View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

JOURNAL OF MULTIVARIATE ANALYSIS 18, 83–92 (1986)

Learn-Merge Invariance of Priors: A Characterization of the Dirichlet Distributions and Processes

W. BÖGE

Institut für Angewandte Mathematik,
Universität Heidelberg, West Germany

AND

J. MÖCKS

Zentralinst. f. Seelische Gesundheit,
Mannheim, West Germany

Communicated by D. A. S. Fraser

Learn-merge invariance is a property of prior distributions (related to postulates introduced by the philosophers W. E. Johnson and R. Carnap) which is defined and discussed within the Bayesian learning model. It is shown that this property in its strong formulation characterizes the Dirichlet distributions and processes. Generalizations towards weaker formulations are outlined. © 1986 Academic Press, Inc.

1. Introduction and Outline

Every Bayesian analysis of the multinomial situation requires choosing a prior from the set of all distributions over the possible parameter vectors, i.e., over the simplex spanned by the observation categories. By applying the principle of natural conjugate priors, the class of all priors is restricted to the Dirichlet-family. However, this principle accounts merely for mathematical convenience and does not seem to be genuinely statistically founded. This raises the question, whether other statistically motivated requirements may be formulated with respect to any potentially chosen prior.

Received September 15, 1981; revised January 24, 1984.

AMS 1980 subject classifications: Primary 62A15; Secondary 03B99, 62G99.

Key words and phrases: prior, Dirichlet distribution, multinomial situation, symmetric measures, inductive learning.

0047-259X/86 $3.00

Copyright © 1986 by Academic Press, Inc.
All rights of reproduction in any form reserved.

In this paper we propose a principle termed "Learn-Merge Invariance" (LMI) which in its strongest formulation like the above principle characterizes the Dirichlet distributions and, if suitably applied, also characterizes the Dirichlet processes introduced by Ferguson (1973). Roughly speaking, the principle of LMI states that it should make no difference in inference, whether one combines (merges) some of the observation categories before or after sampling. This principle has not yet been investigated in the literature. However, Lochner (1975) also stated this as a desirable property, but did not know that it characterizes the Dirichlet-distributions, if arbitrary mergers are allowed. Within non-statistical framework the philosophers W. E. Johnson (1932) and R. Carnap (1958) introduced independently postulates which turn out (in section 2) to be equivalent to the principle of LMI in a special case.

Instead of choosing a prior from the set of all measures on the simplex, one may equivalently choose a symmetric measure on $A^{\mathbb{N}}$, where A denotes the finite set of observation categories. This is possible, since both sets of measures are known to be bijectively related to each other, by virtue of the results of Hewitt-Savage (1955) (i.e. the generalization of De Finetti's representation theorem). This relation is utilized in the approach given here. Given the set of all strictly positive symmetric measures on $A^{\mathbb{N}}$, the principle of LMI (and generalizations of it) may be conveniently formulated and discussed in terms of these measures rather than in terms of random variables (Section 2). In Section 3, the LMI-Measures on $A^{\mathbb{N}}$ and their counter parts on the simplex (the Dirichlet family) are characterized. The generalization to infinite A (characterizing the Dirichlet processes) is carried out in section 4.

The main results in the finite case were already obtained by Böge (1969) and some proofs were shortened by Tremmel (1971). This paper reports still shorter proofs. The presentation given here follows the structure worked out by the second author.

2. The Principle of Learn-Merge-Invariance

First we give some notation and definitions. A denotes a finite set of observation categories, and $A^{\mathbb{N}}$, A^* the set of all infinite respectively finite sequences of elements of A (the latter includes the empty sequence). For $d \in A^*$ let $|d|$ denote the lenght of d. $W(A)$ resp. $W(A^{\mathbb{N}})$ is the space of all probability measures on A resp. $A^{\mathbb{N}}$ (with usual σ-algebra). For $P \in W(A^{\mathbb{N}})$ and $d \in A^*$, $P(d)$ denotes the probability of d according to the marginal distribution of P on the first $|d|$ coordinates. According to Hewitt-Savage (1955), $P \in W(A^{\mathbb{N}})$ will be called symmetric if $P(d) = P(\pi d)$ $\forall d \in A^*$, where π may be any permutation operator on the $|d|$ elements. P is called regular if $P(d) > 0$, $\forall d \in A^*$. Let $S(A^{\mathbb{N}})$ and $R(A^{\mathbb{N}})$ respectively denote the set of all symmetric and of all regular probability measures on $A^{\mathbb{N}}$ and let $SR(A^{\mathbb{N}}) = S(A^{\mathbb{N}}) \cap R(A^{\mathbb{N}})$.

