NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Conditional Probability Spaces

by

Gunnar Taraldsen and Bo Henry Lindqvist

PREPRINT
STATISTICS NO. 6/2007

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2007/S6-2007.pdf

Conditional Probability Spaces

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

28th May 2007

Abstract

Improper priors are used frequently, but often formally and without reference to a sound theoretical basis. The present paper demonstrates that Kolmogorov's (1933) formulation of probability theory admits a minimal generalization which includes improper priors and a general Bayes theorem.

The resulting theory is closely related to the theory of conditional probability spaces formulated by Renyi (1970), but the initial axioms and the motivation differ.

The formulation includes Bayesian and conventional statistics as extreme cases, and suggests that intermediate cases can be considered.

KEY WORDS: Axioms of probability, Propriety of posterior, Sigma-finite, Bayes theorem

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 conventional procedures (Bayard and Berger, 2004). It can safely 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 companion paper (Taraldsen and Lindqvist, 2007) presents similar results, and includes in particular a discussion of the marginalization paradoxes. The purpose here is to give a more precise definition of the theory, and to include some proofs.

2 Conditional probability spaces

The concept of a *conditional probability space* Ω to be defined below is identical with the more standard concept of a *σ-finite measure space*. The first term was introduced by Renyi (1970), and this will be explained below.

Definition 1 (Conditional probability space) *A conditional probability space is a set Ω equipped with a σ-finite measure P defined on a σ-algebra $\mathcal{E}$ of sets in Ω. The members of the σ-algebra are called the events. An event A is an elementary condition if $0 < \mathrm{P}(A) < \infty$. The conditional probability given an elementary condition is defined by*

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \tag{1}$$

A probability space is a conditional probability space such that $\mathrm{P}(\Omega) = 1$. An improper probability is a measure such that $\mathrm{P}(\Omega) = \infty$.

It follows immediately that each elementary conditional probability $\mathrm{P}(\cdot \mid A)$ is a probability measure, and it is in particular normalized: $\mathrm{P}(\Omega \mid A) = 1$. This will be generalized below for more general conditions. General terms and results from probability theory are found in standard textbooks (Rudin, 1987; Doob, 1990; Halmos, 1950; Kolmogorov, 1956). Some terms have been used already above, and some familiarity with measure theory will be assumed also in the following. The following are however so central that they are repeated here:

Definition 2 (σ-algebra) *A σ-algebra is a family $\mathcal{E}$ of subsets of a set Ω which includes Ω and is closed under complements and countable unions. A measurable space is a set equipped with a σ-algebra. The members of the σ-algebra are called the measurable sets.*

Definition 3 (Measure) *A function P is countably additive if $\mathrm{P}(\cup_i A_i) = \sum_i \mathrm{P}(A_i)$ holds for every disjoint countable family $\{A_i\}$ of sets in the domain of P. A measure P is a countably additive function defined on a σ-algebra of sets with $\mathrm{P}(\emptyset) = 0$, and $0 \leq \mathrm{P}(A) \leq \infty$ for all measurable sets A. A measure space is a measurable space equipped with a measure. A measure space is σ-finite if it equals a countable union of measurable sets with finite measure.*

Renyi (1970, Def.2.2.1) gives the following definition:

Definition 4 (Bunch) *A family $\mathcal{B} \subset \mathcal{E}$ of events is a bunch if it is closed under unions, the sure event Ω is a countable union of sets from $\mathcal{B}$, and the empty event $\emptyset$ is not a member of $\mathcal{B}$.*

Renyi (1970, Def.2.2.2) defines the concept of a conditional probability space by a family of conditional probability measures $\{\mathrm{P}(\cdot \,|\, B)\}_{B\in\mathcal{B}}$, where $\mathcal{B}$ is a bunch. His definition is more general than Definition 1, but he proves that every conditional probability space can be extended to a full conditional probability space. The Renyi (1970, Thm.2.2.2) extension coincides with the concept of a conditional probability space given in Definition 1. It can be observed that P is uniquely determined from $\mathrm{P}(\cdot \,|\, B)$ up to a positive constant factor.

