Steffen L. Lauritzen

General Exponential Models for Discrete Observations

Preprint
July
1974
10

Institute of Mathematical Statistics
University of Copenhagen

Steffen L. Lauritzen

GENERAL EXPONENTIAL MODELS FOR DISCRETE OBSERVATIONS

Preprint 1974 No. 10

INSTITUTE OF MATHEMATICAL STATISTICS

UNIVERSITY OF COPENHAGEN

July 1974

GENERAL EXPONENTIAL MODELS FOR DISCRETE OBSERVATIONS[1]

by

Steffen L. Lauritzen*

Institute of Mathematical Statistics
University of Copenhagen

A class of models generalizing exponential families is defined via the algebraic structure of the sufficient statistics. The maximum likelihood estimate for the unknown parameter is shown to exist and be unique.

The sequence of sufficient statistics from successive repetitions of experiments corresponding to a general exponential model is shown to form an extreme family of Markov chains as defined by Lauritzen (1974).

Key Words: commutative semigroups, exponential families, extreme models, maximum likelihood estimation, sufficiency

Abbrev. Title: General Exponential Models

[1]This research was supported in part by Office of Naval Research Contract N00014-67-A-0112-0030, Stanford University and by the Danish Research Council. Reproduction in whole or in part is permitted for any purpose of the United States Government.

*At present visiting Stanford University.

1. Introduction and Summary

The purpose of the present paper is to illustrate the concept of an extreme family as defined by Lauritzen (1974) and to define a class of statistical models for discrete observations generalizing classical exponential families.

In the classical formulation, a discrete exponential family is a family of probability measures $(P_\theta, \theta \in \Theta)$, where the parameter space Θ is a subset of k-dimensional Euclidean vector space, the probability function being given by

$$P_\theta(x) = a(\theta)\, b(x)\, e^{\sum_{i=1}^{k} \theta_i t_i(x)} \tag{1.1}$$

where $x \in E$, a discrete set, t_i are real valued functions and $\theta = (\theta_1,\ldots,\theta_k)$. If one observes independent identically distributed random variables $X_1,\ldots,X_n$ with the common probability for X_i given by (1.1), the joint probability will be given by

$$P_\theta^{(n)}(x_1,\ldots,x_n) = a(\theta)^n \left(\prod_{j=1}^{n} b(x_j)\right) \cdot e^{\sum_{i=1}^{k} \theta_i \left(\sum_{j=1}^{n} t_i(x_j)\right)} \tag{1.2}$$

Now, $P_\theta^{(n)}$ is again an exponential family with the same parameter space as before. Somehow this is not a coincidence. If we try to look closer at the elements of the exponential family, we might understand this fact.

The function b is a common reference measure defining the support of the measures $(P_\theta, \theta \in \Theta)$, and $a(\theta)$ is a normalizing constant. The functions

$(t_1,\ldots,t_k)$ are the sufficient statistics, and as an experiment is repeated, the sufficient statistics for the combined experiment is obtained as the sum of the sufficient statistics for the experiments in the repetition:

$$t_i^{(n)}(x_1,\ldots,x_n) = t_i(x_1) + \ldots + t_i(x_n) \quad . \tag{1.3}$$

The reason for this is that the function

$$g_\theta\colon (t_1,\ldots,t_k) \to e^{\sum_{i=1}^{k} \theta_i t_i}$$

is a homomorphism of the range space of $(t_1,\ldots,t_k)$ into the group $((0,\infty),\cdot)$:

$$g_\theta\Big((t_1,\ldots,t_k) + (s_1,\ldots,s_k)\Big) = g_\theta\Big((t_1,\ldots,t_k)\Big) \cdot g_\theta\Big((s_1,\ldots,s_k)\Big) \quad .$$

The idea in this paper is that most results about exponential families essentially are based on the above properties only. We shall therefore try to define a class of families of distributions via these properties.

If we again look at (1.1), (1.2) and (1.3) we see that we never substract. In fact, we only use that the algebraic operation + is associative and commutative. A set with a composition which is associative and commutative is here called a commutative semigroup. We shall establish some of the simple results about these in Section 2.

Let us consider another family of distributions $(P_\theta, \theta\in\Theta)$, where $\Theta = \{1,2, \ldots ,\}$ and

$$P_\theta\ (x) = \frac{1}{\theta} \cdot \chi_{\{1,\dots,\theta\}}\ (x) \quad ,$$

where $x \in E = \{1,2,\dots\}$ and χ_A is the indicator function of the set A, i.e.

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise .} \end{cases}$$

Consider $X_1,\dots,X_n$ independent indentically distributed as above. Their joint probability is given by

$$P_\theta^{(n)}\ (x_1,\dots,x_n) = \frac{1}{\theta^n}\,\chi_{\{1,\dots,\theta\}}\ (\max\{x_1,\dots,x_n\})$$

The same situation as before is actually present if we replace $t(x_1)+t(x_2)$ by $\max\{t(x_1),\ t(x_2)\}$. Just $g_\theta(x) = \chi_{\{1,\ \dots\ ,\ \theta\}}(x)$ may vanish in this case whereas exponentials are always strictly positive. The support of the measure P_θ in this last example varies with $\theta \in \Theta$, which is not the case in the first example.

In section 7, part I of Barndorff-Nielsen (1973), there is a detailed discussion of problems connected to maximum likelihood estimation in exponential families. The maximum likelihood estimate in regular canonical exponential families is shown to exist iff the observation happens to be so, that the value of the sufficient statistic falls within the interior of the convex hull of the support of the measures in the family, transformed by the sufficient statistics. This means that if the boundary of this convex hull has positive probability, one might very well get an observation from which it is impossible to estimate. To solve this problem it is proposed there to make a suitable extension of the model, the extension being defined for families where the set of possible values of the set of sufficient statistics is assumed to be finite. The extension is called the completion of an exponential family.

The measures in the completion of an exponential family have certainly their support varying with the parameter, and the "fixed support" property therefore does not seem to be essential to the nice results existing for exponential families.

The families defined in the present paper are shown to be "complete" in the sense that the maximum likelihood estimate of the parameter always exists.

In section 3, the families are defined and some examples are discussed. In section 4 we show the existence and uniqueness of the maximum likelihood estimate of the unknown parameter in such families.

In section 5 we show that the family of Markov chains made up by sequences of sufficient statistics from successive independent repetitions of an experiment giving rise to a general exponential model, is in fact an extreme family of Markov chains as defined by Lauritzen (1974).

In section 6 we shall briefly discuss the relation between the models defined in the present paper and the completion of an exponential family as defined by Barndorff-Nielsen (1973).

2. Commutative Semigroups

Commutative semigroups will play an essential role in the present paper. We shall quote the definition.

Definition 2.1 Let M be a set and $*$ a composition rule on M. $(M,*)$ is said to be a commutative semigroup if $*$ is associative and commutative i.e., if

$$\text{i) } \forall a,b,c \in M: \quad a * (b*c) = (a*b) * c \quad ,$$

$$\text{ii) } \forall a,b \in M: \quad a * b = b * a \quad ,$$

As we shall only consider commutative semigroups throughout this paper, we shall just write "semigroup" instead of "commutative semigroup". Examples of such semigroups are

1) $(\underset{\sim}{N},+)$ where $\underset{\sim}{N} = \{1,2, \ldots\}$.
2) $(\underset{\sim}{N},\vee)$, where $x\vee y = \max\{x,y\}$.
3) $(\underset{\sim}{N},\wedge)$, where $x\wedge y = \min\{x,y\}$.
4) $(\underset{\sim}{R}_+,\cdot)$, where $\underset{\sim}{R}_+$ is the set of nonnegative real numbers.

