Improper Priors Are Not Improper

Gunnar Taraldsen,
SINTEF Information and Communication Technology,
N-7465 Trondheim, Norway
(email: Gunnar.Taraldsen@ntnu.no)

Bo Henry Lindqvist
Department of Mathematical Sciences,
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway
(email: Bo.Lindqvist@math.ntnu.no)

January 11, 2010

Author's Footnote:

Gunnar Taraldsen is Research Scientist, SINTEF Information and Communication Technology, N-7465 Trondheim, Norway (email: Gunnar.Taraldsen@ntnu.no).

Bo Henry Lindqvist is Professor, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway (email: Bo.Lindqvist@math.ntnu.no).

Abstract

It is well known that improper priors in Bayesian statistics may lead to proper posterior distributions and useful inference procedures. This motivates us to give an elementary introduction to a theoretical frame for statistics that includes improper priors. Axioms that allow improper priors are given by a relaxed version of Kolmogorov's formulation of probability theory. The theory of conditional probability spaces formulated by Renyi is closely related, but the initial axioms and the motivation differ. One consequence of the axioms is a general Bayes theorem which gives proper posterior distributions, and furthermore, the theory also gives a convenient frame for formulation of non-Bayesian statistical models. The results are in particular relevant for the current usage of improper priors in Markov Chain Monte Carlo methods, and for methods for simulation from conditional distributions given sufficient statistics. This theory gives an alternative to ad hoc arguments without an underlying theory, and removes apparent paradoxes. Readers that acknowledge the need for a theoretical basis for statistical inference including improper priors are urged to consider the theory of conditional probability spaces as presented here.

KEYWORDS: Axioms of probability, Bayesian statistics, Conditional law, Marginalization paradox, Posterior propriety, Admissibility

1. INTRODUCTION

Let the prior knowledge regarding a parameter θ be given by a density $\pi(\theta)$. A statistical model for the observation X with a density $f(x \,|\, \theta)$ gives the posterior density

$$\pi(\theta \,|\, x) = \frac{f(x \,|\, \theta)\pi(\theta)}{f(x)}, \tag{1}$$

where $f(x) = \int f(x \,|\, \theta)\pi(\theta)d\theta$ is the marginal density of X. Typically θ and X are vectors. It is well known how to modify (1) if θ has a discrete distribution, or a more general distribution. Equation (1) is one version of Bayes' theorem.

The algorithm given by equation (1) is well defined as long as $0 < f(x) < \infty$, but the usual proof of the validity is restricted to the case where $\pi(\theta)$ is a probability density. Nonetheless, Bayesian analysis is routinely carried out successfully without assuming $\pi(\theta)$ to be a probability density. In principle $\pi(\theta)$ may be taken as any non-negative and non-null function on the parameter space, and "success" means that the posterior density in (1) makes sense.

The aim of this paper is to present the essential ingredients in a theory which allows densities π which are not probability densities, and in which a simple condition for the "properness" of improper priors can be formulated. This theory has equation (1) as a consequence, including cases where the prior is improper. The idea is simply to allow infinite probabilities in the axioms of Kolmogorov (1933).

Somewhat surprisingly, all conditional distributions, derived using the Radon-Nikodym theorem as in the case of ordinary probability theory, are still proper probability distributions. Halmos and Savage (1949) justified this for finite measures, and it was later proved by Eaton (1982) and Chang and Pollard (1997) that it holds more generally. The hook is just a certain condition that needs to be satisfied for the variable that we condition on: It is only allowed to condition on regular variables. This term was introduced by Renyi (1970). We will instead refer to these variables as σ-finite, as explained in more detail later.

The presentation will not be mathematically complete. Technical terms similar to *measurable* and *almost everywhere* will be avoided. This is not because the terms are unnecessary, but is rather a choice given that it is hoped that readers without familiarity with measure

theory should be able to follow the arguments. On the other hand, the mathematically oriented reader should be able to fill in the necessary qualifiers. For an excellent introduction to measure theory, and in particular the Radon-Nikodym theorem, we refer to Rudin (1987). Kolmogorov (1933), Halmos (1950) and Renyi (1970) present most relevant and readable alternatives with additional focus on probability theory.

