The curved exponential family of a staged tree

Christiane Görgen

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
e-mail: christiane.goergen@mis.mpg.de

Manuele Leonelli

School of Human Sciences and Technology, IE University, Madrid, Spain
e-mail: orlando.marigliano@mis.mpg.de

Orlando Marigliano

KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: orlando.marigliano@mis.mpg.de

Abstract: Staged tree models are a discrete generalization of Bayesian networks. We show that these form curved exponential families and derive their natural parameters, sufficient statistic, and cumulant-generating function as functions of their graphical representation. We give necessary graphical criteria for classifying regular subfamilies and discuss implications for model selection.

MSC2020 subject classifications: Primary 62H99; secondary 68T99.
Keywords and phrases: Bayesian Information Criterion, Chain Event Graph, Curved Exponential Family, Graphical Model, Staged Tree.

1. Introduction

Staged trees define statistical models which can account for a variety of partial and asymmetric conditional independence statements between discrete events (Collazo, Görgen and Smith, 2018). The use of these graphical models in applications is constantly increasing (Barclay, Hutton and Smith, 2013; Collazo and Smith, 2015; Keeble et al., 2017) and free software for practitioners is newly available (Carli et al., 2020). However, their formal properties have only recently been studied for the first time (Görgen and Smith, 2018). We now extend this formal study by proving that in general staged tree models form curved exponential families and by expressing their natural parameters, sufficient statistic and cumulant generating function as a function of the underlying graphical representation. We also give graphical conditions under which they form regular exponential families.

Because exponential families exhibit a multitude of desirable inferential properties (Kass and Vos, 1997), similar characterisations for other graphical models have been studied. For instance, undirected graphical models with no hidden variables are regular exponential families (Lauritzen, 1996), Bayesian networks (e.g. Koller and Friedman, 2009) are curved exponential families, and directed graphical models with hidden variables are stratified exponential families (Geiger et al., 2001). These results are critical for instance for model selection techniques. In particular, Haughton (1988) proved that for curved exponential families the Bayesian information criterion (Schwarz, 1978) is an asymptotically valid rule. For staged trees, to this date model selection has usually been carried out in a Bayesian fashion by selecting the maximum a posteriori model (Freeman and Smith, 2011; Barclay, Hutton and Smith, 2013; Cowell and Smith, 2014; Carli et al., 2020).

By expressing every staged tree model explicitly as a curved exponential family, we in particular achieve such an explicit description for every discrete Bayesian network, which to our knowledge is still missing in the literature. As for Bayesian networks, we can then specify graphical criteria under which the model is a regular exponential family. Furthermore, the results of this paper justify the use of the Bayesian information criterion for model selection. This has already been employed by Silander and Leong (2013) although its asymptotic geometric validity was not assured.

2. Staged Trees

2.1. Probability trees as graphical statistical models

Probability trees are highly intuitive depictions of unfoldings of discrete events (Shafer, 1996) which have been used in a variety of real-world applications (Smith, 2010; Collazo, Görgen and Smith, 2018).

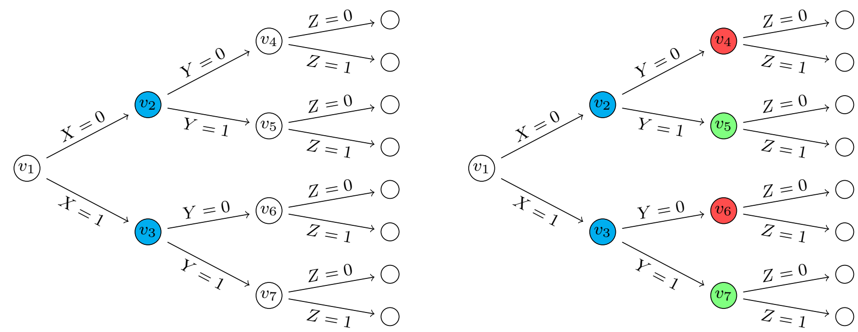
Let $\mathcal{T}$ be a directed, rooted tree graph where every vertex has either no or at least two emanating edges. For simplicity, we number the inner (non-leaf) vertices $v_1, \ldots, v_k$, with v_1 denoting the root, and count the edges emanating from each vertex with indices $j = 1, \ldots, \kappa_i$ for all $i = 1, \ldots, k$. To every edge we assign a positive probability $\theta_{ij} \in (0, 1)$ such that the total sum of all labels belonging to the same vertex is always equal to one, $\sum_{j=1}^{\kappa_i} \theta_{ij} = 1$. Such a labelled tree graph is called a *probability tree*.

