INVARIANTS UNDER MIXING WHICH GENERALIZE DE FINETTI'S THEOREM[1]

By David A. Freedman

University of California, Berkeley

1. Introduction. In 1931, de Finetti published a paper [3] characterizing all stochastic processes which could be represented as mixtures of coin tossing processes—or more precisely, such that the probability measure their marginal distributions induce in the space of sequences of 0's and 1's could be represented as the weighted average of probabilities induced by coin tossing processes. He subsequently [4] generalized this result so as to characterize mixtures of sequences of independent, identically distributed random variables. A concise statement and proof of this result will be found in [8], p. 365.

In 1955, Hewitt and Savage [5] (this paper has a complete bibliography on the subject) made a comprehensive study of this theorem, and obtained results for random variables taking values in quite abstract topological spaces. In the present study, the topology of the range space will be very simple—only natural-number valued random variables will be considered. But the restriction to independence disappears, and with it the consideration of transformations of the base space only which leave the probability fixed.

The basic tool, borrowed from ergodic theory, is the representation theorem of Kriloff and Bogoliouboff [6]. A very elegant presentation of their results will be found in [9]. The generalization is in terms of the "summarizing statistics" of a process, to be defined below. From this point of view, de Finetti's theorem states that a process is a mixture of sequences of independent, identically distributed random variables if and only if it is summarized by the order statistics; that is to say, if and only if any two finite sequences with the same order statistics are assigned the same probability.

The principal generalization is a necessary and sufficient condition for a (stationary) process to be a mixture of (stationary) Markov chains. The condition is that the process be summarized by the transition count; that is, any two finite sequences with the same initial state and the same number of one-step transitions between each pair of states are assigned the same probability. An urn model for this type of process is given in Section 4.

Similar results for some univariate exponential distributions are obtained in Section 5, and more general questions are posed in Section 6. An analysis of the continuous-time case will be made in a future paper.

2. A general theorem. Let $(\mathfrak{S}, B(\mathfrak{S}))$ be a probability space, and let $\{P_\lambda : \lambda \varepsilon \Lambda\}$ be a family of probabilities on $B(\mathfrak{S})$. Take $B(\Lambda)$ to be a σ-algebra

Received May 3, 1961; revised April 2, 1962.

[1] Prepared with the partial support of the Canada Council and of the National Science Foundation, Grant G-14648.

in Λ over which all the λ-functions $P_\lambda(E)$ are measurable, for all $E \varepsilon B(\mathfrak{S})$ (in this circumstance, P_λ is called a $B(\Lambda)$-measurable probability function). If μ is any probability on $B(\Lambda)$, then

DEFINITION 1. *The mixture of P_λ with respect to the probability μ is the function on $B(\mathfrak{S})$ defined by*

$$P_\mu(E) = \int_\Lambda P_\lambda(E)\, d\mu; \qquad E \varepsilon B(\mathfrak{S}).$$

By standard integration theorems, $P\mu$ is a probability on $B(\mathfrak{S})$. From a Bayesian viewpoint, the probability μ is an *a priori* distribution of the unknown parameter λ—see section 1.6 of [7].

This paper is concerned with P_λ induced by the marginal distribution functions of certain classes of stochastic processes, via the Kolmogoroff consistency theorem ([8], p. 93). Hence in the balance of this section, and throughout Sections 3 and 4 the following identifications will be made (Z is the set of natural numbers): $\mathfrak{S}$ is the space of sequences of natural numbers, $B(\mathfrak{S})$ is the σ-field generated by the cylinder sets. The results and proofs hold equally well for the space of two-sided sequences.

The stochastic process $\{X_n : n \varepsilon Z\}$ is defined as the coordinate process,

$$X_n(s) = s(n); \qquad n \varepsilon Z, s \varepsilon \mathfrak{S}, \tag{1}$$

and the shift transformation T which maps $\mathfrak{S}$ onto $\mathfrak{S}$ is defined by

$$(Ts)(n) = s(n+1); \qquad n \varepsilon Z, s \varepsilon \mathfrak{S}.$$

All probabilities P in $\mathfrak{S}$ will be chosen so that the process (1) is stationary, or what is the same, so that P is invariant under T:

$$P(T^{-1}E) = P(E); \qquad E \varepsilon B(\mathfrak{S}).$$

By a slight variant of [6], there is a $B(\mathfrak{S})$-measurable probability function P_s on $\mathfrak{S}$ such that

$$P = \int_{\mathfrak{S}} P_s\, dP \tag{2}$$

and for a set of measure 1 under all stationary probabilities

(i) P_s is metrically transitive ([1], p. 457) i.e., the process (1) is metrically transitive (and the shift T is ergodic) on $(\mathfrak{S}, B(\mathfrak{S}), P_s)$,

(ii) $$P_s(E) = \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n-1} f_E(T^j s)$$

simultaneously for all cylinder sets E, where $f_E(s) = 1, s \varepsilon E; f_E(s) = 0, s \varepsilon / E$.

