BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian Treed Generalized Linear Models

HUGH A. CHIPMAN
University of Waterloo, Canada
hachipman@uwaterloo.ca

EDWARD I. GEORGE
University of Pennsylvania, USA
edgeorge@wharton.upenn.edu

ROBERT E. MCCULLOCH
University of Chicago, USA
robert.mcculloch@gsb.uchicago.edu

SUMMARY

For the standard regression setup, conventional tree models partition the predictor space into regions where the variable of interest Y, can be approximated by a constant. A treed model extends this idea by allowing a functional relationship between Y and the predictors within each region. As opposed to using a single model to describe the global variation of the response, treed models allow for local modeling across the predictor space. In this paper, we consider treed versions of generalized linear models (GLMs) and propose a Bayesian approach to finding and fitting such models. The potential of this approach is illustrated with a treed Poisson regression.

Keywords: BINARY TREES, LAPLACE APPROXIMATION, LOGISTIC REGRESSION, MARKOV CHAIN MONTE CARLO, MODEL SELECTION, POISSON REGRESSION, STOCHASTIC SEARCH.

1. INTRODUCTION

Consider the standard regression setup where Y a variable of interest, and $X_1, \ldots, X_p$ a set of potential explanatory variables or predictors, are vectors of n observations. Suppose it is of interest to estimate the conditional distribution of $Y \mid X$ where $X = (X_1, \ldots, X_p)$. A standard modeling approach is to express this conditional distribution as a member of a single parametric family of models. If such an approach is inadequate, a better alternative may be to partition the predictor space so that within each subset of the partition, the parametric model accurately describes the conditional distribution of the response. Such an alternative can be accomplished by using a treed model.

A treed model is composed of two parts - a recursive partitioning of the predictor space using a tree structure and a distinct model for $Y \mid X$ associated with each subset of the partition. The binary tree T used as a recursive partition is inspired by earlier work (e.g. Morgan and Sonquist 1963, Hawkins and Kass 1982, Breiman, Friedman, Olshen, and Stone 1984, Quinlan 1986 and Clark and Pregibon 1990). An important distinction is that this earlier work assumes that within each subset of the partition, $E(Y \mid X)$ is constant. We shall refer to models making this assumption as "conventional trees". For the second component of the treed model, the subset models considered are all generalized linear models (Nelder and Wedderburn 1972, McCullagh and Nelder 1989, and Dey, Ghosh and Mallick 2000). To fit the two components of the treed GLM, we consider a Bayesian approach. This entails the formulation of prior on the space of trees and on the parameters of the subset models. Efficient Metropolis-Hastings algorithms, obtained using Laplace approximations, are then used to stochastically search for high probability models.

The results in this paper extend the results of Chipman, George and McCulloch (1998, 2002) (hereafter CGM 1998 and CGM 2002) and Denison, Mallick and Smith (1998), (hereafter DMS 1998). Both CGM 1998 and DMS 1998 developed similar Bayesian approaches to conventional trees, the special case of treed models where $E(Y\,|\,X)$ is constant within each subset of the partition. CGM 2002 extended this Bayesian approach to treed regression models where $Y\,|\,X$ follows a normal linear model within the partition subsets.

The idea of treed models is not new. The simplest way to fit such models is to construct a conventional tree, and afterwards replace the piecewise constant model for $E(Y\,|\,X)$ with a richer parametric model. This approach is taken by Quinlan (1992) and Torgo (1997). A shortcoming of such approaches is that the partition is not optimized for the final parametric model. Partitioning the predictor space so that the response is as homogeneous as possible within each node makes linear models less likely to be useful. A better approach would be to grow the tree using a splitting criterion that reflects the model used in each partition. Several papers have suggested such a strategy: Karalič (1992) for multiple regression models, Alexander and Grimshaw (1996) for simple linear regression models, and Chaudhuri, Huang, Loh and Yao (1994) for multiple polynomial regression models. This was extended to incorporate GLMs by Chaudhuri, Lo, Loh, and Yang (1995), and overdispersed treed logistic regressions by Ahn and Chen (1997). Treed models also bear some resemblance to hierarchical mixtures of experts models (Jordan and Jacobs 1994), in which a soft decision rule based on a linear combination of predictors is used in each interior node, and a logistic regression is performed in each terminal node. Further discussion of precedents appears in CGM 2002.

What distinguishes our work from these earlier papers is the Bayesian approach. Rather than use ad hoc penalty criteria for ranking models, the posterior distribution coherently ranks the models by using a likelihood to extract the information provided by the data. Rather than use an ad hoc greedy algorithm to find a model, which can be especially challenging with a treed model, our MCMC algorithm uses the posterior information to guide the stochastic search. Simulation studies in CGM 1998 illustrated that for Bayesian CART, our stochastic search found better trees than a greedy search. CGM 1999 further showed that this MCMC algorithm can find a wider variety of trees than a bootstrapped greedy grow/prune algorithm. Lutsko and Kuijpers (1994) considered a MCMC-like approach using simulated annealing, and also found improvements.

Although the formulation of our methods is entirely model based, in many ways it resembles a machine learning algorithm in the spirit of what Breiman (2001) calls "algorithmic modeling". Our formulation can be construed as an algorithm for discovering structure that is controlled by hyper parameters that can be treated as tuning constants. An advantage of the Bayesian formulation is that it provides natural intrepretability of the hyperparameters thereby facilitating their calibration. It is interesting that Breiman distinguishes between two separate cultures of statistics, a culture that treats data as realizations from a model and a culture that gives primacy to out-of-sample prediction for algorithmic construction. We have used a throughly model-based approach to construct algorithms that provide excellent out-of-sample predictions.

2. TREED GENERALIZED LINEAR MODELS

2.1 *The General Model*

For the purpose of modeling the relationship between a variable of interest Y and a $p \times 1$ vector of predictor variables X, a treed model is a specification of the conditional distribution of $Y \mid X$. Such a model consists of two components - a binary tree T that partitions the domain of X, denoted $\mathcal{X}$, and a parametric model for Y associated with each subset of the partition.

The tree T partitions $\mathcal{X}$ as follows. Each interior node of the tree is associated with a single predictor splitting rule that assigns each (X, Y) observation to one of its two child nodes. For ordered predictors, the assignment is determined by whether or not the predictor is less than a fixed value. For categorical predictors, the assignment is determined by whether or not the predictor belongs to a particular subset of the possible categories. By successive assignments, beginning with the root node, T assigns each (X, Y) observation to one of the b terminal nodes, thereby partitioning the predictor space $\mathcal{X}$ into b disjoint sets.