The positivity assumption in probability trees precludes issues related to validity assumptions as present for instance in Bayesian networks. The sum-to-1 conditions on vertices ensure that the multiplication rule along root-to-leaf paths in a probability tree induces a well-defined probability distribution over the graph.

We write this distribution as $p_\theta(x) = \prod_{i=1}^{k} \prod_{j=1}^{\kappa_i} \theta_{ij}^{\alpha_{ij}(x)}$ where x denotes a root-to-leaf path in the tree, the index $\theta = (\theta_{ij} \mid j = 1 \ldots, \kappa_i; i = 1, \ldots, k)$ denotes the vector of all edge labels, and the exponent $\alpha_{ij}(x)$ is an indicator function which is equal to one if and only if x passes through the jth edge emanating from vertex v_i for $j = 1, \ldots, \kappa_i$ and $i = 1, \ldots, k$. The number of root-to-leaf paths, or single atom events, is denoted n. A *probability tree model* is the set of all probability distributions $\{p_\theta \mid \theta \in \Theta_{\mathcal{T}}\}$ which can be written in this form, for varying values of edge labels. The parameter space $\Theta_{\mathcal{T}} = \times_{i=1}^{k} \Delta^\circ_{\kappa_i - 1}$ of such a model is a $d = \sum_{i=1}^{k} (\kappa_i - 1)$-dimensional product of open $r-1$-dimensional probability simplices $\Delta^\circ_{r-1} = \{(t_1, \ldots, t_d) \in \mathbb{R}^d \mid \sum_{i=1}^{r} t_i = 1 \text{ and } 0 < t_i < 1 \text{ for all } i = 1, \ldots, r\}$, for some positive integers r.

2.2. Probability trees and conditional independence

An event in a probability tree is a set of root-to-leaf paths, often specified by shared vertices or edges. Conditional independence relationships between events can be visu-



(a) A staged tree for three binary random variables representing the Bayesian network $X \to Z \leftarrow Y$.

(b) A staged tree representing the conditional independence relation $X \perp\!\!\!\perp Y$ and $Z \perp\!\!\!\perp X \mid Y$.

FIG 1. *Two discrete graphical models. Vertices which are in the same stage have been assigned the same colour.*

alized using a simple type of colouring of these vertices in the following way.

A *staged tree* is a probability tree together with an equivalence relation on the vertex set such that two vertices are in the same stage if and only if their outgoing edges have the same attached probabilities. A *staged tree model* is a submodel of a probability tree model where probability simplices in the parameter space have been identified with each other according to the equivalence relation.

If a tree depicts the product state space of a vector of discrete random variables, then every root-to-leaf path corresponds to a single event in that space, every non-leaf vertex is associated to a random variable conditional on specific values taken by its ancestors, every edge corresponds to a state of that variable, and every edge can be labelled by the respective conditional probability of the random variable being in that state given the particular ancestor configuration. An identification of emanating edge labels in this case amounts to identifying rows of conditional probability tables. This is illustrated in Fig. 1(a) which shows a staged tree representation for the graphical model $X \to Z \leftarrow Y$ with binary random variables X, Y, Z.

As a consequence, staged tree models include as a special case both discrete Bayesian networks as well as, more generally, context-specific discrete Bayesian networks which may accommodate conditional independence relationships that hold only for a subset of all states of a discrete random variable (Boutilier et al., 1996).

Because probability trees and staged trees may have root-to-leaf paths of different lengths and because there are no constraints on where stages can be imposed, the class of staged tree models is much wider than the one of discrete Bayesian networks. Because staged trees grow quickly even for small problems, whenever they exhibit many symmetries in their subtrees they can further and more compactly be represented by alternative graphical depictions called *chain event graphs*: see the textbook by Collazo, Görgen and Smith (2018) for an in-depth discussion of these points.

