Submitted to the Annals of Statistics

ON BAYES'S THEOREM FOR IMPROPER MIXTURES

By Peter McCullagh*, and Han Han*

The University of Chicago

Although Bayes's theorem demands a prior that is a probability distribution on the parameter space, the calculus associated with Bayes's theorem sometimes generates sensible procedures from improper priors, Pitman's estimator being a good example. However, improper priors may also lead to Bayes procedures that are paradoxical or otherwise unsatisfactory, prompting some authors to insist that all priors be proper. This paper begins with the observation that an improper measure on Θ satisfying Kingman's countability condition is in fact a probability distribution on the power set. We show how to extend a model in such a way that the extended parameter space is the power set. Under an additional finiteness condition, which is needed for the existence of a sampling region, the conditions for Bayes's theorem are satisfied by the extension. Lack of interference ensures that the posterior distribution in the extended space is compatible with the original parameter space. Provided that the key finiteness condition is satisfied, this probabilistic analysis of the extended model may be interpreted as a vindication of improper Bayes procedures derived from the original model.

1. Introduction.

Consider a parametric model consisting of a family of probability distributions $\{P_\theta\}$ indexed by the parameter $\theta \in \Theta$. Each P_θ is a probability distribution on the observation space $\mathcal{S}_1$, usually a product space such as $\mathbb{R}^n$. In the parametric application of Bayes's theorem, the family $\{P_\theta\}$ is replaced by a single probability distribution $P_\pi(d\theta, dy) = P_\theta(dy)\,\pi(d\theta)$ on the product space $\Theta \times \mathcal{S}_1$. The associated projections are the prior π on the parameter space and the marginal distribution

$$P_\pi(\Theta \times A) = \int_\Theta P_\theta(A)\,\pi(d\theta)$$

for $A \subset \mathcal{S}_1$. To each observation $y \in \mathcal{S}_1$ there corresponds a conditional distribution $P_\pi(d\theta \mid y)$, also called the posterior distribution, on Θ.

The joint distribution $P_\pi(d\theta, dy)$ has a dual interpretation. The generative interpretation begins with θ, a random element drawn from Θ with probability distribution π, the second component being distributed according to

---

*Support for this research was provided by NSF Grant DMS-0906592

Keywords and phrases: countable measure, lack of interference, marginalization paradox

the model distribution $Y \sim P_\theta$, now treated as a conditional distribution given θ. In reverse order, the inferential interpretation begins with the observational component $Y \sim P_\pi(\Theta \times dy)$ drawn from the mixture distribution, the parameter component being distributed as $\theta \sim P_\pi(\cdot \mid y)$ from the conditional distribution given $Y = y$. The conditional distribution $P_\pi(\cdot \mid y)$ tells us how to select $\theta \in \Theta$ in order that the joint distribution should coincide with the given joint distribution $P_\pi(d\theta, dy)$.

On the assumption that the marginal measure $P_\nu(dy) = \int_\Theta P_\theta(dy)\, \nu(d\theta)$ is σ-finite, formal application of the Bayes calculus with an improper prior ν yields a posterior distribution $Q(d\theta \mid y)$ satisfying

$$P_\theta(dy)\, \nu(d\theta) = P_\nu(dy)\, Q(d\theta \mid y)$$

[Eaton (1982), Eaton and Sudderth (1995)]. This factorization of the joint measure yields a conditional law that is a probability distribution, in the sense that $Q(\Theta \mid y) = 1$. However, the joint measure is not a probability distribution, so the factorization is not to be confused with Bayes's theorem: it does not offer a probabilistic interpretation of $Q(\cdot \mid y)$ as a family of conditional distributions generated by a joint probability distribution on the product space. As a result, some authors reject the Kolmogorov axiom of total probability, arguing instead for a non-unitary measure theory for Bayesian applications [Hartigan (1983); Taraldsen and Lindqvist (2010)]. The goal of this paper is to show how an improper prior may be accommodated within the standard unitary theory without deviation from the Kolmogorov axioms. A probability space is constructed from the improper measure in such a way that $Q(\cdot \mid y)$ admits a probabilistic interpretation as a family of conditional probability distributions given the observation. Section 6 shows that σ-finiteness is not needed.

It would be inappropriate here to offer a review of the vast literature on improper priors, most of which is not relevant to the approach taken here. Nonetheless, a few remarks are in order. Some statisticians clearly have qualms about the use of such priors, partly because Bayes's theorem demands that priors be proper, partly because the 'degree of belief' interpretation is no longer compelling, and partly because the formal manipulation of improper priors may lead to inferential paradoxes of the sort discussed by Dawid, Stone and Zidek (1973). Lindley (1973) argues correctly that strict adherence to the rules of probability requires all priors to be proper. Even though the Bayes calculus often generates procedures yielding sensible conclusions, he concludes that improper priors must be rejected. Many statisticians, including some who interpret the prior as a 'degree of belief', are inclined to take a less dogmatic view. In connection with Bernoulli trials, Bernardo and Smith (1994) (section 5.2) comment as follows. *It is important to recognize, however, that this is merely an approximation device and in no way justifies [the improper limit $\theta^{-1}(1-\theta)^{-1}$] as having any special significance as a representation of 'prior ignorance'.* In subsequent discussion in section 5.4, they take a more pragmatic view of a reference prior as a mathematical tool generating a reference analysis by the Bayes calculus.