Now, let $(M,*)$ be a semigroup. Consider the set $\hat{M}$ consisting of all homomorphisms $\xi:(M,*) \to (\underset{\sim}{R}_+,\cdot)$, i.e., satisfying for all $a,b\in M$

$$\xi(a)\ \xi(b) = \xi(a*b) \ .$$

If ξ_1, $\xi_2 \in \hat{M}$, the mapping $\xi_1 \cdot \xi_2$ defined by

$$\xi_1 \cdot \xi_2(a) = \xi_1(a)\ \xi_2(a)$$

is obviously in $\hat{M}$ and it is a trivial exercise to verify that $(\hat{M},\cdot)$ is a semigroup. $(\hat{M},\cdot)$ shall be called the dual semigroup to $(M,*)$.

If we have two semigroups $(M,*)$ and $(N,\circ)$ we can form the product of these

$$(M,*) \times (N,\circ) = (M \times N, \circledast) \ ,$$

where

$$(m_1,n_1) \circledast (m_2,n_2) = (m_1 * m_2,\ n_1 \circ n_2) \quad .$$

This is again a semigroup. The dual to a product can easily be obtained from the duals to the elements in the product:

<u>Proposition 2.1</u>: <u>The homomorphisms of</u> $(M \times N, \circledast)$ <u>into</u> $(R_+, \cdot)$ <u>are exactly those of the form</u>

$$\xi(m,n) = \xi_M(m)\ \xi_N(n) \quad ,$$

<u>where</u> $\xi_M \in \hat{M}$ <u>and</u> $\xi_N \in \hat{N}$.

<u>Proof</u>: If $\xi(m,n) = 0$ for all (m,n), ξ is of the form described. We assume that $\xi(m',n') > 0$. The equation

$$\xi(m*m',\ n\circ n') = \xi(m,n)\ \xi(m',n')$$

$$= \xi(m',n)\ \ \xi(m,n') \qquad (2.1)$$

gives as (m',n') is fixed:

$$\xi(m,n) = \frac{\xi(m,n')\ \xi(m',n)}{\xi(m',n')} = g(m)\ h(n) \quad . \qquad (2.2)$$

Further we have for any m, n_1, n_2:

$$\xi(m*m,\ n_1\circ n_2) = g(m*m)\ h(n_1\circ n_2)$$

$$= g(m)^2\ h(n_1)\ h(n_2).$$

As this holds for all m, n_1 and n_2, we must have a constant c_1 such that

$$h(n_1 \circ n_2) = c_1\, h(n_1)\, h(n_2) \quad . \tag{2.3}$$

Analogously there is a c_2 such that

$$g(m_1 * m_2) = c_2\, g(m_1)\, g(m_2) \tag{2.4}$$

From (2.1) and (2.2) we get that $c_1\, c_2 = 1$ and if we define

$$\xi_M(m) = c_2\, g(m)$$

$$\xi_N(n) = c_1\, h(n)$$

we get from (2.3) and (2.4) that $\xi_M \in \hat{M}$ and $\xi_N \in \hat{N}$ and from (2.2) that $\xi(m,n) = g(m)\, h(n) = c_2\, g(m)\, c_1 h(n) = \xi_M(m)\, \xi_N(n)$ which was to be proved.

As mentioned in the introduction, the support of the measures in the families we consider may very often vary with the parameter. This will of course not be in a completely arbitrary fashion but in a fashion compatible with the algebraic structure of the sufficient statistics. To investigate this aspect, the following concept will be of relevance:

Definition 2.2 $F \subseteq M$ is said to be a face of M if

i) F is a subsemigroup of M and

ii) $c \in F \wedge c = a*b \Rightarrow a \in F \wedge b \in F$

The faces of M are exactly the possible positivity regions for elements in $\hat{M}$:

Proposition 2.2: Let $F \subseteq M$. F is a face of M iff there is a $\xi \in \hat{M}$ such that

$$F = \{a \in M: \xi(a) > 0\} .$$

Proof: If $F = \{a \in M: \xi(a) > 0\}$ for some $\xi \in \hat{M}$, then

$$a \in F \wedge b \in F \Rightarrow \xi(a*b) = \xi(a)\,\xi(b) > 0 ,$$

so F is a subsemigroup of M. If $c \in F$ and $c = a*b$, then

$$0 < \xi(a*b) = \xi(a)\,\xi(b)$$

and hence $\xi(a)$ and $\xi(b)$ both must be positive, i.e., $a \in F$ and $b \in F$.

If on the other hand F is a face of M, we can define

$$\xi(a) = \begin{cases} 1 & \text{if } a \in F \\ 0 & \text{otherwise} . \end{cases}$$

ξ is easily seen to be a homomorphism and the result is proved. We also have

Proposition 2.3: M is a face of M

The proof is obvious.

Proposition 2.4. If $(F_i)_{i \in I}$ is a family of faces of M, then $F = \bigcap_{i \in I} F_i$ is a face of M.

Proof: Immediate from the definition.

Remark: From propositions 2.3 and 2.4 it follows that for any $a \in M$ there is a unique smallest face of M, $F(a)$, such that $a \in F(a)$.

Propositions 2.1 and 2.2 enable us to establish a result about faces of product semigroups :

Proposition 2.5: F is a face of $M \times N$ iff $F = F_M \times F_N$, where F_M and F_N are faces of respectively M and N.

Proof: According to proposition 2.1, all homomorphisms ξ of $(M \times N, \circledast)$ into $(\underset{\sim}{R}_+, \cdot)$ are of the form

$$\xi(m,n) = \xi_M(n)\ \xi_N(n) \quad ,$$

where $\xi_M \in \hat{M}$ and $\xi_N \in \hat{N}$. Now

$$\begin{aligned} &\left\{(m,n):\ \ \xi(m,n) > 0\right\} \\ &= \left\{(m,n):\ \ \xi_M(m) > 0 \wedge \xi_N(n) > 0\right\} \\ &= \left\{m:\ \xi_M(m) > 0\right\} \times \left\{n:\ \xi_N(n) > 0\right\} . \end{aligned} \tag{2.1}$$

The proposition 2.2 and equation (2.1) together yield the result.

Example 2.1: Let us consider the semigroup $(\underset{\sim}{N} \cup \{0\}, +)$. Let $\xi(1) = \theta$, some non-negative real number. We must have

$$\xi(n) = \xi(1)^n = \theta^n \quad .$$

It follows that the only faces of $(\underset{\sim}{N} \cup \{0\}, +)$ are $\{0\}$ and $\underset{\sim}{N} \cup \{0\}$

Example 2.2 Let us consider $(\underset{\sim}{N}, \vee)$. Let $n \in \underset{\sim}{N}$ be fixed. The smallest face containing n must contain all integers less than or equal to n as

$$n \vee x = n \quad \text{if} \quad x \leqq n \quad .$$

On the other hand, $\{1,\ldots,n\}$ is obviously a face of $(\underset{\sim}{N},\vee)$. Hence all faces of $(\underset{\sim}{N},\vee)$ are $\underset{\sim}{N}$ itself and subsets of the form $\{1,\ldots,n\}$ for some $n\in\underset{\sim}{N}$.

