UNIFORM AND SUBUNIFORM POSTERIOR ROBUSTNESS: THE SAMPLE SIZE PROBLEM*

by

Anirban DasGupta and Saurabh Mukhopadhyay
Purdue University

Technical Report #92-20C

Department of Statistics
Purdue University

May 1992

* Research was supported by NSF grant DMS-89-230-71 at Purdue University

◇

UNIFORM AND SUBUNIFORM POSTERIOR ROBUSTNESS: THE SAMPLE SIZE PROBLEM

Anirban DasGupta and Saurabh Mukhopadhyay

Purdue University

Abstract

The following general question is addressed: given iid realizations $X_1, X_2, \ldots, X_n$ from a distribution P_θ with parameter θ, where θ has a prior distribution π belonging to some family Γ, is it possible to prescribe a sample size n_0 such that for $n \geq n_0$, posterior robustness is guaranteed to obtain for any actual data we are likely to see or even for all possible data. Formally, we identify a "natural" set C such that $P(\text{The observation vector } X \notin C) \leq \varepsilon$, for all possible marginal distributions implied by Γ, and protect ourselves for all X in the set C. Such a set C typically exists if Γ is tight. The plausibility of such a preposterior guarantee of postexperimental robustness depends on many things: the actual decision problem, the nature of the loss, whether the loss function is known, the variety of priors in Γ, whether the model is regular or nonregular, the dimension of the parameter θ, etc. We explore a variety of these questions.

There are two aspects in these results: one of them is to establish the plausibility itself; this is done by showing uniform convergence to zero of ranges of posterior

quantities. This part forms the mathematical foundation of the program. The second aspect is to provide actual sample size prescriptions for a specific goal to be attained. This forms the application part of the program.

For instance, for testing that the mean of a multinormal distribution belongs to some (measurable) set B, the range of the posterior probability of the hypothesis converges to 0, uniformly for all likely X and uniformly in B, at a rate $\frac{1}{\sqrt{n}}$. In the one dimensional case, the range of the posterior mean converges to 0 uniformly for all likely X, uniformly over the class of all Lipschitz functions, at a rate $\frac{1}{n}$. These assume conjugate priors.

If $\log \pi$ has a bounded gradient, then in any arbitrary dimension, a remarkably strong robustness obtains. For instance, any pair of HPD credible sets of a given level are guaranteed to be visually identical for large n. This is proved by showing that uniformly in X, the Hausdorff distance between the two sets goes to 0 at a rate $\frac{1}{\sqrt{n}}$.

It is demonstrated that much as in classical theory, the rates and the calculations are different in nonregular models. In particular, the classical rate for the regular case can be maintained uniformly over a broad class of loss functions. These results are for the uniform case.

Key words: Prior, Posterior, Uniform robustness, Confidence sets, Risks, Hypothesis tests, Nonregular, Nonconjugate.

AMS Clasification: 62F15, 62C10.

1 Introduction

1.1 The general goal

In finite dimensional parametric problems, considerable evidence has now accumulated that if the observed data are such that the likelihood function is moderately concentrated near a common center of the priors, then posterior robustness is obtained in a broad and general sense. There are things which are not well understood here; for instance, the role of skewness in the priors or the effect of the dimension. It is also expected that in *regular* statistical problems, the quality of posterior robustness will improve with an increase in sample size. Loosely speaking, the central reason for such an expectation is the convergence to normality (independent of the prior) of the posterior distribution under frequently satisfied conditions. Thus posterior robustness may obtain by virtue of a secondary phenomenon, namely, closeness to classical inference. These are broad and vague statements. Mathematics should be much more precise: on grounds of aesthetics, as also for actual practical utility. The general goal of this article is to address and answer the following general question:

Suppose an observable X (let us say a sufficient statistic arising out of n iid realizations $X_1, \ldots, X_n$ from some model) has distribution P_θ, and the parameter θ has a prior (probability) distribution π belonging to a specified family Γ. Can we demonstrate the existence of an explicit sample size n_0 such that posterior robustness is guaranteed to obtain for this sample size (or larger) if the actual observed x when the experiment is conducted happens to be any of the x's we at all expect to see. In other words, we will protect ourselves against all plausible x that may occur once the experiment is conducted and this we attempt to do by only selecting the sample size. The hope, of course, is that the required sample size n_0 is realistic. We will try to show some general structure in these problems; no doubt, the answers will depend on the nature of the problem. But an explicit prescribed sample size for a

specific goal and some concrete mathematical structure in the general problem are our main objectives. We call this the problem of subuniform posterior robustness. If the observed data x are one against which we are not preprotected, we have reasons to worry about the model assumptions we have made, for such data x were not surmised at the design stage. In some cases, we will actually go all the way and demonstrate the presense of posterior robustness for all possible x provided only the sample size is sufficiently large. We will call this uniform posterior robustness. We will solve the simpler problems in this article since this is our first excursion into this area. Some of these problems relate to the concept of stable estimation; see Kadane and Chuang (1978). Also see Meeden and Isaacson (1977) and Lehmann (1986). For a general exposition to Bayesian robustness, see Berger (1985). We suspect some others may have toyed with some of the ideas presented here, and may have even been surreptitiously aware of the plausibility of some of the results, particularly notable among them Herman Rubin.

1.2 An illustrative example

Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$, so that the sufficient statistic $\bar{X} \sim N(\theta, \frac{1}{n})$. Interest lies in testing $H_0 : \theta \leq 0$. Elicitation has produced a family of conjugate normal priors $N(0, \tau^2)$, where $0 < \tau_1^2 \leq \tau^2 \leq \tau_2^2 < \infty$. Robustness goals may be quite different. Suppose we will like to have the posterior probability of H_0 to vary in a range of at most c, where c is a prespecified (small) positive number (c may even depend on n). The question then is what is the smallest sample size for which the goal is realized provided x is any of the values we expect to observe. Formally, let $\varepsilon > 0$ be fixed; suppose C is such that $P(X \notin C) \leq \varepsilon$ for all possible marginal distributions of X under the specified priors on θ. We like to know if there exists n_0 such that $n \geq n_0$ implies

$$\sup P(H_0|x) - \inf P(H_0|x) \leq c; \ \forall x \in C,$$

where the *sup* and *inf* above are naturally with respect to the specified priors. From a practical point of view, it is crucial that we can identify an explicit n_0. Notice a set C with the given property is clearly not unique. In fact in some problems such a set C does not even exist (for instance, when the set of priors is such that the marginals do not form a tight family). But often such C does exist and furthermore there is a unique natural choice of C. For instance, if θ is a location vector and the family of priors Γ is tight, then C exists and can be assumed compact. If Γ is not tight, existence of C is severely geopardized. For instance, if θ is the mean of a normal distribution on the line, and Γ is the class of all symmetric unimodal distributions, then no such C exists.

The level sets of each marginal distribution are symmetric intervals $[-a, a]$ and a natural choice of C is therefore $C = [-a(\varepsilon), a(\varepsilon)]$, where $a(\varepsilon) = z_{\frac{\varepsilon}{2}}\sigma_2$, with $\sigma_2^2 = \frac{1}{n} + \tau_2^2$ and $z_{\frac{\varepsilon}{2}}$ the $100(1-\frac{\varepsilon}{2})th$ percentile of the $N(0,1)$ distribution (notice this is the set of smallest Lebesgue measure among all C with the desired property). The family of posterior distributions is $N(rx, \frac{r}{n})$ with $r_1 \leq r \leq r_2$, where $r_i = \frac{\tau_i^2}{\frac{1}{n} + \tau_i^2}$. We thus need:

$$\sup_{|x| \leq a(\varepsilon)} |\Phi(\sqrt{n\, r_2}x) - \Phi(\sqrt{n\, r_1}x)| \;\leq\; c. \tag{1}$$

By symmetry, it is sufficient to have

$$\Phi(\sqrt{n\, r_2}x) - \Phi(\sqrt{n\, r_1}x) \;\leq\; c \text{ for } 0 \leq x \leq a(\varepsilon) \tag{2}$$

By elementary methods, the LHS of (2) is maximized on the interval $[0, a(\varepsilon)]$ at

$$\begin{aligned} x = x_0 &= \left(\frac{\log \frac{r_2}{r_1}}{n(r_2 - r_1)}\right)^{\frac{1}{2}} \quad \text{if } x_0 \leq a(\varepsilon) \\ &= a(\varepsilon) \qquad\qquad\quad \text{if } x_0 > a(\varepsilon) \end{aligned}$$

The condition $x_0 \leq a(\varepsilon)$ is equivalent to

$$\frac{n(\tau_2^2 - \tau_1^2)}{1 + n\tau_1^2} \geq \frac{1}{z_{\frac{\varepsilon}{2}}^2} \log\left(\frac{\tau_2^2(1 + n\tau_1^2)}{\tau_1^2((1 + n\tau_2^2)}\right). \tag{3}$$

Since the LHS of (3) is monotone increasing and the RHS monotone decreasing in n, and opposite inequalities between the two sides hold at $n = 0, \infty$ respectively, inequality (3) holds for all sufficiently large n. Hence, for large n,

$$(4) \qquad \sup_{0 \le x \le a(\varepsilon)} (\Phi(\sqrt{n\,r_2}x) - \Phi(\sqrt{n\,r_1}x)) \;=\; \Phi(\sqrt{n\,r_2}x_0) - \Phi(\sqrt{n\,r_1}x_0).$$

A patient but easy calculation shows that $\sqrt{n\,r_i}x_0 \to 1$ as $n \to \infty$ for each $i = 1, 2$, and hence for any $c > 0$, (1) is guaranteed to hold for large n. We now go one step further and establish the precise rate at which (4) converges to zero.