For $\varphi\colon A \to B$, $P \in W(A^{\mathbb{N}})$, let P^φ denote the φ-induced measure on $B^{\mathbb{N}}$, and $\Gamma(A)$ the class of all such surjective φ's with arbitrary B. We shall call such mappings "mergers", and $P \to P^\varphi$ the "process of merging" (enlargement of observation categories). $P_d(\cdot)$ ($d \in A^*$, $P \in R(A^{\mathbb{N}})$) denotes the conditional distribution on the remaining coordinates given d. We shall call $P \to P_d$ the "process of learning", since it reflects the updating of P according to the observed data. Note that

$$\begin{gathered} P(d) = P(d_1) \cdot P_{d_1}(d_2) \cdot \cdots \cdot P_{d_1,\dots,d_{n-1}}(d_n) \\ \forall d = d_1,\dots, d_n \in A^* \text{ and } P \in R(A^{\mathbb{N}}). \end{gathered} \tag{2.1}$$

Now the statistician is requested to choose a $P \in SR(A^{\mathbb{N}})$ which reflects all a-priori knowledge. For practical reasons he might be interested to know whether P is Learn-Merge invariant (LMI) for some or all mergers, in the sense that the process of learning is interchangeable with the process of merging. For example in the analysis of categorial data it is common usage to merge the categories after sampling, and it would be undesireable if the conclusions drawn in this case could be different from those obtained–with the same observed frequencies–if the categories had been merged in the beginning. Loosely speaking, LMI states something like "scale" independence of the learning process. Formally: Let $P \in R(A^{\mathbb{N}})$, then all diagrams of the form

$$\begin{tikzcd} P \arrow[r, "d"] \arrow[d, "\varphi"'] & P_d \arrow[d, "\varphi"] \\ P^\varphi \arrow[r, "\varphi d"] & {} \end{tikzcd} \tag{2.2}$$

should commute for a given merger φ and every $d \in A^*$, where φ is defined on A^* in a natural manner. That is $(P_d)^\varphi = P^\varphi_{\varphi d}$ for all $d \in A^*$. We will say that $P \in R(A^{\mathbb{N}})$ is compatible with φ. For this definition one needs only the partition of A induced by φ. Therefore we will interpret $\varphi \in \Gamma(A)$ according convenience either as a mapping or as a partition without mentioning this explicitly. The following proposition states a necessary and sufficient condition for $P \in R(A^{\mathbb{N}})$ to be compatible with a given $\varphi \in \Gamma(A)$.

(2.3) PROPOSITION: *$P \in R(A^{\mathbb{N}})$ is φ-compatible for $\varphi \in \Gamma(A)$ iff $P_d(Y) = P_u(Y)$ for every $Y \in \varphi$ and every $d, u \in A^*$ with $\varphi d = \varphi u$.*

Proof. Assume P to be φ-compatible, let $d, u \in A^*$ where $\varphi d = \varphi u$ and let $|v| = 1$. Setting $\varphi^{-1}v = Y$, the following holds by (2.2):

$$P_d(Y) = (P_d)^{\varphi}\,(v) = P^{\varphi}_{\varphi d}(v) = P^{\varphi}_{\varphi u}(v) = (P_u)^{\varphi}\,(v) = P_u(Y),$$

proving the first part. Conversely it suffices to show

$$(P_{e,d})^{\varphi}(v) = (P_e)^{\varphi}_{\varphi d}(v) \qquad \text{for} \quad e,d \in A^*, \quad |v| = 1.$$

