Bayesian and Empirical Bayesian Forests

Matt Taddy TADDY@CHICAGOBOOTH.EDU
University of Chicago Booth School of Business

Chun-Shen Chen
eBay

Jun Yun
eBay

Abstract

We interpret random forests via the framework of distribution-free nonparametric Bayesian analysis, so that the ensemble average is an approximation to posterior mean inference for the population CART tree. This insight motivates a class of fully Bayesian Forest (BF) algorithms that provide small gains in predictive performance and large gains in interpretability (from a Bayesian perspective) over their classically bagged predecessors. The framework is then applied to derive Empirical Bayesian Forests (EBF), in which a single short tree "trunk" is estimated and Bayesian Forests are fit to the data at each leaf of the trunk. We are able to derive conditions under which this fixed trunk has high posterior probability, and demonstrate that in such settings the EBF performs nearly as well in out-of-sample prediction as the full BF. The advantage of pre-partitioning a fixed trunk is that the EBF ensembles can be fit independently for each initial partition. This implies a novel strategy for fitting tree ensemble predictors on data stored in a distributed file system (such as HDFS), and we show that it strongly outperforms the common strategy of fitting forests to without-replacement data subsets. The work is illustrated on a number of publicly available examples, and we describe work on its deployment at eBay.com.

1. Introduction

Decision trees are a fundamental machine learning tool. They partition the feature (input) space into regions of response homogeneity, such that the response (output) value associated with any point in a given partition can be predicted from the average for that of neighbors. The classification and regression tree (CART) algorithm of (Breiman et al., 1984) is a common recipe for building trees; it grows greedily through a series of partitions on features, each of which maximizes reduction in some measure of impurity at the current tree leaves (terminal nodes; i.e., the implied input space partitioning). The development of random forests (RF) by (Breiman, 2001), which predict through the average of many CART trees fit to bootstrap re-samples of the data, is an archetype for the successful strategy of tree ensemble learning. For prediction problems with training sets that are large relative to the number of inputs (or in conjunction with dimension reduction strategies) properly trained ensembles of trees can predict out-of-the-box as well as any carefully tuned, application-specific alternative.

This article makes three contributions to understanding and application of decision tree ensembles (or, *forests*).

Bayesian forest: A nonparametric Bayesian (npB) point-of-view allows for the straightforward, but not well-recognized, interpretation of forests as a sample from a posterior over trees. Imagine CART applied to a data generating process (DGP) with finite support: the tree greedily splits support to minimize impurity of the partitioned response distributions (terminating at some minimum-leaf-probability threshold). We present a nonparametric Bayesian model for DGPs based on multinomial draws from (large) finite support, and derive the Bayesian forest (BF) algorithm for sampling from the distribution of CART trees implied by the posterior over DGPs. Random forests are an approximation to this BF exact posterior sampler, and we show in examples that BFs provide a small but re-

Taddy is also a scientific consultant at eBay, as well as a Neubauer family faculty fellow at the University of Chicago.

liable gain in predictive performance over RFs.

Posterior tree variability: Based upon this npB framework, we derive results on the approximate stability of CART over different DGP realizations. In particular, we find that, for the data at a given node on the sample CART tree, the probability that the next split for a posterior DGP realization matches the sample split location is

$$\mathrm{p}\left(\text{split matches sample CART}\right) \gtrsim 1 - \frac{p}{\sqrt{n}} e^{-n}, \qquad (1)$$

where p is the number of possible split locations and n the number of observations on the current node. Even if p grows with n, the result indicates that partitioning can be stable conditional on the data being split. This conditioning is key: CART's well known instability is due to its recursive nature, such that a single split different from sample CART at some node removes any expectation of similarity below that node. However, for large samples, (1) implies that we will see little variation at the top hierarchy of trees in a forest. We illustrate such stability in our examples.

Empirical Bayesian forests for Big Data: the npB forest interpretation and tree-stability results lead us to propose empirical Bayesian forests (EBF) as an algorithm for building approximate BFs on massive distributed datasets (e.g., those stored on a Hadoop distributed file system). Traditional empirical Bayesian analysis fixes parameters in high levels of a hierarchical model at their marginal posterior mode, and proceeds to quantify uncertainty for the rest of the model conditional upon these fixed values. EBFs work the same way: we fit a single shallow CART *trunk* to the sample data, and then sample a BF ensemble of *branches* at each terminal node of this trunk. The initial CART trunk thus maps observations to their branch, and each branch BF can be fit in parallel without any communication with the other branches. When we expect little posterior variability about the trunk structure, an EBF sample should look similar to the (much more costly, or even infeasible) full BF sample. In a number of experiments, we compare EBFs to the common distributed-computing strategy of fitting forests to random data subsets, and find that the EBFs lead to a large improvement in predictive performance.

BFs offer small performance advantages over RFs, but their main usefulness is as a Bayesian interpretation for ensembles of CART trees. The intuition gained is especially valuable because there is little frequentest theory available on inference for decision trees. The BF model also motivates the more practical contribution of EBFs, leading to new algorithms for distributed computing and cost savings from avoiding repeatedly sampling model levels about which you have little uncertainty. This type of strategy is the key to efficient machine learning with Big Data: focus the 'Big' on the pieces of models that are most difficult to learn.

Bayesian forests are introduced in Section 2 along with a survey of Bayesian tree models, Section 3 investigates tree stability in theory and practice, and Section 4 presents the empirical Bayesian forest framework. Throughout, we use publicly available data on home prices in California to illustrate our ideas. We also provide a variety of other data analyses to benchmark performance, and close with description of how EBF algorithms are being built and perform in large-scale machine learning at eBay.com.

2. Bayesian forests

Informally, write dgp to represent the random variable defined over an as-of-yet undefined set of possible DGPs. A Bayesian analogue to classical 'distribution-free' nonparametric statistical analysis (e.g., Hollander & Wolfe, 1999) has two components:

1. set a nonparametric statistic $\mathcal{T}(dgp)$ that is of interest in your application regardless of the true DGP,
2. and build a flexible model for the DGP, so that the posterior distribution on $\mathcal{T}(dgp)$ can be derived from posterior distribution on possible DGPs.

