POSTERIOR DISTRIBUTIONS FROM BOOTSTRAPPED LIKELIHOODS

by

Dennis D. Boos
John F. Monahan

Department of Statistics
North Carolina State University
Raleigh, North Carolina 27650
U.S.A.

December 1983

Abstract

Bayesian analysis is subject to the same kinds of misspecification problems which motivate the robustness and nonparametric literature. We present a method of incorporating prior information which performs well without direct knowledge of the error distribution. This is accomplished by replacing the likelihood in Bayes' formula with a bootstrap estimate of the sampling distribution of a robust estimator. The flexibility of the method is illustrated by examples, and its performance relative to standard Bayesian techniques is evaluated in a Monte Carlo study.

Key words: bootstrap, robust, Bayesian analysis, prior, information, kernel density estimation

An earlier version of this paper was presented at Symposium in Inference, Hitotubashi Univ., October 1982. The authors wish to thank Jay Kadane, Joseph Sedransk, Daniel Solomon, and Arnold Zellner for their helpful comments.

1. Introduction

Criticism of Bayesian analysis is often focused on the origin of the prior distribution. This largely philosophical debate overshadows a more critical use of prerogative: the choice of the error distribution. We propose a method of incorporating prior information about a structural parameter θ when the standard distributional assumptions, such as normality, are unwarranted. The procedure is founded upon an estimator $\hat{\theta}$ of the parameter of interest in a semi-parametric framework such as regression with unknown error distribution. Thus θ may be a set of regression coefficients, or, in the simple case, a center of symmetry. Additionally, θ may be a population parameter for which there is a natural estimator $\hat{\theta}$, such as the sample median for a population median. The viewpoint is Bayesian in that the prior and posterior beliefs are expressed in a probability distribution on θ. However, we do not use the likelihood in Bayes' formula

$$\pi(\theta|\text{data}) \propto \pi(\theta)L(\text{data}|\theta)$$

because it reflects too strongly the distributional assumptions inherent in its specification. Since we believe misspecification due to outliers or heavy-tailed distributions to be likely, we replace the likelihood with an estimate of the sampling distribution of $\hat{\theta}$ obtained from the bootstrap. By choosing a robust estimator $\hat{\theta}$, we add no complication to the bootstrap step and expect to obtain sound inference across a wide range of possible error distributions.

Our approach differs substantially from other approaches for robust Bayesian inference. From the subjectivist viewpoint, the possibility that the data may be more heavy-tailed than presumed or contain spurious observations would be expressed in the likelihood and prior. Box and Tiao (1973, Ch. 3) discuss enlarging the family of distributions with a hyperparameter to

accommodate the different shapes reflected in prior beliefs. Ramsay and Novick (1980) suggest using model densities that lead to more robust inference. Jeffreys (1967, pp. 211-212) suggests a nonparametric likelihood for inference on a population median θ,

$$\left(\frac{1}{2}\right)^n \binom{n}{n(1-F_n(\theta))} \approx \left(\frac{2}{\pi n}\right)^{\frac{1}{2}} \exp\{-2n[F_n(\theta) - \tfrac{1}{2}]^2\} ,$$

where F_n denotes the empirical distribution function. We compare this approach with ours in Example 1 in Section 2. Dempster (1975) surveys other subjectivist views of the robustness problem.

Our approach avoids the task of constructing prior distributions on hyper- or even nuisance parameters in which there is little information or interest. The distributional assumptions are quite modest. For example, we will focus on the regression model

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i \qquad i = 1,\cdots,n \qquad (1)$$

and require only that the errors e_i are independent, symmetric about zero (if b_0 is of interest), and have the same distribution. The parameter(s) θ of interest will be a subset of $\{b_0, b_1, \cdots, b_k\}$.

The critical step in constructing our estimated posterior $\hat{\pi}(\theta|\hat{\theta})$ is the estimation of the sampling density of $\hat{\theta}$ using the bootstrap (Efron, 1979). To implement the approach in the regression situation we first estimate the regression function and construct an estimate $\hat{F}_n$ of the distribution function of the errors. Thus $\hat{F}_n$ may be the empirical distribution function of the residuals $Y_i - \hat{Y}_i$ or some natural modification. Next, generate B random samples of size n from $\hat{F}_n$ and calculate $\hat{\theta}_j^*$ from sample j, $j = 1,\cdots,B$. From these values $\hat{\theta}_1^*,\cdots,e_B^*$, we estimate the sampling density of $\hat{\theta}|\theta$. When evaluated at the observed $\hat{\theta}$, this becomes our "likelihood" $\hat{L}_{nB}(\hat{\theta}|\theta)$ to be used in Bayes' formula

$$\hat{\pi}(\theta|\text{data}) \propto \pi(\theta)\hat{L}_{nB}(\hat{\theta}|\theta) \ .$$

The performance of our method parallels that of the estimator $\hat{\theta}$ in the usual tradeoffs between efficiency and robustness, when viewed in the classical estimation setting. The "posterior" $\hat{\pi}(\theta|\text{data})$ has greater dispersion (less information) than the one obtained from a correctly specified likelihood because we sample from $\hat{F}_n$ and not F and because $\hat{\theta}$ is neither sufficient for θ nor optimal. The latter loss is the "insurance premium" often mentioned in robustness. However, when the common distributional assumptions are unwarranted or violated, our approach still extracts sound and substantial information from the data.

Supporting these statements are the results of a Monte Carlo study of the location problem, which is discussed in Section 4. Further description and examples of the location model follow in Section 2. Analysis of regression applications is discussed in Section 3. Section 5 is a brief summary and the Appendix shows the asymptotic validity of the method.

2. The Location Problem

The location model is the simplest case of the regression model (1). Observed are iid random variables $Y_1, Y_2, \cdots, Y_n$ with density $f(y|\theta) = f_0(y-\theta)$, where f_0 is symmetric about zero $f_0(-y) = f_0(y)$ and θ is the (scalar) parameter of interest. Prior information on the parameter θ is expressed in a probability distribution with density $\pi(\theta)$.

The usual Bayesian analysis based on conjugate families of distributions assumes that the errors are normally distributed, $f_0(y) = \phi(y/\sigma)/\sigma$, where ϕ denotes the standard normal density. With the inclusion of the nuisance scale parameter σ, common practice (DeGroot, 1970, Sec. 9.6) is then to specify the prior distribution of θ given σ to be normally distributed with mean θ_0 and variance σ^2/τ_0. In addition, marginally σ^{-2} has an independent gamma prior distribution with shape ν and scale η. The resulting marginal posterior distribution of $(\theta|Y_1, Y_2, \cdots, Y_n)$ is Student's t with $2\nu + n$ degrees of freedom, centered at

$$(\tau_0\theta_0 + n\bar{Y})/(\tau_0 + n)$$

and with scale

$$[\eta + \tfrac{1}{2}\sum(Y_i - \bar{Y})^2 + \tau_0 n(\bar{Y} - \theta_0)^2/2(\tau_0 + n)]/(2\nu + n)(\tau_0 + n).$$

Since $\bar{Y}$ is sufficient for θ in the above normal model, we could have explained the "standard analysis" solely in terms of $\bar{Y}$ and its sampling density. This suggests an alternative approach based on estimators $\hat{\theta}$ and their sampling densities rather than on full likelihoods. The advantage here is that a wealth of robust estimators are available with the property that little information is lost compared to $\bar{Y}$ when f_0 is normal, but much

is gained when f_0 is heavier-tailed than normal. Typical examples include trimmed means, Huber M-estimators, and the Hodges-Lehmann estimator $\hat{\theta}_{HL} = \text{median } \{(Y_i + Y_j)/2,\ 1 \leq i \leq j \leq n\}$.

