y,S_j}| \leq 1$ *for all* $y \in \mathbb{Y}^{\neq}_{S_j}$. *Then*

- *for a generic* θ_i*-covariation scheme* σ

$$\mathcal{D}_{\mathrm{CD}}(\sigma(\mathrm{P}), \mathrm{P}) = \log \max_{i \in S_j} \frac{\tilde{\theta}_i}{\theta_i} - \log \min_{i \in S_j} \frac{\tilde{\theta}_i}{\theta_i} \tag{8}$$

- *for proportional covariation* σ_{pro}

$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{pro}}(\mathrm{P}), \mathrm{P}) = \left| \log \frac{\tilde{\theta}_i}{\theta_i} - \log \frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right|$$

- *for uniform covariation* σ_{uni}

$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{uni}}(\mathrm{P}), \mathrm{P}) = \log \max \left\{ \frac{\tilde{\theta}_i}{\theta_i}, \frac{1 - \tilde{\theta}_i}{(\#S_j - 1) \min_{k \in S_j^{-i}} \theta_k} \right\} - \log \min \left\{ \frac{\tilde{\theta}_i}{\theta_i}, \frac{1 - \tilde{\theta}_i}{(\#S_j - 1) \max_{k \in S_j^{-i}} \theta_k} \right\}$$

- *for linear covariation* σ_{lin}, *where* $\delta_k = -\gamma_k$ *for all* $k \in S_j^{-i}$,

$$\mathcal{D}_{\mathrm{CD}}(\sigma_{\mathrm{lin}}(\mathrm{P}), \mathrm{P}) = \log \max \left\{ \frac{\tilde{\theta}_i}{\theta_i}, \frac{1 - \tilde{\theta}_i}{\min_{k \in S_j^{-i}} \delta_k^{-1} \theta_k} \right\} - \log \min \left\{ \frac{\tilde{\theta}_i}{\theta_i}, \frac{1 - \tilde{\theta}_i}{\max_{k \in S_j^{-i}} \delta_k^{-1} \theta_k} \right\}$$

Proof. Equation (8) follows from equation (5) by imposing the condition $0 \leq |A_{y,S_j}| \leq 1$. Equation (8) then coincides to the CD distance between one conditional probability distribution in BNs and its varied version and the specific form of the distance under different covariation schemes can be derived as in Renooij (2014). □

Corollary 4 generalizes the results of Renooij (2014), which derive the specific form of the sensitivity function for various covariation schemes in BNs, to the case of non-multilinear models for specific choices of varied parameter. Importantly, the form of the CD distance derived in Corollary 4 has the very important consequence that for some varied parameters proportional variation can be shown to be optimal.

Theorem 3. *Under the conditions of Corollary 4, proportional covariation minimizes the CD distance between the original and varied distribution amongst all possible covariation schemes.*

Proof. The theorem follows from equation (8) which is the CD distance between one conditional probability distribution in BNs and its varied version. As proven in Chan and Darwiche (2002) this distance is minimized by proportional covariation. □

Theorem 3 therefore extends the results of Chan and Darwiche (2002) and Leonelli et al. (2017a) which prove the optimality of proportional covariation for BNs and multilinear MMs to specific one-way sensitivity analyses in non-multilinear models.

Example 6. For the non-multilinear staged tree in Figure 3, consider the stage $\{F_{1,A}, F_{1,B}\}$. Suppose there is an additional edge coming out of this stage ending in a leaf (for example by splitting the fail result, into badly failed and moderately fail). Then one could show that the columns associated to the parameters of the stage probability distribution in the A matrix have only zero or one entries. This can also be seen graphically since $F_{1,A}$ and $F_{1,B}$ are not along a same root-to-leaf path. Therefore, by Theorem 3, if one probability from this stage distribution is varied then by proportionally covarying the remaining parameters the CD distance between the original staged tree distribution and the new one is minimized.

5.2. ϕ-divergences in non-multilinear models

Another class of divergences which is often used in practice is the so-called ϕ-divergence (Ali and Silvey, 1966).

Definition 7. The ϕ-divergence from $\tilde{\mathrm{P}}$ to P over a discrete sample space $\mathbb{Y}$ is

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

where Φ is the class of convex functions $\phi(x)$, $x \geq 0$, such that $\phi(1) = 0$, $0\phi(0/0) = 0$ and $0\phi(x/0) = \lim_{x \to +\infty} \phi(x)/x$.

By definition, and conversely to CD distances, ϕ-divergences are not symmetric, i.e. $\mathcal{D}_\phi(\tilde{\mathrm{P}}, \mathrm{P}) \neq \mathcal{D}_\phi(\mathrm{P}, \tilde{\mathrm{P}})$. Notice that this class includes a large number of commonly used divergences, most notably Kullback-Leibler divergence (Kullback and Leibler, 1951) for $\phi(x) = x \log(x)$ and the inverse Kullback-Leibler divergence for $\phi(x) = -\log(x)$.

