Supplementary Material for Incorporating Grouping Information into Bayesian Decision Tree Ensembles

A Gibbs Sampler

Our Gibbs sampler operates on the state space $\{(\mathcal{T}_t, \mathcal{M}_t)_{t=1}^T, W, \pi, \sigma, \alpha, \psi\}$. Metropolis-Hastings updates for $\{(\mathcal{T}_t, \mathcal{M}_t)_{t=1}^T, \sigma\}$ are now standard; details can be found in Kapelner and Bleich (2016).

We update (W, π) using a data-augmentation strategy similar to LDA. Associate to each branch b in the ensemble a latent group indicator Z_b such that $Z_b \sim \text{Categorical}(\pi)$ and the coordinate j in the decision rule $[x_j \le C_b]$ is distributed as $\text{Categorical}(\omega_{g1}, \ldots, \omega_{gP})$ conditional on $Z_b = g$. The full conditional of Z_b is given by $\Pr(Z_b = g \mid -) \propto \pi_g \omega_{gj(b)}$ where $j(b)$ denotes the coordinate currently used to split b. By conjugacy of the Dirichlet distribution to multinomial sampling, we also have the full conditionals $\pi_g \sim \text{Dirichlet}(\alpha\lambda_1 + \sum_b I(Z_b = 1), \ldots, \alpha\lambda_G + \sum_b I(Z_b = G))$ and

$$\omega_g \sim \text{Dirichlet}\left(\psi q_{g1} + \sum_b I(Z_b = g \text{ and } j(b) = 1), \ldots, \psi q_{gP} + \sum_b I(Z_b = g \text{ and } j(b) = P)\right).$$

This leads to the Gibbs sampler given in Algorithm 1

B Proof of Proposition 3.1

First, it is easy to check that $E(s_j) = O(G^{-1})$ so that $\text{Cov}(s_j, s_k) = E(s_j s_k) + O(G^{-2})$. Similarly, we have $\text{Var}(s_j) = E(s_j^2) + O(G^{-2})$. Next, let (A, B) denote the variables selected by the first two branches of the ensembles, and let (C, D) denote the groups selected by the first two branches in the ensemble. Throughout, we will compute probabilities for (A, B, C, D) using the Pólya Urn scheme for a finite-dimensional Dirichlet distribution (Blackwell and MacQueen, 1973; Neal, 2000). First,

$$E(s_j s_k) = \Pr(A = j, B = k) = \sum_{g,h} \Pr(A = j, B = k \mid C = g, D = h)\ \Pr(C = g, D = h).$$

Algorithm 1 One iteration of a Gibbs sampling algorithm which targets the posterior distribution.

```
Update $\{(\mathcal{T}_t, \mathcal{M}_t)_{t=1}^{T}, \sigma\}$ as described by Kapelner and Bleich (2016).
For each branch $b$ in $\{\mathcal{T}_t\}_{t=1}^{T}$, sample $Z_b = g$ with probability
    $\Pr(Z_b = g \mid -) = \frac{\pi_g \, \omega_{gj(b)}}{\sum_{k=1}^{G} \pi_k \, \omega_{kj(b)}}.$
Sample $\pi \sim \text{Dirichlet}(\alpha\lambda_1 + \sum_b I(Z_b = 1), \ldots, \alpha\lambda_G + \sum_b I(Z_b = G))$.
For $g = 1, \ldots, G$, sample
    $\omega_g \sim \text{Dirichlet}\left(\psi q_{g1} + \sum_b I(Z_b = g \text{ and } j(b) = 1), \ldots, \psi q_{gP} + \sum_b I(Z_b = g \text{ and } j(b) = P)\right).$
```

When $g \neq h$ in the above term the Pölya Urn scheme gives $\alpha q_{gj} q_{hk} \lambda_g \lambda_h / (\alpha + 1) = O(G^{-2})$. Because each predictor lies in finitely many groups as $G \to \infty$, summing over $g \neq h$ contributes only an $O(G^{-2})$ term to $E(s_j s_k)$. For $g = h$ we have the terms

$$\frac{\psi q_{gj} q_{gk}}{\psi + 1} \cdot \frac{\lambda_g(\alpha\lambda_g + 1)}{\alpha + 1} = \frac{\psi}{(\alpha + 1)(\psi + 1)} q_{gj} q_{gk} \lambda_g + O(G^{-2}).$$

Hence we have

$$\text{Cov}(s_j, s_k) = \frac{\psi}{G(\psi + 1)(\alpha + 1)} \sum_g q_{gj} q_{gk} \widetilde{\lambda}_g + O(G^{-2}) = \frac{\psi q_j^\top \Lambda q_k}{G(\psi + 1)(\alpha + 1)} + O(G^{-2}).$$

