FINITELY ADDITIVE PROBABILITIES AND PROPER "IMPROPER" PRIORS IN BAYESIAN STATISTICS

Romano Scozzafava[1]

University of North Carolina at Chapel Hill
Department of Statistics

AMS 1980 subject classifications: Primary 62A15; Secondary 60A05

Key Words and Phrases: Finitely additive conditional probabilities, vague initial knowledge, pseudodensity, noninformative priors, generalized Bayes theorem, marginalization paradox, coherence.

[1] On leave from Università di Roma, Facoltà di Ingegneria, Via A. Scarpa 10, 00161 Roma, Italy.

ABSTRACT

Improper densities constitute a patent violation of a fundamental axiom of probability theory ($P(\Omega) = 1$ if Ω is the certain event), so contradicting any aim at coherence when they are used in Bayesian inference. Finitely additive conditional probabilities allow a proper setting through the concept of *pseudodensity*, which is a point function, with no underlying measure, carrying a sort of "comparative probability". A pseudodensity does the same inferential job as an improper density, but contrary to the latter, it has all "civil rights", so avoiding rightful doubts on the correctness of conclusions. Bayes theorem is extended to pseudodensities and, as a by-product of this approach, further insights into invariance problems and the so-called "marginalization paradox" are given.

1. Introduction

Since the pioneering work of Bayes and Laplace, based on the principle of insufficient reason, many attempts have been made in Bayesian inference to define prior densities which add little or no information to that supplied by the likelihood through data. From this point of view, there is a strong intuitive tendency, given a statistical model, to regard as natural (if the parameter of interest has bounded range) the uniform distribution, notwithstanding its vulnerability to changes of the variable. Moreover, for unbounded ranges, improper prior densities are often used in the literature to express the lack of information: but this patent slighting on the probabilistic framework (the violation of the axiom giving probability equal to one to the certain event) clearly contradicts any aim at coherence, which is a typical feature of the Bayesian approach. (As a consequence, this argument has sometimes been used for challenging Bayesian statistics in general!).

The purpose of this paper is to show that, if we do not assume countable additivity as a postulate for probability (in accordance with the fact that any probabilistic and coherence argument leads only to the weaker requirement of finite additivity), the so-called "improper" (uniform or not) priors no longer deserve this attribute (as a synonym of "unsuitable"). Finite additivity allows, for example, a uniform distribution on any subset (bounded or not) or $\mathbb{R}$ and inconsistency is avoided. A typical situation is that of the uniform distribution on the set of integers, with zero *probability* to each point: but, since the aforementioned prior is zero everywhere, this should also be true, by Bayes theorem, for the posterior, which then looks like neither useful nor interesting. Is there a way out of it? For *any* distribution giving zero probability to such "elementary" events, we resort to a device, due to deFinetti (1936) and adopted also in

Scozzafava (1981a), to distinguish between any pair of them through a suitable (finitely additive) conditional probability. Then we introduce the concept of *pseudodensity*, a sort of "weaker" density not carrying a measure but only a kind of "comparative" probability. We prove that multiplying a prior pseudodensity by the likelihood always gives rise to a posterior which is, possibly, a density. In fact a pseudodensity does the same inferential job as an improper density, but contrary to the latter, the former has all "civil rights" in a proper probabilistic framework.

Moreover, as a by-product of this approach, we get a clearer insight into the problem of finding *invariant* prior distributions: in fact a pseudodensity is just a point function (while a density, proper or not, depends also on the underlying measure), and so a different parametrization has no effect on the "comparative" conditional probability. Yet, if it is possible looking on our posterior as a density, clearly the new parameters depend on the old ones in the usual way, through the Jacobian: but this is a question pertaining to elementary calculus and not to statistics!

In the final part of the paper we deal also with the so-called "marginalization paradoxes", giving a resolution in terms of our formulation.

Last (but not least) we lay stress on the fact that our approach to "improper" distributions resorts to extremely simple mathematical tools.

Preliminary versions of the contents of this article are in a contributed paper at the 14th European Meeting of Statisticians, Wroclaw (Scozzafava, 1981b) and in Scozzafava (1982a,b).

2. Vague Distributions

The word "noninformative" is suggestive of a prior (or, in general, any probability distribution) that is expressing not only a sort of "ignorance", but is also difficult (if not quite meaningless) to be determined with adequate accuracy.

Let $Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(m)})$ be a random vector (with values in $\mathbf{R}^m$): given $y = (y^{(1)}, y^{(2)}, \ldots, y^{(m)}) \in \mathbf{R}^m$, the event $\{Y=y\}$ will be also denoted simply by y or by $y^{(1)} \wedge y^{(2)} \wedge \ldots \wedge y^{(m)}$ and called *elementary event* (in $\mathbf{R}^m$).

Definition: The random vector Y has a *vague* distribution when:

(i) there exists on the set E of elementary events $\{Y=y\}$, with $y \in \mathbf{R}^m$, a *reflexive* and *transitive* relation $\cdot\!\!\geq$ such that $x, y \in E$ implies $x \cdot\!\!\geq y$ or $y \cdot\!\!\geq x$ (or both: in this case we write $x \sim y$ and x is said *equivalent* to y; on the other hand, $x \cdot\!\!> y$ means $x \cdot\!\!\geq y$ and $x \not\sim y$);

(ii) $y \cdot\!\!> v$ for any $y \in R_Y$ (the range of Y) and $v \notin R_Y$.

Since any $v \notin R_Y$ is identified with the *impossible event* $\emptyset$, (ii) can be written

$$y \cdot\!\!> \emptyset \quad \text{(for any } y \in R_Y\text{)}.$$

It is clear that to be vague is not a "demanding" condition for a distribution (by the way, any discrete or absolutely continuous *probability* distribution is vague, since the relation $\cdot\!\!\geq$ might be introduced by putting, for a discrete distribution with relevant probability P,

$$\text{(1)} \qquad x \cdot\!\!\geq y \quad \text{iff} \quad P(x) \geq P(y) \ ,$$

or, in the absolutely continuous case,

$$\text{(2)} \qquad x \cdot\!\!\geq y \quad \text{iff} \quad f(x) \geq f(y) \ ,$$

where f is the relevant density. Of course, our interests will not be particularly turned to such distributions).

