The Annals of Statistics
1984, Vol. 12, No. 2, 751–757

A NOTE ON SELECTING PARAMETRIC MODELS IN BAYESIAN INFERENCE[1]

BY WILLIAM S. KRASKER

Harvard University

This note is concerned with how to replace assessment of a "true" prior on a nonparametric family of distributions—which is usually infeasible—by assessment of an approximating prior with support in a parametrized subfamily, in such a way that the posterior derived from the parametric model is close to the "true" posterior. In general it is not sufficient that the approximating prior be close to the true prior in the sense of weak convergence, and we characterize the additional aspect of the true prior that must be considered explicitly.

0. Introduction. This paper is concerned with the situation in which a statistician observes data that are believed to be independently and identically distributed according to one of a family of density functions $\{p(\cdot \mid \omega)\}$, where ω ranges over a parameter space Ω. Ideally, a Bayesian would proceed by placing a prior on ω, and applying Bayes' rule to derive the posterior. All inferences about the true distribution of the data would be based on the posterior for ω.

In many applications it is impossible to execute this plan because Ω is too complex for the statistician to be able to specify the prior distribution completely. (For example, Ω might index the set of all continuous univariate distributions). In these cases one generally restricts attention to a subset of Ω that can be parametrized by a k-vector θ, and places a prior on θ. This trivially induces a prior on Ω that differs from the "true" prior, though one might hope that the approximation is close. (As Diaconis and Ylvisaker (1983) point out, the term "true prior" raises some philosophical issues. We use the term simply to mean the prior that one would place on Ω if placing a general prior on that set were easy.) Two questions arise. First, if the approximate prior is indeed close to the true prior, will the posterior derived from that approximation necessarily be close to the true posterior? Second, what should "close" mean in this context?

In the next section we will argue that the topology of weak convergence (convergence of the integrals of all bounded continuous functions) is a reasonable answer to the second question, for both the priors and posteriors. However, with this choice of topology the answer to the first question is no, and, perhaps surprisingly, this fact does not hinge on the potential complexity of Ω. For example, let $\{p(\cdot \mid \omega)\}$ be the family of univariate normal densities with mean 0 and standard deviation ω. Suppose the observed x equals 0. If the true prior on

Received August 1982; revised December 1983.

[1] This research was supported by the Division of Research, Harvard Business School, and by NSF grant SES-81-12498.

AMS 1980 *subject classifications.* Primary 62G99, secondary 62A15.

Key words and phrases. Nonparametric Bayes models, parametric models, Bayesian inference, approximation of priors.

ω is a unit mass at $\omega = 1$, then obviously so is the true posterior. However, if our approximate prior has mass $1 - \varepsilon$ at 1 (so that it is "near" the true prior in terms of weak convergence) and mass ε at some ω very near 0, then the resulting posterior will differ greatly from the true posterior. A smoothed-out version of this example, satisfying assumptions (A1) – (A4) in Section III, could be made to exhibit the same discontinuity. We will see that with those four assumptions, discontinuities are possible because the density of x, conditional on a δ-neighborhood of ω, need not converge to $p(x \mid \omega)$ *uniformly* in ω as $\delta \to 0$. The main result of the paper—which holds for very general Ω—is that if one's prior is close to the true prior, and places all its mass on a subset of Ω where the convergence just described is uniform, then the posterior will be close to the true posterior.

Previous studies, such as Stein (1965) and Diaconis and Ylvisaker (1983), have considered the question of whether approximately correct priors lead to approximately correct posteriors. However, these studies essentially *define* the separation between priors in terms of the distance between the corresponding posteriors, and were not intended to answer the question posed here, which is how to find guidelines for achieving an adequate approximation.

