Received: 16 February 2019 | Accepted: 16 February 2021

DOI: 10.1111/rssb.12423

ORIGINAL ARTICLE

Variable selection with ABC Bayesian forests

Yi Liu[1] | Veronika Ročková[2] | Yuexi Wang[2]

[1]Department of Statistics, University of Chicago, Chicago, USA

[2]Booth School of Business, University of Chicago, Chicago, USA

Correspondence
Veronika Ročková, Booth School of Business, University of Chicago, Chicago, IL, USA.
Email: veronika.rockova@chicagobooth.edu

Abstract

Few problems in statistics are as perplexing as variable selection in the presence of very many redundant covariates. The variable selection problem is most familiar in parametric environments such as the linear model or additive variants thereof. In this work, we abandon the linear model framework, which can be quite detrimental when the covariates impact the outcome in a non-linear way, and turn to tree-based methods for variable selection. Such variable screening is traditionally done by pruning down large trees or by ranking variables based on some importance measure. Despite heavily used in practice, these ad hoc selection rules are not yet well understood from a theoretical point of view. In this work, we devise a Bayesian tree-based probabilistic method and show that it is consistent for variable selection when the regression surface is a smooth mix of $p > n$ covariates. These results are the first model selection consistency results for Bayesian forest priors. Probabilistic assessment of variable importance is made feasible by a spike-and-slab wrapper around sum-of-trees priors. Sampling from posterior distributions over trees is inherently very difficult. As an alternative to Markov Chain Monte Carlo (MCMC), we propose approximate Bayesian computation (ABC) Bayesian forests, a new ABC sampling method based on data-splitting that achieves higher ABC acceptance rate. We show that the method is robust and successful at finding variables with high marginal inclusion probabilities. Our ABC algorithm provides a new avenue towards approximating the median probability model in non-parametric setups where the marginal likelihood is intractable.

KEYWORDS
approximate Bayesian computation, BART, consistency, spike-and-slab, variable selection

1 | PERSPECTIVES ON NON-PARAMETRIC VARIABLE SELECTION

In its simplest form, variable selection is most often carried out in the context of linear regression (Fan & Li, 2001; George & McCulloch, 1993; Tibshirani, 1996). However, confinement to linear parametric forms can be quite detrimental for variable importance screening, when the covariates impact the outcome in a non-linear way (Turlach, 2004). Rather than first selecting a parametric model to filter out variables, another strategy is to first select variables and then build a model. Adopting this reversed point of view, we focus on developing methodology for the so called ‘model-free’ variable selection (Chipman et al., 2001).

There is a long strand of literature on the fundamental problem of non-parametric variable selection. One line of research focuses on capturing non-linearities and interactions with basis expansions and performing grouped shrinkage/selection on sets of coefficients (Lin & Zhang, 2006; Radchenko & James, 2010; Ravikumar et al., 2009; Scheipl, 2011). Lafferty and Wasserman (2008) propose the RODEO method for sparse non-parametric function estimation through regularization of the derivative expectation operator and provide a consistency result for the selection of the optimal bandwidth. Candes et al. (2018) propose a model-free knock-off procedure, controlling FDR in settings when the conditional distribution of the response is arbitrary. In the Bayesian literature, Savitsky et al. (2011) deploy spike-and-slab priors on covariance parameters of Gaussian processes to erase variables. In this work, we focus on other non-parametric regression techniques, namely trees/forests which have been ubiquitous throughout machine learning and statistics (Breiman, 2001; Chipman et al., 2010). The question we wish to address is whether one can leverage the flexibility of regression trees for effective (consistent) variable importance screening.

While trees are routinely deployed for data exploration, prediction and causal inference (Gramacy & Lee, 2008; Hill, 2011; Taddy et al., 2011a), they have also been used for dimension reduction and variable selection. This is traditionally done by pruning out variables or by ranking them based on some importance measure. The notion of variable importance was originally proposed for CART using overall improvement in node impurity involving surrogate predictors (Breiman et al., 1984). In random forests (RF), for example, the importance measure consists of a difference between prediction errors before and after noising the covariate through a permutation in the out-of-bag sample. However, this continuous variable importance measure is on an arbitrary scale, rendering variable selection ultimately ad hoc. Principled selection of the importance threshold (with theoretical guarantees such as FDR control or model selection consistency) is still an open problem. Simplified variants of importance measures have begun to be understood theoretically for variable selection only very recently (Ishwaran, 2007; Kazemitabar et al., 2017).