2. KOLMOGOROV REVISITED

Probability theory, as formulated by Kolmogorov (1933), identifies every event A with a subset of a fixed underlying abstract space Ω. The family $\mathcal{E}$ of events is assumed to be such that the complementary event is an event, and so that countable intersections and unions of events are events. It is furthermore assumed that Ω is equipped with a fixed law Pr with $\Pr(A) \geq 0$ for $A \in \mathcal{E}$, so that $\Pr(A_1 \cup A_2 \cup \cdots) = \Pr(A_1) + \Pr(A_2) + \cdots$ whenever $A_1, A_2, \ldots$ are pairwise disjoint events. These assumptions are exactly as in the original formulation (Kolmogorov, 1933), but Kolmogorov adds the axiom $\Pr(\Omega) = 1$. This will not be done here, and the case $\Pr(\Omega) = \infty$ will be allowed in the following.

A random quantity X, which takes values in a space Ω_X, is as usual identified with a function $X : \Omega \to \Omega_X$. It is required that $(X \in C) = \{\omega \in \Omega \,|\, X(\omega) \in C\}$ is an event in Ω for any event C in Ω_X. This makes sense since it is also assumed that Ω_X is equipped with a family $\mathcal{E}_X$ of events. Kolmogorov used this very powerful recipe which gives meaning to the concept of random numbers, random vectors, random functions (stochastic processes), and other random quantities that could be of interest. Here, and elsewhere, we use the convention that capital letters are used for random quantities while the corresponding lower case ones are realizations.

The crux of the assumption of a fixed underlying space Ω is that events given simultaneously by various different random objects are well defined. A somewhat theoretical example is given by the event that the trajectory of a stochastic process is continuous and that it avoids a random set. A more common example is that an n-tuple of random numbers is contained in a specified open subset of $\mathbb{R}^n$.

The law $\Pr_X$ of a random quantity X is defined on $\mathcal{E}_X$ by

$$\Pr_X(C) = \Pr(X \in C) \quad \text{(Kolmogorov, 1933, p.21, eq. (1))}, \tag{2}$$

which is well defined since $(X \in C) \in \mathcal{E}$, and Pr is defined on $\mathcal{E}$. The law Pr on Ω is hence lifted to a law $\Pr_X$ for X on Ω_X. Equation (2) will turn out to be one of the heroes in this story. It survives also in the case $\Pr(\Omega) = \infty$, which is necessary in our treatment of improper priors.

3. MARGINAL AND JOINT LAWS

A particular consequence of equation (2) is that the law of $X = (U, V)$ determines the law of U:

$$\Pr_U(A) = \Pr(U \in A) = \Pr(U \in A, V \in \Omega_V) = \Pr_{U,V}(A \times \Omega_V). \tag{3}$$

In this context, the law $\Pr_U$ of U is referred to as a marginal law, and the law $\Pr_{U,V}$ of (U, V) is referred to as the joint law of U and V.

Equation (2), and the particular consequence (3), may seem innocent, but marks a difference between the theory presented here and the alternative theory presented by Hartigan. The following quote demonstrates this explicitly (Hartigan, 1983, p.23):

> *It is assumed therefore that the joint distribution, the conditional distribution, and the marginal distribution are specified separately to follow the axioms of conditional probability.*

Chang and Pollard (1997, p.308) assume that $T = (X, Y, Z)$ is a quantity with uniform distribution on $\Omega_T = (0, 1)^2 \times \mathbb{R}$, and argue along the lines of Hartigan to reach the conclusion:

> *Is there any paradox in (X, Y) appearing to have several different joint distributions? We think not.*

This is in contrast to the point of view presented here: The marginal distribution of a random quantity U is uniquely determined by the joint distribution of a random quantity (U, V) as in equation (3). The distribution of $U = (X, Y)$ is hence in particular uniquely

determined by the distribution of $(U, V) = (X, Y, Z)$, in contrast to the conclusion of Chang and Pollard.

