Bayesian Variable and Transformation Selection in Linear Regression

Jennifer A. HOETING, Adrian E. RAFTERY, and David MADIGAN

This article suggests a method for variable and transformation selection based on posterior probabilities. Our approach allows for consideration of all possible combinations of untransformed and transformed predictors along with transformed and untransformed versions of the response. To transform the predictors in the model, we use a change-point model, or "change-point transformation," which can yield more interpretable models and transformations than the standard Box–Tidwell approach. We also address the problem of model uncertainty in the selection of models. By averaging over models, we account for the uncertainty inherent in inference based on a single model chosen from the set of models under consideration. We use a Markov chain Monte Carlo model composition (MC^3) method which allows us to average over linear regression models when the space of models under consideration is very large. This considers the selection of variables and transformations at the same time. In an example, we show that model averaging improves predictive performance as compared with any single model that might reasonably be selected, both in terms of overall predictive score and of the coverage of prediction intervals. Software to apply the proposed methodology is available via StatLib.

Key Words: Bayesian model averaging; Change-point transformation; Markov chain Monte Carlo model composition; Model uncertainty; Posterior model probability.

1. INTRODUCTION

Variable and transformation selection are basic components of linear regression model building. Variable selection and transformation selection are typically performed in a specific order with only a subset of the possible models being considered. In the methodology described in this article, we select a set of transformations based on the full model, but then we consider all possible subsets of untransformed and transformed predictors. In addition, we also allow for selection of transformations for the response.

Jennifer Hoeting is Associate Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (E-mail: jah@stat.colostate.edu). Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322. David Madigan is Professor, Department of Statistics, 477 Hill Center, Rutgers University, Piscataway, NJ 08855.

Journal of Computational and Graphical Statistics, Volume 11, Number 3, Pages 485–507
DOI: 10.1198/106186002501

Our approach to transformations of the predictors involves "change-point" transformations which produce interpretable transformations of the predictors. A change-point transformation is a piecewise linear transformation of the original predictor, where there is no change in the expected value of the response above (or below) a certain value of the predictor. Change-point transformations can be more interpretable than the standard power transformations of Box and Tidwell (1962). This article shows how easy model and predictor interpretation can be when change-point transformations are included. For mathematical simplicity we use the Box–Cox class of transformations for the response.

We also consider the problem of model uncertainty in the selection of models. For linear regression models, Raftery, Madigan, and Hoeting (1997) showed the usefulness of *averaging* across sets of predictors, as opposed to conditioning on a single model or set of predictors; accounting for this component of model uncertainty provides improved out-of-sample predictive performance. Here we expand that work by incorporating the usually ignored component of model uncertainty due to transformations. The inclusion of transformations typically yields further improvements in predictive performance.

The next section describes the class of transformations we consider for the response and the predictors. Section 3 describes Bayesian variable and transformation selection and model averaging. Section 4 introduces an example using our methodology. Section 5 assesses predictive performance and shows that model averaging improves predictive performance as compared with any individual model that might reasonably have been selected using standard techniques.

2. TRANSFORMATIONS

When the response is transformed, a Jacobian term enters into the likelihood for the untransformed response. The class of power transformations of Box and Cox (1964) leads to an easily computed Jacobian. We used the Box–Cox class of power transformations for the response.

For transformation of the predictors, however, no new distributional assumptions are necessary, and there are other, more interpretable approaches to transformations than the standard Box–Tidwell power transformations (Box and Tidwell 1962). For these reasons, we adopted a different method for transforming the predictors, consisting of an initial exploratory use of ACE, Breiman and Freidman's (1985) algorithm for regression model linearization, followed by change-point transformations if needed.

2.1 Transformation of the Response via Power Transformations

The Box–Cox class of power transformations changes the problem of selecting a transformation into one of estimating a parameter. The model is $Y^{(\rho)} = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$ and

$$y^{(\rho)} = \begin{cases} \frac{y^{\rho}-1}{\rho} & \rho \neq 0 \\ \log(y) & \rho = 0. \end{cases} \tag{2.1}$$

While the class of power transformations is mathematically appealing, power transformations are typically not easily interpretable unless they are limited to a few possible values of ρ. We have therefore limited ρ to the values $(-1, 0, .5, 1)$, so that the transformed predictors can roughly be interpreted as the reciprocal, the logarithm, the square root, and the untransformed response.

2.2 TRANSFORMATION OF PREDICTORS VIA CHANGE-POINT TRANSFORMATIONS

The standard class of transformations of the *independent* variables to achieve linearity is also the class of power transformations, as proposed by Box and Tidwell (1962). An alternative is the class of "change-point transformations." Change-point transformations produce simplified versions of the predictors where there is no change in the expected value of the response above (or below) a certain value of the predictor. Change-point transformations have been applied in other contexts by Raftery, Lewis, and Aghajanian (1995); Raftery, Lewis, Aghajanian, and Kahn (1996); and Raftery and Richardson (1996).

We use a two-step process to identify a change-point transformation for a predictor. First we run the alternating conditional expectation (ACE) algorithm (Breiman and Friedman 1985) using the untransformed response and predictors as input. We use the output from ACE to suggest the form of the transformation. In the second, confirmatory stage, we use Bayes factors to choose the location of the change point.

The ACE algorithm selects nonlinear transformations for both the response and the predictors to produce an additive model. For the model,

$$g(Y) = \theta + \sum_{j=1}^{p} f_j\left(X_j\right) + \epsilon,$$

ACE chooses nonlinear functions g and $f_1, \ldots, f_p$ to maximize the correlation between the transformed response, $g(Y)$, and the sum of the transformed predictors, $\theta + \sum_{j=1}^{p} f_j\left(X_j\right)$, where θ is an unknown constant. In the ACE algorithm, transformations are found iteratively using a nonparametric smoother until this correlation fails to increase.

We do not directly use the transformations provided by ACE. Rather, we use ACE to *suggest* parametric transformations of the predictors and response. The transformations suggested by ACE for individual predictors often have roughly the form of a change point, with no change in the expected value of the response above (or below) a certain value of the predictor. This type of transformation is often more interpretable than the commonly used power transformations discussed earlier. To choose the change point and to determine the evidence for the change point, we use an approximate Bayes factor which is described in the following.

2.3 APPROXIMATE BAYES FACTOR FOR CHANGE-POINT TRANSFORMATIONS

Raftery (1994) suggested a simple Bayesian approach to estimating and testing for a change point. Consider the case where $f(x)$ is a monotonic ACE transformation for a predictor, x, in a linear regression model. Suppose that a plot of x versus $f(x)$ shows the assumption of a single change point to be reasonable. For example, visual inspection of the plots in Figure 1 indicates that the assumption of a single change point is reasonable for predictors $X1$, $X2$, and $X3$. We call these plots "ACE diagnostic plots." (This figure will be described further below).