In the context of this article, $\mathcal{T}(dgp)$ refers to a CART tree. Indeed, trees are useful precisely because they are good predictors regardless of the underlying data distribution – they do not rely upon distributional assumptions to share information across training observations. Our DGP model, detailed below, leads to a posterior for dgp that is represented through random weighting of observed support. A Bayesian forest contains CART fits corresponding to each draw of support weights, and the BF ensemble prediction is an approximate posterior mean.

2.1. Nonparametric model for the DGP

We employ a Dirichlet-multinomial sampling model in nonparametric Bayesian analysis. The approach dates back to Ferguson (1973). Chamberlain & Imbens (2003) provide an overview in the context of econometric problems. Rubin (1981) proposed the Bayesian bootstrap as an algorithm for sampling from versions of the posterior implied by this strategy, and the algorithm has since become closely associated with this model.

Use $\mathbf{z}_i = \{\mathbf{x}_i, y_i\}$ to denote the features and response for observation i. We suppose that data are drawn *independently* from a finite L possible values,

$$dgp = \mathrm{p}(\mathbf{z}) = \sum_{l=1}^{L} \omega_l \mathbb{1}_{[\mathbf{z}=\boldsymbol{\zeta}_l]} \qquad (2)$$

where $\omega_l \geq 0 \forall l$ and $\sum_l \omega_l = 1$. Thus the generating process for observation i draws l_i from a multinomial with probability ω_{l_i}, and this indexes one of the L support points. Since L can be arbitrarily large, and all data are stored as discrete, this so-far implies no restrictive assumptions beyond that of independence.

The conjugate prior for $\boldsymbol{\omega}$ is a Dirichlet distribution, written $\mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\nu}) \propto \prod_{l=1}^{L} \omega_j^{\nu_l-1}$. We will parametrize the prior with a single concentration parameter $\boldsymbol{\nu} = a > 0$, such that $\mathbb{E}[\omega_l] = a/La = 1/L$ and $\mathrm{var}(\omega_l) = (L{-}1)/[L^2(La{+}1)]$. Suppose you have the observed sample $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]'$. For convenience, we allow $\boldsymbol{\zeta}_l = \boldsymbol{\zeta}_k$ for $l \neq k$ in the case of repeated values. Write $l_1 \ldots l_n = 1 \ldots n$ so that $\mathbf{z}_i = \boldsymbol{\zeta}_i$ and $\mathbf{Z} = [\boldsymbol{\zeta}_1 \cdots \boldsymbol{\zeta}_n]'$. Then the posterior distribution for $\boldsymbol{\omega}$ has $\omega_i = a + 1$ for $i \le n$ and $\omega_l = a$ for $l > n$, so that

$$\mathrm{p}(\boldsymbol{\omega}) \propto \prod_{i=1}^{n} \omega_i^{a} \prod_{l=n+1}^{L} \omega_l^{a-1}. \tag{3}$$

This, in turn, defines our posterior for the data generating process through our sampling model in (2).

There are many possible strategies for specification of a and ζ_l for $l > n$.[1] The non-informative prior that arises as $a \to 0$ is a convenient default: in this limit, $\omega_l = 0$ with probability one for $l > n$.[2] We apply this limiting prior throughout, such that our posterior for the data generating process is a multinomial draw from the observed data points, with a uniform $\mathrm{Dir}(\mathbf{1})$ distribution on the $\boldsymbol{\omega} = [\omega_1 \ldots \omega_n]'$ sampling probabilities. We will also find it convenient to parametrize un-normalized $\boldsymbol{\omega}$ via IID exponential random variables: $\boldsymbol{\theta} = [\theta_1 \ldots \theta_n]$, where $\theta_i \overset{ind}{\sim} \mathrm{Exp}(1)$ in the posterior and $\omega_i = \theta_i/|\boldsymbol{\theta}|$ with $|\boldsymbol{\theta}| = \sum_i \theta_i$.

2.2. CART as posterior functional

Conditional upon $\boldsymbol{\theta}$, the population tree $\mathcal{T}(dgp)$ is defined through a weighted-sample CART fit. In particular, given data $\mathbf{Z}^\eta = \{\mathbf{X}^\eta, \mathbf{y}^\eta\}$ in node η, sort through all dimensions of all observations in $\mathbf{Z}^\eta$ to find the split that minimizes the average of some $\boldsymbol{\omega}$-weighted impurity metric across the two new child nodes. For example, in the case of regression trees, the impurity to minimize is weighted-squared error

$$\mathcal{I}(\mathbf{y}^\eta) = \sum_{i \in \eta} \theta_i (y_i - \bar{y}^\eta)^2 \tag{4}$$

with $\bar{y}^\eta = \sum_i \theta_i y_i / |\boldsymbol{\theta}^\eta|$. The split candidates are restricted to satisfy a minimum leaf probability, such that every node in the tree must have $|\boldsymbol{\theta}^\eta|$ greater than some threshold.[3] This procedure is repeated on every currently-terminal tree node until it is no longer possible to split while satisfying the minimum probability threshold. To simplify notation, we refer to the resulting CART tree as $\mathcal{T}(\boldsymbol{\theta})$.

2.3. Posterior inference and Bayesian forests

Following (Rubin, 1981) we can obtain a sample from the posterior on $\mathcal{T}(\boldsymbol{\omega})$ through a simple Bayesian bootstrap: for $b = 1, \ldots, B$,

- draw $\boldsymbol{\omega}^b \sim \mathrm{Exp}(\mathbf{1})$, and set $\boldsymbol{\Omega}^b = \mathrm{diag}(\boldsymbol{\omega}^b)$; then
- run CART to get $\mathcal{T}(\boldsymbol{\omega})$

We call this a Bayesian Forest, and it differs from a traditional random forest only in that the weights are drawn from a Dirichlet distribution rather than a multinomial distribution. To the extent that RF sampling provides a coarse approximation to the BF samples, the former is a convenient (and computationally advantageous, since fewer observations need to be optimized over) approximation.