More generally, the distribution of any quantity $Y = \phi(X)$ is uniquely determined by the distribution of X by equation (2):

$$\Pr_Y(A) = \Pr(Y \in A) = \Pr(\phi(X) \in A) = \Pr_X(\phi \in A). \tag{4}$$

This makes sense with the interpretation $(\phi \in A) = \{x \in \Omega_X \,|\, \phi(x) \in A\}$. The previous special case follows from $\phi(u, v) = u$. Equation (2) ensures that each random quantity has one, and only one, law.

It is not argued that the approach given by Hartigan is in any way wrong, but it is argued that our approach gives a reasonable alternative. The examples used by Hartigan (1983, p.23) and Chang and Pollard (1997, p.308) will be discussed further in Section 7. It will be explained that the theory described by Hartigan rejects our hero: equation (2). He insists instead on one version of the concept of a disintegration, a generalization of product measure, as a basis for the definition of conditional probabilities.

4. CONDITIONAL LAWS

A precise formulation of the concept of conditional laws in probability theory needs a bit of measure theory. The highlight is then the application of the celebrated Radon-Nikodym theorem (Halmos, 1950). It is fortunate that the relaxation of Kolmogorov's axioms to include an improper and σ-finite law Pr allows a straightforward extension of the classical arguments. A law Pr is σ-finite if there exist events $A_1, A_2, \ldots$ with $\Omega = \cup_i A_i$ and $\Pr(A_i) < \infty$ for $i = 1, 2, \ldots$. A random quantity Θ is said to be σ-finite if the law $\Pr_\Theta$ is σ-finite.

Now, a unique conditional probability $\Pr^\theta(A) = \Pr(A|\Theta = \theta)$ for $A \in \mathcal{E}$ can be shown to exist if Θ is a σ-finite random quantity. The conditional law $\Pr^\theta$ is furthermore always a probability law in the sense that $\Pr^\theta(\Omega) = 1$. A sketch of the argument, which requires some knowledge of measure theory, is presented next.

A desired property of the conditional law is the identity

$$\Pr(A \cap (\Theta \in B)) = \int_B \Pr^\theta(A)\ \Pr_\Theta(d\theta). \tag{5}$$

This can in fact be taken as the defining property of $\Pr^\theta(A)$. The Radon-Nikodym theorem states exactly that the function $g(\theta) = \Pr^\theta(A)$ exists and is uniquely defined by (5): The function g is the density of the measure $\mu(B) = \Pr(A \cap (\Theta \in B))$ with respect to the σ-finite $\Pr_\Theta$. The required absolute continuity is satisfied since $\Pr_\Theta(B) = \Pr(\Theta \in B) = 0$ implies $\Pr(A \cap (\Theta \in B)) = 0$. The normalization $\Pr^\theta(\Omega) = 1$ follows from $\Pr(\Theta \in B) = \int_B 1 \Pr_\Theta(d\theta)$.

The key in the above arguments is simply that the Radon-Nikodym theorem allows σ-finite laws, and the above classical argument can be used also in this case. More surprising is perhaps the conclusion that *the resulting conditional law is always a probability law* with $\Pr^\theta(\Omega) = 1$, and in particular also in the case where $\Pr(\Omega) = \infty$. This normalization is also emphasized by Halmos and Savage (1949, p.230). They give an argument essentially as above, but restricted to the case where $\Pr(\Omega)$ is bounded but not necessarily a probability measure. The more general case with a σ-finite law is treated by Eaton (1982) and Chang and Pollard (1997).

The conclusion is that the conditional law $\Pr^\theta$ is well defined and unique for any σ-finite random quantity Θ. It is, however, almost of equal importance to note that the conditional law $\Pr^\theta$ is not defined when Θ is not σ-finite.

5. TOWARD STATISTICS

In some sense, statistics generalizes probability theory by the addition and focus on the concept of parameters. In Bayesian statistics, parameters are furthermore interpreted as random quantities. It is hence natural to represent a parameter θ by a function $\Theta : \Omega \to \Omega_\Theta$, and let $\Pr_\Theta(A) = \Pr(\Theta \in A)$ define the law of Θ just as in equation (2). The law $\Pr_\Theta$ is the prior distribution of Θ.

It is common practice in statistical theory to assume that the sample space Ω_X and the parameter space Ω_Θ are given without any link to an underlying fixed space Ω (Lehmann, 1959; Schervish, 1995). The additional assumption of a fixed underlying abstract space will however soon be demonstrated to be most convenient, just as in the original formulation of probability theory by Kolmogorov (1933).

