Supporting Information for Semiparametric Mixed-Scale Models Using Shared Bayesian Forests

Antonio R. Linero
Department of Statistics
Florida State University
arlinero@stat.fsu.edu

Debajyoti Sinha
Department of Statistics
Florida State University
sinhad@stat.fsu.edu

Stuart R. Lipsitz
Department of Medicine
Brigham and Women's Hospital
slipsitz@partners.org

S.1 MEPS predictors

The following predictors were used in the analysis of the MEPS dataset:

- `age`: the age of the individual as of 2015.
- `smoke`: whether the individual currently smokes.
- `race`: whether the individual is black, white, Hispanic, Asian, or other.
- `insurance`: whether an individual has private, public, or no health insurance.
- `phealth`: the individuals perceived health (1 to 5).
- `income:` the individual's familial income.
- `meds:` the number of prescription medications the individual is taking.
- `bmi:` the body mass index.
- `education:` the level of completed education.
- `diabetes, stroke, cancer, heart_attack, cognitive_limitations, arthritis`: indicators for whether the subject has suffered from any of these conditions.
- `down`: whether an individual feels down/depressed/hopeless.
- `dentist:` the number of dentist visits over the survey period.

S.2 MCMC for the log-normal and gamma hurdle models

We provide details for implementing Markov chain Monte Carlo algorithms for the gamma hurdle and log-normal hurdle models. We consider a Markov chain operating on the $\mathcal{T}_t$'s, $\mathcal{M}_t$'s, and non-tree-specific parameters $\boldsymbol{\omega}$. Both approaches use the same basic approach, which is summarized by Algorithm 1.

Algorithm 1 Bayesian backfitting algorithm

for $t = 1, \dots, T$ **do**
  Propose $\mathcal{T}_t' \sim Q(\mathcal{T}_t \to \mathcal{T}_t')$
  Sample $U \sim \text{Uniform}(0, 1)$ and compute the acceptance ratio

$$\rho(\mathcal{T}_t, \mathcal{T}_t') = \frac{\Lambda(\mathcal{T}_t')Q(\mathcal{T}_t' \to \mathcal{T}_t)}{\Lambda(\mathcal{T}_t)Q(\mathcal{T}_t \to \mathcal{T}_t')}.$$

  If $U \le \rho(\mathcal{T}_t, \mathcal{T}_t')$, set $\mathcal{T}_t \leftarrow \mathcal{T}_t'$, otherwise leave $\mathcal{T}_t$ unchanged.
  Update the leaf node parameters according to (S.1), (7) and/or (8).
end for
Make an update to $\boldsymbol{\omega}$ which leaves its full conditional invariant.
for $i = 1, \dots, n$ **do**
  Sample $Z_i \sim \text{Normal}(\theta_0 + h_\theta(\boldsymbol{x}), 1)$ truncated to $(-\infty, 0)$ or $(0, \infty)$ according as $Y_i = 0$ or $Y_i > 0$.
end for

To compute $\Lambda(\mathcal{T}_t)$, Algorithm 1 requires the expression

$$L_\theta(t, \ell) = (\sqrt{2\pi})^{-n_\ell} \sqrt{\frac{\sigma_\theta^{-2}}{\sigma_\theta^{-2} + n_\ell}} \exp\left(-\frac{\text{SSE}_\ell}{2} - \frac{n_\ell \sigma_\theta^{-2} \bar{R}_\ell^2}{2(n_\ell + \sigma_\theta^{-2})} \right),$$

where

$$n_\ell = |\{i : \boldsymbol{X}_i \rightsquigarrow (t, \ell)\}|, \qquad \text{SSE}_\ell = \sum_{i : \boldsymbol{X}_i \rightsquigarrow (t,\ell)} (R_i - \bar{R}_\ell)^2,$$
$$\bar{R}_\ell = \frac{1}{n_\ell} \sum_{i : \boldsymbol{X}_i \rightsquigarrow (t,\ell)} R_i, \qquad R_i = Z_i - \theta_0 - \sum_{j \ne t} g(\boldsymbol{X}_i; \mathcal{T}_t, \mathcal{M}_{\theta,t}),$$

which is derived (e.g.) in Kapelner and Bleich (2016). Further, we require the full conditional

$$\theta_{t\ell} \sim \text{Normal}\left(\frac{n_\ell \bar{R}_\ell}{n_\ell + \sigma_\theta^{-2}}, \frac{1}{n_\ell + \sigma_\theta^{-2}} \right). \tag{S.1}$$

For the gamma-hurdle model, $\boldsymbol{\omega}$ consists of just the parameter α, which we give the prior $\alpha^{-1/2} \sim \text{Cauchy}_+(0, A)$. Under this prior, the full conditional for α is proportional to

$$\frac{\alpha^{\alpha N}}{\Gamma(\alpha)^N} \left(\prod_{i=1}^n Y_i^\alpha\right) \exp\left[\alpha \sum_i \{\lambda_0 + h_\lambda(\boldsymbol{x})\} - \alpha \sum_i Y_i e^{\lambda_0 + h_\lambda(\boldsymbol{x})}\right] \times \frac{1}{\pi(A\alpha^{3/2} + \alpha^{1/2}/A)},$$

and we update α using slice sampling (Neal, 2003).

