Statistics and Its Interface Volume 2 (2009) 247–254

Bayesian R-estimates in linear models

Xiaojiang Zhan and Thomas P. Hettmansperger*

A Beyesian approach to applying nonparametric rank-based methodology to linear models is discussed. Information in the data is summarized by a rank-based quantity, whose asymptotic distribution is used as a pseudo-likelihood. The posterior distribution (up to a normalizing constant) of the coefficient(s) given the rank-based quantity can be obtained by assuming a prior distribution for the coefficient(s) in the linear model. This posterior distribution, together with simulation methods (typically the Markov Chain Monte Carlo methodology), can then be used for inference.

Keywords and phrases: Robust estimate, Rank estimate, Bayesian analysis, Linear models.

1. INTRODUCTION

In data analysis, common parametric modeling methods force restrictive assumptions regarding the underlying population(s), such as symmetry and unimodality. These parametric modeling approaches are usually based on least squares or maximum likelihood fitting, and are easily impaired by outlying observations. Also, they can be quite inefficient when the hypothesized distribution deviates from the true data distribution in the heaviness of tails, skewness, etc. In many situations, nonparametric rank-based approaches provide more flexible modeling specifications, can be much more efficient relative to the traditional parametric methods when the data distribution differs from the assumed distribution, and do not lose much efficiency when the assumed model is correct. On the other hand, the Bayesian approach to inference is attractive in incorporating prior information into the inference machinery by Bayes' theorem, resulting in a unifying, constructive inference methodology. Hence, it is desired to combine the rank-based methods with the Bayesian approach and utilize their respective advantages in data analysis. This combination of the two methodologies, which we call the Bayesian semiparametric method, can be applied to many models widely used in practice, including the one-sample and two-sample location models, and linear models [12–14]. In this paper we focus on the application to linear models, especially regression models.

Jeffreys [7] pioneered this approach with the simple sign statistic in the one-sample location model. Monahan and Boos [10] argued that, since the binomial distribution of the sign statistic is not a true "likelihood", posterior probability statements may be problematic for small sample sizes. In large samples in which the "likelihood" is replaced by a normal approximation, as is the case in this paper, the problem largely disappears. In addition, Lavine [9] made a case for the use of these "likelihoods" when considering a vector of quantiles, which was later extended by Dunson and Taylor [1].

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ denote the $n \times 1$ vector of observations and let $\boldsymbol{X}$ denote the $n \times p$ design matrix whose ith row is $\boldsymbol{x}_i'$, the ith vector of explanatory variables. We assume $\boldsymbol{X}$ has full column rank p. Then consider the linear model

$$\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\mathbf{1}$ is an $n \times 1$ column vector of ones, α is the scalar intercept parameter, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, and $\boldsymbol{e}$ is an $n \times 1$ vector of i.i.d. errors with a common c.d.f. F, where F is absolutely continuous with a density function $f = F'$ and $F(0) = 1/2$. Hence the median of the distribution of Y_i is $\alpha + \boldsymbol{x}_i'\boldsymbol{\beta}$. Since model (1) includes an intercept parameter α, we can assume without loss of generality that $\boldsymbol{X}$ is centered; namely, the columns of $\boldsymbol{X}$ sum to 0. Then the subspaces spanned by $\mathbf{1}$ and the columns of $\boldsymbol{X}$ are orthogonal, which results in uncorrelated estimates of α and $\boldsymbol{\beta}$. We will concentrate mainly on estimation of $\boldsymbol{\beta}$, since α can be estimated using the Bayesian semiparametric procedures for the one-sample location models [12, 13]. In Section 2 we present the posterior distribution of $\boldsymbol{\beta}$. Then a normal approximation to the posterior distribution is discussed in Section 3 by using the asymptotic linearity result for the negative of the gradient of a pseudo-norm. This approximation clarifies the nature of the Bayesian semiparametric estimates and also facilitates use of the Markov Chain Monte Carlo (MCMC) technique in posterior estimation by providing a good jumping distribution and starting points in the Metropolis algorithm. Posterior inference by the Metropolis algorithm is given in Section 4, and concluding remarks are given in Section 5.

2. POSTERIOR DISTRIBUTION OF β

In order to better understand the estimation procedure, we first give a brief description of the geometry of estimation, which is detailed in Hettmansperger and McKean [5]. Let Ω denote the column space spanned by the columns

*Corresponding author.

of $\boldsymbol{X}$. Given a norm or a pseudo-norm, we can estimate $\boldsymbol{\beta}$ by some $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ minimizes the distance between $\boldsymbol{Y}$ and the subspace Ω. Consider the pseudo-norm

$$\|\boldsymbol{u}\|_\varphi = \sum_{i=1}^{n} a[R(u_i)]u_i,$$

where $a(1) \le a(2) \le \cdots \le a(n)$ is a set of scores generated as $a(i) = \varphi[i/(n+1)]$ for some nondecreasing scores generating function $\varphi(u)$ defined on $(0,1)$ and standardized such that $\int_0^1 \varphi(u)du = 0$ and $\int_0^1 \varphi^2(u)du = 1$. For the above pseudo-norm, we define the dispersion function by

$$\begin{aligned} D_\varphi(\boldsymbol{\beta}) &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_\varphi \\ &= \sum_{i=1}^{n} a[R(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})](Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \end{aligned}$$

