Bayesian R-Estimates

Tom Hettmansperger
Penn State University

Xiaojiang Zhan
Merck

Primary References:

Jeffreys (1998) *Theory of Probability, 3rd ed.*

Hodges and Lehmann (1963) Ann. Math. Statist.

Motivated by discussions with Jaeyoung Lee, Seoul National University.

Suppose $\mathbf{X} = (X_1, \ldots, X_n)^T$ *iid*

$G(x|\theta) = F(x - \theta)$ where $F(\theta) = 1/2$, uniquely.

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ the realized sample

1. Prior $\pi(\theta)$ on Ω
2. Likelihood $L(\mathbf{x}|\theta) = \Pi_{i=1}^n f(x_i - \theta)$
3. Posterior

$$p(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)\pi(\theta)}{\int_\Omega L(\mathbf{x}|\theta)\pi(\theta)d\theta} \propto L(\mathbf{x}|\theta)\pi(\theta)$$

Suppose $X_1, \ldots, X_n$ iid $n(\theta, \sigma^2)$ with σ^2 known and prior is $n(\mu_0, \sigma_0^2)$.

$p(\theta|\mathbf{x}) \propto$

$$\exp\left\{-\frac{1}{2\sigma^2}\Sigma(x_i - \theta)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}\Sigma(\theta - \mu_0)^2\right\}$$

$$\exp\left\{-\frac{n}{2\sigma^2}(\theta - \overline{x})^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}\Sigma(\theta - \mu_0)^2\right\}$$

The **Bayes estimate** (square error loss):

$$E(\Theta|\mathbf{x}) = \left\{\frac{n}{\sigma^2}\overline{x} + \frac{1}{\sigma_0^2}\mu_0\right\} / \left\{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right\}$$

$L(\mathbf{x}|\theta)$ summarizes info about θ contained in data.

$L(\mathbf{x}|\theta)$ updates the prior into the posterior.

Replace $L(\mathbf{x}|\theta)$ by the distribution of some rank based quantity, denoted $T(\mathbf{X},\theta)$ and use this distribution as a pseudo likelihood.

Let $g(T(\mathbf{x},\theta)|\theta)$ denote the pmf of $T(\mathbf{X},\theta)|\theta$ evaluated at the realized data $\mathbf{x}$. Call $g(T(\mathbf{x},\theta)|\theta)$ **pseudo likelihood or the T-likelihood.**

The Sign statistic:

Suppose $T(\mathbf{x},\theta) = \Sigma I(x_i \leq \theta)$, then the T-likelihood is determined by $B(.5,\theta)$.

$$g(T(\mathbf{x},\theta)|\theta) = \binom{n}{T(\mathbf{x},\theta)}\left(\frac{1}{2}\right)^n$$

$$= \binom{n}{0}\left(\frac{1}{2}\right)^n \quad for \ \theta < x_{(1)}$$

$$= \binom{n}{1}\left(\frac{1}{2}\right)^n \quad for \ x_{(1)} \le \theta < x_{(2)}$$

$$\vdots$$

$$= \binom{n}{n}\left(\frac{1}{2}\right)^n \quad for \ x_{(n)} \le \theta$$

The T-likelihood estimate $\widehat{\theta}$ is a value of θ that maximizes $g(T(\mathbf{x},\theta)|\theta)$.

One of these values, $\widehat{\theta} = med(x_i)$, also solves the R-estimating equation,

$$T(\mathbf{x},\theta) = \Sigma I(x_i \le \theta) \simeq n/2.$$

Now use $g(T(\mathbf{x},\theta)|\theta)$ to update the prior $\pi(\theta)$.

