KYBERNETIKA — VOLUME 49 (2013), NUMBER 1, PAGES 23–39

MIXTURE DECOMPOSITIONS OF EXPONENTIAL FAMILIES USING A DECOMPOSITION OF THEIR SAMPLE SPACES

Guido F. Montúfar

We study the problem of finding the smallest m such that every element of an exponential family can be written as a mixture of m elements of another exponential family. We propose an approach based on coverings and packings of the face lattice of the corresponding convex support polytopes and results from coding theory. We show that $m = q^{N-1}$ is the smallest number for which any distribution of N q-ary variables can be written as mixture of m independent q-ary variables. Furthermore, we show that any distribution of N binary variables is a mixture of $m = 2^{N-(k+1)}(1 + 1/(2^k - 1))$ elements of the k-interaction exponential family.

Keywords: mixture model, non-negative tensor rank, perfect code, marginal polytope

Classification: 52B05, 60C05, 62E17

1. INTRODUCTION

The m-mixture of a set of probability distributions $\mathcal{M}$ is the set of all possible convex combinations of m of its points:

$$\mathrm{Mixt}^m(\mathcal{M}) := \Big\{ \sum_{j=1}^m \alpha_j p_j \;\Big|\; p_j \in \mathcal{M},\ \alpha_j \geq 0 \ \text{ for } j \in \{1,\dots,m\},\ \text{and } \sum_{j=1}^m \alpha_j = 1 \Big\} .$$

The numbers $\alpha_j \in \mathbb{R}_{\geq 0}$ are called mixture weights and the summands p_j mixture components. There is an abundant literature on mixture models, see [5, 25, 26, 33]. They arise within probabilistic models that involve latent variables, see for instance [27, 28]. An exponential family on a finite set $\mathcal{X}$, with sufficient statistics $A \in \mathbb{R}^{d\times\mathcal{X}}$ and reference measure $\nu \in \mathbb{R}^{\mathcal{X}}_{>0}$, is the set of probability distributions p_θ, parametrized by $\theta \in \mathbb{R}^d$, of the form

$$p_\theta(x) = \frac{1}{Z_\theta}\nu(x)\exp(\theta^\top A_x) \quad \forall\, x \in \mathcal{X} ,$$

where A_x, $x \in \mathcal{X}$ are the columns of A, and $Z_\theta = \sum_{y\in\mathcal{X}} \nu(y)\exp(\theta^\top A_y)$ is the partition function. See Section 2 for details and [2, 7, 10] for standard references. We consider the following problem:

Problem 1. Given two exponential families $\mathcal{E}$ and $\mathcal{E}'$ on a finite set $\mathcal{X}$, find the smallest natural number $m = m(\mathcal{E}, \mathcal{E}')$, if there is any, for which $\mathrm{Mixt}^m(\mathcal{E}) \supseteq \mathcal{E}'$.

We propose a general approach based on coverings and packings of support sets of probability distributions, combinatorics of convex polytopes, and results from coding theory. We give explicit solutions when $\mathcal{E}$ is the independence model of N finite valued random variables, or a k-interaction exponential family, expressed in terms of the number of variables and the cardinality of their state spaces. When $\mathcal{E}'$ is equal to the convex hull of $\mathcal{E}$, for instance equal to the set $\mathcal{P}$ of strictly positive probability distributions on $\mathcal{X}$, then $m(\mathcal{E},\mathcal{E}')$ is the *Carathéodory number* of $\mathcal{E}$. We address Problem 1 for closures of exponential families as well. The closure of a statistical model $\mathcal{M}$, in the standard topology of the real valued functions, is denoted by $\overline{\mathcal{M}}$. When $\overline{\mathcal{E}}$ is the set of product distributions of N random variables, then $m(\overline{\mathcal{E}},\mathcal{E}')$ is the maximal non-negative outer-product rank of the N-way tables of probabilities described by $\mathcal{E}'$. Problem 1 can be thought of as a tensor decomposition problem.

The problem of representing probability distributions as mixtures of specific models has a long record. A renowned result in this direction is de Finetti's theorem, which states that exchangeable sequences of Bernoulli (i. e., binary) variables, are mixtures of independent and identically distributed Bernoulli variables, see [9, 24]. In general, the expressive power of mixture models is not satisfactorily understood. Until recently it was a long standing problem whether the m-mixture of the set of probability distributions of n independent binary variables had the dimension expected from parameter counting, which is $\min\{n\cdot m+(m-1), 2^n-1\}$. M. Catalisano, A. Geramita, and A. Gimigliano [8] proved that this mixture model indeed has the expected dimension for any combination of m and n, except for $n=4$ and $m=3$ when the dimension is smaller. In connection with this, the identifiability of parameters of mixtures of independent binary variables has been treated, for example in [6]. For mixtures of independent non-binary variables the dimension and parameter identifiability problems are largely unsettled.

When $\overline{\mathcal{E}}$ is the set of probability distributions of two independent variables with values in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, it is known that $\text{Mixt}^m(\overline{\mathcal{E}})$ equals the set $\overline{\mathcal{P}}$ of all possible probability distributions (on $\mathcal{X}_1\times\mathcal{X}_2$) as soon as $m\geq\min\{|\mathcal{X}_1|,|\mathcal{X}_2|\}$, see [16, 31]. This is to say that every non-negative $k\times l$ matrix can be written as the sum of at most $\min\{k,l\}$ non-negative rank-one matrices. If $|\mathcal{X}_1|,|\mathcal{X}_2|>2$, it is known that $\text{Mixt}^2(\overline{\mathcal{E}})\neq\overline{\mathcal{P}}$, see [16]. We generalize these results in Theorem 13:

The smallest m for which any probability distribution on $\{1,\dots,q\}^N$ can be written as the mixture of m product distributions, is q^{N-1} (when q is a prime power).

The result q^{N-1} is larger than expected from naïve parameter counting. In particular, the m-mixture model of $N\geq 5$ independent binary variables has the same dimension as $\mathcal{P}$ whenever $m\geq 2^N/(N+1)$. The m-mixture of a k-interaction model can be viewed as a system of stochastic units including higher-order interactions and a hidden m-valued variable. We show (Theorem 16):

The smallest m for which any probability distribution on $\{0,1\}^N$ can be represented as the mixture of m distributions from the k-interaction model is at most $2^{N-(k+1)}(1+\frac{1}{(2^k-1)})$.

We provide similar, however weaker, results when the variables are not binary, but take values in arbitrary finite sets. We also give a bound on the smallest number of mixtures

of independent binary distributions needed to represent k-interaction models.

Our proofs are based on comparison of the support sets of probability distributions contained in the closures of different exponential families. The support of a probability distribution p is the set $\mathrm{supp}(p) := \{x \in \mathcal{X} \colon p(x) > 0\}$. Combinatorial aspects of support sets of closures of exponential families have been studied in [20, 21, 22, 30]. We add to this analysis and put forward the analysis of a special type of support sets:

Definition 2. Given a set of probability distributions $\mathcal{M}$ on a finite set $\mathcal{X}$ we call $\mathcal{Y} \subseteq \mathcal{X}$ an *S-set* of $\mathcal{M}$ iff every probability distribution p with support $\mathrm{supp}(p) \subseteq \mathcal{Y}$, is contained in $\overline{\mathcal{M}}$.

The "S" in this definition stands for "support" and "simplex", considering that the set of probability distributions p with $\mathrm{supp}(p) \subseteq \mathcal{Y}$ is a simplex (the convex hull of affinely independent points in Euclidian space). A probability distribution p can be decomposed as a mixture of m probability distributions from the closure of $\mathcal{M}$ whenever the support of p is contained in the union of m S-sets of $\mathcal{M}$. This gives rise to the problem: *Given an exponential family on $\mathcal{X}$, find the smallest possible collection of S-sets that covers $\mathcal{X}$.*

In Section 2 we review basics of exponential families, their support sets, and convex supports. Section 3 formalizes our approach and discusses S-sets of exponential families. Section 4 treats coverings and packings using support sets of independence models and k-interaction families, and contains our solutions to Problem 1 for these models. Technical proofs are displaced to the Appendix.

2. EXPONENTIAL FAMILIES AND CONVEX SUPPORTS

We consider a system of $N \in \mathbb{N}$ random variables X_i with values in finite sets $\mathcal{X}_i$ for $i \in [N] := \{1, \ldots, N\}$. The joint *sample space* of this system is $\mathcal{X} := \times_{i=1}^{N} \mathcal{X}_i$. The probability distributions with support $\mathcal{Y} \subseteq \mathcal{X}$ are denoted by $\mathcal{P}(\mathcal{Y})$, or just by $\mathcal{P}$ if $\mathcal{Y} = \mathcal{X}$ is clear. The closure $\overline{\mathcal{P}}(\mathcal{Y})$ is the set of all probability distributions p with $\mathrm{supp}(p) \subseteq \mathcal{Y}$ and is called the probability simplex on $\mathcal{Y}$. The variable X_i is called q-ary when $|\mathcal{X}_i| = q$. For a subset of indices $\lambda \subseteq [N]$, x_λ denotes an element of $\times_{i \in \lambda} \mathcal{X}_i$, or the natural restriction of some $x \in \mathcal{X}$ to the coordinates $i \in \lambda$. The expression $[x_\lambda]$ represents a *cylinder set* of dimension $(N - |\lambda|)$, defined as the set of all $y \in \mathcal{X}$ with $y_\lambda = x_\lambda$. In the binary case $\mathcal{X} = \{0,1\}^N$ the cylinder sets and the (sets of vertices of) faces of the N-dimensional unit cube $[0,1]^N$ are in natural correspondence.

