Semiparametric Mixed-Scale Models Using Shared Bayesian Forests

Antonio R. Linero

Department of Statistics, Florida State University
**email:* arlinero@stat.fsu.edu

and

Debajyoti Sinha

Department of Statistics, Florida State University
**email:* sinhad@stat.fsu.edu

and

Stuart R. Lipsitz

Division of General Internal Medicine, Brigham and Women's Hospital
**email:* slipsitz@partners.org

Summary: This paper demonstrates the advantages of sharing information about unknown features of covariates across multiple model components in various nonparametric regression problems including multivariate, heteroscedastic, and semi-continuous responses. In this paper, we present methodology which allows for information to be shared nonparametrically across various model components using Bayesian sum-of-tree models. Our simulation results demonstrate that sharing of information across related model components is often very beneficial, particularly in sparse high-dimensional problems in which variable selection must be conducted. We illustrate our methodology by analyzing medical expenditure data from the Medical Expenditure Panel Survey (MEPS). To facilitate the Bayesian nonparametric regression analysis, we develop two novel models for analyzing the MEPS data using Bayesian additive regression trees - a heteroskedastic log-normal hurdle model with a "shrink-towards-homoskedasticity" prior, and a gamma hurdle model.

Key words and phrases. Bayesian additive regression trees; Heteroskedastic errors; Hurdle models; Nonparametric Bayes; Variable selection.

1. Introduction

In complex statistical problems it is often of interest to share information across multiple model parameters and components. For studies with multiple responses, the same unknown set of features may be associated with the responses. In our motivating example of medical expenditure data from the Medical Expenditure Panel Survey (MEPS), many individuals record no medical expenditures (zero response) over the course of a year. As a consequence, the distribution of an individual's semi-continuous response of total yearly medical expenditures is a mixture of a point-mass at zero and a continuous distribution on the positive reals. Intuitively, the set of factors which predict whether an individual incurs *no* medical expenditure may also be predictive of the *magnitude* that individual's medical expenditure if one occurs.

An increasingly popular method for modeling nonparametric functions is the Bayesian additive regression trees (BART) framework introduced by Chipman et al. (2010). The BART framework has been successfully applied to a diverse set of problems including survival analysis (Sparapani et al., 2016), causal inference (Hahn et al., 2017; Hill, 2011), analysis of loglinear models (Murray, 2017), imputation of missing predictors (Xu et al., 2016), and high dimensional prediction and variable selection (Linero, 2018).

In this paper, we introduce *shared forests*, which nonparametrically model multiple model components using a single set of trees. By viewing BART as a method for learning data-adaptive basis expansions, shared forests restrict the basis functions across model components to be the same while allowing the corresponding coefficients to be different. A simulation study shows that sharing information across model components in this fashion can be very beneficial, particularly in sparse high-dimensional problems in which variable selection a is necessary step.

In addition to our shared forests model, we make several additional contributions which are of practical interest in their own right. Semi-continuous responses are routinely modeled via two-part mixture models, often called hurdle models in econometrics, with a binary component modeling the probability of a zero response, and a continuous distribution modeling the response given it is non-zero. We present two novel semi-parametric hurdle models for analyzing semi-continuous responses. The first is a type of gamma hurdle model, which are popular for modeling rainfall data (Feuerverger, 1979), in which the mean of the gamma distribution and the probability of a zero response are both modeled nonparametrically. The second model is a log-normal hurdle model (Aitchison, 1955; Xiao-Hua and Tu, 1999) in which the log-mean, log-variance, and the probability of a zero response are all modeled nonparametrically. See Tu (2006) for a review of zero-inflated and hurdle models. To the best of our knowledge, we are first to adapt BART to the mean of a gamma distribution; this requires developing an analog of the usual Bayesian backfitting approach for fitting BART models of Chipman et al. (2010). Additionally, while nonparametric models for the variance have been considered in other Bayesian sum-of-trees approaches (Murray, 2017; Pratola et al., 2017), we are required to develop a different nonparametric approach for the log-variance of our log-normal hurdle model in order to allow the tree structures to be shared across the mean, variance, and probability of a zero response while preserving computational tractability. In order to prevent overfitting on the variance component of the model, our variance modeling framework is designed to be centered at, and to allow shrinking heavily towards, a parsimonious model with constant variance. This allows us to model heteroskedasticity in the data while preserving estimation efficiency when the variance of the response is actually constant.

Our approach has natural connections with several proposals in the machine learning literature on *multi-task learning* and *multi-output* learning; see Borchani et al. (2015) for a review. Related methods include multi-objective decision trees, multi-task boosting, and multi-task kernel methods. Additionally, there are several methods which share information across models in the BART framework. The Bayesian causal forest (BCF) model of Hahn et al. (2017) uses two separate forests to model the distribution of potential outcomes: the first forest accounts for the direct effect of confounders on the potential outcomes, and is identical for both the treatment and controls, while the second forest is unique to the distribution of the treated samples. Another related work by Starling et al. (2018) proposes a model in which a temporally indexed response is modeled using BART, with the forest being shared across time. We discuss these connections in Section 2.3.

We apply our methodology to data from the Medical Expenditure Panel Survey (MEPS). The outcome Y is an individual's total health care expenditure during the course of the year 2015. We show that the heteroskedastic log-normal hurdle model fits this data very well, and we use a shared forest to jointly model (i) the probability of $Y = 0$, (ii) the mean of $\log Y$ given Y is nonzero, and (iii) the variance of $\log Y$ given Y is nonzero. By examining the fit of the mean and variance components, we are able to validate the earlier observation of Blough and Ramsey (2000) that the variance of Y is roughly proportional to $E(Y)^{1.5}$ for MEPS data. However, we are also able to identify sources of heterogeneity which are not explained by this relationship between the variance and the mean.

In Section 2, we introduce our shared forests framework. In Section 3, we develop and give default prior specifications for the gamma hurdle and log-normal hurdle models we use later to analyze the MEPS data. In Section 4, we conduct a simulation study which illustrates the potential benefits of sharing information across model components. In Section 5, we apply the methodology developed here to the MEPS dataset. We conclude in Section 6 with a discussion. Additional computational details, and details about the analysis of the MEPS dataset, are given in a web appendix.

2. Shared Forests

2.1 *Review of Bayesian Additive Regression Trees*

We briefly review the Bayesian Additive Regression Trees (BART) framework (Chipman et al., 2010), an extremely useful tool for constructing highly flexible Bayesian semiparametric models. BART models typically outperform comparable linear models and often outperform machine learning techniques such as boosted decision trees and random forests. We assume that the unknown function of interest $h(\boldsymbol{x})$ can be expressed as a sum of T regression trees depending on tree structures $\mathcal{T}_t$ and leaf-node parameters $\mathcal{M}_t$,

$$h(\boldsymbol{x}) = \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t), \tag{1}$$

where $g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t) = \mu_{t\ell}$ if the predictor value $\boldsymbol{x}$ is associated to leaf node ℓ of tree t. As illustrated in Figure 1, the decision tree $\mathcal{T}_t$ encodes a recursive partition of the predictor space $\mathcal{X} = [0,1]^P$, with $g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t)$ being piecewise-constant. Let $\mathcal{L}_t$ denote the leaf nodes of the tree, so that $\mathcal{M}_t = \{\mu_{t\ell} : \ell \in \mathcal{L}_t\}$.

The prior for $h(\cdot)$ in (1) consists of a prior mass function $\pi_{\mathcal{T}}(\cdot)$ for the tree structures $\mathcal{T}_t$ and a prior on the leaf node parameters $\mathcal{M}_t$. Chipman et al. (2010) propose a branching process prior for $\mathcal{T}_t$. A draw from this prior is obtained by generating, for each node at depth d, two child nodes with probability $q(d) = \gamma(1+d)^{-\zeta}$; otherwise, the node becomes a leaf node (which defines a new equivalence class). This process iterates for $d = 0, 1, 2, \ldots$ until we reach a depth d at which all of the nodes are leaves. Note that $q(d)$ is not a mass function over d, but instead is the prior probability of a given leaf node being converted to a branch node. The case $\zeta = 0$ corresponds to the Galton-Watson process (Athreya and Ney, 2004). A sufficient condition for termination of this branching process is $\zeta > 0$. After the tree topology is generated, each branch node b is associated to a decision rule of the form $[x_j \leqslant C_b]$ where the coordinate $j \in \{1, \ldots, P\}$ is selected independently for each branch with probability s_j. Throughout, we will use the sparsity inducing Dirichlet prior $(s_1, \ldots, s_P) \sim \mathcal{D}(\xi/P, \ldots, \xi/P)$ proposed by Linero (2018). This prior concentrates on neighborhoods of sparse probability vectors, a fact which has been leveraged to perform variable selection in linear models (Bhattacharya et al., 2015), and adapt to irrelevant predictors in Gaussian process models (Bhattacharya et al., 2014). Intuitively, if s_j is very small (e.g., $s_j < 10^{-10}$), then predictor x_j is highly unlikely to appear within any splitting rule, effectively eliminating x_j from the model. The Dirichlet prior encourages these extreme values of s_j, allowing the model to perform fully-Bayesian variable selection. In this paper, we set $C_b \sim \text{Uniform}(L_j, U_j)$ for the cutpoints C_b conditional on the tree topology, selected coordinate j, and the parameters of b's "ancestor nodes". Here $(L_1, U_1) \times \cdots \times (L_P, U_P)$ is the hyperrectangle defined by all values of $\boldsymbol{x}$ which lead to branch b.

