NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Bayes Theorem for Improper Priors

by

Gunnar Taraldsen and Bo Henry Lindqvist

PREPRINT
STATISTICS NO. 4/2007

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2007/S4-2007.pdf

Bayes Theorem for Improper Priors

Gunnar Taraldsen and Bo Henry Lindqvist
SINTEF and NTNU, Trondheim, Norway

April 15, 2007

Abstract

Improper priors are used frequently, but often formally and without reference to a sound theoretical basis. A consequence is the occurrence of seemingly paradoxical results. The most famous example is perhaps given by the marginalization paradoxes presented by Stone and Dawid (1972). It is demonstrated here that the seemingly paradoxical results are removed by a more careful formulation of the theory.

The present paper demonstrates more generally that Kolmogorov's (1933) formulation of probability theory admits a minimal generalization which includes improper priors and a general Bayes theorem. It is interesting that the resulting theory is closely related to the theory formulated by Renyi (1970), but the initial axioms and the motivation differ.

The resulting theory includes improper priors, explains the marginalization paradoxes, and gives conditions which ensure propriety of the resulting posteriors. These results are relevant for the current usage of improper priors.

KEY WORDS: Marginalization, Paradox, Axioms of probability, Propriety of posterior, Sigma-finite, Conditional

1 INTRODUCTION

Berger (1985, p.90) argues that use of non-informative improper priors represents the single most powerful method of statistical analysis. Improper priors are indeed used frequently in Bayesian analysis. The motivation is that they are natural choices for the expression of absence of knowledge (Bayes, 1763; Laplace, 1812; Jeffreys, 1966). It can be viewed as an attempt at making Bayesian analysis objective (Berger, 2006). Jeffreys (1966, p.118) argues that improper priors are necessary from a principal point of view as the first initial prior in a chain of distributions obtained from Bayes formula.

In certain applications it is reasonable to have an analysis which is invariant with respect to choices of measurement scale, or more general group actions. The conclusion is then that the prior must be a Haar measure, and this is often improper. A maximum entropy argument is sometimes intuitively appealing, and this also tends to lead to improper priors (Jaynes, 2003; Berger, 1985). In practical applications it may be difficult to decide on a particular prior distribution, and this typically leads to the choice of a standard improper prior. Finally, and this is most important, use of improper priors can be used to obtain excellent frequentist procedures (Bayard and Berger, 2004). It can be concluded that improper priors are here to stay.

The widespread use of improper priors in practice stands in strong contrast to the theoretical treatment of improper priors in standard textbooks. Berger (1985, p.132) indicates that improper priors can be viewed as limits of proper priors, but concludes: The resulting formal posterior distribution cannot rigorously be considered to be a posterior distribution. The usual approach is to *do the calculations with the improper prior as if it were a proper prior.* Schervish (1995, p.20) use this approach, but notes that it is not a very precise recipe. The conclusion seems to be that many standard textbooks in Bayesian analysis rely on the use of improper priors, but fail to include improper priors in the fundamental description of the theory.

This conclusion is not satisfactory from a theoretical point of view. The possible practical consequences are perhaps even more disturbing. One example is given by the use of improper priors in Markov chain Monte Carlo methods (Gelfand and Sahu, 1999), and a possible consequence is that the resulting posterior is improper. Propriety of the resulting priors is a fundamental question. Hobert and Casella (1996) discuss this in more detail with examples, and give conditions which ensure propriety for an important class of models.

The marginalization paradoxes presented by Stone and Dawid (1972) give additional doubt about the use of improper priors. In a discussion of the marginalization paradoxes the prominent Bayesian D. V. Lindley concludes (Dawid et al., 1973, p.218): *Clearly after tonight's important paper, we should use improper priors no longer. The paradoxes displayed here are too serious to be ignored and impropriety must go. Let me personally retract the ideas contained in my own book.*

The aim in the following is to present the essential ingredients in a minimal theory which explains and avoids the problems encountered above. The easy solution is to avoid improper distributions altogether, but it turns out that improper priors can be included by a slight adjustment of the Kolmogorov axioms. This minimal extension of the Kolmogorov theory can be viewed as a special case of the more general theory presented by Hartigan (1983), and the results presented in the following supplements this theory. The original marginalization paradoxes, and a more recent example, will be presented first for further motivation and ease of reference.

2 THREE MARGINALIZATION PARADOXES

Stone and Dawid (1972) presented two most interesting marginalization paradoxes which will be briefly presented here. A third more recent example (Berger, 2006) will also be discussed.

The reader is encouraged to consult the original presentation for further details and a supplementary discussion. Dawid et al. (1973); Bernardo (1979); Kass and Wasserman (1996); Jaynes (2003); Chang and Pollard (1997), and Bernardo and Ramon (1998) provide further references and discussion of the marginalization paradoxes.