3. Staged trees are curved exponential families

3.1. Exponential families

Following Kass and Vos (1997), a parametric statistical model $\{p_\theta \mid \theta \in \Theta\}$ on a space $\mathcal{X}$ is called an *exponential family* if every distribution in the model can be written in the form

$$p_\theta(x) = h(x)\exp\Big(\eta(\theta)^\top T(x) - \psi(\theta)\Big) \qquad \text{for all } x \in \mathcal{X} \tag{1}$$

where $\eta : \Theta \to \mathbb{R}^d$ is the canonical parameter, $T : \mathcal{X} \to \mathbb{R}^d$ is a minimally sufficient statistic, $^\top$ denotes the transpose operation, and $\psi : \Theta \to \mathbb{R}$ is the cumulant-generating function. The space $\mathcal{N} = \{\eta(\theta) \in \mathbb{R}^d \mid \text{(1) is integrable}\}$ is called the *natural parameter space*. If $\mathcal{N}$ is an open and non-empty subset of $\mathbb{R}^d$ then the model is called a *regular* exponential family of dimension d.

For exponential families, moments of all orders exist and the maximum likelihood estimator exists and is unique. Regular exponential families are closed under linear contraints on the natural parameter space and the log-likelihood function on that space is concave. Under more general constraints on the parameters, these families are more technical to study but can often retain many useful asymptotic properties. The two best-studied subclasses of regular exponential families are so-called curved and, more generally, stratified exponential families. The development of this paper requires only the former type of models which we formally introduce in the third subsection below.

3.2. Regular exponential families

In a first step, we focus on probability trees without imposing a stage structure. In particular, we call the probability tree whose graph is a root vertex connected to its leaves exclusively via single edges a *star*.

Lemma 1. *Every probability tree model on n atoms is equal to the full probability simplex Δ°_{n-1}.*

In particular, the star with n leaves and with probability θ_{1r} attached to edge r, for $r = 1, \dots, n$, represents the Multinomial distribution $\mathrm{Multi}(1, \theta)$ *with one trial and parameters $\theta = (\theta_{11}, \dots, \theta_{1n})$.*

Proof. Every probability tree with n root-to-leaf paths specifies a set of distributions inside the open $n-1$-dimensional probability simplex as outlined in Section 2.1. Conversely, given a fixed tree graph with n leaves, every point $(p_1, \dots, p_n)$ inside the simplex can be interpreted as a vector whose rth component p_r is the probability of going down the rth root-to-leaf path in the tree, $r = 1, \dots, n$. We can pick the label of the jth edge out of vertex v_i to be the fraction $\sum_{b\in[ij]} p_b / \sum_{a\in[i]} p_a$ where $[i], [ij] \subseteq \{1, \dots, n\}$ denote the indices of all root-to-leaf paths passing through vertex v_i or through the tail of its jth outgoing edge, respectively. These labels are conditional probabilities and their product along a root-to-leaf path is equal precisely to the atomic probability of that path. This proves the first claim.

As for the second claim, $p_\theta(x) = \prod_{r=1}^n \theta_{1r}^{\alpha_{1r}(x)}$ where $\sum_{r=1}^n \theta_{1r} = 1$ is the probability distribution induced by the star with edges numbered $r = 1, \dots, n$. This is the stated Multinomial distribution. $\square$

Wishart (1949) derives the exponential form (1) for the Multinomial distribution $\mathrm{Multi}(1, \theta)$ as follows: $h \equiv 1$ is constant, the natural parameters are normalized log-probabilities $\eta_r(\theta) = \log(\theta_{1r}/1-\sum_{s=1}^{n-1}\theta_{1s})$ for $r = 1, \dots, n$, the sufficient statistic is the vector of the first $n-1$ edge indicators $T = (\alpha_{11}, \dots, \alpha_{1,n-1})$, and the cumulant-generating function is the logarithm of the normalizing constant $\psi(\theta) = \log(1 - \sum_{s=1}^{n-1}\theta_{1s})$.

By Lemma 1, all probability trees on the same number of root-to-leaf paths are statistically equivalent in the sense that they all represent the same model, namely the full probability simplex. Thus, *any* probability tree is a graphical representation of the Multinomial distribution, and Wishart's result gives a sufficient statistic, cumulant-generating function, and the natural parameters for any probability tree, possibly after reparametrization.

However and centrally for this paper, it is often useful to retain the extra structure of a probability tree with root-to-leaf paths which are longer than single edges. The reason for this is threefold. First and foremost, there is not necessarily an underlying product state space and there can be a very asymmetric stage structure but the staged tree is always an immediate graphical representation of all modelling assumptions: of the space of events, intricate conditional independence relationships within this space, and local sum-to-one conditions. Second, in applications staged trees can describe the modelled space of events in far more detail than more widely known graphical models such as Bayesian networks (Shafer, 1996; Collazo, Görgen and Smith, 2018). And third, this extra graphical structure opens the door to employ techniques from algebraic geometry for the better understanding of the properties of these models (Duarte and Görgen, 2020; Görgen et al., 2018, and references therein).

