Department of Applied Mathematics and Statistics,
University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark, www.math.ku.dk/ams

Niels Richard Hansen
September 8, 2006

Aspects of Algebraic Statistics

Prediction, algorithms, and geometry.

1 Introduction

Algebraic statistics provides an algebraic point of view on statistical models. In this note we consider only models for discrete (finite) multivariate data, and in this case the most clear relation to algebra is that a large number of models are naturally parameterized via polynomials or rational functions. This is closely related to the concept of log-linear models and exponential families for discrete variables. The statistical models can then be seen geometrically as varieties, that is, as solution sets to a set of polynomial equations. This will, however, play a minor role in the present note, where we focus on some computational issues – in particular in relation to predictions of unobserved variables via maximum a posteriori prediction. The ability to evaluate polynomials efficiently turns out to be important, and for predictions of unobserved variables we need to solve linear programming problems efficiently. The algebra provides a unification of these computational problems via so-called semiring homomorphisms. It also provides a deeper understanding of our ability to construct predictions, and to investigate how the predictor functions change as we vary the parameters in the models.

2 Discrete multivariate statistical models

Throughout we think in terms of observing a number of discrete random variables that are thought to be part of a larger set of discrete random variables, for which the joint distribution can be specified in a convenient way. In particular, we are interested in situations where we can formulate a nice statistical model for the entire ensemble of random variables. The distribution of the observed variables is by definition given by a marginalization procedure, but this rarely produces another nice statistical model. If there are many variables, the marginalization – or equivalently, the mere computation of the likelihood function for a given parameter – can seem quite difficult as we have to sum over a very large set. We are also interested in the computation of the conditional distribution of the unobserved variables given the observed – for a given parameter – and the prediction of the unobserved variables. Since we are working with discrete variables the prediction (we will use) is simply the realization of the unobserved variables with the highest probability in the conditional distribution given the observed variables. This is also called maximum

a posteriori prediction. We are of course interested in efficient algorithms for computing the predictions but also in understanding how different choices of parameters affect the predictions. There is no hope that there is a general efficient algorithm if the variables are allowed to have an arbitrary dependence structure, but for some classes of models efficient algorithms are available. The standard example is the popular class of hidden Markov chains, but more generally the class of models with a dependence structure given by a directed acyclic graph (DAG) is computationally tractable, see Cowell et al. (1999). Directed acyclic graphs have become quite popular as a means of describing dependence structures.

In the following we let $X_1, \ldots, X_n$ denote n variables with values in $E_1, \ldots, E_n$ respectively – all finite but potentially different sets. We also let $Y_1, \ldots, Y_m$ denote m additional variables with values in $F_1, \ldots, F_m$ respectively. We let $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_m)$, $E = E_1 \times \ldots \times E_n$ and $F = F_1 \times \ldots \times F_m$. We think in terms of observing X but not Y.

Thus we are considering a full sample space $E \times F$ and a full observation $(\mathbf{x}, \mathbf{y}) \in E \times F$ – however, we observe only $\mathbf{x}$. The distribution of (X, Y) is given by a vector of point probabilities $p(\mathbf{x}, \mathbf{y})$ in the unit simplex $\Delta(E \times F)$ in $\mathbb{R}^{E \times F}$.

$$\Delta(E \times F) = \{p \in \mathbb{R}^{E \times F} \mid \sum_{\mathbf{x} \in E, \mathbf{y} \in F} p(\mathbf{x}, \mathbf{y}) = 1, \quad p(\mathbf{x}, \mathbf{y}) \geq 0\}.$$

A straight forward – but important – observation is that the distribution of X is given by the marginalization

$$p_1(\mathbf{x}) = \sum_{\mathbf{y} \in F} p(\mathbf{x}, \mathbf{y}),$$

where the vector of point probabilities $p_1(\mathbf{x})$ lies in the unit simplex $\Delta(E)$ in $\mathbb{R}^E$. Moreover, the conditional distribution of the (unobserved) Y given $X = \mathbf{x}$ has point probabilities

$$p_2(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p_1(\mathbf{x})}.$$

It follows that if we observe $\mathbf{x}$ and want to predict the value of Y as the most probable observation in the conditional distribution of Y given $X = \mathbf{x}$, then the predictor is

$$\hat{\mathbf{y}}(\mathbf{x}) = \text{argmax}_{\mathbf{y} \in F} p_2(\mathbf{y}|\mathbf{x}) = \text{argmax}_{\mathbf{y} \in F} p(\mathbf{x}, \mathbf{y}).$$

One needs to choose some way of breaking ties or alternatively give a set-valued prediction in case of ties.

Thus if we want to compute the conditional distribution of Y given $X = \mathbf{x}$ we need to carry out a summation over F – to compute the normalizing constant. If we want to predict Y we need to carry out an optimization over F. Should we be interested in the probability of the predicted value of Y we need to carry out the summation as well as the optimization. Since the size of F is $|F_1| \times |F_2| \times \cdot \times |F_m|$, which is most likely a very large number (unless some of the F-sample spaces are trivial and consist of a single element,

the size of F is at least 2^m), the task of carrying out the summation or optimization seems to be computationally prohibitive for large m. For certain models this is *not* the case, and the content of this note is to show that there is a very close – and from an algebraic point of view – a very interesting algorithmic connection between the two problems.

2.1 Log-linear models

A log-linear model on $E \times F$ is specified by a map

$$T : E \times F \to \mathbb{N}_0^d.$$

Thus for each $(\mathbf{x}, \mathbf{y}) \in E \times F$, $T(\mathbf{x}, \mathbf{y}) = (t_1(\mathbf{x}, \mathbf{y}), \ldots, t_d(\mathbf{x}, \mathbf{y}))^t$ is a d-dimensional vector of non-negative integers. Sometimes we can think of the map as a huge $d \times (E \times F)$ matrix with each of the $E \times F$ columns being a d-dimensional integer vector. The map T is going to be the sufficient transformation for the statistical model, which is given by a parameter set $\Theta \subseteq \mathbb{R}_+^d$, and for $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$ the point probabilities are given as

$$p_\theta(\mathbf{x}, \mathbf{y}) = \frac{1}{c(\theta)} \prod_{i=1}^{d} \theta_i^{t_i(\mathbf{x},\mathbf{y})} = \frac{1}{c(\theta)} \theta^{T(\mathbf{x},\mathbf{y})},$$

where

$$c(\theta) = \sum_{\mathbf{x},\mathbf{y}} \theta^{T(\mathbf{x},\mathbf{y})}$$

is the normalizing constant. We use the multi-index notation $\theta^T = \theta_1^{t_1} \cdot \ldots \cdot \theta_d^{t_d}$ if $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ and $T = (t_1, \ldots, t_d) \in \mathbb{N}_0^d$. As we see from this expression, the logarithm of the point probabilities is linear in the parameter – hence the name, log-linear models. We also observe that this is an exponential family, but this will play no role in this note. The sufficient transformation T completely determines the structure of the model, e.g. the dependencies among the variables.

We observe that the computation of $p_{\theta,1}(\mathbf{x})$ can be written as

$$p_{\theta,1}(\mathbf{x}) = \sum_{\mathbf{y} \in F} \prod_{i=1}^{d} \theta_i^{t_i(\mathbf{x},\mathbf{y})},$$

which should be recognized as an evaluation of a d-variable polynomial in $\theta = (\theta_1, \ldots, \theta_d)$. We can also see that for the optimization we have

$$\begin{aligned} \hat{\mathbf{y}}_\theta(\mathbf{x}) &= \operatorname{argmax}_{\mathbf{y} \in F} p(\mathbf{x}, \mathbf{y}) \\ &= \operatorname{argmax}_{\mathbf{y} \in F} \prod_{i=1}^{d} \theta_i^{t_i(\mathbf{x},\mathbf{y})} \\ &= \operatorname{argmax}_{\mathbf{y} \in F} \sum_{i=1}^{d} \log(\theta_i) t_i(\mathbf{x}, \mathbf{y}). \end{aligned}$$

The last equality is seen to hold because log (or equivalently exp) is a monotonely increasing function. We see that the last formulation is actually an optimization of a linear function over a set of integer vectors. The linear function, given by the log-parameter $\log(\theta)$, is

$$\mathbb{R}^d \ni v = (v_1, \ldots, v_d) \mapsto \sum_{i=1}^{d} \log(\theta_i) v_i.$$

As should be clear from subsequent arguments in this note, the optimization over F above can be reformulated as an optimization of the linear function over the convex hull of the points $t_i(\mathbf{x}, \mathbf{y})$ in $\mathbb{R}^d$, which is then a linear programming problem.