Example 1 Let X and Y be independent exponentially distributed variables with means λ and μ. The parameter of interest is the ratio $r = \lambda/\mu$ of the means. Assume a prior distribution $\pi(r)drd\mu$, where π is a proper density. Bayes formula gives the posterior density $f(r, \mu \,|\, x, y)$, and integration over μ gives the required posterior density: $f(r \,|\, x, y) \propto \pi(r)r/(r + y/x)^3$.

The above inference depends on the data only through the ratio $z = y/x$. This is intuitively appealing, but also suggests an alternative route for the calculation. The density of z depends on r alone, and a calculation gives $f(z \,|\, r) \propto r/(r+z)^2$. Combined with the prior density π this gives the posterior density: $f(r \,|\, z) \propto \pi(r)r/(r + z)^2$.

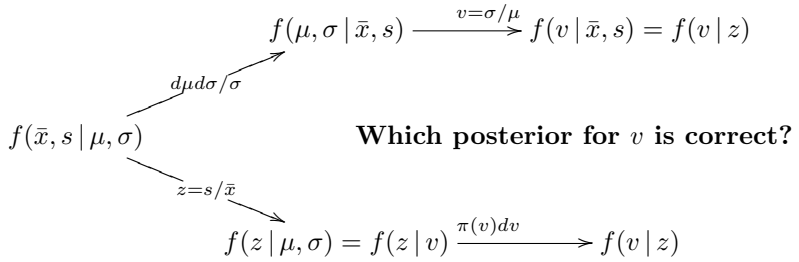
The example is summarized by the diagram:

$$
\begin{array}{ccc}
 & & f(r, \mu \,|\, x, y) \longrightarrow f(r \,|\, x, y) \propto \pi(r)r/(r+z)^3 \\
 & \overset{\pi(r)drd\mu}{\nearrow} & \\
f(x, y \,|\, r, \mu) & & \textbf{Which posterior for } r \textbf{ is correct?} \\
 & \underset{z=x/y}{\searrow} & \\
 & & f(z \,|\, r, \mu) = f(z \,|\, r) \xrightarrow{\pi(r)dr} f(r \,|\, z) \propto \pi(r)r/(r+z)^2
\end{array}
$$

Example 2 Let $X = (X_1, \ldots, X_n)$ be a random sample from a normal distribution with mean μ and variance σ^2. The parameter of interest is the coefficient of variation $v = \sigma/\mu$. Assume that the prior distribution is the conventional right-Haar measure $d\mu d\sigma/\sigma$. A standard calculation gives the posterior density $f(v \,|\, x)$.

The density $f(v \,|\, x)$ depends on the data only through the empirical coefficient of variation $z = s/\bar{x}$. This suggests that it should be possible to base the analysis directly on Z. The distribution of Z depends only on v, and a calculation gives $f(z \,|\, v)$ explicitly. Unfortunately,

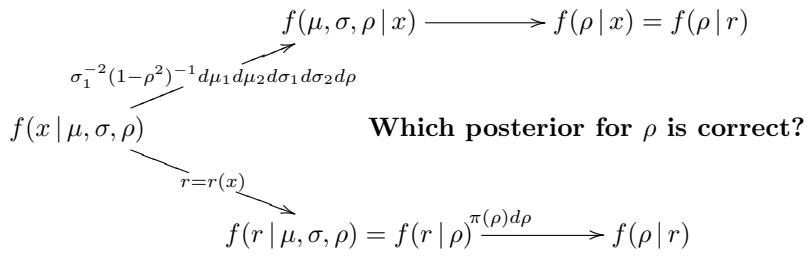
$f(z \mid v)$ as a function of v is not a factor of the expression calculated for $f(v \mid x)$. This example can also be summarized by a diagram:



The conclusion is that it is impossible to find a prior density $\pi(v)$ for v such that the two methods of attack give the same posterior density. The first argument gives a posterior which depends on the observation only through z, but this posterior is not obtainable from the distribution of Z.

Example 3 Let $X = (X_1, \ldots, X_n)$ be a random sample from a bivariate normal distribution with mean $\mu = (\mu_1, \mu_2)$ and 2×2 covariance matrix σ^2. The parameter of interest is the correlation coefficient $\rho = \sigma_{12}^2/(\sigma_1\sigma_2)$. Let the prior distribution be the right-Haar measure $\sigma_1^{-2}(1-\rho^2)^{-1}d\mu_1 d\mu_2 d\sigma_1 d\sigma_2 d\rho$ corresponding to the triangular group. This determines the posterior density $f(\rho \mid x)$.