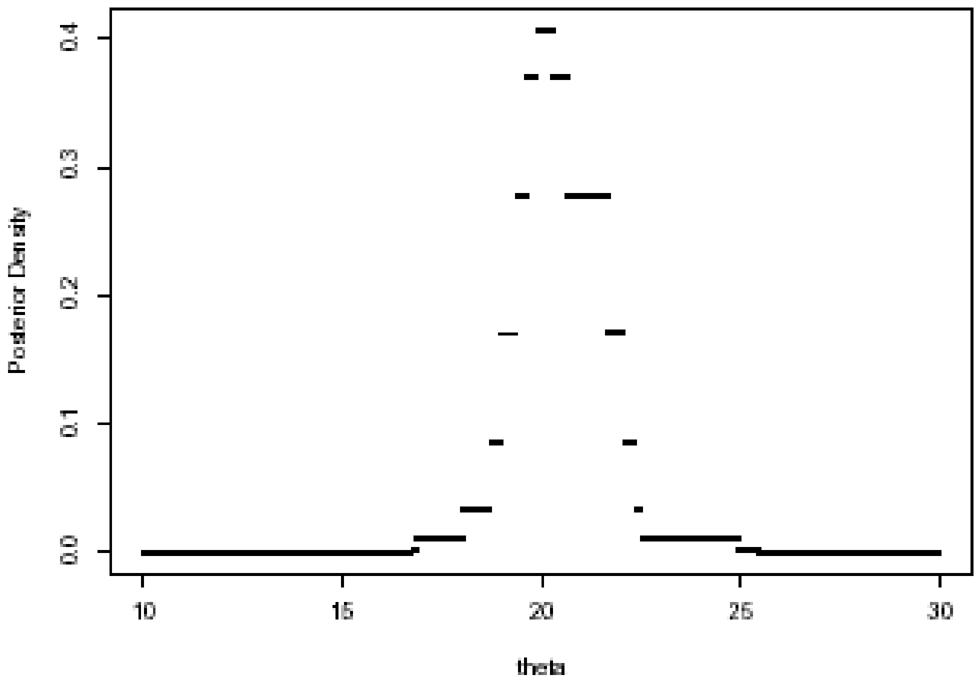
$$p(\theta|T(\mathbf{x},\theta)) = \frac{\binom{n}{0}\left(\frac{1}{2}\right)^n \pi(\theta)}{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \pi(\theta)d\theta} \quad for \;\; \theta <$$

$$\vdots$$

$$= \frac{\binom{n}{1}\left(\frac{1}{2}\right)^n \pi(\theta)}{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \pi(\theta)d\theta} \quad for \;\; x_{(n)} \leq$$

$$= \frac{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n I(x_{(i)} \leq \theta < x_{(i+1)})}{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \pi(\theta)d\theta}$$

- The plot of $p(\theta|T(\mathbf{x},\theta))$ is a segmented version of the prior.
- The $n+1$ segments are determined by the partition of R induced by $x_{(1)} < \ldots < x_{(n)}$.



$$E(\Theta|\mathbf{x}) = \frac{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \theta\pi(\theta)d\theta}{\Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \pi(\theta)d\theta}$$

$$E(\Theta|\mathbf{x}) = \frac{\frac{1}{2}\Sigma_{i=0}^{n}\binom{n}{i}(x_{(i+1)}^2 - x_{(i)}^2)}{\Sigma_{i=0}^{n}\binom{n}{i}(x_{(i+1)} - x_{(i)})} = \Sigma w_i\left(\frac{x_{(i)} + x_{(i+1)}}{2}\right)$$

$$w_j = \binom{n}{j}(x_{(j+1)} - x_{(j)})/\Sigma_{i=0}^{n}\binom{n}{i}(x_{(i+1)} - x_{(i)})$$

Example:

- Generate a sample of size 20 from $n(20, 5^2)$.
- Prior $\pi(\theta)$ is $n(25, 100^2)$ (vague)

	B.S. Estimates				L.S. Estimates	
	Mode	Mean	Median	95% C.S.	$\bar{x}$	95% C.I.
Orig. Data	20.29	20.49	20.41	(18.34, 22.77)	20.48	(18.74, 22.21)
Corrupted	20.29	20.53	20.41	(18.34, 23.11)	21.45	(18.65, 24.25)

Data:

13.87, 13.99, 16.75, 16.87, 18.01, 18.71, 18.99, 19.37, 19.64, 19.88,
20.29, 20.66, 21.68, 22.06, 22.35, 22.49, 24.92, 25.45, 26.40, 27.12

Corruption: $x_{(19)}$ and $x_{(20)}$ were shifted far to the right.

Consider next the **Wilcoxon signed rank statistic and assume underlying distribution** F is symmetric:

$$T(\mathbf{X},\theta) = \sum_{i=1}^{n} R_i(\theta) I(X_i \leq \theta)$$

where $R_i(\theta)$ is the rank of $|X_i - \theta|$ among the absolute values.

