

Additive Main Effect Bayesian Additive Regression Tree interaction models (AMBARTI)

Estevão Batista do Prado

08 January 2021

1 AMBARTI

$$Y_{ij}|\mathbf{x}_{ij}, \mathcal{T}, \mathcal{M}, \Theta, \sigma^2 \sim \text{N} \left(g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t), \sigma^2 \right),$$

where y_{ij} is the yield for genotype i and environment j , and g_i and e_j are the genotype and environment effects, respectively. h is then the individual BART trees with $h(x_{ij}, \mathcal{M}_t, \mathcal{T}_t) = \mu_{t\ell}$. g and e are random effects, so there are variance terms σ_g^2 , σ_e^2 , and residual standard deviation σ^2 , which all need to be estimated, and $\Theta = (\mathcal{G}, \mathcal{E}, \sigma_g^2, \sigma_e^2)^\top$, with $\mathcal{G} = (g_1, \dots, g_I)$ and $\mathcal{E} = (e_1, \dots, e_J)$. In addition, consider the prior distribution for $\mu_{t\ell}$, g_i , e_j , σ_g^2 , σ_e^2 , and σ^2 as

$$\begin{aligned} \mu_{t\ell}|\mathcal{T}_t &\stackrel{\text{iid}}{\sim} \text{N}(\mu_\mu = 0, \sigma_\mu^2), \\ g_i|\mathcal{T}_t &\sim \text{N}(\mu_g, \sigma_g^2), \\ e_j|\mathcal{T}_t &\sim \text{N}(\mu_e, \sigma_e^2), \\ \sigma_g^2 &\sim \text{IG}(a_g, b_g), \\ \sigma_e^2 &\sim \text{IG}(a_e, b_e), \\ \sigma^2 &\sim \text{IG}(a, b), \end{aligned}$$

Similarly to the original BART, the conditional posterior distribution of $\mu_{t\ell}$ in the AMBARTI also depends only on the information provided by all the other trees ($\mathcal{T}_{(t)}$) and means ($M_{(t)}$) via partial residual in the form of

$$\begin{aligned} r_{ij} &= y_{ij} - \left(g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t) \right) \\ r_{ij} &\stackrel{\text{iid}}{\sim} \text{N} \left(g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t), \sigma^2 \right). \end{aligned}$$

Below, we list all full conditionals needed.

The joint posterior distribution of the trees and predicted values is given by

$$\begin{aligned} p(\mathcal{T}, \mathcal{M}, \mathcal{G}, \mathcal{E}, \sigma_g^2, \sigma_e^2, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto p(\mathbf{y} | \mathbf{X}, \mathcal{T}, \mathcal{M}, \mathcal{G}, \mathcal{E}, \sigma_g^2, \sigma_e^2, \sigma^2) p(\mathcal{G}) p(\mathcal{E}) p(\mathcal{M} | \mathcal{T}) p(\mathcal{T}) p(\sigma_g^2) p(\sigma_e^2) p(\sigma^2), \\ p(\mathcal{T}, \mathcal{M}, \Theta, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \left[\prod_{t=1}^m \prod_{\ell=1}^{b_t} \prod_{i: \mathbf{x}_{ij} \in \mathcal{P}_{t\ell}} \prod_{j: \mathbf{x}_{ij} \in \mathcal{P}_{t\ell}} p(y_{ij} | \mathbf{x}_{ij}, \mathcal{T}_t, \mathcal{M}_t, \Theta, \sigma^2) \right] \left[\prod_i p(g_i) \right] \left[\prod_j p(e_j) \right] \times \\ &\times \left[\prod_{t=1}^m \prod_{\ell=1}^{b_t} p(\mu_{t\ell} | \mathcal{T}_t) p(\mathcal{T}_t) \right] p(\sigma_g^2) p(\sigma_e^2) p(\sigma^2). \end{aligned}$$

Let R_t denote the vector of the partial residuals, and $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{e} \in \mathbb{R}^n$ are vectors containing the random effects g_i and e_j for all observations.

$$\begin{aligned} R_t &= \mathbf{y} - \left(\mathbf{g} + \mathbf{e} + \sum_{k \neq t}^T h(\mathbf{X}; \mathcal{T}_k, \mathcal{M}_k) \right), \\ r_{ij}^{(t)} &\stackrel{\text{iid}}{\sim} \text{N} (g_i + e_j + h(\mathbf{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t), \sigma^2). \end{aligned}$$

As all $\mu_{t\ell}$ are independent from each other, it is possible to write $p(\mathcal{M}_t|\mathcal{T}_t, R_t, \sigma^2) = \prod_{\ell=1}^{b_t} p(\mu_{t\ell}|\mathcal{T}_t, R_t, \sigma^2)$. Hence,

$$p(\mu_{t\ell}|\mathcal{T}_t, R_t, \sigma^2) \propto p(R_t|\mathcal{M}_t, \mathcal{T}_t, \sigma^2)p(\mu_{t\ell}),$$

$$\propto \exp\left(-\frac{1}{2\sigma_*^2}(\mu_{t\ell} - \mu_{t\ell}^*)^2\right),$$

which is a

$$N\left(\frac{\sigma^{-2} \sum_{(i,j) \in \mathcal{P}_{t\ell}} r_{ij}^{(t)}}{n_{t\ell}/\sigma^2 + \sigma_\mu^{-2}}, \frac{1}{n_{t\ell}/\sigma^2 + \sigma_\mu^{-2}}\right), \quad (1)$$

where $r_{ij}^{(t)} = y_{ij} - (g_i + e_j + \sum_{k \neq t}^T h(\mathbf{x}_{ij}; \mathcal{T}_k, \mathcal{M}_k))$.