We recall that a "comparative probability" of the kind dealt with by de Finetti (1931), Savage (1954) and many others (for an up-to-date reference, see Wakker (1981)) is a "stronger" structure, since it requires: the validity of (i) for *any* pair of events (and not only for the *elementary* ones), an axiom on the certain event Ω (such as $\Omega \cdot\!\!> \emptyset$) and the condition

$$A \dot{\geq} B \Longleftrightarrow A \vee E \dot{\geq} B \vee E$$

for events A,B,E with $A \wedge E = B \wedge E = \emptyset$. In fact, a comparative probability is the first step toward the assignment of a *numerical* probability P (under suitable conditions assuring the existence of almost uniform partitions, as in Savage (1954)); moreover, the probability P *almost agrees* with $\dot{\geq}$, i.e.

(3) $$A \dot{\geq} B \Longrightarrow P(A) \geq P(B) \quad ,$$

while much stronger conditions (Savage (1954), Wakker (1981)) are needed for the reverse implication (if both implications hold, P *agrees* with $\dot{\geq}$).

3. Diffuse Distributions and Conditional Equivalence.

If the probability distribution P of Y is known on the set $\mathcal{E}$ of elementary events to be

(4) $$P(y) = 0$$

for any $y \in \mathcal{E}$, we say that Y has a *diffuse* distribution. (This terminology is consistent, for a finitely additive P, with that of Dubins and Savage (1965, p.11)).

For example, an absolutely continuous distribution is obviously diffuse, but in general no other informations on P, apart from (4), need be available for a distribution to be diffuse.

In a finitely additive setting, according to theorems by Dubins (1975) and Krauss (1968), P can be extended (not uniquely) as a *full conditional probability* on the algebra $\mathcal{A}$ generated by the set $\mathcal{E}$ of elementary events: so P(A/B) exists for every $A,B \in \mathcal{A}$, with $B \neq \emptyset$. Thus a relation $\dot{\geq}$ can be introduced on the set $\mathcal{E}$ by putting,

(5) $$x \dot{\geq} y \Longleftrightarrow P(x/x \vee y) \geq P(y/x \vee y) \ .$$

Notice that (i), (ii) of Sect. 2 hold, i.e. it is possible to make vague (in many ways) a diffuse (finitely additive) distribution; moreover, since, for distinct $x,y \in E$,

(6) $$P(x/x \vee y) + P(y/x \vee y) = 1 \quad ,$$

the equality $P(x) = P(y) = 0$ *does not* imply that the same is true for the corresponding conditional probabilities: so this last equality does not mean $x \sim y$, which requires either $x = y$ or

(7) $$P(x/x \vee y) = P(y/x \vee y) = \tfrac{1}{2}$$

when $x \neq y$ (a sort of "second order" equiprobability).

To avoid the possible impression that the equivalence between two elementary events $x, y \in E$ induced by (5) has a limited significance (being relative to $x \vee y$), we remark that in fact, when a comparative probability (on the *whole* set of events) is considered in the relevant literature, a *conditional* equivalence $\sim_H$ (given an event $H \neq \emptyset$) is usually defined by putting (as in Savage (1954), p.44):

(8) $$A/H \sim_H B/H \iff A \wedge H \sim B \wedge H$$

(as follows from the corresponding condition on the relation $\succeq_H$). If we adopt the same definition (restricted to *elementary* events $x,y \in E$ and for any "conditioning" $H \neq \emptyset$), it follows in particular, for $H = x \vee y$, that $x,y \in E$ are *conditionally equivalent* (given H) if and only if they are *equivalent*.

This remark has some relevance to the final part of Sect. 6, containing a short discussion of pure likelihood methods.

4. Pseudodensity

Let Y be a random vector and P its probability distribution on the set E of elementary events (possibly $P(y) = 0$ for any $y \in E$). If $F \subset A$ is the set of *finite unions* of elements of E, a finitely additive conditional probability on $\mathcal{A} \times \mathcal{F}^\circ$ (with $\mathcal{F}^\circ = \mathcal{F}-\{\emptyset\}$) can be defined by putting, for $E \in \mathcal{A}$, $A \in \mathcal{F}^\circ$,

$$(9) \qquad P(E/A) = \frac{\sum\limits_{y \in E \wedge A} \alpha(y)}{\sum\limits_{y \in A} \alpha(y)} \quad ,$$

where α is a suitable non negative real function on $\mathcal{E}$, determined only up to a constant factor $k > 0$, with $\alpha(y) = 0$ iff $y = \emptyset$. We can define P on the whole $\mathcal{A} \times \mathcal{A}^\circ$ (though we will not need it outside $\mathcal{F} \times \mathcal{F}^\circ$) by putting, for $E \in \mathcal{A}$, $A \in \mathcal{G} = \mathcal{A}-\mathcal{F}$,

$$(9)' \qquad P(E/A) = \begin{cases} \dfrac{P(E \wedge A)}{P(A)} & , \text{ if } P(A) > 0 \\ 0 & , \text{ if } P(A) = 0,\ E \in \mathcal{F} \\ 1 & , \text{ if } P(A) = 0,\ E \in \mathcal{G}. \end{cases}$$

In particular, for $E = x \in \mathcal{E}$, $A = x \vee y \neq \emptyset$ and $x \neq y \in \mathcal{E}$, (9) gives

$$(9)'' \qquad P(x/x \vee y) = \frac{\alpha(x)}{\alpha(x) + \alpha(y)} \quad .$$

With this choice of P, relations (5), (7) become

$$(10) \qquad x \mathrel{\dot{\geq}} y \iff \alpha(x) \geq \alpha(y) \quad ,$$

$$(10)' \qquad x \sim y \iff \alpha(x) = \alpha(y) \quad ,$$

and, comparing with (2), we see that in the particular case of an absolutely continuous distribution the natural choice of α is $\alpha = kf$ (with arbitrary $k > 0$). Discrete distributions too can be embedded in this framework, since we can take $\alpha = kP$, as follows from comparison between (10) and (1): notice that (9) becomes in this case the usual formula for conditional probabilities.