The density $f(\rho \mid x)$ depends on the data x only through the empirical correlation r. This suggests that it should be possible to base the analysis directly on r. The distribution of r depends only on ρ, and a prior density $\pi(\rho)$ determines $f(\rho \mid r)$. Unfortunately, there exist no prior density $\pi(\rho)$ such that the two arguments give the same posterior. This holds in particular for the natural candidate $\pi(\rho) = (1-\rho^2)^{-1}$. This example can be summarized by a diagram:



This example is similar in principle to the previous two examples, but there are some additional features of the posterior obtained from the first line of argument.

1. The posterior equals the fiducial distribution found by Fisher.
2. The posterior is a confidence distribution: The Bayesian credible sets give exact confidence intervals.

These properties demonstrate that the first posterior is a natural candidate for inference. This posterior depends only on the empirical correlation, but it is not obtainable from the distribution of the empirical correlation and Bayes theorem.

3 KOLMOGOROV REVISITED

Williams (1991, p.23) gives the following intuitive interpretation of an experiment modeled by a random quantity $\boldsymbol{x}$: Tyche, Goddess of Chance, chooses an elementary event ω in Ω according to the law P. The observed result of the experiment is $x = \boldsymbol{x}(\omega)$ in $\Omega_{\boldsymbol{x}}$. The law of the experiment is $\mathrm{P}_{\boldsymbol{x}}$.

Different random quantities are given by different functions, but the set Ω of elementary events with the law P is fixed. The distribution of a random quantity is defined by (Kolmogorov, 1956, p.21)

$$\mathrm{P}_{\boldsymbol{x}}(A) = \mathrm{P}(\boldsymbol{x} \in A) \tag{1}$$

where $(\boldsymbol{x} \in A) := \{\omega \,|\, \boldsymbol{x}(\omega) \in A\}$. Mathematicians, including Kolmogorov, use the notation $\boldsymbol{x}^{-1}(A)$ instead of $(\boldsymbol{x} \in A)$. The classic book by Doob (1990) is recommended for further explanation and motivation for the assumption of a fixed underlying probability space (Ω, P), and natural generalizations of the notation $(\boldsymbol{x} \in A)$.

In Bayesian statistics the observations and the parameters are both modeled as random quantities. A parameter $\boldsymbol{\theta}$ is hence also a function $\boldsymbol{\theta} : \Omega \to \Omega_{\boldsymbol{\theta}}$. *This observation, or rather, choice of definition, is the point of departure from more conventional textbook definitions.* Most of the following are simple mathematical consequences of this definition.

Assume next that the distribution of the parameter $\boldsymbol{\theta}$ is improper. This means that $\mathrm{P}_{\boldsymbol{\theta}}(\Omega_{\boldsymbol{\theta}}) = \infty$, and since this equals $\mathrm{P}(\Omega)$ it follows that P is also unbounded.

Improper priors used in modeling are most often given by densities with respect to either a counting measure or n-dimensional Lebesgue measure. These distributions are σ-finite, so it is natural to assume that the prior $\mathrm{P}_{\boldsymbol{\theta}}$ is σ-finite. This means that $\Omega_{\boldsymbol{\theta}} = A_1 \cup A_2 \cup \cdots$, where $\mathrm{P}_{\boldsymbol{\theta}}(A_i) = \mathrm{P}(\boldsymbol{\theta} \in A_i) < \infty$. The additional observation $\Omega = (\boldsymbol{\theta} \in \cup_i A_i) = \cup_i (\boldsymbol{\theta} \in A_i)$ gives that P is also σ-finite.

The conclusion so far is that the existence of a parameter $\boldsymbol{\theta}$ with a σ-finite unbounded distribution leads to a σ-finite unbounded P. The normalization of P in the Kolmogorov axioms is hence replaced by the more general assumption of σ-finiteness. This is the minimal generalization of the Kolmogorov axioms referred to in the Abstract.

4 BAYES THEOREM

The introduction of the concept of independence, and more generally the concept of conditional distributions, can be regarded as the point where measure theory becomes probability theory (Kolmogorov, 1956, p.8). The assumption that $\boldsymbol{\theta}$ is σ-finite has very convenient consequences: The conditional distribution given $\boldsymbol{\theta} = \theta$ exists, is unique, and is normalized. A sketch of the proof is as follows.