2.3.1. Implementation via scikit-learn

We can implement a BF through simple adjustement of the `ensemble` module of `scikit-learn` (`sklearn`). Upon altering the forest `sample_counts` when bootsrapping to be exponential rather than binomial, we get a bayesian bootstrap. We flag this alternative Bayesian bootstrap by passing `boostrap=2` in any forest class construction (this is documented as a boolean, so we're taking advantage of easy conversion to int).

The forest predictor is a posterior mean for the population CART tree. This makes it clear when we should expect such forests to do a good job in prediction: when this tree provides a good summary of the conditional response distribution. This will be the case when the greedy algorithm leads to a good partitioning – for example, when you have lots of data and a relatively low dimensional feature space. It will not do as well when the greedy algorithm is unreliable; e.g., in high dimensions, especially when many feature dimensions have little or no influence on the response (i.e., in situations were, even if nonparametric partitioning is the correct approach, the greedy algorithm is likely to converge to a minor mode much worse than best). More importantly, a forest predictor will also be the inappropriate choice whenever compared against models with properties reflected in the true data generating process. For example, if the response is a linear function of the features, then knowing the equation for this line through the population is *much more useful* than any partitioning summary.

[1] The unobserved ζ_l act as data we imagine we might have seen, to smooth the posterior away from the data we have actually observed. See (Poirier, 2011) for discussion of how such values can be useful in application.

[2] For $l > n$ the posterior has $\mathbb{E}[\omega_l] = 0$ with variance $\mathrm{var}(\omega_l) = \lim_{a\to 0} a[n{+}a(L{-}1)]/[(n{+}La)^2(n{+}La{+}1)] = 0$.

[3] In practice this can be replaced with thresholds on the minimum number of observations at each leaf.

Or if the idiosynchratic error around the unknown conditional mean function is homoskedastic, we will get better predictions if we use that information when summarizing the conditional response distribution. The use of Random or Bayesian forests should thus be reserved for situations where we *need* to be distribution free – when the conditional response function defies parametric representation.

2.3.2. Motorcycle data illustration

We illustrate the various models with a simple one-dimensional prediction problem: what is the velocity of a motorcyle helmet after impact in a crash? This data, taken from the MASS library for R, provides a series of measurements of crash-test-dummy head acceleration in simulated motorcycle accidents.

2.4. Bayesian tree-as-parameter models

Our multinomial-sampling npB analysis, inspired by classical distribution-free strategies, can be viewed as a minimal-assumption baseline Bayesian treatment. There are plenty of other Bayesian analyses of decision trees in the literature. All of these methods treat the tree as a *parameter* which governs the DGP, rather than a functional thereof, and thus place some set of restrictions on the functional form of the distributional relationship between inputs and outputs. This section will survey these models and estimators, emphasizing that our Bayesian Forest framework allows us to place some intuition underneath the question of when each of these methods will be more appropriate that the completely nonparametric BF (or RF) baseline.

The original Bayesian tree model is the *Bayesian CART* (BCART) of Chipman, George, and McCulloch (1998). BCART treats the set of all possible decision tree fits as the support for their target parameter, and they devise a prior on trees in this set that has tree probability decreasing with its complexity (maximum depth and number of node, such that trees that are equivalent on these properties are equally likely in the prior). The data model is specified for response *y conditional upon* $\mathbf{x}$, and holds that conditional upon a given tree the responses for all observations who are allocated (via $\mathbf{x}$) to the same leaf node are IID from some parametric family. For regression trees, observations in each leaf node are IID Normal with shared variance and mean. For classification trees, each leaf represents single probability function over outcome classes.

BCART is fit through Markov chain Monte Carlo (MCMC). The algorithm draws a (correlated) posterior sample by proposing a number of possible changes to existing tree fit (e.g. grow or prune a given leaf node in the current tree) and accepting those changes (and thus move to new posterior tree draw) with probabilities proportional to the tree prior times leaf likelihood. A natural extension of this framework is to consider alternative leaf models. The original BART authors proposed linear regression at leaves, while the Treed Gaussian Processes of (Gramacy & Lee, 2008) use Gaussian procees regression at each leaf node. The TGP models, in particular, are very appealing in that they use the tree structure (one flexible regression model) to allow for nonstationarity in the Gaussian processes (another flexible regression model). As an alternative to MCMC-based tree frameworks, (Taddy et al., 2011) proposed dynamic regression trees (dynaTree) fit through sequential Monte Carlo (SMC). In particular, dynaTree uses a particle-filtering to sequentially update a particle set of potential trees for the arrival of new data with basic (grow, prune) tree operations – they allow the tree to grow naturally with streaming data.

MCMC and SMC seem naturally suited to analysis of CART, as their sampling of tree space relies upon the same incremental moves that are familiar from the construction of classical CART fits. However, this also means that these sampling algorithms are subject to the same *problems* CART has: they tend to get stuck in small regions of tree-space. For dynaTree SMC, the quality of the particle set degrades quickly (although for truly streaming data, particle rejuvination can help dramatically and has the secondary benefit of allowing tree structure to change in time). The issue is alleviated by having more complex leaf models, as in TGP, since then shorter (easier to sample efficiently) trees are required for fitting the data. Standard Monte Carlo algorithm enhancements, such as multiple parallel chains, can also help. But a full exploration of BCART-style tree space remains very difficult and requires extremely large Markov samples to have any chance of convergence.