In general, a vector Y with a *vague* distribution can be given by a non negative real function α on $\mathcal{E}$, determined only up to a constant factor $k > 0$, with $\alpha(y) = 0$ iff $y = \emptyset$ and *such that* (10) *holds*; this function will be called *pseudodensity* of Y.

Example - Given $E \subseteq \mathbb{R}^m$, we say that the distribution of a random vector Y is *uniform* on E when Y is diffuse and (7) holds for distinct $x,y \in E$ and for all (finitely additive) extensions of P to $A \times A°$. Taking in particular (9), (9)' as conditional probability P, comparison between (9)" and (7) shows that such uniform distribution corresponds to a *constant* (on E) pseudodensity, i.e.

$$(11) \qquad \alpha(x) = \begin{cases} k, & \text{for } x \in E \\ 0, & \text{for } x \notin E \end{cases}$$

with arbitrary $k > 0$. Obviously, if E is (Peano-Jordan) measurable, with finite measure $\mu(E) \neq 0$, we can take $k=1/\mu(E)$ in (11) and the pseudodensity α is a density.

A real (random) number X in [0,1] can be looked on as obtained (in terms of its ternary expansion) by sampling with replacement from an urn containing three balls numbered 0,1,2. But, if the ball 1 is missing, we get X in the Cantor ternary set E, and it is natural to think about a uniform distribution on E: this vague distribution with pseudodensity (11) is compatible with the usual diffuse (not absolutely continuous) distribution determined on [0,1] by Cantor distribution function.

Thanks to the concept of pseudodensity, the ratio of probabilities of *any* two elementary events $x \neq y$ always makes sense, since (9)" allows to *define*

$$(12) \qquad \frac{P(x)}{P(y)} = \frac{P(x/x \vee y)}{P(y/x \vee y)} = \frac{\alpha(x)}{\alpha(y)} \quad .$$

Notice that when the density f of an absolutely continuous distribution is used to define a pseudodensity $\alpha = kf$, we look on f just as a point function, irrespective of its underlying measure: the function α has the only task to "put in order" elementary events through relation (10).

Example of diffuse distributions can also be exhibited by taking as pseudodensities the usual improper distributions ("even more" as point functions: integrals forbidden!).

The need for the use of improper *densities* in the literature (when no proper distribution can express the relevant state of information on Y) depends on the usual assumption concerning the (σ-additive) probability measure P to be dominated by a σ-additive measure (usually, Lebesgue's) and on the ensuing (compulsory!) existence of a density: so it is not easy, at this point, to convey to such a distribution any meaningful idea of being, in some sense, "non informative", since not much freedom (since the very beginning!) is left, like in a "Procrustean bed" ... (de Finetti (1972), p. 201). For example, the use of an improper *density* like $\alpha(y) = 1$ implies that $|y|$ is almost certainly enormously large (while the data will generally imply the contrary!): for the corresponding pseudodensity such a ridiculous situation is clearly avoided, since $\alpha(y)$ is a point function (with no underlying measure) whose only task is to make valid equation (7) for any pair of elementary events $x \neq y$.

5. Finite Additivity and Coherence

According to classical results of deFinetti (in many of his papers: see for example deFinetti (1972), pp. 76-79), *any* finitely additive probability distribution is *coherent* (i.e., there is no possibility of being made a sure loser). For *conditional* distributions (for example, a posterior $\pi(\theta/x)$ in Bayesian inference) coherence amounts to the requirement that, for any *given* x, $\pi(\theta/x)$ is a finitely additive probability distribution (These are essentially the conclusions reached by deFinetti in the 1930s; see also Cornfield (1969), p. 625). As it is well-known, Bayes' theorem turns out to be a necessary and sufficient condition for coherence.

A few remarks seem now pertinent. The well-known definition of coherence given by Heath and Sudderth (1978) requires the absence of a combination of bets which results in a uniformly positive *expected* (with respect to the likelihood) loss: so it is not difficult to envisage situations in which a distribution may be HS-coherent but not dF-coherent: in fact, dF-coherence requires, for all given x, the absence of a combination C_x of bets which results in a uniformly positive loss $L(\theta,x)$, while HS-coherence requires the absence of a combination C of bets which results in a uniformly positive expected loss $\mathbb{E}_x(L(\theta,x))$. Obviously, the latter condition may well hold true even if, for *some* x, deFinetti's condition fails. In this paper, by "coherence" we always intend dF-coherence.

We stress that in a probabilistic framework a pseudodensity is not a stranger, as an improper *density* is, but just a "weaker" notion. In fact, if "informations" on the random vector Y are not "vague", a pseudodensity may be (or become, through Bayes theorem: see next section) a *probability* density; for example, a (Riemann) *integrable* α (with non-zero integral) singles out a density $k\alpha$, with respect to the relevant measure in $\mathbb{R}^m$, with $k = (\int_{R_Y} \alpha dy)^{-1}$: therefore it is certainly coherent. Notice that the integral of a pseudodensity may also be equal to zero (an example is the Cantor distribution or, more generally, any pseudodensity which is positive only on a set of zero measure), or it may be equal to $+\infty$, as in the case of the usual improper *densities*: but, contrary to the latter, α *cannot* be said incoherent, since it defines (as recalled in Section 2) a "weak comparative probability" through a finitely additive conditional probability. Moreover, it is a "candidate" to become a density (for example, going on with sampling, as discussed in Section 6).