In the general case, we thus derive the following result. Here for any probability tree $\mathcal{T}$ we say that the κ_ith edge emanating from vertex v_i points *downwards*, $i = 1, \dots, k$. We then recursively define functions $N_i : \Theta_\mathcal{T} \to \mathbb{R}$ as $N_i(\theta) = 1$ for leaf vertices v_i and else as a product of labels pointing downwards $N_i(\theta) = (1 - \sum_{s=1}^{\kappa_i - 1}\theta_{is})N_{\kappa_i}(\theta)$ for $i = 1, \dots, k$. The shorthand N_{ij} denotes N_r for the vertex v_r which is the tail of the jth edge coming out of vertex v_i.

Proposition 1. *Every probability tree $\mathcal{T}$ with parameters θ represents a regular exponential family where:*

- *the indicators $T_{ij} = \alpha_{ij}$ of the first $j = 1, \dots, \kappa_i - 1$ edges of all inner vertices $i = 1, \dots, k$ are a sufficient statistic,*
- *the natural parameters η_{ij} are locally normalized log-probabilities defined by $\eta_{ij}(\theta) = \log(\theta_{ij}{N_{ij}(\theta)}/{N_i(\theta)})$ for all $j = 1, \dots, \kappa_i - 1$ and $i = 1, \dots, k$, and the natural parameter space is $\mathbb{R}^d$ with $d = \sum_{i=1}^k(\kappa_i - 1)$, and*
- *the cumulant-generating function is the negative log-sum of normalizing constants along the root-to-leaf path whose edges all point downwards, $\psi(\theta) = -\log(N_1(\theta))$.*

Proof. The desired parametrization is equivalent to

$$p_\theta(x) = N_1(\theta) \prod_{i=1}^{k} \prod_{j=1}^{\kappa_i - 1} (\theta_{ij}{}^{N_{ij}(\theta)}/N_i(\theta))^{\alpha_{ij}(x)} \quad \text{for all } x. \tag{2}$$

To prove this equality, we first rewrite the probability mass function introduced in Section 2. For any inner vertex $i = 1, \ldots, k$ in the probability tree, the label of the κ_ith outgoing edge is a function of the first $1, \ldots, \kappa_i - 1$ edges, namely $\theta_{i\kappa_i} = 1 - \sum_{j=1}^{\kappa_i - 1} \theta_{ij}$. The indicator $\alpha_{i\kappa_i}$ of passing along that edge is a function of the indicators of those same edges and equals $\alpha_i(x)(1 - \sum_{j=1}^{\kappa_i - 1} \alpha_{ij}(x))$ where $\alpha_i(x)$ is one if x reaches vertex v_i and zero otherwise.

Thus, the probability mass function of a probability tree can be written as

$$p_\theta(x) = \prod_{i=1}^{k} \prod_{j=1}^{\kappa_i} \theta_{ij}^{\alpha_{ij}(x)} = \prod_{i=1}^{k} \prod_{j=1}^{\kappa_i - 1} \theta_{ij}^{\alpha_{ij}(x)} \left(1 - \sum_{j=1}^{\kappa_i - 1} \theta_{ij}\right)^{\alpha_i(x)(1 - \sum_{j=1}^{\kappa_i - 1} \alpha_{ij}(x))}.$$

And as a consequence, the second claim reduces to:

$$\prod_{i=1}^{k} \left(1 - \sum_{j=1}^{\kappa_i - 1} \theta_{ij}\right)^{\alpha_i(x)(1 - \sum_{j=1}^{\kappa_i - 1} \alpha_{ij}(x))} = N_1(\theta) \prod_{i=1}^{k} \prod_{j=1}^{\kappa_i - 1} \left(\frac{N_{ij}(\theta)}{N_i(\theta)}\right)^{\alpha_{ij}(x)} \tag{3}$$

for all root-to-leaf paths x. We now prove this claim by induction on the structure of the tree.

Let thus in a first step the graph be a star, $k = 1$. If x is the root-to-leaf path which is one of the first $1, \ldots, \kappa_1 - 1$ edges coming out of the root then (3) states that $1 = N_1(\theta) \cdot 1/N_1(\theta) = 1$. If otherwise x equals the κ_1st edge then $1 - \sum_{j=1}^{\kappa_1 - 1} \theta_{1j} = N_1(\theta)$ which is also true.

