Nonparametric Regression using Bayesian Variable Selection

Michael Smith and Robert Kohn
Australian Graduate School of Management
University of New South Wales
Sydney 2052, Australia

First Version 15th June 1994

September 23, 1995

Summary

This paper estimates an additive model semiparametrically, while automatically selecting the significant independent variables and the appropriate power transformation of the dependent variable. The nonlinear variables are modeled as regression splines, with significant knots selected from a large number of candidate knots. The estimation is made robust by modeling the errors as a mixture of normals. A Bayesian approach is used to select the significant knots, the power transformation and to identify outliers using the Gibbs sampler to carry out the computation. Empirical evidence is given that the sampler works well on both simulated and real examples and that in the univariate case it compares favorably with a kernel weighted local linear smoother. The variable selection algorithm in the paper is substantially faster than previous Bayesian variable selection algorithms.

KEY WORDS: Additive model; Power transformation; Gibbs sampler; Regression spline; Robust estimation.

Classification: C11; C15; C22

1 Introduction

A Bayesian approach is used to estimate semiparametrically an additive regression model. Each nonlinear component is modeled as a regression spline using many knots, with a significant subset of the knots chosen by variable selection. Although the discussion is confined to an additive model, the approach extends in a straightforward way to a model with interactions. We note that our approach determines which independent variables enter the regression and so extends to nonparametric regression previous work on variable selection in linear regression models by Mitchell and Beauchamp (1988), George and McCulloch (1993, 1994), and Raftery, Madigan and Hoeting (1993). To help ensure that the regression model is additive and has Gaussian errors we allow the dependent variable to be transformed using a power transformation taking a discrete number of values. We also show how to make the estimation robust to outliers by modeling the errors as a mixture of normals.

In the univariate case, the Bayesian regression spline approach compares favorably with a kernel weighted local linear smoother proposed by Ruppert, Sheather, and Wand (1995); these authors estimate the global bandwidth parameter using direct plugin. In particular, the regression spline approach possesses greater local adaptability to the shape of the function than the local linear estimator which uses a global bandwidth estimate.

Because of the large number of variables involved, the computation is carried out using the Gibbs sampler with the error variance, the regression parameters and the index for the power transformation integrated out. Section 7 shows that our algorithm for Bayesian

[1] Please address all correspondence to Professor Robert Kohn, Australian Graduate School of Management, University of New South Wales, Sydney, NSW 2052, Australia. Tel.: (612) 931 9265, Fax: (612) 662 7621.

[2] Robert Kohn's work was partially supported by an Australian Research Grant. We thank Glen Barnett, Chris Carter, Ed George, Simon Sheather and Steve Marron for useful discussions. Comments by the referees and the associate editor helped improve both the presentation and content of the paper. Thanks also to Matt Wand for making his local linear kernel regression program available to us.

variable selection is considerably faster than previous approaches.

There are many approaches to nonparametric regression; Eubank (1988) gives a useful survey. Currently, the two most popular approaches to smoothing are smoothing splines and kernel based nonparametric regression. Smoothing splines for additive models, using generalized cross-validation to estimate the smoothing parameters, require, in general, $O(n^3)$ operations, where n is the sample size, making computation prohibitive for large samples sizes; e.g. Gu and Wahba (1991). Fast $O(n)$ methods for additive models using the backfitting algorithm are discussed by Hastie and Tibshirani (1990); however, these methods estimate the smoothing (or bandwidth) parameters in a more ad-hoc fashion with the estimates usually based on the values of the independent variables, but not the dependent variable. At present kernel based smoothing with the smoothing parameter(s) estimated by direct plugin seems confined to the univariate case. We also mention Friedman and Silverman (1989) who use regression splines for nonparametric regression and select the knots by a cross-validation procedure. This is computationally very intensive, making it difficult to traverse all possible knot combinations when seeking optimal knot placement. More generally, it seems difficult to make the above approaches to nonparametric regression robust in a practical way. Finally, we note that Bayesian work on robustifying linear regression by modeling the errors as a mixture of normals is discussed by many authors including Box and Tiao (1968) and Verdinelli and Wasserman (1991).

The paper is structured as follows. Section 2 describes variable selection for linear regression and explains how the Gibbs sampler is used to estimate the parameters involved. Section 3 presents our approach to nonparametric regression in the univariate case and empirically compares its performance to kernel weighted local linear smoothing. Section 4 generalizes the treatment in Section 2 to include transformation of the dependent variable as part of the Bayesian analysis. Section 5 shows how to make the estimation robust by modeling the errors as a mixture of normals. Section 6 deals with semiparametric additive regression. Section 7 compares our approach to variable selection with that of George and McCulloch (1993, 1994). Implementation details for variable selection are given in Appendix 1.

2 Variable selection in linear regression

2.1 Description of the model and prior

This section reviews variable selection in linear regression as it forms the basis of the nonparametric approach. Consider the linear regression model

$$y = X\beta + e \tag{2.1}$$

where y is the $n \times 1$ vector of observations, X is the $n \times r$ design matrix, $e \sim N(0, \sigma^2 I_n)$ is the error vector and $\beta = (\beta_1, \ldots, \beta_r)'$ is the $r \times 1$ vector of regression coefficients. Let γ be the $r \times 1$ vector of indicator variables with ith element γ_i such that $\gamma_i = 0$ if $\beta_i = 0$ and $\gamma_i = 1$ if $\beta_i \neq 0$. Given γ, let β_γ consist of all the nonzero elements of β and let X_γ be the columns of X corresponding to those elements of γ that are equal to one. We make the following prior assumptions:

1. Given γ and σ^2, the prior for β_γ is $\beta_\gamma \sim N\left(0, c\sigma^2(X_\gamma' X_\gamma)^{-1}\right)$, where c is a positive scale factor specified by the user. In all our empirical work we standardize the X matrix as in Dongarra et al. (1979) and take $c = 100$. In extensive testing we found that this choice of c works well and the results are insensitive to values of c in the range $10 \leq c \leq 1000$. The reason for choosing a large value of c, and in particular $c = 100$, is that for these values the prior of β_γ, given γ and σ^2, contains very little information about β_γ compared to the likelihood.

   It is not possible to make β diffuse by taking c infinite as then $p(\gamma_i = 1|y) = 0$ for all i; see equation (A.1) of appendix 1. We take the prior variance of β_γ proportional to $\sigma^2(X_\gamma' X_\gamma)^{-1}$ as it results in a fast computing algorithm for the Gibbs sampler which works well in practice. We also think that it is a sensible prior as it is proportional to the variance of the least squares estimate of β_γ.

2. The prior of σ^2 given γ is $p(\sigma^2|\gamma) \propto 1/\sigma^2$. This is a commonly used prior for σ^2 as it makes $\log(\sigma^2)$ uniform.

3. The γ_i are assumed to be apriori independent with $p(\gamma_i = 1) = \pi_i$, $0 \leq \pi_i \leq 1$, for $i = 1, \ldots, r$. In our applications we take the $\pi_i = \frac{1}{2}$ which represents no prior knowledge about whether a variable is included or excluded, and means that each model γ has a prior probability equal to 2^{-r}.

The above prior for $\beta_\gamma|\gamma, \sigma^2$ is related to Zellner's (1986) g-prior; Zellner takes $\beta|\sigma^2 \sim N\{\bar{\beta}, (\sigma^2/g)(X'X)^{-1}\}$, with $\bar{\beta}$ and g prescribed by the user. However, Zellner (1986) does not discuss point priors nor variable selection. George and McCulloch (1993) also consider, but do not emphasize, g-priors; they do not discuss point priors. Raftery et al. (1993) use point priors but not the g-prior structure.

2.2 The Gibbs Sampler

For a given γ, let $q_\gamma = \sum_{i=1}^{r} \gamma_i$ be the number of nonzero elements of β and

$$S(\gamma) = y'y - \frac{c}{1+c} y'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma y \,. \tag{2.2}$$

Then,

$$\begin{aligned} p(y|\gamma) &\propto \int_\sigma \left\{ \int_{\beta_\gamma} p(y|\beta_\gamma, \sigma^2) p(\beta_\gamma|\sigma^2) d\beta_\gamma \right\} p(\sigma^2) d\sigma^2 \\ &\propto (1+c)^{-q_\gamma/2} S(\gamma)^{-n/2}; \end{aligned} \tag{2.3}$$

β_γ is integrated out as a normal integral and σ^2 is integrated out as a inverse gamma integral. The posterior distribution of γ is

$$p(\gamma|y) \propto p(y|\gamma)p(\gamma) \propto (1+c)^{-q_\gamma/2} S(\gamma)^{-n/2} \prod_{i=1}^{r} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \,. \tag{2.4}$$

Because γ takes 2^r distinct values, it is impractical to obtain its posterior by direct enumeration unless r is small. In our applications r is large and we use the following Gibbs sampler (Gelfand and Smith, 1990) to traverse the parameter space.

Gibbs sampler (i) Choose an initial value $\gamma^{[0]} = (\gamma_1^{[0]}, \ldots, \gamma_r^{[0]})$ of γ perhaps by generating it from some distribution. (ii) Successively generate from $p(\gamma_i|y, \gamma_{j\neq i})$, $i = 1, \ldots, r$. Step (ii) is carried out many times and in two stages. The first stage is a warmup period at the end of which it is assumed that the sampler has converged to the joint distribution of $p(\gamma|y)$. The second stage is a sampling period and the γ_i generated during this period are used for inference.

From (2.4) the conditional probability of γ_i is

$$p(\gamma_i|y, \gamma_{j\neq i}) \propto p(y|\gamma)p(\gamma_i) \propto \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}(1+c)^{-q_\gamma/2}S(\gamma)^{-n/2}; \qquad (2.5)$$

because γ_i is binary, the conditional probability $p(\gamma_i|y, \gamma_{j\neq i})$ is obtained by evaluating (2.5) for $\gamma_i = 0$ and $\gamma_i = 1$ and normalizing.

The Gibbs sampler can be executed very efficiently because q_γ will usually be much smaller than r in our problems. Implementation details are given in Appendix 1.

2.3 Estimation

Two estimates of the regression parameters are discussed: the first is based on an estimate of the posterior mode of γ and the second on an estimate of the posterior mean of β. The estimates use the Gibbs iterates of γ during the sampling period, which we write as $\gamma^{[k]}, k = 1, \ldots, K$.

The posterior distribution $p(\gamma|y)$ has support on a parameter space of size 2^r making it difficult to locate its mode by direct enumeration when r is large. We estimate the mode of the posterior density by exploiting the fact that during the sampling period the Gibbs iterates $\gamma^{[k]}$ lie in regions of high probability. We take the the value of $\gamma^{[k]}, k = 1, \ldots, K$, maximizing $p(\gamma|y)$ as the estimate of the posterior mode of $p(\gamma|y)$ and write it as $\hat{\gamma}_M$. From (2.4) computing the posterior mode involves negligible extra computation over that required for the Gibbs sampler. The regression parameters are then estimated by least squares based on the model corresponding to $\hat{\gamma}_M$.

