THE STEIN EFFECT AND BAYESIAN ANALYSIS: A REEXAMINATION

by

Jean-Francois Angers and James O. Berger
Purdue University

Technical Report #85-6

Department of Statistics
Purdue University

May 1985

THE STEIN EFFECT AND BAYESIAN ANALYSIS: A REEXAMINATION

Jean-Francois Angers and James O. Berger

Statistics Department, Purdue University, West Lafayette, IN 47907, U.S.A.

ABSTRACT

The Stein effect, that one could improve frequentist risk by combining "independent" problems, has long been an intriguing paradox to statistics. We briefly review the Bayesian view of the paradox, and indicate that previous justifications of the Stein effect, through concerns of "Bayesian robustness," were misleading. In the course of doing so, several existing robust Bayesian and Stein-effect estimators are compared for a variety of situations.

Key Words and Phrases: Stein effect; Bayesian robustness; independent problems; classes of priors.

1. INTRODUCTION

Suppose $X_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ (σ_i^2 known), and that it is desired to estimate θ_i by an estimator $\delta_i(x_i)$ under squared error loss. The standard estimator $\delta_i^0(x_i) = x_i$ is admissible when frequentist risk (expected loss) is the criterion. Stein (1955) showed, however, that if one combined p such problems, i.e., considered the problem of estimating $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_p)^t$, based on $\underset{\sim}{X} = (X_1,\ldots,X_p)^t$ under sum of squares error loss, then the corresponding natural estimator, $\underset{\sim}{\delta}^0(\underset{\sim}{x}) = (\delta_1^0(x_1),\ldots,\delta_p^0(x_p))^t = \underset{\sim}{x}$, is inadmissible when $p \geq 3$. Indeed, James and Stein (1960) showed that (when all $\sigma_i^2 = 1$) a better estimator is given by

$$\underset{\sim}{\delta}^{J\text{-}S}(\underset{\sim}{x}) = \left(1 - \frac{(p-2)}{|\underset{\sim}{x}|^2}\right) \underset{\sim}{x} ; \qquad (1.1)$$

thus, in terms of frequentist risk

$$R(\underset{\sim}{\theta},\underset{\sim}{\delta}) = E_{\underset{\sim}{\theta}} \, |\underset{\sim}{\delta}(\underset{\sim}{X}) - \underset{\sim}{\theta}|^2,$$

James and Stein (1960) showed that $R(\underset{\sim}{\theta},\underset{\sim}{\delta}^{J\text{-}S}) < R(\underset{\sim}{\theta},\underset{\sim}{\delta}^0)$ for all $\underset{\sim}{\theta}$.

Note that $\underset{\sim}{\delta}^{J\text{-}S}$ involves all the x_j in estimating each θ_i. This is counter-intuitive when the θ_i are from completely different problems, as in the following example.

Example 1. Suppose θ_1 is the mean corn yield per acre in Indiana in 1985, θ_2 is the mass of the universe, and θ_3 is the age of a newly discovered Greek artifact. A large business conglomerate is simultaneously interested in estimating these three quantities, and the directors ascertain that the overall loss to the business in mis-estimation is $[(\theta_1-\delta_1)^2 + (\theta_2-\delta_2)^2 + (\theta_3-\delta_3)^2]$. It

strikes most people as rather odd to combine measurements involving such θ_i in a manner such as (1.1).

This phenomenon, that one can combine independent problems and obtain improved estimators, is called the Stein effect. It is a very general phenomenon in frequentist decision theory (see Berger (1985a) for references). Several explanations for the phenomenon have been proposed. One common explanation (cf. Efron and Morris (1973)) is that if the θ_i are related, as in being i.i.d. observations from a common population, then estimators similar to $\underset{\sim}{\delta}^{J-S}$ can arise as empirical Bayes estimators. However, while such situations are extremely important areas for application of Stein estimation, they are not cases in which the coordinate problems are truly independent; since the θ_i arise from a common population, knowledge concerning θ_j is clearly relevant to estimating θ_i. We will be considering instead situations such as Example 1, where the Stein effect holds, but in which it would be absurd to think of the θ_i as coming from a common population.

A second common "explanation" of the Stein effect is that it is natural because the losses for the different problems have been combined. Indeed, such combination can force a dependence into the final estimate when one has a non-linear utility function. The Stein phenomenon is usually stated, however, in terms of combining the individual losses, $L_i(\theta_i,\delta_i)$, in a linear fashion, to yield

$$L(\underset{\sim}{\theta},\underset{\sim}{\delta}) = \sum_{i=1}^{p} L_i(\theta_i,\delta_i) \tag{1.2}$$

(or perhaps some weighted linear combination), and the situation is then not so clear. It may seem that this forcing of the L_i to be on a common scale might be the cause of a dependence in estimation, but a number of intuitive arguments refute this. The simplest, which we will pursue in this paper, is the Bayesian view of the situation.

Let $\pi(\underset{\sim}{\theta})$ denote a prior (density for convenience) for $\underset{\sim}{\theta}$. The assumption that the problems are completely unrelated can be expressed by the Bayesian condition that the θ_i are apriori independent, i.e., that

$$\pi(\underset{\sim}{\theta}) = \prod_{i=1}^{p} \pi_i(\theta_i). \tag{1.3}$$

Now, for such a prior and with a loss as in (1.2), it is straightforward to verify that the Bayes rule is

$$\underset{\sim}{\delta}^{\pi}(\underset{\sim}{x}) = (\delta^{\pi_1}(x_1),\ldots,\delta^{\pi_p}(x_p))^t,$$

where $\delta^{\pi_i}(x_i)$ is the Bayes rule for estimating θ_i under loss L_i and with respect to the prior π_i. Thus a Bayesian feels that, if the problems are truly unrelated (and the prior is known), then only x_i should be involved in the estimation of θ_i. The combination of losses in a linear fashion, such as (1.2), has no effect on the Bayes rule.