(Then, because of (2.1), $(P_e)^{\varphi}(\varphi d)$ and $P^{\varphi}_{\varphi e}(\varphi d)$ are both equal to $\prod_{i=1}^{n} (P_{e,d_1,\ldots,d_{i-1}})^{\varphi}\,(\varphi d_i)$ where $d = d_1,\ldots, d_n$.) With $\varphi^{-1}v = Y$ and $U = \varphi^{-1}\varphi d$ we have $(P_{e,d})^{\varphi}\,(v) = P_{e,d}(Y)$. On the other hand:

$$(P_e)^{\varphi}(\varphi d, v) = P_e(U \times Y) = \sum_{u \in U} P_e(u \times Y) = \sum_{u \in U} P_e(u) \cdot P_{e,u}(Y)$$
$$= P_e(U) \cdot P_{e,d}(Y) = (P_e)^{\varphi}(\varphi d) \cdot P_{e,d}(Y).$$

Finally $(P_e)^{\varphi}_{\varphi d}(v) = (P_e)^{\varphi}\,(\varphi d, v)/(P_e)^{\varphi}\,(\varphi d) = P_{e,d}(Y) = (P_{e,d})^{\varphi}\,(v)$, completing the proof.

Let $L \subseteq SR(A^{\mathbb{N}})$; we say that L is closed under learning, if for every $P \in L$ and every $d \in A^*$ $P_d \in L$ holds. It is straightforward to show from the definition, that the set of all $P \in SR(A^{\mathbb{N}})$ which are φ-compatible with a specified φ is closed under learning. Using an induction argument and the symmetry of P, we derive from (2.3):

(2.4) COROLLARY: *Let $L \subseteq SR(A^{\mathbb{N}})$ be closed under learning. Every $P \in L$ is φ-compatible if and only if $P_a(Y) = P_b(Y)$ holds $\forall a, b \in A$ with $\varphi a = \varphi b$ and $\forall Y \in \varphi$, $\forall P \in L$.*

P is called an "LMI-measure", if (2.2) holds for all $\varphi \in \Gamma(A)$, that is, P is compatible with every $\varphi \in \Gamma(A)$. Let $CR(A^{\mathbb{N}})$ denote the set of all LMI-measures in $SR(A^{\mathbb{N}})$. It is straightforward to show for $P \in CR(A^{\mathbb{N}})$:

$$P^{\varphi} \in CR(\varphi(A)^{\mathbb{N}}) \qquad \forall \varphi \in \Gamma(A) \tag{2.5}$$

$$P_d \in CR(A^{\mathbb{N}}) \qquad \forall d \in A^* \tag{2.6}$$

The next proposition states an equivalent formulation for $CR(A^{\mathbb{N}})$.

(2.7) PROPOSITION: *Let $P \in SR(A^{\mathbb{N}})$, then $P \in CR(A^{\mathbb{N}})$ iff $P_{d,a}(x) = P_{d,b}(x)$ for every $a, b, x \in A$ with $a \neq x \neq b$, and every $d \in A^*$.*

Proof. Let $P \in CR(A^{\mathbb{N}})$. If $a = b$ there is nothing to prove. Otherwise choose $\varphi \in \Gamma(A)$ with $\{x\} \in \varphi$ and $\{a, b\} \in \varphi$ which is always possible. For $d \in A^*$ we have $P_d \in CR(A^{\mathbb{N}})$ and finally:

$$P_{d,a}(x) = (P_{d,a})^{\varphi}\,(\varphi x) = (P_d)^{\varphi}_{\varphi a}\,(\varphi x) =$$
$$= (P_d)^{\varphi}_{\varphi b}\,(\varphi x) = (P_{d,b})^{\varphi}\,(\varphi x) = P_{d,b}(x).$$

For the converse, let $\varphi \in \Gamma(A)$ be arbitrary, and let $a, b \in A$ with $\varphi a = \varphi b$. Then $P_a(Y) = P_b(Y)$ $\forall Y \in \varphi$. Indeed, if $a, b \notin Y$ this is true by hypothesis. In the other case we have $P_a(Y^c) = P_b(Y^c)$, and the statement follows from $P_a(Y^c) = 1 - P_a(Y) = 1 - P_b(Y)$. The same reasoning holds for P_d, $d \in A^*$, which by (2.4) proves the rest.

Still another equivalent formulation is an easy consequence of (2.7), namely the condition:

$$P_d(x) = P_{d'}(x) \tag{2.8}$$

for every $x \in A$, $d \in A^*$, and all d', which are constructed from $d = d_1, \ldots, d_n$ by arbitrarily replacing any $d_i \in A \setminus \{x\}$ by some element of the same set.