For log-linear models, the problem of computing the conditional distribution is a problem of polynomial evaluation and the problem of prediction is a linear programming problem.

Example 2.1. Let $n = 0$ and take $F_i = \{0, 1\}^m$ (we don't observe anything). Consider the model with $d = 4$ and parameters

$$\theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} \in \mathbb{R}^4_+$$

given by (with $y_0 \in \{0, 1\}$)

$$p(y_1, \ldots, y_n) = \frac{1}{c_m^{y_0}(\theta)} \theta_{y_0 y_1} \cdot \theta_{y_1 y_2} \cdot \ldots \cdot \theta_{y_{m-1} y_m}$$

where

$$c_m^{y_0}(\theta) = \sum_{y_1, \ldots, y_m} \theta_{0 y_1} \cdot \theta_{y_1 y_2} \cdot \ldots \cdot \theta_{y_{m-1} y_m}$$

is the normalizing constant. Note that this looks like a Markov chain, and indeed, we call this model the *toric* Markov chain model on $\{0, 1\}^m$ (with initial distribution being degenerate in y_0). It follows from general results for graphical models (Proposition 3.28, Lauritzen (1996)), but see also the argument below, that under this model, $Y_1, \ldots, Y_m$ is actually a Markov chain, that is, the required conditional independence structure is present. We see that

$$c_m^{y_0}(\theta) = \sum_{y_1} \theta_{y_0 y_1} \left(\sum_{y_2} \theta_{y_1 y 2} \left(\ldots \left(\sum_{y_{m-1}} \theta_{y_{m-2} y_{m-1}} \left(\sum_{y_m} \theta_{y_{m-1} y_m} \right) \right) \right) \right),$$

and this recursive sum-product formula can be rewritten via matrix multiplications;

$$c_m(\theta) = \begin{pmatrix} 1 - y_0 \\ y_0 \end{pmatrix} \theta^m \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

where θ^m is the m'th matrix power of the matrix θ. If we introduce R_m as the diagonal matrix with the row sums of θ in the diagonal, and if we recursively let R_j, $j = m-1, \ldots, 1$

be the diagonal matrix with the row sums of θR_{j+1} in the diagonal, it can be verified that $P_j := R_j^{-1}\theta R_{j+1}$ is a stochastic matrix (positive entries with row sums equal to 1), and that P_j is the matrix of transition probabilities from y_j to y_{j+1}. So this is a Markov chain model – although the transition probabilities are not necessarily time homogeneous. They are, if the row sums of θ are equal. This model falls within the framework of log-linear models with $t_{yy'}(y_1, \ldots, y_m)$ being the number of transitions from y to y' in the sequence $y_1, \ldots, y_m$. Taking $m = 4$ and $y_0 = 0$ this gives the sufficient transformation:

	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111
t_{00}	4	3	2	2	2	1	1	1	2	1	0	0	1	0	0	0
t_{01}	0	1	1	1	1	2	1	1	1	2	2	2	1	2	1	1
t_{10}	0	0	1	0	1	1	1	0	1	1	2	1	1	1	1	0
t_{11}	0	0	0	1	0	0	1	2	0	0	0	1	1	1	2	3

◇

Example 2.2. Let $n = m$, $E_i = F_i = \{0, 1\}$ and $d = 16$. Thus $E \times F = \{0, 1\}^m \times \{0, 1\}^m$. If $x, y, x', y' \in \{0, 1\}$ define (with $x_0, y_0 \in \{0, 1\}$)

$$t_{xyx'y'}(\mathbf{x}, \mathbf{y}) = t_{xyx'y'}(x_1, \ldots, x_m, y_1, \ldots, y_m) = \sum_{k=0}^{m-1} 1(x_k = x, y_k = y, x_{k+1} = x', y_{k+1} = y'),$$

so $t_{xyx'y'}(X, Y)$ is the number of transitions from (x, y) to (x', y') in the (pair) sequence of random variables

$$(x_0, y_0), (X_1, Y_1), \ldots, (X_m, Y_m).$$

Then we consider the log-linear model given by T and with parameter space $\Theta = \mathbb{R}_+^{16}$, which is the (toric) Markov chain model for the (pair) sequence $(X_i, Y_i)_{i=1,\ldots,m}$. Observing only the X-coordinate does not preserve the Markov chain structure, and such a construction is a popular means for introducing a sequence of random variables that has a more complicated dependence structure than an ordinary Markov chain. As in the previous example, we get the class of time homogeneous Markov chains by restricting the parameter set to those positive 4×4 matrices with equal row sums. A further sub-model, which is what we normally refer to as the *hidden Markov chain model*, is given if we further restrict the parameter space to those matrices where

$$\theta_{xyx'y'} = \mu_{yy'}\nu_{y'x'}$$

where μ, ν are two 2×2 matrices with positive entries satisfying

$$\mu_{00} + \mu_{01} = \mu_{10} + \mu_{11} = 1, \quad \nu_{00} + \nu_{01} = \nu_{10} + \nu_{11} = 1.$$

It is possible to verify that this model can also be given as a sub-model of a log-linear model with $d = 8$ (try). The interpretation is that for this model the X variables are conditionally independent given the Y-variables, which in turn marginally form a Markov chain. This is seen directly from the factorization of the θ vector into a μ and a ν factor. Observe that there is a recursive sum-product formula for computing the marginal distribution of

$X_1, \ldots, X_n$;

$$\begin{aligned} p_{\mu,\nu,1}(x_1,\ldots,x_m) &= \sum_{y_1,\ldots,y_m} \mu_{y_0y_1}\nu_{y_1x_1}\mu_{y_1y_2}\nu_{y_2x_2}\cdots\mu_{y_{m-1}y_m}\nu_{y_mx_m} \\ &= \sum_{y_1}\mu_{0y_1}\nu_{y_1x_1}\left(\sum_{y_2}\mu_{y_1y_2}\nu_{y_2x_2}\left(\cdots\left(\sum_{y_m}\mu_{y_{m-1}y_m}\nu_{y_mx_m}\right)\right)\right). \quad (1) \end{aligned}$$

A similar formula holds for the full model with the $\theta_{xyx'y'}$-parameterization. If we introduce the following two diagonal matrices,

$$B(x) = \begin{pmatrix} \nu_{0x} & 0 \\ 0 & \nu_{1x} \end{pmatrix},$$

$x = 0, 1$, then we see from the formula above, that we can re-express $p_{\mu,\nu,1}$ in terms of matrix products;

$$p_{\mu,\nu,1}(x_1,\ldots,x_m) = \begin{pmatrix} 1-y_0 \\ y_0 \end{pmatrix} \mu B(x_1)\mu B(x_2)\ldots\mu B(x_m) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

$\diamond$

3 Algebra

This section provides a brief introduction to some of the algebraic concepts that are useful when studying statistical models for discrete data as considered in the previous section. What we will be particularly interested in is the study of certain semirings and semiring homomorphisms, and the main result of this section is the commutative diagram that we present in Theorem 3.18. It summarizes very well the insight we obtain from the algebraic structures.

