The Canadian Journal of Statistics
Vol. 37, No. 3, 2009, Pages 361–380
La revue canadienne de statistique

Bayesian robust transformation and variable selection: a unified approach

Raphael GOTTARDO[1]* and Adrian RAFTERY[2]

[1]*Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T1Z2*
[2]*Department of Statistics, University of Washington, Seattle, WA 98195-4320, USA*

Key words and phrases: Bayesian model averaging; Box–Cox; generalized regression coefficient; linear regression; Markov chain Monte Carlo; mixture distribution; *t*-distribution.

MSC 2000: Primary 62F15; secondary 62J05.

Abstract: The authors consider the problem of simultaneous transformation and variable selection for linear regression. They propose a fully Bayesian solution to the problem, which allows averaging over all models considered including transformations of the response and predictors. The authors use the Box–Cox family of transformations to transform the response and each predictor. To deal with the change of scale induced by the transformations, the authors propose to focus on new quantities rather than the estimated regression coefficients. These quantities, referred to as generalized regression coefficients, have a similar interpretation to the usual regression coefficients on the original scale of the data, but do not depend on the transformations. This allows probabilistic statements about the size of the effect associated with each variable, on the original scale of the data. In addition to variable and transformation selection, there is also uncertainty involved in the identification of outliers in regression. Thus, the authors also propose a more robust model to account for such outliers based on a *t*-distribution with unknown degrees of freedom. Parameter estimation is carried out using an efficient Markov chain Monte Carlo algorithm, which permits moves around the space of all possible models. Using three real data sets and a simulated study, the authors show that there is considerable uncertainty about variable selection, choice of transformation, and outlier identification, and that there is advantage in dealing with all three simultaneously. *The Canadian Journal of Statistics* 37: 361–380; 2009

Résumé: Nous considérons le problème de la selection de transformations et de variables pour la régression linéaire. Nous proposons une approche Bayesienne à ce problème qui nous permet de faire la moyenne de tous les modèles considerés y compris les transformations de type Box-Cox de la résponse et des prédicteurs. Pour prendre en considération le changement d'unité induit par les transformations, nous proposons d'examiner et d'estimer de nouvelles quantités à la place des coéfficients de regréssion. Ces quantités nouvelles, que nous appellons coefficients de régressions generalisés, peuvent être interpretés comme les coéfficients de régression dans l'unité originale des données, et ne dependent donc pas des transformations sélectionées. En particulier, cela nous permet de faire de l'inference sur la taille des effets associés avec chaque variable, et ce, dans l'unité original des données. En plus des transformations, nous considérons aussi le problème de la détection de valeurs abérrantes, ainsi que l'incertitude associée à cette détection. Pour modéliser ces données abérrantes, nous utilisons une loi de *t* avec un nombre de degrés de liberté inconnu. L'estimation des paramêtres est faite en utilisant un algorithm MCMC efficace qui nous permet de traverser l'espace constitué de tous les modèles possibles. En utilisant trois jeux de données réelles ainsi que des données simulées, nous montrons que l'incertitude associée au choix de variables, de transformations et de données abérrantes est considérable, et qu'il est important que les trois sélections soient considérées en meme temps. *La revue canadienne de statistique* 37: 361–380; 2009

* *Author to whom correspondence may be addressed.*
E-mail: raph@stat.ubc.ca

1. INTRODUCTION

Variable selection in linear regression is an important problem, whose purpose is to select a set of variables that best predict an outcome variable. Given a dependent variable $\boldsymbol{Y}$ and a set of potential independent variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, we wish to compare models of the form $\boldsymbol{Y} = \beta_0 + \boldsymbol{X}_{j_1}\beta_{j_1} + \cdots + \boldsymbol{X}_{j_q}\beta_{j_q} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_q}$ is a selected subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. For this problem, it is common to assume a standard linear model to describe the relationship between the response and independent variables, namely

$$Y_i = \beta_0 + \sum_{k=1}^{q} x_{ij_k}\beta_{j_k} + \epsilon_i, \tag{1}$$

where the β_{j_q}'s are unknown regression coefficients and ϵ_i follows a normal distribution with constant variance ψ^{-1}. In many cases, such assumptions (common variance, additive error structure, normal distribution) might be unrealistic and one solution is to look for transformations of the outcome variable and/or regressors so that (1) is appropriate after transformation.

Box & Cox (1964) discussed the power transformation family of models. In particular, they gave methods for estimating the parameters in the nonlinear model

$$h(\boldsymbol{Y}, \lambda) = \boldsymbol{A}\boldsymbol{\beta} + \psi^{-1/2}\boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ follows a standard normal distribution, ψ is the precision, that is, the reciprocal of the variance, $\boldsymbol{A}$ is a known design matrix, $\boldsymbol{\beta}$ is a vector of parameters, and

$$h(\boldsymbol{Y}, \lambda) = \begin{cases} (\boldsymbol{Y}^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(\boldsymbol{Y}) & \text{otherwise.} \end{cases}$$

Note that this transformation is valid only if $\boldsymbol{Y} > 0$. There exist several methods for estimating the unknown parameters such as maximum likelihood (Box & Cox, 1964) and Bayesian approaches (Box & Cox, 1964; Perrichi, 1981; Sweeting, 1984; Hinkley & Runger, 1984).

There has been some discussion about the correct way to make inference about the regression parameter $\boldsymbol{\beta}$ when the transformation parameter λ is unknown (Bickel & Doksum, 1981; Box & Cox, 1982; Hinkley & Runger, 1984). Bickel & Doksum (1981) showed that the variance of $\hat{\boldsymbol{\beta}}$ is greatly inflated when λ is estimated from the data compared to the case where it is known. Box & Cox (1982) and Hinkley & Runger (1984) argued that linear parameters have meaning only with reference to a particular scale and thus recommended taking a conditional approach, in which inferences from the regression model are conditional on a single selected value of λ. Chen & Lockhart (1997) obtained the Fisher information matrix and its inverse for all the unknown parameters, namely $\boldsymbol{\beta}$, ψ^{-1}, and λ. They showed that the asymptotic distributions of $(\hat{\psi}^{-1}, \hat{\boldsymbol{\beta}})$ conditionally and unconditionally on λ were different and concluded that conditioning on λ was one step short of performing valid analyses. Chen, Lockhart & Stephens (2002) extended these results and argued that inference about the parameter $\psi^{1/2}\boldsymbol{\beta}$ is preferable over inference about the original $\boldsymbol{\beta}$, as the asymptotic distribution of the estimators has better properties. In this article, we consider a different parameter for inference.

Carroll & Ruppert (1981) and Taylor (1986) proposed restricting attention to the predictive distribution of new observations, which can be defined independently of the scale. They studied the properties of the conditional median and conditional mean, respectively. They showed that

there is some cost in estimating λ but that the cost is not severe. However, their approach ignored the regression coefficients, which themselves are often of interest, because they summarize the relationship between the dependent and independent variables.

To avoid scaling issues, Box & Cox (1964) and Smith & Kohn (1996) considered standardizing the transformation so that the scale of the transformed data is approximately the same for each λ, but Dagenais & Dufour (1994) argued against this as there is no clear interpretation of the parameters after such standardization.

Variable selection and choice of transformation are often done sequentially, and the resulting model depends on the order in which they are performed. It would seem more appropriate to do them simultaneously. Since both can be viewed as model selection problems, we can unify them within a Bayesian framework, which would allow us to get more realistic measures of uncertainty by averaging over all models considered. Hoeting, Raftery & Madigan (2001) proposed a simultaneous approach to variable selection and transformation based on a power transformation for the outcome variable and change-point transformations for the independent variables. The change-point transformation has the advantage of not inducing a change of scale. However, the power transformation does, and Hoeting, Raftery & Madigan (2001) did not have to consider this scaling issue. They used an approximate algorithm based on Bayes factors to estimate the cut-points of the change-point transformation and so did not average over all transformations considered. They used the Markov chain Monte Carlo (MCMC) model composition (MC^3) of Madigan & York (1995) to perform variable selection. This can be efficient but it also requires one to integrate out all the model parameters analytically. Geweke (1996) and George & McCulloch (1997) showed that this integration can be avoided.