which is a piecewise linear, continuous and convex function of $\boldsymbol{\beta}$. Then an estimate $\hat{\boldsymbol{\beta}}_\varphi$ of $\boldsymbol{\beta}$ can be chosen such that $\hat{\boldsymbol{\beta}}_\varphi = \text{Argmin} D_\varphi(\boldsymbol{\beta})$. We denote the negative of the gradient (defined almost everywhere) of $D_\varphi(\boldsymbol{\beta})$ by

$$\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{a}[R(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})],$$

where $\boldsymbol{a}[R(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})]' = (a[R(Y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})], \ldots, a[R(Y_n - \boldsymbol{x}_n'\boldsymbol{\beta})])$. Thus $\hat{\boldsymbol{\beta}}_\varphi$ solves the equations

$$\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{a}[R(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})] \doteq \boldsymbol{0}.$$

For the Bayesian semiparametric inference in linear models, we summarize the information in the data via the asymptotic normal distribution of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ (the pseudo-likelihood), which requires certain assumptions on the distribution of the errors, the design matrix and the scores. Before we state a theorem regarding the asymptotic normality of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, we first describe the necessary assumptions below.

First we assume the error density function f has finite Fisher information $I(f)$; that is,

(2)
$$f \text{ is absolutely continuous, } 0 < I(f) = \int_0^1 \varphi_f^2(u)du < \infty,$$

where

$$\varphi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}.$$

In the asymptotic result below, the design matrix $\boldsymbol{X}$ is assumed to be imbedded in a sequence of design matrices satisfying the following two properties:

(3)
$$\lim_{n\to\infty} \max_{1\le i\le n} h_{iin} = 0$$

and

(4)
$$\lim_{n\to\infty} n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Sigma},$$

where h_{iin} are the diagonal elements, subscripted by n, of the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix. Note that condition (3) is a necessary and sufficient design condition for the least squares estimates to have an asymptotic normal distribution, provided that the errors are i.i.d. with finite variance [6].

The following theorem gives the asymptotic normal distribution of the gradient $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ under the hypothesis that $\boldsymbol{\beta}$ is the true vector of parameters. The proof can be found in Hettmansperger and McKean [5].

Theorem 2.1. *Let the scores generating function $\varphi(u)$ be a nondecreasing and square-integrable function defined on $(0,1)$ and standardized such that $\int_0^1 \varphi(u)du = 0$ and $\int_0^1 \varphi^2(u)du = 1$. Under model (1) and conditions (2), (3) and (4),*

(5)
$$n^{-1/2}\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\beta}$ is assumed to be the true vector of parameters.

We follow the suggestion of Jeffreys [7] and treat the distribution of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ as a pseudo-likelihood. Choosing $\boldsymbol{\beta}$ to maximize the probability of the observed value of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is equivalent to solving the rank estimating equation $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$. We go further and use the normal approximation. If f were known then φ_f, defined above, determines the asymptotically most powerful test for $\boldsymbol{\beta}$. In addition, the vector of ranks is asymptotically sufficient; see Hájek, Šidák, and Sen (Section 8.1) [3]. These points along with the robustness of rank methods suggest that we can apply Bayesian updating to the approximate distribution of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. We carry out this plan in the following discussion.

As aforementioned, we summarize the information in the data via the asymptotic normal distribution of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, with $\boldsymbol{\Sigma}$ estimated by $n^{-1}\boldsymbol{X}'\boldsymbol{X}$, and the prior for $\boldsymbol{\beta}$ takes the form of $\mathcal{N}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$, a multivariate normal distribution with mean $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. Then the posterior density function $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$ is proportional to the product of the two multivariate normal density functions, i.e.

(6)
$$\begin{aligned} &p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sqrt{n}}\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right]'\boldsymbol{\Sigma}^{-1}\left[\frac{1}{\sqrt{n}}\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right]\right\} \\ &\quad \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}. \end{aligned}$$

This posterior distribution is a segmented and tilted version of the prior $\mathcal{N}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$. To see this, first note that $R(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, the ranks of the residuals, can only change at the boundaries of the regions defined by the $\binom{n}{2}$ equations $Y_i - \boldsymbol{x}_i'\boldsymbol{\beta} = Y_j - \boldsymbol{x}_j'\boldsymbol{\beta}$. In the interior of these regions the ranks are constant. Put another way, the surface of the

asymptotic normal distribution of $\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$, as a function of $\boldsymbol{\beta}$, is in the form of a step function consisting of flat planes with data-dependent jumps. In the interior of these planes the asymptotic normal densities remain unchanged, and hence the posterior densities $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}))$ within a plane are proportional to the product of the unchanged asymptotic normal density and the changing prior normal densities, which implies that the posterior distribution is a segmented and tilted version of the prior. Unlike the posterior density functions used in the one-sample and two-sample location models, where the detailed changes of the posterior density functions as $\boldsymbol{\beta}$ changes can be characterized [12–14], the complexity of the posterior density function $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}))$ is increased tremendously by its high dimensionality and we can no longer characterize its detailed changes as $\boldsymbol{\beta}$ changes. Consequently, posterior calculation can only be done by simulation methods, typically an MCMC approach—the Metropolis algorithm.