The counting form is more convenient:

$$T(\mathbf{X},\theta) = \sum\sum_{i \leq j} I\left(\frac{X_i + X_j}{2} \leq \theta \right)$$

- $ET(\mathbf{X},\theta) = n(n+1)/2$,
- $VarT(\mathbf{X},\theta) = n(n+1)(2n+1)/24$
- $T(\mathbf{X},\theta)$ is approx normally distributed.

Recall the sign statistic T-likelihood:

$$g(T(\mathbf{x},\theta)|\theta) = \Sigma_{i=0}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^{n} I(x_{(i)} \leq \theta < x_{(i+1)})$$

No closed form for the pmf of Wilcoxon.

$$p_i = P(T(\mathbf{X},\theta) = i \mid \theta) \;\; for \;\; i = 1,\ldots,N = \frac{n(n+1)}{2}$$

Then the Wilcoxon T-likelihood is

$$g(T(\mathbf{x},\theta)|\theta) = \Sigma_{i=0}^{N} p_i I(w_{(i)} \leq \theta < w_{(i+1)})$$

where $w_{(1)} \leq \ldots \leq w_{(N)}$ are the ordered $n(n+1)/2$ pairwise averages $\frac{x_i + x_j}{2}$ $i \leq j$.

We need p_i for computing the posterior.

- $dsignrank$ in R returns the exact values
- p_i can be approximated using the asy normal dist.
- Edgeworth approx will improve the approx.

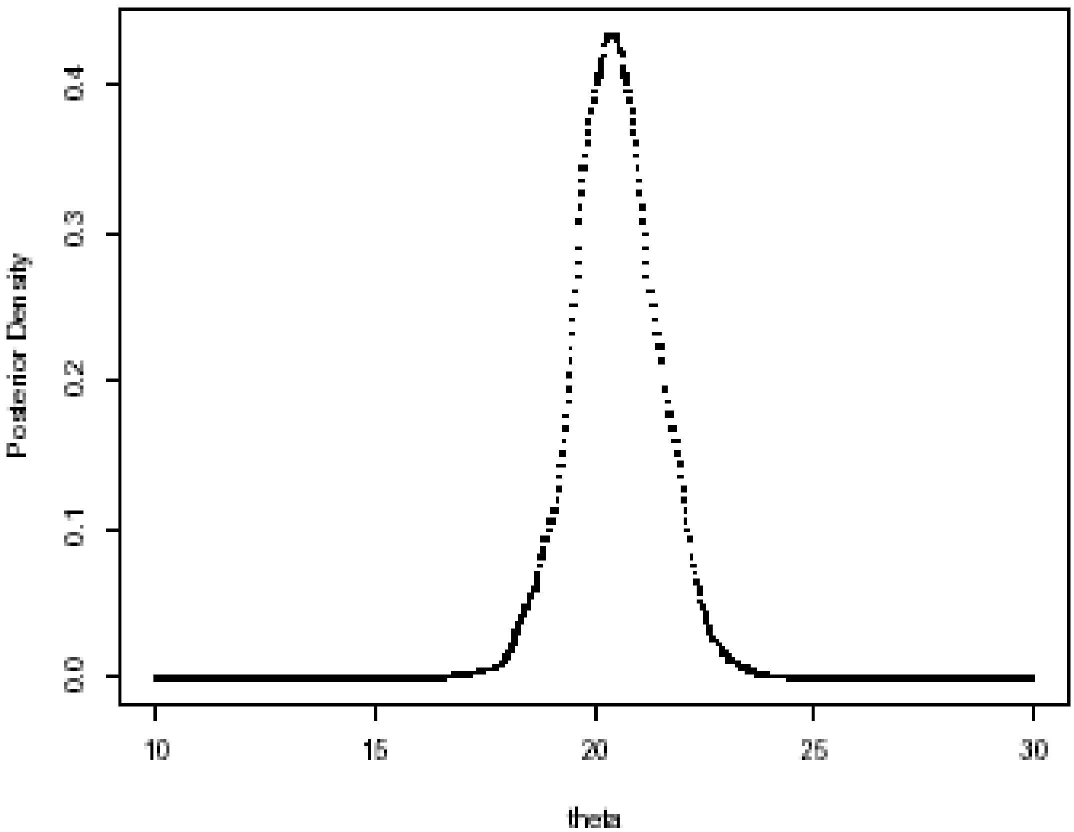
The posterior:

$$p(\theta | T(\mathbf{x}, \theta)) = \frac{\Sigma_{i=0}^{N} p_i I(w_{(i)} \leq \theta < w_{(i+1)})}{\Sigma_{i=0}^{N} p_i \int_{w_{(i)}}^{w_{(i+1)}} \pi(\theta) d\theta}$$

Again the posterior is a segmented version of $\pi(\theta)$ where now the segmentation is determined by the pairwise averages.