The purpose of this note is to offer a purely probabilistic interpretation of an improper prior, in agreement with Lindley's thesis but not with his conclusion. The interpretation that removes the chief mathematical obstacle is that an improper measure on Θ is a probability distribution on the set of subsets of Θ. A proper prior determines a random *element* $\theta \in \Theta$ with distribution π, whereas an improper prior ν determines a random *subset*, a countable collection $\{\theta_i\}$ distributed as a Poisson process with mean measure ν. In the product space $\Theta \times \mathcal{S}_1$, the proper joint distribution P_π determines a random element (θ, Y), whereas the improper distribution P_ν determines a random subset $Z \subset \Theta \times \mathcal{S}_1$, a countable collection of ordered pairs $Z = \{(\theta_i, Y_i)\}$. An observation on a point process consists of a sampling region $A \subset \mathcal{S}_1$ together with the set $\mathbf{y} = Y \cap A$ of events that occur in A. It is critical that the sampling region be specified in such a way that $Y \cap A$ is finite, a condition that puts definite limits on ν and on the set of sampling schemes. Having done so, we obtain the conditional distribution given the observation. The standard Bayesian argument associates with each point $y \in \mathcal{S}_1$ a probability distribution on Θ: the point process argument associates with each finite subset $\mathbf{y} \subset A$ a probability distribution on $\Theta^{\#\mathbf{y}}$. Despite this fundamental distinction, certain aspects of the conditional distribution are in accord with the formal application of the Bayes calculus, treating the mixture as if it were a model for a random element rather than a random subset.

2. Conditional distributions.

Consider a Poisson process with mean measure μ in the product space $\mathcal{S} = \mathcal{S}_0 \times \mathcal{S}_1$. Existence of the process is guaranteed if the singletons of $\mathcal{S}$ are contained in the σ-field, and μ is a countable sum of finite measures, i.e.

$$\mu = \sum_{n=1}^{\infty} \mu_n \quad \text{where} \quad \mu_n(\mathcal{S}) < \infty. \tag{2.1}$$

Kingman's countability condition, also called weak finiteness [Kingman (1993)], is the natural condition for existence because it implies that the marginal measures $\mu_0(B) = \mu(B \times \mathcal{S}_1)$ for $B \subset \mathcal{S}_0$ and $\mu_1(A) = \mu(\mathcal{S}_0 \times A)$ for $A \subset \mathcal{S}_1$

are countable. Consequently, the projected processes exist and are also Poisson.

Unlike σ-finiteness, countability does not imply the existence of a subset $A \subset \mathcal{S}$ such that $0 < \mu(A) < \infty$. If such a set exists, the process is said to be *observable on* A. For example, the measure taking the value ∞ on subsets of positive Lebesgue measure in $\mathbb{R}$ and zero otherwise is countable, but the process is not observable on any subset. Sigma-finiteness is a stronger condition, sufficient for existence but not necessary, and not inherited by the projected marginal measures [Kingman (1993)].

The symbol $Z \sim \mathrm{PP}(\mu)$ denotes a Poisson point process, which is a random subset $Z \subset \mathcal{S}$ such that for each finite collection of disjoint subsets $A_1, \ldots, A_n$ of $\mathcal{S}$, the random variables $\#(Z \cap A_1), \ldots, \#(Z \cap A_n)$ are distributed independently according to the Poisson distribution $\#(Z \cap A_j) \sim \mathrm{Po}(\mu(A_j))$. In much of what follows, it is assumed that $\mu(\mathcal{S}) = \infty$, which implies that $\#Z \sim \mathrm{Po}(\infty)$ is infinite with probability one, but countable on account of (2.1). Since Z is countable and $\mathcal{S}$ is a product set, we may label the events

$$Z = (X, Y) = \{(X_i, Y_i) : i = 1, 2, \ldots\},$$

where $X \subset \mathcal{S}_0$ is a Poisson process with mean measure μ_0 and $Y \subset \mathcal{S}_1$ is a Poisson process with mean measure μ_1. The notation $Z = (X, Y)$ implies that $X \subset \mathcal{S}_0$ and $Y \subset \mathcal{S}_1$ are countable subsets whose elements are in a specific 1–1 correspondence.

To say what is meant by an *observation* on a point process, we must first establish the sampling protocol, which is a test set or sampling region $A \subset \mathcal{S}_1$ such that $\mu_1(A) < \infty$. In this scheme, $\mathcal{S}_0$ is the domain of inference, so X is not observed. The actual observation is the test set A together with the random subset $\mathbf{y} = Y \cap A$, which is finite with probability one. Although we refer to $\mathcal{S}_1$ as the 'space of observations,' it must be emphasized that an observation is not a random *element* in $\mathcal{S}_1$, but a finite random *subset* $\mathbf{y} \subset A \subset \mathcal{S}_1$, which could be empty.