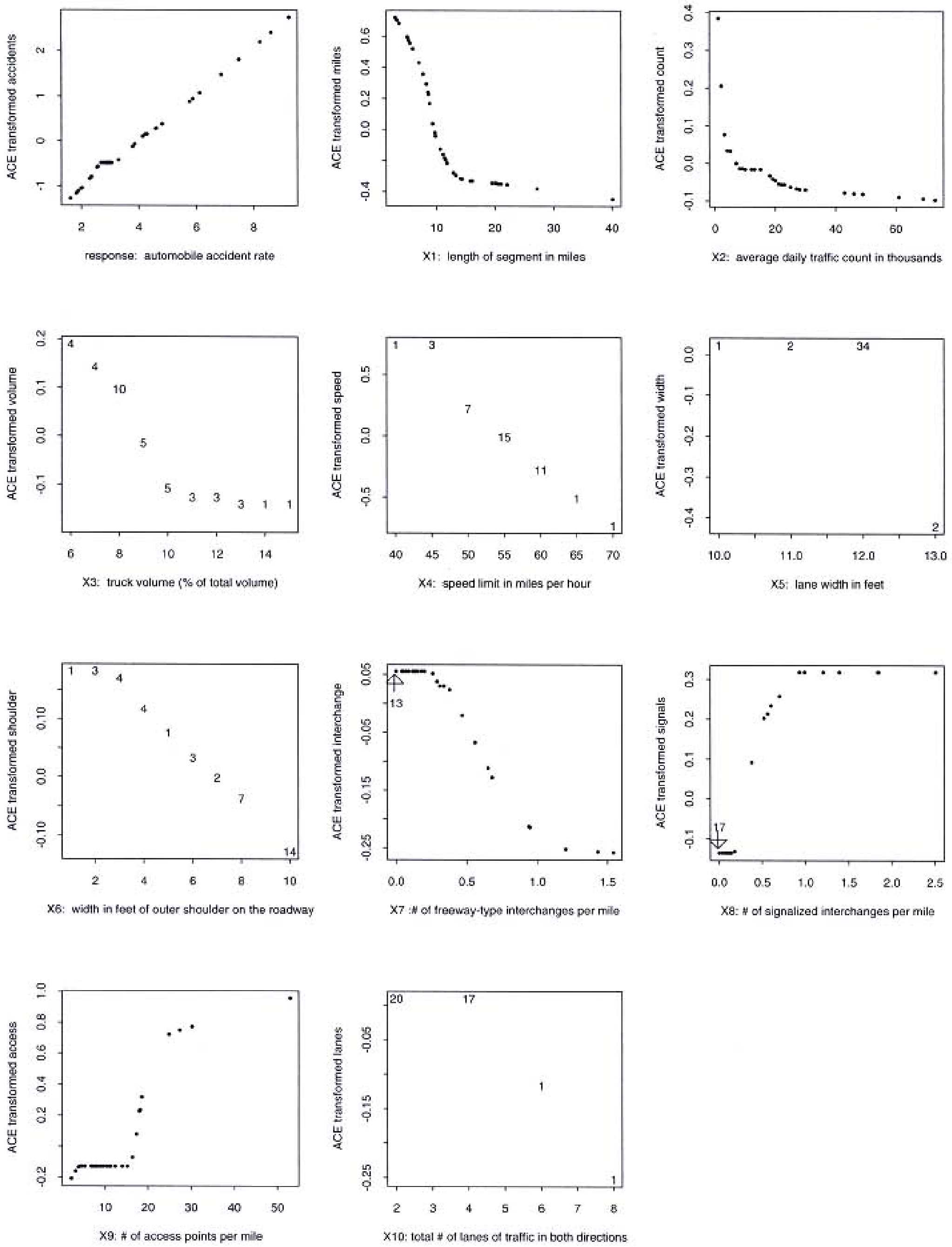

Figure 1. ACE transformations for highway data versus untransformed data. Numbers correspond to the number of observations at those coordinates.

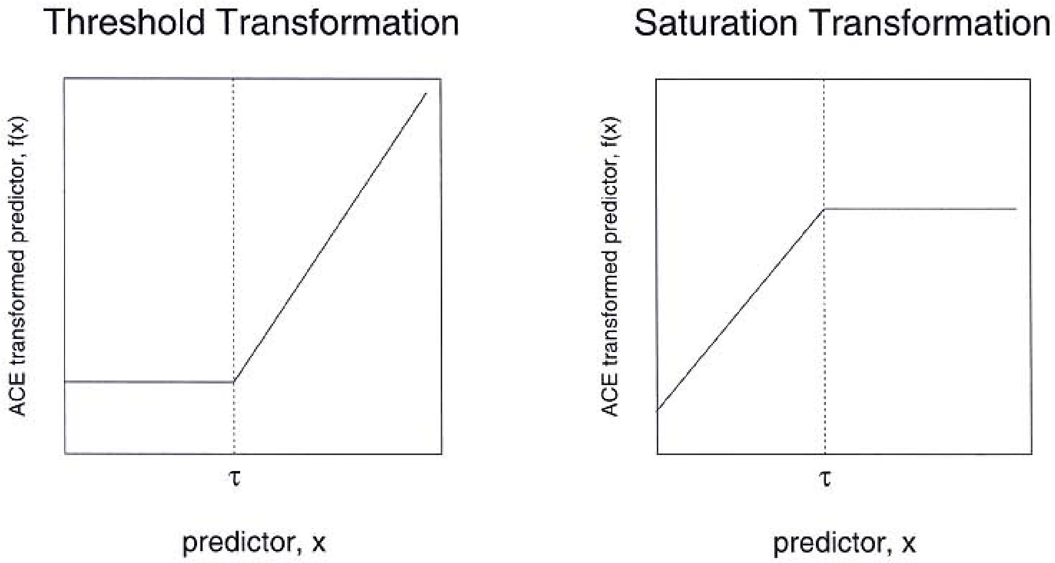

Figure 2. Threshold and saturation transformations.

A change-point model can be rewritten as the linear regression model

$$M_1(\tau): \ f(x) = \beta_0 + \beta_1 x + \beta_2 x_2(\tau) + \epsilon, \tag{2.2}$$

where x is the untransformed predictor, $\epsilon \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, and τ is the change point where

$$x_2(\tau) = \begin{cases} 0 & \text{if} \quad x \leq \tau \\ (x - \tau) & \text{if} \quad x > \tau. \end{cases} \tag{2.3}$$

The no-change-point model is then $M_0 : \beta_2 = 0$.

We call the transformation described in Equation (2.3) a "threshold" transformation where the transformed predictor equals 0 until some threshold at τ. We adopt a threshold transformation if the ACE diagnostic plot has the form of the threshold plot shown in Figure 2. If the ACE diagnostic plot has the form of the saturation plot shown in Figure 2, we use a "saturation" transformation of the form:

$$x_2(\tau) = \begin{cases} (x - \tau) & \text{if} \quad x \leq \tau \\ 0 & \text{if} \quad x > \tau. \end{cases} \tag{2.4}$$

So for a saturation transformation, the transformed predictor equals 0 after some saturation point at τ.

For a given predictor we choose the form of the change-point transformation based on ACE diagnostic plots (saturation or threshold transformation). The next step in the process is to estimate τ, the change-point location. Raftery (1994) showed that if we consider T possible values of a change point at $\tau_i \ \ (i = 1, \ldots, T)$, then an approximate Bayes factor, B_{10}, for comparing the no-change-point model M_0 to the change-point model $M_1(\tau)$, is

$$B_{10} \approx T^{-d/2} \sum_{i=1}^{T} \left\{1 - R^2(\tau_i)\right\}^{-T/2} \text{pr}(\tau_i), \tag{2.5}$$

where d is the number of degrees of freedom involved in the model comparison, $R^2(\tau_i)$ is the multiple coefficient of determination when the change point equals τ_i, and $\mathrm{pr}(\tau_i)$ is the prior probability of a change point at τ_i. We typically consider change points τ equal to the predictor values (excluding the minimum and maximum), so $\tau_i = x_i$ for $i = 1, \ldots, T$ where $T \leq n - 2$.

Following Jeffreys (1961), we adopt the convention that $2\log(B_{10}) \geq 5$ indicates strong evidence for a change point. The most likely observation for the change point to have occurred corresponds to τ_j such that $R^2(\tau_j) = \max_{1 \leq i \leq T} R^2(\tau_i)$. For predictors such that $2\log(B_{10}) \geq 5$, we include the transformed predictor corresponding to $x(\tau)$ as one of the predictors for consideration in potential models.