3.1 Semirings

A semiring is a ring except from the fact that there is no inverse for the addition. A semiring, $(R, +, \cdot)$, is therefore a set, R, endowed with the two operations, $+$ (addition) and $\cdot$ (multiplication), such that the following rules hold for $x, y, z \in R$:

$$\begin{aligned} x + y &= y + x \\ (x + y) + z &= x + (y + z) \\ x \cdot (y \cdot z) &= (x \cdot y) \cdot z \\ x \cdot (y + z) &= x \cdot y + x \cdot z \\ (y + z) \cdot x &= y \cdot x + z \cdot x. \end{aligned}$$

Thus the addition is commutative and associative, and the multiplication is associative and distributive over the addition. In addition, we will assume that there is a zero element, 0, for addition and a unit element, 1, for multiplication, such that

$$\begin{aligned} 0 + x &= x + 0 = x \\ 1 \cdot x &= x \cdot 1 = x. \end{aligned}$$

Finally, a semiring is called commutative if the multiplication is commutative, that is, if

$$x \cdot y = y \cdot x.$$

It would be very convenient if the axioms, as they do in a ring, imply that

$$0 \cdot x = x \cdot 0 = 0,$$

but this is unfortunately not the case[1]. To avoid potential problems we assume that this rule holds, and as one can check, it holds for the semirings considered in this note.

Two of the important semirings in this note are:

Example 3.1. Let $\mathbb{R}_+ = [0, \infty)$ denote the positive real numbers including 0, then $(\mathbb{R}_+, +, \cdot)$ endowed with the usual arithmetic operations (addition and multiplication) is a commutative semiring. We note that 0 is the zero element and 1 is the unit element. $\diamond$

Example 3.2. Let $\mathbb{R}_m = \mathbb{R} \cup \{-\infty\}$ denote the real numbers with $-\infty$ joint. Endow this set with the following operations:

$$\begin{aligned} x \oplus y &= \max\{x, y\} \\ x \odot y &= x + y, \end{aligned}$$

where addition in the last formula is the usual arithmetic addition. Then $(\mathbb{R}_m, \oplus, \odot)$ is a commutative semiring. In this case $-\infty$ is the zero element and 0 is the unit element. $\diamond$

Definition 3.3. *Let $(R_1, +, \cdot)$ and $(R_2, \oplus, \odot)$ denote two semirings. A map, $\varphi : R_1 \to R_2$, is a (semiring) homomorphism if for all $x, y \in R_1$*

$$\varphi(x + y) = \varphi(x) \oplus \varphi(y)$$

and

$$\varphi(x \cdot y) = \varphi(x) \odot \varphi(y).$$

Example 3.4. We define a map $\lambda : \mathbb{R}_+ \to \mathbb{R}_m$ by

$$\lambda(x) = \begin{cases} -\infty & \text{if } x = 0 \\ 0 & \text{if } x > 0, \end{cases}$$

then λ is a homomorphism from $(\mathbb{R}_+, +, \cdot)$ to $(\mathbb{R}_m, \oplus, \odot)$. $\diamond$

[1] A three element (commutative) counter example, due to Kasper K. S. Andersen, can be given by taking $\mathbb{Z}/2$ with the usual ring structure of addition and multiplication and join an element δ such that $x + \delta = \delta + x = \delta \cdot x = x \cdot \delta = \delta$ for $x = 0, 1, \delta$.

3.2 Polynomials

For any commutative semiring $(R, +, \cdot)$ we introduce the set of polynomials in d-variables, denoted $R[d]$, with coefficients in R. This is formally a set of coefficients indexed by the set of positive integer vectors in $\mathbb{N}_0^d$ such that only finitely many of the coefficients are non-zero (do not equal the zero element in the semiring). We use the notation

$$p = \sum_{\alpha} c_\alpha \cdot x^\alpha$$

for a polynomial $p \in R[d]$. The summation is over vectors $\alpha \in \mathbb{N}_0^d$, often called multi-indices, and the coefficient of x^α is $c_\alpha \in R$ with only finitely many of these being different from 0. The notation x^α is at this stage purely formal – a place-holder for the coefficient c_α. When we start to talk about evaluating the polynomial below, the formalism will expand into the following expression:

$$x^\alpha = x_1^\alpha \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_d^{\alpha_d} = \underbrace{x_1 \cdot \ldots \cdot x_1}_{\alpha_1} \cdot \underbrace{x_2 \cdot \ldots \cdot x_2}_{\alpha_2} \cdot \ldots \cdot \underbrace{x_d \cdot \ldots \cdot x_d}_{\alpha_d}$$

for $x = (x_1, \ldots, x_d) \in R^d$ and $\alpha \in \mathbb{N}_0^d$.

We introduce two operations, addition and multiplication, on the set of polynomials $R[d]$ as follows. For $p = \sum_\alpha c_\alpha \cdot x^\alpha, q = \sum_\alpha b_\alpha \cdot x^\alpha \in R[d]$ we define

$$\begin{aligned} p + q &= \sum_{\alpha} (c_\alpha + b_\alpha) \cdot x^\alpha \\ p \cdot q &= \sum_{\gamma} a_\gamma x^\gamma, \qquad a_\gamma = \sum_{\substack{\alpha,\beta \\ \alpha+\beta=\gamma}} c_\alpha + b_\beta. \end{aligned}$$

The first important, structural observation is that these operations make $R[d]$ itself into a semiring. We leave the (boring) proof as an exercise.

Lemma 3.5. *With the addition and multiplication operations introduced above, $(R[d], +, \cdot)$ is a commutative semiring.*

For homomorphisms between semirings there is a natural way to "lift" the homomorphisms to be between the corresponding polynomial rings. The proof is straight forward.

Lemma 3.6. *If $\varphi : R_1 \to R_2$ is a homomorphism between the two semirings $(R_1, +, \cdot)$ and $(R_2, \oplus, \odot)$, then $\Phi : R_1[d] \to R_2[d]$ defined by*

$$\Phi\left(\sum_\alpha c_\alpha \cdot x^\alpha\right) := \bigoplus_\alpha \varphi(c_\alpha) \odot x^{\odot\alpha}$$

is a homomorphism from $(R_1[d], +, \cdot)$ to $(R_2[d], \oplus, \odot)$.

Example 3.7. With $\lambda : \mathbb{R}_+ \to \mathbb{R}_m$ the homomorphism defined in Example 3.4, we denote by $\Lambda : \mathbb{R}_+[d] \to \mathbb{R}_m[d]$ the corresponding, lifted homomorphism between the polynomial rings. The homomorphism works simply by replacing all non-zero coefficients in the polynomial from $\mathbb{R}_+[d]$ with the unit element from $\mathbb{R}_m$ (which is 0). ◇

Example 3.8. In the previous example we introduced the homomorphism $\Lambda : \mathbb{R}_+[d] \to \mathbb{R}_m[d]$. As an application we derive the binomial formula in $\mathbb{R}_m$ from the well known binomial formula. We know that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Taking $d = 2$ and applying the homomorphism Λ on both sides of the equality yields

$$(x \oplus y)^{\odot n} = \bigoplus_{k=0}^{n} x^{\odot k} \odot y^{\odot(n-k)}.$$

Note that the binomial coefficients disappear! The interpretation is that on the left hand side we compute n times the maximum of x and y (which of course is the maximum of nx and ny) by on the right hand side computing the maximum of $kx + (n-k)y$ for $k = 0, \ldots, n$. A slightly silly formula it seems – but never the less correct. ◇

Another homomorphism on polynomial semirings is the evaluation homomorphism.

Definition 3.9. *For $\omega \in R^d$ we define $\varepsilon_\omega : R[d] \to R$ as the evaluation of the polynomial in ω. Thus for $p = \sum_\alpha c_\alpha \cdot x^\alpha$,*

$$\varepsilon_\omega(p) = \sum_\alpha c_\alpha \cdot \omega_1^{\alpha_1} \cdot \cdots \cdot \omega_d^{\alpha_d} \in R$$

is the evaluation of p in ω. This is an element in R.