The distinction between a point process and an observation on the process is the same as the distinction between an infinite process and an observation on that process. An infinite process is a sequence of random variables $Y = (Y_1, Y_2, \ldots)$ indexed by the natural numbers, i.e. a random function $Y\colon \mathbb{N} \to \mathbb{R}$. An observation consists of a sample, a *finite* subset $A \subset \mathbb{N}$, together with the response values $Y[A]$ for the sampled units. Likewise, a point process is a random subset considered as a random function $Y\colon \mathcal{S}_1 \to \{0, 1\}$ indexed by the domain $\mathcal{S}_1$. An observation consists of a sample or sampling region $A \subset \mathcal{S}_1$ together with the restriction $Y[A] = Y \cap A$ of the process to the sample. Usually A is not finite or even countable, but the observation is necessarily finite in the sense that $\#(Y \cap A) < \infty$.

Whether we are talking of sequences or point processes, the domain of inference is not necessarily to be interpreted as a parameter space: in certain applications discussed below, the observation space consists of finite sequences in $\mathcal{S}_1 = \mathbb{R}^n$, and $\mathcal{S}_0 = \mathbb{R}^\infty$ is the set of subsequent trajectories. In this sense, predictive sample-space inferences are an integral part of the general theory (Section 4.2).

We focus here on inferences for the X-values associated with the events $\mathbf{y} = Y \cap A$ that occur in the sampling region, i.e. the subset

$$\mathbf{x} = X[A] = \{X_i : Y_i \in A\} = \{X_i : Y_i \in \mathbf{y}\},$$

in 1–1 correspondence with the observation $\mathbf{y}$. In this formal sense, an inference is a rule associating with each finite subset $\mathbf{y} \subset A$ a probability distribution on $\mathcal{S}_0^{\#\mathbf{y}}$.

Clearly, if $\mathbf{y}$ is empty, $\mathbf{x}$ is also empty, so the conditional distribution is trivial, putting probability one on the event that $\mathbf{x}$ is empty. Without loss of generality, therefore, we assume that $0 < \mu_1(A) < \infty$, that $m = \#\mathbf{y}$ is positive and finite, and that the events are labeled $(Y_1, \ldots, Y_m)$ by a uniform random permutation independent of Z. Given $\#\mathbf{y} = m$, the pairs $(X_1, Y_1), \ldots, (X_m, Y_m)$ are independent and identically distributed random variables with probability density $\mu(dx\, dy)/\mu_1(A)$ in $\mathcal{S}_0 \times A$. Thus the conditional joint density given $Y \cap A = \mathbf{y}$ is equal to

$$p(d\mathbf{x} \mid \mathbf{y}) = \prod_{i=1}^{m} \frac{\mu(dx_i\, dy_i)}{\mu_1(dy_i)} = \prod_{i=1}^{m} \mu(dx_i \mid y_i), \tag{2.2}$$

where $\mu(dx \mid y)$ is the limiting ratio $\mu(dx \times dy)/\mu_1(dy)$ as $dy \downarrow \{y\}$.

The key properties of this conditional distribution are twofold, *conditional independence* and *lack of interference*. First, the random variables $X_1, \ldots, X_m$ are conditionally independent given $Y \cap A = \mathbf{y}$. Second, the conditional distribution of X_i given $\mathbf{y}$ depends only on Y_i, not on the number or position of other events in A. For example, if two or more events occur at the same point $(Y_i = Y_j)$ the random variables X_i, X_j are conditionally independent and identically distributed given $\mathbf{y}$. The test set determines the events on which predictions are made, but beyond that it has no effect. In particular, if $m = 1$, the conditional density of X is $p(dx \mid y) \propto \mu(dx \mid y)$ regardless of the test set.

The observability assumption $\mu_1(A) < \infty$ is not made out of concern for what might reasonably be expected of an observer in the field. On the contrary, finiteness is essential to the mathematical argument leading to (2.2). If the number of events were infinite, countability implies that the values can be labeled sequentially $y_1, y_2, \ldots$ in 1–1 correspondence with the integers. Countability does not imply that they can be labeled in such a way that the infinite sequence is exchangeable. As a result, the factorization (2.2) fails if $\#\mathbf{y} = \infty$.

The remark made above, that the test set has no effect on inferences, is correct but possibly misleading. Suppose that $0 < m < \infty$ and that the observation consists of that information alone without recording the particular values. If $\mu_1(A) = 0$ or $\mu_1(A) = \infty$, no inference is possible beyond the fact that the model is totally incompatible with the observation. If the marginal measure is finite on A, the conditional density is such that the components of $X[A]$ are independent and identically distributed with density $\mu(dx \times A)/\mu_1(A)$, which does depend on the choice of test set. In the context of parametric mixture models with $\Theta \equiv \mathcal{S}_0$, each sequence with distribution P_θ has probability $P_\theta(A)$ of being recorded. Thus, before observation begins, the restriction to $A \subset \mathcal{S}_1$ effectively changes the measure to $P_\theta(A)\,\nu(d\theta)$, which is finite on Θ, but depends on the choice of A.

3. Improper mixtures. Consider a parametric statistical model consisting of a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ on the observation space $\mathcal{S}_1$, one distribution for each point θ in the parameter space. Each model distribution determines a random *element* $Y \sim P_\theta$. A probability distribution π on Θ completes the Bayesian specification, and each Bayesian model also determines a random element $(\theta, Y) \in \Theta \times \mathcal{S}_1$ distributed as $\pi(d\theta)P_\theta(dy)$. The observational component is a random element $Y \in \mathcal{S}_1$ distributed as the mixture $Y \sim P_\pi$, and the conditional distribution given $Y = y$ is formally the limit of $\pi(d\theta)P_\theta(dy)/P_\pi(\Theta, dy)$ as $dy \downarrow \{y\}$.

