MCW Biostatistics Technical Report 63:
BART with logGamma errors using
a convolution mixture of Normals

Purushottam Laud, Rodney Sparapani, Brent Logan and Robert McCulloch
February 2016

Bayesian Additive Regression Trees (BART) [1] allows one to regress a continuous outcome, \( y \), on a vector of covariates, \( x \), via an arbitrarily flexible function, \( f \), i.e. \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \). However, suppose \( \epsilon \) follows some other distribution function, say \( G(\epsilon) \), then we can still employ BART provided that \( g(\epsilon) \) can be reliably approximated by a mixture of Normals, i.e. \( g(\epsilon) \approx \sum_i p_i N(\mu_i, \sigma_i^2) \) where \( \{p_i, \mu_i, \sigma_i\} \) are known. We extend BART to the situation where \( \epsilon \) follows the logGamma distribution with distribution function \( G_\alpha(y) \) and density function \( g_\alpha(y) \).

If \( x \) follows the Gamma distribution, then \( y = \log x \) follows the logGamma distribution, i.e. \( x \sim \text{Gamma}(\alpha, 1) \) where \( \alpha > 0 \) implies \( y \sim \log \text{Gamma}(\alpha, 1) \) and \( g_\alpha(y) = \Gamma(\alpha)^{-1} \exp(-e^{\alpha} - e^{\alpha-y}) \).

Fruhwirth-Schnatter, Fruhwirth, Held, and Rue (FFHR) [2] show how to reliably approximate the logGamma distribution, \( \log \text{Gamma}(\alpha, \beta) \) with a mixture of Normals. However, their method (which we also refer to as FFHR) is not readily applicable to our work since we need \( \alpha < 1 \) while FFHR requires that \( \alpha \geq 1 \). Therefore, we extend FFHR to meet our needs with similar high-degree of accuracy so we can approximate any \( \log \text{Gamma}(\alpha, \beta) \) routinely via \( \log \beta x_\alpha = \log \beta + y_\alpha \) where \( \alpha > 0 \) and \( \beta > 0 \).

FFHR employ the Kullback-Leibler divergence [3] to determine the accuracy of the approximation: \( \delta_{KL}(\theta_\alpha) = \int_{-\infty}^{\infty} g_\alpha(y) \log \frac{g_\alpha(y)}{g_{\hat{\alpha}}(y, \theta_\alpha)} \, dy \) where \( \theta_\alpha = (p_{\alpha}, \mu_{\alpha}, \sigma_{\alpha}) \). Then, they use the Nelder-Mead simplex method [4] to minimize the following objective function: \( \Delta_{KL}(\theta_\alpha) = \delta_{KL}(\theta_\alpha) + 10^9(1 - \sum_{i} p_{\alpha,i})^2 \) subject to the constraints \( 0 < p_{\alpha,i} < 1 \) and \( 0 < \sigma_{\alpha,i} \).

We adopt the following notation: \( y_\alpha \) is a random variable with the same distribution as \( \log x_\alpha \) where \( x_\alpha \sim \text{Gamma}(\alpha, 1) \). Notice that we can decompose \( y_\alpha \) as \( y_\alpha = y_{\alpha+1} + w_\alpha \) where \( w_\alpha \sim \text{Exp}(\alpha) \) and \( y_{\alpha+1} \perp w_\alpha \). This result can be extended to what could be called a “distributional factorial” property of the logGamma distribution: \( y_\alpha = y_{\alpha+n} + \sum_{i=1}^{n} w_{\alpha+i-1} \) where \( w_{\alpha+i-1} \sim \text{Exp}(\alpha + i - 1) \) and \( y_{\alpha+n} \perp w_{\alpha+i-1} \) for \( i = 1, \ldots, n \).

Now, let us return to where our interest resides: \( \alpha < 1 \). We found it difficult to approximate \( \log \text{Gamma} \) for \( \alpha < 1 \) with \( \alpha \geq 1 \) in one step via \( y_\alpha = y_{\alpha+1} + w_\alpha \). However, the approximation with two steps is satisfactory, i.e. substitute \( y_{\alpha+1} = y_{\alpha+2} + w_{\alpha+1} \) into \( y_\alpha = y_{\alpha+1} + w_\alpha \) which yields \( y_\alpha = y_{\alpha+2} + w_{\alpha+1} + w_\alpha \). We
accomplish this by approximating the distribution for each of these 3 terms by their own mixture of Normals the composite of which we call a convolution mixture of Normals.