Example 3.10. If $p = \bigoplus_\alpha c_\alpha \odot x^{\odot\alpha} \in \mathbb{R}_m[d]$ with $c_\alpha \in \mathbb{R}_m$ and only finitely many of them $> -\infty$ we see that for $\omega \in \mathbb{R}_m$

$$\varepsilon_\omega(p) = \max_\alpha \sum_{i=1}^{d} \omega_i \alpha_i + c_\alpha.$$

If all the c_α's that are $> -\infty$ are $= 0$ (as when $p = \Lambda(p')$ with $p' \in \mathbb{R}_+[d]$), then $\varepsilon_\omega(p)$ is the maximum of the linear function given by ω over the set of integer points α with $c_\alpha = 0$. ◇

3.3 Convex polytopes

For the purpose of this note, a convex polytope in $\mathbb{R}^d$ is simply the convex hull of a finite number of vectors from $\mathbb{R}^d$. Thus if $v_1, \ldots, v_n \in \mathbb{R}^d$ are such a finite number of vectors, the convex hull is

$$\operatorname{conv}\{v_1, \ldots, v_n\} = \{\lambda_1 v_1 + \ldots + \lambda_n v_n \mid \lambda_i \geq 0, \quad \sum_{i=1}^{d} \lambda_i = 1\}.$$

We denote by $\mathrm{CP}[d]$ the set of d-dimensional convex polytopes. On this set we introduce two operations as follows; for $A, B \in \mathrm{CP}[d]$ define

$$A \oplus B := \mathrm{conv}(A, B) = \{\lambda v + (1-\lambda)w \mid \lambda \in [0,1],\ v \in A,\ w \in B\}$$

and

$$A \odot B := A + B = \{v + w \mid v \in A,\ w \in B\}.$$

The first operation, $\oplus$, is the convex hull of the two sets, that is the smallest convex set that contains both of the sets A and B. The second operation, $\oplus$, is called the Minkowski sum of the two sets, and is the set consisting of all sums of elements from the two sets.

Lemma 3.11. *If $A, B \in \mathrm{CP}[d]$ then $A \oplus B, A \odot B \in \mathrm{CP}[d]$. Moreover, with these two operations the set of convex polytopes forms a semiring with the empty set, $\emptyset$, as the zero element and the set $\{0\}$ consisting of 0 as the unit element.*

Remark 3.12. Why is the Minkowksi sum of two polytopes a polytope? Its easy to see that it is again a compact convex set, but which finite set of points span the sum? Recall that an *extremal point* of a convex set is a point that can not be written as a (non-degenerate) convex combinations of other points in the set. Polytopes obviously have only finitely many extremal points, and these are to be found among the points that span the polytope. Suppose that $u = v + w$ is an extremal point in $A \odot B$ for $A, B \in \mathrm{CP}[d]$ with $v \in A$ and $w \in B$, and suppose that v, say, is *not* extremal in A. Then $v = \lambda v_1 + (1-\lambda)v_2$ with $\lambda \in (0,1)$ and $v_1 \neq v_2$. Therefore $u_1 := v_1 + w \neq u_2 := v_2 + w$ and

$$u = \lambda u_1 + (1-\lambda)u_2$$

is a non-degenerate convex combination that gives u. Thus we have a contradiction, and v must be an extremal point after all. Likewise must w be extremal in B, and we conclude that the extremal points in the Minkowski sum $A \odot B$ are always sums of extremal points[2]. It is an important fact about compact convex sets that they are spanned by their extremal points, see e.g. Theorem 2.1.9 in Hörmander (1994). For the the Minkowski sum of polytopes we therefore know that it is spanned by (some) of the finite number of sums of extremal points.

We define a map from the semiring $\mathbb{R}_+[d]$ to the semiring $\mathrm{CP}[d]$ as follows: If $p = \sum_\alpha c_\alpha \cdot x^\alpha$ then

$$\mathrm{New}(p) = \mathrm{New}\left(\sum_\alpha c_\alpha \cdot x^\alpha\right) = \mathrm{conv}\{\alpha \mid c_\alpha \neq 0\}.$$

The polytope, $\mathrm{New}(p)$, spanned by the exponents in the polynomial with non-zero coefficients is called the Newton polytope associated with the polynomial.

Lemma 3.13. *The map*

$$\mathrm{New} : (\mathbb{R}_+[d], +, \cdot) \to (\mathrm{CP}[d], \oplus, \odot)$$

is a semiring homomorphism.

[2] the sum of two arbitrary extremal points is, however, not always extremal, but the extremal points need to found among such sums.

Proof: If $p = \sum_\alpha c_\alpha \cdot x^\alpha, q = \sum_\alpha b_\alpha \cdot x^\alpha \in \mathbb{R}_+[d]$, the sum $p+q$ has non-zero coefficients at all places where one of the polynomials have a non-zero coefficient (note the coefficients are all ≥ 0). Therefore

$$\begin{aligned} \mathrm{New}(p+q) &= \mathrm{conv}\left\{\alpha \mid c_\alpha \neq 0 \text{ or } b_\alpha \neq 0\right\} \\ &= \mathrm{conv}\left\{\mathrm{conv}\left\{\alpha \mid c_\alpha \neq 0\right\}, \mathrm{conv}\left\{\alpha \mid b_\alpha \neq 0\right\}\right\} \\ &= \mathrm{New}(p) \oplus \mathrm{New}(q). \end{aligned}$$

For the multiplication we note that the non-zero coefficients in $p \cdot q$ are those where the exponent is of the form $\alpha + \beta$ where $c_\alpha, b_\beta > 0$. Since all the coefficients are in $\mathbb{R}_+$ there are no terms that can cancel out, thus

$$\begin{aligned} \mathrm{New}(pq) &= \mathrm{conv}\left\{\alpha+\beta \mid c_\alpha \neq 0 \text{ and } b_\beta \neq 0\right\} \\ &= \mathrm{conv}\left\{\alpha \mid c_\alpha \neq 0\right\} \odot \mathrm{conv}\left\{\beta \mid b_\beta \neq 0\right\} \\ &= \mathrm{New}(p) \odot \mathrm{New}(q), \end{aligned}$$

where the second equality follows by the fact that all the elements $\alpha + \beta$ that span the convex polytope are by definition in the Minkowski sum of $\mathrm{New}(p)$ and $\mathrm{New}(q)$ and that the extremal points of the Minkowski sum (that spans the Minkowski sum) are to be found among the sums $\alpha + \beta$. This shows that New also preserves the multiplication and thus that New is a homomorphism. □

Example 3.14. Again we take a look at the binomial formula – this time in $\mathrm{CP}[d]$. Taking two polynomials $p, q \in \mathbb{R}_+[d]$, then

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k \cdot q^{n-k},$$

so applying the homomorphism New to both sides of the equality gives

$$(\mathrm{New}(p) \oplus \mathrm{New}(q))^{\odot n} = \bigoplus_{k=0}^{n} \mathrm{New}(p)^{\odot k} \odot \mathrm{New}(q)^{\odot(n-k)}.$$

This equality tells us that if we take the convex hull of $\mathrm{New}(p)$ and $\mathrm{New}(q)$ and then the Minkowski sum of this set with itself n times (the n'th Minkowski power), then the resulting polytope can be computed by first forming the $n+1$ Minkowski sums $\mathrm{New}(p)^{\odot k} \odot \mathrm{New}(q)^{\odot(n-k)}$ and then taking their convex hull. If $p = x^\alpha$ and $q = x^\beta$ for $\alpha, \beta \in \mathbb{N}_0^d$ then

$$\mathrm{New}(p) \oplus \mathrm{New}(q) = \mathrm{conv}\{\alpha, \beta\}$$

is the line segment from α to β. But since $\mathrm{New}(p) = \alpha$ and $\mathrm{New}(q) = \beta$ we see that $\mathrm{New}(p)^{\odot k} = k\alpha$ and $\mathrm{New}(q)^{\odot(n-k)} = (n-k)\beta$, thus

$$\mathrm{New}(p)^{\odot k} \odot \mathrm{New}(q)^{\odot(n-k)} = k\alpha + (n-k)\beta.$$