Now let $\xi\in\hat{\underset{\sim}{N}}$ be positive exactly on $\{1,\ldots,n\}$; it must satisfy

$$\xi(x\vee n) = \xi(x)\ \xi(n) = \xi(n) \quad \text{for} \quad x \leqq n.$$

As $\xi(n)$ is strictly positive, we get

$$\xi(x) = 1 \quad \text{for } x \leqq n \quad ,$$

and hence that the <u>only</u> homomorphisms of $(\underset{\sim}{N},+)$ are indicator functions of faces,

$$\xi(x) = \chi_{\{1,\ldots,n\}}(x)$$

for some $n\in \underset{\sim}{N}\cup\{\infty\}$.

<u>Example 2.3</u>: If we now form the product $(\underset{\sim}{N},\vee)\times(\underset{\sim}{N}\cup\{0\},+)$ it follows from proposition 2.1 that all homomorphisms are of the form

$$\xi(x,y) = \chi_{\{1,\ldots,n\}}(x)\cdot\theta^{y}$$

for some $n\in\underset{\sim}{N}\cup\{\infty\}$ and $\theta\geqq 0$. From proposition 2.5, we get that the faces of this monoid are

$$\{1,\ldots,n\}\times\underset{\sim}{N}\cup\{0\} \quad n\in\underset{\sim}{N}$$

$$\{1,\ldots,n\}\times\{0\}, \quad n\in\underset{\sim}{N}$$

$$\underset{\sim}{N}\times\underset{\sim}{N}\cup\{0\} \text{ and } \underset{\sim}{N}\times\{0\}.$$

3. General Exponential Models

In the following we shall consider an at most denumerable set E, a semigroup $(M,*)$ and a mapping $t: E \to M$. We shall think of E as the sample space and of t as a sufficient statistic. Let $M_1 = t(E)$ and define recursively

$$M_n = M_1 * M_{n-1}, \quad \text{for } n = 2,3,\ldots \quad .$$

This is done for the following reason: if we make n independent observations of a random variable on E, we shall assume that the sufficient statistic will be

$$t^{(n)}(x_1,\ldots,x_n) = t(x_1) * \cdots * t(x_n) \quad ,$$

and hence $M_n = t^{(n)}(E^n)$.

We shall assume that we can infer the size of the experiment from the statistic, or, in other words, that

$$M_m \cap M_n = \emptyset, \quad \text{whenever } n \neq m \quad .$$

For convenience we don't want M to be bigger than necessary, hence we assume that

$$M = \bigcup_{n=1}^{\infty} M_n \quad .$$

Let ν be a σ-finite measure on E so that $\nu(x)$ is positive for all $x \in E$. Let $\hat{M}_\nu$ denote the <u>normalized dual</u> to M:

$$\hat{M}_\nu = \left\{ \xi \in \hat{M}: \sum_{x \in E} \nu(x)\ \xi(t(x)) = 1 \right\}$$

and assume that $\hat{M}_\nu$ is non-empty.

A statistical model for a random variable X taking values in E is a family $\mathcal{P}$ of probability measures on E.

Definition 3.1: $\mathcal{P}$ is said to be a general exponential model if there exist M, t and ν as above, such that

$$P \in \mathcal{P} \iff \exists \xi \in \hat{M}_\nu : \; P\{X = x\} = \nu(x)\, \xi(t(x)) \quad .$$

Remark: It is no restriction to assume that for any $s \in M$, there is a $\xi \in \hat{M}_\nu$ such that $\xi(s) > 0$, since if this is not the case, we can look at

$$M' = \{s \in M : \; \exists \xi \in \hat{M}_\nu : \; \xi(s) > 0\}$$

which clearly is a semigroup and

$$E' = t^{-1}(t(E) \cap M')$$

$$t' = t|E' \; , \quad \nu' = \nu|E' .$$

Since this clearly gives us an equivalent statistical model we shall assume this throughout the paper.

Let us first see in what sense this looks like a "classical" exponential model. Suppose we have observed n independent random variables from the above distribution. The joint probability function is

$$P_\xi^{(n)}(x_1, \ldots, x_n) = \left(\prod_{i=1}^{n} \nu(x_i) \right) \cdot \xi(t(x_1)) \cdots \xi(t(x_n))$$

$$= \left(\prod_{i=1}^{n} \nu(x_i) \right) \xi(t(x_1) * \cdots * t(x_n)) \qquad (3.1)$$

with $\xi \epsilon \hat{M}$. If we compare (3.1) with (1.2) in the introduction, we note, that ν plays the same role as b, the common reference measure. The statistic t corresponds to $(t_1,\ldots,t_k,n)$, i.e. the sufficient statistic plus a "counting variable" indicating the size of the experiment. $\xi \epsilon \hat{M}_\nu$ corresponds to the function

$$(t_1,\ldots,t_k,n) \to a(\theta)^n e^{\sum_{i=1}^{k} \theta_i t_i}$$

so the normalizing constant $a(\theta)$ is taken into ξ and the experiment size into the statistic t.

The above defined models differ from the exponential models in several respects. First the range space of the statistic is a semigroup instead of a subset of a vector space, the parameter space is the normalized dual of this semigroup instead of a subset of a vector space, and there is no assumption of anything like finite dimension. Furthermore we shall see that in general the support of the measures in the family will depend on the parameter ξ, as it will not always be positive. As derived in the previous sections, the possible positivity regions for ξ will be the faces of the semigroup $(M,*)$. From (3.7) and the Neyman factorization theorem it immediately follows that

$$t^{(n)}(X_1,\ldots,X_n) = t(X_1)*\cdots* t(X_n)$$

is sufficient for the parameter ξ from observation of $X_1,\ldots,X_n$.

The relation between these models and the classical exponential models should hopefully be more apparent from the examples below.

Example 3.1 (The Bernoulli Distribution)

Let $E = \{0,1\}$ and $\nu(0) = \nu(1) = 1$. Let $(M,*)$ be the subsemigroup of

$$(M',*) = (\underset{\sim}{N} \cup \{0\},+) \times (\underset{\sim}{N},+)$$

given by

$$M = \bigcup_{n=1}^{\infty} M_n, \quad \text{where} \quad M_n = \{(x,y): x+y = n\} \quad .$$

Let $t(1) = (1,0)$ and $t(0) = (0,1)$. The elements of $\hat{M}$ are all of the form

$$F_{\theta,\eta}(x,y) = \theta^x \eta^y, \quad \theta \geqq 0, \quad \eta \geqq 0.$$

We immediately get that

$$F_{\theta,\eta} \in \hat{M}_\nu \iff \theta^1\eta^0 + \theta^0\eta^1 = 1$$

$$\iff \eta = 1 - \theta \quad .$$

Hence, the model

$$P_\theta\ \{X = x\} = \nu(x)\ F_{\theta,1-\theta}(t(x)) = \begin{cases} \theta & \text{if} \quad x = 1 \\ 1-\theta & \text{if} \quad x = 0 \end{cases} ,$$

where $0 \leqq \theta \leqq 1$, is a general exponential model. The difference between this model and the classical exponential family version of the Bernoulli distribution is that $\theta = 0$ and $\theta = 1$ are included in the model.