Indeed, a desired property of the conditional distribution would be

$$\mathrm{P}(A \cap (\boldsymbol{\theta} \in B)) = \int_B \mathrm{P}(A \,|\, \boldsymbol{\theta} = \theta)\, \mathrm{P}_{\boldsymbol{\theta}}(d\theta) \tag{2}$$

and this can be taken as the defining property of $\mathrm{P}^{\theta}(A) = \mathrm{P}(A \,|\, \boldsymbol{\theta} = \theta)$. But then, since $\mathrm{P}_{\boldsymbol{\theta}}$ is σ-finite, the Radon-Nikodym theorem (Halmos, 1950) states exactly that the function $\theta \mapsto \mathrm{P}(A|\boldsymbol{\theta} = \theta)$ is uniquely defined. This follows since equation (2) can be used to identify $g(\theta) = \mathrm{P}(A \,|\, \boldsymbol{\theta} = \theta)$ as the density of the measure $\mu(B) = \mathrm{P}(A \cap (\boldsymbol{\theta} \in B))$ with respect to $\mathrm{P}_{\boldsymbol{\theta}}$. The required absolute continuity is fulfilled since $\mathrm{P}_{\boldsymbol{\theta}}(B) = 0$ implies $\mathrm{P}(A \cap (\boldsymbol{\theta} \in B)) = 0$. This is a consequence of $A \cap (\boldsymbol{\theta} \in B) \subset (\boldsymbol{\theta} \in B)$. The normalization follows from the case $A = \Omega$.

The conditional distribution of a random quantity $\boldsymbol{x}$ can be defined by $\mathrm{P}^{\theta}_{\boldsymbol{x}}(A) = \mathrm{P}^{\theta}(\boldsymbol{x} \in A)$. In summary the notational conventions here are that subscripts indicate the distributions of random quantities, and superscripts indicate conditions. Kolmogorov used a similar convention, but reversed the role of superscripts and subscripts. The choice of the opposite convention here is dictated by the very common usage of the notation P_X, F_X, f_X, and similar for other important quantities derived from a random quantity X.

A statistical experiment is modeled by a random quantity $\boldsymbol{x}$, but conditionally given $\boldsymbol{\theta} = \theta$. The conditional distribution $\mathrm{P}^{\theta}_{\boldsymbol{x}}$ is usually referred to as the model, and the distribution $\mathrm{P}_{\boldsymbol{\theta}}$ is the prior. In a concrete model the family $\mathrm{P}^{\theta}_{\boldsymbol{x}}$ is specified, and this is the starting point for conventional statistics (Lehmann and Casella, 1998). Bayesian statistics (Berger, 1985) require the additional

specification of the distribution $\mathrm{P}_{\boldsymbol{\theta}}$. *The result so far is a common theoretical frame for both Bayesian and conventional statistics which includes improper distributions.*

The model and the prior determine the distribution of the random quantity $(\boldsymbol{x}, \boldsymbol{\theta})$ by $\mathrm{P}_{\boldsymbol{x},\boldsymbol{\theta}}(dx, d\theta) = \mathrm{P}^{\theta}_{\boldsymbol{x}}(dx)\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta)$. This notation means in particular that if $A \subset \Omega_{\boldsymbol{x}}$ and $B \subset \Omega_{\boldsymbol{\theta}}$, then

$$\begin{aligned} \mathrm{P}_{\boldsymbol{x},\boldsymbol{\theta}}(A \times B) &= \int_B [\int_A \mathrm{P}^{\theta}_{\boldsymbol{x}}(dx)]\ \mathrm{P}_{\boldsymbol{\theta}}(d\theta) \\ &= \int_B \mathrm{P}^{\theta}_{\boldsymbol{x}}(A)\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta) \end{aligned} \tag{3}$$

Again, the Radon-Nikodym theorem gives that the conditional distribution $\mathrm{P}^{\theta}_{\boldsymbol{x}}$ exists and is uniquely determined by $\mathrm{P}_{\boldsymbol{x},\boldsymbol{\theta}}$ and the previous factorization if $\mathrm{P}_{\boldsymbol{\theta}}$ is σ-finite. This follows since equation (3) can be used to identify $g(\theta) = \mathrm{P}^{\theta}_{\boldsymbol{x}}(A)$ as the density of the measure $\mu(B) = \mathrm{P}(\boldsymbol{x} \in A, \boldsymbol{\theta} \in B)$ with respect to $\mathrm{P}_{\boldsymbol{\theta}}$. This notation transforms equation (3) into $\mu(d\theta) = g(\theta)\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta)$, and the existence and uniqueness of g is the Radon-Nikodym theorem for σ-finite measures (Halmos, 1950). The required absolute continuity is fulfilled since $\mathrm{P}_{\boldsymbol{\theta}}(B) = 0$ implies $\mathrm{P}_{\boldsymbol{x},\boldsymbol{\theta}}(A \times B) = 0$. This is a consequence of $(\boldsymbol{x} \in A, \boldsymbol{\theta} \in B) \subset (\boldsymbol{\theta} \in B)$.