Let $\boldsymbol{\omega}$ be a vector of non-tree-specific parameters, such as the variance $\sigma^2 = \text{Var}(Y_i \mid \boldsymbol{X}_i)$ for a regression model with constant variance. Our model for the response Y_i in this setting is expressed as $(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}, h, \boldsymbol{\omega}) \sim f\{y \mid h(\boldsymbol{x}), \boldsymbol{\omega}\}$ where $\{f(\cdot \mid \mu, \boldsymbol{\omega})\}$ is a parametric family. Conditional on the trees $\mathcal{T}_1, \ldots, \mathcal{T}_T$, the leaf node parameters $\{\mu_{t\ell} : \ell \in \mathcal{L}_t, 1 \leqslant t \leqslant T\}$ are given iid priors $\mu_{t\ell} \sim \pi_\mu$. Usually π_μ is chosen to ensure that the integrated likelihood

$$\Lambda(\mathcal{T}_t) = \pi_{\mathcal{T}}(\mathcal{T}_t) \int \prod_{i=1}^{n} f\{Y_i \mid h(\boldsymbol{X}_i), \boldsymbol{\omega}\} \prod_{\ell \in \mathcal{L}_t} \pi_\mu(\mu_{t\ell}) \, d\mu_{t\ell} \tag{2}$$

has a closed form expression. For example, in the regression setting, a popular choice for π_μ is the $\mathcal{N}(0, \sigma_\mu^2)$ density. When using Markov Chain Monte Carlo (MCMC) to conduct Bayesian inference, $\mathcal{T}_t$ can be updated using Metropolis-Hastings, with $\Lambda(\mathcal{T}_t)$ used to compute the acceptance probability; see Pratola (2016) for further details. In the original paper of Chipman et al. (2010), both $\Lambda(\mathcal{T}_t)$ and the full conditionals for the $\mu_{t\ell}$'s are calculated using Bayesian backfitting.

A large reason for the success of BART is the existence of highly effective "default" priors which can be expected to provide a reasonable baseline level of performance without requiring tuning by the user. As a default, we set $\gamma = 0.95$, $\zeta = 2$, and $\xi/(\xi + P) \sim \text{Beta}(0.5, 1)$; other prior specifications are model specific. Additionally, BART models have highly desirable theoretical properties (Linero and Yang, 2018; Rockova and van der Pas, 2017); in particular, in regression problems, certain BART models attain near-minimax optimal rates of estimation for functions $h(\boldsymbol{x})$ with low-order interactions. For an in-depth review of Bayesian regression tree methods, see Chipman et al. (2013) and Linero (2017).

A fact which will be useful for specifying priors later, and for making connections with other approaches, is that, under the conditions $E(\mu_{t\ell}) = 0$ and $\mathrm{Var}(\mu_{t\ell}) = \sigma_\mu^2/T$, the prior converges to a Gaussian process as $T \to \infty$. To see this heuristically, note that we can write $h(\boldsymbol{x}) = T^{-1/2} \sum_{t=1}^{T} \mu_{t\ell}^{\star}$ where the $\mu_{t\ell}^{\star}$'s are iid with mean 0 and variance σ_μ^2. In general, for fixed $\boldsymbol{x}, \boldsymbol{x}'$, we have $\mathrm{Cov}\{h(\boldsymbol{x}), h(\boldsymbol{x}')\} = \sigma_\mu^2 \Pr(\boldsymbol{x} \sim \boldsymbol{x}')$ where the event $[\boldsymbol{x} \sim \boldsymbol{x}']$ denotes that $\boldsymbol{x}$ and $\boldsymbol{x}'$ share the same terminal node in $\mathcal{T}_1$. An application of the multivariate central limit theorem then establishes convergence of the finite dimensional marginals to a multivariate normal distribution. A natural alternative method is to simply apply Gaussian processes in practice. BART has both practical and theoretical benefits over Gaussian process models. First, the computational complexity of Gaussian process methods typically is $O(n^3)$ where n is the number of samples; conversely, in practice, BART has computations which scale slightly faster than $O(nT)$ (Chipman et al., 2010). Second, the recent works of Rockova and van der Pas (2017) and Linero and Yang (2018) show that BART models are capable of adapting to low-order interactions in the covariates. Third, as we will see in Section 4, empirically, BART with a finite T tends to perform better than the limiting model as $T \to \infty$.

2.2 *The shared forests model*

We consider a generalization of (1) and set $(Y_i \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\omega}) \sim f\{y \mid \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{\omega}\}$, where $\boldsymbol{h} = (h_1, \ldots, h_M)$ is a collection of m functions and each $h_m(\boldsymbol{x})$ is modeled nonparametrically as a sum of regression trees. Note that, as illustrated in Figure 1, we are effectively modeling $h_m(\boldsymbol{x})$ in terms of a basis function expansion from an overcomplete family of basis functions

$$h_m(\boldsymbol{x}) = \sum_{t=1}^{T} \sum_{\ell \in \mathcal{L}_t^{(m)}} \mu_{t\ell}^{(m)} \psi_{t\ell}^{(m)}(\boldsymbol{x}), \qquad \psi_{t\ell}^{(m)}(\boldsymbol{x}) = I[\boldsymbol{x} \rightsquigarrow (t, \ell)],$$

where $[\boldsymbol{x} \rightsquigarrow (t, \ell)]$ occurs if $\boldsymbol{x}$ is associated to leaf ℓ of tree $\mathcal{T}_t^{(m)}$ and $I(A)$ is the indicator that the event A occurs. We can then view $\{\psi_{t\ell}^{(m)} : 1 \leqslant t \leqslant T, 1 \leqslant m \leqslant M, \ell \in \mathcal{L}_t^{(m)}\}$ as a collection of features which are adaptively learned from the data to approximate $\{h_1(\boldsymbol{x}), \ldots, h_M(\boldsymbol{x})\}$.

Our proposed shared forest framework assumes that these basis functions are shared across M model components; that is, we assume $\psi_{t\ell}^{(m)}(\boldsymbol{x}) \equiv \psi_{t\ell}(\boldsymbol{x})$ for $m = 1, \cdots, M$. Equivalently, we assume that the features which are useful for approximating $h_m(\boldsymbol{x})$ are the same features that are useful for approximating $h_{m'}(\boldsymbol{x})$. Note, however, that a given feature $\psi_{t\ell}(\boldsymbol{x})$ has different unknown coefficients (effects) $\mu_{t\ell}^{(1)}$ and $\mu_{t\ell}^{(2)}$ respectively for $h_1(\boldsymbol{x})$ and $h_2(\boldsymbol{x})$.

This shared basis function framework is imposed by assuming that the $h_m(\boldsymbol{x})$'s are modeled using T regression trees with the same collection of trees for all M model components that are potentially affected by the covariate vector $\boldsymbol{x}$. That is, we assume

$$h_m(\boldsymbol{x}) = \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t^{(m)}) \ , \tag{3}$$

where $\boldsymbol{\mu}_{t\ell} = (\mu_{t\ell}^{(m)} : 1 \leqslant m \leqslant M)$. We will assume the multivariate prior density $\boldsymbol{\mu}_{t\ell} \sim \pi_{\boldsymbol{\mu}}$, potentially allowing dependence across parameters $\mu_{t\ell}^{(m)}$ for different values of m.

Example 1: Consider a mixed-scale response $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Z_i)$ in which $Y_{ij} = h_j(\boldsymbol{x}) + \epsilon_j$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, $Z_i \sim \text{Bernoulli}[\Phi\{h_3(\boldsymbol{x})\}]$, and $\boldsymbol{h}$ is modeled with a shared forest with $\boldsymbol{\mu}_{t\ell} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\mu / T)$. We consider a variant of this model in Section 4.

Example 2: Consider a semicontinuous response $(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\omega})$ where $Y_i > 0$ occurs with probability $\Phi\{h_1(\boldsymbol{x})\}$ and $(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\omega}) \sim \text{Gam}[\alpha, \alpha \exp\{h_2(\boldsymbol{x})\}]$. We refer to this model as the gamma hurdle model; see Section 3.2.