Proposition 1. *Let* $\mathrm{P} \in MM(A, \theta, S)$ *and suppose the parameter* θ_i *is varied, where* $i \in S_j$*. Then for a generic* θ_i*-covariation scheme* σ

$$\mathcal{D}_\phi(\sigma(\mathrm{P}), \mathrm{P}) = \sum_{y \in \mathbb{Y}^{\neq}_{S_j}} \theta^{A_y} \phi\left(\frac{\tilde{\theta}_{S_j}^{A_{y,S_j}}}{\theta_{S_j}^{A_{y,S_j}}}\right). \tag{9}$$

Proof. Notice that

$$\mathcal{D}_\phi(\sigma(\mathrm{P}), \mathrm{P}) = \sum_{y \in \mathbb{Y}} \theta^{A_y} \phi\left(\frac{\tilde{\theta}^{A_y}}{\tilde{\theta}^{A_y}}\right) = \sum_{y \in \mathbb{Y}} \theta^{A_y} \phi\left(\frac{\tilde{\theta}_{S_j}^{A_{y,S_j}}}{\tilde{\theta}_{S_j}^{A_{y,S_j}}}\right) = \sum_{y \in \mathbb{Y}^{\neq}_{S_j}} \theta^{A_y} \phi\left(\frac{\tilde{\theta}_{S_j}^{A_{y,S_j}}}{\theta_{S_j}^{A_{y,S_j}}}\right)$$

where the last equality follows by noting that for all $y \in \mathbb{Y}^{=}_{S_j}$ the term in the summation is $0\phi(0/0)$ which by definition is equal to zero. □

Notice that as for BNs and multilinear MMs, ϕ-divergences do not depend on the parameter vector θ_{S_j} of the varied parameter only, but on the full θ. Therefore, their computation in practice is more expensive than for CD distances. Furthermore, due to this extra complexity, ϕ-divergences do not simplify greatly for specific covariation schemes. To see this, the ϕ-divergence under proportional covariation can be written as

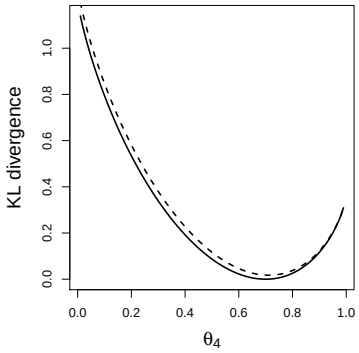
$$\mathcal{D}_\phi(\sigma_{\mathrm{pro}}(\mathrm{P}), \mathrm{P}) = \sum_{y \in \mathbb{Y}^{\neq}_{S_j}} \theta^{A_y} \phi\left(\left(\frac{\tilde{\theta}_i}{\theta_i}\right)^{A_{y,i}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i}\right)^{|A_{y,S_j^{-i}}|}\right),$$

which still depends on the full parameter vector θ. The specific form of the ϕ-divergence under other covariation schemes can be easily deduced by plugging-in their definition into equation (9).

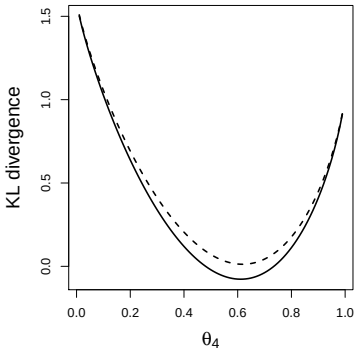
Example 7. The Kullback-Leibler divergences for proportional and uniform covariation and any possible varied value of θ_4 in the trees of Section 2.2 are reported in Figure 5. The form and the value of the divergences for the two trees are similar. Notice that for this example the Kullback-Leibler divergence is always smaller for proportional covariation than uniform covariation, although there is no theoretical guarantee that this is always the case. For instance, under the same setting of Example 4, uniform covariation has a smaller KL divergence than proportional covariation for both variations of θ_1 to 0.2 and 0.4, assuming the original probabilities were $\theta_1 = 0.33$, $\theta_2 = 0.33$ and $\theta_3 = 0.34$.

6. Discussion

The representation of probabilistic graphical models in terms of defining atomic monomial probabilities has proven useful in sensitivity analysis. Here a general approach for this type of analyses in models whose atomic probabilities are non-multilinear, including DBNs, hidden Markov models and staged trees, is introduced. The form of the sensitivity functions and various distances/divergences is derived here for a variety of covariation schemes, and their



(a) Multilinear tree.



(b) Non-multilinear tree.

Figure 6: KL divergence for the staged trees of Section 2.2 for variations of θ_4. Full lines: proportional covariation; Dashed lines: uniform covariation.

properties studied. In general these are different to their counterparts in multilinear MMs and exhibit a more complex structure. One optimality result for proportional covariation is also presented, which can guide the choice of covariation scheme in non-multilinear models in some specific cases.