By the fundamental theorem of calculus,

$$(5) \qquad \Phi(\sqrt{n\,r_2}x_0) - \Phi(\sqrt{n\,r_1}x_0) \;=\; O\left((\sqrt{n\,r_2}x_0 - \sqrt{n\,r_1}x_0)\,\phi(1)\right),$$

where $\phi(.)$ denotes standard normal density. However, (5) is easily seen to be $O\left(\dfrac{\tau_2^2 - \tau_1^2}{2\,\tau_1^2\,\tau_2^2}\dfrac{\phi(1)}{n}\right)$, implying that the maximum possible range of the posterior probability of $H_0 : \theta \le 0$ converges to zero at the rate of $\frac{1}{n}$ and indeed,

$$(6) \qquad n \sup_{x \in C}(\sup P(H_0|x) - \inf P(H_0|x)) \to \frac{\tau_2^2 - \tau_1^2}{2\,\tau_1^2\,\tau_2^2}\phi(1)$$

Formally, (6) is valid even if $\tau_1^2 = 0$ or $\tau_2^2 = \infty$; notice the interesting fact that (1) cannot hold for large n if the degenerate normal distribution is a possible prior. In this case, one has the amusing fact that (1) holds only for sufficiently small n!

The smallest sample size n_0 for which (1) holds is reported in Table 1 for various τ_1^2, τ_2^2, c and ε. Two features immediately stand out: if τ_1, τ_2 are are both small or both large, the required sample size n_0 is astonishingly small. If τ_1 is small and τ_2 is large, as expected, n_0 is large. For instance, if $\varepsilon = 0.1$, $c = 0.01$, $\tau_1 = 1$ and $\tau_2 = 5$, then one only needs a sample of size 12 for (1) to hold. This example is an elementary illustration of the problems we address in the rest of the article.

1.3 Overview

As commented before, the plausibility of a preposterior guarantee of subuniform or uniform posterior robustness will depend on many features of the problem:

a. the nature of the priors: are they all concentrated, or all flat, or some concentrated and some flat? It will be seen that in the latest case a preposterior guarantee is the hardest to provide.

b. the nature of the decision problem: for instance, estimation or testing? It will be seen that a guarantee is easier in testing problems than in estimation problems.

c. the nature of the loss function: the effect of the loss is particularly prominent on the rate of convergence to zero of ranges of posterior quantities. For some losses, the rate can be frustratingly slow, especially in high dimensional problems, making the guarantee largely ornamental.

d. is the loss function fully known: if it is not (indeed many think a loss is more difficult to ascertain than a prior), then one needs a guarantee simultaneously over a broad class of loss functions.

e. is the prescribed sample size for the use of one individual or many or the community as a whole? The larger the client group, the larger is the potential family of problems we need to be protected for and the harder it is to provide a preposterior guarantee. This is related to the point made in part d above.

f. is the model regular or nonregular: from our training in classical inference, we should certainly expect to confront quite different calculations in the two cases.

h. are we content with subuniform robustness or have the ambition of uniform robustness.

All of these issues will be addressed to the extent that it is possible to do so in this article.

1.4 Outline

Section 2.1 treats the one dimensional normal problem with conjugate priors. The multivariate normal case with conjugate priors is considered in section 2.2. In section 3, we consider a natural nonregular problem, namely the $U[0,\theta]$ case. Section 4 describes a family of problems with nonconjugate priors where full fledged uniform robustness can be guaranteed. Section 5 gives a concise summary and discussion. Within each individual section, a variety of problems is addressed. Our principal goal is to give the reader an initial but broad insight into this general problem. The specific problems are discussed in more detail within the individual sections.

2 Normal likelihood

2.1 The univariate case

The illustrative example in subsection 1.2 deals with a common testing problem. We will treat the point estimation problem here. The problem of estimating the mean is of primary importance. We will consider more general functions including the mean as a special case. The main mathematical goal is to establish the possibility of a preposterior guarantee and the appropriate rate of convergence. The main practical goal is to provide an explicit sample size prescription for the benefit of the user.

Theorem 1 *Let* $X \sim N\left(\theta, \frac{1}{n}\right)$ *and let* θ *have a prior* π *belonging to the collection* Γ *of* $N(0,\tau^2)$ *distributions, with* $0 < \tau_1^2 \leq \tau^2 \leq \tau_2^2 < \infty$. *Let* $h(.)$ *be any function such that*

i. $h(-\theta) = h(\theta)$,

ii. $h(\theta)$ *is nondecreasing for* $\theta > 0$,

iii. $h(\theta)$ *is everywhere twice continuously differentiable with* $h'(\theta) \neq 0$ *if* $\theta \neq 0$.

Let C be the interval $\left[-z_{\frac{\varepsilon}{2}}\sigma_2, z_{\frac{\varepsilon}{2}}\sigma_2\right]$ where $\sigma_2^2 = \frac{1}{n} + \tau_2^2$. Then for squared error loss,

$$(7) \quad n \sup_{x \in C}\left[\sup_{\pi} E(h(\theta)|x) - \inf_{\pi} E(h(\theta)|x)\right] \to \frac{\tau_2^2 - \tau_1^2}{\tau_1^2\tau_2^2} z_{\frac{\varepsilon}{2}}\tau_2 \left|h'\left(z_{\frac{\varepsilon}{2}}\sigma_2\right)\right| \quad \text{as } n \to \infty$$

In particular, subuniform posterior robustness obtains with the maximum range of posterior expectation of $h(.)$ converging to 0 at a rate of $\frac{1}{n}$.

Discussion: Again notice that if a point prior is entertained, a preposterior guarantee can not be given. The result stated above automatically handles all functions of the form θ^{2k+1} for nonnegative integers k. For even moments, the rate is the same, but a different proof is required.

The following general notation will be used repeatedly: for $\tau_1^2 \le \tau^2 \le \tau_2^2$, define

$$\begin{aligned} \frac{1}{n} + \tau_i^2 &= \sigma_i^2, \\ \frac{\tau_i^2}{\frac{1}{n} + \tau_i^2} &= r_i, \\ \frac{\tau^2}{\frac{1}{n} + \tau^2} &= r, \text{ and} \\ (8) \qquad z_{\frac{\varepsilon}{2}}\sigma_2 &= a(\varepsilon) \end{aligned}$$

Proof of Theorem 1: If θ has the prior $N(0, \tau^2)$, then it has the posterior $N(rx, \frac{r}{n})$, with $r_1 \le r \le r_2$. By virtue of property *(i)* of $h(\theta)$, it is enough to consider only $0 \le x \le a(\varepsilon)$. For any given $0 \le x \le a(\varepsilon)$,

$$\begin{aligned} E(h(\theta)|x, r) &= \frac{\sqrt{n}}{\sqrt{2\pi r}} \int_{-\infty}^{\infty} e^{-\frac{n}{2r}(\theta - rx)^2} h(\theta)\, d\theta \\ &= \frac{\sqrt{n}}{\sqrt{2\pi r}} \int_{0}^{\infty} \left\{e^{-\frac{n}{2r}(\theta - rx)^2} - e^{-\frac{n}{2r}(\theta + rx)^2}\right\} h(\theta)\, d\theta, \end{aligned}$$

which because of property *(ii)* of h is monotone nondecreasing in r for given x by standard monotone likelihood ratio arguments. Hence, for given $0 \le x \le a(\varepsilon)$,

$$\sup_{\pi} E(h(\theta)|x) - \inf_{\pi} E(h(\theta)|x) = \frac{\sqrt{n}}{\sqrt{2\pi r_2}} \int_{-\infty}^{\infty} e^{-\frac{n}{2r_2}(\theta - r_2 x)^2} h(\theta)\, d\theta$$

$$(9) \qquad - \frac{\sqrt{n}}{\sqrt{2\pi r_1}} \int_{-\infty}^{\infty} e^{-\frac{n}{2r_1}(\theta - r_1 x)^2} h(\theta)\, d\theta$$