Under mild regularity conditions, these robust estimators are approximately normally distributed in large samples,

$$\sqrt{n}\ (\hat{\theta} - \theta) \xrightarrow{L} N(0,v^2)\ ,$$

where v^2 can be estimated. This asymptotic normal distribution can be used as a likelihood for Bayesian analysis (see Pratt, Raiffa, and Schlaifer, 1965, Sec. 19.4). That is, if the prior $\pi(\theta)$ is normal with mean θ_0 and variance σ_0^2, then the posterior of θ is approximately normal with mean

$$(\tau_n\hat{\theta} + \tau_0\theta_0)/(\tau_n + \tau_0)$$

and variance $(\tau_n + \tau_0)^{-1}$, where $\tau_n = n/v^2$ and $\tau_0 = 1/\sigma_0^2$. In large samples, this is a very natural way of doing robust Bayesian analysis. From a practical point of view, however, prior information and Bayesian analysis are most valuable in small samples. As a result, we propose to estimate the density of $\hat{\theta}|\theta$ using the bootstrap (Efron, 1979) instead of relying on the normal approximation.

Basically, the idea is to first estimate the distribution function of the data using the empirical distribution function $\hat{F}_n$ of the Y_i's. Next, generate B random samples, each of size n from $\hat{F}_n$ (now taken as fixed), and calculate $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$. From these B "observations," compute the kernel density estimator

$$\hat{g}_{nB}(u) = (Bh_B)^{-1} \sum_{j=1}^{B} k((u - (\hat{\theta}_j^* - \hat{\theta}))/h_B) \qquad (2)$$

as an approximation for the density of $\hat{\theta} - \theta$. Then $\hat{g}_{nB}$ is used in the same manner as the asymptotic normal distribution to get the approximate posterior of $\theta|\hat{\theta}$. That is, if we put $u = x - \theta$ in (2), then $\hat{g}_{nB}(x - \theta)$ is an estimate of the density of $\hat{\theta}|\theta$, and at $x = \hat{\theta}$ this function of θ is our bootstrapped likelihood

$$\hat{L}_{nB}(\hat{\theta}|\theta) = (Bh_B)^{-1} \sum_{j=1}^{B} k((2\hat{\theta} - \theta - \hat{\theta}_j^*)/h_B) \quad . \tag{3}$$

Our resulting posterior $\hat{\pi}(\theta|\hat{\theta})$ is proportional to $\pi(\theta)\hat{L}_{nB}(\hat{\theta}|\theta)$ and the normalizing constant is found by numerical integration. The kernel method is preferred over other nonparametric density estimates because it is conceptually simple and yields estimates that are nonnegative and integrate to one when a density is used as the kernel k. Alternatives to the empirical distribution function $\hat{F}_n$ have been considered, including smoothed and symmetrized versions. Further discussion follows in Section 4.

Our method is designed so that distributional assumptions can be as flexible as possible. Although the "iid" portion is sacred, the symmetry assumption is really only one way of clearly defining θ. We can replace symmetry by another assumption as long as θ is well defined and actually the parameter of interest. For example, instead of assuming that f_0 is symmetric about 0, we can assume that θ is the median of the Y population. Then the sample median is the natural estimator of θ. In general, the symmetry assumption can be replaced by the more nonparametric assumption that $\hat{\theta} \xrightarrow{P} \theta$ as long as θ so defined is intrinsically the important characteristic of the distribution of the Y_i's. For illustration consider the following example.

Example 1. A colleague, Dr. A. R. Manson, has studied county property tax assessments in North Carolina in order to determine cases of unfair taxation. The available data are all "arm's-length" transactions recorded in a county for a given year. The relevant variable is then $Y \equiv$ the ratio of the assessed value to the selling price for each piece of property sold during the year. Because of the skewed distribution of Y, the median θ of the population has been determined (in court) as the parameter that represents the "level of tax assessment." In the first stage of the analysis, a small sample of the arm's-length transactions is obtained. If an important difference is detected between the median level of tax assessment and the actual assessment of the client, all of the transactions (or a sample of 500) are obtained to construct a narrow confidence interval for the median. Our approach can be very useful in the first stage by incorporating previous years' studies of that county.

A 1977 study of Gaston County, North Carolina, estimated the median θ of the ratio with the sample median $\hat{\theta} = .556$ and a 95% nonparametric confidence interval of (.533, .578) (see David, 1981, p. 15). Incorporating this information with inflation, which affects selling price, we obtained a prior distribution $\pi(\theta)$ for θ in 1978 that is normal with mean $\theta_0 = .526$ and standard deviation $\sigma_0 = .024$. For reanalysis of Gaston County in 1978, if we were to have taken a first stage sample of $n = 50$ of transactions in 1978, we would combine these data with our prior $\pi(\theta)$ to obtain a posterior distribution for θ and decide whether the client may be unfairly taxed.

For illustration, let us analyze the sample of transactions from 1978 given in Table 1. The sample median is $\hat{\theta} = (.502 + .508)/2 = .505$ and a 95.1% nonparametric confidence interval is $(Y_{(18)}, Y_{(32)}) = (.459, .572)$.

Table 1. Ratio of Assessed Value to Sales Price for 50 Arms-Length Transactions in Gaston County, 1978.

.049	.270	.366	.434	.486	.508	.569	.587	.628	.840
.056	.316	.385	.439	.487	.518	.572	.589	.635	.892
.111	.319	.389	.459	.488	.526	.575	.599	.671	1.079
.247	.323	.393	.471	.494	.534	.581	.605	.692	1.100
.249	.329	.412	.473	.502	.538	.585	.627	.833	1.232

Figure 1 gives the prior $\pi(\theta)$ and three posteriors. The first two posteriors are from our analysis $\hat{\pi}(\theta|\hat{\theta})$ using both the empirical distribution function (edf) and a smoothed version generating the bootstrap samples. The posterior means and 95% probability regions are .515 and (.482, .544), and .514 and (.480, .548), respectively. Compared to the confidence interval (.459, .572), our 95% regions are only about half as long. The third posterior follows the analysis proposed by Jeffreys mentioned in the Introduction. Its posterior mean is .520 and 95% region is (.487, .562) and appears shifted to the right and more dispersed than ours.