We see that $(\mathrm{New}(p) \oplus \mathrm{New}(q))^{\odot n}$ is the convex hull of the points $n\alpha + (n-k)\beta$ for $k = 0, \ldots, n$. Observe, however, that

$$k\alpha + (n-k)\beta = \frac{k}{n} n\alpha + \left(1 - \frac{k}{n}\right) n\beta,$$

thus all the points are actually convex combinations of the two points $n\alpha$ and $n\beta$. We conclude that if $p = x^\alpha$ and $q = x^\beta$ then $(\mathrm{New}(p) \oplus \mathrm{New}(q))^{\odot n}$ is the line segment from $n\alpha$ to $n\beta$. ◇

Example 3.15. We want to compute the Newton polytope for the polynomial

$$\begin{aligned} c_m^{y_0}(\theta) &= \sum_{y_1 \cdots y_m} \theta_{y_0 y_1} \cdot \ldots \cdot \theta_{y_{m-1} y_m} \\ &= \sum_{y_1} \theta_{y_0 y_1} \left(\sum_{y_2} \theta_{y_1 y_2} \left(\cdots \left(\sum_{y_{m-1}} \theta_{y_{m-2} y_{m-1}} \left(\sum_{y_m} \theta_{y_{m-1} y_m} \right) \right) \right) \right) \end{aligned}$$

that gives the normalizing constant in the toric Markov chain model in Example 2.1. Let

$$e_{00} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad e_{01} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad e_{10} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad e_{11} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

regarded as elements in $\mathbb{N}_0^4$, then since $\mathrm{New}(\theta_{yy'}) = e_{yy'}$ it follows by the homomorphism property of New that

$$\begin{aligned} V(y_0, m) &:= \mathrm{New}(c_m^{y_0}) \\ &= \bigoplus_{y_1} e_{y_0 y_1} \odot \left(\bigoplus_{y_2} e_{y_1 y_2} \odot \left(\cdots \odot \left(\bigoplus_{y_{m-1}} e_{y_{m-2} y_{m-1}} \odot \left(\bigoplus_{y_m} e_{y_{m-1} y_m} \right) \right) \right) \right). \end{aligned}$$

We see that this recursive formula can also be written as

$$\begin{aligned} V(0, m) &= (e_{00} \odot V(0, m-1)) \oplus (e_{01} \odot V(1, m-1)) \\ V(1, m) &= (e_{10} \odot V(0, m-1)) \oplus (e_{11} \odot V(1, m-1)). \end{aligned}$$

These two recursions can form the basis for an inductive proof that for $n \geq 1$

$$V(0, 2n) = \mathrm{conv}\left\{ \begin{pmatrix} 2n & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 2n-1 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & n \\ n-1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n-1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 2n-2 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 2n-1 \end{pmatrix} \right\}$$

$$V(0, 2n+1) = \mathrm{conv}\left\{ \begin{pmatrix} 2n+1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 2n & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & n \\ n & 0 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n & 1 \end{pmatrix}, \begin{pmatrix} 0 & n+1 \\ n & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 2n-1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 2n \end{pmatrix} \right\}$$

$$V(1, 2n) = \mathrm{conv}\left\{ \begin{pmatrix} 2n-1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 2n-2 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n & 0 \end{pmatrix}, \begin{pmatrix} 1 & n-1 \\ n & 0 \end{pmatrix}, \begin{pmatrix} 0 & n-1 \\ n & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 2n-1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 2n \end{pmatrix} \right\}$$

$$V(1, 2n+1) = \mathrm{conv}\left\{ \begin{pmatrix} 2n & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 2n-1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n+1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & n \\ n & 1 \end{pmatrix}, \begin{pmatrix} 1 & n \\ n & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 2n \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 2n+1 \end{pmatrix} \right\}.$$

The induction start is given by observing that $V(y_0, 1) = \text{conv}\{e_{y_0 0}, e_{y_0 1}\}$. Note that for $n = 1$ there is some redundancy in the representations above, but for $n \geq 2$ the Newton polytopes always have precisely these seven vertices. Actually writing out the details, that is, forming the four different steps in the induction proof and in each of these four steps computing first the 14 different *potential* vertices and then reducing them to the desired seven, is not a task for humans. You may compare this with the sufficient transformation given in Example 2.1 for $m = 4$. These 16 integer vectors span the convex polytope $V(0, 4)$, but only seven of them are vertices. It is also possible to get a geometrical understanding of the polytopes as the convex hull of the integer points contained in two parallel planes intersected by a 3-dimensional simplex. We refer to Chapter 10 in Pachter & Sturmfels (2005) and to Kuo (2006) for additional details. $\diamond$

For a given convex polytope $A \in \text{CP}[d]$ and a vector $\omega \in \mathbb{R}^d_m$ we let

$$\gamma_\omega(A) = \max_{v \in A} \sum_{i=1}^{d} \omega_i v_i$$

denote the maximum of the linear function defined by ω over the polytope A. This introduces a map

$$\gamma_\omega : \text{CP}[d] \to \mathbb{R}_m.$$

Linear programming is precisely the study of this map γ_ω and for given polytopes (or classes of polytopes) it is of interest to find efficient algorithms for the evaluation of the map. Often the polytopes are then given in an alternative form, namely via a number of linear inequalities. It is a main theorem, cf. Theorem 1.1. in Ziegler (1995), in the theory of polytopes that a bounded intersection of halfspaces (i.e. a bounded solution set to a system of linear inequalities) is a polytope as we have defined it (as the convex hull of a finite set of points) and vice versa. From our point of view we will just note the following fact about γ_ω.

Lemma 3.16. *The map*

$$\gamma_\omega : (\text{CP}[d], \oplus, \odot) \to (\mathbb{R}_m, \oplus, \odot)$$

is a semiring homomorphism.

Proof: Note that the maximum of a linear function over a line segment is attained at one of the ends, that is, if $v, w \in \mathbb{R}^d$ then

$$\max_{\lambda \in [0,1]} \left\{ \sum_{i=1}^{d} \omega_i (\lambda v_i + (1 - \lambda) w_i) \right\} = \max \left\{ \sum_{i=1}^{d} \omega_i v_i, \sum_{i=1}^{d} \omega_i w_i \right\}.$$

Therefore for $A, B \in \mathrm{CP}[d]$ we have that

$$\begin{aligned}
\gamma_\omega(A \oplus B) &= \max_{u \in A \oplus B} \left\{ \sum_{i=1}^{d} \omega_i u_i \right\} \\
&= \max_{v \in A, w \in B, \lambda \in [0,1]} \left\{ \sum_{i=1}^{d} \omega_i (\lambda v_i + (1-\lambda) w_i) \right\} \\
&= \max \left\{ \max_{v \in A} \left\{ \sum_{i=1}^{d} \omega_i w_i \right\}, \max_{w \in B} \left\{ \sum_{i=1}^{d} \omega_i v_i \right\} \right\} \\
&= \gamma_\omega(A) \oplus \gamma_\omega(B).
\end{aligned}$$

Likewise

$$\begin{aligned}
\gamma_\omega(A \odot B) &= \max_{v \in A, w \in B} \left\{ \sum_{i=1}^{d} \omega_i (v_i + w_i) \right\} \\
&= \max_{v \in A} \left\{ \sum_{i=1}^{d} \omega_i v_i \right\} + \max_{w \in B} \left\{ \sum_{i=1}^{d} \omega_i w_i \right\} \\
&= \gamma_\omega(A) \odot \gamma_\omega(B).
\end{aligned}$$

This shows that γ_ω is a semiring homomorphism. □

Remark 3.17. It is noteworthy to remember that $\gamma_\omega(A)$ is always attained at one of the extremal points of A because a linear function restricted to a line segment always attains its maximum at one of the ends.

3.4 The commutative diagram

The acquired wisdom from the previous sections on polynomial and polytope semirings can be summarized in the following commutative diagram.

Theorem 3.18. *For any fixed $\omega \in \mathbb{R}_m^d$ the following diagram*

$$\begin{array}{ccc}
\mathbb{R}_+[d] & \xrightarrow{\text{New}} & \mathrm{CP}[d] \\
{\scriptstyle \Lambda}\downarrow & & \downarrow{\scriptstyle \gamma_\omega} \\
\mathbb{R}_m[d] & \xrightarrow[\varepsilon_\omega]{} & \mathbb{R}_m
\end{array}$$

is a commutative diagram of semiring homomorphisms.