Same data: $n(20, 5^2)$ and prior $n(25, 100^2)$

Advantage: many more segments



	Bayesian Semiparametric Estimates			
	Mode	Mean	Median	95% C.S.
$\mathcal{N}(25, 100^2)$ prior	20.47	20.52	20.46	(18.47, 22.48)
$\mathcal{N}(18, 1)$ prior	19.69	19.29	19.33	(17.80, 20.65)

Recall $\overline{x} = 20.48$

General Scores:

$$T(\mathbf{X}, \theta) = \sum_{i=1}^{n} a[R_i(\theta)] I(X_i \le \theta)$$

where scores $a(i) = a_i$ are generated as $a(i) = \varphi[i/(n+1)]$ and $\varphi(u)$ is nondecreasing and square-integrable on $(0, 1)$.

Counting form:

$$T(\mathbf{X}, \theta) = \sum\sum_{i \le j} (a_{j-i+1} - a_{j-i}) I\left(\frac{X_{(i)} + X_{(j)}}{2} \le \theta \right)$$

- $ET(\mathbf{X}, \theta) = \frac{1}{2} \Sigma a_i$,
- $VarT(\mathbf{X}, \theta) = \frac{1}{4} \Sigma a_i^2$
- $T(\mathbf{X}, \theta)$ is approx normally distributed

Unlike the Wilcoxon statistic with integer support points, general scores typically have many more support points. There will be roughly 2^n such points.

The normal approximation works very well in this case since the distribution of the score statistic is symmetrically distributed.

Example: normal scores with $\varphi(u) = \Phi^{-1}(\frac{u+1}{2})$ where $\Phi(x)$ is the standard normal cdf.

Assume a $n(\mu_0, \sigma_0^2)$ prior. Use approx normality of the scores statistic to approximate the posterior.

$$p(\theta \mid T(\mathbf{x},\theta) \propto \exp\left\{-\frac{(T(\mathbf{x},\theta)-\frac{1}{2}\Sigma a_i)^2}{2\frac{1}{4}\Sigma a_i^2}\right\}\exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right.$$

We sample this posterior using the Metropolis algorithm.

Apply this to Wilcoxon for comparison:

	MCMC Estimates		
	Mean	Median	95% C.S.
$\mathcal{N}(25, 100^2)$ prior	20.48	20.45	(18.51, 22.45)
$\mathcal{N}(18, 1)$ prior	19.29	19.35	(17.72, 20.65)

And compare to original:

	Bayesian Semiparametric Estimates			
	Mode	Mean	Median	95% C.S.
$\mathcal{N}(25, 100^2)$ prior	20.47	20.52	20.46	(18.47, 22.48)
$\mathcal{N}(18, 1)$ prior	19.69	19.29	19.33	(17.80, 20.65)

A further approximation: linearization of the statistic

$$0 = \frac{T(\mathbf{X}, \widehat{\theta}) - \frac{n}{2}\int \varphi(u)du}{\sqrt{\frac{n}{4}\int \varphi^2(u)du}}$$

$$= \frac{T(\mathbf{X}, \theta) - \frac{n}{2}\int \varphi(u)du}{\sqrt{\frac{n}{4}\int \varphi^2(u)du}} + \tau\sqrt{n}\,(\widehat{\theta} - \theta) + o_p(1)$$

where

$$\tau^{-1} = \frac{\int \varphi(u)\varphi_f(u)du}{\sqrt{\frac{1}{4}\int \varphi^2(u)du}}$$

and

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

Wilcoxon: $\varphi(u) = u,\ \tau^{-1} = \sqrt{12}\int f^{\,2}(x)dx$

Then the posterior can be approximated by:

$$p(\theta \mid T(\mathbf{x},\theta) \propto \exp\left\{-\frac{n(\theta-\widehat{\theta})^2}{2\tau^2}\right\}\exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\}$$

With approx Bayes solution:

$$E(\Theta|\mathbf{x}) \simeq \{\frac{n\widehat{\theta}}{\tau^2} + \frac{1}{\sigma_0^2}\mu_0\}/\{\frac{n}{\tau^2} + \frac{1}{\sigma_0^2}\}$$

But now τ must be estimated.