As a solution to these mixing issues, the original BCART authors proposed Bayesian Additive Regression Trees (Chipman et al., 2010), which replaces the single tree model with the sum of many small trees. Thus, a single observation is allocated via its features to a leaf node in each of, say, T trees; the associated response is then distributed with mean equal to the average of each leaf node value and with a *shared variance across observations*. The original BART model assumes Normal response, and later extensions allow alternative response families (including nonparametric mixtures of Normals). BART solves the mixing issue by only working with very short (depth 2-5) trees. This stubby tree space is not difficult to explore, and the authors provide a fast and well-mixing MCMC algorithm. However, this resolution comes at a steep price: by specifying a data model that has shared variance for all observations, BART loses an important feature of CART, BCART, and the others: the ability to fit fitting heteroskedastic data. While the other BCART-style Bayesian trees made assumptions about the parametric distribution at each leaf node, they all (as CART does) allow completely differ-

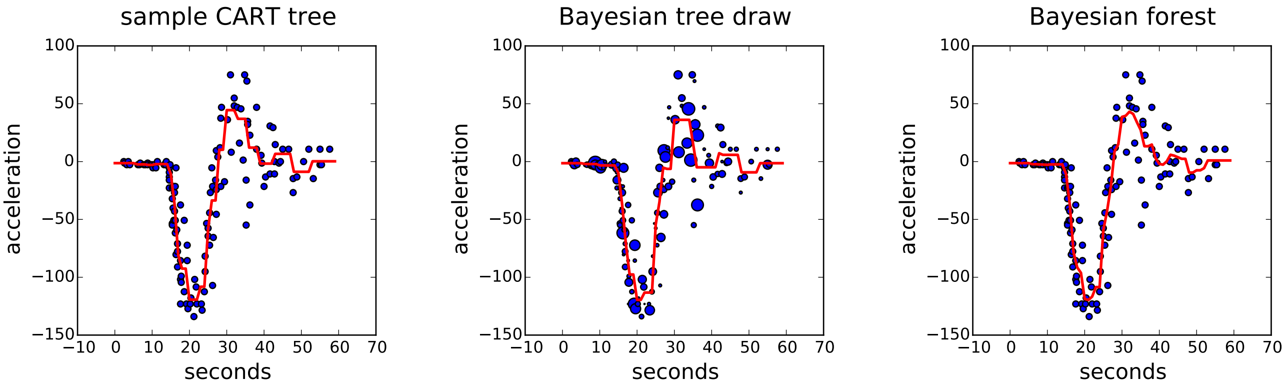


Figure 1. Motorcycle data illustration. The lines show each of a CART predictor fit to the data sample, CART fit to a sample with weights drawn IID from an exponential (point size is proportional to weight), and the BF predictor which averages over 100 weighted CART fits.

ent specifications (e.g., distinct mean and variance) at each leaf. BART's shared variance removes this property.

Despite the assumed homoskedasticity, extensive simulations have shown BART to outperform alternative flexible prediction rules. The empirical evidence of BARTS performance is strong enough that it has gained support as a default Bayesian method for flexible prediction (Hill, 2011). From our perspective, this success occurs when an estimate of the BART model on the population, with its homoskedastic Normal error variance, is a good summary of the population conditional response distribution. Many datasets, especially those analyzed by academics (and after the transformations, e.g., log, that academic statisticians apply to get better behaved response), have the property that they are well fit by flexible regression with homoskedastic errors. For the same reasons, the BCART/TGP/dynaTree models have been shown repeatedly to outperform RF in academic bake-offs; this occurs on data were their parametric response model (e.g. Normal) fits well enough to be a better summary of the data than the posterior mean CART tree. The purpose of the current article is not at all to argue against such strong performance. Rather, we wish to add intuition about the models that would allow you to consider your data applications and predict, given some basic knowledge of the DGP, whether you can leverage any of the semiparametric assumptions in BART, BCART, TGP, dynaTree, etc; or whether a distribution free summary (BF or RF) will be more useful.

Finally, we note that the Bayesian bootstrap framework can also applied to parameters of an assumed parametric DGP, and thus can apply to tree-as-parameter models. In such a framework, outlined as Bayesian bagging in (Clyde & Lee, 2001), the bootstrap mean (and related bootstrap summaries) have interpretation as *Bayesian model averaging*. However, this framework becomes difficult to interpret when the target parameter is infinite dimensional (as it is for tree models), and the MCMC or SMC sampling strategies mentioned above are much more common. The Clyde+Lee paper does consider a Bayesian bagging (model averaging) treatment of CART, but their desire to interpret the tree as a model parameter leads to quite different algorithm and analysis.

2.5. Compare and contrast

The spectrum of Bayesian analysis of trees moves from distribution free BF and RF, through the Bayesian nonparametrics-via-many-parameters approach of BCART, TGP, and dynaTree, to semi-parametric BART which makes truly restrictive assumptions about the response distribution. Instead of blindly picking the method which has performed best in past studies of arbitrary data (e.g., bake-offs in academic papers), we can use this information to predict which methods will suit our application.

In this section, we illustrate with two examples that have very different properties. The first is a simulation study with known homoskedastic errors in the target. The second considers raw dollar-value home prices in california as the response of interest. On these examples, we fit a CART, RF, BF, BART, and BCART as described above. We also include the extremely random trees (ET) of (Geurts et al., 2006), which are similar to random forests except that (a) instead of optimizing greedy splits, candidate splits are chosen randomly and the best is used and (b) all of the data is used to fit each tree (there is no bootstrap resampling or reweighting). ETs perform well on small datasets, where CART has a high tendancy to overfit without careful pruning. On such small datasets (e.g. our Friedman example) the restriction of population support to observed support (assumed in our nonparametric analysis) becomes less acceptable and we might expect the Random and Bayesian forests to suffer.

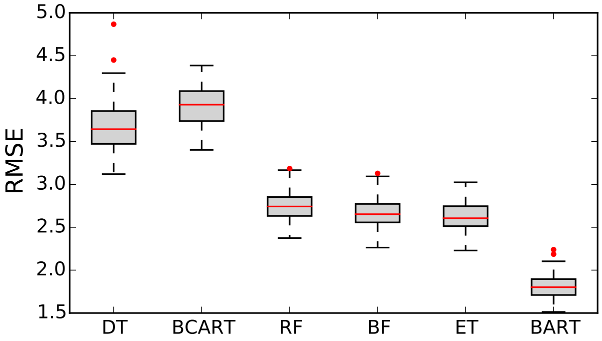


Figure 2. Friedman data: prediction Root Mean Square Error.

