SAMPLING-BASED CALCULATIONS FOR LIKELIHOOD METHODS

by

George Casella and Martin T. Wells
Cornell University

Martin A. Tanner
University of Rochester

BU-1274-M December 1994

SAMPLING-BASED CALCULATIONS FOR LIKELIHOOD METHODS

George Casella[1] and Martin T. Wells
Cornell University

Martin A. Tanner
University of Rochester

December 1994

In this paper we discuss and illustrate sampling based computations that are useful for likelihood methods. We make use of convenient representations of various conditional and unconditional distributions for the sampling based computations. Once the distributional relations are established we use the Gibbs sampler and other resampling schemes to simulate the distributions of various likelihood quantities. We apply the method to Barndorff-Nielsen's formula for the distribution of maximum likelihood estimates, non-normal regression, and to *MANOVA*.

1. Introduction

In this article we address the relationships between resampling based methods and their use in modern likelihood theory. We will consider several problems with the format of the article being that of a case study.

One of the main limitations of classical parametric likelihood theory is its dependence on the large sample sizes and the normal distribution for calculating the distribution theory for likelihood estimators. There has been much progress in recent years in extending the classical $\sqrt{n}$ theory to higher order approximations. Although the better approximations promise *small sample* distribution theory, they may be difficult to compute in practice.

At the same time as the developments in likelihood theory were taking place, there were great advances in Bayesian computations. Aside from the subjective dogma inherent in the Bayesian paradigm, a major criticism had been the lack of methods for the complete implementation of a Bayesian solution. The new methods of sampling based calculations

[1] Research supported by National Science Foundation Grant No. DMS9100839 and National Security Agency Grant No. 90F-073.

This is technical report BU-1274-M in the Biometrics Unit, Cornell University

have gone a long way in remedying this problem, allowing exact solutions in difficult problems.

In this article we apply the sampling based calculations, which the Bayesian have developed, to some likelihood problems. It is known that there are many instances where the Bayesian and likelihood answers agree. This usually happens when one assumes a *flat* prior distribution for the parameter of interest, a phenomenon which often has a group theoretic explanation (Casella and Wells, 1992). However here we are not taking the approach of solving likelihood problems via Bayesian methods, but rather we study the likelihood quantities on their own merit without any *lurking* Bayesian theory. Note that we are interested in Monte Carlo procedures for the assessment of the properties of likelihood estimates. This goal is different than that in Wei and Tanner (1990), Gelfand and Carlin (1991) or Carlin and Gelfand (1992), where the objective was to use Monte Carlo methods in the construction of likelihood estimates in non-standard problems.

In the next section we review some of the current methods of resampling based on distribution theory. In Section 3.1 we study the location-scale problem, while in Section 3.2 we look into the non-normal multiple regression problem, a problem that was initially studied by Fraser, Lee and Ried (1990). In these examples we study the sampling based computations of marginal distributions and p- values. In Section 3.3 we apply the resampling based methodology to the MANOVA problem, and finally, Section 4 contains a short discussion.

2. *Sampling Based Calculations.*

2.1 Gibbs Sampling

The *Gibbs Sampler* is one of the more popular Markov Chain simulation methods for constructing random variables from a specified distribution. The history and recent developments are reviewed by Casella and George (1992). To use the Gibbs sampler one must be able to generate random samples from the complete set of conditional distributions. Given this complete set of conditionals, the Gibbs Sampler is one of the simpler Markov Chain simulation methods available for generating samples from a joint distribution. To generate a sample from a distribution with density $f(\boldsymbol{x}) = f(x_1, \ldots, x_k)$ the Gibbs sampler begins with starting values $\left(x_1^{(0)}, \ldots, x_k^{(0)}\right)$, then iterates through the following loop:

1.) Sample $x_1^{(i+1)}$ from $f\left(x_1 | x_2^{(i)}, \ldots, x_k^{(i)}\right)$
2.) Sample $x_2^{(i+1)}$ from $f\left(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_k^{(i)}\right)$
$\vdots$
k.) Sample $x_k^{(i+1)}$ from $f\left(x_k | x_1^{(i+1)}, \ldots, x_{k-1}^{(i+1)}\right)$.

Under certain regularity conditions it can be shown that as the number of iterations tend to infinity the distribution of the generated sample $\left(x_1^{(i+1)}, \ldots, x_k^{(i+1)}\right)$ converges to a sample from $f(\boldsymbol{x})$ at a geometric rate. See Casella and George (1992), Tanner (1991), Tierney (1991) and Smith and Roberts (1992) for further details and a complete set of many references on the literature for the Gibbs sampler.

2.2 The Weighted Bootstrap

The implementation of the Gibbs Sampler requires that one must be able to generate random samples from the complete set of conditional distributions. However, these conditionals may not always be available. Hence we need to have generation methods that are not dependent on knowledge of the full set of conditionals. The *weighted bootstrap*, a simple modification of importance sampling, gives such a method. We may approximately resample from $f(\boldsymbol{x}) = h(\boldsymbol{x})/\int h(\boldsymbol{x})\, d\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \ldots, x_k)$ as follows. Suppose that we have a density g in hand which resembles f and is easy to sample from. Given $\boldsymbol{x}_i$, $i = 1$,

..., m, a sample from g, calculate $\omega_i = h(x_i)/g(x_i)$ and then $q_i = \omega_i/\sum_{j=1}^{m} \omega_j$. Draw $\boldsymbol{x}^*$, from the discrete $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ placing mass q_i on $\boldsymbol{x}_i$. Then $\boldsymbol{x}^*$ is approximately distributed according to f with the approximation *improving* as m increases. Note that this procedure is just a variant of the familiar *bootstrap* resampling procedure (Efron, 1982), although it is used here for a different purpose. The usual bootstrap provides equally likely resampling of the $\boldsymbol{x}_i$, while here we have weighted resampling with weights determined by the ratio of h to g. Rubin (1988) refers to this non-iterative sampling based procedure as *Sampling/Importance Resampling.*