Except for terminology this is the formulation close to the postulates stated by Johnson (1932) and Carnap (1958). Within statistics (2.8) was introduced by Good (1965, 1967), who called (2.8) "Johnson's sufficiency (resp. sufficientness) postulate". However, his statistical applications and generalizations are along other lines than developed here. The first of the present authors, not knowing W. E. Johnson's paper (1932), adopted (2.8) from Carnap et al. (1958, *Ax* 9 or *NA* 14) and proved its equivalence to the LMI-principle in the mid-sixties. Therefore also the LMI-principle was called "Carnap's postulate" in Böge (1969) and in subsequent papers by Tremmel (1971), Möcks-Stöckel (1972), and Kursetz (1973). (2.8) means that $P_d(x)$ is a function of the number of x's in d, $|d|$, and x. In Johnson's postulate the dependence on the latter argument was omitted, while in Carnap's system the dependence is excluded by an additional axiom (compare also Humburg (1971) and Stegmüller (1973)).

3. Learn-Merge-Invariant Measures

In this section we give the proof, that the measures fulfilling (2.8) (and thus by section 2 the LMI-measures) are defined by a simple learning formula (3.3), which in turn is characterizing the Dirichlet distributions. For convenience we start with the latter. Let $m \in W(A)$ with $m > 0$ (i.e. $m(a) > 0$ for every $a \in A$) and $\rho \in \mathbb{R}$ with $\rho > 0$. Consider the Dirichlet distribution $D_{\rho,m}$ defined by the parameter vector $\tilde{m} = \rho^{-1} \cdot m \in \mathbb{R}^{|A|}$. Fix some $b \in A$, then the density of $D_{\rho,m}$ with respect to $\bigotimes_{a \in A \setminus \{b\}} d\lambda_a$ can be written (for a detailed definition compare Wilks (1962), Ferguson (1973)):

$$f_{\rho,m}(\lambda) = D(\tilde{m})^{-1} \prod_{a \in A} \lambda_a^{\tilde{m}(a) - 1} 1_S(\lambda) \tag{3.1}$$

where $\lambda = (\lambda_a)_{a \in A} \in \mathbb{R}^{|A|}$, $D(\tilde{m}) = \prod_{a \in A} \Gamma(\tilde{m}(a)) / \Gamma(\sum_a \tilde{m}(a))$ (here Γ

denotes the usual Γ-function) and 1_S represents the indicator function of the $|A|-1$ dimensional simplex

$$S := \left\{\lambda | \lambda_a > 0,\ \forall a \in A,\ \sum_{a \in A} \lambda_a = 1\right\}.$$

For $|A| = 2$ this reduces to the Beta Distribution. Let $\mathscr{D}_A$ denote the class of all Dirichlet distributions enlarged by the degenerate cases $D_{0,m} = \varepsilon_m$ ($m \in S \subseteq W(A)$), ε_x denoting the measure giving mass 1 to x. Hewitt-Savage (1955) showed that for every $P \in S(A^{\mathbb{N}})$ there exists a unique $Q \in W(W(A))$, the set of all probability measures on $W(A)$ endowed with the usual σ-algebra, such that

$$P = \int_{W(A)} \bigotimes_{\mathbb{N}} \lambda\ Q(d\lambda). \tag{3.2}$$

Conversely, if $Q \in W(W(A))$ then the above P is in $S(A^{\mathbb{N}})$. Taking $Q = D_{\rho,m} \in \mathscr{D}_A$ we get $P_{\rho,m} \in SR(A^{\mathbb{N}})$,

$$P_{\rho,m} = \int_{W(A)} \bigotimes_{\mathbb{N}} \lambda\ D_{\rho,m}(d\lambda).$$

In the case $\rho = 0$ we have: $P_{0,m} = \bigotimes_{\mathbb{N}} m$. For the finite dimensional marginal distribution of $P = P_{\rho,m}$ with $\rho > 0$ and given $d \in A^*$, $d = d_1, ..., d_n$ we get:

$$P(d) = \int_{W(A)} \prod_{i=1}^{n} \lambda_{d_i} D_{\rho,m}(d\lambda) = D(\tilde{m}_d)/D(\tilde{m})$$

