Exponential Families and Semigroups

Steffen Lauritzen, University of Oxford

Harmony of Gröbner Bases and the Modern Industrial Society

Osaka, Japan, 2010

Let $X_1, \ldots, X_n$ be independent and identically distributed with density $f(x;\theta)$ and $\theta \in \Theta$ unknown.

Fisher (1925) showed that if $t_n(x_1, \ldots, x_n)$ is minimal sufficient, it satisfies

$$t_{m+n}(x_1, \ldots, x_{m+n}) = \phi_{mn}\{t_m(x_1, \ldots, x_m), t_n(x_{m+1}, \ldots, x_{m+n})\}$$

so that *the statistic of a combined sample can be found from those of individual subsamples.*

This fundamental property of *recursive computability* has received less attention than it deserves. It rules out many statistics as minimal sufficient, as Fisher also points out. For example, the *median cannot be a minimal sufficient statistic.*

Using the exchangeability of $X_1, \ldots, X_n$ it follows that

$$\begin{aligned} \phi_{mn}(t_m, t_n) &= \phi_{nm}(t_n, t_m) \\ \phi_{l,m+n}\{t_l, \phi_{mn}(t_m, t_n)\} &= \phi_{l+m,n}\{\phi_{lm}(t_l, t_m), t_n\} \end{aligned}$$

which in essence says that *minimal sufficient statistics combine as Abelian semigroups.* So we can write

$$t_{m+n}(x_1, \ldots, x_{m+n}) = t(x_1, \ldots, x_m) \oplus t(x_{m+1}, \ldots, x_{m+n})$$

and eventually

$$t_{m+n}(x_1, \ldots, x_{m+n}) = t(x_1) \oplus \cdots \oplus t(x_{m+n})$$

where $\oplus$ denotes a semigroup sum.

Exponential family theory has traditionally focused on the case where $t(x) \in \mathbb{R}^d$, and $\oplus$ is vector addition, but this is not necessarily the most appropriate.

Terminology here essentially follows Clifford and Preston (1961). $(S, +)$ is an *Abelian semigroup with unit* if the composition $+$ is associative and commutative so that

$$s + t = t + s, \quad (s + t) + u = s + (t + u) \quad e + s = s + e = s,$$

where e is the unit.

Examples are $(\mathbb{N}_0, +)$, $(\mathbb{N}, \max)$, $(\mathbb{N} \cup \{\infty\}, \min)$, $(\mathbb{R}_0^+, \cdot)$, $(\mathbb{Z}, +)$, $(\mathbb{R}, \cdot)$, ...

If the semigroup does not have a unit, one can always be adjoined, so there is no real loss of generality. In the following we shall say semigroup for semigroup with unit.

The *free semigroup* $\mathbb{F}(\mathcal{X})$ over a set $\mathcal{X}$ consists of all 'frequency tables' or 'histograms' $(\nu_x, x \in \mathcal{X})$ where $\nu_x \in \mathbb{N}_0$ and $\sum_x \nu_x < \infty$, composed in the obvious way

$$(\mu_x, x \in \mathcal{X}) + (\nu_x, x \in \mathcal{X}) = (\mu_x + \nu_x, x \in \mathcal{X}).$$

So essentially this is where all statistics begins: the statistician looks at a collection of objects of some type and counts how may objects there are of each type.

In other words, the *empirical distribution* of a sample lives in the free semigroup.

A (non-negative) *character* ρ on $(S, +)$ is a homomorphism into $(\mathbb{R}_0^+, \cdot)$ with $\rho(e) = 1$, i.e. they satisfy

$$\rho(s + t) = \rho(s)\rho(t)$$

and are thus *exponential functions* on $(S, +)$.

The characters on $(S, +)$ form a semigroup themselves under multiplication $(S^*, \cdot)$. This is the *dual* semigroup.

We shall assume that the semigroup S is *strongly separative* meaning that the bounded non-negative characters separate points.

(Hewitt and Zuckerman 1956) shows that this holds for all complex characters if and only if

$$s + s = s + t = t + t \implies s = t.$$

For non-negative real characters further conditions need to be satisfied (Lauritzen 1988), we omit the details.

A semigroup is strongly separative if and only if it is a subsemigroup of a disjoint union of torsion free groups, ie. in the finite case groups of the form $\mathbb{Z}^d$.

We can assume without loss of generality that the semigroup is strongly separative by taking quotients of the congruence

$$s \sim t \iff \rho(s) = \rho(t) \text{ for all } \rho \in S^*.$$

If the original semigroup is not strongly separative, the quotient semigroup will be.