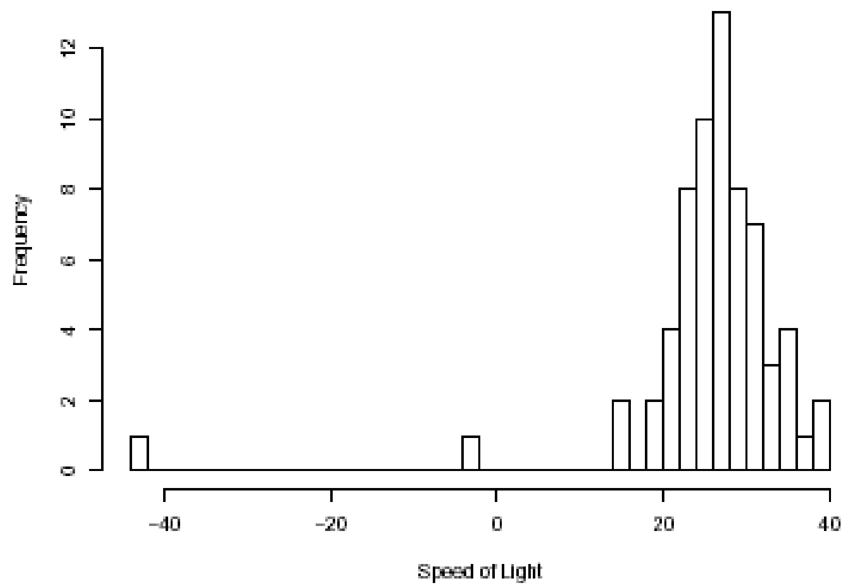
Wilcoxon case: estimate $\int f^2(x)dx$.

Some combination of:

- Normal prior,
- approximate normality of $T(\mathbf{X},\theta)$,
- linearization,
- MCMC methods

is used in extensions to regression.