where $\tilde{m}_d = \tilde{m} + \varepsilon_d^*$, and $\varepsilon_d^* = \sum_{i=1}^{|d|} \varepsilon_{d_i}$. $\tilde{m}_d$ is the parameter measure of the Dirichlet distribution corresponding to P_d, since $P_d(e) = P(d, e)/P(d) = D(\tilde{m}_{d,e})/D(\tilde{m}_d)$ for $e \in A^*$. From the properties of the Γ-function we get in particular for $|d| = 1$

$$P(a) = \tilde{m}(a)/\|\tilde{m}\| = m(a).$$

Thus for $U \subseteq A$ we have

$$P_d(U) = \frac{\tilde{m}_d(U)}{\|\tilde{m}_d\|} = \frac{\tilde{m}(U) + \varepsilon_d^*(U)}{\|\tilde{m}\| + |d|},$$

so that $P = P_{\rho,m}$ obeys the following learning formula:

$$P_d(U) = \frac{m(U) + \rho \cdot \varepsilon_d^*(U)}{1 + \rho \cdot |d|} \qquad (U \subseteq A). \tag{3.3}$$

This is also fulfilled for $P=P_{0,m}$. Conversely, given ρ, m $(\rho>0, m>0)$, we may define $P' \in SR(A^{\mathbb{N}})$ as follows

(1) $P'_d(a)$ by (3.3) for $d \in A^*$, $a \in A$

(2) For the definition of $P'(d)$ use (2.1).

We see that $P=P'$, since $P(d)=P'(d)$ for every $d \in A^*$. Let $C \subseteq SR(A^{\mathbb{N}})$ denote the class of all such P obeying (3.3) with some $m>0$, $\rho>0$ (which is bijectively related to $\mathscr{D}_A$). Note that C is closed under learning and for $P=P_{\rho,m}$ and $\varphi \in \Gamma(A)$ we have $P^\varphi = P_{\rho,m^\varphi}$.

We will show that $C=CR(A^{\mathbb{N}})$ (as defined in section 2) provided that $|A| \geqslant 3$. The first part $C \subseteq CR(A^{\mathbb{N}})$ is straightforward and may be obtained for instance using (3.3) and (2.7). For the second part $CR(A^{\mathbb{N}}) \subseteq C$, we have to show that every LMI-measure fulfills (3.3) with some ρ, m. This was already derived from (2.8) by Johnson (1932) and Carnap (1958) in the special case (see section 2), allowing m only to be the equidistribution on A. As far as we know the proof for the general case was first given by Böge (1969). Stegmüller (1973), using Carnap's philosophical terminology, published independently a very complicated proof. The connection of (3.3) with the Dirichlet distributions was already known to Good (1965), but apparently not to Johnson, Carnap, and Stegmüller.

(3.4) PROPOSITION: *Let $|A| \geqslant 3$. Then $CR(A^{\mathbb{N}})$ coincides with the class C of measures defined by* (3.3) *with some $m>0$, $\rho \geqslant 0$.*

Proof. It remains to show that every $P \in CR(A^{\mathbb{N}})$ obeys (3.3). Let $P \in CR(A^{\mathbb{N}})$, and define $m(U)=P(U)$ for $U \subseteq A$. By $m_d (d \in A^*)$ we denote the measure on A induced by P_d. By symmetry of P, $m(x)\, m_x(y) = m(y)\, m_y(x)$. Thus $m_x(y)/m(y) = m_y(x)/m(x)$ is symmetric in (x, y). By (2.7), for $x \neq y$ the left hand side does not depend on x and the right hand side does not depend on y. We call this non-zero constant $(1+\rho)^{-1}$ (note ρ depends on P). Hence $m_a(x)=(1+\rho)^{-1}\, m(x)$ for $x \neq a$ and therefore $m_a = (1+\rho)^{-1}\,(m+\rho \cdot \varepsilon_a)$, since $m_a(A)=1$. (This is also true for $|A|=2$). Thus (3.3) is true for $n=1$. Let it be true for $n \geqslant 1$, that means $m_d \propto$ (proportional to) $m+\rho\varepsilon_d^*$ for $d=d_1,\ldots, d_n \in A^n$ with the same constant ρ. Applying our result for $n=1$ to P_d instead of P we get $m_{d,a} \propto m+\rho\varepsilon_d^* + \rho'\varepsilon_a$ where $\rho'=\rho'(d)$ possibly depends on d. Let $d'=d_1,\ldots, d_{n-1}$. Then

$$m+\rho\varepsilon_{d'}^* + \rho\varepsilon_{d_n} + \rho'(d)\,\varepsilon_a \propto m_{d,a} = m_{d',a,d_n}$$

$$\propto m+\rho\varepsilon_{d'}^* + \rho\varepsilon_a + \rho'(d', a)\,\varepsilon_{d_n}$$

If $|A| \geqslant 3$ and $a \neq d_n$, then $m+\rho\varepsilon_{d'}^* > 0$, ε_a, ε_{d_n} are linearly independent.