If $k > 1$ then x may have length greater one and we distinguish two analogous cases. Hereby the induction hypotheses holds for trees smaller than the one we consider and so, in particular, the claim is true for its subtrees. Without loss, we number the vertex at the tail of the first edge of x as v_2. For simplicity, A denotes the left hand side of (3) and B the right hand side of that equation. Then:

Case 1: The edge (v_1, v_2) is one of the first $1, \ldots, \kappa_1 - 1$ edges coming out of the root. Then

$$\begin{aligned}
A &= 1 \cdot \prod_{i=2}^{k} \left(1 - \sum_{j=1}^{\kappa_i - 1} \theta_{ij}\right)^{\alpha_i(x)(1 - \sum_{j=1}^{\kappa_i - 1} \alpha_{ij}(x))} \\
&= N_2(\theta) \prod_{i=2}^{k} \prod_{j=1}^{\kappa_i - 1} \left(\frac{N_{ij}(\theta)}{N_i(\theta)}\right)^{\alpha_{ij}(x)} \\
&= N_1(\theta) \frac{N_2(\theta)}{N_1(\theta)} \prod_{i=2}^{k} \prod_{j=1}^{\kappa_i - 1} \left(\frac{N_{ij}(\theta)}{N_i(\theta)}\right)^{\alpha_{ij}(x)} = B
\end{aligned}$$

where the final step is true because $N_2(\theta) = N_{12}(\theta)$.

Case 2: The edge (v_1, v_2) is the κ_1st edge coming out of the root. Then

$$\begin{aligned} A &= \left(1 - \sum_{j=1}^{\kappa_1 - 1} \theta_{1j}\right) \prod_{i=2}^{k} \left(1 - \sum_{j=1}^{\kappa_i - 1} \theta_{ij}\right)^{\alpha_i(x)(1 - \sum_{j=1}^{\kappa_i - 1} \alpha_{ij}(x))} \\ &= \left(1 - \sum_{j=1}^{\kappa_1 - 1} \theta_{1j}\right) N_2(\theta) \prod_{i=2}^{k} \prod_{j=1}^{\kappa_i - 1} \left(\frac{N_{ij}(\theta)}{N_i(\theta)}\right)^{\alpha_{ij}(x)} \\ &= N_1(\theta) \prod_{i=2}^{k} \prod_{j=1}^{\kappa_i - 1} \left(\frac{N_{ij}(\theta)}{N_i(\theta)}\right)^{\alpha_{ij}(x)} = B \end{aligned}$$

where the final step is true because $\prod_{j=1}^{\kappa_1 - 1} \left({N_{1j}(\theta)}/{N_1(\theta)}\right)^{\alpha_{1j}(x)} = 1$.

This proves (2), and every probability tree represents an exponential family in this parametrization.

The regularity claim follows since the map η is a diffeomorphism between the space of model parameters Θ and $\mathbb{R}^d$. Indeed, it has a smooth inverse obtained by first computing p_η for a given $\eta \in \mathbb{R}^d$ according to the given parametrization and then computing $\theta \in \Theta$ such that $p_\eta = p_\theta$ as described in the proof of Lemma 1. $\square$

Proposition 1 enables us to simply read the parametrisation of the underlying exponential family directly from a given probability tree.

Example 1. *Consider Fig. 1 and the probability tree which has the same graph as the staged trees in Fig. 1(a) and 1(b). For simplicity, in this binary tree we do not use double indices but label the upwards edges going out of vertex v_i as θ_i and the downward edges as $1 - \theta_i$ for $i = 1, \dots, 7$.*

The indicator functions α_{i1} of the upwards edges $i = 1, \dots, 7$ are a sufficient statistic for this tree. The corresponding natural parameters are then log-ratios of the downwards labels derived as $\eta_1(\theta) = \log(\theta_1{}^{(1-\theta_2)(1-\theta_5)}/{}_{(1-\theta_1)(1-\theta_3)(1-\theta_7)})$, $\eta_2(\theta) = \log(\theta_2{}^{(1-\theta_4)}/{}_{(1-\theta_2)(1-\theta_5)})$, $\eta_3(\theta) = \log(\theta_3{}^{(1-\theta_6)}/{}_{(1-\theta_3)(1-\theta_7)})$, and $\eta_i(\theta) = \log({}^{\theta_i}/{}_{1-\theta_i})$ for $i = 4, \dots, 7$. And the cumulant-generating function is the logarithm of the product of the labels along the downward root-to-leaf path coming out of the root vertex $\psi(\theta) = -\log\left[(1 - \theta_1)(1 - \theta_3)(1 - \theta_7)\right]$.