Example: Simon Newcomb's speed of light data



66 measurements in 1882

Deviations from 24,800 nanoseconds

Normal prior $n(26, 100^2)$

1. Wilcoxon (assume symmetry)
2. Traditional Bayes with normal likelihood

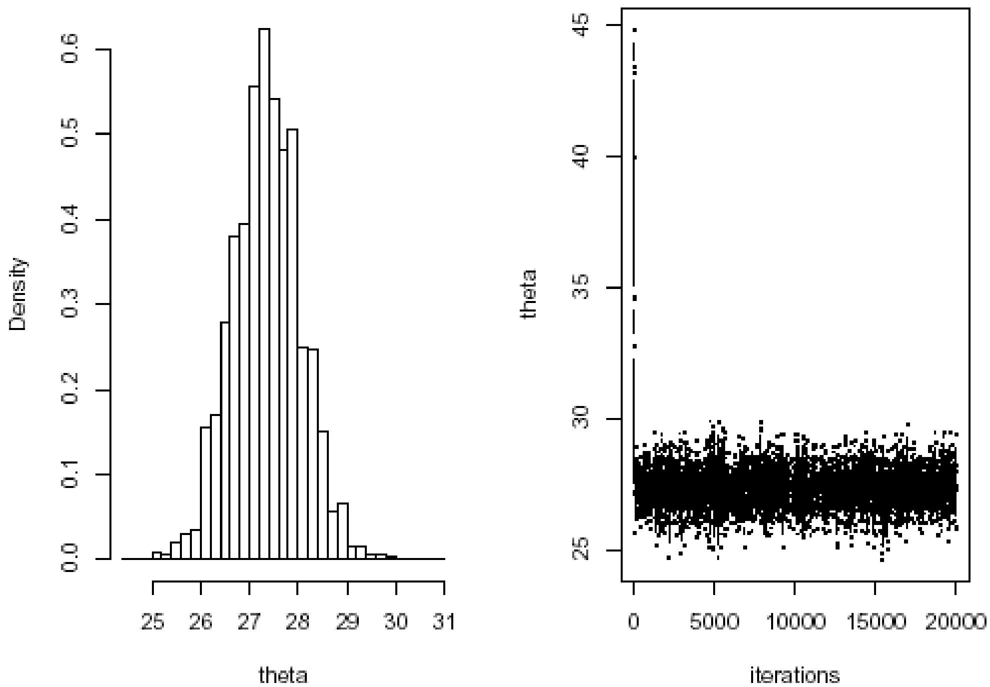


Figure 2.11: Bayesian semiparametric analysis with $T_W(\boldsymbol{X}, \theta)$ and MCMC for Example 2.2.8: Histogram and time series plot of one simulated sequence with the starting point $\theta^0 = 40$. The histogram is based on 15,000 (after discarding the first 5,000) iterations, while the time series plot records all the 20,000 simulations.

Methods	Estimate	95% C.S.	S.E.
Nonparametric (Wilcoxon)	27.5	(26.0, 28.5)	0.62
Bayesian Semiparametric MCMC (Wilcoxon)	27.4	(26.0, 28.7)	0.67
Bayesian Jeffreys' Prior	26.2	(23.6, 28.8)	1.34
Bayesian Normal-Gamma Prior	26.2	(23.6, 28.8)	1.32

1. Select a starting point θ^0, which may be a sample estimate of the location parameter θ, such as the sample median or mean. It could also be the prior mean μ_0 chosen in the presence of substantive prior knowledge about θ.

2. For $t = 1, 2, \ldots$:

   - Sample a candidate point θ^* from a jumping distribution at time t, $J_t(\theta^*|\theta^{t-1})$. The jumping distribution must be symmetric; namely, $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$ for all θ_a, θ_b and t. In our simulation, we use a normal distribution $\mathcal{N}(\theta^{t-1}, v^2)$, where the standard deviation v is specified so that a stable and converging sequence can be attained.

   - Calculate the ratio of the densities,

$$r = \frac{p(\theta^*|T_W(\boldsymbol{x}, \theta^*))}{p(\theta^{t-1}|T_W(\boldsymbol{x}, \theta^{t-1}))}$$

   - Set

$$\theta^t = \begin{cases} \theta^*, & \text{with probability } \min(r, 1); \\ \theta^{t-1}, & \text{otherwise.} \end{cases}$$

Testing: $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$

- Let $\pi_0 = P(H_0 \textit{ is true})$
- Suppose the mass on H_A is spread out according to the density $h(\theta)$.

The marginal distribution of $T(\mathbf{X}, \theta)$ is then

$$m(T(\mathbf{x})) = \pi_0 g(T(\mathbf{x}, \theta_0)|\theta_0) + (1 - \pi(\theta_0)) \int_{(\theta \neq \theta_0)} g(T(\mathbf{x}, \theta)|\theta) h(\theta)$$

Then the posterior probability

$$P(H_0 |\mathbf{x}) = \frac{\pi_0 g(T(\mathbf{x}, \theta_0)|\theta_0)}{m(T(\mathbf{x}))}$$

$$= \left[1 + \frac{(1 - \pi_0)}{\pi_0} \cdot \frac{m(T(\mathbf{x}))}{g(T(\mathbf{x}, \theta_0)|\theta_0)} \right]^{-1}$$

Example:

The sign statistic $T(\mathbf{x},\theta) = \Sigma I(x_i \leq \theta)$

$$P(H_0 \mid \mathbf{x}) = \frac{\pi_0 g(T(\mathbf{x},\theta_0)|\theta_0)}{m(T(\mathbf{x}))}$$

$$= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{\Sigma_{i=0}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \int_{x_{(i)}}^{x_{(i+1)}} \pi(\theta) d\theta}{\Sigma_{i=0}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n I(x_{(i)} \leq \theta_0 < x_{(i+1)})} \right.$$

Could take $\pi(\theta)$ to be $n(\theta_0, \sigma_0^2)$.

Summary:

1. Bayesian perspective can be incorporated in the semiparametric location model.

2. This can be combined with the usual nonparametric rank statistics.

3. Resulting Bayesian R-estimates are more robust than traditional Bayes estimates based on specific likelihoods.

4. For general scores we use a normal approximation to the T-likelihood and approximate the posterior distribution using MCMC methods.

5. This general approach can then be extended to regression models which include the two-sample location model as a special case.

6. Testing is also possible.