3. BAYESIAN MODEL SELECTION AND MODEL AVERAGING

3.1 BAYESIAN FRAMEWORK AND SELECTION OF PRIOR DISTRIBUTIONS

Our definition of a model includes the response (with the transformation identified) and the predictors. The predictors may include the untransformed predictors and the change-point transformations of the predictors. This set-up allows a predictor to be included in the model in both its original and transformed forms.

Each model we consider is of the form:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon = X\beta + \epsilon, \tag{3.1}$$

where the observed data on the predictors are contained in the $n \times (p+1)$ matrix X and the observed data on the dependent variable are contained in the n-vector Y. The quantities Y and X in (3.1) may be transformed as described earlier, and p satisfies $0 \leq p \leq 2k$, where k is the original number of untransformed predictors. We assign to ϵ a normal distribution with mean 0 and variance σ^2 and assume that the ϵ's in distinct cases are independent. We consider the $(p+1)$ parameter vector β and σ^2 to be unknown.

Where possible, informative prior distributions for β and σ^2 should be elicited and incorporated into the analysis—see Kadane et al. (1980) and Garthwaite and Dickey (1992). In the absence of expert opinion we seek to chose prior distributions which reflect uncertainty about the parameters and also embody reasonable a priori constraints. We use prior distributions that are proper but reasonably flat over the range of parameter values that could plausibly arise. These represent the common situation where there is some prior information, but rather little of it, and put us in the "stable estimation" case where results are relatively insensitive to changes in the prior distribution (Edwards, Lindman, and Savage 1963). We use the standard normal-gamma conjugate class of priors,

$$\beta \sim \mathrm{N}(\mu, \sigma^2 V),$$

$$\frac{\nu\lambda}{\sigma^2} \sim \chi^2_{\nu}.$$

Here ν, λ, the $(p+1) \times (p+1)$ matrix V and the $(p+1)$-vector μ are hyperparameters to be chosen.

For noncategorical predictor variables we take the individual β's to be independent a priori. This assumption simplifies the selection of hyperparameters and reduces the problem to one of specifying the diagonal elements of V. We center the distribution of β on zero (apart from β_0) and choose $\mu = (\hat{\beta}_0, 0, 0, \ldots, 0)$ where $\hat{\beta}_0$ is the ordinary least squares estimate of β_0. The covariance matrix V is diagonal with entries $(s_Y^2, \phi^2 s_1^{-2}, \phi^2 s_2^{-2}, \ldots, \phi^2 s_p^{-2})$ where s_Y^2 denotes the sample variance of Y, s_i^2 denotes the sample variance of X_i for $i = 1, \ldots, p$, and ϕ is a hyperparameter to be chosen. While using prior distributions which depend on the actual data may at first seem contrary to the idea of a prior, our goal was to use priors that lead to posterior distributions that are similar to those of a person with little prior information. Examples considered to date using similar set-ups (e.g., Hoeting 1994; Hoeting, Raftery, and Madigan 1996; Raftery 1996; and Raftery et al. 1997) suggest that we achieved this objective. Hoeting, Madigan, Raftery, and Volinsky (1999, pp. 412–413) provided an overview of this and other options for prior parameters in this context.

We use the hyperparameter values $\nu = 2.58$, $\lambda = 0.28$, and $\phi = 2.85$. These hyperparameters were chosen to meet the objective of maximizing $\Pr(\sigma^2 \leq 1)$, while keeping $\Pr(\beta_1, \ldots, \beta_p)$ reasonably flat over the unit hypercube $[-1, 1]^p$ and $\Pr(\sigma^2)$ reasonably flat over $(a, 1)$ for some small a. For more details on the Bayesian framework we have adopted here, including details on our treatment of categorical predictor variables and on our choice of hyperparameter values, see Raftery et al. (1997). Fernández, Ley, and Steel (1997, 1998) offered an alternative prior structure.

The integrated likelihood of the response Y based on the proper priors discussed earlier is a noncentral Student's t distribution with ν degrees of freedom, mean $X\mu$, and variance $[\nu/(\nu-2)]\lambda\left(I + XVX^t\right)$. See Raiffa and Schlaiffer (1961) and Hoeting (1994) for details.

Variable and transformation selection is based upon the comparison of posterior model probabilities. The posterior model probability for model M_k is given by

$$\mathrm{pr}(M_k \mid D) = \frac{\mathrm{pr}(D \mid M_k)\mathrm{pr}(M_k)}{\sum_{l=1}^{K} \mathrm{pr}(D \mid M_l)\mathrm{pr}(M_l)},$$

where $\mathrm{pr}(D \mid M_k)$ is the marginal likelihood of model M_k and $\mathrm{pr}(M_k)$ is the prior probability for model M_k. We assume that all models are equally likely a priori.

3.2 ACCOUNTING FOR MODEL UNCERTAINTY

A typical approach to data analysis is to carry out a model selection exercise leading to a single "best" model and to then make inferences as if the selected model were the true model. However, this ignores a major component of uncertainty, namely uncertainty about the model itself (Leamer 1978; Draper 1995; Raftery 1996). As a consequence, uncertainty about quantities of interest can be underestimated. For striking examples of this see Regal and Hook (1991), Raftery (1996), and Kass and Raftery (1995).

The standard Bayesian solution to the problem of model uncertainty involves averaging over all models under consideration. If $\mathcal{M} = \{M_1, \ldots, M_K\}$ denotes the set of all models being considered and if Δ is the quantity of interest such as a future observable or the utility of a course of action, then the posterior distribution of Δ given the data D is

$$\mathrm{pr}(\Delta \mid D) = \sum_{k=1}^{K} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D) \tag{3.2}$$

(Leamer 1978, p. 117). This is an average of the posterior distribution under each model, weighted by the corresponding posterior model probabilities. We call this "Bayesian model averaging" (BMA).

BMA can be difficult for two reasons. First, $\mathrm{pr}(M_k|D)$ involves integrals that can be hard to compute. This problem is resolved analytically in the Bayesian framework described in Section 3.1. Second, the number of terms in (3.2) can be enormous. We overcome this problem using the sampling approach described below. Hoeting et al. (1999) provided an overview of BMA for several classes of models.

In this article we account for the uncertainty in choosing subsets of the predictors and in choosing transformations of the response. We consider the model space $\mathcal{M}$ to be the set of all possible combinations of untransformed and change-point transformed predictors along with untransformed and Box–Cox transformed versions of the response. We chose not to account for the uncertainty involved in choosing the locations for the change points. However, this uncertainty could be accounted for by integrating over all possible change-point locations (or by using the discrete approximation to this by averaging over the change points that are considered).

3.3 Markov Chain Monte Carlo Model Composition (MC^3)

For small problems, it is possible to compute all of the posterior model probabilities for inclusion in the BMA. For large problems the number of possible models can be enormous and it is not feasible to compute the posterior probability of each model. To address this problem, we have adapted the Markov chain Monte Carlo model composition (MC^3) approach of Madigan and York (1995) to do BMA over the space of all variables and transformations.

Let $\mathcal{M}$ denote the space of models under consideration, including all possible combinations of untransformed and transformed predictors along with untransformed and transformed versions of the response. For each model $M \in \mathcal{M}$ in model space, we define a neighborhood as the set of models with either one predictor more or one predictor less than the model M itself, or a different response transformation from a set of Box–Cox transformations. For example, consider a model space with four predictors (1, 2, 3, T3), where T3 is the change-point transformation of predictor 3, and with four possible transformations of the response corresponding to $\rho = (-1, 0, .5, 1)$ in (2.1). If the algorithm is currently visiting the model with response equal to $Y^{(1/2)}$ and predictors 1 and 2, then the neighborhood of this model includes the models shown in Table 1. Other neighborhood constructions are possible, but in our experience this type of approach is easy to program and leads to satisfactory exploration of the model space (Hoeting et al. 1996; Raftery et al. 1997).

