Leonelli, M., Görgen, C. and Smith, J. Q. (2017) Sensitivity analysis in multilinear probabilistic models. *Information Sciences*, 411, pp. 84-97. (doi:10.1016/j.ins.2017.05.010)

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/140752/

Deposited on: 08 May 2017

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Sensitivity analysis in multilinear probabilistic models

Manuele Leonelli

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

Christiane Görgen and Jim Q. Smith

Department of Statistics, The University of Warwick, Coventry, UK

Abstract

Sensitivity methods for the analysis of the outputs of discrete Bayesian networks have been extensively studied and implemented in different software packages. These methods usually focus on the study of sensitivity functions and on the impact of a parameter change to the Chan-Darwiche distance. Although not fully recognized, the majority of these results rely heavily on the multilinear structure of atomic probabilities in terms of the conditional probability parameters associated with this type of network. By defining a statistical model through the polynomial expression of its associated defining conditional probabilities, we develop here a unifying approach to sensitivity methods applicable to a large suite of models including extensions of Bayesian networks, for instance context-specific ones. Our algebraic approach enables us to prove that for models whose defining polynomial is multilinear both the Chan-Darwiche distance and any divergence in the family of ϕ-divergences are minimized for a certain class of multi-parameter contemporaneous variations when parameters are proportionally covaried.

Keywords: Bayesian networks, CD distance, Interpolating Polynomial, Sensitivity Analysis, ϕ-divergences.

1. Introduction

Many discrete statistical problems in a variety of domains are nowadays often modeled using *Bayesian networks* (BNs) [32]. There are now thousands of practical applications of these models [4, 24, 26], which have spawned many useful technical de-

velopments: including a variety of fast exact, approximate and symbolic propagation algorithms for the computation of probabilities that exploit the underlying graph structure [14, 16, 17]. Some of these advances have been hard-wired into software [6, 27] which has further increased the applicability and success of these methods.

However, BN modeling would not have experienced such a widespread application without tailored methodologies of *model validation*, i.e. checking that a model produces outputs that are in line with current understanding, following a defensible and expected mechanism [19, 34]. Such techniques are now well established for BN models [11, 27, 34]. These are especially fundamental for expert elicited models, where both the probabilities and the covariance structure are defined from the suggestions of domains experts, following knowledge engineering protocols tailored to the BN's building process [30, 35]. We can broadly break down the validation process into two steps: the first concerns the auditing of the underlying graphical structure; the second, assuming the graph represents a user's beliefs, checks the impact of the numerical elicited probabilities within this parametric family on outputs of interest. The focus of this paper lies in this second validation phase, usually called a *sensitivity analysis*.

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. Analyses where more than one parameter at a time are varied are usually referred to as *multi-way*. In both cases a complete sensitivity analysis for discrete BNs often involves the study of *Chan-Darwiche (CD) distances* [6, 7, 8] and *sensitivity functions* [13, 40]. The CD distance is used to quantify global changes. It measures how the overall distribution behaves when one (or more) parameter is varied. A significant proportion of research has focused on identifying parameter changes such that the original and the 'varied' BN distributions are close in CD distance [8, 37]. This is minimized when, after a single arbitrary parameter change, other covarying parameters, e.g. those from the same conditional distribution, have the same proportion of the residual probability mass as they originally had. Sensitivity functions, on the other hand, model local changes with respect to an output of interest. These describe how that output probability varies as one (or potentially more) parameter is allowed to be changed. Although both these concepts can be applied to generic Bayesian analyses, they have been discussed and

applied almost exclusively within the BN literature (see [9, 10, 36] for some exceptions). This is because the computations of both CD distances and sensitivity functions are particularly straightforward for BN models.

In this paper we introduce a unifying comprehensive framework for certain multi-way analyses, usually called in the context of BNs *single full conditional probability table (CPT) analyses* - where one parameter from each CPT of one vertex of a BN given each configuration of its parents is varied. Using the notion of an interpolating polynomial [33] we are able to describe a large variety of models based on their polynomial form. Then, given this algebraic characterization, we demonstrate that one-way sensitivity methods defined for BNs can be generalized to single full CPT analyses for any model whose interpolating polynomial is multilinear, for example context-specific BNs [3] and stratified chain event graphs [12, 39]. Because of both the lack of theoretical results justifying their use and the increase in computational complexity, multi-way methods have not been extensively discussed in the literature: see [2, 7, 21] for some exceptions. This paper aims at providing a comprehensive theoretical toolbox to start applying such analyses in practice.

Importantly, our polynomial approach enables us to prove that single full CPT analyses in any multilinear model are optimal under proportional covariation in the sense that the CD distance between the original and the varied distributions is minimized. The optimality of this covariation method has been an open problem in the sensitivity analysis literature for quite some time [7, 37]. However, we are able to provide further theoretical justifications for the use of proportional covariation in single full CPT analyses. We demonstrate below that for any multilinear model this scheme minimizes not only the CD distance, but also any divergence in the family of ϕ-divergences [1, 15]. The class of ϕ-divergences include a very large number of divergences and distances (see e.g. [31] for a review), including the famous Kullback-Leibler (KL) divergence [28]. The application of KL distances in sensitivity analyses of BNs has been almost exclusively restricted to the case when the underlying distribution is assumed Gaussian [20, 21], because in discrete BNs the computation of such a divergence requires more computational power than for CD distances. We demonstrate below that this additional complexity is a feature shared by any divergence in the family of ϕ-divergences.

The paper is structured as follows. In Section 2 we define interpolating polynomials and demonstrate that commonly used models entertain a polynomial representation. In Section 3 we review a variety of divergence measures. Section 4 presents a variety of results for single full CPT sensitivity analyses in multilinear models, namely the derivation of sensitivity functions and the proof of optimality of proportional covariation. We conclude with a discussion.

2. Multilinear and polynomial parametric models

In this section we first provide a generic definition of a parametric statistical model together with the notion of interpolating polynomial. We then discuss parametric models whose interpolating polynomial is multilinear.

2.1. Parametric models and interpolating polynomials

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ be a random vector with an associated discrete and finite sample space $\mathbb{Y}$, with $\#\mathbb{Y} = q$. Although our methods straightforwardly applies when the entries of $\boldsymbol{Y}$ are random vectors, for ease of notation, we henceforth assume its elements are univariate.

Definition 1. Denote by $\mathbb{P}_{\boldsymbol{\theta}} = \{p_{\boldsymbol{\theta}}(\boldsymbol{y}) \mid \boldsymbol{y} \in \mathbb{Y}\}$ the values of a probability mass function $p_{\boldsymbol{\theta}} : \mathbb{Y} \to [0, 1]$ which depends on a choice of parameters $\boldsymbol{\theta} \in \mathbb{R}^k$. The entries of $\mathbb{P}_{\boldsymbol{\theta}}$ are called *atomic probabilities* and the elements $\boldsymbol{y} \in \mathbb{Y}$ *atoms*.

Definition 2. A discrete *parametric statistical model* on $q \in \mathbb{N}$ atoms is a subset $\mathbb{P}_\Psi \subseteq \Delta_{q-1}$ of the $q - 1$ dimensional probability simplex, where

$$\Psi \; : \quad \mathbb{R}^k \to \mathbb{P}_\Psi, \; \boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}, \tag{1}$$

is a bijective map identifying a particular choice of parameters $\boldsymbol{\theta} \in \mathbb{R}^k$ with one vector of atomic probabilities. The map Ψ is called a *parametrisation* of the model.

The above definition is often encountered in the field of *algebraic statistics*, where properties of statistical models are studied using techniques from algebraic geometry and commutative computer algebra, among others [18, 38]. We next follow [22] in extending some standard terminology.

Definition 3. A model $\mathbb{P}_\Psi \subseteq \Delta_{q-1}$ has a *monomial parametrisation* if

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}^{\boldsymbol{\alpha}_{\boldsymbol{y}}}, \quad \text{for all } \boldsymbol{y} \in \mathbb{Y},$$

where $\boldsymbol{\alpha}_{\boldsymbol{y}} \in \mathbb{N}_0^k$ denotes a vector of exponents and $\boldsymbol{\theta}^{\boldsymbol{\alpha}_{\boldsymbol{y}}} = \theta_1^{\alpha_{1,\boldsymbol{y}}} \cdots \theta_k^{\alpha_{k,\boldsymbol{y}}}$ is a monomial. Then equation (1) is a monomial map and $\boldsymbol{\theta}^{\boldsymbol{\alpha}_{\boldsymbol{y}}} \in \mathbb{R}_k[\Theta]$, for all $\boldsymbol{y} \in \mathbb{Y}$. Here $\Theta = \{\theta_1, \ldots, \theta_k\}$ is the set of indeterminates and $\mathbb{R}_k[\Theta]$ is the polynomial ring over the field $\mathbb{R}$.

For models entertaining a monomial parametrisation the network polynomial we introduce in Definition 4 below concisely captures the model structure and provides a platform to answer inferential queries [17, 23].