For any given posterior distribution, one can implement the Metropolis algorithm in an infinite number of ways. Even after reparameterizing, there are still endless choices for the jumping distribution, from which candidates are sampled. If a normal approximation to the posterior distribution can be constructed and then used as the jumping distribution, the resulting Metropolis algorithm is often more efficient and outperforms alternative random walk Metropolis algorithms with other jumping distributions. Therefore, we discuss in the next section a normal approximation to $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}))$, which can also be used to obtain rough estimates of the posterior quantities.

3. NORMAL APPROXIMATION TO THE POSTERIOR DISTRIBUTION WITH ASYMPTOTIC LINEARITY

In order to construct a normal approximation to the posterior distribution, we use the asymptotic linearity result for the gradient $\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$ [5], which can be written as follows for our purpose:

$$\begin{aligned}(7)\quad &\frac{1}{\sqrt{n}}\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\varphi)\\ &\quad=\frac{1}{\sqrt{n}}\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})-\tau_\varphi^{-1}\boldsymbol{\Sigma}\sqrt{n}(\hat{\boldsymbol{\beta}}_\varphi-\boldsymbol{\beta})+o_p(1),\end{aligned}$$

where $\boldsymbol{\beta}$ is the true vector of parameter, $\hat{\boldsymbol{\beta}}_\varphi$ is the classical rank-based estimate defined previously and the scale parameter τ_φ is defined as

$$\tau_\varphi^{-1}=\int\varphi(u)\varphi_f(u)du.$$

Since $\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\varphi)\doteq\boldsymbol{0}$, the above linearity yields

$$(8)\qquad \frac{1}{\sqrt{n}}\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\doteq\tau_\varphi^{-1}\boldsymbol{\Sigma}\sqrt{n}(\hat{\boldsymbol{\beta}}_\varphi-\boldsymbol{\beta}).$$

Then by (6),

$$\begin{aligned}(9)\quad &p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}))\\ &\propto\exp\left\{-\frac{1}{2}\left[\tau_\varphi^{-1}\boldsymbol{\Sigma}\sqrt{n}(\hat{\boldsymbol{\beta}}_\varphi-\boldsymbol{\beta})\right]'\boldsymbol{\Sigma}^{-1}\left[\tau_\varphi^{-1}\boldsymbol{\Sigma}\sqrt{n}(\hat{\boldsymbol{\beta}}_\varphi-\boldsymbol{\beta})\right]\right\}\\ &\qquad\cdot\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}\\ &\qquad\propto\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}^{*-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\},\end{aligned}$$

where

$$(10)\qquad \boldsymbol{\Sigma}^*=(\tau_\varphi^{-2}n\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_0^{-1})^{-1},$$

$$(11)\qquad \boldsymbol{\beta}^*=\boldsymbol{\Sigma}^*(\tau_\varphi^{-2}n\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_\varphi+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0)$$

and the last row of (9) is obtained by expanding the exponents, collecting terms and then completing the quadratic form in $\boldsymbol{\beta}$. Therefore, the posterior $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}))$ can be approximated by a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\beta}^*,\boldsymbol{\Sigma}^*)$. The scale parameter τ_φ can be estimated by several methods discussed in Section 3.7.1 of Hettmansperger and McKean [5]. We estimate τ_φ in a frequentist perspective and do not integrate its estimation into the Bayesian semiparametric inference machinery. This is because the normal approximation, in which τ_φ appears, is not used in the regular construction of the posterior and not for general inference purposes. As discussed later, use of the normal approximation is restricted to providing starting points and serving as the jumping distribution in the Metropolis algorithm. In the above approximation, the posterior mean $\boldsymbol{\beta}^*$ is a weighted average of the rank-based estimate and the prior mean, with weights given by the data and prior precision matrices, $\tau_\varphi^{-2}n\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0^{-1}$, respectively. This is because, by the nonparametric theory, the asymptotic distribution of the classical rank-based estimate $\hat{\boldsymbol{\beta}}_\varphi$ is given by

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\varphi-\boldsymbol{\beta})\xrightarrow{\mathcal{L}}\mathcal{N}_p(\boldsymbol{0},\tau_\varphi^2\boldsymbol{\Sigma}^{-1}).$$

Example 3.1. In this example we generate two posterior surface plots using the normalized posterior densities and the approximate multivariate normal densities, respectively, based on simulated data with 25 observations ($n=25$). Suppose $\boldsymbol{X}^*=(\boldsymbol{X}_1^*,\boldsymbol{X}_2^*)$ is a 25×2 design matrix consisting of two explanatory variables $\boldsymbol{X}_1^*$ and $\boldsymbol{X}_2^*$, where $\boldsymbol{X}_1^*=(X_{1,1},\ldots,X_{1,25})'$ is a random sample of size 25 drawn from $\mathcal{N}(20,5^2)$ and $\boldsymbol{X}_2^*=(X_{2,1},\ldots,X_{2,25})'$ is a random sample of size 25 drawn with replacement from the sequence $1,1.5,2,2.5,\ldots,50$. Let $\boldsymbol{Y}=(Y_1,\ldots,Y_{25})'$ be generated such that $Y_i=10+6X_{1,i}^*+20X_{2,i}^*+e_i,\ i=1,\ldots,25$, where the errors e_i follow a $\mathcal{N}(0,10^2)$ distribution. Let $\boldsymbol{X}$ be the centered version of $\boldsymbol{X}^*$, i.e. $\boldsymbol{X}=\boldsymbol{X}^*-\boldsymbol{1}(\bar{x}_1^*,\bar{x}_2^*)$, where $\bar{x}_i^*,\ i=1,2$, is the mean of the ith column of $\boldsymbol{X}^*$ and

