Extreme Point Models in Statistics [with Discussion and Reply]

Author(s): Steffen L. Lauritzen, Ole E. Barndorff-Nielsen, A. P. Dawid, Persi Diaconis and Søren Johansen

Source: *Scandinavian Journal of Statistics*, 1984, Vol. 11, No. 2 (1984), pp. 65-91

Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics

Stable URL: http://www.jstor.com/stable/4615945

Scand J Statist 11: 65–91, 1984

Extreme Point Models in Statistics

STEFFEN L. LAURITZEN

Aalborg University Centre

ABSTRACT. We give a survey of the general theory of extreme point models in statistics, i.e. statistical models that are given as the extreme points of the convex set of probability measures satisfying (in a general sense) a symmetry condition. Special emphasis is payed to examples, some of which are only partially solved, some are classical and some are recent.

Key words: exchangeability, de Finetti's theorem, Rasch models, repetitive structures, sufficiency

0. Introduction and summary

The present paper surveys the general theory of extreme point models as developed in Lauritzen (1982), (in the following denoted [EF]) with special emphasis on examples. The proofs of almost all results are omitted from this presentation and appropriate references are given instead.

It is our aim to point out that a number of statistical models have the common structure that they are given as the extreme points of a certain convex set of probability measures, given by symmetry properties (in a general sense). We show that such models have various desirable properties.

We thereby want to emphasize that the relation between a statistical analysis and a statistical model is a kind of duality, in the sense that the model can be "generated" by the analysis and vice versa, reflecting the perception that the model is a particular "representation" of the analysis.

The probably best known example is the Bernoulli model for repeated tosses of a coin corresponding to de Finetti's (1931) theorem:

Let $X_1, \ldots X_n, \ldots$ be a sequence of random variables taking values in $\{0,1\}$ and suppose their joint distribution is *exchangeable*, i.e. that

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$$

for all $n \in \mathbf{N}$ and all permutations $\pi \in S(n)$, the symmetric group of order n. Here $X \stackrel{\mathcal{D}}{=} Y$ means that X and Y have the same distribution.

Then there is a unique probability measure μ on $[0,1]$ such that for all $n \in \mathbf{N}$,

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \int_0^1 \theta^{\Sigma_{i=1}^n x_i} (1-\theta)^{n-\Sigma_{i=1}^n x_i} \mu(d\theta). \tag{0.1}$$

Moreover, the limit

$$\bar{X}_\infty = \lim_{n \to \infty} \frac{X_1 + \ldots + X_n}{n}$$

exists almost surely and μ *is the distribution of* $\bar{X}_\infty$. We can twist this result slightly by realising that

$$(X_1, \ldots, X_n) \stackrel{\mathcal{D}}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}) \quad \forall \pi \in S(n)$$

if and only if for all $t \in \{0, \ldots, n\}$

$$P\{X_1 = x_1, \ldots, X_n = x_n | X_1 + \ldots + X_n = t\} = \frac{1}{\binom{n}{t}} 1_{\{t\}}(x_1 + \ldots + x_n), \tag{0.2}$$

and thereby noticing that the class of exchangeable probability measures on $\{0,1\}^{\mathbf{N}}$ is *the largest class of probabilities* for which

$$t_n: \{0,1\}^n \to \{0,1,\ldots,n\}$$
$$t_n(x_1, \ldots, x_n) = x_1 + x_2 + \ldots + x_n$$

for each n is *sufficient* with (0.2) as conditional distributions given $t_n(X_1, \ldots, X_n) = t$.

Further, this class is a *convex set*, with the independent Bernoulli measures $P_\theta\{\cdot\}$ as *extreme points*. We have a unique integral representation as

$$P\{\cdot\} = \int_{[0,1]} P_\theta\{\cdot\}\, \mu(d\theta)$$

where

$$P_\theta\{X_1 = x_1, \ldots, X_n = x_n\} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n - \sum_{i=1}^n x_i},$$

so that, according to $P_\theta, X_1, X_2, \ldots$ are just independent Bernoulli trials with probability of "success" equal to θ.

Moreover, because the measure μ is "the limiting distribution of the sufficient statistic", the *probability P can be identitied from complete observation of the entire sequence* $(X_1, X_2, \ldots)$ *if and only if P is an extreme point* itself. In fact, observing a single realization of the process $X_1, X_2 \ldots$ gives rise to just one value of $\bar{X}_\infty$. This value identifies the mixing measure μ if and only if μ is degenerate.

The fact that the independent Bernoulli measures are the extreme points of the convex set of measures, for which $x_1 + \ldots + x_n$ for all n is sufficient and (0.2) are the corresponding conditional distributions, will in the sequel be expressed as "the model of independent identical Bernoulli trials is *an extreme point model*".

We shall see that such extreme point models in fact occur quite commonly in statistics and that results like the above integral representation and interpretation of the representing measure as the limiting distribution of the sufficient statistic, are of a quite general nature.

Generalizations of de Finetti's theorem can be made in several directions. One is to exchange the spaces $\{0,1\}$ with more general measure spaces, as done by Hewitt & Savage (1955).

Another is to consider invariance of distributions under the action of other groups than the permutation group like e.g. the group of rotations of $\mathbf{R}^n$ where the random variables $X_1, \ldots, X_n$ all take values in $\mathbf{R}$, cf. e.g. Kingman (1972).

Yet another method is to specify other sufficient statistics than $x_1 + \ldots + x_n$.

All these generalizations are special cases of the same theory, to be surveyed in the following.

Section 1, 2 and 3 are devoted to a survey of the general results, 4 to a discussion of the relation between the theory and inference problems, whereas the remaining sections are devoted to examples.

The theory is intimately related to considerations in statistical mechanics, see e.g.

Preston (1979), although slightly different in technique and very different in scope, since the "purpose" of statistical mechanical models is to explain observable macroscopic behaviour from microscopic properties, whereas statistical models are used to infer about macroscopic behaviour from observable microscopic phenomena.

A genuine survey of the connection of this line of research to other area's would itself be interesting but is outside the scope of the present paper.

A survey paper which has some slight overlap with the present but mostly deals with different aspects and examples is given by Diaconis & Freedman (1982).

1. Preliminaries. Markov kernels

Throughout the paper sample spaces will be assumed *Polish*, i.e. topological spaces that are metrizable, complete and separable. All maps are considered continuous and all probability measures regular Borel measures. If $\mathscr{X}$ is a Polish space, $\mathscr{B}(\mathscr{X})$ denotes the σ-algebra of Borel sets of $\mathscr{X}$ and $\mathbf{P}(\mathscr{X})$ the (Polish) space of probability measures on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ endowed with the weak topology, i.e.

$$P_\alpha \to P \Leftrightarrow \int f dP_\alpha \to \int f dP$$

for all continuous and bounded real-valued functions on $\mathscr{X}$.

A Markov kernel Π from $\mathscr{X}$ to $\mathscr{Y}$ is a function

$$\Pi: \mathscr{X} \times \mathscr{B}(\mathscr{Y}) \to \mathbf{R}$$

such that

(i) $\forall x \in \mathscr{X}: \Pi(x, \cdot) \in \mathbf{P}(\mathscr{Y})$
(ii) the map $x \to \Pi(x, \cdot)$ is continuous.

If Π_1 is a Markov kernel from $\mathscr{X}$ to $\mathscr{Y}$ and Π_2 is a Markov kernel from $\mathscr{Y}$ to $\mathscr{Z}$ the composition

$$\Pi = \Pi_2 \Pi_1$$

is the Markov kernel from $\mathscr{X}$ to $\mathscr{Z}$ given as

$$\Pi(x, C) = \int_{\mathscr{Y}} \Pi_2(y, C) \Pi_1(x, dy).$$

A Markov kernel Π from $\mathscr{X}$ to $\mathscr{Y}$ induces a continuous map from $\mathbf{P}(\mathscr{X})$ to $\mathbf{P}(\mathscr{Y})$ by

$$(\Pi\mu)(B) = \int_{\mathscr{X}} \Pi(x, B) \mu(dx)$$

and the composition above $\Pi = \Pi_2 \Pi_1$ is just the composition of the corresponding maps.

If $t: \mathscr{X} \to \mathscr{Y}$ is a continuous map, it induces a Markov kernel Π_t as

$$\Pi_t(x, B) = 1_B(t(x)).$$

We shall instead of Π_t just write t and we do thus not distinguish between a map and the Markov kernel induced by it. The application of the Markov kernel induced by t to a probability $\mu \in \mathbf{P}(\mathscr{X})$ gives the *lifted* measure $\mu \cdot t^{-1}$:

$$(t\mu)(B) = (\Pi_t \mu)(B) = \int_{\mathscr{X}} \Pi_t(x, B)\mu(dx) = \int_{\mathscr{X}} 1_B(t(x))\mu(dx) = \mu(t^{-1}(B)).$$

If μ is a probability on $\mathscr{X}$ and $t: \mathscr{X} \to \mathscr{Y}$ is a map, a Markov kernel Q from $\mathscr{Y}$ to $\mathscr{X}$ is a regular conditional probability given t if

(i) $Qt\mu = \mu$
(ii) $tQ = I_y$

where the first condition in more conventional terms says

$$\mu(B) = \int_{\mathscr{Y}} Q(y, B)\mu \cdot t^{-1}(dy), \forall y \in \mathscr{Y}.$$

2. Repetitive structures. Sufficiency

The basis for our further investigations is the notion of a repetitive structure, introduced in the discrete setting by P. Martin-Löf (1974). A *repetitive structure* is a projective system of continuous maps and Polish spaces. More precisely it is given by:

(1) A partially ordered ($<$) set I which is directed to the right, i.e.

$\forall i, j \in I\ \exists k \in I$: $i < k, j < k$.

(2) A family $(\mathscr{X}_i, i \in I)$ of Polish spaces.
(3) A family $(p_{ij})_{i<j}$ of continuous surjective maps

$p_{ij}: \mathscr{X}_j \to \mathscr{X}_i$

Satisfying for $i < j < k$:

$p_{ij} p_{jk} = p_{ik}$.

These maps are called *projections*.

We think of I as describing a family of experiments with corresponding sample spaces $\mathscr{X}_i$ and p_{ij} define the relation between these, i.e. in which sense i is a "subexperiment" of j. For technical reasons we assume throughout that I has a *cofinal sequence*, i.e. there is a sequence $(i_n)_{n \in \mathbf{N}} \subset I$ such that

$\forall i \in I\ \exists n \in \mathbf{N}$: $i < i_n$.