Comparing coefficients we get $\rho'(d)=\rho$ (and $\rho'(d',a)=\rho$). Thus for all $a\in A$

$$m_{d,a} \propto m+\rho\varepsilon_d^*+\rho\varepsilon_a=m+\rho\varepsilon_{d,a}^*.$$

Finally, a suitably chosen $d\in A^*$ shows, that ρ may not be smaller than zero.

This proves that $CR(A^{\mathbb{N}})$ is bijectively related to $\mathscr{D}_A$ and hence the Dirichlet distributions are characterized by means of the principle of Learn-Merge Invariance. In the case $|A|=2$ we have $CR(A^{\mathbb{N}})=SR(A^{\mathbb{N}})$, and hence (3.4) is not true.

For $\rho>0$ Carnap used $\lambda=\rho^{-1}=\|\tilde{m}\|$ instead of ρ as a parameter. λ has the nice property to become $\lambda'=\lambda+1$ after a further observation. Johnson (1932) used a parameter similar to ρ and interpreted it as the "weight" which is given to the observation relative to the prior guess. We prefer ρ, since with ρ it is easy to formulate (3.3) also in the degenerate case $\rho=0$.

4. The Case of a General Measurable Space

We first consider the case of a non regular measure on $A^{\mathbb{N}}$ with A finite, where we define the set of symmetric LMI-measures on $A^{\mathbb{N}}$ as the set $C(A^{\mathbb{N}})$ of all measures P for which a subset $A_+\subseteq A$ with $P\in CR(A_+^{\mathbb{N}})$ exists, namely $A_+=\{a\in A\colon P(a)>0\}$. Thus $CR(A^{\mathbb{N}})=C(A^{\mathbb{N}})\cap R(A^{\mathbb{N}})$, and, by prop. (3.4), $C(A^{\mathbb{N}})$ consists of the set $C'(A^{\mathbb{N}})$ of all measures P defined by (2.1) and (3.3) with arbitrary $\rho\geqslant 0$ and probability measure $m\geqslant 0$ on A, and the set of all other symmetric probability measures P on some $A_+^{\mathbb{N}}$ with $\varnothing\neq A_+\subseteq A$ and $|A_+|\leqslant 2$. $C'(A^{\mathbb{N}})$ will be called the set of all refinable symmetric LMI-measures, because it can also be defined as follows:

> $C'(A^{\mathbb{N}})$ consists of all measures which are equal to P^{φ} for arbitrary large finite B, some map $\varphi\colon B\to A$ and some $P\in CR(B^{\mathbb{N}})$.

From this last definition $C'(A^{\mathbb{N}})\subseteq C(A^{\mathbb{N}})$ follows, because for $P'\in C'(A^{\mathbb{N}})$ and $A_+=\{a\in A\colon P'(a)>0\}=\varphi(B)\subseteq A$, we have $P'\in C'R(A_+^{\mathbb{N}}):=C'(A_+^{\mathbb{N}})\cap R(A_+^{\mathbb{N}})\subseteq CR(A_+^{\mathbb{N}})\subseteq C(A^{\mathbb{N}})$ by (2.5). Moreover, using the results of the last section, since P fulfills (3.3) with some $\rho\geqslant 0$, $m>0$, it follows that P' fulfills (3.3) with this ρ and $m'=m^{\varphi}\geqslant 0$. We get all of these.