Proof: We have showed that all the maps that enter are homomorphisms, so all we need is to show that the diagram is commutative. However, for the evaluation in ω of a polynomial in $\mathbb{R}_m[d]$, see Example 3.10, we know that in the image of Λ this is precisely the

maximization of the linear function given by ω over the set of exponents in the polynomial. Thus, if $p = \sum_\alpha c_\alpha \cdot x^\alpha$ then

$$\begin{aligned} \varepsilon_\omega \circ \Lambda(p) &= \max_{\alpha : c_\alpha \neq 0} \left\{ \sum_{i=1}^{d} \omega_i \alpha_i \right\} \\ &= \max_{v \in \mathrm{New}(p)} \left\{ \sum_{i=1}^{d} \omega_i v_i \right\} \\ &= \gamma_\omega \circ \mathrm{New}(p). \end{aligned}$$

The second equality follows from the fact discussed in Remark 3.17 that the maximum over a convex polytope equals the maximum over the extremal points of the polytope, and the extremal points of $\mathrm{New}(p)$ are indeed to be found among the points α with $c_\alpha \neq 0$ that span $\mathrm{New}(p)$. □

A commutative diagram is a nice algebraic result, but the real understanding lies in the interpretation of the diagram. It tells us how we can translate (via the homomorphisms) an efficient algorithm from the semiring $\mathbb{R}_+[d]$ to the other rings in the diagram. By an efficient algorithm we mean algorithms that use as few algebraic operations as possible. If we want to minimize the actual computation time, one needs to know the precise complexity of the two different algebraic operations, but in many cases we talk about reducing e.g. an exponentially growing number of additions to a polynomially growing number of additions and multiplications. The crux of the matter is that a homomorphism preserves the semiring operations. Thus if you can write a polynomial in $\mathbb{R}_+[d]$ as (most likely recursive) sums and products of more elementary polynomials, any homomorphism to another ring preserves this structure. Such a rewriting of the polynomial is often referred to as an algorithm, and thus the homomorphisms transfer algorithms between semirings. Instead of computing the Newton polytope of a polynomial by collecting the (huge) number of exponents and taking the convex hull, we may be able to compute the Newton polytope by a (recursive) application of Minkowski sums and convex hull operations on more simple polytopes. Likewise, we can find the prediction by a (recursive) application of additions and maximizations[3].

Example 3.19. We continue Example 2.2 using the μ-ν-parameterization. Note that the eight μ and ν parameters take the role as the variables in the polynomial semiring, and we consider $p_{\mu,\nu,1}(x_1, \ldots, x_m)$ as a polynomial in $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}, \nu_{00}, \nu_{01}, \nu_{10}, \nu_{11}$. To emphasize this point, we let

$$q_{\mathbf{x}}(\mu, \nu) = p_{\mu,\nu,1}(\mathbf{x})$$

[3]at least we can compute the maximum. It is often possible rather easily to find the maximizer more or less directly from the algorithm. It is, however, not so clear how to formulate this fact from the point of view of the algebra we consider here.

denote this polynomial. We introduce,

$$e_{000} = \left(\begin{array}{cc} 1 & 0 \\ 0 & 0 \\ \hline 1 & 0 \\ 0 & 0 \end{array}\right), \quad e_{001} = \left(\begin{array}{cc} 1 & 0 \\ 0 & 0 \\ \hline 0 & 1 \\ 0 & 0 \end{array}\right), \quad e_{010} = \left(\begin{array}{cc} 0 & 1 \\ 0 & 0 \\ \hline 0 & 0 \\ 1 & 0 \end{array}\right), \ \ldots \ e_{111} = \left(\begin{array}{cc} 0 & 0 \\ 0 & 1 \\ \hline 0 & 0 \\ 0 & 1 \end{array}\right),$$

Disregarding how these zeroes and ones are organized (as matrices – just to show the systematic), the e_{ijk}'s are regarded as 8-dimensional vectors with 0-1-entries, thus as elements in $\mathbb{N}_0^8$. We also see that for the monomial $\mu_{yy'}\nu_{y'x}$ we have

$$\mathrm{New}(\mu_{yy'}\nu_{y'x}) = e_{yy'x}.$$

Applying New to $q_{\mathbf{x}}$ and using the formula (1) therefore gives that

$$\mathrm{New}(q_{\mathbf{x}}) = \bigoplus_{y_1} e_{y_0 y_1 x_1} \odot \left(\bigoplus_{y_2} e_{y_1 y_2 x_2} \odot \left(\ldots \odot \left(\bigoplus_{y_{m-1}} e_{y_{m-2} y_{m-1} x_m} \odot \left(\bigoplus_{y_m} e_{y_{m-1} y_m x_m} \right) \right) \right) \right).$$

From the formula we see that we have to make a convex hull of (two) points, then Minkowski adding *points* to the two polytopes, which are simple translations in space of the convex polytopes, then again convex hull operations (of two polytopes each time), translations, convex hull, etc. Each time we have two polytopes, we make four translations, and we put the resulting four polytopes together two and two by the convex hull operation giving two polytopes again. We see that the crucial operation is the convex hull operation, and an efficient implementation of this is needed if we want to carry out the actual computation of the Newton polytope $\mathrm{New}(q)$. ◇

4 The few prediction functions theorem

For a given log-linear model with sufficient statistic T and parameter space Θ, the prediction function $\hat{\mathbf{y}}_\theta : E \to F$ for $\theta \in \Theta$ assigns to any observed $\mathbf{x} \in E$ a prediction $\mathbf{y} \in F$. The subject of this section is to study the number of prediction functions that we obtain by varying $\theta \in \Theta$. The approach is essentially a matter of transforming the problem into a problem of counting vertices of Newton polytopes. It is, however, a slightly tricky matter to handle the θ's that give non-unique predictions. Remember that in such cases we can regard the predictor function as set valued – predicting the *set* of optimal y's – or alternatively we have selected an a priori (parameter independent) choice function, which assign to any set of (optimal) y's a unique representation. This last case is sometimes referred to as having a "consistent way of breaking ties". We will prefer to think in this way to make the predictor function take values in F, but we will *not* be interested in all the different ways to break ties, i.e. to select a choice function. For *any* such given choice function we are interested in the number of prediction functions we can obtain by varying θ – or more precisely an upper bound of this number.

First we fix an observation $\mathbf{x} \in E$ and give bounds on the number of elements in the set

$$\Pr(\mathbf{x}) = \{\hat{\mathbf{y}}_\theta(\mathbf{x}) \mid \theta \in \Theta\}$$

in terms of T. Second, we give a similar result on the total number of prediction functions in

$$\Pr = \{\hat{\mathbf{y}}_\theta : E \to F \mid \theta \in \Theta\}.$$

The noteworthy fact is that the bounds we obtain are typically much smaller – and of polynomial order in n and m – than the total number of functions from E to F (which is $|F|^{|E|}$, and for $E_i = F_i = \{0, 1\}$ this number is 2^{m2^n}, which grows doubly exponentially).

In the following let $T : E \times F \to \mathbb{N}_0^d$ be the sufficient statistic and

$$m(T) = \max_{i,\mathbf{x},\mathbf{y}} t_i(\mathbf{x}, \mathbf{y})$$

denote the maximal coordinate value for this statistic. This implies that any of the x-polytopes,

$$A(\mathbf{x}) = \text{conv}\{T(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in F\},$$

given by T is contained in the cube $[0, m(T)]^d$, which contains $(m(T) + 1)^d$ integer points.