I. Formal setup and notation. Our formal framework is just the standard one for Bayesian statistics (see Lindley, 1972, page 1), consisting of a measurable space $(S, \mathscr{S})$, called the sample space, and a collection of probability distributions on that space indexed by a set Ω. It is customary to assume that those distributions have density functions $p(\cdot \mid \omega)$ with respect to some measure μ, in the sense that

$$\Pr\{x \in B\} = \int_B p(x \mid \omega)\, d\mu(x), \quad (\omega \in \Omega). \tag{1.1}$$

In a Bayesian analysis one also assumes that there is a σ-algebra $\mathscr{B}$ on Ω, and a prior P on $\mathscr{B}$. Having observed $x \in S$, one applies Bayes' rule to form the posterior P_x on $(\Omega, \mathscr{B})$ according to

$$P_x(A) = \frac{\int_A p(x \mid \omega)\, dP(\omega)}{\int_\Omega p(x \mid \omega)\, dP(\omega)} = \frac{p(x \mid A)}{p(x)} P(A) \quad (A \in \mathscr{B}), \tag{1.2}$$

where

$$p(x \mid A) \equiv \frac{1}{P(A)} \int_A p(x \mid \omega)\, dP(\omega) \tag{1.3}$$

and

$$p(x) \equiv p(x \mid \Omega). \tag{1.4}$$

As we mentioned in our informal introduction, the difficulty in applications is that Ω may be too complex for the statistician to be able to specify fully the prior P. Consequently, he is forced to approximate P by a distribution $\hat{P}$ that is easier to describe and update. Most often he simply restricts attention to a subset $\{\omega_\theta\} \subset \Omega$ parametrized by $\theta \in \mathbb{R}^k$, so that $\hat{P}$ is completely specified by a prior on θ. (For example, Ω might index the set of all continuous distributions on $\mathbb{R}$, while

$\{\omega_\theta\}$ indexes only the normal distributions.) Typically the support of $\hat{P}$ will have P-probability zero. An alternative is to try to approximate P by a Dirichlet process (see Ferguson, 1973, 1974), which is quite restrictive, or by a mixture of Dirichlet processes (which is very general; see Dalal and Hall, 1980). The support of such a prior is large, but this is not an advantage in the present context (and in fact makes it harder to meet the hypotheses of Proposition 1). Though the support of the true prior P is usually all of Ω, we see no reason why it is inherently desirable that the support of $\hat{P}$ be large. The sole criterion for choosing $\hat{P}$ should be whether or not $\hat{P}_x$ is near P_x.

The theory presented in the next two sections grew out of an attempt to understand the controversy over Bayesian robustness. Huber (1980, 1981), Hampel (1973), and Rubin (1977) have criticized the so-called Bayesian approach to robustness, which, according to Huber (1981, page vi), "confounds the subject with admissible estimation in an ad hoc supermodel, and still lacks reliable guidelines on how to select the supermodel and the prior so that we end up with something robust." Our original goal was to develop guidelines for selecting robust models, but, partly as a result of Dempster's (1975) arguments, we have decided that the Huber-Hampel robustness theory is not relevant to the Bayesian case. Notions like influence, sensitivity, and breakdown point, which comprise much of the theory of robust parametric estimation, violate the likelihood principle by involving the sample space at points other than the observations. They therefore seem out of place in a Bayesian analysis, even one that explicitly recognizes that the model is only an approximation. (Diaconis and Freedman (1983) take an intermediate position, arguing that consideration of the effects of hypothetical samples on the posterior is helpful for determining whether a particular prior is really a correct quantification of current knowledge.) The possibility of gross errors affects a Bayesian analysis only through its effect on the true prior P. In determining whether $\hat{P}$ is an adequate approximation to P, the sample space should enter only through the observed x.

II. Priors and posteriors. We will metrize both the set of priors on $(\Omega, \mathscr{B})$ and the set of posteriors with the Prohorov metric, which induces the topology of weak convergence. In order to do this we will assume that Ω is a separable metric space (Ω, d) with Borel sets $\mathscr{B}$. For any $A \subset \Omega$ and $\varepsilon > 0$, let

$$A^\varepsilon = \{\omega \in \Omega: d(\omega, A) \leq \varepsilon\}. \tag{2.1}$$

The Prohorov metric π is then defined by

$$\pi(P, \hat{P}) = \inf\{\varepsilon > 0: P(A) \leq \hat{P}(A^\varepsilon) + \varepsilon \text{ for all measurable } A\}. \tag{2.2}$$

(See Huber (1981) for a proof that π is symmetric, positive definite, and satisfies the triangle inequality). The statement $\pi(P, \hat{P}) \leq \varepsilon$ is equivalent to the statement that

$$\hat{P}(A^{c\varepsilon c}) - \varepsilon \leq P(A) \leq \hat{P}(A^\varepsilon) + \varepsilon \tag{2.3}$$

for all measurable A, where a superscript c denotes complementation; this shows the sense in which $\hat{P}$ approximates P.