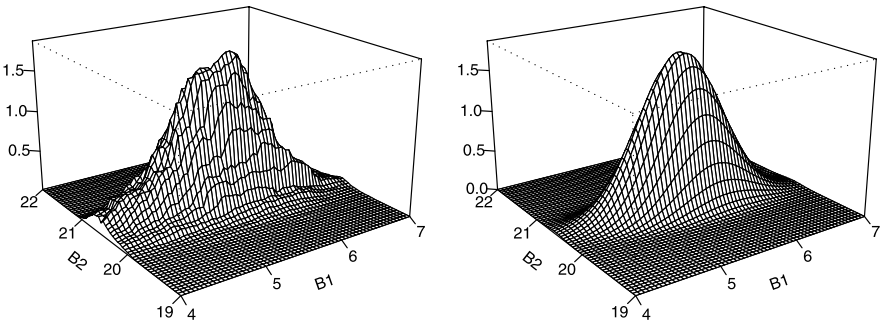

Figure 1. Posterior surface plots for Example 3.1. The plot on the left corresponds to the normalized posterior densities and the one on the right corresponds to the approximate bivariate normal densities.

$\mathbf{1}$ is a 25×1 column vector of ones. The scores generating function we use in the example is $\varphi(u) = \sqrt{12}(u - 1/2)$, which is used to generate the Wilcoxon scores. The corresponding classical rank-based estimate is $\hat{\boldsymbol{\beta}}_\varphi = (5.49, 20.38)'$ and an estimate of τ_φ is 12.70 calculated by the RGLM package (see http://www.stat.wmich.edu/slab/RGLM/). A $\mathcal{N}_2(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ distribution is taken as a noninformative prior, with $\boldsymbol{\beta}_0 = (0, 0)$ and

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 10^4 & 0 \\ 0 & 10^4 \end{pmatrix}.$$

While the posterior $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$ in (6) is unnormalized, we need a normalized posterior density function to take the scale into account and produce a valid visualization of the posterior surface. The details of approximating the normalizing constant are discussed in Zhan [12], where the Laplace's method [8] is used to approximate the marginal density of $\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Figure 1 shows the posterior surface plots using the normalized posterior densities and the approximate bivariate normal densities, respectively, based on the above simulated data. A visual comparison suggests that the normal approximation to the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$ using asymptotic linearity works fairly well even for this moderate sample size.

Our experience with the preceding normal approximation suggests that its performance deteriorates when substantive prior information that conflicts with the data is available, especially for large samples. As the sample size increases, the data precision matrix $\tau_\varphi^{-2} n \boldsymbol{\Sigma}$ quickly dominates $\boldsymbol{\beta}^*$ over the prior precision matrix $\boldsymbol{\Sigma}_0^{-1}$ (see (11)), leaving the approximate posterior mean $\boldsymbol{\beta}^*$ little influenced by the prior mean even in the presence of a fairly strong prior. For this reason, we recommend estimating the posterior mode $\tilde{\boldsymbol{\beta}}$ rather than using $\boldsymbol{\beta}^*$ as the center of the normal approximation, especially given substantive prior information that conflicts with the data. The estimation of $\tilde{\boldsymbol{\beta}}$ can be achieved by various numerical methods solving optimization problems. Noticing that the asymptotic linearity result (8) makes it possible to calculate the first and second derivatives of the log posterior density, one may attempt to estimate $\tilde{\boldsymbol{\beta}}$ by the Newton-Raphson method. It turns out, however, that the Newton-Raphson estimate is also prone to the inaccuracy in approximation under certain circumstances because the algorithm, together with asymptotic linearity, essentially uses (9) as an approximate posterior density and thus finds $\boldsymbol{\beta}^*$ as the posterior mode. Therefore, given substantive prior information that conflicts with the data, we should resort to optimization methods other than the Newton-Raphson algorithm and other algorithms requiring evaluation of derivatives. Meanwhile, we limit use of the normal approximation to the jumping distribution and starting points in the Metropolis algorithm, for which the approximation inaccuracy in question doesn't seem to be a problem and the normal approximation still provides good choices of the jumping distribution and starting points. Of course, when the sample size is moderate or large, the choice of the jumping distribution is seldom an issue since many effective jumping distributions can be found easily.

4. POSTERIOR INFERENCE BY MCMC SIMULATION

Given the normal approximation to the posterior distribution using asymptotic linearity considered in the previous section, posterior inference can proceed by using the approximate normal distribution as the jumping distribution in the Metropolis algorithm. Being aware that there is no fully satisfactory method for drawing simulations in general, we approximate the posterior quantities by the following algorithm, which is often successful for simulating from $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$ in our context.