A *filter* is a subsemigroup of $(S, \cdot)$ with

$$s + t \in F \implies s, t \in F.$$

So filters correspond to faces of convex sets.

A *prime ideal* I is a subsemigroup of S so that

$$a + b \in I \implies a \in I \text{ or } b \in I.$$

Clearly, *F is a filter if and only if $S \setminus F$ is a prime ideal.*

Since the intersection of filters always are filters, *we have for every element $s \in S$ a unique smallest filter F_s containing s.*

The *filter components* are the equivalence classes of the relation $s \sim t \iff F_s = F_t$.

The *support* F^ρ of a character ρ is a filter:

$$F^\rho = \{s \mid \rho(s) > 0\}.$$

Indeed, *all filters are support filters for some character,* since the indicator $\chi_F(s)$ of a filter F is a character.

Classic exponential families have fixed support. The Fisher-Darmois-Koopman-Pitman theorems says that *if the support is independent of the parameter* and the sufficient statistic *has fixed dimension,* then the statistical model must be exponential.

In some sense the focus on 'fixed dimension' and 'fixed support' is off the point.

With a slight twist compared to Lauritzen (1975, 1988), a *full, canonical exponential family* of densities $f(x;\theta)$ on a space $\mathcal{X}$ has the form

$$f(x;\theta) = \frac{dP_\theta(x)}{d\mu} = \theta\{t(x)\}c(\theta)^{-1}, \quad \theta \in \Theta$$

where $(S,+)$ is a semigroup generated by $t(\mathcal{X})$ and

$$\Theta = \left\{ \theta \in S^* \mid c(\theta) = \int \theta\{t(x)\}\,\mu(dx) < \infty, \right\}$$

i.e. the domain of the *Laplace transform* $\hat{\nu}$ of $\nu = \mu \circ t^{-1}$:

$$c(\theta) = \hat{\nu}(\theta) = \int \theta\{t(x)\}\,\mu(dx).$$

The Laplace transform for semigroups has been studied by Berg et al. (1984).

Note that semigroup exponential families are typically *different from the extended exponential families* introduced by Barndorff-Nielsen (1973, 1978).

A standard canonical exponential family is defined as $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ where

$$f(x;\theta) = \frac{dP_\theta(x)}{d\mu} = e^{\theta^\top t(x)} c(\theta)^{-1}, \quad \theta \in \Theta$$

with $\Theta \subseteq \mathbb{R}^d$ the domain of the Laplace transform.

For discrete and finite sample spaces the *extended* exponential family is simply defined as the weak closure $\overline{\mathcal{P}}$ of $\mathcal{P}$.

It can be shown to consist of *all distributions of the form* $P_\theta(\cdot \mid F)$, *where* F *is a face of the convex support of* $\mathcal{P}$ *and* $P_\theta \in \mathcal{P}$.

It can be shown (Lauritzen 1982, 1988, Ressel 1985) that *if a probability distribution on $\mathcal{X}$ is summarised by a semigroup statistic t such that*

$$p(x_1, \ldots, x_n) = \phi\{t(x_1) + \cdots + t(x_n)\}$$

then it is a mixture of exponential family distributions in the sense defined:

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_{S^*} c(\theta)^{-n} \theta\{t(x_1) + \cdots + t(x_n)\}\, M(d\theta) \\ &= \int_{S^*} \prod_{i=1}^{n} c(\theta)^{-1} \theta\{t(x_i)\}\, M(d\theta) \end{aligned}$$

so that, conditionally on θ, X_i are independent and identically distributed.

The likelihood function for a sample $X_1 = x_1, \ldots, X_n = x_n$ is

$$L(\theta) = \frac{\theta\{t(x_1) + \cdots + t(x_n)\}}{c(\theta)^n} = \frac{\theta\{t_n\}}{c(\theta)^n}.$$

If we extend L to the compact set

$$\{(\lambda, \theta) : 0 \leq \lambda \leq 1, \theta \in D\}$$

by letting

$$L(\lambda, \theta) = \lambda L(\theta)$$

L is continuous and attains therefore its maximum, which must happen for $\lambda = 1$.

Hence *L attains its maximum over* Θ.

There is a unique $\hat{\theta} \in \Theta$ so that L attains its maximum at $\hat{\theta}$.