Definition 4. The *network polynomial* of a model $\mathbb{P}_\Psi$ with monomial parametrisation Ψ is given by

$$c_{\mathbb{P}_\Psi}(\boldsymbol{\theta}, \lambda) = \sum_{\boldsymbol{y} \in \mathbb{Y}} \lambda_{\boldsymbol{y}} \boldsymbol{\theta}^{\boldsymbol{\alpha}_{\boldsymbol{y}}},$$

where $\lambda_{\boldsymbol{y}}$ is an indicator function for the atom $\boldsymbol{y}$.

Probabilities of events in the underlying sigma-field can be computed from the network polynomial by setting equal to one the indicator function of atoms associated to that event. In the following it will be convenient to work with a special case of the network polynomial where all the indicator functions are set to one.

Definition 5. The *interpolating polynomial* of a model $\mathbb{P}_\Psi$ with monomial parametrisation Ψ is given by the sum of all atomic probabilities,

$$c_{\mathbb{P}_\Psi}(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha} \in \mathbb{A}} \boldsymbol{\theta}^{\boldsymbol{\alpha}},$$

where $\mathbb{A} = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \mid \boldsymbol{y} \in \mathbb{Y}\} \subset \mathbb{N}_0^k$.

2.2. Multilinear models

In this work we focus on parametric models whose interpolating polynomial is multilinear.

Definition 6. We say that a parametric model $\mathbb{P}_\Psi$ is *multilinear* if its associated interpolating polynomial is multilinear, i.e. if $\mathbb{A} \subseteq \{0,1\}^k$.

We note here that a great portion of well-known non-dynamic graphical models are multilinear. We explicitly show below that this is the case for BNs and context-specific BNs [3]. In [23] we showed that certain chain event graph models [39] have multilinear interpolating polynomial. In addition, decomposable undirected graphs and probabilistic chain graphs [29] can be defined to have a monomial parametrisation whose associated interpolating polynomial is multilinear. An example of models not entertaining a monomial parametrisation in terms of atomic probabilities are non-decomposable undirected graphs, since their joint distribution can then only be written as a rational function of multilinear functions [9].

2.2.1. Bayesian networks

For an $m \in \mathbb{N}$, let $[m] = \{1, \dots, m\}$. We denote with Y_i, $i \in [m]$, a generic discrete random variable and with $\mathbb{Y}_i = [m_i]$ its associated sample space. For an $A \subseteq [m]$, we let $\boldsymbol{Y}_A = (Y_i)_{i \in A}$ and $\mathbb{Y}_A = \times_{i \in A} \mathbb{Y}_i$. Recall that for three random vectors $\boldsymbol{Y}_i$, $\boldsymbol{Y}_j$ and $\boldsymbol{Y}_l$, we say that $\boldsymbol{Y}_i$ is conditional independent of $\boldsymbol{Y}_j$ given $\boldsymbol{Y}_l$, and write $\boldsymbol{Y}_i \perp\!\!\!\perp \boldsymbol{Y}_j \mid \boldsymbol{Y}_l$, if $\Pr(\boldsymbol{Y}_i = \boldsymbol{i} \mid \boldsymbol{Y}_j = \boldsymbol{j}, \boldsymbol{Y}_l = \boldsymbol{l}) = \Pr(\boldsymbol{Y}_i = \boldsymbol{i} \mid \boldsymbol{Y}_l = \boldsymbol{l})$, for every $\boldsymbol{i} \in \mathbb{Y}_i$, $\boldsymbol{j} \in \mathbb{Y}_j$ and $\boldsymbol{l} \in \mathbb{Y}_l$.

Definition 7. A BN over a discrete random vector $\boldsymbol{Y}_{[m]}$ consists of

- $m-1$ *conditional independence* statements of the form $Y_i \perp\!\!\!\perp \boldsymbol{Y}_{[i-1]} \mid \boldsymbol{Y}_{\Pi_i}$, where $\Pi_i \subseteq [i-1]$;
- a *directed acyclic graph* (DAG) $\mathcal{G}$ with vertex set $V(\mathcal{G}) = \{Y_i : i \in [m]\}$ and edge set $E(\mathcal{G}) = \{(Y_i, Y_j) : j \in [m], i \in \Pi_j\}$;
- conditional probabilities $\mathrm{P}(Y_i = j \mid \boldsymbol{Y}_{\Pi_i} = \boldsymbol{\pi})$ for every $j \in \mathbb{Y}_i$, $\boldsymbol{\pi} \in \mathbb{Y}_{\Pi_i}$ and $i \in [m]$.

The vector $\boldsymbol{Y}_{\Pi_i}$, $i \in [m]$, includes the *parents* of the vertex Y_i, i.e. those vertices Y_j such that there is an edge (Y_j, Y_i) in the DAG $\mathcal{G}$ of the BN. For a vertex Y with parents $\boldsymbol{Y}_\Pi$, let $\theta_{y\boldsymbol{\pi}} = \mathrm{P}(Y = y | \boldsymbol{Y}_\Pi = \pi)$. From [7] we know that for any atom $\boldsymbol{y} \in \mathbb{Y}_{[m]}$ its associated

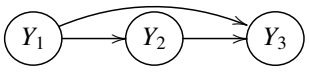


Figure 1: A BN model for the medical problem in Example 1.

monomial in the network polynomial can be written as

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \prod_{\boldsymbol{y}\sim\{y,\boldsymbol{\pi}\}} \lambda_y \theta_{y\boldsymbol{\pi}},$$

where $\sim$ denotes the compatibility relation among instantiations.

Lemma 1. *From [17, 23], the interpolating polynomial of a BN model can be written as*

$$c_{\mathrm{BN}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in\mathbb{Y}_{[m]}} \prod_{\boldsymbol{y}\sim\{y,\boldsymbol{\pi}\}} \theta_{y\boldsymbol{\pi}}. \tag{2}$$

From Equation (2) we can immediately deduce the following.

Proposition 1. *A BN is a multilinear parametric model, whose interpolating polynomial is homogeneous with monomials of degree m.*

Example 1. Suppose a newborn is at risk of acquiring a disease and her parents are offered a screening test (Y_1) which can be either positive ($Y_1 = 2$) or negative ($Y_1 = 1$). Given that the newborn can either severely ($Y_2 = 3$) or mildly ($Y_2 = 2$) contract the disease or remain healthy ($Y_2 = 1$), her parents can then decide whether or not to give her a vaccine to prevent a relapse ($Y_3 = 2$ and $Y_3 = 1$, respectively). We assume that the parents' decision about the vaccine does not depend on the screening test if the newborn contracted the disease.

The above situation can be described, with some loss of information, by the BN in Figure 1, with probabilities, for $i, l \in [2]$ and $j \in [3]$,

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

Its associated interpolating polynomial has degree 3 and equals

$$c_{\mathrm{BN}}(\boldsymbol{\theta}) = \sum_{i\in[2]} \sum_{j\in[3]} \sum_{l\in[2]} \theta_i \theta_{ji} \theta_{lji}.$$

Its specific form can also be seen as the sum of the monomials reported in Table 1.

$\theta_1\theta_{11}\theta_{111}$	$\theta_1\theta_{11}\theta_{211}$	$\theta_1\theta_{21}\theta_{121}$	$\theta_1\theta_{21}\theta_{221}$	$\theta_1\theta_{31}\theta_{131}$	$\theta_1\theta_{31}\theta_{231}$
$\theta_2\theta_{12}\theta_{112}$	$\theta_2\theta_{12}\theta_{212}$	$\theta_2\theta_{22}\theta_{122}$	$\theta_2\theta_{22}\theta_{222}$	$\theta_2\theta_{32}\theta_{132}$	$\theta_2\theta_{32}\theta_{232}$

Table 1: Monomials in the interpolating polynomial of the BN in Figure 1.

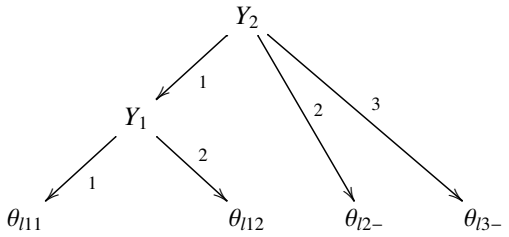


Figure 2: CSI-tree associated to vertex Y_3 of the BN in Figure 1 of Example 1, where $\theta_{l2-} = \Pr(Y_3 = l|Y_2 = 2)$, $\theta_{l3-} = \Pr(Y_3 = l|Y_2 = 3)$, $\theta_{l12} = \Pr(Y_3 = l|Y_2 = 1, Y_1 = 2)$ and $\theta_{l11} = \Pr(Y_3 = l|Y_2 = 1, Y_1 = 1)$, for $l \in [2]$.

2.2.2. Context-specific Bayesian networks

In practice it has been recognized that often conditional independence statements do not hold over the whole sample space of certain conditioning variables but only for a subset of this, usually referred to as a *context*. A variety of methods have been introduced to embellish a BN with additional independence statements that hold only over contexts. A BN equipped with such embellishments is usually called *context-specific BN*. Here we consider the representation known as *context specific independence (CSI)-trees* and introduced in [3].