For the posteriors, the Prohorov metric can be justified on decision-theoretic grounds. For example, if a decision maker's utility function $U(D, \omega)$ is bounded and satisfies a Lipschitz condition in ω, uniformly in the decision D, then a decision that is optimal relative to $\hat{P}$ will be nearly optimal relative to P, if $\pi(P, \hat{P})$ is small. This fact is a straightforward corollary of our Lemma 1; stronger results are contained in Kadane and Chuang (1978).

For the set of priors, our choice of the Prohorov metric derives from consideration of the limits of one's ability to approximate a true prior P by a parametric model. Note that if $\hat{P}$ is specified by a parametric model $\{\omega_\theta\}$ and a prior on θ, $\hat{P}(\{\omega_\theta\})$ will always be one even though $P(\{\omega_\theta\})$ will generally be zero. This shows that it is too much to ask that $\hat{P}(A)$ be near $P(A)$ for every measurable A, or to put it differently, it is generally not feasible to approximate P closely in a strong topology like total-variation distance. On the other hand, with sufficiently detailed modeling one might hope to be able to satisfy the inequalities in (2.3) for all measurable A, even if ε is small.

III. Guidelines for model selection. Throughout this section, in which we develop some guidelines for model selection (Proposition 1), the observation x is assumed fixed.

For $\delta > 0$ and $\omega \in \Omega$, define $N_\delta(\omega)$, the "δ-neighborhood of ω," by $N_\delta(\omega) = \{\omega' \in \Omega: d(\omega, \omega') < \delta\}$. Also, for any $Z \in \mathscr{B}$, denote by P_Z the conditional distribution of ω, given Z. It is easy to show that $\pi(P_Z, P) \leq 1 - P(Z)$. Denote by $P_{Z,x}$ the distribution of ω given Z and x, and note that

$$P_{Z,x}(A) = \int_A p(x \mid \omega)\, dP_Z(\omega) \Big/ \int_\Omega p(x \mid \omega)\, dP_Z(\omega).$$

We will make the following regularity assumptions about the true prior.

(A1) $P(N_\delta(\omega)) > 0$ for all $\omega \in \Omega$ and $\delta > 0$.

(A2) $\lim_{\delta \to 0} p(x \mid N_\delta(\omega)) = p(x \mid \omega)$ for almost all $\omega \in \Omega$.

(A3) For any $\delta > 0$, $p(x \mid N_\delta(\omega))$ is bounded and satisfies a Lipschitz condition in ω.

(A4) $p(x) > 0$.

Assumptions (A1) and (A4) guarantee the existence of P_x and $p(x \mid N_\delta(\omega))$, as defined by (1.2) and (1.3). (A3) is a smoothness condition on P whose use stems from the property proved in Lemma 1.

Condition (A2) will certainly hold if $p(x \mid \omega)$ is continuous in ω; for example if the sample space is discrete and the metric d on Ω induces the topology of weak convergence (or a stronger topology) on the associated probability mass functions $p(\cdot \mid \omega)$. (Indeed, in this case $p(x \mid \omega)$ is uniformly continuous in ω.) However, if $\{p(\cdot \mid \omega)\}_{\omega \in \Omega}$ is the set of continuous distributions on $\mathbb{R}$, then $p(x \mid \omega)$ will not be continuous in ω. One would nevertheless expect the true prior P to satisfy the weaker condition (A2). Actually, (A2) enters our analysis only through one of its well-known implications, that there are subsets of Ω of arbitrarily high P-

probability (and hence of arbitrarily high P_x-probability) on which $p(x \mid N_\delta(\omega))$ converges to $p(x \mid \omega)$ *uniformly* as $\delta \to 0$.