2.5.1. Friedman example

A common simulation experiment in evaluating flexible predictors is based around the so-called Friedman function, first proposed for this purpose in (Friedman, 1991) MARS paper. The function is

$$\begin{aligned} y &= f(\mathbf{x}) + \varepsilon \qquad (5)\\ &= 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon \end{aligned}$$

where $\varepsilon \sim \mathrm{N}(0,1)$ and $x_j \sim \mathrm{U}(0,1)$. We include as features for training the spurrious $x_6 \ldots x_p$, matching Friedman with MARS and Chipman et al with BART. Each candidate regression model is fit to 100 random draws from the Friedman function, and tested at 1000 new x locations (simulated uniformly as in the training data). We calculate the root mean square error between predicted values and true $f(\mathbf{x})$ as a measure predictive performance.

As predicted, the only model that assumes the (true) homoskedastic error structure, BART, well outperforms all others. The two forests, BF and RF, are both a large improvement over a single decision tree. The fully Bayesian BF is only about 1% better than the approximately Bayesian RF, as Bayesian and classical bootstrapping weights differ little in practice. Both are outperformed slightly by the extremely random trees (ET), which might be expected due to the small sample size (for which the observed support approximation to population support, assumed in our forest interpretation, is poor). The only suprise for us is the very poor performance of BCART (even worse than a single decision tree); we hypothesis that this is due to the notoriously poor mixing of the BCART MCMC, such that this fit is neither finding a posterior mean (as it is intended to) or optimizing to a local posterior mode (as DT does).

2.5.2. California housing data

For the next example, we consider prediction of median home price by census block in california, based upon eight features of each region (location, income, housing stock). The data are taken from http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html. This problem has a response distribution that is difficult to summarize parametrically. Standard analysis takes the log price as the response of interest, which at least tames some of the error heteroskedasticity. Instead, we will attempt to model the conditional expection of raw dollar home values. This mimics the setting common in the analysis of online transaction data (e.g. clicks or dollars spent), where the variable average effects and predictive performance on raw $ or click scale are of primary interest.

It appears from the histogram (and investigating the raw values) that the data has been capped at 500k. In any case, the unconditional response has a long right tail, but is much more regular than what we commonly see in digital commerce applications (e.g., see Taddy et al, 2014, for transaction data with massive spikes at zero and at psychological price thresholds, as well as a tail that includes values 50k times larger than the mean).

Note that bart fit takes around 75 sec in R, vs 10 sec for BF run in serial.

The results are now reversed (except for DT and BCART, which still underperform all others). The methods which place no assumptions on the data generating process, RF and BF, do much better than BART and it's restrictive error model. The implicit regularization of extratrees is no longer of any benefit, as the larger sample size means that our finite support approximation is solid. As always, BF offers a real but small gain over RF.

2.6. Empirical Bayesian Forests

The previous sections re-interpret forests (random or Bayesian – they are roughly equivalent) as samples from the distribution-free posterior over greedy CART fits. This interpretation is now applied to guide strategies for forest fit on massive data, by treating that problem as one of *approximate* sampling from the posterior.

Empirical Bayes (EB) is an established framework with a successful track record in fast approximate Bayesian inference; see, e.g., (Efron, 2010) for an overview. In parametric models, EB can be interpreted as fixing at their marginal posterior maximum (MAP) the parameters at higher levels of a *hierarchical model*. For example, in the simple setting of many group means (the average student test score for each of many schools) shrunk to an overall global center (the outcome for an imaginary 'average school'), an EB procedure would first find the overall average test score

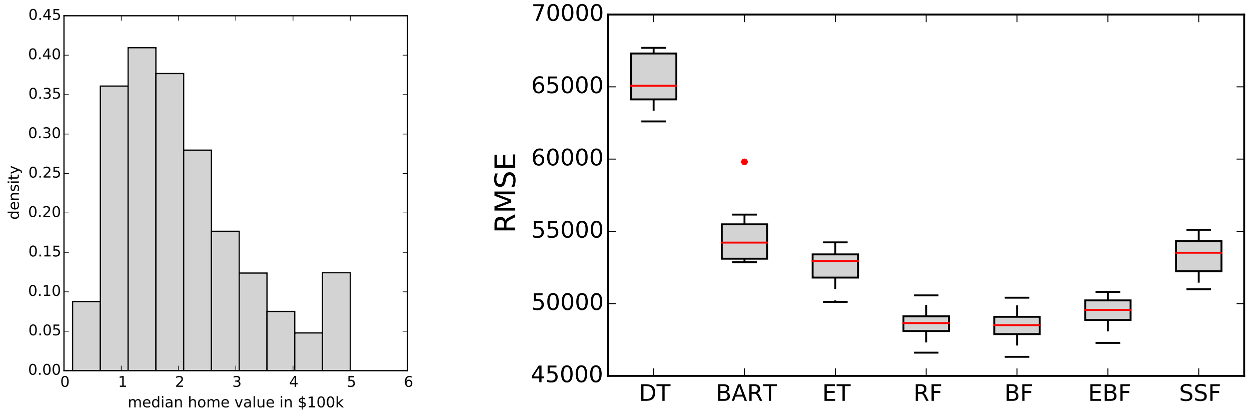


Figure 3. California housing data

across all schools and then shrink posterior estimated score means towards this value within each individual school. (Kass & Steffey, 1989) investigate such procedures and show that, under fairly general conditions, the conditional posterior (conditioning on a marginal MAP overall average) for each group mean (school score average) quickly approaches the fully Bayesian unconditional posterior as the sample size grows.

CART trees are not a parametric model, but they are hierarchical and admit an interpretation similar to those studied in (Kass & Steffey, 1989). The data which reaches any given interior node is a function of the partitioning implied by nodes shallower in the tree. Moreover, due to the *greedy* algorithm through which CART grows, a given shallow tree is unaffected by changes to the tree structure below. Finally (this point is investigated in detail in the next section), the partitioning implied by shallower sub-trees is much more stable (has lower posterior variability) than the fully grown trees.

An Empirical Baysian Forest (EBF) takes advantage of this hierarchical struture by fixing the highest levels in the hierarchy – the earliest CART splits – at a single estimate of high posterior probability. We refer to this fixed sub tree as the *trunk*, and will generally take it as a large-leaf (i.e., short) CART fit to the entire dataset. This is fit to unweighted data, which corresponds to the mean of our nonparametric Bayesian posterior over DGPs. Based upon this trunk, data is partitioned into leaves of manageable size, and a forest is fit to each. That is, EBF replaces the BF posterior sample of trees with a conditional posterior sample that takes the pre-fit trunk as fixed at its marginal MAP. Since this trunk has relatively low variance, the EBF should provide predictions similar to that of the full BF.

