BAYESIAN ROBUSTNESS WITH ASYMPTOTICALLY NICE CLASSES OF PRIORS

by

N. D. Shyamalkumar
Purdue University

Technical Report #96-22

Department of Statistics
Purdue University
West Lafayette, IN USA

June 1996

*

BAYESIAN ROBUSTNESS WITH ASYMPTOTICALLY NICE CLASSES OF PRIORS

N. D. Shyamalkumar

Department Of Statistics, Purdue University, U.S.A.*

Abstract

In the literature on robust Bayesian analysis which deals with uncertainty with respect to the prior, it is usual to model this uncertainty in the form of a class of priors. It is then natural to ask under what simple conditions on the class does the normalized posterior converge in some sense to a normal distribution uniformly over the class of priors. We observe such a simple sufficient condition and discuss its utility in defining classes of priors when one has vague knowledge about his prior. Examples will be discussed where robust Bayesian analysis can be easily done with such asymptotically well behaved classes.

KEY WORDS: Sensitivity, ϵ-contamination, Density ratio class

1 Introduction

The theory of Bayesian Robustness involves assessment of the sensitivity of the posterior measures of interest to uncertainty in some or all of the elicited quantities. It is common in

*Research supported by FrederickN. Andrews fellowship and partially supported by the National Science Foundation, Grants DMS-8923071 and DMS-9303556.

the area to restrict oneself to uncertainty in the prior. In this case one usually embeds the *best guess* prior in a class of priors, which may or may not be topological in nature, and tries to find the range of values that the posterior measures of interest assumes when the prior is allowed to vary in the chosen class. Common such classes are the epsilon-contamination class with unrestricted or restricted contaminants, density ratio classes and quantile classes.

It is shown in this paper that if we model prior uncertainty by some of the above mentioned classes, the range of the posterior measure of interest might not converge to zero, even as we let the sample size to grow to infinity. We give a simple condition, which when satisfied by the class of priors, guarantees convergence of the range of posterior measures to zero with increasing sample size.

Specifically we consider the case of posterior credible intervals. The asymptotic behavior desired not only depends on the class of priors, but also on the likelihood and the loss function. Here, as usual, we restrict our attention to squared error loss and we assume that the likelihood satisfies the usual regularity conditions. Under such assumption we observe that the proof of asymptotic normality of the posterior given in Walker (1969) can be modified to yield a simple condition for the classes of priors to satisfy in order to get the desired asymptotic behavior of the ranges.

We note that, for the case of the density ratio neighborhood, $C(p_1, p_2)$,

$$C(p_1, p_2) = \{\pi : p_1(\theta) \leq \pi(\theta) \leq p_2(\theta), \forall\theta; \int \pi(d\theta) < \infty\},$$

where p_1 and p_2 are two continuous functions, the range of the posterior probability of , say, the 95%, posterior confidence interval under a fixed prior in $C(p_1, p_2)$ tends to 0.95, if and only if p_1 and p_2 are identical at the true parameter. The above fact can be derived from results in De Robertis (1978). Hence the density ratio neighborhoods are too big to give us the behavior we desire.

On many occasions it is difficult to come up with the *best guess* prior. In these cases one has very vague knowledge about the prior and the classes should contain reasonably flat priors. It is clear that an attempt to ignore the robustness question and the use of non-informative priors has it's own perils and so does using small classes of improper priors.

In Pericchi & Walley (1991) a strong case is made for the use of the class of all double exponential distributions, with the scale parameter bounded away from zero by a constant, when there is very little prior information. One of the properties of this class is that it is translation invariant. This class is reasonably tight to allow the range of the posterior probability of credible intervals to converge to a singleton. In fact it is, in some sense, the family with the sharpest tail among all translational invariant families with the above property. We suggest some other classes, which like the above, are translation invariant and which behave in the desired manner.