Recall now the point of depature of the present study, which is to study the consequences of

allowing improper priors. Thus, assume that $\Pr_\Theta(\Omega_\Theta) = \infty$, with the additional requirement that Θ is σ-finite. It can be recalled that this means that there is a disjoint partition $\Omega_\Theta = A_1 \cup A_2 \cup \cdots$ where $\Pr_\Theta(A_i) < \infty$ for all i. The additional observation that $\Omega = (\Theta \in \cup_i A_i) = \cup_i (\Theta \in A_i)$ gives that Pr is also by necessity improper and σ-finite, since $\Pr(\Omega) = \Pr_\Theta(\Omega_\Theta) = \infty$ and $\Pr(\Theta \in A_i) = \Pr_\Theta(A_i) < \infty$.

Thus, by allowing the law of Θ to be improper and σ-finite, *we are forced to assume that* Pr *is improper and σ-finite.* Note, however, that the σ-finiteness of Pr does not imply that every random quantity defined on Ω is σ-finite. As we shall see later, this issue is closely related to the problem of propriety of posterior distributions.

A simple example of a random quantity X with a law which is not σ-finite follows. Assume as above that $\Pr(\Omega) = \infty$ and that Pr is σ-finite. Let $X(\omega) = 1$ for all $\omega \in \Omega$. It follows that $\Pr_X(A)$ equals 0 or ∞, with $\Pr_X(A) = \infty$ if $1 \in A$, and no countable partition of the sample space Ω_X into sets with finite measure can be found. Note also that $\Pr(\Omega) = \infty$ implies more generally that every random quantity has an improper law. This represents, however, no hindrance for doing statistical inference since every conditional law turns out to be a probability law with $\Pr^\theta(\Omega) = 1$, and the focus will be on the conditional laws.

6. STATISTICAL MODELS

The conditional law $\Pr_X^\theta$ of a random quantity X given a σ-finite random quantity Θ is defined by

$$\Pr_X^\theta(A) = \Pr^\theta(X \in A). \tag{6}$$

Equation (6) can be seen as the second hero in this story as it generalizes equation (2). The assumption of a fixed underlying space Ω ensures that both X and Θ are identified with functions on Ω, and this makes it possible to use equation (6) as a definition since $\Pr^\theta$ is a law on Ω by equation (5) and $(X \in A)$ is an event in Ω. It follows in particular that $\Pr_X^\theta(\Omega_X) = \Pr^\theta(X \in \Omega_X) = \Pr^\theta(\Omega) = 1$, so $\Pr_X^\theta$ is a probability law.

The resulting family $\{\Pr_X^\theta\}$ of conditional laws in (6) corresponds exactly to what is usually referred to as a *statistical model*, where X is the observation and θ is the model parameter. As noted in the introduction, Ω_X and Ω_Θ are usually subsets of finite dimensional vector spaces,

but non-parametric analysis is included since any space equipped with a family of events is allowed. In non-Bayesian analysis the family $\{\Pr_X^\theta\}$ is usually specified. Bayesian analysis requires the additional specification of the law $\Pr_\Theta$, which is then the *prior* law for Θ.

The important point is that in the framework defined here, the laws for X for given parameter values θ are always probability laws. On the other hand, prior laws are allowed to be improper, but are required to be σ-finite.

Having thus seen how the *model* and the *prior law* are represented in a natural way in our framework, it remains to consider the *posterior law*. In the notation introduced above, the posterior law of Θ given $X = x$ is simply $\Pr_\Theta^x$. This is well defined if X is σ-finite.

It is important to stress that the required σ-finiteness of X is not guaranteed, and hence has to be checked in each case. This is in effect precisely what is done in papers proving posterior propriety. Berger et al. (2005, p.617) give good examples which demonstrate that posterior propriety and admissibility properties are determined by a study of the marginal law of the data X. Their proofs of propriety are actually also proofs of the σ-finiteness of the marginal law $\Pr_X$ of the data X. Eaton (2004) considers explicitly cases with a σ-finite X, and explains a most interesting characterization of admissibility in terms of recurrence of an associated symmetric Markov chain (Hobert et al., 2007).