Table 1. Model Neighborhood for Transformations. The model space includes four predictors (1, 2, 3, T3) and 4 transformations of the response corresponding to $\rho = (-1, 0, .5, 1)$ in (2.1). The neighborhood of the model with predictors 1 and 2 and with response equal to $Y^{(1/2)}$ is given below.

Predictors	*Response*
1	$Y^{(1/2)}$
2	$Y^{(1/2)}$
1, 2, 3	$Y^{(1/2)}$
1, 2, T3	$Y^{(1/2)}$
1, 2	$Y^{(-1)}$
1, 2	$Y^{(0)}$
1, 2	Y

Once we have defined a neighborhood $\text{nbd}(M)$ for each model $M \in \mathcal{M}$, we define a transition matrix q by setting $q(M \rightarrow M') = 0$ for all $M' \notin \text{nbd}(M)$ and $q(M \rightarrow M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \rightarrow M')$. The new state is then accepted with probability

$$\min\left\{1, \frac{\text{pr}(M' \mid D)}{\text{pr}(M \mid D)}\right\}.$$

Otherwise the chain stays in state M.

Under this set-up, we can construct a Markov chain $\{M(t), t = 1, 2, \ldots\}$ with state space $\mathcal{M}$ and equilibrium distribution $\text{pr}(M_i \mid D)$. We simulate this Markov chain to obtain observations $M(1), \ldots, M(N)$. Under certain regularity conditions, for any function $h(M_i)$ defined on $\mathcal{M}$, the average

$$\frac{1}{N}\sum_{t=1}^{N} h(M(t)) \tag{3.3}$$

converges almost surely to $E(h(M))$ as $N \rightarrow \infty$ (Smith and Roberts 1993). To compute (3.2) in this fashion set $h(M) = \text{pr}(\Delta \mid M, D)$.

3.4 Strategy

Our strategy for Bayesian transformation and variable selection or model averaging is as follows:

1. Run ACE allowing for nonmonotonic transformation of the response and all continuous predictors. Determine from plots of the ACE transformations versus the untransformed values whether monotonic transformations are reasonable.
2. Run ACE again, with monotonic constraints for all variables except those for which strongly non-monotonic relationships were identified in Step 1. Plot the ACE transformations against the untransformed values. For the ACE diagnostic plots that show an approximately linear relationship, no transformation is indicated.
3. If the relationship between the ACE transformed response and the untransformed response is nonlinear, then there is evidence that the response should be transformed.

Table 2. Predictors of Highway Accident Rate

1	length of the segment in miles
2	average daily traffic count in thousands
3	truck volume as a percent of the total volume
4	speed limit (in 1973, before the 55 mph limits)
5	lane width in feet
6	width in feet of outer shoulder on the roadway
7	number of freeway-type interchanges per mile in the segment
8	number of signalized interchanges per mile in the segment
9	number of access points per mile in the segment
10	total number of lanes of traffic in both directions
11	1 if federal aid interstate highway, 0 otherwise
12	1 if principal arterial highway, 0 otherwise
13	1 if major arterial highway, 0 otherwise
14	1 if major collector highway, 0 otherwise

In this case use the MC^3 algorithm described in Section 3.3 with four possible transformations of the response corresponding to $\rho = (-1, 0, .5, 1)$. If the ACE diagnostic plot for the response is roughly linear, transformation of the response is considered to be unnecessary.

4. For an individual predictor, if the relationship between the ACE transformed predictor and the untransformed predictor is nonlinear, then we consider transforming it. For such a predictor, first determine if a threshold transformation or a saturation transformation is appropriate by comparing the ACE diagnostic plot to Figure 2. Then use the approximation to the Bayes factor described in Section 2.3 to choose the change-point value. Finally, include the transformed and untransformed predictors as potential predictors in a model selection procedure or a model averaging procedure.

4. HIGHWAY ACCIDENTS

We consider the highway accident data from Weisberg (1985, Table 8.1). The dependent variable is the automobile accident rate on 39 highway sections, and there are 14 potential predictor variables (Table 2).

Weisberg (1985) hypothesized that the categorical variables for type of highway (predictors 11–14) might be important predictors of accident rate because the type of highway is defined by the source of financial support used by the Highway Department to maintain the roads. Weisberg allowed only three of the dummy variables in the model to avoid having linearly dependent columns in the predictor matrix. He included or excluded the first three dummy variables (predictors 11–13) as a group. In contrast, we used all four possible dummy variables for type of highway as inputs in MC^3, but allow a maximum of three of them to be in any one model. Our approach is more flexible than Weisberg's (1985) method of considering only models with either all three dummy variables included or all three dummy variables excluded, because it allows for consideration of any subset of the dummy variables. This could potentially lead to models that are easier to interpret if, say,

Table 3. Transformed Predictors of Highway Accident Rate. A "T" in front of a number indicates that the predictor has been transformed. 2 log (B_{10}) is the overall Bayes factor for a change point versus no change point (over the range of that predictor).

	Predictor	*Change point τ*	*2 log (B_{10})*
T1	length of the segment in miles	12.91	148
T2	average daily traffic count in thousands	4.00	89
T3	truck volume as a percent of the total volume	11.00	21
T7	# freeway-type interchanges per mile	0.20	85
T8	# signalized interchanges per mile	0.70	60
T9	# access points per mile	10.30	57

only one of the dummy variables enters into the model.

Unlike Weisberg (1985), we did not force the length of the segment in miles (predictor 1) into all models. He contended that this variable would be important because it should be negatively correlated with accident rate. That is, if you increase the length of a segment by one mile, it is unlikely that any accidents would have occurred in a short segment as accidents are rare, however, the rate of accidents would be lowered because of the increase of the length of segment. We have chosen not to force this variable into all models and instead to allow the data themselves to determine whether it should be included.

4.1 HIGHWAY DATA: CHANGE-POINT TRANSFORMATIONS

To choose transformations for the highway data, we first ran ACE in S-Plus© without monotonicity constraints. Plots of the transformed variables from the ACE output against the original variables indicate that monotonic transformations are reasonable for the response and the continuous predictors. Next we ran ACE with all transformations constrained to be monotonic. The ACE diagnostic plots are shown in Figure 1. The ACE diagnostic plots for the response, and predictors 4, 6, and 10 exhibit an approximately linear trend, suggesting that no transformation is necessary. The ACE diagnostic plot for predictor 5 is not linear. However, there are only two observations that deviate from a linear relationship, so we chose to consider this relationship to be linear.

For predictors 1, 2, 3, 7, 8, and 9, we concluded that the assumption of a single change point, as described in equation (2.2), is reasonable. As suggested by Figure 1, we modeled predictors 1, 2, 3, and 8 using a saturation effect and predictors 7 and 9 using a threshold effect. To choose a change point for each transformation we used the Bayes factor approach of Section 2.3. For each predictor we considered each value of the predictor (excluding the maximum and minimum values) to be a potential change point. We assumed that all change point locations were equally likely a priori. The overall Bayes factor for a change point versus no change point indicated strong evidence for a transformation for all six predictors (Table 3).

As an example, we show the values of $R^2(\tau)$ for length of the segment in Figure 3. The change point indicated for this predictor, which corresponds to the maximum R^2 value, is 12.9 miles. The change-point values corresponding to the maximum R^2 are given in Table 3.