A countable measure ν such that $\nu(\Theta) = \infty$ does not determine a *random element* $\theta \in \Theta$, but it does determine an infinite *random subset* $X \subset \Theta$. Furthermore, the joint measure $\nu(d\theta)P_\theta(dy)$ is countable, so there exists a random subset $Z = (X, Y) \subset \Theta \times \mathcal{S}_1$, distributed according to the Poisson process with mean measure $\nu(d\theta)P_\theta(dy)$. If this interpretation is granted, it is necessary first to specify the sampling region $A \subset \mathcal{S}_1$, in such a way that $P_\nu(A) < \infty$ to ensure that only finitely many events $\mathbf{y} = Y \cap A$ occur in A. To each observed event $Y_i \in \mathbf{y}$ there corresponds a parameter point $\theta_i \in X[A]$ such that $(\theta_i, Y_i) \in Z$. Parametric inference consists in finding the joint conditional distribution given $Y \cap A = \mathbf{y}$ of the particular subset of parameter values $\theta_1, \ldots, \theta_m$ corresponding to the events observed.

This probabilistic interpretation forces us to think of the parameter and the observation in a collective manner, as sets rather than points. Taken literally, the improper mixture is not a model for a random element in $\Theta\times\mathcal{S}_1$, but a model for a random subset $Z = (X, Y) \subset \Theta \times \mathcal{S}_1$. If $\nu(\Theta) < \infty$, as in a proper mixture, it is sufficient to take $A = \mathcal{S}_1$ and to record the entire subset $\mathbf{y} \subset \mathcal{S}_1$, which is necessarily finite. However, if $\nu(\Theta) = \infty$, it is necessary to sample the process by first establishing a test set $A \subset \mathcal{S}_1$ such that $P_\nu(A) < \infty$, and then listing the finite set of values $\mathbf{y} = Y \cap A$ that occur in A. Generally speaking, this finiteness condition rules out many sampling schemes that might otherwise seem reasonable. In the special case where $\#\mathbf{y} = 1$, $X[A]$ is a random subset consisting of a single point, whose conditional density at $x \in \Theta$ is

$$\mathrm{pr}(X[A] \in dx \mid \mathbf{y} = \{y\}) = \frac{\nu(dx)\, p_x(y)}{\int_\Theta p_\theta(y)\, \nu(d\theta)}, \tag{3.1}$$

where $p_\theta(y)$ is the density of P_θ at y. The finiteness condition on A ensures that the integral in the denominator is finite, and the occurrence of an event at y implies that P_ν assigns positive mass to each open neighborhood of y.

Provided that $0 < P_\nu(A) < \infty$, this purely probabilistic conclusion may be interpreted as a vindication of the formal Bayes calculation associated with an improper prior. However, the two versions of Bayes's theorem are quite different in logical structure; one implies a single random element, the other infinitely many. Accordingly, if a statistical procedure is to be judged by a criterion such as a conventional loss function, which presupposes a single observation and a single parameter, we should not expect optimal results from a probabilistic theory that demands multiple observations and multiple parameters. Conversely, if the procedure is to be judged by a criterion that allows for multiple sequences each with its own parameter, we should not expect useful results from a probabilistic theory that recognizes only one sequence and one parameter. Thus, the existence of a joint probability model associated with an improper prior does not imply optimality in the form of coherence, consistency or admissibility. For example, in the MANOVA example of Eaton and Sudderth (1995), the Poisson point process interpretation yields the classical posterior, which is incoherent in de Finetti's sense and is strongly inconsistent in Stone's sense.

The observability condition implies that the restriction of P_ν to A is finite, and hence trivially σ-finite. The role of the finiteness condition is illustrated by two examples in Sections 4 and 6. For the Gaussian model P_ν is countable for every $n \geq 0$ and σ-finite for $n \geq 2$, which guarantees the existence of a sampling region if $n \geq 2$. For the Bernoulli model P_ν is countable for each $n \geq 0$ but not σ-finite for any n. Nonetheless, the finiteness condition for observability is satisfied by certain subsets $A \subset \{0, 1\}^n$ for $n \geq 2$.

4. Gaussian point process.

4.1. *Parametric version.* Consider the standard model for a Gaussian sequence with independent $N(\theta, \sigma^2)$ components. Let p be a given real number, and let the prior measure be $\nu(d\theta\, d\sigma) = d\theta\, d\sigma/\sigma^p$ on the parameter space $\Theta = \mathbb{R} \times \mathbb{R}^+$. For all p, both ν and the joint measure on $\Theta \times \mathbb{R}^n$ satisfy the countability condition. Consequently a Poisson point process $Z = (X, Y) \subset \Theta \times \mathbb{R}^n$ exists in the product space. For $n > 2 - p$, the marginal measure P_ν has a density in $\mathbb{R}^n$

$$\lambda_n(y) = \frac{\Gamma((n+p-2)/2) 2^{(p-3)/2} \pi^{-(n-1)/2} n^{-1/2}}{(\sum_{i=1}^n (y_i - \bar{y})^2)^{(n+p-2)/2}}, \tag{4.1}$$

