The Markov Moment Problem and de Finetti's Theorem: Part I 15 June 2003

by Persi Diaconis and David Freedman

Dedicated to the memory of Sergei Kerov

Abstract

The Markov moment problem is to characterize sequences $s_0, s_1, s_2, \ldots$ admitting the representation $s_n = \int_0^1 x^n f(x)\,dx$, where $f(x)$ is a probability density on $[0, 1]$ and $0 \le f(x) \le c$ for almost all x. There are well-known characterizations through complex systems of non-linear inequalities on $\{s_n\}$. Necessary and sufficient linear conditions are the following: $s_0 = 1$, and

$$0 \le (-1)^{n-j} \binom{n}{j} \Delta^{n-j} s_j \le c/(n+1)$$

for all $n = 0, 1, \ldots$ and $j = 0, 1, \ldots, n$. Here, Δ is the forward difference operator. This result is due to Hausdorff. We give a new proof with some ancillary results, for example, characterizing monotone densities. Then we make the connection to de Finetti's theorem, with characterizations of the mixing measure.

Introduction

We begin by reviewing the Hausdorff moment problem. Then we take up the Markov moment problem, with a solution due to Hausdorff (1923). Although this work was discussed in an earlier generation of texts (Shohat and Tamarkin, 1943, pp. 98–101; Widder, 1946, pp. 109–12; Hardy, 1949, pp. 272–3), it seems less well known today than the one due to the Russian school. Next, we sketch some generalizations and the connection to de Finetti's theorem. We close with some historical notes, including a brief statement of the Russian work. We believe that our Theorem 4 is new, along with the local theorems, the applications to Bayesian statistics (Theorems 8 and 9), and the characterization of measures with monotone densities (Theorem 10). Many of the results in this paper can be seen as answers to one facet or another of the following question: what can you learn about a measure from the moments, and how is it to be done?

The Hausdorff moment problem

Let $s_0, s_1, s_2, \ldots$ be a sequence of real numbers. When is there a probability measure μ on the unit interval such that s_n is the nth moment of μ? In other words, we seek the necessary and sufficient conditions on $\{s_n\}$ for there to exist a probability μ with

$$s_n = \int_0^1 x^n \,\mu(dx) \quad \text{for} \quad n = 0, 1, \ldots.$$

This is the Hausdorff moment problem.

To state Hausdorff's solution, let $\Delta t_n = t_{n+1} - t_n$ be the forward difference operator. Define an auxiliary sequence as

$$s_{n,j} = (-1)^{n-j} \binom{n}{j} \Delta^{n-j} s_j \tag{1}$$

for $n = 0, 1, \ldots$ and $j = 0, 1, \ldots, n$. By convention, $\Delta^0 s_j = s_j$. Thus,

$$
\begin{aligned}
s_{j,j} &= s_j, \\
s_{j+1,j} &= (j+1)(s_j - s_{j+1}), \\
s_{j+2,j} &= \tfrac{1}{2}(j+1)(j+2)(s_{j+2} - 2s_{j+1} + s_j),
\end{aligned}
$$

and so forth. The reason for introducing the binomial coefficients will be discussed later.

Theorem 1. *Given a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1)*. There exists a probability measure μ on* $[0, 1]$ *such that $\{s_n\}$ is the moment sequence of μ if and only if $s_0 = 1$, and $0 \le s_{n,j}$ for all n and j. Then μ is unique.*

This theorem is due to Hausdorff (1921), but Feller (1971, pp. 224–28) may be more accessible; the proof will not be repeated here. The "Hausdorff moment condition" is that $0 \le s_{n,j}$ for all n and j.

The Markov moment problem

The "Markov moment problem" is to characterize moments of probabilities that have uniformly bounded densities, which constrains μ in Theorem 1 to have the form $\mu(dx) = f(x)\,dx$, where $f \le c$ a.e. Of course, $f \ge 0$ a.e. and $\int_0^1 f\,dx = 1$, so $c \ge 1$. Hausdorff's solution is presented as Theorem 2.

Theorem 2. *Given a positive real number c, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1)*. There exists a probability measure μ on* $[0, 1]$ *such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *μ is absolutely continuous, and*
(iii) *$d\mu/dx$ is almost everywhere bounded above by c,*

if and only if $s_0 = 1$, and $0 \le s_{n,j} \le c/(n+1)$ for all n and j. Then μ is unique.

Our proof will use the following lemma.

Lemma 1. *Suppose $\{s_n\}$ is the moment sequence of the probability μ on* $[0, 1]$*; define the auxiliary sequence by* (1)*. Then*

(a) $s_{n,j} = \binom{n}{j} \int_0^1 x^j (1-x)^{n-j}\,\mu(dx)$.
(b) *If μ is Lebesgue measure, then $s_n = 1/(n+1)$.*
(c) *If μ is Lebesgue measure, then $s_{n,j} = s_{n,n} = s_n = 1/(n+1)$.*

Proof. Claim (a). Induction on $n = j, j+1, \ldots$.
Claim (b). Integration.
Claim (c). This just depends on the beta integral (Feller, 1971, p. 47):

$$
\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \text{for positive real } \alpha, \beta. \tag{2}
$$

Remarks. (i) Property (b) characterizes Lebesgue measure, in view of the uniqueness part of Theorem 1. Likewise, $s_{n,j} = s_{n,n}$ for all n and $j = 0, \ldots, n$ is a characterization, as in (c). Indeed,

$$\sum_{j=0}^{n} \binom{n}{j} x^j (1-x)^{n-j} = 1$$

for all x in the unit interval—after all, $[x + (1-x)]^n = 1$. Lemma 1a implies

$$\sum_{j=0}^{n} s_{n,j} = 1.$$

If the $s_{n,j}$ are equal for all $j = 0, 1, \ldots, n$, each must be $1/(n+1)$, so $s_n = s_{n,n} = 1/(n+1)$ for all $n = 0, 1, \ldots$. In essence, this characterization of the uniform distribution on $[0, 1]$ is due to Bayes (1764): see Stigler (1986, pp. 128–9).

(ii) Without the binomial coefficients in (1), the upper bound on $s_{n,j}$ in Theorem 2 would be more cumbersome to state. A deeper justification may be given by formulas (1.8) and (3.7) in Feller (1971, pp. 221, 225).

(iii) The condition $s_0 = 1$ may be dropped in Theorems 1 and 2; then μ is a finite positive measure, of total mass s_0. Indeed, $\sum_{j=0}^{n} s_{n,j} = s_0$ for any sequence $\{s_n\}$; this can be proved directly, or see (1.9) in Feller (1971, p. 221).

Proof of Theorem 2. Suppose conditions (i), (ii), and (iii) hold true. The conditions on s follow from Lemma 1. Conversely, suppose the conditions on s hold true. Theorem 1 shows the existence (and uniqueness) of a probability measure μ whose moment sequence is $\{s_n\}$. What remains to be seen is that μ is absolutely continuous, having a density bounded by c. If g is a non-negative continuous function on $[0, 1]$, its nth approximating Bernstein polynomial is by definition

$$B_{n,g}(x) = \sum_{j=0}^{n} g\left(\frac{j}{n}\right) \binom{n}{j} x^j (1-x)^{n-j}.$$

We claim that

$$\int_0^1 B_{n,g}(x)\, \mu(dx) \le c \int_0^1 B_{n,g}(x)\, dx.$$

Indeed, the left side is $\sum_j g(j/n) s_{n,j}$ by Lemma 1a, and the right side is $\sum_j g(j/n)[c/(n+1)]$ by Lemma 1c; finally, use the condition that $s_{n,j} \le c/(n+1)$.

Of course, $B_{n,g}$ converges to g uniformly as $n \to \infty$: see Feller (1971, pp. 222–4), or Lorentz (1966) for a more detailed discussion. So, for all non-negative continuous g,

$$\int_0^1 g(x)\, \mu(dx) \le c \int_0^1 g(x)\, dx. \tag{3}$$

Let $\mathcal{G}$ be the set of Borel measurable functions g on $[0, 1]$ with $0 \le g \le 1$. Let $\mathcal{G}_1$ consist of the $g \in \mathcal{G}$ for which (3) holds. Then $\mathcal{G}_1$ contains all the continuous functions in $\mathcal{G}$ and is closed under pointwise limits, so $G_1 = G$. Put $g = 1_A$, the indicator function of a Borel set A, to conclude that $\mu(A) \leq c\lambda(A)$, where λ is Lebesgue measure. Now μ is absolutely continuous; denote the Radon-Nikodym derivative $d\mu/dx$ by f. Let $A = \{x : 0 \leq x \leq 1 \,\&\, f(x) > c\}$. If $\lambda(A) > 0$, then

$$\mu(A) = \int_A f(x)\,dx > c\lambda(A).$$

But we have already seen that $\mu(A) \leq c\lambda(A)$. This contradiction shows that $\lambda(A) = 0$, proving Theorem 2.

Example 1. Let $f(x) = 1/(2\sqrt{x})$ on $(0, 1]$. This density is unbounded, but its nth moment is $s_n = 1/(2n+1) \leq 1/(n+1)$. Thus, the simple condition $s_n \leq c/(n+1)$ is not sufficient to make the density bounded: auxiliary conditions are needed. For our f, $(n+1)s_{n,j}$ is unbounded. Indeed, $s_{n,j}$ can be computed explicitly, using Lemma 1a and the formula for the beta integral (2):

$$\begin{aligned}
s_{n,j} &= \frac{1}{2}\binom{n}{j}\int_0^1 x^j(1-x)^{n-j}\,\frac{1}{\sqrt{x}}\,dx \\
&= \frac{1}{2}\binom{n}{j}\frac{\Gamma(j+\frac{1}{2})\Gamma(n-j+1)}{\Gamma(n+\frac{3}{2})} \\
&= \frac{1}{2}\,\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\,\frac{\Gamma(n+1)}{\Gamma(n+\frac{3}{2})}.
\end{aligned}$$

By Stirling's formula, $\log\Gamma(x) = (x-\frac{1}{2})\log(x+k) - x + O(1)$ as x gets large, for any constant k. Hence

$$\log(n+1) + \log\Gamma(n+1) - \log\Gamma(n+\tfrac{3}{2}) = \tfrac{1}{2}\log(n+1) + O(1).$$

So

$$\lim_{n\to\infty}(n+1)s_{n,j} = \infty$$

for any fixed j. The boundedness condition of Theorem 2 is not satisfied.