Consider a strictly positive function ν on $\mathcal{X}$, and a linear subspace V of the space $\mathbb{R}^{\mathcal{X}}$ of real valued functions on $\mathcal{X}$. The exponential family $\mathcal{E}_{\nu,V}$ is defined as the image of $V \to \mathcal{P} \subset \mathbb{R}^{\mathcal{X}}$; $f \mapsto \nu \exp(f) / \sum_{x \in \mathcal{X}} \nu(x) \exp(f(x))$. For simplicity we set $\nu \equiv 1$ and omit the subscript, as the results contained in this paper hold for any strictly positive ν. A matrix $A \in \mathbb{R}^{d \times \mathcal{X}}$ with row span V is called a *sufficient statistics* of $\mathcal{E}_V$. The rows of A are functions on $\mathcal{X}$ called *observables*. Denoting the columns by $A_x, x \in \mathcal{X}$, the probability distributions in $\mathcal{E}_V$ can be written as $p_\theta(x) = \frac{1}{Z_\theta} \exp(\theta^\top A_x)\ \forall\, x \in \mathcal{X}\ \forall\, \theta \in \mathbb{R}^d$, where $Z_\theta := \sum_y \exp(\theta^\top A_y)$. For simplicity we always denote a sufficient statistics by A and the corresponding exponential family by $\mathcal{E}$. The parametrization given above depends on A, but $\mathcal{E}$ itself only depends on V (modulo the constant functions). We

assume, without loss of generality, that $\mathbb{1} := (1, \dots, 1)$ is a row of A. The map $\theta \mapsto p_\theta$ is bijective and $\mathcal{E}$ has dimension $\mathrm{rk}(A) - 1$ exactly when the rows of A (including $\mathbb{1}$), are linearly independent, see for instance [2]. The elements of an exponential family $\mathcal{E}$ are strictly positive. The closure $\overline{\mathcal{E}}$ includes probability distributions with support strictly contained in $\mathcal{X}$.

Given a collection of sets $\Delta \subseteq 2^{[N]}$, the *hierarchical model* $\mathcal{E}_\Delta$ is the exponential family defined by $V_\Delta := \{\sum_{\lambda\in\Delta} f_\lambda \colon f_\lambda \in \mathbb{R}^{\mathcal{X}} \text{ with } f_\lambda(x_\lambda, x_{[N]\setminus\lambda}) = f_\lambda(x_\lambda, \tilde{x}_{[N]\setminus\lambda})$ $\forall\, x, \tilde{x} \in \mathcal{X},\ \forall\, \lambda \in \Delta\}$. The *$k$-interaction exponential family* $\mathcal{E}^k$ is the hierarchical model $\mathcal{E}_{\Delta_k}$ with $\Delta_k := \{\lambda \subseteq [N] : |\lambda| \leq k\}$. The special case $\mathcal{E}^1$ is called *independence model*. The independence model consists of all strictly positive independent distributions, or product distributions, of the variables $X_1, \dots, X_N$. There is a natural hierarchy of nested models $\mathcal{E}^1 \subset \mathcal{E}^2 \subset \cdots \subset \mathcal{E}^N = \mathcal{P}$, see details in [1, 3]. The dimension of $\mathcal{E}_\Delta$ is $\dim(\mathcal{E}_\Delta) = \sum_{\lambda\in\Delta} \prod_{i\in\lambda}(|\mathcal{X}_i| - 1) - 1$, see [19]. The binary k-interaction model has dimension $\dim(\mathcal{E}^k_{N,\mathrm{bin}}) = \sum_{i=1}^k \binom{N}{i}$. The sufficient statistics of any binary hierarchical model $\mathcal{E}_\Delta$ can be chosen as $A = (\sigma_{\lambda,x})_{\lambda\in\Delta, x\in\{0,1\}^N}$, where

$$\sigma_{\lambda,x} := (-1)^{|\mathrm{supp}(x)\cap\lambda|} \quad \forall\, x \in \{0,1\}^N \quad \forall\, \lambda \in 2^{[N]}\,.$$

The rows of $\sigma = (\sigma_{\lambda,x})_{\lambda\in 2^{[N]}, x\in\{0,1\}^N}$ with labels λ from an inclusion complete set $\Delta \subseteq 2^{[N]}$ are an orthogonal basis of $V_\Delta \subseteq \mathbb{R}^{\mathcal{X}}$, $\mathcal{X} = \{0,1\}^N$. In particular, σ is a Hadamard matrix.

The *convex support* of $\mathcal{E}$, as realized from a sufficient statistics A, is the image of the *moment map*, $\pi \colon \overline{\mathcal{P}} \to \mathbb{R}^d\,;\ p \mapsto A \cdot p$. This is the following convex polytope (the convex hull of finitely many points in Euclidian space):

$$Q := \mathrm{conv}\{A_x\}_{x\in\mathcal{X}}\,.$$

The moment map π defines a homeomorphism of $\overline{\mathcal{E}}$ and Q, and $A \cdot p$ is called the *expectation parameter* vector of the point $p \in \overline{\mathcal{E}}$, see [2, 10] and further details in the Appendix. A *face* of the polytope Q is the intersection of Q with a hyperplane in $\mathbb{R}^d$ such that all points of Q lie on one of the closed halfspaces defined through that hyperplane. In particular, Q is a face of itself. The dimension of a face F is defined as the dimension of its affine hull $\dim(F) := \dim \mathrm{aff}(F)$. The *combinatorial type* of Q is the set of all its faces, denoted by $\mathcal{F}(Q)$, together with the partial order of inclusion. For any $0 \leq g \leq \dim(Q) - 1$ the union of g-dimensional faces $\cup_{F\in\mathcal{F}(Q):\dim(F)=g} F$ contains all vertices of Q [18, Theorem 15.1.2]. Any nonsingular affine transformation of a polytope preserves its combinatorial type [17, Theorem 3.2.3]. In turn, the combinatorial type of Q depends only on the row span of A (modulo the constant functions).

A set $\mathcal{Y} \subseteq \mathcal{X}$ is called a *facial* set of the exponential family $\mathcal{E}$ iff $\mathcal{Y} = \{x \in \mathcal{X} \colon A_x \in F\}$ for some face F of Q. The set of all facial sets of $\mathcal{E}$ is denoted by $\mathcal{F}(\mathcal{E}) \subseteq 2^{\mathcal{X}}$. It is well known that $\mathcal{F}(\mathcal{E})$ and $\mathcal{F}(Q)$ are in one-to-one correspondence (see, for example [14, 30]): A set $\mathcal{Y} \subseteq \mathcal{X}$ is the support of a distribution $p \in \overline{\mathcal{E}}$ if and only if $\mathcal{Y}$ is a facial set of $\mathcal{E}$.

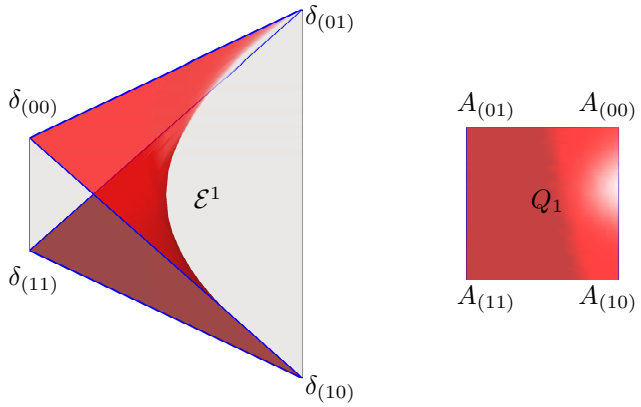

Fig. 1. Left: The set of probability distributions of two independent binary variables $\overline{\mathcal{E}^1}$ (red surface) within the three-dimensional simplex of probability distributions on $\{0,1\}^2$. The vertices of the probability simplex are the point measures δ_x, $x \in \{0,1\}^2$ (distributions with a single support point $\{x\}$). Right: The convex support of $\mathcal{E}^1$ realized as the convex hull of the sufficient statistics A from Example 3.

Example 3. Consider the set of strictly positive product distributions of two binary variables, $p(x_1,x_2) = p_1(x_1)p_2(x_2)$ for all $(x_1,x_2) \in \{0,1\}^2$, where p_1 and p_2 are strictly positive distributions on $\{0,1\}$. This is an exponential family $\mathcal{E}^1$ with sufficient statistics

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \qquad \begin{matrix} \\ (00) & (01) & (10) & (11) \end{matrix}$$

whereby $p_1(x_1)p_2(x_2) = \frac{1}{Z}\exp(\theta^\top A_{(x_1x_2)})$ when $\exp(\theta_2) = p_1(0)/p_1(1)$ and $\exp(\theta_3) = p_2(0)/p_2(1)$. The parameter θ_1 is irrelevant. The convex support $Q_1 = \operatorname{conv}\{A_x\}_{x\in\{0,1\}^2}$ is a square. See Figure 1. The two-mixture $\operatorname{Mixt}^2(\overline{\mathcal{E}^1})$ is the union of all line segments $\{\alpha p + (1-\alpha)q \colon \alpha \in [0,1]\}$ connecting pairs $p, q \in \overline{\mathcal{E}^1}$. The support sets of distributions in $\overline{\mathcal{E}^1}$ are $\{0,1\}^2$, the pairs $\{(00),(01)\}$, $\{(01),(11)\}$, $\{(11),(10)\}$, $\{(10),(00)\}$, and all the points $\{(00)\},\{(01)\},\{(10)\}$, $\{(11)\}$. All support sets are S-sets, except for $\{0,1\}^2$. Figure 1 reveals that every point in the probability simplex is a mixture of two distributions with supports in the S-sets $\{(0,0),(0,1)\}$ and $\{(1,0),(1,1)\}$. These two S-sets cover the entire sample space $\{0,1\}^2$.