To obtain a Bayesian explanation for the Stein effect in a setting such as Example 1, we must, therefore, leave the pure Bayesian framework. A natural and practically relevant direction to look is towards concerns of Bayesian robustness. Indeed, Berger (1982a) so approached the problem, imagining the specification

of a prior π^0, of the form (1.3), and advocating consideration of the class of priors

$$\Gamma = \{\pi = (1-\varepsilon)\ \pi^0 + \varepsilon q;\ q \text{ an arbitrary distribution}\}, \qquad (1.4)$$

where $0 < \varepsilon < 1$ is a constant reflecting the accuracy of the specification of π^0. The idea is that, in practice, π^0 will merely be an approximate quantification of prior beliefs, and that priors "close to" π^0 are also plausible. The robust Bayesian seeks to conduct an analysis which is likely to be satisfactory for all such "close" priors. The class in (1.4) is a simple and useful class of close priors. (It might be argued that allowing the "contamination," q, to be arbitrary allows too much deviation from π^0, but for the purposes of this paper such a concern is irrelevant; the same results can be shown to hold as long as there are priors in Γ corresponding to q which give arbitrarily small mass to compact sets).

In Berger (1982a, 1984), it was shown that the Stein effect is valuable in achieving Bayesian robustness with respect to classes such as (1.4); indeed, it was shown to be possible (in symmetric situations) to construct Stein-type estimators which are nearly optimal in terms of Bayes risk

$$r(\pi,\underset{\sim}{\delta}) = E^{\pi}\ R(\underset{\sim}{\theta},\underset{\sim}{\delta}) = \int R(\underset{\sim}{\theta},\underset{\sim}{\delta})\ \pi(\underset{\sim}{\theta})\ d\underset{\sim}{\theta}, \qquad (1.5)$$

for an elicited prior π^0, and yet are also highly robust with respect to Γ, in the sense that

$$r_{\Gamma}(\underset{\sim}{\delta}) = \sup_{\pi\in\Gamma} r(\pi,\underset{\sim}{\delta}) \qquad (1.6)$$

is also nearly optimal. The criteria (1.5) and (1.6) are, of course, partly frequentist in nature, and are of interest to a Bayesian only in that they can indicate good average posterior expected loss for $\underset{\sim}{\delta}$. (In Berger (1984, 1985b) the value of such criteria to a Bayesian is more fully discussed.) It suffices here to note that these criteria are of interest, and that Berger (1982a, 1982b) showed that a Stein-type estimator can do much better by these criteria than can robust coordinatewise independent estimators of the form

$$\underset{\sim}{\delta}(\underset{\sim}{x}) = (\delta_1(x_1),\ldots,\delta_p(x_p))^t. \tag{1.7}$$

Thus there appeared to be value in the Stein effect, even from a Bayesian perspective.

The intuitive paradox has caused us to continue to examine the issue, and we noticed a disturbing step in the above argument. The step was the inclusion of priors in (1.4) that are coordinatewise dependent; even though π^0 is of the form (1.3), many (indeed most) of the priors in (1.4) will have dependent coordinates. And, although it may often be the case that dependence among the θ_i is possible (in which case (1.4) might be reasonable), our stated goal was to investigate the value of the Stein effect for completely unrelated problems. Thus, in Example 1, while we might be quite uncertain as to the choice of π^0, our prior belief in the independence of the π_i may be so strong that only priors satisfying (1.3) would be deemed reasonable.

A sensible class of priors for such a situation is

$$\Gamma_1 = \{\pi = \prod_{i=1}^{p}[(1-\varepsilon)\,\pi_i^0(\theta_i) + \varepsilon q_i(\theta_i)],\ \text{the } q_i \text{ arbitrary}\}. \tag{1.8}$$

(Again, the only essential requirement for the following results is that Γ_1 contain priors corresponding to q_i which give arbitrarily small mass to compact sets.) This class contains priors which are close to π^0 and which preserve the independence structure. The obvious question to ask is--Does the Stein effect allow improved performance, from a robust Bayesian perspective, for Γ_1? We will see that the answer is, essentially, no; there does not appear to be value to the Stein effect in combining independent problems.

To avoid possible confusion, we again emphasize that we are not considering the formally related empirical Bayes situation where π is of the form (1.3), and the coordinatewise priors, π_i, have unknown and partially common features. The coordinate problems are not then independent to a Bayesian (placing a second stage prior on the common features and integrating out clearly reveals the dependence), and a Bayesian will naturally employ coordinatewise dependent estimators.

2. PRELIMINARIES

For convenience, we restrict ourselves to the symmetric situation with independent $X_i \sim \eta(\theta_i,\sigma^2)$, $i=1,\ldots,p$, σ^2 known, loss $L_i(\theta_i,\delta_i) = (\theta_i-\delta_i)^2$ in estimating θ_i by δ_i, and overall loss $L(\underset{\sim}{\theta},\underset{\sim}{\delta}) = \sum_{i=1}^{p} (\theta_i-\delta_i)^2$ in estimating $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_p)^t$ by $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_p)^t$. We imagine that the θ_i are apriori independent, and that rough quantification of the prior information results in a $\eta(\mu_i,\tau^2)$ prior, to be denoted π_i^0, for each θ_i. Here we presume that τ^2 and $\underset{\sim}{\mu} = (\mu_1,\ldots,\mu_p)^t$ are subjectively specified numbers. Thus the "base" prior for the Bayesian robustness investigation is $\pi^0(\underset{\sim}{\theta}) = \prod_{i=1}^{p} \pi_i^0(\theta_i)$, the $\eta_p(\underset{\sim}{\mu},\tau^2 I)$ distribution. Note that, since τ^2 is assumed known, this is not an empirical Bayes situation. It is a situation where, by coincidence, we have p completely independent problems which have similar known variances. This is not being put forth as a realistic practical situation (although the amount of theoretical

literature on the situation is quite astounding), but rather because it is the situation most conducive to improvements via the Stein effect; if the Stein effect is of no or limited value to Bayesian robustness here, it is unlikely to be helpful in more realistic scenarios involving completely independent problems with very different variances.