3.3. Curved exponential families

As stated in Section 2.2, every discrete Bayesian network has a corresponding staged tree representation. Since for instance the collider graph does not represent a regular exponential family (Koller and Friedman, 2009), staged trees cannot in general be regular exponential families either. They rather form what is called a *curved* exponential family: a submodel of a regular exponential family whose parameter space is a smooth manifold (Efron, 1978).

Theorem 1. *Staged tree models are curved exponential families.*

Proof. By Proposition 1, every staged tree model is a submodel of a regular exponential family. We prove that the natural parameter space is always a smooth manifold of the right dimension by showing it is the image of a certain linear subspace of Θ under the diffeomorphism η.

Let $\mathcal{T}_0$ denote a probability tree with parameter space $\Theta_0 = \times_{i=1}^{k}\Delta^{\circ}_{\kappa_i-1}$. Let $\mathcal{T}$ denote the same tree graph together with an imposed stage structure and parameter space $\Theta_{\mathcal{T}} = \times_{r\in R}\Delta^{\circ}_{\kappa_r-1}$ for some index set $R \subseteq \{1,\ldots,k\}$. The parameter space of the staged tree model is the kernel of the parameter space of the saturated model under the linear function $h_{\mathcal{T}} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_0-d}$ which encodes the $d_0 - d$ identifications of edge labels as

$$h_{\mathcal{T}}(\theta) = \big(\theta_{ij} - \theta_{st} \;\mid\; \text{for all } j = t = 1,\ldots,\kappa_j \text{ and all } v_i \text{ and } v_s \text{ in the same stage}\big) \tag{4}$$

where $d = \sum_{r\in R}(\kappa_r - 1)$ is the number of free parameters in the staged tree, and $d_0 = \sum_{i=1}^{k}(\kappa_i - 1)$ is the number of free parameters in the probability tree. By construction, $h_{\mathcal{T}}(\theta) = 0$ if and only if θ fulfills the stage constraints in $\mathcal{T}$, or rather, the parameter space of the staged tree model equals the kernel of the map (4), so $h_{\mathcal{T}}^{-1}(0) = \Theta_{\mathcal{T}}$.

Since $h_{\mathcal{T}}$ is surjective, its kernel is d-dimensional. Thus, the parameter space of the staged tree model is a d-dimensional linear space in $\mathbb{R}^{d_0}$. As a consequence, the natural parameter space, obtained as the image of the diffeomorphism η given in Proposition 1, is a smooth manifold of the same dimension. The claim follows. □

As an alternative proof strategy for Theorem 1 one may use the characterization of staged tree models as the solution set of a collection of polynomial equations and inequalities provided by Duarte and Görgen (2020). Whenever these polynomials do not exhibit any algebraic singularities inside the probability simplex, the model is a curved exponential family: compare the discussion of implicit model representations given in Geiger et al. (2001) and of algebraic exponential families in Drton and Sullivant (2007).

In particular, whilst the equations coding stage constraints in the conditional probabilities as in (4) are always linear, those coding the same constraints formulated in terms of the natural parameters in general are not. We can however state the following graphical results.

Proposition 2. *Using the parametrization given in Proposition 1, two vertices v_i and v_s, for some $i, s \in \{1,\ldots,k\}$, are in the same stage if and only if for all $j, t = 1,\ldots,\kappa_i$:*

$$\eta_{ij}(\theta)\eta_{st}(\theta)P_{ij}(\theta)P_{st}(\theta) = \eta_{sj}(\theta)\eta_{it}(\theta)P_{sj}(\theta)P_{it}(\theta) \tag{5}$$

where $P_{ij}(\theta) = \sum_{r\in[>ij]}\prod_{ab}\eta_{ab}(\theta)^{\alpha_{ab}(r)}$ is the sum of all (unnormalized) probabilities in the subtree rooted at the tail of the jth edge coming out of vertex i, $[> ij] \subseteq \{1,\ldots,n\}$ denotes the respective index set, and $\eta_{i\kappa_i}$ is defined to be 1.