3. S-SETS OF EXPONENTIAL FAMILIES

We assess the expressive power of mixture models, comparing the support sets of distributions from different models. In this section we formalize the idea, and relate S-sets of

exponential families to simplex faces of their convex supports.

Given an exponential family $\mathcal{E}$ on $\mathcal{X}$ we consider the following function, which gives the minimal cardinality of a facial packing of any set $\mathcal{Z} \subseteq \mathcal{X}$:

$$\kappa^f_{\mathcal{E}}: 2^{\mathcal{X}} \to \mathbb{N}\,;\; \mathcal{Z} \mapsto \min\{n \in \mathbb{N}\colon \exists\, \mathcal{Y}_1, \ldots, \mathcal{Y}_n \in \mathcal{F}(\mathcal{E}) \text{ with } \cup_i \mathcal{Y}_i = \mathcal{Z}\}\,.$$

We set $\kappa^f_{\mathcal{E}}(\mathcal{Z}) = \infty$ if there does not exist a facial packing of $\mathcal{Z}$. All $\mathcal{Y}_i$ in this definition are required to be subsets of $\mathcal{Z}$. For many exponential families, including hierarchical models (with $\cup_{\lambda\in\Delta}\lambda = [N]$), every $\{x\}$ is a facial set. In particular $\kappa^f_k := \kappa^f_{\mathcal{E}^k} < \infty$ for all $k > 0$. We also consider the smallest number of S-sets that cover $\mathcal{Z}$, which is the following function:

$$\kappa^s_{\mathcal{E}}: 2^{\mathcal{X}} \to \mathbb{N}\,;\; \mathcal{Z} \mapsto \min\{n \in \mathbb{N}\colon \exists\, \mathcal{Y}_1, \ldots, \mathcal{Y}_n \; S\text{-sets with } \cup_i \mathcal{Y}_i \supseteq \mathcal{Z}\}\,,$$

whereby we set $\kappa^s_{\mathcal{E}}(\mathcal{Z}) = \infty$ if there does not exist an S-set covering of $\mathcal{Z}$. If κ S-sets cover $\mathcal{X}$, then at most κ S-sets are needed for packing any $\mathcal{Z} \subseteq \mathcal{X}$, because any subset of an S-set is an S-set. We abbreviate $\kappa^s_{\mathcal{E}}(\mathcal{X})$ with $\kappa^s_{\mathcal{E}}$. Finally, given two exponential families $\mathcal{E}$ and $\mathcal{E}'$, we consider the maximum of $\kappa^f_{\mathcal{E}}$ restricted to the facial sets of $\mathcal{E}'$:

$$\kappa^f_{\mathcal{E},\mathcal{E}'} := \max_{\mathcal{Z}\in\mathcal{F}(\mathcal{E}')} \kappa^f_{\mathcal{E}}(\mathcal{Z})\,.$$

The functions $\kappa^f_{\mathcal{E}}$ and $\kappa^s_{\mathcal{E}}$ can be defined for any model $\mathcal{M} \subseteq \overline{\mathcal{P}}$ in the place of the exponential family $\mathcal{E}$ by simply replacing "facial sets" with "support sets of distributions within $\overline{\mathcal{M}}$." We have the following:

Lemma 4. Consider two exponential families $\mathcal{E}, \mathcal{E}' \subseteq \mathcal{P}(\mathcal{X})$.

- If $m \geq \kappa^s_{\mathcal{E}} < \infty$, then $\mathrm{Mixt}^m(\mathcal{E}) = \mathcal{P}$.
- $\mathrm{Mixt}^m(\overline{\mathcal{E}}) \supseteq \overline{\mathcal{E}'}$ implies $m \geq \kappa^f_{\mathcal{E},\mathcal{E}'}$.

Proof. See Appendix. □

Remark 5. If $\mathrm{Mixt}^m(\mathcal{E}) = \mathcal{P}$, then also $\mathrm{Mixt}^m(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$, and if $\mathrm{Mixt}^m(\overline{\mathcal{E}}) \neq \overline{\mathcal{P}}$, then $\overline{\mathcal{P}} \setminus \mathrm{Mixt}^m(\mathcal{E})$ has a non-empty interior. If $\mathrm{Mixt}^m(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$, then $m \geq \max \kappa^f_{\mathcal{E}}$, and if $\kappa^f_{\mathcal{E},\mathcal{E}'} = \infty$, then $\mathrm{conv}(\mathcal{E}) \not\supset \mathcal{E}'$. Lemma 4 can be formulated for arbitrary models as well. In that case however, the implication of the first item holds only for the closures: If $m \geq \kappa^s_{\mathcal{M}}$, then $\mathrm{Mixt}^m(\overline{\mathcal{M}}) = \overline{\mathcal{P}}$.

Example 6. Any distribution p with support in a cylinder set $[y_{\Lambda^c}]$, $\Lambda \subseteq [N]$, $|\Lambda| = k$ is contained in the closure of the k-interaction family $\overline{\mathcal{E}^k}$. Indeed, if $p \in \overline{\mathcal{P}}$ is arbitrary with support $[y_{\Lambda^c}]$, then $p(x) = \lim_{\alpha\to\infty} \exp(f(x_\Lambda) - \alpha \sum_{j\in\Lambda^c} g_j(x_j))/Z$, where Z is a normalization constant, $f(x) = f(x_\Lambda)$ is a function of the variables $X_i, i \in \Lambda$ with $f(x_\Lambda) = \log(p(x)) + \log(Z)\ \forall\, x \in [y_{\Lambda^c}]$, and g_j is a function of X_j, only, taking value 0 for $x_j = y_j$ and 1 otherwise. Therefore, every k-dimensional cylinder set is an S-set of $\mathcal{E}^k$. In particular, if $\mathcal{X} = \{1, \ldots, q\}^N$, then $\kappa^s_{\mathcal{E}^k} \leq q^{N-k}$ and $\mathrm{Mixt}^{q^{N-k}}(\mathcal{E}^k) = \mathcal{P}$.

Lemma 7. Consider an exponential family $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$ and some $\mathcal{Y} \subseteq \mathcal{X}$. The following items are equivalent:

- $\overline{\mathcal{E}} \supseteq \overline{P}(\mathcal{Y})$, i. e., $\mathcal{Y}$ is an *S-set*.
- $\mathrm{conv}\{A_y\}_{y\in\mathcal{Y}}$ is a $(|\mathcal{Y}|-1)$-dimensional simplex face of the convex support Q.
- $\mathrm{supp}(m^{\pm}) \not\subset \mathcal{Y}$ for all $m \in \ker(A) \subset \mathbb{R}^{\mathcal{X}} \setminus \{0\}$, where $m^{\pm}(x) := \max\{0, \pm m(x)\}$ $\forall\, x \in \mathcal{X}$.

P r o o f . The first item implies the second, because the moment map defines a bijection between $\mathcal{P}(\mathcal{Y})$ and $\mathrm{conv}\{A_y\}_{y\in\mathcal{Y}}$. For the other direction: The matrix $A_{\mathcal{Y}} := (A_y)_{y\in\mathcal{Y}}$ defines an exponential family $\mathcal{E}_{\mathcal{Y}} = \overline{\mathcal{E}} \cap \mathcal{P}(\mathcal{Y})$, because $\mathcal{Y}$ is facial. If $\mathrm{conv}\{A_y\}_{y\in\mathcal{Y}}$ is a $(|\mathcal{Y}|-1)$-simplex, then all columns of $A_{\mathcal{Y}}$ are linearly independent ($\mathbb{1}$ is a row of A), and hence $\ker A_{\mathcal{Y}} = \{0\}$. As a consequence, any $p \in \overline{\mathcal{P}}(\mathcal{Y})$ trivially satisfies $\prod_x (p(x))^{m^+(x)} - \prod_x (p(x))^{m^-(x)} = 0 \ \forall\, m \in \ker A_{\mathcal{Y}}$, which implies $p \in \overline{\mathcal{E}}_{\mathcal{Y}}$ [14, 30]. The third item is equivalent to: $\mathcal{Y}$ is facial, see [30], and additionally $\mathrm{supp}(m) \not\subset \mathcal{Y}$ $\forall\, m \in \ker(A) \setminus \{0\}$. This implies $\ker A_{\mathcal{Y}} = \{0\}$. □

Remark 8. By Lemma 7, $\overline{\mathcal{E}}$ contains any p with $|\mathrm{supp}(p)| < |\mathrm{supp}(m^+)|$ for all $m \in \ker(A) \setminus \{0\}$, and there always exists some $q \in \overline{\mathcal{P}}(\mathcal{X}) \setminus \overline{\mathcal{E}}$ with

$$|\mathrm{supp}(q)| = \min_{m\in\ker(A)\setminus\{0\}} |\mathrm{supp}(m^+)|.$$

When every column A_x of the sufficient statistics is a vertex of Q and κ simplex faces of Q contain all A_x, then $\mathrm{Mixt}^{\kappa}(\overline{\mathcal{E}}) = \mathrm{conv}(\overline{\mathcal{E}})$. When all $A_x, x \in \mathcal{X}$ are distinct vertices of Q, then $\overline{\mathcal{E}}$ contains all possible point measures, $\kappa^{\mathrm{s}}_{\mathcal{E}}$ is the smallest number of simplex faces that contain all vertices, and $\mathrm{Mixt}^{\kappa^{\mathrm{s}}_{\mathcal{E}}}(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$. Computing $\kappa^{\mathrm{s}}_{\mathcal{E}}$ can be difficult, in general. Two examples of related problems are: Finding minimum clique coverings, which is a graph-theoretical NP-complete problem, and describing perfect covering codes on $\{0,1\}^N$, which so far are not completely understood (see [11]).