<u>Example 3.2</u> (The Poisson distribution).

Let $E = \underset{\sim}{N} \cup \{0\}$ and $\nu(x) = \frac{1}{x!}$. Let

$$(M,*) = (\underset{\sim}{N} \cup \{0\},+) \times (\underset{\sim}{N},+)$$

We have

$$M = \bigcup_{n=1}^{\infty} M_n, \quad \text{where} \quad M_n = \{(x,y);\ y = n\}.$$

Let $t(x) = (x,1)$. We get

$$F_{\theta,\eta} \in \hat{M}_\nu \iff \sum_{x=0}^{\infty} \frac{1}{x!} \theta^x \eta = 1$$

$$\iff \eta = e^{-\theta}$$

Hence, the model

$$P_\theta\{X = x\} = \nu(x)\, F_{\theta,e^{-\theta}}(t(x)) = \frac{\theta^x}{x!} e^{-\theta} \quad ,$$

where $\theta \geqq 0$ is a general exponential model. Again the inclusion of $\theta = 0$ is the only difference between this and the classical approach.

So far, the examples considered have basically been exponential models in the classical sense apart from adding some degenerate distributions. The following examples show that the models in fact can be quite different from the classical exponential models.

Example 3.3 (The uniform distribution).

Let $E = \underset{\sim}{N}$ and $\nu(x) = 1$ for all $x \in E$. Let $(M,*)$ be the subsemigroup of

$$(M',*) = (\underset{\sim}{N},\nu) \times (\underset{\sim}{N},+) .$$

given by

$$M = \bigcup_{n=1}^{\infty} M_n, \quad \text{where} \quad M_n = \{(x,y) : y = n\}$$

Let $t(x) = (x,1)$. The elements of $\hat{M}$ are all of the form

$$F_{\theta,\eta}(x,y) = \chi_{\{1,\ldots,\theta\}}(x)\eta^y, \quad \theta \in E, \; \eta \geqq 0 .$$

We have

$$F_{\theta,\eta} \in \hat{M}_\nu \iff \sum_{x=1}^{\infty} \eta \, \chi_{\{1,\ldots,\theta\}}(x) = 1$$

$$\iff \eta = \frac{1}{\theta}, \quad \theta \in \underset{\sim}{N}$$

Hence, the model

$$P_\theta\{X = x\} = \nu(x) \, F_{\theta,\frac{1}{\theta}}(t(x)) = \frac{1}{\theta} \cdot \chi_{\{1,\ldots,\theta\}}(x) ,$$

where $\theta = 1,2,\ldots,$ is a general exponential model.

Combining examples 3.1 - 3.3 we get the following:

Example 3.4 (Doubly truncated geometric distribution with unknown truncation points). Let $E = \underset{\sim}{N}$ and $\nu(x) = 1$ for all $x \in E$. As our semigroup $(M,*)$ we choose the subsemigroup of

$$(M',*) = (\underset{\sim}{N}, \vee) \times (\underset{\sim}{N} \cup \{\infty\}, \wedge) \times (\underset{\sim}{N},+) \times (\underset{\sim}{N},+) \quad ,$$

given by $M = \bigcup_{n=1}^{\infty} M_n$, where

$$M_1 = \{(x,x,x,1) : x \in \underset{\sim}{N}\} \quad ,$$

and M_n is recursively defined as

$$M_n = M_1 * M_{n-1} \quad \text{for} \quad n = 2,3,\ldots$$

Let $t(x) = (x,x,x,1)$. The elements of $\hat{M}$ are all of the form

$$F_{\theta,\eta,\lambda,\mu}(x,y,z,n) =$$

$$\chi_{\{1,\ldots,\theta\}}(x)\ \chi_{\{\eta,\ldots,\infty\}}(y)\ \lambda^z \mu^n \quad ,$$

where $\theta \in \underset{\sim}{N} \cup \{\infty\}$, $\eta \in \underset{\sim}{N} \cup \{\infty\}$, $\lambda \geqq 0$ and $\mu \geqq 0$.

We get

$$F_{\theta,\eta,\lambda,\mu} \in \hat{M}_\nu \iff \sum_{x=\eta}^{\theta} \lambda^x \mu = 1$$

$$\iff \eta \leqq \theta < \infty,\ \lambda = 1 \ \text{ and } \ \frac{1}{\mu} = \theta - \eta + 1$$

$$\text{or} \quad \eta \leqq \theta < \infty,\ \lambda \neq 1 \ \text{ and } \ \frac{1}{\mu} = \frac{\lambda^{\theta+1} - \lambda^{\eta}}{\lambda - 1} \quad .$$

$$\text{or} \quad \eta < \infty,\ \theta = \infty,\ \lambda < 1 \ \text{ and } \ \frac{1}{\mu} = \frac{\lambda^{\eta}}{1-\lambda}$$

Hence the model

$$P_{\theta,\eta,\lambda}\{X = x\} = \phi(\lambda,\eta,\theta)\ \lambda^{x}\chi_{\{\eta,\ \dots\ ,\ \theta\}}(x) \quad ,$$

where $\lambda \geqq 0$, $\theta \geqq \eta$, $\eta \in \underset{\sim}{N}$ and

$$\phi(\lambda,\eta,\theta) = \begin{cases} \dfrac{1}{\theta-\eta+1} & \text{if } \lambda = 1 \text{ and } \theta < \infty \\ \dfrac{\lambda-1}{\lambda^{\theta+1}-\lambda^{\eta}} & \text{if } \lambda \neq 1 \text{ and } \theta < \infty \\ \dfrac{1-\lambda}{\lambda^{\eta}} & \text{if } \lambda < 1 \text{ and } \theta = \infty \end{cases}$$

is a general exponential model.

Finally we shall consider an example, where the general exponential model is different from a classical one in the sense of infinite-dimensionality of the parameter space.

Example 3.5 (The completely free distribution). Let E be any denumerable set, and $\nu(x) = 1$ for all $x \in E$. Let $(M,*) = {}^{E}\underset{\sim}{N} \cup \{0\}$, consisting of all mappings f from E to $\underset{\sim}{N} \cup \{0\}$ where $\{x: f(x) \neq 0\}$ is finite and non-empty. The composition rule is pointwise addition

$$(f*g)\ (x) = f(x) + g(x)\ .$$

We have the partitioning of $M = \bigcup_{n=1}^{\infty} M_n$, where

$$M_n = \left\{f: \sum_{x \in E} f(x) = n\right\}\ .$$

If we let $t(x) = \chi_{\{x\}}$, we can see, that the sufficient reduction of a sample of size n becomes the "frequency table", i.e., $t^{(n)}(x_1,\dots,x_n)$ is the function in ${}^{E}N \cup \{0\}$, having the value n_x in x iff x occurs exactly n_x times in the sample $(x_1,\dots,x_n)$. $\hat{M}$ consists of the elements

$$g_\theta(f) = \prod_{x \in E} \theta(x)^{f(x)} ,$$

where θ is any mapping from E into the non-negative real numbers. We have

$$g_\theta \in \hat{M}_\nu \iff \sum_{x \in E} \left(\prod_{y \in E} \theta(y)^{\chi_{\{x\}}(y)} \right) = 1$$

$$\iff \sum_{x \in E} \theta(x) = 1 .$$

Hence, the model

$$P_\theta\{X = x\} = \nu(x)\, g_\theta(\chi_{\{x\}}) = \theta(x) ,$$

where θ satisfies

$$\left. \begin{array}{l} \theta(x) \geqq 0 \quad \text{for all} \quad x \in E \quad \text{and} \\ \sum_{x \in E} \theta(x) = 1 \end{array} \right\} ,$$

is a general exponential model. Other examples could be generated ad libitum.