2 Results

Let Θ denote the parameter space, which we assume to be a subset of the real line. We denote the sample density w.r.t. a σ-finite measure μ, by $f(x, \theta)$, when the parameter value is θ. In the sequel we deal only with priors which are absolutely continuous w.r.t. the Lebesgue measure on the real line.

The proof of the following result is based on the proof of asymptotic normality of the posterior given in Walker (1969). The set of needed conditions is split into three sets of which the sets A and B are identical to that of Walker(1969). These latter conditions are given at the end for the sake of completeness. Condition C is given below.

Condition C

The class $\mathcal{F}$ satisfies the condition that for each fixed θ_0 in Θ and $\epsilon > 0$, there exists a neighborhood $N(\theta_0)$ around θ_0 such that

$$\sup_{\pi \in \mathcal{F}} \left| \frac{\pi(\theta)}{\pi(\theta_0)} - 1 \right| < \epsilon \quad \forall\, \theta \in N(\theta_0).$$

Theorem 2.1 *Suppose that the conditions A-C hold. Let θ_0 be an interior point of Θ and let the data $X_1, X_2, \ldots, X_n$ be i.i.d. with density $p(x|\theta_0)$. Let σ_n be the positive square root of $-1/L_n''(\hat{\theta}_n)$ whenever this exists. Then, for real constants, a and b, the posterior probability of the interval $\hat{\theta}_n + b\sigma_n < \theta < \hat{\theta}_n + a\sigma_n$ converging in P_{θ_0} probability to the standard normal probability of the interval (a, b), uniformly over the class of priors $\mathcal{F}$, as $n \to \infty$.*

PROOF. An outline of the proof is given in the last section. □

Lemma 2.1 *Under the same conditions as the above theorem,*

$$\sup_{\pi_1,\pi_2\in\mathcal{F}} \int |\pi_1(\theta|\underset{\sim}{x}_n) - \pi_2(\theta|\underset{\sim}{x}_n)|d\theta \rightarrow 0$$

in P_{θ_0} probability.

PROOF. The proof of this stronger statement follows in the same way as that of the above theorem. □

Remark 2.1 The above result is similar to one in Ghosh et al (1992). They prove almost sure convergence, unlike our weaker, in "probability" convergence.

Lemma 2.2 *If the class of prior densities, $\mathcal{F}$, is contained in the set of densities*

$$\{\pi : |\frac{\pi'}{\pi}| \leq K\} \textit{ for some } K > 0,$$

then $\mathcal{F}$ satisfies condition C.

PROOF. By a Taylor series expansion,

$$|\pi(\theta) - \pi(\theta_0)| \leq |\theta - \theta_0| \cdot |\pi'(\theta^*)| \leq K\,|\theta - \theta_0|\,\pi(\theta^*),$$

Using the gradient condition in the above yields,

$$|\pi(\theta) - \pi(\theta_0)| \leq K\,|\theta - \theta_0|\,\pi(\theta_0)\exp\{K|\theta_0 - \theta^*|\} \leq K\,|\theta - \theta_0|\,\pi(\theta_0)\exp\{K\,|\theta - \theta_0|\}.$$

It is clear that, by choosing a positive δ such that $K\,\delta\exp\{K\,\delta\} \leq \epsilon$,

$$\sup_{\pi\in\mathcal{F}} |\frac{\pi(\theta)}{\pi(\theta_0)} - 1| < \epsilon \;\forall\; \theta \text{ satisfying } |\theta - \theta_0| < \delta.$$

□

Remark 2.2 It is interesting to note that, under such a condition on the prior and some conditions on the tail of the likelihood, it was shown in Mukhopadhyay and DasGupta (1993) that if $\pi_c(\theta) = \frac{1}{c}\pi(\frac{x}{c})$, then the posterior converges uniformly over the sample to the likelihood as $c \rightarrow \infty$. The convergence is in L_1, but they also show that uniform convergence is possible if we assume further conditions on the likelihood.