We can now state our main result.

PROPOSITION 1. *Assume* (A1) − (A4). *Let* $\varepsilon > 0$, *and let* Z *be a measurable subset of* Ω, *with* $P_x(Z) \geq 1 - \varepsilon$, *on which* $p(x \mid N_\delta(\omega)) \to p(x \mid \omega)$ *uniformly in* ω *as* $\delta \to 0$. *There exists* $\eta > 0$ *such that if* $\pi(\hat{P}, P_Z) \leq \eta$ *and* $\hat{P}(Z) = 1$, *then* $\pi(\hat{P}_x, P_x) \leq 2\varepsilon$.

If S is discrete then, because $p(x \mid \omega)$ is uniformly continuous in ω as mentioned earlier, one can take $Z = \Omega$. Hence, for a discrete sample space the discontinuity that motivated this paper does not arise and so it suffices to have $\hat{P}$ close to P to ensure that $\hat{P}_x$ is close to P_x. As the example in the introduction demonstrates, this is not true for a continuous sample space. However, the proposition provides guidelines for choosing a parametric model $\{\omega_\theta\}$ and a prior on θ by showing what additional aspect of the true prior P must be considered explicitly. (Note that, although the approximating prior $\hat{P}$ chosen as indicated by the proposition will always be π-close to P, it will depend on the observation x, which is taken as given. However, $\hat{P}$ is really just a device for facilitating the computation of a posterior that is close to the true posterior. The true prior P does *not* depend on x.) If, say, Ω indexes the set of all continuous univariate probability distributions, then P might be assumed to be such that the convergence of $p(x \mid N_\delta(\omega))$ to $p(x \mid \omega)$ is slow only if the density $p(\cdot \mid \omega)$ has a "spike" at x. One could therefore let Z index those densities that are both bounded by K and satisfy a Lipschitz condition with constant K, letting K be large enough that $P_x(Z)$ is greater than, say, $\varepsilon/2$. According to the proposition, if P_Z is closely approximated by a parametric model $\hat{P}$ satisfying $\hat{P}(Z) = 1$, then $\pi(\hat{P}_x, P_x)$ will be $\leq \varepsilon$. More specifically, in cases in which the true density is thought to be nearly normal with high probability, one could let the parametric model be the family of normal distributions, restricting the scale parameter to be bounded away from zero so that $\hat{P}(Z) = 1$.

The following lemma, which is used in the proof of Proposition 1, is probably well known (and is certainly closely related to the fact that the Prohorov and bounded-Lipschitz metrics generate the same topology; see Huber, 1981, page 33). However, we have not seen it stated in the literature.

LEMMA 1. *Let* $f: \Omega \to \mathbb{R}$ *be nonnegative, bounded by* M, *and satisfy a Lipschitz condition* $|f(\omega) - f(\omega')| \leq Kd(\omega, \omega')$. *If* P *and* $\hat{P}$ *are probability measures on* $(\Omega, \mathscr{B})$ *satisfying* $\pi(\hat{P}, P) \leq \varepsilon$, *then*

$$\int_A f \, dP \leq \int_{A^\varepsilon} f \, d\hat{P} + \varepsilon(K + M)$$

for all measurable A.

PROOF. Let χ_A be the characteristic function of the set A. Then

$$\int_A f\,dP = \int f\chi_A\,dP = \int_0^\infty P(f\chi_A \geq t)\,dt = \int_0^M P(f\chi_A \geq t)\,dt.$$

The hypotheses of the lemma imply $\{f\chi_A \geq t\}^\varepsilon \subset \{(f + K\varepsilon)\chi_{A^\varepsilon} \geq t\}$. Using the definition of the Prohorov metric it follows that

$$P(f\chi_A \geq t) \leq \hat{P}(\{f\chi_A \geq t\}^\varepsilon) + \varepsilon \leq \hat{P}((f + K\varepsilon)\chi_{A^\varepsilon} \geq t) + \varepsilon.$$