4. Estimation in general exponential models

We shall consider the following estimation problem:

Let $X_1, \ldots, X_n$ be independent and identically distributed on E with

$$P_\xi\{X = x\} = \nu(x)\, \xi(t(x)) ,$$

where ν and t are known and as in the previous section and $\xi \in \hat{M}_\nu$ is unknown. Our sample space is E^n, the parameter space is $\hat{M}_\nu$ and the likelihood function becomes

$$L(x_1, \ldots, x_n, \xi) = \prod_{x=1}^{n} \nu(x_i) \quad \xi(t(x_1) * \cdots * t(x_n)).$$

As mentioned earlier, $t^{(n)}$ given by

$$t^{(n)}(x_1,\ldots,x_n) = t(x_1)*\cdots*t(x_n)$$

is sufficient for ξ and $\hat{\xi}$ is clearly a maximum likelihood estimator of ξ iff

$$\hat{\xi}(t_o) = \sup_{\xi\in\hat{M}_\nu} \xi(t_o) \quad ,$$

where $t_o = t(x_1)*\cdots*t(x_n)$.

In the following we shall establish the existence and uniqueness of $\hat{\xi}$ for any n and $x_1,\ldots,x_n$.

First we prove a lemma:

Lemma 4.1 *Let* $\hat{M}^*_\nu = \left\{\xi\in\hat{M}\colon \sum_{x\in E} \nu(x)\, \xi(t(x)) \leqq 1\right\}$. $\hat{M}^*_\nu$ *is compact in the pointwise topology*.

Proof: Let $\xi_1, \xi_2,\ldots$ be a sequence of elements in $\hat{M}^*_\nu$. As $[0, \infty]^M$ is compact, we can always find a subsequence $\xi_{n_1}, \xi_{n_2},\ldots$ so that for any $s\in M$,

$$\xi_{n_i}(s) \xrightarrow[i\to\infty]{} \xi(s) \quad ,$$

where $0 \leqq \xi\,(s) \leqq \infty$.

We have to show that this limit ξ in fact is an element of $\hat{M}^*_\nu$. From Fatou's lemma, we get

$$\sum_{x\in E} \nu(x)\,\xi(t(x)) \leqq \liminf_{i\to\infty} \sum_{x\in E} \nu(x)\, \xi_{n_i}(t(x)) \leqq 1 \qquad (4.1)$$

We shall now just prove that $\xi(s) < +\infty$ and that

$$\xi(s * t) = \xi(s)\,\xi(t) .$$

But as $t(E) = M_1$ and $\nu(x) > 0$ for all $x \in E$, (4.1) gives that $\xi(s) < +\infty$ for all $s \in M_1$. Now, if $s \in M_n = M_1 * \cdots * M_1$, where M_1 appears n times, we have

$$\xi(s) = \lim_{i\to\infty} \xi_{n_i}(s) = \lim_{i\to\infty} (\xi_{n_i}(s_1)\cdots\xi_{n_i}(s_n)) = \xi(s_1)\cdots\xi(s_n) ,$$

where $s_1,\dots,s_n \in M$, and $s_1 * \cdots * s_n = s$. This gives that $\xi(s) < \infty$ for all $s \in M$ since

$$\xi(e) = \lim_{i\to\infty} \xi_{n_i}(e) = 1,$$

and also that $\xi \in \hat{M}$. The lemma is proved.

We can now show the existence of the maximum likelihood estimate for any $n, x_1,\dots,x_n$:

<u>Proposition 4.1</u>: <u>For all</u> $s_0 \in M$, <u>there is a</u> $\hat{\xi} \in \hat{M}_\nu$, such that $\hat{\xi}(s_0) = \sup_{\xi \in \hat{M}_\nu} \xi(s_0) > 0$.

<u>Proof</u>:

As $\hat{M}^*_\nu$ is compact and the mapping $\xi \to \xi(s)$ is continuous, there is a $\xi^* \in \hat{M}^*_\nu$ so that

$$\xi^*(s_0) = \sup_{\xi \in \hat{M}^*_\nu} \xi(s_0) .$$

$\xi^*(s_0)$ must be strictly positive (see the remark to Definition 3.1). But if

$$\sum_{x \in E} \nu(x)\, \xi^*(t(x)) = c < 1$$

then $\hat{\xi}: M \to \underset{\sim}{R}_+$ defined as

$$\hat{\xi}(s) = \xi^*(s)\left(\frac{1}{c}\right)^n \quad \text{for} \quad s \in M_n$$

is in $\hat{M}^*_\nu$ and $\hat{\xi}(s_0) > \xi^*(s_0)$, which is a contradiction. Hence we must have $c = 1$, $\hat{\xi}(s) = \xi^*(s)$, $\hat{\xi} \in \hat{M}_\nu$ and

$$\hat{\xi}(s_0) = \sup_{\xi \in \hat{M}_\nu} \xi(s_0) \quad ,$$

which was to be proved.

Next we prove the uniqueness of the maximum likelihood estimate.

Proposition 4.2: *If* $\hat{\xi}_1(s_0) = \hat{\xi}_2(s_0) = \sup_{\xi \in \hat{M}_\nu} \xi(s_0)$, then $\hat{\xi}_1 = \hat{\xi}_2$.

Proof: For $s \in M$ let

$$\xi(s) = \sqrt{\xi_1(s)\ \xi_2(s)}\ .$$

Define $\hat{\xi}: M \to \underset{\sim}{R}_+$ by

$$\hat{\xi}(s) = \frac{\xi(s)}{\left(\sum_{x \in E} \nu(x)\ \xi(t(x))\right)^k} \qquad \text{for } s \in M_k\ .$$

Obviously $\hat{\xi} \in \hat{M}_\nu$. If $\hat{\xi}_1 = \hat{\xi}_2$ for all $s \in M_1$, $\hat{\xi}_1 = \hat{\xi}_2$ for all $s \in M$. Cauchy-Schwarz inequality gives

$$\sum_{x \in E} \nu(x)\ \xi(t(x)) \leqq \left(\sum_{x \in E} \nu(x)\ \hat{\xi}_1(t(x))\right)\left(\sum_{x \in E} \nu(x)\ \hat{\xi}_2(t(x))\right) = 1,$$

as $\hat{\xi}_1$ and $\hat{\xi}_2$ are in $\hat{M}_\nu$. We therefore have

$$\hat{\xi}_1 \neq \hat{\xi}_2 \Rightarrow \sum_{x \in E} \nu(x)\ \xi(t(x)) < 1.$$

But then

$$\hat{\xi}(s_0) > \sqrt{\hat{\xi}_1(s_0)\ \hat{\xi}_2(s_0)} = \hat{\xi}_1(s_0) = \hat{\xi}_2(s_0) \quad ,$$

which is a contradiction. Hence $\hat{\xi}_1 = \hat{\xi}_2 = \hat{\xi}$ which was to be proved.