We will work with two classes of prior distributions. The first has already been discussed, namely Γ_1 in (1.8). The second class is somewhat less conservative, allowing at most one of the coordinates of $\underset{\sim}{\theta}$ to have a contaminated prior. The class is

$$\Gamma_2 = \{\pi(\underset{\sim}{\theta}) = \frac{1}{p}\sum_{j=1}^{p}([(1-\epsilon)\pi_j^0(\theta_j) + \epsilon q_j(\theta_j)]\prod_{i\neq j}\pi_i^0(\theta_i)), \text{ the } q_j \text{ arbitrary}\}. \quad (2.1)$$

Note that $\Gamma_2 \subset \Gamma_1$, so that demanding robustness with respect to Γ_2 is less stringent than demanding it for Γ_1.

By looking at $r(\pi^0,\underset{\sim}{\delta})$, $r_{\Gamma_1}(\underset{\sim}{\delta})$, and $r_{\Gamma_2}(\underset{\sim}{\delta})$ (see (1.5) and (1.6)), we hope to obtain reasonable indications of the performance of an estimator $\underset{\sim}{\delta}$. It is actually somewhat more conveninet to normalize $r(\pi^0,\underset{\sim}{\delta})$ and $r_\Gamma(\underset{\sim}{\delta})$ as follows: define

$$RSR(\pi^0,\underset{\sim}{\delta}) = \frac{r(\pi^0,\underset{\sim}{\delta}) - r(\pi^0,\underset{\sim}{\delta}^{\pi^0})}{r(\pi^0,\underset{\sim}{\delta}^0) - r(\pi^0,\underset{\sim}{\delta}^{\pi^0})} = \frac{(\sigma^2+\tau^2)}{p\sigma^4}\, r(\pi^0,\underset{\sim}{\delta}) - \frac{\tau^2}{\sigma^2}, \quad (2.2)$$

and

$$r_\Gamma^*(\underset{\sim}{\delta}) = \frac{(\sigma^2+\tau^2)}{p\sigma^4}\, r_\Gamma(\underset{\sim}{\delta}) - \frac{\tau^2}{\sigma^2}. \quad (2.3)$$

The "relative savings risk" of $\underset{\sim}{\delta}$, $RSR(\pi^0, \underset{\sim}{\delta})$, measures the proportion of available improvement over $\underset{\sim}{\delta}^0(\underset{\sim}{x}) = \underset{\sim}{x}$ that is achieved by $\underset{\sim}{\delta}$ (for the prior π^0). RSR near 0 is optimal, while RSR near 1 indicates performance as bad as that of $\underset{\sim}{\delta}^0$. (Efron and Morris (1971) first introduced this concept.) The main purpose in using these scaled versions of risk is that the scaled versions are often independent of σ^2, τ^2, and $\underset{\sim}{\mu}$.

Note also that r_Γ (or r_Γ^*) can be used to define Γ-minimaxity. A Γ-minimax estimator is an estimator which minimizes r_Γ (or, equivalently, r_Γ^*) over all possible estimators.

3. THE ESTIMATORS AND RISKS

The estimators that are to be considered are various robustified versions of the Bayes estimator with respect to π^0, which is given for the situation of Section 2 by

$$\underset{\sim}{\delta}^{\pi^0}(\underset{\sim}{x}) = (\delta_1^{\pi^0}(x_1), \ldots, \delta_p^{\pi^0}(x_p))^t,$$

where

$$\delta_i^{\pi^0}(x_i) = x_i - \frac{\sigma^2}{(\sigma^2 + \tau^2)}(x_i - \mu_i).$$

Clearly $RSR(\pi^0, \underset{\sim}{\delta}^{\pi^0}) = 0$, but it is easy to see that $r_{\Gamma_1}^*(\underset{\sim}{\delta}^{\pi^0}) = r_{\Gamma_2}^*(\underset{\sim}{\delta}^{\pi^0}) = \infty$, so that $\underset{\sim}{\delta}^{\pi^0}$ is highly suspect from a robustness viewpoint. (Note that $r_\Gamma^*(\underset{\sim}{\delta}^0) = 1$ for any Γ, so that values of r_Γ^* larger than one can be avoided.)

The most natural way to robustify $\underset{\sim}{\delta}^{\pi^0}$ is on a coordinatewise basis. An appealing choice is the limited translation estimator of Efron and Morris (1971), given by

$$\underset{\sim}{\delta}^{M,I}(\underset{\sim}{x}) = (\delta_1^{M,I}(x_1),\ldots,\delta_p^{M,I}(x_p))^t, \tag{2.4}$$

where

$$\delta_i^{M,I}(x_i) = x_i - \min\left\{1, \frac{[M(\sigma^2 + \tau^2)]^{\frac{1}{2}}}{|x_i - \mu_i|}\right\} \frac{\sigma}{(\sigma^2 + \tau^2)}(x_i - \mu_i).$$

A robustified version of $\underset{\sim}{\delta}^{\pi^0}$ that involves the Stein effect is given in Berger (1982b), and can be written as

$$\underset{\sim}{\delta}^{M,S}(\underset{\sim}{x}) = \underset{\sim}{x} - \min\{1, \rho_{M,p}(|\underset{\sim}{x} - \underset{\sim}{\mu}|^2/(\sigma^2 + \tau^2))\} \frac{\sigma^2}{(\sigma^2 + \tau^2)}(\underset{\sim}{x} - \underset{\sim}{\mu}), \tag{2.5}$$

where, for p = 3,5, and 7 (the 3 cases to be considered in this paper), $\rho_{M,p}(r)$ is given by

$$\rho_{M,3}(r) = \frac{\sqrt{M}}{\sqrt{r}} + \frac{2}{r}, \quad \rho_{M,5}(r) = \frac{M}{2 + \sqrt{Mr}} + \frac{6}{r},$$

$$\rho_{M,7}(r) = \frac{M(2 + \sqrt{Mr})}{12 + 6\sqrt{Mr} + Mr} + \frac{10}{r}.$$