The story concerning Gaston County does not stop in 1978. All properties were reassessed in 1979 to "full value." Again we wish to take an initial sample of n = 50 transactions in 1980 to decide whether a full study is worthwhile. Our prior information is based on inflation and also on experience which indicates that "full value" is really 85-90% assessment. Using this information, we arrived at a normal prior for θ in 1980 with mean $\theta_0 = .7945$ and standard deviation $\sigma_0 = .0285$. For illustration, we took 5 samples of size n = 50 (from a data base of 173 transactions) and report the posterior means and standard deviations and also the sample medians in Table 2. The first posterior, which bootstraps from the usual edf, appears to have slightly smaller dispersion but it is more variable. The mean of Jeffrey's posterior is smaller than the other two in 4 of the 5 samples. The similarity of the three posteriors confirms that our approach is giving reasonable results. Again for the record, the median of the full set of 173 transactions is $\hat{\theta} = .806$.

Figure 1

MEDIAN - GASTON COUNTY 1978

RATIO OF ASSESSED VALUE TO SALES PRICE

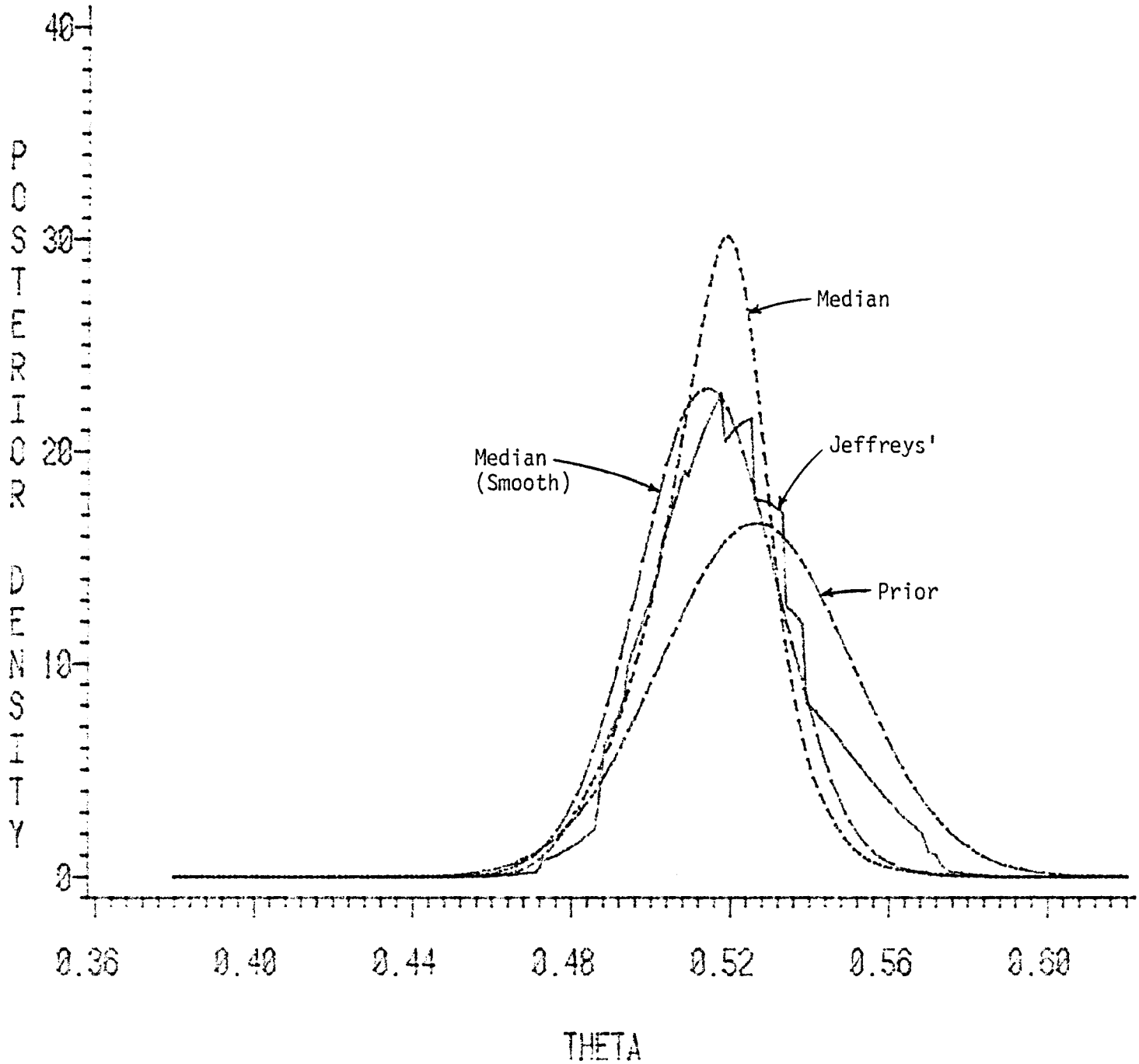


Table 2. Posterior Means and Standard Deviations from 5 Samples of n = 50 from Gaston County, 1980.

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Usual edf	.806	.018	.781	.020	.810	.016	.805	.016	.789	.026
Smoothed edf	.808	.019	.780	.020	.309	.021	.802	.022	.788	.022
Jeffreys'	.804	.022	.785	.021	.801	.018	.797	.022	.786	.022
Sample Median	$\hat{\theta}$ = .819		$\hat{\theta}$ = .769		$\hat{\theta}$ = .816		$\hat{\theta}$ = .808		$\hat{\theta}$ = .776	

3. Regression

In the regression case, we observe $Y = (Y_1, Y_2, \cdots, Y_n)^T$ arising from

$$Y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i \ ,$$

where the errors e_i are iid random variables centered at zero. Symmetry of the errors is only required for our approach if b_0 is a parameter of interest. The design matrix X has elements $(X)_{ij} = x_{ij}$ which are fixed and known, and we let $b = (b_0, b_1, \cdots, b_k)^T$. Let us first review a common Bayesian approach.