The typical example of a repetitive structure has I as a subset of the *set of subsets of a countable set T*, ordered by inclusion, and if $A \in I$, then

$$\mathscr{X}_A = \underset{t \in A}{\times} \mathscr{X}_t$$

and the projections, p_{AB} are just coordinate projections, i.e. for $A \subset B$:

$p_{AB}(x_t, t \in B) = (x_t, t \in A)$.

Associated with such a projective system we have the *projective limit*

$$\mathscr{X} = \varprojlim_{i \in I} \mathscr{X}_i = \{(x_i, i \in I) \mid \forall i < j : p_{ij}(x_j) = x_i\}$$

which is Polish (when I has a cofinal sequence) when equipped with the topology of pointwise convergence. In the standard example of a repetitive structure, $\mathscr{X}$ can be identified with the product $\times_{t\in T}\mathscr{X}_t$.

We have the *canonical projections*

$$p_j: \mathscr{X} \to \mathscr{X}_j \quad \text{given as } p_j(x_i,\ i\in I) = x_j$$

and these satisfy $p_{ij}p_j = p_i$ for $i<j$.

Since later $\mathscr{X}$ shall be equipped with families of probability measures, we introduce the "random variables" taking values in $\mathscr{X}_i$

$$X_i(x) = p_i(x), \quad \text{where now } p_{ij}(X_j) = X_i.$$

The version of Kolmogorov's consistency theorem given in Bourbaki (1969), ensures a one-to-one correspondence between a probability measure μ on $\mathscr{X}$ and a consistent family $(\mu_i, i\in I)$ where $\mu_i \in \mathbf{P}(\mathscr{X}_i)$ and $p_{ij}\mu_j = \mu_i$, $i<j$. We shall thus write $\mu=(\mu_i, i\in I)$ without ambiguity. μ_i can be interpreted as the distribution of X_i induced by μ.

If $\{\mu(\theta),\ \theta\in\Theta\}$ is a parametrized family of probability measures on $\mathscr{X}$ we shall say that a system of continuous surjective maps

$$t_i: \mathscr{X}_i \to \mathscr{Y}_i, \quad i\in I$$

is *sufficient* if there exist Markov kernels

$$\mathscr{Y}_i \overset{Q_i}{\to} \mathscr{X}_i$$

such that the following conditions are satisfied

$$\left.\begin{array}{ll} \text{(i)} & Q_i\, t_i\, \mu_i(\theta) = \mu_i(\theta) \quad \forall\theta\in\Theta,\ \forall i\in I \\ \text{(ii)} & t_i\, Q_i = I_{\mathscr{Y}_i} \quad \forall i\in I \\ \text{(iii)} & Q_i\, t_i\, p_{ij}\, Q_j = p_{ij}\, Q_j \quad \forall i<j. \end{array}\right\} \tag{2.1}$$

Letting $Y_i = t_i(X_i)$, (i) and (ii) together ensure that for $A\in\mathscr{B}(\mathscr{X}_i)$, $y\in\mathscr{Y}_i$

$$Q_i(A|y) = \mu(\theta)\{X_i\in A \,|\, Y_i = y\},$$

i.e. that t_i is sufficient in the usual sense, and (iii) ensures that X_i and Y_j are *conditionally independent* given Y_i, i.e. that t_i is *transitive* in the sense of Bahadur (1954), or more precisely, that for $B\in\mathscr{B}(\mathscr{Y}_j)$

$$Q_i(A|y) = \mu(\theta)\{X_i\in A \,|\, Y_i = y,\ Y_j\in B\}.$$

For details on this, see [EF].

It follows ([EF], p. 205) that for any increasing sequence $i_1<i_2<\ldots<i_n<\ldots$, the process made up by the corresponding values of the sufficient statistics

$$Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}, \ldots$$

is a *Markov process*.

Of special interest to us is the tail-σ-algebra

$$\mathscr{A}_t = \bigcap_{i\in I} \sigma(Y_j,\ j>i)$$

and also the *extended* tail σ-algebra

$$\mathcal{A}_{\mu(\theta)} = \{A \in \mathcal{B}(\mathcal{X}) | \forall i \in I : \mu(\theta)\{A | X_i = x\} \overset{\text{a.s.}}{=} \mu(\theta)\{A | Y_i = t_i(x)\}\},$$

$\mathcal{A}_{\mu(\theta)}$ is the σ-algebra of events that are conditionally independent of X_i given Y_i for all $i \in I$. The transitivity of the sequence of statistics ensures that we have

$$\mathcal{A}_{\mu(\theta)} \supseteq \mathcal{A}_t.$$

3. Maximal and extremal families. Basic limit theorems

Suppose now that we have given a repetitive structure, a family of continuous surjective maps $t_i: \mathcal{X}_i \to \mathcal{Y}_i$, $i \in I$ and a family Q_i, $i \in I$ of Markov kernels from $\mathcal{Y}_i$ to $\mathcal{X}_i$ such that (ii) and (iii) of (2.1) are satisfied. We could then ask for the family of all probability measures $\mu \in \mathbf{P}(\mathcal{X})$ such that also (i) is satisfied, i.e. such that

$$Q_i t_i \mu_i = \mu_i, \quad \forall i \in I.$$

We denote this family by $\mathcal{M}$ and call it the *maximal family* corresponding to $(t_i)_{i \in I}$ and $(Q_i)_{i \in I}$. $\mathcal{M}$ is maximal with the property that $(t_i)_{i \in I}$ is a sufficient system and Q_i are the conditional distributions given $t_i(X_i)$, in other words, (t_i) will be sufficient for any family of the type

$$\{\mu(\theta),\ \theta \in \Theta\} \subseteq \mathcal{M}.$$

$\mathcal{M}$ is a convex set and we shall by $\mathcal{E}$ denote the *extreme points* of $\mathcal{M}$, i.e.

$$\mu \in \mathcal{E} \Leftrightarrow [\mu = \lambda\mu_1 + (1-\lambda)\mu_2 \wedge \lambda \in]0, 1[\Rightarrow \mu = \mu_1 = \mu_2],$$

where of course both μ, μ_1 and μ_2 are elements of $\mathcal{M}$. $\mathcal{E}$ is called the *extremal family* corresponding to $(t_i)_{i \in I}$, $(Q_i)_{i \in I}$ and the repetitive structure. A model of the type having

$$\Theta = \mathcal{E},\ \mu_i(\theta) = \theta_i, \quad i \in I$$

is called a *canonical model* or an *extreme point model.*

We have the following results ([EF] Prop. IV, 1.1).

3.1. Proposition. *$\mathcal{E}$ and $\mathcal{M}$ are Polish. $\mathcal{M}$ is a simplex in the sense that for all $\mu \in \mathcal{M}$ there is exactly one $P_\mu \in \mathbf{P}(\mathcal{E})$ such that*

$$\mu(A) = \int_{\mathcal{E}} e(A) P_\mu(de), \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

Further, if we for any $\mu \in \mathcal{M}$, any fixed cofinal and increasing sequence $\mathbf{i} = (i_1 < i_2 < \ldots < i_n < \ldots)$ and any fixed $i \in I$, $A \in \mathcal{B}(\mathcal{X}_i)$ define the sequence of random variables

$$Z_n^{\mathbf{i}}(A) = Q_i(p_{ii_n}^{-1}(A) | Y_{i_n})$$

$$(= \mu\{X_i \in A | Y_{i_n}\}), \quad i_n > i$$

we have ([EF], pp. 58–59, p. 208):

3.2. Proposition. *$Z_n^i(A)$ is a bounded reverse martingale w.r.t. the σ-algebras $\sigma(Y_{i_m}, m \geq n)$. Further we have*

$$E_\mu Z_n^i(A) = \mu(A)$$

$$Z_n^i(A) \underset{n\to\infty}{\overset{\text{a.s.}\,\mu}{\to}} Z_\infty(A) = \mu\{A \mid \mathcal{A}_t\}.$$

The extreme points are characterized by the following main theorem ([EF], p. 209):

3.3. Proposition. *For $\mu \in \mathcal{M}$ the following are equivalent:*

(i) $\mu \in \mathcal{E}$
(ii) $\mathcal{A}_t$ is μ-trivial ($A \in \mathcal{A}_t \Rightarrow \mu(A) \in \{0, 1\}$)
(iii) $\mathcal{A}_\mu$ is μ-trivial
(iv) $Z_\infty(A) = \mu(A)$ a.s. μ.

And as corollaries we have

3.4. Corollary. *P_μ in Proposition 3.1. defines the distribution of $Z_\infty(\cdot)$ in the sense that for all $B \in \mathcal{B}(\mathcal{E})$*

$$\mu\{Z_\infty(\cdot) \in B\} = P_\mu(B)$$

To see this, realise that because everything is countably generated, (iv) can be extended to hold for all $A \in \mathcal{B}(\mathcal{X})$ simultaneously and

$$\mu\{Z_\infty(\cdot) \in B\} = \int_{\mathcal{E}} e(Z_\infty(\cdot) \in B) P_\mu(de) = \int_{\mathcal{E}} 1_B(e) P_\mu(de) = P_\mu(B). \qquad \square$$

3.5. Corollary. $P_\mu\{Z_\infty(\cdot) \in \mathcal{E}\} = 1$, which is obtained by letting $B = \mathcal{E}$ in Corollary 3.4. $\square$

Interpreting these results we can think of $Z_n^i(A)$ as a canonical estimate of $\mu(A)$ and Proposition 3.3 (iv) tells us that if $\mu \in \mathcal{E}$ this estimate is strongly consistent, i.e.

$$\mu_n(A) = Z_n^i(A) \underset{n\to\infty}{\to} Z_\infty(A) = \mu(A) \quad \text{a.s. } \mu.$$

and further that this even characterizes the measures in $\mathcal{E}$ among those in $\mathcal{M}$.

In other words, the parameter in an extreme point model is always consistently estimable, *and is defined in terms of the observations* ($\theta = Z_\infty$).

In specific examples it is in general very difficult to identify $\mathcal{E}$ in a reasonably explicit way. We shall in the following give various examples where this has been done at least in part.