The examples presented suggest that proportional covariation minimizes both CD distances and ϕ-divergences under much milder conditions than the ones given in Theorem 3. However, it is currently unknown under which conditions proportional covariation is optimal in general. General conditions of optimality in multilinear models have been derived only recently in Leonelli and Riccomagno (2018). The identification of these in the more general non-multilinear case is the subject of ongoing research.

Software for carrying out sensitivity analysis in practice is still very limited (see samIam, for a notable exception). A package for sensititivity analysis in BNs, and more generally for MMs, in the open-source R software (R Core Team, 2018) is currently under development. The development of such a package is critical and could be of great benefit for the whole AI community.

Acknowledgements

The author kindly thanks Christiane Görgen and Jim Q. Smith for comments on previous versions of the manuscript.

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society Series B*, 28:131–142, 1966.

S. B. Amsalu, A. Homaifar, and A. C. Esterline. A simplified matrix formulation for sensitivity analysis of hidden Markov models. *Algorithms*, 10 (3):97, 2017.

C. Boutilier, N. Friedman, M Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

B. Brandherm and A. Jameson. An extension of the differential approach for Bayesian network inference to dynamic Bayesian networks. *International Journal of Intelligent Systems*, 19(8):727–748, 2004.

E. Castillo and U. Kjærulff. Sensitivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique. *Reliability Engineering & System Safety*, 79(2):139–148, 2003.

E. Castillo, J. M. Gutiérrez, and A. S. Hadi. Parametric structure of probabilities in Bayesian networks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 89–98. Springer, 1995.

E. Castillo, J. M. Gutiérrez, and A. S. Hadi. Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 27(4):412–423, 1997.

H. Chan and A. Darwiche. When do numbers really matter? *Journal of Artificial Intelligence Research*, 17:265–287, 2002.

H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. *International Journal of Approximate Reasoning*, 38: 149–174, 2005a.

H. Chan and A. Darwiche. Sensitivity analysis in Markov networks. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1300–1305, 2005b.

T. Charitos and L. C. van der Gaag. Sensitivity analysis of Markovian models. In *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference*, pages 806–811, 2006a.

T. Charitos and L. C. van der Gaag. Sensitivity analysis for threshold decision making with DBNs. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 72–79, 2006b.

R.A. Collazo, C. Görgen, and J.Q. Smith. *Chain event graphs*. Chapmann & Hall, 2018.

V. M. H. Coupé and L. C. Van Der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36(4):323–356, 2002.

A. Darwiche. A differential approach to inference in Bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.

G. de Cooman, F. Hermans, and E. Quaeghebeur. Sensitivity analysis for finite Markov chains in discrete time. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 129–136, 2008.

G. Freeman and J. Q. Smith. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis*, 102:1152–1165, 2011.

M. A. Gómez-Villegas, P. Main, and R. Susi. The effect of block parameter perturbations in Gaussian Bayesian networks: sensitivity and robustness. *Information Sciences*, 222:429–458, 2013.

C. Görgen and M. Leonelli. Model-preserving sensitivity analysis for families of Gaussian distributions. *arXiv:1809.10794*, 2018.

C. Görgen, M. Leonelli, and J. Q. Smith. A differential approach for staged trees. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 346–355. Springer, 2015.

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

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

K. B. Laskey. Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(6): 901–909, 1995.

M. Leonelli and E. Riccomagno. A geometric characterization of sensitivity analysis in monomial models. *arXiv:*, 2018.

M. Leonelli, C. Görgen, and J. Q. Smith. Sensitivity analysis in multilinear probabilistic models. *Information Sciences*, 411:84–97, 2017a.

M. Leonelli, E. Riccomagno, and J. Q. Smith. A symbolic algebra for the computation of expected utilities in multiplicative influence diagrams. *Annals of Mathematics and Artificial Intelligence*, 81(3-4):273–313, 2017b.

D. Nur, D. Allingham, J. Rousseau, K. L. Mengersen, and R. McVinish. Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis. *Computational Statistics & Data Analysis*, 53(5):1873–1882, 2009.

M. Oberguggenberger, J. King, and B. Schmelzer. Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. *International Journal of Approximate Reasoning*, 50(4):680–693, 2009.

C. A. Pollino, O. Woodberry, A. Nicholson, K. Korb, and B. T. Hart. Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. *Environmental Modelling & Software*, 22(8):1140–1152, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

S. Renooij. Efficient sensitivity analysis in hidden Markov models. *International Journal of Approximate Reasoning*, 53(9):1397–1414, 2012.

S. Renooij. Co-variation for sensitivity analysis in Bayesian networks: properties, consequences and alternatives. *International Journal of Approximate Reasoning*, 55:1022–1042, 2014.

samIam. *Sensitivity analysis, modeling, inference and more*. URL `http://reasoning.cs.ucla.edu/samiam/`.

J. Q. Smith. *Bayesian decision analysis: principles and practice*. Cambridge University Press, 2010.

J.Q. Smith and P.E. Anderson. Conditional independence and chain event graphs. *Artificial Intelligence*, 172:42–68, 2008.

L. Uusitalo. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203(3-4):312–318, 2007.