1. Construct the approximate (multivariate) normal posterior $\mathcal{N}_p(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}^*)$ based on an estimated posterior mode $\tilde{\boldsymbol{\beta}}$. Draw $J \geq 2$ samples from the approximate normal posterior distribution and use them as starting points to run J independent sequences of the following Metropolis algorithm. For each of these starting points, we perform the steps below.

2. Select a starting point $\boldsymbol{\beta}^0$. For $t = 1, 2, \ldots$:
   - Sample a candidate point $\boldsymbol{\beta}^+$ from the approximate normal posterior distribution with mean $\boldsymbol{\beta}^{t-1}$ and covariance matrix $\boldsymbol{\Sigma}^*$.
   - Calculate the ratio of the densities,
   $$r = \frac{p(\boldsymbol{\beta}^+|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^+))}{p(\boldsymbol{\beta}^{t-1}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{t-1}))}$$
   - Set
   $$\boldsymbol{\beta}^t = \begin{cases} \boldsymbol{\beta}^+, & \text{with probability } \min(r, 1); \\ \boldsymbol{\beta}^{t-1}, & \text{otherwise.} \end{cases}$$
3. Run the above iterative simulation until approximate convergence appears to have been reached, in the sense that the statistic $\sqrt{\hat{R}}$, defined below, is near 1 for each component of $\boldsymbol{\beta}$.
4. Let K be the length of each sequence after discarding a certain percent of the early iterations. Summarize inference about the posterior distribution by collecting the $J \times K$ samples and treating them as samples from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$.

An approach to monitoring convergence of the estimand (or each scalar estimand for the multiparameter case) and at the same time estimating the posterior variance of the estimand was proposed in Gelman et al. [2]. Let K be the length of each sequence after discarding a certain number of the early simulations. We label the m^{th} component of each p-dimensional draw from J independent sequences of length K as β_{ij}^m $(i = 1, \ldots, K;\ j = 1, \ldots, J)$, and compute B and W, the between- and within-sequence variances:

$$B = \frac{K}{J-1}\sum_{j=1}^{J}(\bar{\beta}_{\cdot j}^m - \bar{\beta}_{\cdot\cdot}^m)^2,$$

where

$$\bar{\beta}_{\cdot j}^m = \frac{1}{K}\sum_{i=1}^{K}\beta_{ij}^m, \quad \bar{\beta}_{\cdot\cdot}^m = \frac{1}{J}\sum_{j=1}^{J}\bar{\beta}_{\cdot j}^m,$$

and

$$W = \frac{1}{J}\sum_{j=1}^{J}s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{K-1}\sum_{i=1}^{K}(\beta_{ij}^m - \bar{\beta}_{\cdot j}^m)^2.$$

Note that if only one sequence is simulated, B cannot be computed. We proceed to estimate $Var(\beta^m|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$, the marginal posterior variance of β^m, by a weighted average of W and B, i.e.

$$\widehat{Var}(\beta^m|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})) = \frac{K-1}{K}W + \frac{1}{K}B,$$

which "overestimates the marginal posterior variance assuming the starting distribution is appropriately overdispersed, but is unbiased under stationarity (that is, if the starting distribution equals the target distribution), or in the limit $K \to \infty$" [2]. As recommended in Gelman et al. [2], we can monitor convergence of the MCMC simulation by a potential scale reduction $\sqrt{R}$ estimated by

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{Var}(\beta^m|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))}{W}},$$

which declines to 1 as $K \to \infty$. The potential scale reduction $\sqrt{R}$ is the factor by which the scale of the current distribution for β^m might be reduced if the simulations were continued as $K \to \infty$. If $\sqrt{R}$ is high, proceeding with further simulations may increase the accuracy of the posterior inference.

We next illustrate the above simulation procedure in the following example by a simple linear regression problem, for which the robustness against outliers will also be considered for convenience of visualizing the regression line in a simple linear model.

Example 4.1. Consider a simple linear regression model

$$\boldsymbol{Y} = \alpha + \beta\boldsymbol{X} + \boldsymbol{e},$$

where $\boldsymbol{X} = (X_1, \ldots, X_{30})$ is a random sample drawn with replacement from the sequence $1, 1.5, 2, 2.5, \ldots, 50$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_{30})$ were generated such that $Y_i = 10 + 6X_i + e_i$, $i = 1, \ldots, 30$, where the errors e_i follow a $\mathcal{N}(0, 10^2)$ distribution. We manually corrupted the response values of three observations and fit the model using different methods: the classical rank-based procedure and the Bayesian semiparametric procedure, with both using the Wilcoxon scores and the corrupted data, and the least squares approach with and without the three corrupted observations, respectively. For the Bayesian semiparametric approach, the prior distribution is $\mathcal{N}(6.22, 300^2)$, where 6.22 is the classical rank-based estimate of β and also the estimate of the posterior mode. The estimate of τ_φ is 21.19, which is calculated by the RGLM package and is required in the computation of $\boldsymbol{\Sigma}^*$. Of course, the choices of the starting points and the jumping distribution are not critical for a simple regression model of this sample size. Different specifications of the starting points and the jumping distribution can lead to a successful convergence of the simulation sequences. Table 1 displays the estimates for the aforementioned methods. The Bayesian semiparametric estimates are quite consistent with the traditional rank-based estimates, both of which are robust against the three unusual observations. The robustness can be seen more easily in Figure 2. The least squares line with outliers is conspicuously pulled towards the outliers, while the Bayesian semiparametric (and the rank-based) equations stay closer to the least squares line without outliers. As mentioned in Section 1, the intercept

Table 1. Estimates for Example 4.1

Method	$\hat{\beta}^{\dagger}$	SE‡	$\hat{\alpha}$
Bayesian Semiparametric	6.22	0.41	8.50
Rank-based	6.22	0.26	8.54
Least Squares (w/ Outliers)	6.87	0.60	6.11
Least Squares (w/o Outliers)	5.97	0.19	10.64

Note: $\sqrt{\hat{R}} = 1.000$.
†Posterior median for the Bayesian semiparametric method.
‡Posterior SD for the Bayesian semiparametric method.