These remarks paraphrase Sections 1 and 2 of [9]. Indeed, let $\hat{\mathfrak{S}}$ be the space of sequences of natural numbers and ∞; where $Z \cup \{\infty\}$ is the one-point compactification of Z with the discrete topology. Then $\hat{\mathfrak{S}}$ with the product topology is compact and metrizable, so that the Kriloff-Bogoliouboff theory applies to it.

But $\mathfrak{S}$ is a Borel measurable subset of $\hat{\mathfrak{S}}$, and $P_s(\mathfrak{S}) = 1$ on a Borel set of probability 1 for any invariant P with $P(\mathfrak{S}) = 1$. Further, f_E is continuous on $\hat{\mathfrak{S}}$.

In order to obtain Theorem 1, it is necessary to consider certain sequences of statistics defined on the process X_n. The conditions on these sequences are stated as

DEFINITION 2. *Let U_n be a function on Z^n for all n in Z. The sequence $\{U_n : n \varepsilon Z\}$ has S-structure if and only if*

$$U_n(j_1, \cdots, j_n) = U_n(k_1, \cdots, k_n)$$

and

$$U_m(r_1, \cdots, r_m) = U_m(s_1, \cdots, s_m)$$

together imply

$$U_{n+m}(j_1, \cdots, j_n, r_1, \cdots, r_m) = U_{n+m}(k_1, \cdots, k_n, s_1, \cdots, s_m),$$

for all n and m in Z and all sequences $j.$, $k.$, $r.$, and $s.$ in Z.

For example, the order statistics clearly have S-structure. So does the sequence T_n which will be used to analyze mixtures of Markov processes. The statistic T_n maps Z^n into $Z \times Z^{Z \times Z}$, in such a way that almost all coordinates vanish. If $j.$ is a sequence in Z, then

$$T_n(j_1, \cdots, j_n) = (j_1, t_{rs} : r, s \varepsilon Z)$$

where t_{rs} is the number of one-step transitions from r to s among $(j_1, \cdots, j_n)$. Thus, $t_{11}(1, 1, 2, 1) = 1$.

To demonstrate that T_n has S-structure, note that

$$T_n(j_1, \cdots, j_n) = T_n(k_1, \cdots, k_n)$$

implies $k_1 = j_1$. But then $k_n = j_n$, for the sequence ends with j_1 if and only if

$$\sum_r t_{rj_1} = \sum_s t_{j_1 s};$$

while it ends with $j_0 \neq j_1$ if and only if

$$\sum_r t_{rj_0} = 1 + \sum_s t_{j_0 s}.$$

The assertion is then immediate.

The concept of a summarizing statistic is made clear in the following definition.

DEFINITION 3. *A probability P in $\mathfrak{S}$ is summarized by $\{U_n : n \varepsilon Z\}$ if and only if*

$$U_n(j_1, \cdots, j_n) = U_n(k_1, \cdots, k_n)$$

implies

$$P(X_1 = j_1, \cdots, X_n = j_n) = P(X_1 = k_1, \cdots, X_n = k_n).$$

For convenience, a cylinder set of the form $\{s: X_{n_j}(s) = k_j, 1 \leqq j \leqq m\}$ with $n_1 < n_2 < \cdots < n_m$ will be called a pattern, with first state k_1, and last state k_m. For each $j < m$, it has an $(n_{j+1} - n_j)$-step transition from k_j to k_{j+1}. A

sequence A is a pattern with $n_j = j$; its length is m. The corresponding point in Z^m, $\{k_j, 1 \leqq j \leqq m\}$, will be denoted by $\hat{A}$.

In this terminology, the relationship between Definitions 2 and 3 may be stated as follows. Suppose P is summarized by $\{U_n : n \varepsilon Z\}$, which has S-structure. Let A and B be sequences of length n, C and D sequences of length m. Then $U_n(\hat{A}) = U_n(\hat{B})$ and $U_m(\hat{C}) = U_m(\hat{D})$ imply $P(A \cap T^{-(n+d)}C) = P(B \cap T^{-(n+d)}D)$. When P is summarized by $\{T_n : n \varepsilon Z\}$, a somewhat stronger result holds. If A and B are two patterns which begin with the same state and have the same j-step transitions between each pair of states, for all j, then $P(A) = P(B)$.