which is finite at all points $y \in \mathbb{R}^n$ except for the diagonal set. Provided that $n \geq 2$ and $n > 2 - p$, there exists in $\mathbb{R}^n$ a subset A such that $P_\nu(A) < \infty$, which serves as the region of observation. In fact, these conditions are sufficient for σ-finiteness in this example. To each observation $\mathbf{y} = Y \cap A$ and to each event $y \in \mathbf{y}$ there corresponds a conditional distribution on Θ with density

$$p(\theta, \sigma \,|\, Y \cap A = \mathbf{y}, y \in \mathbf{y}) = \phi_n(y; \theta, \sigma)\, \sigma^{-p} / \lambda_n(y),$$

where $\phi_n(y; \theta, \sigma)$ is the Gaussian density at y in $\mathbb{R}^n$. The conditional distribution (2.2) of the parameter subset $X[A] \subset \Theta$ given $Y \cap A = \mathbf{y}$ is a product of factors of this type, one for each of the events in $\mathbf{y}$. It should be emphasized here that the information in the conditioning event is not simply that $\mathbf{y} \subset Y$, but also that Y contains no other events in A.

4.2. *Non-parametric version.* Let $\mathbb{N}$ be the set of natural numbers, and let $\mathcal{S} = \mathbb{R}^{\mathbb{N}}$ be the collection of real-valued sequences,

$$\mathcal{S} = \mathbb{R}^{\mathbb{N}} = \{y = (y_1, y_2, \ldots) : y_i \in \mathbb{R},\ i \in \mathbb{N}\},$$

with product σ-field $\mathscr{R}^{\mathbb{N}}$. We construct directly in this space a Poisson process $Z \subset \mathcal{S}$ whose mean measure Λ is uniquely determined by its finite-dimensional projections Λ_n with density (4.1). By their construction, these measures are finitely exchangeable and satisfy the Kolmogorov consistency condition $\Lambda_{n+1}(A \times \mathbb{R}) = \Lambda_n(A)$ for each integer $n \geq 0$ and $A \in \mathscr{R}^n$. In keeping with the terminology for random sequences, we say that the point process $Z \sim \mathrm{PP}(\Lambda)$ is infinitely exchangeable if each Λ_n is finitely exchangeable.

Let $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_0$, where $\mathcal{S}_1 = \mathbb{R}^n$ is the projection onto the first n coordinates, and $\mathcal{S}_0 \cong \mathcal{S}$ is the complementary projection onto the subsequent coordinates. Each event $z \in Z$ is an ordered pair, so we write $Z = (Y, X) \subset \mathcal{S}$ as a countable set of ordered pairs (Y_i, X_i) in which the marginal process $Y \subset \mathcal{S}_1$ is Poisson with parameter Λ_n, and $X \sim \mathrm{PP}(\Lambda)$ has the same distribution as Z. Provided that the set $A \subset \mathcal{S}_1$ has finite Λ_n-measure, the observation $\mathbf{y} = Y \cap A$ is finite. To each event $y \in \mathbf{y}$ there corresponds an event $z = (y, x) \in Z$, so that $y = (z_1, \ldots, z_n)$ is the initial sequence, and $x = (z_{n+1}, \ldots)$ is the subsequent trajectory. The conditional distribution (2.2) is such that the subsequent trajectories $X[A]$ are conditionally independent and non-interfering given $Y \cap A = \mathbf{y}$. For each event $y \in \mathbf{y}$, the k-dimensional joint density at $x = (x_1, \ldots, x_k)$ of the subsequent trajectory is

$$p(dx \mid Y \cap A = \mathbf{y}, y \in \mathbf{y}) = \frac{\lambda_{n+k}(y, x)\, dx}{\lambda_n(y)}, \tag{4.2}$$

which is the k-dimensional exchangeable Student t density [Kotz and Nadarajah (2004), page 1] on $\nu = n + p - 2 > 0$ degrees of freedom.

For any continuous location-scale model with finite p-th moment and improper prior density proportional to $d\mu\, d\sigma/\sigma^p$ with $p > 0$, the initial segment $Y \subset \mathbb{R}^2$ is a Poisson process with intensity

$$\lambda_2(y) \propto \frac{1}{|y_1 - y_2|^p}.$$

Otherwise if $p \leq 0$ the initial segment of length $n > 2 - p$ is a Poisson process with intensity

$$\lambda_n(y) \propto \frac{1}{(\sum_{i=1}^n (y_i - \bar{y})^2)^{(n+p-2)/2}}.$$

The prescription (4.2) extends each event $y \in Y$ to an infinite random sequence in such a way that the set of extended sequences $Z \subset \mathbb{R}^{\mathbb{N}}$ is a Poisson process with mean measure Λ. Given $Y \subset \mathbb{R}^2$, these extensions are conditionally independent, non-interfering, and each extension is an exchangeable sequence. In the Gaussian case, (4.2) is equivalent to the statement that each initial sequence with $s_n \neq 0$ is extended according to the recursive Gosset rule

$$y_{n+1} = \bar{y}_n + s_n \epsilon_n \sqrt{\frac{n^2 - 1}{n(n + p - 2)}},$$

where $\bar{y}_n, s_n^2$ are the sample mean and variance of the first n components, and $\epsilon_n \sim t_{n+p-2}$ has independent components. The resulting extension is an exchangeable sequence whose k-dimensional joint density at $(y_3, \ldots, y_{k+2})$ is $\lambda_{k+2}(y_1, \ldots, y_{k+2})/\lambda_2(y_1, y_2)$.