The framework presented here gives the probably most general result available regarding posterior propriety: *Posterior propriety is ensured if the marginal law of the data is σ-finite.* This is an "if and only if result" in the sense that the posterior law is even not defined when the data has a marginal law which is not σ-finite. Properness of a prior law can hence be defined by σ-finiteness, and for Bayesian analysis by the additional property that the resulting marginal law of the data is σ-finite.

7. ARE MARGINALS UNIQUE?

Yes, marginal laws are uniquely determined by the joint law. As explained previously, the hero in our story, equation (2) has the law of propagation of laws equation (4) as a consequence, and the uniqueness of the marginal law in equation (3) is a special case of this.

It was also mentioned earlier that the answer is no if the alternative axioms presented by

Hartigan (1983) are chosen as a basis instead. A consequence is hence that Hartigan abandons the law of propagation of laws.

The difference between the two alternative sets of axioms will be elaborated further by a reconsideration of three illustrative examples found in the literature. The first is our favorite, and the next two are included as promised at the end of Section 3.

Gelfand and Sahu (1999, p.250) give the following example to demonstrate possible problems with marginal laws for improper priors. The example demonstrates nicely how seemingly reasonable arguments may give contradictions if there is no reference to an underlying theory. It also demonstrates some further properties of the theory.

Let (X, Y) be a random vector with uniform law on $\Omega_{X,Y} = \{1, 2\} \times \mathbb{N}$. We assume then that $\Pr_{X,Y}(\{x, y\}) = 1$ for all points (x, y) in $\Omega_{X,Y}$. Intuitively, according to Gelfand and Sahu, it seems that the marginal law of X is the uniform probability on the two point set $\Omega_X = \{1, 2\}$.

Define $(U, V) = \phi(X, Y)$, where $\phi(1, y) = (1, y)$, $\phi(2, 2y) = (2, y)$, and $\phi(2, 2y-1) = (3, y)$. This gives a one-one mapping from $\Omega_{X,Y}$ onto $\Omega_{U,V} = \{1, 2, 3\}\times\mathbb{N}$. The uniform law for (X, Y) is mapped into a uniform law for (U, V) by the rule for propagation of laws in equation (4). As explained earlier, this is due to the hero in our story: equation (2).

The intuition referred to above now gives that the marginal law of U is the uniform probability on $\Omega_U = \{1, 2, 3\}$. The event $(X = 1)$ equals the event $(U = 1)$, but the above intuitive argument gives the seemingly paradoxical result:

$$\text{“}1/2 = \Pr(X = 1) = \Pr(U = 1) = 1/3\text{“}.$$

It will now be explained how this marginalization paradox can be resolved.

As previously explained, every unconditional law is improper if Pr is improper: If Z is a random quantity, then $\Pr_Z(\Omega_Z) = \Pr(\Omega) = \infty$. This holds in particular for the random variables X, Y, U, V in the example. Intuitively, and this intuition can be proved to be correct, the laws of X , Y, U, and V are all uniform by symmetry. Since both Ω_X and Ω_U are finite this implies that $\Pr_X\{x\} = \infty = \Pr_U\{u\}$. The event $(X = 1)$ equals the event $(U = 1)$, and the particular result $\Pr(X = 1) = \infty = \Pr(U = 1)$ partly resolves the previous marginalization paradox.

Both X and U exemplify existence of variables with laws of the form $\infty \cdot \mu$, where μ is counting measure. The theory presented by Hartigan (1983) does not allow laws like this, but we are forced to do so by insisting on the validity of equation (2).

The variables X and U are hence not σ-finite, but the variables Y and V are σ-finite: $\Pr(Y = y) = \sum_x \Pr(Y = y, X = x) = 2$, and similarly $\Pr(V = v) = 3$. The σ-finiteness ensures that both conditional laws $\Pr^y$ and $\Pr^v$ exist, and a calculation gives that $\Pr^y_X$ and $\Pr^v_U$ are uniform probability laws on $\{1, 2\}$ and $\{1, 2, 3\}$, respectively. The result is

$$1/2 = \Pr^y(X = 1) \neq \Pr^v(U = 1) = 1/3,$$

which is not paradoxical at all.