Under the usual unweighted bootstrap, $\boldsymbol{x}^*$ has probability element equal to

$$\lim_{m\to\infty} P(x^* \in A) = \lim_{m\to\infty} \sum_{i=1}^{m} \frac{1}{m}\, 1_A(x_i) = E_g\, 1_A(x) = \int_A g(x)\, dx$$

so that $\boldsymbol{x}^*$ is approximately distributed as an observation from $g(x)$. Similarly, under the weighted bootstrap, $\boldsymbol{x}^*$ has has probability element equal to

$$\lim_{m\to\infty} P(x^* \in A) = \lim_{m\to\infty} \sum_{i=1}^{m} q_i\, 1_A(x_i) = \lim_{m\to\infty} \frac{\frac{1}{m}\sum_{i=1}^{m} \omega_i\, 1_A(x_i)}{\frac{1}{m}\sum_{i=1}^{m} \omega_i}$$

$$= \frac{E_g\, \dfrac{h(x)}{g(x)}\, 1_A(x)}{E_g\, \dfrac{h(x)}{g(x)}} = \frac{\int_A h(x)\, dx}{\int_{R^k} h(x)\, dx} = \int_A f(x)\, dx$$

so that $\boldsymbol{x}^*$ is approximately distributed as an observation from f. Note that the sample size, m, under such resampling can be as large as desired, with the approximation improving with m. An important caveat is that the less f resembles g the larger the sample size m will need to be in order that the distribution of $\boldsymbol{x}^*$ approximates f well. The match in the tails is particularly important.

Finally, the fact that either the Gibbs Sampler or the weighted bootstrap allows f to be known only up to proportionality constant, that is, only through h, is crucial, since we wish to avoid the integration required to standardize h. Note that if the normalizing constant is required in some calculation $m^{-1} \sum_{i=1}^{m} \omega_i$ provides a consistent estimator. For more on the weighted bootstrap in Bayesian analysis see the discussion by Smith and

Gelfand (1991).

Various researchers have used *importance sampling* (*IS*) and its modifications in Bayesian calculations. Wei and Tanner (1990) use *IS* in the context of data augmentation. Zeger and Karim (1991) use *acceptance/rejection* in the context of the Gibbs sampler for generalized linear models. Ritter and Tanner (1992) use an approximate cumulative distribution function to sample in the context of the Gibbs sampler. The ideas of *IS* are not at all new, see Hammersley and Handscomb (1964) for the history and further details.

2.3 Monte Carlo Marginalization

Once the multivariate set of observations from the distribution with density $f(\boldsymbol{x})$ have been generated we need to have a method for marginalization. When using the Gibbs Sampler marginalization is usually done by summing over the appropriate conditional distribution, as suggested in Gelfand and Smith (1990). Due to the complexity of some problems one may instead wish to marginalize using a Monte Carlo method. The method introduced below is convenient since we do not always have a closed form expression for the conditional density $f(x_r|x_0)$, where we use the notation x_r as the component of interest and x_0 as the remaining component.

Let S denote the support of the full joint density of $\boldsymbol{x} = (x_0, x_r)$, and let S_0, S_r be the supports of the distribution x_0 and the conditional distribution $x_r|x_0$, respectively. Therefore the marginal of x_r is

$$f(x_r) = \int_{S_0} f(x_0, x_r)\, d\, x_0.$$

To estimate $f(x_r)$ using a generated sample $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m} = \left\{x_0^{(i)}, x_r^{(i)}\right\}_{i=1}^{m}$ from $f(\boldsymbol{x}) = f(x_0, x_r)$, we calculate

$$\hat{f}(x_r) \equiv \frac{1}{m} \sum_{i=1}^{m} \phi\left(x_r^{(i)} \mid x_0^{(i)}\right) f(x_r, x_0^{(i)}) / f(x_0^{(i)}, x_r^{(i)}), \qquad (2.1)$$

where the weight function, ϕ, is any conditional density defined on S_r. Note that if $\phi(x_r|x_0) = f(x_r|x_0)$, then

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\phi(x_r^{(i)}|\,x_0^{(i)})\,f(x_r^{(i)},\,x_0^{(i)})}{f(x_r^{(i)},\,x_0^{(i)})}$$

$$=\frac{1}{m}\sum_{i=1}^{m}\frac{f(x_r^{(i)}|\,x_0^{(i)})\,f(x_r,\,x_0^{(i)})}{f(x_r^{(i)},\,x_0^{(i)})}$$

$$=\frac{1}{m}\sum_{i=1}^{m}\frac{f(x_r,\,x_0^{(i)})}{f(x_0^{(i)})}$$

$$=\frac{1}{m}\sum_{i=1}^{m}f(x_r|\,x_0^{(i)}),$$

which is the standard Gibbs Sampler approximation. Moreover, since

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}f(x_r|\,x_0^{(i)})=\int_{S_0}f(x_r|x_0)f(x_0)\,dx_0=f(x_r),$$

we have a good approximation to the marginal distribution of x_r.

The interesting fact is that the above argument holds even if $\phi(x_r|x_0)\neq f(x_r|x_0)$. Indeed, ϕ may be any function, as long as it is a conditional density (a sufficient condition for the following argument to hold). Thus, the function ϕ may be chosen to have a convenient form.

If we use (2.1) to approximate $f(x_r)$, then for any x_r^*,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\frac{\phi(x_r^{(i)}|\,x_0^{(i)})\,f(x_r^*,\,x_0^{(i)})}{f(x_r^{(i)},\,x_0^{(i)})}$$

$$=\int_{S}\frac{\phi(x_r|\,x_0)\,f(x_r^*,\,x_0)}{f(x_r,\,x_0)}f(x_r,\,x_0)\,dx_r\,x_0$$

$$=\int_{S_r}\phi(x_r|x_0)\,dx_r\int_{S_0}f(x_r^*,\,x_0)\,dx_0$$

$$=f(x_r^*),$$

since ϕ is a conditional density defined on S_r. The more closely ϕ resembles $f(x_r| x_0)$, the better the estimate, but any conditional density ϕ on S_r will work. The convergence result follows from the convergence of the Markov Chain sampling scheme or from the weighted bootstrap. A nice feature of this marginalization technique is that one does not have to develop new algorithms. For Bayesian applications of marginalization techniques see Gelfand, Smith and Lee (1991) and Chen (1992). The general theory of Monte Carlo marginalization is a consequence of the Conditional Monte Carlo method discussed in Hammersley and Handscomb (1964).

2.4 Monte Carlo Acceleration

There is also a simple method for accelerating the convergence of the Monte Carlo marginalization method, thereby reducing the number of draws required in order to achieve a preassigned level of numerical accuracy. One of the methods is an application of the theory of regression estimators from sampling theory, see Cochran (1977, Ch. 7) for more details.