Example 2. Consider the medical problem in Example 1. Using the introduced notation, we notice that by assumption, for each $l \in [2]$, the probabilities θ_{l2i} are equal for all $i \in [2]$ and called θ_{l2-}. Similarly, θ_{l3i} are equal and called θ_{l3-}, $i, l \in [2]$. These constraints can be represented by the CSI-tree in Figure 2, where the inner nodes are random variables and the leaves are entries of the CPTs of one vertex. The tree shows that, if $Y_2 = 2$ or $Y_2 = 3$ then no matter what the value of Y_1 is, the CPT for $Y_3 = l$ will be equal to θ_{l2-} and θ_{l3-} respectively. In our polynomial approach, context-specific independences can be straightforwardly imposed in the interpolating polynomial rep-

resentation of the model. In fact the interpolating polynomial for the model in this example corresponds to the polynomial in equation (2) where the appropriate indeterminates are substituted with θ_{l2-} and θ_{l3-}. This polynomial is again multilinear and homogeneous, just like for all context-specific BNs embellished with CSI-trees.

We notice here that the interpolating polynomial of a multilinear model is not necessarily homogenous, as for example the one associated to certain chain event graph models, as shown in [23].

3. Divergence measures

In sensitivity analyses for discrete parametric statistical models we are often interested in studying how far apart from each other are two vectors of values of two probability mass functions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ from the same model $\mathbb{P}_{\Psi}$. Divergence measures are used to quantify this dissimilarity between probability distributions. In this section we provide a brief introduction to these functions within the context of our discrete parametric probability models.

Definition 8. A *divergence measure* $\mathcal{D}$ within a discrete parametric probability model $\mathbb{P}_{\Psi}$ is a function $\mathcal{D}(\cdot,\cdot) : \mathbb{P}_{\Psi} \times \mathbb{P}_{\Psi} \to \mathbb{R}$ such that for all $\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}} \in \mathbb{P}_{\Psi}$:

- $\mathcal{D}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) \geq 0$;
- $\mathcal{D}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = 0$ iff $\mathbb{P}_{\boldsymbol{\theta}} = \mathbb{P}_{\tilde{\boldsymbol{\theta}}}$.

The larger the divergence between two probability mass functions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$, the more dissimilar these are. Notice that divergences are not formally metrics, since these do not have to be symmetric and respect the triangular inequality. We refer to divergences with these two additional properties as *distances*.

The divergence most commonly used in practice is the KL divergence [28].

Definition 9. The *KL divergence* between $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}} \in \mathbb{P}_{\Psi}$, $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$, is defined as

$$\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \sum_{\boldsymbol{y} \in \mathbb{Y}} p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y}) \log\left(\frac{p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y})}{p_{\boldsymbol{\theta}}(\boldsymbol{y})}\right), \tag{3}$$

assuming $p_{\boldsymbol{\theta}}(\boldsymbol{y}), p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in \mathbb{Y}$.

Notice that the KL divergence is not symmetric and thus $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) \neq \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ in general. However both divergences can be shown to be a particular instance of a very general family of divergences, called ϕ-divergences [1, 15].

Definition 10. The *ϕ-divergence* between $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}} \in \mathbb{P}_{\Psi}$, $\mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$, is defined as

$$\mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \sum_{\boldsymbol{y} \in \mathbb{Y}} p_{\boldsymbol{\theta}}(\boldsymbol{y}) \phi\left(\frac{p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y})}{p_{\boldsymbol{\theta}}(\boldsymbol{y})}\right), \qquad \phi \in \Phi, \tag{4}$$

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

So for example $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ for $\phi(x) = x \log(x)$ and $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ for $\phi(x) = -\log(x)$. Many other renowned divergences are in the family of ϕ-divergences: for example J divergences [25] and total variation distances (see [31] for a review).

The distance usually considered to study the dissimilarity of two probability mass functions in sensitivity analyses for discrete BNs is the aforementioned Chan-Darwiche distance. This distance is not a member of the ϕ-divergence family.

Definition 11. The *CD distance* between $\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}} \in \mathbb{P}_{\Psi}$, $\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})$, is defined as

$$\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \log \max_{\boldsymbol{y} \in \mathbb{Y}} \frac{p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y})}{p_{\boldsymbol{\theta}}(\boldsymbol{y})} - \log \min_{\boldsymbol{y} \in \mathbb{Y}} \frac{p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y})}{p_{\boldsymbol{\theta}}(\boldsymbol{y})}, \tag{5}$$

where $0/0$ is defined as 1.

Notice that $\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ since CD is formally a distance and not a divergence. It has been noted that in sensitivity analysis in BNs, if one parameter of one CPT is varied, then the CD distance between the original and the varied BN equals the CD distance between the original and the varied CPT [8]. This distributive property, and its associated computational simplicity, has lead to a wide use of the CD distance in sensitivity studies in discrete BNs.

4. Sensitivity analysis in multilinear models

We can now formalize sensitivity analysis techniques for multilinear parametric models. We focus on an extension of single full CPT analyses from BNs to generic

multilinear models. Standard one-way sensitivity analyses can be seen as a special case of single full CPT analyses when only one parameter is allowed to be varied. We demonstrate in this section that all the results about one-way sensitivity analysis in BN models extend to single full CPT analyses in multilinear parametric models and therefore hold under much weaker assumptions about the structure of both the sample space and the underlying conditional independences. Before presenting these results we review the theory of *covariation*.

4.1. Covariation

In one-way analyses one parameter within a parametrisation of a model is varied. When this is done, then *some* of the remaining parameters need to be varied as well to respect the sum-to-one condition, so that the resulting measure is a probability measure. In the binary case 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 the user needs to take into account, as reviewed below.

Let $\theta_i \in \Theta$ be the parameter varied to $\tilde{\theta}_i$ and suppose this is associated to a random variable Y_C in the random vector $\boldsymbol{Y}$. Let $\Theta_C = \{\theta_1, \ldots, \theta_i, \ldots, \theta_r\} \subseteq \Theta$ be the subset of the parameter set including θ_i describing the probability distribution of Y_C and whose elements need to respect the sum to one condition. For instance Θ_C would include the entries of a CPT for a fixed combination of the parent variables in a BN model or the entries of a CPT associated to the conditional random variable from a leaf of a CSI-tree as in Figure 2. Suppose further these parameters are indexed according to their values, i.e. $\theta_1 \leq \cdots \leq \theta_i \leq \cdots \leq \theta_r$. From [37] we then have the following definition.

Definition 12. Let $\theta_i \in \Theta_C$ be varied to $\tilde{\theta}_i$. A *covariation* scheme $\sigma(\theta_j, \tilde{\theta}_i) : [0,1]^2 \to [0,1]$ is a function that takes as input the value of both $\tilde{\theta}_i$ and $\theta_j \in \Theta_C$ and returns an updated value for θ_j denoted as $\tilde{\theta}_j$.

Different covariation schemes may entertain different properties which, depending on the domain of application, might be more or less desirable. We now list some of these properties from [37].

Definition 13. In the notation of Definition 12, a covariation scheme $\sigma(\theta_j, \tilde{\theta}_i)$ is

- *valid*, if $\sum_{j\in[r]} \sigma(\theta_j, \tilde{\theta}_i) = 1$;
- *impossibility preserving*, if for any parameter $\theta_j = 0$, $j \neq i$, we have that $\sigma(\theta_j, \tilde{\theta}_i) = 0$;
- *order preserving*, if $\sigma(\theta_1, \tilde{\theta}_i) \leq \cdots \leq \sigma(\theta_j, \tilde{\theta}_i) \leq \cdots \leq \sigma(\theta_r, \tilde{\theta}_i)$;
- *identity preserving*, if $\sigma(\theta_j, \theta_i) = \theta_j$, $\forall\, j \in [r]$;
- *linear*, if $\sigma(\theta_j, \tilde{\theta}_i) = \gamma_j \tilde{\theta}_i + \delta_j$, for $\gamma_j \in [0, 1]$ and $\delta_j \in (-1, 1)$.

Of course any covariation scheme needs to be valid, otherwise the resulting measure is not a probability measure and any inference from the model would be misleading. Applying a linear scheme is very natural: if for instance $\delta_j = -\gamma_j$, then $\sigma(\theta_j, \tilde{\theta}_i) = \delta_j(1 - \tilde{\theta}_i)$ and the scheme assigns a proportion δ_j of the remaining probability mass $1 - \tilde{\theta}_i$ to the remaining parameters. Following [37] we now introduce a number of frequently applied covariation schemes.