The next result giving some more detailed information about the maximum likelihood estimate should be compared to the results in section 7, part I of Barndorff-Nielsen (1973).

<u>Proposition 4.3</u>: <u>The positivity region of</u> $\hat{\xi}$ <u>where</u> $\hat{\xi}(s_0) = \sup_{\xi \in \hat{M}_\nu} \xi(s_0)$ <u>is exactly the face</u> $F(s_0)$.

<u>Proof:</u> As $\hat{\xi}(s_0) > 0$, we have from proposition 2.2 that

$$M^+(\hat{\xi}) = \{s \in M: \hat{\xi}(s) > 0\} \supseteq F(s_0) \ .$$

Suppose that $M^+(\hat{\xi}) \neq F(s_0)$, i.e., there is an s' in $M^+(\hat{\xi}) \setminus F(s_0)$. Then $s' \in M_n$ for some n and $s' = s_1 * s_2 * \cdots * s_n$. As $M^+(\hat{\xi})$ is a face, we then have

$$s_1, \ldots, s_n \in M^+(\hat{\xi}) \cap M_1 \quad .$$

At least one of them, say s_1, must be outside $F(s_0)$ since $F(s_0)$ is a submonoid and the sum is outside $F(s_0)$. Now let

$$\xi'(s) = \begin{cases} \hat{\xi}(s) & \text{for } s \in F(s_0) \\ 0 & \text{otherwise} \end{cases}$$

and define $\xi\colon\ M \to \underset{\sim}{R}_+$ by

$$\xi(s) = \frac{\xi'(s)}{\left(\sum_{x \in E} \nu(x)\xi'(t(x))\right)^n} \quad \text{for} \quad s \in M_n .$$

Clearly $\xi \in \hat{M}_\nu$ and

$$\sum_{x \in E} \nu(x)\ \xi'(t(x)) < 1.$$

Therefore $\xi(s_0) > \hat{\xi}(s_0)$, which is a contradiction. Hence we must have

$$M^+(\hat{\xi}) = F(s_0) \ ,$$

which was to be proved.

So, the support of the estimated measure is closely tied to the way the observations can occur. If s_0 is observed after n experiments, $t(x_i)$ _must_ be in the smallest face containing s_0 for all $i = 1,\ldots,n$ as

$$s_0 = t(x_1)*\cdots*t(x_n) \ .$$

The estimate contains this information as the support of $P_{\hat{\xi}}$ is reduced to the subset of E, where $t(x) \in F(s_0)$.

5. Extreme Families of Random Walks on Monoids

First we introduce the definition of an extreme family of Markov chains as given by Lauritzen (1974).

Let $(E_n,\ n=1,2,\ldots)$ be a family of discrete, at most denumerable spaces and $Q = (Q_{mn})_{m \leqq n}$ a family of matrices with elements $q_{mn}(x,y)$, $x \in E_m$, $y \in E_n$, satisfying

$$\left.\begin{array}{ll} q_{mn}(x,y) \geqq 0, & \sum_{x \in E_m} q_{mn}(x,y) = 1 \\ & \\ Q_{mn}\ Q_{np} = Q_{mp} & \text{for} \quad m \leqq n \leqq p \ . \end{array}\right\}$$

and

$\mathfrak{M}(Q)$ denotes the set of sequences of probability measures $\mu = (\mu_n, n = 1,2,\ldots)$ such that μ_n is a probability measure on E_n and

$$\mu_m = Q_{mn}\, \mu_n \quad \text{for all} \quad m \leqq n.$$

$\mathfrak{M}(Q)$ is a convex set and $\mathcal{E}(Q)$ shall mean the extreme points of $\mathfrak{M}(Q)$. The family of Markov chains on $\Pi_{n=1}^{\infty} E_n$ defined by the initial distributions

$$P^{\mu}\{X_1 = x\} = \mu_1(x)$$

and the transition probabilities for $m \leqq n$

$$P^{\mu}\{X_n = y | X_m = x\} = \begin{cases} q_{mn}(x,y)\,\dfrac{\mu_n(y)}{\mu_m(x)} & \text{for} \quad \mu_m(x) \neq 0 \\ \mu_n(y) & \text{otherwise} \end{cases}$$

where μ takes all values in $\mathcal{E}(Q)$, is called the extreme family generated by Q.

For all $\mu \in \mathfrak{M}(Q)$, the matrices Q_{mn} define the "backward conditional probabilities", i.e.

$$P^{\mu}\{X_m = x | X_n = y\} = q_{mn}(x,y) \quad \text{for} \quad m \leqq n$$

and μ_n the marginal distributions of X_n, i.e.

$$P^{\mu}\{X_n = x\} = \mu_n(x) \quad .$$

Now consider the sequence of spaces M_n, n=1,2,... where $M = \cup_{n=0}^{\infty} M_n$ corresponding to a general exponential model. Let

$$Y_n = t(X_1)*\cdots*t(X_n),$$

where $X_1,\ldots,X_n$ are independent and identically distributed as

$$P_\xi\ \{X = x\} = \nu(x)\ \xi(t(x))\ ,$$

where $\xi \in \hat{M}_\nu$ is unknown. Let

$$\alpha(s) = \sum_{x\in E:t(x)=s} \nu(x)\quad .$$

We have $\alpha(s) > 0$ for all $s \in M_1$. Define the n'th convolution α^{*n} of α as $\alpha^{*1}(s) = \alpha(s)$ and

$$\alpha^{*n}(s) = \sum_{a*b=s} \alpha(a)\ \alpha^{*(n-1)}\ (b)\quad \text{for}\quad m=2,3,\ldots$$

We have $\alpha^{*n}(s) > 0$ for all $s \in M_n$.

$Y_1, Y_2,\ldots$ forms a Markov chain on $\Pi_{n=1}^{\infty} M_n$ and we have for $m \leqq n$ and $P_\xi\ \{Y_n=y\} > 0$

$$P_\xi\ \{Y_m = x|Y_n = y\} = \frac{P_\xi\{Y_n=y|Y_m=x\}\cdot P_\xi\{Y_m=x\}}{P_\xi\{Y_n=y\}}$$

$$= \frac{\left(\sum_{a:\,a*x=y} \alpha^{*(n-m)}\ (a)\ \xi(a)\right)\ \alpha^{*m}(x)\ \xi(x)}{\alpha^{*n}(y)\ \xi(y)}$$

$$= \frac{\alpha^{*m}(x)}{\alpha^{*n}(y)} \sum_{a:\,a*x=y} \alpha^{*(n-m)}(a)\ .$$

We shall now consider the system of backward conditional distributions $(Q_{mn})_{m\leqq n} = Q$ with elements $q_{mn}(x,y)$ $x \in M_m, y \in M_n$ given by

$$q_{mn}(x,y) = \frac{\alpha^{*m}(x)}{\alpha^{*n}(y)} \sum_{a:\,a*x=y} \alpha^{*(n-m)}\ (a)$$

We shall find $\mathcal{E}(Q)$ and in fact show that

$$\mathcal{E}(Q) = \left\{\mu:\ \exists\, \xi \in \hat{M}_\nu:\quad \mu_n(x) = \alpha^{*n}(x)\ \xi(x)\right\}$$

i.e. exactly the family of Markov chains made up by sequences of sufficient statistics from successive repetitions of experiments giving rise to random variables following a general exponential model.