Hoeting & Ibrahim (1998) proposed taking a predictive Bayesian viewpoint as the basis for variable and transformation selection. They used the Box–Cox family of transformations, but did not have to worry about scaling issues as they did not average over all transformations considered but instead selected the best one. Their variable selection approach requires one to compute all possible models, which may not be practical even for moderate sized problems. Liu et al. (2003) also considered the problem of transformation and variable selection, but in the context of clustering, and thus were not concerned with scaling issues.

In this article, we introduce a Bayesian model for variable and transformation selection in linear regression. Our transformation selection approach is based on the Box–Cox family of transformations but could be generalized to other types of transformation. Parameter estimation is carried out using an efficient MCMC algorithm, which allows us to move around the space of all possible models (including transformations). To deal with the change of scale induced by the transformations, we focus on new quantities, which we call generalized regression coefficients. These have a similar interpretation to the usual regression coefficients on the original scale of the data, but do not depend on the transformations selected. Finally, in addition to variable and transformation selection, there is also uncertainty associated with the identification of outliers in regression. We include the identification of outliers in our methodology; our approach is based on t-distributions with unknown degrees of freedom.

Section 2 starts by describing our basic model and prior specification for variable and transformation selection, and then extends it to deal with outliers. Section 2 also discusses our solution to the scaling problem. In Section 3, we introduce the MCMC algorithm used for parameter estimation. In Section 4, we illustrate our methodology using three real data sets and a simulated one. In Section 5, we evaluate the predictive performance of our model compared to models that do not include transformation and outlier selection. Finally, in Section 6 we discuss our findings and possible extensions.

2. A MODEL FOR VARIABLE AND TRANSFORMATION SELECTION

Given a dependent variable $\boldsymbol{Y}$ and a set of potential regressors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, we wish to compare models of the form $g(\boldsymbol{Y}) = \beta_0 + g_{j_1}(\boldsymbol{X}_{j_1})\beta_{j_1} + \cdots + g_{j_q}(\boldsymbol{X}_{j_q})\beta_{j_q} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_q}$ is a selected subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, and g and the g_{j_q}'s are transformations from a predefined set of real functions, $\mathcal{T}$. Here the set of possible transformations $\mathcal{T}$ is a parametric family, $\mathcal{T} = \{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d\}$. When the response is transformed, a Jacobian term enters into the likelihood for the untransformed response, and so we require $g_{\boldsymbol{\theta}}$ to be a diffeomorphism for each $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. Note that this assumption is not required for the independent variables; see, for example, Hoeting, Raftery & Madigan (2001) where the authors used a change-point transformation for the independent variables. Here, for simplicity, we assume that the set of possible transformations is the same for the outcome variable and the independent variables.

We use the Box–Cox family of transformations, which we define as

$$\mathcal{T} = \{g_\lambda(x) \equiv (x^\lambda - 1)/\lambda : \lambda \in \Lambda \subset \mathbb{R}\},$$

where Λ is a given subset of $\mathbb{R}$. Note that the methodology presented here could be extended to other parametric transformations. In particular, if the data are not positive, then one could use the shifted power transformation of Box & Cox (1964) or the extended power transformation family of Bickel & Doksum (1981). The idea of transforming the independent variables was proposed by Box & Tidwell (1962).

2.1. A Basic Model

We now assume that a standard linear model can be used to describe the relationship between the transformed response and independent variables, namely

$$g_\lambda(Y_i) = \beta_0 + \sum_{j=1}^{p} g_{\lambda_j}(x_{ij})\beta_j + \epsilon_i, \tag{3}$$

$$(\epsilon_i|\psi) \sim \mathrm{N}(0, \psi^{-1}),$$

where the $\boldsymbol{\beta}_i$'s are the unknown regression coefficients, and λ and λ_j's are the transformation parameters.

In order to allow each variable to be either in or out of the model, we model each regression coefficient as a mixture of a Normal distribution and a point mass at zero, as follows:

$$(\beta_j|\lambda, \lambda_j, \sigma_\beta) \sim (1-w)\delta_0 + w\mathrm{N}\left(0, \frac{S^2_{g_\lambda(\boldsymbol{Y})}}{S^2_{g_{\lambda_j}(\boldsymbol{X}_j)}}\sigma^2_\beta\right), \tag{4}$$

where w is the prior probability of being in the model for each variable, $S^2_{\boldsymbol{z}}$ denotes the empirical variance of $\boldsymbol{z}$, and σ^2_β is a common variance parameter. Note that the prior distribution for the regression coefficients is allowed to depend on the scales of the variables; this is to account for the change of scale induced by the transformations. We do not view this as being in contradiction with the Bayesian paradigm, but rather as an approximation to the prior information of an investigator who knows something, but not much, about the problem at hand. The prior distributions of the regression coefficients are as spread out as they can reasonably be given the marginal standard deviations of the variables. Note that data-dependent priors have been used by other researchers in this context (Box & Cox, 1964; Sweeting, 1984).

The parameter w is assumed to have a Beta distribution, $\mathrm{Beta}(w_0\kappa_0, (1-w_0)\kappa_0)$, where w_0 and κ_0 are fixed hyperparameters. The mean of this distribution is w_0 and the variance is

$w_0(1 - w_0)/(1 + \kappa_0)$ so that w_0 can be seen as a "best guess" for w, whereas κ_0 controls the spread around this prior guess. Putting a prior on w allows us to estimate the proportion of variables, which can be seen as a Bayesian solution to the multiple testing problem arising when the number of variables to be included is large (Scott & Berger, 2006). Throughout this article we use $w_0 = 0.5$ and $\kappa_0 = 4$, that is, a fairly noninformative Beta (2,2) prior. Note that this prior makes every model (marginally) equally likely a priori.

The prior for the variance parameter σ_β^2 is taken to be uniform in the interval [0, 1]. The rationale for this is as follows. If all the variables are standardized by dividing by their standard deviations, then (4) implies that the prior variance of β_j is σ_β^2. On this scale, β_j rarely exceeds 1 in absolute value. It never does so when $P = 1$, by the Cauchy–Schwarz inequality, and empirical evidence that it rarely does so when $P > 1$ was given by Raftery, Madigan & Hoeting (1997). Thus, $\sigma_\beta^2 = 1$ will almost always be more than large enough to cover the range of values of β_j, and much smaller values can easily be appropriate. Thus, a U[0, 1] prior for σ_β^2 covers the range of possibilities fairly well. The results are typically insensitive to reasonable changes in this prior.

The prior for the scaling parameter ψ is taken to be improper, $\pi(\psi) \propto \psi^{-1}$. The prior for the intercept β_0 is $(\beta_0|\lambda) \propto 1/S_{g_\lambda(\boldsymbol{Y})}$, to account for the change of scale induced by λ. Our rational for this prior follows from an idea used in Box & Cox (1964). Suppose that for a fixed value λ_1, the transformation over the range of observation is nearly linear, $g_\lambda(y) \approx \text{const} + l_\lambda g_{\lambda_1}(y)$, where l_λ is a rescaling constant (depending on λ). Suppose furthermore that for λ_1, the regressors have little effect on the response, that is, $E[\boldsymbol{g}_{\lambda_1}(Y)|\boldsymbol{X}] \approx \beta_0$, and using the approximate linear relationship $E[\boldsymbol{g}_\lambda(Y)|\boldsymbol{X}] \approx l_\lambda \beta_0$. Choosing $\lambda_1 = 1$, a simple estimate for l_λ is $S_{g_\lambda(\boldsymbol{Y})}$. The prior for λ only reflects the Jacobian term coming from the linear transformation. Again, this prior is data dependent and is in the spirit of the one used in Box & Cox (1964). While simple, this prior accounts for the change of scale induced by the transformation of the response variable. We have found this prior to give reasonable results in practice. Finally, we used a uniform prior on $[-1, 1]$ for the transformation parameters λ and λ_j's.