Definition 14. In the notation of Definition 12, we define

- the *proportional* covariation scheme, $\sigma_{\text{pro}}(\theta_j, \tilde{\theta}_i)$, as

$$\sigma_{\text{pro}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{1-\tilde{\theta}_i}{1-\theta_i}\theta_j, & \text{otherwise.} \end{cases}$$

- the *uniform* covariation scheme, $\sigma_{\text{uni}}(\theta_j, \tilde{\theta}_i)$, for $r = \#\Theta_C$, as

$$\sigma_{\text{uni}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{1-\tilde{\theta}_i}{r-1}, & \text{otherwise.} \end{cases}$$

- the *order preserving* covariation scheme, $\sigma_{\text{ord}}(\theta_j, \tilde{\theta}_i)$, for $i \neq r$, as

$$\sigma_{\text{ord}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{\theta_j}{\theta_i}\tilde{\theta}_i, & \text{if } j < i \text{ and } \tilde{\theta}_i \leq \theta_i, \\ \frac{-\theta_j(1-\theta_{\text{suc}})}{\theta_{\text{suc}}\theta_i}\tilde{\theta}_i + \frac{\theta_j}{\theta_{\text{suc}}}, & \text{if } j > i \text{ and } \tilde{\theta}_i \leq \theta_i, \\ \frac{\theta_j}{\theta_{\max}-\theta_i}(\theta_{\max} - \tilde{\theta}_i), & \text{if } j < i \text{ and } \tilde{\theta}_i > \theta_i, \\ \frac{\theta_j-\theta_{\max}}{\theta_{\max}-\theta_i}(\theta_{\max} - \tilde{\theta}_i) + \theta_{\max}, & \text{if } j > i \text{ and } \tilde{\theta}_i > \theta_i, \end{cases}$$

  where $\theta_{\max} = 1/(1 + r - i)$ is the upper bound for $\tilde{\theta}_i$ and $\theta_{\text{suc}} = \sum_{k=i+1}^{r} \theta_k$ is the original mass of the parameters succeeding θ_i in the ordering.

Scheme/Property	valid	imp-pres	ord-pres	ident-pres	linear
Proportional	✓	✓	✗	✓	✓
Uniform	✓	✗	✗	✗	✓
Order Preserving	✓	✓	✓	✓	✓

Table 2: Summary of the covariation schemes and the properties these entertain.

Table 2 summarizes which of the properties introduced in Definition 13 the above schemes entertain (see [37] for more details). Under proportional covariation, to all the covarying parameters, i.e. those parameters $\theta_j \in \Theta_C \setminus \{\theta_i\}$, is assigned the same proportion of the remaining probability mass as these originally had. Although this scheme is not order preserving, it maintains the order among the covarying parameters. The uniform scheme on the other hand gives the same amount of the remaining mass to all covarying parameters. In addition, although the order preserving scheme is the only entertaining the order preserving property, this limits the possible variations allowed. Note that this scheme is piece-wise linear, i.e. a function composed of straight-line sections. All the schemes in Definition 14 are domain independent and therefore can be applied with no prior knowledge about the application of interest. Other schemes, for instance domain dependent or non-linear, have been defined, but these are not of interest for the theory we develop here.

4.2. Sensitivity functions

We now generalize one-way sensitivity methods in BNs to the single full CPT case for general multilinear models. This type of analysis is simpler than other multi-way methods since the parameters varied/covaried never appear in the same monomial of the BN interpolating polynomial. So we now find an analogous CPT analysis in multilinear models which has the same property. Suppose we vary n parameters $\theta_{1_i}, \ldots, \theta_{n_i}$ and denote by $\Theta_j = \{\theta_{j_1}, \ldots, \theta_{j_{r_j}}\}$, $j \in [n]$, the set of parameters including θ_{j_i} and associated to the same (conditional) random variable: thus respecting the sum to one condition. Assume these sets are such that $\cap_{j\in[n]}\Theta_j = \emptyset$. Note that a collection of such sets can not only be associated to the CPTs of one vertex given different parent configurations,

but also, for instance, to the leaves of a CSI-tree as in Figure 2 or to the positions along the same *cut* in a CEG [39].

We start by investigating sensitivity functions. These describe the effect of the variation of the parameters $\theta_{1_i}, \dots, \theta_{n_i}$ on the probability of an event $\mathbb{Y}_T \subseteq \mathbb{Y}$ of interest. A sensitivity function $f_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ equals the probability $\mathrm{P}(\boldsymbol{Y} \in \mathbb{Y}_T) \triangleq p_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{y}_T)$ and is a function in $\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}$, where $\theta_{1_i}, \dots, \theta_{n_i}$ are varied to $\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}$. Our parametric definition of a statistical model enables us to explicitly express these as functions of the covariation scheme for any multilinear model. Recall that $\mathbb{A} = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}\}$ and let $\mathbb{T} = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}_T\}$. Let $\mathbb{A}_j, \mathbb{T}_j \subset \{0,1\}^k$ be the subsets of $\mathbb{A}$ and $\mathbb{T}$ respectively including the exponents where the entry associated to an indeterminate in Θ_j is not zero, $\mathbb{A}_{j_s} \subseteq \mathbb{A}_j$ and $\mathbb{T}_{j_s} \subseteq \mathbb{T}_j$ be the subsets including the exponents such that the entry relative to θ_{j_s} is not zero, $j \in [n]$, $s \in [r_j]$. Formally,

$$\mathbb{A}_j = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}, \alpha_{j_s,\boldsymbol{y}} \neq 0, s \in [r_j]\}, \qquad \mathbb{T}_j = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}_T, \alpha_{j_s,\boldsymbol{y}} \neq 0, s \in [r_j]\},$$
$$\mathbb{A}_{j_s} = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}, \alpha_{j_s,\boldsymbol{y}} \neq 0\}, \qquad \mathbb{T}_{j_s} = \{\boldsymbol{\alpha}_{\boldsymbol{y}} \,|\, \boldsymbol{y} \in \mathbb{Y}_T, \alpha_{j_s,\boldsymbol{y}} \neq 0\}.$$

Let $\mathbb{A}_{-j_s}, \mathbb{T}_{-j_s} \subseteq \{0,1\}^{k-1}$ be the sets including the elements in $\mathbb{A}_{j_s}$ and $\mathbb{T}_{j_s}$, respectively, where the entry relative to $\theta_{j_s} \in \Theta_j$ is deleted. Lastly, let $\boldsymbol{\theta}_{-j_s} = \prod_{\theta_k \in \Theta \setminus \{\theta_{j_s}\}} \theta_k$.

Example 3. To illustrate the notation introduced, consider the medical application of Example 1 described by the BN in Figure 1. For this example

$$\Theta = \{\theta_1, \theta_2, \theta_{11}, \theta_{21}, \theta_{31}, \theta_{12}, \theta_{22}, \theta_{32}, \theta_{111}, \theta_{211}, \theta_{121}, \theta_{221}, \theta_{131}, \theta_{231}, \theta_{112}, \theta_{212}, \theta_{122}, \theta_{222}, \theta_{132}, \theta_{232}\},$$

and the elements of the associated set $\mathbb{A}$ are reported in Table 3. For example we can see that the top-left element of Table 3 is associated to the monomial $\theta_1\theta_{11}\theta_{111}$. Now suppose we vary the parameters θ_{21} and θ_{22}. Let $\Theta_1 = \{\theta_{11}, \theta_{21}, \theta_{31}\}$ and $\Theta_2 = \{\theta_{12}, \theta_{22}, \theta_{32}\}$ be the two sets of parameters that need to respect the sum to one condition after parameters' variations. Then $\mathbb{A}_1$ and $\mathbb{A}_2$ simply correspond to the left and right column of Table 3, respectively, since for instance the left column has non-zero exponents for the elements in Θ_1. Then, for instance, the set A_{1_1} comprising all exponents with a non-zero entry for θ_{11} corresponds to the first two entries on the left column of Table 3. Conversely, the set $\mathbb{A}_{-1_1}$ includes the first two rows in the left column of Table 3, but for each of these the third vector entry, that associated to θ_{11}, is deleted. Last consider

$(1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)$	$(0,1,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0)$
$(1,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0)$	$(0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0)$
$(1,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)$	$(0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0)$
$(1,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0)$	$(0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0)$
$(1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0)$	$(0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0)$
$(1,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0)$	$(0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1)$

Table 3: Elements of the set $\mathbb{A}$ for the BN in Figure 1.

the event $Y_1 = 1$, i.e. that the screening test is negative. The associated set $\mathbb{T}$ then corresponds to the left column of Table 3. The sets $\mathbb{T}_j$ and $\mathbb{T}_{j_s}$ can then be deduced by similar observations as for $\mathbb{A}_j$ and $\mathbb{A}_{j_s}$.