First we need a lemma. For $\mu = (\mu_n, n=1,2,\dots) \in \mathfrak{M}(Q)$, $x \in M_k$, $k=1,2,\dots$ define the sequence $T_{x,k}\,\mu$ by

$$T_{x,k}\,\mu_n(a) = \begin{cases} \dfrac{\mu_{n+k}(a*x)}{\mu_k(x)}\ \dfrac{\alpha^{*k}(a)\ \alpha^{*n}(x)}{\alpha^{*(n+k)}(a*x)} & \text{if}\quad \mu_k(x) > 0 \\ \mu_n(a) & \text{otherwise} \end{cases}$$

<u>Lemma 4.1</u> $\mu \in \mathfrak{M}(Q) \Rightarrow T_{x,k}\mu \in \mathfrak{M}(Q)$.

<u>Proof:</u>

Clearly, if $\mu_k(x) = 0$, $T_{x,k}\,\mu = \mu$ and hence $T_{x,k}\,\mu \in \mathfrak{M}(Q)$.

If $\mu_k(x) \neq 0$, we have

$$\sum_{a \in M_n} T_{x,k}\,\mu_n(a) = \frac{1}{\mu_k(x)} \cdot \sum_{b \in M_{n+k}}\ \sum_{a:a*x=b} \frac{\alpha^{*k}(x)\ \alpha^{*n}(a)}{\alpha^{*(n+k)}(b)}\ \mu_{n+k}(b)$$

$$= \frac{1}{\mu_k(x)} \cdot \sum_{b \in M_{n+k}} q_{n,n+k}(x,b)\ \mu_{n+k}(b) = 1 \qquad (5.1)$$

as μ was known to be in $\mathfrak{M}(Q)$.

Further, we get

$$\sum_{b \in M_n} q_{mn}(a,b)\ T_{x,k}\,\mu_n(b)$$

$$= \sum_{b \in M_n}\ \sum_{c:\,c*a=b} \frac{\alpha^{*m}(a)\ \alpha^{*(n-m)}(c)}{\alpha^{*n}(b)}\ \frac{\mu_{n+k}(b*x)}{\mu_k(x)}\ \frac{\alpha^{*n}(b)\ \alpha^{*k}(x)}{\alpha^{*(n+k)}(b*x)}$$

$$= \frac{\alpha^{*m}(a)\ \alpha^{*k}(x)}{\alpha^{*(m+k)}(a*x)} \cdot \frac{1}{\mu_k(x)} \cdot \sum_{b \in M_n} \sum_{c:c*a=b} \frac{\alpha^{*(m+k)}(a*x)\ \alpha^{*(n-m)}(c)}{\alpha^{*(n+k)}(b*x)} \mu_{n+k}(b*x)$$

Now

$$\{c:c*a = b\} \subseteq \{c:c*a*x = b*x\}$$

and

$$\{b*x:\ x \in M_n\} \subseteq M_{n+k},$$

so we have the inequality

$$\sum_{b \in M_n} q_{mn}(a,b)\ T_{x,k}\ \mu_n(b)$$

$$\leq \frac{\alpha^{*m}(a)\ \alpha^{*k}(x)}{\alpha^{*(m+k)}(a*x)} \frac{1}{\mu_k(x)} \cdot \sum_{d \in M_{n+k}} \sum_{c:c*(a*x)=d} \frac{\alpha^{*(m+k)}(a*x)\ \alpha^{*(n-m)}(c)}{\alpha^{*(n+k)}(d)} \mu_{n+k}(d)$$

$$= \frac{\alpha^{*m}(a)\ \alpha^{*k}(x)}{\alpha^{*(m+k)}(a*x)} \frac{1}{\mu_k(x)} \sum_{d \in M_{n+k}} q_{m+k,\ n+k}(a*x,d)\ \mu_{n+k}(d)$$

$$= T_{x,k}\ \mu_m(a)\ ,$$

or in short,

$$T_{x,k}\ \mu_m(a) \geq \sum_{b \in M_n} q_{mn}(a,b)\ T_{x,k}\ \mu_n(b)\ . \qquad (5.2)$$

But by (5.1) both sides of (5.2) add up to one when summing over $a \in M_m$, and hence we must have equality and therefore $T_{x,k}\ \mu \in \mathfrak{M}(Q)$, which was to be proved.

Proposition 4.1

$$\mu \in \mathcal{E}(Q) \iff \exists\, \xi \in \hat{M}_\nu : \quad \mu_n(x) = \alpha^{*n}(x)\ \xi(x)\ .$$

Or, in words, the extreme family generated by Q consist of "random walks",

$$Y_n = t(X_1)*\cdots*t\ (X_n)$$

where the $t(X)$'s are independent and identically distributed with the distribution of $Y_1 = t(X_1)$ given by

$$\mu_1(x) = \alpha(x)\ \xi(x) \quad \text{for some} \quad \xi \in \hat{M}_\nu\ .$$

Proof: The proof consists of the following steps. First we use lemma 4.1 to obtain a representation of any $\mu \in \mathcal{M}(Q)$ as a convex combination of other elements $(T_{x,k}\ \mu)$ in $\mathcal{M}(Q)$. If μ then is extreme, μ must be equal to these other elements, which gives us an equation. This equation is essentially the homomorphism equation and we can then establish "$\Rightarrow$". The proof of this and of lemma 4.1 is a direct generalization of the proof of an equivalent result for M being the k-dimensional integer lattice in Neveu (1964) and Martin-Löf (1973). To prove "$\Leftarrow$" we show that a proper mixture of homorphisms, cannot be a homomorphism.

Suppose now that $\mu \in \mathcal{M}(Q)$ is extreme, i.e. $\mu \in \mathcal{E}(Q)$. We note that the equation for $\mu \in \mathcal{M}(Q)$.

$$\mu_n(a) = \sum_{b \in M_{n+k}} \ \sum_{c:c*a=b} \frac{\alpha^{*n}(a)\ \alpha^{*k}(c)}{\alpha^{*(n+k)}(a*c)}\ \mu_{n+k}\ (a*c) \qquad (5.3)$$

implies that

$$\mu_n(a)=0 \Rightarrow \mu_{n+k}(a*c) = 0 \quad ,$$

as $q_{n,n+k}(a,b) > 0$ for all $b \in M_{n+k}$. Hence (5.3) can be rearranged to

$$\mu_n(a) = \sum_{b \in M_{n+k}} \sum_{c:c*a=b} \mu_k(c) \cdot T_{c,k}\, \mu_n(a)$$

This gives μ as a convex combination of $T_{c,k}\,\mu$, $c \in M_k$ for all $k = 1,2,\ldots$ and as μ was supposed to be extreme,

$$\mu_n(a) = T_{c,k}\, \mu_n(a) \quad \text{for all } k = 1,2,\ldots \text{ and } c \in M_k.$$

Thus, for all $c \in M_k$, $k = 1,2,\ldots$ such that $\mu_k(c) > 0$, we must have

$$\frac{\mu_{n+k}(a*c)}{\alpha^{*(n+k)}(a*c)} = \frac{\mu_n(a)}{\alpha^{*n}(a)} \; \frac{\mu_k(c)}{\alpha^{*k}(c)} \tag{5.4}$$

If we let

$$h_n(a) = \frac{\mu_n(a)}{\alpha^{*n}(a)} \quad ,$$

(5.4) becomes

$$h_{n+k}(a*c) = h_n(a)\, h_k(c) \tag{5.5}$$

But as

$$\mu_k(c) = 0 \Rightarrow \mu_{n+k}(a*c) = 0$$

(5.5) must hold for <u>all</u> n, k, $a \in M_n$, $c \in M_k$. As $M_n \cap M_m = \emptyset$ for $m \neq n$, we can define a mapping ξ from M to $\underset{\sim}{R}_+$ by

$$\xi(a) = h_n(a) \quad \text{for} \quad a \in M_n ,$$

and by (5.5) $\xi \in \hat{M}$.