The treed model then associates a parametric model for the distribution of $Y \mid X$ with each of the terminal nodes of T. More precisely, for X values that are assigned to the ith terminal node of T, the conditional distribution of Y is given by a parametric model $Y \mid X \sim p(y \mid X, \theta_i)$ indexed by θ_i. Letting $\Theta = (\theta_1, \ldots, \theta_b)$, a treed model is then fully specified by the pair (Θ, T). By treating the observed data as realizations from a treed model, we can then compute the posterior distribution over (Θ, T). This parametric modeling sets the stage for a Bayesian analysis. In contrast, early tree formulations were essentially proposed as data analysis tools rather than models.

CGM 1998 and DMS 1998 proposed Bayesian approaches to finding and fitting conventional trees. In contrast to treed models, conventional trees use terminal node distributions for $Y \mid X$ that are not functions of X. For example, with a continuous response, a conventional tree would assume $Y \mid X \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for X values assigned to the ith terminal node of T. Such models correspond to step functions for the expected value of $Y \mid X$, and may require large trees to approximate an underlying distribution $Y \mid X$ whose mean is continuously changing in X. By using a richer structure at the terminal nodes, treed models can transfer structure from the tree to the terminal nodes. When such structure exists, smaller and hence more interpretable trees may be used to describe the distributions for $Y \mid X$.

Finally, we should point out that although we use the one symbol "X" for notational simplicity, one can decide to restrict attention to one subset of the components of X for the splitting rules in T and to a different subset for the terminal node models $p(y \mid X, \theta)$. These subsets need not be disjoint.

2.2. *Terminal Node GLMs*

Each tree T induces a partition $T_1, \ldots, T_b$ of the predictor space $\mathcal{X}$, where T_i is the subset of $\mathcal{X}$ corresponding to the ith terminal node of T. Note that $\bigcup T_i = \mathcal{X}$ and $T_i \bigcap T_{i'} = \emptyset$ for $i \neq i'$. For a given T, specification of the terminal node models for Y is facilitated by using a double indexing scheme where (x_{ij}, y_{ij}) denotes each of the $j = 1, \ldots, n_i$ observations of (X, Y) assigned to T_i. All the data assigned to T_i is denoted $x_i = (x_{i1}, \ldots, x_{in_i})$ and $y_i = (y_{i1}, \ldots, y_{in_i})$, and the entire data set is denoted $x = (x_1, \ldots, x_b)$ and $y = (y_1, \ldots, y_b)$. The overall sample size is denoted $n = \sum n_i$. This indexing scheme is conditional on T, and will be different for trees that induce a different partition of $\mathcal{X}$. Finally, in order to accommodate an intercept term in all of the terminal node models below, we shall assume throughout that the first component variable in every x_{ij} is identically equal to 1.

Perhaps the most natural and tractable case of a treed model is the treed regression model, CGM 2002, which is obtained by associating an independent normal linear model with each terminal node subset T_i. Using the notation above, this can be expressed as

$$Y_{ij} \mid x, \Theta, T \sim \mathrm{N}(x_{ij}^T \beta_i, \sigma_i^2) \tag{1}$$

with all Y_{ij} conditionally independent given (x, Θ, T). Here, β_i is an unknown $p \times 1$ vector of regression coefficients and $\theta_i = (\beta_i, \sigma_i^2)$. Under this model both the mean $E(Y_{ij} \mid x, \Theta, T)$ and the variance $Var(Y_{ij} \mid x, \Theta, T)$ functions can change across the terminal node subsets T_i.

In this paper, we consider generalized linear models (GLMs) as the terminal node models. For a given T, such models are of the form

$$p(y_{ij} \mid x, \Theta, T) = \exp\left\{\phi_i^{-1}[y_{ij}\eta_{ij} - \psi(\eta_{ij})] + c(y_{ij}, \phi_i)\right\}. \tag{2}$$

where for some strictly increasing function $h(\cdot)$,

$$\eta_{ij} = h(x_{ij}^T \beta_i), \tag{3}$$

$j = 1, \ldots, n_i$ and $i = 1, \ldots, b$. We will also assume throughout that all the Y_{ij} are conditionally independent given (x, Θ, T). Here, $\theta_i = \beta_i$ or $\theta_i = (\beta_i, \phi_i)$ according to whether the dispersion parameter ϕ_i is treated as known.

Letting $\mu_{ij} \equiv E(Y_{ij} \mid x, \Theta, T)$ denote the conditional mean of Y_{ij}, μ_{ij} is related to $x_{ij}^T \beta_i$ by $g(\mu_{ij}) = x_{ij}^T \beta_i$, where $g(\cdot)$ is called the link function. Because $\mu_{ij} = \psi'(\eta_{ij})$, the link function is implicitly determined by the relationship between η_{ij} and $x_{ij}^T \beta_i$. Indeed, $h(\cdot) = \psi'^{-1}(g^{-1}(\cdot))$ in (3). When $g^{-1} = \psi'$ so that $\eta_{ij} = x_{ij}^T \beta_i$, g is called the canonical link function. Note that the variance of Y_{ij}, $\sigma_{ij}^2 \equiv Var(Y_{ij} \mid x, \Theta, T) = \phi_i \psi''(\eta_{ij})$ depends on η_{ij} and ϕ_i.

The normal linear model (1) is the special case of (2) where g is the identity transform so that $\mu_{ij} = x_{ij}^T \beta_i$, $\sigma_{ij}^2 = \phi_i$ and $\psi(\eta_{ij}) = \eta_{ij}^2/2$. Other exponential family distributions for Y are easily subsumed by (2). When Y_{ij} are 0-1 random variables with means $\mu_{ij} = \pi_{ij}$, the logistic regression model is obtained with the logit transformation $g(\pi_{ij}) = \log[\pi_{ij}/(1 - \pi_{ij})] = x_{ij}^T \beta_i$, $\phi_i \equiv 1$ and $\psi(\eta_{ij}) = \log(1 + \exp(\eta_{ij}))$. In this case, the Y_{ij} are conditionally independent Bernoulli random variables with means $\pi_{ij} = g^{-1}(x_{ij}^T \beta_i) = \exp(x_{ij}^T \beta_i)/(1 + \exp(x_{ij}^T \beta_i))$. When Y_{ij} are counts with mean $\mu_{ij} = \lambda_{ij}$, the Poisson regression model is obtained with the log transformation $g(\lambda_{ij}) = \log \lambda_{ij} = x_{ij}^T \beta_i$, $\phi_i \equiv 1$ and $\psi(\eta_{ij}) = \exp(\eta_{ij})$. In this case, the Y_{ij} are conditionally independent Poisson random variables with means $\lambda_{ij} = g^{-1}(x_{ij}^T \beta_i) = \exp(x_{ij}^T \beta_i)$. In each of these cases, g is a canonical link and $\eta_{ij} = x_{ij}^T \beta_i$.