As a last remark, notice that the adoption, as a general norm, of finite additivity does not prevent taking into account (when this turns out to be suitable) distributions that are (possibly with respect to some particular subfamily of events) countably additive: what is essential is that the latter property be not seen as a characteristic of *any* distribution.

6. Generalized Bayes Theorem

Let the possible "observations" be represented by a random vector $X = (X_1, X_2, \ldots, X_n)$, whose pseudodensity (possibly a density, in particular a probability) is given and depends on some parameters $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(r)}$: in other words, we have a family of (sampling) pseudodensities $p(x/\theta)$ indexed by $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(r)})$. The latter too is represented by a random vector Θ, with pseudodensity $\pi(\theta)$.

In the two particular cases in which X and Θ are both discrete or both absolutely continuous, the product

$$(13) \qquad \alpha(\theta,x) = \pi(\theta)p(x/\theta)$$

is, respectively, the joint probability distribution or the joint density of the vector (Θ,X). We show that (13) is the natural *joint pseudodensity* to be ascribed to the vector (Θ,X) in the general case.

In fact, according to the meaning of $p(x/\theta)$ (analogous to that of a "nonconditional" pseudodensity, conveyed in (10)'), given two *conditional* elementary events x/θ and y/θ on the same "line" $\Theta=\theta$, we have

$$(14) \qquad x/\theta \sim y/\theta \iff p(x/\theta) = p(y/\theta) \ .$$

The natural extension of (14) to *any* pair of conditional elementary events x/θ_1 and y/θ_2 is

$$(14)' \qquad x/\theta_1 \sim y/\theta_2 \iff p(x/\theta_1) = p(y/\theta_2) \ .$$

So the ratio between the corresponding conditional probabilities (possibly they are equal to zero) can be defined, as in (12), by

$$(12)' \qquad \frac{P(x/\theta_1)}{P(y/\theta_2)} = \frac{p(x/\theta_1)}{p(y/\theta_2)} \quad .$$

It follows easily

$$\frac{\pi(\theta_1)}{\pi(\theta_2)} \; \frac{p(x/\theta_1)}{p(y/\theta_2)} = \frac{P(\theta_1)}{P(\theta_2)} \; \frac{P(x/\theta_1)}{P(y/\theta_2)} = \frac{P(\theta_1 \wedge x)}{P(\theta_2 \wedge y)} = \frac{\alpha(\theta_1,x)}{\alpha(\theta_2,y)} \quad ,$$

where α denotes the joint pseudodensity of (Θ,X). Therefore, as claimed, we have $\alpha(\theta,x) = \pi(\theta)p(x/\theta)$.

Now, given an observation $x \in \mathbb{R}^n$, to assign accordingly a conditional pseudodensity $\pi(\theta/x)$ to $\Theta/\{X = x\}$ we start by exchanging the roles of Θ and X in (14), i.e.

$$\theta_1/x \sim \theta_2/x \iff \pi(\theta_1/x) = \pi(\theta_2/x).$$

Then (8) and (13) give

$$\theta_1/x \sim \theta_2/x \iff \theta_1 \wedge x \sim \theta_2 \wedge y \iff \pi(\theta_1)p(x/\theta_1) = \pi(\theta_2)p(x/\theta_2) \quad ,$$

and so a consistent assignment of the sought pseudodensity of $\Theta/\{X=x\}$ is $\pi(\theta)p(x/\theta)$ *for fixed* x. Since a pseudodensity is determined only up to a constant (with respect to θ) factor $K(x) > 0$, it follows

$$(15) \qquad \pi(\theta/x) = K(x)\pi(\theta)p(x/\theta) \quad ,$$

which is *Bayes theorem* for pseudodensities.

In particular, if $\pi(\theta)p(x/\theta)$ is (as function of θ) integrable, then the posterior $\pi(\theta/x)$ can be interpreted, for

$$K(x) = \left(\int_{\mathcal{R}_\Theta} \pi(\theta)p(x/\theta)d\theta\right)^{-1} \quad ,$$

as a density, and (15) takes the form of the usual Bayes theorem.

If the prior is *uniform*, given any two elementary events one has, obviously,

$$\frac{\pi(\theta_1)}{\pi(\theta_2)} = 1 \ ,$$

and from (15) it follows that the corresponding ratio of posteriors is

$$\frac{\pi(\theta_1/x)}{\pi(\theta_2/x)} = \frac{p(x/\theta_1)}{p(x/\theta_2)} \ ,$$

i.e. it equals the *likelihood ratio*. For example, in the case of a *diffuse* distribution, both prior and posterior *probabilities* are zero everywhere, yet the prior ones are of the same "order", while it is up to the likelihood to compare posterior "orders".

An interesting feature in the use of (15) is that for all the most common statistical models the likelihood $p(x/\theta)$ (which is a density in x, for fixed θ) is *integrable in* θ too (as it is easy to check: anyway, this is shown in Scozzafava (1982a), Section 14): so the uniform distribution with pseudodensity $\pi(\theta) = 1$ (irrespective of the range of Θ) gives rise to a posterior pseudodensity which actually singles out a density, i.e. the standardized likelihood

$$\pi(\theta/x) = \frac{p(x/\theta)}{\int_{R_\Theta} p(x/\theta)d\theta} \ .$$

So under this condition (integrability of the likelihood with respect to θ) it is not meaningless, for example, to speak (for $r > 1$) of "marginal" likelihoods.

The usefulness of a posterior $\pi(\theta/x)$ being interpreted as a density lies in the possibility of evaluating probabilities of relevant subsets of R_Θ, i.e. the so-called "composite hypotheses": but often the best way to convey to the experimenter what the data tell him about Θ is to draw a graph of the posterior distribution, and this procedure is still significant when $\pi(\theta/x)$ is only a pseudodensity. Anyway, one can start again sampling (taking now

$\pi(\theta/x)$ as prior pseudodensity) in order to make more "informative" this vague distribution (recall our discussion in Section 5).