Standard Bayesian analysis (DeGroot, 1970, Sec. 11.10) based on conjugate distributions assumes that $e_i \sim N(0, \sigma^2)$ and that the prior would follow the normal-gamma structure. Given σ, the prior on b is $MVN(\beta, \sigma^2 \tau_0^{-1})$ and the marginal prior on σ^{-2} is gamma (ν, η). Here τ_0 is a $(k + 1)$ square positive definite matrix. The marginal posterior of $b|Y$ is then multivariate Student's t with $2\nu + n$ degrees of freedom centered at

$$\bar{b} = (X^T X + \tau_0)^{-1} (X^T Y + \tau_0 \beta)$$

with dispersion matrix

$$(X^T X + \tau_0)^{-1} \ 2\eta_1 / (2\nu + n) \ ,$$

where $\eta_1 = \eta + \frac{1}{2} Y^T (Y - X\bar{b}) + \frac{1}{2} (\beta - \bar{b})^T \tau_0 \beta$. To focus on a subset $\theta \subset b$, we partition b, $\bar{b}$, and $(X^T X + \tau_0)$ in the same way: $b = (\theta, \phi)^T$, $\bar{b} = (\bar{\theta}, \bar{\phi})^T$, and

$$(X^T X + \tau_0) = \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix} \ .$$

Then the marginal distribution of θ is multivariate Student's t with $2\nu + n$ degrees of freedom centered at $\bar{\theta}$ with dispersion matrix

$$\left(\frac{2n_1}{2\nu + n}\right) \cdot (\tau_{11} - \tau_{12}\tau_{22}^{-1}\tau_{21})^{-1} \quad .$$

Our approach in this regression situation is basically the same as for simple location. That is, we estimate the sampling density of $\hat{\theta}|\theta$ for a given estimation procedure and use that density evaluated at $\hat{\theta}$ as a likelihood for θ. The standard bootstrap method (c.f. Efron, 1979, Freedman, 1981) is to generate bootstrap errors $e^* = (e_1^*,\ldots,e_n^*)$ from an empirical distribution function of the residuals $\hat{e} = Y - X\hat{b}$. Then calculate $\hat{b}^*$ from $Y^* = X\hat{b} + e^*$. The process is repeated B times and these realizations of $\hat{\theta}^* \subset \hat{b}^*$ are used to estimate the sampling density of $\hat{\theta} - \theta$ and $\hat{L}(\hat{\theta}|\theta)$. In the following examples we use M-estimation with the Huber ψ function with $k = 1.0$ and Proposal 2 for scaling (see Huber, 1981, Ch. 7).

In small samples it makes sense to adjust the residuals before forming the empirical distribution function in order to account for the fitting process. Two obvious ways are to replace $\hat{e}_i$ by $[n/(n - k - 1)]^{\frac{1}{2}}\hat{e}_i$ or $\hat{e}_i/(1 - h_{ii})^{\frac{1}{2}}$, where h_{ii} is the ith diagonal element of the "hat" matrix, $H = X(X^TX)^{-1}X^T$. We prefer the latter adjusted residuals since all the residuals then have the same variance when using least squares and approximately the same variance when using M-estimation.

Example 2. Randolph, et al. (1977) studied reproduction in cotton rats and presented a scatterplot of total litter weight gain during the first 12 days after birth versus the number in each litter. We extracted the data in Table 3 from their scatterplot. One rat gave birth to two heavier-than-average litters which subsequently were the largest Y's in Table 3, 68.4 and 77.1. Although this rat appeared average in every other respect, it was

Table 3. Litter Size and Weight Gain of Cotton Rats. From Randolph, et al. (1977).

Litter Size (X)	Weight Gain (grams) (Y)
2	31.1
3	36.9
4	41.6
4	46.1
4	48.4
4	48.4
5	30.1
5	44.4
5	46.8
5	54.0
6	48.9
6	50.1
6	51.5
6	56.2
6	68.4
7	77.1

clear that her litters have a large effect on the least squares fits. Thus the researchers decided to report least squares regressions with and without those two points. Our reproduced data yield least squares lines $y = 25.6 + 4.2x$ with the last two pairs deleted and $y = 15.6 + 6.8x$ for all 16 pairs. M-estimation using a Huber ψ function with $k = 1.0$ and Proposal 2 for scale yields $y = 20.4 + 5.8x$ based on all 16 points. Using the results of a previous study by Kilgore (1970), we constructed a joint prior distribution on (b_0, b_1, σ^2) of the normal-gamma form with $\beta = (20.4, 8.01)^T$, scale matrix

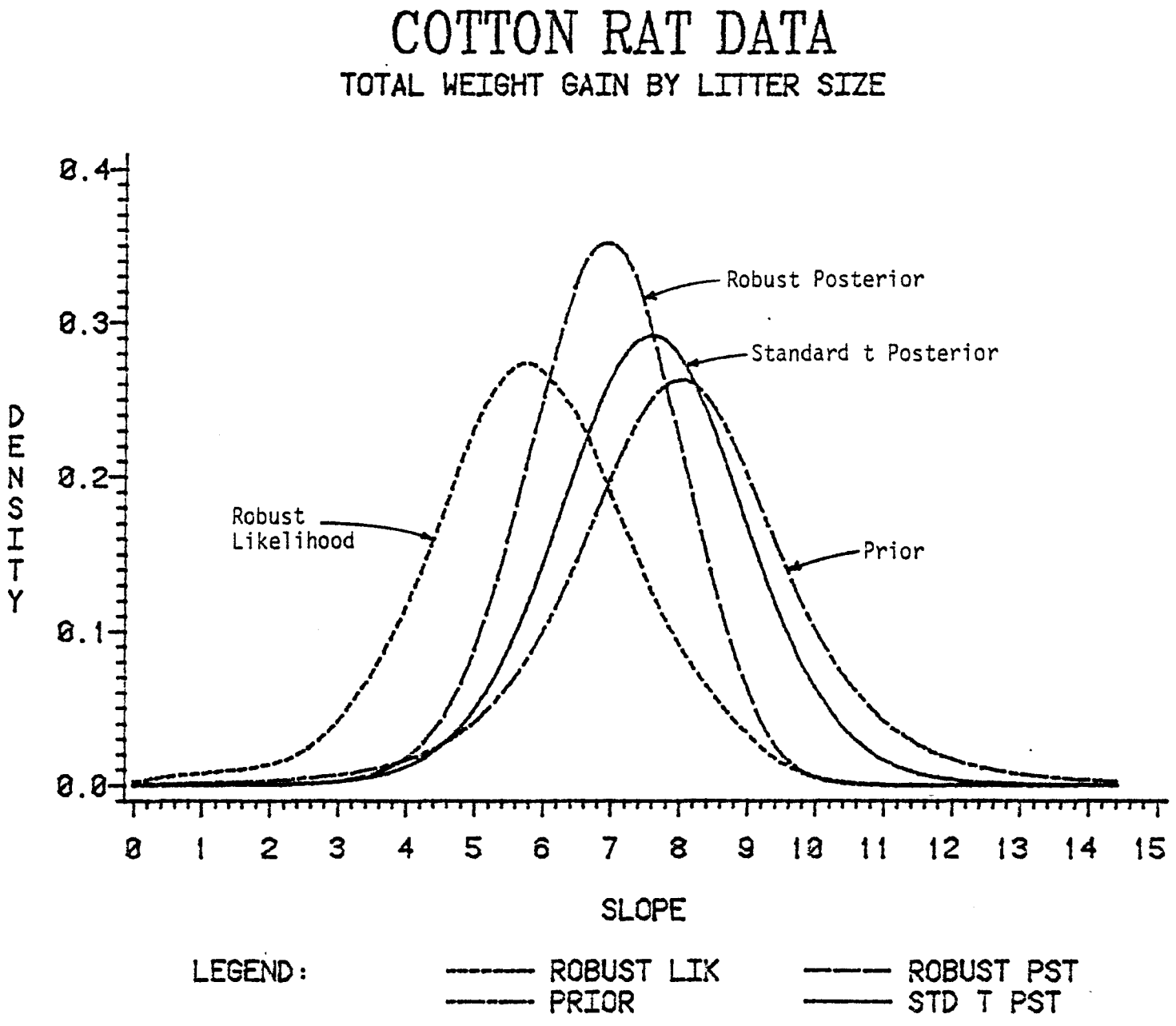
$$\tau_0^{-1} = \begin{pmatrix} 3.39 & -.56 \\ -.56 & .10 \end{pmatrix},$$