The assumption that dispersion is fixed at $\phi_i \equiv 1$ in some GLMs can render such models inadequate for data that exhibits greater dispersion. One solution might be to simply consider an overdispersed version of such models where $\phi_i > 1$. Another popular possibility is to go outside the GLM family by using mixture model elaborations such as the beta-binomial or the gamma-Poisson. Other useful alternatives have been proposed by Efron (1986), Jorgensen (1987) and Gelfand and Dalal (1990).

3. PRIOR SPECIFICATIONS FOR TREED GLMS

Since a treed model is identified by (Θ, T), a Bayesian analysis of the problem proceeds by specifying a prior probability distribution $p(\Theta, T)$. This is most easily accomplished by specifying a prior $p(T)$ on the tree space, a conditional prior $p(\Theta \,|\, T)$ on the parameter space, and then combining them via $p(\Theta, T) = p(\Theta \,|\, T)p(T)$.

3.1. *Specification of* $p(T)$

For $p(T)$, we recommend the specification proposed in CGM 1998 for conventional trees. This prior is implicitly defined by a tree-generating stochastic process that grows trees from a single-node tree by randomly splitting terminal nodes. A tree's propensity to grow under this process is controlled by a two-parameter node splitting probability $P(\text{node splits} \,|\, \text{depth} = d) = \alpha(1 + d)^{-\gamma}$, where the root node has depth 0. The parameter α is a base probability of growing a tree by splitting a current terminal node and γ determines the rate at which the propensity to split diminishes as the tree gets larger. The specification of (α, γ) can be guided by the marginal prior distribution on the number of terminal nodes (the tree size), which can be easily simulated. For example, using such marginals, (α, γ) can be chosen to express the belief that reasonably small trees should yield adequate fits to the data. The tree prior $p(T)$ is completed by specifying a prior on the splitting rules assigned to intermediate nodes. We use a prior that is uniform on available variables at a particular node, and within a given variable, uniform on all possible splits for that variable.

3.2. *Specification of* $p(\Theta \,|\, T)$

Turning to the specification of $p(\Theta \,|\, T)$, we note that while $p(T)$ above is sufficiently general for all treed model problems, the specification of $p(\Theta \,|\, T)$ will necessarily be tailored to the particular form of the model $p(y \,|\, x, \theta)$ under consideration. However, some aspects of the $p(\Theta \,|\, T)$ specification should be generally considered. When reasonable, an assumption of iid components of Θ reduces the choice to that of a single prior $p(\theta)$ for $\theta_1, \ldots, \theta_b$. However, even with this simplification, the specification of $p(\theta)$ can be difficult and crucial. In particular, a key consideration is to avoid conflict between $p(\Theta \,|\, T)$ and the likelihood information from the data. On the one hand, if we make $p(\theta)$ too tight (i.e. with very small spread around the prior mean) the prior may be too informative and overwhelm the information in the data corresponding to a terminal node. On the other hand, if $p(\theta)$ is too diffuse (spread out), $p(\Theta \,|\, T)$ will be even more so, particularly for large values of b (large trees) given our iid model for the θ_i. Excessively diffuse prior priors can "wash out" the likelihood in the sense of the Bartlett-Lindley paradox, Bartlett (1957), pushing the posterior of T towards concentration on very small trees. Such a relationship between tree size and dispersion of the θ prior is illustrated in Table 1 of Section 5.1.

For simplicity, we shall assume $\theta_i = \beta_i$ throughout, treating the dispersion parameter ϕ_i as known and to be used as a tuning parameter in applications. More elaborate techniques, such as those mentioned at the end of Section 2, would be required to deal with unknown ϕ_i. Instead we take $\phi_i = \phi$ and investigate the effect of various fixed ϕ values. This strategy is illustrated in Section 5.1 where ϕ_i is seen to be very influential in the modeling process, playing a similar role to the residual variance in least squares regression.

For the prior on the β_is, we consider the simple choice that they are iid multivariate normal conditionally on T,

$$\beta_1, \ldots, \beta_b \,|\, T \;\; iid \;\; \sim N_p(\bar{\beta}, A^{-1}). \tag{4}$$

This further reduces the specification problem to the choice of values for the hyperparameters $\bar{\beta}$ and the $p \times p$ inverse covariance matrix A. We recommended such a normal prior for treed normal regression models in CGM 2002 because of the relative transparency for hyperparameter selection and because it was conjugate, yielding an exact closed form expression for $p(y \,|\, T)$. Although (4) will not be conjugate for other GLMs, it does allow for relatively straightforward use of an effective Laplace approximation of $p(y \,|\, T)$ as will be seen in Section 4.1.

We use information about the distribution of the transformed mean, $g(\mu_{ij})$, to guide the choice of hyperparameter values for $\bar{\beta}$ and A. To simplify notation, we restrict attention here to canonical link models where $g(\mu_{ij}) = x_{ij}^T \beta_i = \eta_{ij}$. Suppose, for the moment, that plausible values were available for $\eta_{\min}, \overline{\eta}, \eta_{\max}$, the minimum, central, and maximum values for the η_{ij}. Automatic choices for such values are discussed at the end of the section.

To further simplify hyperparameter values selection, we also standardize the last $(p-1)$ components of x_{ij} to each have mean 0 and range 1. (Recall that the first component of x_{ij} is always 1 so that an intercept term is included in every terminal node model).

We are now ready to consider the choice of $\bar{\beta}$. To get started, it may be useful to consider this choice under the assumption that a tree is not needed, and a single GLM model is appropriate for the complete data. In this case, a natural default choice is $\bar{\beta} = (\overline{\eta}, 0, \ldots, 0)^T$. This choice guards against an unreasonable value for the intercept while incorporating the neutral value 0 for the remaining components of β, indicating indifference between positive and negative values. Note that standardization of the predictors to have mean 0 decouples the global relationship between the intercept and the other coefficients. However, given T, the mean values of predictors will generally not be 0 within subsets of the partition.