EBF is contrasted to a common existing algorithm, which for Big Data simply splits the data into sub-samples and fits a forest to each (cite?). This strategy can be justified from standard sampling theory as yeilding a tree average that is unbiased for the population tree average, but it discards information available in the full Big Data sample. Given the high uncertainty and low effective sample sizes associated with deep tree structure (otherwise, we wouldn't need to bother with Big Data in the first place), such sub-sampling is ignoring potentially useful information. Indeed, the next section argues that so long as the initial tree Trunk is not overfit, then EBF should outperform forest fits averaged across data subsets. Either approach can be used in conjuntion with further techniques for distribution of the tree fit, e.g. via the PLANET MapReduce scheme of (Panda et al., 2009).

2.6.1. Uncertainty about trees

Theory on decision trees is sparse. The original CART book of (Breiman et al., 1984) provides results on the consistency of tree partitioners; they show that any partitioner that which is able to eventually learn enough to partition into very small-diameter leaves relative to the DGP probability function will be able to reproduce the conditional response distribution of that DGP. However, this result says little about the structure of the underlying trees, nor does it say little about the ability of a tree to predict when there is not enough data to reproduce a finely partition the DGP. Another set of theory focuses on the frequentist properties of individual split decisions. In the original CHAID work of (Kass, 1980), split decisions (on soley categorical variables) are based upon χ^2 tests of the contingency table at each leaf node. (Loh, 2002) and (Hothorn et al., 2006) are example generalizations, both of which combat the various multiple testing and other biases inherrent in tree-building through a sequence of hypothesis tests. However, such contributions provide little intuition about the variability of an entire decision tree or in the setting where we are not working from a no-split null hypothesis distribution.

Despite this lack of theory, it is widely recognized that there is a large amount of uncertainty (sampling, or posterior) about the structure of decision trees. For example, (Geurts & Wehenkel, 2000) present a large amount of simulation-based empirical evidence of tree uncertainty, and they find that the locations and order of split points in trees is sometimes no-better than random (indeed, this motivates work by the same authors on extremely random trees). The intuition behind such randomness is clear: the probability of a tree having any given branch structure is the product of conditional probabilities for each successive split. After a enough steps any specific tree approaches probability of zero. This uncertainty argues for the inherrently 'Big' data requirements of flexible tree regression: there is enough model complexity and uncertainty that we need to bring as much data as possible to bear on learning. This was observed by, e.g., the authors of (Panda et al., 2009) when they compare RFs fit to the full dataset to those averaged across tree fits on without-replacement sub-samples of the data (which is a common technique for fitting RFs in distribution).

However, it is possible that elements of the tree structure are relatively stable. For example, in the context of boosting, (Appel et al., 2013) argue that the *conditionally* optimal split locations for internal nodes (they call stumps) are learnable from subsets of the full data alocated to each node. They use this to propose a faster boosting algorithm. In this paper we make a related claim: the top of a tree (its trunk) has structure that is stable enough that it can be estimated in advance and fixed at a posterior mode. Our argument for the potential stability of tree trunks begins with a result on the mean and variance of the difference, at each internal node, betwen MSE associated with the modal CART partioning and the MSE that results after a split at any other location. The uncertainty associated with these statistics, the maximum of which determines our tree, is shown to shrink with e^{-n} under our npB model. This is reasurring. Moreover, we can lower bound the probability of selecting the modal CART split among p possible as $1 - \frac{p}{\sqrt{n}}e^{-n}$, which will be quickly close to one even if p grows with n.

2.6.2. Probability of the sample CART tree

We focus on regression trees for this example, wherein node impurity is measured as the sum of squared errors. Consider the simplified setup where each $x_j \in \{0,1\}$ is a binary random variable (possibly created as a discretization of a continuous input or through dummy expansion of a categorical input). Say that $\mathtt{f}_j = \{i : x_{ij} = 0\}$ and $\mathtt{t}_j = \{i : x_{ij} = 1\}$ are the corresponding implied partitions. We will derive the distribution for the weighted SSE resulting from a split on any such variable,

$$\sigma_j^2(\boldsymbol{\theta}) = \frac{1}{n}\sum_i \theta_i \left[y_i - \mu_j(x_{ij})\right]^2$$

where $\mu_j(0) = \sum_{i\in\mathtt{f}_j} y_i\,\theta_i / \left|\boldsymbol{\theta}_{\mathtt{f}_j}\right|$ and $\mu_j(1) = \sum_{i\in\mathtt{t}_j} y_i\,\theta_i / \left|\boldsymbol{\theta}_{\mathtt{t}_j}\right|$.

We can simulate the posterior for σ_j implied by the gamma posterior on weights $\boldsymbol{\theta}$, but an exact analytic expression is not available. Instead, we follow the approach used in (Lancaster, 2003), (Poirier, 2011), and (Taddy et al., 2014): we derive a first-order Taylor approximation to the function σ_j and describe the posterior for that related functional. In particular, the $1 \times n$ gradient of σ_j with respect to $\boldsymbol{\theta}$ is

$$\nabla\sigma_j^2 = \nabla\frac{1}{n}\left[\sum_i \theta_i y_i^2 - \frac{1}{|\boldsymbol{\theta}_{\mathtt{f}}|}\left(\mathbf{y}_{\mathtt{f}}'\boldsymbol{\theta}_{\mathtt{f}}\right)^2 - \frac{1}{|\boldsymbol{\theta}_{\mathtt{t}}|}\left(\mathbf{y}_{\mathtt{t}}'\boldsymbol{\theta}_{\mathtt{t}}\right)^2\right] \tag{6}$$

which has elements $\nabla_i\sigma_j^2 = \left(y_i - \mu(x_{ij})\right)^2/n$

The Taylor approximation is then

$$\sigma_j^2 \approx \tilde\sigma_j^2 = \sigma_j^2(\mathbf{1}) + \nabla\sigma^2\big|_{\boldsymbol{\theta}=\mathbf{1}}(\boldsymbol{\theta}-\mathbf{1}) = \frac{1}{n}\sum_i \theta_i\left(y_i - \bar y_j(x_{ij})\right)^2$$

with $\bar y_j(0) = \frac{1}{n_{\mathtt{f}_j}}\sum_{i\in\mathtt{f}_j} y_i$ and $\bar y_j(t) = \frac{1}{n_{\mathtt{t}_j}}\sum_{i\in\mathtt{t}_j} y_i$ are the observed response averages in each partition.