The condition of σ-finiteness of P ensures that the set of elementary conditions is a bunch. This gives a natural motivation for the σ-finiteness condition, and it demonstrates that the common condition $\mathrm{P}(\Omega) = 1$ can be too restrictive. Another, and quite different approach (Taraldsen and Lindqvist, 2007), will also show that the condition of σ-finiteness is natural and that the condition $\mathrm{P}(\Omega) < \infty$ can be too restrictive. The required theory will be presented next.

3 Random quantities and statistics

The main motivation historically, and also to day, for the concept of probability is its use in modeling. In statistics the first part is given by the assumed identification of some observed results with the realization of a random quantity. It is assumed that the random quantity has a distribution, but this distribution is unknown. Roughly speaking, the purpose of probability theory is to characterize the outcome and consequences of the experiment when the distribution is known, but in statistics the purpose is to characterize the distribution based on the observations.

The standard parametric set-up is to assume that the experiment is described by a family $\{\mathrm{P}_X^\theta\}$ of distributions indexed by a parameter θ. This is referred to as the statistical model. The quantity X corresponds to the observations, and the purpose is to characterize the unknown parameter θ.

In Bayesian statistics it is assumed that the parameter is also a random quantity. The parameter values are not observed, but it is assumed that the parameter has a known distribution $\mathrm{P}_{\boldsymbol{\theta}}$. This prior distribution together with the statistical model gives the joint distribution of $(X, \boldsymbol{\theta})$, and hence also the posterior distribution $\mathrm{P}_{\boldsymbol{\theta}}^x$ of the parameter given the data.

In applications it is often required to consider cases where $\mathrm{P}_{\boldsymbol{\theta}}$ is improper, but then the standard probability theory is not sufficient. The common approach is to simply calculate $\mathrm{P}_{\boldsymbol{\theta}}^x$ by common rules and ignore that there is a lack of underlying theory for these rules. It will be shown here that the conditional probability space gives an underlying theory, and that the above conventional and Bayesian set-up follows as consequences. The theory gives also the correct rules for the calculation of $\mathrm{P}_{\boldsymbol{\theta}}^x$, but some 'common rules' fail.

Definition 5 (Random quantity) *A random quantity X in a set Ω_X is a measurable function $X : \Omega \to \Omega_X$ from the conditional probability space Ω into the measurable space Ω_X. The distribution P_X of X is defined by $\mathrm{P}_X(A) = \mathrm{P}(X \in A)$, where $(X \in A) = \{\omega \,|\, X(\omega) \in A\}$. The random quantity is called σ-finite if its distribution is σ-finite. A random variable X is a random quantity such that Ω_X is the set of real numbers equipped with the Borel σ-field. The expectation of a random variable is defined by $EX = \int X(\omega)\,\mathrm{P}(d\omega)$ if it exists. A random variable X is locally integrable if $\int_A |X(\omega)|\,\mathrm{P}(d\omega) < \infty$ for all events A with $\mathrm{P}(A) < \infty$.*

If X is σ-finite, then it follows that (Ω_X, P_X) is a conditional probability space. The previous definition makes perfect sense also when P is only required to be a measure, but it follows then that P is σ-finite if there exist a σ-finite random quantity. The standard probabilistic set-up is to assume that there is a fixed underlying probability space, and that every random quantity is based on this space. The generalization here is simply given by a replacement of the probability space by a conditional probability space as suggested by Renyi. The reward of the standard set-up, and the set-up here, is that the joint distribution of any family of random quantities is well defined. The definition of more advanced random quantities such as random functions, random sets, or random linear operators are naturally given. A main issue of discussion in these case are the definition, and characterization of the family of events in Ω_X. An additional issue here in the case of an underlying conditional probability space is to prove σ-finiteness.