The second estimate of β is the mixture estimate of the posterior mean $E(\beta|y)$,

$$\hat{\beta} = K^{-1} \sum_{k=1}^{K} E\left(\beta|y, \gamma^{[k]}\right) .$$

This conditional expectation is evaluated exactly at each iteration as β is conditionally multivariate student t.

3 Univariate nonparametric regression

3.1 Regression Splines

Suppose

$$y_i = f(x_i) + e_i \quad i = 1, \ldots, n \tag{3.1}$$

where y_i is the ith observation, e_i is an independent $N(0, \sigma^2)$ error sequence, and $f(x)$ is a smooth function. We propose to approximate $f(x)$ by the cubic regression spline

$$b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \sum_{k=1}^{m} \beta_k (x - \tilde{x}_k)_+^3, \tag{3.2}$$

where $\tilde{x}_1, \ldots, \tilde{x}_m$ are the m 'knots' placed along the domain of the independent variable x, such that $\min(x_i) < \tilde{x}_1 < \ldots < \tilde{x}_m < \max(x_i)$, while $(z)_+ = \max(0, z)$. By replacing $f(x)$ in (3.1) by its approximation (3.2), the nonparametric regression can be rewritten as a linear regression. Let $r = m + 4$, $\beta = (b_0, b_1, b_2, b_3, \beta_1, \ldots, \beta_m)'$, let $\mathbf{x} = (x_1, \ldots, x_n)'$, let $\mathbf{1}$ be a $n \times 1$ vector of ones, and let the $n \times r$ matrix $X = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, (\mathbf{x} - \mathbf{1}\tilde{x}_1)_+^3, \ldots, (\mathbf{x} - \mathbf{1}\tilde{x}_m)_+^3)$. Then (3.1) can be written as the linear model (2.1) if $f(x)$ replaced by (3.2).

The most important question associated with fitting regression splines is the choice of both the number and location of the knots $\tilde{x}_1, \ldots, \tilde{x}_m$. If the knots are badly located, details of the curve can be missed, while if too many knots are included the fitted spline based on these knots will have high local variance. One way solve the problem is to introduce a large number of potential knots from which a significant subset can be selected, e.g. Friedman and Silverman (1989, pp. 9-11). The problem then becomes one of variable selection where each knot corresponds to a column of a design matrix from which a significant subset is to be determined. Although the number of knots introduced, m, will typically be large so that r will be large, the number of significant variables q required to obtain a good approximation will usually be quite small. This is what makes our algorithm so fast.

In the univariate case we usually place a knot after every three to five of the sorted values of the independent variable, up to a maximum of 40 knots. This results in the knots following the density of the independent variable and helps ensure that a knot is at, or near, the positions required to capture the curvature in the regression function.

3.2 Simulated Examples

The performance of the Bayesian nonparametric estimators is now studied using simulated data. The Bayesian estimators are also compared to a kernel weighted local linear estimator with its global bandwidth parameter estimated by direct plugin as in Ruppert, Sheather and Wand (1995). These authors analyze the large sample properties of their bandwidth estimator and show both theoretically and by simulation its excellent performance.

The following three regression functions are used in the simulation,

$$f_1(x) = 2x - 1, \quad f_2(x) = \sin(10\pi x), \quad f_3(x) = \phi(x, 0.15, 0.05)/4 + \phi(x, 0.6, 0.2)/4\,; \quad (3.3)$$

$\phi(x, \mu, \sigma)$ is the normal density, with mean μ and standard deviation σ, evaluated at x. These three functions are selected as they typify the range of functions met in univariate regression: f_1 is linear, f_2 is nonlinear but should be estimated well by a smoother using a global bandwidth, and f_3 seems to require a smoother with an adaptive bandwidth.

For each function, $n = 100$ observations were generated from (3.1) with the errors e_i independent $N(0, \sigma^2)$ and the knots x_i generated uniformly on $(0, 1)$. For each function, three levels of the error standard deviation are used: (i) low noise with σ equal to 1/8 the range of the function, (ii) medium noise with σ equal to 1/4 the range of the function, and (iii) high noise with σ equal to 1/2 the range of the function. The knots are chosen to follow the density of the independent variable, one every four observations giving a total of $m = 24$ knots and $r = 28$ columns in the design matrix X. The Gibbs sampler was run for a warmup period of 100 iterations and a sampling period of 1500 iterations, with arbitrary initial condition $\gamma^{[0]} = (1, 0, \ldots, 1, 0, 1)'$. Convergence, as indicated by $q^{[k]} = \sum_{i=1}^{k} \gamma_i^{[k]}$ and $\log p(\gamma^{[k]}|y)$, occurred within a dozen iterations for each of the three functions and all selected variables are significant at the 1% level as judged by their respective t-statistics.

The posterior mode estimate of the regression function is obtained by plugging the least squares fit of the the model $\hat{\gamma}_M$ into (3.2) and the posterior mean estimate of the regression function is obtained by plugging the posterior mean estimate of β into (3.2). Figure 1 shows the simulation output when the error standard deviation is 1/4 of the range of the function. The left three panels plot the data, the true function, and the Bayesian posterior mean and posterior mode estimates against the independent variable for all three functions. The right three panels are similar plots for the local linear regression estimator.

Both Bayesian estimators perform well for the linear function as it is in the basis of the regression spline; in particular, the Bayesian mode estimate suggests that the regression function is a straight line. The local linear regression estimator also performs well for f_1 and f_2 as they both require a global bandwidth estimator; its performance is comparable to that of the Bayesian estimators. For f_3, the Bayesian estimates are far smoother, possessing a degree of local adaptability that is difficult to obtain with a kernel smoother using a global bandwidth.

Figure 1 about here

Let $f(x)$ be the unknown regression function and $\hat{f}(x)$ its estimate. One numerical criterion for comparing nonparametric estimators is integrated squared error (ISE) which is the integral of $\{f(x) - \hat{f}(x)\}^2$ over the unit interval. We approximate this integral by taking a grid of 400 equally spaced points, $z_i = i/400, i = 1, \ldots, 400$, and compute the ISE as

$$ISE = \frac{1}{400} \sum_{i=1}^{400} \left(f(z_i) - \hat{f}(z_i) \right)^2 .$$

Table 1 gives the $\log(ISE)$ for the 3 estimators, together with the number of variables $\hat{q}$ used by the posterior mode estimate. We note that in all our calculations we use the logarithm to the base e.

function	r	Bayes mode	Bayes mean	kernel	$\hat{q}$
		$\log(ISE)$	$\log(ISE)$	$\log(ISE)$	
$f_1(x)$	28	-5.8585	-3.9733	-4.0416	2
$f_2(x)$	28	-3.0589	-2.6349	-2.5711	11
$f_3(x)$	28	-4.4664	-4.2775	-3.8409	5

Table 1: $\log(ISE)$ for the three estimators, and number of variables $\hat{q}$ selected by the posterior mode estimate for the results in Figure 1

The above simulations were repeated 100 times for each function and for the three noise levels and are summarized in Figure 2 by boxplots of $\log(ISE)$. Figure 2 confirms that the results in Figure 1 are typical and shows that for the examples considered both Bayesian estimators compare favorably with the local linear estimator. Figure 2 also shows that the relative performance of the local linear estimator improves as the error standard deviation increases. In general the Bayesian mode estimator performs better than the Bayesian mean estimator if the regression function is in the basis of the regression spline, as in the case of f_1.

4 Data transformation

4.1 Gibbs sampler and estimation

Sections 3 assumes an additive regression model with normally distributed homoscedastic errors. To achieve this Box and Cox (1964) advocate taking a power transformation y^λ of y. An important question when carrying out a Bayesian analysis on transformed data is the specification of the prior $p(\beta, \sigma^2, \gamma|\lambda)$; many researchers argue that this prior ought to depend on λ because the center and spread of the y_i^λ is different for each λ. However, the specification of this prior is more controversial; Box and Cox (1964) use a prior that depends on the observations y_i, while Sweeting (1985) gives a more recent discussion of the problem of choosing an appropriate prior. Our approach is different to that of Box and Cox (1964) and Sweeting (1985), but is related to the normalized power transformations discussed by Emerson and Stoto (1983). We consider transformations of the form $w_i(y;\lambda) = a(y;\lambda) + b(y;\lambda)y_i^\lambda$, with the scalars $a(y;\lambda)$ and $b(y;\lambda)$ chosen so that the median and the interquartile range of the $w_i(y;\lambda)$ are approximately the same for all λ. This makes it plausible to take $p(\beta, \sigma^2, \gamma|\lambda)$ independent of λ. We find our approach to transformations works well in practice, but do not compare it with other approaches.

Using the Gibbs sampler, λ is first estimated by the mode of a mixture estimate of $p(\lambda|y)$. The regression function is then estimated as in Section 2, conditioning on the estimated value of λ as advocated by Box and Cox (1982). Because of the high dependence between λ and γ in the posterior distribution, it is necessary to integrate out λ when generating the γ_i. We do so by allowing λ to take on just a small set of values denoted by Λ; in our examples we take $\Lambda = \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2\}$, which is adequate for most applications.

Our approach is now explained in more detail. We assume that the observations y_i are positive; if they are not then, as in Box and Cox (1964), a positive constant can be added to each of them. Let $y_{i,\lambda} = y_i^\lambda$ if $\lambda > 0$, $y_{i,\lambda} = \log(y_i)$ if $\lambda = 0$, and $y_{i,\lambda} = -y_i^\lambda$ if $\lambda < 0$. For $i = 1, \ldots, n$, let $w_i(y;\lambda) = a(y;\lambda) + b(y;\lambda)y_{i,\lambda}$ be a one-to-one transformation of y and let $y_{(i)}, y_{(i),\lambda}$, and $w_{(i)}(y;\lambda)$ be the ith smallest values of $y_i, y_{i,\lambda}$, and $w_i(y;\lambda)$. The values of $a(y;\lambda)$ and $b(y;\lambda)$ are determined as the solution to the pair of equations

$$y_{(n/2)} = w_{(n/2)}(y;\lambda) \quad \text{and} \quad y_{(3n/4)} - y_{(n/4)} = w_{(3n/4)}(y;\lambda) - w_{(n/4)}(y;\lambda)\,;$$

i.e.

$$b(y;\lambda) = \{y_{(3n/4)} - y_{(n/4)}\}/\{y_{(3n/4),\lambda} - y_{(n/4),\lambda}\} \quad \text{and} \quad a(y;\lambda) = y_{(n/2)} - b(y;\lambda)y_{(n/2),\lambda}\,.$$

We interpret $(n/2)$ to be the largest integer less than or equal to $n/2$, with $(n/4)$ and $(3n/4)$ interpreted similarly. Let $J(\lambda) = |\det\{\partial w(y;\lambda)/\partial y'\}|$ be the Jacobian of the transformation. We assume that $y_{(n/4)}, y_{(n/2)}$ and $y_{(3n/4)}$ correspond to distinct observations; if this is not the case then the y_i are perturbed by a tiny amount to make it so. It is then straightforward to evaluate $J(\lambda)$ as the matrix $\partial w(y;\lambda)/\partial y'$ is very sparse with only its diagonal and the three columns corresponding to $y_{(n/4)}$, $y_{(n/2)}$, and $y_{(3n/4)}$ not equal to zero.