Bayesian trees and forests select variables based on probabilistic considerations. The BART procedure (Chipman et al., 2010) can be adapted for variable selection by forcing the number of available splits (trees) to be small, thereby introducing competition between predictors. BART then keeps track of predictor inclusion frequencies and outputs a probabilistic importance measure: an average proportion of all splitting rules inside a tree ensemble that split on a given variable, where the average is taken over the Markov Chain Monte Carlo (MCMC) samples. This measure cannot be directly interpreted as the posterior variable inclusion probability in anisotropic regression surfaces, where wigglier directions require more splits. Bleich et al. (2014) consider a permutation framework for obtaining the null distribution of the importance weights. Zhu et al. (2015) implement reinforcement learning for selection of splitting variables during tree construction to encourage splits on fewer more important variables. All these developments point to the fact that regularization is key to enhancing performance of trees/forests in high dimensions. Our approach differs in that we impose regularization from *outside* the tree/forest through a spike-and-slab wrapper.

Spike-and-slab variable selection consistency results have relied on analytical tractability (approximation availability) of the marginal likelihood (Castillo et al., 2015; Johnson & Rossell, 2012; Narisetty & He, 2014). Nicely tractable marginal likelihoods are ultimately unavailable in our framework, rendering the majority of the existing theoretical tools inapplicable. For these contexts, Yang and Pati (2017) characterized general conditions for model selection consistency, extending the work of Lember and van der Vaart (2007) to non *iid* setting. Exploiting these developments, we show variable selection consistency of our non-parametric spike-and-slab approach when the regression function is a smooth mix of covariates. Building on Ročková and van der Pas (2020), our paper continues the investigation of missing theoretical properties of Bayesian CART and BART. We show model selection consistency when the smoothness is known as well as joint consistency for both the regularity level *and* active variable set when the smoothness is not known and when $p > n$. These results are the first model selection consistency results for Bayesian forest priors.

The absence of a tractable marginal likelihood complicates not only theoretical analysis, but also computation. We turn to approximate Bayesian computation (ABC) (Csillery et al., 2010; Marin et al., 2012; Plagnol & Tavaré, 2004) and propose a procedure for model-free variable selection. Our ABC method *does not* require the use of low-dimensional summary statistics and, as such, it *does not* suffer from the known difficulty of ABC model choice (Robert et al., 2011). Our method is based on sample splitting where at each iteration (a) a random subset of data is used to come up with a proposal draw and (b) the rest of the data is used for ABC acceptance. This new data-splitting approach increases ABC effectiveness by increasing its acceptance rate. ABC Bayesian forests relate to the recent line of work on combining machine learning with ABC (Jiang et al., 2017; Pudlo et al., 2015). We propose dynamic plots that describe the evolution of marginal inclusion probabilities as a function of the ABC selection threshold.

The paper is structured as follows. Section 2 introduces the spike-and-slab wrapper around tree priors. Section 3 develops the ABC variable selection algorithm. Section 4 presents model selection consistency results. Section 5 demonstrates the usefulness of the ABC method on simulated data and Section 6 wraps up with a discussion.

1.1 | Notation

With $\|\cdot\|_n$ we denote the empirical L^2 norm. The class of functions $f(\boldsymbol{x})\colon [0,1]^p \to \mathbb{R}$ such that $f(\cdot)$ is constant in all directions excluding $\mathcal{S}_0 \subseteq \{1, \cdots, p\}$ is denoted with $\mathcal{C}(\mathcal{S}_0)$. With $\mathcal{H}_p^\alpha$, we denote α-Hölder continuous functions with a smoothness coefficient α. $a \lesssim b$ denotes a is less or equal to b, up to a multiplicative positive constant, and $a \approx b$ denotes $a \lesssim b$ and $b \lesssim a$. The ε-covering number of a set Ω for a semimetric d, denoted by $N(\varepsilon; \Omega; d)$, is the minimal number of d-balls of radius ε needed to cover set Ω.