and $\nu = 5/2$ and $n = 51.3$. In Figure 2 are given the marginal prior density of the slope parameter b_1 and the marginal posterior of b_1 from the standard Bayesian analysis of the 16 observations. Also pictured are the estimated likelihood $\hat{L}(b_1|\text{data})$ using M-estimation and $B = 1000$ bootstrap samples and the resultant estimated posterior $\hat{\pi}(b_1|\text{data})$. Our approach has allowed us to incorporate valuable prior information in a situation where outliers and nonnormality were suspected.

Example 3. Ramsay and Novick (1980) give values of three psychological variables measured on 29 children. The variables are $DL = 100 \sin^{-1}\sqrt{p}$, where p is the proportion correct on a dichotic listening task; VI = a test of verbal intelligence; and PI = a test of performance intelligence. The dependent variable Y is DL and b_1 and b_2 are regression coefficients for VI and PI, respectively. Ramsay and Novick give for the normal-gamma prior means $\beta = (43, .31, .31)^T$, scale matrix

$$\tau_0^{-1} = \begin{pmatrix} 69.44444 & -.20833 & -.20833 \\ -.20833 & .00694 & -.00556 \\ -.20833 & -.00556 & .00694 \end{pmatrix},$$

Figure 2

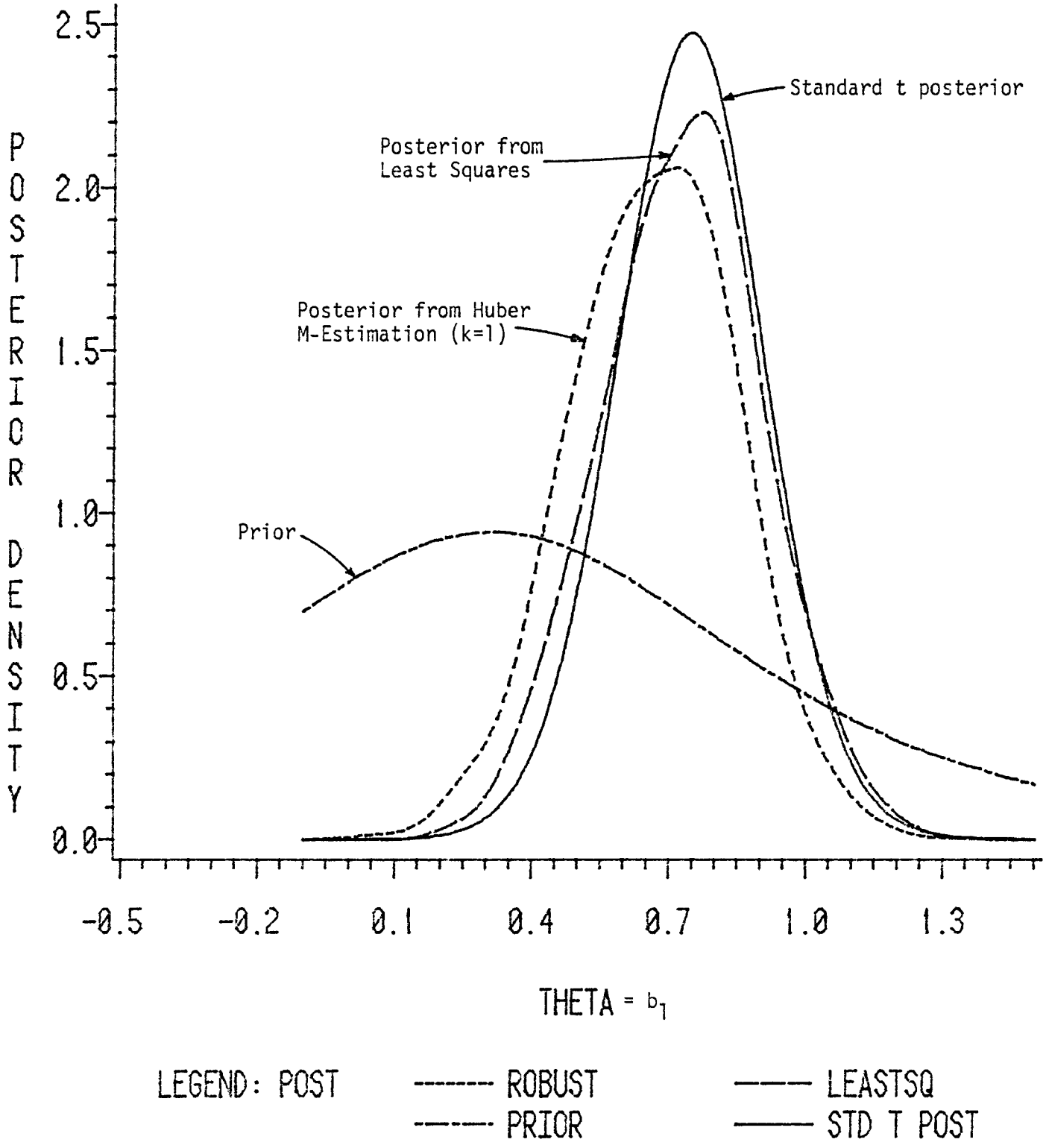

COTTON RAT DATA
TOTAL WEIGHT GAIN BY LITTER SIZE
DENSITY
0.4
0.3
0.2
0.1
0.0
Robust Posterior
Standard t Posterior
Robust Likelihood
Prior
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SLOPE
LEGEND:
ROBUST LIK
PRIOR
ROBUST PST
STD T PST

and $\nu = 1.5$ and $\eta = 75$ for the parameters of the gamma density of σ^{-2}. We focus on b_1, the coefficient of verbal intelligence (VI). Figure 3 shows the prior, our estimated posterior using least squares and Huber's M-estimation as in the previous example, and the t posterior resulting from the usual conjugate analysis.