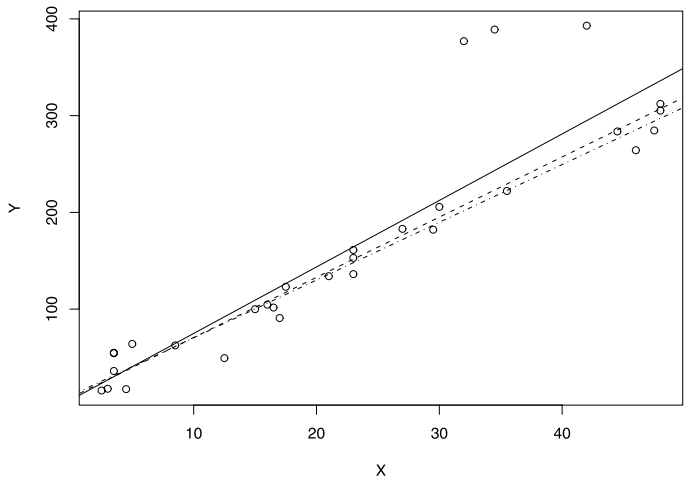

Figure 2. Scatter plot and regression equation lines for Example 4.1. The solid line represents the least squares equation with the three outliers included, the dashed line in the middle represents the Bayesian semiparametric equation, and the dot-dashed line represents the least squares equation with the three outliers excluded.

α can be estimated using the procedures discussed in Zhan [12] and Zhan and Hettmansperger [13] for the one-sample location models. Specifically, $\hat{\alpha}$ is the median of the Walsh averages of the residuals provided that we assume symmetry of the error distribution. If the symmetry assumption is not reasonable, on the other hand, then $\hat{\alpha}$ can be simply taken as the median of the residuals. Since estimation of α is not the problem we address in this paper, we only report the point estimates of α in Table 1 assuming symmetry of the error distribution.

Example 4.2. We consider a real data example in which proper priors are in order. The example comes from forest ecology in which measurements are taken on trees with the goal of estimating the volume of timber in a given forest. It is important to have a quick way to estimate the volume of wood in any tree. It is not too difficult to measure the height of a tree and even easier to measure the circumference and hence the diameter. A sample of black cherry trees

Table 2. Data for Example 4.2

Volume	10.3	10.3	10.2	16.4	18.8	19.7	15.6	18.2	22.6	
	19.9	24.2	21.0	21.4	21.3	19.1	22.2	33.8	27.4	
	25.7	24.9	34.5	31.7	36.3	38.3	42.6	55.4	55.7	
	58.3	51.5	51.0	77.0						
Diameter	8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	
	11.2	11.3	11.4	11.4	11.7	12.0	12.9	12.9	13.3	
	13.7	13.8	14.0	14.2	14.5	16.0	16.3	17.3	17.5	
	17.9	18.0	18.0	20.6						
Height	70	65	63	72	81	83	66	75	80	75
	79	76	76	69	75	74	85	86	71	64
	78	80	74	72	77	81	82	80	80	80 87

from the Allegheny National Forest was selected and cut. The volume, height and diameter (at 4.5 feet above ground) were recorded; see Table 2. If we treat the usable wood in a tree as roughly a cylinder then the formula for the volume is $V = \pi(d/2)^2 h$, where d is the diameter and h is the height. Further, $\ln V = \ln(\pi/4) + 2\ln d + \ln h$. Let $Y = \ln V$, $X_1 = \ln d$, and $X_2 = \ln h$. Then $E(Y) = \ln(\pi/4) + 2X_1 + X_2$. Hence, we take as a prior distribution for $\boldsymbol{\beta}$ a bivariate normal distribution with mean vector $\boldsymbol{\beta}_0 = (2, 1)$ and diagonal covariance matrix

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Choice of either prior variance reflects the strength of belief that the cut tree is roughly a cylinder. For example, if a tree tapers substantially then the cylinder is more like a truncated cone and we could take a relatively large value for σ_1^2. The use of the rank-based methods protects against outliers in y-space but not x-space. In the least squares analysis there are two observations flagged as having large residuals. Three cases of prior variances were considered in our analysis. Case 1: If we believe that β_1 (for diameter) is between 1.5 and 2.5, then σ_1 may be set to 1/6 (we used "the 99.7 rule" and put 6 standard deviations over the interval between 1.5 and 2.5) and $\sigma_1^2 = 0.028$. As we may feel more confident about the height, we may take β_2 between 0.75 and 1.25, and hence take $\sigma_2 = 0.5/6 = 1/12$ and $\sigma_2^2 = 0.007$. Case 2: If we are less confident about β_1, then we may take $\sigma_1 = 4/6 = 2/3$ and $\sigma_1^2 = 0.444$; we keep σ_2 the same as Case 1. Case 3: For comparison purposes, we also assume a noninformative prior by taking $\sigma_1^2 = \sigma_2^2 = 10^4$. As shown in Table 3, the Bayesian semiparametic method with informative priors (Cases 1 and 2) yield narrower interval estimates than the rank-based and least squares methods, particularly for β_2 for which a strong prior (small prior variance) is used.