It is important to note here that it is straightforward to choose the scalars $a(y;\lambda)$ and $b(y;\lambda)$ so the median and the interquartile range of the $w_i(y;\lambda)$ are exactly the same for all λ. However, the computation of the Jacobian is more complicated.

We now explain how to estimate the mode of $p(\lambda|y)$ and the regression curve. Instead of (2.1) we consider the model

$$w(y;\lambda) = X\beta + e$$

with the prior for β, σ^2 and γ as in Section 2 and with the elements of Λ having equal prior probability. To estimate the mode of $p(\lambda|y)$ we run the Gibbs sampler as in Section 2 by generating from $p(\gamma_i|y,\gamma_{j\neq i})$, $i = 1,\ldots,r$. To evaluate $p(\gamma|y)$ we note that

$$p(\gamma|y) = \sum_{\lambda\in\Lambda} p(\lambda,\gamma|y) \propto \sum_{\lambda\in\Lambda} p(y|\lambda,\gamma)p(\lambda)p(\gamma)$$

and $p(y|\lambda,\gamma) = p\{w(y;\lambda)|\lambda,\gamma\}J(\lambda)$, where $J(\lambda)$ is the Jacobian of the transformation. From (2.2) and (2.3),

$$p(y|\lambda,\gamma) \propto (1+c)^{-q_\gamma/2} S(\lambda,\gamma)^{-n/2} J(\lambda)\,,$$

where

$$S(\lambda,\gamma) = w(y;\lambda)'w(y;\lambda) - \frac{c}{1+c} w(y;\lambda)'X_\gamma \left(X_\gamma' X_\gamma\right)^{-1} X_\gamma' w(y;\lambda).$$

Let $\gamma^{[1]},\ldots,\gamma^{[K]}$ be the iterates of γ in the sampling period. Then

$$\hat{p}(\lambda|y) = \frac{1}{K}\sum_{k=1}^{K} p(\lambda|y,\gamma^{[k]})$$

is the mixture estimate of $p(\lambda|y)$; we take the $\lambda\in\Lambda$ that maximizes $\hat{p}(\lambda|y)$ as the estimate of the posterior mode of of $p(\lambda|y)$ and call it $\hat{\lambda}_M$. The probability $p(\lambda_i|y,\gamma)$ for $\lambda_i \in \Lambda$ is computed using the expression

$$p(\lambda_i|y,\gamma) = \frac{p(y|\lambda_i,\gamma)p(\lambda)}{\sum_{\lambda\in\Lambda} p(y|\lambda,\gamma)p(\lambda)}$$

Once $\hat{\lambda}_M$ is obtained the Bayesian nonparametric estimates of the regression function are obtained by running the Gibbs sampler again as in Section 2.

We find it necessary to integrate out λ when generating γ as the Gibbs sampler which generates $\gamma_i|y, \gamma_{j\neq i}, \lambda$, $i = 1, \ldots, r$ and $\lambda|y, \gamma$ tends to get stuck, because of the high correlation between the λ and γ iterates.

Remark 4.1 The transformations used by Emerson and Stoto (1983) are somewhat different than ours. They choose a transformation $T(y; \lambda)$ so that the medians are the same for all λ and the first derivative of $T(y; \lambda)$ with respect to y is unity at the median of the y_i.

4.2 Simulated examples

The performance of the above approach for simultaneously estimating the power transformation and the regression function is illustrated using the data discussed in the previous section and shown in Figure 1. For the data generated from f_1, the observations y_i were transformed to $\tilde{y}_i = \exp(y_i)$, for the data generated from f_2 the y_i were transformed to $\tilde{y}_i = (2.2 - y_i)^{-1/2}$, and for data generated from f_3 the observations are transformed to $\tilde{y}_i = (y_i + 1.5)^2$. Figure 3 plots the transformed data $\tilde{y}$ in the left panels. The data sets $\tilde{y}$ are used as the dependent variable in a regression and the location and scale corrected power transformations were applied as described above.

The correct back transformations $\tilde{y}_i \rightarrow a_1 + b_1 \log(\tilde{y}_i)$, $\tilde{y}_i \rightarrow a_2 - b_2(\tilde{y}_i)^{-2}$, and $\tilde{y}_i \rightarrow a_3 + b_3(\tilde{y}_i)^{1/2}$ were selected with posterior probabilities $0.999, 0.7666$ and 0.998 respectively. The right hand panels in Figure 3 show the back transformed data, along with the posterior mean fits. It is clear that, at least for these realizations, the nonparametric approach with data transformations performs very well. Further simulations indicated that this combined approach is highly effective.

Figure 3 about here

5 Robust nonparametric regression

5.1 Sampling scheme and estimation

We show how to make the Bayesian nonparametric estimates robust to outliers in the dependent variable by modeling the errors as a mixture of normals. For simplicity, data transformation is not discussed, but the results below are easily extended to include data transformations as in Section 4.

Consider the model (2.1) with $e \sim N(0, \sigma^2\Omega)$, where Ω is a diagonal matrix with ith diagonal element ω_i such that $\omega_i = 1$ if the ith observation is not an outlier and $\omega_i = \kappa$, with κ large and positive, if the ith observation is an outlier. Let $\omega = (\omega_1, \ldots, \omega_n)$; we will often write $\Omega = \Omega_\omega$ to indicate that it is a function of ω. The following prior assumptions are analogous to those in Section 2.1:

1. Given γ, ω, and σ^2, the prior for $\beta_\gamma \sim N\left(0, c\sigma^2(X_\gamma' \Omega_\omega^{-1} X_\gamma)^{-1}\right)$, with $c = 100$.
2. The prior for σ^2 given γ and ω, and the prior for γ given ω, are the same as in Section 2.1.
3. The ω_i are independent apriori with $p(\omega_i = \kappa) = \pi_e$. In the examples we take $\kappa = 100$ and $\pi_e = 0.05$. The value $\pi_e = 0.05$ reflects our belief that outliers are fairly rare but our method works well even if the actual percentage of outliers is substantially higher. Our empirical work shows that taking $\kappa = 100$ works well for outliers of varying sizes.

Because the posterior distribution of γ and ω has support on 2^{n+r} points it is infeasible to enumerate it directly for values of n and r that are moderate to large. Instead, the Gibbs sampler is again used for estimation. The Gibbs sampler is initialized by $(\gamma^{[0]}, \omega^{[0]})$, and γ and ω are then successively generated from the following two sets of conditionals distributions: (a) $p(\gamma_i | y, \gamma_{j \neq i}, \omega)$, $i = 1, \ldots, r$; (b) $p(\omega_i | y, \gamma, \omega_{j \neq i})$, $i = 1, \ldots, n$.

The conditional probabilities required for the Gibbs sampler are evaluated as follows. Let

q_γ be the number of nonzero elements of β, and let

$$S(\gamma,\omega) = y'\Omega_\omega^{-1}y - y'\Omega_\omega^{-1}X_\gamma \left(X_\gamma'\Omega_\omega^{-1}X_\gamma\right)^{-1} X_\gamma'\Omega_\omega^{-1}y\,.$$

Then, as in (2.3), $p(y|\gamma,\omega) \propto (1+c)^{-q_\gamma/2}S(\gamma,\omega)^{-n/2}$, so that

$$\begin{aligned}
p(\gamma,\omega|y) &\propto \left(\prod_{i=1}^{n}\omega_i\right)^{-\frac{1}{2}} \prod_{i=1}^{r}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}(1+c)^{-q_\gamma/2}S(\gamma,\omega)^{-n/2} && (5.1)\\
p(\gamma_i|y,\gamma_{j\neq i},\omega) &\propto (1+c)^{-q_\gamma/2}S(\gamma,\omega)^{-n/2}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} && (5.2)\\
p(\omega_i|y,\gamma,\omega_{j\neq i}) &\propto \omega_i^{-\frac{1}{2}}S(\gamma,\omega)^{-n/2}\,. && (5.3)
\end{aligned}$$

The probability $p(\gamma_i|y,\gamma_{j\neq i},\omega)$ is obtained by evaluating (5.2) for $\gamma_i = 0$ and $\gamma_i = 1$ and normalizing. The probability $p(\omega_i|y,\gamma,\omega_{j\neq i})$ is obtained similarly from (5.3).

Let $(\gamma^{[k]},\omega^{[k]}), k = 1,\ldots,K$, be the Gibbs iterates of (γ,ω) in the sampling period. Then, if $(\hat\gamma_M,\hat\omega_M)$ are the pair $(\gamma^{[k]},\omega^{[k]})$ that maximize $p(\gamma,\omega|y)$, with $p(\gamma,\omega|y)$ evaluated by (5.1), we take $(\hat\gamma_M,\hat\omega_M)$ as the estimate of the posterior mode of (γ,ω). Let $\hat\beta_M = \left(X_\gamma'\Omega_\omega^{-1}X_\gamma\right)^{-1} X_\gamma'\Omega_\omega^{-1}y$ be the generalized least squares of estimate of β_γ evaluated at $(\gamma,\omega) = (\hat\gamma_M,\hat\omega_M)$. We call the regression spline based on $\hat\beta_M$ the posterior mode estimate of the regression curve. The mixture estimate of the posterior mean of β is $K^{-1}\sum_{k=1}^{K}E(\beta|y,\gamma^{[k]},\omega^{[k]})$; we call the regression spline based on this estimate the posterior mean estimate of the regression curve. The conditional expectation is evaluated exactly for each k as β is conditionally distributed as a multivariate t.

5.2 Simulated examples

The performance of the robust nonparametric Bayesian estimators is illustrated empirically using three simulated examples. A comparison is also made with the estimates obtained using the (nonrobust) kernel weighted local linear estimator of Ruppert et al. (1995), and a local quadratic regression estimator called loess (Cleveland and Grosse 1991, Cleveland, Grosse and Shyu 1992) which is intended to be more robust. In implementing loess, care was taken to ensure that the iteratively reweighted least squares algorithm that underpins each robust local fit had ample opportunity to converge by increasing the number of iterations from the default of four to fifteen. However, we used the default smoothing parameter given in Splus as we are not aware of any data driven choice of bandwidth for loess. The results show that in the presence of large outliers the robust Bayesian estimators perform well, whereas the other two methods can be severely affected. Our numerical measure of performance is again integrated squared error, defined in Section 3.2.

We introduced ten extra outlying observations into the three univariate simulated data sets discussed in Section 3.2 and plotted in Figure 1. The ten values of the independent variable were generated uniformly on $(0, 1)$. For f_1, the extra observations were generated from a uniform distribution on $(-15, 15)$. For f_2, the extra observations were obtained by generating from a Cauchy distribution and multiplying the resulting variate by two. For f_3, the extra observations were generated from $N(0, 8^2)$. The outliers are deliberately generated from distributions that are different from the way they are modeled in order to provide a more challenging test of our robust approach.