For the log-normal hurdle model, $\boldsymbol{\omega}$ consists of the baseline standard deviation $\sigma_0 = \exp(-\lambda_0/2)$. Let $\nu_i = \sigma^2(\boldsymbol{X}_i)/\sigma_0^2$. Then the full conditional of σ_0 is proportional to

$$\sigma_0^{-N/2} \exp\left[-\frac{1}{\sigma_0^2}\sum_{i=1}^n \nu_i\{Y_i - \mu(\boldsymbol{X}_i)\}^2\right] \frac{1}{\pi(1+\sigma_0^2)},$$

where $N = |\{i : Y_i > 0\}$. As before, σ_0 can be updated via slice sampling.

S.3 Metropolis-Hastings Algorithm

We construct Metropolis-Hastings the proposal from updating $\mathcal{T}_t$ using the general strategies outlined, for example, Kapelner and Bleich (2016), Chipman et al. (1998), and Pratola (2016). As outlined in the main article, we assume that marginal likelihood

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

can be computed in closed form. Given a transition kernel $q(\mathcal{T}_t \to \mathcal{T}_t')$ for updating $\mathcal{T}_t$, the Metropolis-Hastings acceptance probability is given by

$$A(\mathcal{T}_t \to \mathcal{T}_t') = \frac{\Lambda(\mathcal{T}_t')q(\mathcal{T}_t' \to \mathcal{T}_t)}{\Lambda(\mathcal{T}_t)q(\mathcal{T}_t \to \mathcal{T}_t')}.$$

Our choice of $q(\cdot)$ is a mixture of three possible moves: a `Birth` proposal, a `Death` proposal, and a `Change` proposal. The `Birth` step consists of the following steps.

1. Select a leaf node ℓ to become a branch.

2. Select a predictor j to construct the split, according to s.

3. Sample $C_b \sim \text{Uniform}(L_j, U_j)$, with (L_j, U_j) defined as in Section 2.1.

By a retrospective-sampling argument, a valid transition probability associated to this move is given by

$$q(\mathcal{T}_t \to \mathcal{T}_t') = \frac{p_{\texttt{Birth}}(\mathcal{T}_t)}{L_t}$$

where $p_{\texttt{Birth}}(\mathcal{T}_t)$ is the user-specified probability of proposing a `Birth` move (which may depend on the tree structure, as `Death` moves are not possible from the root).

The inverse of the `Birth` transition is a `Death` transition. This requires selecting the node ℓ, which is a branch node in $\mathcal{T}_t'$. This occurs with probability

$$q(\mathcal{T}_t' \to \mathcal{T}_t) = \frac{p_{\texttt{Death}}(\mathcal{T}_t')}{B+1}$$

where B is the number of branches which are not grandparents (i.e., both children are leaves).

The `Death` transition involves the following steps.

1. Select a branch node b, which is not a grandparent.

2. Delete the two child nodes of b, making it a leaf.

We again have the following forward/backward transition probabilities

$$q(\mathcal{T}_t \to \mathcal{T}_t') = \frac{p_{\texttt{Death}}(\mathcal{T}_t)}{B} \qquad \text{and} \qquad q(\mathcal{T}_t' \to \mathcal{T}_t) = \frac{p_{\texttt{Death}}(\mathcal{T}_t)}{L_t - 1}.$$

Finally, the `Change` transition is carried out as follows:

1. Select a branch node b which is not a grandparent.

2. Select a new predictor j according to s.

3. Sample a new cut point $C_b \sim \text{Uniform}(L_j, U_j)$.

As noted by Kapelner and Bleich (2016), the transition probability simplifies substantially in this case as

$$A(\mathcal{T}_t \to \mathcal{T}_t') = \frac{\Lambda(\mathcal{T}_t')}{\Lambda(\mathcal{T}_t)} \wedge 1.$$

S.4 Proof of Lemma 1

Proof. To show the mapping is surjective, for any $F \in \mathscr{M}$, we can take $\pi(\boldsymbol{x}) = F_{\boldsymbol{x}}(\{0\})$ and $G_{\boldsymbol{x}}(A) = F_{\boldsymbol{x}}(A \cap \{0\}^c)/[1 - \pi(\boldsymbol{x})]$ when $F_{\boldsymbol{x}}$ has an atom at 0, and take $\pi(\boldsymbol{x}) = 0$ and $G_{\boldsymbol{x}} = F_{\boldsymbol{x}}$ otherwise.

To show the mapping is injective, consider any $(\pi, G) \neq (\pi', G')$ in $\mathscr{P} \times \mathscr{G}$, and define $F_{\boldsymbol{x}} = \pi(\boldsymbol{x})\delta_0 + [1 - \pi(\boldsymbol{x})]G_{\boldsymbol{x}}$. Define $F'_{\boldsymbol{x}}$ similarly. First, if $\pi \neq \pi'$ then there exists an $\boldsymbol{x}$ such that $F_{\boldsymbol{x}}(\{0\}) \neq F'_{\boldsymbol{x}}(\{0\})$. Conversely, suppose $\pi = \pi'$ but $G \neq G'$. Then there exists a set A and a $\boldsymbol{x}$ such that $G_{\boldsymbol{x}}(A) \neq G'_{\boldsymbol{x}}(A)$; because G and G' do not have atoms at 0, we can assume without loss of generality that $0 \notin A$. But then, noting that $1 - \pi(\boldsymbol{x}) \neq 0$, $F_{\boldsymbol{x}}(A) = [1 - \pi(\boldsymbol{x})]G_{\boldsymbol{x}}(A) \neq [1 - \pi(\boldsymbol{x})]G'_{\boldsymbol{x}}(A) = F'_{\boldsymbol{x}}(A)$, so $F \neq F'$. $\square$

References

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

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