We will discuss this acceleration in the context of Monte Carlo marginalization. Suppose, as in the discussion above, we have the sample $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m} = \{x_0^{(i)}, x_r^{(i)}\}_{i=1}^{m}$ from $f(\boldsymbol{x}) = f(x_0, x_r)$. Suppose there is also a function g, defined on the support of the distribution of $\boldsymbol{x}$, which has values highly correlated with $f(x_r|x_0)$ and whose integral over the support of the distribution x_0, $\bar{g}$ is known. Define the vector v as = $\{v(\boldsymbol{x}^{(i)})\}_{i=1}^{m}$ to have components

$$\{v(x^{(i)})\}_{i=1}^{m} = \{\phi(x_r^{(i)} \mid x_0^{(i)}) f(x_r, x_0^{(i)}) / f(\boldsymbol{x}_0^{(i)})\}_{i=1}^{m},$$

and define the Monte Carlo evaluation of the function g as

$$\hat{g}(x_r) = \frac{1}{m} \sum_{i=1}^{m} g(x_r, x_0^{(i)}).$$

Now compute the linear regression slope $\hat{\beta}(x_r)$ of v on the vector $\{g(x_r, x_0^{(i)})\}_{i=1}^{m}$, and compute the regression estimator of $f(x_r)$, $\hat{f}_R(x_r)$, by

$$\hat{f}_R(x_r) = \hat{f}(x_r) + \hat{\beta}(x_r)\,(\bar{g}(x_r) - \hat{g}(x_r)).$$

It is easy to see that the limiting sampling variance of $\hat{f}_R(x_r)$ equals $(1 - \rho^2)$ times the limiting sampling variance of $\hat{f}$, where ρ is the correlation between the function evaluations of f and g. Note that if the selection of g is quite poor the regression estimate will reduce to $\hat{f}$ since $\hat{\beta}(x_r)$ will be approximately equal to zero. Therefore it is clear that when these are highly correlated it is possible to decrease the sampling variance dramatically, that is, a fewer number of draws are required in order to achieve a preassigned level of numerical accuracy.

Another simple method for accelerating the convergence of the Monte Carlo marginalization method is to use common random numbers. This entails reusing the random number stream during the marginalization procedure. This is an example of the idea of *blocks* from the design of experiments. For more details and ideas on various Monte Carlo *swindles* for variance reduction and hence on reductions in the number of draws that are required in order to achieve a preassigned level of numerical accuracy see Ripley (1987, Ch. 5) and Tierney (1991).

3. *Applications of Gibbs Sample and the Weighted Bootstrap to Frequentist Inference.*

In this section we apply the Gibbs sampler and weighted bootstrap to several problems in frequentist inference. We first study the exact distribution of the likelihood estimates in the location-scale problem. Secondly we consider the problem of computing of observed significance levels in non-normal multiple regression. Lastly, we examine the computation of probabilities for Wilks' criterion for testing in MANOVA.

3.1 *The Location-Scale Problem*

In this example we can do the calculations exactly therefore we can assess the merits of our procedure. Upon simplifying the calculations in the linear model

$$y_i = \mu + \varepsilon_i \sigma \qquad i = 1, \ldots, n, \ \varepsilon_i \sim iid\, f(\cdot),$$

and applying Barndorff-Nielsen's (1983) formula (also see Fisher, 1934) we find the joint distribution of the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$, given the ancillary $a_i = (y_i - \hat{\mu})/\hat{\sigma}$ to be

$$p(\hat{\mu}, \hat{\sigma}|\mu, \sigma, a) = \frac{\hat{\sigma}^{n-2} \prod_{i=1}^{n} f((y_i - \mu)/\sigma)}{\iint \hat{\sigma}^{n-2} \prod_{i=1}^{n} f((y_i - \mu)/\sigma)\, d\hat{\mu}\, d\hat{\sigma}} .$$

By writing

$$\begin{aligned} \frac{(y_i - \mu)}{\sigma} &= \frac{(y_i - \hat{\mu})}{\hat{\sigma}}\left(\frac{\hat{\sigma}}{\sigma}\right) + \left(\frac{\hat{\mu} - \mu}{\sigma}\right)\left(\frac{\hat{\sigma}}{\sigma}\right) \\ &\equiv a_i v + t v \end{aligned}$$

and transforming to (t, v) we get

$$p(t, v|a) = \frac{v^{n-1} \prod_{i=1}^{n} f(v(a_i + t))}{\iint v^{n-1} \prod_{i=1}^{n} f(v(a_i + t))dv\, dt} .$$

In the usual case the error distribution is taken to be unit normal, and it then it follows that

$$p(t, v|a) = \frac{n^{n/2}}{\sqrt{\pi}} \left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}} v^{n-1} \, exp\{-v^2 \, (1 + t^2)/2\}$$

since $\sum_{i=1}^{n} a_i^2 = n$. Upon further calculation it can be shown that the conditional distribution of t given a has a $t(n-1)/\sqrt{n-1}$ distribution and v given a has a $\sqrt{\chi^2(n-1)/n}$ distribution. It can also be shown that the conditional distribution of t given v and a has a $N(0, 1)/(\sqrt{n}\ v)$ distribution and the distribution of v given t and v has a $\sqrt{\chi^2_{(n)}/n(1+t^2)}$ distribution.

As an experiment we studied the accuracy of two resampling methods for the normal location scale problem. The first is Gibbs sampling while the second is the weighted bootstrap. Using the conditionals of $(t|v, a)$ and $(v|t, a)$ it is easy to generate a random sample from the joint distribution of $(t, v|a)$. Note that to use the Gibbs sampler we need to have closed form expressions for $(t|v, a)$ and $(v|t, a)$ whereas for the weighted bootstrap we only need to know the joint distribution up to a multiplicative constant. The result for the Gibbs sample of normal location-scale problem are reported in Figures 3.1 and 3.2. Figure 3.1 shows that the approximation to the density of the location estimate is very good in the tails and misses a bit in the center. Figure 3.2 shows that the approximation to the density of the scale estimate is spectacular throughout the entire range. Both these were run with $n = 15$, by taking the last random quantity in a stream of *4000* and repeat this m = *200* times. Marginalization is given by summing over the appropriate conditional distribution, as suggested in Gelfand and Smith (1990). The time taken was approximately 4 minutes on an *IBM 486.*

When using the weighted bootstrap we originally choose the approximate joint density g (as discussed above) of $(t, v|a)$ as a Normal$(0, 1/n)$ times the square root of a $\chi^2(n)/n$. The results of the weighted bootstrap for the normal location-scale problem are reported in Figures 3.3 and 3.4 plotting the differences of the true and simulated pdf's of the location and scale estimates, respectively. The story is similar to the one for Gibbs sampling. We used 800 samples the get one observation, this was replicated 8000 times. Marginalization was done via the Monte Carlo marginalization scheme discussed in Section 2.3. The time taken was approximately 3 minutes on an *IBM 486.*

Comparing the two sampling schemes we find the weighted bootstrap to be much more convenient for this problem. We did not need to have a closed form for the conditionals and it took much less time.