Bayesian inference for the shared forest model of (3) can be conducted by extending (2) to incorporate priors on the parameters for the leaf nodes across the multiple model components giving the integrated likelihood

$$\begin{aligned}\Lambda(\mathcal{T}_t) &= \pi_{\mathcal{T}}(\mathcal{T}_t) \int \prod_{i=1}^{n} f\{Y_i \mid \boldsymbol{h}(\boldsymbol{X}_i), \boldsymbol{\omega}\} \left[\prod_{\ell \in \mathcal{L}_t} \pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}_{t\ell}) \, d\boldsymbol{\mu}_{t\ell}\right] \\ &= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \int \prod_{i : \boldsymbol{X}_i \rightsquigarrow (t,\ell)} f\{Y_i \mid \boldsymbol{h}(\boldsymbol{X}_i), \boldsymbol{\omega}\} \, \pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}_{t\ell}) \, d\boldsymbol{\mu}_{t\ell} \ .\end{aligned} \tag{4}$$

As before, if (4) has a closed form then one can update $\mathcal{T}_t$ within an MCMC algorithm using standard Metropolis-Hastings proposals.

2.3 *Related methods*

There are several proposals for BART models which are related to our shared forests model. Hahn et al. (2017) consider a related structure in the context of causal inference; given a binary treatment z, they model potential outcomes $Y_i(z)$ as $Y_i(z) = h(\boldsymbol{X}_i) + z\alpha(\boldsymbol{X}_i) + \epsilon_i$, with both $h(\boldsymbol{x})$ and $\alpha(\boldsymbol{x})$ modeled using BART priors. This is referred to as a *Bayesian causal forest* (BCF). The function $h(\boldsymbol{x})$ represents the prognostic effects of the covariates $\boldsymbol{X}_i$, which is shared across both potential outcomes, while $\alpha(\boldsymbol{x})$ represents a treatment-covariate interaction, which is unique to the treated group. This differs from the shared forests framework we present in that we only require sharing the tree topologies across model components, while the BCF model shares the entire function $h(\boldsymbol{x})$. Alternatively, one may view the BCF model for $h(\boldsymbol{x})$ as a shared forest in which the values in leaf ℓ, given by $(\mu_{\ell,0}, \mu_{\ell,1})$, are perfectly correlated. In the context of causal inference, this separation of the effect into a perfectly-shared forest $h(\boldsymbol{x})$ and a completely separate treatment effect $\alpha(\boldsymbol{x})$ is desirable because it gives the analyst a great deal of control over the prior information expressed about individual-level treatment effects.

In the context of functional regression, Starling et al. (2018) model a temporally-observed response using a BART model as $Y_i(t) = h_t(\boldsymbol{X}_i) + \epsilon_i(t)$ where here $t \in \mathscr{T}$ indexes the observation time. The parameters of the leaf nodes of the trees in the ensemble are then modeled as random functions $\mu_\ell(t)$, with Gaussian process priors. The distinction between how t and $\boldsymbol{x}$ are incorporated into their model is referred to as *targeted smoothing*, as the model induces a higher degree of smoothing over t than $\boldsymbol{x}$. This approach can also be cast as a type of shared forest model in which the collection of regression functions $\{h_t(\boldsymbol{x}) : t \in \mathscr{T}\}$ share the same tree topology. The dependence between $h_t(\boldsymbol{x})$ and $h_{t'}(\boldsymbol{x})$ induced using Gaussian processes is analogous to using the multivariate normal prior described in Example 1.

Shared forests have natural connections to many proposals for *multi-task* or *multi-output* methods in machine learning. The most immediate connections are with multi-objective decision trees (MODTs) initially proposed by De'Ath (2002). MODTs are grown in a CART-like fashion, but use a multivariate purity function for evaluating the quality of splits. In this way, splits are useful for predicting all outputs simultaneously. Our shared forests model with $T = 1$ is essentially a Bayesian version of a MODT, as the marginal likelihood (4) will be large when $\mathcal{T}_t$ gives good predictions across all tasks. MODTs can be ensembled in the usual ways via the bagging and random forests algorithms (Kocev et al., 2007).

Our characterization of the shared forests model in terms of a shared basis function expansion is similar to the assumption of the multi-task feature learning approach of Argyriou et al. (2007). Our approach can also be related to the FIRE algorithm for fitting rule ensembles proposed by Aho et al. (2012). Each $\psi_{t\ell}(\boldsymbol{x})$ in the ensemble is a "rule" and the $\mu_{t\ell}$'s are task-specific weights assigned to each rule. Additionally, in the same way that BART is analogous to gradient boosted decision trees (Chipman et al., 2010; Freund et al., 1999), the shared forests model is analogous to the boosted multi-task learning approach of Chapelle et al. (2011).

Using the connection between BART and Gaussian processes, we can also interpret the shared forests model in terms of multi-task kernel methods. Recall from Section 2.1 that BART can be thought of as approximately implementing Gaussian process regression when the number of trees T is large, with kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \sigma_\mu^2 \Pr(\boldsymbol{x} \sim \boldsymbol{x}')$. If the leaf nodes of the ensemble across the tasks are endowed with the prior $\boldsymbol{\mu}_{t\ell} \sim \mathcal{N}(0, \Sigma_\mu / T)$ then the cross-task kernel function is given by $\operatorname{Cov}\{h_j(\boldsymbol{x}), h_k(\boldsymbol{x})\} = \Sigma_{ij} \Pr(\boldsymbol{x} \sim \boldsymbol{x}')$. This matches the proposal of Bonilla et al. (2008) for multi-task Gaussian processes. Sharing of information across tasks for the shared forest occurs at a deeper level still, however, as in the finite-T case the rule-sharing interpretation of our approach still applies even if $\Sigma_{ij} = 0$; as we show in the simulation study of Section 4, substantial gains are possible even with $\Sigma_{ij} = 0$.

3. Models for semicontinuous data

3.1 *Probit-based hurdle models*

Motivated by the MEPS dataset, we present two models for analyzing zero-inflated responses. Throughout, let $\pi(\boldsymbol{x}) = \Pr(Y_i > 0 \mid \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\omega})$ denote the probability of a non-zero response. The gamma hurdle and log-normal hurdle models below are special cases of the probit-based hurdle model, where $\pi(\boldsymbol{x}) = \Phi\{\theta_0 + h_\theta(\boldsymbol{x})\}$, and $(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \boldsymbol{\omega}) \sim f\{y \mid \boldsymbol{h}_u(\boldsymbol{x}), \boldsymbol{\omega}\}$. Here, $\{f(\cdot \mid \mu, \boldsymbol{\omega})\}$ is a parametric family of densities for the positive part of Y_i. We model $\boldsymbol{h} = (h_\theta, \boldsymbol{h}_u)$ with a shared forest. Let $\theta_{t\ell}$ denote the parameter associated to leaf ℓ of $\mathcal{T}_t$ for h_θ and $u_{t\ell}$ the parameter associated to leaf ℓ of $\mathcal{T}_t$ for $\boldsymbol{h}_u$. We use independent priors for the $\theta_{t\ell}$'s and $u_{t\ell}$'s and, following Chipman et al. (2010), set $\theta_{t\ell} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\theta^2)$.

For the sake of computational convenience, we do not use (4) directly, but instead augment the data with latent variables $Z_i \overset{\text{indep}}{\sim} \mathcal{N}\{\theta_0 + h_\theta(\boldsymbol{X}_i), 1\}$ such that $Y_i > 0$ if-and-only-if $Z_i > 0$ (Albert and Chib, 1993). Before computing (4) we first sample the Z_i's from a $\mathcal{N}\{\theta_0 + h_\theta(\boldsymbol{X}_i), 1\}$ distribution, truncated to $(-\infty, 0)$ or $(0, \infty)$ according as

$Y_i = 0$ or $Y_i > 0$. We then compute the integrated likelihood

$$\begin{aligned}\Lambda(\mathcal{T}_t) &= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} \Bigg[\int \prod_{i : \boldsymbol{X}_i \rightsquigarrow (t,\ell)} \mathcal{N}\{Z_i \mid \theta_0 + h_\theta(\boldsymbol{X}_i), 1\} \, \mathcal{N}(\theta_{t\ell} \mid 0, \sigma_\theta^2) \, d\theta_{t\ell} \\ &\qquad \times \int \prod_{i : Z_i > 0, \boldsymbol{X}_i \rightsquigarrow (t,\ell)} f\{Y_i \mid \boldsymbol{h}_u(\boldsymbol{X}_i), \boldsymbol{\omega}\} \, \pi_u(u_{t\ell}) \, du_{t\ell} \Bigg] \\ &= \pi_{\mathcal{T}}(\mathcal{T}_t) \prod_{\ell \in \mathcal{L}_t} L_\theta(t,\ell) \cdot L_u(t,\ell) \ .\end{aligned} \tag{5}$$

Notice that $L_\theta(t,\ell)$ does not depend on our choice for the distribution of the non-zero Y_i's and can be computed in closed form; an expression for $L_\theta(t,\ell)$ is given in the web appendix. Hence, all that must be done to apply the probit-based hurdle model is to be able to compute $L_u(t,\ell)$ in closed form.