The robust Bayesian estimator was applied without data transformation to each of the generated data sets with model selection made from 31 terms; one knot every four observations plus the four polynomial terms. The sampler was run for 2000 iterations to find the modal estimate of γ and ω and a further 500 iterations were run to calculate the mixture estimates. Convergence, as measured by the value of the posterior probability $p(\gamma, \omega|y)$ at the Gibbs iterates $\gamma^{[k]}, \omega^{[k]}$, the number of included terms, and the number of identified outliers, occurred within a handful of iterations for each function. Figure 4 presents the data and the result of the simulation for all three functions. The top three panels show the data; the circled observations are identified as outliers by the posterior mode estimate of ω. The bottom three panels show the true curve, the posterior mode and mean estimates, the local linear estimate, and the loess estimate. The corresponding values of ISE for the four estimates are given in Table 2. Figure 4 and Table 2 suggest that the two Bayesian estimators are robust and considerably outperform the other two estimators.

To confirm that the above comparisons are representative, each of the above simulations was repeated 100 times and boxplots of the $\log(ISE)$ are given in Figure 5 for each of the

	f_1	f_2	f_3
Bayesian modal	0.0043	0.0420	0.001
Bayesian mixture	0.0085	0.0838	0.012
kernel	0.3586	0.574	0.828
loess	0.3207	0.5817	0.309

Table 2: ISE of each estimator for the three sets of simulated data using the functions f_1, f_2 and f_3.

estimators and for each of the three functions. For example, Figure 5(a) gives boxplots for the posterior mode (MOD), posterior mean (MIX), local linear (KER2) and loess (LO2) estimators for the function f_1. To obtain a more absolute measure of the performance of the robust Bayesian estimators, the functions f_1, f_2, and f_3 were also estimated nonparametrically by the local linear estimator of Ruppert et al. and loess using only the 100 non-outlying observations. The resulting boxplot of the ISE for the Ruppert et al. estimator (KER1) and loess (LO1) are also given in Figures 5(a)–5(c). The boxplots indicate that the Bayesian estimates have good absolute performance as well as being more robust than the other two estimators. Qualitatively similar results were obtained for different sample sizes and signal to noise ratios.

Figures 4 and 5 about here

6 Additive semiparametric regression

6.1 Introduction

Because regression splines are linear models it is possible to use them to estimate an additive model by constructing a single design matrix made up of the columns of the individual design matrices of the type outlined in Section 3.1, but using only a single intercept. Variable selection can then be performed on the columns of the design matrix as outlined in Sections 2, 4 and 5. Section 6.1 illustrates the additive capability of the Bayesian approach by applying it to the Boston housing data discussed by Harrison and Rubinfield (1978) and Breiman and Friedman (1985), the latter using the ACE algorithm. The analysis of the Boston housing data shows that the Bayesian approach to outliers not only robustifies the analysis, but also identifies unusual points in the data. Section 6.2 uses a simulated example to illustrate how the Bayesian approach simultaneously transforms the data, selects the correct independent variables and robustifies against outliers.

6.2 Analysis of the Boston housing data

There are 506 observations on the dependent variable Y and thirteen independent variables, $X_1, \ldots, X_{13}$, which are described below.

Y =MV The median value of owner-occupied homes.

X_1 =CRIM per capita crime rate by town.

X_2 =ZN proportion of town's residential land zoned for lots greater than 25,000 square feet.

X_3 =INDUS proportion of nonretail business acres per town.

X_4 =CHAS Charles river dummy variable with value 1 if tract bounds the Charles river.

X_5 =NOX nitrogen oxide concentration (parts per hundred million)

X_6 =RM Average number of rooms.

X_7 =AGE proportion of owner-occupied units prior to 1940

X_8 =DIS logarithm of the weighted distances to five employment centers in the Boston region.

X_9 =RAD logarithm of the rate of accessibility to radial highways.

X_{10} =TAX full-value property tax rate (per $10,000.)

X_{11} =PTRATIO pupil teacher ratio by town

X_{12} =B=$(\text{Bk}-0.63)^2$ where Bk is the proportion of blacks in the population.

X_{13} =STAT proportion of the population that is lower status.

Looking at plots of the data, residuals of a least squares fit and previous analysis we concluded that X_5, X_6, X_8, X_{10} and X_{13} were the most important variables to model as non-linear. Thus, the model used is

$$\begin{aligned} w(Y;\lambda) &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + f_5(X_5) + f_6(X_6) + \beta_7 X_7 + f_8(X_8) + \beta_9 X_9 + \\ &\quad f_{10}(X_{10}) + \beta_{11} X_{11} + \beta_{12} X_{12} + f_{13}(X_{13}) + e\,; \end{aligned}$$

the functions f_i, for $i = 5, 6, 8, 10, 13$, are modeled using regression splines. Along the domain of these five variables, 8,6,8,5 and 8 knots, respectively, were placed so that they followed the density of the observations. It was found that additional knots tended to produce a matrix $X_\gamma' \Omega^{-1} X_\gamma$ which is almost singular, and therefore difficult to invert. Nevertheless, this number of knots seems sufficient to ensure that a knot is at, or near, all important locations required to capture the non-linear nature of the functions. Considering the three polynomial terms also included for each of these five variables, the other eight linear terms (β_i, $i = 1, 2, 3, 4, 7, 9, 11, 12$), and the intercept α, resulting in a design matrix with $r = 59$ columns.

The Gibbs sampler was applied to estimate this semiparametric additive model with a warmup period of three hundred iterations; two hundred iterations were then used to find the posterior mode estimate of λ, which was the transformation $y \rightarrow a + by^{1/2}$. Convergence was extremely rapid from any initial point. This is confirmed in Figure 6 by plots of the log posterior probabilities $\log p(\gamma^{[k]}, \omega^{[k]}|y)$, the number of included variables $q^{[k]} = \sum_{i=1}^{r} \gamma_i^{[k]}$, and the number of outliers at each iteration during the warmup period; these plots stabilize after 10 to 20 iterations. Similar plots were produced for all examples in this paper, providing strong empirical evidence for the rapid convergence of the Gibbs sampler.

Figure 6 about here

Because the results are very similar to those obtained on the original scale we choose, for simplicity, to present them on the original scale and use only the posterior mode estimate of γ and ω. The sampler was again run for a warmup period of 300 iterations, with a sampling period of 3,000 iterations to compute the posterior mode estimate of γ and ω. Of the $r = 59$ columns of the design matrix, 28 were selected by $\hat{\gamma}_M$. Of the linear terms including the intercept, only β_2, β_3 and β_4 were not selected. The regression splines f_5, f_6, f_8, f_{10} and

f_{13} based on $\hat{\gamma}_M$ include 5,5,4,4 and 4 terms respectively, the plots of which are given in Figures 7(a)–7(e). The function estimates largely correspond to those given by Breiman and Friedman's ACE procedure (1985, pp. 484-487).

Outliers are identified in Figure 7 by either a cross or a circle. Analysis of the outlying observations reveals an interesting profile for ten of the sixteen data points–numbers 367, 369-373, 377, 392, 408 and 410–which are indicated in Figure 7 by a cross. Figure 7(d) shows these ten observations are subject to a high property tax rate, Figure 7(c) shows the same ten are very close to five employment centers, Figure 7(e) suggests that these properties are located in high status neighborhoods, and Figure 7(b) shows that these observations do not necessarily represent large properties. Looking at the original census tracts (Belsley, Kuh and Welsch, 1980, p.230) these ten properties are from adjacent districts of central Boston (Back Bay, Beacon Hill, Charlestown, East Boston, Downtown and Roxbury) and appear to possess all the features of inner city executive apartments. The model was also estimated non-robustly, with the component estimates similar to the robust case except for the estimate of $f_8(X_8)$, shown in Figure 7(f), which is severely affected by outliers for low values of X_8.

Figure 7 about here

The linear least squares regression of $\Omega_\omega^{-1/2}y$ on $\Omega_\omega^{-1/2}X_\gamma$, with $(\gamma,\omega) = (\hat{\gamma}_M, \hat{\omega}_M)$, produced well behaved residuals with an $R^2 = 0.94$, compared with an $R^2 = 0.895$ for the non-robust Bayesian fit. A least squares fit using the linear regression model in Harrison and Rubinfeld (1978) gave an $R^2 = 0.73$, while the robust linear estimate of Harrison and Rubinfeld (1978) has a (weighted) $R^2 = 0.80$. The ACE analysis given by Breiman and Friedman (1985) produced an $R^2 = 0.89$. This comparison of R^2 values for the various methods is only a rough measure of their relative performance; it does not take into account that the Bayesian estimates condition on $\hat{\gamma}_M$ and $\hat{\omega}_M$, and that ACE seeks a transformation that maximizes R^2.

6.3 Simulated Example

This example shows how robust Bayesian regression performs in determining an additive nonparametric model in the presence of a data transformation. Two hundred observations of four independent variables, $X_1, \ldots, X_4$, were generated from uniform distributions on the interval [0,1]. Two hundred observations were then generated from the model

$$y = \alpha + f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + e,$$

where $\alpha = 7.5$, and the errors e are independent $N(0, 0.75^2)$. The function $f_1(x) = 0$ if $x < 1/4$, and $f_1(x) = 1.5\sin(2\pi(1.25 - x)^2)$, if $1/4 \leq x \leq 1$. The function $f_2(x) = 0$ if $0 < x < 1/4$ and $f_2(x) = 1.5\cos(4\pi(x - 0.25)) - 1.5$ if $1/4 < x < 3/4$. The function $f_3(x) = -2.5x$ and the function $f_4(x)$ is null. Twenty outliers were added to the simulated data, and the dependent variable of the resulting data then transformed $y \to \tilde{y} = 1/y^2$. We use $\tilde{y}$ as the dependent variable in the regression and estimate the appropriate data transformation as in Section 4. The data is plotted on its original scale against the four independent variables in Figures 8(a)–8(d), the circled observations being the twenty outliers. Plots of the transformed data, which are not included here, show that it is difficult to determine either the outlying observations, or the underlying functional forms, on the wrong scale.

The full Bayesian procedure was applied to robustly estimate the model. Each of the four functions, $f_1, \ldots, f_4$ was modeled using a regression spline with 14 knots (one every fifteen observations.) Along with the 12 associated polynomial terms and the intercept α, this resulted in robust model selection being performed from a total of 69 terms. We use a smaller number of knots in the multivariate case to avoid numerical instabilities in the design matrix. The sampler was first run to estimate the mode of $p(\lambda|y)$; the correct power transformation, $\hat{\lambda}_M = -1/2$, being selected. The back transformed data $a_j + b_j(\tilde{y})^{-1/2}, j = 1, 2, 3, 4$, are plotted in Figures 8(a)–(d). The sampler was then rerun with $\lambda = -1/2$ to obtain the posterior mode and the posterior mean function estimates. The posterior mode estimate includes 12 of the 69 variables with all twenty outliers, and no other observations, being correctly identified; these are the circled observations in Figures 8(a)–(d).