Another monotone likelihood ratio argument treating $0 \leq x \leq a(\varepsilon)$ as the parameter establishes (9) as monotone nondecreasing in x, implying that

$$\begin{aligned} & \sup_{x \in C} \left[\sup_{\pi} E(h(\theta)|x) - \inf_{\pi} E(h(\theta)|x) \right] \\ (10) \qquad = \; & E\left[(h(\theta)|\theta \sim N(r_2 a(\varepsilon), \frac{r_2}{n}) \right] - E\left[(h(\theta)|\theta \sim N(r_1 a(\varepsilon), \frac{r_1}{n}) \right] \end{aligned}$$

The assertion of the theorem now follows on a two term Taylor expansion of h around $z_{\frac{\varepsilon}{2}}\tau_2$ and using the fact that $r_i \to 1$ and $n(r_2 - r_1) \to \frac{\tau_2^2 - \tau_1^2}{\tau_1^2 \tau_2^2}$ as $n \to \infty$.

Corollary 1 *For any given $c > 0$, and π as in Theorem 1,*

$$\sup_{x \in C} \left[\sup_{\pi} E(h(\theta)|x) - \inf_{\pi} E(h(\theta)|x) \right] \leq c$$

for all large n.

For $h(\theta) = \theta$, the actual prescribed sample sizes n_0 are given in Table 1 for various combinations of $\tau_1, \tau_2, \varepsilon$ and c. The result above shows that for a broad class of skew–symmetric functions, the range of the posterior mean converges to zero at the rate of $\frac{1}{n}$. The result, however, is not uniform over the functions h. We will now show that $\frac{1}{n}$ is in fact the rate of convergence *uniformly* over another broad class of functions, the function $h(\theta) = \theta$ being a particular member of this class.

Theorem 2 *Let $\mathcal{F}$ be the class of all functions with Lipschitz norm $\leq M$, i.e.,*

$$|h(u) - h(v)| \leq M|u - v|,$$

where $M < \infty$ is fixed. Then,

$$\sup_{x \in C} \sup_{h \in \mathcal{F}} \left[\sup_{\pi} E(h(\theta)|x) - \inf_{\pi} E(h(\theta)|x) \right] = O\left(\frac{1}{n}\right).$$

Remark. In analysis, consideration of the Lipschitz class allows looking at functions with many zigzags. The function $h(\theta) = |\theta|$, which is of statistical interest, is one example. See Hewitt and Stromberg (1978).

For the proof of Theorem 2, we need the following facts. We will let $M = 1$ without loss of generality.

Lemma 1 *Given two probability measures P_1 and P_2,*

$$\sup_{h \in \mathcal{F}} \left| \int h \, dP_1 - \int h \, dP_2 \right| = \inf \{ E|X - Y| : X \sim P_1, Y \sim P_2 \}.$$

Proof: See Kantorovich and Rubinstein (1958).

Remark. The infimum on the RHS of the above lemma is with respect to all joint distributions having P_1 and P_2 as marginals.

Lemma 2 *If P_1, P_2 are distributions on the real line with corresponding CDF's F_1 and F_2, then*

$$\inf \{ E|X - Y| : X \sim P_1, Y \sim P_2 \} = \int_{-\infty}^{\infty} |F_1(x) - F_2(x)| dx.$$

Proof: See Dall'Aglio (1956). Also see Dudley (1968) and Rachev (1984).

Combining the two Lemmas, one gets that

$$\sup_{h \in \mathcal{F}} \left| \int h \, dP_1 - \int h \, dP_2 \right| = \int_{-\infty}^{\infty} |F_1(x) - F_2(x)| \, dx.$$

Proof of Theorem 2: For a fixed pair of posterior distributions, say, $P_r = N(rx, \frac{r}{n})$ and $P_s = N(sx, \frac{s}{n})$, and for a fixed x, we will prove that $\sup_{h \in \mathcal{F}} \left| \int h \, dP_r - \int h \, dP_s \right|$ is $O\left(\frac{1}{n}\right)$. Even

though r, s and x all vary in compact sets, this is not enough to establish the assertion of the theorem. However, the $\frac{1}{n}$ rate for fixed r, s and x will be proved by demonstrating an expansion for $\sup_{h \in \mathcal{F}} \left| \int h \, dP_r - \int h \, dP_s \right|$ in powers of $\frac{1}{n^{\frac{1}{2}}}$, with the leading term as $\frac{1}{n}$ and coefficients that can be universally bounded. This is enough for proving the theorem in its full strength. The following are the main steps.

Step 1. If F_r, F_s denote the CDF's of P_r, P_s respectively, then

$$\begin{aligned}
\int_{-\infty}^{\infty} |F_r(\theta) - F_s(\theta)| \, d\theta &= \int_{-\infty}^{\infty} \left| \Phi\left(\sqrt{\frac{n}{r}}(\theta - rx)\right) - \Phi\left(\sqrt{\frac{n}{s}}(\theta - sx)\right) \right| d\theta \\
&= \int_{-\infty}^{-x\sqrt{rs}} \left(\Phi\left(\sqrt{\frac{n}{s}}(\theta - sx)\right) - \Phi\left(\sqrt{\frac{n}{r}}(\theta - rx)\right) \right) d\theta \\
&+ \int_{-x\sqrt{rs}}^{\infty} \left(\Phi\left(\sqrt{\frac{n}{r}}(\theta - rx)\right) - \Phi\left(\sqrt{\frac{n}{s}}(\theta - sx)\right) \right) d\theta
\end{aligned}$$

Step 2. On the first of the two integrals, separate and write as

$$\int_{-\infty}^{-x\sqrt{rs}} \Phi\left(\sqrt{\frac{n}{s}}(\theta - sx)\right) d\theta - \int_{-\infty}^{-x\sqrt{rs}} \Phi\left(\sqrt{\frac{n}{r}}(\theta - rx)\right) d\theta$$

(both are finite). On the first of these, make the change of variable $\sqrt{\frac{n}{s}}(\theta - sx) = u$, write $\Phi(u)$ as $\int_{-\infty}^{u} \phi(t)dt$ and use Fubini's theorem to obtain the integral as

$$\left(-x\sqrt{rs} + sx\right) \Phi\left((\sqrt{r} - \sqrt{s})\sqrt{n}x\right) + \sqrt{\frac{s}{n}} \phi\left((\sqrt{r} - \sqrt{s})\sqrt{n}x\right) = T_1 + T_2 \quad \text{(say)}$$

The second integral similarly gives

$$\left(-x\sqrt{rs} + rx\right) \Phi\left((\sqrt{s} - \sqrt{r})\sqrt{n}x\right) + \sqrt{\frac{r}{n}} \phi\left((\sqrt{s} - \sqrt{r})\sqrt{n}x\right) = T_3 + T_4 \quad \text{(say)}$$

Step 3. $T_2 - T_4$ gives an expansion with terms $\frac{1}{n^{k+\frac{1}{2}}}$, $k \geq 1$, since $\phi\left((\sqrt{r} - \sqrt{s})\sqrt{n}x\right) \to \phi(0) = \frac{1}{\sqrt{2\pi}}$ and the multiplier $\sqrt{\frac{s}{n}} - \sqrt{\frac{r}{n}}$ has the stated expansion. All coefficients can be bounded independent of r, s and x.

Step 4. $T_1 - T_3$ similarly admits an expansion with terms $\frac{1}{n^k}$, $k \geq 1$. This is because $\pm(\sqrt{r}-\sqrt{s})\sqrt{n}x \to 0$ and the outside multiplier $(s-r)x$ admits the stated expansion. Again, all coefficients can be bounded.

Step 5. The second integral

$$\int_{-x\sqrt{rs}}^{\infty} \left(\Phi\left(\sqrt{\frac{n}{r}}(\theta - rx)\right) - \Phi\left(\sqrt{\frac{n}{s}}(\theta - sx)\right)\right) d\theta$$

can be handled by the same argument on transforming θ to $-\theta$ so that the interval of integration is again a neighborhood of $-\infty$ and thus the integrals can be separated. Combining steps 4 and 5, the theorem is obtained.