A polytope P is called *K-neighborly*, when the convex hull of any K, or less, of its vertices is a face (see [23, 32]). If the convex support of $\mathcal{E}$ is K-neighborly and $\overline{\mathcal{E}}$ contains all point measures, then every $\mathcal{Y} \subseteq \mathcal{X}$ with $|\mathcal{Y}| \leq K$ is an S-set of $\mathcal{E}$. It is known that the convex support Q_k of the k-interaction family is (2^k-1)-neighborly, see [20]. In Section 4 we will study the *simpliciality* of Q_k and corresponding vertex set coverings using simplex faces. A polytope P is K-simplicial if all its K-dimensional faces are simplices (this does not mean that any $(K+1)$ vertices define a face of P).

Example 9. The convex support of the two-interaction family $\mathcal{E}^2$ on $\mathcal{X} = \{0,1\}^4$ is a polytope with 16 vertices and dimension 10. We computed the face lattice of Q_2 (using the software `Polymake` [13]). We found 56 facets (proper faces of maximal dimension, 9), out of which 16 are simplices. One of them is $\mathrm{conv}\{A_x\}_{x\in\mathcal{Y}}$, $\mathcal{Y} = \{(0000), (1000), (0100), (0010), (1001), (0101), (0011), (1101), (1011), (0111)\}$. In total 8 S-sets of $\mathcal{E}^2$ contain 6 binary vectors with an even number of ones, and 8 contain 6 vectors with

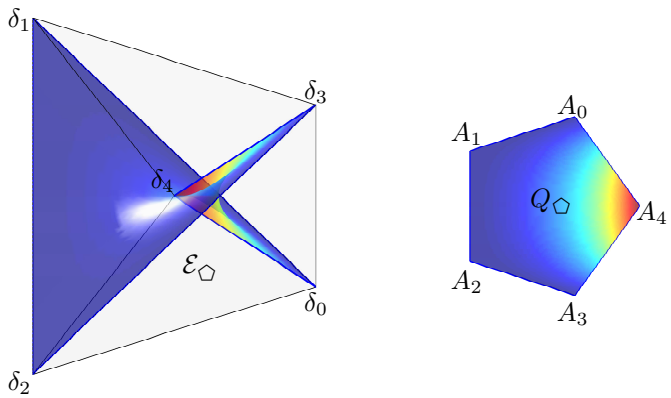

Fig. 2. Schlegel diagram of the four-dimensional probability simplex on $\{0,\dots,4\}$ and corresponding projection of the two-dimensional exponential family $\mathcal{E}_{\pentagon}$ with convex support $Q_{\pentagon}$ a regular pentagon. The color indicates the value that the distributions take on $x=4$; blue for $p(4)=0$ and red for $p(4)=1$. The uniform distribution $\frac{\mathbb{1}}{5}$ and the point measure δ_4 are projected into the same point.

an odd number of ones. The other 40 facets have each 12 vertices. Denote the S-sets (of cardinality 10) by $F_i, i=1,\dots,16$ and the facial sets of cardinality 12 by $G_i, i=1,\dots,40$. We found that $F_i\cup F_j\neq\mathcal{X}\ \forall i,j$ and $F_i\cup G_j\neq\mathcal{X}\ \forall i,j$. Since all faces (and in particular all simplex faces) are subsets of some facet, at least 3 S-sets of $\mathcal{E}^2$ are needed to cover $\mathcal{X}$.

Example 10. Let $\mathcal{X}=\{0,\dots,n-1\}$ and let $\mathcal{E}$ be an exponential family with convex support an n-gon (a polygon with n vertices). We call this family an n-gon exponential family. It is two-dimensional and contains all point measures δ_x in its closure. n-gon exponential families have been studied in the context of model design in [4]. Assume that the boundary of the convex support Q of an n-gon family is the polyline $A_0A_1\cdots A_{n-1}A_0$. The facial sets are: $\mathcal{X}$, the pairs $\{i,i+1\}\mod n$, and the points $\{i\}$ for $i\in\mathcal{X}$. All facial sets, except $\mathcal{X}$, are S-sets. The sample space $\mathcal{X}$ is covered by $\kappa^s_{\mathcal{E}}=\lceil\frac{n}{2}\rceil$ S-sets, while the packing of any set $\mathcal{Y}\subseteq\mathcal{X}$ requires at most $\max\kappa^f_{\mathcal{E}}=\lfloor\frac{n}{2}\rfloor$ facial sets. By Lemma 4 the smallest m for which $\mathrm{Mixt}^m(\mathcal{E})=\mathrm{conv}(\mathcal{E})=\mathcal{P}$ satisfies $\lfloor\frac{n}{2}\rfloor\leq m\leq\lceil\frac{n}{2}\rceil$. For $n=5$ (see Figure 2 right) we show that $m\geq 2=\lfloor\frac{n}{2}\rfloor$ is necessary and sufficient, see below.

Proposition 11. If $\mathcal{E}_{\pentagon}$ is an exponential family on $\mathcal{X}=\{0,1,2,3,4\}$ with pentagonal convex support, then $\mathrm{Mixt}^2(\mathcal{E}_{\pentagon})=\mathcal{P}(\mathcal{X})$.

Proof. See Appendix. □

Remark 12. Example 10 shows that in general $\kappa^f_{\mathcal{E}} \neq \kappa^s_{\mathcal{E}}$. In such a case $m = \kappa^s_{\mathcal{E}}$ is not necessarily the smallest m for which $\mathrm{Mixt}^m(\mathcal{E}) = \mathcal{P}$. For pentagonal exponential families $\kappa^s_{\mathcal{E}}$ is off by one, and the same likely happens for all n-gon exponential families with odd n greater or equal to five. However, in the next section we show that $\kappa^f_{\mathcal{E}}$ equals $\kappa^s_{\mathcal{E}}$ for many independence models, and we believe that this generalizes to many interaction models.

4. MIXTURES OF HIERARCHICAL MODELS

4.1. Independence Models

The Hamming distance between two vectors $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ is $d_H(x, y) := |\{i \in [N] \colon x_i \neq y_i\}|$. A set $\mathcal{Y} \subseteq \mathcal{X}$ has minimum distance d if the smallest Hamming distance between two distinct points $x, y \in \mathcal{Y}$ is at least d. The independence model of N variables with joint sample space $\mathcal{X} = \times_{i\in[N]}\mathcal{X}_i$ is:

$$\mathcal{E}^1 = \Big\{ p \in \mathcal{P} : p(x_1, \ldots, x_N) = \prod_{i\in[N]} p_i(x_i) \;\forall\, (x_1, \ldots, x_N) \in \mathcal{X}, \text{ and } p_i \in \mathcal{P}(\mathcal{X}_i) \;\forall\, i \in [N] \Big\} . \tag{1}$$

For binary variables the convex support of $\mathcal{E}^1$ is a combinatorial N-cube, the facial sets are the cylinder sets (including those of dimension zero), and the S-sets are the pairs of vectors with Hamming distance one to each other, plus all individual binary vectors. In general, the convex support is a Cartesian product $Q_1 = \times_{i\in[N]} S_i$, where S_i is a $(|\mathcal{X}_i| - 1)$-dimensional simplex for every $i \in [N]$. The facial sets are:

$$\mathcal{F}(\mathcal{E}^1) = \{ \times_{i\in[N]} \mathcal{Y}_i : \mathcal{Y}_i \subseteq \mathcal{X}_i \text{ for all } i \in [N] \} , \tag{2}$$

and the S-sets are the subsets of one-dimensional cylinders; i. e., the sets

$$\{y_1\} \times \cdots \times \{y_{i-1}\} \times \mathcal{Y}_i \times \{y_{i+1}\} \times \cdots \{y_N\} \tag{3}$$

with $y_j \in \mathcal{X}_j$ for all $j \in [N] \setminus \{i\}$ and $\mathcal{Y}_i \subseteq \mathcal{X}_i$ for $i \in [N]$.