As the next example we consider the exponential location-scale problem. If the error distribution is taken to be unit exponential then it follows that $\hat{\mu} = y_{(1)}$ the sample minimum, $\hat{\sigma} = \bar{y} - \hat{\mu}$ and $a_i = (y_i - \hat{\mu}) / \hat{\sigma} = (y_i - y_{(1)}) / (\bar{y} - \hat{\mu})$. Then

$$p(t, v|a) = \frac{n^n}{(n-2)!} \; v^{n-1} \, exp\{-n \, v \, (1 + t)\}$$

since here $\sum_{i=1}^{n} a_i = n$. Upon further calculation it can be shown that the conditional distribution of t given a has an $F(2, 2n - 2) / (n - 1)$ distribution and v given a has a $Gamma(n - 1, n^{-1})$ distribution. It can also be shown that the conditional distribution of t given v and a has a $Exponential((n\ v)^{-1})$ distribution and the distribution of v given t and v has a $Gamma(n, [n(1 + t)]^{-1})$ distribution.

Since we have the full conditionals we could use the Gibbs Sampler, although we do not since the weighted bootstrap turns out to be more efficient. In this case we use the weighted bootstrap with approximate joint density of $(t, v|a)$ as a $Gamma(1, n)$ times an $Exponential(n)$. Marginalization was done via the Monte Carlo marginalization scheme discussed in Section 2.3. The results of this example are reported in Figures 3.5 – 3.8. Figures 3.5 and 3.7 plot the differences of the true and simulated of the location parameter estimate's pdf and cdf, respectively. Figures 3.6 and 3.8 plot the differences of the true and simulated of the scale parameter estimate's pdf and cdf, respectively. We used 1600 samples the get one observation, this was replicated 32000 times. The time taken was approximately 6 minutes on an *IBM 486*.

3.2 *Non-Normal Multiple Regression*

Consider the linear regression model for non-normal errors of the form $y = X\beta + \sigma\varepsilon$, where ε has a density $f_\lambda(\varepsilon)$ on R^n and X is an $n \times p$ design matrix with regression parameter vector β. In principle f_λ may depend on unknown parameters $\lambda \in R^q$, however, we will assume that λ is fixed. Standard methods of conditioning in transformation models (Fraser, 1979, p. 113) use the one-to-one change of variables from y to $\{\hat{\beta}(y), \hat{\sigma}(y), d(y)\}$, where $\hat{\beta}(y)$ and $\hat{\sigma}(y)$ are estimates of location and scale, respectively $d(y) = (y - X\hat{\beta}(y))/\hat{\sigma}(y)$, is a standardized residual. The conditional joint distribution of $\hat{\beta}(y) = b$ and $\hat{\sigma}(y) = s$ given d by Fraser (1979, p. 114) as

$$p(b, s|d)\, db\, ds = h_{\lambda}^{-1}(d)\, f_{\lambda} \left[(X(b - \beta) + sd)\sigma^{-1} \right] (s/\sigma)^{n}\, s^{-(p+1)}\, |X^TX|^{1/2}\, db\, ds$$

on $R^p \times R^+$. The ancillary statistic $d(y)$ are standardized residuals and have distribution

$$h_{\lambda}(d) = \int_{R^p \times R^+} f_{\lambda}(Xb + sd)\, s^{n-p-1}\, |X^TX|^{1/2}\, db\, ds.$$

By applying the transformation $w = log\ \hat{\sigma}(y)$, the density on $R^p \times R^+$ is given by

$$p(b,\ w|d)\ db\ dw = h_{\lambda}^{-1}(d)\, f_{\lambda}\left[e^{-\ log\sigma} (X(b - \beta) + e^{w}d) \right] e^{n(w\ -\ log\sigma)} e^{-\ pw}\, |X^TX|^{1/2}\ db\ dw.$$

Note that the desired quantities are

$$t = (\hat{\beta}(y) - \beta)/\hat{\sigma}(y) \quad \text{and} \quad w = w(y) - log\ \sigma = log\,(\hat{\sigma}(y)\ /\ \sigma). \qquad (3.1)$$

The conditional joint distribution of (t, w) is

$$p(t, w|d)\, dt\, dw = h_{\lambda}^{-1}(d)\, f_{\lambda}\ \left[(Xt + d)\, e^{w}\right] e^{nw}\, |X^TX|^{1/2}\, dt\, dw. \qquad (3.2)$$

When we are interested in inference on a single parameter, say β_p, thus inference will be based on the marginal density which is obtained by integrating out the appropriate components. That is, we want to compute

$$f(t_p) = \int f\,(t,\ w|d)\ dt_1\ dt_2 \ldots dt_{p-1}\ dw.$$

the pivotal density that provides inference for β_p through the pivot $t_p = (\hat{\beta}_p(y) - \beta_p)/\hat{\sigma}(y)$.

As in Fraser *et al.* (1990) we propose to approximate the marginal density by a one-dimensional conditional density. Fraser *et al.* use an analytic approximation method, while we will use a sampling-based approach. We will use the notation x_r as the component of interest and x_0 as the remaining component. In particular, we wish to test H_o: $\beta_p = \beta_{po}$, and, for the regression problem, using the notation discussed earlier, let $(x_0, x_r) = (w, t_1, t_2, ..., t_{p-1}, t_p)$ where $x_r = t_p = (\hat{\beta}_p(y) - \beta_{po})/s$. Although we will only discuss examples with scaler x_r, the component of interest, the component may be a vector.

