Supplementary Material to Gibbs Priors for Bayesian Nonparametric Variable Selection with Weak Learners

S.1 Bayesian Backfitting for Random Basis Function Expansions

Consider the model

$$Y_i = \sum_{m=1}^{M} \beta_m \, b(X_i; \gamma_m) + \epsilon_i, \qquad \epsilon_i \sim \text{Normal}(0, \sigma^2),$$

with $\beta_m \sim \text{Normal}(0, \sigma_\beta^2)$. To update the parameters of a single function (β_m, γ_m) we consider the residual $R_i \sim \text{Normal}\{\beta_m \, b(X_i; \gamma_m), \sigma^2\}$. Routine calculations show that the marginal likelihood of $\boldsymbol{R}$ given γ_m (integrating out β_m) is given by

$$\Lambda(\gamma_m) = (2\pi\sigma^2)^{-N/2} \left(1 + \frac{\sigma_\beta^2}{\sigma^2} \boldsymbol{b}^\top \boldsymbol{b}\right)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\boldsymbol{R}^\top \boldsymbol{R} - \frac{(\boldsymbol{R}^\top \boldsymbol{b})^2 \sigma_\beta^2}{\sigma^2 + \sigma_\beta^2 \, \boldsymbol{b}^\top \boldsymbol{b}} \right] \right\},$$

where $\boldsymbol{b} = (b(X_1; \gamma_m), \ldots, b(X_N; \gamma_m))^\top$. We can then use $\Lambda(\gamma)$ to construct a Metropolis-Hastings algorithm with β_m integrated out. To update β_m, we simply notice that the full conditional of β_m is given by

$$\beta_m \sim \text{Normal}\left(\frac{\sigma^{-2} \, \boldsymbol{b}^\top \boldsymbol{R}}{\sigma_\beta^{-2} + \sigma^{-2} \, \boldsymbol{b}^\top \boldsymbol{b}}, \frac{1}{\sigma_\beta^{-2} + \sigma^{-2} \, \boldsymbol{b}^\top \boldsymbol{b}} \right).$$

Using these results we (i) update γ_m using a Metropolis-Hastings algorithm and (ii) sample β_m form its full-conditional, where any changes to the x_j's used in γ_m are sampled according

to Proposition 2. For the multivariate adaptive regression splines (MARS) basis we use the following proposals to modify γ_m.

`BIRTH` Select a covariate j according to Proposition 2 and sample a new basis function $\max(0, x_j - C)$ by sampling $C \sim \text{Uniform}(0, 1)$; add this to the basis function by taking $b(x; \gamma_m) \leftarrow b(x; \gamma_m) \max(0, x_j - C)$.

`DEATH` Randomly select one of the basis components $\max(0, x_j - C)$ and remove it from $b(x; \gamma_m)$.

`SWAP` Randomly select one of the basis components $\max(0, x_j - C)$ and sample a new predictor j' according to Proposition 2 and $C' \sim \text{Uniform}(0, 1)$. Then swap $\max(0, x_j - C)$ with $\max(0, x_{j'} - C')$ in $b(x; \gamma_m)$.

As with BART, the `BIRTH` and `DEATH` moves are inverses of one another, while the `SWAP` move is its own inverse. The acceptance probabilities for these moves can be derived along similar lines as Proposition 4.

S.2 Proof of Proposition 3

This follows by induction from the proof of Proposition 2 and the fact that Gibbs priors correspond to exchangeable random partition processes.

S.3 Proof of Proposition 4

We compute the Metropolis-Hastings ratio as the product of (i) the tree construction ratio $\pi_{\mathcal{T}}(\mathcal{T}')/\pi_{\mathcal{T}}(\mathcal{T}_t)$, (ii) the likelihood ratio $\Lambda(\mathcal{T}')/\Lambda(\mathcal{T}_t)$, (iii) and the transition ratio $q(\mathcal{T}_t \mid \mathcal{T}')/q(\mathcal{T}' \mid \mathcal{T}_t)$.