For if this were not the case and

$$L(\theta_1) = L(\theta_2) = L(\hat{\theta})$$

for $\theta_1 \neq \theta_2$, we would have

$$L(\sqrt{\theta_1\theta_2}) = \frac{\sqrt{\theta_1(t_n)\theta_2(t_n)}}{c(\sqrt{\theta_1\theta_2})^n} > \sqrt{L(\theta_1)L(\theta_2)} = L(\hat{\theta}).$$

The inequality follows because

$$c(\sqrt{\theta_1\theta_2}) = \sum_x \sqrt{\theta_1\{t(x)\}\theta_2\{t(x)\}}\mu(x) < \sqrt{c(\theta_1)c(\theta_2)}.$$

Let $t_n = t(x_1) + \cdots + t(x_n)$. It then holds that *the support of* $\hat{\theta}$ *is the smallest filter containing* t_n*:*

$$F^{\hat{\theta}} = F_{t_n}.$$

For else $L(\hat{\theta} \cdot \chi_{F_{t_n}}) \geq L(\hat{\theta})$. Further *if* θ^* *satisfies*

$$\log \frac{\theta^*(t_n)}{\eta(t_n)} = n\mathbb{E}_{\theta^*} \left\{ \log \frac{\theta^*\{(t(X)\}}{\eta\{(t(X)\}} \right\} \text{ for all } \eta \in S^* \text{ with } F^{\eta} = F_{t_n}, \tag{1}$$

then $\hat{\theta} = \theta^*$.

A condition similar to *steepness* is needed in general to ensure existence of a solution to (1). In the case of $\mathcal{X}$ being finite, this is easily seen to be satisfied.

In the finite case these facts *essentially reduce the estimation problem to the calculation of* F_{t_n}.

Once this is done, the estimation problem is essentially done in the classical exponential subfamily of distrbutions with support exactly on F_{t_n}.

However, *this is not necessarily easy* in general and in some cases identical to identifying faces of convex polytopes.

However, *it might be possible that modern computational algebraic geometry could help solving this problem?*

For $(S, +) = (\mathbb{N}, \max)$ the characters are indicator functions

$$\theta(x) = \chi_{[0,\theta]}(x)$$

and the exponential family is the set of uniform distributions on $\{1, \ldots, \theta\}$ for $\theta \in \Theta = \mathbb{N}$:

$$f(x; \theta) = \theta^{-1} \chi_{[0,\theta]}(x).$$

The maximum likelihood estimate based on $X_1, \ldots, X_n$ is

$$\hat{\theta} = \max\{X_1, \ldots, X_n\} = X_{(n)}.$$

Note, as it should be, $F^{\hat{\theta}} = [0, \hat{\theta}]$.

Similarly, for $(S, +) \subseteq (\mathbb{R}^2, (\min, \max))$ we have a two-parameter family of characters

$$\theta = (\lambda, \mu), \lambda < \mu \text{ with } \theta(x) = \chi_{]\lambda,\mu]}(x)$$

and the associated exponential family

$$f(x; \theta) = \frac{1}{\mu - \lambda} \chi_{]\lambda,\mu]}(x)$$

and MLE equal to

$$\hat{\theta} = (\min\{X_1, \ldots, X_n\}, \max\{X_1, \ldots, X_n\}) = (X_{(1)}, X_{(n)}).$$

$(\mathbb{N}, \cdot)$ is isomorphic to the free semigroup $\mathbb{F}(\mathbb{P})$ over the prime numbers $\mathbb{P}$ through the representation of any integer as the product of its prime factors:

$$x = \prod_{\pi \in \mathbb{P}} \pi^{\nu_\pi(x)} \sim \{\nu_\pi(x), \pi \in \mathbb{P}\}.$$

A character θ can be represented as

$$\theta = \{\theta_\pi, \pi \in \mathbb{P}), \quad \theta(x) = \prod_{\pi \in \mathbb{P}} \theta_\pi^{\nu_\pi(x)}$$

so that the probability mass function becomes

$$p(x; \theta) = \phi(\theta)^{-1} \prod_{\pi \in \mathbb{P}} \theta_\pi^{\nu_\pi(x)}$$

where

$$\phi(\theta) = \prod_{\pi \in \mathbb{P}} (1 - \theta_\pi).$$

For the Laplace transform $\phi(\theta)$ to be finite, θ must satisfy

$$\sum_{\pi} \theta_\pi < \infty.$$

The corresponding exponential family has $\nu_\pi(X)$ geometrically distributed with parameter θ_π, independently for all π. From observations $X_1, \ldots, X_n$, the MLE is

$$\hat{\theta}_\pi = \frac{\nu_\pi(X_1) + \ldots + \nu_\pi(X_n)}{n + \nu_\pi(X_1) + \ldots + \nu_\pi(X_n)} = \frac{\nu_\pi(X_1 \cdots X_n)}{n + \nu_\pi(X_1 \cdots X_n)}.$$