4. Implications for statistical inference

The theory of extreme point models *per se* has no direct relation to statistical inference as it is usually discussed, since it concerns an analysis of the *model* rather than the inference procedures themselves, the model being regarded as fixed. We believe that our considerations do have consequences for statistical practice, where after all the establishing of the model *is* an important part of the inference procedure. Many conditioning arguments used in statistical inference can actually be seen as modifications of the model rather than the

inference procedures. We hope to throw more light on this when discussing the examples. The results briefly surveyed in the previous two sections play a reasonably prominent role in the modern Bayesian approach to statistics, cf. de Finetti (1975). See also the discussion of intersubjectivistic parameters in Dawid (1979).

Consider a given repetitive structure. A Bayesian would to this associate a personal, subjective probability $\mu_0 \in \mathbf{P}(\mathscr{X})$, describing her uncertainty about all possible outcomes of the various experiments. She might also specify a system of statistics $t_i: \mathscr{X}_i \to \mathscr{Y}_i$ that are interesting for predictive or other purposes. Markov kernels Q_i can then be calculated as

$$Q_i(A|Y) = \mu_0\{X_i \in A | Y_i = y_i\}$$

and these will automatically satisfy

$$Q_i\, t_i\, \mu_0 = \mu_0, \quad t_i\, Q_i = I_{\mathscr{Y}_i}.$$

If the system of statistics is reasonably well behaved, we would also have the transitivity condition satisfied:

$$Q_i t_i p_{ij} Q_j = p_{ij} Q_j.$$

Because of these relations μ_0 will now automatically satisfy $\mu_0 \in \mathscr{M}$, where $\mathscr{M}$ is the maximal family corresponding to $(t_i)_{i\in I}$, $(Q_i)_{i\in I}$.

The parameter can now be *defined* as the random variable

$$\theta = Z_\infty$$

which takes values in $\Theta = \mathscr{E}$.

The *prior* distribution of θ is simply given as

$$\mu_0\{\theta \in B\} = P_{\mu_0}(B)$$

where P_{μ_0} is the measure defining the integral representation of μ_0 as

$$\mu_0(\cdot) = \int_{\mathscr{E}} e(\cdot) P_{\mu_0}(de).$$

The inference on the unknown, random parameter θ is now performed by computing the *posterior* distribution of θ given the observations, i.e.

$$\mu_0\{\theta \in B | X_i = x\} = \mu_0\{\theta \in B | Y_i = t_i(x)\}$$

If we denote these posterior distributions by $P_{X_i}(\cdot)$, these become random variables taking values in $\mathbf{P}(\Theta) = \mathbf{P}(\mathscr{E})$ and we can show that the Bayesian posteriors converge almost surely (w.r.t. the subjective μ_0) to the "true" value, i.e. the measure degenerate at θ:

4.1. Proposition. *For all $B \in \mathscr{B}(\mathscr{E}) = \mathscr{B}(\Theta)$ and all cofinal increasing sequences $(i_n)_{n\in\mathbf{N}}$*

$$P_{X_{i_n}}(B) \underset{n\to\infty}{\to} 1_B(\theta) \quad \text{a.s. } \mu_0.$$

Proof. Let $Z = 1_B(\theta)$. Z is bounded and $\mathscr{B}(\mathscr{X})$ measurable and further

$$Z_n = E_{\mu_0}(Z | X_{i_n}) = P_{X_{i_n}}(B).$$

The family of σ-algebras $\sigma(X_{i_n})$ is increasing and $\mathscr{B}(\mathscr{X})=\lim_{n\in\mathbf{N}}\uparrow\sigma(X_{i_n})$. Thus $Z_n=P_{X_{i_n}}(B)$ is a martingale and

$$Z_n\to Z=1_B(\theta) \quad \text{a.s.} \quad \mu_0. \qquad \square$$

Note that the main reason for this result to be true is that θ actually *is a function of the observations,* a fact that we find a central part of the theory.

Another way of looking at the theory is given as follows. Suppose we have a given model for the repetitive structure, a model that might have emerged from probabilistic and other mathematical considerations that are external to the repetitive structure itself. Then we could find a system of $(t_i)_{i\in I}$, such that this is sufficient and transitive. Ways of finding such a system is indicated in ch. II.2 of [EF]. This gives then automatically the kernels $(Q_i)_{i\in I}$ defined by the conditional distributions of X_i given t_i and we can (in principle) find $\mathscr{E}$ and $\mathscr{M}$. If our first family $\{\mu(\theta),\ \theta\in\Theta\}$ is identical to $\mathscr{E}$, we know that the model is canonical. If not, we might expect difficulties and might want to modify the model or rather change the model to the canonical. It is worth noting that observations could *never* contradict the canonical model if it did not contradict the first one, since it is *in principle impossible to distinguish* whether or not Z_∞ is random, even asymptotically, from only one observation from the projective limit.

Finally we might along the lines of Martin-Löf (1974), use the procedure as a tool for model building, by specifying the statistics t_i as a family of interesting data reductions and the conditional distributions from elementary symmetry considerations. The construction of the extreme point (canonical) model can then be performed and we obtain a statistical model which is consistent with the symmetry considerations and where the parameter in a natural way is defined as the limit of the interesting statistics in an infinitely large experiment.

In the discrete case the use of *uniform* distributions for the conditional distribution of the observations given the statistic can also be justified by a maximum entropy argument rather than symmetry, an approach which is often used in statistical mechanics.

It is tempting to think of the probabilities defining the conditional distributions of the observations given the value of the statistics as *descriptive* in the sense that a statistician chooses to describe a data set x by the *reduction* of x to the value $t(x)$ and $Q(\cdot\,|t(x))$. In this sense the extreme point model is the canonical reduction or *description* of the infinitely large experiment or "population". This line of thinking was implicitely contained in the excellent notes of Martin-Löf (1970).

Finally, we might in the situation, where we have only one sample space $\mathscr{X}$, find it useful to embed $\mathscr{X}$ into a repetitive structure and use this for constructing models or analysing a given model.

The main problem is that the actual identification of $\mathscr{E}$ is hard in any concrete case. However the number of cases, where $\mathscr{E}$ is known in a reasonably explicit way is gradually growing and various hints to solve similar problems can be taken from these examples, some of which are described in detail in [EF].

5. Exponential families

The most well-known class of extreme point models are given by the exponential families corresponding to the simplest possible repetitive structure (independent identical repetitions). In the discrete case this looks as follows. The partially ordered set I is the set of integer intervals $\{1,\dots,n\}$, $n\in\mathbf{N}$. The sample spaces are given as

$$\mathscr{X}_{\{1,\ldots,n\}} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_n$$

where $\mathscr{X}_i = \mathscr{X}$ is a fixed discrete, at most countable set. We shall use the notation $\mathfrak{X}_n$ for $\mathscr{X}_{\{1,\ldots,n\}}$. The projections are coordinate projections and the projective limit can be identified with the infinite product space

$$\mathfrak{X} = \mathscr{X}_1 \times \ldots \times \mathscr{X}_n \times \ldots .$$

If $t: \mathscr{X} \to S$ is a fixed function into an Abelian semigroup $(S, \oplus)$, a canonical model is given as

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} (a(x_i)\phi(\theta)^{-1}\theta(t(x_i)), \quad \theta \in D \tag{5.1}$$

where $a: \mathscr{X} \to]0, \infty[$ is a fixed function and D is a subset of $\mathrm{EXP}(S)$, *the exponential functions* on $(S, \oplus)$, i.e. those non-negative real-valued functions f satisfying

$$f(s \oplus t) = f(s)f(t) \quad \forall s, t \in S \tag{5.2}$$

and

$$D = \left\{ \theta \in \mathrm{EXP}(S) | \phi(\theta) = \sum_{x \in \mathscr{X}} a(x)\theta(t(x)) < \infty \right\}.$$

In other words, according to $p(\cdot|\theta)$, the coordinate random variables are independent and identically distributed with distributions

$$p(x|\theta) = a(x)\phi(\theta)^{-1}\theta(t(x)).$$

Combining (5.1) and (5.2) we get

$$p(x_1, \ldots, x_n|\theta) = \left(\prod_{i=1}^{n} a(x_i) \right) \phi(\theta)^{-n}\theta(t(x_1) \oplus \ldots \oplus t(x_n)),$$

so that clearly

$$t_n(x_1, \ldots, x_n) = t(x_1) \oplus \ldots \oplus t(x_n)$$

is sufficient. That these models in fact are extreme point models is shown in [EF] Ch. III, see also Lauritzen (1975).

In the special case where $(S, \oplus) \subseteq (\mathbf{N}^d, +)$ we get the usual exponential families, *extended* as in Barndorff-Nielsen (1973, 1978), as also shown by Martin-Löf (1974).

The generalized exponential families described above have not necessarily "finite-dimensional statistics", i.e. $(S, \oplus)$ is not necessarily finitely generated, and also the support of $p(\cdot|\theta)$ vary with θ.

Examples include the family of arbitrary distributions on a fixed set $\mathscr{X}$ corresponding to S being the set of positive integer valued measures on $\mathscr{X}$ with finite support, $a(x) = 1$ and t being

$$t(x) = \varepsilon_x$$

where ε_x is the measure degenerate at x. Then

$$t_n(x_1, \ldots, x_n) = \varepsilon_{x_1} + \ldots + \varepsilon_{x_n}$$

is *the empirical distribution.*

Also the family of uniform distributions on $\{1, \ldots, \theta\}$, $\theta \in \mathbf{N}$ is exponential in the above sense with $S=(\mathbf{N}, \vee)$, where

$$m \vee n = \max\{m, n\},$$

$t(x)=x$ and $a(x)=1$. It is interesting to note that the estimation theory for these generalized exponential families is at least as elegant as for the usual exponential families, see [EF] Ch. III, where also further examples are discussed.

As discussed in e.g. Jaynes (1957), usual exponential families can be justified from maximum entropy considerations. This is also true for the generalized exponential families as we shall now show in the finite case.

Let $\mathscr{X}$ be discrete and finite and consider the exponential family

$$p_\theta(x) = \theta(t(x))/\phi(\theta), \quad \theta \in \mathrm{EXP}(S)$$

where $t: \mathscr{X} \to (S, \oplus)$ is a given statistic.