We perform a similar analysis for the second moment. We have

$$E(s_j^2) = \sum_{g,h} \Pr(A = j, B = j \mid C = g, D = h) \, \Pr(C = g, D = h).$$

As before, the terms with $g \neq h$ can be shown to contribute $O(G^{-2})$ to the summation. Again applying the Pölya Urn scheme, the $g = h$ terms are given by

$$\frac{q_{gj}(\psi q_{gj} + 1)}{(\psi + 1)} \cdot \frac{\lambda_g(\alpha\lambda_g + 1)}{\alpha + 1} = \frac{\lambda_g}{(\psi + 1)(\alpha + 1)} [\psi q_{gj}^2 + q_{gj}] + O(G^{-2}).$$

Summing over g, and again noting that only a constant number of summation terms are non-zero as G grows, gives

$$\text{Var}(s_j) = \frac{\psi q_j^\top \Lambda q_j + q_j^\top \Lambda \mathbf{1}}{G(\psi + 1)(\alpha + 1)} + O(G^{-2}).$$

Method	Hyperparameters	Deviance
OG-BART	$(1,1)$	620
	$(1,10)$	614
	$(10,1)$	591
	$(10,10)$	590
SBART	$(1,-)$	646
	$(10,-)$	609

Table 1: Results for different hyperparameter settings. For OG-BART, the setting (a,b) denotes that $\alpha \sim \text{Exp}(a)$ and $\psi = b$ For SBART, the setting $(a,-)$ denotes that $\alpha \sim \text{Exp}(a)$; there is no analogous quantity ψ.

Hence the correlation is

$$\text{Cor}(s_j, s_k) = \frac{\psi q_j^\top \Lambda q_k + O(G^{-1})}{\sqrt{\psi q_j^\top \Lambda q_j + q_j^\top \Lambda \mathbf{1} + O(G^{-1})}\sqrt{\psi q_k^\top \Lambda q_k + q_k^\top \Lambda \mathbf{1} + O(G^{-1})}}.$$

Letting $G \to \infty$ establishes the result.

C Breast Cancer Results for Other Hyperparameter Values

We consider the breast cancer dataset for several additional priors for α and ψ. For comparison, we also consider the SBART model with different choices of hyperparameter α. We replicate the cross-validation experiment for the breast cancer dataset to assess predictive performance. Results are given in Table 1. We see that generally higher values of the hyperparameters result in better predictive performance. Better predictive performance is also obtained for SBART with higher values of α, however its performance is still well below a similar OG-BART model. We prefer the OG-BART model with $\alpha \sim \text{Exp}(1)$ and $\psi = 1$ for analysis, as the models it produces are sparser, leading to simpler model interpretation.

D Misspecified Groups and False Positives

In the simulation study, we observed a curious behavior: when the grouping structure is misspecified and we have bi-level sparsity, the number of false positives *decreased*. We now illustrate that this behavior is to be expected. Consider the normal means problem

$$\begin{aligned} Y_i &\sim \text{Normal}(\theta_i, 1), \\ \theta_i &\sim \text{Bernoulli}(\pi_{g_i}), \\ \pi_g &\sim \text{Uniform}(0,1), \end{aligned}$$

π_1	Correct	Incorrect
0.3	(30, 54)	(15, 17)
0.8	(80, 388)	(117, 217)

Table 2: Results of normal means simulation. "Correct" denotes that the correct grouping structure was used; "Incorrect" denotes that an incorrect grouping structure was used. The result (30, 54), for example, indicates that there were an average of 30 false positives and 54 true positives.

where $g = (g_1, \dots, g_n)$ records the group that θ_i is in. Our goal is to detect if $\theta_i = 0$ (negative) or $\theta_i = 1$ (positive). We consider $n = 1000$ and simulate from this model with two equal-sized groups, the second of which is pure noise ($\pi_2 \equiv 0$). We consider a sparse-within-group setting with $\pi_1 = 0.3$ and a dense-within-group setting with $\pi_1 = 0.8$. We classify θ_i as a positive if the posterior probability of $\theta_i = 1$ exceeds 0.5. We replicate this experiment 20 times for each setting. To obtain a misspecified structure, we randomly assign each i to one of the groups. We then fit the model using the JAGS software package.

Results are given in Table 2. We again see the behavior of an incorrect group structure leading to fewer false positives. Intuitively, this occurs because this model will estimate $\pi_1 \approx \pi_2 \approx 0.15$; so that all θ_i's are equally penalized, and this is sufficient to control false positives. On the other hand, a correctly specified model will estimate $\pi_1 \approx 0.3$ and $\pi_2 \approx 0$; this applies a smaller penalty to the noise θ_i's in group 1, allowing them to enter into the model more easily.

The upshot of using a correct grouping structure is that we obtain far more true positives, leading to a higher F_1 score. This occurs regardless of the value of π_1.

References

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355.

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. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.