When the sample size is relatively small (e.g. 10–12 observations per slope parameter), the choices of the starting points and the jumping distribution may be crucial and need to be generated as discussed in the above algorithm. Oth-

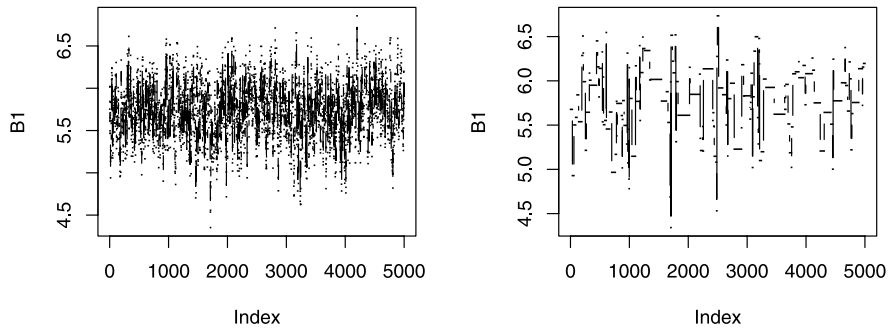

Figure 3. Time series plots of β_1 after a 5000 iteration burn-in for the regression model ($n = 50$) in Example 3.1. A multivariate normal jumping distribution with covariance matrix $\boldsymbol{\Sigma}^*$ and $diag(5, 5)$ is used for the plots on the left and right, respectively, which shows a more efficient simulation with the use of $\boldsymbol{\Sigma}^*$.

Table 3. Estimates for Example 4.2

Method	$\hat{\beta}_1(95\%\, CI^\dagger)$	$\hat{\beta}_2(95\%\, CI^\dagger)$	$\hat{\alpha}^\ddagger$
BS Case 1	2.00 (1.88, 2.13)	1.02 (0.86, 1.17)	−6.26
BS Case 2:	2.00 (1.87, 2.15)	1.02 (0.86, 1.17)	−6.26
BS Case 3:	1.98 (1.80, 2.17)	1.13 (0.59, 1.65)	−6.69
Rank-based	1.96 (1.81, 2.10)	1.15 (0.76, 1.54)	−6.70
LS	1.98 (1.83, 2.14)	1.12 (0.70, 1.54)	−6.63

Note: BS = Bayesian Semiparametric; LS = Least Squares; $\sqrt{\hat{R}} = 1.000$ for both β_1 and β_2; $\hat{\tau}_\varphi = 0.08$.
†95% Credible set for the Bayesian semiparametric method.
‡Based on the median of the residuals for the Bayesian semiparametric and rank-based methods.

erwise the simulation sequences may not converge successfully, yielding bumpy times series plots for the sequences. On the other hand, the specifications of the starting points and the jumping distribution need not adhere strictly to the above algorithm in the presence of moderate or large sample sizes (e.g. greater than 15 observations per slope parameter), since many other specifications may also yield convergent sequences. However, specifying a jumping distribution (i.e. a multivariate normal distribution with a suitably chosen covariance matrix) other than the one discussed above may result in a less efficient algorithm in that the consequent jumps may be rejected too frequently and the random walk may waste too much time standing still. This is illustrated in Figure 3, where we implement the Metropolis algorithm for the same regression model in Example 3.1, but the sample size is now $n = 50$. The times series plot of β_1 on the left is based on a multivariate normal jumping distribution with covariance matrix $\boldsymbol{\Sigma}^*$, while the plot on the right is based on a multivariate normal jumping distribution with covariance matrix

$$\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}.$$

Clearly, it takes longer for the random walk in the right plot to range over the parameter space of β_1, since the jumps are rejected too frequently and the random walk wastes too much time standing still. In contrast, the random walk in the left plot is much more efficient and moves quickly.

Sample sizes play an important role in the performance of the Bayesian semiparametric method in linear regression models, particularly when the Metropolis algorithm is used to obtain posterior inference. Convergence of the simulation sequences can be obtained quickly and easily for large samples and the resulting estimates of the posterior variances for components of $\boldsymbol{\beta}$ are reasonably small. However, the estimates for small samples may not be stable and may require more effort in monitoring the simulation process. As shown in Zhan [12], by using the Metropolis algorithm, the Bayesian semiparametric approach with a noninformative prior has excellent frequentist operating characteristics for moderate to large samples, whereas its performance may be worrisome for small samples. When the number of observations per slope parameter falls below 10, the Metropolis algorithm typically breaks down and fails to provide convergent simulation sequences.