3.2 *Gamma hurdle models*

Our semiparametric gamma hurdle model sets $Y_i \sim \text{Gam}[\alpha, \alpha \exp\{\lambda_0 + h_\lambda(\boldsymbol{x})\}]$ conditional on $Y_i > 0$ and $\boldsymbol{X}_i = \boldsymbol{x}$, where $\text{Gam}(\alpha, \beta)$ is parameterized to have mean α/β and variance α/β^2. We model $h_\theta(\boldsymbol{x})$ and $h_\lambda(\boldsymbol{x})$ with a shared forest, $h_\theta(\boldsymbol{x}) = \sum_{t=1}^T g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_{\theta,t})$, and $h_\lambda(\boldsymbol{x}) = \sum_{t=1}^T g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_{\lambda,t})$. Note that, under this model, we have

$$\begin{aligned} E(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \alpha) &= \exp\{-\lambda_0 - h_\lambda(\boldsymbol{x})\}, \\ \text{Var}(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{h}, \alpha) &= \frac{\exp[-2\{\lambda_0 + h_\lambda(\boldsymbol{x})\}]}{\alpha}, \end{aligned} \tag{6}$$

so that the conditional standard deviation of Y_i is proportional to its mean.

The leaf-specific parameters for $h_\lambda(\boldsymbol{x})$ are given log-gamma priors $\lambda_{t\ell} \sim \log \text{Gam}(\alpha_\lambda, \beta_\lambda)$. The log-gamma prior is chosen because it is conjugate to the gamma likelihood and makes computation of (5) tractable. Under this prior for the leaf parameters, the gamma hurdle model is immediately applicable provided that we can compute the likelihood factor

$$L_\lambda(t,\ell) = \int \prod_{i \in \ell : Y_i > 0} \text{Gam}[Y_i \mid \alpha, \alpha \exp\{\lambda_0 + h_\lambda(\boldsymbol{x})\}] \, \log \text{Gam}\{\lambda_{t\ell} \mid \alpha_\lambda, \beta_\lambda\}.$$

To do this, similar to Murray (2017) for loglinear models, we define $\eta_i = \alpha \exp\{\lambda_0 + h_\lambda(\boldsymbol{X}_i) - g(\boldsymbol{X}_i; \mathcal{T}_t, \mathcal{M}_{\lambda,t})\}$. By analogy with the usual Bayesian backfitting algorithm of Chipman et al. (2010), the η_i's play the role of the backfitted response. Let $A_\ell = \{i : i \in \ell, Y_i > 0\}$ and $N_\ell = |A_\ell|$. Then

$$\prod_{i \in A_\ell} \text{Gam}[Y_i \mid \alpha, \alpha \exp\{\lambda_0 + h_\lambda(\boldsymbol{X}_i)\}] = \left(\prod_{i \in A_\ell} \frac{(Y_i \eta_i)^\alpha}{Y_i \Gamma(\alpha)} \right) \exp\left(\alpha N_\ell \lambda_{t\ell} - \sum_{i \in A_\ell} Y_i \eta_i e^{\lambda_{t\ell}} \right).$$

Integrating against the $\log \mathrm{Gam}(\lambda_{t\ell} \mid \alpha_\lambda, \beta_\lambda)$ density gives

$$L_\lambda(t, \ell) = \frac{\prod_{i \in A_\ell} \eta_i^\alpha Y_i^{\alpha - 1}}{\Gamma(\alpha)^{N_\ell}} \cdot \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)} \cdot \frac{\Gamma(\alpha_\lambda + N_\ell \alpha)}{(\beta_\lambda + \sum_{i \in A_\ell} Y_i \eta_i)^{\alpha_\lambda + N_\ell \alpha}}.$$

Hence (5) can be computed in closed form. Additionally, by conjugacy of the log-gamma distribution, we have the full conditionals

$$\lambda_{t\ell} \sim \log \mathrm{Gam}\left(\alpha_\lambda + \alpha N_\ell, \beta_\lambda + \sum_{i \in A_\ell} Y_i \eta_i\right). \tag{7}$$

A detailed Markov chain Monte Carlo algorithm is given in the web appendix.

3.3 *Log-normal hurdle model*

A shortcoming of the gamma hurdle model is that the relationship between $\boldsymbol{x}$ and the variance is captured entirely through the mean. As an alternative, we propose the heteroskedastic log-normal hurdle model with $\pi(\boldsymbol{x}) = \Phi\{\theta_0 + h(\boldsymbol{x})\}$ and $(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x}, h, mu, \sigma^2) \sim \log \mathcal{N}\{\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\}$. We again use a shared forest to model the three functions

$$\mu(\boldsymbol{x}) = \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_{\mu,t}), \qquad \sigma^{-2}(\boldsymbol{x}) = \exp\left\{\lambda_0 + \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_{\lambda,t})\right\},$$

$$h(\boldsymbol{x}) = \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_{\theta,t}).$$

The resulting model for the mean and variance of $(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x})$ is given by $m(\boldsymbol{x}) = \exp\{\mu(\boldsymbol{x}) + \sigma^2(\boldsymbol{x})/2\}$ and $s^2(\boldsymbol{x}) = m(\boldsymbol{x})^2[\exp\{\sigma^2(\boldsymbol{x})\} - 1]$.

Like the gamma hurdle model, when a *homoskedastic* model for $\log Y_i$ is used, we find that the mean $m(\boldsymbol{x})$ is proportional to the standard deviation $s(\boldsymbol{x})$. By modeling $\sigma^2(\boldsymbol{x})$ nonparametrically, however, we allow for more complex relationships between $m(\boldsymbol{x})$ and $s(\boldsymbol{x})$. Our heteroskedastic model for the $\log Y_i$'s is similar to the heteroskedastic BART models proposed by Murray (2017) and Pratola et al. (2017), but our model differs in two respects. First, the trees used to model the mean and variance functions are shared, which is helpful because the variance function $\sigma^2(\boldsymbol{x})$ is generally much more weakly identified than the mean $\mu(\boldsymbol{x})$. Second, our choice of prior for $\sigma^2(\boldsymbol{x})$ will explicitly shrink the posterior model towards a homoskedastic model; see Section 3.4.

Let $\mu_{t\ell}$ and $\lambda_{t\ell}$ be the leaf parameters associated to leaf ℓ of $\mathcal{T}_t$ for $\mu(\cdot)$ and $\sigma(\cdot)$ respectively and let $\tau_{t\ell} = \exp(\lambda_{t\ell})$. We use a normal-gamma prior for $(\mu_{t\ell}, \tau_{t\ell})$, i.e., $\tau_{t\ell} \sim \mathrm{Gam}(\alpha_\lambda, \beta_\lambda)$ and $\mu_{t\ell} \sim \mathcal{N}\{0, 1/(\kappa \tau_{t\ell})\}$. This normal-gamma prior allows for computation of the likelihood factor

$$L_{\mu,\lambda}(t,\ell) = \int \prod_{i:\boldsymbol{X}_i \rightsquigarrow (t,\ell), Y_i > 0} \log \mathcal{N}\{Y_i \mid \mu(\boldsymbol{X}_i), \sigma^2(\boldsymbol{X}_i)\}$$
$$\times \mathcal{N}\{\mu_{t\ell} \mid 0, 1/(\kappa \tau_{t\ell})\} \; \mathrm{Gam}(\tau_{t\ell} \mid \alpha_\lambda, \beta_\lambda) \; d\mu_{t\ell} \; d\tau_{t\ell}.$$

Let $W_i = \log Y_i$ and suppose $\boldsymbol{X}_i \rightsquigarrow (t,\ell)$. Then, conditional on $Y_i > 0$, we have $W_i \sim \mathcal{N}(\eta_i + \mu_{t\ell}, \frac{1}{\nu_i \tau_{t\ell}})$, where $\eta_i = \sum_{j \neq t} g(\boldsymbol{X}_i; \mathcal{T}_j, \mathcal{M}_{\mu,j})$, and $\nu_i = \exp\left\{\lambda_0 + \sum_{j \neq t} g(\boldsymbol{X}_i; \mathcal{T}_j, \mathcal{M}_{\lambda,j})\right\}$. Let $Q_i = W_i - \eta_i$ and $A(\ell) = \{i : \boldsymbol{X}_i \rightsquigarrow (t,\ell), Y_i > 0\}$; Q_i and ν_i are analogous to the backfitted response in the usual Bayesian backfitting algorithm. We have

$$\prod_{i \in A(\ell)} \mathcal{N}\left(W_i \mid \eta_i + \mu_{t\ell}, \frac{1}{\nu_i \tau_{t\ell}}\right) = \prod_{i \in A(\ell)} \left(\frac{\nu_i \tau_{t\ell}}{2\pi}\right)^{1/2} \exp\left\{-\frac{\nu_i \tau_{t\ell}}{2}(Q_i - \mu_{t\ell})^2\right\}.$$

This likelihood is conjugate to the normal-gamma prior for $(\mu_{t\ell}, \tau_{t\ell})$, and routine computations give the expression

$$\left(\prod_i \sqrt{\frac{\nu_i}{2\pi}}\right) \sqrt{\frac{\kappa}{\kappa + w_\ell}} \frac{\beta_\lambda^{\alpha_\lambda} \Gamma(\alpha_\lambda + N_\ell/2)}{\Gamma(\alpha_\lambda)} \left(\beta_\lambda + \frac{S_\ell^2}{2} + \frac{\kappa w_\ell \bar{Q}_\ell^2}{2(\kappa + w_\ell)}\right)^{-(\alpha_\lambda + N_\ell/2)},$$

for $L_{\mu,\lambda}(t,\ell)$ where

$$\bar{Q}_\ell = \frac{\sum_{i \in A(\ell)} \nu_i Q_i}{\sum_{i \in A(\ell)} \nu_i}, \qquad w_\ell = \sum_{i \in A(\ell)} \nu_i, \qquad \text{and} \qquad S_\ell^2 = \sum_{i \in A(\ell)} \nu_i (Q_i - \bar{Q}_\ell)^2.$$