The elementary marginalization paradox given by the Gelfand and Sahu (1999, p.250) example has now been explained, and there is no paradox left. The more famous marginalization paradoxes presented by Stone and Dawid (1972) and Dawid et al. (1973) can be given a similar non-paradoxical explanation. A more recent and important example related to the correlation coefficient in a bivariate normal distribution is presented by Berger and Sun (2008). The marginalization paradox in that case can also be explained as above.

An alternative view of the marginalization paradoxes is given by Hartigan (1983) and Chang and Pollard (1997), but their conclusions are also that the seemingly paradoxical results are explained as non-paradoxical. We hope that the readers can appreciate that there really are some advantages of having a well defined theory as a basis for argumentation. These advantages are shared by the approach presented here and the alternative approach of Hartigan. This can be contrasted with the more loose approaches which are commonly followed.

Hartigan (1983, p.23) considers factorizations of the form $p(x, y) = p(x \,|\, y)p(y)$ for a law $\Pr(X = x, Y = y) = p(x, y)$ on $\Omega_{X,Y} = \mathbb{N} \times \mathbb{N}$. The uniform case $p(x, y) = 1$ gives the law $p(y) = \infty$ in our theory, and the conditional law $p(x \,|\, y)$ does not exist. Hartigan takes the opposite point of view and requires that the factorization $p(x, y) = p(x \,|\, y)p(y)$ remains valid, and concludes that the marginal law $p(y)$ can be specified arbitrarily as long as the factorization remains valid. The particular choice $p(y) = 1$ gives $p(x \,|\, y) = 1$. It can also be

noted from this example that the conditional density $p(x \mid y)$ in the sense of Hartigan needs not be a probability density.

As described earlier, Chang and Pollard (1997, p.308) considered a quantity (X, Y, Z) with uniform law on $(0,1)^2 \times \mathbb{R}$, and concluded that the marginal law of (X, Y) could be rather arbitrary. This can be understood in the sense of Hartigan since there exist many factorizations of the form $\Pr_{X,Y,Z}(dx, dy, dz) = g(z \mid x, y) f(x, y)\, dxdydz$ with $g(z \mid x, y) f(x, y) = 1$.

8. DISCUSSION

The most common case in applications is given by $\Pr_X^{\theta}(dx) = f(x \mid \theta)dx$ and $\Pr_{\boldsymbol{\theta}}(d\theta) = \pi(\theta)d\theta$. Equation (2) implies $\Pr_X(A) = \int_A f(x)dx$, where $f(x) = \int f(x \mid \theta)\pi(\theta)d\theta$. Furthermore, it follows from the uniqueness of the conditional law defined by equation (5) that the posterior law $\Pr_{\Theta}^{x}$ can be computed from (1) as promised, but X must be σ-finite.

It should be noted that ∞ is in general a possible value for $f(x)$. It can be proved that a necessary and sufficient condition for σ-finiteness of $\Pr_X$ is that $f(x) < \infty$ for (almost) all x. This condition is the one that is usually looked for in Bayesian applications, and the theory presented here gives a theoretical basis for this.

We have already emphasized that all marginal laws in the setup are improper, while all conditional laws are probability laws. If we would still like to include a finite law as a prior, then we need to define it as a conditional law given some other parameter. Depending on the context this can be referred to as a hyper-parameter.

It was explained in the previous section that the mere existence of a factorization $\Pr_{X,Y}(dx, dy) = f(x \mid y) f(y)\, dxdy$ is not sufficient to conclude that $f(y)$ is the marginal density of Y, and that a conditional density $f(x \mid y)$ exists. This view is shared by Halmos and Savage (1949, p.230) as they explain in a footnote. Consider in particular the case where the prior law of the parameter (Θ_1, Θ_2) is given by the product law $\pi_1(\theta_1)d\theta_1\, \pi_2(\theta_2)d\theta_2$. It does not follow generally that $\pi_1(\theta_1)d\theta_1$ is the law of Θ_1; this follows only if π_2 is a probability density. In that case it also follows that π_2 is the conditional density of Θ_2 given $\Theta_1 = \theta_1$.