Figures 8(e)-(g) plot the posterior mode and the posterior mean estimates for the functions f_1, f_2 and f_3. Also included are plots of the true functional forms on the appropriate scale; that is, $a_j + b_j f_j(x), j = 1, 2, 3$. The function f_4 is not estimated as none of the terms associated with its estimate is included in $\hat{\gamma}_M$; the variable X_4 is therefore correctly rejected from the model. Figure 8(h) is a normal probability plot of the residuals when $\Omega_\omega^{-1/2} y$ is regressed on $\Omega_\omega^{-1/2} X_\gamma$, where (γ, ω) is set to the posterior mode estimate $(\hat{\gamma}_M, \hat{\omega}_M)$; the plot suggests that the residuals are normally distributed.

Figure 8 about here

7 Discussion of related work

Differences in approach to Bayesian model selection revolve primarily around the specification of the conditional prior $\beta | \gamma, \sigma^2$, because it introduces the indicator variables into the model. Mitchell and Beauchamp (1988, p.1024) use a point null prior, $\beta_i = 0$, for $\gamma_i = 0$, and a uniform prior for $\gamma_i = 1$, i.e. $\beta | \gamma, \sigma^2 \sim \text{Uniform}(-a_i, a_i)$, with a_i large for each i. The decision of how large to choose the values of a_i is left to the user.

George and McCulloch (1993) use the nonconjugate normal prior $\beta_i | \gamma, \sigma^2 \sim N(0, \tau_i^2)$ if $\gamma_i = 0$ and $\beta_i | \gamma, \sigma^2 \sim N(0, c_i^2 \tau_i^2)$ if $\gamma_i = 1$. The constants τ_i and c_i are chosen so that τ_i is small and c_i is large. George and McCulloch (1993) make some suggestions on suitable choices for c_i and τ_i and use the following Gibbs sampler to generate models of high probability: Generate from (a) $p(\beta | y, \sigma^2, \gamma)$; (b) $p(\sigma^2 | y, \beta, \gamma)$; (c) $p(\gamma_i | y, \beta, \sigma^2, \gamma_{j \neq i})$ for $i = 1, \ldots, r$. We have found this sampler difficult to implement for our problems because of the high correlation between β and γ. If τ_i is chosen too small then the sampler is nearly degenerate and tends to get stuck. If τ_i is chosen too large, significant terms are omitted and high local bias is experienced. We note that this sampler requires $O(r^3)$ operations to generate β which can be considerably slower than our algorithm if q is much smaller than r.

George and McCulloch (1994) consider the conjugate prior $\beta_i | \gamma, \sigma^2 \sim N(0, \sigma^2 \tau_i^2)$ if $\gamma_i = 0$ and $\beta_i | \gamma, \sigma^2 \sim N(0, \sigma^2 c_i^2 \tau_i^2)$ if $\gamma_i = 1$ and obtain $p(\gamma | y)$ by integrating out β and σ^2. The parameters τ_i and c_i are again specified by the user. We call the prior $\beta_i = 0$ a point null, and the prior $\beta_i|\gamma, \sigma^2 \sim N(0, \sigma^2\tau_i^2)$, $\tau_i > 0$, a non-point null. A Gibbs sampler based on the non-point null prior can be considerably slower than that based on the point null; the ratio of speeds is $O(r^2/q^2)$. The difference in speed between the two priors is even more pronounced in the robust case, with the ratio of speeds being $O(r^3/q^3)$ because the Cholesky decomposition of the matrix $X_\gamma' \Omega_\omega^{-1} X_\gamma$ is recalculated when generating each ω_i; see Appendix 1 for details. Raftery et al. (1993) also use a point null prior, but do not take advantage of it in their computations.

The motivation for the non-point null prior in George and McCulloch (1994) is that if $\beta_i \leq 2\sigma\tau_i$, then β_i is 'small enough' and the corresponding variable can be dropped from the regression. If information on what is 'small enough' is available , then the prior in George and McCulloch (1994) is appropriate. In the applications considered in the present paper, and in many other variable selection problems, such information is usually unavailable. To assess the difference in speed between the algorithm based on the point null prior and the non-point null prior, we report the time to complete 1000 iterations for both the non-robust and robust cases. For the nonrobust case we use the data in Figure 1, which consists of $n = 100$ observations generated from the model (3.1) using each of the regression functions (3.3) and with the error standard deviation equal to 1/4 the range of the function. For the robust case we use the corresponding data in Figure 2.

Table 3 reports the time (in seconds) and the average number of variables included ($\bar{q}$) when the point null prior is used. The table shows that using the point null can be substantially faster than using the non-point null. In most of the examples we have examined 2000 iterations appear to be sufficient to obtain a good estimate of the posterior mode and 300 iterations are sufficient to obtain a good estimate of the posterior mean.

	non-robust case				robust case			
function	r	$\bar{q}$	time (secs) point null	time (secs) non-point null	r	$\bar{q}$	time (secs) point null	time (secs) non-point null
f_1	28	5.34	2.9	36.3	31	7.05	44	650
f_2	28	12.75	8.33	36.3	31	13.74	120	650
f_3	28	7.58	4.17	36.3	31	7.79	50	650

Table 3: The average number of variables used ($\bar{q}$) and the time (in seconds) per 1000 iterations for point and non-point null priors

Appendix 1 Implementation of the Gibbs sampler

We outline how to efficiently implement the Gibbs sampler described in Section 2; extensions to the data transformation and robustness are then briefly discussed. Before running the Gibbs sampler the terms $y'y, X'y$ and $X'X$ are computed. Because γ_i takes on only two values, it is not difficult to show that $p(\gamma_i = 1|y, \gamma_{j\neq i}) = 1/(1+h)$, where

$$h = \frac{1-\pi_i}{\pi_i}(c+1)^{\frac{1}{2}}\left(\frac{S(\gamma^1)}{S(\gamma^0)}\right)^{\frac{n}{2}}, \tag{A.1}$$

$\gamma^1 = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \ldots, \gamma_r)$ and $\gamma^0 = (\gamma_1, \ldots, \gamma_{i-1},\ \gamma_i = 0, \gamma_{i+1}, \ldots, \gamma_r)$. Suppose that $\gamma = \gamma^0$ before γ_i is generated. Then $S(\gamma^0)$ is known and it is only necessary to obtain $S(\gamma^1)$. The main computational difficulty in obtaining $S(\gamma^1)$ is evaluating $y'X_{\gamma^1}\left(X'_{\gamma^1}X_{\gamma^1}\right)^{-1}X'_{\gamma^1}y$. This is done by factoring $X'_{\gamma^1}X_{\gamma^1}$ as $L_1L'_1$, where L_1 is lower triangular, using the Cholesky decomposition and then computing $L_1^{-1}X'_{\gamma^1}y$. We note that $X'_{\gamma^1}X_{\gamma^1}$ and $X'_{\gamma^0}X_{\gamma^0}$ differ by only one row and column so that L_1 can be readily obtained from L_0, where $L_0L'_0$ is the Cholesky decomposition of $X'_{\gamma^0}X_{\gamma^0}$; see Dongarra, Moler, Bunch and Stewart (1979, Ch. 10). If $\gamma = \gamma^1$ before γ_i is generated, then L_0 can similarly be obtained from L_1. From Dongarra et al. (1979), generating γ_i requires $q^2_{\gamma^1}$ operations. Hence generating γ requires $O(rq^2)$ operations, where q is the typical number of regressors required. We refer the reader to Dongarra et al. (1979) for a discussion of fast and stable methods for updating a Cholesky decomposition.

When the dependent is transformed, the terms $w(y;\lambda)'w(y;\lambda)$ and $X'w(y;\lambda)$ are obtained for each value of $\lambda \in \Lambda$, as well as $X'X$. Fast calculation of $S(\lambda,\gamma)$ is done as above.

In the robust case, to generate each $\omega_i, i = 1,\ldots,n$, requires updating $y'\Omega_\omega^{-1}y, y'\Omega_\omega^{-1}X_\gamma$, and $X_\gamma'\Omega_\omega^{-1}X_\gamma$, and recalculating the Cholesky decomposition of the matrix $X_\gamma'\Omega_\omega^{-1}X_\gamma$. Given the Cholesky decomposition of the matrix $X_\gamma'\Omega_\omega^{-1}X_\gamma$, the indicator vector γ is generated efficiently as outlined above.

References

Belsley D. A., E., Kuh, and R. Welsch, 1980, Regression Diagnostics. New York: Wiley.

Box, G.E.P. and D.R. Cox, 1964, An analysis of transformations. Journal of the Royal Statistical Society, Ser. B 26, 211-252.

Box, G.E.P. and D.R. Cox , 1982, An analysis of transformations revisited, rebutted. Journal of the American Statistical Association 77, 209-210.

Box, G. E. P., and G.C. Tiao , 1968, A Bayesian approach to some outlier problems, Biometrika 55, 119-129.

Breiman, L., and J.H. Friedman , 1985, Estimating optimal transformations for multiple regression and Correlation. Journal of the American Statistical Association 80, 580-598

Cleveland, W.S. and E. Grosse , 1991, Computational methods for local regression. Statistics and Computing 1, 47-62

Cleveland, W.S., E. Grosse, and W.M. Shyu , 1992, Local regression models, in Chambers, J.M. and Hastie, T. J. , Statistical Models in S, eds. Chambers, J.M. and Hastie, T. J., Pacific Grove, California: Wadsworth, pp. 309-373.

Dongarra, J. J., C.B. Moler, J.R. Bunch and G.W. Stewart, 1979, Linpack Users' Guide, Philadelphia: Siam.

Emerson, J.D. and M.A. Stoto , 1983, Transforming data. In 'Understanding robust and exploratory data analysis', 97-128, Eds. Hoaglin, D.C., F. Mosteller, and J.W. Tukey, New York: John Wiley and Sons.

Eubank, R.L., 1988, Spline smoothing and nonparametric regression. New York, NY: Marcel Dekker.

Friedman J. H. and B.W. Silverman , 1989, Flexible parsimonious smoothing and additive modeling. Technometrics 31, 3-39.

Gelfand, A. E., and Smith, A. F. M., 1990, Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398-409.

George, E. I. and R.E. McCulloch , 1993, Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881-889

George, E. I. and R.E. McCulloch , 1994, Fast Bayes Variable Selection. Preprint.

Gu, C. and G. Wahba , 1991, Minimising GCV/GML scores with multiple smoothing parameters via the Newton method. SIAM Journal of Scientific and Statistical Computing 12, 383-398.

Harrison, D. and D.L. Rubinfeld , 1978, Hedonic housing prices and demand for clean air. Journal of Environmental Economics and Management 5, 81-102

Hastie, T.J. and R.J. Tibshirani , 1990, Generalized additive models. New York: Chapman Hall.