We again have a closed form for (5). Moreover, we also have the following full conditionals for the leaf parameters:

$$\tau_{t\ell} \sim \mathrm{Gam}(\widehat{\alpha}_\ell, \widehat{\beta}_\ell), \qquad \text{and} \qquad \mu_{t\ell} \sim \mathcal{N}\{\widehat{\mu}_\ell, 1/(\widehat{\kappa}_\ell \tau_{t\ell})\}, \tag{8}$$

where

$$\widehat{\alpha}_\ell = \alpha_\lambda + N_\ell/2, \qquad \widehat{\beta}_\ell = \beta_\lambda + \frac{S_\ell^2}{2} + \frac{\bar{Q}_\ell^2 \kappa w_\ell}{2(\kappa + w_\ell)},$$
$$\widehat{\kappa}_\ell = \kappa + w_\ell, \qquad \widehat{\mu}_\ell = \frac{\sum_{i \in A(\ell)} v_i Q_i}{\widehat{\kappa}}.$$

Additional details for the various steps of the MCMC algorithm are deferred to the web appendix.

3.4 *Prior specification*

An advantage of the BART framework is that there exist standard "default" priors which have proven to work remarkably well in practice. In particular, very little tuning is required to obtain an acceptable baseline level of performance. We develop default priors for the gamma hurdle and log-normal hurdle models we consider here. For both models, we will use the default prior recommended by Chipman et al. (2010) for the $\theta_{t\ell}$'s.

Additionally, we apply a quantile normalization separately to each column of the design matrix $\boldsymbol{X}$ so that the predictors are distributed approximately uniformly on $[0, 1]$.

We first give a prior specification for the log-normal hurdle model. As a preprocessing step, we work with $W_i = \log Y_i$; further, we standardize the finite W_i's to have mean 0 and standard error 1. In order for the prior to be stable as the number of trees is increased, we choose the hyperparameters so that $E(\lambda_{t\ell}) = 0$ and $\mathrm{Var}(\lambda_{t\ell}) = a_\lambda^2/T$, and similarly for $\mu_{t\ell}$ and $\theta_{t\ell}$. This ensures that the stochastic process $\sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t)$ converges to a Gaussian process as $T \to \infty$ so that the prior is stable under adding additional trees.

Appropriate values for $(\alpha_\lambda, \beta_\lambda)$ can be obtained by solving the equations

$$E(\lambda_{t\ell}) = \psi(\alpha_\lambda) - \log \beta_\lambda = 0, \tag{9}$$

$$\mathrm{Var}(\lambda_{t\ell}) = \psi'(\alpha_\lambda) = a_\lambda^2/T. \tag{10}$$

Noting that $\psi'(\alpha) \approx \alpha^{-1}$, (10) implies that for moderate values of T we will have $\alpha \approx T/a_\lambda^2$. Additionally, noting that $\psi(\alpha) \approx \log(\alpha)$, (9) implies that $\alpha_\lambda \approx \beta_\lambda$; in particular, both α_λ and β_λ are roughly proportional to T.

As there is typically less information in the data about the second order effect $\sigma^2(\boldsymbol{x})$ than the first order effect $\mu(\boldsymbol{x})$, it is sensible to shrink our model towards a homoskedastic model. Note that if all the $\lambda_{t\ell}$'s are equal to 0 then the variance function reduces to $\sigma^2(\boldsymbol{x}) = \exp(-\lambda_0)$ so that the model is homoskedastic. Accordingly, we place a half-Cauchy$(0, 1)$ on the baseline standard deviation $\sigma_0 = \exp(-\lambda_0/2)$ and shrink the $\lambda_{t\ell}$'s heavily to zero. As a default, we have found $a_\lambda = 0.5$ to work well in practice; alternatively, one might set $a_\lambda \sim$ half-Cauchy$(0, 1)$ to allow the model to adaptively determine the amount of heteroskedasticity in the data.

Next, by analogy with the prior specification of Chipman et al. (2010), we ensure that the $\mu_{t\ell}$'s marginally have mean 0 and standard deviation $3/(k_\mu\sqrt{T})$ by noting that $\mathrm{Var}(\mu_{t\ell}) = \beta/\{(\alpha - 1)\kappa\}$. As noted above, for moderate T we will have $\alpha_\lambda \approx \beta_\lambda \propto T$, so that $\mathrm{Var}(\mu_{t\ell}) \approx \kappa^{-1}$. This suggests setting $\kappa^{-1/2} = k_\mu/\sqrt{T}$ (or giving $\kappa^{-1/2}$ a prior centered at this value). Here k_μ is a tuning parameter which controls the signal-to-noise ratio and by default we set $k_\mu = 1.5$.

We recommend a similar default prior for the gamma model. We first scale the non-zero Y_i's to have mean 1. As before, we impose the restrictions $E(\lambda_{t\ell}) = 0$ and $\mathrm{Var}(\lambda_{t\ell}) = a_\lambda^2/T$ so that $h_\lambda(\boldsymbol{x}) \mathrel{\dot{\sim}} \mathcal{N}(0, a_\lambda^2)$. This can be accomplished by solving the system of equations (9, 10). As a default, we set $a_\lambda = k_\lambda\sqrt{\mathrm{Var}(\log Y_i \mid Y_i > 0)}$ where k_λ is a user-specified tuning parameter which we set to 1.5. Additionally, we require a prior for the shape parameter α. From (6) we see that $1/\alpha$ is a dispersion parameter. We use a weakly informative half-Cauchy prior $\alpha^{-1/2} \sim \text{half-Cauchy}(0, A)$ for some $A > 0$. For the MEPS data in particular we set $A = 1$ to encourage small values of α, as medical expenditures are highly right-skewed.

3.5 *Identifiability of the model components*

Given that the hurdle models we have proposed are mixture models, there is a question of whether the models we have defined here are identifiable. Let $\mathcal{X}$ denote a predictor space and $(\mathcal{Y}, \mathscr{B})$ be a measurable space. Given a class $\{F_\theta : \theta \in \Theta\}$ where $F_\theta : \mathscr{B} \times \mathcal{X} \to [0, 1]$ is a probability distribution on $(\mathcal{Y}, \mathscr{B})$ for every $\boldsymbol{x} \in \mathcal{X}$, the parameter θ is called *identifiable* if the mapping $\theta \mapsto F_\theta$ is one-to-one (Lehmann and Casella, 2006, Definition 1.5.2). General forms of the hurdle model may not be identifiable, particularly when we are mixing a point mass at 0 with a distribution that also is supported at 0, or if the positive part has probability 0. The following lemma shows that this is essentially the only case in which we might run into identifiability issues. A proof this result is given in the web appendix.

Lemma 1: *Let $\mathcal{X}$ denote an arbitrary set, $(\mathbb{R}, \mathscr{B})$ the Borel measurable reals, and δ_0 the point-mass distribution at 0. Let $\mathscr{G}$ denote the set of conditional distributions with no atoms at 0*

$$\mathscr{G} = \left\{ \begin{matrix} G : \mathscr{B} \times \mathcal{X} \to [0,1] & \textit{such that} & G_{\boldsymbol{x}}(\cdot) \textit{ is a probability distribution} \\ & \textit{and} & G_{\boldsymbol{x}}(\{0\}) = 0 \textit{ for all } \boldsymbol{x} \in \mathcal{X} \end{matrix} \right\}$$

and let $\mathscr{P}$ be the collection of conditional probabilities which are bounded by 1, $\mathscr{P} = \{\pi : \mathcal{X} \to [0, 1)\}$. Let $\mathscr{M}$ denote the collection of conditional distributions on $(\mathbb{R}, \mathscr{B})$ which are not identically 0,

$$\mathscr{M} = \left\{ \begin{matrix} M : \mathscr{B} \times \mathcal{X} \to [0,1] & \textit{such that} & F_{\boldsymbol{x}}(\cdot) \textit{ is a probability distribution} \\ & \textit{and} & F_{\boldsymbol{x}}(\{0\}) \neq 1 \textit{ for all } \boldsymbol{x} \in \mathcal{X} \end{matrix} \right\}.$$

Then the mapping $\mathscr{G} \times \mathscr{P} \to \mathscr{M}$ given by $(G, \pi) \mapsto \pi(\boldsymbol{x})\delta_0 + \{1 - \pi(\boldsymbol{x})\}G_{\boldsymbol{x}}$, is a bijection.