These two remarks are proved using the same argument. The required probabilities may be computed by filling in the gaps in all possible ways and summing. But the definitions then apply to each summand.

Using this machinery, it is possible to prove the following theorem.

THEOREM 1. *A probability P is summarized by the sequence $\{U_n : n \varepsilon Z\}$ which has S-structure if and only if it may be represented as a mixture of metrically transitive probabilities which are summarized by $\{U_n : n \varepsilon Z\}$.*

PROOF. The "if" part is clear, and has nothing to do with S-structure. The "only if" part will be proved using the representation (2).

Suppose A and B are two sequences of length m, with $U_m(\hat{A}) = U_m(\hat{B})$. Then $P_s(A) = P_s(B)$ a.e. $[P]$. Indeed,

$$P_s(A) = \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n-1} f_A(T^j s) \text{ a.e. } [P],$$

and

$$P_s(B) = \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n-1} f_B(T^j s) \text{ a.e. } [P].$$

Since the quantities on the right lie in $[0, 1]$, the convergence is L^2, and

$$\begin{aligned} E(P_s(A) - P_s(B))^2 = \lim_{n\to\infty} n^{-2} \sum_{j=1}^{n} \sum_{k=1}^{n} \{&E(f_A(T^j s) f_A(T^k s)) \\ &+ E(f_B(T^j s) f_B(T^k s)) - E(f_A(T^j s) f_B(T^k s)) \\ &- E(f_B(T^j s) f_A(T^k s))\}. \end{aligned}$$

For $d = |j - k| \geqq m$, each summand vanishes since

$$P(A \cap T^{-d}A) = P(B \cap T^{-d}B) = P(A \cap T^{-d}B) = P(B \cap T^{-d}A),$$

and these evaluate the four expectations. Moreover, the relative frequency of summands with $|j - k| < m$ goes to 0, and each lies in $[-2, 2]$. Hence $E(P_s(A) - P_s(B))^2 = 0$, and $P_s(A) = P_s(B)$ a.e. $[P]$ as required.

Since there are only a countable number of pairs of finite subsets of Z, after subtracting a countable number of null sets from $\mathfrak{S}$, $\{U_n : n \varepsilon Z\}$ will summarize

all the probabilities P_s with s in the remaining measure 1 set. This completes the proof and the general discussion.

3. Mixtures of Markov chains. In order to apply this theory to the characterization of mixtures of Markov chains, it is only necessary to investigate metrically transitive probabilities which are summarized by $\{T_n : n \in Z\}$. This is done in

THEOREM 2. *A metrically transitive probability P summarized by $\{T_n : n \in Z\}$ is Markov, i.e., with respect to it the process $\{X_n\}$ is Markov.*

PROOF. An invariant probability P is metrically transitive if and only if

$$\lim_{n\to\infty} (1/n) \sum_{j=1}^{n} P(A \cap T^{-j}B) = P(A)P(B) \tag{3}$$

for all sequences A and B (see [8], Theorem C of p. 435). In particular

$$\lim_{n\to\infty}(1/n) \sum_{j=1}^{n} P(X_1 = k, X_{1+j} = k) = P(X_1 = k)^2. \tag{4}$$

Next, let A be a sequence of length m whose last state is k, and B a sequence whose first state is k. Then if $d \geqq 1$, $n \geqq 1$,

$$\begin{aligned} &P\{[X_1 = k] \cap T^{-n}[A \cap T^{-(d+m)}B]\} \\ &\qquad = P\{[X_1 = k, X_{d+2} = k] \cap T^{-(n+d+1)}[A \cap T^{-(m-1)}B]\}, \end{aligned} \tag{5}$$

since the patterns inside braces on both sides of (5) have the same first state and the same transition count. Now let $n \to \infty$ and apply (3) to obtain

$$P[X_1 = k]P(A \cap T^{-(d+m)}B) = P[X_1 = k, X_{d+2} = k]P(A \cap T^{-(m-1)}B). \tag{6}$$

Then the Markov property follows in the form

$$\begin{aligned} &P[X_i = j_i, 1 \leqq i \leqq N]P[X_1 = j_N, X_2 = j_{N+1}] \\ &\qquad = P[X_1 = j_N]P[X_i = j_i, 1 \leqq i \leqq N + 1]. \end{aligned} \tag{7}$$