For general discussion of the properties of these estimators, see Berger (1982b). Their Bayesian robustness is indicated by the fact that they equal $\underset{\sim}{\delta}^{\pi^0}$ if $|\underset{\sim}{x} - \underset{\sim}{\mu}|$ is small (i.e., the prior information appears to be correct), and yet are not allowed to deviate excessively from $\underset{\sim}{\delta}^0(\underset{\sim}{x}) = \underset{\sim}{x}$ if $|\underset{\sim}{x} - \underset{\sim}{\mu}|$ is large.

For various ε and Γ_1 or Γ_2 (defined in (1.8) or (2.1), respectively), we will calculate and compare the "most robust" estimators of the forms (2.4) and (2.5). This will be done by simply minimizing $r^*_\Gamma(\underset{\sim}{\delta}^{M,I})$ of $r^*_\Gamma(\underset{\sim}{\delta}^{M,S})$ over M (for the various ε and Γ). For notational convenience, M* will be used to denote the optimal value of M in each situation. (Note that this will, in general, depend on the choice of Γ, ε, and the form of the estimator.)

The interest in $\underset{\sim}{\delta}^{M^*,I}$ is that it can be shown to be nearly optimal among <u>all</u> coordinatewise independent estimators (i.e., those of the form (1.7)), in terms of minimizing $r^*_{\Gamma_1}$ or $r^*_{\Gamma_2}$. Likewise, $\underset{\sim}{\delta}^{M^*,S}$ is nearly optimal among all "Stein-effect" estimators of the form $\underset{\sim}{\delta}(\underset{\sim}{x}) = \underset{\sim}{x} - \rho(|\underset{\sim}{x} - \underset{\sim}{\mu}|^2)(\underset{\sim}{x} - \underset{\sim}{\mu})$. (The arguments are similar to those in Berger and Berliner (1984) or Marazzi (1985), and will be omitted.) Thus a comparison of the performance of these estimators should provide a strong indication as to the value of the Stein effect in the situations considered. The following lemmas present risk formulas needed for the calculations and comparisons.

<u>Lemma 1</u>. The estimator $\underset{\sim}{\delta}^{M,I}$ has

$$RSR(\pi^0, \underset{\sim}{\delta}^{M,I}) = [1 - \psi_1(M)][1 + M] - \left(\frac{2M}{\pi}\right)^{\frac{1}{2}} e^{-M/2}, \qquad (2.6)$$

where ψ_1 is the c.d.f. of the chi-square distribution with one degree of freedom. Also

$$r^*_{\Gamma_1}(\underset{\sim}{\delta}^{M,I}) = (1 - \varepsilon)RSR(\pi^0, \underset{\sim}{\delta}^{M,I}) + \varepsilon(M + 1) \tag{2.7}$$

and

$$r^*_{\Gamma_2}(\underset{\sim}{\delta}^{M,I}) = (1 - \frac{\varepsilon}{p})RSR(\pi^0, \underset{\sim}{\delta}^{M,I}) + \frac{\varepsilon}{p}(M + 1). \tag{2.8}$$

Proof. Equation (2.6) is given in a slightly different form in Efron and Norris (1971). To verify (2.7), observe that, since Γ_1 contains only priors of the form (1.3),

$$\sup_{\pi \in \Gamma_1} r(\pi^0, \underset{\sim}{\delta}^{M,I}) = \sup_{\pi \in \Gamma_1} \sum_{i=1}^{p} r(\pi_i, \delta_i^{M,I})$$

$$= \sup_{\pi \in \Gamma_1} \sum_{i=1}^{p} (1 - \varepsilon)r(\pi_i^0, \delta_i^{M,I}) + \varepsilon r(q_i, \delta_i^{M,I})$$

$$= (1 - \varepsilon)r(\pi^0, \underset{\sim}{\delta}^{M,I}) + \varepsilon \sum_{i=1}^{p} \sup_{q_i} r(q_i, \delta_i^{M,I})$$

$$= (1 - \varepsilon)r(\pi^0, \underset{\sim}{\delta}^{M,I}) + \varepsilon \sum_{i=1}^{p} \sup_{\theta_i} R(\theta_i, \delta_i^{M,I})$$

$$= (1 - \varepsilon)r(\pi^0, \underset{\sim}{\delta}^{M,I}) + \varepsilon p\left(\frac{M\sigma^4}{(\sigma^2 + \tau^2)} + \sigma^2\right),$$

the last fact following from a result of Efron and Morris (1971). Using the

definitions of RSR and $r^*_{\Gamma_1}$ and simplifying, yields (2.7). Equation (2.8) is similarly established. □

Lemma 2. The estimator $\underset{\sim}{\delta}^{M,S}$ has

$$RSR(\pi^0, \underset{\sim}{\delta}^{M,S}) = [1 - \psi_p(y)][1 + \frac{M}{p}] - \frac{(y/2)^{p/2} e^{-y/2}}{\Gamma(1 + p/2)}, \qquad (2.9)$$

where ψ_p is the c.d.f. of the chi-square distribution with p degrees of freedom and y is the solution to the equation

$$\rho_{M,p}(y) = 1.$$

Also,

$$r^*_{\Gamma_1}(\underset{\sim}{\delta}^{M,S}) = (1 - \varepsilon)^p \, RSR(\pi^0, \underset{\sim}{\delta}^{M,S}) + [1 - (1 - \varepsilon)^p](\frac{M}{p} + 1), \qquad (2.10)$$

and

$$r^*_{\Gamma_2}(\underset{\sim}{\delta}^{M,S}) = (1 - \varepsilon) \, RSR(\pi^0, \underset{\sim}{\delta}^{M,S}) + \varepsilon(\frac{M}{p} + 1). \qquad (2.11)$$