A consequence of this result is that the semiparametric hurdle models developed in Section 3.2 and Section 3.3 are also identifiable when the model parameters are understood to be the nonparametrically-specified functions $\boldsymbol{h}$ and the parametric component $\boldsymbol{\omega}$ (noting that $(\boldsymbol{h}, \boldsymbol{\omega})$ maps in a one-to-one fashion to (G, π)). The individual trees in the ensemble are, however, not identifiable, as the collection of possible regression trees is an overcomplete basis. In practice, we are usually only interested in recovering $\boldsymbol{h}$ rather than the individual trees, so that this lack of identifiability is not a concern.

4. Simulation study

In this section, we examine the benefits of sharing information across related tasks using a simple simulation study. We consider a mixed response

$$\Pr(Z_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) = \Phi\{\sigma_\theta \, h(\boldsymbol{x})\}, \qquad (Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}) \sim \mathcal{N}\{h(\boldsymbol{x}), \sigma^2\}, \qquad (11)$$

with $(Z_i \perp Y_i \mid \boldsymbol{X}_i = \boldsymbol{x})$. This is similar to the zero-inflated response setting, but with the continuous portion of the distribution always observed (see also Example 1). Note that the information in $\boldsymbol{X}_i$ is captured by the one-dimensional summary $h(\boldsymbol{X}_i)$ which is shared across both models. We emphasize that the structure (11) is not assumed by the shared forest model - only the basis functions are shared - and must effectively be learned from the data. We consider the benchmark function given by Friedman (1991)

$$h(\boldsymbol{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5.$$

We sample $\boldsymbol{X}_i$ uniformly distributed on $[0,1]^P$; if $P > 5$ then the predictors $X_{i6}, \dots, X_{iP}$ have no influence on the response. We compare the shared forest to an approach which fits separate BART models to $(Z_i \mid \boldsymbol{X}_i = \boldsymbol{x})$ and $(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x})$ so that information is not shared across tasks. Our focus is on how well these models estimate $\Pr(Z_i = 1 \mid \boldsymbol{X}_i) = \pi(\boldsymbol{X}_i)$ as measured by the cross-entropy between the true and estimated π's,

$$\text{Loss} = \int \left[\pi(\boldsymbol{x}) \log\left(\frac{\pi(\boldsymbol{x})}{\widehat{\pi}(\boldsymbol{x})}\right) + \{1 - \pi(\boldsymbol{x})\} \log\left(\frac{1 - \pi(\boldsymbol{x})}{1 - \widehat{\pi}(\boldsymbol{x})}\right)\right] \, d\boldsymbol{x},$$

which is computed by Monte Carlo integration. We focus on the setting in which the continuous response Y_i is relatively informative while the information contained in Z_i is relatively weak by fixing $\sigma^2 = 1$. We consider a training set of size $n = 250$ for both the Y_i's and the Z_i's.

Results are given in Figure 2, with 20 replications per simulation setting. In the left panel, we fix $\sigma_\theta = 4$ (roughly corresponding to $\pi(\boldsymbol{X}_i) \sim \text{Uniform}(0,1)$) and examine how sharing impacts the loss as P varies from $P = 5$ to $P = 250$. We see that, as the variable selection task becomes more difficult, the model which does not share information is far more sensitive to irrelevant predictors than the model which does share. This is because the Y_i's are much more informative about the relevant predictors than the Z_i's, so that the shared model can do a much better job of selecting the relevant predictors. In the right panel, we fix $P = 20$ and vary the signal level σ_θ from 1 to 20. In this case, the gain from sharing is essentially constant, with higher losses for higher signal levels.

A potentially important tuning parameter in BART models is the number of trees T

used in the ensemble. Other works have supported the following overall trend: predictive performance of BART models is typically insensitive to the number of trees included, provided that we include sufficiently many. We find that this behavior holds for the shared forests model as well, with Figure 3 summarizing the results of the simulation experiment with $P = 20$ and $\sigma_\theta = 4$ fixed as a function of T. As before, results are based on 20 replications of the experiment for each setting. As expected we see a slight decrease in performance as T increases from the optimal choice.

5. Analysis of MEPS data

Our motivating example is from the 2015 Medical Expenditure Panel Survey (MEPS). The MEPS study (Natarajan et al., 2008) was developed to estimate national and regional health care use and expenditures in the United States. We first illustrate the capability of the proposed model to effectively capture heteroskedasticity in the MEPS data. We analyze data from 10,729 adult females who participated in the survey, focusing on the total medical expenditure during 2015, denoted Y_i. Previous analyses of this dataset have suggested taking $\text{Var}(Y_i) = \phi E(Y_i)^{1.5}$ (Blough and Ramsey, 2000; Natarajan et al., 2008). We consider a list of predictors including, among other things, age, race, family income, whether the individual smokes, perceived health, body mass index, and number of visits to the dentist over the survey period; a full list of predictors is given in the web appendix.

We fit the log-normal hurdle and gamma hurdle models to the data. We examine the fit of these models to the positive part of the data $(Y_i \mid Y_i > 0, \boldsymbol{X} = \boldsymbol{x})$ by considering the generalized residuals (Cox and Snell, 1968) $r_i = \Phi^{-1}\{\widehat{F}_{\boldsymbol{X}_i}(Y_i)\}$ where $\widehat{F}_{\boldsymbol{x}}$ is an estimate of the cdf of $(Y_i \mid Y_i > 0, \boldsymbol{X}_i = \boldsymbol{x})$ obtained from the model. In the case of the log-normal hurdle model, r_i is equivalent to the usual standardized residual $\{\log Y_i - \widehat{\mu}_i(\boldsymbol{x})\}/\widehat{\sigma}(\boldsymbol{x})$; for comparison, we also consider the raw residuals $\log Y_i - \widehat{\mu}_i(\boldsymbol{x})$ to examine the effect of heteroskedasticity on the model fit.

Quantile-quantile plots of the residuals compared to a reference Gaussian distribution are given in Figure 4. We see that the log-normal hurdle model fits the data very well. Additionally, we see that ignoring heteroskedasticity causes a poor fit in the left tail of the data, corresponding to individuals with lower healthcare costs. By comparison, the gamma model fits poorly. We also consider a generalized gamma distribution (Stacy,

1962) which models Y_i^{ϕ} with a gamma distribution, where ϕ is learned from the data. The generalized gamma model fits roughly as well as a homoskedastic log-normal model, but is inferior to the heteroskedastic log-normal model due to the stringent relationship between the mean and the variance implied by the generalized gamma model.

In addition to fitting the data well, the heteroskedastic log-normal model provides several interesting insights into the nature of the heteroskedasticity in the data. Let $\widehat{m}(\boldsymbol{x})$ and $\widehat{s}(\boldsymbol{x})$ denote the posterior mean of $m(\boldsymbol{x})$ and $s(\boldsymbol{x})$ given in Section 3.3. The top panel of Figure 5 gives a plot of $\widehat{m}(\boldsymbol{X}_i)$ against $\widehat{s}(\boldsymbol{X}_i)$ on the log-log scale. To aide visualization, points with similar values of $(\widehat{m}(\boldsymbol{X}_i), \widehat{s}(\boldsymbol{X}_i))$ are grouped into hexagonal tiles and are shaded according to the *average number of dentist visits per subject* within each tile.

There are several interesting features of the top panel of Figure 5. First, there is near-linear relationship between $\log \widehat{m}(\boldsymbol{X}_i)$ and $\log \widehat{s}(\boldsymbol{X}_i)$. An ordinary least squares (OLS) fit of $\log \widehat{m}(\boldsymbol{X}_i)$ to $\log \widehat{s}(\boldsymbol{X}_i)$ has slope 0.7556 and an R^2 of 82%. Hence, the OLS fit suggests the approximation $\widehat{s}^2(\boldsymbol{X}_i) \propto \widehat{m}(\boldsymbol{X}_i)^{1.511}$, which agrees nearly exactly with Blough and Ramsey (2000). Second, by shading the hexagonal tiles by the number of dentist visits, we see clearly that the mean *does not* account for all of the heteroskedasticity due to the predictors. We see, for example, that individuals with lower numbers of visits to the dentist tend to have a standard deviation which is higher than what would be predicted by the mean alone. To understand this relationship better, we let $\delta = \log \widehat{s}(\boldsymbol{X}_i) - 0.7556 \log \widehat{m}(\boldsymbol{X}_i) - 2.672$ denote the residual in predicting $\log \widehat{s}(\boldsymbol{X}_i)$ with $\log \widehat{m}(\boldsymbol{X}_i)$ by OLS. The bottom panel of Figure 5 shows how the distribution of δ varies across the number of dentist trips and the individual's perceived health status. We see first that individuals with fewer dentist trips have standard deviations which are larger than what would be predicted using only the mean; similarly, individuals with higher perceived health status scores (corresponding to *lower* perceived health) also tend to have higher variability than would be predicted by the mean alone.