2.2 The multivariate case

In this section we will look at the so called symmetric normal problem, i.e., we let $X \sim N_p(\theta, \frac{1}{n}I)$ and $\theta \sim N_p(0, \tau^2 I)$, with $\tau_1^2 \leq \tau^2 \leq \tau_2^2$. Then, using the notation of (8), the posterior distributions are $N_p(rx, \frac{r}{n}I)$, with $r_1 \leq r \leq r_2$. The set C is now the sphere

$$C = \{x : ||x||_2 \leq \sigma_2 \chi_\varepsilon(p)\}, \tag{11}$$

where $\chi_\varepsilon^2(p)$ is the $100(1-\varepsilon)th$ percentile of the chi–square distribution with p degrees of freedom. We consider the question of a preposterior guarantee in four different problems. This are now addressed one at a time. The results are all valid, with modifications, if the covariance matrix Σ of θ satisfies $\Sigma_1 \leq \Sigma \leq \Sigma_2$, Σ_1, Σ_2 p.d.

2.2.1 Point estimation of the mean

The result stated in this section is entirely trivial and is stated merely for the purpose of direct reference. The theorem stated below shows that the diameter of the set of posterior means converges to zero at the rate of $\frac{1}{n}$ irrespective of the dimension p.

Theorem 3 *Under the structure assumed above,*

$$n \sup_{x \in C} diam(S(x)) \to \frac{\tau_2^2 - \tau_1^2}{\tau_1^2 \tau_2^2} \tau_2 \chi_\epsilon(p) \quad as\ n \to \infty, \tag{12}$$

where $diam(S(x))$ denotes the diameter in Euclidean distance of the set of posterior means for given x.

Proof: Trivial.

Discussion. Thus again we can give a guarantee of the posterior means being very close together for all large samples simultaneously for all data likely to be observed. Notice also that Theorem 3 implies that in very high dimensional problems, the sample size needs to grow only at the rate of $\sqrt{p}$ to ensure subuniform posterior robustness. Again, the actual prescribed sample sizes are given in Table 2. The sample sizes make the diameter of $S(x)$ less than or equal to $2\sqrt{\frac{p}{n}}$ uniformly in $x \in C$. The accuracy index $2\sqrt{\frac{p}{n}}$ comes from the standard error of the classical estimate, since $2\sqrt{\frac{p}{n}}$ is the distance between $-\frac{1}{\sqrt{n}}\mathbf{1}$ and $\frac{1}{\sqrt{n}}\mathbf{1}$.

2.3 Range of risks

Apart from keeping the posterior means close together, it may be important to keep the range of the posterior risks small. Indeed, some have argued that only the posterior risk needs to be robust. Here we state and prove a theorem on the maximum range of the posterior risk for general power losses of the form $\|\theta - a\|_2^k, k > 0$. We will explicitly demonstrate the effect of the actual value of k on the problem at hand.

Theorem 4 *Let the likelihood and the priors be as in Theorem 3. For estimating the mean θ using the loss $\|\theta - a\|_2^k, k > 0$, let $r(\pi, x)$ denote the posterior risk for a fixed prior π. Then,*

$$n^{\frac{k}{2}+1} \sup_{x \in \Re^p} [\sup_{\pi} r(\pi, x) - \inf_{\pi} r(\pi, x)] \to \frac{k 2^{\frac{k}{2}-1} \Gamma(\frac{k+p}{2})(\tau_2^2 - \tau_1^2)}{\Gamma(\frac{p}{2}) \tau_1^2 \tau_2^2}, as\ n \to \infty. \tag{13}$$

Discussion. Several points, although elementary, are worth noting. First, the posterior robustness in risk is fully uniform, a gift of the conjugate structure (although, also see Section 4). Second, the rate of convergence to zero is $\frac{1}{n^{\frac{k}{2}+1}}$. Thus the faster the loss goes to zero at zero, the easier it is to provide a guarantee of posterior robustness – an expected phenomenon. Finally, for very high dimensional problems, a straightforward calculation using Stirling's formula and (13) gives the fact that as $p \to \infty$, the maximum range of the posterior risks is $O(\frac{p^{\frac{k}{2}}}{n^{\frac{k}{2}+1}})$, so that n needs to grow at the rate of $p^{\frac{k}{2+k}}$ to ensure uniform robustness of posterior risks.

Proof of Theorem 4: Under the $N(rx, \frac{r}{n}I)$ posterior, the posterior risk equals

$$E(\|\theta - rx\|^k \mid \theta \sim N(rx, \frac{r}{n}I)) = (\frac{r}{n})^{\frac{k}{2}}\ 2^{\frac{k}{2}}\ \frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{p}{2})}. \tag{14}$$

(14) is maximized at $r = r_2$ and minimized at $r = r_1$. From here, the theorem follows on elementary calculations.

2.4 Hypothesis testing

The example in Section 1.2 demonstrates that for the common one sided testing problem in one dimension, the range of the posterior probability converges to zero uniformly over x in C at the rate of $\frac{1}{n}$. However, keeping in mind the point we made in section 1.3 (point e), if the prescribed sample size is for the use of many individuals or the community as a whole, we cannot reasonably assume that all clients would want to test the same hypothesis. Indeed, in such a case, we have little control on which hypotheses may be tested. In this section we prove the surprising result that for testing an arbitrary hypothesis $H_0 : \theta \in B$, *uniformly over all measurable* B, it is possible to give a preposterior guarantee of robustness in the posterior probability of H_0. The price to pay is a slower rate of convergence. But we show that irrespective of the dimension p, the rate of the convergence is $\frac{1}{\sqrt{n}}$. Thus a *common sample size* can be prescribed for all arbitrary hypothesis testing problems that guarantees a

range of posterior probability smaller than any prespecified (small) number simultaneously for all x we are likely to see. While this result is mathematicaly attractive, the actual sample size prescriptions in Table 1 show that the price to pay for such an ambitious all engulfing posterior robustness is indeed high.

Theorem 5 *Let the likelihood and the prior be as in Theorem 3. Consider the hypothesis testing problem $H_0 : \theta \in B$, where B is any measurable subset of $\Re^p$. Then,*

$$\begin{aligned} &\sup_{x \in C} \sup_{B} [\sup_{\pi} P(\theta \in B|x) - \inf_{\pi} P(\theta \in B|x)] \\ &\leq \frac{(\tau_2^2 - \tau_1^2)\chi_\varepsilon(p)}{\tau_1\sqrt{2}} \sqrt{\frac{n}{(1+n\tau_1^2)(1+n\tau_2^2)}} + \frac{p^{\frac{3}{2}} 2^{p+1}(\tau_2^2 - \tau_1^2)}{\tau_1^2} \frac{1}{1+n\tau_2^2}. \end{aligned} \tag{15}$$

Discussion. From Theorem 5, it is clear that the LHS of (15) converges to zero at least as fast as $\frac{1}{\sqrt{n}}$. We will later prove that it can not go to zero faster than $\frac{1}{\sqrt{n}}$, which will establish $\frac{1}{\sqrt{n}}$ as the correct rate of convergence. Thus, specializing to specific hypotheses (such as $H_0 : \theta \leq 0$ as in section 1.2) may lead to a faster rate of convergence. The following two fundamental lemmas are useful for proving Theorem 5.

Lemma 3 *Let Q_1 and Q_2 denote the $N_p(\mu_1, I)$ and $N_p(\mu_2, I)$ distributions respectively. Then,*

$$\sup_B |Q_1(B) - Q_2(B)| \leq 2^{-\frac{1}{2}} \|\mu_1 - \mu_2\|.$$

Proof: Easy on using the well known fact that

$$\sup_B |Q_1(B) - Q_2(B)| = \frac{1}{2} \int_{-\infty}^{\infty} |q_1(x) - q_2(x)| dx,$$

where q_i is the density of Q_i.

Lemma 4 *Let Q_1 and Q_2 denote the $N_p(0, I)$ and $N_p(0, \Sigma)$ distributions respectively. Then,*

$$\sup_B |Q_1(B) - Q_2(B)| \leq p\ 2^{p+1} \|\Sigma - I\|_2,$$

where $\|A_{p\times p}\|_2^2$ denotes $trA'A$.

Proof: See Pfanzagl (1973).

Proof of Theorem 5: We will first prove that for any fixed pair of posteriors P_1 and P_2 (each normal), $\sup_{x \in C} \sup_B |P_1(B) - P_2(B)|$ satisfies the bound given in (15). This will imply that uniformly in x belonging to C, and for any fixed B, any two members of the set of numbers $\{P_\alpha(B)\}$ are at a distance equal to at most the RHS of (15), where $\{P_\alpha\}$ denotes the total collection of posteriors. Since the RHS of (15) is a universal constant, the required assertion will follow from this.