Example 1: Let $\mathcal{F}$ be the class of all double exponential densities with arbitrary median and scale parameter which is bounded away from zero. That is, for some $c > 0$,

$$\mathcal{F} = \{\pi : \pi(\theta) = \frac{e^{\frac{-|\theta-\mu|}{\sigma}}}{\sigma}, \sigma > c \; \mu \in \mathbb{R}\}.$$

This class would satisfy condition C, in view of Lemma 2.2. Similarly, the class consisting of Cauchy densities defined in a similar way,

$$\mathcal{F} = \{\pi : \pi(\theta) = \frac{\sigma^2}{\pi(\sigma^2 + (\theta - \mu)^2)}, \; \sigma > c, \; \mu \in \mathbb{R}\},$$

would satisfy condition C. The above two facts can be proved by checking that the gradient of the densities in each case is bounded by a finite constant, which would depend on c, uniformly over each class, and then appealing to the above lemma. The former class was considered by Pericchi & Walley (1991). □

The above classes have the property that they are translation invariant. This is a nice feature for a class to have, when you are trying to model very little prior information on the parameter. The gradient condition of the lemma says that the sharpest tails allowed are exponential. In a sense, these families are extremal. Note the lemma gives only a sufficient condition. So one could possibly take any density function satisfying the above gradient condition, create a location scale family and bound the scale parameter away from zero to obtain a translation invariant family with the desired asymptotic behavior.

Note that the condition in the above lemma is global whereas condition C is local in nature. Below we prove that a suitably defined sub-class of the normal family satisfies condition C, whereas the condition of the above lemma is clearly seen to be violated.

Example 2: Consider the class of densities, $\mathcal{F}$, defined as

$$\mathcal{F} = \{\pi : \pi(\theta) = \frac{e^{\frac{-(\theta-\mu)^2}{2\cdot\sigma^2}}}{\sqrt{2\cdot\pi}\sigma}, \; |\mu - \mu_0| < K \cdot \sigma^2, \; \sigma^2 > c\},$$

where K and c are some real positive constants. This class was previously considered by Perricchi & Walley (1991). We shall below check condition C directly. Fix $\theta_0 \in \Theta$. Let $N(\theta_0)$ be a δ ball around θ_0. We now show how to choose the value of δ. Note that, for

$\pi = \mathcal{N}(\mu, \sigma^2) \in \mathcal{F}$ and $\theta \in N(\theta_0)$,

$$\begin{aligned} & |\pi(\theta) - \pi(\theta_0)| \\ \leq\ & |\theta - \theta_0| \cdot |\pi'(\theta^*)| \ = \ |\theta - \theta_0| \cdot \pi(\theta^*) \cdot \tfrac{|\theta^* - \mu|}{\sigma^2} \\ =\ & |\theta - \theta_0| \cdot \tfrac{|\theta^* - \mu|}{\sigma^2} \cdot \pi(\theta_0) \cdot \tfrac{\pi(\theta^*)}{\pi(\theta_0)} \end{aligned}$$

by application of a Taylors series expansion. Note that

$$\begin{aligned} & \pi(\theta^*)/\pi(\theta_0) \\ =\ & \exp(-\tfrac{(\theta^* - \mu)^2 - (\theta_0 - \mu)^2}{2 \cdot \sigma^2}) \ \leq\ \exp(|\theta^* - \theta_0| \cdot [\tfrac{|\theta^* - \theta_0|}{2 \cdot \sigma^2} + \tfrac{|\theta_0 - \mu|}{\sigma^2}]) \\ \leq\ & \exp(\delta \cdot [\tfrac{\delta}{2 \cdot \sigma^2} + \tfrac{|\theta_0 - \mu_0| + |\mu - \mu_0|}{\sigma^2}]) \\ \leq\ & \exp(\delta \cdot (K + \tfrac{\delta}{2 \cdot c} + \tfrac{|\theta_0 - \mu_0|}{c})). \end{aligned}$$