Proof. Equation (2.9) follows from Berger (1982b) or Chen (1983). To establish (2.10), note first that any distribution of the form

$$\pi^*(\theta) = [\underset{\text{some indices}}{\Pi} \pi_i^0(\theta_i)] \, [\underset{\text{other indices}}{\Pi} q_i(\theta_i)] \qquad (2.12)$$

satisfies

$$\sup_{\{q_i\}} r(\pi^*, \underset{\sim}{\delta}^{M,S}) = \sup_{\underset{\sim}{\theta}} R(\underset{\sim}{\theta}, \underset{\sim}{\delta}^{M})$$

$$= p\sigma^2 + \frac{M\sigma^4}{(\sigma^2 + \tau^2)} .$$

(These equalities follow by letting one of the involved q_i give mass to a point θ_i^*, letting $|\theta_i^*| \to \infty$, and using results about the behavior of $\rho_{M,p}$ from Berger (1982b).) But any $\pi \in \Gamma_1$ can be written

$$\pi(\underset{\sim}{\theta}) = (1 - \varepsilon)^p \, \pi^0(\underset{\sim}{\theta}) + \sum_{\ell} \pi_\ell^*(\underset{\sim}{\theta}),$$

where the π_ℓ^* are of the form (2.12) and have total mass $[1 - (1 - \varepsilon)^p]$; thus

$$\sup_{\pi \in \Gamma_1} r(\pi, \underset{\sim}{\delta}^{M,S}) = (1 - \varepsilon)^p \, r(\pi^0, \underset{\sim}{\delta}^{M,S}) + \sup_{\{qi\}} \sum_{\ell} r(\pi_\ell^*, \underset{\sim}{\delta}^{M,S})$$

$$= (1 - \varepsilon)^p \, r(\pi^0, \underset{\sim}{\delta}^{M,S}) + [1 - (1 - \varepsilon)^p][p\sigma^2 + \frac{M\sigma^4}{(\sigma^2 + \tau^2)}].$$

Using the definitions of RSR and $r_{\Gamma_1}^*$ and simplifying, yields (2.10). The verification of (2.11) is similar and will be omitted. □

Having derived explicit formulas for $r_{\Gamma_1}^*$ and $r_{\Gamma_2}^*$ for the estimators being

considered, it is a sample matter to numerically minimize over M to find the optimal M*. Table 1 presents the values of M* for various ε and p = 3,5,7. Thus, when ε = .2, p = 3, and Γ_2 is the class of priors considered, the estimator of the form (2.5) which minimizes $r^*_{\Gamma_2}$ is $\underset{\sim}{\delta}^{.42,S}$ (i.e., M* = .42).

Table 2 compares the robustness of $\underset{\sim}{\delta}^{M*,I}$ and $\underset{\sim}{\delta}^{M*,S}$, as measured by $r^*_{\Gamma_1}$, for various ε and p = 3, 5, and 7. (The estimator $\underset{\sim}{\delta}^{T}$ will be discussed shortly.) Table 3 presents the analogous results for robustness as measured by $r^*_{\Gamma_2}$. In both cases, $\underset{\sim}{\delta}^{M*,I}$ is superior to $\underset{\sim}{\delta}^{M*,S}$. Thus greater overall robustness appears to be available from the class of coordinatewise independent estimators (i.e. (1.7)), than from the class of Stein-effect estimators. This is in sharp contrast to the results obtained in Berger (1982a, b, 1984) for the "dependent" class Γ in (1.4).

Of course, $r^*_{\Gamma_1}$ and $r^*_{\Gamma_2}$ do not tell the whole story, in that they essentially measure only the "worst" that can happen. It is thus interesting to also look at the "best" that can happen in using an estimator $\underset{\sim}{\delta}$; this best will be roughly given by RSR(π^0,$\underset{\sim}{\delta}$), which again can be thought of as the percentage of available Bayesian gains sacrificed by using $\underset{\sim}{\delta}$ instead of $\underset{\sim}{\delta}^{\pi^0}$. Table 4 gives values of RSR for $\underset{\sim}{\delta}^{M*,I}$ and $\underset{\sim}{\delta}^{M*,S}$, when M* is the optimal choice with respect to Γ_1, while Table 5 gives the corresponding values of RSR, when M* is the optimal choice with respect to Γ_2. Observe that the Stein-effect estimators seem to do better than the coordinatewise independent estimators. Thus, while $\underset{\sim}{\delta}^{M*,I}$ is generally better than $\underset{\sim}{\delta}^{M*,S}$ in the "worst case scenario," the reverse seems to be true in the "best case."

To attempt a finer comparison of the coordinatewise independent and Stein-effect estimators in terms of RSR, it was decided to replace M* in $\underset{\sim}{\delta}^{M*,I}$ by M' the value of M for which $\underset{\sim}{\delta}^{M',I}$ and $\underset{\sim}{\delta}^{M*,S}$ are equally robust (i.e., for which $r^*_\Gamma(\underset{\sim}{\delta}^{M',I}) = r^*_\Gamma(\underset{\sim}{\delta}^{M*,S})$). The RSR of $\underset{\sim}{\delta}^{M',I}$ was then calculated, and is given in

Tables 4 and 5. The results are unclear. For small ε the coordinatewise independent estimator, $\underset{\sim}{\delta}^{M',I}$, tends to be better, while for large ε the Stein-effect estimator, $\underset{\sim}{\delta}^{M*,S}$, has smaller RSR. This is, perhaps, not surprising, in that, at the extreme $\varepsilon = 1$, only minimax estimators are allowed, and only Stein-effect estimators can be minimax and have small RSR.