Towards this end, let P_1 be the $N_p(rx, \frac{r}{n}I)$ distribution and let P_2 be the $N_p(sx, \frac{s}{n}I)$ distribution; here x is now fixed, and assume without loss of generality that $r_1 \leq s \leq r \leq r_2$. For notational convenience, we also denote the variational distance $\sup_B |P_1(B) - P_2(B)|$ by $\|P_1 - P_2\|$.

Hence,

$$
\begin{aligned}
\|P_1 - P_2\| &= \|N(rx, \frac{r}{n}I) - N(sx, \frac{s}{n}I)\| \\
&\leq \|N(rx, \frac{r}{n}I) - N(sx, \frac{r}{n}I)\| + \|N(sx, \frac{r}{n}I) - N(sx, \frac{s}{n}I)\| \\
&\quad \text{(triangular inequality)} \\
&= \sup_B |P(Y_1 \in B | Y_1 \sim N(rx, \frac{r}{n}I)) - P(Y_2 \in B | Y_2 \sim N(sx, \frac{r}{n}I))| \\
&+ \sup_B |P(Y_1 \in B | Y_1 \sim N(sx, \frac{r}{n}I)) - P(Y_2 \in B | Y_2 \sim N(sx, \frac{s}{n}I))| \\
&= \sup_B |P(Z_1 \in B\sqrt{\frac{n}{r}} | Z_1 \sim N(\sqrt{\frac{n}{r}}rx, I)) - P(Z_2 \in B\sqrt{\frac{n}{r}} | Z_2 \sim N(\sqrt{\frac{n}{r}}sx, I))| \\
&+ \sup_B |P(Y_1 \in B | Y_1 \sim N(0, \frac{r}{n}I)) - P(Y_2 \in B | Y_2 \sim N(0, \frac{s}{n}I))| \\
&\quad \text{(change of variables)} \\
&= \sup_B |P(Z_1 \in B | Z_1 \sim N(\sqrt{\frac{n}{r}}rx, I)) - P(Z_2 \in B | Z_2 \sim N(\sqrt{\frac{n}{r}}sx, I))| \\
&+ \sup_B |P(Z_1 \in B\sqrt{\frac{n}{r}} | Z_1 \sim N(0, I)) - P(Z_2 \in B\sqrt{\frac{n}{r}} | Z_2 \sim N(0, \frac{s}{r}I))| \\
&\quad (\ B\sqrt{\tfrac{n}{r}} \text{ and } B \text{ form the same collection of sets plus change of variable})
\end{aligned}
$$

$$
\begin{aligned}
&= \|N(\sqrt{\frac{n}{r}}rx, I) - N(\sqrt{\frac{n}{r}}sx, I)\| + \|N(0,I) - N(0, \frac{s}{r}I)\| \\
&\leq 2^{-\frac{1}{2}}\|\sqrt{\frac{n}{r}}rx - \sqrt{\frac{n}{r}}sx\| + p\ 2^{p+1}\|I - \frac{s}{r}I\|_2 \\
&\quad \text{(Lemmas 3 and 4)} \\
&= \frac{\sqrt{n}}{\sqrt{2r}}(r-s)\|x\| + p^{\frac{3}{2}}\ 2^{p+1}\frac{r-s}{r} \\
&\leq \frac{\sqrt{n}}{\sqrt{2r_1}}(r_2 - r_1)\sigma_2\chi_\varepsilon(p) + p^{\frac{3}{2}}\ 2^{p+1}\frac{r_2 - r_1}{r_1} \qquad (16)
\end{aligned}
$$

(15) now follows from (16) on using the definitions of r_1, r_2 and σ_2 given in (8). This proves Theorem 5.

We will now show that the rate of convergence of the LHS of (15) cannot be faster than $\frac{1}{\sqrt{n}}$.

Theorem 6 *For $r > s$,*

$$
\begin{aligned}
\|N_1(rx, \frac{r}{n}) - N_1(sx, \frac{s}{n})\| &= \Phi(\sqrt{\frac{n}{s}}(a - sx)) + \Phi(\sqrt{\frac{n}{s}}(a + sx)) \\
&\quad - \Phi(\sqrt{\frac{n}{r}}(a - rx)) - \Phi(\sqrt{\frac{n}{r}}(a + rx)), \qquad (17)
\end{aligned}
$$

where $a^2 = \frac{rs}{n(r-s)}\log\frac{r}{s} + rsx^2$.

Proof: It is well known that

$$
\begin{aligned}
&\|N_1(rx, \frac{r}{n}) - N_1(sx, \frac{s}{n})\| \\
&= \frac{1}{2}\int_{-\infty}^{\infty} |\frac{\sqrt{n}}{\sqrt{2\pi r}}e^{-\frac{n}{2r}(\theta - rx)^2} - \frac{\sqrt{n}}{\sqrt{2\pi s}}e^{-\frac{n}{2s}(\theta - sx)^2}|d\theta \\
&= \frac{1}{2}\int_{-\infty}^{\infty} |f_r(\theta) - f_s(\theta)|d\theta \quad \text{(say)}. \qquad (18)
\end{aligned}
$$

¿From (18), the Theorem follows on straightforward integration on using the fact that $f_r(\theta) \geq f_s(\theta)$ iff $|\theta| \geq a$.

Corollary 2 *The LHS of (15) cannot converge to zero at a rate faster than $\frac{1}{\sqrt{n}}$.*

Proof: Since the LHS of (15) for $p = 1$ is greater than or equal to $\|N_1(rx, \frac{r}{n}) - N_1(sx, \frac{s}{n})\|$ for any fixed r, s, and $x \in C$, the corollary will follow for $p = 1$ if it can be proved that (17) converges to 0 at the rate $\frac{1}{\sqrt{n}}$. This, however, follows on noting that $\Phi(\sqrt{\frac{n}{s}}(a - sx)) - \Phi(\sqrt{\frac{n}{r}}(a - rx))$ determines the rate of convergence of (17) and a two term Taylor expansion around zero estabilishes this rate to be $\frac{1}{\sqrt{n}}$. Since the variational distance cannot go to zero faster than $\frac{1}{\sqrt{n}}$ in one dimension, the same is true on considering rectangles $B \times \Re \times \ldots \times \Re$ in any dimension.

Remark. Inequality (15) is used in Table 1 to provide prescribed sample sizes for making the LHS of (15) smaller than c for various choices of $\tau_1, \tau_2, \varepsilon$ and c.

2.5 Construction of robust confidence sets

For estimating a multivariate normal mean, the classical confidence set

$$\{\theta \mid \|\theta - x\| \leq \frac{\chi_\alpha(p)}{\sqrt{n}}\} \tag{19}$$

covers (in the frequentist sense) θ with a probability of $1 - \alpha$ for all θ and has the property that its volume converges to zero at the rate of $\frac{1}{n^{\frac{p}{2}}}$. In order to be competitive, it may therefore be desirable to construct a confidence set for θ which has a posterior probability of $1 - \alpha$ for all priors π under consideration and whose volume goes to zero at the classical rate. We will prove that this is indeed possible and demonstrate such a set. We will also provide the usual preposterior guarantee for the volume to be smaller than a specified number for all x we are likely to see.

Theorem 7 *Let the likelihood and the prior be as in Theorem 3. Consider the confidence set*

$$S(x) = \{\theta : \|\theta - r_0 x\| \leq c\}$$

with $r_0 = \sqrt{r_1 r_2}$ *and* $c = \frac{r_2 F^{-1}(1-\alpha)}{n}$, *where* $F(.)$ *denotes the CDF of a noncentral chi-square distribution with* p *degrees of freedom and noncentrality parameter*

$$\delta = n\sigma_2^2 \chi_\varepsilon^2(p) \frac{(r_2 - r_0)^2}{r_2}. \tag{20}$$

Then, $P(\theta \in S(x)|x) \geq 1-\alpha$ *for all priors* π *under consideration and furthemore,* $\sup_{x \in C} vol(S(x))$ *goes to zero at the rate* $\frac{1}{n^{\frac{p}{2}}}$.

Discussion: The problem of determining the confidence set that actually minimizes $vol(S(x))$ under the restriction that $\inf_\pi P(\theta \in S(x)|x) \geq 1 - \alpha$ is hard. Some results of this type are known for suitable prior families. See DasGupta (1991). Notice the curious fact that instead of centering the suggested set at $\frac{r_1+r_2}{2}x$, we are centering at $\sqrt{r_1 r_2}x$. This has a mathematical advantage. Any statistical benefit is unknown.