Combining these inequalities yields

$$\begin{aligned} & |\pi(\theta) - \pi(\theta_0)| \\ \leq\ & \delta \cdot (K + \tfrac{|\theta_0 - \mu_0| + \delta}{c}) \cdot \exp(\delta \cdot (K + \tfrac{\delta}{2 \cdot c} + \tfrac{|\theta_0 - \mu_0|}{c})) \cdot \pi(\theta_0), \quad \forall \theta \in N(\theta_0). \end{aligned}$$

It is clear that δ can be chosen to make $\delta \cdot (K + \frac{|\theta_0 - \mu_0| + \delta}{c}) \cdot \exp(\delta \cdot (K + \frac{\delta}{2 \cdot c} + \frac{|\theta_0 - \mu_0|}{c}))$ arbitrarily small. Hence we see that this class satisfies condition C. □

Remark 2.3 In Berger (1990) the class Γ_C, defined as

$$\Gamma_C = \{N(\mu, \sigma^2) \text{ distributions } : -0.2 \leq \mu \leq 0.2 \ \text{ and } \ 0.7 \leq \sigma^2 \leq 1.3\},$$

is mentioned as a useful conjugate class for a normal likelihood when it is known apriori that the median and quartiles are 0 and ± 0.675, respectively. As seen in the above example, this class would give the desired behavior asymptotically.

Lemma 2.3 *Let $\mathcal{F}$ be a class which satisfies condition C and let $\mathcal{C}(\mathcal{F})$ denote the convex hull of $\mathcal{F}$. Then $\mathcal{C}(\mathcal{F})$ also satisfies condition C. Further, if $\mathcal{F}$ can be indexed by a parameter, say $\gamma \in \Gamma$, then the set $\mathcal{G}$ defined as*

$$\mathcal{G} = \{\eta : \eta = \int \pi_\gamma \, dP(\gamma), \ P \text{ a probability on } \Gamma\},$$

also satisfies condition C.

PROOF. We shall only prove the second statement. Note that

$$\begin{aligned}
&|\frac{\eta(\theta)}{\eta(\theta_0)} - 1| \\
&= |\frac{\int \pi_\gamma(\theta)\ dP(\gamma)}{\int \pi_\gamma(\theta_0)\ dP(\gamma)} - 1| \quad = \quad |\frac{\int (\pi_\gamma(\theta) - \pi_\gamma(\theta_0))\ dP(\gamma)}{\int \pi_\gamma(\theta_0)\ dP(\gamma)}| \\
&\leq \frac{\int |\pi_\gamma(\theta) - \pi_\gamma(\theta_0)|\ dP(\gamma)}{\int \pi_\gamma(\theta_0)\ dP(\gamma)} \quad \leq \quad \sup_{\gamma\in\Gamma} \frac{|\pi_\gamma(\theta) - \pi_\gamma(\theta_0)|}{\pi_\gamma(\theta_0)}.
\end{aligned}$$

As the class $\mathcal{F}$ satisfies condition C, we see from the above that so does $\mathcal{G}$. □

Example 3: Let $\mathcal{F}$ be any class which satisfies the condition C. Then one can define an ϵ-contamination class around a base prior, say π_0, with contaminants in $\mathcal{F}$ as

$$\Gamma_{\epsilon,\mathcal{F}} = \{\pi : \pi = (1-\epsilon)\pi_0 + \epsilon\eta, \eta \in \mathcal{F}\}.$$

If we assume that the prior π_0 is continuous and positive, then it can be shown that the above class also satisfies condition C, by applying Lemma 2.3 twice. Trying to find the extrema of posterior quantities w.r.t. the above class reduces to the problem of finding extrema of expressions of the type

$$\frac{A + \int \phi\, d\eta}{B + \int \psi\, d\eta}$$

as η varies over $\mathcal{F}$, where A,B are constants and ϕ,ψ are functions of the parameter. For example, in the mixture case, when the index takes values in $\mathbb{R}^n$, the problem reduces to an n-dimensional extremal problem. □