The Gosset extension is such that the sequence $(\bar{y}_n, s_n^2)$ is Markovian and has a limit. Given a single sequence in the sampling region, the joint distribution of the limiting random variables $(\bar{y}_\infty, s_\infty)$ is

$$p(\bar{y}_\infty, s_\infty \,|\, Y \cap A = \mathbf{y}, y \in \mathbf{y}) = \phi_n(y; \bar{y}_\infty, s_\infty)\, s_\infty^{-p}/\lambda_n(y),$$

which is the posterior density on Θ as computed by the Bayes calculus with improper prior.

5. Cauchy sequences. Consider the standard model for a Cauchy sequence having independent components with parameter $\theta \in \mathbb{R} \times \mathbb{R}^+$. For $p > 0$, the prior measure $\nu(d\theta) = d\theta_1\, d\theta_2/\theta_2^p$ satisfies the countability condition, which implies that a Poisson process $X = (X, Y) \subset \Theta \times \mathbb{R}^n$ with mean measure P_ν exists in the product space. If $0 < p < n$ and $n \geq 2$, the marginal measure in $\mathbb{R}^n$ has a density which is finite at all points $y \in \mathbb{R}^n$ whose components are distinct. The density satisfies the recurrence formula

$$\lim_{y_n \to \pm\infty} \pi y_n^2 \lambda_{n,p}(y_1, \ldots, y_n) = \lambda_{n-1,p-1}(y_1, \ldots, y_{n-1}).$$

For integer $p \geq 2$, the density is

$$\lambda_{n,p}(y) = \begin{cases} \dfrac{(-1)^{(n-p+1)/2}}{\pi^{n-2} 2^{n-p+1}} \displaystyle\sum_{r \neq s} \dfrac{|y_s - y_r|^{n-p}}{d_r d_s} & (n-p) \text{ odd;} \\[2ex] \dfrac{(-1)^{(n-p)/2}}{\pi^{n-1} 2^{n-p}} \displaystyle\sum_{r \neq s} \dfrac{(y_s - y_r)^{n-p} \log|y_s - y_r|}{d_r d_s} & (n-p) \text{ even;} \end{cases} \tag{5.1}$$

where $d_r = \prod_{t \neq r}(y_t - y_r)$. For example, $\lambda_{2,1}(y) = 1/(2|y_1 - y_2|)$ and

$$\lambda_{3,2}(y) = \frac{1}{2\pi|(y_1 - y_2)(y_2 - y_3)(y_1 - y_3)|}.$$

For $n > p$, there exists a subset $A \subset \mathbb{R}^n$ such that $\Lambda_n(A) < \infty$, which serves as the region of observation. The Poisson process determines a probability distribution on finite subsets $\mathbf{y} \subset A$, and to each point $y \in \mathbf{y}$ it also associates a conditional distribution on Θ with density

$$\frac{P_\nu(d\theta \times dy)}{\Lambda_n(dy)} = \frac{f_n(y; \theta)\theta_2^{-p}}{\lambda_n(y)}, \tag{5.2}$$

where $f_n(y; \theta)$ is the Cauchy density at $y \in \mathbb{R}^n$.

In the nonparametric version with Θ replaced by $\mathbb{R}^k$, the conditional distribution extends each point $y \in A$ to a sequence $(y, X) \in \mathbb{R}^{n+k}$, with conditional density $X \sim \lambda_{n+k}(y, x)/\lambda_n(y)$. The extension is infinitely exchangeable. The tail trajectory of the infinite sequence is such that, if $T_k\colon \mathbb{R}^k \to \Theta$ is Cauchy-consistent, $T_{n+k}(y, X)$ has a limit whose density at $\theta \in \Theta$ is (5.2).

6. Binary sequences. Consider the standard model for a Bernoulli sequence with parameter space $\Theta = (0,1)$. The prior measure $\nu(d\theta) = d\theta/(\theta(1-\theta))$ determines a Poisson process with intensity $\theta^{n_1(y)-1}(1-\theta)^{n_0(y)-1}$ at (y,θ) in the product space $\mathcal{S}_1 \times \Theta$. Here $\mathcal{S}_1 = \{0,1\}^n$ is the space of sequences of length n, $n_0(y)$ is the number of zeros and $n_1(y)$ is the number of ones in y. The marginal measure on the observation space is

$$\Lambda_n(\{y\}) = \begin{cases} \Gamma(n_0(y))\Gamma(n_1(y))/\Gamma(n) & n_0(y), n_1(y) > 0 \\ \infty & \text{otherwise,} \end{cases}$$

which is countable but not σ-finite. Any subset $A \subset \mathcal{S}_1$ that excludes the zero sequence and the unit sequence has finite measure and can serve as the region of observation. Given such a set and the observation $\mathbf{y} = Y \cap A$ recorded with multiplicities, the conditional distribution (2.2) associates with each $y \in \mathbf{y}$ the beta distribution

$$P_\nu(\theta \mid Y \cap A = \mathbf{y}, y \in \mathbf{y}) = \frac{\theta^{n_1(y)-1}(1-\theta)^{n_0(y)-1}\Gamma(n)}{\Gamma(n_1(y))\Gamma(n_0(y))}$$

on the parameter space.