2.2. A Robustified Model

It has been shown that transformation selection can be heavily influenced by the presence of a few outliers (Carroll & Ruppert, 1982, 1985; Cook & Wang, 1983; Atkinson, 1988; Hinkley & Wang, 1988; Cheng, 2005). As with transformation and variable selection, the order in which outlier and transformation selection are done tends to lead to different answers. Thus, once again, we wish to do everything simultaneously, and our approach is based on t distributions with unknown degrees of freedom. We introduce a more robust version of (3), as follows:

$$
\begin{aligned}
g_\lambda(Y_i) &= \beta_0 + \sum_{j=1}^{p} g_{\lambda_j}(x_{ij})\beta_j + \frac{\epsilon_i}{\sqrt{\varpi_i}}, \qquad (5)\\
(\epsilon_i|\psi) &\sim \mathrm{N}(0, \psi^{-1}),\\
(\varpi_i|\nu) &\sim \mathrm{Gamma}(\nu/2, \nu/2),
\end{aligned}
$$

where the β_i's are the unknown regression coefficients and the ϖ_i's are independent of the ϵ_i's. Since the ϖ's are independent of the ϵ's, we have $\epsilon_i/\sqrt{\varpi_i} \sim t_{(\nu,0,\psi^{-1})}$, that is, the errors have a t distribution with ν degrees of freedom and scale parameter ψ^{-1}. Note that conditionally on the ϖ_i, the likelihood is again Gaussian and the ϖ_i can be interpreted as weights, a small value for ϖ_i would downweight the influence of y_i in the likelihood calculation.

The prior for the degrees of freedom, ν, is taken to be uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. All other priors remain the same. As we will see in Section 4, the

accommodation of outliers can have a substantial influence on both variable and transformation selection.

2.3. Generalized Regression Coefficients

In regression analysis, the β_j's themselves are also of interest as they summarize the relationship between the dependent variable and the independent variables. One difficulty with transformations is that the scale, and thus the interpretation, of each β_j depends on the transformations currently being applied. As a result, the usual MCMC posterior summaries for the β_j's, such as means and standard deviations, are meaningless. As a solution we focus on a different quantity, which is defined independently of the transformation and has a similar interpretation to β_j on the original scale of the response and independent variables. When there is no transformation it is equal to β_j.

On the original scale (i.e., when no transformation is applied), we have, for each observation (omitting the observation index i),

$$\text{med}(Y) = \beta_0 + \sum_j \beta_j X_j,$$

and as a result $\text{d}[\text{med}(Y)]/\text{d}X_j = \beta_j$. Similarly, after transformation of Y and X,

$$\text{med}\left(\frac{Y^\lambda - 1}{\lambda}\right) = \beta_0 + \frac{\sum_j \beta_j (X_j^{\lambda_j} - 1)}{\lambda_j}.$$

Thus by invariance of the median we get

$$\text{med}(Y) = \left[1 + \lambda \left\{\beta_0 + \frac{\sum_j \beta_j (X_j^{\lambda_j} - 1)}{\lambda_j}\right\}\right]^{1/\lambda},$$

and so

$$\frac{\text{d}[\text{med}(Y)]}{\text{d}X_j} = \beta_j X_j^{\lambda_j - 1} \left[1 + \lambda \left\{\beta_0 + \frac{\sum_j \beta_j (X_j^{\lambda_j} - 1)}{\lambda_j}\right\}\right]^{1/\lambda - 1}.$$

The quantity $\text{d}[\text{med}(Y)]/\text{d}X_j$ does not depend on the transformations and has a similar interpretation to β_j on the original scale. Thus we use the sample average of values of this quantity, namely

$$\beta_j^{\text{G}} \equiv \frac{1}{n} \sum_i \beta_j X_{ij}^{\lambda_j - 1} \left[1 + \lambda \left\{\beta_0 + \frac{\sum_j \beta_j (X_{ij}^{\lambda_j} - 1)}{\lambda_j}\right\}\right]^{1/\lambda - 1}, \tag{6}$$

as a measure of the marginal change in Y when X_j is varied, all else equal. This has a similar interpretation across transformations, and it is equal to β_j on the original scale of the data. We call the quantity in (6) the generalized regression coefficient, as it generalizes the usual regression coefficient in the presence of transformations. The generalized regression coefficients can easily be estimated from the MCMC output.

Note that the quantity $1 + \lambda\{\beta_0 + \sum_j \beta_j (X_{ij}^{\lambda_j} - 1)/\lambda_j\}$ might be negative, in which case (6) is not defined. However, this is unlikely to happen in practice since the observations are required to be positive. In fact, the probability of $1 + \lambda\{\beta_0 + \sum_j \beta_j (X_{ij}^{\lambda_j} - 1)/\lambda_j\}$ being negative goes

to zero as the sample size increases. The generalized regression coefficients, given by (6), were all well defined for all the examples analysed here. However, if β_j^{G} is not defined for a few possible values of the β_k parameters, one solution would be to find the posterior distribution of β_j^{G} conditional on its being well defined.

3. PARAMETER ESTIMATION AND MCMC COMPUTATION

3.1. MCMC Algorithm

It can be difficult to devise good MCMC algorithms in the context of transformation and variable selection due to the change of scale induced by the transformation at each iteration (Smith & Kohn, 1996). Here, we introduce an efficient algorithm that accommodates such changes of scale. The basic idea is to rescale the regression coefficients and the error variance while shifting the intercept every time a new transformation parameter is proposed in order to increase the acceptance rate. Our MCMC algorithm cycles through the following steps:

1. Block update $(\lambda, \beta_0, \boldsymbol{\beta}, \psi)$ by Metropolis-Hastings.
2. Update β_0 by Gibbs sampling.
3. For $j = 1$ to p:
   (a) If $\beta_j > 0$ block update $(\lambda_j, \beta_0, \beta_j, \psi)$ by Metropolis-Hastings.
   (b) Update β_j by Gibbs sampling.
4. Update σ_β^2 by Metropolis-Hastings.
5. Update ψ by Gibbs sampling.
6. Block update $\boldsymbol{\varpi}$ and ν by Gibbs sampling (for the model with t distributed errors).

For move 1, we first randomly select a candidate transformation λ^* using a symmetric proposal centred at the current value λ. Based on this selected value, we then carry out a deterministic update for the β_j, β_0, and ψ: first we rescale β_j to β_j^* in light of Equation (4), and then choose β_0^* and ψ^* to be those that maximize the likelihood ratio over the set of possible transformations $\beta_0^* = \beta_0 + \delta$ and $\psi^* = \alpha\psi$, where δ and α are constants in $\mathbb{R}$ and $\mathbb{R}^+$, respectively. We then use the Metropolis-Hastings algorithm to decide whether or not to accept these new candidate values; see the Appendix for more details. Rescaling the regression coefficients and variance, and shifting the intercept is not, in theory, necessary. However, it noticeably increases the acceptance rate and the overall performance of the MCMC algorithm. This implies that the intercept and the regression coefficients are updated several times in each MCMC cycle.

Move 3a is similar to move 1, and is thus described in the Appendix with the other moves.

3.2. Posterior Model Probability and Rao-Blackwellization

In each case, the posterior model probability can be estimated from the MCMC output as the proportion of time spent in the corresponding model. The marginal posterior probability that the coefficient for each predictor does not equal zero, namely $\Pr(\beta_j \neq 0|\boldsymbol{y})$, can be obtained by summing the posterior model probabilities across models for each predictor. However, note that in (7), w_j^* corresponds to the probability that $\beta_j \neq 0$ given everything else in the model. Thus, one can obtain more efficient estimates by Rao-Blackwellization when averaging the w_k^* values computed at each iteration at no extra cost.

4. EXAMPLES

We now give results for three real data examples and a simulated example. All results from our method were obtained using an MCMC algorithm based on one million iterations thinned by 100 after a burn-in period of 1,000 iterations.