Proposition 2. *Consider a multilinear model $\mathbb{P}_\Psi$ where the parameters $\theta_{j_i} \in \Theta_j$ are varied to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ is covaried according to a valid scheme $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, $j \in [n]$, $s \in [r_j] \setminus \{j_i\}$. The sensitivity function $f_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i})$ can then be written as*

$$f_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i}) = \sum_{j\in[n]} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_i}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_i} \tilde{\theta}_{j_i} + \sum_{j\in[n]} \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s} \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}}. \quad (6)$$

Proof. The probability of interest can be written as

$$\begin{aligned} p_\theta(\boldsymbol{y}_T) &= \sum_{\boldsymbol{\alpha}\in\mathbb{T}} \theta^{\boldsymbol{\alpha}} = \sum_{j\in[n]} \sum_{s\in[r_j]} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s} \theta_{j_s} + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}} \\ &= \sum_{j\in[n]} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_i}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_i} \theta_{j_i} + \sum_{j\in[n]} \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s} \theta_{j_s} + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}}. \end{aligned}$$

The result follows by substituting the varying parameters with their varied version. □

From Proposition 2 we can deduce that for a multilinear model, under a linear covariation scheme, the sensitivity function is multilinear.

Corollary 1. *Under the conditions of Proposition 2 and the linear covariation schemes $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) = \gamma_{j_s}\tilde{\theta}_{j_i} + \delta_{j_s}$, the sensitivity function $f_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i})$ equals*

$$f_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i}) = \sum_{j\in[n]} a_j \tilde{\theta}_{j_i} + b, \quad (7)$$

where

$$a_j = \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_i}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_i} + \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\gamma_{j_s}, \qquad b = \sum_{j\in[n]} \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\delta_{j_s} + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}}. \quad (8)$$

Proof. The result follows by substituting the definition of a linear covariation scheme into equation (6) and then rearranging. □

Therefore, under a linear covariation scheme, the sensitivity function is a multilinear function of the varying parameters $\tilde{\theta}_{j_i}$, $j \in [n]$. This was long known for BN models [5, 37, 40]. However, we have proven here that this feature is shared amongst all models having a multilinear interpolating polynomial. In BNs the computation of the coefficients a_j and b is particularly fast since for these models computationally efficient propagation techniques have been established. But these exist, albeit sometimes less efficiently, for other models as well (see e.g. [14] for chain graphs). Within our symbolic definition, we note however that once the exponent sets $\mathbb{T}_{-j_s}$, $s \in [r_j]$, are identified, then one can simply plug-in the values of the indeterminates to compute these coefficients.

We now deduce the sensitivity function when parameters are varied using the popular proportional scheme.

Corollary 2. *Under the conditions of Proposition 2 and proportional covariation schemes* $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) = \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}}\theta_{j_s}$*, the sensitivity function,* $f_{\mathbf{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i})$ *can be written in the multilinear form of equation (7), where*

$$a_j = \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_i}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_i} - \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{j_s}} \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}{1-\theta_{j_i}}, \qquad b = \sum_{j\in[n]} \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{j_s}} \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}{1-\theta_{j_i}} + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}}.$$

Proof. For a proportional scheme the coefficients in the definition of a linear scheme equals $\gamma_{j_s} = -\theta_{j_s}/(1-\theta_{j_i})$ and $\delta_{j_s} = \theta_{j_s}/(1-\theta_{j_i})$. By substituting these expressions into equation (8) we have that

$$a_j = \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_i}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_i} - \sum_{s\in[r_j]\setminus\{j_i\}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s} \frac{\theta_{j_s}}{1-\theta_{j_i}}, \quad b = \sum_{\substack{j\in[n] \\ s\in[r_j]\setminus\{j_i\}}} \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s} \frac{\theta_{j_s}}{1-\theta_{j_i}} + \sum_{\boldsymbol{\alpha}\in\mathbb{T}\setminus\cup_{k\in[n]}\mathbb{T}_k} \boldsymbol{\theta}^{\boldsymbol{\alpha}}.$$

By noting that $\sum_{\boldsymbol{\alpha}\in\mathbb{T}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\theta_{j_s} = \sum_{\boldsymbol{\alpha}\in\mathbb{T}_{j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}$ the result then follows. □

$\theta_1 = 0.6,$	$\theta_{11} = 0.5,$	$\theta_{21} = 0.4,$	$\theta_{12} = 0.24,$	$\theta_{22} = 0.35$	
$\theta_{111} = 0.8,$	$\theta_{121} = 0.5,$	$\theta_{131} = 0.2,$	$\theta_{112} = 0.6$	$\theta_{122} = 0.5$	$\theta_{132} = 0.2.$

Table 4: Probability specifications for Example 2.

It is often of interest to investigate the conditional probability of a target event ($\boldsymbol{Y} \in \mathbb{Y}_T$) given that an event ($\boldsymbol{Y} \in \mathbb{Y}_O$) has been observed, $\mathbb{Y}_T, \mathbb{Y}_O \subseteq \mathbb{Y}$. This can be represented by the *conditional* sensitivity function $f^{\boldsymbol{y}_O}_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i})$ describing the probability $\mathrm{P}(\boldsymbol{Y} \in \mathbb{Y}_T \mid \boldsymbol{Y} \in \mathbb{Y}_O)$ as a function of the varying parameters $\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i}$.

Corollary 3. *Under the conditions of Corollary 1, a conditional sensitivity function* $f^{\boldsymbol{y}_O}_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i})$ *can be written as the ratio*

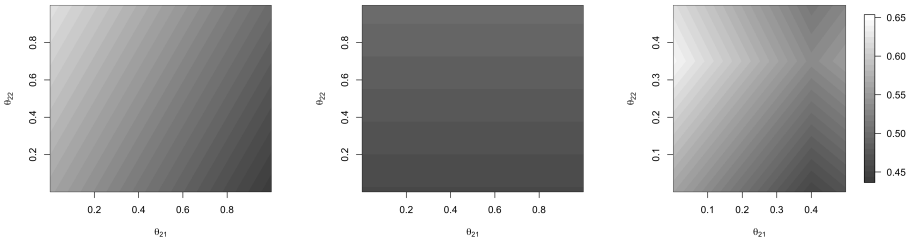
$$f^{\boldsymbol{y}_O}_{\boldsymbol{y}_T}(\tilde{\theta}_{1_i}, \ldots, \tilde{\theta}_{n_i}) = \frac{\sum_{j\in[n]} c_j \tilde{\theta}_{j_i} + d}{\sum_{j\in[n]} e_j \tilde{\theta}_{j_i} + f}, \tag{9}$$

where $c_j, e_j \in [0, 1]$, $j \in [n]$, *and* $d, f \in (-1, 1)$.

Proof. The result follows from equation (7) and by noting that $\mathrm{P}(\boldsymbol{Y} \in \mathbb{Y}_T \mid \boldsymbol{Y} \in \mathbb{Y}_O) = \mathrm{P}(\boldsymbol{Y} \in \{\mathbb{Y}_T \cap \mathbb{Y}_O\})/\mathrm{P}(\boldsymbol{Y} \in \mathbb{Y}_O)$. □

The form of the coefficients in Corollary 3 can be deduced by simply adapting the notation of equation (6) to the events $\mathrm{P}(\boldsymbol{Y} \in \{\mathbb{Y}_T \cap \mathbb{Y}_O\})$ and $\mathrm{P}(\boldsymbol{Y} \in \mathbb{Y}_O)$ for the numerator and the denominator, respectively, of equation (9). Sensitivity functions describing conditional probabilities in BNs have been proven to entertain the form in equation (9). Again, Corollary 3 shows that this is so for any model having a multilinear interpolating polynomial.

Example 4. Suppose the BN model definition in Example 1 is completed by the probability specifications in Table 4. Suppose we are interested in the event that the parents do not decide to vaccine. Figure 3 shows the sensitivity functions for this event when θ_{21} (on the x-axis) and θ_{22} (on the y-axis) are varied and the other covarying parameter are changed with different schemes. We can notice that for uniform and proportional covariation the sensitivity function is linear in its arguments, whilst for order preserving covariation this is piece-wise linear. Notice that the resulting probabilities after



(a) Proportional. (b) Uniform. (c) Order-preserving.

Figure 3: Sensitivity functions for Example 4 under different covariation schemes.

full single CPT variations are significantly different for different covariation schemes. Thus, without formal justifications to prefer one scheme over the others, any inference resulting from such sensitivity analyses might not be tenable.

4.3. The Chan-Darwiche distance

Whilst sensitivity functions study local changes, CD distances describe global variations in distributions [8]. These can be used to study by how much two vectors of atomic probabilities vary in their distributional assumptions if one arises from the other via a covariation scheme. We are then interested in the global impact of that local change.

We next characterize the form of the CD distance for multilinear models in single full CPT analyses, first generalizing its form, again derived in [37] for BN models. We demonstrate that the distance depends only on the varied and covaried parameters: thus very easy to compute.