Theorem 4.1. *For all $x \in E$*

$$|\Pr(\mathbf{x})| \leq 1 + \sum_{j=0}^{d-1} \binom{(m(T)+1)^d}{j},$$

and if $m(T) \to \infty$ we find that the dominant term is $(m(T) + 1)^{d(d-1)}/(d - 1)!$, thus

$$|\Pr(\mathbf{x})| = O(m(T)^{d(d-1)}).$$

Theorem 4.2. *For the number of prediction functions we have the following bound;*

$$|\Pr| \leq 2 \sum_{l=0}^{d-1} \binom{(2m(T)+1)^d}{l} \sum_{j=0}^{d-l-1} \binom{(2m(T)+1)^d - l - 1}{j}.$$

If $m(T) \to \infty$ the dominant term in the upper bound is of order $m(T)^{d(d-1)}$, thus for $m(T) \to \infty$

$$|\Pr| = O(m(T)^{d(d-1)}).$$

To prove these results we need to develop a few tools. A convex subset, $F \subseteq A$, of a polytope A is called a *face* if for any non-degenerate convex combination, $x = \lambda_1 x_1 + \ldots \lambda_l x_l$, with $x_1, \ldots, x_l \in A$, $\lambda_i > 0$, and $\sum_i \lambda_i = 1$, it holds that if $x \in F$ then $x_1, \ldots, x_l \in F$. Thus elements in a face, F, of A can not be a non-degenerate convex combination including points from A outside of F. Moreover, we see that since any element in A is a convex combination of the extremal points, a face must contain a subset of the extremal points, $f_1, \ldots, f_r$, say. Therefore, since F is convex, it also contains the convex hull of $f_1, \ldots, f_r$.

Finally we conclude that the face must actually equal this convex hull, for if any point x outside this convex hull is in the face F, then x is a non-degenerate convex combination of extremal points with at least one not being in $\{f_1, \ldots, f_r\}$, and this leads to a contradiction. We conclude that a face of A is always the convex hull of extremal points in A, and as such convex polytopes themselves, but we also note that it is certainly not all convex hulls of extremal points that are faces. For a face F of A we define the dimension of the face as the dimension of the affine subspace of $\mathbb{R}^d$ that the polytope F spans. We note that the 0-dimensional faces of A are precisely the extremal points. These faces are also called vertices. Likewise, the 1-dimensional faces are called edges. Counting (or bounding) the number of faces of any dimension for a polytope is paramount for bounding the number of prediction functions. In Appendix A we give some of the useful such bounds.

Lemma 4.3. *For any linear function*

$$v \mapsto \omega^t v = \sum_{i=1}^{d} \omega_i v_i$$

with $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ the subset of A where this function attains its maximum is a face of A.

Proof: Let F denote the subset of A where the linear function attains its maximum. Its not empty, as this is a maximization of a continuous function over a compact set, and it is trivially convex being an intersection of a hyperplane and A. Take $x \in F$ and let

$$x = \lambda_1 x_1 + \ldots \lambda_l x_l$$

with $x_1, \ldots, x_l \in A$, $\lambda_i > 0$, and $\sum_i \lambda_i = 1$. Since $\omega^t x$ is maximal over A we have $\omega^t x_i \leq \omega^t x$ for $i = 1, \ldots, l$ and if we have strict inequality in any case, the linearity of the function gives a contradiction since all the λ_i's are strictly positive. Thus $x_i \in F$ for $i = 1, \ldots, l$ and F is a face. □

Proof of Theorem 4.1: By Lemma 4.3 the linear function

$$(v_1, \ldots, v_d) \mapsto \sum_{i=1}^{d} \log(\theta_i) v_i$$

for any $\theta \in \Theta$ takes its maximum on a face of the polytope

$$A(\mathbf{x}) = \text{conv}\{T(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in F\}.$$

The total number of faces of $A(x)$ therefore bounds the number of different prediction functions. Letting $n(\mathbf{x})$ denote the number of vertices of $A(\mathbf{x})$ we get from Lemma A.1 that

$$|\text{Pr}(\mathbf{x})| \leq 1 + \sum_{j=0}^{d-1} \binom{n(\mathbf{x})}{j},$$

since $\binom{n(\mathbf{x})}{j}$ bounds the number of j-dimensional faces for $j = 0, \ldots, d-1$ and there is (at most) one d-dimensional face. Since $A(\mathbf{x}) \subseteq [0, m(T)]^d$ the maximal number of (integer) vertices that $A(\mathbf{x})$ can have is $(m(T)+1)^d$. □

Remark 4.4. Note that for fixed $x \in E$ we can obtain a (potentially smaller) x-dependent bound by replacing $m(T)$ with $\max_{i,\mathbf{y}} t_i(\mathbf{x}, \mathbf{y})$. It is quite likely that one can obtain better bounds where the a growth for $m(T) \to \infty$ is of a lower polynomial order. It will not change the fact that the theorem provides a polynomial – though most likely pessimistic – bound in terms of $m(T)$.

Lemma 4.3 implies that for the linear function given by $\omega \in \mathbb{R}^d$ there is a "face" function

$$\text{face}_\omega : \text{CP}[d] \to \text{CP}[d]$$

such that $\text{face}_\omega(A)$ for $A \in \text{CP}[d]$ is the face of A where the linear function takes its maximum. The face function works well together with Minkowski addition.

Lemma 4.5. *If $A, B \in \text{CP}[d]$, then for any $\omega \in \mathbb{R}^d$*

$$\text{face}_\omega(A \odot B) = \text{face}_\omega(A) \odot \text{face}_\omega(B).$$

Moreover, the summands on the right hand side are unique in the sense that if

$$\text{face}_\omega(A) \odot \text{face}_\omega(B) = \text{face}_{\omega'}(A) \odot \text{face}_{\omega'}(B)$$

then $\text{face}_\omega(A) = \text{face}_{\omega'}(A)$ *and* $\text{face}_\omega(B) = \text{face}_{\omega'}(B)$.

Proof: Suppose that $v \in \text{face}_\omega(A)$ and $w \in \text{face}_\omega(B)$, then $\omega^t(v+w) = \omega^t v + \omega^t w \geq \omega^t v' + \omega^t w' = \omega^t(v'+w')$ for any $v' \in A$ and $w' \in B$, which show "$\supseteq$". If $u \in \text{face}_\omega(A \odot B)$ we can write $u = v + w$ where $v \in A$ and $w \in B$, and then if $u' = v' + w' \in \text{face}_\omega(A) \odot \text{face}_\omega(B)$ we have

$$\omega^t v + \omega^t w = \omega^t u \geq \omega^t u' = \omega^t v' + \omega^t w'$$

but since v' and w' are in the ω-faces of A and B respectively, we must have equality throughout. This shows "$\subseteq$". Suppose then that

$$\text{face}_\omega(A) \odot \text{face}_\omega(B) = \text{face}_{\omega'}(A) \odot \text{face}_{\omega'}(B),$$

so for any $v' \in \text{face}_{\omega'}(A)$ and $w' \in \text{face}_{\omega'}(B)$ we have $v' + w' = v + w$ where $v \in \text{face}_\omega(A)$ and $w \in \text{face}_\omega(B)$. Since $v' \in A$ and $w' \in B$ we have $\omega^t v \geq \omega^t v'$ and $\omega^t w \geq \omega^t w'$ but

$$\omega^t v + \omega^t w = \omega^t v' + \omega^t w',$$

which shows that $v' \in \text{face}_\omega(A)$ and $w' \in \text{face}_\omega(B)$. Interchanging the roles of ω and ω' shows the other way around, and we conclude that the face summands on both sides are unique. □

Proof of Theorem 4.2: We show below that

$$|\text{Pr}| \leq \sum_{j=0}^{d} f_j \left(\bigodot_{\mathbf{x} \in E} A(\mathbf{x}) \right), \tag{2}$$

that is, the number of prediction functions is bounded by the total number of faces in the Minkowski sum $\bigodot_{\mathbf{x}\in E} A(\mathbf{x})$. Since any $A(\mathbf{x}) \subseteq [0, m(T)]^d$ the number of non-parallel edges among all the summands does not exceed $(2m(T)+1)^d$. Theorem A.4 (in the Appendix) and 2 gives the desired bound on $|\Pr|$. Since the bound is a polynomial in $(2m(T)+1)^d$ of degree $d-1$ (check), it follows that the dominant term is of order $m(T)^{d(d-1)}$ for $m(T) \to \infty$.