Mitchell, T. J. and J.J. Beauchamp , 1988, Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023-1036

Raftery, A., D. Madigan, and J. Hoeting , 1993, Model selection and accounting for model uncertainty in linear regression models. Working paper, University of Washington.

Ruppert D., Sheather S.J., and Wand M. P., 1995, An effective bandwidth selector for local least squares regression. To appear in the Journal of the American Statistical Association.

Sweeting, T.J., 1985, Consistent prior distributions for transformed models. In 'Bayesian Statistics 2', 755-762, Eds. Bernardo, J.M., D.V. Lindely, and A.F.M. Smith. Amsterdam; Elsevier Science Publishers B.V. (North-Holland).

Verdinelli, I. and L. Wasserman , 1991, Bayesian analysis of outlier problems using the Gibbs sampler. Statistics and Computing 1, 105-117.

Zellner, A., 1986, On assessing prior distributions and Bayesian regression analysis with g-prior distributions, In 'Bayesian Inference and Decision Techniques-Essays in Honor of Bruno de Finetti', 233-243, Eds. P.K. Goel and A. Zellner, Amsterdam: North-Holland.

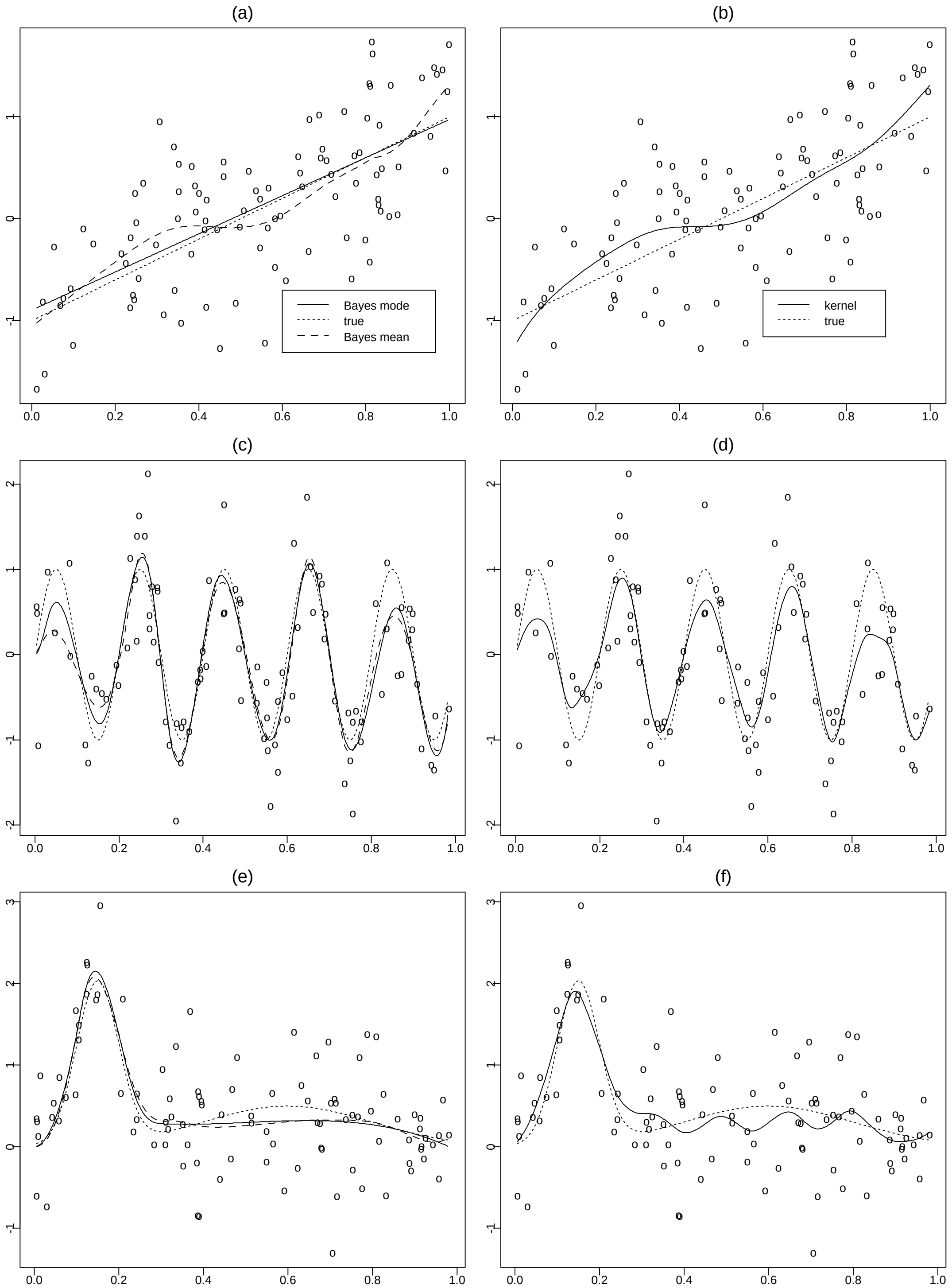

Figure 1: Panels (a),(c), and (e): Bayesian modal estimate (solid line), Bayesian mean estimate (dashed line), the true regression curve (dotted line), and the data (scatter plot) for each of the three functions. Panels (b), (d), and (f): Local linear estimate (solid line), true regression function (dotted line), and the data (scatter plot).

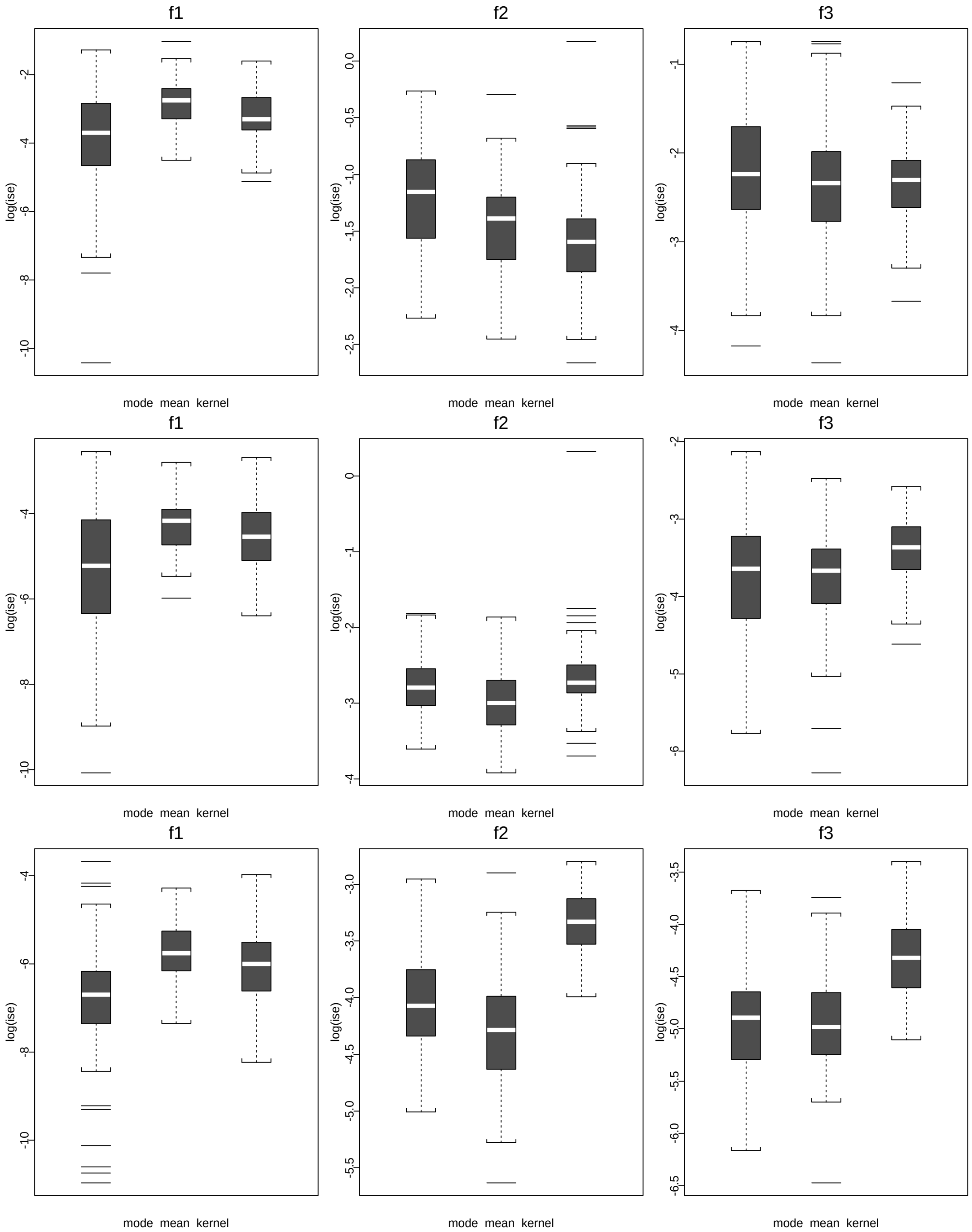

Figure 2: Boxplots of $\log ISE$. The left panels refer to f_1, the middle panels to f_2 and the right panels to f_3. The bottom 3 panels refer to the low noise case, the middle three panels to the medium noise case, and the top three panels to the high noise case. In each panel, the the left, middle, and right boxplots refer to the posterior mode estimate, posterior mean estimate, and the local linear estimate.