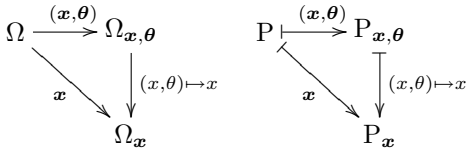
The relation $\mu(d\theta) = g(\theta)\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta)$ can also be written $g(\theta) = \mu(d\theta)/\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta)$. This later formulation indicates a constructive approach for the calculation of a conditional probability as a limit of elementary conditional probabilities. This is actually possible with some care (Rudin, 1987, p.143)(Doob, 1990, p.343).

The conditional $\mathrm{P}^{\boldsymbol{\theta}}_{\boldsymbol{x}}$ is defined above by two different approaches. The first approach relies on $\mathrm{P}^{\boldsymbol{\theta}}$ on Ω, and defines $\mathrm{P}^{\boldsymbol{\theta}}_{\boldsymbol{x}} = (\mathrm{P}^{\boldsymbol{\theta}})_{\boldsymbol{x}}$. The second approach defines $\mathrm{P}^{\boldsymbol{\theta}}_{\boldsymbol{x}}$ directly on $\Omega_{\boldsymbol{x}}$. The uniqueness part of the Radon-Nikodym ensures consistency between these two definitions. There exist simple examples which demonstrates that $\mathrm{P}^{\boldsymbol{\theta}}_{\boldsymbol{x}}$ may exist even if $\mathrm{P}^{\boldsymbol{\theta}}$ does not exist.

Equation (3) with $A = \Omega_{\boldsymbol{x}}$ gives that the conditional distribution must be normalized:

$$\mathrm{P}^{\theta}_{\boldsymbol{x}}(\Omega_{\boldsymbol{x}}) = 1 \tag{4}$$

It should also be noted that the distribution of the random quantity $(\boldsymbol{x}, \boldsymbol{\theta})$ gives the distribution of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$. The proof for $\boldsymbol{x}$ follows from the diagrams



This observation also explains why it is not necessary in applications to specify P, but rather a sufficiently rich family of joint distributions. The distribution $\mathrm{P}_{\boldsymbol{x}}$ is the marginal distribution in a Bayesian theory.

The factorization which defines the conditionals can be applied twice if both $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are σ-finite. This gives

$$\mathrm{P}^{\theta}_{\boldsymbol{x}}(dx)\,\mathrm{P}_{\boldsymbol{\theta}}(d\theta) = \mathrm{P}_{\boldsymbol{x},\boldsymbol{\theta}}(dx, d\theta) = \mathrm{P}^{x}_{\boldsymbol{\theta}}(d\theta)\,\mathrm{P}_{\boldsymbol{x}}(dx) \tag{5}$$

The conclusion is that a prior $\mathrm{P}_{\boldsymbol{\theta}}$ and a model $\mathrm{P}^{\theta}_{\boldsymbol{x}}$ which give a σ-finite marginal $\mathrm{P}_{\boldsymbol{x}}$ determine the posterior $\mathrm{P}^{x}_{\boldsymbol{\theta}}$. *This is the Bayes theorem corresponding to the title of this paper.*

The most common case in applications is given by $\mathrm{P}^{\theta}_{\boldsymbol{x}}(dx) = f(x\,|\,\theta)\mu(dx)$ and $\mathrm{P}_{\boldsymbol{\theta}}(d\theta) = f(\theta)\nu(d\theta)$, where μ and ν are counting measure or n-dimensional Lebesgue measure. The marginal distribution is given by

$$\begin{aligned} \mathrm{P}_{\boldsymbol{x}}(A) &= \mathrm{P}(\boldsymbol{x} \in A, \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}) \\ &= \int\int [x \in A] f(x\,|\,\theta) f(\theta) \nu(d\theta) \mu(dx) \\ &= \int_A f(x) \mu(dx) \end{aligned} \tag{6}$$

and the marginal density is hence

$$f(x) = \int f(x \,|\, \theta) f(\theta) \nu(d\theta) \tag{7}$$

It should be noted that ∞ is a possible value for $f(x)$.

The case $f(x) = 1/|x|$ provides an example where $\boldsymbol{x}$ is σ-finite, but $f(x)$ takes the value ∞ and is furthermore not integrable on $(-\epsilon^2, \epsilon^2)$. The marginal is σ-finite by definition iff there exists a countable partition $\{A_i\}$ of $\Omega_{\boldsymbol{x}}$ with $\mathrm{P}(\boldsymbol{x} \in A_i) < \infty$. A more convenient necessary and sufficient condition for σ-finiteness is that $\mu(f = \infty) = 0$. This holds in general for distributions on the form $f(x)\mu(dx)$, where μ is σ-finite.