We consider the maximal cardinality of subsets $\mathcal{Y} \subseteq \mathcal{X} = \times_{i\in[N]}\mathcal{X}_i$ with minimum distance two:

$$\mathcal{A}_{\mathcal{X}} := \max\{|\mathcal{Y}| : \mathcal{Y} \subseteq \mathcal{X} \text{ and } d_H(x, y) \geq 2 \;\forall\, x, y \in \mathcal{Y}, x \neq y\} . \tag{4}$$

For example, the sets of binary vectors of length N with an even (odd) number of ones, $Z_{\pm} := \big\{x \in \mathcal{X} = \{0, 1\}^N \colon \prod_{i\in[N]} (-1)^{x_i} = \pm 1\big\}$, have the largest possible cardinality $|Z_{\pm}| = 2^{N-1}$ among all sets of length-N binary vectors of minimum distance two. For mixtures of independence models we have:

Theorem 13. The mixture model $\mathrm{Mixt}^m(\overline{\mathcal{E}^1})$ contains every probability distribution with support in a union of m one-dimensional cylinder sets, and does not contain any probability distribution supported by a set of cardinality more than m and minimum distance (at least) two. Furthermore,

- If $m \geq |\mathcal{X}|/\max\{|\mathcal{X}_i|\}_{i\in[N]}$, then $\mathrm{Mixt}^m(\mathcal{E}^1) = \mathcal{P}$.
- If $\mathrm{Mixt}^m(\overline{\mathcal{E}^1}) \supseteq \mathcal{P}$, then $m \geq \mathcal{A}_{\mathcal{X}} \geq \max\{\frac{s^N}{1+N(s-1)}, q^{N-1}\}$, where $s = \min\{|\mathcal{X}_i|\}_{i\in[N]}$ and q is the largest prime power smaller or equal to s.

In particular, when $\mathcal{X} = \{1,\dots,q\}^N$ and q is a prime power, then

$$\mathrm{Mixt}^m(\mathcal{E}^1) = \mathcal{P} \quad \text{ if and only if } \quad m \geq q^{N-1}\,.$$

Proof. For the first statement: Any face of Q_1 which has more than one vertex contains edges. An edge of Q_1 is the convex hull of a column pair A_x, A_y of the sufficient statistics, with $d_H(x,y) = 1$. Therefore, any facial set contained in a set that does not contain pairs of Hamming distance one, has cardinality one. For the remaining statements, assume (without loss of generality) $|\mathcal{X}_1| = \max\{|\mathcal{X}_i|\}_{i\in[N]}$ and $|\mathcal{X}_N| = \min\{|\mathcal{X}_i|\}_{i\in[N]}$. The first bullet is by Lemma 4, covering the sample space with the S-sets $\{(x_1, y_2, \dots, y_N)\colon x_1 \in \mathcal{X}_1\}$ for all $(y_2,\dots,y_N) \in \times_{i\in[N]\setminus\{1\}}\mathcal{X}_i$. For the second bullet we use the first part of this theorem, and the fact that $\overline{\mathcal{P}}(\{1,\dots,s\}^N)$ is contained in $\overline{\mathcal{P}}(\mathcal{X})$. For any q the *maximal cardinality of a q-ary code of length N and minimum distance d*, defined as $\mathcal{A}_q(N,d) := \max\{|\mathcal{Y}| : \mathcal{Y} \subseteq \{1,\dots,q\}^N \text{ and } d_H(x,y) \geq d\ \forall\, x,y \in \mathcal{Y}, x \neq y\}$, is familiar in coding theory. It is known that $\mathcal{A}_q(N,2) \geq \frac{q^N}{\sum_{j=0}^{d-1}\binom{N}{j}(q-1)^j}$ (*Gilbert-Varshamov bound* [15, 34]), and that when q is any power of a prime number, $\mathcal{A}_q(N,d) \geq q^k$, where k is the largest integer with $q^k < \frac{q^N}{\sum_{j=0}^{d-2}\binom{N-1}{j}(q-1)^j}$. Evaluating these bounds for $d = 2$ completes the proof. □

Corollary 14. Let $1 \leq k \leq N-1$. If $\mathcal{X} = \{0,1\}^N$ and $\mathrm{Mixt}^m(\overline{\mathcal{E}}) \supseteq \mathcal{E}^k$, then

$$m \geq \max\{|\mathcal{Z}| : \mathcal{Z} \in \mathcal{F}(\mathcal{E}^k),\ \mathcal{Z} \subseteq Z_{\pm}\} \geq 2^k - 1\,.$$

Proof. The first inequality is by Lemma 4, since $\kappa^f_{1,k} \geq \max\{|\mathcal{Z}| : \mathcal{Z} \in \mathcal{F}(\mathcal{E}^k)$. The second one follows from Lemma 18. □

Example 15. The first inequality in Corollary 14 is useful when we have information about the support sets of $\overline{\mathcal{E}^k}$. The second inequality improves the bound

$$m \geq (\dim(\mathcal{E}^k)+1)/(N+1) = \sum_{j=0}^{k}\binom{N}{j}/(N+1)$$

that can be derived comparing the dimension of both models, when k is close to N. For instance:

- When $\mathcal{X} = \{0,1\}^4$, $\mathcal{E}^2$ has S-sets of cardinality 6, contained in Z_+ (see Example 9). Hence, if $\mathrm{Mixt}^m(\overline{\mathcal{E}^1}) \supseteq \mathcal{E}^2$, then $m \geq 6$.
- If $\mathcal{X} = \{0,1\}^4$ and $\mathrm{Mixt}^m(\overline{\mathcal{E}^1}) \supseteq \mathcal{E}^3$, then $m \geq 7$. For comparison, counting parameters yields only $m \geq \lceil(\dim(\mathcal{E}^3)+1)/(4+1)\rceil = 3$.

4.2. Interaction Models

Theorem 16. Consider a hierarchical model $\mathcal{E}_\Delta$ on $\mathcal{X} = \times_{i\in[N]}\mathcal{X}_i$ with $\Delta \supseteq \Delta_k$, $1 \leq k < N$.

- The mixture model $\mathrm{Mixt}^m(\overline{\mathcal{E}_\Delta})$ contains any probability distribution $p \in \overline{\mathcal{P}}(\mathcal{X})$ when m is larger or equal to $\min\{n\colon \exists\, \mathcal{Y}_1, \ldots, \mathcal{Y}_n$ k-cylinder sets of $\mathcal{X}$ with $\mathrm{supp}(p) \subseteq \cup_{i=1}^n \mathcal{Y}_i\}$. Furthermore, $\mathrm{Mixt}^m(\mathcal{E}_\Delta) = \mathcal{P}$ whenever $m \geq |\mathcal{X}|/\max\{\prod_{i\in\lambda}|\mathcal{X}_i|\colon \lambda \subseteq [N], |\lambda| = k\}$.
- In the case of binary variables, the convex support Q_Δ is $(2^k{-}1)$-neighborly, $(2^{k+1}{-}3)$-simplicial, and all its vertices are contained in the union of $2^{N-(k+1)}(1+\frac{1}{2^k-1})$ simplex faces. Moreover, $\mathrm{Mixt}^m(\mathcal{E}_\Delta) = \mathcal{P}$ whenever $m \geq 2^{N-(k+1)}(1+\frac{1}{2^k-1})$.

The first item of Theorem 16 follows from the observation that all k-cylinders are S-sets of $\mathcal{E}^k$, see Example 6. The (2^k-1)-neighborliness of Q_Δ was shown in [20]. The $(2^{k+1}{-}3)$-simpliciality follows from a classic result of convex polytopes, which states that if P a K-neighborly d-dimensional polytope, then every face F of dimension less than $2K$ is a simplex, see [17, Theorem 7.4.3]. In order to prove the remaining statements of the theorem, we need to find the $(2^{k+1}-3)$-dimensional faces of Q_Δ and show that at most $2^{N-(k+1)}(1+\frac{1}{2^k-1})$ of them cover all vertices. Before proving this, some remarks are appropriate:

Remark 17. Regarding the upper bound $2^{N-(k+1)}(1+\frac{1}{2^k-1})$ on the minimal cardinality of an S-set covering of $\{0,1\}^N$ (second item of Theorem 16): When $k=1$ the bound equals 2^{N-1} and is tight by Theorem 13. When $N=4$ and $k=2$ the bound is $\lceil 2^{4-(2+1)}/(1-2^{-2})\rceil = 3$ and is tight in view of Example 9. When $k = N-1$ the bound equals 2 and is tight, because $\mathrm{Mixt}^1(\mathcal{E}_\Delta) = \mathcal{E}_\Delta \neq \mathcal{P}$. In spite of this, the characterization of the simplex faces of convex support polytopes and the computation of the smallest simplex-face-vertex-set coverings for hierarchical models with general interaction sets Δ and non-binary variables, is not fully accomplished at this point. In particular we believe that the bound provided in the first item can be further improved, as for binary variables the second item provides a much better bound.

Note that any facial set of $\mathcal{E}^k$ is a facial set of $\mathcal{E}_\Delta$, $\Delta \supseteq \Delta_k$. For $0 < k < N$, any $(k+1)$-dimensional cylinder set $[y_{\lambda^c}], \lambda \subseteq [N], |\lambda| = k+1$ is a facial set of $\mathcal{E}^k$ (for example by similar arguments as in Example 6). Hence the vertices of Q_Δ can be covered by $2^{N-(k+1)}$ disjoint faces $\{F_i\}_i$ corresponding to $(k+1)$-dimensional cylinder sets. These F_i are not simplices, but they contain $(2^{k+1}-3)$-dimensional simplex faces (see below), which we can arrange in a convenient way to cover all vertices of Q_k disjointly. We use the following Lemma 18, which subsumes various ideas and remarks from [17, 19, 22].

A d-dimensional *cyclic polytope* with v vertices (see [12]) is defined as the convex hull of v distinct points on the *d-moment curve*: $C(v,d) := \mathrm{conv}\{x(t_i)\}_{i=1,\ldots,v}$, where $v \geq d+1$, $t_1 < \cdots < t_v \in \mathbb{R}$, and $x(t) = (t, t^2, \ldots, t^d) \in \mathbb{R}^d$.