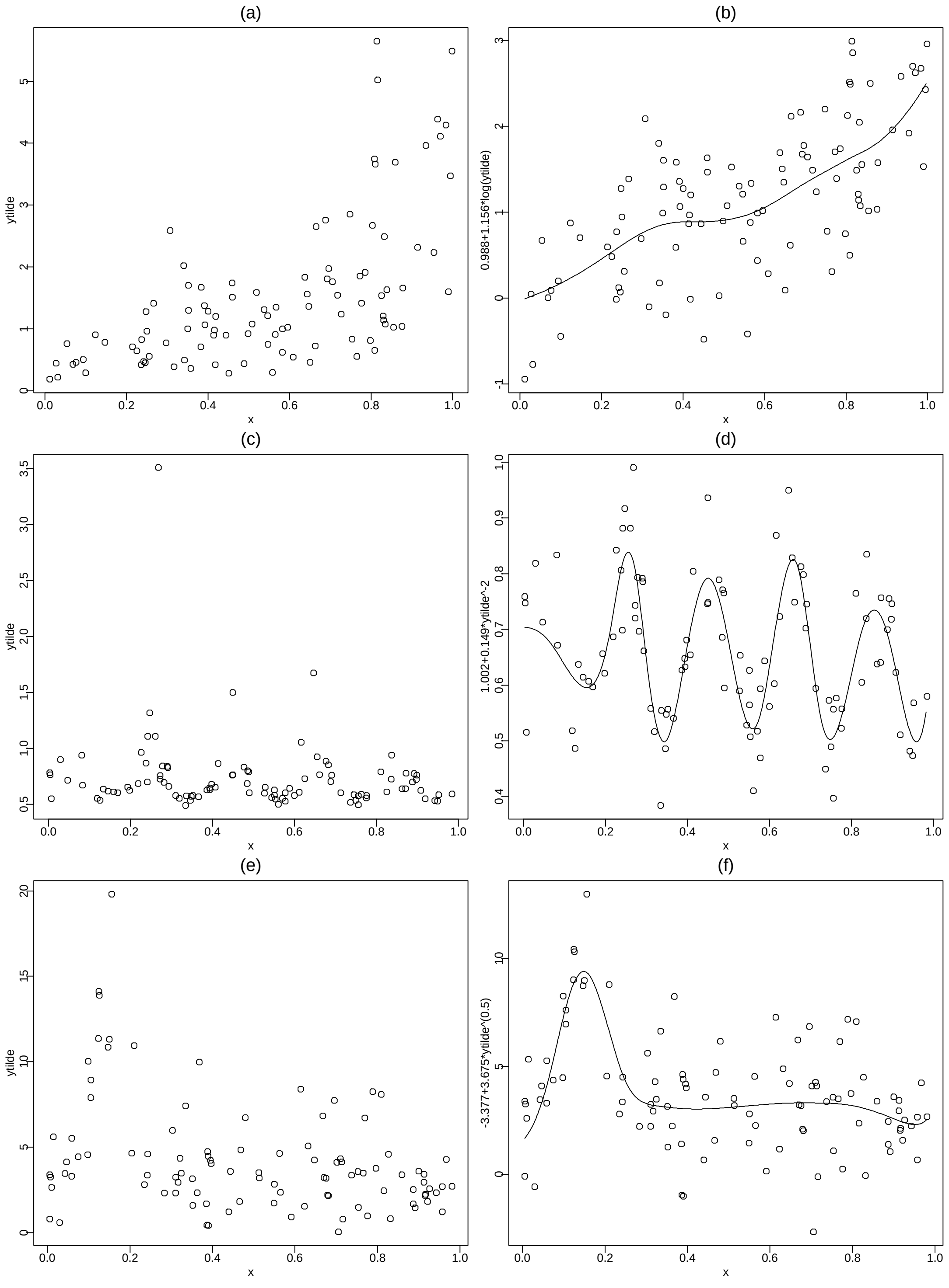

Figure 3: Panels (a), (c), and (e) are plots of the transformed data; panels (b), (d), and (f) are plots of the back transformed data (scatter plot) and the Bayesian mean estimate of the regression curve (solid line).

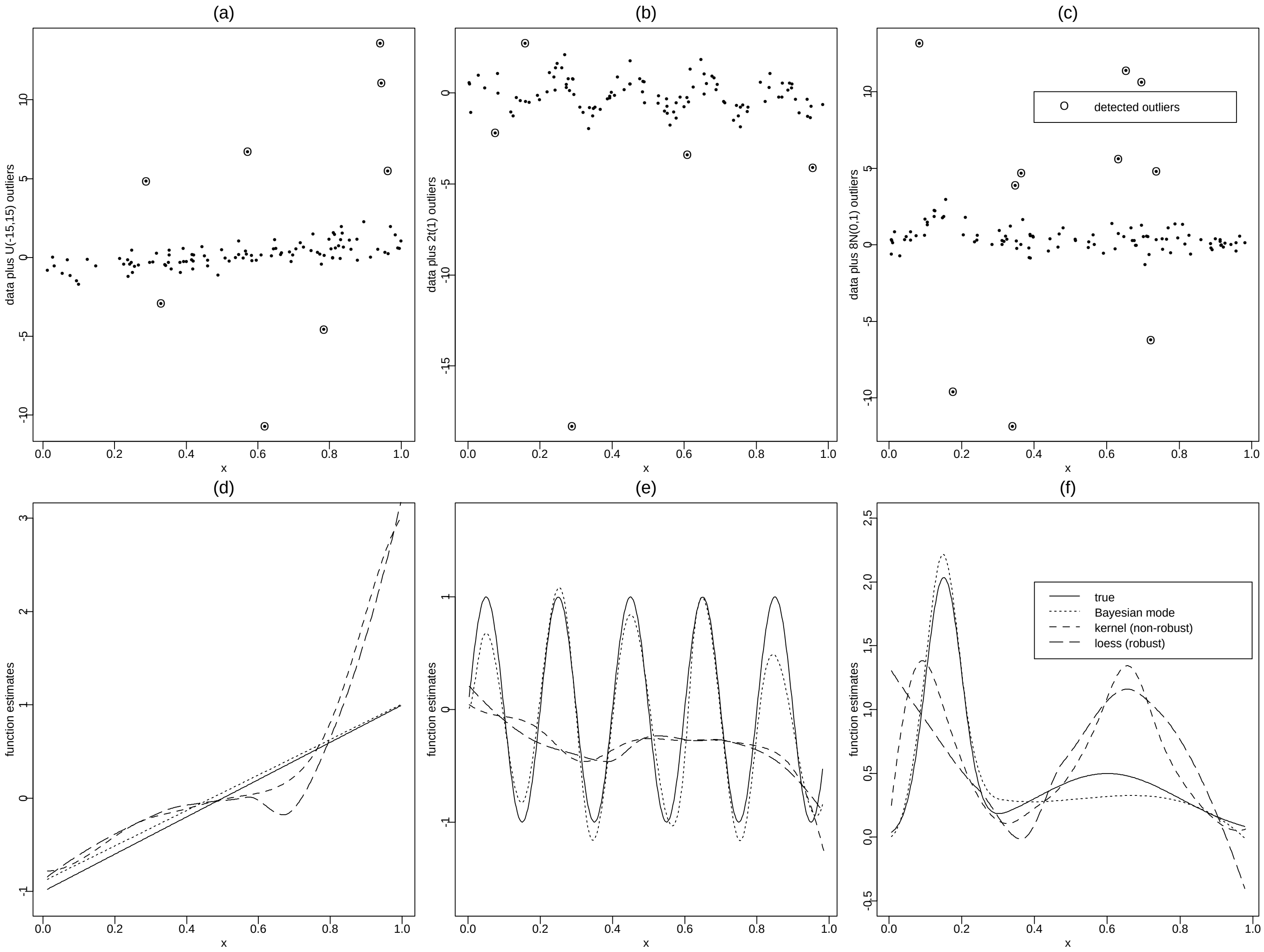

Figure 4: Top three panels are scatter plots of the data with the circled points identified as outliers by the Bayesian mode estimator; bottom three panels are plots of the true regression curve (solid line), Bayesian mode (dotted line), local linear (short dashes), and loess (long dashes).

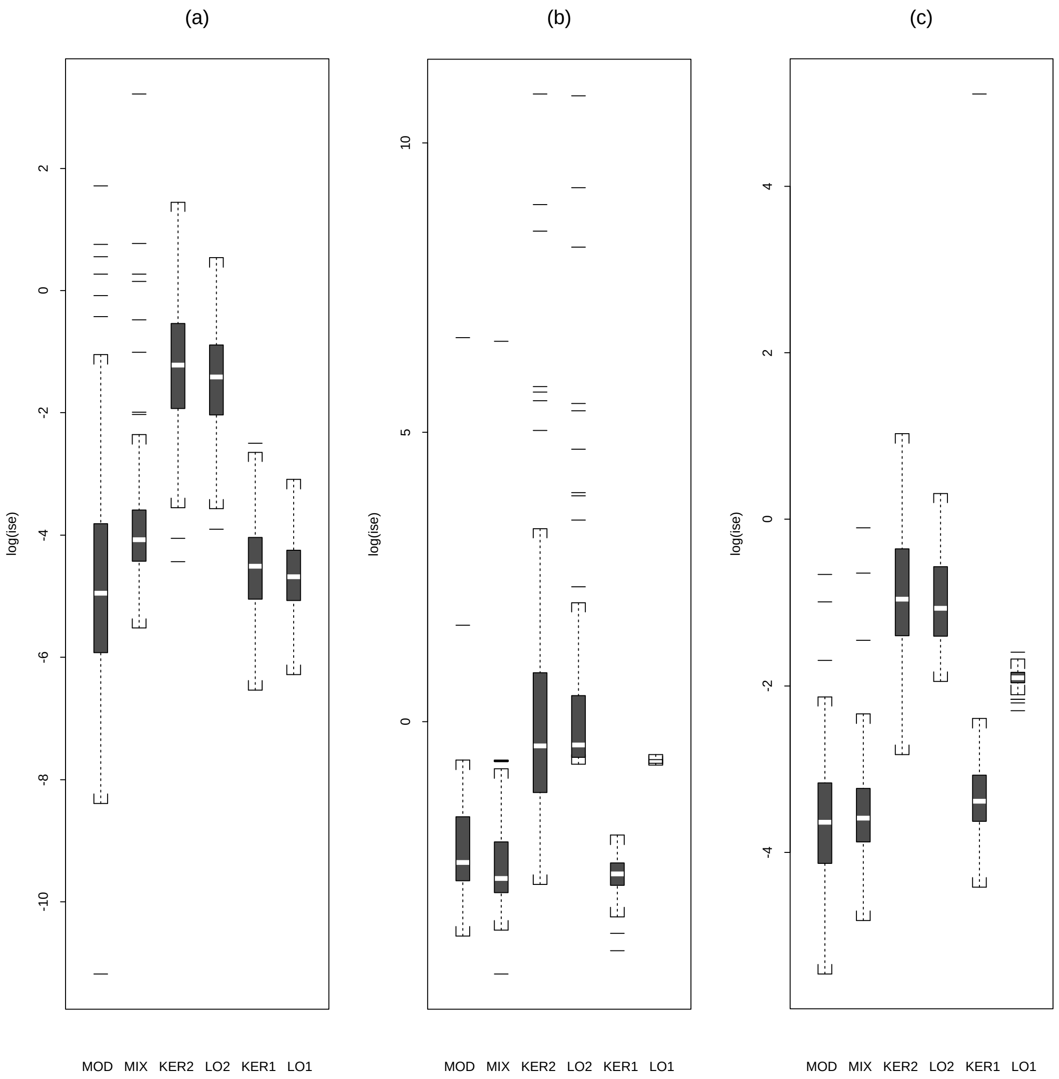

Figure 5: Boxplots of $\log ISE$. Panels (a)–(c) correspond to the functions f_1, f_2 and f_3. In each panel the boxplots from left to right correspond to the posterior mode estimate (MOD), mixture estimate (MIX), the local linear estimate with outliers (KER2), the loess estimate with outliers (LO2), the local linear estimate with no outliers (KER1) and the loess estimate with no outliers (LO1).