Figure 3

RAMSAY-NOVICK REGRESSION PROBLEM

VERBAL INTELLIGENCE COEFFICIENT



4. Performance in the Location Problem

So far in this paper we have attempted to show that our approach is a natural and practical way of doing robust Bayesian inference in small samples. In this section, the *performance* of our estimated posteriors in the simple location problem is evaluated by Monte Carlo methods. We hope to show that when the data is normally distributed our method works nearly as well as the standard Bayesian analysis. Moreover, when the data comes from a heavier-tailed distribution, our method does much better than the standard (normal) analysis.

For simplicity, we fix the variance $\sigma^2 = 1$. If the data Y consists of n independent normal $(\theta,1)$ random variables and the prior $\pi(\theta)$ is normal (θ_0,τ_0^{-1}), then the posterior is normal with mean $(\tau_0\theta_0 + n\bar{Y})/(\tau_0 + n)$ and variance $(\tau_0 + n)^{-1}$. This posterior can be obtained from the Section 2 results by letting $\nu = \eta \to \infty$. For comparison we implement our approach for 5 estimators:

1) $\bar{Y}$, the sample mean

2) 10% trimmed mean (mean of middle 80% of ordered data)

3) 20% trimmed mean

4) the sample median

5) the Hodges-Lehmann estimator, HL $\equiv$ median $\{(Y_i + Y_j)/2,\ 1 \leq i \leq j \leq n\}$.

These estimators provide a range of performance in efficiency and robustness.

When the errors are normal, our $\hat{L}(\hat{\theta}|\theta)$ has less information than the full likelihood $L(Y|\theta)$. This loss of information has two components:

A) The loss due to reducing to $\hat{\theta}$, an insufficient statistic.

B) The loss due to estimating the density of $\hat{\theta}|\theta$.

For evaluating these components, it is useful to introduce a third "likelihood,"

$L(\hat{\theta}|\theta)$, which is the sampling density of the estimator evaluated at the observed value $\hat{\theta}$. We then wish to compare the following three posteriors

$$\pi(\theta|Y) \propto \pi(\theta)L(Y|\theta)$$

$$\pi(\theta|\hat{\theta}) \propto \pi(\theta)L(\hat{\theta}|\theta)$$

$$\hat{\pi}(\theta|\hat{\theta}) \propto \pi(\theta)\hat{L}(\hat{\theta}|\theta) \ .$$

Thus, (A) represents the comparison between $\pi(\theta|Y)$ and $\pi(\theta|\hat{\theta})$ and (B) represents the comparison between $\pi(\theta|\hat{\theta})$ and $\hat{\pi}(\theta|\hat{\theta})$.

To evaluate (A) we have computed the expected information (Lindley, 1956) in the posteriors $\pi(\theta|\hat{\theta})$ for the five estimators ($\pi(\theta|\bar{Y}) = \pi(\theta|Y)$ since $\bar{Y}$ is sufficient) and normally distributed priors $\pi(\theta)$. The results reported in Boos and Monahan (1982) can be easily summarized: the loss of information is comparable to $\tau_0 + 1/\mathrm{Var}\hat{\theta}$ versus $\tau_0 + n$. This concise summary can be attributed to the correspondence in the location case between information and dispersion and to the near normality of the sampling distributions of $\hat{\theta}$ (Monahan, 1982). Thus, the loss in (A) can be made appropriately small by choosing $\hat{\theta}$ to have a suitable variance at the normal.

Monte Carlo methods were employed to evaluate (B), the loss due to estimating the density of $\hat{\theta}|\theta$. We measure this loss by comparing the mean squared errors (MSE) of the posterior means from $\hat{\pi}(\theta|\hat{\theta})$ and $\pi(\theta|\hat{\theta})$. The Monte Carlo loop repeats for N = 200 samples, $Y_1,\ldots,Y_n$, iid from N(0,1) for sample sizes n = 5, 10, 20, 40. For each sample, $\hat{F}_n$ is computed and B bootstrap resamples $Y_1^*,\ldots,Y_n^*$, are generated from $\hat{F}_n$. Two different versions of $\hat{F}_n$ were studied:

1) F_n, the usual empirical distribution function. Note that F_n has mean $\bar{Y} = \int y dF_n(y)$ and variance $n^{-1}\sum(Y_i - \bar{Y})^2$. An iid sample from F_n may be drawn by selecting *with replacement* from the set $\{Y_1,\cdots,Y_n\}$.

2) F_{ns}, a smoothed version of F_n · F_{ns} is actually a kernel estimate with the bandwidth parameter h chosen so that the variance is $(n - 1)^{-1}\sum(Y_i - \bar{Y})^2$. F_{ns} also has mean $\bar{Y}$. An iid sample from F_{ns} consists simply of $Y_1^* + hU_i,\cdots,Y_n^* + hU_n$ where $Y_1^*,\cdots,Y_n^*$ is an iid sample from F_n and $U_1,\cdots,U_n$ is an iid sample from the kernel density which defines F_{ns}.

The kernel density used for F_{ns} as well as for $\hat{g}_{nB}$ in (2) is the density of the sum of three independent uniform deviates. It is smooth, easy to compute, and has finite support. The smoothing parameter for $\hat{g}_{nB}$, $h_B = 1.059B^{-1/5}$ is optimal in terms of mean integrated squared error for this kernel when estimating the normal distribution. Unless otherwise noted, the bootstrap replication size is B = 400. Note that taking $B \to \infty$ would eliminate one source of error but add considerable computing cost.

Table 4 gives ratios of MSE's (over Monte Carlo replications) of the posterior means (over θ). The entries are the estimated MSE's from our $\hat{\pi}(\theta|\hat{\theta})$ divided by the estimated MSE's from $\pi(\theta|\hat{\theta})$. Recall that $\pi(\theta|\hat{\theta})$ is obtained from the true sampling density of $\hat{\theta}$. The columns labled "F_n" refer to bootstrapping from F_n and those labeled "F_{ns}" refer to bootstrapping from F_{ns}. Note that by n = 20 the estimated posterior $\hat{\pi}(\theta|\hat{\theta})$ is comparable to $\pi(\theta|\hat{\theta})$ for all estimators except the median. Smoothing of the errors (using F_{ns}) is helpful in small samples. Other prior distributions and different bootstrap replication sizes (B = 100, B = 1000) gave similar results.

Table 4. Ratio of Estimated MSE of Posterior Means: $\hat{\pi}(\theta|\hat{\theta})$ vs. $\pi(\theta|\hat{\theta})$.