We need to show (2). Observe that the prediction functions for θ and θ' in Θ are equal if $\mathrm{face}_{\log(\theta)}(A(\mathbf{x})) = \mathrm{face}_{\log(\theta')}(A(\mathbf{x}))$ for all $\mathbf{x} \in E$. If

$$\bigodot_{\mathbf{x}\in E} \mathrm{face}_{\log(\theta)}\left(A(\mathbf{x})\right) = \bigodot_{\mathbf{x}\in E} \mathrm{face}_{\log(\theta')}\left(A(\mathbf{x})\right),$$

Lemma 4.5 implies that the summands that enter are unique, and if θ and θ' therefore give rise to the same face on $\bigodot_{\mathbf{x}\in E} A(\mathbf{x})$, they give rise to the same face on all of the polytopes $A(\mathbf{x})$, and the corresponding prediction functions are equal. We conclude that there can be at most as many prediction functions as there are faces of $\bigodot_{\mathbf{x}\in E} A(\mathbf{x})$, and this shows (2). □

Remark 4.6. From the bound on $|\Pr(x)|$ one obtains the trivial bound on $|\Pr|$ to be

$$\prod_{x\in E} |\Pr(x)| \leq \left(1 + \sum_{j=0}^{d-1} \binom{(m(T)+1)^d}{j}\right)^{|E|}.$$

It is quite remarkable that the much better bound obtained above for $|\Pr|$ is of the same order for $m(T) \to \infty$ as the bound on each of the factors $|\Pr(x)|$ themselves.

Example 4.7. Consider the setup in Example 2.2, where it is clear by the definition of T that $m(T) = m$. Thus in the full model we have $d = 16$ and the bound we get on the number of prediction functions is at most of the order $m^{16\cdot 15} = m^{240}$, which is hardly impressive. For the hidden Markov model, which is a sub-model, we can take $d = 8$ and get at most of the order $m^{8\cdot 7} = m^{56}$ prediction functions, which is better, but still a terrible polynomial growth order. ◇

Remark 4.8. Note that some bounds found in the literature, e.g. Chapter 9 in Pachter & Sturmfels (2005), Pachter & Sturmfels (2004*b*) and Pachter & Sturmfels (2004*a*), seem to be better. These bounds, or at least their proofs, are dubious, as they do not really take into account the non-uniqueness problems for the prediction functions, and insist that it is enough to count vertices. For the bound on $|\Pr(\mathrm{x})|$ this is true *if* we make the extra assumption that the choice function that resolves the non-uniqueness always chooses a prediction that corresponds to a vertex. Assuming so we can bound $|\Pr(\mathrm{x})|$ by $(m(T)+1)^d$ and a slightly better bound can be found as Theorem 7 in Pachter & Sturmfels (2004*b*). For the bound obtained in Theorem 4.2 it is not so easy to obtain a bound where we count vertices only. This is claimed in Chapter 9 in Pachter & Sturmfels (2005) in the proof of Theorem 9.3, but the problem is that even if we resolve the non-uniqueness by choosing vertices always, one cannot combine several polytopes (via Minkowski addition) as we do

in proof without the potential risk that some of the j-dimensional faces for $j \geq 1$ introduce prediction functions that do not correspond to any of the vertices of the Minkowski sum. Note, however, that the asymptotic growth rate obtained in Theorem 4.2 is the same as in Theorem 9.3 in Pachter & Sturmfels (2005).

A Counting faces

We present a few results here – without proof – on how to count or bound the number of faces for a convex polytopes. Let $f_j(A)$ denote the number of faces of dimension j for $j = 0, \ldots, d-1$ for a polytope $A \in \mathrm{CP}[d]$. A polytope is called *simplicial* if every face of dimension $< d$ is a simplex. A j-dimensional simplex (in $\mathbb{R}^d$ for $j \leq d$) is by definition the convex hull of $j+1$ affinely independent points, see Ziegler (1995), Example 0.3. Thus for a simplicial polytope, any j-dimensional face with $j < d$ has $j+1$ vertices, which are also vertices of A, thus for simplicial polytopes there is the natural upper bound

$$f_j(A) \leq \binom{n}{j+1}$$

for $j = 0, 1, \ldots, d-1$. By Lemma 8.24 in Ziegler (1995) the face-numbers for any polytope can be bounded by the face-numbers of a simplicial polytope with the same number of vertices. This gives the following result:

Lemma A.1. *For any polytope $A \in \mathrm{CP}[d]$ with n vertices it holds that*

$$f_j(A) \leq \binom{n}{j+1}$$

for $j = 0, 1, \ldots, d-1$.

The other results in this appendix deal with the face-numbers for Minkowski sums of polytopes. First, a precise formula for the number of vertices in a so-called *zonotope* is given. A zonotope is by definition the Minkowski sum of a number of line segments. Thus if $v_i, w_i \in \mathbb{R}^d$, $i = 1, \ldots, m$, and

$$A_i = \{\lambda v_i + (1-\lambda) w_i \mid \lambda \in [0,1]\},$$

then

$$Z := A_1 \odot \cdots \odot A_m$$

is a zonotope. We say that the line segments A_i, $i = 1, \ldots, m$, in $\mathbb{R}^d$ are in *general position* if any subset of at most d of the line segments forms a linearly independent set of lines.

Theorem A.2 (Gritzmann & Sturmfels (1993))**.** *If the line segments $A_1, \ldots, A_m$ are in general position, then the number of l-dimensional faces for the zonotope Z is*

$$2\binom{m}{l} \sum_{j=0}^{d-l-1} \binom{m-l-1}{j}.$$

Remark A.3. Note that if the dimension d is larger than or equal to the number of line segments in the zonotope, the formula for vertices ($l = 0$) simply reads

$$2\sum_{j=0}^{d-1}\binom{m-1}{j} = 2\sum_{j=0}^{m-1}\binom{m-1}{j} = 2(1+1)^{m-1} = 2^m, \quad d \geq m.$$

Thus if $d \geq m$ each addition of a line doubles the number of vertices.

Zonotopes generated from line segments in general position represent "the worst" polytopes that can arise as Minkowski sums, in the sense that their face-numbers grow as fast as possible in terms of the number of edges that enter in the Minkowski sum. Remember that the edges of a polytope are the 1-dimensional faces. Theorem 2.1.10 in Gritzmann & Sturmfels (1993) states that:

Theorem A.4. *If $A_1, \ldots, A_k$ are polytopes in* CP$[d]$*, and if m denotes the number of non-parallel edges among all the edges of the polytopes, then the number of l-dimensional faces of their Minkowski sum $A_1 \odot A_2 \odot \cdots \odot A_k$ is bounded by*

$$2\binom{m}{l}\sum_{j=0}^{d-l-1}\binom{m-l-1}{j}$$

References

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999), *Probabilistic networks and expert systems*, Statistics for Engineering and Information Science, Springer-Verlag, New York.

Gritzmann, P. & Sturmfels, B. (1993), 'Minkowski addition of polytopes: computational complexity and applications to Gröbner bases', *SIAM J. Discrete Math.* **6**(2), 246–269.

Hörmander, L. (1994), *Notions of convexity*, Vol. 127 of *Progress in Mathematics*, Birkhäuser Boston Inc., Boston, MA.

Kuo, E. H. (2006), 'Viterbi sequences and polytopes', *J. Symbolic Comput.* **41**(2), 151–163.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17 of *Oxford Statistical Science Series*, The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

Pachter, L. & Sturmfels, B. (2004*a*), 'Parametric inference for biological sequence analysis', *Proc. Natl. Acad. Sci. USA* **101**(46), 16138–16143 (electronic).

Pachter, L. & Sturmfels, B. (2004*b*), 'Tropical geometry of statistical models', *Proc. Natl. Acad. Sci. USA* **101**(46), 16132–16137 (electronic).

Pachter, L. & Sturmfels, B., eds (2005), *Algebraic statistics for computational biology*, Cambridge University Press, New York.

Ziegler, G. M. (1995), *Lectures on polytopes*, Vol. 152 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.