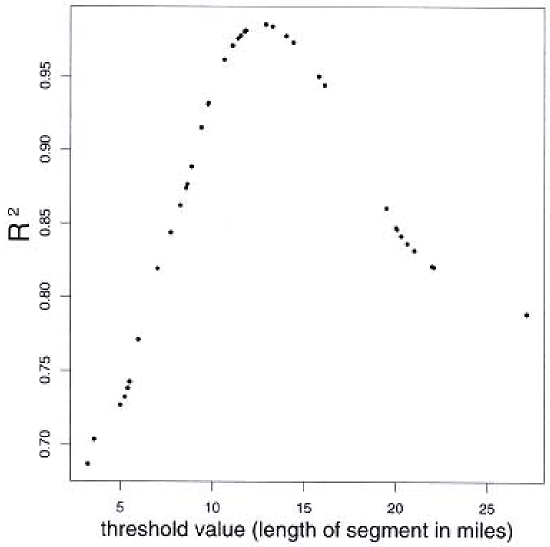

Figure 3. R^2 *for change-point model for length of segment in miles (predictor 1).*

These change points are shown in the context of the ACE diagnostic plots in Figure 4. All of the change points indicated by the Bayes factor approach appear to be reasonable, with the possible exception of the threshold for predictor 9.

For predictor 9 we would have anticipated a change point at 15.2, indicated by the dotted line in Figure 4. The R^2 values were quite similar for these two change points with $R^2(10.3) = 0.84$ and $R^2(15.2) = 0.82$. In fact the R^2 values for predictor 9 were quite similar across the entire range of the data. This lack of distinction between the R^2 values is probably due to the nonlinear pattern exhibited to the right of the dotted line in Figure 4. Nonetheless, we used 10.3 which is the value indicated by the Bayes factor approach ($2\log(B_{10})$) so that our method for choosing a change point would be uniform for all predictors. We will discuss this issue further in Section 7.

Some of the change-point transformations (see Table 3) have practical interpretations. For example, in Figure 4 the predictor for truck volume as a percent of total volume (3) exhibits a saturation effect with a change point at 11%. This can be interpreted to mean that after truck volume increases over 11% of the total traffic volume, accident rate is constant as a function of truck volume. Predictor 7 exhibits a threshold effect, with accident rate constant until the number of freeway-type interchanges per mile reaches 0.18. Above 0.18 there is a linear relationship between predictor 7 and accident rate. The other transformations can be interpreted similarly.

4.2 Highway Data: Transformation and Variable Selection

Weisberg's highway data includes 14 candidate predictors of accident rate. In addition to the 14 candidate predictors, we also included the six transformed predictors T1, T2, T3, T7, T8, and T9. Thus, we used 20 potential predictors as inputs in MC^3. Standard diagnostic checking (e.g., Weisberg 1985) for the full model, including transformed and untransformed predictors, did not reveal any gross violations of the assumptions underlying normal linear regression.

For MC^3, 3156 different models were visited in 20,000 iterations. The models with the ten largest posterior model probabilities for MC^3 are given in Table 4. The posterior model

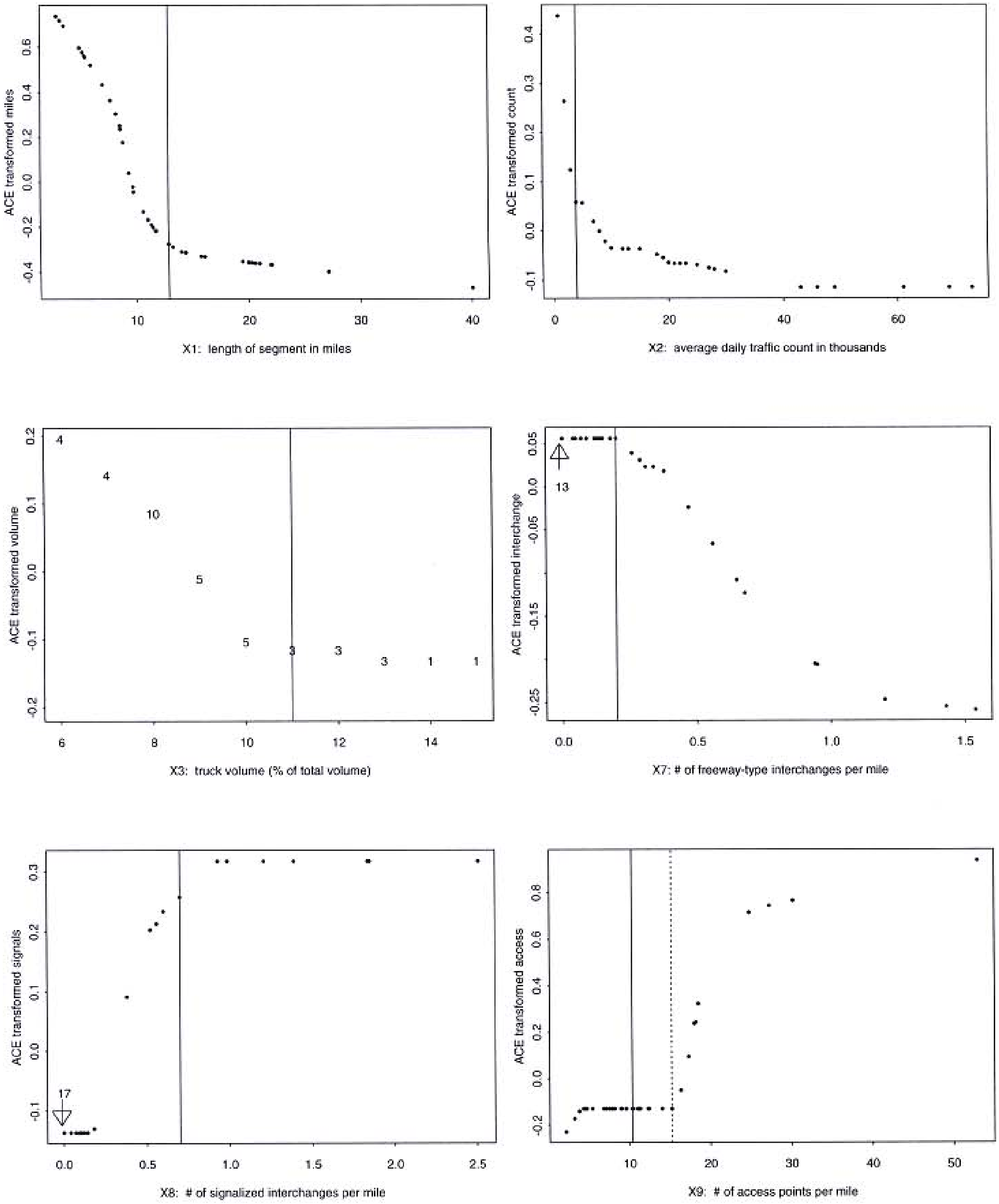

Figure 4. ACE transformations for highway data versus untransformed data. The change point indicated by the Bayes factor for change-point transformations is indicated by the solid vertical line. The dotted line in the plot for X9 is the anticipated change point. See text for details. Numbers correspond to the number of observations at those coordinates.

Table 4. Highway Data, Full Dataset: Models with the ten highest MC^3 posterior model probabilities.

									Number		*Posterior*
									of	R^2	*Model*
				Predictor					*Predictors*	*(%)*	*Prob. (%)*
4				12	T1		T8	T9	5	82	4
4			9	12	T1		T8		5	83	4
			9	12	T1		T8		4	79	2
4				12	T1		T8		4	79	2
4				12	T1	T2	T8	T9	6	84	2
4		8		12	T1		T8	T9	6	83	1
4			9	12	T1		T8	T9	6	83	1
4		8		12	T1	T2	T8	T9	7	85	1
4			9	12	T1	T2	T8		6	84	1
	6		9	12	T1		T8		5	77	1

probabilities for MC^3 were relatively small indicating a great deal of model uncertainty for this dataset.