Example 2. The moments of the "Cantor measure" may be of interest in connection with Theorem 2. The Cantor measure is the distribution of $2\sum_{j=1}^{\infty} X_j/3^j$, the X_j being independent and identically distributed, $X_j = 0$ with probability 1/2 and $X_j = 1$ with probability 1/2. This measure is uniform on the Cantor set, and is therefore purely singular. For $n \geq 2$, the nth moment is

$$s_n > \frac{1}{2}\Big(1-\frac{1}{n}\Big)^n\frac{1}{n^{\log_3 2}}.$$

Indeed, the Cantor measure assigns mass 2^{-m} to the interval $[1-3^{-m}, 1]$, so $s_n \geq 2^{-m}(1-3^{-m})^n$ for any positive integer m. Now choose m with $\log_3 n \leq m < 1 + \log_3 n$. In particular,

$$\lim\,(n+1)s_n = \infty.$$

See Grabner and Prodinger (1996) for more detailed estimates.

L_p densities

Theorem 2 characterizes the moment sequences of probabilities with L_∞ densities on $[0, 1]$. The next result (also due to Hausdorff) characterizes L_p densities for $p > 1$. To state the theorem, define

$$c_n = \left\{ \frac{1}{n+1} \sum_{j=0}^{n} \left[(n+1)s_{n,j} \right]^p \right\}^{1/p}. \tag{4}$$

Theorem 3. *Given real numbers $p > 1$ and $0 < c < \infty$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1)*, and c_n by* (4)*. There exists a probability measure μ on* $[0, 1]$ *such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *μ is absolutely continuous, and*
(iii) *$d\mu/dx$ is in L_p with p-norm at most c,*

if and only if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $c_n \leq c$.

So far, absolute continuity is defined relative to Lebesgue measure, but Lebesgue measure can be replaced by any other probability ν on $[0, 1]$. To avoid trivial complications, suppose ν assigns positive mass to the open unit interval $(0, 1)$. Let t_n be the moment sequence of ν, and $t_{n,j}$ the corresponding auxiliary sequence defined by (1) with t_n in place of s_n. Lemma 1a confirms that $t_{n,j} > 0$. Replace the definition (4) by

$$c_n = \left\{ \sum_{j=0}^{n} t_{n,j} \left(\frac{s_{n,j}}{t_{n,j}} \right)^p \right\}^{1/p}. \tag{5}$$

Theorem 4. *Let ν be a probability on* $[0, 1]$*, assigning positive mass to* $(0, 1)$*. Given real numbers $p > 1$ and $0 < c < \infty$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1)*, and c_n by* (5) *rather than* (4)*. There exists a probability measure μ on* $[0, 1]$ *such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *$\mu << \nu$, and*
(iii) *$d\mu/d\nu$ is in L_p with p-norm at most c,*

if and only if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $c_n \leq c$.

In (iii), the p-norm of $d\mu/d\nu$ is relative to ν, i.e., $\left(\int (d\mu/d\nu)^p \, d\nu \right)^{1/p}$. Theorem 3 is a special case of Theorem 4; our proof of the latter depends on the connection with de Finetti's theorem, which is explained next. Let $X_1, X_2, \ldots$ be random variables taking only the values 0 and 1. The sequence is "exchangeable" if the joint distribution is invariant under finite permutations, for example,

$$P\{X_1 = 1, X_2 = 0, X_3 = 1\} = P\{X_1 = 0, X_2 = 1, X_3 = 1\}.$$

Either the random variables can be permuted, or the values.

Theorem 5. *Let* $e_1, e_2, \ldots$ *be* 0 *or* 1. *The* 0–1 *valued random variables* $X_1, X_2, \ldots$ *are exchangeable if and only if there is a probability measure* μ *on* $[0, 1]$ *such that*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i}\,\mu(d\theta), \tag{6}$$

for all n *and* e_i. *Then* μ *is unique.*

This theorem is due to de Finetti (1931,1937); for a review, see Hewitt and Savage (1955). The "if" part is straightforward. Necessity is more subtle because μ must be constructed, but Hausdorff's theorem can be used (Feller, 1971, pp. 228–9). The proof of Theorem 5 will not be detailed here. Before applying the theorem, we explain how the auxiliary sequence (1) connects to (6). Suppose the X_i are exchangeable, and $S_n = X_1 + X_2 + \cdots + X_n$. Let s_n be the moment sequence of μ in Theorem 5, and define $s_{n,j}$ by (1). Fix n and j with $0 \leq j \leq n$. Fix some particular finite sequence $e_1, e_2, \ldots, e_n$ of 0s and 1s whose sum is j. Then

$$P\{S_n = j\} = \binom{n}{j} P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \binom{n}{j}\int_0^1 x^j(1-x)^{n-j}\,\mu(dx).$$

By Lemma 1a,

$$P\{S_n = j\} = s_{n,j}. \tag{7}$$

The notation is flawed, in that s_n is a moment of μ rather than a value of S_n.

Proof of Theorem 4. If $s_0 = 1$ and $0 \leq s_{n,j}$ for all n and j, there is a probability μ on $[0, 1]$ whose moment sequence is $\{s_n\}$. For the rest, the "if" and "only if" assertions can be proved together: the issue is to determine from the moments whether μ is absolutely continuous with respect to ν, and $d\mu/d\nu \in L_p(\nu)$. We begin by constructing an exchangeable sequence $X_1, X_2, \ldots$ of 0–1 valued random variables that satisfy (6): write P_μ for P. Define P_ν in the analogous way. Let $S_n = X_1 + \cdots + X_n$. Let $\mathcal{F}_n$ be the field generated by $X_1, \ldots, X_n$, and $\mathcal{F}$ the σ-field generated by all the X's, so $\mathcal{F}_n \uparrow \mathcal{F}$. Let H_n be the random variable whose value is $P_\mu\{S_n = j\}/P_\nu\{S_n = j\}$ on the set $\{S_n = j\}$. Then H_n is the Radon-Nikodym derivative of P_μ with respect to P_ν, both restricted to $\mathcal{F}_n$. Thus, H_n is a martingale relative to P_ν, and c_n is the p-norm of H_n relative to P_ν. By Jensen's inequality,

$$c_n \text{ in (5) are non-decreasing.} \tag{8}$$

From this point on, we use the standard martingale theory for differentiating measures. The key martingale fact is Theorem 4.1 on pp. 319–20 in Doob (1953); the application to differentiating measures is summarized in Freedman (1983, pp. 345–6): for more discussion, see Hewitt and Stromberg (1969, pp. 369–75). We conclude that

$$H_n \to H_\infty \text{ a.e. } [P_\mu + P_\nu], \tag{9}$$

with

$$H_\infty = dP_\mu/dP_\nu \tag{10}$$

for the full σ-field $\mathcal{F}$: the limit is infinite on the part of the space where P_μ is singular with respect to P_ν. Moreover,

$$c_n = [\mathrm{E}_\nu(H_n^p)]^{1/p} \uparrow [\mathrm{E}_\nu(H_\infty^p)]^{1/p}, \tag{11}$$

where E_ν denotes expectation relative to P_ν. In particular, if $\sup_n c_n \leq c < \infty$, then $H_\infty \in L_p(\nu)$ and $\|H_\infty\|_p \leq c$. On the other hand, if c_n is unbounded, then $H_\infty \notin L_p(\nu)$. The next (and last) step in the proof is perhaps worth isolating as a proposition, which writes H for H_∞.

Proposition 1. *Let $L = \lim_n S_n/n$, which exists a.e. relative to $P_\mu + P_\nu$. Let $h = d\mu/d\nu$, and $H = dP_\mu/dP_\nu$, with the understanding that $h = \infty$ on the part of the unit interval where μ is singular with respect to ν; similarly for H on its domain. Then*

(i) $P_\mu L^{-1} = \mu$.
(ii) $P_\nu L^{-1} = \nu$.
(iii) $H = h(L)$ *a.e. relative to* $P_\mu + P_\nu$.

Proof. Only claim (iii) is argued. To begin with, we impose the side condition that $\mu << \nu$. Let P_θ be the distribution when a θ-coin is tossed, so

$$P_\theta\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i},$$

the e_i being 0 or 1. Furthermore,

$$P_\mu = \int_0^1 P_\theta\, \mu(d\theta), \quad P_\nu = \int_0^1 P_\theta\, \nu(d\theta).$$

For any $A \in \mathcal{F}$,

$$\begin{aligned}
\int_A h(L)\, dP_\nu &= \int_0^1 \left(\int_A h(L)\, dP_\theta\right) \nu(d\theta) \\
&= \int_0^1 \left(\int_A h(\theta)\, dP_\theta\right) \nu(d\theta) \\
&= \int_0^1 P_\theta(A)\, h(\theta)\, \nu(d\theta) \\
&= \int_0^1 P_\theta(A)\, \mu(d\theta) = P_\mu(A).
\end{aligned}$$

The second equality holds because $P_\theta(L = \theta) = 1$ by the strong law of large numbers. The fourth equality holds by the side condition $\mu << \nu$, with $h = d\mu/d\nu$. Thus, $h(L)$ is a version of dP_μ/dP_ν. This proves claim (iii) under the side condition, but the general case follows: notice that H and h depend affinely on μ, then replace μ by $(\mu + \nu)/2$. This completes the argument, and the proof of Theorem 4.

Remarks. (i) Theorem 4 holds as stated when $p = \infty$, if we redefine c_n in (5) as

$$c_n = \max_{j=0,\ldots,n} s_{n,j}/t_{n,j}.$$

This is Corollary 3.1 in Knill (1997).

(ii) The case $p = 1$ is more problematic. We can show that $\mu << \nu$ iff the martingale H_n is uniformly P_ν-integrable, but this is little more than a restatement of the definition of absolute continuity, and uniform integrability may not be any easier to check in applications than absolute continuity.

(iii) The conditions we have considered in Theorems 2–4 are of the form $f_n(s_0, s_1, \ldots, s_n) \leq k_n$, where f_n is a specified continuous function on R^{n+1}, k_n is a constant, and $s_0, s_1, \ldots$ a sequence that may—or may not—be the moment sequence of a probability that is being characterized in some way. No condition of this form can describe the moment sequences of absolute continuous probabilities, because the set of absolutely continuous probabilities is not weak-star closed.