Proposition 3. *Let $\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}} \in \mathbb{P}_\Psi$, where $\mathbb{P}_\Psi$ is a multilinear parametric model and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ arises from $\mathbb{P}_{\boldsymbol{\theta}}$ by varying θ_{j_i} to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ to $\tilde{\theta}_{j_s} = \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, where $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$ is a valid covariation scheme, $j \in [n], s \in [r_j] \setminus \{j_i\}$. Then the CD distance between $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ is equal to*

$$\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \log \max_{\substack{j\in[n]\\ s\in[r_j]}} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}} - \log \min_{\substack{j\in[n]\\ s\in[r_j]}} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}. \tag{10}$$

Proof. For a multilinear parametric model the CD distance can be written as

$$\begin{aligned}\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) &= \log \max_{\boldsymbol{\alpha}\in\mathbb{A}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}} - \log \min_{\boldsymbol{\alpha}\in\mathbb{A}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}} \\ &= \log \max \left\{ \max_{\substack{j\in[n]\\ \boldsymbol{\alpha}\in\mathbb{A}_j}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}, \max_{\boldsymbol{\alpha}\in\mathbb{A}\setminus\cup_{l\in[n]}\mathbb{A}_l} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}} \right\} - \log \min \left\{ \min_{\substack{j\in[n]\\ \boldsymbol{\alpha}\in\mathbb{A}_j}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}, \min_{\boldsymbol{\alpha}\in\mathbb{A}\setminus\cup_{l\in[n]}\mathbb{A}_l} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}} \right\}.\end{aligned}$$

If $\boldsymbol{\alpha} \in \mathbb{A} \setminus \cup_{l\in[n]}\mathbb{A}_l$, then $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and thus $\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}/\boldsymbol{\theta}^{\boldsymbol{\alpha}} = 1$. Because of the validity of the covariation scheme note that $\max_{\boldsymbol{\alpha}\in\mathbb{A}_j} \tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}/\boldsymbol{\theta}^{\boldsymbol{\alpha}} \geq 1$ and $\min_{\boldsymbol{\alpha}\in\mathbb{A}_j} \tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}/\boldsymbol{\theta}^{\boldsymbol{\alpha}} \leq 1$, for all $j \in [n]$. Thus

$$\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \log \max_{\substack{j\in[n]\\ \boldsymbol{\alpha}\in\mathbb{A}_j}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}} - \log \min_{\substack{j\in[n]\\ \boldsymbol{\alpha}\in\mathbb{A}_j}} \frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}.$$

Now note that $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_{-j_s}\tilde{\theta}_{j_s}$, for a $\theta_{j_s} \in \Theta_j$, $j \in [n]$, since no two parameters in $\cup_{j\in[n]}\Theta_j$ can have exponent non-zero in the same monomial. Thus $\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}/\boldsymbol{\theta}^{\boldsymbol{\alpha}} = \tilde{\theta}_{j_s}/\theta_{j_s}$ since $\boldsymbol{\alpha} \in \{0,1\}^k$ and the result follows. □

We can now prove that the proportional covariation scheme is optimal for single full CPT analyses. This is important since a set of parameters might be varied to change an uncalibrated probability of interest, but a user might want to achieve this by choosing a distribution as close as possible to the original one. Several authors have posed this problem for BNs without finding a definitive answer [7, 37]. Here, exploiting our polynomial model representation, we can prove the optimality of the proportional scheme not only for BN models, but also for multilinear ones in single full CPT analyses.

Theorem 1. *Under the conditions of Proposition 3 and proportional covariations $\sigma_j(\theta_{j_s}, \theta_{j_i})$, the CD distance between $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ is minimized and can be written in closed form as*

$$\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \log \max_{j\in[n]} \left\{ \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}}, \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}} \right\} - \log \min_{j\in[n]} \left\{ \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}}, \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}} \right\}. \tag{11}$$

Proof. First note that we can write equation (10) as

$$\mathcal{D}_{\mathrm{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \log \max \left\{ \max_{s\in[r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \max_{s\in[r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\} - \log \min \left\{ \min_{s\in[r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \min_{s\in[r_s]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\}. \tag{12}$$

Now, let $\bar{\theta}_{j_i} = \tilde{\theta}_{j_i}$ and suppose $\bar{\theta}_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ is obtained via a valid covariation scheme, $j \in [n]$, $s \in [r_j]$. We want to prove that $\mathcal{D}_{\text{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\bar{\boldsymbol{\theta}}}) \geq \mathcal{D}_{\text{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})$. Suppose now the proportional scheme is optimal for one-way sensitivity analyses. If this is true, we must have that, for all $j \in [n]$,

$$\max_{s\in[r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \geq \max_{s\in[r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}, \qquad \text{and} \qquad \min_{s\in[r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \leq \min_{s\in[r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}.$$

Therefore,

$$\max\left\{\max_{s\in[r_1]} \frac{\bar{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \max_{s\in[r_n]} \frac{\bar{\theta}_{n_s}}{\theta_{n_s}}\right\} \geq \max\left\{\max_{s\in[r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \max_{s\in[r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}}\right\},$$

and

$$\min\left\{\min_{s\in[r_1]} \frac{\bar{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \min_{s\in[r_n]} \frac{\bar{\theta}_{n_s}}{\theta_{n_s}}\right\} \leq \min\left\{\min_{s\in[r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \ldots, \min_{s\in[r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}}\right\},$$

from which the optimality condition follows.

We thus have to prove that for a single parameter change, the proportional covariation scheme minimizes the CD distance in any multilinear model. The proof follows similar steps to the ones in [6] for BNs. Fix $j \in [n]$ and note that if either $\theta_{j_i} = 0$ or $\theta_{j_i} = 1$ then the distance is infinite under both covariation schemes and the result holds. Consider now $\theta_{j_i} \in (0, 1)$ and suppose $\bar{\theta}_{j_i} = \tilde{\theta}_{j_i} > \theta_{j_i}$. Under a proportional scheme, we have that

$$\max_{s\in[r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}} = \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}} \quad \text{and} \quad \min_{s\in[r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}} = \min_{s\in[r_j]\setminus\{j_i\}} \frac{\theta_{j_s}(1-\tilde{\theta}_{j_i})}{(\theta_{j_s}(1-\theta_{j_i}))} = \frac{(1-\tilde{\theta}_{j_i})}{(1-\theta_{j_i})}.$$

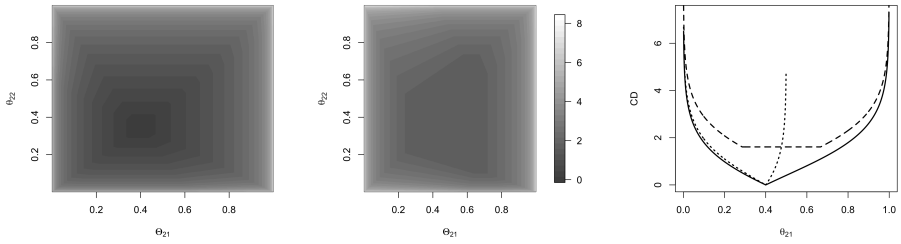
Conversely, for the generic covariation scheme $\sigma(\theta_{j_s}, \bar{\theta}_{j_i})$ we have that

$$\begin{aligned}\frac{1-\bar{\theta}_{j_i}}{1-\theta_{j_i}} &= \frac{\sum_{s\in[r_j]\setminus\{j_i\}} \bar{\theta}_{j_s}}{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}} = \frac{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}(\bar{\theta}_{j_s}/\theta_{j_s})}{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}} \\ &\geq \frac{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}(\min_{k\in[r_j]} \bar{\theta}_k/\theta_k)}{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}} = \min_{s\in[r_j]} \frac{\bar{\theta}_s}{\theta_s}.\end{aligned}$$

Thus since $(1-\bar{\theta}_{j_i})/(1-\theta_{j_i}) = (1-\tilde{\theta}_{j_i})/(1-\theta_{j_i})$ we have that $\min_{s\in[r_j]} \tilde{\theta}_{j_s}/\theta_{j_s} \geq \min_{s\in[r_j]} \bar{\theta}_{j_s}/\theta_{j_s}$. Furthermore,

$$\max_{s\in[r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \geq \frac{\bar{\theta}_{j_i}}{\theta_{j_i}} = \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}} = \max_{s\in[r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}.$$

It then follows that $\mathcal{D}_{\text{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\bar{\boldsymbol{\theta}}}) \geq \mathcal{D}_{\text{CD}}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})$ when $\tilde{\theta}_{j_i} > \theta_{j_i}$ for one-way analyses. For the case $\tilde{\theta}_{j_i} < \theta_{j_i}$ the proof mirrors the one presented here. The explicit form of



(a) Proportional. (b) Uniform. (c) One-way analysis.

Figure 4: CD distances for Example 5 under different covariation schemes: proportional (black), uniform (dashed), order-preserving (dotted).

the distance under proportional covariation schemes in equation (11) follows by noting that the maximum and the minimum can either be $\tilde{\theta}_{j_i}/\theta_{j_i}$ or $(1-\tilde{\theta}_{j_i})/(1-\theta_{j_i})$. □

Example 5. In Figure 4 we plot the CD distance between the varied and the original probability distributions for Example 1 when θ_{21} (x-axis) and θ_{22} (y-axis in 4a and 4b) are varied for the covariation schemes so far considered. From Figures 4a and 4b we can see intuitively why the distance under proportional covariation is smaller than in the uniform case. This becomes clearer when we only let θ_{21} vary as shown in Figure 4c since then the solid line representing proportional covariation is always underneath the others. Notice that in Figure 4c the CD distance for order preserving covariation is computed up to 0.5, the value associated to $\theta_{\max}$ in this example.