If $\boldsymbol{x}$ is σ-finite, then the posterior distribution is given by $\mathrm{P}^x_{\boldsymbol{\theta}}(d\theta) = f(\theta \,|\, x)\nu(d\theta)$, where

$$f(\theta \,|\, x) = \frac{f(x \,|\, \theta) f(\theta)}{f(x)} \tag{8}$$

The proof follows by inspection of

$$\begin{aligned} \mathrm{P}^\theta_{\boldsymbol{x}}(dx)\, \mathrm{P}_{\boldsymbol{\theta}}(d\theta) &= f(x \,|\, \theta) f(\theta) \mu(dx) \nu(d\theta) \\ &= f(x) f(\theta \,|\, x) \mu(dx) \nu(d\theta) \\ &= \mathrm{P}^x_{\boldsymbol{\theta}}(d\theta)\, \mathrm{P}_{\boldsymbol{x}}(dx) \end{aligned} \tag{9}$$

The conclusion is that the elementary version of Bayes theorem remains valid. The key requirement is that the marginal distribution of the data $\boldsymbol{x}$ is σ-finite.

5 THE MARGINALIZATION PARADOXES

Example 2 continued: The marginal $\boldsymbol{x}$ is σ-finite iff $n > 1$. It follows that the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ given $\boldsymbol{x}$ is a well defined proper distribution, and this gives a well defined proper posterior distribution for the coefficient of variation $\boldsymbol{v} = \boldsymbol{\sigma}/\boldsymbol{\mu}$. This can be calculated explicitly as explained by Stone and Dawid (1972).

The alternative calculation suggested by Stone and Dawid (1972) is given by consideration of the conditional distribution given $\boldsymbol{v}$. This is however an invalid approach since $\boldsymbol{v}$ is not σ-finite. The distribution is rather trivial

$$\mathrm{P}(a < \boldsymbol{v} < b) = \mathrm{P}(a < \boldsymbol{\sigma}/\boldsymbol{\mu} < b) = \infty \tag{10}$$

for $a < b$, and more generally $\mathrm{P}_{\boldsymbol{v}}(dv) = \infty dv$. Incidentally, this provides an example of a measure which is not σ-finite, but absolutely continuous with respect to Lebesgue measure. The rules $0 \cdot \infty = 0$ and $a \cdot \infty = \infty$ for $a > 0$ are used here, and elsewhere.

An alternative point of view is that the prior information for the two approaches differs. The first approach assumes $\mathrm{P}_{\boldsymbol{\mu},\boldsymbol{\sigma}}(d\mu, d\sigma) = [\sigma > 0]\sigma^{-1}\, d\mu d\sigma$, and the second approach assumes $\mathrm{P}_{\boldsymbol{v}}(v) = \pi(v)dv$ with a σ-finite $\boldsymbol{v}$. The first assumption leads to a distribution for $\boldsymbol{v}$ which is not σ-finite, and it is hence not compatible with the second approach. The conclusion then is that the two approaches are based on different prior information, and it is not paradoxical that the conclusions differ. This explanation is similar in spirit to the one presented by Jaynes (2003), but the mathematical approach differs.

□

Example 1 continued: The prior distribution $\mathrm{P}_{\boldsymbol{r},\boldsymbol{\mu}}(dr, d\mu) = \pi(r)dr d\mu$ is σ-finite. Integration of the joint distribution over r and μ proves that the marginal $(\boldsymbol{x}, \boldsymbol{y})$ is σ-finite. It follows that the posterior distribution of $(\boldsymbol{r}, \boldsymbol{\mu})$ given $(\boldsymbol{x}, \boldsymbol{y})$ is well defined, and furthermore explicitly given as stated earlier.

The alternative calculation suggested by Stone and Dawid (1972) is given by consideration of the conditional distribution given $\boldsymbol{r}$. This is an invalid approach since $\boldsymbol{r}$ is not σ-finite. The distribution is again rather trivial, and in this case given by $\mathrm{P}_{\boldsymbol{r}}(r) = \infty\pi(r)dr$.

An alternative point of view is that the prior information for the two approaches differ. The first approach assumes $\mathrm{P}_{\boldsymbol{r},\boldsymbol{\mu}}(dr, d\mu) = \pi(r)drd\mu$ and the second approach assumes $\mathrm{P}_{\boldsymbol{r}}(dr) = \pi(r)dr$. These assumptions are not compatible since the first assumption leads to a distribution for $\boldsymbol{r}$ which is not σ-finite, while the second approach assumes that $\boldsymbol{r}$ has a proper distribution. The first approach has an unbounded P and the second approach has a normalized P. There is no reason to expect that the two different priors should give the same posterior.