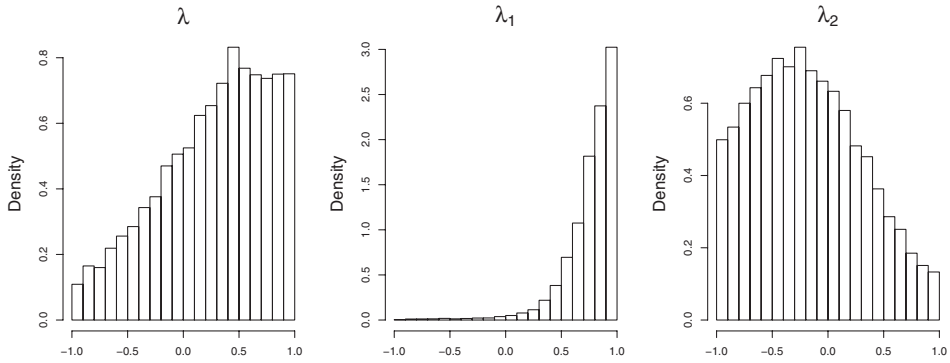

FIGURE 1: Histograms of MCMC samples from the posterior distributions of the transformation parameters, λ and λ_k's for the Hald data.

4.1. The Hald Data

We first apply our methodology to the Hald cement data. These data have been analysed by many researchers; see, for example, Cook (1977), George & McCulloch (1993), and Hoeting & Ibrahim (1998). A full description of the data can be found in Draper & Smith (1981). There are four predictors, each one measuring the percentage composition of a particular ingredient in samples of cement concrete. The response is the heat evolved in calories per gram of cement.

Here we use model (3), as it has been noted that the Hald data were well-behaved with no influential observations (Cook, 1977). This was also confirmed by fitting the robust model (5), and observing a large estimated number of degrees of freedom. Figure 1 shows histograms of the marginal posterior distribution of the transformation parameters for the response and the most significant variables. There is evidence that the second independent variable, $\boldsymbol{X}_2$, needs to be transformed and some evidence that the outcome variable needs to be transformed as well. These results are in agreement with the estimated transformation parameters given by Hoeting & Ibrahim (1998) even though these authors did not consider transformation of the outcome variable. However, there is substantial uncertainty about the transformations, and such uncertainty is naturally taken into account with our approach, which is not the case for the approach of Hoeting & Ibrahim (1998).

Figure 2 shows that there are obvious changes of scale in the regression coefficients induced by the transformations. As a result, the usual ergodic averages from the MCMC output are meaningless, as explained in Section 2. The bottom graphs of Figure 2 show the trace plots of the generalized regression coefficients. The trace plot looks much better, with no obvious changes of scale, and the usual MCMC estimates can now be used. This fact is also illustrated in Figure 3 where it can be seen that the sampled values for β_2 clearly depend on the transformation parameters λ and λ_2. However, the generalized regression coefficients at each iteration do not depend on the transformation parameters.

The posterior model probabilities are given in Table 1 and are in broad agreement with previous results (George & McCulloch, 1993; Hoeting & Ibrahim, 1998). Note that the exact posterior model probabilities computed by George & McCulloch (1993) and Hoeting & Ibrahim (1998) differ from ours. This is because the models used are slightly different and because we consider transformations of both the outcome and the independent variables.

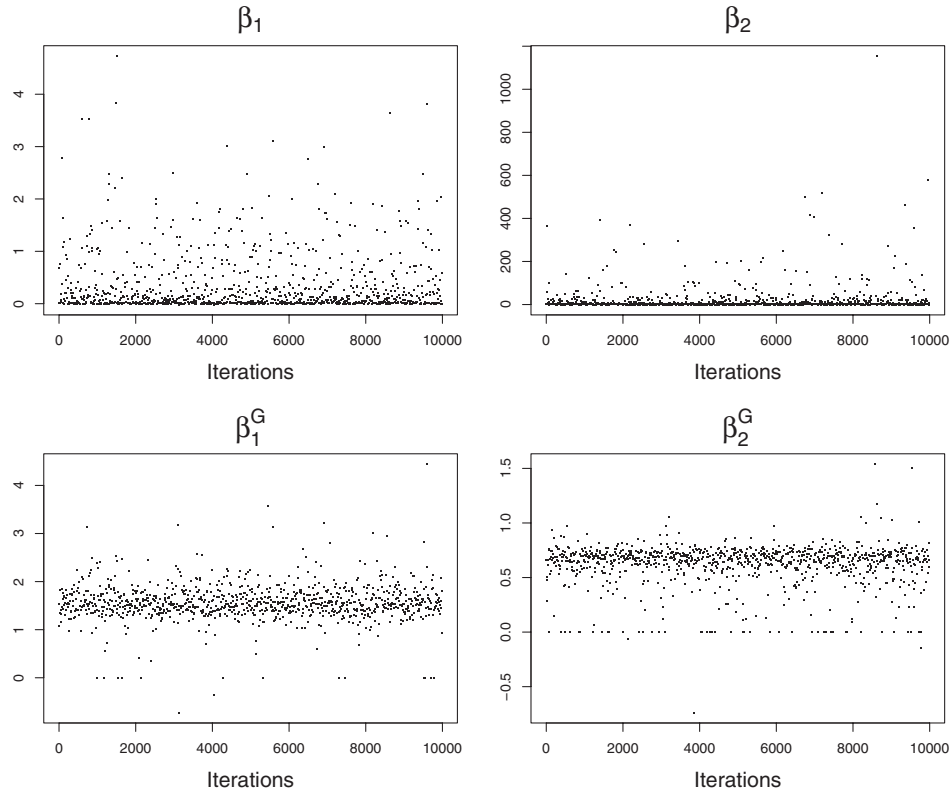

FIGURE 2: Trace plots of the regression coefficients, β_k's (top row), and the generalized regression coefficients (bottom row) for the Hald data. Note the y-axis scales: the regression coefficients show large changes of scale, while the generalized regression coefficients do not.

4.2. The US Crime Data

We now turn to the US crime data (Ehrlich, 1973), a larger data set with 47 observations and 15 independent variables and so potentially $2^{15} = 32,768$ different models. The variable names are given in Table 2. Ehrlich's analysis concentrated on the relationship between crime rate and predictors 14 and 15 (probability of imprisonment and average time served in state prisons). In his original analysis, Ehrlich (1973) focused on two regression models, consisting of the predictors (9, 12, 13, 14, 15) and (1, 6, 9, 10, 12, 13, 14, 15), respectively, which were chosen in advance based on theoretical grounds.

After logarithmic transformations of the data, Raftery, Madigan & Hoeting (1997) stated that standard diagnostic checking (e.g., Draper & Smith, 1981) did not reveal any gross violations of the assumptions underlying normal linear regression. Thus, we use the model with Gaussian errors given in (3). We consider transformation of all variables except $\boldsymbol{X}_2$, which is binary.

Figure 4 shows histograms of the marginal posterior distribution of each transformation parameter. The histogram for the transformation parameter of the outcome variable, λ, supports the logarithmic transformation originally used by Ehrlich (1973). However, our analysis suggests that independent variables 13 and 14 might need to be transformed. Again, there is a great deal of uncertainty about the transformation; and this uncertainty is naturally taken into account. As with the Hald data, there are obvious changes of scale in the regression coefficients induced by the transformations; see trace plots in Supplementary Material.