Proof of Theorem 7: For notational convenience, a random variable with a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter δ will itself be denoted by $NC\chi^2(p, \delta)$. Also, let $N_p(sx, \frac{s}{n}I)$ be a typical posterior distribution and corresponding probabilities are denoted as $P_s(.)$. Then,

$$\begin{aligned}
P_s((\theta - r_0 x)'(\theta - r_0 x) \leq c) &= P_s(\frac{n}{s}(\theta - r_0 x)'(\theta - r_0 x) \leq \frac{nc}{s}) \\
&= P(NC\chi^2(p, \lambda) \leq \frac{nc}{s}) \\
&\quad \text{(where } \lambda = \tfrac{n}{s}(s - r_0)^2 \|x\|^2\text{)} \\
&\geq P(NC\chi^2(p, \lambda) \leq \frac{nc}{r_2}) \text{ since } s \leq r_2 \\
&\geq P(NC\chi^2(p, \frac{n(r_2 - r_0)^2}{r_2}\|x\|^2 \leq \frac{nc}{r_2}) \\
&\quad \text{(since noncentral chi-square distributions are stochastically} \\
&\quad \text{increasing in } \lambda \text{ and } \tfrac{(s-r_0)^2}{s} \leq \tfrac{(r_2-r_0)^2}{r_2} \text{ by calculus)} \\
&\geq P(NC\chi^2(p, \delta) \leq \frac{nc}{r_2}) \qquad (21) \\
&\quad \text{(since } \tfrac{n(r_2-r_0)^2}{r_2}\|x\|^2 \leq \delta \text{ for } x \in C\text{)}
\end{aligned}$$

The first assertion in the Theorem now follows from (21).

To prove the second assertion, it is enough to show that $F^{-1}(1-\alpha) = O(1)$ for any $0 < \alpha < 1$, since $r_2 = O(1)$. This will follow if we can exhibit $M \geq F^{-1}(1-\alpha)$ such that $M = O(1)$. However, this follows immediately on choosing $M = \frac{2p}{\alpha}$ from Chebyshev's inequality

$$P(Y > M) \leq \frac{p+\delta}{M}, \tag{22}$$

since an easy calculation shows that $\delta = O(\frac{1}{n})$. This completes the proof of the Theorem.

Remark. Again, for practical utility, an explicit prescribed sample size n_0 is necessary. This is provided in Table 1.

3 Nonregular cases

The classical asymptotic theory for nonregular distributions provides interesting departures from the regular case. For instance, for n iid observations from the $U[0,\theta]$ distribution, the MLE of θ, when normalized, converges to an exponential distribution (this is not surprising in view of the well known extreme value theory: see Galambos (1987)). Furthermore, the normalizing constant is n rather than $\sqrt{n}$. This departure from the regular case permeates into the present family of problems. We will demonstrate this by considering the $U[0,\theta]$ case. Again, we look at a number of problems.

The results in this section assume the following common structure:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0,\theta],$$

θ has a Pareto distribution with density

$$\pi(\theta) = \frac{\alpha a^\alpha}{\theta^{\alpha+1}}, \qquad \theta \geq a.$$

We assume a is known and let $\alpha_1 \leq \alpha \leq \alpha_2 < \infty$. Without loss of generality, we will work with the likelihood function

$$f(x|\theta) = \frac{n\, x^{n-1}}{\theta^n}, \qquad 0 < x \leq \theta.$$

The posterior distribution of θ is then another Pareto with density

$$d\nu_\alpha(\theta) = \frac{(n+\alpha)(x \vee a)^{n+\alpha}}{\theta^{n+\alpha+1}}, \qquad \theta \geq x \vee a.$$

3.1 Hypothesis testing and variational distance

As in the regular normal case, we will consider the possibility of a preposterior guarantee of posterior robustness simultaneously for all possible null hypotheses.

Theorem 8 *Let the likelihood and the priors be as above. Then,*

$$\begin{aligned} &\sup_{x \geq 0} \sup_{B} \left[\sup_{\pi} P(\theta \in B|x) - \inf_{\pi} P(\theta \in B|x) \right] \\ &= \left(\frac{\alpha_1 + n}{\alpha_2 + n} \right)^{\frac{\alpha_1 + n}{\alpha_2 - \alpha_1}} - \left(\frac{\alpha_1 + n}{\alpha_2 + n} \right)^{\frac{\alpha_2 + n}{\alpha_2 - \alpha_1}} \end{aligned} \tag{23}$$

Discussion. There are two principal features worth noting. First, the uniform posterior robustness. In fact, the proof of the Theorem will reveal that for any fixed x, the variational diameter of the posteriors is the RHS of (23). Secondly, the RHS of (23) is $O(\frac{1}{n})$. There is thus a difference in the rate of convergence as compared to the regular normal case.

Proof of Theorem 8: As before, we will evaluate

$$\frac{1}{2} \int |d\nu_\alpha(\theta) - d\nu_\beta(\theta)| d\theta \quad \text{for} \quad \alpha_1 \leq \beta < \alpha \leq \alpha_2 \tag{24}$$

and then evaluate its supremum over α, β and $x \geq 0$. The maximization over x will be seen to be redundant.

(24) equals, for fixed x,

$$\frac{1}{2}\int_{x\vee a}^{\infty}|\frac{(n+\alpha)(x\vee a)^{n+\alpha}}{\theta^{n+\alpha+1}}-\frac{(n+\beta)(x\vee a)^{n+\beta}}{\theta^{n+\beta+1}}|d\theta$$

$$=\frac{1}{2}\int_{1}^{\infty}|\frac{n+\alpha}{z^{n+\alpha+1}}-\frac{n+\beta}{z^{n+\beta+1}}|dz$$

$$=\frac{1}{2}\int_{1}^{R}\left(\frac{n+\alpha}{z^{n+\alpha+1}}-\frac{n+\beta}{z^{n+\beta+1}}\right)dz+\int_{R}^{\infty}\left(\frac{n+\beta}{z^{n+\beta+1}}-\frac{n+\alpha}{z^{n+\alpha+1}}\right)dz,$$

(where $R=\left(\frac{\alpha+n}{\beta+n}\right)^{\frac{1}{\alpha-\beta}}$),

$$(25)\qquad =\left(\frac{\beta+n}{\alpha+n}\right)^{\frac{\beta+n}{\alpha-\beta}}-\left(\frac{\beta+n}{\alpha+n}\right)^{\frac{\alpha+n}{\alpha-\beta}}$$

Notice (25) is free of x.

To see that (25) is maximized when $\beta=\alpha_1$ and $\alpha=\alpha_2$, first hold α fixed and let $u=\frac{\alpha-\beta}{\alpha+n}$. Then (25) equals $(1-u)^{\frac{1}{u}}.\ \frac{u}{1-u}$ on algebra. Since u cannot go outside of $[0,1]$, and $(1-u)^{\frac{1}{u}}.\ \frac{u}{1-u}$ is monotone nondecreasing on $[0,1]$, it follows that given α, (25) is maximized when $\beta=\alpha_1$ (recall $\beta<\alpha$). Symmetry gives that given β, (25) is maximized when $\alpha=\alpha_2$. The two statements now give (23).

Corollary 3 *Under the assumed model,*

$$\sup_{x\geq 0}\sup_{B}[\sup_{\pi}P(\theta\in B|x)-\inf_{\pi}P(\theta\in B|x)]=O(\frac{1}{n}).$$

Proof: Simple on using (23).

3.2 Point estimation with an uncertain loss function

As we commented in section 1, it is important to keep in mind that the loss function is as hard to elicit as a prior, perhaps even more. Theory of utility implies a bounded utility function. With this as the motivation, we will assume that we only know that we have an invariant loss $L(\theta,\delta)=W(\frac{\delta}{\theta})$ where $0\leq W(t)\leq 1$ is a nondecreasing function of $|t-1|$. The truncated invariant quadratic loss (and many others) satisfies this requirement. Thus

there is little that is assumed about the functional form of the loss function. The following problem is addressed here: take a reasonable and common point estimator. Can one then prescribe an explicit sample size which will guarantee a small range of posterior risks for this estimate simultaneously for all priors, all losses, and all x one is likely to see? We will take the MLE of θ as the procedure and demonstrate that even this towering goal is attainable. First we need to identify a set of x one is likely to observe.