The total posterior probability that the coefficient for each predictor does not equal 0 for MC^3 is given in Table 5. $\Pr(\beta_i \neq 0|D)$ is the sum of the posterior model probability across models with predictor i. For a predictor i, the probability that the coefficient for the transformed predictor i or the untransformed predictor Ti is not equal to 0, $\Pr(\beta_i \neq 0 \cup \beta_{Ti} \neq 0|D)$, is given in the last column in Table 5. For the untransformed predictors, speed limit (4) and the indicator variable for principal arterial (12) received strong support from the data. For the transformed predictors, length of segment (T1) and the number of signalized interchanges (T8) had high $\Pr(\beta_i \neq 0|D)$. Predictors 1, 8, and 9 appeared in models with high posterior probabilities either as transformed or untransformed predictors. These results indicate that predictors T1, 4, T8, 9 or T9, and 12 are important

Table 5. Highway Data, Full Dataset: $\Pr(\beta_i \neq 0 \mid D)$, expressed as a percentage for MC^3. $\Pr(\beta_i \neq 0 \mid D)$ is the sum of the posterior model probability across models for predictor i. The last column gives the probability of the coefficient that the coefficient for transformed and/or untransformed predictor is not equal to 0, $\Pr(\beta_i \neq 0 \cup \beta_{Ti} \neq 0|D)$.

Predictor number	*Predictor*	*Untrans-formed*	*Trans-formed*	*Trans. or untrans.*
1	length of segment	5	100	100
2	average daily traffic	5	29	33
3	truck volume	13	15	26
4	speed limit	62	–	62
5	lane width	3	–	3
6	shoulder width	26	–	26
7	# interchanges	6	6	11
8	# signalized interchanges	27	91	93
9	# access points	46	59	92
10	# lanes	4	–	4
11	interstate highway	8	–	8
12	principal arterial	75	–	75
13	major arterial	9	–	9
14	major collector	7	–	7

Table 6. Highway Data, Full Dataset: Mean and Standard Deviation of the BMA Marginal Posterior Distribution for the Regression Coefficients. The results given here are based on standardized data (columns have means equal to 0 and variances equal to 1).

Predictor		*Mean*	*SD*
number	*Predictor*	$\beta\|D$	$\beta\|D$
1	length of segment		
	untransformed	−0.0004	0.03
	transformed	−0.3759	0.09
2	average daily traffic		
	untransformed	−0.0033	0.03
	transformed	−0.0489	0.09
3	truck volume		
	untransformed	−0.0154	0.06
	transformed	−0.0207	0.07
4	speed limit	−0.1737	0.16
5	lane width	−0.0017	0.02
6	shoulder width	−0.0535	0.11
7	# interchanges		
	untransformed	0.0010	0.07
	transformed	-0.0064	0.08
8	# signalized interchanges		
	untransformed	−0.0736	0.17
	transformed	0.3583	0.19
9	# access points		
	untransformed	0.1217	0.24
	transformed	0.2154	0.25
10	# lanes	−0.0032	0.03
11	interstate highway	0.0064	0.05
12	principal arterial	−0.2225	0.16
13	major arterial	−0.0033	0.05
14	major collector	0.0064	0.03

predictors of highway accident rate. Indeed, these predictors typically appear in the models with the highest posterior model probabilities (Table 4). It is interesting to note that individually, there is not strong support for nonzero coefficients for predictors 9 and T9, untransformed and transformed values for the number of access points in a segment. However, there is indeed support for the inclusion of the number of access points in some form as $\Pr(\beta_9 \neq 0 \cup \beta_{T9} \neq 0|D)$=0.92.

The mean and standard deviation of the BMA marginal posterior distribution for each of the coefficients is given in Table 6. Each posterior distribution is a mixture of noncentral Student's t distributions. The estimates in Table 6 directly incorporate model uncertainty. Taking account of model uncertainty tends to shrink the parameter estimates towards 0 and tends to increase the standard deviation of the estimate.

The shrinkage effect can be demonstrated via closer examination of the BMA posterior distribution associated with the coefficient for speed limit (4). Figure 5 shows the

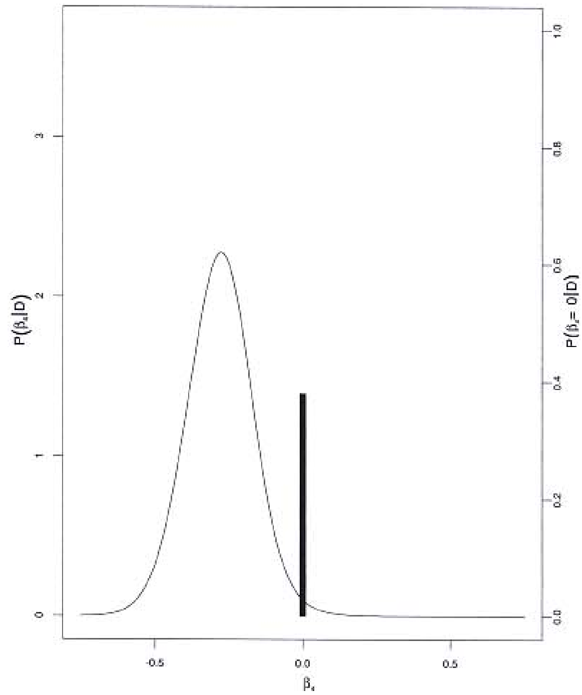

Figure 5. The BMA marginal posterior density for β_4, the coefficient for speed limit. The spike corresponds to $P\big(\beta_4 = 0|D\big) = 0.38$. The vertical axis on the left corresponds to the posterior distribution for β_4 and the vertical axis on the right corresponds to the posterior distribution for β_4 equal to 0. The density is scaled so that the maximum density is equal to $P\big(\beta_4 \neq 0|D\big)$ on the right axis.

marginal posterior distribution for the coefficient for β_4. The spike in the plot of the posterior distribution corresponds to $\mathrm{P}\big(\beta_4 = 0|D\big) = 0.38$. This spike is an artifact of our approach as we consider models with a predictor fully excluded from the model. The averaged curve in Figure 5 is centered at -0.28. Weighting this by $\mathrm{P}\big(\beta_4 \neq 0|D\big) = 0.62$ gives $-0.28 \times 0.62 = -0.17$. This is the mean estimate of the parameter shown in Table 6. Thus, accounting for model uncertainty has the effect that the parameter estimate is closer to 0. This shrinkage effect is similar to other shrinkage estimates such as ridge regression.

The inclusion of transformations in the model selection procedure can lead to a better understanding of the relationship between the predictors and the response. Predictors T1 and T8 are both coded as saturation transformations. For predictor T1, the relationship between segment length and automobile accident rate per segment can be described as linear up to a threshold at 12.9 miles. For larger segment lengths, the expected value of the response does not change. The relationship between the number of signalized interchanges per mile and the response can be described similarly.

5. ASSESSMENT OF PREDICTIVE PERFORMANCE

5.1 Overview

We compare the quality of the predictions from model averaging with that of the predictions from any single model that an analyst might reasonably have selected. To select the single models we use three standard variable selection techniques: Efroymson's stepwise method (Miller 1990), minimum Mallow's C_p, and maximum adjusted R^2. Efroymson's stepwise method is like forward selection except that when a new variable is added to the subset, partial correlations are considered to see if any of the variables currently in the subset should be dropped. Similar hybrid methods of forward selection are found in most standard statistical computer packages. For the stepwise methods, we used significance levels of .05 and of .15; the latter corresponds roughly to choosing the model with the best value of AIC.