We begin with the `BIRTH` step. First, the probability $q(\mathcal{T}_t \mid \mathcal{T}')$ is given by the probability of choosing a `DEATH` move associated to the node ℓ proposed from the `BIRTH` step. This is

given by $\frac{q_{\texttt{DEATH}}(\mathcal{T}')}{|\,\mathrm{NOG}(\mathcal{T}')|}$. The probability $q(\mathcal{T}' \mid \mathcal{T}_t)$ is given by the probability of (i) choosing the leaf node ℓ, (ii) choosing the splitting coordinate j_ℓ according to (5), and (iii) sampling $C_\ell \sim \mathrm{Uniform}(A_\ell, B_\ell)$. The probability density of this move is given by $\frac{q_{\texttt{BIRTH}}(\mathcal{T}_t)\,\psi_{j_\ell}(\mathcal{T}_t)}{(B_\ell - A_\ell)\,|\mathcal{L}(\mathcal{T}_t)|}$, where $\psi_{j_\ell}(\mathcal{T}_t)$ is given by (5). Putting these together we get

$$\frac{q_{\texttt{DEATH}}(\mathcal{T}')\,(B_\ell - A_\ell)\,|\mathcal{L}(\mathcal{T}_t)|}{|\,\mathrm{NOG}(\mathcal{T}')|q_{\texttt{BIRTH}}(\mathcal{T}_t)\,\psi_j(\mathcal{T}_t)}.$$

Next, the prior ratio is given by $\pi(\mathcal{T}')/\pi(\mathcal{T}_t)$. Note that the probability of $\mathcal{T}'$ is the same as the probability of $\mathcal{T}_t$ except that the leaf node ℓ is instead chosen to be a branch, its two children are both made leaves, j_ℓ is chosen according to (5), and $C_\ell \sim \mathrm{Uniform}(A_\ell, B_\ell)$. Hence when we compute the ratio we get

$$\frac{\pi(\mathcal{T}')}{\pi(\mathcal{T})} = \frac{\rho(d)\{1-\rho(d+1)\}^2\psi_{j_\ell}(\mathcal{T}_t)}{(B_\ell - A_\ell)\{1-\rho(d)\}}.$$

We now observe that the term $\psi_{j_\ell}(\mathcal{T}_t)/(B_\ell - A_\ell)$ cancels when the prior ratio and transition ratio are multiplied together. Hence multiplying the three ratios together gives the result for $R_{\texttt{BIRTH}}$.

The argument for $R_{\texttt{DEATH}}$ is the same as the argument for $R_{\texttt{BIRTH}}$ with the roles of $\mathcal{T}'$ and $\mathcal{T}_t$ switched. Finally, for $R_{\texttt{PRIOR}}$ we have $q(\mathcal{T}' \mid \mathcal{T}_t) = \pi(\mathcal{T}')$ and $q(\mathcal{T}_t \mid \mathcal{T}') = \pi(\mathcal{T}_t)$, so the prior and transition ratios cancel.

S.4 Traceplots for Real Data Examples

In Figure S.1 and Figure S.2 we provide the traceplots for the MCMC schemes of the DART and Gibbs priors when fit to the Hitters and WIPP datasets respectively. We monitor (i) the error standard deviation σ, (ii) the log-likelihood of the data "Loglik", and (iii) the size of the model, given by the number of predictors used on a particular iteration. For both datasets we ran four parallel chains for 10,000 warmup iterations and 10,000 sampling iterations with