Yet the procedure of singling out a *density* from the posterior can be applied *in most known cases* of improper distributions, looked on as prior pseudodensities (a list of some improper distributions together with an interesting criterion for their choice is in Piccinato (1977); Bernardo (1979) is a relevant reference as well). Though an improper density does the same operational job as a pseudodensity, the slighting on the probabilistic framework contained in the former raises rightful doubts on the correctness of the results, that are often accepted as sensible conclusions.

The fact that, in the case of a *uniform* prior (and only in this case), our posterior essentially coincides with the likelihood, allows a partial reconciliation (from a practical point of view) with some aspects of R.A. Fisher's approach to inference; but notice that, since we are working in a more general setting through the concept of *vague* distribution, the situation can be seen as analogous to the following well-known "reconciliation" in mathematics: while it is forbidden to write $\sqrt{-1}$ in the real number system, nevertheless it is possible to write $\sqrt{z}$, where z is any complex number, for example z = -1. On the other hand, our approach cannot undergo the usual criticism levelled at pure likelihood methods, since we *always* remain in a probabilistic framework and *we do not acknowledge any privileged status to the uniform distribution.*

Moreover, when $\pi(\theta/x)$ is only a pseudodensity (a case not often met in practice), it makes complete (and not only pairwise) comparison among elementary events, as already noticed in the last remarks of Section 3.

Different aspects of the use of finitely additive distributions as priors in Bayesian inference are considered in an interesting paper by Hill (1980). A vague suggestion for a finitely additive approach in this context is also in Pratt (1976). Some relevant theory is in Heath and Sudderth (1978): they give conditions for rescuing only those improper priors yielding a posterior which could also have been obtained from a proper finitely additive prior (recall that their definition of coherence is different from ours).

7. Invariance and Noninformative Priors

The real meaning of noninformative priors in Bayesian inference and the relevant (through consideration of the structure of the problem) notion of invariance are controversial matters in the literature on these subjects. We look on invariance taking into account no features of the mathematical structure of the problem other than those expressible through the notion of pseudodensity: actually, measure theoretic ingredients are often a useful mathematical tool, yet not always conveying a proper statistical meaning.

Let Θ be a random vector with pseudodensity $\pi(\theta)$, and $\gamma = \psi(\theta)$ a one-to-one transformation of the range R_Θ of Θ. The vector $\Gamma = \psi(\Theta)$ refers to the same situation, with a different parametrization: so, given any two (distinct) elementary events

$$\gamma_1 = \psi(\theta_1) \ , \qquad \gamma_2 = \psi(\theta_2) \ ,$$

a natural requirement is

$$\frac{P(\gamma_1/\gamma_1 \vee \gamma_2)}{P(\gamma_2/\gamma_1 \vee \gamma_2)} = \frac{P(\theta_1/\theta_1 \vee \theta_2)}{P(\theta_2/\theta_1 \vee \theta_2)} \ , \tag{16}$$

since we are dealing with (a different representation of) the very *same* events. It follows, denoting by $\pi_0(\gamma)$ the pseudodensity of the random vector Γ,

$$\frac{\pi_0(\psi(\theta_1))}{\pi_0(\psi(\theta_2))} = \frac{\pi(\theta_1)}{\pi(\theta_2)}$$

and hence, for all $\theta \in R_\Theta$,

$$(17) \qquad \pi_0(\gamma) = \pi_0(\psi(\theta)) = K_0\pi(\theta)$$

for any $K_0 > 0$.

Let R_Θ represent the parameter space in Bayesian inference: we shall refer, for the sake of simplicity, to random numbers (one-dimensional random vectors), but no essential modifications are needed for the general case. If $\pi(\theta)$ is the relevant prior, we get for the posterior corresponding to the parameter γ (notations as in Section 6), taking into account (17):

$$(17)' \qquad \begin{aligned} \pi_0(\gamma/x) &= K(x)\pi_0(\gamma)p(x/\psi^{-1}(\gamma)) = \\ &= K(x)K_0(x)\pi(\theta)p(x/\theta) = K_0(x)\pi(\theta/x) \ , \end{aligned}$$

so that relation (17) still holds for the two posteriors. In particular, $\pi_0(\gamma/x)$ could be a density (even if $\pi_0(\gamma)$ is not, as noted in Section 6), and so in this case (assuming differentiability of ψ)

$$(17)'' \qquad 1 = \int_{\mathbb{R}} \pi_0(\gamma/x)d\gamma = \int_{\mathbb{R}} \pi_0(\psi(\theta)/x)\psi'(\theta)d\theta = K_0(x)\int_{\mathbb{R}} \pi(\theta/x)\psi'(\theta)d\theta \ .$$

We claim that these elementary remarks pertain to calculus and not to statistics: measure-theoretic aspects should not be the protagonists on the scene of inference, since the differences between formulas (17)' and (17)" are due to the fact (which has only an obvious mathematical relevance) that the underlying measures of $\pi_0(\gamma/x)$ and $\pi(\theta/x)$ are different.

Anyway, coming back to equation (17), let us consider a *family* Φ of one-to-one tranformations ψ of R_Θ: condition (16) entails that (17) must hold for *all* ψ in the family Φ, and so if Φ contains the *identity* transformation, it follows

$$(18) \qquad \pi_0(\theta) = K\pi(\theta) \ .$$

(We denote, here and in the sequel, by the same letter K (sometimes K_1) any constant (with respect to θ) positive factor). The above equation (18) holds for any $\theta \in R_\Theta$; therefore, for any $\psi \in \Phi$,

$$\pi_0(\psi(\theta)) = K\pi(\psi(\theta)),$$

and using this in (17) gives

$$\pi(\psi(\theta)) = K\pi(\theta) \tag{19}$$

for any $\theta \in R_\Theta$ and $\psi \in \Phi$.