Prior Mean	Prior Variance		n=5 F_n	n=5 F_{ns}	n=10 F_n	n=10 F_{ns}	n=20 F_n	n=20 F_{ns}	n=40 F_n	n=40 F_{ns}
0.0	0.1	$\bar{Y}$	2.02	1.70	1.18	1.08	1.00	.98	.97	.97
		10% Trim			1.14	1.19	.96	1.00	.96	.97
		20% Trim	2.26	1.98	1.18	1.16	.98	1.01	.96	.98
		Median	3.62	2.70	1.67	1.58	1.25	1.30	1.09	1.14
		HL	2.29	1.94	1.17	1.15	.99	1.02	.96	.97
0.0	0.2	$\bar{Y}$	1.50	1.32	1.08	1.01	.99	.98	.98	1.00
		10% Trim			1.05	1.05	.97	1.00	.98	.99
		20% Trim	1.56	1.50	1.06	1.07	.98	1.00	.97	.99
		Median	2.23	1.99	1.28	1.34	1.12	1.22	1.07	1.14
		HL	1.64	1.49	1.06	1.07	.98	1.01	.98	.99

Note: Monte Carlo replications N = 200. Bootstrap replications B = 400. Data and prior are normally distributed. F_n and F_{ns} refer to the type of bootstrap used.

Up to this point, we have seen the costs of our approach under normality. Now we would like to compare our estimated posterior $\hat{\pi}(\theta|\hat{\theta})$ to the "reported" posterior $\pi(\theta|Y)$ (still based on the assumption of normal errors) when in fact the data are not normal. We consider three alternative error distributions: uniform, Laplace and t_3 (Student's t distribution with 3 degrees of freedom), each scaled to have unit variance. Table 5 lists the ratio of the estimated MSE of the posterior mean (over θ) of $\hat{\pi}(\theta|\hat{\theta})$ to that of the "reported" posterior $\pi(\theta|Y)$ for n = 20 and for all 5 estimators. As expected, the robust estimators give a substantial improvement at the Laplace and t_3 distributions. The optimality of the median for the Laplace is demonstrated here by a 1/3 reduction in MSE. At t_3 the 20% trimmed mean and HL are best, reducing the MSE of the posterior mean by roughly 60%. It is interesting that our approach with $\bar{Y}$ does much better at t_3 than the normal theory posterior when the prior is fairly peaked. As expected, when the prior is diffuse, $\bar{Y}$ and the normal theory posterior perform about the same. If short tails are a possibility, the superiority of the reported posterior at the uniform suggests the use of an adaptive estimator which performs well over a wide range of distributions including both short and long tails. Note that with our approach, using a data based adaptive procedure changes neither the method nor the analysis.

Finally, we are concerned about the ability of our technique to *correctly* express information in the posterior. In the sampling theoretic framework, one often checks whether confidence intervals have the correct coverage probabilites. In the Bayesian situation, we can count the proportion of times that a posterior high probability region contains the true θ for repeated realizations from $\pi(\theta)$. The "Hits" column of Table 6 gives these counts for nominal 90% posterior

Table 5. Ratio of Estimated MSE of Posterior Means: $\hat{\pi}(\theta|\hat{\theta})$ vs. Normal Theory Posterior at n = 20.

Prior Mean	Prior Variance		Model Distributions: Uniform F_n	Uniform F_{ns}	Normal F_n	Normal F_{ns}	Laplace F_n	Laplace F_{ns}	t_3 F_n	t_3 F_{ns}
0.0	0.1	$\bar{Y}$	1.05	1.01	1.00	.98	.91	.88	.64	.63
		10% Trim	1.15	1.20	.96	1.00	.84	.84	.47	.45
		20% Trim	1.25	1.31	.97	1.00	.76	.74	.47	.45
		Median	1.81	1.86	1.25	1.29	.68	.67	.53	.57
		HL	1.04	1.05	.95	.98	.87	.85	.45	.43
0.0	0.2	$\bar{Y}$	1.02	1.00	.99	.98	.93	.91	.64	.64
		10% Trim	1.20	1.24	.98	1.01	.84	.84	.44	.43
		20% Trim	1.34	1.41	1.00	1.03	.73	.73	.38	.37
		Median	1.85	1.98	1.25	1.36	.64	.65	.49	.54
		HL	1.07	1.09	.96	.99	.84	.83	.41	.39
0.0	∞	$\bar{Y}$	1.01	.99	1.00	1.01	1.00	1.00	.97	1.00
		10% Trim	1.34	1.38	1.05	1.07	.87	.88	.41	.41
		20% Trim	1.66	1.72	1.11	1.14	.74	.74	.37	.37
		Median	2.88	3.04	1.67	1.78	.66	.69	.49	.53
		HL	1.18	1.19	1.03	1.05	.84	.84	.37	.36

Note: Monte Carlo replications N = 200. Bootstrap replications B = 400. Priors are normally distributed. F_n and F_{ns} refer to the type of bootstrap used.

regions in 200 replications. Each replication consists of drawing a random θ' from $\pi(\theta)$, then drawing a sample $Y_1,\cdots,Y_n$ from the particular error distribution centered at θ', constructing the posterior region, and checking to see if θ' is in the region. In Table 6 we see that except for the median, the counts hover nicely around the expected 180 counts. The median's performance is improved when using the smoothed edf F_{ns} (not shown).

A more comprehensive check on the validity of the posteriors can be obtained by computing $\alpha(\theta') = \int_{-\infty}^{\theta'}\hat{\pi}(\theta|\hat{\theta})d\theta$, the value of the posterior cumulative at θ'. A simple argument shows that unconditionally $\alpha(\theta')$ should be uniformly distributed on (0,1). Deviations from uniformity would indicate that posterior probability regions have the wrong coverage probabilities. Table 6 gives the values of two goodness-of-fit statistics based on the 200 values of $\alpha(\theta')$. The first is the well-known Kolmogorov-Smirnov (KS) statistic. It has little power in this situation because of the symmetry of the data and prior. The second is the sum of the second and fourth components of the Neyman-Barton (NB) test statistics (see Quesenberry and Miller, 1979). It is specifically designed to pick up scale and kurtosis alternatives and has a null distribution which is approximately chi-squared with two degrees of freedom. It strongly rejects the median in all the tabled situations and also flags $\bar{Y}$ at the Laplace with the peaked prior and the normal theory posterior at t_3 with the less informative prior.

In general Table 6 supports the claim that our posteriors are giving honest assessments. It also shows that the normal theory posterior is reasonably honest in its assessments. Of course, at long-tailed distributions the robust estimator posterior will be better than the normal theory posterior in terms of the *length* of probability regions just as they are in Table 5 in terms of point estimation MSE.