In either of the following cases, (5) is a system of linear equations:

- *if the downwards pointing paths in the subtrees rooted at v_i and v_s have the same labels, in the sense that $N_i(\theta) = N_s(\theta)$ and $N_{ij}(\theta) = N_{sj}(\theta)$ for all $j = 1,\ldots,\kappa_i$,*
- *if the staged tree is* simple*: that is, if for any two vertices in the same stage their outgoing subtrees represent the same statistical model; or equivalently if $P_{ij}(\theta) = P_{sj}(\theta)$ for all $j = 1,\ldots,\kappa_i$,*

- *if the staged tree is* balanced*: that is, if the polynomial condition $P_{ij}(\theta)P_{st}(\theta) = P_{sj}(\theta)P_{it}(\theta)$ is satisfied for all $j, t = 1, \dots, \kappa_i$, or*
- *if the staged tree has an underlying Bayesian network which is decomposable.*

Proof. Let v_i and v_s be vertices in the same stage, for some $i, s \in \{1, \dots, k\}$. Duarte and Görgen (2020) show that their outgoing labels θ_{ij} and θ_{sj} are equal for all $j = 1, \dots, \kappa_i$ if and only if the equation

$$p_{[ij]}p_{[st]} = p_{[sj]}p_{[it]} \tag{6}$$

is true for all $j, t = 1, \dots, \kappa_i$. Here, we use shorthand notation for the sum $p_{[ij]} = \sum_{r\in[ij]} p_r$ of atomic probabilities of root-to-leaf paths passing through the jth edge coming out of vertex i, and for the respective index set $[ij] \subseteq \{1, \dots, n\}$. This result is formulated in terms of atomic probabilities and is true independently of the choice of parametrization. Now if we parametrize the staged tree using Proposition 1, all summands in $p_{[ij]}$ share the factor $N_1(\theta)$ as well as the labels $\eta_\bullet(\theta)$ appearing along the root-to-v_i path in the tree. Because the analogous observation is true for $p_{[st]}, p_{[sj]}, p_{[it]}$, these factors cancel out in (6). This equation thus simplifies to the statement in (5).

The four itemized criteria are, respectively: a straightforward result of Proposition 1, proven in Duarte and Görgen (2020), proven in Duarte and Ananiadi (2020), and proven by Geiger, Meek and Sturmfels (2006). □

In natural language, we now know: first that the stage constraints translate into crossproducts and sums of natural parameters, and second that a very regular stage structure on a probability tree preserves the regularity of the probability tree model.

In particular, staged trees represent regular exponential families if they give rise to a class of chain event graphs with a very compact graphical representation (Collazo, Görgen and Smith, 2018), if they fulfill a computable polynomial or geometric criterion (Duarte and Görgen, 2020), or if they belong to the well-known class of decomposable Bayesian networks which is amenable to fast computations (Koller and Friedman, 2009).

For curved exponential families, the equations on the natural parameters are non-trivial and have to the best of the authors' knowledge not been explicitly derived in the literature of Bayesian networks or undirected graphical models. The formulation we found in this paper shows that they are functions of a staged tree's subgraphs for every conditional independence relation and has only been possible thanks to the probability tree's expressiveness of the underlying space of events.

Example 2. *Figure 1(a) is a staged tree representation of the collider Bayesian network $X \to Z \leftarrow Y$ which is not a regular exponential family. We can see here that indeed in our parametrization, the linear stage identifications $\theta_2 = \theta_3$ do not give rise to linear constraints on the natural parameters $\eta_2(\theta)$ and $\eta_3(\theta)$. Instead, $\eta_2(1+\eta_4)(1+\eta_7) = \eta_3(1+\eta_5)(1+\eta_6)$ by Proposition 2.*

The staged tree in Fig. 1(b) however fulfills all of the criteria listed in Proposition 2. In particular, because $\eta_4 = \eta_6$ and $\eta_5 = \eta_7$, the above equation also simplifies to $\eta_2 = \eta_3$. Its linear stage constraints $\theta_2 = \theta_3$, $\theta_4 = \theta_6$, and $\theta_5 = \theta_7$ thus give rise to linear (equality) constraints on the natural parameters. The same simplification would *occur were* v_4 *and* v_5*, and* v_6 *and* v_7 *in the same stage, respectively, rather than* v_4 *and* v_6*, and* v_5 *and* v_7*.*

Acknowledgement

The authors are grateful to Giovanni Pistone and to Piotr Zwiernik for discussions during the early stages of this project and to Eva Riccomagno for comments on an earlier version of this paper.