As in the preceding section, we may by-pass the parameter space and proceed directly by constructing a Poisson process with mean measure Λ in the space of infinite binary sequences. The values assigned by Λ to the infinite zero sequence and the infinite unit sequence are not determined by $\{\Lambda_n\}$, and can be set to any arbitrary value, finite or infinite. Regardless of this choice, (2.2) may be used to predict the subsequent trajectory of each of the points $\mathbf{y} = Y \cap A$ provided that $\Lambda_n(A) < \infty$. In particular, the conditional distribution of the next subsequent component is

$$\mathrm{pr}(y_{n+1} = 1 \mid Y \cap A = \mathbf{y}, y \in \mathbf{y}) = n_1(y)/n.$$

This is the standard Polya urn model [Durrett (2005)] for which the infinite average of all subsequent components is a beta random variable with parameters $(n_0(y), n_1(y))$, in agreement with the parametric analysis.

7. Interpretation. The point-process interpretation of an improper measure on Θ forces us to think of the parameter in a collective sense as a random subset rather than a random point. One interpretation is that a proper prior is designed for a specific scientific problem whose goal is the estimation of a particular parameter about which something may be known, or informed guesses can be made. An improper mixture is designed for a generic class of problems, not necessarily related to one another scientifically, but all having the same mathematical structure. Logistic regression models, which are used for many purposes in a wide range of disciplines, are generic in this sense. In the absence of a specific scientific context, nothing can be known about the parameter, other than the fact that there are many scientific problems of the same mathematical type, each associated with a different parameter value. In that wider sense of a generic mathematical class, it is not unnatural to consider a broader framework encompassing infinitely many scientific problems, each with its own parameter. The set of parameters is random but not indexed in an exchangeable way.

A generic model may be tailored to a specific scientific application by coupling it with a proper prior distribution π that is deemed relevant to the scientific context. If there is broad agreement about the model and the relevance of π to the context, subsequent calculations are uncontroversial. Difficulties arise when no consensus can be reached about the prior. According to one viewpoint, each individual has a personal prior or belief; Bayes's theorem is then a recipe for the modification of personal beliefs [Bernardo and Smith (1994), chapter 2]. Another line of argument calls for a panel of so-called experts to reach a consensus before Bayes's theorem can be used in a mutually agreeable fashion [Weerhandi and Zidek (1981); Genest, McConway and Schervish (1986)]. A third option is to use proper but flat or relatively uninformative priors. Each of these options demands a proper prior on Θ in order that Bayes's theorem may be used.

This paper offers a fourth option by showing that it is possible to apply Bayes's theorem to the generic model. Rather than forcing the panel to reach a proper consensus, we may settle for an improper prior as a countable sum of proper, and perhaps mutually contradictory, priors generated by an infinite number of experts. Although Bayes's theorem can be used, the structure of the theorem for an improper mixture is not the same as the structure for a proper prior. For example, improper Bayes estimators need not be admissible.

Finiteness of the restriction of the measure to the sampling region is needed in our argument. If the restriction to the sampling region is σ-finite, we may partition the region into a countable family of disjoint subsets of finite measure, and apply the extension subset by subset. The existence of a Poisson point process on the sampling region is assured by Kingman's superposition theorem. Lack of interference implies that these extensions are mutually consistent, so there is no problem dealing with such σ-finite restrictions. This is probably not necessary from a statistical perspective, but it does not create any mathematical problems because the extension does not depend on the choice of the partition of the region.

8. Marginalization paradoxes. The unBayesian characteristic of an improper prior distribution is highlighted by the marginalization paradoxes discussed by Stone and Dawid (1972), and by Dawid, Stone and Zidek (1973). In the following example from Stone and Dawid (1972), the formal marginal posterior distribution calculated by two methods demonstrates the inconsistency.

EXAMPLE 8.1. *The observation consists of two independent exponential random variables $X \sim \mathcal{E}(\theta\phi)$ and $Y \sim \mathcal{E}(\phi)$, where θ and ϕ are unknown parameters. The parameter of interest is the ratio θ.*

Method 1. The joint density is

$$\mathrm{pr}(dx, dy \,|\, \theta, \phi) = \theta\phi^2 e^{-\phi(\theta x + y)}\, dx\, dy.$$

Given the improper prior distribution $\pi(\theta)\, d\theta\, d\phi$, the marginal posterior distribution for θ is

$$\pi(\theta \,|\, x, y) \propto \frac{\pi(\theta)\theta}{(\theta x + y)^3} \propto \frac{\pi(\theta)\theta}{(\theta + z)^3}, \tag{8.1}$$

where $z = y/x$.