2 | BAYESIAN SUBSET SELECTION WITH TREES

We will work within the purview of non-parametric regression, where a vector of continuous responses $\boldsymbol{Y}^{(n)} = (Y_1, \cdots, Y_n)'$ is linked to fixed (rescaled) predictors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in [0,1]^p$ for $1 \leq i \leq n$ through

$$Y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for} \quad 1 \leq i \leq n, \tag{1}$$

where $f_0(\cdot)$ is the regression mixing function and $\sigma^2 > 0$ is a scalar. It is often reasonable to expect that only a small subset $\mathcal{S}_0$ of $q_0 = |\mathcal{S}_0|$ predictors actually exert influence on $\boldsymbol{Y}^{(n)}$ and contribute to the mix. The subset $\mathcal{S}_0$ is seldom known with certainty and we are faced with the problem of variable selection. Throughout this paper, we assume that the regression surface is smoothly varying (α-Hölder continuous) along the active directions $\mathcal{S}_0$ and constant otherwise, that is we write $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$.

Unlike linear models that capture the effect of a single covariate with a single coefficient, we permit non-linearities/interactions and capture variable importance with (additive) regression trees. By doing so, we hope to recover non-linear signals that could be otherwise missed by linear variable selection techniques.

As with any other non-parametric regression method, regression trees are vulnerable to the curse of dimensionality, where prediction performance deteriorates dramatically as the number of variables p increases. If an oracle were to isolate the active covariates $\mathcal{S}_0$, the fastest achievable estimation rate would be $n^{-\alpha/(2\alpha+|\mathcal{S}_0|)}$. This rate depends only on the intrinsic dimensionality $q_0 = |\mathcal{S}_0|$, not the actual dimensionality p which can be much larger than n. Recently, Ročková and van der Pas (2020) showed that with *suitable regularization*, the posterior distribution for Bayesian CART and BART actually concentrates at this fast rate (up to a log factor), adapting to the intrinsic dimensionality and smoothness. Later in Section 4, we continue their theoretical investigation and focus on consistent *variable selection* that is estimation of $\mathcal{S}_0$ rather than $f_0(\cdot)$. Spike-and-slab regularization plays a key role in obtaining these theoretical guarantees.

2.1 | Trees with spike-and-slab regularization

Many applications offer a plethora of predictors and some form of redundancy penalization has to be incurred to cope with the curse of dimensionality. Bayesian regression trees were originally conceived for prediction rather than variable selection. Indeed, original tree implementations of Bayesian CART (Chipman et al., 1998; Denison et al., 1998) do not seem to penalize inclusion of redundant variables aggressively enough. As noted by Linero (2018), the prior expected number of active variables under the Bayesian CART prior of Chipman et al. (1998) satisfies $\lim_{p\to\infty}\mathbb{E}[q] = K - 1$ as $p \to \infty$ where K is the fixed number of bottom leaves. This behaviour suggests that (in the limit) the prior forces inclusion of the maximal number of variables while splitting on them only once. This is far from ideal. To alleviate this issue, we deploy the so-called *spike-and-forest priors* that is spike-and-slab wrappers around sum-of-trees priors (Ročková & van der Pas, 2020). As with the traditional spike-and-slab priors, the specification starts with a prior distribution over the 2^p active variable sets:

$$\mathcal{S} \sim \pi(\mathcal{S}) \quad \text{for each} \quad \mathcal{S} \subseteq \{1, \ldots, p\}. \tag{2}$$

We elaborate on the specific choices of $\pi(\mathcal{S})$ later in Sections 3.2 and 4.

Given the pool of variables $\mathcal{S}$, a regression tree/forest is grown using *only* variables inside $\mathcal{S}$. This prevents the trees from using too many variables and thereby from overfitting. Recall that each individual regression tree is characterized by two components: (1) a tree-shaped K-partition of $[0,1]^p$, denoted with $\mathcal{T}$, and (2) bottom node parameters (step heights), denoted with $\boldsymbol{\beta} \in \mathbb{R}^K$. Starting with a parent node $[0,1]^p$, each K-partition is grown by recursively dissecting rectangular cells at chosen internal nodes along one of the active coordinate axes, all the way down to K terminal nodes. Each tree-shaped K-partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ consists of K partitioning rectangles $\Omega_k \subset [0,1]^p$.