Fraser *et al.* (1990) approximate the density of $f(x_r)$ by the one-dimensional conditional density which is proportional to $f(\hat{x}_0(x_r), x_r)$ where $\hat{x}_0\,(x_r)$ is the value of x_0 that maximizes $f(x_0, x_r)$ for each x_r. To improve the approximation of the marginal density by the conditional density Fraser *et al.* (1990) modify the original point distribution in a way that does not change the marginal distribution of the component of interest but may improve the approximations of the conditional distribution to the marginal. This involves the construction of a pseudo-model for the data that has the same marginal density for the component of interest. Let

$$\hat{j} = \left[-\frac{d}{dxdx^T}\, \ell(x) \right]_{(\hat{x}_0, \hat{x}_r)}$$

be the x_r Hessian of $\ell(x) = log\, f(x_0, x_r)$ evaluated at the overall maximum $(\hat{x}_0, \hat{x}_r)$, and let

$$\hat{j}\,(x_r) = \left| \frac{\partial^2}{\partial x_0 x_0^T}\, \ell(x_0, x_r) \right|_{\hat{x}_0(x_r)}$$

be the $(r - 1) \times (r - 1)$ negative Hessian of $\ell(x_0, x_r)$ for fixed x_r at its restricted maximum. By defining the new variable $z_0 = \hat{j}^{1/2}(x_r)\,[\hat{x}_0 - \hat{x}_0\,(x_r)]$ one can see that the joint density of (z_0, x_r) is proportional to $g(z_0, x_r) = |\hat{j}\,(x_r)|^{-1/2} f\,(\hat{x}_0\,(x_r) + \hat{j}^{1/2}(x_r)\, z_0, x_r)$. Hence the observed level of significance $P(x_r \geq x_r^{obs})$ may be approximated by

$$\int_{x_r^{obs}}^{\infty} g\,(0, x_r)\; d\,x_r \Big/ \int_{-\infty}^{\infty} g\,(0, x_r)\; d\,x_r.$$

Fraser *et al.* (1990) refer to this as the ridge observed level of significance.

In this example we compare our technique with Fraser *et al.* (1990) using one of their examples found in their article. Here we combine the marginalization technique with the weighted bootstrap method. We consider the regression model $y = \alpha + \beta x + \sigma\varepsilon$, where ε is a standardized Student random variable on λ degrees of freedom. For 25 values of the explanatory variable x at unit step size from -12 to $+12$, the response values for y are recorded in Table 3.1. The data were generated with $\alpha = 20$, $\beta = 1$, $\sigma = 1.1966$ and $\lambda = 3$. Let t_1 and t_2 be the "t-statistics" for α and β, respectively, and let be as in (3.1).

Table 3.1. A sample of 25 regression responses with *Student (3)* error

7.9042	16.2425	9.9128	10.0184	12.8359
12.8607	15.1697	16.0589	16.6068	18.5075
19.1212	19.8824	21.3117	21.6194	21.6348
23.2321	23.0110	24.7835	23.3734	26.7593
29.1283	24.6564	29.9679	31.4070	32.6893

Suppose we are interested in testing $\beta = 1$. The first thing that needs to be done is to generate data from the joint distribution of the two regression parameter test statistics and the scale statistics. We use the weighted bootstrap, for this will approximate the joint density $N(0,\ 1/n) \times N(0,\ 1) \times \sqrt{\chi^2(n)/n}$. We used 1600 samples to get one observation, this was replicated 32000 times. Marginalization was done via the Monte Carlo marginalization scheme discussed in Section 2.3, and was used to find the marginal density of the test statistic $t_2 = (\hat{\beta} - 1)/s$. The conditional density $\phi(\cdot|\cdot)$ weight in (2.1) was chosen to be a univariate normal with mean equal to zero and variance equal to the inverse of the sample variance of the independent variables. Recall that the choice of $\phi(\cdot|\cdot)$ need only be a conditional density on the support of t_1 given the normalized t_2 and w. Therefore our choice is a matter of convenience. A better weight function may be possible, however ours seems to work well in practice. We also used the regression acceleration method discussed in Section 2.4, with the regression function g equal to the product $N(0,\ 1/n) \times N(0,\ 1) \times \sqrt{\chi^2(n)/n}$. Again the choice of this function is out of convenience.

The observed level of significance was then computed with Monte Carlo integration. Table 3.2 compares the approximate value of Fraser *et al.* with our Monte Carlo method. The time taken was approximately 10 minutes. We see that the results are quite similar to the approach of Fraser *et al.*; whose approximation is much less computationally intensive, but requires more analytic calculations in addition to a one-dimensional Monte Carlo integration.

Table 3.2: Observed level of significance for β using a Student *(λ)* analysis with $\lambda = 1, 3, 6, \infty$.

	$\lambda = 1$	$\lambda = 3$	$\lambda = 6$	$\lambda = \infty$
Fraser *et al.* ols	.3113	.2607	.1643	.0238
Monte Carlo ols	.3297	.2724	.1762	.0361

3.3 Wilks' Likelihood Ratio Test

A p-dimensional multivariate analysis of variance (MANOVA) has error sum of square matrix S_e with n degrees of freedom and a hypothesis sum of squares S_n with q degrees of freedom. To test the usual MANOVA hypothesis one rejects when the likelihood ratio

$$\Lambda_{p,\, q,\, n} = |S_e + S_n|^{-1}\, |S_e|$$

is smaller than a certain critical value, where $|\cdot|$ denotes the determinant. Under the usual normal sampling distribution assumptions S_e and S_n are independent Wishart matrices with a common parameter matrix under the null hypothesis. Anderson (1984) shows that under the null hypothesis

$$\Lambda_{p,\, q,\, n} \overset{\mathcal{D}}{=} \prod_{i=1}^{p} B_e[(n-i+1)/2,\, q/2], \tag{3.1}$$

on (0, 1] where the $B[\cdot, \cdot]$'s are independent beta random variables with the indicated degrees of freedom.

Butler, Huzurbazar and Booth (1992) use this characterization (and others) to derive a saddlepoint approximation to the null distribution of log $\Lambda_{p,\, q,\, n}$. This approximation is very accurate. However, as with any saddlepoint approximation in order to implement the