□

Example 3 continued:

The seemingly conflicting conclusions are explained as in the previous examples. A particular observation is that the prior $\sigma_1^{-2}(1-\rho^2)^{-1}d\mu_1 d\mu_2 d\sigma_1 d\sigma_2 d\rho$ gives the prior density $\infty \cdot (1-\rho^2)^{-1}$ for ρ, and the distribution of the correlation is not σ-finite. It can in particular not be concluded that $(1-\rho^2)^{-1}$ is the density of ρ.

□

A general comment is that if a conditional distribution given x happens to depend on only $z = \phi(x)$, then the conditional distribution coincide with the conditional distribution given z. This theorem can be proven as in the case of proper distributions, but it must be assumed that both x and z have σ-finite distributions. The above distribution of ρ is not σ-finite, and this explains why the conclusion of the theorem fails in the example.

The marginalization paradoxes have now been resolved with reference to an underlying theory.

6 DISCUSSION

A possible and natural interpretation of an improper P is as a relative degree-of-belief given by $\mathrm{P}(A)/\mathrm{P}(B)$. Consideration of this leads naturally to the elementary definition of $\mathrm{P}(A \mid B)$, and to the point-of-view that the conditional probability is the more fundamental concept.

Renyi (1970) starts with an axiomatic definition of a conditional probability space given by objects $\mathrm{P}(A \mid B)$. The family of sets $B \subset \Omega$ is a *bunch*, where the prototype of a bunch is given by the sets which fulfill $0 < \mathrm{P}(B) < \infty$ for a σ-finite P. The structure theorem (Renyi, 1970, p.40) for a conditional probability space and the completeness theorem (Renyi, 1970, p.43) show that a σ-finite P and the elementary definition of $\mathrm{P}(A \mid B)$ give all possible conditional probability spaces.

Renyi (1970, p.73) classifies random variables as regular if they transfer the conditional probability space into a conditional probability space on the real line. It follows that a random variable is regular in the sense defined by Renyi if and only if it is σ-finite as defined here.

In the present paper the need for improper priors is the main practical motivation for the introduction of a σ-finite P. The arguments given by Renyi (1970) give a more fundamental motivation. The resulting theory is similar, but includes now also statistical models of both Bayesian and conventional type.

A direct interpretation of $\mathrm{P}(A)/\mathrm{P}(B)$ in terms of relative frequencies and a law of large numbers fails since a sequence of i.i.d. variables fails to exist in the case of an improper distribution P. This is related to the marginalization paradoxes presented and explained above. Further investigations into this kind of interpretation lead naturally to the consideration of sequences of exchangeable variables and the related law of large numbers. This is most interesting, but will not be discussed further here.

Existence of a σ-finite random quantity implies that the sample space Ω is σ-finite. This is equivalent with the existence of a normalizing random variable. This is a variable $\boldsymbol{n} > 0$ with $E\boldsymbol{n} = 1$ (Rudin, 1987, p.121). The measure μ defined by $\mu(d\omega) = \boldsymbol{n}(\omega)\,\mathrm{P}(d\omega)$ is then a proper probability on Ω. This gives a device for generalization of well known results such as the Radon-Nikodym theorem to the case of unbounded probabilities. The measure μ is equivalent with P in the sense that $\mathrm{P}(A) = 0$ is equivalent with $\mu(A) = 0$.

A normalizing variable can also be useful for the practical problem of Monte Carlo simulation from an improper distribution. A random sequence $x_1, x_2, \ldots$ from the proper distribution $\mu(dx) = n(x)\,\mathrm{P}_{\boldsymbol{x}}(dx)$ gives a weighted random sequence $(x_1, w_1), (x_2, w_2), \ldots$ from $\mathrm{P}_{\boldsymbol{x}}$, where the weight is

given by $w_i = 1/n(x_i)$. The claim $E\phi(X) = \lim[\phi(x_1)w(x_1) + \cdots + \phi(x_n)w(x_n)]/n$ follows from the law of large numbers for μ. It is interesting to notice that weighted random sequences and improper priors are important also in conditional Monte Carlo simulations (Trotter and Tukey, 1956; Lindqvist and Taraldsen, 2005).

The conditional distribution P^x is said to be a regular conditional distribution if it is a measure for almost all x. It is well known that it is impossible to represent a general conditional distribution P^x by a regular conditional distribution (Doob, 1990, p.624). A sufficient condition for the existence of a regular conditional distribution is that Ω is a complete separable metric space (a Polish space) equipped with the Borel field, or it's completion.