(iv) Theorems 2–4 can be extended in a straightforward way from the unit interval to the unit cube in R^d.

(v) Hausdorff was working with finite signed measures. Theorems 2–5 can be extended to cover that case, although the interpretation of de Finetti's theorem for signed priors remains a little mysterious, at least for elderly statisticians; also see Feynman (1987). For multi-dimensional signed measures, see Knill (1997); for an application to de Finetti's theorem, see Jaynes (1986).

Local theorems

Theorem 2 can be modified if we desire only that μ should be absolutely continuous on the interval $[a, b]$, with $0 \leq a < b \leq 1$, and $d\mu/dx \leq c$ on $[a, b]$; off this interval, μ has no special features. We begin with the sufficiency part of Theorem 2, only sketching the development.

Theorem 6. *Given real numbers a, b, c with $0 \leq a < b \leq 1$ and $c > 0$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation (1). There exists a probability measure μ on $[0, 1]$ such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *μ is absolutely continuous on the interval $[a, b]$, and*
(iii) *$d\mu/dx$ is almost everywhere bounded above by c on the interval $[a, b]$,*

if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $s_{n,j} \leq c/(n+1)$ for all n and j with $a \leq j/n \leq b$. Then μ is unique.

Here is a generalization of the sufficiency part of Theorem 4.

Theorem 7. *Given a positive real number c, and a, b with $0 \leq a < b \leq 1$, and a probability ν on $[0, 1]$ that assigns positive mass to (a, b), and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequences $s_{n,j}$ and $t_{n,j}$ by applying equation (1) to μ and ν respectively. Define c_n as follows:*

$$c_n = \Big\{ \sum_{an \leq j \leq bn} t_{n,j} \Big(\frac{s_{n,j}}{t_{n,j}} \Big)^p \Big\}^{1/p}. \tag{12}$$

There exists a probability measure μ on $[0, 1]$ such that

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *μ is absolutely continuous with respect to ν on the interval $[a, b]$, and*

(iii) $d\mu/dx \in L_p(\nu)$ *on the interval* $[a, b]$*, with norm at most* c*,*

if $s_0 = 1$*, and* $0 \le s_{n,j}$ *for all* n *and* j*, and* $c_n \le c$*. Then* μ *is unique.*

Proofs are straightforward, using Hausdorff's theorem to get μ and techniques described earlier in the paper to characterize $d\mu/dx$. For example, take Theorem 6. We can prove (3) for all continuous functions on the interval $[a, b]$, then for all Borel functions g on $[a, b]$ with $0 \le g \le 1$. The balance of the argument is unchanged. The conditions, however, are not necessary, as will be shown by example.

Example 3. To see why the upper bound in Theorem 6 cannot be a necessary condition, take $a = 0$ and $b = 1/2$. Let μ assign mass 1/2 to $[0, 1/2]$, with density bounded above by c; let μ assign the remaining mass 1/2 to $1/2 + h$. Choose n large and even, then $h > 0$ small. Consider $P_\mu\{S_n = n/2\}$. The part of μ on $[0, 1/2]$ contributes at most $c/(n+1)$ to $P_\mu\{S_n = n/2\}$. But—if $h = 0$—the other piece of $P_\mu\{S_n = n/2\}$ is of order $1/\sqrt{n}$. If $h > 0$ is small, this other piece can therefore be much larger than $c/(n+1)$.

For Theorem 6, the necessary and sufficient upper bound condition on $s_{n,j}$ would be $s_{n,j} \le c/(n+1) + \exp(-2\delta^2 n)$ for all δ with $0 < \delta < (b-a)/2$ and all n, j with $a + \delta \le j/n \le b - \delta$. See (3.5) in Diaconis and Freedman (1990). Example 3 indicates why the term $\exp(-2\delta^2 n)$ is needed, and the restriction to $a + \delta \le j/n \le b - \delta$. The characterization of L_p densities relative to Lebesgue measure is also relatively straightforward. For other base measures, we do not have clean results.

Applications to Bayesian statistics

Theorems on moment sequences can be translated in a straightforward way into theorems characterizing the mixing measure μ in Theorem 5. We give two examples. Recall that P_θ is the distribution when a θ-coin is tossed, so

$$P_\theta\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i},$$

the e_i being 0 or 1. Furthermore,

$$P_\mu = \int_0^1 P_\theta\, \mu(d\theta). \tag{13}$$

Theorem 8. *Let* X_i *be* 0–1 *valued random variables on the probability triple* $(\Omega, \mathcal{F}, P)$*. Let* c *be a positive real number. Then* $\{X_i\}$ *admits the representation*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i} f(\theta)\, d\theta$$

for all n *and* $e_i = 0$ or 1*, and* $0 \le f \le c$ *a.e., iff*

(i) *the* X_i *are exchangeable, and*

(ii) $P_\mu\{S_n = j\} \le c P_\lambda\{S_n = j\}$ *for all* $n = 0, 1, \ldots$ *and* $j = 0, 1, \ldots, n$*, where* λ *is Lebesgue measure on* $[0, 1]$*, and* $S_n = X_1 + \cdots + X_n$*.*

Then f is unique.

This is immediate from (7) and Theorem 2. The analog of Theorem 4 is as follows.

Theorem 9. *Let X_i be* 0–1*-valued random variables on the probability triple* $(\Omega, \mathcal{F}, P)$. *Let ν be a probability on* $[0, 1]$, *assigning positive mass to* $(0, 1)$. *Let $p > 1$ and $0 < c < \infty$. Then $\{X_i\}$ admits the representation*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i} f(\theta)\, \nu(d\theta)$$

for all n and $e_i = 0$ or 1, and $f \in L_p(\nu)$ has norm at most c, iff

(i) *the X_i are exchangeable, and*
(ii) *$c_n \le c$ for all $n = 0, 1, \ldots,$ where*

$$c_n = \left[\sum_{j=0}^{n} P_\nu\{S_n = j\} \left(\frac{P_\mu\{S_n = j\}}{P_\nu\{S_n = j\}} \right)^p \right]^{1/p} \tag{14}$$

and $S_n = X_1 + \cdots + X_n$.

Then f is unique.

Theorem 9 can be extended to the case $p = \infty$ by redefining c_n as follows:

$$c_n = \max_{j=0,\ldots,n} P_\mu\{S_n = j\}/P_\nu\{S_n = j\}.$$

There are yet more general theorems characterizing partially exchangeable processes with L_p densities, in the setting of Diaconis and Freedman (1984): we will explore such results in Part II of this paper. In the abstract setting, the proofs are more transparent (although the setting itself may seem a little strange).

Monotone densities

In some applications, it is desired to characterize monotone densities in terms of their moments; see, for instance, Diaconis and Kemperman (1996). Theorem 10 gives a result for densities that are non-decreasing. We will need the following lemma, which expresses a monotone function as a mixture of the extreme step functions.

Lemma 2. *Let F be a non-negative, right-continuous, non-decreasing function on* $[0, 1)$; *we allow $F(0) > 0$ and $F(1-) = \infty$. Let $f_\theta(x) = 0$ for $0 \le x < \theta$ and $f_\theta(x) = 1/(1-\theta)$ for $\theta \le x < 1$, so f_θ is a probability density for $0 \le \theta < 1$. Then*

$$F = \int_{[0,1)} f_\theta\, \nu(d\theta),$$

where the measure ν on $[0, 1)$ *is defined as follows: $\nu(d\theta) = (1-\theta)F(d\theta)$, with $F(d\theta)$ assigning mass $F(0)$ to 0. Finally, the total mass in ν is $\int_0^1 F(x)\, dx$.*

Proof. The calculation will seem trite, but it is easy to get lost if you start at the wrong place. Let $H_\theta = 0$ on $[0, \theta)$ and $H_\theta = 1$ on $[\theta, 1)$. Then

$$F(x) = \int_{[0,1)} H_\theta(x)\, F(d\theta) = \int_{[0,1)} f_\theta(x)\ (1-\theta) F(d\theta) = \int_{[0,1)} f_\theta(x)\, \nu(d\theta).$$

To evaluate the mass in ν, integrate over $x \in [0, 1)$. The proof is complete.

Theorem 10. *Given a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence $s_{n,j}$ by equation* (1). *There exists a probability measure μ on* $[0, 1]$ *such that*

(i) $\{s_n\}$ *is the moment sequence of μ, and*
(ii) μ *is absolutely continuous on* $[0, 1)$*, and*
(iii) $d\mu/dx$ *is non-decreasing on* $[0, 1)$*,*

if and only if $s_0 = 1$, and $0 \le s_{n,j}$ for all n and j, and $s_{n,j}$ is nondecreasing in j for all n. The probability μ has a possible atom at 1*, but $\mu\{1\} = 0$ iff $s_n \to 0$.*

Proof. Suppose μ satisfies conditions (i), (ii), and (iii). Then μ is a convex combination of point mass at 1, and an absolutely continuous probability on [0, 1] with a non-decreasing density. If $\mu\{1\} = 1$, it is clear that $s_{n,j}$ is non-decreasing with j. Suppose on the other hand that μ is absolutely continuous on [0, 1] and $d\mu/dx$ is non-decreasing. As in Lemma 2,

$$d\mu/dx = \int_{[0,1)} f_\theta\, \nu(d\theta).$$

(In this application, ν is a probability measure.)

Since $s_{n,j}$ is affine in μ by Lemma 1a, it suffices to consider the θ's one at a time, i.e., we can take ν to be point mass at θ. Let $0 \le j < n$. We claim that $s_{n,j} \le s_{n,j+1}$, that is,

$$\binom{n}{j} \int_0^1 x^j (1-x)^{n-j} f_\theta(x)\, dx \le \binom{n}{j+1} \int_0^1 x^{j+1}(1-x)^{n-j-1} f_\theta(x)\, dx, \tag{15}$$

which is to say,

$$(j+1) \int_\theta^1 x^j (1-x)^{n-j}\, dx \le (n-j) \int_\theta^1 x^{j+1}(1-x)^{n-j-1}\, dx. \tag{16}$$

Let $G(\theta)$ be the right hand side of (16) minus the left hand side, namely,

$$G(\theta) = \int_\theta^1 x^j (1-x)^{n-j-1} g(x)\, dx,$$

where

$$g(x) = (n-j)x - (j+1)(1-x).$$

Now

$$G'(\theta) = \theta^j (1-\theta)^{n-j-1} h(\theta),$$

where

$$h(\theta) = -g(\theta) = (j+1)(1-\theta) - (n-j)\theta.$$

Clearly, $h(\theta) > 0$ for $0 \le \theta < (j+1)/(n+1)$ and $h(\theta) < 0$ for $(j+1)/(n+1) < \theta \le 1$. Thus, G increases from 0 at 0—see Lemma 1c—to its maximum at $(j+1)/(n+1)$, and then decreases to 0 at 1. In short, $G > 0$ except at 0 and 1, where G vanishes. Thus, (16) holds for $0 \le \theta \le 1$, and (15) must hold for $0 \le \theta < 1$, completing the proof of the "only if" part of the theorem. The converse follows from Proposition 2 below, with $p_{n,j} = s_{n,j}$. The convergence of μ_n is discussed in the remarks following the proposition.