Indeed, if $P[X_1 = j_N] = 0$, both sides of (7) vanish. Otherwise, by (6), if $n \geqq N$

$$\begin{aligned} &P\{[X_i = j_i, 1 \leqq i \leqq N] \cap [X_n = j_N, X_{n+1} = j_{N+1}]\} \\ &= P[X_1 = j_N]^{-1}P[X_i = j_i, 1 \leqq i \leqq N + 1]P[X_1 = j_N, X_{n-N+1} = j_{N+1}]. \end{aligned} \tag{8}$$

Now let $n \to \infty$. Using (3), the left side of (8) goes $(C, 1)$ to the left side of (7); while (4) implies that the right side of (8) converges $(C, 1)$ to the right side of (7). This completes the proof.

The last theorem of this section is an immediate consequence of Theorems 1 and 2.

THEOREM 3. *The necessary and sufficient condition for a probability to be a mixture of Markov probabilities is that it be summarized by $\{T_n : n \in Z\}$.*

4. An urn model. The following urn model for mixtures of Markov chains of two states was developed in conversation with Professors Blackwell and Dubins.

Start with two Pólya urns ([2], V. 2), U_0 and U_1. Each contains some balls marked 0, some marked 1. An urn U is selected according to some probability distribution. Then a ball is selected at random from U. Define $X_1 = 0$ or 1 according as the ball is marked 0 or 1. Replace the ball, together with another marked the same, in U. Then select a ball at random from U_{X_1}, to determine X_2, etc. The process $\{X_j\}$ is a mixture of (perhaps nonstationary) Markov chains (with stationary transition probabilities). In this simple case, the assumption of stationarity is dispensable. If the initial compositions of U_0 and U_1 differ, $\{X_j\}$ will not be a mixture of coin-tossing processes.

5. Exponential distributions. This section will characterize mixtures of sequences of independent random variables having a common distribution drawn from an exponential family. The mixture will be over some specified parameter of this family.

De Finetti's theorem will be used in the following form. If a probability P is summarized by the order statistics, then almost all $[P]$ of the probabilities P_s in the representation (2) are power product measures; i.e., with respect to them, the process X_n is a sequence of independent, identically distributed random variables. This follows easily by Theorem 1 and equation (3). This may be extended to random variables taking real values (or, e.g., values in a compact metric space) by a trivial discretization argument and, say, Alaoglu's theorem, see section 9 of [8a].

Now suppose a probability P and a sequence $\{U_n : n \varepsilon Z\}$ with S-structure are given. Suppose, moreover, that there exist functions $h(\cdot)$ and $f(\cdot, \cdot)$ such that P factors as

$$(9) \qquad P[X_i = j_i, 1 \leqq i \leqq n] = \left[\prod_{i=1}^{n} h(j_i)\right] f(n, U_n(j_1, \cdots, j_n)).$$

Then almost all $[P]$ of the probabilities P_s in (2) factor in the same way, namely, as

$$(10) \qquad P_s[X_i = j_i, 1 \leqq i \leqq n] = \left[\prod_{i=1}^{n} h(j_i)\right] f_s(n, U_n(j_1, \cdots, j_n)).$$

Minor alterations in the proof of Theorem 1 produce this slight strengthening.

Finally, suppose that the sequence U_n is additive. That is, U_1 is a function from Z to R^k, and $U_n(j_1, \cdots, j_n) = \sum_{i=1}^{n} U_1(j_i)$. Then $\{U_n : n \varepsilon Z\}$ has S-structure, and if the factorization (9) holds then the order statistics summarize P, so that the preceding remark and de Finetti's theorem both apply. That is to say, P is a mixture of power product measures each of which factors as (10). And this implies the following functional equation:

$$(11) \qquad f_s\left(n, \sum_{i=1}^{n} U_1(j_i)\right) = \prod_{i=1}^{n} f_s(1, U_1(j_i)),$$

whenever $\prod_{i=1}^{n} h(j_i) > 0$.