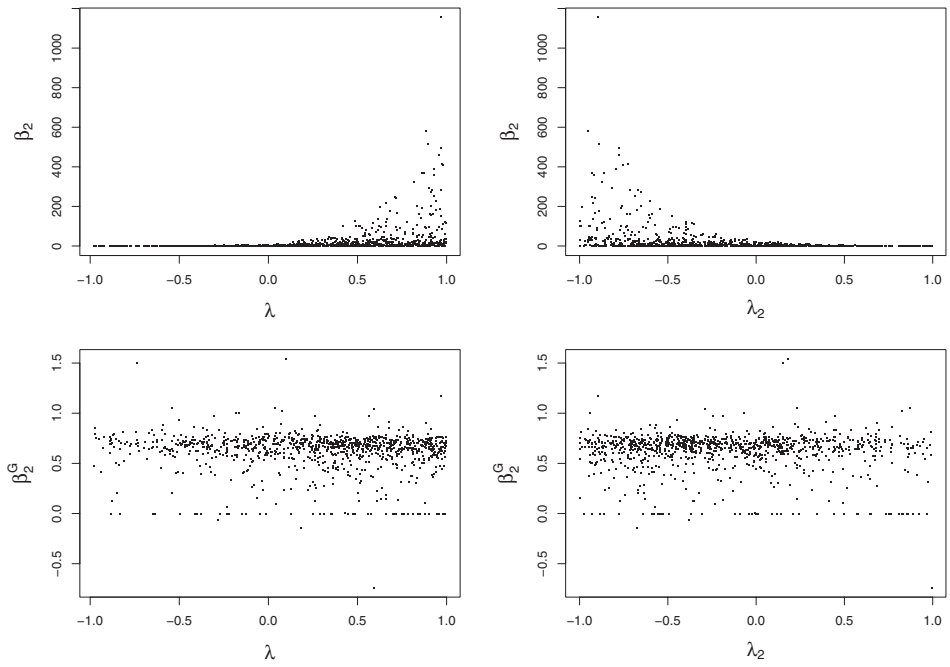

FIGURE 3: Scatter plots of one regression coefficient, β_2 (top row), and the generalized regression coefficients, β_2^G (bottom row), as a function of the transformation parameters λ and λ_2 for the Hald data. Each point corresponds to a specific MCMC sample. The regression coefficients are heavily dependent on the transformation parameters, while such dependence is absent for the generalized regression coefficients.

TABLE 1: Estimated posterior model probabilities for the Hald data.

X_1	X_2	X_3	X_4	Probability
•	•			0.478
•	•		•	0.205
•	•	•		0.137
•	•	•	•	0.132
•			•	0.023
•		•	•	0.013
0.988	0.959	0.288	0.374	1

Only models with posterior probabilities greater than 0.01 are displayed. The marginal posterior probabilities of each variable being included in the model are shown in the last row.

The marginal posterior model probabilities of inclusion computed from our model are given in Table 2. For comparison, we have also included the posterior probabilities computed by MCMC model composition as described in Raftery, Madigan & Hoeting (1997). Overall the probabilities are in broad agreement. For example, the most significant variables are the same (and in the same order), all with probability greater than 0.8. However, the results are slightly different because we consider variable transformation, which they did not, and because we estimate the

TABLE 2: Estimated marginal posterior probabilities (MPP) of inclusion for the crime data.

Predictor	MPP	MC^3	Step.	Ehrlich's Models	
$\boldsymbol{X}_1$: Percentage of males age 14–24	71	79	$*$		$\times$
$\boldsymbol{X}_2$: Indicator variable for southern state	41	17			
$\boldsymbol{X}_3$: Mean years of schooling	87	98	$*$		
$\boldsymbol{X}_4$: Police expenditure in 1960	79	72	$*$		
$\boldsymbol{X}_5$: Police expenditure in 1959	70	50			
$\boldsymbol{X}_6$: Labor force participation rate	30	6			$\times$
$\boldsymbol{X}_7$: Number of males per 1,000 females	36	7			
$\boldsymbol{X}_8$: State population	38	23			
$\boldsymbol{X}_9$: Number of non-Whites per 1,000 people	67	62		$+$	$\times$
$\boldsymbol{X}_{10}$: Unemployment rate of urban males age 14–24	31	11	$*$		$\times$
$\boldsymbol{X}_{11}$: Unemployment rate of urban males age 35–39	51	45	$*$		
$\boldsymbol{X}_{12}$: Wealth	52	30	$*$	$+$	$\times$
$\boldsymbol{X}_{13}$: Income inequality	98	100	$*$	$+$	$\times$
$\boldsymbol{X}_{14}$: Probability of imprisonment	83	83	$*$	$+$	$\times$
$\boldsymbol{X}_{15}$: Average time served in state prisons	43	22		$+$	$\times$

The MC^3 columns correspond to the MCMC model composition results presented in Hoeting, Raftery & Madigan (1996). The last three columns indicate the predictors selected by stepwise regression ($*$) and the predictors included in the two models considered by Ehrlich. $+$ corresponds to Ehrlich model 1 and $\times$ corresponds to Ehrlich model 2.

proportion of independent variables to be included in the model. Raftery, Madigan & Hoeting (1997) considered every model equally likely a priori, which implicitly fixes the proportion of independent variables to be included in the model at 0.5. In our case, the estimated posterior mean for the proportion parameter w is 0.57, which suggests that more variables should be included. We have also included the models selected by Ehrlich as well as the best model obtained by stepwise regression. Most of the variables selected by the three models agree with our marginal posterior probabilities. All the model posterior probabilities computed by our model were less than 5%, and we do not show them here, given the large number of models. This indicates that, as noted by Raftery, Madigan & Hoeting (1997), there is a great deal of uncertainty and that selecting one model is not ideal. This will be illustrated in Section 5 where we will compare the predictive performance of several models including stepwise regression when there is no model averaging.

4.3. The Scottish Hill Racing Data

Our third example involves data supplied by the Scottish Hill Runners Association. This example has been used by many researchers to illustrate the influence of outliers in linear regression (Atkinson, 1986, 1988; Hoeting, Raftery & Madigan, 1996). Here, we use it to illustrate the influence of outliers on both variable and transformation selection. The purpose of the study was to investigate the relationship between record times of 35 hill races and two predictors: distance, defined as the total length of the race measured in miles, and climb, defined as the total elevation gained in the race, measured in feet; see Atkinson (1986) for further details.

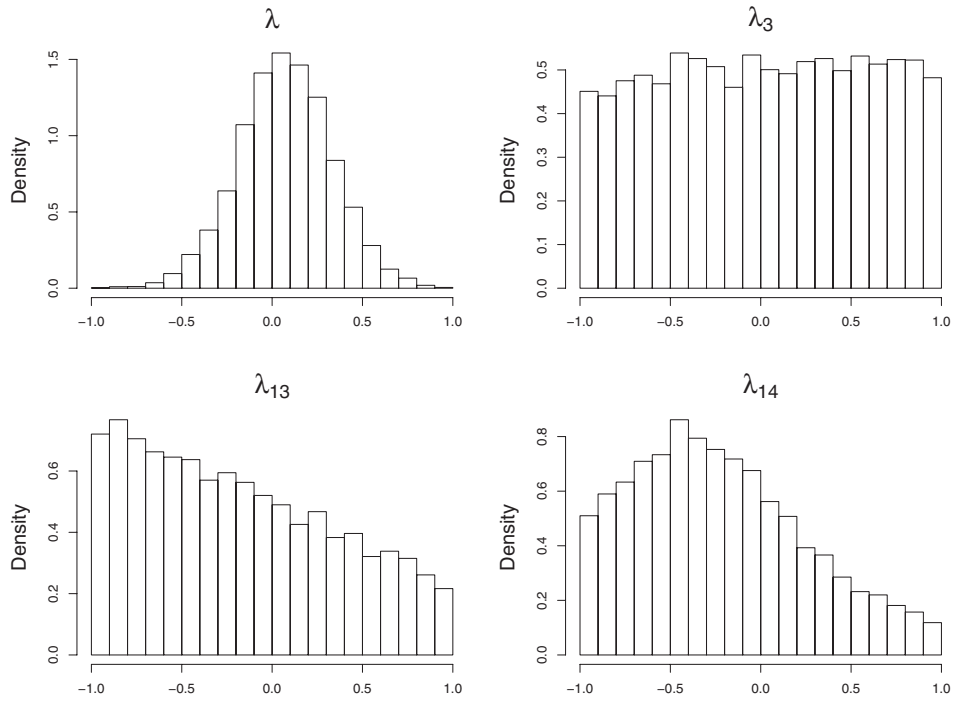

FIGURE 4: Histograms of the posterior distributions of the transformation parameters, λ and λ_k's, for the crime data. Only λ_k's for which the corresponding posterior probability for the regression coefficient of being nonzero is greater than 0.8 are displayed.