Proposition 2. *Let the probability* μ_n *on* $[0,1]$ *assign mass* $p_{n,j}$ *to* j/n *for* $j = 0, 1, \ldots, n$, *with* $0 \le p_{n,0} \le p_{n,1} \le \ldots \le p_{n,n}$ *and* $\sum_{j=0}^{n} p_{n,j} = 1$. *Suppose* $\mu_n \to \mu$ *weak-star. Let* F *be the distribution function of* μ. *Then* F *is convex on* $[0,1]$, *hence absolutely continuous on* $[0,1)$ *with nondecreasing density* F'. *There is a possible atom at* 1.

Proof. Take the convolution of μ_n with the uniform distribution on $[-\frac{1}{2n}, \frac{1}{2n}]$, in effect replacing the point masses with their histogram. The resulting measure has distribution function F_n which is convex—because F_n' is monotone—and still converges weak-star to F. Let D be the set of discontinuity points of F. Then $D \cup D/2 \cup D/3 \cup \cdots$ is countable. So, there are small positive h with $jh \in D$ for no integer j: after all, $jh \in D$ iff $h \in D/j$. Next, F_n converges pointwise to F on the h-skeleton $h, 2h, \ldots$, because F is continuous there. Since F_n is convex on this skeleton, so is F. But h can be arbitrarily small. Therefore, F is convex on $(0,1)$. In particular, F is continuous on $(0,1)$, even absolutely continuous, and its density F' is increasing. Suppose by way of contradiction that 0 were an atom with mass $\delta > 0$. For any $x, h > 0$ with $0 < x < 1-h$, we would have $\mu[x, x+h] = \lim_n F_n(x+h) - F_n(x) \ge \limsup_n F_n(h) - F_n(0) \ge \delta$, which is impossible; the first inequality holds because F_n' is monotone; the second, because $p_{0,n} \le 1/(n+1)$ so $F_n(0) - F_n(-h) \to 0$ while $\mu\{0\} = \delta$. Thus, F is continuous even at 0, with $F(0) = 0$. This finishes the proof of Proposition 2, and hence of Theorem 10.

Remarks. (i) Decreasing densities can be characterized in a similar way, although the possible atom moves to 0, and can be excluded by requiring $s_{n,0} \to 0$.

(ii) The existence of the density in Theorem 10 follows from the monotonicity of $s_{n,j}$, but the density need not be bounded.

(iii) Why does μ_n converge? Hausdorff proved Theorem 1 by showing directly that μ_n converges weak-star to the desired μ: see Feller (1971, pp. 225–26). For us, it may seem more natural to prove the relevant law of large numbers. The convergence of μ_n would follow, along with Hausdorff's moment theorem, the convergence of the Bernstein polynomials, and de Finetti's theorem. In essence, that is the path followed by de Finetti (1937). Compactness arguments are also feasible.

(iv) Theorem 10 completes Bayes' observation that a uniform density corresponds to a uniform distribution for S_n: the uniform density is non-decreasing and non-increasing, so the resulting distribution of S_n has the same features. Of course, there are familiar arguments that are more direct: see Lemma 1 and the remarks that follow it.

(v) Suppose μ is absolutely continuous on $[0,1)$, and $d\mu/dx$ is non-decreasing on $[0,1)$. Unless $d\mu/dx$ is constant, $s_{n,j}$ will be strictly increasing with j. Indeed, the inequality in (16) is strict unless $\theta = 0$ or 1; the inequality in (15) is therefore strict unless $\theta = 0$, corresponding to a density that is constant. On the other hand, if μ has an atom at 1, then $s_{n,n-1} < s_{n,n}$.

Historical notes

Hausdorff

Hausdorff's work on the moment problem was motivated by summability theory (Hausdorff, 1921, 1923). In brief, let $S = \{s_{n,j} : n = 0, 1, \ldots, j = 0, \ldots, n\}$ be a triangular matrix of real numbers. The "S-limit" of a sequence $\{x_i\}$ is $\lim_n \sum_{j=0}^{n} s_{n,j} x_j$. A summability method S is "regular" if $\lim x_i = x_\infty$ implies that the S-limit is x_∞. Familiar examples include Cesàro's method, where $s_{n,j} = 1/(n+1)$, and Euler's E_p method with

$$s_{n,j} = \binom{n}{j} p^j (1-p)^{n-j}.$$

Hausdorff introduced a more general scheme, defining

$$s_{n,j} = \binom{n}{j} \int_0^1 p^j (1-p)^{n-j}\, \mu(dp) \tag{17}$$

where μ is a finite signed measure on $[0, 1]$. For instance, setting μ to Lebesgue measure gives us Cesàro's method: see Lemma 1b. If μ is point mass at p, we get E_p. Among many other things, Hausdorff showed that a summability method defined by (17) is regular iff $\mu\{0\} = 0$ and $\mu(0, 1] = 1$; this is more or less obvious from (7). However, μ need not be a probability measure: its negative part need not vanish. Methods defined by (17) are now called "Hausdorff methods." For additional discussion, see Widder (1946) or Hardy (1949).

Some notes on Hausdorff (1923) may be of interest. The auxiliary sequence, with the binomial coefficients, is introduced in equation (5) on p. 223; the positivity condition is (A) on the same page. The solution to the moment problem is Satz I on p. 226. The condition for an L_p density is (C) on p. 234, and the theorem is Satz III on p. 236. The condition for L_∞ is (D) on the same page, and the solution to the Markov moment problem is Satz IV on p. 237. The hitherto-unmentioned Satz II on p. 232 characterizes moment sequences of finite signed measures: his necessary and sufficient condition (B) is, in our notation, $\sup_n \sum_j |s_{n,j}| < \infty$.

The Russian School

Solutions to the Markov moment problem, and similar results for the half-line and the whole line, were among the great achievements of the Russian school. Perhaps the history begins with Chebychev, who gave a rigorous proof of the Central Limit Theorem using the method of moments, with connections to the theory of continued fractions, orthogonal polynomials, and numerical quadrature. His student Markov formulated the moment problem we have been discussing (along with many other contributions in other areas).

Let $\{s_n\}$ be a given sequence of real numbers, and c a given positive real. When is $s_n = \int_0^1 x^n f(x)\, dx$ for all n, with f a probability density bounded above by c? To answer this question, Markov expanded

$$\exp\left[\frac{1}{c}\left(\frac{s_0}{z} + \frac{s_1}{z^2} + \frac{s_2}{z^3} + \frac{s_3}{z^4} + \cdots\right)\right] \tag{18}$$

as a continued fraction, and showed that positivity of certain coefficients was a necessary condition. The condition turned out to be sufficient as well.

There were later developments by Ahiezer and Krein (1962), and Krein and Nudelman (1977). One theorem in Ahiezer and Krein (1962, p. 71) can be stated this way: $s_n = \int_0^1 x^n f(x)\,dx$ for all n, with $0 \le f \le c$ a.e., iff t_n satisfies Hausdorff's condition, where t_n is defined by a formal series expansion of (18) in powers of $1/z$:

$$\exp\left[\frac{1}{c}\left(\frac{s_0}{z}+\frac{s_1}{z^2}+\frac{s_2}{z^3}+\frac{s_3}{z^4}+\cdots\right)\right] = 1+\frac{t_1}{z}+\frac{t_2}{z^2}+\frac{t_3}{z^3}+\cdots \tag{19}$$

Of course, the t_n are polynomial functions of s_n/c. For example,

$$t_1 = \frac{s_0}{c}, \quad t_2 = \frac{s_1}{c}+\frac{1}{2}\frac{s_0^2}{c^2}, \quad t_3 = \frac{s_2}{c}+\frac{s_0 s_1}{c^2}+\frac{1}{6}\frac{s_0^3}{c^3}.$$

In general,

$$t_n = \frac{1}{n!}\sum_{\pi}\prod_{j=1}^{n}(j s_{j-1}/c)^{a_j(\pi)} \tag{20}$$

where π runs through the permutations of length n, and $a_j(\pi)$ is the number of cycles in π of length j. Here, $s_0 = \int f$: if $s_0 = 1$, then f is a probability density.

Sergei Kerov made several remarkable contributions to this theory. For instance, (19) sets up a one-to-one correspondence between the moments $\{s_n\}$ of a density bounded by c, and the moments $\{t_n\}$ of an auxiliary measure ν on $[0, 1]$. Given f, Kerov showed how to pick a random point from ν, by generating a nested sequence of random intervals

$$[0,1] \supset [X_1, Y_1] \supset [X_2, Y_2] \supset \cdots$$

that shrink to a point. Despite the complexity of (19), Kerov's algorithm is elegance itself. At stage $n+1$, pick a point U at random in $[X_n, Y_n]$. Then flip a coin that lands heads with probability $f(U)/c$, or tails with the remaining probability $1-[f(U)/c]$. If the coin lands heads, $X_{n+1} = U$ and $Y_{n+1} = Y_n$. But if the coin lands tails, $X_{n+1} = X_n$ and $Y_{n+1} = U$. Probabilities have to be bounded between 0 and 1: that is where the condition $0 \le f \le c$ comes in.

Kerov found striking connections between his algorithm and Young tableaux, as well as eigenvalues of random matrices, and the zeroes of orthogonal polynomials. Recently, expansions connected to the Markov moment problem—like (19) and (20)—have found applications in Bayesian non-parametric statistics: Cifarelli and Regazzini (1990), Diaconis and Kemperman (1996).