Method 2. Notice that the posterior distribution depends on (x, y) only through z. For a given θ, z/θ has an $F_{2,2}$ distribution, that is,

$$\mathrm{pr}(z \,|\, \theta) \propto \frac{\theta}{(\theta + z)^2}.$$

Using the implied marginal prior $\pi(\theta)\, d\theta$, as if it were the limit of a sequence of proper priors, we obtain

$$\pi(\theta \,|\, z) \propto \frac{\pi(\theta)\theta}{(\theta + z)^2} \tag{8.2}$$

which differs from (8.1). It has been pointed out by Dempster and in the author's rejoinder [Dawid, Stone and Zidek (1973)], that no choice of $\pi(\theta)$ could bring the two analysis into agreement.

From the present viewpoint, the improper prior determines a random subset of the parameter space and a random subset of the observation space $(\mathbb{R}^+)^2$. Under suitable conditions on π, the bivariate intensity

$$\lambda(x, y) = 2\int_0^\infty \frac{\theta\, \pi(\theta)\, d\theta}{(\theta x + y)^3}$$

is finite on the interior of the observation space, so the bivariate process is observable. Equation (2.2) associates with each event (x, y) the conditional distribution (8.1) in agreement with the formal calculation by method 1. Each event (x, y) determines a ratio $z = y/x$, and the set of ratios is a Poisson point process in $(0, \infty)$. However, the marginal measure is such that $\Lambda_z(A) = \infty$ for sets of positive Lebesgue measure, and zero otherwise. This measure is countable, but the marginal process is not observable. Thus, conclusion (8.2) deduced by method 2 does not follow from (2.2), and there is no contradiction.

Conversely, if the prior measure $\pi(d\theta)\rho(d\phi)$ is multiplicative with $\rho(\mathbb{R}^+) < \infty$ and π locally finite, the marginal measure on the observation space is such that

$$P_\nu(\{a < x/y < b\}) < \infty$$

for $0 < a < b < \infty$. Thus, the ratio $z = x/y$ is observable, and the conditions for method 2 are satisfied. The point process model associates with each ratio $0 < z < \infty$ the conditional distribution with density

$$\pi(\theta, \phi \,|\, z) \propto \frac{\rho(\phi)\,\theta}{(\theta + z)^2}$$

in agreement with (8.2). However, the conditional distribution given (x, y)

$$\pi(\theta, \phi \,|\, (x, y)) \propto \theta\phi^2 e^{-\phi(\theta x + y)}\, \pi(\theta)\, \rho(\phi)$$

is such that the marginal distribution of θ given (x, y) is not a function of z alone. Once again, there is no conflict with (8.2).

All of the other marginalization paradoxes in Dawid, Stone and Zidek (1973) follow the same pattern.

Jaynes (2003) asserts that "an improper pdf has meaning only as the limit of a well-defined sequence of proper pdfs." On this point, there seems to be near-universal agreement, even among authors who take diametrically opposed views on other aspects of the marginalization paradox (Akaike (1980), Dawid, Stone and Zidek (1973) and Wallstrom (2007)). No condition of this sort occurs in the point-process theory. However, a sequence of measures μ_n such that $\mu_n(A) < \infty$, each of which assigns a conditional distribution (2.2) to every $y \in A$, may have a weak limit $\mu_n \to \mu$ such that $\mu(A) = \infty$ for which no conditional distribution exists.

References.

Akaike, H. (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distribution. *J.R. Statist. Soc. B* **42** 46–52.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* Wiley, Chichester.

DAWID, A., STONE, M. and ZIDEK, J. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J.R. Statist. Soc. B* **35** 189–233.

DURRETT, R. (2005). *Probability: Theory and Examples*. Thomson Brooks/Cole, California.

EATON, M. (1982). A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics*.

EATON, M. and SUDDERTH, W. (1995). The formal posterior of a standard flat prior in MANOVA is incoherent. *Journal of the Italian Statistical Society* **2** 251–270.

GENEST, C., MCCONWAY, K. and SCHERVISH, M. (1986). Characterization of externally Bayesian pooling operators. *Ann. Statist.* **14** 487–501.

HARTIGAN, J. (1983). *Bayes Theory*. Springer, New York.

JAYNES, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, New York.

KINGMAN, J. (1993). *Poisson Processes*. Clarendon Press, Oxford.

KOTZ, S. and NADARAJAH, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, New York.

LINDLEY, D. (1973). Discussion of Dawid, Stone and Zidek (1973). *J.R. Statist. Soc. B* **35** 218–219.

STONE, M. and DAWID, A. (1972). Un-Bayesian implications of improper Bayes inference in routine posterior statistical problems. *Biometrika* **59** 369–375.

TARALDSEN, G. and LINDQVIST, B. (2010). Improper priors are not improper. *The American Statistician* **64** 154–158.

WALLSTROM, T. (2007). The marginalization paradox and the formal Bayes' law. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, edited by K. Knuth et al., AIP Conference Proceedings* **954** 93–100.

WEERHANDI, S. and ZIDEK, J. (1981). Multi-Bayesian statistical decision theory. *J. Roy. Statist. Soc. Ser. A* **144** 85–93.

PETER MCCULLAGH
DEPARTMENT OF STATISTICS
5734 SOUTH UNIVERSITY AVE
CHICAGO, IL 60637
E-MAIL: pmcc@galton.uchicago.edu

HAN HAN
DEPARTMENT OF STATISTICS
5734 SOUTH UNIVERSITY AVE
CHICAGO, IL 60637
E-MAIL: han@galton.uchicago.edu