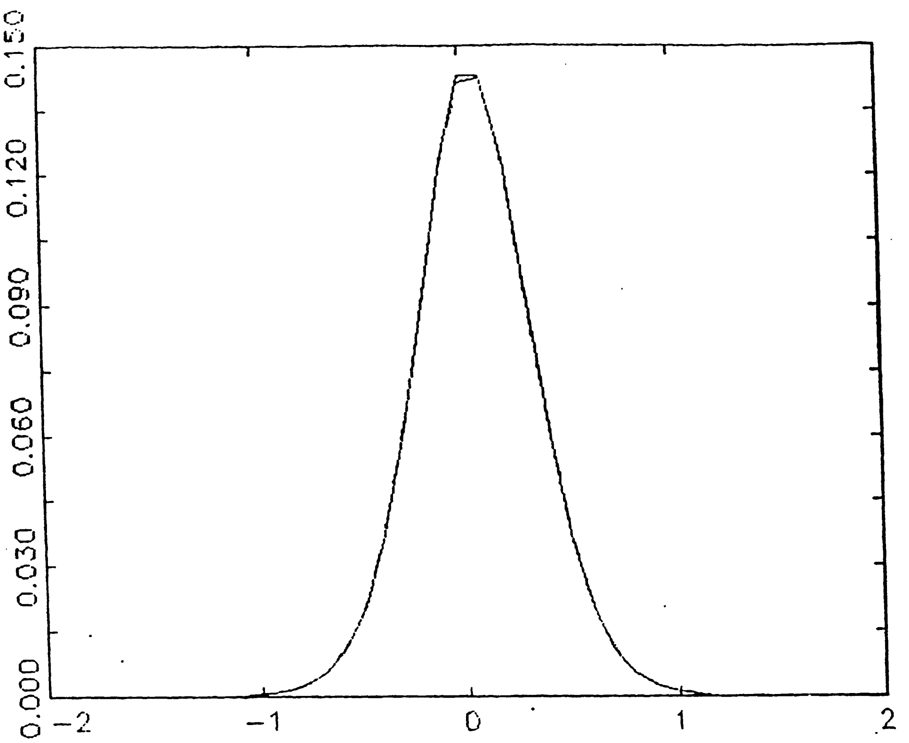

Figure 3.1: True versus estimated pdf for the location estimate problem using Gibbs for Normal data.

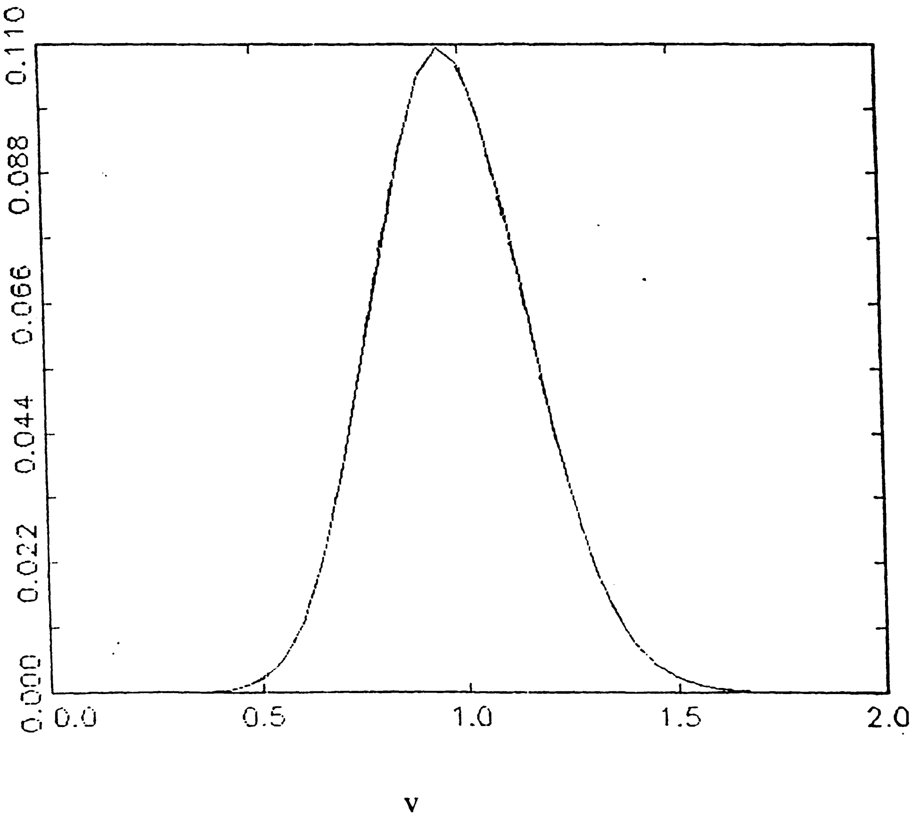

Figure 3.2: True versus estimated pdf for the scale estimate using Gibbs for Normal data.

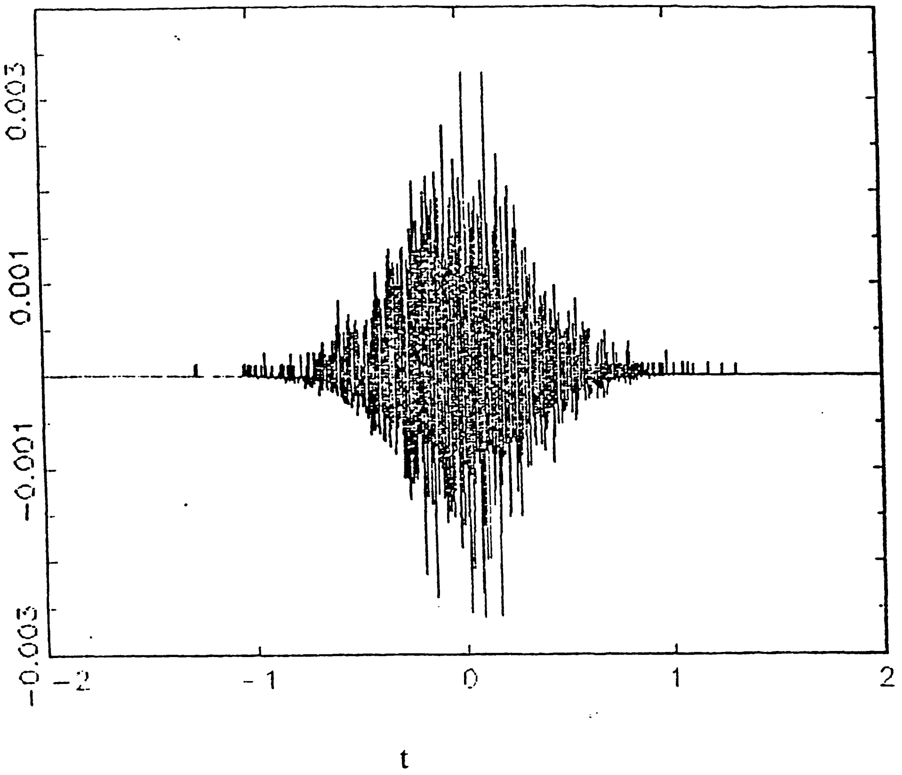

Figure 3.3: Difference in the estimated and true pdf's of the location estimate using the weighted bootstrap for normal data.