In the above definition we may equivalently replace "arbitrary large finite B" by "some finite B with $|B|\geqslant 3$". Further note:

$$C'R(A^{\mathbb{N}})=CR(A^{\mathbb{N}}) \qquad \text{for} \quad |A|\geqslant 3. \tag{4.1}$$

$$P^{\psi}\in C'((\psi A)^{\mathbb{N}}) \qquad \text{for} \quad P\in C'(A^{\mathbb{N}}),\ \psi\in\Gamma(A) \tag{4.2}$$

To cover the general case of an arbitrary set A (endowed with a σ-algebra) we define the set $C'(A^{\mathbb{N}})$ of refinable symmetric LMI-measures as the set of all measures P on $A^{\mathbb{N}}$ with $P^{\psi} \in C'((\psi A)^{\mathbb{N}})$ for all measurable ψ such that ψA is finite. In the finite case this reduces to the usual definition; $P \in C'(A^{\mathbb{N}})$ follows with $\psi = id_A$, the converse by (4.2). If m is induced on A by P, then P^{ψ} obeys (3.3) with $m' = m^{\psi} \geqslant 0$, where $\rho \geqslant 0$ does not depend on ψ. Thus for $\rho = 0$ $P = m^{\mathbb{N}}$ while for $\rho > 0$ P (using (3.2)) belongs to the distribution Q of the Dirichlet process with parameter measure $\tilde{m} = \rho^{-1} m$, as is well-known (see Ferguson (1973)). Conversely for each such Q the corresponding P belongs to $C'(A^{\mathbb{N}})$.

5. Outlook

It is shown that only the Dirichlet distributions are Learn-Merge invariant for all mergers. Therefore Dirichlet distributions appear as a satisfactory recommendation, if the categories are on a truly nominal level. However, in practice we are often faced with some structure on the observation set, e.g. a natural ordering of the categories, and a Dirichlet prior will not take account of any "proximity" relation. This has been critizised by several authors, c.f. Good (1965, 1967) Lindley (1970), and Lochner (1975). As an approach to this problem within the thinking of LMI, we may state the postulate (2.2) for such mergers, which are in some sense compatible with the structure of the observation categories. This might result in larger classes of priors fitted to the given structure. Some work has already been done in this direction, concerning ordering-structures (R. Tremmel (1971)), unweighted-undirected graphs and finite Cartesian products (J. Möcks/M. Stöckel (1972), Kursetz (1973)). We should note that no enlargement was achieved in the case of ordering-structures. The results just mentioned will be communicated in forthcoming papers.

Acknowledgment

The authors would like to thank Dr. W. Armbruster for helpful remarks improving the presentation of the paper.

Note added in proof. See also Zabell (1982) for further discussions on statistical issues of Johnson's and Carnap's philosophical system.

References

[1] Böge, W. (1969). *Statistikvorlesung.* Mimeo, Heidelberg.

[2] Carnap, R., and Stegmüller, W. (1958). *Induktive Logik und Wahrscheinlichkeit.* Springer-Verlag, Wien.

[3] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

[4] Good, I. J. (1965). *The Estimation of Probabilities.* MIT Press, Cambridge, Mass.

[5] Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B* **29**, 399–431.

[6] Hewitt, E., and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.

[7] Humburg, J. (1971). Die Problematik apriorischer Wahrscheinlichkeiten im System der Induktiven Logik von Rudolf Carnap. *Arch. Math. Logik Grundlag* **14**, 135–147.

[8] Johnson, W. E., and Braithwaite, R. B. (1932). Probability: Deductive and inductive problems (Appendix). *Mind* **41**, 421–423.

[9] Kursetz, R. (1973). *Carnapmaße auf dem cartesischen Produkt von zwei Alpabeten,* Diplomarbeit, Heidelberg.

[10] Lindley, D. (1970). *Bayesian Statistics - A Review.* Soc. Indus. Appl. Math., Philadelphia.

[11] Lochner, R. H. (1975). A generalized Dirichlet distribution in Bayesian life testing. *J. Roy. Statist. Soc. Ser. B* **37**, 103–113.

[12] Möcks, J., and Stöckel, M. (1972). *Carnapmaße über endliche cartesische Produkte,* Diplomarbeit, Heidelberg.

[13] Stegmüller, W. (1973). *Personelle und Statistische Wahrscheinlichkeit, Erster Halbband.* Teil II, Kap. 13, Ex: Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie, Bd. IV.

[14] Tremmel, R. (1971). *Rationale Lernprozesse und ihre Approximation durch endliche Automaten.* Diplomarbeit, Heidelberg.

[15] Wilks, S. (1962). *Mathematical Statistics.* Wiley, New York.

[16] Zabell, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *Ann. Statist.* **10**, 1091–1099.