A different method for achieving a type of Bayesian robustness was proposed in Stein (1981) (see also Dey and Berger (1983)). The method, when applied to our situation, results in the estimator

$$\underset{\sim}{\delta}^{T}(\underset{\sim}{x}) = \underset{\sim}{x} - \min\left\{\frac{\sigma^2}{(\sigma^2+\tau^2)}, \frac{2(\ell-2)}{|\underset{\sim}{z}|^2}\right\}\underset{\sim}{z};$$

here $\underset{\sim}{z} = (z_1,\ldots,z_p)^t$, $z_i = \text{sgn}(y_i)\min\{|y_i|,|y|_{(\ell)}\}$, $y_i = (x_i - \mu_i)$, and $|y|_{(\ell)}$ is the ℓ^{th} order statistic of the sequence $\{|y_1|,|y|_2,\ldots,|y|_p\}$. The choice of ℓ should correspond roughly to a guess as to the number of θ_i for which prior misspecification is feared; because of the small p in this study and because Γ_2 corresponds to the case where at most one coordinate is misspecified, the choice $\ell = p-1$ was made.

Using results of Dey and Berger (1983), RSR $(\pi^0,\underset{\sim}{\delta}^T)$ can be calculated. For $p = 5$ and $p = 7$, the values are .383 and .337, respectively. Furthermore $r^*_{\Gamma_1}(\underset{\sim}{\delta}^T)$ and $r^*_{\Gamma_2}(\underset{\sim}{\delta}^T)$ can be calculated in a manner similar to (but more complicated than) Lemmas 1 and 2. The key idea is to recognize that, if q_i gives probability one to θ_i^* and $|\theta_i^*| \to \infty$, then $\underset{\sim}{\delta}^T$ effectively becomes the estimator with $z_j = y_j$ for $j \neq i$, and $z_i = \text{sgn}(y_i)\max_{j\neq i}|y_j|$; although this is the worst case, a Stein effect still remains. If, on the other hand, one is maximizing over two contaminations q_i and q_j, they can be chosen to give probability one to

θ_i^* and θ_j^*, respectively, with $|\theta_i^*| \to \infty$ and $|\theta_j^*| \to \infty$; $\underset{\sim}{\delta}^T$ will then collapse back to $\underset{\sim}{\delta}^0(\underset{\sim}{x}) = \underset{\sim}{x}$. Thus, in the expansion of terms in $r^*_{\Gamma_1}$ or $r^*_{\Gamma_2}$, those involving at most one q_i yield reduced Bayes risk, while those involving two or more of the q_i result in the minimax risk. Expressions in Dey and Berger (1983) can be used to calculate all terms.

Tables 2 and 3 give values of $r^*_{\Gamma_1}$ and $r^*_{\Gamma_2}$ for $\underset{\sim}{\delta}^T$, when p = 5 or 7. (For p = 3, $\underset{\sim}{\delta}^T(\underset{\sim}{x}) = \underset{\sim}{\delta}^0(\underset{\sim}{x}) = \underset{\sim}{x}$, since $\ell = p - 1 = 2$.) Even for Γ_2 (a class of priors for which $\underset{\sim}{\delta}^T$ would seem ideally suited, since $\underset{\sim}{\delta}^T$ truncates precisely one "bad" coordinate here), the performance of $\underset{\sim}{\delta}^T$ seems inferior to that of the robust coordinatewise independent estimator. Its value of r^*_Γ is always larger, and its RSR (.383 or .337 for p = 5 or 7) is better only for quite large ε. Observe that $\underset{\sim}{\delta}^T$ does seem to be better than $\underset{\sim}{\delta}^{M^*,S}$ in terms of r^*_Γ, unless ε is small, but its RSR is never better than that of $\underset{\sim}{\delta}^{M^*,S}$.

4. CONCLUSIONS

The numerical results in Section 3 indicate that a Bayesian who is quite certain that several problems are unrelated, but is otherwise somewhat uncertain about the prior specification, will not see any clear value in utilizing the Stein effect; superior robustness can be achieved through use of, say, coordinatewise independent limited translation estimators (except possibly for large ε, where no clear conclusions emerged from the limited criteria considered). The coordinatewise independent estimators also have several additional appealing propoerties. First, in contrast with the Stein-effect estimators (though not necessarily the $\underset{\sim}{\delta}^T$ versions), the coordinatewise independent estimators do not suffer from the problem that one or more badly specified π_i^0 can cause the entire estimator to collapse back to $\underset{\sim}{x}$. (A single large $|x_i - \mu_i|$ will cause $\underset{\sim}{\delta}^{M,S}$ to

be approximately $\underset{\sim}{x}$; one bad apple can thus ruin the pie.) A related type of "additional robustness," of the coordinatewise independent estimators, is robustness with respect to the loss. The coordinatewise risk of $\underset{\sim}{\delta}^{M^*,S}$ can be quite large (cf. Efron and Morris (1973)), while the coordinatewise risk of $\underset{\sim}{\delta}^{M^*,I}$ is inherently controlled; thus, if there is uncertainty about the "weights" that should be assigned to each component of the loss, $\underset{\sim}{\delta}^{M^*,I}$ can be substantially more robust. ($\underset{\sim}{\delta}^{T}$ will be substantially more robust than $\underset{\sim}{\delta}^{M^*,S}$ in terms of this type of loss robustness, but, again, $\underset{\sim}{\delta}^{T}$ sacrifices a good deal in terms of r_Γ^* and RSR.) Related to this loss robustness is the fact that there is no need to determine weights for the coordinatewise losses, $L_i(\theta_i,\delta_i)$, in calculating $\underset{\sim}{\delta}^{M^*,I}$ (just do each coordinate separately), while the optimal Stein-effect estimator will depend crucially on such weights. From a practical perspective this might be of paramount concern, since it is often very difficult to convince practitioners to compare and select weights for the losses from the separate coordinate problems.