Christiane Görgen and Manuele Leonelli were supported by the programme “Oberwolfach Leibniz Fellows” of the Mathematisches Forschungsinstitut Oberwolfach in 2017. Orlando Marigliano was supported by the International Max Planck Research School.

References

BARCLAY, L. M., HUTTON, J. L. and SMITH, J. Q. (2013). Refining a Bayesian network using a chain event graph. *International Journal of Approximate Reasoning* **54** 1300 - 1309.

BOUTILIER, C., FRIEDMAN, N., GOLDSZMIDT, M. and KOLLER, D. (1996). Context-specific independence in Bayesian networks. In *12th Conference on Uncertainty in Artificial Intelligence (UAI 96)* (E. HORVITZ and F. JENSEN, eds.). *Uncertainty in Artificial Intelligence* 115 - 123. Morgan Kaufmann Publishers Inc., San Francisco.

CARLI, F., LEONELLI, M., RICCOMAGNO, E. and VARANDO, G. (2020). The R Package `stagedtrees` for Structural Learning of Stratified Staged Trees. Preprint available from `arXiv:2004.06459 [stat.ME]`.

COLLAZO, R. A., GÖRGEN, C. and SMITH, J. Q. (2018). *Chain Event Graphs*. *Computer Science & Data Analysis Series*. Chapman & Hall.

COLLAZO, R. A. and SMITH, J. Q. (2015). A new family of non-local priors for chain event graph model selection. *Bayesian Analysis* **11** 1165 - 1201.

COWELL, R. G. and SMITH, J. Q. (2014). Causal discovery through MAP selection of stratified chain event graphs. *Electronic Journal of Statistics* **8** 965 - 997.

DRTON, M. and SULLIVANT, S. (2007). Algebraic statistical models. *Statist. Sinica* **17** 1273–1297. MR2398596

DUARTE, E. and ANANIADI, L. (2020). Gröbner bases for staged trees. Preprint available from `arXiv:1910.02721 [math.AC]`.

DUARTE, E. and GÖRGEN, C. (2020). Equations defining probability tree models. *Journal of Symbolic Computation* **99** 127–146.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

FREEMAN, G. and SMITH, J. Q. (2011). Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis* **102** 1152-1165.

GEIGER, D., MEEK, C. and STURMFELS, B. (2006). On the toric algebra of graphical models. *Ann. Statist.* **34** 1463–1492. MR2278364 (2007m:60035)

GEIGER, D., HECKERMAN, D., KING, H. and MEEK, C. (2001). Stratified exponential families: graphical models and model selection. *Ann. Statist.* **29** 505 - 529.

GÖRGEN, C. and SMITH, J. Q. (2018). Equivalence Classes of Staged Trees. *Bernoulli* **24** 2676–2692.

GÖRGEN, C., BIGATTI, A., RICCOMAGNO, E. and SMITH, J. Q. (2018). Discovery of statistical equivalence classes using computer algebra. *Internat. J. Approx. Reason.* **95** 167–184. MR3772779

HAUGHTON, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16** 342–355.

KASS, R. E. and VOS, P. W. (1997). *Geometrical foundations of asymptotic inference*. *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley & Sons, Inc., New York A Wiley-Interscience Publication. MR1461540

KEEBLE, C., THWAITES, P. A., BAXTER, P. D., BARBER, S., PARSLOW, R. C. and LAW, G. R. (2017). Learning through chain event graphs: the role of maternal factors in childhood type I diabetes. *American Journal of Epidemiology* **186** 1204–1208.

KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic graphical models*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA Principles and techniques. MR2778120

LAURITZEN, S. L. (1996). *Graphical models*. Clarendon Press, Oxford.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6** 461–464.

SHAFER, G. (1996). *The Art of causal Conjecture*. *Artificial Management*. MIT Press, Cambridge.

SILANDER, T. and LEONG, T. Y. (2013). A Dynamic Programming Algorithm for Learning Chain Event Graphs. In *Discovery Science*, (J. Fürnkranz, E. Hüllermeier and T. Higuchi, eds.). *Lecture Notes in Computer Science* **8140** 201–216. Springer Berlin Heidelberg.

SMITH, J. Q. (2010). *Bayesian decision analysis: principles and practice*. Cambridge University Press, Cambridge; New York.

WISHART, J. (1949). Cumulants of multivariate multinomial distributions. *Biometrika* **36** 47–58. MR0033996