Let for $A \subseteq V$

$$\mathcal{X} = \times_{v \in V} \mathcal{X}_v, \quad \mathcal{X}_A = \times_{v \in A} \mathcal{X}_A$$

and

$$t_A : \mathcal{X} \to \mathbb{F}(\mathcal{X}_A), \quad t_A(x) = \delta_{x_A}.$$

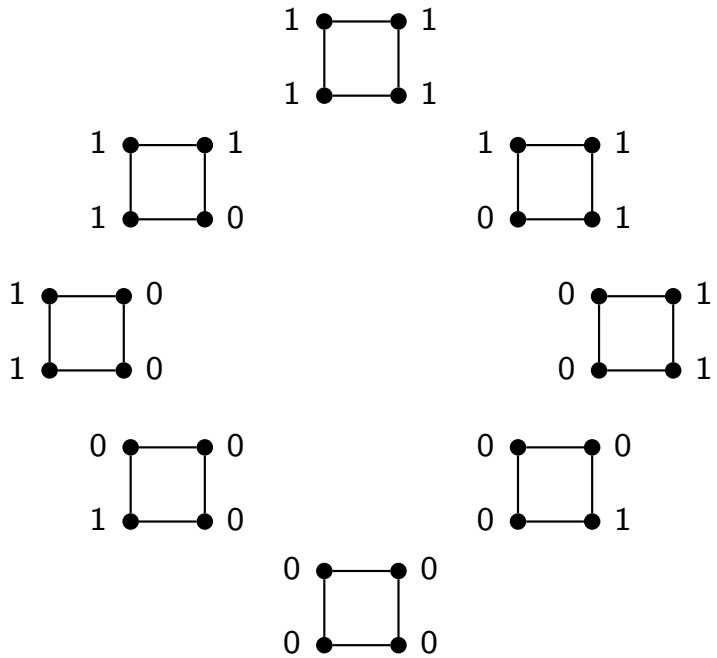
Define further

$$t_{\mathcal{A}} = (t_A, A \in \mathcal{A}).$$

The corresponding exponential family is a *hierarchical log-linear model.* Most characters have the form

$$\theta = \{\theta_A, A \in \mathcal{A}\}, \quad \theta(t_A, A \in \mathcal{A}) = \prod_{A \in \mathcal{A}} \prod_{x_A \in \mathcal{X}_A} \theta_A(x_A)^{t_A(x_A)}$$

but some do not factorize.



The uniform on these 8 configurations is part of the exponential family for the hierarchical log-linear model with generating class $\mathcal{A} = \{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$ and the free semigroup over this set is a filter in the image semigroup of the statistic.

Its density does not factorize but the distribution is limit of factorizing distributions (Moussouris 1974).

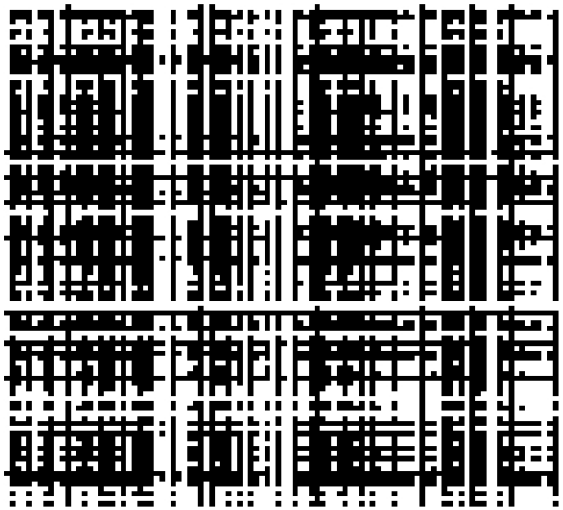
For a binary matrix, Rasch (1960) considered models for intelligence testing based on the fundamental assumption that the row sums and column sums are sufficient. Each row would typically correspond to an individual who scored either 1 or 0 for a number of tasks, represented by the columns.