Hence

$$\begin{aligned}\int_A f\,dP &\leq \int_0^M \hat{P}((f + K\varepsilon)\chi_{A^\varepsilon} \geq t) + \varepsilon\,dt \\ &\leq \int_{A^\varepsilon} f + K\varepsilon\,d\hat{P} + \varepsilon M \leq \int_{A^\varepsilon} f\,d\hat{P} + \varepsilon(K + M). \quad \square\end{aligned}$$

PROOF OF PROPOSITION 1. From the hypothesis $P_x(Z) \geq 1 - \varepsilon$ it follows that $\pi(P_{Z,x}, P_x) \leq 1 - P_x(Z) \leq \varepsilon$. It therefore suffices to show that

$$\frac{\int_A p(x \mid \cdot)\,dP_Z}{\int_\Omega p(x \mid \cdot)\,dP_Z} \leq \frac{\int_A p(x \mid N_\delta(\cdot))\,dP_Z}{\int_\Omega p(x \mid N_\delta(\cdot))\,dP_Z} + \frac{\varepsilon}{3} \tag{3.1}$$

$$\leq \frac{\int_{A^\varepsilon} p(x \mid N_\delta(\cdot))\,d\hat{P}}{\int_\Omega p(x \mid N_\delta(\cdot))\,d\hat{P}} + \frac{2\varepsilon}{3} \tag{3.2}$$

$$\leq \frac{\int_{A^\varepsilon} p(x \mid \cdot)\,d\hat{P}}{\int_\Omega p(x \mid \cdot)\,d\hat{P}} + \varepsilon \tag{3.3}$$

for all measurable A and some δ when $\hat{P}$ is close enough to P_Z and $\hat{P}(Z) = 1$. The outer expressions in this chain of inequalities imply that $\pi(\hat{P}_x, P_{Z,x}) \leq \varepsilon$, from which $\pi(\hat{P}_x, P_x) \leq 2\varepsilon$ follows. So let $\{\hat{P}_n\}$ be an arbitrary sequence that is π-convergent to P_Z and satisfies $\hat{P}_n(Z) = 1$; we have to show that for some sequence $\{\delta_n\}$, (3.1)–(3.3) hold for large n. Using (A3) and Lemma 1, choose δ_n going to zero slowly enough to preserve the inequalities in line (3.2) and also keep the denominator in (3.2) bounded away from zero. Due to the uniform convergence on Z, (3.1) and (3.3) will also hold for large n. $\square$

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

DALAL, S. R. and HALL, G. J. JR. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8** 664–672.

DEMPSTER, A. P. (1975). A subjectivist look at robustness. *Bull. I.S.I. Proc. 40th Session* **46** *Book 1* 349–374.

DIACONIS, P. and FREEDMAN, D. (1983). Frequency properties of Bayes rules. In *Scientific Inference, Data Analysis, and Robustness.* 105–115. Eds. G. E. P. Box, T. Leonard, C. F. Wu. Academic, New York.

DIACONIS, P. and YLVISAKER, D. (1983). Quantifying prior opinion. Stanford University Technical Report.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrsch. verw. Gebiete* **27** 87–104.

HUBER, P. J. (1980). Comments on "Sampling and Bayes' inference in scientific modelling and robustness" by G. E. P. Box. *J. Roy. Statist. Soc. A* **143** 418–420.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KADANE, J. B. and CHUANG, D. T. (1978). Stable decision problems. *Ann. Statist.* **6** 1095–1110.

LINDLEY, D. V. (1972). *Bayesian Statistics, A Review.* Society for Industrial and Applied Mathematics, Philadelphia.

RUBIN, H. (1977). Robust Bayesian Estimation. In *Statistical Decision Theory and Related Topics, II.* Eds. S. S. Gupta and D. S. Moore. Academic, New York.

STEIN, C. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli, Bayes, Laplace* 217–240. Eds. J. Neyman and L. Lecam. Springer, New York.

GRADUATE SCHOOL OF BUSINESS ADMIN.
HARVARD UNIVERSITY
MORGAN 339
SOLDIERS FIELD
BOSTON, MASSACHUSETTS 02163