The discussion given by Berger and Sun (2008) demonstrates the fact that the Bayesian algorithm with improper priors is of considerable importance also in non-Bayesian analysis.

Incidentally, the original motivation for the work presented here gives another example of the usage of improper priors in a strictly non-Bayesian problem: Improper priors are essential in the algorithm for simulation from the conditional distribution given a sufficient statistic as presented by Lindqvist and Taraldsen (2005), with calculation of exact p-values as one main example.

Renyi (1970) introduced and motivated the concept of conditional probability spaces independently of the previous arguments. His initial motivation is not given by statistical inference, but rather the intuition that conditional probability is the fundamental concept. Consequently he gives a definition of a conditional probability space based on a family of objects $\Pr(A \mid B)$. This is in line with the arguments given by Jeffreys (1961), in which all probabilities are conditional probabilities. Renyi's theory is, however, not too well known by statisticians.

An elaboration of the close connection between the theory of Renyi and the approach here requires the precise language of measure theory, and will not be explained further. It follows from this elaboration, however, that the theory of conditional probability spaces naturally leads to the modification of the axioms of probability presented in Section 2.

It can hence be concluded that the theory of Renyi is a generalization of the theory of Kolmogorov which gives a natural theoretical frame for the formulation of general statistical models. A characteristic feature of this frame is that both parameters and observations are represented by functions defined on a common underlying space Ω. Because of this we will suggest that the term *conditional probability space* is used for Ω in the case described in Section 2. The case $\Pr(\Omega) = \infty$ where Pr is a σ-finite law is allowed, and $\Pr(\Omega) = \infty$ is necessary if improper priors are to be included.

References

Berger, J. O., Strawderman, W., and Tang, D. (2005), "Posterior Propriety and Admissibility of Hyperpriors in Normal Hierarchical Models," *The Annals of Statistics*, 33, 606–646.

Berger, J. O. and Sun, D. (2008), "Objective Priors for the Bivariate Normal Model," *The Annals of Statistics*, 36, 963–82.

Chang, J. and Pollard, D. (1997), "Conditioning as Disintegration," *Statistica Neerlandica*, 51, 287–317.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973), "Marginalization Paradoxes in Bayesian and Structural Inference," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 35, 189–233.

Eaton, M. L. (1982), "A Method for Evaluating Improper Prior Distributions." in *Statistical Decision Theory and Related Topics III*, eds. Gupta, S. S. and Berger, J. O., New York: Academic Press, pp. 329–352.

— (2004), "Evaluating Improper Priors and the Recurrence of Symmetric Markov Chains: an Overview," in *A Festschrift for Herman Rubin*, ed. DasGupta, A., Beachwood, Ohio, USA: Institute of Mathematical Statistics, pp. 5–20.

Gelfand, A. E. and Sahu, S. K. (1999), "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," *Journal of the American Statistical Association*, 94, 247–253.

Halmos, P. (1950), *Measure Theory*, New York: Springer-Verlag (1974).

Halmos, P. and Savage, L. J. (1949), "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics," *Annals of Mathematical Statistics*, 20, 225–241.

Hartigan, J. (1983), *Bayes theory*, New York: Springer.

Hobert, J. P., Tan, A., and Liu, R. (2007), "When is Eaton's Markov Chain Irreducible?" *Bernoulli*, 13, 641–652.

Jeffreys, H. (1961), *Theory of Probability*, Oxford: Clarendon Press.

Kolmogorov, A. (1933), *Foundations of the Theory of Probability*, New York: Chelsea Publishing (1956).

Lehmann, E. (1959), *Testing Statistical Hypotheses*, New York: Springer (1997).

Lindqvist, B. H. and Taraldsen, G. (2005), "Monte Carlo Conditioning on a Sufficient Statistic," *Biometrika*, 92, 451–464.

Renyi, A. (1970), *Foundations of Probability*, Amsterdam: North-Holland.

Rudin, W. (1987), *Real and Complex Analysis*, New York: McGraw-Hill.

Schervish, M. (1995), *Theory of Statistics*, New York: Springer.

Stone, M. and Dawid, A. P. (1972), "Un-Bayesian Implications of Improper Bayes Inference in Routine Statistical Problems," *Biometrika*, 59, 369–375.