Note that the above conclusions have some force for non-Bayesians also, because of the nature of Stein estimation. It can be argued (see Berger and Berliner (1984) and the discussion by Brown in Berger (1984)) that (i) prior information must be utilized to select an alternative to $\underset{\sim}{\delta}^{0}$; (ii) the most appealing method, in Stein estimation, of utilizing such information is to specify a prior distribution, π^0, and seek to minimize $r(\pi^0,\underset{\sim}{\delta})$ subject to some frequentist risk restriction (say, $R(\underset{\sim}{\theta},\underset{\sim}{\delta}) \leq C$); and (iii) a more intuitively accessible mathematical formulation of this last problem is the Γ-minimax problem of selecting a Γ as in (1.4), (1.8), or (2.1) (with ε related to C), and finding $\underset{\sim}{\delta}$ to minimize $r_\Gamma^*(\underset{\sim}{\delta})$. Thus following a natural path within frequentist Stein-estimation theory leads to consideration of the issues we have been discussing.

Of course, the conclusion that the Stein effect is of no clear value in

combining completely unrelated problems does not say that it is of no use in combining related problems. And, from a practical perspective, the only problems in which practitioners would even think of using a combined estimator are those in which the parameters might well be related. The message, therefore, is simply that there is no "paradoxical" feature of Stein estimation to a Bayesian. If there is reason to suspect some prior relationship between the θ_i, a Bayesian might well use a (suitably tailored) Stein-effect estimator; otherwise he will not. The choice would involve careful consideration of prior knowledge.

Even when the θ_i are clearly independent apriori, a Bayesian consultant can utilize the Stein-effect when dealing with frequentist clients. The use of $\underset{\sim}{\delta}^{M^*,S}$ can be shown to be preferable to the use of $\underset{\sim}{\delta}^{0}(\underset{\sim}{x}) = \underset{\sim}{x}$ from a Bayesian perspective (even conditionally); $\underset{\sim}{\delta}^{M^*,S}$ can thus be recommended when frequentist justification is an externally imposed requirement. Of course, a Bayesian might feel somewhat silly in recommending the combining of completely unrelated problems in such a scenario, but that is the price that would have to be paid for agreeing to guarantee standard frequentist validity.

BIBLIOGRAPHY

Berger, J., (1982a). Bayesian robustness and the Stein effect. J. Amer. Statist. Assoc., 77, 358-368.

Berger, J., (1982b). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In Statistical Decision Theory and Related Topics III, S. S. Gupta and J. Berger (Eds.). Academic Press, New York.

Berger, J., (1984). The robust Bayesian viewpoint. In Robustness in Bayesian Statistics, J. Kadane (Ed.). North -Holland, Amsterdam.

Berger, J., (1985a). The Stein effect. In Encyclopedia of Statistical Sciences, S. Kotz and N. L. Johnson (Eds.). Wiley, New York.

Berger, J., (1985b). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Berger, J., and Berliner, L. M., (1984). Bayesian input in Stein-estimation and a new minimax empirical Bayes estimator. J. of Econometrics, 25, 87-108.

Chen, S. Y., (1983). Restricted risk Bayes estimation. Ph.D. Thesis, Department of Statistics, Purdue University, West Lafayette.

Dey, D. K., and Berger, J., (1983). On truncation of shrinkage estimators in simultaneous estimation of normal means. J. Amer. Statist. Assoc., 78, 865-869.

Efron, B., and Morris, C., (1971). Limiting the risk of Bayes and empirical Bayes estimators--Part I: the Bayes case. J. Amer. Statist. Assoc., 66, 807-815.

Efron, B., and Morris, C., (1973). Stein's estimation rule and its competitors--an empirical Bayes approach. J. Amer. Statist. Assoc., 68, 117-130.

James, W., and Stein, C., (1960). Estimation with quadratic loss. In Proc. Fourth Berkeley Symposium on Math. Statist. and Prob., 1, 361-379. University of California Press, Berkeley.

Marazzi, A., (1985). On constrained minimization of the Bayes risk for the linear model. Statistics and Decisions.

Stein, C., (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 197-206. University of California Press, Berkeley.

Stein, C., (1981). Estimation of the mean of a multivariate normal distribution. Ann. Statist., 9, 1135-1151.

Table 1. Values of M*.

ε \ p	$\Gamma_1, \underset{\sim}{\delta}^{M^*,I}$	$\Gamma_2, \underset{\sim}{\delta}^{M^*,I}$	$\Gamma_1, \underset{\sim}{\delta}^{M^*,S}$			$\Gamma_2, \underset{\sim}{\delta}^{M^*,S}$		
	all p	all p	3	5	7	3	5	7
0.0	∞	∞	∞	∞	∞	∞	∞	∞
0.1	1.30	2.38	0.17	0.002	0*	0.50	0.19	0.05
0.2	0.74	1.67	0.05	0*	0*	0.24	0.06	0*
0.3	0.47	1.30	0.02	0*	0*	0.14	0.02	0*
0.4	0.30	1.06	0.007	0*	0*	0.09	0.002	0*
0.5	0.19	0.88	0.001	0*	0*	0.05	0*	0*
0.6	0.11	0.74	0.001	0*	0*	0.03	0*	0*
0.7	0.06	0.63	0*	0*	0*	0.01	0*	0*
0.8	0.03	0.54	0*	0*	0*	0.006	0*	0*
0.9	0.006	0.47	0*	0*	0*	0.001	0*	0*
1.0	0	0	0	0	0	0	0	0

0* means smaller than 1×10^{-4}.