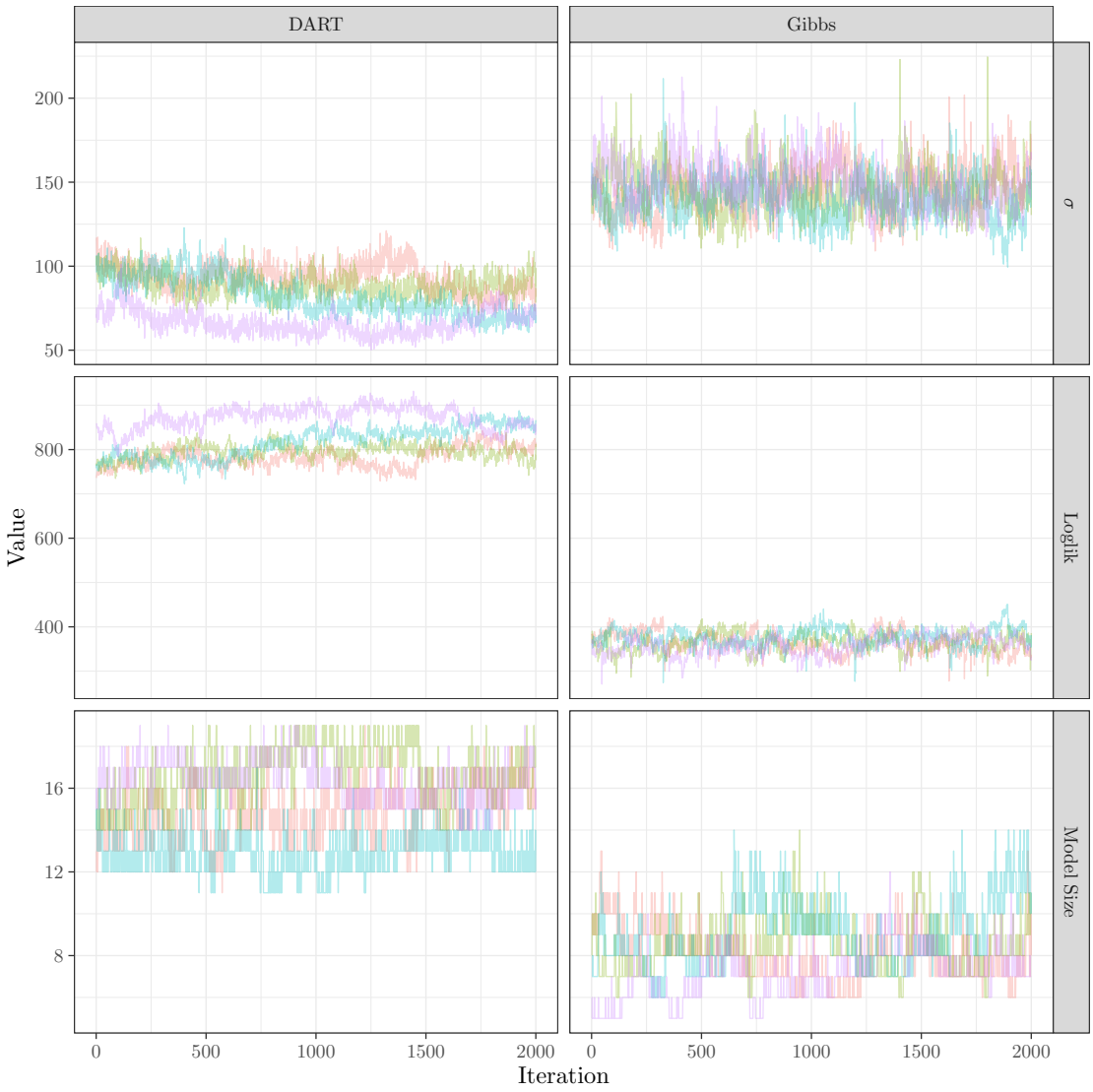

Figure S.1: Traceplots for the Hitters dataset.

a thinning interval of 5 (so 2,000 samples saved in total).

We see that, generally speaking, the Gibbs prior tends to mix better and, surprisingly, this carries over to the mixing of both σ and "Loglik." Interestingly, the data suggests that the DART prior may also be overfitting substantially - the training data log-likelihood is substantially larger and the error is substantially smaller than the corresponding quantities for the Gibbs prior. That this is due to overfitting is suggested by the fact that the Gibbs prior outperforms DART once cross-validation is applied.