Table 6. Tests of Uniformity and Coverage at n = 20

Prior Mean	Prior Variance		Normal KS	Normal NB	Normal Hits	Laplace KS	Laplace NB	Laplace Hits	t_3 KS	t_3 NB	t_3 Hits
0.0	0.1	$\bar{Y}$	.044	.20	179	.087	5.20	176	.041	.62	178
		10% Trim	.064	.37	179	.060	1.74	182	.040	.04	181
		20% Trim	.063	1.82	175	.060	3.44	184	.040	3.61	173
		Median	.075	35.00	158	.096	23.51	168	.107	42.41	156
		HL	.079	1.06	177	.062	4.46	182	.040	.91	174
		Normal	.034	.61	186	.078	2.63	181	.039	.10	181
0.0	1.0	$\bar{Y}$	.041	.75	180	.065	.61	175	.044	.01	180
		10% Trim	.051	1.04	183	.076	1.80	182	.059	.88	182
		20% Trim	.057	3.81	173	.082	.51	186	.047	.81	175
		Median	.076	49.67	158	.095	7.95	168	.092	41.24	157
		HL	.036	1.36	179	.085	.54	183	.053	1.55	177
		Normal	.039	1.04	186	.065	.11	180	.082	6.16	182

Monte Carlo replications N = 200. Bootstrap replications from the usual edf with B = 400. Critical values for Kolmogorov-Smirnov (KS) are .086 and .096 for levels .10 and .05. For the Neyman-Barton (NB) test, the values are 4.61 and 5.99. Large values are underlined. "Hits" are the counts for nominal 90% probability regions.

5. Summary

In this paper we have presented a robust method of Bayesian inference: Bayesian, because we use prior information and treat unknown parameters as random variables; robust, because our method performs well over a wide range of error distributions. We have demonstrated its usefulness in simple location problems and in regression. If robust estimators are employed, our approach leads to decisions that lose only a little to the optimal when the error distribution is normal and gain over standard methods when the errors are not. In most cases, the posteriors we compute give an honest appraisal of the information about a parameter.

Appendix

The focus of the paper is on small samples, but it is useful to know that the method is asymptotically valid. Here we want to show that the asymptotic distribution of $\theta|\hat{\theta}$ from our estimated posterior is an appropriate normal distribution. Consider estimators θ such that

$$L_n(t) = P(n^{\frac{1}{2}}(\hat{\theta}-\theta) \leq t|\theta) \to L(t),$$

where $L(t)$ is a normal distribution function centered at zero with density $\ell(t)$. For simplicity we consider the location problem of Section 2 where our estimated posterior is

$$\hat{\pi}(\theta|\hat{\theta}) = \pi(\theta)\hat{L}_{nB}(\hat{\theta}|\theta)/\hat{S}_n$$

$$= \pi(\theta)\hat{g}_{nB}(\hat{\theta}-\theta)/\hat{S}_n ,$$

with $\hat{S}_n = \int\pi(\theta)\hat{g}_{nB}(\hat{\theta}-\theta)d\theta$. Recall that $\hat{g}_{nB}(x)$ is an estimator of the density of $\hat{\theta}-\theta$. In order to obtain an asymptotic distribution for $\theta|\hat{\theta}$ we need to look at the standardized random variable $T|\hat{\theta} = n^{\frac{1}{2}}(\theta-\hat{\theta})|\hat{\theta}$ which has density

$$\hat{f}_n(t)=\hat{\pi}(n^{-\frac{1}{2}}t + \hat{\theta}|\hat{\theta}) = \pi(n^{-\frac{1}{2}}t + \hat{\theta})n^{-\frac{1}{2}}\hat{g}_{nB}(-n^{-\frac{1}{2}}t)/\hat{S}_n .$$

The following theorem says that the limiting distribution of $n^{\frac{1}{2}}(\theta-\hat{\theta})|\hat{\theta}$ is the same as that of $n^{\frac{1}{2}}(\hat{\theta}-\theta)|\theta$.

THEOREM. Suppose that $\pi(\theta)$ is a bounded and continuous prior density on R_1. Let $\hat{\theta} \xrightarrow{wp1} \theta_1$. If $n^{-\frac{1}{2}}\hat{g}_{nB}(n^{-\frac{1}{2}}x) \xrightarrow{wp1} \ell(x)$ at each x, then

$$\hat{f}_n(t) \xrightarrow{wp1} \ell(-t) = \ell(t) \text{ at each } t.$$

The "wpl" convergences can be weakened to "in probability" convergences and θ can be generalized to belong to R^k. Likewise, the pointwise convergences can be changed to $L_1^{(r)}$ convergence meaning $\int t|^r|f-g|dt \to 0$. A general proof of the theorem is given by Theorem 8 of Boos (1983). The form of $\hat{f}_n(t)$ makes the proof fairly transparent since the numerator obviously converges to $\pi(\theta_1)\ell(-t)$, and a little work gives the convergence of $\hat{S}_n$ to $\pi(\theta_1)$. The crucial assumption of the theorem is the convergence of $n^{-\frac{1}{2}}\hat{g}_{nB}(n^{-\frac{1}{2}}x)$, our bootstrap estimate of the density of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$. Conditions for this latter convergence are given by Theorem 7 of Boos (1983) and further discussion in Section 3 of that paper.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R. Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press.

Boos, D. D. (1983), "On Convergence of Densities of Translation and Scale Statistics," Institute of Statistics Mimeo Series #1625, Raleigh, North Carolina.

Boos, D. D., and Monahan, J. F. (1983), "The Bootstrap for Robust Bayesian Analysis: An Adventure in Computing," in *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, eds. Heiner, K.W., Sacher, R. S., and Wilkinson, J. W., 101-107.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, Mass.: Addison-Wesley.

David, H. A., *Order Statistics*, 2nd Edition, New York: Wiley.

DeGroot, M. H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

Dempster, A. P. (1975), "A Subjectivist Look at Robustness," *Proceedings of the 40th Session of the International Statistical Institute*, 1, 349-374.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1-26.

Freedman, D. A. (1981), "Bootstrapping Regression Models," *Annals of Statistics*, 9, 1218-1228.

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

Jeffreys, H. (1967), *Theory of Probability*, Third Edition, Oxford University Press.

Kilgore, D.L. (1970), "The Effects of Northward Dispersal on Growth Rate of Young at Birth and Litter Size in *Sigmodon Hispidus*," *American Midland Naturalist*, 84, 510-520.

Lindley, D. V. (1956), "On a Measure of the Information Provided by an Experiment," *Annals of Mathematical Statistics*, 27, 986-1005.

Monahan, J. F. (1983), "The Sampling Distribution of Some Robust Estimates," *Proceedings of the Pacific Area Statistical Conference* (to appear).

Pratt, J. W., Raiffa, H., and Schlaifer, R. (1965), *Introduction to Statistical Decision Theory*, New York: McGraw-Hill.

Quesenberry, C. P., and Miller, F. L., Jr. (197), "Power Studies of Some Tests for Uniformity, II," *Communications in Statistics B, Simulation and Computation*, 8, 271-290.

Ramsay, J. O, and Novick, M. R. (1980), "PLU Robust Bayesian Decision Theory: Point Estimation," *Journal of the American Statistical Association*, 75, 901-907.

Randolph, P. A., Randolph, J. C., Mattingly, K., and Foster, M. M. (1977), "Energy Costs of Reproduction in the Cotton Rat, *Sigmodon Hispious*," *Ecology*, 58, 31-45.