To measure performance we did a cross-validation analysis. We omitted each observation, in turn, and ran MC^3 or a standard model selection procedure using the remaining observations. We assess performance by comparing the predicted value for each hold-out observation to the observed value.

We use two measures of performance, both of which are based on the posterior predictive distribution. The first measure of predictive ability is the coverage for 95% prediction intervals. We define predictive coverage as the proportion of observations that fall in their corresponding 95% posterior prediction intervals.

The second measure of predictive ability is the logarithmic scoring rule of Good (1952) where for each event A which occurs, a score of $-\log\{\mathrm{pr}(A)\}$ is assigned. For our example, the "predictive log score" is based on the posterior predictive distribution suggested by Geisser (1980). The predictive log score is a combined measure of predictive bias (a systematic tendency to predict on the low side or the high side) and calibration (a systematic tendency to over- or understate predictive accuracy). The smaller the predictive log score for a given model or model average, the better the predictive performance. See Raftery et al. (1997) for details of the computation of predictive coverage and of the predictive log score.

5.2 HIGHWAY DATA: PREDICTIVE PERFORMANCE WHEN TRANSFORMATIONS ARE CONSIDERED

For all methods, we used the set of transformations that were selected for the full dataset (Table 3). Initial diagnostics indicated that these transformations were reasonable for the cross-validation analysis. To perform cross validation for the model averaging approach, we omitted each of the 39 observations in this dataset, in turn, and ran MC^3 for 20,000 iterations using the remaining observations. For the standard methods, we performed a similar analysis, omitting each observation, in turn, and then selecting a single best model for each standard model selection approach.

For MC^3 and models selected using standard approaches, we computed a 95% posterior prediction interval for each omitted observation. We report the number of observations that fell in their corresponding 95% prediction intervals (Table 7). For the model averaging approach, 90% of the observations in the performance set fell in the MC^3 95% prediction interval. Standard variable selection methods had worse predictive performance with predictive coverage as low as 79%.

Table 7. Highway Data: Cross-Validated Predictive Performance. Comparison between model averaging and standard variable selection techniques. The "Number of Observations" column is the number of observations that fell inside the corresponding 95% prediction interval. The percentage values shown for the stepwise procedures correspond to the significance levels for the F-to-enter and F-to-delete values. For example, F = 3.84 corresponds approximately to the 5% level.

Method	*Predictive Coverage*	*Number of Observations*	*Predictive log score*
Bayesian model averaging via MC^3	90	35	43.2
Stepwise regression (5%)	85	33	54.6
Stepwise regression (15% – AIC)	82	32	50.0
Mallow's C_p	79	31	53.3
Adjusted R^2	79	31	50.9

We also computed the predictive log score for all methods. The model averaging approach has superior predictive performance as measured by the predictive log score. The standard methods that select single models produce higher predictive log scores. The log score can also be interpreted on a "per observation" basis. A difference in predictive log score of 6.8 can be interpreted as an improvement in predictive performance per observation by a factor of $\exp(6.8/39) = 1.19$ or by about 19%. Thus, the model averaging approach predicts the accident rate about 19% more effectively than single models chosen using stepwise regression (with F-to-enter and F-to-delete values corresponding to the 15% significance level). Similarly, on a per-observation basis, the model averaging approach predicts the accident rate about 22%, 30% and 34% better than adjusted R^2, Mallow's C_p, and stepwise regression (with F-to-enter and F-to-delete values corresponding to the 5% significance level), respectively.

5.3 Highway Data: Predictive Performance When Transformations are not Considered

To investigate whether including both transformed and original predictors in model selection leads to over-fitting the data, we compared MC^3 results when transformations are included and excluded. Instead of doing a leave-one-out cross-validation analysis, we randomly selected 30 observations (the "training set") and used these to run MC^3. For the training set, we used ACE diagnostic plots to investigate transformations as described in Section 4.1. Transformation of predictors 7 and 8 was indicated. The threshold values for predictors 7 and 8 corresponding to the maximum R^2 in equation (2.5) were the same as the threshold values indicated for the entire dataset (see Table 3). To investigate predictive performance, we computed the predictive log score on the 9 observations that were not included in the training set, namely observation numbers 9, 10, 21, 24, 25, 27, 33, 35, and 39, as listed in Weisberg (1985, Table 8.1).

Table 8 gives the predictive log score for MC^3 and the top 10 models by posterior model probability for the case when transformations are included. For this split of the data, MC^3 produced better out-of-sample predictions than most of the models with high posterior

Table 8. Highway Data With Transformations, 30 Observations in Training Set: MC^3 Predictive Performance. Models with the ten highest posterior model probabilities for MC^3.

									Posterior	Predictive
									Model	Log
				Predictor					Prob. (%)	Score
	4	7	8	9					2	14.4
1	4		8			12			2	13.1
	4		8	9			T7		1	14.2
	4		8						1	14.8
	4	7	8	9			T7		1	13.9
1	4	7	8	9		12			1	11.8
	4		8	9	11				1	12.3
	4	7		9				T8	1	11.8
1				9		12		T8	1	11.4
1	4		8	9	11				1	12.5
MC^3 model averaging										10.5

model probability. For example, MC^3 outperformed the model with the highest posterior probability by 3.9 points.

Table 9 is similar to Table 8 except that transformations were not included in the analysis. Comparison of these tables reveals that the predictive log score for MC^3 model averaging when transformations are included is slightly better than the predictive log score when transformations are not included. If over-fitting was occurring, one would expect the score to be worse.

These results indicate that modest gains in the predictive log score are realized when transformations are considered. We would expect the gains would be larger in examples where the models with the highest model probabilities included transformations.

Table 9. Highway Data With no Transformations, 30 Observations in Training Set: Models with the ten highest posterior model probabilities for MC^3.

									Posterior	Predictive
									Model	Log
				Predictor					Prob. (%)	Score
	4	7	8	9					4	14.4
1	4		8				12		4	13.1
	4		8						2	14.8
1	4		8	9			12		2	11.8
	4		8	9		11			2	12.3
1	4		8	9		11			2	12.5
1				9					1	10.8
1	4		8	9			12	13	1	13.7
	4		8	9					1	11.2
	4			9	10				1	11.7
MC^3 model averaging										10.7

6. SOFTWARE FOR IMPLEMENTING MC^3

BMA is a set of S-Plus© functions which can be obtained free of charge via the World Wide Web at http://lib.stat.cmu.edu/S/bma, or by sending the e-mail message "send BMA from S" to statlib@stat.cmu.edu.

The program MC3.REG performs Markov chain Monte Carlo model composition for linear regression allowing for Bayesian variable selection, outlier identification, and transformation selection. The set of programs implements the MC^3 algorithm described in Section 3.3.

7. DISCUSSION

Averaging over models which include transformations produces better out-of-sample predictive performance than models chosen using standard techniques, in the example. Change-point transformations can lead to better understanding of the relationship between the predictors and the response.

In addition to variable and transformation selection, there is also uncertainty involved in the identification of outliers in regression. In Hoeting et al. (1996), we considered variables and outliers. We showed in an example that accounting for this uncertainty via BMA improves predictive performance as compared with any single set of variables and outliers that could reasonably be selected. To broaden the flexibility of our proposed methodology as well as to improve our ability to account for model uncertainty, a variable, transformation, and outlier selection approach which combines variable selection, outlier identification, and transformation selection has also been proposed (Hoeting 1994).

Clyde, DeSimone, and Parmigiani (1996) proposed a method for model mixing based on a re-expression of the space of models in terms of an orthogonalization of the design matrix. George and McCulloch (1993) developed the stochastic search variable Selection (SSVS) method which is similar in spirit to MC^3. So far, this has been applied to variable selection in regression but not to transformations.