4.4. ϕ-divergences

Although the CD distance is widely used in sensitivity analyses, comparisons between two generic distributions are usually performed by computing the KL divergence. For one-way sensitivity analysis in BNs, the KL divergence equals the KL divergence between the original and varied conditional probability distribution of the manipulated parameter times the marginal probability of the conditioning parent configuration [8]. This means that one way sensitivity analyses based on KL distances can become computationally infeasible, since this constant term might need to be computed an arbitrary large number of times. In Proposition 4 below we demonstrate that

this property is common to any ϕ-divergence for any multilinear model and single full CPT analyses.

Proposition 4. *Let $\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}} \in \mathbb{P}_\Psi$, where $\mathbb{P}_\Psi$ is a multilinear parametric model and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ arises from $\mathbb{P}_{\boldsymbol{\theta}}$ by varying θ_{j_i} to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ to $\tilde{\theta}_{j_s} = \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, where $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$ is a valid covariation scheme, $j \in [n], s \in [r_j] \setminus \{j_i\}$. Then the ϕ-divergence between $\mathbb{P}_{\tilde{\boldsymbol{\theta}}}$ and $\mathbb{P}_{\boldsymbol{\theta}}$ is equal to*

$$\mathcal{D}_\phi(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \sum_{j\in[n]} w_j \mathcal{D}_\phi(\mathbb{P}^j_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}^j_{\boldsymbol{\theta}}), \tag{13}$$

where $\mathbb{P}^j_{\boldsymbol{\theta}}$ denotes the vector of atomic probabilities in Θ_j and $w_j \in [0, 1]$.

Proof. For a model with monomial parametrisation the ϕ-divergence can be written as

$$\mathcal{D}_\phi(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \boldsymbol{\theta}^{\boldsymbol{\alpha}} \phi\left(\frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}\right) = \sum_{j\in[n]} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_j} \boldsymbol{\theta}^{\boldsymbol{\alpha}} \phi\left(\frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}\right) + \sum_{\boldsymbol{\alpha}\in\mathbb{A}\setminus\cup_{j\in[n]}\mathbb{A}_j} \boldsymbol{\theta}^{\boldsymbol{\alpha}} \phi\left(\frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}\right).$$

Notice that for $\boldsymbol{\alpha} \in \mathbb{A} \setminus \cup_{j\in[n]}\mathbb{A}_j$, $\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}/\boldsymbol{\theta}^{\boldsymbol{\alpha}} = 1$. Thus, since $\phi(1) = 0$, we then have that

$$\begin{aligned}\mathcal{D}_\phi(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) &= \sum_{j\in[n]} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_j} \boldsymbol{\theta}^{\boldsymbol{\alpha}} \phi\left(\frac{\tilde{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}}\right) = \sum_{j\in[n]} \sum_{s\in[r_j]} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\theta_{j_s} \phi\left(\frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\tilde{\theta}_{j_s}}{\boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}\theta_{j_s}}\right) \\ &= \sum_{j\in[n]} \sum_{s\in[r_j]} \theta_{j_s} \phi\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}. \end{aligned} \tag{14}$$

Now notice that by construction $\sum_{\boldsymbol{\alpha}\in\mathbb{A}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}$ is equal for all $s \in [r_j]$ and is a number between zero and one that we denote w_j. This is because the probability distribution of any random vector $\boldsymbol{Y}$ can be written as

$$\begin{aligned}\Pr(\boldsymbol{Y} = \boldsymbol{y}) &= \Pr(Y_m = y_m|\boldsymbol{Y}_{[m-1]} = \boldsymbol{y}_{[m-1]})\Pr(Y_{m-1} = y_{m-1}|\boldsymbol{Y}_{[m-2]} = \boldsymbol{y}_{[m-2]}) \cdots \Pr(Y_1 = y_1) \\ &= \theta_{y_m \boldsymbol{y}_{[m-1]}}\theta_{y_{m-1}\boldsymbol{y}_{[m-2]}} \cdots \theta_{y_1}.\end{aligned}$$

Notice that, for instance, any parameter $\theta_{y_m\boldsymbol{y}_{[m-1]}}$, for any $y_m \in \mathbb{Y}_m$ but for a fixed $\boldsymbol{y}_{[m-1]} \in \mathbb{Y}_{[m-1]}$, is multiplied with the same linear combination of parameters in the interpolating polynomial. This linear combination of probabilities corresponds to the term $\sum_{\boldsymbol{\alpha}\in\mathbb{A}_{-j_s}} \boldsymbol{\theta}^{\boldsymbol{\alpha}}_{-j_s}$ in equation (14). A statistical model then simply imposes some equality constraints on these probabilities, but each parameter will still be multiplied by the same linear combination of probabilities. From this observation the result then follows. □

Example 6. To illustrate the form of ϕ-divergences in full single CPT analyses reported in equation (13), we now derive it for the KL divergence in Definition 9. Consider the BN model in Example 1 and suppose again that θ_{21} and θ_{22} are varied. The KL divergence can be computed as

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}}) = \theta_1(\theta_{111} + \theta_{211})\tilde{\theta}_{11} \log\left(\frac{\tilde{\theta}_{11}}{\theta_{11}}\right) + \theta_1(\theta_{121} + \theta_{221})\tilde{\theta}_{21} \log\left(\frac{\tilde{\theta}_{21}}{\theta_{21}}\right) \\ + \theta_1(\theta_{131} + \theta_{231})\tilde{\theta}_{31} \log\left(\frac{\tilde{\theta}_{31}}{\theta_{31}}\right) + \theta_2(\theta_{112} + \theta_{212})\tilde{\theta}_{12} \log\left(\frac{\tilde{\theta}_{12}}{\theta_{12}}\right) \\ + \theta_2(\theta_{122} + \theta_{222})\tilde{\theta}_{22} \log\left(\frac{\tilde{\theta}_{22}}{\theta_{22}}\right) + \theta_2(\theta_{132} + \theta_{232})\tilde{\theta}_{32} \log\left(\frac{\tilde{\theta}_{32}}{\theta_{32}}\right).\end{aligned}$$

We can notice in the above expressions that each element of $\mathcal{D}_{\phi}(\mathbb{P}^j_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}^j_{\boldsymbol{\theta}})$, i.e. $\tilde{\theta}_{lj} \log(\tilde{\theta}_{lj}/\theta_{lj})$, is multiplied by $\theta_j(\theta_{1lj} + \theta_{2lj})$ for $l \in [3]$ and $j \in [2]$. However, since $\theta_{1lj} + \theta_{2lj} = 1$ for every $l \in [3]$ and each $j \in [2]$, then every element in $\mathcal{D}_{\phi}(\mathbb{P}^i_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}^i_{\boldsymbol{\theta}})$ is multiplied by the same probability: in this specific example θ_j. Notice that the parameter θ_j, $j \in [2]$, corresponds to the marginal probability of the conditining parent configuration. This property is of course expected to hold since the underlying model is a BN [8].

The additional complexity of having to compute the constant terms w_j, $j \in [n]$, in equation (13) has limited the use of KL divergences, and more generally ϕ-divergences, in both practical and theoretical sensitivity investigations in discrete BNs. However, looking at probabilistic models from a polynomial point of view, we are able here to establish an additional strong theoretical justification for the use proportional covariation even in single full CPT analyses, since this also minimizes any ϕ-divergence.

Theorem 2. *Under the conditions of Proposition 4, $\mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ is minimized by the proportional covariation scheme.*

Proof. Since w_j in equation (13) is a positive constant, $\mathcal{D}_{\phi}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ is minimized if each $\mathcal{D}_{\phi}(\mathbb{P}^j_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}^j_{\boldsymbol{\theta}})$ attains its minimum. Fix a $j \in [n]$. We use the method of Lagrange multipliers to demonstrate that $\mathcal{D}_{\phi}(\mathbb{P}^j_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}^j_{\boldsymbol{\theta}})$ is minimized by proportional covariation, subject to the constraint that $\sum_{s \in [r_j]} \tilde{\theta}_{j_s} - 1 = 0$. Define

$$L = \sum_{s \in [r_j]} \theta_{j_s} \phi\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) - \lambda\left(\sum_{s \in [r_j]} \tilde{\theta}_{j_s} - 1\right).$$

Taking the first derivative of L with respect to $\tilde{\theta}_{j_s}$ and equating it to zero gives

$$\frac{\partial}{\partial \tilde{\theta}_{j_s}} L = \phi'\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) = \lambda,$$

where ϕ' denotes the derivative of ϕ. By inverting we then deduce that

$$\tilde{\theta}_{j_s} = \phi'(\lambda)^{-1}\theta_{j_s}. \tag{15}$$

Since equation (15) holds for every $s \in [r_j] \setminus \{j_i\}$ we have that

$$\sum_{s\in[r_j]\setminus\{j_i\}} \tilde{\theta}_{j_s} = \phi'(\lambda)^{-1} \sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s} \tag{16}$$