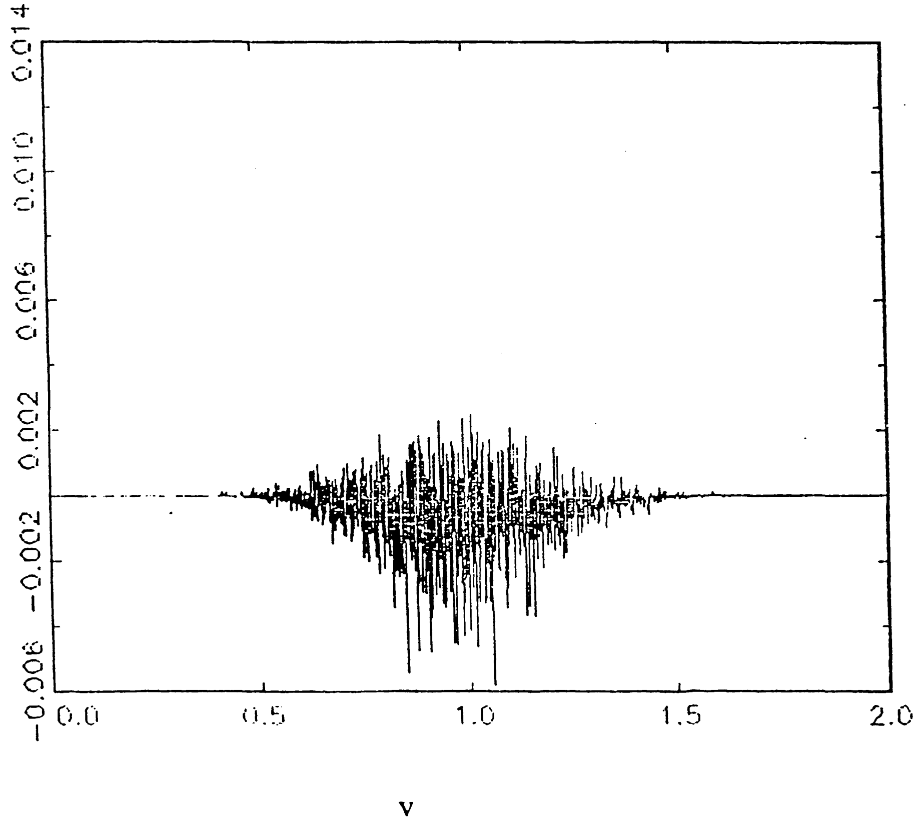

Figure 3.4: Difference in the estimated and true pdf's of the scale estimate using the weighted bootstrap for Normal data.

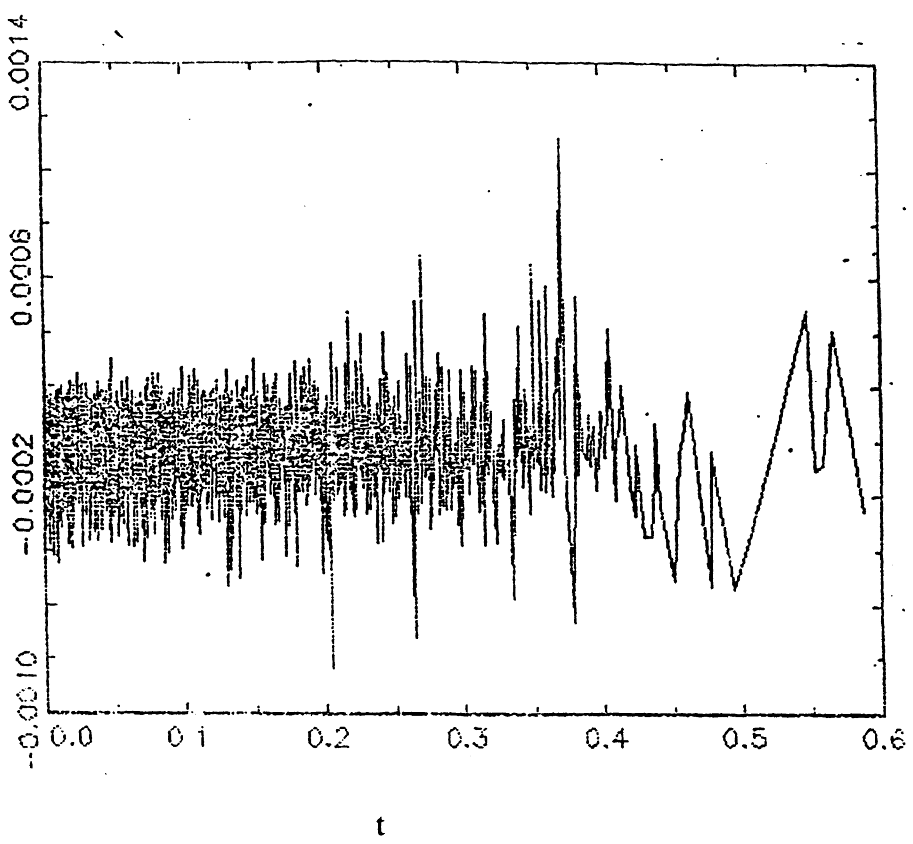

Figure 3.5: Differences in the estimated and true pdf's of the location estimate using the weighted bootstrap for exponential data.