Acknowledgments

We thank Christian Berg for suggesting a beautiful alternative proof for the "if" part of Theorem 2. The s_n are moments of a probability on $[0, 1]$, call it μ, by Hausdorff's theorem. Let λ be Lebesgue measure on $[0, 1]$, and $\{t_n\}$ the moment sequence of $c\lambda - \mu$, with auxiliary sequence $\{t_{n,j}\}$ defined by the analog of (1). Then

$$t_{n,j} = \binom{n}{j}\int_0^1 x^j(1-x)^{n-j}\,d(c\lambda-\mu) = c/(n+1) - s_{n,j}$$

by Lemma 1, so $\{t_n\}$ satisfies the Hausdorff condition,

$$(-1)^{n-j}\binom{n}{j}\Delta^{n-j}t_j \geq 0,$$

although $t_0 = 1$ is unlikely. Hence, $c\lambda \geq \mu$, again by Hausdorff's theorem. We also thank Jon McAuliffe for a number of helpful comments, and a very careful anonymous referee.

References

Ahiezer, N. I. and Krein, M. (1962). *Some Questions in the Theory of Moments.* Amer. Math. Soc. Providence.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Cifarelli, D. and Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *Ann. Statist.* 18: 429–442.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nationale dei Lineii Ser. 6, Memorie, classe di Scienze, Fisiche, Matematiche e Naturali* 4: 251–99.

de Finetti, B. (1937). La prévision: ses lois logiques ses sources subjectives. *Ann. Inst H. Poincaré* 7: 1–68. Translated in H. Kyburg, H. Smokler, eds. *Studies in Subjective Probability*. Wiley, New York (1964).

Diaconis, P. and Freedman, D. (1984). Partial exchangeability and sufficiency. In J. K. Ghosh, J. Roy, eds. *Proc. Indian Statist. Assoc. Golden Jubilee: Applications And New Directions.* Indian Statist. Assoc., Calcutta, 205–236.

Diaconis, P. and Freedman, D. (1990), On the uniform consistency of Bayes estimates for multinomial probabilities. *Ann. Statist.* 18: 1317–27.

Diaconis, P. and Kemperman, J. (1996). Some new tools for Dirichlet priors. In J. Bernardo et al, eds. *Bayesian Statistics* 5, Oxford University Press, 97–106.

Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York. Reprinted 1990, Wiley Classics Library.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications.* Vol. II, Second edition, Wiley, New York.

Feynman, R. (1987). Negative probabilities. In B. J. Hiley, F. D. Peat, eds. *Quantum Implications.* Routledge & Kegan-Paul.

Freedman, D. A. (1983). *Markov Chains.* Springer, New York. First published in 1971 by Holden Day, San Francisco.

Grabner, P. J. and Prodinger, H. (1996). Asymptotic analysis of the moments of the Cantor distribution. *Statistics & Probability Letters* 26: 243–8.

Hardy, G. H. (1949). *Divergent Series.* Oxford. Reprinted 1991 by Chelsea, New York.

Hausdorff, F. (1921). Summationsmethoden und Momentfolgen I, II. *Math. Zeitschrift* 9: 74–109, 280–299.

Hausdorff, F. (1923). Momentprobleme für ein endliches intervall. *Math. Zeitschrift* 16: 220–48.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *TAMS* 80: 470–501.

Hewitt, E. and Stromberg, K. (1969). *Real and Abstract Analysis.* Springer, New York.

Jaynes, E. I. (1986). Some applications and extensions of the de Finetti representation theorem. Pp. 31–42 in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Eds. P. K. Goel and A. Zellner, North-Holland, Amsterdam.

Kerov, S. (1993). Transition probabilities for continual Young diagrams and the Markov moment problem. *Functional Analysis and its Applications* 27: 104–117.

Kerov, S. (1998). Interlacing Measures. *Amer. Math. Soc. Transl.* Ser. 2 Vol. 181, pp. 35–83. *Advances in the Mathematical Sciences,* 35, G. I. Olshanski, ed.

Knill, O. (1997). On Hausdorff's moment problem in higher dimensions. Technical report, Department of Mathematics, Harvard University, http://abel.math.harvard.edu/~knill/preprints/

Krein, M. G. and Nudelman A. (1977). *The Markov Moment Problem and Extremal Problems.* Amer. Math. Soc. Providence.

Lorentz, C. (1966). *Bernstein Polynomials.* University of Toronto Press, Toronto.

Shohat, J. and Tamarkin, J. (1943). *The Problem of Moments.* Amer. Math. Soc. Providence.

Stigler, S. M. (1986). *The History of Statistics.* Harvard University Press.

Widder, D. V. (1946). *The Laplace Transform.* Princeton University Press. First printing, 1941.

Affiliations

Persi Diaconis is Professor of Mathematics and of Statistics, Stanford University, Stanford, CA 94305.

David A. Freedman is Professor of Statistics and of Mathematics, University of California, Berkeley, CA 94720-3860, freedman@stat.berkeley.edu

June 2003
Technical Report No. 631, Part I
Statistics Department
University of California
Berkeley, CA 94720-3860
www.stat.berkeley.edu/~census/631.pdf
To appear in *Mathematische Zeitschrift*

The Markov Moment Problem and de Finetti's Theorem: Part II 15 June 2003

by Persi Diaconis and David Freedman

Abstract

This paper gives an abstract version of de Finetti's theorem that characterizes mixing measures with L_p densities. The general setting is reviewed; after the theorem is proved, it is specialized to coin tossing and to exponential random variables. Laplace transforms of bounded densities are characterized, and inversion formulas are discussed.

Introduction

In part I of this paper, we discussed the Hausdorff moment problem on the unit interval, and explained how such problems can be translated into questions about the prior or "mixing" measure in Bayesian statistics. Our object here is to give a version of de Finetti's theorem that characterizes mixing measures with L_p densities, in the general setting described by Diaconis and Freedman (1984), which covers "partial exchangeability." We begin by reviewing the setup and proving general theorems; then we give some examples, showing how the general theory specializes to normal variables, coin tossing, and exponential variables. In connection with the latter, we characterize Laplace transforms of bounded densities and discuss inversion formulas. As will be seen, the abstract theory gives a generalized procedure for inverting probability transforms. Finally, there is a brief literature review. Theorems 2–4 and their corollaries are thought to be new.

The abstract setup can be described as follows. For $i = 1, 2, \ldots,$ let Ω_i be a Polish space equipped with the Borel σ-field $\mathcal{F}_i$. Let $\Omega = \prod_{i=1}^{\infty} \Omega_i$ and $\mathcal{F} = \prod_{i=1}^{\infty} \mathcal{F}_i$. Let X_i be the ith coordinate function on Ω. The nth "sufficient statistic" T_n is a Borel mapping from $\prod_{i=1}^{n} \Omega_i$ to a Polish space W_n equipped with its Borel σ-field $\mathcal{B}_n$. In principle, T_n does not act on Ω, although $T_n(X_1, \ldots, X_n)$ does. For each n and $t \in W_n$, let $Q_{n,t}$ be a probability on $\left(\prod_{i=1}^{n} \Omega_i, \prod_{i=1}^{n} \mathcal{F}_i\right)$. It is assumed that $t \to Q_{n,t}$ is Borel.

To illustrate the setup, suppose the X_i are independent normal random variables with common mean 0 and variance $\sigma^2 > 0$. Then Ω_i would be the real line, W_n would be the set of positive real numbers, and $T_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i^2$. In this example, $Q_{n,t}$ is uniform on the n-tuples of real numbers $(x_1, \ldots, x_n)$ with $\sum_{i=1}^{n} x_i^2 = t$. Geometrically, this set of n-tuples is the sphere centered at 0 having radius $\sqrt{t}$. Statistically, $Q_{n,t}$ is the conditional distribution of the sample, given the sufficient statistic.

We return to the abstract setting, and define M_Q, the partially exchangeable probabilities, as the set of P on $(\Omega, \mathcal{F})$ such that for each n, given $T_n(X_1, \ldots, X_n) = t$, a regular conditional P-distribution for $X_1, \ldots, X_n$ is $Q_{n,t}$. Informally, $Q_{n,t}$ is the distribution of the data given that the sufficient statistic took the value t. This does not depend on the parameters, i.e., is the same for all $P \in M_Q$. Said another way, M_Q is the set of P for which $Q_{n,t}$ works as advertised. In our normal example, M_Q will turn out to be the set of probability distributions faced by a textbook Bayesian statistician, who is going to observe (by assumption) a sequence of independent normal random variables with mean 0 and variance $\sigma^2 > 0$, and is contemplating all possible prior probabilities for σ^2. That is the content of Theorem 1 below.

We impose the following regularity conditions (which are obvious for the normal, once you decipher the notation).

(1) $Q_{n,t}\{T_n = t\} = 1$.
(2) If $T_n(x_1, \ldots, x_n) = T_n(x_1{}', \ldots, x_n{}')$ then $T_{n+1}(x_1, \ldots, x_n, x) = T_{n+1}(x_1{}', \ldots, x_n{}', x)$ for all $x \in \Omega_{n+1}$.
(3) For each $s \in W_n$ and $t \in W_{n+1}$, relative to $Q_{n+1,t}$, the kernel $Q_{n,s}$ is a regular conditional distribution for $(X_1, \ldots, X_n)$ given $T_n(X_1, \ldots, X_n) = s$ and $X_{n+1} = x$. Here, the X_i are viewed as the coordinate functions on $\prod_{i=1}^{n+1} \Omega_i$.

We define the partially exchangeable σ-field $\hat{\Sigma}$ as

$$\hat{\Sigma} = \bigcap_{n=1}^{\infty} \hat{\Sigma}(n),$$

where $\hat{\Sigma}(n)$ is spanned by $T_n(X_1, \ldots, X_n), X_{n+1}, X_{n+2}, \ldots$. The main theorem proved in Diaconis and Freedman (1984) is the following.