Turning to the choice of A, we make the simplifying reduction that $A = \text{diag}(1/\sigma_0^2, 1/\sigma_\beta^2, \ldots, 1/\sigma_\beta^2)$ where σ_0 is the prior standard deviation of the intercept, and σ_β is the prior standard deviation of the other regression coefficients. This reduces the specification to the choice of two scalars, σ_0 and σ_β. As noted above, the choice of these hyperparameters is crucial. Essentially, the challenge is to choose σ_0 and σ_β large enough to accommodate all plausible values of β, but no larger than that. For choosing σ_0, suppose that all slope coefficients except the intercept were zero. Then we would want σ_0 such that $6\sigma_0 \approx \eta_{\max} - \eta_{\min} \equiv \Delta$, corresponding to the belief that a substantial mass of the normal prior lies within $\overline{\eta} \pm 3\sigma_0$. With this in mind, we treat σ_0 as a tuning parameter and consider various values around $\Delta/6$ as default choices to be explored.

For choosing σ_β, note that by standardizing the range of the predictors to have range 1, a full range increase in a predictor with a regression coefficient equal to Δ would lead to full range increase in $g(\mu_{ij}) = x_{ij}^T \beta_i$ when all the other predictors remained unchanged. However, such reasoning is not completely satisfying in at least two ways. First, for a given tree T, the predictor observations will generally no longer be standardized within each subset of the partition (although they would have a range less than 1, and a mean between -1 and 1). Second, the presence of multicollinearity can necessitate substantially larger coefficient values. However, if severe multicollinearity is present we can usually shrink coefficients towards zero without appreciable damage to the fit. In fact, such shrinkage often stabilizes calculations and even improves predictions. Given these considerations, we also treat σ_β as a tuning parameter and consider various values around $\Delta/6$ as default choices to be explored. Generally, smaller σ_β values will result in estimated coefficients that are shrunk, and trees with fewer terminal nodes.

We conclude this section with a brief discussion of automatic choices of $\eta_{\min}, \overline{\eta}$ and $\eta_{\max}$. In the absence of prior information, a natural automatic choice for $\overline{\eta}$ is $\overline{\eta} = g(\overline{y})$ where $\overline{y}$ is the overall mean of the y_{ij} values. Choice of $\eta_{\min}$ and $\eta_{\max}$ is more challenging. We have found it reasonable to fit a single GLM to the data and to use the minimum and maximum of the MLEs $\hat{\eta}_{ij}$ to estimate $\eta_{\min}$ and $\eta_{\max}$ respectively. A potential drawback is that if the single GLM severely underfits the data, then the range of $\hat{\eta}_{ij}$ may be too small, yielding too tight a prior on β. Despite this drawback, using predictions from a GLM model is certainly superior to using the range of the observed y_{ij}, which can lead to unrealistic bounds. For example with Poisson regression, an observed count $y_{ij} = 0$ would yield $\eta_{\min} = \log(0) = -\infty$. Finally, we note that genuine prior information may well exist in many applications. For example, in modeling of insurance claim counts, actuaries have a good idea of the lowest and highest possible accident rates among all rating groups. Such information might be used exclusively, or combined with automatic choices.

4. POSTERIOR COMPUTATION AND EXPLORATION

Given a set of training data, a Bayesian analysis would ideally proceed by computing the entire posterior distribution $p(\Theta, T \,|\, y, x)$. Unfortunately, in problems such as this, the size of the model space is so large that exhaustive calculation of the posterior is simply not feasible. However, posterior information can still be obtained by using a combination of analytical simplification or approximation together with MCMC sampling from the posterior. For example, the general strategy used in CGM 1998, 2002 was to first eliminate Θ by obtaining a closed form expression for the marginal likelihood

$$p(y \,|\, x, T) = \int p(y \,|\, x, \Theta, T) p(\Theta \,|\, T) d\Theta, \tag{5}$$

and then to use a Metropolis-Hasting (MH) algorithm to simulate a Markov chain sample from $p(T \,|\, y, x) \propto p(y \,|\, x, T)p(T)$. Because the Markov chain simulation tends to gravitate towards higher posterior probability trees, it can effectively be used as a stochastic search algorithm.

CGM 1998, 2002 were able to analytically perform the integration in (5) because conjugate priors were used. However, for GLMs (2) other than the normal linear model, analytical integration is unavailable with the normal prior (4). Instead, we use a Laplace approximation described below to obtain $\tilde{p}(y \,|\, x, T) \approx p(y \,|\, x, T)$, and then apply an MH algorithm to simulate a Markov chain sample from

$$\tilde{p}(T \,|\, y, x) \propto \tilde{p}(y \,|\, x, T)p(T).$$

A similar strategy in the context of model averaging of survival models was successfully used by Raftery, Madigan and Volinsky (1996).

4.1. *Laplace Approximation of* $p(y \,|\, x, T)$

For our treed GLMs, we express the integral in (5) as

$$p(y \,|\, x, T) = \prod_{i=1}^{b} \int L(\beta_i \,|\, x_i, y_i, T) p(\beta_i \,|\, T) d\beta_i \tag{6}$$

where $L(\beta_i \,|\, x_i, y_i, T) = \prod_{j=1}^{n_i} p(y_{ij} \,|\, x_{ij}, \beta_i, T)$ is the likelihood of β_i from (2) and (3). (Recall that we treat ϕ_i as known). To approximate $p(y \,|\, x, T)$, it thus suffices to approximate each of the integrals in (6), and for this purpose we use a Laplace approximation, see Tierney and Kadane (1986).