It is quite natural to identify the prior distribution $\mathrm{P}_{\boldsymbol{\theta}}$ in Bayesian statistics with the distribution of a random quantity $\boldsymbol{\theta} : \Omega \to \Omega_{\boldsymbol{\theta}}$. It follows as a consequence that P is improper if $\mathrm{P}_{\boldsymbol{\theta}}$ is improper. The prior distribution in applications are usually σ-finite, and as noted above this leads to a σ-finite P.

The following two results give characterizations of σ-finiteness which are most useful. Some applications are discussed by Taraldsen and Lindqvist (2007).

Proposition 1 *A measure μ is σ-finite if and only if the measure $\nu(dx) = n(x)\mu(dx)$ is a probability measure for some measurable function $n > 0$.*

Proof. Let $A_k = (n > 1/k)$. It follows that $\Omega_X = \cup_k A_k$ because $n > 0$. The equalities $1 \geq \nu(A_k) \geq (1/k)\mu(A_k)$ give $\mu A_k \leq k < \infty$, and proves that μ is σ-finite.

The second part of the proof is a slight modification of the proof given by Rudin (1987, Lemma 6.9): Let $A_1, A_2, \ldots$ be a countable partition of Ω_X such that $0 < \mu A_k < \infty$. Define $n(x) = \sum_k [x \in A_k]/(2^k \mu A_k)$. It follows that $n > 0$, and that $\nu(dx) = n(x)\mu(dx)$ is a probability measure. □

Proposition 2 *Let $\mu(dx) = f(x)\nu(dx)$, where the measure ν is σ-finite and the measurable function f is positive: $f \geq 0$. The measure μ is σ-finite if and only if $\nu(f = \infty) = 0$.*

Proof. The σ-finiteness of μ follows if it can be proven that the restriction to $(0 < f < \infty)$ is σ-finite since $\mu(0 < f < \infty)^c = 0$. It can and will hence be assumed that $\Omega_X = (0 < f < \infty)$ in the first part of the proof. Let $m > 0$ be such that $m(x)\nu(dx)$ is a probability measure. It follows that $n(x)\mu(dx)$ with $n = m/f > 0$ is a probability measure, and hence that μ is σ-finite.

Let $n > 0$ be such that $n(x)\mu(dx)$ is a probability measure. It follows that $1 = \int n(x)f(x)\nu(dx) \geq (1/k)\nu(n > 1/k, f = \infty) \cdot \infty$, and hence $\nu(n > 1/k, f = \infty) = 0$. This proves the claim $\nu(f = \infty) = 0$, since $(f = \infty) = \cup_k (n > 1/k, f = \infty)$ from $n > 0$. □

The prior distribution in Bayesian analysis is most often specified on the form $\mathrm{P}_{\boldsymbol{\theta}}(d\theta) = \pi(\theta)\nu(d\theta)$ where ν is counting measure or Lebesgue measure on $\mathbb{R}^d$. The condition $\nu(\pi = \infty) = 0$ ensures that the prior distribution is σ-finite.

The joint distribution of $(X, \boldsymbol{\theta})$ is given by $f(x \,|\, \theta)\pi(\theta)\mu(dx)\nu(d\theta)$ when the statistical model is given by $\mathrm{P}^{\theta}_X(dx) = f(x \,|\, \theta)\mu(dx)$. Proposition 2 can again be used to verify that $(X, \boldsymbol{\theta})$ is σ-finite.

The marginal distribution of X equals the distribution of X, and is given by $\mathrm{P}_X(dx) = f(x)\mu(dx)$ where

$$f(x) = \int f(x \,|\, \theta)\pi(\theta)\nu(d\theta) \tag{2}$$

It is here easy to find examples where the necessary and sufficient condition $\mu(f = \infty) = 0$ in Proposition 2 is not fulfilled. This means that Proposition 2 gives the crucial test for verification of the σ-finiteness of the marginal. This is most important since σ-finiteness of the marginal ensures existence and uniqueness of the main result in a Bayesian analysis: *The posterior distribution* $\mathrm{P}^x_{\boldsymbol{\theta}}$. This claim follows as a special case of the following Theorem:

Theorem 1 *Let X be a locally integrable random variable, and let T be a σ-finite random quantity. The conditional expectation $E^t(X) = E(X \,|\, T = t)$ is defined to be a measurable function of t such that*

$$E(X\,[T \in A]) = E(E^T(X)[T \in A]) \tag{3}$$

holds for all measurable sets A. The conditional expectation exists, is uniquely defined P_T almost everywhere, and is normalized.