Proposition III.4.5 and III.4.8 of [EF] ensures the existence and uniqueness of a $\hat{\theta} \in D$ such that for a given value $t(x_0)=t_0$

$$\log\hat{\theta}(t_0) - \log\eta(t_0) = E_{\hat{\theta}} \log\frac{\hat{\theta}(t(X))}{\eta(t(X))} \quad \forall \eta \in D_{F_{t_0}} \tag{5.3}$$

where $D_{F_{t_0}}$ is a specified subset of $\mathrm{EXP}(S)$, see [EF] for details. If we assume $F_{t_0}=S$ (which is equivalent to $\hat{\theta}(x)>0 \ \forall s \in S$), $D_{F_{t_0}}$ is the set of strictly positive exponential functions, and we get by considering $\eta_0 \equiv 1$ and a simple manipulation that (5.3) is equivalent to

$$\log\eta(t_0) = E_{\hat{\theta}} \log\eta(t(X)) \quad \forall \eta \in D_{F_{t_0}}. \tag{5.4}$$

Let now $\mathbf{P}_{t_0}$ be the set of probability measures on $\mathscr{X}$ such that the equation analogous to (5.4) is satisfied, i.e.

$$\mu \in \mathbf{P}_{t_0} \Leftrightarrow \log\eta(t_0) = \sum_{x \in \mathscr{X}} \mu(x) \log\eta(t(x)) \quad \forall \eta \in D_{F_{t_0}}. \tag{5.5}$$

Then clearly $p_{\hat{\theta}}$ *is in* $\mathbf{P}_{t_0}$ and we shall show that $p_{\hat{\theta}}$ has *maximal entropy among the measures in* $\mathbf{P}_{t_0}$, where the entropy is defined as

$$En(\mu) = -\sum_{x \in \mathscr{X}} \mu(x) \log\mu(x).$$

We get

$$\begin{aligned} En(p_{\hat{\theta}}) - En(\mu) &= \sum_x \mu(x) \log\mu(x) - \sum_x p_{\hat{\theta}}(x) \log p_{\hat{\theta}}(x) \\ &= \sum_x \mu(x) \log\mu(x) - E_{\hat{\theta}} \log\hat{\theta}(t(X)) + \log\phi(\hat{\theta}) \\ &= \sum_x \mu(x) \log\mu(x) - \log\hat{\theta}(t_0) + \log\phi(\hat{\theta}) \end{aligned}$$

$$= \sum_x \mu(x) \log \mu(x) - \sum_x \mu(x) \log \hat{\theta}(t(x)) + \log \phi(\hat{\theta})$$

$$= \sum_x \mu(x) \log \frac{\mu(x)}{p_{\hat{\theta}}(x)} \geqslant 0$$

where the last inequality is the information inequality. (5.5) represents a definition of the *expectation* of a semigroup-valued statistic as an additive functional on EXP(S)

$$[E_\mu t(X), \eta] \stackrel{\text{def}}{=} E_\mu \log \mu(t(X))$$

and $p_{\hat{\theta}}$ is then the probability measure on $\mathscr{X}$ maximizing the entropy subject to the constraint

$$E_\mu t(X) = \bar{t}_0$$

where $[\bar{t}_0, \eta] = \log \eta(t_0)$.

It should be noted that the fact that exponential families are extreme point models depends heavily on the particular simple repetitive structure and the symmetry in the conditional distributions. If we e.g. consider the family of distributions given by $X_1, \ldots, X_n, \ldots$ being independent with

$$p(x_n|\theta) = \pi_n^{x_n} e^{\theta x_n}/(1+\pi_n e^{\theta}), \quad \theta \in \mathbf{R}$$

where $x_n \in \{0, 1\}$ and $(\pi_n)_{n \in \mathbf{N}}$ is a fixed and known sequence of positive real numbers, we get for the joint distribution of $X_1, \ldots, X_n$

$$p(x_1, \ldots, x_n|\theta) = \left(\prod_{i=1}^n \pi_i^{x_i} \right) \prod_{i=1}^n (1+\pi_i e^{\theta})^{-1} e^{\theta \sum_{i=1}^n x_i},$$

i.e. for all n we have an exponential family with

$$t_n(x_1, \ldots, x_n) = x_1 + \ldots + x_n$$

being minimal sufficient (and transitive). The conditional distributions of the observations given the statistics are

$$q(x_1, \ldots, x_n|t) = \left(\prod_{i=1}^n \pi_i^{x_i} \right) \Big/ \gamma_t(\pi_1, \ldots, \pi_n)$$

if $x_1 + \ldots + x_n = t$, and zero otherwise, where γ_t are the elementary symmetric functions

$$\gamma_t(\pi_1, \ldots, \pi_n) = \sum_{\substack{x_1 + \ldots + x_n = t \\ x_i \in \{0,1\}}} \left(\prod_{i=1}^n \pi_i^{x_i} \right).$$

It follows from the results of Pitman (1978), see also [EF] pp. 91 ff., that this family corresponds to an extreme point model if and only if

$$\sum_{n=1}^{\infty} \pi_n/(1+\pi_n)^2 = \infty. \tag{5.5}$$

This again reflects the fact that θ is consistently estimable from one realization of $X_1, \ldots, X_n, \ldots$ if and only if (5.5) is fulfilled. As is easily seen, (5.5) is equivalent to

$$\lim_{N\to\infty} V_\theta\left(\sum_{n=1}^{N} X_n\right) = \sum_{n=1}^{\infty} \pi_n e^{\theta}(1+\pi_n e^{\theta})^{-2} = \infty.$$

A related example is the model given by $X_1, \ldots, X_n, \ldots$ independent and Poisson distributed with $E_\theta X_n = \theta^n$ for $\theta \in]0, \infty[$. The point probabilities become

$$p(x_1, \ldots, x_n|\theta) = \left(\prod_{i=1}^{n} x_i!\right)^{-1} \theta^{\sum_{i=1}^{n} i x_i} e^{-\sum_{i=1}^{n} \theta^i}.$$

Again we have for each n an exponential family with

$$t_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} i x_i$$

being minimal sufficient (and transitive) and the conditional distributions of the observations given the statistic being

$$q(x_1, \ldots, x_n|t) = \begin{cases} c_n(t)^{-1} & \text{if } x_1+2x_2+\ldots+nx_n = t \\ 0 & \text{otherwise} \end{cases}$$

where

$$c_n(t) = \sum_{\nu=0}^{\infty} \frac{1}{\nu!} \sum_{x_1+\ldots+nx_n=t} \binom{\nu}{x_1, \ldots, x_n}.$$

The situation is here somewhat more delicate than in the previous example but one can show ([EF] pp. 119ff.) that

(a) $p(\cdot|\theta)$ is an extreme point if and only if $\theta \geq 1$;
(b) the distributions $\tilde{p}(\cdot|y)$, $y \in \mathbf{N}$ obtained by conditioning on Y_∞:

$$\tilde{p}(\cdot|y) = p(\cdot|Y_\infty = y, \theta_0)$$

for some $\theta_0 \in]0, 1[$, where

$$Y_\infty = \sum_{i=1}^{\infty} iX_i,$$

are also extreme points.

It is unfortunately an open problem whether there are other extreme points but we conjecture that this is not the case.

6. Exponential families for Markov chains

Consider a sequence of random variables $X_0, X_1, \ldots, X_n, \ldots$, taking values in a discrete and at most countable set $\mathscr{X}$. Consider statistics

$$t_n(x_0, \ldots, x_n) = (x_0, \{n_{xy}\}_{(x,y)\in\mathscr{X}\times\mathscr{X}}, x_n)$$

where n_{xy} are the *transition counts*

$$n_{xy} = \#\{k|(x_k, x_{k+1})) = (x, y)\}.$$

Diaconis & Freedman (1980) have shown that the extreme points of the class of probabilities for which t_n is sufficient for all n and the conditional distribution of the observations given the statistics is uniform on the set of strings $(x_0, \ldots, x_n)$ with given first and last value and given transition counts, fall into three classes

(1) recurrent Markov chains
(2) processes starting with a fixed string of transient states and continuing as recurrent Markov chains
(3) totally transient processes.

If $\mathscr{X}$ is finite, only the types (1) and (2) apply. Results of Höglund (1974) indicate that a similar result can be obtained by considering the statistics

$$t_n(x_0, \ldots, x_n) = (x_0, t(x_0, x_1) \oplus \ldots \oplus t(x_{n-1}, x_n), x_n)$$

where

$$t: \mathscr{X} \times \mathscr{X} \to S$$

is a fixed statistic into an Abelian semigroup $(S, \oplus)$. In the finite case the non-degenerate extreme points then correspond to recurrent Markov chains with transition probabilities

$$P_\theta\{X_{n+1} = y | X_n = x\} = e_\theta(y)\,\theta(t(x, y))/(e_\theta(x)\,\phi(\theta)), \quad \theta \in D^+$$

where e_θ is a positive eigenvector corresponding to the maximal eigenvalue $\phi(\theta)$ of the positive matrix:

$$\{\theta(t(x, y))\}_{(x,y)\in\mathscr{X}\times\mathscr{X}}$$

$$\sum_{y \in \mathscr{X}} \theta(t(x, y))\, e_\theta(y) = \phi(\theta)\, e_\theta(x)$$

and D^+ is the set of positive exponential functions. This gives joint probabilities having $X_0 = x_0$ degenerate and

$$p(x_1, \ldots, x_n) = e_\theta(x_n)/e_\theta(x_0)\,\phi(\theta)^{-n}\theta(t(x_0, x_1) \oplus \ldots \oplus t(x_{n-1}, x_n)).$$

Unfortunately we do not know a clear proof of this result at present, nor do we know how it extends to the infinite case.

Note that the example studied by Diaconis & Freedman corresponds to the case where S is the semigroup of positive integer valued measures on $\mathscr{X}\times\mathscr{X}$ with $t(x, y)$ being the measure degenerate at (x, y), thus being an analogue of the family of arbitrary distributions, as considered in the previous section.

7. Linear normal models

The case of projective systems of linear normal models can fortunately be solved completely. Let I be a directed set as usual and let

$$\mathscr{X}_i = \mathbf{R}^{n_i}$$

where $n_i<n_j$ for $i<j$ and $n_i<\infty$. As projections we take a system of linear maps A_{ij}, $i<j$ satisfying

$$A_{ij}A_{jk}=A_{ik} \quad \text{for } i<j<k$$
$$A_{ij}A_{ij}^*=I_i, \quad i<j$$

where * denotes transpose and I_i the identity on $\mathbf{R}^{n_i}=\mathscr{X}_i$. The projective limit can be identified with $\mathbf{R}^{\mathbf{N}}$.