The stability issue with the Metropolis algorithm for small samples (i.e. 8–12 observations per slope parameter) seems to be attributed to the jumping (or rejection) rule of the algorithm, which makes it possible for the random walk to jump to points with trivial posterior densities (i.e. points where the posterior density is not concentrated) in the parameter space of $\boldsymbol{\beta}$. Once a candidate point with a trivial posterior density is accepted, the random walk may wander around the parameter space with trivial posterior densities, and it may or may not come back to the region where the posterior density $p(\boldsymbol{\beta}|\boldsymbol{S}_\varphi(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}))$ is concentrated because the data are not strong enough to pull it back. A computation alternative, which does not encompass any jumping (or rejection) rule and thus offers stabler estimates for small samples, is the adaptive importance sampling (AIS) scheme proposed by Oh and Berger [11]. See Zhan [12] for the details of the posterior calculation by AIS.

5. CONCLUDING REMARKS

This paper deals with an implementation of the rank-based methods in the Bayesian framework for linear models, with the discussion focused on the linear regression problem. It has been found that the proposed Bayesian semiparametric analysis makes it possible to utilize prior information about the parameters in the rank-based methods, and the resulting estimates inherit robustness against potential outliers in y-space (but not X-space) from the classical rank-based estimates. Further, posterior calculation by the Metropolis algorithm (or other simulation methods such as AIS), while requiring more effort in monitoring convergence for small samples, provides stable estimates quickly for moderate to large samples in terms of the number of observations per slope parameter. All this makes the proposed approach a robust and handy tool in a Bayesian analysis of linear models.

The prior distribution used throughout the discussion is the (multivariate) normal distribution, which allows an appealing interpretation of the posterior mean as a compromise between the classical rank-based estimate and the prior mean. Although we selected a normal prior due to its reasonableness and mathematical convenience, other non-normal prior distributions may be assumed in the Bayesian semiparametric inference. For instance, using a (multivariate) t distribution as an informative prior in the presence of substantive prior information can help prevent unusual observations from exerting an undue influence on posterior inference. In the event that a noninformative prior is desired, its specific form should not matter, provided that the prior chosen is vague and approximates fairly well a priori ignorance of the parameters of interest. Caution should be exercised, however, when one attempts to adopt uniform distributions as noninformative priors. Use of an improper prior—a uniform distribution which is proportional to a constant on the parameter space—will result in an improper posterior distribution and thus should be avoided. The impropriety of the posterior distribution is due to the fact that the pseudo-likelihood, as a function of the parameter(s), remains constant as the parameter(s) go to positive or negative infinity. In other words, all values of the parameter(s) beyond the range of the data are equally likely according to the pseudo-likelihood. A proper posterior can be obtained, though, by using a bounded support uniform prior that stays constant within certain bounds, and the determination of the bounds depends upon scientific knowledge about the parameter(s) or a sensible guess of the extreme limits.

Various scores generating functions can be used in the Bayesian semiparametric analysis. Different scores generating functions have their own strengths in terms of nonparametric robustness, i.e. robustness against outliers, for different underlying distributions. Discussion pertinent to choices among a variety of scores generating functions can be found in Hettmansperger [4]. Although different scores are compared for different underlying distributions based on the asymptotic relative efficiency which is defined in classical nonparametrics, we have found empirically that the nonparametric efficiency results also offer working guidelines for selecting suitable scores to achieve robustness in the Bayesian semiparametric analysis.

Received 5 September 2008

REFERENCES

[1] Dunson, D. B. and Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400. MR2129840

[2] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman & Hall/CRC, London. MR1385925

[3] Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of rank tests*, 2nd edition. Academic Press, San Diego, CA. MR1680991

[4] Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. John Wiley & Sons, New York. MR0758442

[5] Hettmansperger, T. P. and McKean, J. W. (1998). *Robust nonparametric statistical methods*. Arnold, London. MR1604954

[6] Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, New York. MR0606374

[7] Jeffreys, H. (1961). *Theory of probability*, 3rd edition. Oxford University Press, London. MR0187257

[8] Kass, R. E. and Raftery, A. E. (1995). *Journal of the American Statistical Association*, **90**, 773–795

[9] Lavine, M. (1995). On an approximate likelihood for quantiles. *Biometrika*, **82**, 220–222. MR1332852

[10] Monahan, J. R. and Boos, D. D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, **79**, 271–278. MR1185129

[11] Oh, M.-S. and Berger, J. O. (1992). Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, **41**, 143–168. MR1276184

[12] Zhan, X. (2005). *Bayesian semiparametric inference based on ranks*. Unpublished Ph.D. dissertation, Department of Statistics, The Pennsylvania State University.

[13] Zhan, X. and Hettmansperger, T. P. (2005). Bayesian semiparametric inference based on ranks in one-sample location models. *Journal of Statistical Research*, **39**, 61–75. MR2195208

[14] Zhan, X. and Hettmansperger, T. P. (2006). Bayesian R-estimates in two-sample location models. *Computational Statistics & Data Analysis*, **51**, 5077–5089. MR2370708

Xiaojiang Zhan
Merck & Co., Inc.
RY34-A316
126 E. Lincoln Ave.
Rahway, NJ 07065, USA
E-mail address: xiaojiang_zhan@merck.com

Thomas P. Hettmansperger
Department of Statistics
The Pennsylvania State University
University Park, PA 16802, USA
E-mail address: tph@stat.psu.edu