Theorem 1. *Conditions* (1), (2), *and* (3) *are in force. Then* M_Q *is convex, and there is a set* $G \in \hat{\Sigma}$ *with the following properties.*

(i) $P(G) = 1$ *for all* $P \in M_Q$.
(ii) *For each* $\omega \in G$, *the sequence of probabilities* $Q_{n,T_n(\omega)}$ *converges weak-star to a limiting probability* $Q_\omega \in M_Q$, *which is* 0–1 *on* $\hat{\Sigma}$.
(iii) *As* ω *ranges over* G, *the kernels* Q_ω *range over the extreme points of* M_Q.
(iv) *For any* $P \in M_Q$, *the kernel* Q_ω *is a regular conditional* P*-distribution for* $X_1, X_2, \ldots$ *given* $\hat{\Sigma}$, *and*

$$P = \int_G Q_\omega \, \hat{P}(d\omega), \tag{4}$$

with $\hat{P}$ *the restriction of* P *to* $\hat{\Sigma}$. *The representation* (4) *is unique, i.e.,* $\hat{P} \leftrightarrow P$. *Moreover,* P *is extreme iff* P *is* 0–1 *on* $\hat{\Sigma}$, *i.e.,*

$$P\{\omega : \omega \in G \;\&\; Q_\omega = P\} = 1. \tag{5}$$

Remarks. (i) The σ-field $\hat{\Sigma}$ may be restricted even further, to the σ-field $\check{\Sigma}$ spanned by $\omega \to Q_\omega$. Then $\hat{P}$ is replaced by $\check{P}$, the restriction of P to $\check{\Sigma}$, the advantage being that $\check{\Sigma}$ is a Borel σ-field equivalent to the inseparable $\hat{\Sigma}$ up to sets that have measure 0 for all $P \in M_Q$.

(ii) In this context, P is the mixture and $\check{P}$ is the mixing measure. Equation (4) becomes

$$P = \int_G Q_\omega \, \check{P}(d\omega). \tag{6}$$

(iii) If we identify all points in the same atom of $\check{\Sigma}$, the resulting quotient space $\mathcal{X}$ is analytic. The quotient of $\check{\Sigma}$ is the Borel σ-field in $\mathcal{X}$ and the quotient π of $\check{P}$ is a probability on that σ-field. Then

$$P = \int_{\mathcal{X}} Q_x \, \pi(dx). \tag{7}$$

This may be a more convincing analog to de Finetti's theorem for coin-tossing. To define Q_x, choose any ω in the fiber corresponding to x—it doesn't matter which—and set $Q_x = Q_\omega$. Let X map ω in G to $x \in \mathcal{X}$, so that $\pi = PX^{-1}$, i.e., π is the limiting distribution of the random measures Q_{n,T_n}. Among other things, the mixing measure has been recovered from the mixture. This is a generalized inversion formula; the application to Laplace transforms will be detailed below. Technically, $\mathcal{X}$ can be realized as the image of G under X, and is then an analytic subset of the set of probabilities on Ω; in applications, $\mathcal{X}$ will be a homelier object.

(iv) In our normal example, G can be taken as the set where $(X_1^2 + \cdots + X_n^2)/n$ converges to a finite positive limit L. Then Q_ω makes the coordinates independent normal random variables, with variance $L(\omega)$. The quotient space $\mathcal{X}$ in (iii) is $(0, \infty)$, and the quotient probability π is the prior on σ^2, viz., the distribution of L. Said with less formality, $X_1, X_2, \ldots$ have an orthogonally invariant distribution iff they are scale mixtures of independent normal variables with mean 0 and variance 1. (It is the distributions that are being mixed, not the random variables; the customary informal language is, well, informal.)

In view of Theorem 1(iii) and (5),

$$Q_{\omega'}\{\omega : \omega \in G \;\&\; Q_\omega = Q_{\omega'}\} = 1 \text{ for all } \omega' \in G, \tag{8}$$

an equation that will be used later. Condition (1) implies that $t \to Q_{n,t}$ is 1–1. Hence T_n and Q_{n,T_n} span the same σ-field, and we may view Q_{n,T_n} as the sufficient statistic instead of T_n.

Bounded densities

Our first result characterizes mixtures where the mixing measure has a bounded density. It is the abstract version of Theorem 4 for L_∞ in Part I. Let $P_i \in M_Q$ for $i = 0, 1$, let n_0 be a positive integer, and let c be a positive constant. Conditions (1), (2), and (3) are in force; we use the notation of Theorem 1.

Theorem 2. *Let $P^{(n)}$ be the restriction of P to the σ-field spanned by $X_1, X_2, \ldots, X_n$, and let $U_n = T_n(X_1, \ldots, X_n)$ map Ω to W_n. The following conditions are equivalent.*

(i) $\check{P}_1 \le c\check{P}_0$.

(ii) $P_1 \le cP_0$.

(iii) $P_1^{(n)} \le cP_0^{(n)}$ *for all* $n = n_0, n_0 + 1, \ldots$.

(iv) $P_1 U_n^{-1} \le cP_0 U_n^{-1}$ *for all* $n = n_0, n_0 + 1, \ldots$.

Proof. Plainly, (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv), the first implication being immediate from (6). Next, (iv) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i). The first implication results from sufficiency: if $P \in M_Q$, then

$$P^{(n)} = \int_{W_n} Q_{n,t}\, PU_n^{-1}(dt). \tag{9}$$

The proof of Theorem 2 is complete. From the present perspective, if condition (iii) holds for any n_0, it plainly holds for $n_0 = 1$; this will be helpful in one of the applications below, where the dependence of the conditions on n_0 will be less transparent. Theorem 2 characterizes mixtures with bounded densities, and in the next section, we turn to L_p densities.

Theorem 3. *$P_1 << P_0$ iff $\check{P}_1 << \check{P}_0$, and then $d\check{P}_1/d\check{P}_0$ is a version of dP_1/dP_0.*

Proof. If $\check{P}_1 << \check{P}_0$, then $P_1 << P_0$ by (6). The converse is obvious from the fact that $\check{P}$ restricts P to a smaller σ-field. To compute the Radon-Nikodym derivative, suppose $\check{P}_1 << \check{P}_0$. Let $\phi = d\check{P}_1/d\check{P}_0$, and fix $A \in \mathcal{F}$. Then

$$\int_A \phi\, dP_0 = \int_G \left(\int_A \phi(\omega)\, Q_{\omega'}(d\omega) \right) \check{P}_0(d\omega'). \tag{10}$$

But $Q_{\omega'}$ concentrates on the $\check{\Sigma}$-atom containing ω' by (8), and ϕ is $\check{\Sigma}$-measurable, so

$$Q_{\omega'}\{\phi = \phi(\omega')\} = 1,$$

and we may replace $\phi(\omega)$ on the right side of (10) by $\phi(\omega')$. Thus

$$\begin{aligned}
\int_A \phi\, dP_0 &= \int_G \left(\int_A \phi(\omega')\, Q_{\omega'}(d\omega) \right) \check{P}_0(d\omega') \\
&= \int_G \left(\int_A Q_{\omega'}(d\omega) \right) \phi(\omega')\, \check{P}_0(d\omega') \\
&= \int_G Q_{\omega'}(A)\, \check{P}_1(d\omega') \\
&= P_1(A).
\end{aligned}$$

Corollary 1. *Let $\phi = d\check{P}_1/d\check{P}_0$, with $\phi = \infty$ on the part of the space where $\check{P}_1$ is singular with respect to $\check{P}_0$. Define Φ in the analogous way for P_0 and P_1. Then $\Phi = \phi$ a.e. $[P_0 + P_1]$.*

Proof. This is immediate from Theorem 3, on replacing P_0 by $\frac{1}{2}(P_0 + P_1)$. In principle, Φ need only be $\mathcal{F}$-measurable; in fact, however, Φ is $\check{\Sigma}$-measurable up to null sets.

Recall that $U_n = T_n(X_1, \ldots, X_n)$. Suppose

$$P_1 U_n^{-1} << P_0 U_n^{-1} \text{ for all } n. \tag{11}$$

Let $h_n = dP_1 U_n^{-1}/dP_0 U_n^{-1}$, a Borel function on W_n. Let

$$c_n = \left(\int_{W_n} h_n^p\, dP_0 U_n^{-1} \right)^{1/p} \tag{12}$$

and

$$H_n = h_n(U_n). \tag{13}$$

Recall that $P^{(n)}$ is the restriction of P to the σ-field spanned by $X_1, \ldots, X_n$, so that $P_1^{(n)} << P_0^{(n)}$ by (9) and (11); and $H_n = dP_1^{(n)}/dP_0^{(n)}$ is a martingale relative to P_0. The proof of the next theorem is omitted as a routine application of differentiation theory (Hewitt and Stromberg, 1969, pp. 369–75).

Theorem 4. *Assume* (11), *and definitions (*12–13*). Then* c_n *is non-decreasing as* n *increases. Moreover,* H_n *converges a.e.* $P_0 + P_1$ *to a limit* H*, which is infinite on the part of the space where* P_1 *is singular with respect to* P_0*, and* dP_1/dP_0 *on the part of the space where* $P_1 << P_0$*. Finally,* $\lim c_n = (\int H^p \, dP_0)^{1/p}$.

Corollary 2. $P_1 << P_0$ *with an* L_p *density having norm at most* c *iff* $\sup c_n \leq c$.

Remarks. (i) $P_1 << P_0$ iff H_n is uniformly P_0-integrable; but this amounts to little more than restating the definition of absolute continuity.

(ii) Corollary 2 is the abstract version of Theorem 9 in Part I. Theorems 2 and 3 here capture the reasoning for Theorem 4 in Part I.

Examples

Example 1: Coin tossing. To make contact with de Finetti's original result for coin tossing—Theorem 5 in Part I—let $\Omega_i = \{0, 1\}$ and $W_n = \{0, \ldots, n\}$, with the discrete topology on both. Let $T_n(x_1, \ldots, x_n) = x_1 + \cdots + x_n$. Informally, 1 is heads, 0 is tails; $\omega \in \Omega$ is the record of an infinite number of coin tosses, $X_i(\omega)$ being the outcome on the ith toss; $T_n(X_1, \ldots, X_n)$ is the number of heads in the first n tosses of the coin. For $j = 0, \ldots, n$, let $Q_{n,j}$ be the uniform distribution on the $\binom{n}{j}$ sequences of 0s and 1s of length n whose sum is j. It takes only a few (tedious) minutes to verify the following:

M_Q consists of all the exchangeable probabilities on $(\Omega, \mathcal{F})$.

Conditions (1), (2), and (3) are satisfied.

In Theorem 1, the σ-field $\hat{\Sigma}$ consists of the Borel sets invariant under finite permutations of coordinates.

G can be taken as the set where $(X_1 + \cdots + X_n)/n$ converges as $n \to \infty$; call the limit L.