In particular, Atkinson (1986) and Hadi (1990) concluded that races 7 and 18 are outliers. After they removed observations 7 and 18, their methods indicated that observation 33 is also an outlier. Thus, they concluded, observations 7 and 18 mask observation 33.

We start by fitting the Gaussian model (3). The top graphs of Figure 5 show histograms of the marginal posterior distribution of each transformation parameter for the model with Gaussian errors. All three histograms, except perhaps that of λ_1, clearly suggest transformation of the corresponding variables. The bottom graphs of Figure 5 show histograms of the marginal posterior distribution of each transformation parameter for the model with t errors. Now, there is not much evidence that the first independent variable $\boldsymbol{X}_1$ should be transformed and less evidence that the outcome variable should be transformed as well. From the Gaussian model, the posterior mean of λ is 0.12 and a 95% credible interval is (–0.34, 0.52), while from the t model the posterior mean of λ is 0.48 and a 95% credible interval is (0.33, 0.64). This suggests that transformations for both $\boldsymbol{X}_1$ and $\boldsymbol{Y}$ are heavily influenced by the presence of outliers.

Table 3 shows the estimated posterior medians of the ϖ_i's, which can be interpreted as weights. Observations 18, 19, and 24 have small weights, suggesting that they are outliers. In particular, the corresponding 90% credible intervals do not contain 1 whereas all other intervals contain 1. On the other hand, observation 7 is slightly downweighted while observation 33 is not downweighted at all, suggesting that transformations of the response and the independent variables are enough to accommodate such outliers. For comparison, we fitted the same model without transforming the independent variables, and observations 7, 18, and 33 had the three smallest weights, all smaller than 0.4 (results not shown). Note, however, that transformations of the independent variables

 DOI: 10.1002/cjs

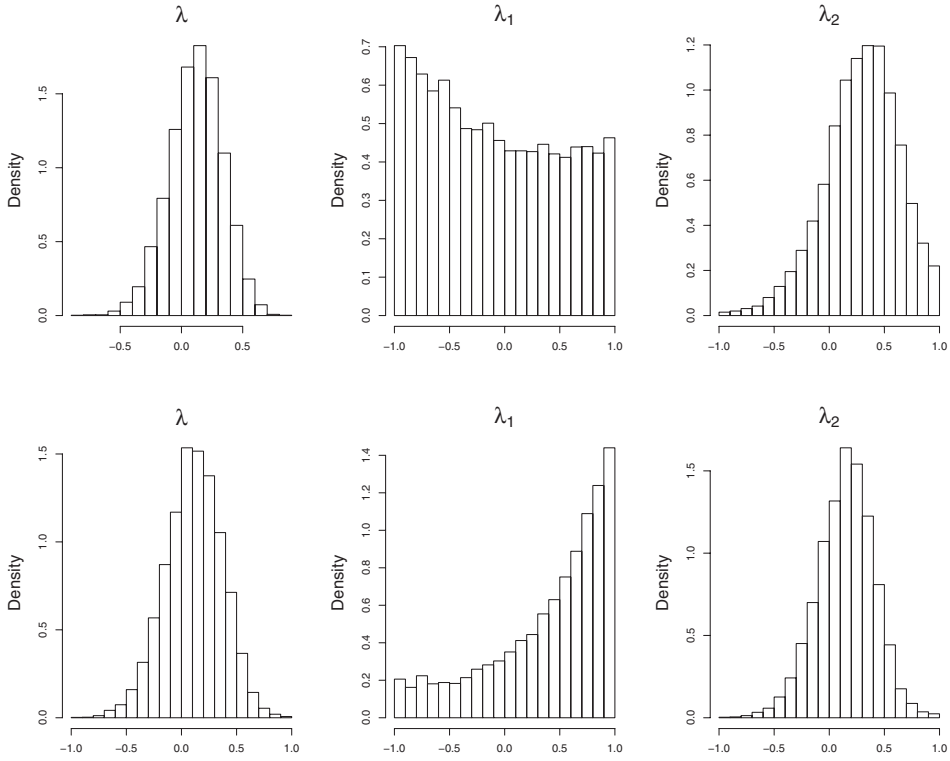

FIGURE 5: Histograms of the transformation parameters, λ and λ_k's, for the Scottish hill race data for the Gaussian model (top) and t model (bottom).

were not considered by Atkinson (1986). Overall, this indicates that transformation selection and treatment of outliers affect one another, and suggests that the two should be done simultaneously.

Finally, the posterior model probabilities are given in Table 4. There are substantial differences between the Gaussian and t models. After accounting for potential outliers the posterior probability of $\boldsymbol{X}_1$ being included in the model increased substantially.

In the three previous examples considered, the true answer is unknown. We have also used simulated data to evaluate the performance of our methodology when the truth is known, under

TABLE 3: Posterior weights, that is, posterior medians of the ϖ's, associated with each observation of the Scottish hill race data.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
Weight	0.80	0.98	0.89	1.01	1.05	0.74	0.51	1.08	1.08	0.93	0.73	0.66
Obs.	13	14	15	16	17	**18**	**19**	20	21	22	23	**24**
Weight	0.94	0.61	0.90	0.90	0.93	**0.03**	**0.28**	1.08	1.00	0.68	0.96	**0.30**
Obs.	25	26	27	28	29	30	31	32	33	34	35	
Weight	0.92	0.96	0.98	0.97	0.90	0.42	0.99	0.92	0.95	1.10	0.73	

Observations with small weights are downweighted. Observations 18, 19, and 24 have the smallest weights and are highlighted in bold, suggesting that they might be outliers.

TABLE 4: Estimated posterior model probabilities for the Scottish hill race Data.

Gaussian			t		
X_1	X_2	Prob.	X_1	X_2	Prob.
•	•	0.359	•	•	0.765
	•	0.641		•	0.235
0.359	1.000	1	0.767	1.000	1

Only models with posterior probabilities greater than 0.01 are displayed. The marginal posterior probabilities are given by the last row.

various model missspecifications. The simulation details and results of the simulation study are shown in on-line Supplementary Material.

5. ASSESSMENT OF PREDICTIVE PERFORMANCE

In addition to variable selection, it is of interest to look at the predictive performance of our approach compared to others. We use the predictive ability of the selected models for future observations to measure the effectiveness of a model selection strategy. Our specific objective is to compare the quality of the predictions based on model averaging with the quality of predictions based on any single model that an analyst might reasonably have selected. For comparison with other standard variable selection techniques, we included the same variants as in our simulation study (see Supplementary Material), namely transformation (T), outlier selection (O), and variable selection (VS). In order to be fair, when no transformation was selected, the transformation for the response was set to the maximum likelihood estimate of the Box–Cox parameter from the full model. Finally, we have included the simple normal model on the transformed response where the variable to be included were first selected by stepwise regression based on that same transformed response as previously described.

To measure performance, we randomly split the complete data set into two subsets. A 50/50 split was chosen here so that each portion would contain enough data to be a representative sample. We ran our model using half of the data, called the training data, and evaluated the prediction performance using the remaining half of the data, called the test data. A calibration plot was used to determine if the predictions were well calibrated. A model is well calibrated if, for example, 70% of the observations in the test data set are less than or equal to the 70th percentile of the posterior predictive distribution. The calibration plot shows the degree of calibration for different models for various posterior percentiles.

For each percentile considered, the performance was measured by the absolute difference between the posterior percentile and the proportion of observation that falls below that same value. A well-calibrated model would have an expected absolute difference of zero. Here, we used four random splits and the prediction performance was averaged over the four splits. Predictive performance was assessed for the crime data, the racing data, and two simulated data sets. The Hald data contained only 13 observations and thus was excluded here.