Bahadur and Bickel (1968) prove that there always exist a version of the conditional expectation such that $E^t\phi(\boldsymbol{t}, \boldsymbol{x}) = E^t\phi(t, \boldsymbol{x})$. The proof generalizes verbatim to the case of a σ-finite P. The special case $\mathrm{P}^t(\boldsymbol{t} = t) = 1$ is of particular intuitive importance: The conditional distribution is concentrated on $(\boldsymbol{t} = t)$. If $\Omega_{\boldsymbol{x}}$ is a Polish space and $\boldsymbol{t} = \tau(\boldsymbol{x})$, then it can be proven that there exist a regular $\mathrm{P}^t_{\boldsymbol{x}}$ so that $\mathrm{P}^t_{\boldsymbol{x}}(\tau = t) = 1$. It is known that in general it is impossible to find a regular $\mathrm{P}^t_{\boldsymbol{x}}$ so that $\mathrm{P}^t_{\boldsymbol{x}}(\tau = t) = 1$ holds identically for all t (Bahadur and Bickel, 1968, p.378).

Chang and Pollard (1997) provide a discussion of conditional distributions with particular emphasis on the property $\mathrm{P}^t_{\boldsymbol{x}}(\tau = s) = 0$ for $t \neq s$. They allow also σ-finite $\mathrm{P}^t_{\boldsymbol{x}}$. These correspond to the more general concept of conditional distributions considered by Hartigan (1983).

The notation in equation (5) indicates that it should be possible to integrate with respect to $\mathrm{P}^x_{\boldsymbol{\theta}}$. Kolmogorov (1956, p.54) shows that it is possible to extend the conventional definition of the integral to allow integration with respect to conditional distributions in general. This limiting procedure gives a version of the conditional expectation defined as a more general integral and justifies the notation in equation (5) in full generality. It is hence noteworthy that conditional distributions give a more general concept than an indexed family of probability measures.

7 CONCLUSION

It has been demonstrated that a generalized Kolmogorov probability theory which includes improper priors can be developed with relatively little extra effort. The resulting formulation is essentially equivalent with the theory formulated by Renyi, but the motivation is different. The theory includes a general Bayes theorem.

This theory explains the marginalization paradoxes, and may also prove to be of use in connection with other problems such as the development of a general theory for confidence distributions. The results are relevant for Markov chain Monte Carlo methods, and gives in particular a general sufficient condition for propriety of the resulting posteriors.

Improper priors may lead to inadmissible inference procedures (Stone and Dawid, 1972) and paradoxical inference (Taraldsen, 2006). The presented theory does not remove the relevance of these examples and the marginalization paradoxes.

References

Bahadur, R. and P. Bickel (1968). Substitution in conditional expectation. *Ann. Math. Statist. 39*(2), 377–378.

Bayard, M. J. and J. O. Berger (2004). The interplay of bayesian and frequentist analysis. *Statistical Science 19*(1), 58–80.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Trans Roy Soc 53*, 370–418. (reproduced in Biometrika, 45 (3/4), 1958).

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer (second edition).

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis 1*(3), 385–402.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian-inference. *Journal of the Royal Statistical Society Series B-Methodological 41*(2), 113–147.

Bernardo, J. M. and J. M. Ramon (1998). An introduction to bayesian reference analysis: inference on the ratio of multinomial parameters. *Journal of the Royal Statistical Society Series D-the Statistician 47*(1), 101–135.

Chang, J. and D. Pollard (1997). Conditioning as disintegration. *Statistica Neerlandica 51*(3), 287–317.

Dawid, A. P., M. Stone, and J. V. Zidek (1973). Marginalization paradoxes in bayesian and structural inference. *Journal of the Royal Statistical Society Series B-Statistical Methodology 35*(2), 189–233.

Doob, J. L. (1990). *Stochastic Processes*. Wiley.

Gelfand, A. and S. Sahu (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*.

Halmos, P. (1950). *Measure Theory*. Van Nostrand Reinhold.

Hartigan, J. (1983). *Bayes theory*. Springer.

Hobert, J. and G. Casella (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1966). *Theory of probability* (Third ed.). Oxford.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(435), 1343–70.

Kolmogorov, A. (1956). *Foundations of the theory of probability*. Chelsea.

Laplace, P. (1812). *Theorie Analytique des Probabilites*. Paris: Courcier.

Lehmann, E. and G. Casella (1998). *Theory of point estimation*. Springer.

Lindqvist, B. and G. Taraldsen (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika 92*, 451–464.

Renyi, A. (1970). *Foundations of Probability*. Holden-Day.

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

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

Stone, M. and A. Dawid (1972). Un-Bayesian Implications of Improper Bayes Inference in Routine Statistical Problems. *Biometrika 59*(2), 369–375.

Taraldsen, G. (2006). Instrument resolution and measurement accuracy. *Metrologia 43*, 539–44.

Trotter, H. and J. Tukey (1956). Conditional Monte Carlo for normal samples. *Symposium on Monte Carlo Methods.* H.A. Meyer, Ed. Wiley, New York, 64–79.

Williams, D. (1991). *Probability with martingales*. Cambridge.