Let further L_i, $i\in I$ be a family of linear subspaces of $\mathscr{X}_i$ satisfying

$$A_{ij}(L_j)=L_i,$$

and take now as sufficient statistics

$$t_i(x)=(B_ix, \|x\|^2)$$

where B_i is the orthogonal projection onto L_i and $\|\cdot\|$ is the standard Euclidean distance. The conditional distributions $q(\cdot|(y,s^2))$ should be taken uniform on the sphere

$$\{x\in \mathbf{R}^{n_i}|B_ix=y, \ \|x\|^2=s^2\}$$

If we let $\tilde{A}_{ij}$ be the restriction of A_{ij} to L_j, $\{(L_i, i\in I), \tilde{A}_{ij}, i<j\}$ is a projective system too, and we can find the projective limit

$$L=\varprojlim_{i\in I} L_i$$

with canonical projections $\tilde{A}_i$. If we now consider the model given as

$$\Theta=L\times]0,\infty[$$

and the distribution of X_i given $\theta=(\xi,\sigma^2)$ as

$$X_i\sim N(\tilde{A}_i\xi, \sigma^2I_i),$$

we can show ([EF] pp. 217ff.) that this is an extreme point model if and only if both of the following are satisfied:

(a) $n_i-\dim(L_i)\to\infty \ \ i\to\infty$
(b) $\forall i: A_{ij}B_jA_{ij}^*\to 0 \ \ j\to\infty$

Condition (a) says that the degrees of freedom available for estimating σ^2 should tend to infinity and (b) that the maximum likelihood estimate of $\tilde{A}_i\xi=E_\xi X_i$ should be consistent. To see the latter we note that based on X_j, the maximum likelihood estimate of $\tilde{A}_j\xi$ is B_jX_j and since $\tilde{A}_i=A_{ij}\tilde{A}_j$, $A_{ij}B_jX_j$ is the maximum likelihood estimate of $\tilde{A}_i\xi$ based on X_j. But this has variance equal to

$$(A_{ij}B_j)(A_{ij}B_j)^*=A_{ij}B_jA_{ij}^*$$

since $B_jB_j^*=B_j^2=B_j$ because B_j is an orthogonal projection. As a special case we have the additivity models for two-way classification, i.e. where $I=\mathbf{N}\times\mathbf{N}$,

$$\mathscr{X}_{m,n}=\{(x_{ij})_{i=1,\ldots,m,j=1,\ldots,n}|x_{ij}\in\mathbf{R}\}$$
$$L_{m,n}=\{(x_{ij})|x_{ij}=\alpha_i+\beta_j, \ i=1,\ldots,m; j=1,\ldots,n\}$$

and A_{ij} are coordinate projections.

If we consider the models defined as $(X_{ij})_{(i,j)\in\mathbf{N}\times\mathbf{N}}$ being independent and normally distributed with

$$X_{ij}\sim N(\xi_{ij},\sigma^2)$$

where

$$(\xi_{ij})_{i=1,\ldots,m,j=1,\ldots,n}\in L_{m,n}\quad \forall m,n,\ \sigma^2\geq 0,$$

this is an extreme point model as shown in [EF] pp. 223–24.

Another special case is determined by the linear regression problems with $I=\mathbf{N}$, $\mathscr{X}_n=\mathbf{R}^n$ and

$$L_n=\{(x_1,\ldots,x_n)|x_i=\alpha+\beta t_i\}$$

where $(t_1,\ldots,t_n,\ldots)$ is a fixed and known sequence of real numbers, the model having $(X_n)_{n\in\mathbf{N}}$ independent and

$$X_n\sim N(\alpha+\beta t_n,\sigma^2)\quad \alpha,\beta\in\mathbf{R},\ \sigma^2>0.$$

This is an extreme point model if and only if

$$SSD_t^n=\sum_{i=1}^{n}(t_i-\bar{t}^n)^2\to\infty,\quad n\to\infty \tag{7.1}$$

where $\bar{t}^n=(t_1+\ldots+t_n)/n$, see [EF] p. 225 for a proof of this.

These results reflect the fact that the row—and column effects in the additive model for the two-way classification are consistently estimable when the number of rows and columns both tend to infinity. Also that the parameters in the regression model are consistently estimable if and only if condition (7.1) is fulfilled.

8. Models for 0−1 matrices

An interesting class of examples different from the usual class of exponential families are extreme point models for 0−1 matrices.

Aldous (1981) has investigated the class of doubly infinite arrays $(X_{ij})_{i,j\in\mathbf{N}}$ of random variables that are *row-column exchangeable* (RCE-arrays), i.e. satisfying for all $m,n,\pi\in S(m)$, $\sigma\in S(n)$

$$(X_{ij})_{i=1,\ldots,m;\,j=1,\ldots,n}\overset{\mathscr{D}}{=}(X_{\pi(i)\sigma(j)})_{i=1,\ldots,m;j=1,\ldots,n}$$

where $S(m)$ is the group of permutations of m elements. If we consider the special case where X_{ij} takes values in $\{0,1\}$ this corresponds to the repetitive structure, where $I=\mathbf{N}\times\mathbf{N}$, $\mathscr{X}_{[m,n]}$ is the set of $m\times n$ matrices with elements either zero or one, $t_{[m,n]}$ being a maximal invariant under the action of $S(m)\times S(n)$ on $\mathscr{X}_{[m,n]}$ defined by permutation of the indices. Finally the row-column exchangeability corresponds to the conditional distribution of the matrix given the statistic being uniform of the corresponding orbit.

The extreme points of the class of RCE-distributions are those that are *dissociated*, i.e. where

$$(X_{ij})_{i\leq m;j\leq n}\quad\text{and}\quad(X_{ij})_{i>m;j>n}$$

are independent for all m and n.

Further, any such array can be matched in distribution by choosing

(i) a measurable function $h:]0,1[^3 \to \{0,1\}$
(ii) independent sequences $(\xi_i)_{i\in\mathbf{N}}$, $(\eta_j)_{j\in\mathbf{N}}$, $(\lambda_{ij})_{i,j\in\mathbf{N}}$ of i.i.d. uniform $]0,1[$ random variables, and letting

$$X_{ij}^* = h(\xi_i, \eta_j, \lambda_{ij}),$$

i.e. X_{ij}^* is composed by h from a random row-effect ξ_i, a random column-effect η_j and an interaction λ_{ij}. An alternative formulation lets (X_{ij}^*) be conditionally independent given $(\xi_i)_{i\in\mathbf{N}}$, $(\eta_j)_{j\in\mathbf{N}}$ and the conditional probabiliity of (X_{ij}^*) being equal to 1 is now $\phi(\xi_i, \eta_j)$. The "parameter" is here h (or ϕ) but it is unfortunately not clear in which sense the model is overparametrized, since different choices of h can give the same ϕ and different choices of ϕ the same distribution of X_{ij}^*.

A different but related model is *Rasch's model for item analysis*. In this model, the random variables X_{ij} should be interpreted as the response of a person j to a question i and the model lets X_{ij} all be independent with

$$P_{\alpha,\beta}\{X_{ij}=1\} = 1-P_{\alpha,\beta}\{X_{ij}=0\} = \frac{\alpha_i\beta_j}{1+\alpha_i\beta_j}$$

where $\boldsymbol{\alpha}=(\alpha_i)_{i\in\mathbf{N}}$ and $\boldsymbol{\beta}=(\beta_j)_{j\in\mathbf{N}}$ are sequences of unknown non-negative parameters. The marginal point probabilities for $(x_{ij})_{i\le m, j\le n}$ are then

$$P_{\alpha,\beta}\{X_{[m,n]} = x_{[m,n]}\} = \prod_{i=1}^{m}\prod_{j=1}^{n}\frac{(\alpha_i\beta_i)^{x_{ij}}}{1+\alpha_i\beta_j} = \frac{\prod_{i=1}^{m}\alpha_i^{r_i}\prod_{j=1}^{n}\beta_j^{s_j}}{\prod_i\prod_j(1+\alpha_i\beta_j)} \tag{8.1}$$

Where $r_i=\Sigma_{j=1}^n x_{ij}$, $s_j=\Sigma_{i=1}^m x_{ij}$ are the row- and column sums of the matrix $x_{[m,n]}$.

The set of row- and column sums is a sufficient statistic and the conditional distribution of the matrix given the statistic is uniform on the set of matrices having the given row- and column sums.

In the repetitive structure described in connection with the RCE-arrays it is not clear at present what the corresponding extreme point model is. However one can show ([EF] Ch. IV.7.) that the following condition is *necessary* for $P_{\alpha,\beta}$ to be an extreme point

$$\text{A:}\sum_{n=1}^{\infty}\frac{\alpha_n}{(1+\alpha_n)^2} = \sum_{n=1}^{\infty}\frac{\beta_n}{(1+\beta_n)^2} = \infty.$$

Further the condition below is *sufficient* for $P_{\alpha,\beta}$ to be an extreme point:

$$\text{B:}\sum_{n=1}^{\infty}\frac{\alpha_n\beta_n}{(1+\alpha_n)(1+\beta_n)(1+\alpha_n\beta_n)} = \infty.$$

This reflects the fact that if A is not satisfied, the parameters are not consistently estimable from observation of the infinite matrix.

If e.g. $\Sigma_{n=1}^{\infty}\beta_n(1+\beta_n)^{-2}<\infty$ the α-s are not estimable, if $\Sigma_{n=1}^{\infty}\alpha_n(1+\alpha_n)^{-2}<\infty$ the β-s are not. On the other hand, if B is satisfied, both the α-s and β-s are consistently estimable up

to a constant factor that can be multiplied on the α-s and divided into the β-s. In the special case where all the $\alpha - s$ and $\beta - s$ are contained in a compact interval not containing zero, condition B is trivially fulfilled and consistent estimation is possible. Haberman (1977) investigated asymptotic properties of maximum likelihood estimates under this compactness assumption and showed *uniform* consistency and asymptotic normality along cofinal sequences, where the number of rows and columns tend to infinity at not too different speeds. It would be interesting to see asymptotic theory under condition B (or A).