Now take the first partial derivative of L with respect to λ and equate it to zero. This gives

$$\frac{\partial}{\partial \lambda} L = \sum_{s\in[r_j]} \tilde{\theta}_{j_s} = 1 \quad \Longrightarrow \quad \sum_{s\in[r_j]\setminus\{j_i\}} \tilde{\theta}_{j_s} = 1 - \tilde{\theta}_{j_i} \tag{17}$$

Plugging the right hand side of (17) into (16), we deduce that

$$\phi'(\lambda)^{-1} = \frac{1-\tilde{\theta}_{j_i}}{\sum_{s\in[r_j]\setminus\{j_i\}} \theta_{j_s}} = \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}}. \tag{18}$$

Thus, by plugging (18) into (15) we conclude that

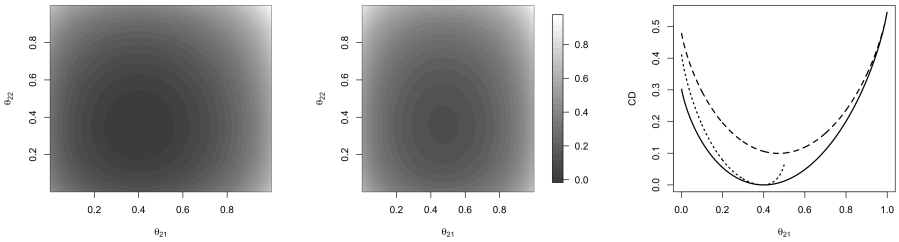
$$\tilde{\theta}_{j_s} = \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}}\theta_{j_s}.$$

This is guaranteed to be a minimum by the convexity of the function ϕ. □

Example 7. As an example of a ϕ-divergence, in Figure 5 we have plotted $\mathrm{KL}(\mathbb{P}_{\tilde{\boldsymbol{\theta}}}, \mathbb{P}_{\boldsymbol{\theta}})$ for Example 1 when θ_{21} (x-axis) and θ_{22} (y-axis in 5a and 5b) are varied for the covariation schemes so far considered. From Figures 5a and 5b we can see why the KL divergence under proportional covariation is smaller than in the uniform case. This becomes clearer when we only let θ_{21} vary as shown in Figure 5c since again the solid line, associated to proportional covariation, is always underneath the others.

5. Discussion

The polynomial representation of discrete statistical models based on the interpolating polynomial does not only represent an elegant characterization of a large class



(a) Proportional. (b) Uniform. (c) One-way analysis.

Figure 5: KL divergences for Example 7 under different covariation schemes: proportional (solid), uniform (dashed), order-preserving (dotted).

widely-used graphical models, it also provides a platform to answer a variety of sensitivity queries. Exploiting this representation we have been able to extend one-way sensitivity analysis results known for BNs only to a wide array of other models, for instance context-specific BNs and chain event graph amongst others. However, this technology allowed us to prove new optimality results of the proportional covariation operator for all multilinear models, including BNs. We showed that proportional covariation does not only minimize the CD distance in single full CPT sensitivity analyses, but also any divergence in the class of ϕ-divergences.

We notice that the flexibility of the interpolating polynomial representation might enable us to investigate even larger classes of models, for instance dynamic BNs, extending sensitivity methods to dynamic settings. The interpolating polynomial of such models is not necessarily multilinear. Preliminary results seem to suggest that in this framework both sensitivity functions and CD distances exhibit different properties than in the simpler multilinear case, with the potential of even more informative sensitivity investigations.

A different extension of this work would conversely look at more general multi-way sensitivity analyses, where varied parameters might appear in the same monomial of the interpolating polynomial. Intuitively, for such analyses sensitivity functions will not simply be multilinear as for single full CPT analyses, but also include interaction terms. Similarly, both CD distances and ϕ-divergences will be affected by such interactions.

For this reason the proportional covariation scheme for generic multi-way analyses might not be always optimal. However the polynomial representation of probabilities in BNs and related models gives us a promising starting point to start investigating this class of problems.

Acknowledgements

M. Leonelli was supported by Capes, C. Görgen was supported by the EPSRC grant EP/L505110/1 whilst J.Q. Smith was partly supported by EPSRC grant EP/K039628/1 and The Alan Turing Institute under EPSRC grant EP/N510129/1.

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, 28:131–142, 1966.

[2] J. H. Bolt and L. C. van der Gaag. Balanced tuning of multi-dimensional Bayesian network classifiers. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 210–220. Springer, 2015.

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

[4] R. Cano, C. Sordo, and J. M. Gutiérrez. Applications of Bayesian networks in meteorology. In *Advances in Bayesian Networks*, pages 309–328. Springer, 2004.

[5] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. Sensitivity analysis in discrete Bayesian networks. *IEEE T. Syst. Man Cyb.*, 27:412–423, 1997.

[6] H. Chan and A. Darwiche. When do numbers really matter? *J. Artificial Intelligence Res.*, 17:265–287, 2002.

[7] H. Chan and A. Darwiche. Sensitivity analysis in Bayesian networks: from single to multiple parameters. In *Proocedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 317–325, 2004.

[8] H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. *Internat. J. Approx. Reason.*, 38:149–174, 2005.

[9] H. Chan and A. Darwiche. Sensitivity analysis in Markov networks. In *Prooceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1300–1305, 2005.

[10] T. Charitos and L. C. van der Gaag. Sensitivity analysis of Markovian models. In *Proceedings of the FLAIRS Conference*, pages 806–811, 2006.

[11] S. H. Chen and C. A. Pollino. Good practice in Bayesian network modelling. *Environ. Modell. Softw.*, 37:134–145, 2012.

[12] R. Collazo, C. Görgen, and J. Q. Smith. *Chain event graphs*. Chapman Hall (to appear), 2017.

[13] V. M. H. Coupé and L. C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Ann. Math. Artif. Intell.*, 36:323–356, 2002.

[14] R. G. Cowell, A. P. Dawid, Lauritzen S. L., and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, New York, 1999.

[15] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:85–108, 1963.

[16] P. Dagum and E. Horvitz. A Bayesian analysis of simulation algorithms for inference in belief networks. *Networks*, 23:499–516, 1993.

[17] A. Darwiche. A differential approach to inference in Bayesian networks. *J. ACM*, 3:280–305, 2003.

[18] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Birkhäuser Verlag, Basel, 2009.

[19] S. French. Modelling, making inferences and making decisions: the roles of sensitivity analysis. *Top*, 11:229–251, 2003.

[20] M. A. Gómez-Villegas, P. Main, and R. Susi. Sensitivity analysis in Gaussian Bayesian networks using a divergence measure. *Comm. Statist. Theory Methods*, 36:523–539, 2007.

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

[22] C. Görgen and J. Q. Smith. Equivalence classes of staged trees. *Bernoulli (Forthcoming)*, 2017.

[23] 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.

[24] D. Heckerman, A. Mamdani, and M. P. Wellman. Real-world applications of Bayesian networks. *Commun. ACM*, 38:24–26, 1995.

[25] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A*, 186:453–461, 1946.

[26] M. I. Jordan. Graphical models. *Statist. Sci.*, 19:140–155, 2004.

[27] K. Korb and A. E. Nicholson. *Bayesian artificial intelligence*. CRC Press, Boca Raton, 2010.

[28] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.

[29] S. L. Lauritzen. *Graphical models*. Oxford University Press, Oxford, 1996.

[30] M. Neil, N. Fenton, and L. Nielson. Building large-scale Bayesian networks. *Knowl. Eng. Rev.*, 15:257–284, 2000.

[31] L. Pardo. *Statistical inference based on divergence measures*. CRC Press, Boca Raton, 2005.

[32] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan-Kauffman, San Francisco, 1988.

[33] G. Pistone, E. Riccomagno, and H. P. Wyn. Gröbner bases and factorisation in discrete probability and Bayes. *Stat. Comput.*, 11:37–46, 2001.

[34] J. Pitchforth and K. Mengersen. A proposed validation framework for expert elicited Bayesian networks. *Expert Syst. Appl.*, 40:162–167, 2013.

[35] E. Rajabally, P. Sen, S. Whittle, and J. Dalton. Aids to Bayesian belief network construction. In *Proceedings of the 2nd International Conference on Intelligence Systems*, pages 457–461, 2004.

[36] S. Renooij. Efficient sensitivity analysis in hidden Markov models. *Internat. J. Approx. Reason.*, 53:1397–1414, 2012.

[37] S. Renooij. Co-variation for sensitivity analysis in bayesian networks: properties, consequences and alternatives. *Internat. J. Approx. Reason.*, 55:1022–1042, 2014.

[38] E. Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.

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

[40] L. C. van der Gaag, S. Renooij, and V. M. H. Coupé. Sensitivity analysis of probabilistic networks. In *Advances in Probabilistic Graphical Models*, pages 103–124. Springer, 2007.