A parallel equation for improper densities is given, in the particular case $\psi(\theta) = \theta+c$, in Berger (1980, p. 73), resorting to partly similar motivations.

Notice that (19) is essentially a consequence of (16): but if we want to convey in $\pi(\theta)$ the idea of being "noninformative", condition (16) seems too weak, since invariance requirements do not relate specifically to "ignorance".

In fact, $\pi(\theta)$ could be said *noninformative* when, for *any* one-to-one transformation ψ of R_Θ and for any $\theta \in R_\Theta$,

$$\theta \sim \psi(\theta) \ , \tag{20}$$

where $\sim$ has the usual meaning (7), (10)'. It follows

$$P(\theta/\theta \vee \psi(\theta)) = P(\psi(\theta)/\theta \vee \psi(\theta)) = \tfrac{1}{2} \ ,$$

that is

$$\pi(\theta) = \pi(\psi(\theta)) \ . \tag{21}$$

This condition appears as a particular case of (i.e. stronger than) equation (19), requiring the choice of just one value for K (namely, K = 1) and its validity for *any* one-to-one ψ. So the conflict between the two objectives involved in generating noninformative priors, namely expressing ignorance and obtaining distributions with some invariance properties (clearly stressed in Zellner (1973), quoting Jeffreys), is avoided.

The dimension of Θ played no role in deducing (21), whose only solution is (for any $K_1 > 0$)

(22) $$\pi(\theta) = K_1 \ .$$

Then a uniform pseudodensity "can" always (in multiparameter problems too) be taken as a noninformative prior: actually, if one does not dislike to define noninformative priors through condition (20), not only it "can" be chosen, but really it "must" (by the way, (20) could be relaxed requiring its validity only for suitable families Φ, as in the study of invariance).

Anyhow, the uniform pseudodensity (22) comes to the fore being a solution (for K=1) of equation (19). Coming back to this equation, we consider now three classical families of transformations, corresponding respectively to location, scale and location-scale parameters.

Let $R_\Theta = \mathbb{R}$. If

(23) $$\psi(\theta) = \theta+c \quad \text{(for any real c)} \ ,$$

substituting (23) into (19) gives

(19)' $$\pi(\theta+c) = K\pi(\theta) \ .$$

It is not difficult to see that equation (19)', besides the uniform pseudodensity (22) (in accordance with well-known invariance results for a location parameter), has many other solutions (cfr. Berger (1980), p. 73), for example (if $K = e^{-\lambda c}$, $\lambda > 0$)

(24) $$\pi(\theta) = K_1\lambda e^{-\lambda\theta}$$

for any $K_1 > 0$.

If $R_\Theta = \mathbb{R}^+$ and

(25) $$\psi(\theta) = c\theta \quad (\text{for any } c > 0) \ ,$$

equation (19) becomes

(19)'' $$\pi(c\theta) = K\pi(\theta) \ .$$

A solution of (19)'' (in keeping with usual invariance prescriptions for a scale parameter) is, if $K = 1/c$,

(26) $$\pi(\theta) = K_1/\theta \ .$$

Other solutions (choosing $K = c^\alpha$) are

(27) $$\pi(\theta) = K_1\theta^\alpha \quad (\text{for any } \alpha \in \mathbb{R})$$

(the *uniform* distribution corresponding to $\alpha = 0$).

If $R_\Theta = \mathbb{R} \times \mathbb{R}^+$ and $\psi(\theta,\sigma) = (\theta+c, b\sigma)$, with $c \in \mathbb{R}$, $b \in \mathbb{R}^+$, equation (19) becomes

(19)''' $$\pi(\theta+c, b\sigma) = K\pi(\theta,\sigma) \ ,$$

and so a class of solutions is

(28) $$\pi(\theta,\sigma) = K_1\sigma^\alpha \qquad (\alpha \in \mathbb{R}) \ .$$

A last example, for $R_\Theta = [0,1]$, is obtained by considering the particular transformation (besides the identity one) $\psi(\theta) = 1-\theta$ (symmetry with respect to $\theta = \frac{1}{2}$). Equation (19) then takes the form

$$\pi(1-\theta) = K\pi(\theta) \quad ,$$

which is satisfied for $K = 1$ by

$$\pi(\theta) = K_1[\theta(1-\theta)]^\alpha \quad (\text{for any } \alpha \in \mathbb{R}) \ .$$

So (21) too holds for this ψ. Widely used "ignorance" priors in this class are those corresponding to $\alpha = -\frac{1}{2}$ and $\alpha = -1$.

Notice that compelling since the beginning the prior $\pi(\theta)$ to be a *density*, we would get (assuming differentiability of ψ)

(29) $$\pi(\psi(\theta))\psi'(\theta) = \pi(\theta)$$

instead of (19). For example, if the family Φ is that given by (23), it follows from (29)

(30) $$\pi(\theta+c) = \pi(\theta) \quad ,$$

whose *only* solution is $\pi(\theta) = K$: then solutions like (24) are lost and, moreover, the uniform distribution is not so legitimate as (22), since it is necessarily, as a density, nonzero only on a set of *finite* measure, so contradicting (30).

Similarly, for the set of transformations $\psi(\theta) = c\theta$, equation (29) gives

(30)' $$\pi(c\theta)\cdot c = \pi(\theta) \quad ,$$

which is satisfied (formally) by (27) only for $\alpha = -1$; but to avoid a contradiction with the assumption that $\pi(\theta)$ is a probability density, we should obviously restrict the range of Θ to a suitable (bounded) set (and so (30)' is no longer valid). Such "ad hoc" procedure, taking a bounded interval (a,b) and then the limits for $a \to 0$, $b \to \infty$, is successfully adopted, for example, in Berger (1980, pp. 71-72) for the "table entry" first digit problem (yet a simple approach to this problem is possible also in terms of finitely additive probability measures, as in Scozzafava (1981a)).