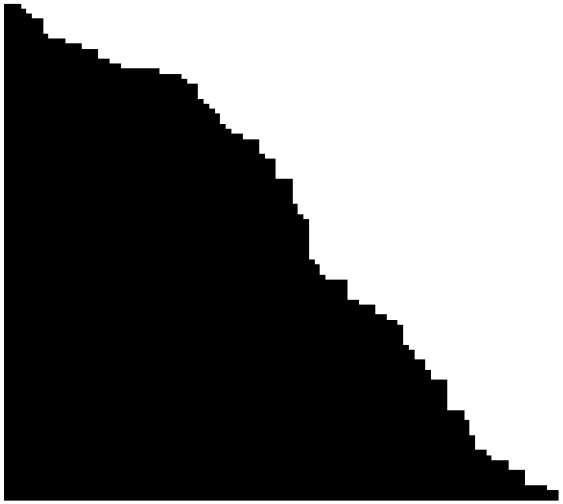
In an exchangeable version, a more natural summary statistic would be the empirical distribution of the row- and column sums:

$$t(x) = (\rho_i, \sigma_j, i, j = 1, 2, \ldots)$$

where ρ_i is the number of rows with sum i and σ_j the number of columns with sum j.

The strictly positive characters of the image semigroup were found in Lauritzen (2003) and correspond to random effect Rasch models, whereas there are some nasty but interesting characters associated with partially degenerate distributions (some individuals never being able to do some tasks etc.).

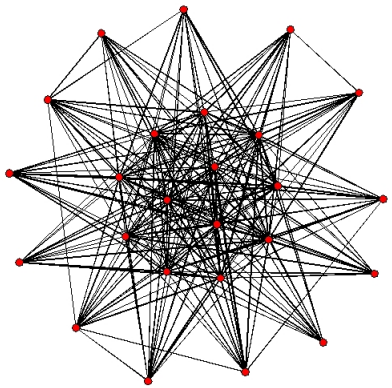




If the binary matrix is considered to be the adjacency matrix of a random (bipartite) graph, the statistic

$$t(x) = (\rho_i, \sigma_j, i, j = 1, 2, \ldots)$$

would correspond to the empirical distribution of the in- and out-degrees. The positive characters would then correspond to latent space network models (Hoff et al. 2002), see Lauritzen (2008).



Barndorff-Nielsen, O. E.: 1973, *Exponential Families and Conditioning*, John Wiley and Sons, Copenhagen, Denmark.

Barndorff-Nielsen, O. E.: 1978, *Information and Exponential Families in Statistical Theory*, John Wiley and Sons, New York.

Berg, C., Christensen, J. P. R. and Ressel, P.: 1984, *Harmonic Analysis on Semigroups*, Springer-Verlag, Heidelberg.

Clifford, A. H. and Preston, G. B.: 1961, *The Algebraic Theory of Semigroups*, Vol. I, American Mathematical Society, Providence, Rhode Island.

Fisher, R. A.: 1925, Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.

Hewitt, E. and Zuckerman, H. S.: 1956, The l_1-algebra of a commutative semigroup, *Transactions of the American Mathematical Society* **83**, 70–97.

Hoff, P. D., Raftery, A. E. and Handcock, M. S.: 2002, Latent space approaches to social network analysis, *Journal of the American Statistical Association* **97**, 1090–1098.

Lauritzen, S. L.: 1975, General exponential models for discrete observations, *Scandinavian Journal of Statistics* **2**, 23–33.

Lauritzen, S. L.: 1982, *Statistical Models as Extremal Families and Systems of Sufficient Statistics*, Aalborg University Press, Aalborg, Denmark.

Lauritzen, S. L.: 1988, *Extremal Families and Systems of Sufficient Statistics*, Vol. 49 of *Lecture Notes in Statistics*, Springer-Verlag, Heidelberg.

Lauritzen, S. L.: 2003, Rasch models with exchangeable rows and columns, *in* J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds), *Bayesian Statistics 7*, Clarendon Press, pp. 215–233.

Lauritzen, S. L.: 2008, Exchangeable Rasch models, *Rendiconti di Matematica, Serie VII* **28**, 83–95.

Moussouris, J.: 1974, Gibbs and Markov random systems with constraints, *Journal of Statistical Physics* **10**, 11–33.

Rasch, G.: 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Vol. 1 of *Studies in Mathematical Psychology*, Danmarks Pædagogiske Institut, Copenhagen.

Ressel, P.: 1985, de Finetti-type theorems: An analytical approach, *The Annals of Probability* **13**, 898–922.