If μ is extreme, we then have

$$\frac{\mu_n(a)}{\alpha^{*n}(a)} = \xi(a) \Longleftrightarrow \mu_n(a) = \alpha^{*n}(a)\ \xi(a).$$

As μ is a probability, we have

$$\sum_a \mu_n(a) = \sum_a \alpha^{*n}(a)\ \xi(a) = 1 ,$$

i.e., that $\xi \in \hat{M}_\nu$. We have proved "$\Rightarrow$".

Now suppose that

$$\mu_n^{\xi_o}(x) = \alpha^{*n}(x)\ \xi_o(x) \quad \text{for some} \quad \xi_o \in \hat{M}_\nu .$$

All elements in $\mathfrak{M}(Q)$ are mixtures of the elements in $\mathcal{E}(Q)$. It follows from what we proved before that the set of sequences

$$\{\mu^\xi,\ \xi \in \hat{M}_\nu\} ,$$

where

$$\mu_n^\xi(x) = \alpha^{*n}(x)\ \xi(x) , \tag{5.6}$$

<u>contains</u> $\mathcal{E}(Q)$. A fortiori any $\mu \in \mathfrak{M}(Q)$ can be represented as a mixture of elements of the form (5.6) This is in particular true for μ^{ξ_o}. Hence, there is a probability measure P on $\hat{M}_\nu$ so that for all $x \in M$

$$\alpha^{*n}(x)\ \xi_o(x) = \int_{\xi \in \hat{M}_\nu} \alpha^{*n}(x)\ \xi(x)\ d\ P(\xi)\ . \qquad (5.7)$$

But (5.7) is equivalent to

$$\xi_o(x) = \int_{\xi \in \hat{M}_\nu} \xi(x)\ d\ P(\xi) \quad \text{for all} \quad x \in M$$

Using the homomorphism property, we have

$$\left(\int_{\xi \in \hat{M}_\nu} \xi(x)\ d\ P(\xi)\right)^2 = (\xi_o(x))^2 = \xi_o(x * x)$$

$$= \int_{\xi \in \hat{M}_\nu} \xi(x * x)\ d\ P(\xi) = \int_{\xi \in \hat{M}_\nu} (\xi(x))^2\ d\ P(\xi) \qquad (5.8)$$

But (5.8) implies that $P\{\xi_o\} = 1$ and hence that ξ_o is extreme. The proof is complete.

6. Additional Comments

The families defined in the present paper are sometimes identical to the completion of a regular canonical exponential family as defined by Barndorff-Nielsen (1973).

Let T be a finite subset of the k-dimensional integer lattice. Define T_n by

$$T_1 = T \quad \text{and} \quad T_n = T + T_{n-1} \quad \text{for} \quad n = 2,3,\ldots$$

Let $(M,*)$ be the semigroup

$$M = \{(t,n): \; t \in T_n, \; n \in \underset{\sim}{N}\} \quad ,$$

with the composition

$$(s, m) * (t, n) = (s + t, m + n) \quad .$$

If we have a general exponential model on a space E with $t(x) = (g(x), 1)$, where g is a mapping from E onto T, this model can be identified with the completion of the canonical exponential family generated by ν and g in exactly the same way as in Martin-Löf (1973). This situation is present in example 3.1 of the present paper.

If $g(E) = T$ contains more than integer lattice points, this is not necessarily the case as the following example shows.

Example 6.1 Let $E = \{(0,0), (1,0), (0,1), (\sqrt{2}/4, 1/2)\}$. Let $M = \cup_{n=1}^{\infty} M_n$, where

$$M_n = \{(x,y,n): \; (x,y) \in E_n\}$$

where E_n is recursively defined as

$$E_1 = E \text{ and } E_n = E + E_{n-1} \text{ for } n = 2,3,\ldots$$

and the composition on $(M,*)$ is defined as

$$(x,y,n) * (x', y', n') = (x + x', y + y', n + n').$$

Let $\nu(x,y) = 1$ for all (x,y) in E and

$$t(x,y) = (x,y,1) \quad .$$

The completion of the exponential family generated by ν and t would consist of all probability measures of exponential type with support equal to E, $\{(0,0), (0,1)\}$, $\{(0,0), (1,0)\}$ or $\{(0,1), (1,0)\}$ as well as the probabilities degenerate at $(1,0)$, $(0,0)$ and $(0,1)$. Because $(\sqrt{2}/4, 1/2)$ is in the interior of the convex hull of E, no probability in the family would be degenerate at this point.

The subset F of M given by

$$F = \left\{\left(n \frac{\sqrt{2}}{4}, \frac{n}{2}\right) : n \in \underset{\sim}{N}\right\} \tag{6.8}$$

is obviously a face of M. Hence the general exponential model corresponding to ν, t and M <u>will contain</u> the probability degenerate in $(\sqrt{2}/4, 1/2)$.

This example illustrates the essential difference between the completions defined by Barndorff-Nielsen (1973) and the models in this paper: the "completions" are defined via geometrical concepts in $\underset{\sim}{R}^k$ or via topological considerations whereas the general exponential models are derived via algebraic structure in the statistics, thus letting the actual observations play a more prominent role. If one after n experiments in the above example obtains the value of the statistic to be $(n\sqrt{2}/4, n/2)$, this <u>must</u> be because $(\sqrt{2}/4, 1/2)$ was observed n times. This is reflected in the estimated probability measure, which will be degenerate at $(\sqrt{2}/4, 1/2)$ as can be seen from proposition 4.3.

Acknowledgments

I am grateful to all colleagues at the Institute of Mathematical Statistics, University of Copenhagen for several inspiring discussions on the topics of the present paper. My sincere thanks are due to Ann Mitchell, Imperial College, Søren Johansen and Niels Keiding, Copenhagen, T. W. Anderson, Stanford and Per Martin-Löf, Stockholm for reading various versions of this paper and suggesting important alterations in the manuscript. Finally, this paper would never exist if Hans Brøns, Copenhagen ever missed an opportunity to point out, that algebraic properties of sufficient statistics are essential features of the models they correspond to and vice versa.

References

Barndorff-Nielsen, O. (1973). Exponential Families and Conditioning. Copenhagen, Wiley.

Lauritzen, S. L. (1974). Sufficiency, Prediction and Extreme Models. Scandinavian Journal of Statistics, **1**.

Martin-Löf, P. (1973). Repetitive Structures. Department of Mathematical Statistics. University of Stockholm. (Mimeographed report).

Neveu, J. (1964). Chaînes de Markov et théorie du potentiel. Ann. Faculté des Sciences, Clermont, **3**, 37-65.

Steffen L. Lauritzen
Institute of Mathematical Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen
DENMARK