Lemma 18. Let $0 < k < N$ and $\mathcal{X} = \{0,1\}^N$. Any $(k+1)$-dimensional cylinder set $\mathcal{Y}$ is a facial set of $\mathcal{E}^k$ and the corresponding face F of the convex support Q_k is a simplicial

polytope, combinatorially equivalent to the cyclic polytope $C(2^{k+1}, 2^{k+1}-2)$. There are exactly 2^{2k} S-sets of cardinality $(2^{k+1}-2)$ contained in $\mathcal{Y}$; namely $\{\mathcal{Z} \subset \mathcal{Y} : \mathcal{Y} \cap Z_\pm \not\subseteq \mathcal{Z}\}$. In particular, if a set $Z \subseteq \mathcal{X}$ contains $\mathcal{Y} \cap Z_\pm$ but does not contain $\mathcal{Y}$, then $\mathcal{Z}$ is not facial.

Proof. See Appendix. □

Proof of Theorem 16. Let $x_i^{i+k} := (x_i, \dots, x_{i+k}) \in \{0,1\}^{\{i,\dots,i+k\}}$. Consider the following partition of $\{0,1\}^N$ into $(k+1)$-dimensional cylinder sets:

$$C_y := \{(x_1^{k+1}, x_{k+2}^N) \in \{0,1\}^N : x_{k+2}^N = y\} \quad \text{for all} \quad y \in \{0,1\}^{N-(k+1)} .$$

By Lemma 18 the elements of any C_y can be disjointly covered by:
(i) An S-set of $\overline{\mathcal{E}^k}$ of cardinality $2^{k+1}-2$. We denote this set by G_y.
(ii) A pair of vectors differing in one entry:

$$E_y := \{(z_1^k, x_{k+1}, y) \in \{0,1\}^N : z_1^k \text{ fixed}\} . \tag{5}$$

The vector z in eq. (5) can be chosen equal for all E_y, such that the S-sets $\{G_y\}_y$ satisfy:

$$\bigcup_{y \in \{0,1\}^{N-(k+1)}} G_y \quad = \quad \{0,1\}^N \setminus \tilde{C}_{N-k} ,$$

where $\tilde{C}_{N-k}$ is the following $(N-k)$-dimensional cylinder set:

$$\tilde{C}_{N-k} = \bigcup_{y \in \{0,1\}^{N-(k+1)}} E_y \quad = \quad \{(z_1^k, \tilde{y}_1^{N-k}) : z_1^k \text{ fixed}\} .$$

The cylinder set $\tilde{C}_{N-k}$ can be considered as a new sample space which still has to be covered using as few S-sets as possible. If $N-k < k+1$, only one S-set is required. Iteration of the previous idea until exhausting all coordinates yields that κ, the minimal number of simplex faces of Q_k covering all vertices, is not more than:

$$\kappa \leq 1 + \sum_{0 \leq i \leq \frac{N-(k+1)}{k}} \frac{2^{N-ik}}{2^{k+1}} = \left\lceil \frac{2^N}{2^{k+1}} \sum_{i=0}^{\infty} \frac{1}{(2^k)^i} \right\rceil = \left\lceil \frac{2^{N-(k+1)}}{1-2^{-k}} \right\rceil .$$

□

We conclude this section with a few observations on S-sets of hierarchical models. From Lemma 18 we can derive a rough cardinality upper bound for the S-sets of $\mathcal{E}^k$. Let $K(N, k+1)$ denote the smallest cardinality of a set $\mathcal{Y} \subseteq \{0,1\}^N$ which intersects all $(k+1)$-dimensional cylinder sets, and let $B_{N,k+1}$ denote a Hamming ball in $\{0,1\}^N$ of radius $k+1$.

Proposition 19. If $\mathcal{Y} \subseteq \mathcal{X} = \{0,1\}^N$ is an S-set of $\mathcal{E}^k$, then

$$|\mathcal{Y} \cap Z_\pm| \leq 2^{N-1} - K(N, k+1) \leq 2^{N-1}(1 - 2/|B_{N,k+1}|)$$

and $|\mathcal{Y}| \leq |\Delta_k|$. Furthermore,

$$|\mathcal{Y}| \leq 2^N - 2K(N, k+1) \leq 2^N(1 - 2/|B_{N,k+1}|),$$

since $\mathcal{X}$ is disjointly covered by the two sets Z_+ and Z_-.

P r o o f . See Appendix. □

Example 20. When $\mathcal{X} = \{0,1\}^4$, by Proposition 19 any S-set of $\mathcal{E}^2$ intersects Z_+, or Z_-, at most at $8 - 2 = 6$ points. This bound is attained exactly, in view of Example 9.

It is worthwhile mentioning that, if a collection of index sets $\Delta \subseteq 2^{[N]}$ is symmetric with respect to a permutation $\pi : [N] \to [N]$, then the convex support Q_Δ of the associated exponential family, also has this symmetry. In particular, if $\mathcal{Y}$ is an S-set of $\mathcal{E}^k$, then $\pi(\mathcal{Y}) := \{(x_{\pi(1)}, \ldots, x_{\pi(N)}) : x \in \mathcal{Y}\}$ is also an S-set, for any permutation π. Furthermore, we have:

Proposition 21. If $\mathcal{E}$ is an exponential family with sufficient statistics

$$A = ((-1)^{|\mathrm{supp}(x) \cap \lambda|})_{\lambda \in \Delta, x \in \mathcal{X}},\ \Delta \subseteq 2^{[N]},\ \mathcal{X} = \{0,1\}^N,$$

then $\mathcal{Y}$ is an S-set if and only if $x * \mathcal{Y} := \{x + y \mod 2 : y \in \mathcal{Y}\}$ is an S-set $\forall\, x \in \mathcal{X}$, and moreover, $\mathcal{Y} \subseteq \mathcal{X}$ is a facial set if and only if $x * \mathcal{Y}$ is a facial set $\forall\, x \in \mathcal{X}$.

P r o o f . See Appendix. □

The sets $\{x * \mathcal{Y}\}_{x \in \mathcal{X}}$ are not necessarily all different from each other, but they are if $|\mathcal{Y}|$ is odd, or if $\mathcal{Y}$ is a Hamming ball. The orbit $\{x * z : x \in \mathcal{X}\}$ of any $z \in \mathcal{X}$, covers $\mathcal{X}$. In particular, $\cup_{x \in \mathcal{X}} x * \mathcal{Y} = \mathcal{X}$ and $|x * \mathcal{Y}| = |\mathcal{Y}|$ for any $x \in \mathcal{X}$, $\mathcal{Y} \subseteq \mathcal{X}$, $\mathcal{Y} \neq \emptyset$. These observations have interesting relations to coding theory; for example, any binary hierarchical model has a convex support which is the convex hull of a binary linear code, see [22].

APPENDIX

P r o o f o f L e m m a 4.

1. Let $\{\mathcal{Y}_i\}_{i=1}^m$ be an S-set covering of $\mathcal{X}$. W.l.o.g. $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset\ \forall\, i \neq j$. Any $p \in \overline{\mathcal{P}}$ can be written as $\sum_{i=1}^m \alpha_i f_i$ and $f_i \in \overline{\mathcal{E}}$ choosing f_i with $\mathrm{supp}(f_i) \subseteq \mathcal{Y}_i$, $f_i = p|_{\mathcal{Y}_i} / \sum_{x \in \mathcal{Y}_i} p(x)$ and $\alpha_i = \sum_{x \in \mathcal{Y}_i} p(x)$. This shows $\mathrm{Mixt}^m(\overline{\mathcal{E}}) = \overline{\mathcal{P}}$.

For strictly positive distributions: The convexity of $\mathcal{P}$ implies $\mathrm{Mixt}^m(\mathcal{E}) \subseteq \mathcal{P}$ for all $m \geq 1$. The direction "$\supseteq$" is a bit more elaborate. By the first part of this proof the set $\mathrm{Mixt}^m(\mathcal{E})$ is dense in $\overline{\mathcal{P}}$; we need to show that within $\overline{\mathcal{P}}$ only the boundary $\partial\mathcal{P} := \overline{\mathcal{P}} \setminus \mathcal{P}$ is not contained in $\mathrm{Mixt}^m(\mathcal{E})$. We use topological arguments. Let $Y_i := \overline{\mathcal{P}}(\mathcal{Y}_i)$, $i = 1, \ldots, m$ be disjoint faces of $\overline{\mathcal{P}}$ containing all point measures $\{\delta_x\}_{x \in \mathcal{X}}$. Let $p_\eta := (A|_{\overline{\mathcal{E}}})^{-1}(\eta)$ denote the distribution in $\overline{\mathcal{E}}$ with expectation parameter η. The mixture map $\phi : D := \overline{\mathcal{P}}_m \times (\times_{i=1}^m Q) \to \overline{\mathcal{P}}\ ;\ (\alpha, \eta_1, \ldots, \eta_m) \mapsto \sum_{i=1}^m \alpha(i) p_{\eta_i}$ is surjective. Restricting the domain D to the subset $C := \partial(\mathcal{P}_m \times (\times_{i=1}^m (A \cdot Y_i)))$ we get a

continuous bijection $\phi|_C\colon C \to \partial\mathcal{P}$ between the compact domain C and the Hausdorff codomain $\partial\mathcal{P}$. Therefore, $\phi|_C$ is a homeomorphism and induces isomorphisms between the homotopy groups of C and those of $\partial\mathcal{P} \simeq S^{|\mathcal{X}|-2}$, the $(|\mathcal{X}|-2)$-sphere. Denote by $\mathring{D}$ the relative interior of the polytope D. Note that $\phi(\mathring{D}) \subseteq \mathcal{P}$. For any $\epsilon > 0$ there is a continuous deformation $C \to \tilde{C} \subseteq \mathring{D}$ which is mapped by ϕ into a continuous deformation $\partial\mathcal{P} \to \phi(\tilde{C}) \subset \mathcal{P} \setminus \mathcal{P}^\epsilon$, $\mathcal{P}^\epsilon := \{p \in \mathcal{P}\colon p(x) \geq \epsilon\ \forall\, x \in \mathcal{X}\}$. If $\phi(\mathring{D})$ does not contain $\mathcal{P}^\epsilon$, then $\phi(\tilde{C})$ is not contractible in $\phi(\mathring{D})$, in contradiction to the contractibility of $\mathring{D}$ (which is a convex set). Since any element of $\mathcal{P}$ belongs to some $\mathcal{P}^\epsilon$, this shows $\mathrm{Mixt}^m(\mathcal{E}) \supseteq \mathcal{P}$.