Q_ω is the probability on Ω making the coordinates X_i independent tosses of a p-coin, where $p = L(\omega)$.

The quotient space $\mathcal{X}$ in Remark (iii) is the unit interval; the quotient σ-field is the Borel σ-field; and the quotient probability is the distribution of L, which is the mixing measure μ on $[0, 1]$ in Theorem 5 of Part I.

Theorem 5 in part I is therefore a special case of Theorem 1 here. Of course, a direct proof is easier. But Theorem 1 does provide a unified framework for de Finetti's theorem and many variations. Corollary 2 here gives Theorem 9 in Part I, and the present Theorem 2 does L_∞. At least for us, the abstract setup makes the structure of the proofs easier to see.

Example 2: Exponential random variables. The random variable $X > 0$ has the exponential distribution with parameter λ if $P(X > x) = \exp(-\lambda x)$ for $x > 0$. Here, $0 < \lambda < \infty$. Mixtures of independent exponentials with a common parameter were characterized by Freedman (1963). Informally, a sequence of positive random variables is a mixture of exponentials iff the sums are sufficient statistics, and given the sum, the summands are uniformly distributed over the simplex. To make the connection with Theorem 1, we take $\Omega_i = W_n = (0, \infty)$ with the Borel σ-field.

Let $T_n(x_1, \ldots, x_n) = x_1 + \cdots + x_n$. For $0 < t < \infty$, let $Q_{n,t}$ be the uniform distribution on the positive, finite $x_1, \ldots, x_n$ whose sum is t. Let P_λ be the probability on $(\Omega, \mathcal{F})$ according to which the coordinates X_i are independent exponentials with the common parameter λ. If μ is a probability on $(0, \infty)$, let $P_\mu = \int_0^\infty P_\lambda \, \mu(d\lambda)$, that being the mixture we want to characterize. Abstractly, P on $(\Omega, \mathcal{F})$ admits the representation $P = P_\mu$ iff $P \in M_Q$, and then μ is unique. The G in Theorem 1 is again the set where $(X_1 + \cdots + X_n)/n$ converges to a finite positive limit; denote the latter by L. And $Q_\omega = P_{1/L(\omega)}$ makes the X_i independent exponentials with common parameter $1/L(\omega)$: the inverse results from the fact that an exponential distribution with parameter λ has mean $1/\lambda$. The quotient space $\mathcal{X}$ in Remark (iii) is $(0, \infty)$, the quotient σ-field is the Borel σ-field, and the quotient probability is the distribution of $1/L$, namely, μ. If μ is allowed to have positive mass at 0, the argument is a little more complicated, because the X_i will be infinite with probability $\mu\{0\}$.

Suppose μ and ν are two probabilities on $(0, \infty)$, and c is a positive real number. It is almost obvious from Theorem 2 that $P_\mu \le c P_\nu$ iff $\mu \le c\nu$. In the next section, we restate the condition in terms of the Laplace transform, which may be more interesting. We also characterize μ with a bounded density: this is (a little) beyond the scope of our previous theorems, since Lebesgue measure is infinite on $(0, \infty)$.

In these examples, the "sufficient statistic" is the sum, and the conditional distribution is uniform—on $\{0, 1, \ldots, n\}$ for the coin and the simplex for the exponential. In other situations, the sufficient statistic and the conditional will be more complicated: see Diaconis and Freedman (1984) for more examples and discussion.

Laplace transforms

Let μ be a probability on $[0, \infty)$. Its Laplace transform is

$$\phi(x) = \int_0^\infty e^{-\lambda x} \, \mu(d\lambda). \tag{14}$$

We use λ as the variable of integration, in keeping with Example 2, and write ϕ_μ for ϕ if there is any ambiguity. According to a celebrated theorem of Bernstein, Laplace transforms of probabilities on $[0, \infty)$ are characterized as being "completely monotone," and taking the value 1 at $x = 0$; furthermore, μ in (14) is unique. See Widder (1946, pp. 144–163) or Feller (1971, pp. 233, 439). For these purposes, ϕ on $[0, \infty)$ is completely monotone if the nth derivative $\phi^{(n)}$ exists on $[0, \infty)$ for all n, and these functions alternate in sign, so that $(-1)^n \phi^{(n)} \ge 0$ for all $n = 0, 1, \ldots$. Of course, $\phi^{(n)}$ is continuous because $\phi^{(n+1)}$ exists. At 0, continuity and differentiability are from the right: ϕ may not be defined to the left of 0. By convention, $\phi^{(0)} = \phi$.

To avoid technical nuisances, we assume until further notice that $\mu\{0\} = 0$. Recall that $X_1, \ldots, X_n$ are independent exponential random variables relative to P_λ, with common parameter λ. The density of $X_1 + \cdots + X_n$ is

$$x \to \frac{x^{n-1}}{(n-1)!} e^{-\lambda x} \lambda^n \tag{15}$$

for $n = 1, 2, \ldots$. This is a well known formula (Feller, 1971, p. 11), and is easy to verify directly. To get the density of the sum relative to P_μ we just the integrate (15) with respect to $\mu(d\lambda)$:

$$\frac{x^{n-1}}{(n-1)!} \int_0^\infty e^{-\lambda x} \lambda^n \, \mu(d\lambda) = (-1)^n \frac{x^{n-1}}{(n-1)!} \phi^{(n)}(x) \tag{16}$$

for $n = 1, 2, \ldots$. The equality in (16) follows by differentiating (14) under the integral sign, n times.

Lemma 1. *Let μ and ν be two probabilities on $(0, \infty)$. Let c be a positive constant. Then $\mu \leq c\nu$ iff $(-1)^n \phi_\mu^{(n)}(x) \leq c(-1)^n \phi_\nu^{(n)}(x)$ for all $n = 0, 1, \ldots$ and $x > 0$. For sufficiency, the upper bound is needed only for large positive n.*

Proof. Combine Theorem 2 and (16), the latter giving the density of the sufficient statistic with respect to P_μ or P_ν. This is where we use n_0 in Theorem 2.

Corollary 3. *Let ν be exponential with parameter h. Then $\mu \leq c\nu$ iff*

$$(-1)^n \phi_\mu^{(n)}(x) \leq cn!h/(x+h)^{n+1}$$

for all $n = 0, 1, \ldots$ and $x > 0$. For sufficiency, the upper bound is needed only for large positive n.

Proof. The Laplace transform of ν is $\phi_\nu(x) = h/(h+x)$, so $(-1)^n \phi_\nu^{(n)}(x) = hn!/(h+x)^{n+1}$.

Theorem 5. *Let ϕ be a given function on $[0, \infty)$, and c a given positive real number. Then ϕ is the Laplace transform of a probability μ on $(0, \infty)$ such that μ is absolutely continuous, with a density bounded above by c, iff $\phi(0) = 1$ and*

$$0 \leq (-1)^n \phi^{(n)}(x) \leq cn!/x^{n+1} \tag{17}$$

for all $n = 0, 1, \ldots$ and $x > 0$. Furthermore, μ is unique. For sufficiency, the upper bound is needed only for large positive n.

Proof. For uniqueness, (14) determines μ according to Bernstein's theorem. Suppose that ϕ is the Laplace transform of a probability μ on $(0, \infty)$ with $d\mu/dx \leq c$; the conditions on ϕ and its derivatives follow by routine calculus, proving necessity. For sufficiency, μ exists by Bernstein's theorem. Let

$$\psi(x) = \int_0^\infty e^{-\lambda x}\, e^{-\lambda} \mu(d\lambda) = \phi(x+1). \tag{18}$$

Plainly, $(-1)^n \psi^{(n)}(x) \leq cn!/(1+x)^{n+1}$. Corollary 3 shows that $e^{-\lambda}\mu(d\lambda) \leq ce^{-\lambda} d\lambda$, which completes the proof.

Essentially this theorem can be found in Widder (1946, p. 315) or Feller (1971, p. 440); also see Hirschman and Widder (1955, chap. 7). There are similar—albeit more complicated—results for L_p: see Widder (1946, pp. 288, 312–14). Rather than pursuing this topic, we turn to inversion formulas for the Laplace transform (Widder, 1946, p. 288; Feller, 1971, p. 440). These have always seemed mysterious, at least to us; the theory developed here may help. In Example 2, the P_μ density of $n/(X_1 + \cdots + X_n)$ converges weak-star to μ: indeed, $n/(X_1 + \cdots + X_n)$ converges a.e. $[P_\lambda]$ to λ, by the strong law. The density of the denominator $X_1 + \cdots + X_n$ was computed from the Laplace transform ϕ of μ, in (16). By a change of variables ($y = n/x$), the density of $n/(X_1 + \cdots + X_n)$ is seen to be

$$f_n(y) = (-1)^n \frac{1}{n!} \left[\phi^{(n)}\left(\frac{n}{y}\right)\right] \left(\frac{n}{y}\right)^{n+1}. \tag{19}$$

As noted above, $f_n(y)dy$ converges to $\mu(dy)$ as n grows, which gives the basic inversion formula (Widder, 1946, p. 288). The convergence is better for smoother μ, but weak-star convergence always holds. We have assumed $\mu\{0\} = 0$. Otherwise, the contribution from 0 needs to be assessed separately: the distribution of $n/(X_1 + \cdots + X_n)$ picks up an atom at 0, whose mass is—naturally—$\mu\{0\}$.

Example 3. Let $0 < \alpha < \infty$. The Γ-density $\lambda \to \lambda^{\alpha-1}e^{-\lambda}/\Gamma(\alpha)$ has Laplace transform $x \to 1/(1+x)^\alpha$ for $0 \le x < \infty$. Now let $f(\lambda) = 1$ for λ near 0, while $f(\lambda) = 1/\sqrt{|1-\lambda|}$ for λ near 1, and $f(\lambda) = e^{-\lambda}$ for λ near ∞. The definition of f on $(0, \infty)$ can be completed so that f is a positive density, and C_∞ except at 1. It is routine to show that the Laplace transform ϕ of f is approximately 1 near 0 and $1/x$ near ∞. Plainly, f is unbounded. In short, the condition $\phi(x) \le c/x$ does not establish the boundedness of f in Theorem 5. In this example, the upper bound in (17) will hold for $n = 0, \ldots, n_0$, although c will depend on n_0. As $n \to \infty$, however, (19) suggests that $(-1)^n\phi^{(n)}(x)x^{n+1}/n!$ will be unbounded for x near n, so the upper bound in (17) fails. We have not verified this directly, but see Widder (1946, p. 288).