To assess whether there is a benefit of using the shared forests methodology for the MEPS data, we compute the log pseudo-marginal likelihood for the shared forest model and an equivalent model which does not share the trees across model components. Specifically, we fit the heteroskedastic BART (HBART) model of Pratola et al. (2017), which sets $Y_i = g(\boldsymbol{X}_i) + s(X_i)\epsilon_i$ with $g(\boldsymbol{X}_i)$ and $\log s(\boldsymbol{X}_i)$ given BART priors. HBART was fit to the non-zero observed Y_i's. The probability of a zero response was modeled using a binary BART model with a probit link, i.e., $\Pr(Y_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}) = \Phi\{h(\boldsymbol{x})\}$ where

$h(\cdot)$ was given a BART prior. The log pseudo-marginal likelihood (LPML) is given by $\text{LPML} = \sum_{i=1}^{n} \log \text{CPO}_i$ where $\text{CPO}_i = f(Y_i \mid \boldsymbol{Y}_{-i}, \boldsymbol{X})$ is the predictive density of the i^{th} observation given $\boldsymbol{Y}_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ (Geisser and Eddy, 1979). Results are given in Table 1. The LPML was computed using the Markov chain output using the `loo` package in `R` (Vehtari et al., 2017). Multiple fits of the model using different seeds for the MCMC algorithm give qualitatively similar results.

We see that the shared forest gives a substantial boost in LPML for the binary component of the model. This suggests that the features learned from the continuous part of the model are helpful in determining whether an individual incurs any medical expense. We also observe a less substantial, but still large, improvement in performance for the regression model.

6. Discussion

In this paper we introduced shared forests and demonstrated their usefulness on both simulated data and data from the MEPS dataset. Additionally, we introduced two novel models for semicontinuous data: a gamma hurdle model and a heteroskedastic log-normal hurdle model.

There are several promising areas for future work. First, there are other possibilities for sharing information across nonparametric components. Here we have restricted the components to share the same basis function expansion. To make the models more tightly coupled, one might consider shrinking together the coefficients of these expansions; an example where this might be useful is in meta-regression, where one would expect both that features across different studies will exert similar (but not necessarily identical) effects on the outcome. In the other direction, one might allow the models to share a *subset* of the basis functions; for example, each model component might consist of a shared forest combined with an *innovation forest* which is specific to each task. This structure is likely to be useful if only a subset of relevant features are shared across nonparametric components. A special case of such a construction is given by Hahn et al. (2017) to estimate heterogeneous causal effects; in our terminology, their model consists of a shared forest which captures the prognostic features of covariates which are shared across treatment levels $z = 1$ and $z = 0$ and an innovation forest which is specific to the treatment $z = 1$.

Additionally, Linero and Yang (2018) recently demonstrated that the discrete nature of decision trees can lead to suboptimal performance on both a theoretical and practical level, and that this can be corrected by replacing the usual decision trees with *smooth* decision trees. The shared forests framework can easily be extended to allow for smooth decision trees for the homoskedastic log-normal hurdle model, but non-trivial modifications are required to apply this strategy to the heteroskedastic log-normal and gamma hurdle models.

Acknowledgements

This work was partially supported by NSF grant DMS-1712870 and by the Office of the Secretary of Defense, Directorate of Operational Test and Evaluation and the Test Resource Management Center under the Science of Test research consortium.

References

Aho, T., Ženko, B., Džeroski, S., and Elomaa, T. (2012). Multi-target regression with rule ensembles. *Journal of Machine Learning Research*, 13(1):2367–2407.

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the american statistical association*, 50(271):901–908.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pages 41–48.

Athreya, K. B. and Ney, P. E. (2004). *Branching processes*. Courier Corporation.

Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of Statistics*, 42(1):352.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Blough, D. K. and Ramsey, S. D. (2000). Using generalized linear models to assess medical care costs. *Health Services and Outcomes Research Methodology*, 1(2):185–202.

Bonilla, E. V., Chai, K. M., and Williams, C. (2008). Multi-task gaussian process prediction. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pages 153–160.

Borchani, H., Varando, G., Bielza, C., and Larrañaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233.

Chapelle, O., Shivaswamy, P., Vadrevu, S., Weinberger, K., Zhang, Y., and Tseng, B. (2011). Boosted multi-task learning. *Machine Learning*, 85(1-2):149–173.

Chipman, H., George, E. I., Gramacy, R. B., and McCulloch, R. (2013). Bayesian treed response surface models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):298–305.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B*, pages 248–275.

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, 83(4):1105–1117.

Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika*, 66(3):655–658.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 4(5):771–780.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Hahn, P. R., Murray, J. S., and Carvalho, C. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Kocev, D., Vens, C., Struyf, J., and Džeroski, S. (2007). Ensembles of multi-objective decision trees. In *Proceedings of the 18th European Conference on Machine Learning*, pages 624–631. Springer.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Linero, A. R. (2017). A review of tree-based bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society, Series B*, 80(5).

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arXiv:1701.01503*.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G., and Gonin, R. (2008). Variance estimation in complex survey sampling for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(1):75–87.

Pratola, M. (2016). Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3):885–911.

Pratola, M., Chipman, H., George, E., and McCulloch, R. (2017). Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*.

Rockova, V. and van der Pas, S. (2017). Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1078.08734*.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R., and Scott, J. G. (2018). Functional response regression with funbart: an analysis of patient-specific stillbirth risk. *arXiv preprint arXiv:1805.07656*.

Tu, W. (2006). Zero-inflated data. *Encyclopedia of environmetrics*, 6.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Xiao-Hua, Z. and Tu, W. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics*, 55(2):645–651.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602.

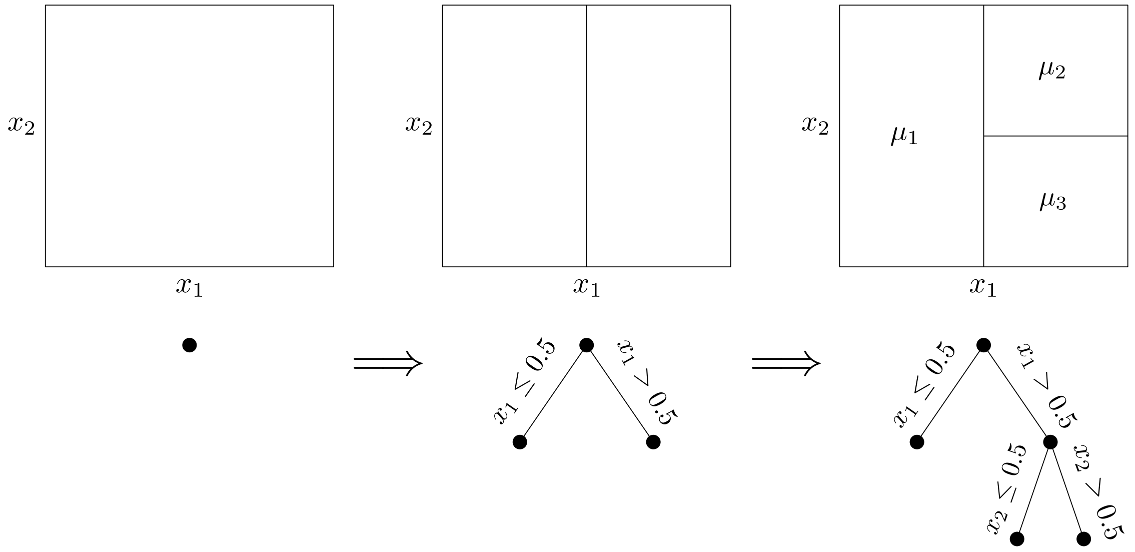

Figure 1. Schematic illustration of the construction of a decision tree (bottom) with the induced recursive partitioning of the predictor space $\mathcal{X} = [0, 1]^2$. After the decision tree is constructed, parameters associated to leaf node ℓ are given a mean parameter μ_ℓ.

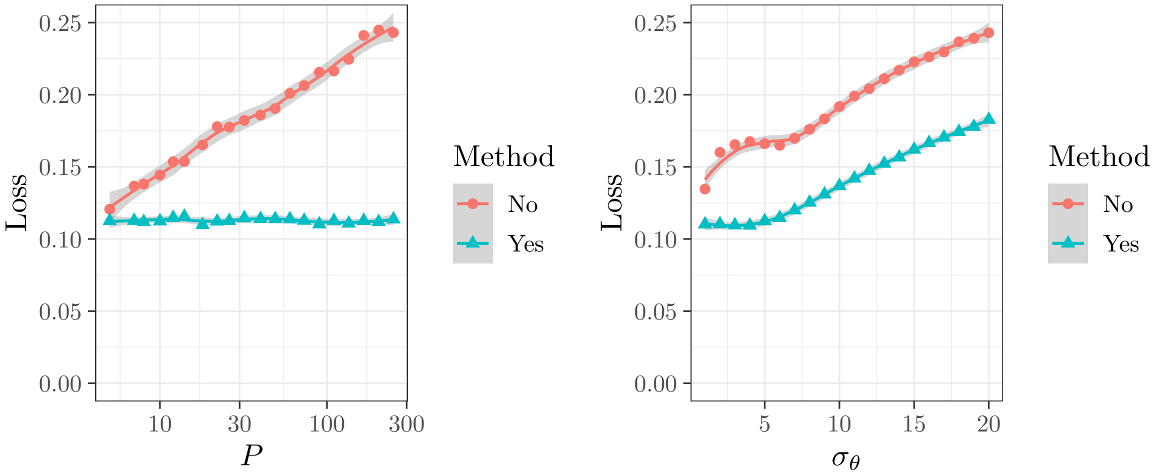

Figure 2. Left: average value of Loss for $\log P \in (\log 5, \log 250)$ averaged over 20 replications. Right: average value of Loss for σ_θ ranging from 1 to 20. "No" indicates the single BART model while "Yes" indicates use of the shared forest. This figure appears in color in the electronic version.