8. Marginalization "Paradoxes" and Sufficient Statistics

Given, as in Section 6, a random vector $X = (X_1,X_2,\ldots,X_n)$ with pseudo-density $p(x/\theta)$, let $T = t(X)$ be a *statistic* and

$$A_t = \{x \in R_X : t(x) = t\} \quad , \quad t \in R_T \quad ,$$

the elements of the corresponding partition of R_X. Clearly, any A_t is an elementary event relative to the vector T; so, denoting A_t simply by t, it is meaningful to speak of the (conditional) pseudodensity $p_0(t/\theta)$ of T (given θ). In particular, if $p(x/\theta)$ is, for any θ, a *density* in x and there exists the density $p_0(t/\theta)$, the procedure for its determination from $p(x/\theta)$ is standard.

Now, let us define (as in Section 6) the pseudodensity $\pi(\theta/x)$ and, similarly (if the pseudodensity $p_0(t/\theta)$ of $T/\{\Theta = \theta\}$ is given), the pseudodensity $\pi(\theta/t)$ (by conditioning to $\{T = t\}$ in place of $\{X = x\}$). We shall say that T is *sufficient* for Θ when, for any x and for *all* prior pseudodensities $\pi(\theta)$,

$$(31) \qquad \pi(\theta/x) = k(x)\pi(\theta/t) \quad ,$$

with $t = t(x)$ and a suitable $k(x)$ (equal to 1 when the posteriors are densities).

It is not difficult to prove the following "one-way" generalized Neyman factorization theorem: if T is sufficient for Θ, then there exist two functions g (on $R_T \times R_\Theta$) and h (on R_X) such that the likelihood $p(x/\theta)$ can be written

$$(32) \qquad p(x/\theta) = g(t(x),\theta)h(x) \ .$$

In fact, by the generalized Bayes theorem (15), from (31) we get

$$K(x)p(x/\theta) = k(x)K_0(t(x))p_0(t(x)/\theta),$$

and so (32) holds with

$$g(t(x),\theta) = p_0(t(x)/\theta), \quad h(x) = k(x)K_0(x)/K(x) \ .$$

Actually, the function g is the pseudodensity (or, in particular, the density) of the vector $T/\{\Theta=\theta\}$.

It follows that, if $p(x/\theta)$ is *not* of the form

$$(32)' \qquad p(x/\theta) = p_0(t(x)/\theta)h(x) \quad ,$$

equality (31) does not hold (for any prior, as it is easily seen). This remark is the starting point for a simple resolution of the so-called marginalization paradox, which has its source in the use of non-sufficient statistic: this aspect has not been considered by other authors (and they were right, since improper prior *densities* are involved in the paradox, while as yet Bayes and factorization theorems were known only for ordinary, proper densities).

We illustrate the situation presenting in our own notation one of the examples given in Dawid, Stone and Zidek (1973).

Data (x,y) have been obtained which give rise to the likelihood

$$p(x,y/\theta,\gamma) = \theta\gamma^2 e^{-\gamma(\theta x+y)} \quad , \tag{33}$$

and to make inference about the vector parameter $(\theta,\gamma) \in \mathbb{R}^+ \times \mathbb{R}^+$ the prior

$$\pi(\theta,\gamma) = f(\theta) \tag{34}$$

is chosen by a statistician B_1, where $f(\theta)$ is a proper density in θ (notice $\pi(\theta,\gamma)$ is improper, i.e. a "strict" pseudodensity in our setting). Then, by Bayes theorem

$$\pi(\theta,\gamma/x,y) = Kf(\theta)\theta\gamma^2 e^{-\gamma(\theta x+y)} \quad . \tag{35}$$

But suppose B_1 is interested only in making inference about θ: then, integrating out γ gives easily the marginal in θ

$$\pi(\theta/x,y) = \frac{K\theta f(\theta)}{(\theta x+y)^3} \quad , \tag{36}$$

which could be written, putting $z = y/x$, $k = K/x^3$,

$$\pi(\theta/x,y) = \frac{k\theta f(\theta)}{(\theta+z)^3} \quad . \tag{37}$$

At this point, a statistician B_2 notices that the posterior distribution of θ is proportional to a function of the statistic z only: therefore he goes back to (33) to find the probability density of z, obtaining, by straightforward calculations

$$(38) \qquad p_0(z/\theta,\gamma) = \frac{\theta}{(\theta+z)^2} \quad ,$$

i.e. $p_0(z/\theta,\gamma)$ is a function of θ only. Now, taking the (proper) prior

$$(34)' \qquad \pi(\theta) = f(\theta) \ ,$$

where $f(\theta)$ is the same as in (34), B_2 gets

$$(37)' \qquad \pi(\theta/z) = \frac{k_0\theta f(\theta)}{(\theta+z)^2}$$

which is not a multiple of (37). This is (or "should be") the marginalization paradox.

But notice:

(i) the factorization property (32)' is not enjoyed by the likelihood (33), as it is clear from formula (38);

(ii) then for any prior $\pi(\theta,\gamma)$

$$(39) \qquad \pi(\theta,\gamma/x,y) \neq k\pi(\theta,\gamma/z) \ ;$$

(iii) integrating out γ (whose "fortuitous absence" in the right hand side has no mathematical relevance) cannot, in general, turn (39) into an equality.

Apart from this resolution based on (i), (ii), (iii), i.e. on the generalized factorization theorem, we essentially agree with Jaynes (1980) (notwithstanding our quite different formal approach to the problem) that the discrepancy between (37) and (37)' arises not from any defect of improper priors (which, on the other hand, are "legitimized" in our setting), but from a rather suble failure of B_2 to take into account all the relevant information when z is used in place of (x,y). In fact, z is not a sufficient statistic. Another relevant paper is that by Regazzini

(1982), where an improper prior and the ensuing marginalization paradox are justified via Rényi theory of generalized densities.