Remarks. (i) We think that Widder (1946, p. 288, Definition 6) omitted a factor $1/k!$ in the definition of $L_{k.t}$; if so, our (19) matches up; otherwise, we cannot verify the calculations following his definition.

(ii) From the present perspective, Bernstein's theorem can be derived from Hausdorff's solution to the little moment problem—Theorem 2 in Part I. The connection is made by the mapping $\lambda \to -\log\lambda$, which takes the unit interval to the half-line. Bernstein seems to have been unaware of Hausdorff's work; Widder confesses to having rediscovered it for a third time (Widder, 1946, p. 144). With respect to Hausdorff's solution to Markov's problem, we might be in fourth place.

Example 4: Normal random variables with mean 0. In connection with Theorem 1, we considered scale mixtures of normal random variables with common mean 0. There, we used variance as the parameter; here, it will be more convenient to use the "natural parameter" $\lambda = 1/\sigma^2$. See Lehmann (1991, p. 57). Let $\Omega_i = (-\infty, \infty)$, and let P_λ on $\Omega = \prod_i \Omega_i$ make the coordinates X_i independent normal random variables, with mean 0 and common variance $1/\lambda$. For any probability μ on $(0, \infty)$, let $P_\mu = \int P_\lambda\,\mu(d\lambda)$. When does μ have a bounded density with respect to Lebesgue measure? with respect to Haar measure $d\lambda/\lambda$? The nth sufficient statistic will be taken as $T_n = \frac{1}{2}(X_1^2 + \cdots + X_n^2)$. Let $x \to \psi_{\mu,n}(x)$ be the density of T_n with respect to P_μ. By excluding a set of measure 0 with respect to all P_μ, we can assume that our X_i never vanish, so $T_n > 0$. Let $m = n/2$. For $n = 1, 2, \ldots$, the density of T_n with respect to P_λ is

$$x \to \frac{x^{m-1}}{\Gamma(m)} e^{-\lambda x}\lambda^m.$$

See (Feller, 1971, pp. 47–48). The density with respect to P_μ is therefore

$$x \to \psi_{\mu,2m}(x) = \frac{x^{m-1}}{\Gamma(m)}\int_0^\infty e^{-\lambda x}\lambda^m\,\mu(d\lambda) = (-1)^m\frac{x^{m-1}}{(m-1)!}\phi_\mu^{(m)}(x)$$

for $m = 1, 2, \ldots$, where $\phi_\mu(x)$ is the Laplace transform of μ; the second equality holds by (16). Now $x^2\psi_{\mu,2m}(x)/m = (-1)^m x^{m+1}\phi_\mu^{(m)}(x)/m!$, and Theorem 5 shows

(20) The mixing measure μ in Example 4 is absolutely continuous with a density bounded above by c iff $\psi_{\mu,2m}(x) \le cm/x^2$ for all positive x and $m = 1, 2, \ldots$.

Interestingly, this constrains T_n only for positive even n: we are not ready for fractional derivatives, nor is $\psi_{\mu,0}$ defined. (Among other things, $\Gamma(0) = \infty$ and $T_0 = 0$ if it is to be defined at all.) Of course, if $\dot{\mu} \le c$, then $\phi_\mu(x) \le c/x$; but going in, the upper bound is unavailable for the Laplace transform itself. That is why we wanted a version of Theorem 5 that requires the upper bound only for derivatives.

Of course, if $\phi(x)$ is the Laplace transform of $\mu(d\lambda)$, then $-\phi'(x)$ is the Laplace transform of $\lambda\mu(d\lambda)$, and Theorem 5 can be applied to the latter. Indeed,

$$(-1)^m \frac{x^{m+1}}{m!} \frac{\partial^m}{\partial x^m}(-\phi_\mu) = (-1)^{m+1} \frac{x^{m+1}}{m!} \frac{\partial^{m+1}}{\partial x^{m+1}} \phi_\mu = x\psi_{\mu,2m+2}(x).$$

Consequently,

(21) μ is absolutely continuous with a density bounded above by $\lambda \to c/\lambda$ iff $\psi_{\mu,2m+2}(x) \le c/x$ for all positive x and $m = 0, 1, \ldots$.

Example 2 can be handled in a similar way. This is not surprising, since the sum of m exponential variables is distributed as 1/2 times a χ^2 variable with $2m$ degrees of freedom.

(22) The mixing measure μ in Example 2 is absolutely continuous with a density bounded above by c iff the density of the sufficient statistic $X_1 + \cdots + X_m$ is bounded above by cm/x^2 for all positive x and $m = 1, 2, \ldots$.

There is an entertaining geometrical consequence to the connection between the χ^2 and the exponential distributions. Let $X_1, X_2, \ldots, X_{2n-1}, X_{2n}$ be independent normal random variables, with mean 0 and variance 1. Then $(X_1^2 + X_2^2)/2, \ldots, (X_{2n-1}^2 + X_{2n}^2)/2$ are independent standard exponential variables. Given $X_1^2 + X_2^2 + \cdots + X_{2n-1}^2 + X_{2n}^2$, we have on the one hand that

$$X_1, X_2, \ldots, X_{2n-1}, X_{2n}$$

is uniformly distributed over a sphere in R^{2n}; on the other hand,

$$(X_1^2 + X_2^2)/2, \ldots, (X_{2n-1}^2 + X_{2n}^2)/2$$

is uniformly distributed over a simplex in the positive orthant of R^n. Consequently,

(23) Pick a point $(x_1, x_2, \ldots, x_{2n-1}, x_{2n})$ uniformly at random on the surface of a sphere in $2n$-dimensional Euclidean space. Then the point $\big((x_1^2 + x_2^2), \ldots, (x_{2n-1}^2 + x_{2n}^2)\big)$ is uniformly distributed over the corresponding simplex in the positive orthant of n-dimensional space.

In general, as is well known, the partitioned sum of squares has a Dirichlet distribution on the simplex.

Brief Literature Review

The proof of Theorem 1 is given in Diaconis and Freedman (1984). This follows Oxtoby (1952), who gave a masterful exposition of the Krilov-Bogolioubov theory, presenting stationary processes as mixtures of ergodic processes. Similar techniques were used by Hunt (1960) to develop the Martin boundary for transient Markov chains. The Scandinavian school has worked on such problems from a slightly different perspective: see Martin-Löf (1974), Lauritzen (1988), or Kallenberg (1999). There has been an extensive development of such theories in statistical mechanics; see Ruelle (1984) and Georgii (1988). Aldous (1985) discusses applications to probability theory; and Schervish (1995), to Bayesian statistics. Many other examples, discussed from the perspective of semigroups and Choquet theory, will be found in Berg, Christensen, and Ressel (1984); the connection to de Finetti's theorem is explained in Ressel (1985). The characterization of mixtures of normals appears in Freedman (1963). It is often attributed to Schoenberg (1938a): see especially Theorem 2 on p. 817, also see Schoenberg (1938b). But the translation is not without difficulty.

References

Aldous, D. J. (1985). *Exchangeability and related topics.* École d'été de probabilités de Saint-Flour XIII. edité par P. L. Hennequin. Springer-Verlag. Lecture notes in mathematics 1117.

Berg, C., Christensen, J. P. R., Ressel, P. (1984). *Harmonic analysis on semigroups: theory of positive definite and related functions*. Springer Graduate Texts in Mathematics, vol. 100.

Diaconis, P. and Freedman, D. (1984). Partial exchangeability and sufficiency. In J. K. Ghosh, J. Roy, eds. *Proc. Indian Statist. Assoc. Golden Jubilee: Applications And New Directions.* Indian Statist. Assoc., Calcutta, 205–236.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications.* Vol. II, Second edition, Wiley, New York.

Freedman, D. A. (1963). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Stat.* 34: 1194–1216.

Georgii, H. O. (1988). *Gibbs Measures and Phase Transitions.* de Gruyter, Berlin.

Hewitt, E. and Stromberg, K. (1969). *Real and Abstract Analysis.* Springer, New York.

Hirschman, I. I. and Widder, D. V. (1955). *The Convolution Transform.* Princeton University Press.

Hunt, G. (1960). Markoff chains and Martin boundaries. *Ill. J. Math.* 4: 313–40.

Kallenberg, O. (1999). Multivariate sampling and the estimation problem for exchangeable arrays. *J. Theor. Prob.* 12: 859–83.

Lanford, O. and Ruelle, D. (1969). Observables at infinity and states with short range correlations in statistical mechanics. *Comm. Math. Phys.* 13: 194–215.

Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics.* Springer Lecture Notes in Statistics, No. 49.

Lehmann, E. L. (1991). *Testing Statistical Hypotheses*. 2nd ed. Wadsworth & Brooks/Cole.

Martin-Löf, P. (1974). Repetitive structures and the relation between canonical and micro-canonical distributions in statistics and statistical mechanics. In O. Barndorff-Nielsen, P. Blaesild, G. Schou, eds. *Proceedings of Conference on Foundational Questions in Statistical Inference.* Aarhus.

Oxtoby, J. C. (1952). Ergodic sets. *Bull. Amer. Math. Soc.* 58: 116–36.

Ressel, P. (1985). De Finetti-type theorems: An analytical approach. *Ann. Prob.* 13: 898–922

Ruelle, D. (1984). *Thermodynamic Formalism: The Mathematical Structures of Classical Equilibrium Statistical Mechanics.* Cambridge University Press.

Schervish, M. J. (1995). *Theory of Statistics.* Springer.

Schoenberg, I. (1938a). Metric spaces and completely monotone functions. *Ann. Math.* 39: 811–41.

Schoenberg, I. (1938b). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.* 44: 522–36.

Widder, D. V. (1946). *The Laplace Transform.* Princeton University Press. First printing, 1941.

Affiliations

Persi Diaconis is Professor of Mathematics and of Statistics, Stanford University, Stanford, CA 94305.

David A. Freedman is Professor of Statistics and of Mathematics, University of California, Berkeley, CA 94720-3860, freedman@stat.berkeley.edu

June 2003
Technical Report No. 631, Part II
Statistics Department
University of California
Berkeley, CA 94720-3860
www.stat.berkeley.edu/~census/631.pdf
To appear in *Mathematische Zeitschrift*