FFHR provide a high-degree of accuracy for approximation at the integers of \( \alpha \in \{1, 2, 3, 4\} \) with mixtures of 10 Normals, i.e.
\[
y^\text{approx} \sim \sum_{i=1}^{10} p_{y^i} N(\mu_{y^i}, \sigma_{y^i}^2)
\]
(note that the FFHR weights, \( p_{y^i} \), are identical, \( p_{y^i} = p_{y^i'} \), for \( \alpha \neq \alpha' \in \{1, 2, 3, 4\} \), however, this will not be the case for our approximations for \( \alpha \in (2, 3) \)). To create mixtures for non-integer \( \alpha \in (2, 3) \), we estimate a starting point via linear interpolation starting with \( \alpha = 2.5 \) as \( 0.5\hat{\theta}_2 + 0.5\hat{\theta}_3 \). Then we plug this starting point into the subplex algorithm \([5]\) where we minimize \( \Delta_{KL}(\theta_{2.5}) \) arriving at the solution \( \hat{\theta}_{2.5} = \arg\min_{\theta_{2.5}} \Delta_{KL}(\theta_{2.5}) \). The subplex algorithm (a variant of the Nelder-Mead simplex method) is more efficient and robust than simplex while retaining the latter’s facility with discontinuous objectives. Now, we proceed to fill in the grid of 129 points: \( 2 + \frac{1}{128}, \ldots, 2 + \frac{127}{128}, 3 \); i.e. create a linear interpolation starting point \( 0.5\hat{\theta}_2 + 0.5\hat{\theta}_{2.5} \) to plug into the subplex method arriving at \( \theta_{2.25} \), etc. Once you reach this grid level, linear interpolation between the grid points provides sufficiently accurate approximations.

Finally, we approximate the distribution of \( w \sim \text{Exp}(1) \) by a mixture of 20 Normals. Since the Exponential distribution is restricted to positive values and the Normal distribution is not, we can not employ the Kullback-Leibler divergence. Instead, we rely on integrated squared error: \( \delta_{ISE}(\zeta) = \int_0^{\infty} (g_w(y) - \hat{g}_w(y; \zeta))^2 dy \) and use the objective function \( \Delta_{ISE}(\zeta) = \delta_{ISE}(\zeta) + 10^9(1 - \sum_i p_{wi})^2 \). Accuracy of approximations achieved can be seen visually in the plots of densities (and log-densities) of \(-y, y \sim \log \Gamma((, \alpha), 1)\) for various choices of \( \alpha \) in the range of interest on the next several pages.

References


alpha = 0.9

KL = 0.00038 on \( \mu - 3\sigma \) to \( \mu + 7\sigma \)

alpha = 0.8

KL = 0.00038 on \( \mu - 3\sigma \) to \( \mu + 7\sigma \)
\textbf{alpha = 0.7} \\
\includegraphics[width=0.4\textwidth]{alpha_0.7_log_density}

\textit{KL = 0.00037 on } \textit{mu−3sd to mu+7sd}

\textbf{alpha = 0.6} \\
\includegraphics[width=0.4\textwidth]{alpha_0.6_log_density}

\textit{KL = 0.00035 on } \textit{mu−3sd to mu+7sd}
\begin{align*}
\text{alpha} &= 0.5 \\
\text{KL} &= 0.00032 \text{ on } \mu - 3\sigma \text{ to } \mu + 7\sigma \\
\text{alpha} &= 0.4 \\
\text{KL} &= 0.00029 \text{ on } \mu - 3\sigma \text{ to } \mu + 7\sigma
\end{align*}
alpha = 0.3

KL = 0.00023 on mu−3sd to mu+7sd

alpha = 0.2

KL = 0.00018 on mu−3sd to mu+7sd
$\alpha = 0.1$

$KL = 0.00015$ on $\mu-3\sigma$ to $\mu+7\sigma$

$\alpha = 0.09$

$KL = 0.00015$ on $\mu-3\sigma$ to $\mu+7\sigma$
alpha = 0.08

KL = 0.00016 on mu−3sd to mu+7sd

alpha = 0.07

KL = 0.00016 on mu−3sd to mu+7sd
alpha = 0.06

KL = 0.00017 on mu−3sd to mu+7sd

alpha = 0.05

KL = 0.00019 on mu−3sd to mu+7sd
\( \alpha = 0.04 \)