Note that as a consequence we see that the set of extreme points is *not* closed in the weak topology. If we define $\boldsymbol{\alpha}^{(N)}$, $\boldsymbol{\beta}^{(N)}$ by

$$\alpha_i^{(N)} = \beta_i^{(N)} = \begin{cases} i^2 & \text{if } i \leqslant N \\ 1 & \text{if i} > N, \end{cases}$$

$P_{\boldsymbol{\alpha}^{(N)}, \boldsymbol{\beta}^{(N)}}$ are all extreme since condition B is satisfied and

$$P_{\boldsymbol{\alpha}^{(N)}, \boldsymbol{\beta}^{(N)}} \to P_{\boldsymbol{\alpha}, \boldsymbol{\beta}}$$

where $\alpha_i = i^2 = \beta_i$. But $P_{\boldsymbol{\alpha}, \boldsymbol{\beta}}$ is not extreme since condition A is violated.

If we consider the Rasch model in the different repetitive structure where the number of rows m is fixed and only the number of columns n is allowed to tend to infinity, $P_{\boldsymbol{\alpha}, \boldsymbol{\beta}}$ is never extreme. The corresponding extreme point model is then obtained by conditioning on the sequence $\mathbf{s} = (s_1, s_2, \ldots)$ of column sums and considering this as a fixed "parameter". We then get the probabilities

$$P_{\boldsymbol{\alpha}, \mathbf{s}}\{X_{ij} = x_{ij},\ i \leqslant m,\ j \leqslant n\} = \prod_{j=1}^{n} \left[\gamma_{s_j}(\alpha_1, \ldots, \alpha_m)^{-1} \prod_{i=1}^{m} \alpha_i^{x_{ij}} \right]$$

where $\gamma_s(\cdot, \ldots, \cdot)$ are the elementary symmetric functions. If the sequence $\mathbf{s}$ has infinitely many coordinates that are not equal to 0 or m, $P_{\boldsymbol{\alpha}, \mathbf{s}}$ is extreme and the remaining extreme points consist of various degenerate measures, see [EF], pp. 99ff.

Note that the idea of considering the extreme point model here leads to a conditional inference procedure which is also supported by other arguments, since the conditional model is free of the *nuisance* parameters $(\beta_1, \ldots, \beta_n, \ldots)$, see e.g. Andersen (1973).

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Mult. Anal.* **11**, 581–598.

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Mentalhygieinisk Forskningsinstitut, Copenhagen.

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25**, 423–462.

Barndorff-Nielsen, O. (1973). Exponential families and conditioning. Wiley, Copenhagen.

Barndorff-Nielsen, O. (1978). *Information, exponential families and conditioning*. Wiley, New York.

Bourbaki, N. (1969). *Integration*, Ch. IX. Hermann, Paris.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc.* **41**, 1–31.

Diaconis, P. & Freedman, D. (1980). de Finetti's theorem for Markov chains. *Ann. Prob.* **8**, 115–130.

Diaconis, P. & Freedman, D. (1982). Partial exchangeability and sufficiency. Tech. Rep. No. 190. Dept. of Statistics, Stanford. (To appear in *Sankhya*.)

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale del Lincei*, Ser. 6. Memorie, Classe di Scienze Fisiche, Mathematiche e Naturali, **4**, 251–299.

de Finetti, B. (1975). *Theory of probability*, Vol. I+II. Wiley, New York.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815–841.
Hewitt, E. & Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.
Höglund, T. H. (1974). Central limit theorems and statistical inference for Markov chains. *Z. f. Warscheinlichkeitstheorie u. v. Gebiete* **19**, 123–151.
Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630.
Kingman, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika* **59**, 492–493.
Lauritzen, S. L. (1975). General exponential models for discrete observations. *Scand J. Statist.* **2**, 23–33.
Lauritzen, S. L. (1982). *Statistical models as extremal families*. Aalborg University Press.
Martin-Löf, P. (1970). Statistiska Modeller. Notes by Rolf Sundberg. Mimeographed. (Swedish.)
Martin-Löf, P. (1974). Repetitive structures. In *Proceedings of Conference of Foundational Questions in Statistical Inference*. Aarhus 1974, Memoirs 1 (ed. O. Barndorff-Nielsen, P. Blæsild and G. Schou).
Pitman, J. W. (1978). An extension of de Finetti's theorem. *Adv. Appl. Prob.* **10**, 268–269.
Preston, C. (1979). Canonical and microcanonical Gibbs states. *Z. f. Wahrscheinlichkeitstheorie u. w. Gebiete* **46**, 125–158.

Received March 1981, in final form January 1984

Steffen L. Lauritzen, Institute of Electronic Systems,
Aalborg University Centre, Strandvejen 19, 9000 Aalborg, Denmark

DISCUSSION OF STEFFEN LAURITZEN'S PAPER

Ole E. Barndorff-Nielsen (University of Aarhus)

This impressive paper brings a variety of questions to mind, but the following comments are restricted to two of these.

It appears that there are some intriguing relations between the concept of extreme point models and the Fisherian ideas of information and conditionality.

Thus, for the model function

$$p(x_1,\ldots,x_n|\theta)=\prod_{i=1}^{n}\pi_i^{x_i}\prod_{i=1}^{n}(1+\pi_i e^{\theta})^{-1}e^{\theta\sum_{i=1}^{n}x_i},$$

considered in section 5, we have that the observed Fisher information is

$$j_n(\theta)=\sum_{i=1}^{n}\pi_i e^{\theta}/(1+\pi_i e^{\theta})^2.$$

For any fixed value of θ we find, writing $j_\infty(\theta)$ for $\lim_{n\to\infty} j_n(\theta)$,

$$j_\infty(\theta)=\infty \Leftrightarrow \sum_{i=1}^{\infty}\pi_i/(1+\pi_i)^2=\infty;$$

in other words, observed information tends to ∞ if and only if the model is an extreme point model. (Since the model is regular exponential we have $j_n(\theta)=V_\theta(\Sigma_{i=1}^n x_i)$ so that the above double-implication is a paraphrase of a remark in section 5.)

Similarly, for the sequence $x_1, x_2, \ldots, x_n, \ldots$ of Poisson variates with $E_\theta x_n=\theta^n$ we have that observed information based on $x_1,..,x_n$ is given by

$$j_n(\theta) = \sum_{i=2}^{n} i(i-1)\theta^{i-2},$$

and again the model is extreme if and only if $j_\infty(\theta)=\infty$.

As a third example, relating to the Bernoulli model discussed in section 1, suppose that θ in that model follows the beta distribution

$$P(\theta;\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\theta^{\alpha-1}(1-\theta)^{\alpha-1}.$$

The observed information on α provided by $x_1,\ldots,x_n$ is

$$j_n(\alpha) = 2\psi'(\alpha)-4\psi'(2\alpha)-\psi'\left(\sum_{i=1}^{n} x_i+\alpha\right)-\psi'\left(n-\sum_{i=1}^{n} x_i+\alpha\right)+4\psi'(n+2\alpha)$$

and (with probability 1)

$$j_\infty(\alpha) = 2\psi'(\alpha)-4\psi'(2\alpha).$$

The finiteness of $j_\infty(\alpha)$ is a reflection of the fact that whatever the value of α the law of the Bernoulli sequence is not an extreme point.

In considering observed information in the Rasch model, as given by (8.1), there is a slight simplification in reparametrising to $\varphi_i=\log\alpha_i$ and $\psi_j=\log\beta_j$. Observed information on φ_i, given the other parameters, is

$$\sum_{j=1}^{n} \frac{e^{\varphi_i+\psi_j}}{(1+e^{\varphi_i+\psi_j})^2},$$

from which it is immediately plausible that condition A, of section 8, should be a necessary, but not in general a sufficient, condition. It seems not unlikely that condition B has an interpretation in terms of the observed information matrix for the full parameter set.

The author has already indicated a relation to conditional inference, by the last paragraph of the paper. What considerations does the viewpoint of extreme point models lead to in connection with the so-called nonergodic exponential models (cf. Basawa & Scott (1983) and references therein)? For example, suppose $x_1,x_2,\ldots,x_n,\ldots$ follows the autoregressive process of order 1

$$x_n = \beta x_{n-1}+u_n, \quad n=1,2,\ldots,$$

where $x_0=0$ and $u_1,u_2,\ldots,u_n,\ldots$ is a sequence of independent and $N(0,1)$-distributed random variables while the regression parameter β can take any real value. The model for $(x_1,\ldots,x_n)$ is then a $(2,1)$ exponential model, of the nonergodic type, and inference on β based on $x_1,\ldots,x_n$ should in principle be performed conditional on an ancillary statistic. (The signed likelihood ratio ancillary or the affine ancillary can be used as approximate ancillaries, cf. for instance Barndorff-Nielsen (1980)). Is there a derived extreme point model for β in this case and how does the inference from this compare with the more traditional conditional approach?

References

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Basawa, I. V. & Scott, D. J. (1983). *Asymptotic optimal inference for non-ergodic models*. Springer, Heidelberg.

A. P. Dawid (University College, London)

For some years now I have followed with great interest and admiration the work that Steffen Lauritzen has so elegantly surveyed here. It has two qualities in particular which I appreciate: the beauty and depth of its mathematics, and the importance of its underlying statistical concepts. Here, at last, is a topic truly worthy of the much abused title "Mathematical Statistics".

On the mathematical side, we have a rich vein of deep and difficult problems from which, so far, only a few nuggets have been mined. The statistical side is, perhaps, even more challenging, since it addresses the most basic question in Statistics: Where does the model come from? In this we have a philosophical challenge to the very foundations of our subject.

The meaning and interpretation of a statistical model are all too rarely questioned. Is there a "true model"? How are probability assignments to be verified or falsified? A simple but instructive example is the following (Dawid, 1984). Consider a model for a non-replicable time-series $(X_t: 1 \leqslant t < \infty)$, under which the X_t are jointly normal with $E(X_t)=\mu$, $\mathrm{var}(X_t)=\sigma^2$, $\mathrm{cov}(X_s, X_t)=\varrho\sigma^2$ $(s \neq t)$. Here μ, σ^2, $\varrho \geqslant 0$ are arbitrary parameters. If $\varrho > 0$, this model asserts that different X's are positively correlated. What does this mean? How can it be established? Can ϱ be estimated? It turns out that none of the parameters is consistently estimable. In particular, the mle of ϱ is always 0. What is going on here?