Suppose that $\sigma_1(\mathbf{1}) < \sigma_j(\mathbf{1})\ \forall j$, so that variable 1 is that selected to dictate partitioning by CART. Then we can consider variability about this selection by looking at the difference betwen the approximate SSE for alternative partitions and this modal value,

$$\Delta_j(\boldsymbol{\theta}) = \tilde\sigma_1^2 - \tilde\sigma_j^2 = \frac{1}{n}\sum_i \theta_i\left[\bar y_1^2(x_{i1}) - \bar y_j^2(x_{ij}) - 2y_i\left(\bar y_1(x_{i1}) - \bar y_j(x_{ij})\right)\right]$$

Say $\mathbf{d}_j = [d_{j1}\ldots d_{jn}]'$ is the vector of squared error differences. Then the total difference has mean $\mathbb{E}\Delta_j = \bar d_j$ and variance $\mathrm{var}\Delta_j = \mathbf{d}_j'\mathbf{d}_j/n^2$. Since Δ_j is the mean of independent Gamma random variables with known means and variances, the central limit theorem applies so that delta converges in distribution to a Gaussian

$$\sqrt{n}(\Delta_j(\boldsymbol{\theta}) - \bar d_j) \rightsquigarrow \mathrm{N}(0,\ \mathbf{d}_j'\mathbf{d}_j/n).$$

Thus the full-sample CART split occurs if all $\Delta_j < 0$, which occurs with probability (note $\bar d_j < 0$)

$$\mathrm{p}\left(\Delta_j < 0:\ j = 2 \dots p\right) \geq 1 - \sum_{j=2}^{p} \mathrm{p}(\Delta_j > 0) \rightsquigarrow 1 - \sum_{j=2}^{p} \Phi\left(- \right. \qquad \frac{\exp(-nz_j^2/2)}{z_j\sqrt{n}}$$

where $z_j = \left|\bar{d}_j\right| \left(\mathbf{d}_j' \mathbf{d}_j / n\right)^{-\frac{1}{2}}$ is sample mean increase in impurity over the sample standard deviation of impurity. This ratio is bounded in probability, so that that the exponential bound goes to zero very quickly with n. In particular, ignoring variation in z_j across variables we get

$$\mathrm{p}\left(\text{split matches sample CART}\right) \gtrsim 1 - \frac{p}{\sqrt{n}} e^{-n}.$$

Thus, even in a more difficult setting where $p \approx n \times d$, with d some underlying continuous variable dimension and p the input dimension discretization on these variables, the probability of the observed split goes to one at order $\mathcal{O}(n^{-1})$ if $d < e^n / n^{3/2}$.

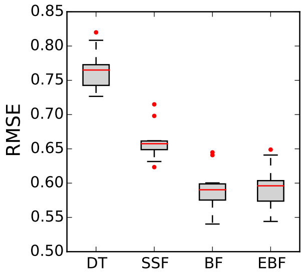
Given this, why is there any uncertainty at all about trees? The answer is recursion: each split is conditionally stable given the sample at the current node, but the probability of any specific sequence of splits is roughly (we don't actually have independence) the product of individual node split probabilities. This can get small as we move deeper down the tree. Moreover, given one split different from the sample CART tree, the rest of the tree will grow aribitrarily far from this modal structure. In addition, the sample size going into our probability bounds is shrinking exponentially with each partition, whereas the dimension of eligible split variables is reduced only by one at each level.

2.6.3. Tree stability in California

We'll illustrate trunk stability and the idea of EBFs on the California housing data from above. To begin, consider a *trunk* with no less than 3500 census blocks (out of 20640 total) in each leaf partition. Greedy CART leads to a five node tree.

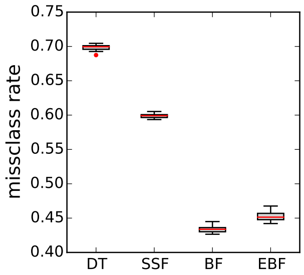
Shallow tree variance The tree above represents the optimal (greedy) five partition CART. As predicted by theory, it turns out to be very stable. Running a random forest of trees which stop at this minimum leaf size, we get

We can also repeatedly sample 90% of the data, and the CART fit with `min_sample_leaf=3500` is always similar: from visual inspection, each fold fit splits on the same variables, just on slightly different values. Moreover, a Bayesian forest of trees limited to this node size does little better in OOS prediction (only 1%), offering additional evidence that the single CART fit is close to the posterior mean at this depth.



RMSE	% WTB	
0.5905	0.0	BF
0.5953	0.8	EBF
0.6607	11.9	SSF
0.7648	29.5	DT

Figure 5. Wine Data: 10-fold OOS prediction experiment Root Mean Square Error and % Worse than Best for the mean RMSE.



MCR	% WTB	
0.4341	0.0	BF
0.4531	4.4	EBF
0.5989	38.0	SSF
0.6979	60.8	DT

Figure 6. Beer Data: 10-fold OOS prediction experiment Miss-Classification Rate and % Worse than Best for the average MCR.

OOS experiment Finally, we consider OOS prediction for the EBF, using a trunk fixed at the shallow trees from above, and averaging of Bayesian forests fit to sub-samples of comparable size. The EBF does around 8% better than random sub-sampling, and only around 2% worse than the BF fit to the entire dataset.

```
2.0\% worse than full sample BF
8.0\% better than random sub sampling
```

2.6.4. Beer and Wine

As another test of OOS performance, we consider the Vino Verde dataset of http://www3.dsi.uminho.pt/pcortez/wine/. There are 4898 observations on 11 continuous attributes (physiochemical properties of the wine) plus wine color (red or white) as inputs, with an 'expert' quality ranking on the scale of 0-10 as response. Here, we find that averaging forests fit to 5 random subsets of the data do 10% worse in OOS prediction than the EBF conditional on pre-tree partitioning into 5 leaves. Here, the EBF does close to as well as a full forest: it is only 1% more expensive than the full BF fit.