\( X \)

\( \text{KL} = 0.00023 \) on \( \mu - 3 \text{sd} \) to \( \mu + 7 \text{sd} \)

\( \alpha = 0.03 \)

\( X \)

\( \text{KL} = 3 \times 10^{-4} \) on \( \mu - 3 \text{sd} \) to \( \mu + 7 \text{sd} \)
alpha = 0.02

KL = 0.00047 on mu−3sd to mu+7sd

alpha = 0.01

KL = 0.00085 on mu−3sd to mu+7sd
\[ \alpha = 0.009 \]

\[ \text{KL} = 0.00115 \text{ on } \mu \pm 3\sigma \text{ to } \mu \pm 7\sigma \]

\[ \alpha = 0.008 \]

\[ \text{KL} = 0.00164 \text{ on } \mu \pm 3\sigma \text{ to } \mu \pm 7\sigma \]
alpha = 0.007

KL = 0.00226 on mu−3sd to mu+7sd

alpha = 0.007

KL = 0.00226 on mu−3sd to mu+7sd

alpha = 0.006

KL = 0.00283 on mu−3sd to mu+7sd

alpha = 0.006

KL = 0.00283 on mu−3sd to mu+7sd
alpha = 0.005

KL = 0.00337 on mu−3sd to mu+7sd

alpha = 0.004

KL = 0.00409 on mu−3sd to mu+7sd
\begin{align*}
\text{alpha} &= 0.003 \\
\text{KL} &= 0.00516 \text{ on mu−3sd to mu+7sd}
\end{align*}

\begin{align*}
\text{alpha} &= 0.002 \\
\text{KL} &= \text{NaN} \text{ on mu−3sd to mu+7sd}
\end{align*}
alpha = 0.001

KL = NaN on mu−3sd to mu+7sd

log density

density
\[ \alpha = 2.03125 \]

\[ \alpha = 2.0390625 \]
\( \alpha = 2.0625 \)

\( \log \) density

\( \alpha = 2.0703125 \)

\( \log \) density

\( \alpha = 2.0625 \)

density

\( \alpha = 2.0703125 \)

density
alpha = 2.09375

\begin{align*}
\log \text{density} & \\
\text{x} & \end{align*}

alpha = 2.1015625

\begin{align*}
\log \text{density} & \\
\text{x} & \end{align*}
alpha = 2.15625

alpha = 2.1640625

log density

density

x

x
\[
\text{alpha} = 2.21875
\]

\[
\alpha = 2.2265625
\]
alpha = 2.234375

alpha = 2.2421875
\begin{align*}
\text{alpha} &= 2.34375 \\
\text{alpha} &= 2.34375 \\
\text{alpha} &= 2.3515625 \\
\text{alpha} &= 2.3515625
\end{align*}
alpha = 2.40625

![Graph of log density vs. x with alpha = 2.40625](image1)

alpha = 2.40625

![Graph of density vs. x with alpha = 2.40625](image2)

alpha = 2.4140625

![Graph of log density vs. x with alpha = 2.4140625](image3)

alpha = 2.4140625

![Graph of density vs. x with alpha = 2.4140625](image4)
\[ \text{density} \]

\[ \text{log density} \]

\[ \alpha = 2.4375 \]

\[ \alpha = 2.4453125 \]
alpha = 2.453125

alpha = 2.453125

alpha = 2.4609375

alpha = 2.4609375
alpha = 2.515625

alpha = 2.5234375
alpha = 2.53125

alpha = 2.5390625
alpha = 2.546875

alpha = 2.5546875
alpha = 2.578125

\[ \log \text{density} \]

\[ x \]

alpha = 2.5859375

\[ \log \text{density} \]

\[ x \]
alpha = 2.65625

alpha = 2.6640625
alpha = 2.671875

alpha = 2.671875

alpha = 2.6796875

alpha = 2.6796875
alpha = 2.703125

alpha = 2.703125

alpha = 2.7109375

alpha = 2.7109375
alpha = 2.71875

alpha = 2.7265625
alpha = 2.765625

alpha = 2.7734375
alpha = 2.8125

alpha = 2.8125

alpha = 2.8203125

alpha = 2.8203125
alpha = 2.921875

alpha = 2.9296875
\[
\text{alpha} = 2.9375
\]

\[
\text{alpha} = 2.9453125
\]
alpha = 2.96875

alpha = 2.9765625