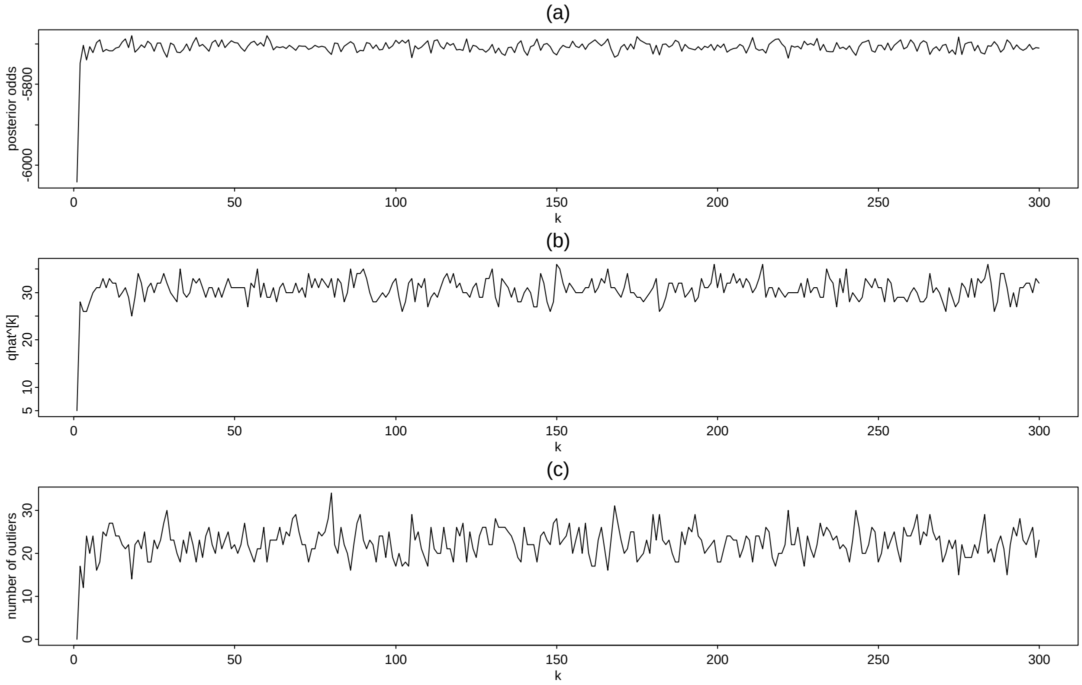

Figure 6: Panels (a)-(c) Trace of the $\log p(\hat{\gamma}^{[k]}, \hat{\omega}^{[k]}|y)$, number of significant regressors, and the number of outliers for each of the generated values $(\hat{\gamma}^{[k]}, \hat{\omega}^{[k]})$ during the warmup period where $k = 1, \ldots, 300$

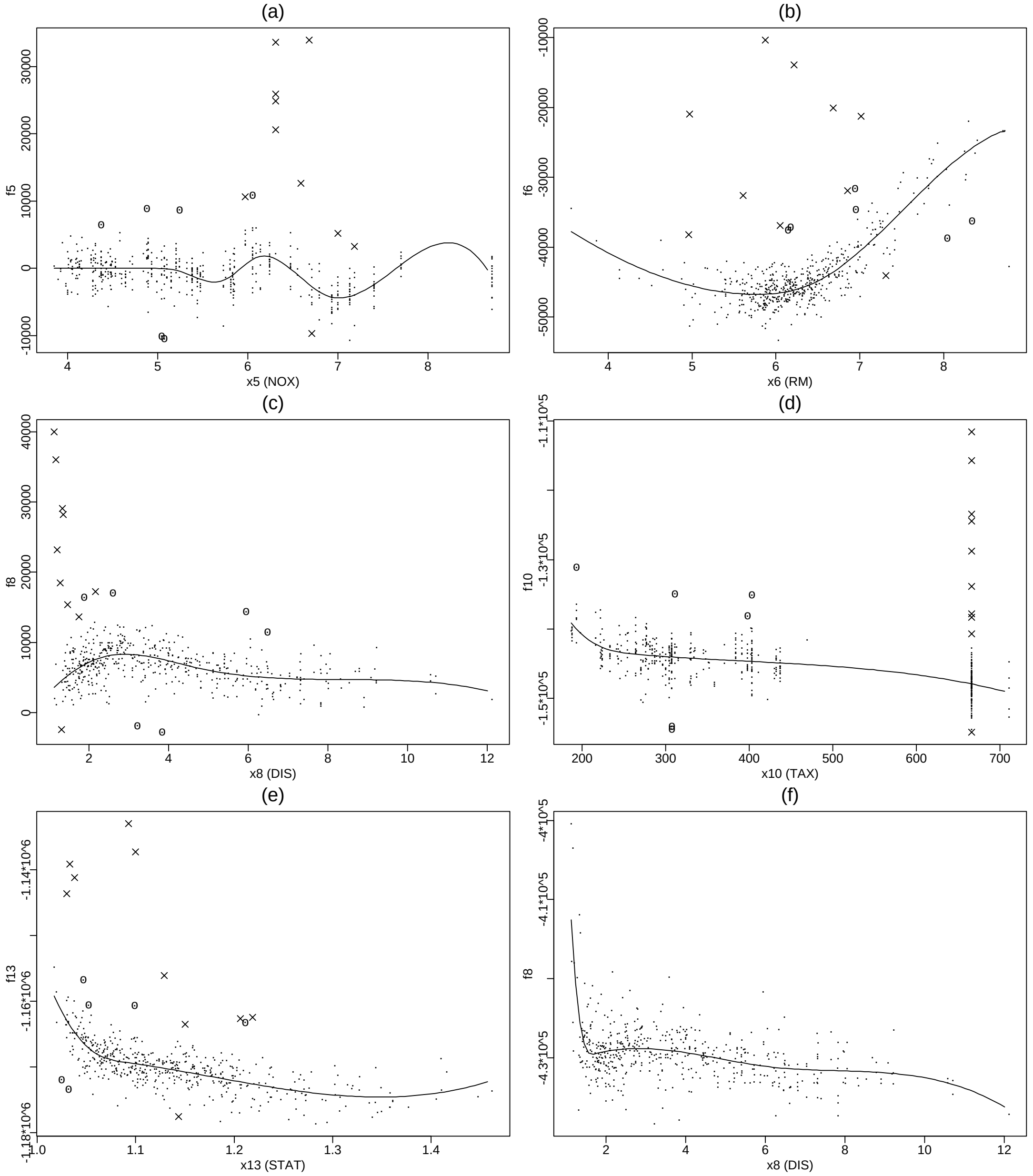

Figure 7: Panels (a)-(e) plot the robust posterior mean estimates of $f_5(X_5), f_6(X_6), f_8(X_8), f_{10}(X_{10})$, and $f_{13}(X_{13})$. The panels also show the data minus the effects of all the other variables as estimated by our fit. Outliers discussed in the text are indicated by a cross and the remaining outliers are circled. Panel (f) is a plot of the nonrobust estimate of $f_8(X_8)$.

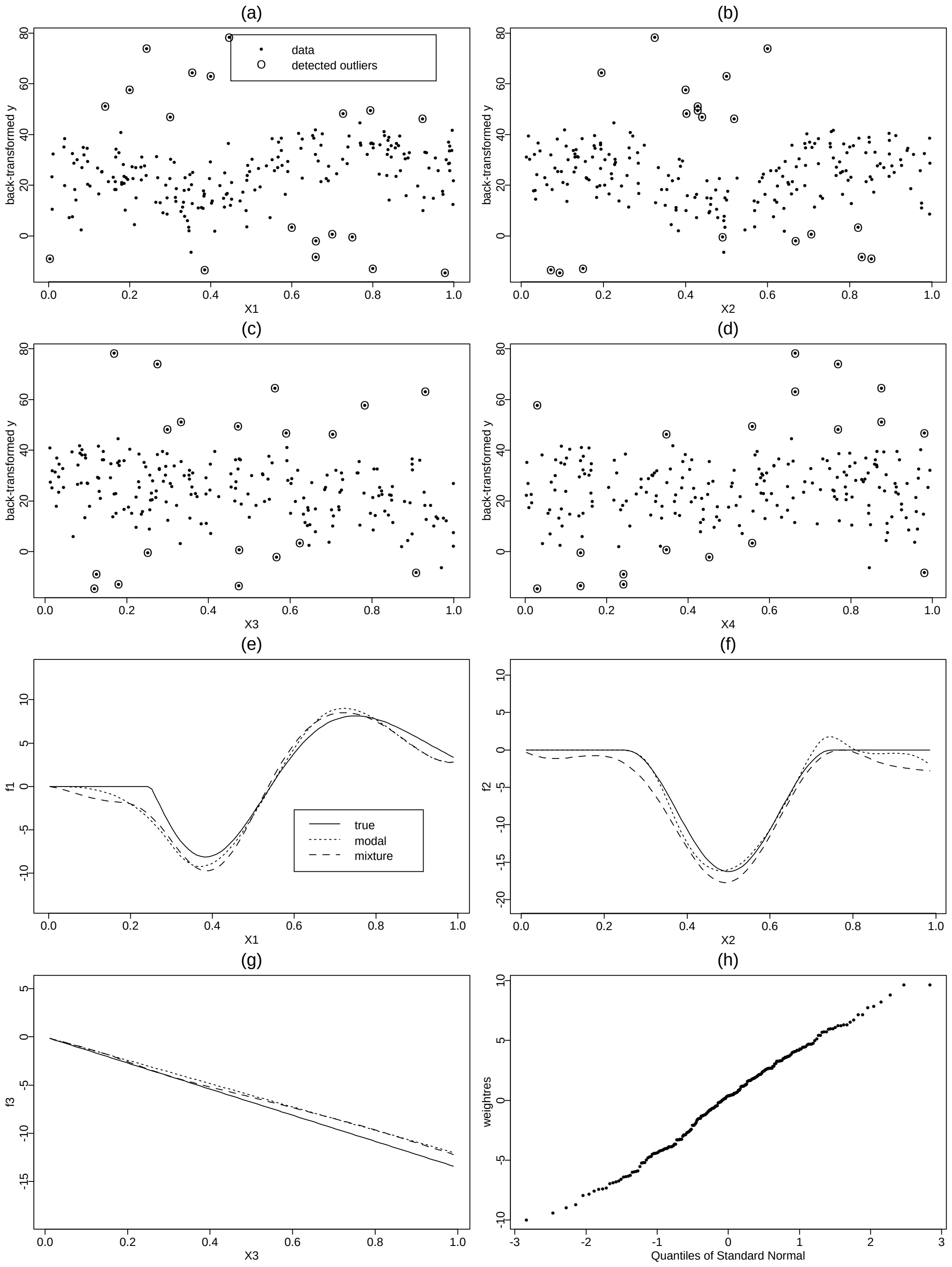

Figure 8: Panels (a)-(d): plots of the generated data on the original scale. The observations detected as outliers by the posterior mode estimate are circled. Panels (e)-(g): plots of the true functions under the appropriate linear transformation (solid line), the posterior mode estimate (dotted line), and posterior mean estimate (dashed line). Panel (h): normal probability plot of generalized least squares residuals