3 Proof

Condition A

1. Θ is a closed set of points.
2. The set of points $\{x : f(x|\theta) > 0\}$ is independent of θ; we denote this by $\mathcal{X}$.
3. If θ_1, θ_2 are two distinct points of Θ,

$$\mu\{x : f(x|\theta_1) \neq f(x|\theta_2)\} > 0.$$

4. Let $x \in \mathcal{X}$, $\theta' \in \Theta$. Then for all θ such that $|\theta - \theta'| < \delta$, with δ sufficiently small,

$$|\log f(x|\theta) - \log f(x|\theta')| < H_\delta(x, \theta'),$$

where $\lim_{\delta \to 0} H_\delta(x, \theta') = 0$ and, for any $\theta_0 \in \Theta$,

$$\lim_{\delta \to 0} \int_{\mathcal{X}} H_\delta(x, \theta') f(x|\theta_0)\, d\mu = 0.$$

5. If Θ is not bounded, then for any $\theta_0 \in \Theta$ and sufficiently large η,

$$\log f(x|\theta) - \log f(x|\theta_0) < K_\eta(x, \theta_0),$$

whenever $|\theta| > \eta$, where

$$\lim_{\eta \to \infty} \int_{\mathcal{X}} K_\eta(x, \theta_0) f(x|\theta_0)\, d\mu < 0.$$

Condition B

Let θ_0 be an interior point of Θ.

1. $\log f(x|\theta)$ is twice differentiable with respect to θ in some neighborhood of θ_0.

2. Let

$$J(\theta_0) = \int_{\mathcal{X}} f_0 \left(\frac{\partial f_0}{\partial \theta_0}\right)^2 d\mu,$$

where f_0 denotes $f(x|\theta_0)$. Then $0 < J(\theta_0) < \infty$.

3.

$$\int_{\mathcal{X}} \frac{\partial f_0}{\partial \theta_0}\, d\mu = \int_{\mathcal{X}} \frac{\partial^2 f_0}{\partial \theta_0^2}\, d\mu = 0.$$

4. If $|\theta - \theta_0| < \delta$, where δ is sufficiently small, then

$$\left|\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} - \frac{\partial^2 \log f(x|\theta_0)}{\partial \theta_0^2}\right| < M_\delta(x, \theta_0),$$

where

$$\lim_{\delta \to 0} \int_{\mathcal{X}} M_\delta(x, \theta_0) f(x|\theta_0)\, d\mu = 0.$$

Outline of the Proof of Theorem 2.1: It is clear that the behavior of the priors outside a neighborhood of the true parameter point, say θ_0, does not matter. This is because of the relative exponential decay of the likelihood outside the neighborhood. In the neighborhood, by using Taylor series we know that the likelihood behaves as a normal density with the mean at the MLE. It is in this neighborhood that we need to replace the prior density by a constant, that given by the density value at θ_0. It is here that we need the Condition C. The probability statements involve only the measure given by the μ-density $f(\cdot|\theta_0)$. The proof is otherwise the same as that given in Walker (1969). □

References

Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis.* J. Kadane,(ed.), North-Holland, Amsterdam.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer Verlag, New York, second edition.

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference* **25**, 303-328.

DeRobertis, L. (1978). The Use Of Partial Prior Knowledge In Bayesian Inference. Doctoral dissertation submitted to Yale University.

Ghosh, J.K., Ghosal, S. & Samanta, T. (1994). Stability and Convergence of the Posterior in Non-Regular Problems. *Statistical Decision Theory and Related Topics V*, ed. Gupta S.S. & Berger J.O., 183-199.

Mukhopadhyay S. & DasGupta A. (1993). Uniform Approximation of Bayes Solutions and Posteriors: Frequentistly valid Bayes Inference. Technical Report No. 93-12C, Purdue University.

Pericchi, L. R. & Walley, P. (1991). Robust Bayesian Credible Intervals and Prior Ignorance. International Statistical Review (1991), **58** 1, pp 1-23.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31**, 80-88.

Department of Statistics
Purdue University
West Lafayette, Indiana 47907