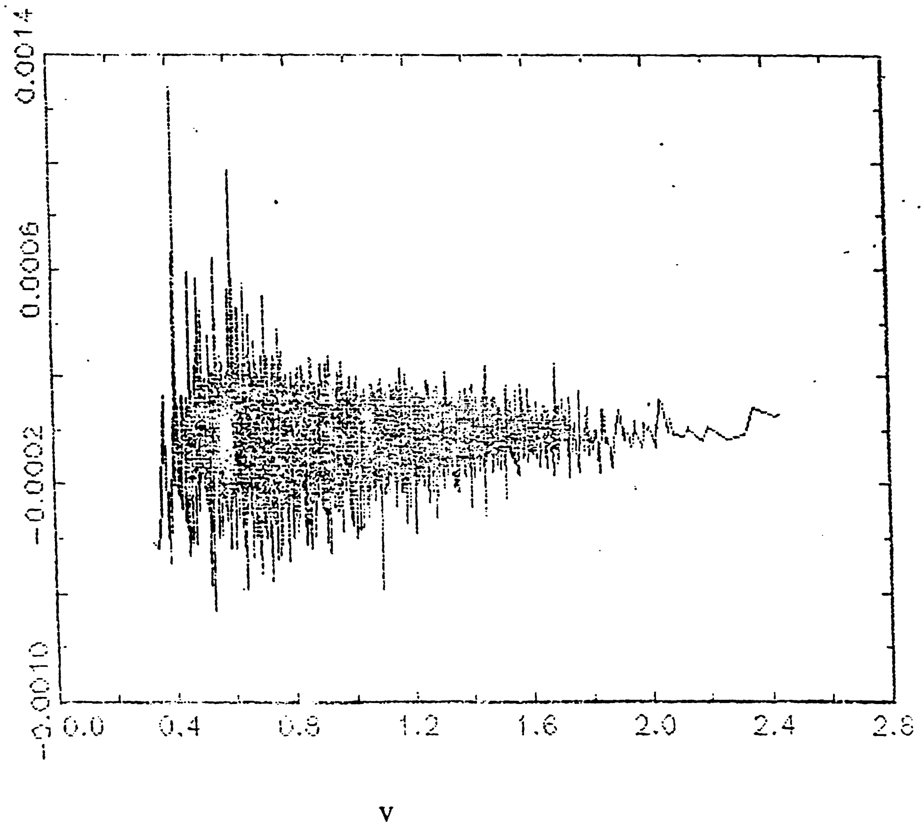

Figure 3.4: Differences in the estimated and true pdf's of the scale estimate using the weighted bootstrap for exponential data.

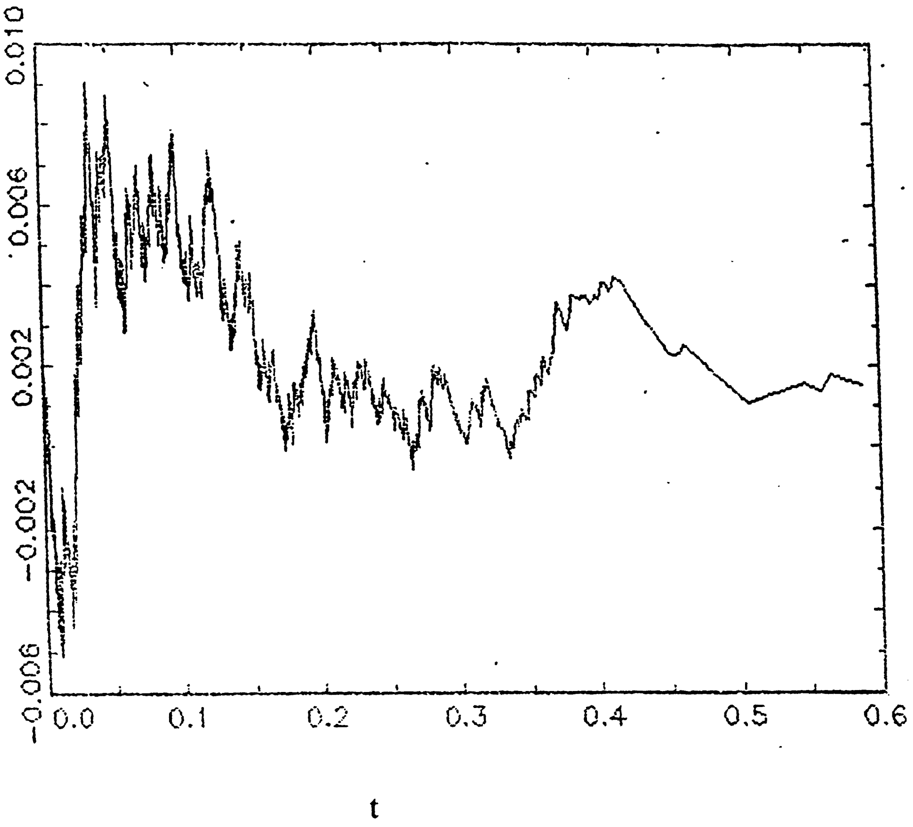

Figure 3.7: Differences in the estimated and true pdf's of the location estimate using the weighted bootstrap for exponential data.

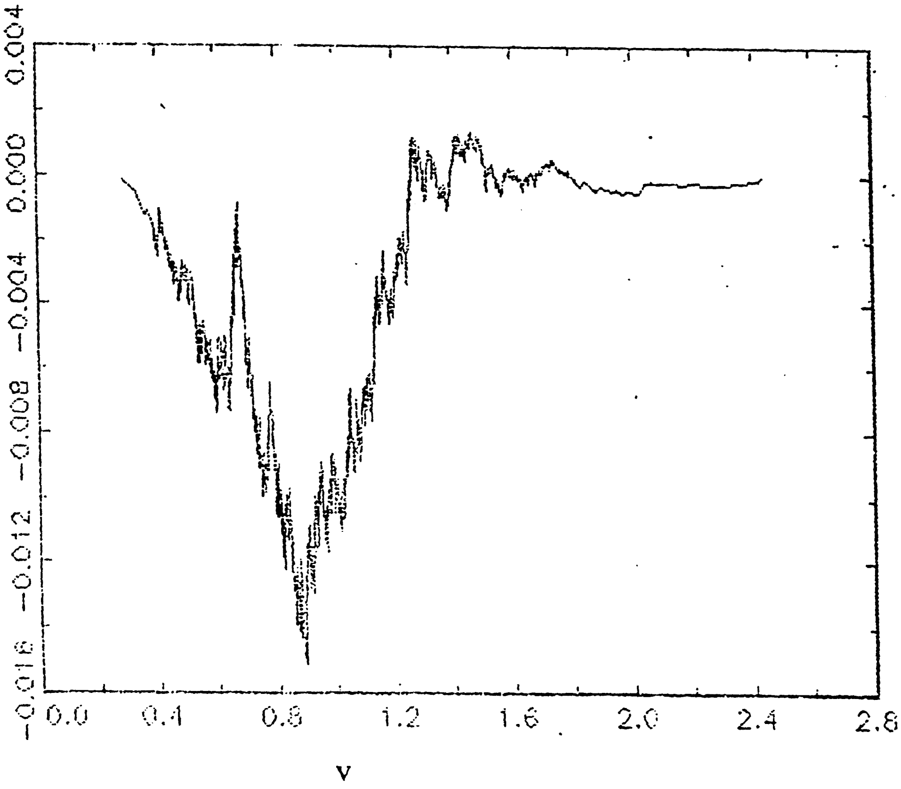

Figure 3.8: Differences in the estimated and true pdf's of the scale estimate using the weighted bootstrap for exponential data.