Let $\mathcal{L}(\beta_i) \equiv \log[L(\beta_i \,|\, x_i, y_i, T) p(\beta_i \,|\, T)]$ denote the log posterior of β_i (up to a norming constant). For notational convenience, we suppress the dependence on x_i, y_i, T in $\mathcal{L}(\beta_i)$. Using a quadratic approximation of $\mathcal{L}(\beta_i)$ around the posterior mode β_i^*, each of the integrals in (6) can be approximated as

$$\begin{aligned} \int L(\beta_i \,|\, x_i, y_i, T) p(\beta_i \,|\, T) d\beta_i &= \int \exp\{\mathcal{L}(\beta_i)\} d\beta_i \\ &\approx \int \exp\left\{\mathcal{L}(\beta_i^*) - \frac{1}{2}(\beta_i - \beta_i^*)^T(-\mathcal{L}''(\beta_i^*))(\beta_i - \beta_i^*)\right\} d\beta_i \\ &= \exp\{\mathcal{L}(\beta_i^*)\}\ (2\pi)^{p/2}\ |-\mathcal{L}''(\beta_i^*)|^{-1/2} \end{aligned} \tag{7}$$

where $\mathcal{L}''(\beta_i^*)$ is the $p \times p$ matrix of second derivatives of $\mathcal{L}$ evaluated at β_i^*.

Now, under our normal prior $\beta_i \sim \mathrm{N}(\bar{\beta}, A^{-1})$ in (4), the log posterior $\mathcal{L}$ can be conveniently expressed, (up to a norming constant), as

$$\mathcal{L}(\beta_i) \equiv l(\beta_i) - \frac{1}{2}(\beta_i - \bar{\beta})^T A (\beta_i - \bar{\beta}) \tag{8}$$

where $l(\beta_i) \equiv \log L(\beta_i \,|\, x_i, y_i, T)$ denotes the log likelihood of β_i. It also follows that

$$\mathcal{L}'(\beta_i) = l'(\beta_i) - A(\beta_i - \bar{\beta}) \tag{9}$$

and

$$\mathcal{L}''(\beta_i) = l''(\beta_i) - A. \tag{10}$$

Using (8) and (10), the Laplace approximation (7) with the normal prior on β_i can be expressed, (up to a norming constant), as

$$\begin{aligned} \int L(\beta_i \,|\, x_i, y_i, T) p(\beta_i \,|\, T) d\beta_i &\approx \exp\{\mathcal{L}(\beta_i^*)\}\ (2\pi)^{p/2}\ |-l''(\beta_i^*) + A|^{-1/2} \\ &= \frac{|A|^{1/2}}{|-l''(\beta_i^*) + A|^{1/2}}\ \exp\left\{l(\beta_i^*) - \frac{1}{2}(\beta_i^* - \bar{\beta})^T A (\beta_i^* - \bar{\beta})\right\}. \end{aligned} \tag{11}$$

This approximation depends on the data through β_i^*, $l(\beta_i^*)$ and $l''(\beta_i^*)$.

From (2) and (3), the general form for the log likelihood for each terminal node GLM is

$$l(\beta_i) = \phi_i^{-1} \sum_{j=1}^{n_i} \left[y_{ij} h(x_{ij}^T \beta_i) - \psi(h(x_{ij}^T \beta_i))\right].$$

The first and second derivatives of $l(\beta_i)$ are

$$l'(\beta_i) = \phi_i^{-1} \sum_{j=1}^{n_i} \Big[[y_{ij} - \psi'(h(x_{ij}^T \beta_i))] h'(x_{ij}^T \beta_i) \Big] x_{ij}$$

and

$$l''(\beta_i) = \phi_i^{-1} \sum_{j=1}^{n_i} \Big[[y_{ij} - \psi'(h(x_{ij}^T \beta_i))] h''(x_{ij}^T \beta_i) - \psi''(h(x_{ij}^T \beta_i))(h'(x_{ij}^T \beta_i))^2 \Big] x_{ij} x_{ij}^T$$

Special cases of l, l' and l'' are easily obtained for particular models based on $\psi(\cdot), h(\cdot)$ and ϕ_i. For example, in the logistic regression setting where $\psi(\eta) = \log(1 + \exp(\eta))$, $h(x) = x$ and $\phi_i \equiv 1$, we obtain $l(\beta_i) = \sum_{j=1}^{n_i} [y_{ij}\eta_{ij} - \log(1 + \exp(\eta_{ij}))]$, $l'(\beta_i) = \sum_{j=1}^{n_i} [y_{ij} - \pi_{ij}]\, \eta_{ij} x_{ij}$, and $l''(\beta_i) = \sum_{j=1}^{n_i} -x_{ij} x_{ij}^T \pi_{ij}(1 - \pi_{ij})$, since $\psi'(\eta) = e^\eta/(1+e^\eta) = \pi$, $\psi''(\eta) = e^\eta/(1+e^\eta)^2$, $h'(x) \equiv 1$ and $h''(x) \equiv 0$. In the Poisson regression setting where $\psi(\eta) = e^\eta$, $h(x) = x$ and $\phi_i \equiv 1$, we obtain $l(\beta_i) = \sum_{j=1}^{n_i} [y_{ij}\eta_{ij} - \lambda_{ij}]$, $l'(\beta_i) = \sum_{j=1}^{n_i} [y_{ij} - \lambda_{ij}]\eta_{ij} x_{ij}$, and $l''(\beta_i) = \sum_{j=1}^{n_i} -\lambda_{ij}\ x_{ij} x_{ij}^T$, since $\psi'(\eta) = \psi''(\eta) = e^\eta = \lambda$, $h'(x) \equiv 1$ and $h''(x) \equiv 0$.

Finally, to compute (11) for each i using these expressions, the posterior mode β_i^* is needed. To find β_i^*, we use a simple Newton-Raphson algorithm

$$\beta^{(k+1)} = \beta^{(k)} - (\mathcal{L}''(\beta^{(k)}))^{-1}(\mathcal{L}'(\beta^{(k)}))$$

where $\mathcal{L}'$ and $\mathcal{L}''$ are obtained from (9) and (10).

One implementation detail is worth noting. Throughout this section, norming constants have been omitted for the sake of clarity. These constants can actually affect the posterior probability, since each terminal node GLM will have its own norming constants, and the number of terminal nodes varies across trees. All norming constants are retained in our program code except where they are known to cancel.

4.2. *Markov Chain Monte Carlo Posterior Exploration*

We use MCMC to stochastically search for high posterior trees T by using the following Metropolis-Hastings algorithm which simulates a Markov chain $T^0, T^1, T^2, \ldots$ with limiting distribution $\tilde{p}(T \,|\, y, x) \propto \tilde{p}(y \,|\, x, T)p(T)$, where $\tilde{p}(y \,|\, x, T)$ is the Laplace approximation to $p(y \,|\, x, T)$ proposed above. Starting with an initial tree T^0, this algorithm iteratively simulates the transitions from T^i to T^{i+1} by the two steps:

1. Generate a candidate value T^* with probability distribution $q(T^i, T^*)$.
2. Set $T^{i+1} = T^*$ with probability

$$\alpha(T^i, T^*) = \min\left\{ \frac{q(T^*, T^i)}{q(T^i, T^*)} \frac{\tilde{p}(y \,|\, x, T^*)p(T^*)}{\tilde{p}(y \,|\, x, T^i)p(T^i)}, 1 \right\}. \tag{12}$$

Otherwise, set $T^{i+1} = T^i$.