For the crime and racing data (Figure 6), our robust model performed best for most percentile values (shaded grey areas). It can be seen that taking into account the transformation and variable selection uncertainty was important. Outlier selection was also important, particularly for the Racing data where there are several outliers. We have also looked at the impact of model

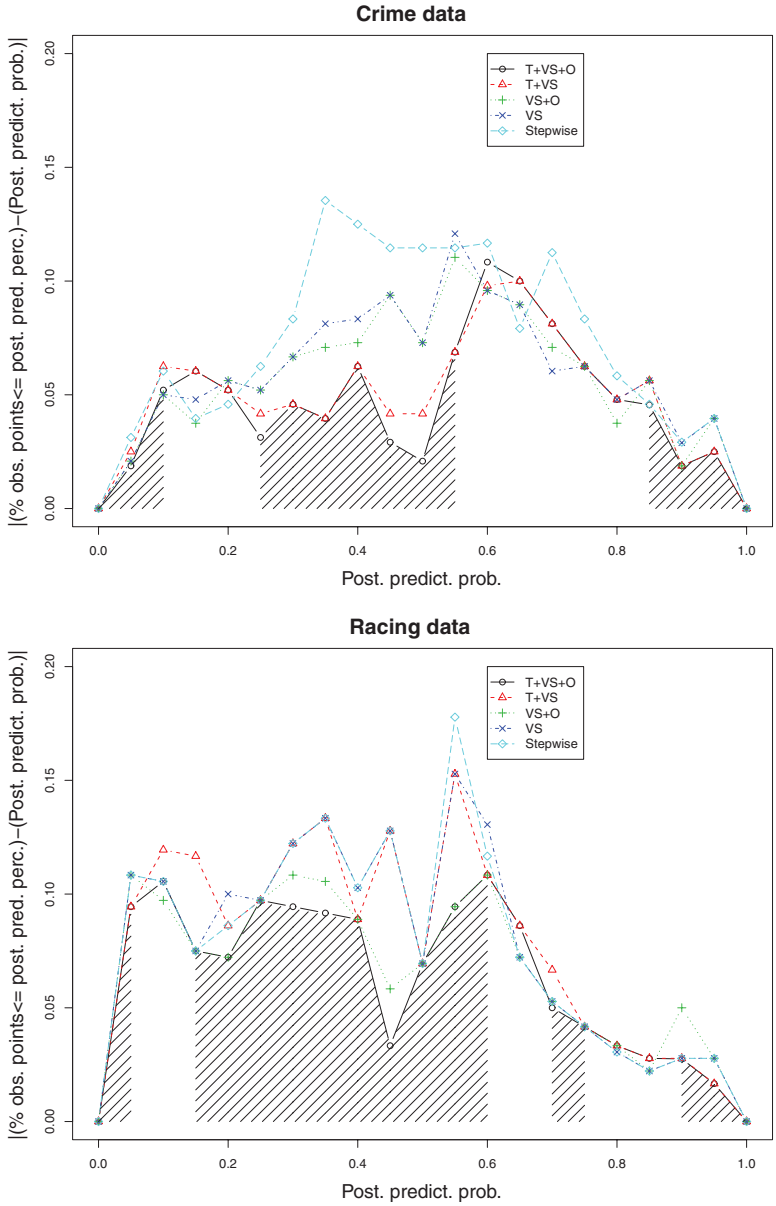

FIGURE 6: Predictive performance curves for each different variants of our model: transformation selection (T), outlier selection (O), variable selection (VS) and stepwise on both the crime data (top) and racing data (bottom). For each percentile value, the performance was measured by the absolute difference between the posterior percentile and the proportion of observation that falls below that same value. The smaller the absolute difference, the better the performance. Areas shaded in grey show where our robust model with transformation selection (T + VS + O) performs the best. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

mis-specification on the calibration plots as shown in on-line Supplementary Material. The simulations along with the real data suggest that the price of including transformation and outlier selection when there are no transformations and no outliers is not great, but that the converse is not true.

6. DISCUSSION

We have introduced a unified approach to the problems of choice of transformations, variable selection, and outlier identification in regression. Using three real examples and a simulated one, we have shown that there can be considerable uncertainty about each of these three modeling choices, and that all three should be done simultaneously. The approach deals with the change of scale induced by each transformation and makes inference (including probabilistic statements) about the size of the effect associated with each predictor.

Similar to Box & Cox (1964) and Hinkley & Runger (1984), our prior formulation for the regression coefficients relies on the transformation parameters to accommodate any change of scale. We have found a data-dependent prior based on the scales of the variables to work well in practice, but other prior formulations have been proposed (Perichi, 1981; Sweeting, 1984).

Here we chose a continuous prior for the Box–Cox transformation parameters. For scientific reasons, such as interpretation of the transformation parameters, an investigator might want to restrict each transformation to a finite number of values. Our approach can be adapted to the case where the parameters are restricted to a finite set. However, we have found our MCMC algorithm to mix poorly in the discrete case due to the possible large difference in likelihood between two very different transformation parameters. It would be possible to derive more efficient algorithms to overcome this problem, such as simulated tempering algorithms (Geyer, 1991; Marinari & Parisi, 1992). However, we have introduced new parameters that we call generalized regression coefficients, which have the same interpretation as the regression coefficients on the original scale of the data. Thus, using our approach one can make inference about the effect of each variable on the original scale, whose interpretation remains valid regardless of the transformation selected. In addition, as pointed out by Carroll (1982) in the context of maximum likelihood inference, restricted (discrete) and unrestricted inferences can lead to significantly different answers, and in such a case the unrestricted approach might be preferable. Finally, in order to test if a transformation should be used, it could be possible to use a mixture of a point mass at one and a continuous distribution for the prior of the transformation parameter. However, as with the discrete prior, this would significantly increase the complexity of our MCMC algorithm, and our simulation suggests that it would not necessarily improve predictive performance. It could have advantages from the point of view of interpretation, however.

We have focused here on Box–Cox transformations, because they remain the most used transformations in regression analysis and because previous literature has studied the assessment of uncertainty when such transformations are used, which is also our concern. Peter McCullagh, in his discussion of Chen, Lockhart & Stephens (2002), has pointed out a more fundamental reason for focusing on power transformations, namely that they are the only transformations that are consistent under aggregation. However, the essence of our methodology is also applicable to other approaches to accommodating nonlinearity in regression, particularly those that have a parametric flavour, including fractional polynomials (Royston & Altman, 1994) and spline regression (Hansen & Kooperberg, 2002). The basic idea of a generalized regression coefficient, defined as $\mathrm{d}[\mathrm{med}(Y)]/\mathrm{d}X_j$, seems potentially even more generally applicable, including nonparametric approaches such as generalized additive models (Hastie & Tibshirani, 1990).

We used a univariate update Gibbs proposal for the regression coefficient, which we found to work well in the examples explored here. However, for very large problems with multicollinearity, an algorithm that performs block updates might be desirable. For example, this could be done by adding an extra step to our MCMC scheme such as the block update described in Nott & Green (2004) or in Gottardo & Raftery (2008).

 DOI: 10.1002/cjs

APPENDIX: MCMC ALGORITHM

The rescaled coefficients used in move 1 are defined as $\beta_j^* = S_{g_{\lambda^*}(\boldsymbol{Y})}/S_{g_\lambda(\boldsymbol{Y})}\beta_j$. Then the candidate values β_0^* and ψ^*, which maximize the likelihood ratio over the set of possible transformations $\beta_0^* = \beta_0 + \delta$ and $\psi^* = \alpha\psi$, where δ and α are given by

$$\delta = \frac{\sum_i \varpi_i\{g_{\lambda^*}(y_i) - g_\lambda(y_i) - \sum_j(\beta_j^* - \beta_j)g_{\lambda_j}(X_{ij})\}}{\sum_i \varpi_i}$$

and

$$\alpha = \frac{\sum_i \varpi_i\{g_\lambda(y_i) - \beta_0 - \sum_j \beta_j g_{\lambda_j}(X_{ij})\}^2}{\sum_i \varpi_i\{g_{\lambda^*}(y_i) - \beta_0^* - \sum_j \beta_j^* g_{\lambda_j}(X_{ij})\}^2}.$$

Note that these are given for the t-distributed model, and the corresponding estimates for the Gaussian model can be obtained by setting $\varpi_i \equiv 1$. The Metropolis-Hastings algorithm is used to decide whether or not to accept the move. The proposal induces a Jacobian term in the acceptance ratio equal to $\alpha\prod_{\{j:\beta_j\neq 0\}}[S_{g_{\lambda^*}(\boldsymbol{Y})}/S_{g_\lambda(\boldsymbol{Y})}]$, due to the change of scale in $\boldsymbol{\beta}$ and ψ.