Theorem 9 *Under the assumed structure,*

a. *Each marginal distribution of X is unimodal about a.*

b. $P(X < a) \to 0$ *as* $n \to \infty$ *uniformly over all marginals.*

c. *Given* $\varepsilon > 0$, $P(a \le X \le ka) \ge 1 - \varepsilon$ *uniformly over all marginals, where*

$$k = \left(\frac{n}{\varepsilon(n + \alpha_1) - \alpha_1} \right)^{\frac{1}{\alpha_1}}, \tag{26}$$

provided $n \ge \frac{1-\varepsilon}{\varepsilon} \cdot \alpha_1$.

Discussion. We will take the interval $C = [a, ka]$ to be our set of X. Notice the very curious fact that the marginals all have a common mode but the probability of being smaller than the mode is uniformly small! The interval $[a, ka]$ is thus *not* a level set of any of the marginals. But taking the set of x to be on one side of a saves an enormous amount of unnecessary technical warfare.

Proof of Theorem 9:

a. On direct calculation, the marginal distributions have densities of the form

$$m(x|\alpha) = \frac{\alpha n a^{\alpha}}{n + \alpha} \cdot \frac{x^{n-1}}{(x \vee a)^{n+\alpha}}, \qquad x \ge 0, \tag{27}$$

which are unimodal with mode at a.

b. ¿From (27), $P(X \leq a) = \frac{\alpha}{n+\alpha}$, which converges to zero uniformly in α, for $\alpha_1 \leq \alpha \leq \alpha_2 < \infty$.

c. ¿From (27),

$$P(a \leq X \leq ka) = 1 - \frac{\alpha + \frac{n}{k^\alpha}}{n+\alpha} \tag{28}$$

Thus it is sufficient to have $k \geq \left(\frac{n}{\varepsilon(n+\alpha)-\alpha}\right)^{\frac{1}{\alpha}}$. This holds by construction of k.

The following well known fact is needed due to the multiplicity of loss functions.

Lemma 5 *Let P_1, P_2 be any two probability measures on a measurable space $[S, \mathcal{B}]$. Let Ω be a family of measurable functions $W(\cdot)$ on S. Define the family of measurable sets $\mathcal{F} = \{B: B = W^{-1}(x, 1]: 0 \leq x \leq 1, W \text{ in } \Omega\}$. Then,*

$$\sup_{W \in \Omega} \left| \int W dP_1 - \int W dP_2 \right| = \sup_{B \in \mathcal{F}} |P_1(B) - P_2(B)|.$$

Theorem 10 *Let the likelihood, prior and the loss be as described before. Consider the MLE of θ, namely, $\delta(X) = X$. Then,*

$$\begin{aligned} &\sup_{x \in C} \sup_{W} \left[\sup_{\pi} r(\pi, W, \delta) - \inf_{\pi} r(\pi, W, \delta) \right] \\ &= \left(\frac{\alpha_1 + n}{\alpha_2 + n}\right)^{\frac{\alpha_1+n}{\alpha_2-\alpha_1}} - \left(\frac{\alpha_1 + n}{\alpha_2 + n}\right)^{\frac{\alpha_2+n}{\alpha_2-\alpha_1}} \end{aligned} \tag{29}$$

where $r(\pi, W, \delta)$ denotes the posterior expected loss of $\delta(x)$ when the prior is π and the loss function is W.

Discussion. Notice the remarkable fact that (29) simply equals the maximum range of posterior probability of a null hypothesis as given in (23)! This connects the point estimation problem with an uncertain loss with a hypothesis testing problem with an unspecified hypothesis. The added mystery is that the coincidence occurs only for the MLE. Note that we already know that (29) converges to zero (at a rate of $\frac{1}{n}$) and thus we can provide a

preexperimental guarantee of posterior robustness for all likely x even when little is assumed about the form of the loss function.

Proof of Theorem 10: Fix any two posteriors ν_α and ν_β.

Then the difference in the posterior risk of the MLE under ν_α and ν_β equals

$$|\int W(\frac{x}{\theta})d\nu_\alpha(\theta) - \int W(\frac{x}{\theta})d\nu_\beta(\theta)|;$$

recall $\theta \geq x \geq 0$.

Using the notation of Lemma 5, the family $\mathcal{F}$ consists of sets

$$\begin{aligned} B &= \{\theta: |\frac{x}{\theta} - 1| \geq t\} \\ &= \left[\frac{x}{1-t}, \infty\right) \end{aligned}$$

since $\theta \geq x$. Here $0 \leq t \leq 1$. Thus, by virtue of Lemma 5, it is enough to evaluate

$$\sup_{x \in C} \sup_{\alpha_1 \leq \beta < \alpha \leq \alpha_2} \sup_{0 \leq t \leq 1} |P_\alpha(\theta \geq \frac{x}{1-t}|x) - P_\beta(\theta \geq \frac{x}{1-t}|x)| \tag{30}$$

Since $x \geq a$ for $x \in C$,

$$|P_\alpha(\theta \geq \frac{x}{1-t}|x) - P_\beta(\theta \geq \frac{x}{1-t}|x)| = |(1-t)^{n+\alpha} - (1-t)^{n+\beta}|. \tag{31}$$

For given α, β, (31) is maximized at

$$t = 1 - \left(\frac{\beta + n}{\alpha + n}\right)^{\frac{1}{\alpha - \beta}}. \tag{32}$$

Substitution into (31) and a repetition of the argument following (25) establishes the Theorem.

3.3 Construction of robust confidence intervals

The purpose here is to construct an interval I in analogy with the regular normal case such that $\inf_\pi P(\theta \in I|x) \geq 1 - \gamma$ where $0 < \gamma < 1$ is specified and such that the length of I goes

to zero (uniformly in x belonging to C) at the classically attainable rate. The classically attainable rate is $\frac{1}{n}$ since the standard classical interval $[x, x\gamma^{-\frac{1}{n}}]$ has a length converging to zero at the rate $\frac{1}{n}$ if x is in a compact set independent of n (which the interval $C = [a, ka]$ is).

Theorem 11 *Let the likelihood and the prior be as in Theorem 10. Define the interval*

$$I = \left[x, x\gamma^{-(n+\alpha_1)}\right].$$

Then $\inf_{\pi} P[\theta \in I|x] \geq 1 - \gamma$, *and*

$$\begin{aligned} sup_{x\in C}\ Length\ (I) &= \left(\frac{n}{\varepsilon(n+\alpha_1)-\alpha_1}\right)^{\frac{1}{\alpha_1}} \cdot (\gamma^{-(n+\alpha_1)} - 1) \cdot a \\ &= O(\frac{1}{n}). \end{aligned}$$

Proof of the Theorem 11: Straightforward calculation gives everything except the rate of convergence. The $O(\frac{1}{n})$ rate is proved by using that $\left(\frac{n}{\varepsilon(n+\alpha_1)-\alpha_1}\right)^{\frac{1}{\alpha_1}}$ is $O(1)$ and $(\gamma^{-(n+\alpha_1)}-1)$ is $O(\frac{1}{n})$.

Remark. Again notice the departure from the corresponding problem in the regular case, where the rate of convergence was $\frac{1}{\sqrt{n}}$.

4 Nonconjugate priors

The material in sections 2 and 3 assumed conjugate priors. This made explicit calculations possible and easier. Very often, however, protection against nonconjugate priors consistent with elicited information is desirable. Many results with nonconjugate priors are available in Mukhopadhyay and DasGupta (1992). We will only describe two results with nonconjugate priors here. The robustness obtained in these results is uniform over all x and brings out a novel connection between the earlier classical works of Strawderman and Cohen (1971),

Brown and Hwang (1982) etc. and the issues of posterior robustness presented here. We discuss this at more depth after the following results.

Theorem 12 *Let $X_1, \ldots, X_n$ be iid $N(\theta, I)$ and let θ have a prior density $\pi(\theta)$ belonging to some family Γ. Assume that*

$$\sup_{\pi \in \Gamma} \sup_{\theta} \frac{||\nabla \pi(\theta)||}{\pi(\theta)} \leq K < \infty. \tag{33}$$

Then

a. *There exists a universal constant M_1 such that*

$$\sup_{x \in \Re^p} \sup_{B} \left[\sup_{\pi} P(\theta \in B|x) - \inf_{\pi} P(\theta \in B|x) \right] \leq \frac{M_1}{\sqrt{n}}. \tag{34}$$

b. *For any two priors π_1 and π_2 belonging to Γ, if $S_1(x)$ and $S_2(x)$ denote the corresponding HPD credible sets for θ of level $1 - \alpha$, then there exists a universal constant M_2 such that*

$$\sup_{x \in \Re^p} \sup_{\pi_1, \pi_2 \in \Gamma} d(S_1(x), S_2(x)) \leq \frac{M_2}{\sqrt{n}}, \tag{35}$$

where $d(S_1, S_2)$ denotes Hausdorff distance between S_1 and S_2 (see Dugundji (1975)).