In (12), $q(T, T^*)$ is the kernel which generates T^* from T by randomly choosing among four steps: GROW, PRUNE, CHANGE, and SWAP. Details of these steps are given in CGM 1998, 2002. Although these moves better explore the posterior than a greedy grow/prune algorithm, the chain may get trapped in local maxima. Multiple restarts may be employed to efficiently explore the posterior on trees. See CGM 1998, 2002 for additional details.

5. AN APPLICATION

5.1 *A Wave Soldering Experiment*

To illustrate and assess our Bayesian treed GLM approach, we applied it to the Poisson regression dataset `solder2`, which is available in S and is described in Chapter 1 of Chambers and Hastie (1992). These data were originally collected by Comizzoli, Landwehr, and Sinclair (1990) as part of an experiment to investigate a wave soldering procedure for mounting electrical components on circuit boards. The response, `skips`, is a visual count of the number of skips in solder applied to a circuit board. In 623 of the 750 observations, `skips` had a value 0. The mean and maximum value of `skips` were 1.19 and 32, respectively. The five categorical predictors are:

- `Opening (S/M/L)`: amount of clearance around the mounting pad;
- `Solder (Thick/Thin)`: amount of solder;
- `Mask (5 levels)`: type and thickness of the material used for the solder mask;
- `PadType (10 levels)`: the geometry and size of the mounting pad; and
- `Panel (1/2/3)`: each board was divided into three panels, with three runs on a board.

We began by fitting two Poisson regressions with log link functions. The first model contained indicator variables for all main effects, while the second (suggested in Chambers and Hastie 1992) contained main effects plus the three two-way interactions `Opening:Solder, Opening:Mask` and `Mask:Solder`. The mean deviance for the first model was 1.52 with 731 df and the mean deviance for the second was 1.24 with 719 df. Mean deviances greater than 1 suggest either lack of fit or overdispersion.

As an alternative to these Poisson GLMs, we proceeded to consider a treed Poisson GLM, and elected to consider all predictors in both the splitting rules of interior nodes and in the terminal node GLMs. As described in Section 3.2, we set the prior mean of the intercept to be $\hat{\beta}_0 = g(\bar{y}) = \log(\bar{y})$ and set the prior means of the slope components of β equal to 0. To gauge the choices for the hyperparameters σ_0 and σ_β, we fitted the main effects Poisson regression mentioned above and found the range of predicted values of $\hat{\eta}_{ij} = \log \hat{\mu}_{ij}$ to be $\Delta = 17.8$. This suggests $\sigma_0 = \sigma_\beta = \Delta/6 \approx 3$. We considered two settings for the prior standard deviation in a neighborhood of 3, namely slopes $\sigma_\beta = 2, 4$. The prior standard deviation of the intercept σ_0 was fixed at 4. A value of 4 was used instead of 3 because it was felt that additional dispersion in the intercept was less likely to have an impact on the treed model. Because of the overdispersion we observed in the Poisson GLMs, it seemed reasonable to consider that the dispersion parameter ϕ would exceed 1 (no overdispersion) and be no more than 3. With this in mind, we treated ϕ as a tuning parameter and considered three settings $\phi = 1.5, 2, 3$. As will be seen below, varying this dispersion parameter ϕ plays an important role in the analysis. The modeling procedure was run separately for the six combinations of σ_β and ϕ.

The MH search was run with tree prior parameters $\alpha = 0.25$ and $\gamma = 2$, giving prior mass of approximately (0.75,0.22,0.03) on trees with 1, 2, and ≥ 3 terminal nodes, respectively. We set the Markov transition kernel $q(T^i, T^*)$ to randomly choose one of the four steps with probabilities P(GROW) = P(PRUNE)= 0.1 and P(CHANGE) = P(SWAP)= 0.4. For each estimation, one chain with 2500 steps was used, taking approximately 11 minutes to execute on a PentiumIII/1GHz computer. Although many trees are visited by the MH algorithm, the "best" tree is determined as follows: The most frequently visited tree size is identified, and then for this size we choose the tree with the highest log integrated likelihood. While the alternative of ranking trees by log posterior probabilities may seem appealing, it suffers from the dilution phenomenon discussed in CGM 1998.

We also considered a conventional tree with constant means at terminal nodes, fit with a Poisson likelihood via greedy search and cross-validation. Compared to our treed Poisson regressions, this conventional model relies much more on the tree structure to explain the variation of the response `skips`. This tree was fit to the data with the `rpart` implementation (Therneau and Atkinson 1997) in Splus and R. Sensitivity to overdispersion is not a substantial issue when cross-validation is used to determine tree complexity. For this reason, we did not consider different parameter settings of the `rpart` procedure.

To compare the different modeling methods, as well as the various tuning parameter choices, we used a repeated 10-fold cross-validation. Within each 10-fold cross-validation, each data point is predicted out-of-sample once (i.e., in 9/10 of the cases it will be used for training and 1/10 for testing). Five different replications of 10-fold cross-validation were carried out. All methods were compared with the same five replications, thereby removing blocked effects. For the `rpart` tree, two cross-validations are in fact being performed: cross-validation internal to the `rpart` code is determining an appropriate tree size, and our cross-validation algorithm is assessing out-of-sample performance.

Figure 1 compares the performance of the different methods using mean and median deviance contributions, where the deviance contribution from observed response y_i with out-of-sample prediction $\hat{\mu}_i$ given by $d_i = 2\left[(\hat{\mu}_i - y_i) + y_i(\log y_i - \log \hat{\mu}_i)\right]$. The log posterior could have also been used as a performance measure, but due to comparisons with likelihood based methods (`glm, rpart`), we used deviance measures instead.