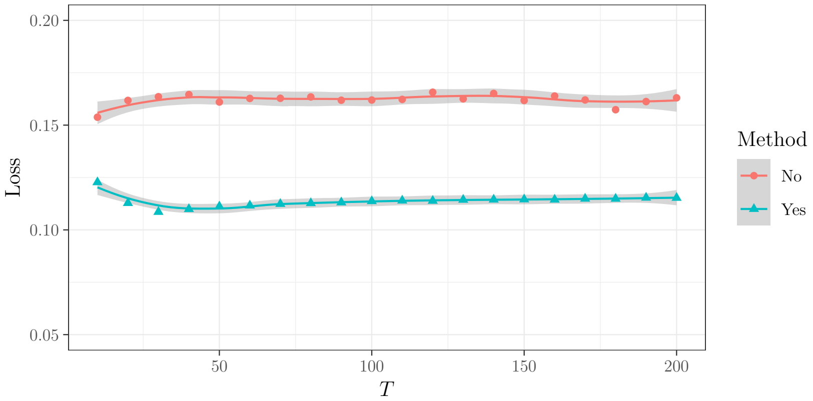

Figure 3. Average values of Loss for T evenly space in $[10, 200]$ averaged over 20 replications. This figure appears in color in the electronic version.

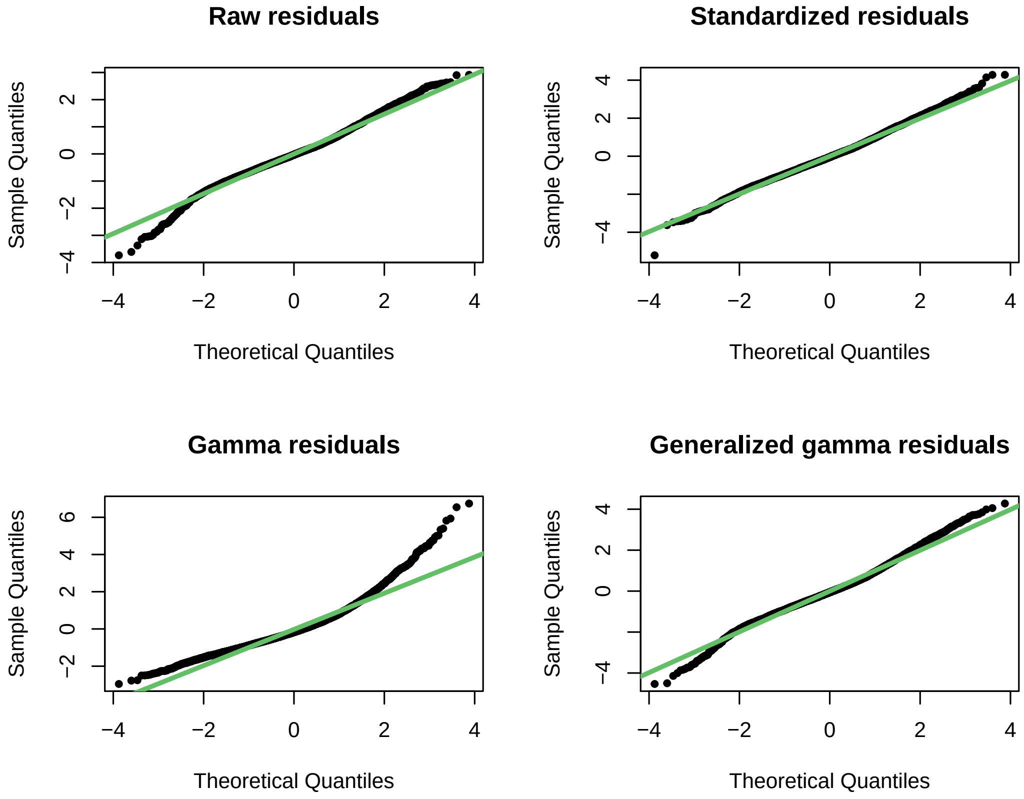

Figure 4. Quantile-quantile plots comparing the residuals r_i for each model to a reference normal distribution. The top panels give the raw residuals $\log Y_i - \widehat{\mu}(\boldsymbol{X}_i)$ (left) and standardized residual r_i for the log-normal hurdle model (right). The bottom panels give the residuals r_i for the gamma hurdle (left) and the generalized gamma hurdle (right) models. This figure appears in color in the electronic version.

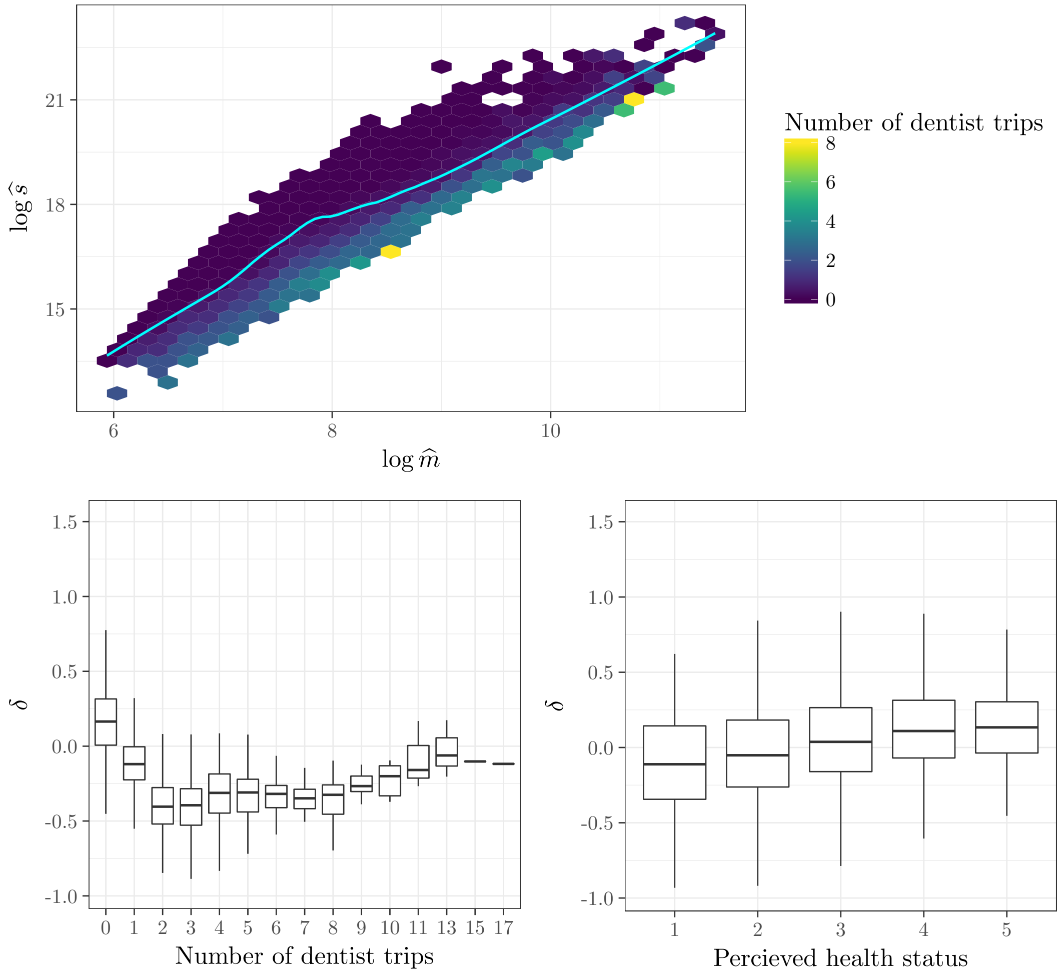

Figure 5. Top: Plot of $\widehat{s}^2(\boldsymbol{X}_i)$ against $\widehat{m}(\boldsymbol{X}_i)$ on the log-log scale; individual points are binned into hexagons, which are shaded according to the number of dentist visits the subject has. Bottom: boxplots of δ for the number of dentist trips and perceived health status. This figure appears in color in the electronic version.

	Shared	Not Shared
Regression	-15166.9	-15267.2
Binary	-1552.8	-2069.7
Total	-16719.7	-17336.9

Table 1

LPML of the model when the forests are shared across the mean, variance, and hurdle components, compared with the LPML when the forests are not shared. The row "Regression" gives the LPML contribution obtained from $[Y_i \mid Y_i > 0]$, while the row "Binary" give the LPML contribution obtained from $I(Y_i > 0)$; "Total" gives the final LPML.