2. Consider some $p \in \overline{\mathcal{E}'}$ with $\mathrm{supp}(p) = \mathcal{Z} \in \mathcal{F}(\mathcal{E}')$. If p is written as a mixture of elements from $\overline{\mathcal{E}}$, then every mixture component with positive mixture weight must have a support $\mathcal{Y} \in \mathcal{F}(\mathcal{E})$, $\mathcal{Y} \subseteq \mathcal{Z}$. Furthermore, the union of the support sets of these summands must equal $\mathcal{Z}$. The minimal number of summands is, by definition, equal to $\kappa^f_{\mathcal{E}}(\mathcal{Z})$. □

Proof of Proposition 11. Consider any exponential family $\mathcal{E}$, and assume (without loss of generality) that the sufficient statistics contains the row $\mathbb{1}$. The image of the moment map $\pi\colon p \mapsto A \cdot p$ is the convex support $Q = \mathrm{conv}\{A_x\}_{x\in\mathcal{X}}$. Since π is continuous, $\overline{\mathcal{E}}$ is compact, and Q is Hausdorff, this bijective map is in fact a homeomorphism. We denote by $p_\eta = (\pi|_{\overline{\mathcal{E}}})^{-1}(\eta)$ the unique preimage of $\eta \in Q$ by the moment map restricted to $\overline{\mathcal{E}}$. The m-mixture of $\overline{\mathcal{E}}$ is parametrized by a *mixture map* $\phi\colon D := \overline{\mathcal{P}}_m \times Q^m \to \overline{\mathcal{P}}\ ;\ (\alpha, \eta_1, \ldots, \eta_m) \mapsto \sum_{i=1}^m \alpha_i p_{\eta_i}$. Consider the *normal space* $\mathcal{N} = \ker A$ of $\mathcal{E}$. For any $p \in \overline{\mathcal{P}}$ the set $\mathcal{N}_p := \{q \in \overline{\mathcal{P}} : p - q \in \mathcal{N}\}$ is a polytope of dimension $\dim\ker A$ which intersects $\overline{\mathcal{E}}$ at a unique point $p_{\mathcal{E}} \in \overline{\mathcal{E}} \cap \mathcal{N}_p$ (see [29, Theorem 2.16]). Hence $\overline{\mathcal{P}} = (\overline{\mathcal{E}} + \ker A) \cap \overline{\mathcal{P}} = \cup_{p\in\overline{\mathcal{E}}}\mathcal{N}_p$. The boundary of $\mathcal{N}_p$ is contained in the boundary of $\mathcal{P}$.

In the case of $\mathcal{E}_{\pentagon}$ $\dim\ker A = 2$. Furthermore, any subset of $\mathcal{X} = \{0,1,2,3,4\}$ of cardinality 4 is contained in the union of two S-sets. Hence $\mathrm{Mixt}^2(\overline{\mathcal{E}}) \supset \partial\mathcal{P} := \overline{\mathcal{P}} \setminus \mathcal{P}$, and the restriction $\phi|_C\colon C := \partial(\overline{\mathcal{P}}_2 \times Q^2) \to \partial\mathcal{P}$ is a continuous surjection. Now, for any $p \in \mathcal{E}_{\pentagon}$ we consider the set $B_p = \phi^{-1}(\mathcal{N}_p) = \{(\alpha, \eta_1, \eta_2) \in D\colon \sum_{i=1}^2 \alpha_i\eta_i = \pi(p)\}$. This set is mapped by ϕ into the set of convex combinations of 2 elements of $\overline{\mathcal{E}}_{\pentagon}$ which have the same expectation parameter as p. We consider also $\partial B_p = B_p \cap (\overline{\mathcal{P}}_2 \times (\partial Q)^2)$, which corresponds to the same kind of mixtures, but with mixture components from the boundary $\partial\mathcal{E}_{\pentagon} := \overline{\mathcal{E}_{\pentagon}} \setminus \mathcal{E}_{\pentagon}$. We have that $\phi\colon \partial B_p \to \partial\mathcal{N}_p$ is surjective and has degree 2! (the cardinality of the preimage of a regular value, which arises from the freedom to permute the mixture components). The set ∂B_p is parametrized by an angle, say γ, and $\phi|_{\partial B_p}(\gamma)$ circulates $\partial\mathcal{N}_p$ twice. Using that B_p is contractible, it follows that $\phi|_{B_p} = \mathcal{N}_p$ and $\mathrm{Mixt}^2(\overline{\mathcal{E}_{\pentagon}}) = \overline{\mathcal{P}}$. For strictly positive distributions the claim follows from the fact that $\phi(\mathring{B}_p) \subseteq \mathcal{P}$, and that the image of an ε-retraction of B_p, $(1-\varepsilon)(B_p - p) + p$, can be made such that it contains any δ-retraction of $\mathcal{N}_p$, $(1-\delta)(\mathcal{N}_p - p) + p$. □

Proof of Lemma 18. By Lemma 7 $\mathcal{Y}$ is not an S-set $\Leftrightarrow \exists m \in \ker A \setminus \{0\}$ with $\mathrm{supp}(m^+) \subseteq \mathcal{Y}$. If $\mathrm{supp}(m^+) = \mathcal{Y}$, then $\mathcal{Y}$ is not facial, see [14, 30]. Consider the sufficient statistics $A = (\sigma_{\lambda,x})_{\lambda\in\Delta_k, x\in\mathcal{X}}$. The kernel of this matrix is spanned by the rows of the matrix $(\sigma_{\lambda,x})_{\lambda\in 2^{[N]}\setminus\Delta_k, x\in\mathcal{X}}$, which can be written as $(\sigma_{\lambda,x})_{\lambda\in\Delta_{N-k}, x\in\mathcal{X}}\,\mathrm{diag}(\sigma_{[N],x})_{x\in\mathcal{X}}$.

The row span of $(\sigma_{\lambda,x})_{\lambda\in\Delta_{N-k},x\in\mathcal{X}}$ contains any function of $(N-k)$ variables, including the indicator function $\mathbb{1}_{\mathcal{Y}}(x)$ of any $(k+1)$-cylinder set $\mathcal{Y}$. This corresponds to a kernel element of A with entries $m(x)=\mathbb{1}_{\mathcal{Y}}(x)\sigma_{[N],x}, x\in\mathcal{X}$ and $\operatorname{supp}(m^+)=Z_+\cap\mathcal{Y}$.

Since not all subsets of $\mathcal{Y}$ are facial, the corresponding face F of the convex support, which has 2^{k+1} vertices, is not a simplex and has dimension less than $2^{k+1}-1$. By [17, Theorem 7.4.3] and the (2^k-1)-neighborliness of Q_k [20], F is $(2^{k+1}-3)$-simplicial and has dimension less than $2^{k+1}-2$ (otherwise it would be a simplex). The combinatorial equivalence of F and the cyclic polytope $C(2^{k+1},2^{k+1}-2)$ follows from the fact that *any $2n$-dimensional, n-neighborly polytope with $v\le 2n+3$ vertices is combinatorially equivalent to the cyclic polytope $C(v,2n)$* [17, Theorem 7.2.3].