2.7. Discussion

Tree-based learning algorithms are massively useful and very popular. However, the lack of parametric structure

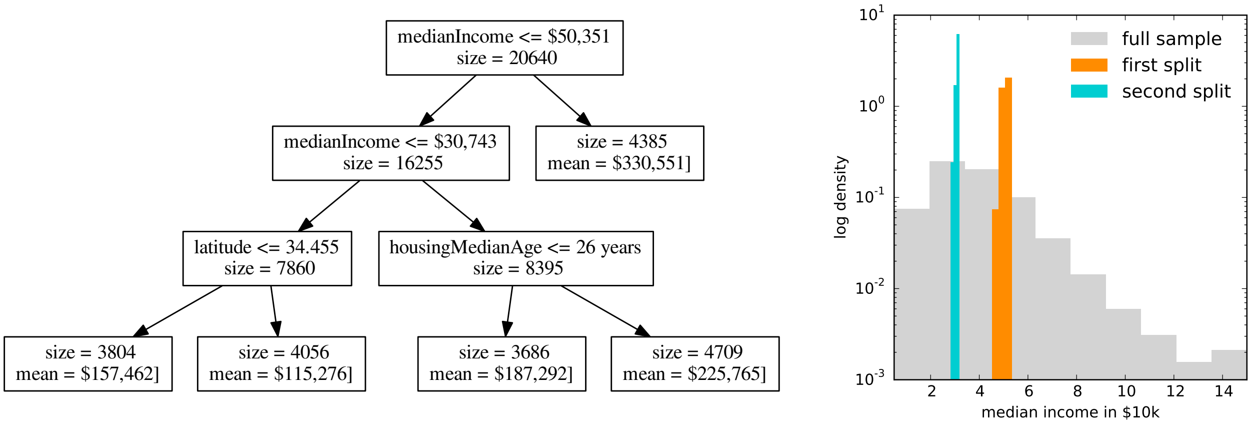


Figure 4. California housing data

makes it difficult to find meaningful theoretical guidance for their performance in different settings. This leads to an exclusive reliance upon cross validation experiments (and simulation studies) for evaluating candidate algorithms, which while informative is less helpful for a practitioner who needs to know which algorithm is best suited to particular data. Bayesian modelling and interpretation of tree models is a help here, as the model assumptions and properties are transparent to anyone who is able to understand probabalistic models. Our Bayesian Forests bring this intuition to one of the most heavily used tree-based algorithms, Random Forests, and thus allow it to be realated easily to other candidate Bayesian methods.

The other contribution here, which likely has more practical implications, is the proposal of Empirical Bayesian forests. Given our Baysian interpretation of Forests, the common idea of fixing high-level parameters with little uncertainty is applied in an algorithm that yields posterior mean predictions close to that of a forest fit to the full dataset while working efficiently on small sub-samples. Even if one is not willing to assume in advance how deep a tree will be stable, in environments where forests are repeatedly re-fit (to incorporate incomming information) one can monitor the stability at higher levels and use this information to fix some tree elements in future runs.

References

Appel, Ron, Fuchs, Thomas, Dollr, Piotr, and Perona, Pietro. Quickly boosting decision trees-pruning underachieving features early. In *JMLR Workshop and Conference Proceedings*, volume 28, pp. 594–602. JMLR, 2013.

Breiman, Leo. Random forests. *Machine Learning*, 45: 5–32, 2001.

Breiman, Leo, Friedman, Jerome, Olshen, Richard, and Stone, Charles. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

Chamberlain, Gary and Imbens, Guido W. Nonparametric applications of bayesian inference. *Journal of Business & Economic Statistics*, 21:12–18, 2003.

Chipman, Hugh A., George, Edward I., and McCulloch, Robert E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4:266–298, 2010.

Clyde, Merlise and Lee, Herbert. Bagging and the bayesian bootstrap. In *Artificial Intelligence and Statistics*, 2001.

Efron, Bradley. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. IMS Statistics Monographs. Cambridge, 2010.

Ferguson, T.S. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

Friedman, Jerome. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67, 1991.

Geurts, Pierre and Wehenkel, Louis. Investigation and reduction of discretization variance in decision tree induction. In *Machine Learning: ECML 2000*, pp. 162–170. Springer, 2000.

Geurts, Pierre, Ernst, Damien, and Wehenkel, Louis. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

Gramacy, Robert B. and Lee, Herbert K. H. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103:1119–1130, 2008.

Hill, Jennifer L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, January 2011.

Hollander, Myles and Wolfe, Douglas. *Nonparametric Statistical Methods*. Wiley, 2nd edition, 1999.

Hothorn, Torsten, Hornik, Kurt, and Zeileis, Achim. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674, 2006.

Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127, 1980.

Kass, Robert E. and Steffey, Duane. Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84:717, 1989. ISSN 01621459.

Lancaster, Tony. A note on bootstraps and robustness. Technical report, Working Paper, Brown University, Department of Economics, 2003. URL http://www.econstor.eu/handle/10419/80157.

Loh, Wei-Yin. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12: 361–386, 2002.

Panda, Biswanath, Herbach, Joshua S., Basu, Sugato, and Bayardo, Roberto J. Planet: massively parallel learning of tree ensembles with mapreduce. *Proceedings of the VLDB Endowment*, 2(2):1426–1437, 2009.

Poirier, Dale J. Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed bayesian bootstrap. *Econometric Reviews*, 30(4):457–468, 2011.

Rubin, Donald. The bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

Taddy, M. A., Gramacy, R. B., and Polson, N. G. Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106:109–123, 2011.

Taddy, Matt, Gardner, Matt, Chen, Liyun, and Draper, David. Heterogeneous treatment effects in digital experimentation. *arXiv preprint arXiv:1412.8563*, 2014.