Now the above model implies that the (X_t) are exchangeable. So, by de Finetti's Theorem, they will become independent after conditioning on the tail σ-field. In fact, letting $Y=\lim_{n\to\infty} n^{-1}\Sigma_{t=1}^{n} X_t$, $Z_t=X_t-Y$, we have $X_t=Y+Z_t$ with $Y \sim N(\mu, \varrho\sigma^2)$, $Z_t \sim N(0, (1-\varrho)\sigma^2)$ $(t=1,2,\ldots)$, all independently. The corresponding extreme-point model is obtained by conditioning on Y, thus recovering the submodel of the original model for which $\varrho=0$. No sequence of data could distinguish between a distribution with $\varrho>0$ and one with $\varrho=0$. What then is the meaning of the "correlation" ϱ? It seems clear to me that any use of the original model, rather than its extreme-point version, is fraught with danger.

Another important philosophical application of the ideas of extreme-point modelling is the following (Dawid, 1982*a*, *b*). Consider a collection X of variables, and suppose that it can be generally agreed that uncertainty about gX is the same as that about X, for all transformations g belonging to some *symmetry group* G. The archetypal example is exchangeability, with g a permutation of elements, but other important applications arise in experimental design layouts, for example. As with exchangeability, such problems can usually be embedded in an appropriate invariant repetitive structure. The extreme points of the family of all invariant probability distributions constitute a family which has every right to be called "the" appropriate statistical model under the assumed symmetries. (From exchangeability of events we can thus derive the Bernoulli model.) In other words, we have conjured probability distributions out of thin air by invoking ideas of symmetry alone. This magical process has, I believe, deep implications for the philosophy of Probability and Statistics.

References

Dawid, A. P. (1982*a*). Intersubjective statistical models. In *Exchangeability in probability and statistics* (ed. G. Koch and F. Spizzichino), pp. 217–232. North-Holland Publishing Company.

Dawid, A. P. (1982*b*). Probability, symmetry and frequency. Research Report 13, Department of Statistical Science, University College, London.

Dawid, A. P. (1984). A Bayesian view of statistical modelling. *Bayesian inference and decision techniques with applications: Essays in honor of Bruno de Finetti* (ed. P. K. Goel and A. Zellner). North-Holland Publishing Company (to appear).

Persi Diaconis (Stanford University)

It is a pleasure to congratulate Lauritzen on this superb summary of a rapidly expanding area. I would also like to acknowledge how much I appreciate his continued contributions to this area. Lauritzen realized the importance of Martin-Löf's treatment before anyone else. He made needed basic changes (extending the conditional distribution from uniform on the inverse images to a general projective system). All of this was done well before the articles of Dynkin & Föllmer made this area fashionable. I only lament that some of his research remains in unpublished technical reports and urge him to broader distribution.

I want to make a comment, add a reference, and ask a question.

Comment. The example in Section 5 concerning the extreme points of the model generated by independent Poisson variables with parameter θ^i is closely related to an important technical tool in the study of random permutations. Let S_n denote the symmetric group on n letters. The cycle vector of $\pi \in S_n$ is the vector $a=(a_1(\pi), a_2(\pi), \ldots)$ where $a_i(\pi)$ is the number of cycles in π of length i, when π is written as a product of cycles.

If π is chosen at random on S_n, then a is a random vector. Many features of permutations can be read from a. For example, a_1 is the number of fixed points of π; the distribution of a_1 is one of the oldest in probability—Monmort showed in 1708 that $P\{a_1=0\} \doteq 1-1/e$. The order of π is the smallest integer n such that $\pi^n=\mathrm{id}$; the order of π is the least common multiple of those i such that $a_i \neq 0$. Erdös & Turan have shown that as n tends to infinity, log order (π) has mean $(\log n)^2/2$, variance $(\log n)^3/3$, and a limiting standard normal distribution. Lloyd & Shepp discuss the distribution of the number of cycles and the maximum cycle length.

All of the distributions involved can nowadays be read off of a probabilistic construction closely related to Lauritzen's example:

Let N have a geometric distribution with parameter p. Choose a permutation at random by first choosing N and then choosing π uniformly in S_n. Lloyd & Shepp show that under this model, the random variables $a_i(\pi)$ have independent Poisson distributions with parameter P^i/i for $i=1,2,\ldots$. Here the sufficient statistic is $\Sigma\, ia_i$ as in Lauritzen's example, and the conditional distribution is

$$P\{a_1,\ldots,a_n | \Sigma\, i\, a_i = n\} = \begin{cases} \prod \dfrac{1}{i^{a_i}\, a_i!} & \text{if } \Sigma\, ia_i = n \\ 0 & \text{otherwise.} \end{cases}$$

This is of course the distribution of the vector a from a randomly chosen permutation in S_n.

In applying these facts to probability problems one computes the distribution of functionals under the independent Poisson model and then argues that this is the correct asymptotic distribution by using a Tauberian theorem. It would be of great interest if there were any other extreme points which could be used systematically.

Reference. In discussing Aldous' theorem on zero-one matrices, it is remarked that the extreme points are not completely known. This problem has been solved by Hoover who showed the only indeterminancy is measure preserving transformations of the coordinates of h. Alas, Hoover's proof makes heavy use of modem logic (model theory) and I do not know any probabilist who understands the result.

Question. One of Lauritzen's main contributions to this field is the explicit recognition that non-uniform distributions were needed as conditional distributions to get the usual

models. I would dearly love to hear a conversation between Lauritzen and Martin-Löf on this point. I wonder if he can provide us with a discussion of the many issues involved.

References

Hoover, D. A. (1981). Relations on probability spaces and arrays of random variables. Preprint, Institute for advanced Study, Princeton, New Jersey.

Shepp, L. A. & Lloyd, S. P. (1966). Ordered cycle length in a random permutation. *Trans. Amer. Math. Soc.* **121**, 340–357.

Søren Johansen (University of Copenhagen)

First I would like to congratulate Steffen Lauritzen on a good and readable paper which tries to tackle a very important problem in statistical inference namely the concept of model building.

The basic idea is that statistical models have more structure than is usually reflected in the discussion of inference principles.

By making this structure, i.e. the repetitive structure, more explicit one can embed the model into a family of models which is natural from the frequentist point of view.

An important aspect is the interpretation of the parameter, and in fact the definition of the parameter as the limit of a statistic in a large experiment.

My comments will concentrate mainly on this point by considering three examples from the paper.

1. Binary exchangeable variables.
2. Independent Poisson variables with $EX_n=\theta^n$.
3. Independent Gaussian variables with $EX_n=\varrho t_n$,

$$VX_n = \sigma^2 \quad \text{and} \quad \sum_{n=1}^{\infty} t_n^2 < \infty.$$

In the first example de Finetti's result shows that the model depends on a probability measure μ on $[0, 1]$, which can then be considered the parameter.

The family is then a maximal family and not extremal. The extremal model is found by restricting the parameter to be a one point measure on $[0, 1]$.

The implication of the theory is that μ it not identifiable from a realization of $\{X_n\}$. If the parameter is restricted to be a one point measure then it can be identified at least asymptotically and one gets that $\bar{X}_n \to \theta$ a.s. P_θ.

Thus the extremal family can be used to analyse identifiability and shows to what extent the parameter can be estimated consistently. In a sense nothing is lost by going to the extremal family as along as only one realization of the process is observed.

In the second example we have independent Poisson variables with mean $EX_n=\theta^n$.

If $\theta>1$ these measures are extremal and if $\theta<1$ then they are not. One should instead condition on $Y_\infty=\Sigma_{n=1}^{\infty} nX_n$, where $VY_\infty=\Sigma_{n=1}^{\infty} n^2\theta^n<\infty$, $\theta<1$.

Thus again the extremal family indicates exactly what can be estimated consistently, namely θ if $\theta>1$ otherwise just Y_∞.

The price for going to the extremal family is of course that we give up θ (if $\theta<1$) and replace it by Y_∞. Now Y_∞ may have an operational meaning, in the sense that we can estimate it, but θ may still have an outside meaning even if it is less than 1. The experiment performed does not help us in determining θ consistently, still an estimate of θ can be made and confidence intervals can also be made.

The final example is typical of this situation: In this case we have $\hat{\beta}_n=\Sigma_{i=1}^n X_i\, t_i/\Sigma_{i=1}^n t_i^2 \sim N(\beta,\, \sigma^2/\Sigma_{i=1}^n t_i^2)$ which converges to $\beta_\infty \sim N(\beta,\, \sigma^2/\Sigma_{i=1}^\infty t_i^2)$ which is easy to use for inference purposes. The extremality considerations imply that one should condition on β_∞ and consider the distributions of $X_1, \ldots, X_n$ given β_∞ which does not even involve the parameter β. Thus the parameter β should be removed from the statistical problem alltogether and yet β may have an outside interpretation, whereas the extremality considerations only give β an operational meaning in relation to a specific design.

My last comment will be on a comparison of the fixed effects and random effects model for one way analysis of variance.

Let the data be given by X_{ij}, $i=1, \ldots, k$, $j=1, \ldots, n$.

If we can take more measurements by sampling from each of the k groups, i.e. by letting $n \to \infty$, we perform the data reduction given by

$$\bar{X}_{i.} = \sum_{j=1}^{n} X_{ij}/n \quad \text{and} \quad s_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij}-\bar{X}_{i.})^2/k(n-1).$$

The extreme point model underlying this is clearly that of the X_{ij}'s being independent and $X_{ij} \sim N(\xi_i, \sigma^2)$.

If instead we want to extend our observation by including more groups then we condense the data further to

$$\bar{X} = \sum_{i=1}^{k} \sum_{j=1}^{n} X_{ij}/kn; \quad s_1^2 = \sum_{i=1}^{k} (\bar{X}_{i.}-\bar{X})^2/(k-1); \quad s_0^2.$$

The extreme point model generated by this data reduction is that the X_{ij}'s are independent between groups and that within a group we have $EX_{ij}=\xi$, $VX_{ij}=\omega^2$, $V(X_{ij}, X_{im})=\nu$, $j \neq m$, where $\nu \geqslant -\omega^2/(n-1)$.

If we allow both n and k to go to infinity we get that $\nu \geqslant 0$ and in this case we have the usual representation of the model:

$$X_{ij} = Y_i + U_{ij}$$

where the Y's and the U's are independent such that $EY_i=\xi$, $VY_i=\nu$, $EU_{ij}=0$ and $VU_{ij}=\sigma^2$, hence $\omega^2=\sigma^2+\nu$.