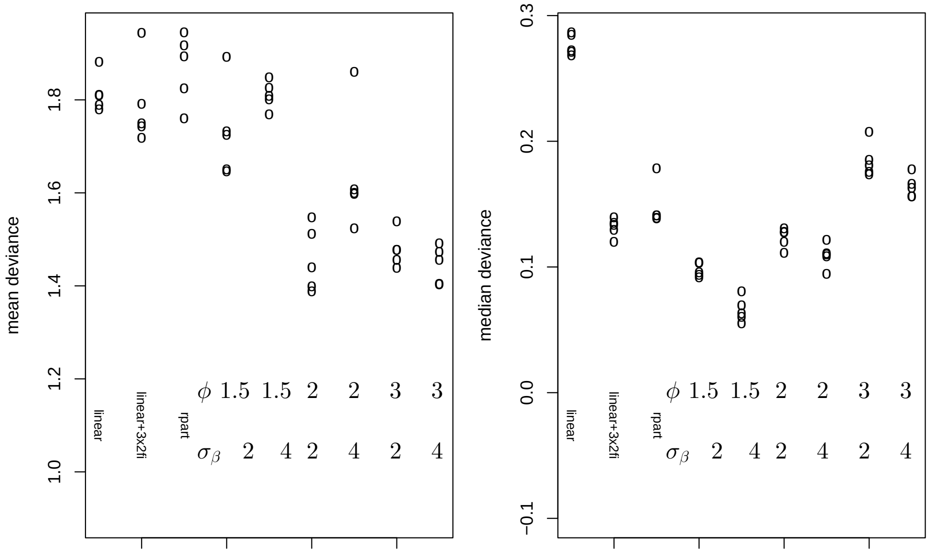

Figure 1. *Comparison of two GLMs (main effects models and main effects with three two-way interactions),* `rpart` *(conventional tree with Poisson data) and six treed models (with various parameter settings).*

We see in Figure 1 that the best performers in terms of mean deviance are treed models with $\phi = 2$ or $\phi = 3$. However, the distribution of the d_i's turned out to be very long-tailed, with a few influential large values. For this reason we also considered the median deviance. For this measure, we see an even more substantial difference, as well as reduced variability in the values across the five simulations.

An understanding of the relationship between tree size and the parameters ϕ and σ_β may be useful in interpreting Figure 1. Table 1 shows the mean size of the "best" tree reported in the 50 runs (5 permutations × 10 folds) for each of the six (ϕ, σ_β) settings considered. As overdispersion increases, the tree size decreases, since the log posterior is divided by the factor ϕ. Increasing σ_β also makes smaller trees more likely. Evidently, a tree of 4 or more nodes is overfitting, and a tree with 2 nodes may be slightly underfitting (as can be seen from the median deviance contribution).

Table 1. *Mean size of tree across 10 folds of cross-validation, and 5 permutations of data. Note that this "mean" is not across the posterior, since only one tree is reported for each of the 50 runs.*

ϕ	σ_β	mean tree size
1.5	2	4.94
1.5	4	4.54
2.0	2	3.18
2.0	4	3.04
3.0	2	2.44
3.0	4	2.00

From the cross-validation results, it appeared that reasonably good and robust performance was obtained with the settings $\phi = 2, \sigma_\beta = 2$. We thus ran our treed model search algorithm using all the data with these settings. From a run of 2500 steps, we selected a "best" tree, which is given in Figure 2. This tree splits first on `Opening`, and subsequently on `PadType` in one node. Since the GLM with interactions involving `Opening` was an improvement over the main effects GLM, it is not surprising that this variable was split upon. Note however, that not all interactions with `Opening` are fit by this tree, since two categories (`Opening = middle, large`) are kept together. The subsequent split on PadType is suggestive of a three-way interaction between `Opening, PadType` and other variables.

In each terminal node a separate Poisson regression model is fit to the data. The degrees of freedom used by each model varies from one node to the next because some predictors are constant within nodes. This is obvious for variables used as splitting rules, such as `Opening` in Node 1, and the experimental design also eliminates some categorical variable levels in some of the terminal nodes.

The coefficients of the GLMs in each node are plotted in Figure 3. Although the actual degrees of freedom used in each node varies, the informative prior makes it possible to calculate posterior means for all 19 regression coefficients, even when some predictors are constant in terminal nodes. This somewhat restricts interpretation of this plot, since within a node, some estimates are aliased with others. We can see however, that the effect of solder thickness is large when the opening is small and near zero otherwise, an interaction noted in the original analysis. We see also that in Node

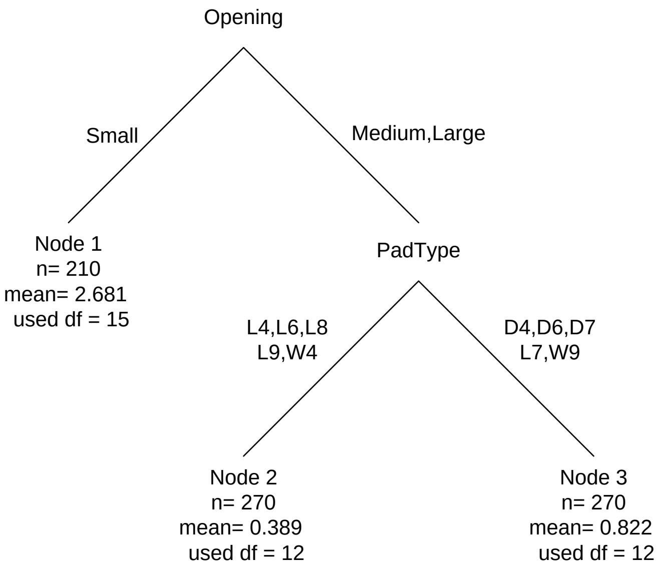

Figure 2. *The best tree found with $\phi = 2, \sigma_\beta = 2$.*

3 (`Openings`=`Med/Large`, `PadType`=`D4,D6,D7,L7,W9`) coefficients for `L7` and `W9` are especially large. Mean numbers of `skips` and sample sizes for a partitioning of the data into four groups are given in Table 2.