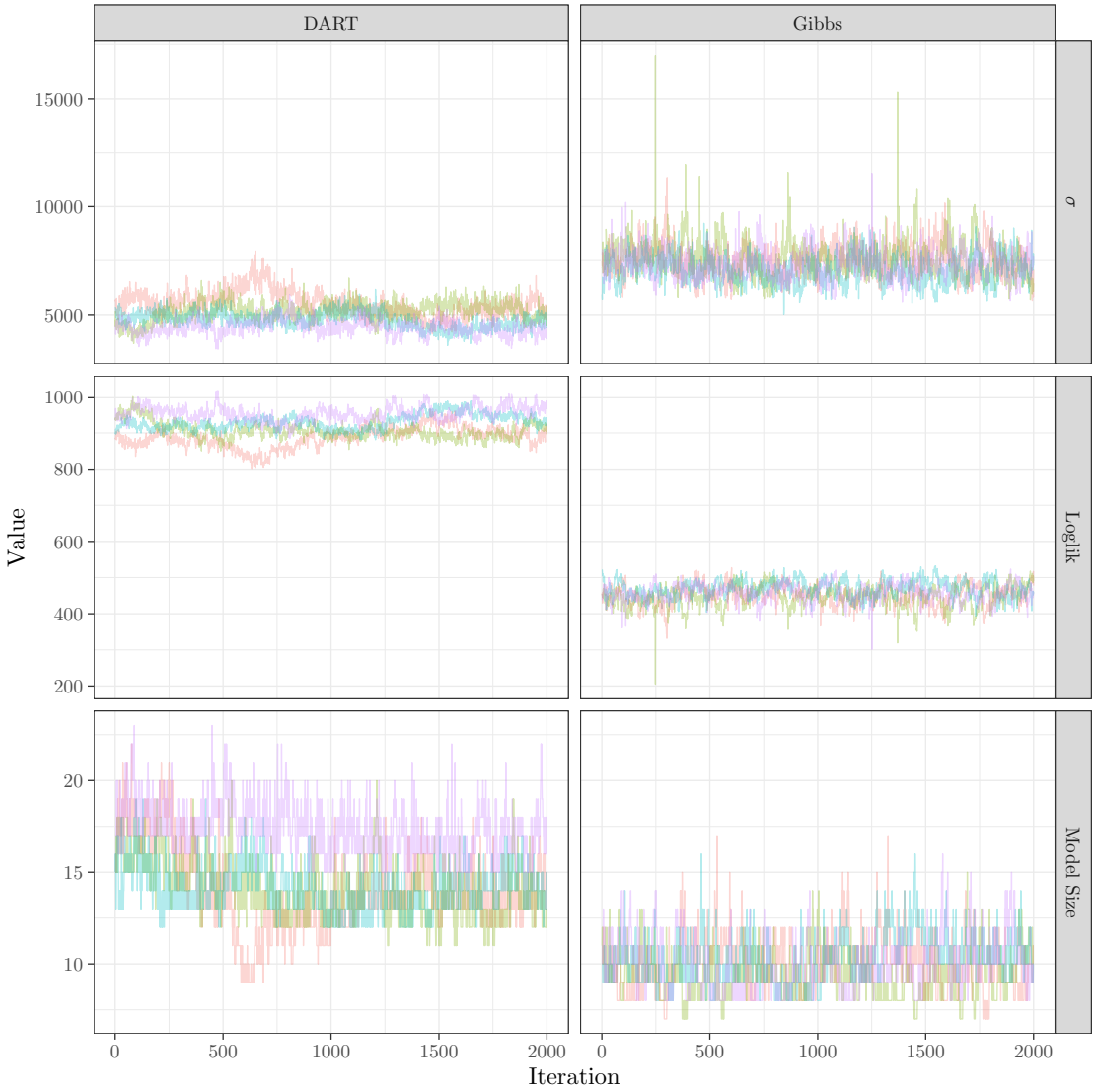

Figure S.2: Traceplots for the WIPP dataset.