Volinsky (1997) and others have noted the relationship between BMA and ridge regression. By shrinking regression parameter estimates towards zero, ridge regression accounts for over-confidence in the full model. In contrast, in the BMA estimates of the regression parameters shrinkage occurs via the posterior model probabilities. Volinsky showed that ridge regression can outperform BMA under certain conditions in simulation studies. He proposes combining BMA and ridge regression by using a "ridge regression prior" in BMA. This corresponds to the prior $\beta \sim N\left(0, \frac{\sigma^2}{k} I\right)$, where k is the ridge regression shrinkage parameter. Clyde and George (1999) have shown that a closely related approach, empirical Bayes BMA, works well for nonparametric regression using wavelets.

In this article we assessed convergence of the Markov chain by comparing the results of MC^3 to models selected using standard model selection procedures. Assessing convergence of the MC^3 procedure is still an open problem. Available software such as CODA (Best, Cowles, and Vines 1995) is not completely applicable bewcause it focuses on convergence of parameters. A main concern here is whether or not the model space has been explored. While use of CODA diagnostics on Equation (3.3) is possible, this software is not well suited to address the question of whether or not the model space has been adequately covered.

The first step in the overall strategy in Section 3.4 involves running the ACE algorithm to determine whether monotonic transformations of the predictors is a reasonable assumption. If this is not reasonable, then the change point approach to transform the predictors may not be reasonable. In our experience, the monotonic assumption for ACE transformations is often appropriate. If it is not, another approach would be to include power transformations of the predictors in the MC^3 algorithm in a similar manner to that proposed for the response.

In this article we use an automatic method to choose change-point locations that is based on the Bayes factor. In the example above, the change point for predictor 9 indicated by the automatic approach was somewhat questionable. We used the change point indicated by the Bayes factor, however, to maintain the fully automatic approach for choosing change-point locations for all examples. One could use the Bayes factor approach as only a guide to choosing the location for a change point in conjunction with diagnostic plots like those in Figure 1. This choice is up to the user.

We have found ACE and the change-point transformations useful in several contexts, but other more elaborate approaches such as Friedman and Silverman's (1989) adaptive algorithm to optimize over the number and location of spline knots (TURBO), Friedman's (1991) multivariate adaptive regression splines (MARS), and Hastie and Tibshirani's (1990) adaptive backfitting algorithm for generalized additive models (BRUTO) might also be useful. Our change-point transformations for multiple regression models could also be considered to be a special case of the regression splines developed by Kooperberg, Stone, and Truong (1995). They used a linear spline approach to estimate the conditional hazard function of censored response data with one or more covariates. Our approach is also similar to Smith and Kohn's (1996) work on nonparametric regression. These authors suggested a Bayesian approach to select regression spline knots, variables and Box–Cox transformations of the response variable. None of these authors accounted for model uncertainty in their work, but our approach could be generalized to account for model uncertainty in these contexts.

There also exists a substantial computer science literature addressing feature selection and transformation usually under the heading "Constructive Induction." For an overview we refer the reader to a special issue of *IEEE Intelligent Systems*, Vol. 13, No. 2, March/April 1998 and the references therein. Much of this literature concerns itself with improving predictive performance through the inclusion of new predictors that are functions of multiple existing predictors; see, for example, Pazzani (1996). Genetic algorithms are commonly used to deal with the resultant search problem.

ACKNOWLEDGMENTS

Research was supported by ONR contract N-00014-91-J-1074 (Raftery and Hoeting), NSF grant DMS 92111627 (Madigan), and NSF grant DMS-9806243 (Hoeting).

[Received June 1999. Revised May 2001.]

REFERENCES

Best, N. G., Cowles, M. K., and Vines, S. K. (1995), *CODA Convergence Diagnosis and Output Analysis Software for Gibbs Sampler Output*,

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Series B, 26, 211–246.

Box, G. E. P., and Tidwell, P. W. (1962), "Transformation of the Independent Variables," *Technometrics*, 4, 531–550.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–619.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction via Orthoganalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208.

Clyde, M., and George, E. I. (1999), "Empirical Bayes Estimation in Wavelet Nonparameteric Regression," in *Bayesian Inference in Wavelet Based Models*, eds. P. Muller, and B. Vidakovic, New York: Springer-Verlag, pp. 309-322.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society*, Series B, 57, 45–97.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Fernandez, C., Ley, E., and Steel, M. F. J. (1997), "Statistical Modeling of Fishing Activities in the North Atlantic," Technical Report, Department of Econometrics, Tilburg University, the Netherlands.

——— (1998), "Benchmark Priors for Bayesian Model Averaging," Technical Report, Department of Econometrics, Tilburg University, the Netherlands.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–141.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additvie Modelling" (with discussion), *Technometrics*, 31, 3–39.

Garthwaite, P. H., and Dickey, J. M. (1992), "Elicitation of Prior Distributions for Variable-Selection Problems in Regression," *Annals of Statistics*, 20, 1697–1719.

Geisser, S. (1980), Discussion of "Sampling and Bayes' Inference in Scientific Modeling and Robustness" by G.E.P. Box, *Journal of the Royal Statistical Society*, Series A, 143, 416–417.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Series B, 14, 107–114.

Hahn, G. J., and Meeker, W. Q. (1993), "Assumptions for Statistical Inference," *The American Statistician*, 47, 1–11.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hoeting, J. A. (1994), "Accounting for Model Uncertainty in Linear Regression," PhD thesis, University of Washington.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–417; corrected version of the paper available at http://www.stat.washington.edu/www/research/online/hoetingl999.pdf.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1996), "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression," *Journal of Computational Statistics and Data Analysis*, 22, 251–271.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), London: Oxford University Press.

Kadane, J. B., Dickey, J. M., Winker, R. L., Smith, W. S., and Peters, S. C. (1980), "Interactive Elicitation of Opinion for a Normal Linear Model," *Journal of the American Statistical Association*, 75, 845–854.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995), "Hazard Regression," *Journal of the American Statistical Association*, 90, 78–94.

Leamer, E. E. (1978), *Specification Searches*, New York: Wiley.

Madigan, D., and York, J. (1995),"Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Miller, A. J. (1990), *Subset Selection in Regression*, London: Chapman and Hall.

Pazzani, M. (1996), "Constructive Induction of Cartesian Product Attributes," in *Information, Statistics and Induction in Science*, Melbourne, Australia.

Raftery, A. E. (1994), "Change Point and Change Curve Modeling in Stochastic Processes and Spatial Statistics," *Journal of Applied Statistical Science*, 1, 403–424.

——— (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models," *Biometrika*, 83, 251–266.

Raftery, A. E., Lewis, S. M., and Aghajanian, A. (1995), "Demand or Ideation? Evidence From the Iranian Marital Fertility Decline," *Demography*, 32, 159–182.

Raftery, A. E., Lewis, S. M., Aghajanian, A., and Kahn, M. J. (1996), "Event History Analysis of World Fertility Survey Data," *Mathematical Population Studies*, 6, 129–153.

Raftery, A. E., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Raftery, A. E., and Richardson, S. (1996), "Model Selection for Generalized Linear Models via Glib, with Application to Epidemiology," in *Bayesian Biostatistics*, eds. D. A. Berry and D. K. Stangi, New York: Dekker, pp. 321–354.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge, MA: The MIT Press.

Regal, R., and Hook, E. B. (1991), "The Effects of Model Selection on Confidence Intervals for the Size of a Closed Population," *Statistical Medicine*, 10, 717–721.

Smith, A. F. M., and Roberts, G. 0. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society*, Series B, 55, 3–23.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

Volinsky, C. T. (1997), "Bayesian Model Averaging for Censored Survival Models," PhD thesis, Univ. of Washington.

Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), New York: Wiley.