Discussion. Notice how only uniform boundedness of the gradient of the logarithm of the prior results in two extremely broad posterior robustness results. The first result says that irrespective of which data may be observed, posterior robustness in any testing problem can be preexperimentally guaranteed by simply choosing the sample size large. The second result says that irrespective of which data may be observed, any two credible sets will be visually near identical. A small Lebesgue measure of the symmetric difference of $S_1(x)$ and $S_2(x)$ does *not* guarantee visual similarity. A small Hausdorff distance, however, does. The central assumption (33) is essentially a flatness condition. Normal priors do not satisfy it. In one dimension, double exponential or flatter priors satisfy (33). The importance of (33) in

frequentist decision theory, e.g., admissibility results, has been emphasized by many workers in the area.

Proof of Theorem 12: Both results follow, on a clever transformation, from earlier results in Mukhopadhyay and DasGupta (1992).

For part **a.**, we use the result that if $n = 1$, $\pi(\theta)$ satisfies (33), and we define a scaled prior $\pi_\tau(\theta) = \frac{1}{\tau}\pi(\frac{\theta}{\tau})$, then

$$\sup_{x \in \Re^p} \int |\pi_\tau(\theta|x) - \phi(||\theta - x||)|d\theta \leq \frac{M_0}{\tau} \tag{36}$$

for some universal constant M_0. (34) follows from (36) on using the fact that for probability measures P_1, P_2, the variational distance $\sup_B |P_1(B) - P_2(B)|$ equals $\frac{1}{2}\int |dP_1(\theta) - dP_2(\theta)|d\theta$, where dP_i denotes the density of P_i, and on using the transformation $\theta \to \theta\sqrt{n}$ in our problem. Then, formally, $\sqrt{n}$ can be identified with τ in the result given in (36) and n can be taken as 1 (loosely, τ and $\sqrt{n}$ are switchable). Finally use triangular inequality after using (36) once for P_1 and once for P_2.

For part **b.**, again use the same transformation in the result that if $S_\tau(x)$ is HPD for the scaled prior $\pi_\tau(\theta)$, then there exist two positive universal constants N_1 and N_2 such that

$$\begin{aligned} S_0(x) &= \{\theta: ||\theta - x|| \leq \chi_\alpha(p) - \frac{N_1}{\tau}\} \\ &\subseteq S_\tau(x) \\ &\subseteq \{\theta: ||\theta - x|| \leq \chi_\alpha(p) + \frac{N_2}{\tau}\} = S^0(x) \end{aligned} \tag{37}$$

uniformly in $x \in \Re^p$.

Interchanging τ and $\sqrt{n}$ as before (which is valid), it follows from (37) that $d(S_0, S^0)$ converges to zero at the rate of $\frac{1}{\sqrt{n}}$ uniformly in x. Hence $d(S_1, S_2)$ must go to zero uniformly in x at the same rate since S_1, S_2 *both* satisfy the inclusion property $S_0 \subseteq S_1$, $S_2 \subseteq S^0$. This proves the Theorem.

5 Summary

The main goal of this article was to demonstrate that much as we have always done in classical statistics, for example in power calculations, it is possible to prescribe a sample size which will guarantee a prespecified level of posterior robustness for any data we are likely to see. We have discussed a large variety of problems and have given evidence that the answer will depend on the problem. Many of the results indicate that such a preexperimental guarantee may be possible under broad flexibility, for instance, even when the loss and the prior are simultaneously uncertain. We are continuing our work in this particular area for semi and nonparametric priors.

Acknowledgement. Our deepest appreciation goes to Teena Seele and Brani Vidakovic for their magnanimous effort in helping finish this article in a very short time. Herman Rubin was a helpful listener on a number of occasions and we are glad to thank him.

6 Appendix

Table 1

p	ε	τ_1	τ_2	c	α	n_{01}	n_{02}	n_{03}
1	.1	.5	2	.3	.05	25		
1	.1	.5	2	.1	.05	59		
1	.1	1	5	.3	.05	21		
1	.1	1	5	.1	.05	50		
3	.1	.5	2	.3	.05	41		
3	.1	.5	2	.1	.05	99		
3	.1	1	5	.3	.05	35		
3	.1	1	5	.1	.05	89		
10	.1	.5	2	.3	.05	77		
10	.1	.5	2	.1	.05	203		
10	.1	1	5	.3	.05	68		
10	.1	1	5	.1	.05	192		

Table 1 (continued)

p	ε	τ_1	τ_2	c	α	n_{01}	n_{02}	n_{03}
1	0	.5	2	.1			3	
1	0	.5	2	.05			8	
1	0	.5	2	.01			44	
1	0	1	5	.1			1	
1	0	1	5	.05			2	
1	0	.5	2	.01			12	
1	.1	.5	2	.15				3576
1	.1	.5	2	.1				7903
1	.2	.5	2	.15				2245
1	.2	.5	2	.1				4911
1	.1	1	5	.15				1435
1	.1	1	5	.1				3193
1	.2	1	5	.15				891
1	.2	1	5	.1				1968
3	.1	.5	2	.15				11596
3	.1	.5	2	.1				23399
3	.2	.5	2	.15				9502
3	.2	.5	2	.1				18770

n_{01} = sample size needed for existence of $100(1-\alpha)\%$ robust confidence set with radius $\leq c$.

n_{02} = sample size needed for range of $P(\theta \leq 0|x) \leq c$.

n_{03} = sample size needed for range of $P(\theta \in B|x) \leq c$ uniformly in B.

Table 2

p	ε	τ_1	τ_2	n_{04}
1	.1	.5	2	30
1	.2	.5	2	15
1	.1	2	5	1
1	.2	2	5	1
2	.1	.5	2	25
2	.2	.5	2	14
2	.1	2	5	1
2	.2	2	5	1
3	.1	.5	2	22
3	.2	.5	2	14
3	.1	2	5	1
3	.2	2	5	1
10	.1	.5	2	15
10	.2	.5	2	11
10	.1	2	5	1
10	.2	2	5	1

n_{04} = sample size needed to make diameter of set of posterior means $\leq 2\sqrt{\frac{p}{n}}$.

References

[1] BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. Springer Verlag, New York.

[2] BROWN, L. D. and HWANG, J. T. (1982). A unified admissibility proof. In *Statistical Decision Theory and Related Topics*, **1** , Academic Press, New York.

[3] DALL'AGLIO, G. (1956). Sugli estremi deli momenti delle funzioni di ripartizione doppia. *Ann. Scuola Norm. Sup. Pisa* (Ser. 3), **10**, pp. 35–74.

[4] DASGUPTA, A. (1991). Diameter and volume minimizing confidence sets in Bayes and classical problems. *Ann.Statist.*, **19**, pp. 1225-1243.

[5] DUDLEY, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.*,**39**, pp. 1563 – 1572.

[6] DUGUNDJI, J. (1975). *Topology.* Prentice Hall, London.

[7] GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics.* 2nd ed. Malabar, Florida , Krieger.

[8] HEWITT, E. and STROMBERG, K. (1978). *Real and Abstract Analysis.* Springer Verlag, New York.

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

[10] KANTOROVICH, L. V. and RUBINSTEIN, G. SH. (1958). On a space of completely additive functions. *Vestnik Leningrad Univ.*, **13**, no. 7, Ser. Mat. Astron. Phys. 2: pp. 52–59.

[11] LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. John Wiley, New York.

[12] MEEDEN, G. and ISAACSON, D. (1977). Approximate behavior of the posterior distribution for a large observation. *Ann.Statist.*,**5**, pp. 899 – 908.

[13] MUKHOPADHYAY, S. and DASGUPTA, A. (1992). Classically acceptable Bayesian inference. *Tech. Report, Dept. of Statistics, Purdue University*

[14] PFANZAGL, J. (1973). The accuracy of the normal approximation for estimates of vector parameters. *Z. Wahrsch. verw. Gebiete* , **25**, pp. 171– 198.

[15] RACHEV. S. T. (1984). The Monge – Kantorovich mass transference problem and its stochastic applications. *Theor. Prob. and Appl.*, **29**, pp. 647–676.

[16] STRAWDERMAN, W. E. and COHEN, A. (1971). Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. *Ann. Math. Statist.*, **42**, pp. 270–296.

Department of Statistics
Mathematical Sciences Building
Purdue University
West Lafayette, Indiana 47907, USA