To complete the proof we use Gale's Evenness Criterion: *A d-tuple $V_J=\{x(t_j)\}_{j\in J}$ $J\subset[v],|J|=d$ of vertices of $C(v,d)$, spans a facet iff between any two elements of J there is an even number of elements in $[v]\setminus J$* [17, Theorem 4.7.2]. In our case $v=2^{k+1}$ and $d=2^{k+1}-2$. The combinatorial structure of the cyclic polytope is independent of the map $i\mapsto t_i$ and we may choose $I=[v]:=\{1,\dots,2^{k+1}\}\subset\mathbb{N}$. The sets $V_J,|J|=2^{k+1}-2$ satisfying the evenness criterion are exactly the complements of pairs $\{i^e,i^o\}\subset[v]$, where i^e is even and i^o is odd. There are 2^{2k} such pairs, and hence facets. This is the same as the number of sets respecting the condition on S-sets, $\mathcal{Z}\not\supseteq\mathcal{Y}\cap Z_\pm$, shown at the beginning of this proof. Therefore, all sets $\mathcal{Z}$ with $\mathcal{Z}\not\supseteq\mathcal{Y}\cap Z_\pm$ correspond to facets of $C(2^{k+1},2^{k+1}-2)$ and are indeed S-sets. □

Proof of Proposition 19. Let $\mathcal{Y}$ be any S-set of $\mathcal{E}^k$ and let C be any $(k+1)$-dimensional cylinder set. By Lemma 18 $|(C\cap Z_\pm)\setminus\mathcal{Y}|\ge 1$. Therefore, the maximal cardinality of an S-set $\mathcal{Y}\subseteq Z_\pm$ is upper bounded by $|Z_\pm|-K(N,k+1)$, where $K(N,k+1)$ is the smallest cardinality of a set that intersects every $(k+1)$-dimensional cylinder set. The union of all $(k+1)$-cylinder sets that contain a point x equals the Hamming ball $B_{N,k+1}(x)\subseteq\mathcal{X}$ of radius $k+1$ centered at x. Hence $K(N,k+1)$ is the minimal cardinality of a binary code of covering radius $k+1$. If $R<N\le 2R+1$, then $K(N,R)=2$, but in general computing $K(N,R)$ is hard (see [11]). A crude lower bound is the *sphere-covering bound*: $K(N,R)\ge 2^N/|B_{N,R}|$. Here $|B_{N,R}|=\sum_{i=0}^R\binom{N}{i}$. On the other hand, the cardinality of an S-set of $\mathcal{E}^k$ can not exceed $\dim Q_k+1=|\Delta_k|=|B_{N,k}|$, by parameter counting arguments. □

Proof of Proposition 21. Consider the sufficient statistics $A=(\sigma_{\lambda,x})_{\lambda\in\Delta,x\in\mathcal{X}}$. We abbreviate $(\sigma_{\lambda,x})_{\lambda\in\Delta,x\in\mathcal{Y}}$ by $\sigma(\Delta,\mathcal{Y})$. A set $\mathcal{Y}\subseteq\mathcal{X}$ is an S-set of $\mathcal{E}$ if and only if *(i)* $\operatorname{rk}\sigma(\Delta,\mathcal{Y})=|\mathcal{Y}|$, (i. e., $\mathcal{Y}$ describes a $(|\mathcal{Y}|-1)$-simplex), and *(ii)* there exists a vector $c\in\mathbb{R}^{|\Delta|}$ for which $\langle c,\sigma(\Delta,y)\rangle=0\;\forall\,y\in\mathcal{Y}$ and $\langle c,\sigma(\Delta,x)\rangle\ge 1\;\forall\,x\in\mathcal{X}\setminus\mathcal{Y}$, (i. e., $\mathcal{Y}$ is a facial set). We show that $\mathcal{Y}$ satisfies these properties if and only if $x*\mathcal{Y}$ does. We have that

$$\sigma(\lambda,x*y) \;=\; (-1)^{|(\operatorname{supp}(x)\triangle\operatorname{supp}(y))\cap\lambda|}=(-1)^{|\operatorname{supp}(x)\cap\lambda|}(-1)^{|\operatorname{supp}(y)\cap\lambda|}$$
$$\forall\,x\in\mathcal{X},\;\lambda\in 2^{[N]},\;y\in\mathcal{X}$$

and thus $\sigma(\Delta,x*\mathcal{Y})=\operatorname{diag}(\sigma(\Delta,x))\cdot\sigma(\Delta,\mathcal{Y})$. Hence $\operatorname{rk}\sigma(\Delta,\mathcal{Y})=\operatorname{rk}\sigma(\Delta,x*\mathcal{Y})$. Consider, on the other hand, the vector $\tilde{c}:=\operatorname{diag}(\sigma(\Delta,x))\cdot c$. We have $\langle\tilde{c},\sigma(\Delta,x*y)\rangle=\langle c,\sigma(\Delta,y)\rangle=0\;\forall\,y\in\mathcal{Y}$, and $\langle\tilde{c},\sigma(\Delta,z')\rangle\ge 1\;\forall\,z'\in\mathcal{X}\setminus y*\mathcal{Y}$. □

ACKNOWLEDGMENT

I am grateful to Johannes Rauh, Thomas Kahle, and Nihat Ay for many valuable discussions and comments. Furthermore, I am grateful to Shun-ichi Amari for valuable discussions. I thank Jason Morton for help in proof-reading the manuscript. I am grateful to anonymous reviewers for very helpful suggestions. This work was carried out mostly at MPI-MIS, Leipzig, Germany; partly at RIKEN BSI, Hirosawa, Saitama, Japan; and partly at PennState, supported by DARPA grant FA8650-11-1-7145.

(Received October 27, 2011)

REFERENCES

[1] S. Amari: Information geometry on hierarchical decomposition of stochastic interactions. IEEE Trans. Inform. Theory *47* (1999), 1701–1711.

[2] S. Amari and H. Nagaoka: Methods of information geometry, Vol. *191*. Oxford University Press, 2000. Translations of mathematical monographs.

[3] N. Ay and A. Knauf: Maximizing multi-information. Kybernetika *42* (2006), 517–538.

[4] N. Ay, G. F. Montúfar, and J. Rauh: Selection criteria for neuromanifolds of stochastic dynamics. In: Advances in Cognitive Neurodynamics (III). Springer, 2011.

[5] C. M. Bishop: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York 2006.

[6] C. Bocci and L. Chiantini: On the identifiability of binary segre products. J. Algebraic Geom. *5* (2011).

[7] L. Brown: Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayworth 1986.

[8] M. V. Catalisano, A. V. Geramita, and A. Gimigliano: Secant varieties of $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ (n-times) are not defective for $n \geq 5$. J. Algebraic Geom. *20* (2011), 295–327.

[9] P. Diaconis: Finite forms of de Finetti's theorem on exchangeability. Synthese *36* (1977), 271–281.

[10] B. Efron: The geometry of exponential families. Ann. Statist. *6* (1978), 2, 362–376.

[11] S. L. G. Cohen, I. Honkala, and A. Lobstein: Covering Codes. Elsevier, 1997.

[12] D. Gale: Neighborly and cyclic polytopes. In: Convexity: Proc. Seventh Symposium in Pure Mathematics of the American Mathematical Society 1961, pp. 225–233.

[13] E. Gawrilow and M. Joswig: Polymake: a framework for analyzing convex polytopes. In: Polytopes – Combinatorics and Computation (G. Kalai and G. M. Ziegler, eds.), Birkhäuser 2000, pp. 43–74.

[14] D. Geiger, C. Meek, and B. Sturmfels: On the toric algebra of graphical models. Ann. Statist. *34* (2006), 1463–1492.

[15] E. Gilbert: A comparison of signalling alphabets. Bell System Techn. J. *31* (1052), 504–522.

[16] Z. Gilula: Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. Biometrika *66* (1979), 2, 339–344.

[17] B. Grünbaum: Convex Polytopes. Second edition. Springer-Verlag, New York 2003.

[18] M. Henk, J. Richter-Gebert, and G. M. Ziegler: Basic Properties of Convex Polytopes. CRC Press, Boca Raton 1997.

[19] S. Hoşten and S. Sullivant: Gröbner bases and polyhedral geometry of reducible and cyclic models. J. Combin. Theory Ser. A *100* (2002), 2, 277–301.

[20] T. Kahle: Neighborliness of marginal polytopes. Contrib. Algebra Geometry *51* (2010), 45–56.

[21] T. Kahle and N. Ay: Support sets of distributions with given interaction structure. In: Proc. WUPES'06, 2006.

[22] T. Kahle, W. Wenzel, and N. Ay: Hierarchical models, marginal polytopes, and linear codes. Kybernetika *45* (2009), 189–208.

[23] G. Kalai: Some aspects of the combinatorial theory of convex polytopes. 1993.

[24] J. F. C. Kingman: Uses of exchangeability. Ann. Probab. *6* (1978), 2, 183–197.

[25] B. G. Lindsay: Mixture models: theory, geometry, and applications. NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, 1995.

[26] G. McLachlan and D. Peel: Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley, 2000.

[27] G. F. Montúfar and N. Ay: Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. Neural Comput. *23* (2011), 5, 1306–1319.

[28] G. F. Montúfar, J. Rauh, and N. Ay: Expressive power and approximation errors of restricted Boltzmann machines. In: Advances in Neural Information Processing Systems *24* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), MIT Press, 2011, pp. 415–423.

[29] J. Rauh: Finding the Maximizers of the Information Divergence from an Exponential Family. Ph. D. Thesis, Universität Leipzig, 2011.

[30] J. Rauh, T. Kahle, and N. Ay: Support sets of exponential families and oriented matroids. Internat. J. Approximate Reasoning *52* (2011), 5, 613–626.

[31] R. Settimi and J. Q. Smith: On the geometry of Bayesian graphical models with hidden variables. In: Proc. Fourteenth conference on Uncertainty in artificial intelligence, UAI'98, Morgan Kaufmann Publishers 1998, pp. 472–479.

[32] I. Shemer.: Neighborly polytopes. Israel J. Math. *43* (1982), 291–311.

[33] D. Titterington, A. F. M. Smith, and U. E. Makov: Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, 1985.

[34] R. Varshamov: Estimate of the number of signals in error correcting codes. Dokl. Akad. Nauk SSSR *117* (1957), 739–741.

Guido Montúfar, Department of Mathematics, Pennsylvania State University, University Park, PA 16802. U.S.A.

e-mail: gfm10@psu.edu