If the image under U_1 of the set $\{j\colon h(j) > 0\}$ is a reasonable subset of R^k, (11) may be solved to give

$$f_s\left(n, \sum_{i=1}^{n} U_1(j_i)\right) = a_s^n \exp \sum_{i=1}^{n} c_s \cdot U_1(j_i),$$

where $a_s > 0$ and $c_s \varepsilon R^k$. In particular

THEOREM 4. *The necessary and sufficient condition that a probability P may be represented as a mixture of probabilities for which the process $\{X_n\}$ is a sequence of independent random variables with common distribution*

(i) *Poisson $P(\lambda)$; mix over λ:*
(ii) *Binomial $B(N, p)$; mix over p:*
(iii) *Inverse Binomial $IB(N, p)$; mix over p:*

is

$$\text{(i)} \qquad P[X_i = j_i, 1 \leqq i \leqq n] = \left[\prod_{i=1}^{n} (j_i\,!)^{-1}\right] f\left(n, \sum_{i=1}^{n} j_i\right),$$

$$\text{(ii)} \qquad P[X_i = j_i, 1 \leqq i \leqq n] = \left[\prod_{i=1}^{n} \binom{N}{j_i}\right] g\left(n, \sum_{i=1}^{n} j_i\right),$$

$$\text{(iii)} \qquad P[X_i = j_i, 1 \leqq i \leqq n] = \left[\prod_{i=1}^{n} \binom{N + j_i - 1}{N - 1}\right] h\left(n, \sum_{i=1}^{n} j_i\right).$$

6. Further questions. The methods of this paper seem to give a fairly satisfactory classification of the ergodic components of integer-valued processes. By limiting arguments, they also give some information about real-valued processes. There the situation is much more complicated. A possible generalization of Theorem 1 to this case is outlined below. Even this deeper result, however, does not lead to conditions for a process to be a mixture of stationary, real-valued Markov processes.

Let $\mathfrak{S} = R^Z$, $B(\mathfrak{S}) = \prod_1^\infty B(R)$, $B(R)$ being the Borel sets of R, and define $X_n(s) = s(n)$ for s in $\mathfrak{S}$. Only probabilities in $\mathfrak{S}$ which are invariant under the shift will be considered. In other words, the marginal distribution functions of stationary processes (with time parameter in Z) are under consideration.

If Q is a measure in $\mathfrak{S}$, its restriction to R^n will be denoted by $Q^{(n)}$. Let P be a probability, and consider its representation (2). Suppose there is a sequence ν_n of totally σ-finite measures in R^n, and a sequence B_n of sub-fields, $B_n \subset B(R)^n$, such that $[P]$ for almost all s

(i) $P_s^{(n)} \ll \nu_n$,
(ii) B_n is sufficient for s, so that ([7], pp. 47–50),

$$dP_s/d\nu_n = h_n f_s(n, \cdot) \tag{12}$$

where h_n is a nonnegative $B(R)^n$-measurable function on R^n, and $f_s(n, \cdot)$ is a nonnegative B_n-measurable function on R^n. It is possible to show that $f.(n, \cdot)$

is a $B(\mathfrak{S}) \times B_n$-measurable function; if $f_p(n, \cdot) = \int_{\mathfrak{S}} f_s(n, \cdot)\, dP(s)$, then

$$dP/d\nu_n = h_n f_p(n, \cdot), \tag{13}$$

and $f_p(n, \cdot)$ is B_n-measurable.

The basic question is: When is the converse true? That is, under what conditions on the sequences ν_n, h_n, and B_n does the factorization (13) guarantee the factorization (12) a.e. $[P]$?

From this point of view, Theorem 1 derives (12) from (13) provided

(i) $\nu_n(E)$ is the number of n-tuplets of natural numbers in E,
(ii) $h_n \equiv 1$,
(iii) the fields B_n are induced by a sequence of statistics having S-structure.

Moreover, even granting (i) and (ii), simple examples show that Theorem 1 fails, unless some condition like (iii) is imposed.

7. Acknowledgments. The author wishes to thank Professor W. Feller for suggesting the problem, and Professor G. A. Barnard for his patient and critical attention to its unfolding.

REFERENCES

[1] Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.
[2] Feller, William (1960). *An Introduction to Probability Theory and Its Applications* **1** 2nd ed. Wiley, New York.
[3] de Finetti, Bruno (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Ser.* 6, Memorie, Classe di Scienze Fisiche, Mathematiche e Naturali **4** 251–299.
[4] de Finetti, Bruno (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68.
[5] Hewitt, Edwin and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501.
[6] Kriloff, N. and Bogoliouboff, N. (1937). La théorie générale de la mesure dans son application à l'étude des systèmes dynamiques de la mécanique non-linéaire. *Ann. Math., 2nd Ser.* **38** 65–113.
[7] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
[8] Loève, Michel (1960). *Probability Theory*, 2nd ed. Van Nostrand, Princeton.
[8a] Loomis, Lynn H. (1953). *An Introduction to Abstract Harmonic Analysis*. Van Nostrand, Princeton.
[9] Oxtoby, John C. (1952). Ergodic sets. *Bull. Amer. Math. Soc.* **58** 116–130.