Table 2. Values of $r^*_{\Gamma_1}$ for $\underset{\sim}{\delta}^{M^*,I}$, $\underset{\sim}{\delta}^{M^*,S}$, and $\underset{\sim}{\delta}^{T}$.

	$r^*_{\Gamma_1}(\underset{\sim}{\delta}^{M^*,I})$	$r^*_{\Gamma_1}(\underset{\sim}{\delta}^{M^*,S})$			$r^*_{\Gamma_1}(\underset{\sim}{\delta}^{T})$	
ε \ p	all p	3	5	7	5	7
0.0	0	0	0	0	0.38	0.34
0.1	0.33	0.39	0.45	0.54	0.50	0.49
0.2	0.51	0.60	0.70	0.80	0.62	0.67
0.3	0.64	0.74	0.84	0.92	0.74	0.82
0.4	0.75	0.84	0.93	0.97	0.84	0.91
0.5	0.83	0.91	0.97	0.99	0.91	0.97
0.6	0.89	0.95	0.99	1.00	0.96	0.99
0.7	0.94	0.98	1.00	1.00	0.99	1.00
0.8	0.97	0.99	1.00	1.00	1.00	1.00
0.9	0.99	1.00	1.00	1.00	1.00	1.00
1.0	1.00	1.00	1.00	1.00	1.00	1.00

Table 3. Values of $r^*_{\Gamma_2}$ for $\underset{\sim}{\delta}^{M^*,I}$, $\underset{\sim}{\delta}^{M^*,S}$, and $\underset{\sim}{\delta}^{T}$.

	$r^*_{\Gamma_2}(\underset{\sim}{\delta}^{M^*,I})$			$r^*_{\Gamma_2}(\underset{\sim}{\delta}^{M^*,S})$			$r^*_{\Gamma_2}(\underset{\sim}{\delta}^{T})$	
ε \ p	3	5	7	3	5	7	5	7
0.0	0	0	0	0	0	0	0.38	0.34
0.1	0.15	0.10	0.08	0.19	0.14	0.12	0.40	0.35
0.2	0.25	0.17	0.14	0.31	0.25	0.22	0.42	0.37
0.3	0.33	0.23	0.18	0.42	0.35	0.32	0.44	0.38
0.4	0.40	0.28	0.22	0.52	0.44	0.42	0.46	0.40
0.5	0.46	0.33	0.26	0.61	0.54	0.51	0.48	0.41
0.6	0.51	0.37	0.30	0.70	0.63	0.61	0.50	0.42
0.7	0.56	0.41	0.33	0.78	0.72	0.71	0.52	0.44
0.8	0.61	0.45	0.36	0.85	0.81	0.81	0.54	0.45
0.9	0.65	0.48	0.39	0.93	0.91	0.90	0.56	0.47
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4. Values of $RSR(\pi^0, \underset{\sim}{\delta})$ for $\underset{\sim}{\delta}^{M^*,I}$, $\underset{\sim}{\delta}^{M^*,S}$, and $\underset{\sim}{\delta}^{M',I}$, when M* is optimal for Γ_1.

	$RSR(\pi^0, \underset{\sim}{\delta}^{M^*,I})$	$RSR(\pi^0, \underset{\sim}{\delta}^{M^*,S})$			$RSR(\pi^0, \underset{\sim}{\delta}^{M',I})$		
ε \ p	all p	3	5	7	3	5	7
0.0	0	0	0	0	0	0	0
0.1	0.11	0.10	0.07	0.03	0.03	0.02	0.01
0.2	0.20	0.11	0.07	0.03	0.08	0.04	0.03
0.3	0.29	0.21	0.07	0.03	0.10	0.08	0.06
0.4	0.38	0.24	0.07	0.03	0.18	0.13	0.11
0.5	0.47	0.27	0.07	0.03	0.26	0.20	0.19
0.6	0.57	0.28	0.07	0.03	0.35	0.30	0.29
0.7	0.66	0.29	0.07	0.03	0.46	0.42	0.42
0.8	0.77	0.29	0.07	0.03	0.60	0.58	0.58
0.9	0.88	0.29	0.07	0.03	0.77	1.00	0.77
1.0	1.00	0.29	0.07	0.03	1.00	1.00	1.00

Table 5. Values of $RSR(\pi^0, \underset{\sim}{\delta})$ for $\underset{\sim}{\delta}^{M^*,I}$, $\underset{\sim}{\delta}^{M^*,S}$, and $\underset{\sim}{\delta}^{M',S}$, when M* is optimal for Γ_2.

	$RSR(\pi^0, \underset{\sim}{\delta}^{M^*,I})$			$RSR(\pi^0, \underset{\sim}{\delta}^{M^*,S})$			$RSR(\pi^0, \underset{\sim}{\delta}^{M',I})$		
ε \ p	3	5	7	3	5	7	3	5	7
0.0	0	0	0	0	0	0	0	0	0
0.1	0.04	0.03	0.02	0.04	0.03	0.02	.009	.003	.001
0.2	0.08	0.05	0.04	0.08	0.05	0.03	.018	.005	.002
0.3	0.11	0.07	0.06	0.11	0.06	0.03	.025	.007	.002
0.4	0.14	0.09	0.08	0.14	0.07	0.03	.032	.008	.002
0.5	0.17	0.11	0.09	0.17	0.07	0.03	.038	.009	.003
0.6	0.20	0.13	0.10	0.20	0.07	0.03	.044	.010	.003
0.7	0.23	0.15	0.11	0.22	0.07	0.03	.049	.011	.003
0.8	0.26	0.17	0.12	0.25	0.07	0.03	.054	.011	.003
0.9	0.29	0.19	0.14	0.27	0.07	0.03	.059	.012	.003
1.0	1.00	1.00	1.00	0.29	0.07	0.03	1.00	1.00	1.00