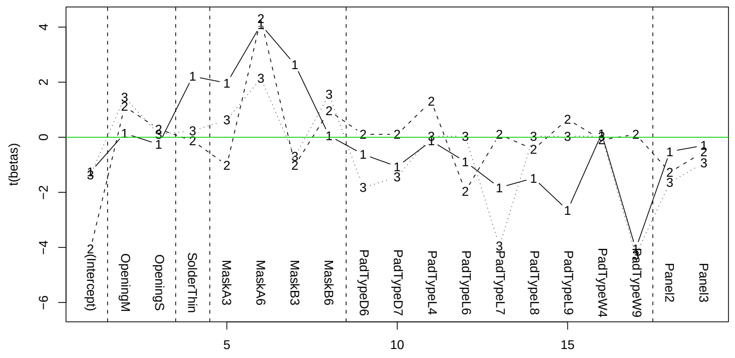

Figure 3. *Posterior mean regression coefficients for the best tree (tree given in Figure 2). Coefficients for each node are joined with a line, and the numeric plotting symbols correspond to the node numbers given in Figure 2.*

The table indicates a very low rate of skips for `PadType` = `L7,W9` and `M` or `L` levels of `Opening`. Evidently the split on `PadType` was chosen because of quite different mean levels, but within nodes 2 and 3, there are also differences by `PadType`. The differing coefficients in nodes 2 and 3 for `Mask` suggest that this may be a secondary reason for the choice of this particular split.

Table 2. *Summaries of the response for subsets of the data indicated by particular values of the predictors*

Group	mean `skips`	n
`Opening` = `S`	2.6810	210
`Opening` = `M,L`, `PadType` = `L7`	0.0185	54
`Opening` = `M,L`, `PadType` = `W9`	0.0000	54
`Opening` = `M,L`, `PadType` ≠ `L7,W9`	0.7546	432
total	1.1867	750

5.2. *Simulation Study of the Null Case*

A potential criticism of any flexible model is that it finds complicated structure when there is none. In the `solder2` example, out-of-sample validation indicates that complex structure actually is present. But how will the Bayesian search for treed models perform when the true model is a single GLM?

To study this null case, we simulated data from a single GLM, using the same predictor values as in the `solder2` dataset. Response values were simulated using regression coefficients similar to the MLEs from a single GLM fit to the original data. An overdispersion component was incorporated into some of the simulated data sets by adding a random effect to η_{ij} before generating the observed response. We generated 60 data sets, 20 with no overdispersion, 20 with moderate overdispersion and 20 with severe overdispersion. For each of these data sets, we ran our procedure with the same settings as in the previous section, except with four settings of $\phi = 1, D/2, D, 2D$ where D was the observed mean deviance of a Poisson GLM fit to the data. We considered these choices of ϕ to explore the effect of calibrating ϕ to the data.

In the interest of brevity, we give but a precis of our findings, which were very favorable. First of all, we were most interested to see how often our approach incorrectly partitioned the data by using a tree with more than one node. For data simulated from a single GLM with no overdispersion, the selected trees had a mean size of just over 1 for $\phi = 1, D, 2D$. With moderate overdispersion, average tree size was smallest at just over 1 when $\phi = 2D$ and around 2 when $\phi = D$. With severe overdispersion, average tree size was smallest, between 2 and 3, when $\phi = 2D$. In terms of fit to the data, our treed models were very competitive with a single GLM fit to the data, usually achieving a similar value for out-of-sample deviance. It is interesting to note that in most cases when the tree size was larger than 1, the treed model fits were not dramatically worse than those of a single GLM. Overfitting only became a problem when excessively small ϕ values were used.

REFERENCES

Ahn, H. and Chen, J.J. (1997). Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. *Biometrics* **53**, 435-455.

Alexander, W. P. and Grimshaw, S. D. (1996). Treed regression. *J. Comp. Graph. Statist.* **5**, 156–175.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533-534.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statist. Sci.* **16** , 199 – 231 (with discussion).

Breiman, L., Friedman, J. Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Pacific Drove, CA: Wadsworth.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93**, 935–960 (with discussion).

Chipman, H. A., George, E. I. and McCulloch, R. E. (1999). Making sense of a forest of trees. *Proceedings of the 30th Symposium on the Interface*, S. Weisberg, Ed., Interface Foundation of North America, 84–92.

Chipman, H., George, E. I., and McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning* **48**, 299–320.

Chambers, J.M. and Hastie, T.J. (eds.) (1992) *Statistical Models in S.* Boca Raton: CRC Press.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica* **4**, 143–167.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica* **5**, 641–666.

Clark, L. A. and Pregibon, D. (1992). Tree-Based Models. *Statistical Models in S.* (J. M. Chambers, and T. J. Hastie, eds.) Boca Raton: CRC Press, 377-420.

Comizzoli, R. B, Landwehr, J. M., and Sinclair, J. D. (1990). Robust materials and processes: key to reliability. *AT&T Technical Journal* **69** , 113–128.

Denison, D., Mallick, B. and Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363-377.

Dey, D.K., Ghosh, S.K. and Mallick, B.K. (2000). *Generalized Linear Models: A Bayesian Perspective.* New York: Marcel Dekker.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81**, 709-721.

Gelfand, A.E. and Dalal S. (1990). A note on overdispersed exponential families. *Biometrika* **77**, 55-64.

Hawkins, D. M. and Kass, G. V. (1982). Automatic interaction detection. *Topics in Applied Multivariate Analysis.* (D. M. Hawkins, ed.).Cambridge: University Press.

Jorgensen, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. B* **49**, 127-162 (with discussion).

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181–214.

Karalič, A. (1992). Employing linear regression in regression tree leaves. *Proceedings of ECAI-92.* Chichester: Wiley, 440–441.

Lutsko, J. F. and Kuijpers, B. (1994). Simulated annealing in the construction of near-optimal decision trees. *Selecting Models from Data: AI and Statistics IV.* (P. Cheeseman and R. W. Oldford, Eds.). Berlin: Springer, 453–462.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* London: Chapman and Hall

Morgan, J. N. , and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58**, 415-434.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc. A* **135**, 370-384.

Quinlan, J. R. (1986). Induction of decision trees, *Machine Learning* **1**, 81–106.

Quinlan, J. R. (1992). Learning with continuous classes, in *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 343–348, World Scientific.

Raftery, A. E., Madigan, D. and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 323-349.

Therneau, T.M. and Atkinson, E.J. (1997). An introduction to recursive partitioning using the RPART routines, *Tech. Rep.*, Department of Health Science Research, Mayo Clinic, Rochester, Minnesota. Available online at `http://www.mayo.edu/hsr/techrpt/61.pdf`

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Torgo, L. (1997). Functional models for regression tree leaves. in *Proceedings of the International Machine Learning Conference (ICML-97).* San Mateo, CA: Morgan Kaufmann, 385-393.