The remark (already made by Dawid, Stone and Zidek) that the paradox can be avoided if B_1 uses, instead of (34), the prior

$$\pi(\theta,\gamma) = \frac{f(\theta)}{\gamma} \quad , \tag{40}$$

does not conflict with our argument, since the two different sides of (39), obtained through the use of a non-sufficient statistic, correspond to a *same* prior. With regard to their proof (see also Sudderth (1980)) that, if B_1 employs a *proper* prior $\pi_1(\theta,\gamma)$ and B_2 its marginal $\pi_2(\theta)$, the paradox does not arise, it does not conflict with our argument as well, for similar reasons. So this last result could be re-read as follows: since z is not a sufficient statistic (because (32) does not hold), B_2 is not actually using all available data, and so he cannot get the same posterior as B_1 by adopting the same prior $\pi_1(\theta,\gamma)$; yet B_2 can restore the situation carrying out a suitable choice of a different prior, namely (in this particular case where the density of z is a function of θ only) the marginal $\pi_2(\theta)$ of $\pi_1(\theta,\gamma)$.

As a matter of fact, this is not a peculiar merit of proper distributions, as the prior (40) reveals. Moreover, notice that a *proper* prior like

$$\pi(\theta,\gamma) = f(\theta)\lambda e^{-\gamma\lambda} \tag{41}$$

gives, for $\lambda = y$ (recall that x,y are the observed data), instead of (35) and (36),

$$\pi(\theta,\gamma/x,y) = Kf(\theta)\theta\gamma^2 y\ e^{-\gamma(\theta x+2y)} \quad ,$$

$$\pi(\theta/x,y) = \frac{K\theta f(\theta)}{(\theta x+2y)^3} \quad ,$$

and so (37) becomes

$$\pi(\theta/x,y) = \frac{k\theta f(\theta)}{(\theta+2z)^3} \quad ,$$

that no longer agrees with B_2's solution (37)' obtained using as prior the marginal f(θ) of (41). Clearly, the choice of the prior (41) is beyond reproach from a "syntactic" point of view (it could have been made by a third statistician B_3 arriving later than B_1 and B_2 on the scene of inference!). It is not so, in general, from a "semantic" point of view, but priors depending more or less on sample data (for example, through the likelihood) are not rare in the literature (we just mention the well-known locally uniform priors of Box and Tiao (1973) and the operational reference priors of Bernardo (1979), who also deals with the marginalization paradox).

REFERENCES

BERGER, J.O. (1980). *Statistical Decision Theory.* Springer Verlag, New York.

BERNARDO, J.M. (1979). Reference Posterior Distributions for Bayesian Inference. *J. Roy. Statist. Soc. Ser. B, 41*, 113-128.

BOX, G.E.P. and TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics, 25*, 617-657.

DAWID, A.P., STONE, M. and ZIDEK, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B., 35*, 189-213.

DE FINETTI, B. (1931). Sul significato soggettivo della probabilità. *Fund. Math., 17*, 298-329.

DE FINETTI, B. (1936). Les probabilités nulles. *Bull. Sci. Math., 60*, 275-288.

DE FINETTI, B. (1972). *Probability, Induction, Statistics.* Wiley, New York.

DUBINS, L.E. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *Ann. Probab., 3*, 89-99.

DUBINS, L.E. and SAVAGE, L.J. (1965). *How To Gamble If You Must.* McGraw-Hill, New York.

HEATH, D. and SUDDERTH, W.D. (1978). On finitely additive priors, coherence and extended admissibility. *Ann. Statist., 6*, 333-345.

HILL, B.M. (1980). On some statistical paradoxes and non-conglomerability. *Bayesian Statistics* (Proc. Intern. Meeting, Valencia, 1979, Eds. Bernardo, J.M., De Groot, M.H., Lindley, D.V. and Smith, A.F.M.) 39-49. University Press, Valencia.

JAYNES, E.T. (1980). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics* (Ed. A. Zellner) 43-78. North-Holland, Amsterdam.

KRAUSS, P.H. (1968). Representation of conditional probability measures on Boolean algebras. *Acta Math. Acad. Sci. Hungar, 19*, 229-241.

PICCINATO, L. (1977). Predictive distributions and noninformative priors. *Trans. 7th Prague Conf. and 1974 European Meeting of Statisticians,* 399-407. Reidel, Dordrecht.

PRATT, J.W. (1976). Comment to Stone, M.: Strong inconsistency from uniform priors. *J. Amer. Statist. Ass., 71,* 119-120.

REGAZZINI, E. (1982). Considerazioni intorno ad un particolare paradosso di marginalizzazione. *Scritti in onore di I. Gasparini,* Vol. II, 869-879. Giuffrè, Milano.

SAVAGE, L.J. (1954). *The Foundations of Statistics.* Wiley, New York (2nd Ed., Dover, New York, 1972).

SCOZZAFAVA, R. (1981a). Non-conglomerability and the first digit problem. *Statistica, 41,* 561-565.

SCOZZAFAVA, R. (1981b). Improper priors and finite additivity (Abstract). *14th Europ. Meeting of Statist.*, Wroclaw, 271-272.

SCOZZAFAVA, R. (1982a). Il ruolo della probabilità comparativa, finitamente additiva, nella statistica bayesiana. *Pubbl. Ist. Mat. Appl. Univ. Roma,* n. 246.

SCOZZAFAVA, R. (1982b). Distribuzioni non informative finitamente additive. *2^ Rassegna di Metodi Statistici e Appl.,* Univ. Cagliari - C.N.R., 199-212.

SUDDERTH, W.D. (1980). Finitely additive probabilities, coherence and the marginalization paradox. *J. Roy. Statist. Soc. Ser. B, 42,* 339-341.

WAKKER, P. (1981). Agreeing probability measures for comparative probability structures. *Ann. Statist., 9,* 658-662.

ZELLNER, A. (1973). Discussion on Dawid, A.P., Stone, M. and Zidek, J.V.: Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B., 35,* 229-230.