Thus the difference between the three models comes out very clearly if one includes in the model formulation in what way one wants to extend the set of observations.

Reply by Steffen Lauritzen

First I would like to thank the participants in the discussion for their interesting comments and questions.

A. P. Dawid points to the possibility of deriving statistical models from symmetry considerations. I should like to point out explicitly that in a mathematical sense, the "symmetry" approach is almost equivalent to the "sufficiency" approach. To each sufficient system of statistics, there is a system of transformation groups (those preserving the value of the statistics). To each system of groups, there is a system of statistics (the maximal invariants). The difference between the approaches is of a nonmathematical nature, and due to the fact that in certain cases one can describe groups in simple terms having an immediate intuitive appeal (row-column exchangeability), whereas the corresponding maximal invariants are strange objects. Conversely, it might be the statistics that

are immediately understood (row-column sums), whereas the groups seem artificial and appear as a "coincidence". It seems an interesting task to try to *combine* the approaches systematically. For example, referring to section 8 of the paper, it seems reasonable to *conjecture* that if an array $X=(X_{ij})_{i,j\in\mathbf{N}}$ of 0−1 valued random variables is row-column exchangeable *and* the conditional distribution of $X_{[m,n]}$ given its row and column sums is uniform, as in the Rasch model, then the array has the same distribution as (X_{ij}^*) where X_{ij}^* are independent *given* $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ with

$$P\{X_{ij}^*=1\,|\,\boldsymbol{\xi},\boldsymbol{\eta}\}=\frac{\xi_i\,\eta_j}{1+\xi_i\,\eta_j}$$

and $\boldsymbol{\xi}=(\xi_i)_{i\in\mathbf{N}}$ are i.i.d. with distribution F and independent of $\boldsymbol{\eta}=(\eta_j)_{j\in\mathbf{N}}$ that are i.i.d. with distribution G. In many respects this is a more reasonable model than the one discussed by Aldous and the Rasch model itself.

S. Johansen points in several examples to the fact that parameters might have a meaning outside the repetitive structure under consideration, and that a strict application of the idea of "always using the extreme point model" may lead to a total removal of the parameter of interest from the problem. Still one can make confidence intervals etc. for the parameter in question. An important distinction to be made here has been done explicitly by Dawid (1982), who introduces the notion of *extrinsic* and *intrinsic* parameters. A parameter is extrinsic, if it has a meaning from a well defined context different from the experiment under investigation, and if this is not the case, it is intrinsic. A very clear example of an intrinsic parameter is the "difficulty" of a question and the "ability" of a person in the Rasch model. A precise definition of an intrinsic parameter is one which is the limit of the sufficient statistic in a particular repetitive structure, in other words the canonical parameter in an extreme point model. Now if, as in the regression example discussed by Johansen, an extrinsic parameter β happens *not* to be a function of the intrinsic parameter $(\hat{\beta}_\infty)$, we should get a strong suspicion that *the design* $(\Sigma t_n^2<\infty)$ *is inadequate*. Also one should be extremely careful with confidence intervals since these typically refer to an (in this circumstance) unjustified frequency interpretation of certain probabilities.

O. Barndorff-Nielsen points to a possible connection between the notion of extremality and infinite Fisher information and asks for the connection to work of Basawa, Feigin, Heyde and Scott on non-ergodic exponential models. There is some connection, although this is not as clear cut as the examples might suggest. Let me first describe the relations in verbal terms and then through some examples give a more precise statement. The notion of extremality is a *strict* version of ergodicity. Infinite Fisher information is a *weak* concept and could be termed *first-order ergodicity*. The ergodicity considered by Basawa et al. is similarly a property of *weak* type and could be termed *second-order ergodicity*. Under suitable regularity conditions (ensuring that the information is well defined) extremality will imply ergodicity of first and second order, whereas the converse will only be true in special cases. Let us recall that Basawa et al. terms a model *non-ergodic* if the observed information

$$j_n(\theta,X_1,\ldots,X_n)=-D^2\log L_n(\theta;X_1,\ldots,X_n)$$

properly normalized, converges to a non-degenerate random variable. The Neyman factorization theorem ensures that the observed information is a function of the sufficient statistic and thus that this limiting random variable is measurable w.r.t. the tail σ-algebra of the sufficient statistics. Thus all extreme point models are second-order ergodic, provided that the observed information is well defined. It is, however, rarely so that this

limiting random variable *generates* the tail σ-algebra. In e.g. exponential families, the observed information about the canonical parameter does not depend on the observation at all. In general the observed information depends heavily on the parametrization, and it is not clear to me whether this is also the case for second-order ergodicity. Basawa et al. have formulated a "tail conditionality principle" involving a conditioning on the tail σ-algebra generated by the observed information, treating the limiting random variable as a parameter. This is obviously now related to the idea of investigating the extreme point model, although the latter involves a much stronger conditioning.

And now to the examples: Suppose $(X_n)_{n\in\mathbf{N}}$ are independent Poisson variables with

$$EX_1 = e^{\theta}, \quad EX_n = e^{2\theta}, \quad n \geq 2; \quad \theta \in \mathbf{R}.$$

The statistic

$$T_n = X_1 + 2\sum_{i=1}^{n} X_i$$

is sufficient and the observed and expected Fisher information are equal and equal to

$$i_n(\theta) = j_n(\theta, x_1, \ldots, x_n) = 4(n-1)e^{2\theta} + e^{\theta}$$

The model is clearly both first and second-order ergodic but the event

$$A = \{T_n \text{ is even infinitely often}\}$$

$$\underset{\text{a.s.}}{=} \{X_1 \text{ is even}\}$$

is tail-measureable with $P_\theta(A) \notin \{0, 1\}$, and the model is *not* an extreme point model. The corresponding extreme point model is obtained by introducing an extra parameter to describe the events A and A^c and conditioning on this.

If we reparametrize by letting $\theta = \log\lambda$, we get

$$j_n(\lambda, x_1, \ldots, x_n) = 2(n-1) + t_n\lambda^{-2}$$

and the tail σ-algebra generated by the observed information happens to be identical to that generated by T_n. It now depends on what is meant by "proper normalization" whether or not the model is second-order ergodic in this parametrization.

A less pathological example is a simple Galton-Watson process:

$$X_{n+1} = X_n + \sum_{i=1}^{X_n} Y_{in}, \quad X_0 = 1$$

where (Y_{in}) are i.i.d. with a geometric distribution:

$$P_\theta\{Y_{in} = y\} = (1-\theta)\theta^y, \quad \theta > 0, \; y \in 0, 1, 2, \ldots$$

The likelihood function is proportional to

$$L_n(\theta; x_1, \ldots, x_n) = (1-\theta)^{x_0 + \ldots + x_{n-1}}\theta^{x_n - x_0} = (1-\theta)^{z_n}\theta^{x_n - 1}$$

where $z_n = 1 + x_1 + \ldots + x_{n-1}$. The observed and expected Fisher information are

$$j_n(\theta; x_1, \ldots, x_n) = z_n(1-\theta)^{-2} + (x_n - 1)\theta^{-2}$$

$$i_n(\theta) = ((1-\theta)^{-n} - 1)\theta^{-2}(1-\theta)^{-1}$$

The model is first-order ergodic, but not ergodic in the sense of Basawa et al. see e.g. Basawa & Scott (1983), and therefore not an extreme point model. In fact

$$X_n(1-\theta)^n \xrightarrow{\text{a.s.}} W(\theta)$$

where $W(\theta)$ has an exponential distribution. Conditioning on $W=w$, X_n-X_{n-1} are independent and Poisson distributed with geometrically increasing expectations, and the model is therefore closely related to the Poisson example considered in section 5 of the paper.

Finally Barndorff-Nielsen raises a question concerning the extreme point model corresponding to an explosive autoregressive process. The question is certainly interesting, but it would demand a publication on its own and a certain amount of hard work to give a full answer.

As a partial answer, let me say the following. The corresponding minimal totally sufficient statistic is

$$T_n = \left(\sum_{i=1}^{n-1} X_i^2, \sum_{i=1}^{n} X_i X_{i-1}, X_n \right).$$

The *maximal* family will contain measures given as

$$X_{n+1} = \beta X_n + \gamma \beta^{-n} + \varepsilon_n,$$

where ε_n are i.i.d. $N(0, \tau^2)$, $\tau^2>0$ and γ and β are real-valued parameters. Thus an extreme point consideration seems in the first place to give rize to an *extension* of the model. In the case $|\beta|<1$ these measures are probably extreme, whereas this is *not* the case when $|\beta|>1$, since then ΣX_i^2 can be normalized to converge to a non-degenerate random variable. I hope to be able to answer this question more completely in the future.

P. Diaconis asks for the status of the non-uniform conditional distributions. I can hardly give a conversation between Martin-Löf and myself at this place, but I think the following is a correct evaluation. It was an important and interesting point made by Martin-Löf that surprisingly many interesting statistical models *can* be generated by uniform distributions. It requires a slightly more general notion of uniformity, cf. Martin-Löf (1975), than that described in Martin-Löf's earlier work (1970, 1974). It also requires some ingenuity, see how the Poisson model is derived in Martin-Löf (1974). I am not too convinced that non-uniform distributions are really *needed* as suggested by Diaconis. I introduced the non-uniform distributions as a technical convenience to make the theory more flexible and to be able to treat more examples without having to be as ingenious as was otherwise necessary.

References

Basawa, I. V. & Scott, D. J. (1983). *Asymptotic optimal inference for non-ergodic models*. Lecture Notes in Statistics, 17. Springer, Heidelberg.

Dawid, A. P. (1982). Probability, symmetry and frequency. Research report. University College, London.

Martin-Löf, P. (1970). Statistiska modeller. Notes by Rolf Sundberg. Mimeographed (Swedish).

Martin-Löf, P. (1974). Repetitive structures. In *Proceedings of Conference on Foundational Questions in Statistical Inference*. Aarhus 1974. Memoirs 1 (ed. O. Barndorff-Nielsen, P. Blæsild and G. Schou).

Martin-Löf, P. (1975). Reply to Sverdrup's polemic article "Tests without power". *Scand. J. Statist.* **3**, 161–165.