Proof. The expectation values involved in equation (3) are both defined by integration with respect to P. The right-hand side is however also equal to an integral with respect to P_T

$$\int_A E^t(X)\ \mathrm{P}_T(dt) \tag{4}$$

The proof of this claim follows from the general Change-of-variables formula

$$\int \phi(T(\omega))\ \mathrm{P}(d\omega) = \int \phi(t)\ \mathrm{P}_T(dt) \tag{5}$$

which remains valid for improper distributions P. The proof follows from consideration of limits of simple functions ϕ.

The measure determined by $A \mapsto E(X[T \in A])$ is σ-finite and absolutely continuous with respect to P_T, and the Radon-Nikodym theorem ensures existence and uniqueness of the conditional expectation identified as the density with respect to P_T.

The measure is σ-finite since X is assumed to be locally integrable, and because P_T is assumed to be σ-finite. The absolute continuity follows since $\mathrm{P}_T(A) = 0$ implies $E(X[T \in A]) = 0$.

The calculation

$$\begin{aligned} \int_A E^t(1)\ \mathrm{P}_T(dt) = E(1\,[T \in A]) &= \int [T(\omega) \in A]\ \mathrm{P}(d\omega) \\ &= \int_A 1\ \mathrm{P}_T(dt) \end{aligned} \tag{6}$$

proves the normalization. This is the almost everywhere equality $E^t(1) = 1$. □

4 Discussion and conclusion

The conditional distribution P^t is defined from the above by $\mathrm{P}^t(A) = E^t[A]$. The indicator function $X = [A]$ is locally integrable. As explained by Kolmogorov (1956, p.54) it is possible to extend the conventional definition of the integral to allow integration with respect to conditional distributions in general, and this holds also for the more general case of a conditional probability space. It is however well known that a conditional distribution can be identified with a family of distributions if Ω is a complete separable metric space equipped with the Borel field, and the proof of this generalizes verbatim to the case of a conditional probability space considered here.

Theorem 1 gives a generalization of conditioning $(\cdot \,|\, A)$ with respect to a elementary condition A to the case of conditioning $(\cdot \,|\, T = t)$ with respect to a σ-finite random quantity. The link between the definitions is obtained from consideration of the random quantity $T = \sum t_i[A_i]$, where A_i is a countable partition of Ω into elementary conditions with $A_1 = A$.

The conditional distribution P_X^θ corresponding to a σ-finite $\boldsymbol{\theta}$ can be defined by either $(\mathrm{P}_X)^\theta$ based on $\mathrm{P}_{(X,\boldsymbol{\theta})}$ or by $(\mathrm{P}^\theta)_X$, but the result is the same. This gives a convenient formulation which includes both conventional statistics and Bayesian statistics. Conventional statistics is here characterized by avoiding any specification of the distribution $\mathrm{P}_{\boldsymbol{\theta}}$, and Bayesian statistics represents the other extreme in that it is assumed that the distribution $\mathrm{P}_{\boldsymbol{\theta}}$ is completely known. This suggests also that it could be reasonable in some concrete problems to consider statistical inference where $\mathrm{P}_{\boldsymbol{\theta}}$ is unknown, but restricted by some conditions such as symmetry.

It is convenient to have a theoretical model which includes both conventional and Bayesian statistics, and it has been demonstrated here that Renyi (1970)'s theory of conditional probability accomplishes this. This theory includes in particular improper priors as used in Bayesian statistics.

References

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.

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.

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

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

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. and B. Lindqvist (2007). Bayes theorem for improper priors. Technical Report S4-2007, Norwegian university of science and technology. http://www.math.ntnu.no/preprint/statistics/2007/S4-2007.pdf.