$$p(g_i|\mathbf{y}, \sigma_g^2, \sigma^2) \propto p(\mathbf{y}|g_i, \sigma_g^2, \sigma^2)p(g_i),$$

$$\propto \exp\left(-\frac{1}{2\sigma_{g*}^2}(g_i - g_i^*)^2\right),$$

which is a

$$N\left(\frac{\sum_j [y_{ij} - e_j - \hat{\mu}_{ij}]/\sigma^2 + \mu_g/\sigma_g^2}{n_{g_i}/\sigma^2 + 1/\sigma_g^2}, \frac{1}{n_{g_i}/\sigma^2 + 1/\sigma_g^2}\right), \quad (2)$$

where $\hat{\mu}_{ij} = \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{T}_t, \mathcal{M}_t)$, $r_{ij} \in R$, and n_{g_i} is the number of observations that belong to g_i .

$$p(e_j|\mathbf{y}, \sigma_e^2, \sigma^2) \propto p(\mathbf{y}|e_j, \sigma_e^2, \sigma^2)p(e_j),$$

$$\propto \exp\left(-\frac{1}{2\sigma_{e*}^2}(e_j - e_j^*)^2\right),$$

which is a

$$N\left(\frac{\sum_i [y_{ij} - g_i - \hat{\mu}_{ij}]/\sigma^2 + \mu_e/\sigma_e^2}{n_{e_j}/\sigma^2 + 1/\sigma_e^2}, \frac{1}{n_{e_j}/\sigma^2 + 1/\sigma_e^2}\right). \quad (3)$$

$$p(\sigma_g^2|\mathcal{G}) \propto p(\mathcal{G}|\sigma_g^2)p(\sigma_g^2)$$

$$\propto (\sigma_g^2)^{-\left(\frac{XX+a_g}{2}+1\right)} \exp\left(-\frac{YY+2b_g}{2}\right),$$

which is an

$$\text{IG}\left(\frac{|\mathcal{G}|}{2} + a_g, \frac{\sum_{i=1}^I (g_i - \mu_g)^2}{2} + b_g\right). \quad (4)$$

$$p(\sigma_e^2|\mathcal{G}) \propto p(\mathcal{E}|\sigma_e^2)p(\sigma_e^2)$$

$$\propto (\sigma_e^2)^{-\left(\frac{XX+a_e}{2}+1\right)} \exp\left(-\frac{YY+2b_e}{2}\right),$$

which is an

$$\text{IG}\left(\frac{|\mathcal{E}|}{2} + a_e, \frac{\sum_{j=1}^E (e_j - \mu_e)^2}{2} + b_e\right). \quad (5)$$

Then, after generating all predicted values for all trees, σ^2 can be updated based on

$$\begin{aligned} p(\sigma^2|T, M, \mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, T, M, \sigma^2)p(\sigma^2) \\ &\propto (\sigma^2)^{-\left(\frac{n+\nu}{2}+1\right)} \exp\left(-\frac{S+\nu\lambda}{2\sigma^2}\right), \end{aligned} \quad (6)$$

where $S = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $\hat{y}_i = \mathbf{g} + \mathbf{e} + \sum_{t=1}^T h(\mathbf{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t)$. The expression in (6) is an $\text{IG}((n + \nu)/2, (S + \nu\lambda)/2)$, and drawing samples from it is straightforward.

$$\begin{aligned} p(\mathcal{T}_t|R_t, \sigma^2) &\propto p(\mathcal{T}_t) \int p(R_t|\mathcal{M}_t, \mathcal{T}_t, \sigma^2)p(\mathcal{M}_t|\mathcal{T}_t)d\mathcal{M}_t, \\ &\propto p(\mathcal{T}_t)p(R_t|\mathcal{T}_t, \sigma^2), \\ &\propto p(\mathcal{T}_t) \prod_{\ell=1}^{b_t} \left[\left(\frac{\sigma^2}{\sigma_\mu^2 n_{t\ell} + \sigma^2} \right)^{1/2} \exp \left(\frac{\sigma_\mu^2 [n_{t\ell} \bar{R}_\ell]^2}{2\sigma^2(\sigma_\mu^2 n_{t\ell} + \sigma^2)} \right) \right], \end{aligned}$$

where $\bar{R}_\ell = \sum_{(i,j) \in \mathcal{P}_{t\ell}} (r_{ij}^{(t)} - g_i - e_j)/n_{t\ell}$, $r_{ij}^{(t)} \in R_t$ and $n_{t\ell}$ is the number of observations that belong to $\mathcal{P}_{t\ell}$. This sampling is carried out through a Metropolis-Hastings step, as the expression does not have a known distributional form;