For move 3a, we first select a candidate transformation λ_j^* using a symmetric proposal centred at the current value λ_j. Based on this candidate value, we carry out a deterministic update for β_j given by $\beta_j^* = S_{g_j(\boldsymbol{X}_j)}/S_{g_j^*(\boldsymbol{X}_j)}\beta_j$ and β_0^* and ψ^* by maximizing the likelihood ratio over the set of possible transformations $\beta_0^* = \beta_0 + \delta$ and $\psi^* = \alpha\psi$, where δ and α are constants in $\mathbb{R}$ and $\mathbb{R}^+$, respectively. Straightforward calculations lead to the optimal values

$$\delta = \frac{\sum_i \varpi_i\{\beta_j g_{\lambda_j}(X_{ij}) - \beta_j^* g_{\lambda_j^*}(X_{ij})\}}{\sum_i \varpi_i}$$

and

$$\alpha = \frac{\sum_i \varpi_i\{g_\lambda(y_i) - \beta_0 - \sum_j \beta_j g_{\lambda_j}(X_{ij})\}^2}{\sum_i \varpi_i\{g_\lambda(y_i) - \beta_0^* - \sum_{k\neq j}\beta_k g_{\lambda_k}(X_{ik}) - \beta_j^* g_{\lambda_j^*}(X_{ij})\}^2}.$$

We then use the Metropolis-Hastings algorithm to decide whether or not to accept these new candidate values. Again this proposal induces a Jacobian term in the acceptance ratio equal to $\alpha S_{g_{\lambda_j}(\boldsymbol{X}_j)}/S_{g_{\lambda_j^*}(\boldsymbol{X}_j)}$, due to the change of variable.

The Gibbs sampler step 3b is performed using the full conditionals for the β_k's given by

$$(\beta_j|\ldots) \sim (1 - w_j^*)\delta_0 + w_j^*\mathrm{N}\left(\frac{\psi\sum_i r_{ij}}{\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j}}, \left(\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j}\right)^{-1}\right), \tag{7}$$

where

$$w_j^* = 1 - \frac{1-w}{1 - w + w\sqrt{\psi_\beta/(\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j})}\exp(0.5(\psi\sum r_{ij})^2/(\psi\sum_i \varpi_i g_{\lambda_j}(X_{ij})^2 + \psi_{\beta_j}))},$$

the residual r_{ij} is defined by

$$r_{ij} = \varpi_i(g_\lambda(y_i) - \beta_0 - \sum_{k \neq j} \beta_k g_{\lambda_k}(X_{ik}))g_{\lambda_j}(X_{ij}),$$

and $\psi_{\beta_j} = (S^2_{g_\lambda(\boldsymbol{Y})}/S^2_{g_{\lambda_j}(\boldsymbol{X}_j)}\sigma^2_\beta)^{-1}$. All other updates are straightforward, involving Gibbs or random walk type proposals, and are not described here.

On-Line Supplementary Material

Additional Supplementary Material can be downloaded at http://www.rglab.org/download/BayesianVSandT_Supp.pdf.

ACKNOWLEDGEMENTS

Both the authors were supported by NIH grant 8 R01 EB002137-02. Raftery was also supported by NICHD grant 1 R01HD O54511, NSF grant IIS0534094, and NSF grant ATM0724721. The authors are grateful to the associate editor and three anonymous referees whose comments greatly improved the article.

BIBLIOGRAPHY

A. C. Atkinson (1986). Aspects of diagnostic regression analysis (discussion of "Influential observations, high leverage points and outliers in linear regression," by S. Chatterjee, A. S. Hadi). *Statistical Science*, 1, 397–402.

A. C. Atkinson (1988). Transformations unmasked. *Technometrics*, 30, 311–318.

P. J. Bickel & K. A. Doksum (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296–311.

G. E. P. Box & D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

G. E. P. Box & D. R. Cox (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209–210.

G. E. P. Box & P. W. Tidwell (1962). Transformation of the independent variables. *Technometrics*, 4, 531–550.

R. J. Carroll (1982). Prediction and power transformation when the choice of power is restricted to a finite set. *Journal of the American Statistical Association*, 77, 908–915.

R. J. Carroll & D. Ruppert (1981). On prediction and the power transformation family. *Biometrika*, 68, 609–615.

R. J. Carroll & D. Ruppert (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, 10, 429–441.

R. J. Carroll & D. Ruppert (1985). Transformations in regression: A robust analysis. *Technometrics*, 27, 1–12.

G. Chen & R. A. Lockhart (1997). Box-Cox transformed linear models: A parameter-based asymptotic approach. *Canadian Journal of Statistics*, 25, 517–529.

G. Chen, R. A. Lockhart & M. A. Stephens (2002). Box-Cox transformations in linear models: Large sample theory and test of normality (with discussion). *Canadian Journal of Statistics*, 30, 177–234.

T. Cheng (2005). Robust regression diagnostics with data transformations. *Computational Statistics & Data Analysis*, 49, 875–891.

R. D. Cook (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15–18.

R. D. Cook & P. C. Wang (1983). Transformations and influencial cases in regression. *Technometrics*, 25, 337–343.

M. G. Dagenais & J. Dufour (1994). Pitfalls of rescaling regression models with Box-Cox transformations. *Review of Economics and Statistics*, 76, 571–575.

N. R. Draper & H. S. Smith (1981). Applied regression analysis. Wiley, New York.

I. Ehrlich (1973). Participation in illegitimate activities: A theoritical and empirical investigation. *Journal of Political Economy*, 81, 521–565.

E. I. George & R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881–889.

E. I. George & R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339–374.

J. Geweke (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5—Proceedings of the Fifth Valencia International Meeting*, pp. 609–620.

C. Geyer (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Sumposium on the Interface*, pp. 156–163.

R. Gottardo & A. E. Raftery (2008). Markov chain Monte Carlo computations with mixture of mutually singular distributions. *Journal of Computational and Graphical Statistics*, 17, 917–943.

A. S. Hadi (1990). A stepwise procedure for identifying multiple outliers in linear regression. *American Statistical Association Proceedings of the Statistical Computing Section*, 137–142.

M. H. Hansen & C. Kooperberg (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science*, 17, 2–51.

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

D. V. Hinkley & G. Runger (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302–309.

D. V. Hinkley & S. Wang (1988). More about transformations and influential cases in regression. *Technometrics*, 30, 435–440.

J. Hoeting, A. E. Raftery & D. Madigan (2001). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics*, 11, 485–507.

J. A. Hoeting & J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in linear model. *Computational Statistics & Data Analysis*, 28, 87–103.

J. A. Hoeting, A. E. Raftery & D. Madigan (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 22, 251–270.

J. Liu, J. L. Zhang, M. J. Palumbo & C. E. Lawrence (2003). Bayesian clustering with variable and transformation selection. In Bayesian statistics 7, J. M. Bernardo, et al. (eds.). Oxford University Press, Oxford, UK, pp. 249–275.

D. Madigan & J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63, 215–232.

E. Marinari & G. Parisi (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19, 451–458.

D. Nott & P. J. Green (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13, 141–157.

L. R. Perrichi (1981). A Bayesian approach to tranformations to normality. *Biometrika*, 68(1), 35–43.

A. E. Raftery, D. Madigan & J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179–191.

P. Royston & D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Applied Statistics*, 43, 429–467.

J. G. Scott & J. O. Berger (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136, 2144–2162.

M. Smith & R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317–343.

T. J. Sweeting (1984). On the choice of prior distribution for the box-cox transformed linear model. *Biometrika*, 71, 127–134.

J. Taylor (1986). The retransformed mean after a fitter power transformation. *Journal of the American Statistical Association*, 81, 114–118.

Received 30 April 2007
Accepted 26 February 2009