While Bayesian CART approximates $f_0(\boldsymbol{x})$ with a single tree mappings $f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k$, Bayesian additive regression trees (BART) use an aggregate of T mappings

$$f_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x})$$

where $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ is an ensemble of tree partitions and $\boldsymbol{B} = [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^T]$ is an ensemble of step coefficients. In a fully Bayesian approach, prior distributions have to be specified over the set of tree structures $\mathcal{E}$ and over terminal node heights $\boldsymbol{B}$. The spike-and-forest (SF) construction can accommodate various tree prior options.

To assign a prior over $\mathcal{E}$ for a given T, one possibility is to first pick the number of bottom nodes, independently for each tree, from a prior

$$K^t \sim \pi(K) \quad \text{for} \quad K = 1, \ldots, n, \tag{3}$$

such as the Poisson distribution (Denison et al., 1998). Given the vector of tree sizes $\mathbf{K} = (K^1, \ldots, K^T)'$ and a set of covariates $\mathcal{S}$, we assign a prior over so-called valid ensembles/forests $\mathcal{VE}_{\mathcal{S}}^{\mathbf{K}}$. We say that a tree ensemble $\mathcal{E}$ is valid if it consists of trees that have non-empty bottom leaves. One can pick a tree partition ensemble from a uniform prior over *valid* forests $\mathcal{E} \in \mathcal{VE}_{\mathcal{S}}^{\mathbf{K}}$, i.e.

$$\pi(\mathcal{E} \mid \mathcal{S}, \mathbf{K}) = \frac{1}{\Delta(\mathcal{VE}_{\mathcal{S}}^{\mathbf{K}})} \mathbb{I}\left(\mathcal{E} \in \mathcal{VE}_{\mathcal{S}}^{\mathbf{K}}\right), \tag{4}$$

where $\Delta(\mathcal{VE}_{\mathcal{S}}^{\mathbf{K}})$ is the number of valid tree ensembles characterized by $\mathbf{K}$ bottom leaves and split directions $\mathcal{S}$. The prior (3) and prior (4) were deployed in the Bayesian CART implementation of Denison et al. (1998) (with $T = 1$) and it was studied theoretically by Ročková and van der Pas (2020). Another related Bayesian forest prior (implemented in the BART procedure and studied theoretically by Ročková and Saha (2019)) consists of an independent product of branching process priors (one for each tree) with decaying split probabilities (Chipman et al., 1998). The implementation is very similar to the one of Denison et al. (1998).

Finally, given the partitions $\mathcal{T}^t$ of size K^t for $1 \leq t \leq T$, one assigns (independently for each tree) a Gaussian product prior on the step heights

$$\pi(\boldsymbol{\beta}^t \mid K^t) = \prod_{k=1}^{K^t} \phi(\beta_k^t; \sigma_{\beta}^2), \tag{5}$$

where $\phi(x;\sigma_\beta^2)$ denotes a Gaussian density with mean zero and variance $\sigma_\beta^2 = 1/T$ (as suggested by Chipman et al. 2010). The prior for σ^2 can be chosen as inverse chi-squared with hyperparameters chosen based on an estimate of the residual standard deviation of the data (Chipman et al., 2010).

The most crucial component in the SF construction, which sets it apart from existing BART implementations, is the active set $\mathcal{S}$ which serves to mute variables by restricting the pool of predictors available for splits. The goal is to learn which set $\mathcal{S}$ is most likely (a posteriori) and/or how likely each variables is to have contributed to f_0. Unlike related tree-based variable selection criteria, the spike-and-slab envelope makes it possible to perform variable selection directly by evaluating posterior model probabilities $\Pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$ or marginal inclusion probabilities $\Pi(j \in \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)})$ for $1 \leq j \leq p$. RF (Breiman, 2001) also mute variables, but they do so from within the tree by randomly choosing a small subset of variables for each split. The spike-and-slab approach mutes variables externally rather than internally. Bleich et al. (2014) note that when the number of trees is small, the Gibbs sampler for BART can get trapped in local modes which can destabilize the estimation procedure. On the other hand, when the number of trees is large, there are ample opportunities for the noise variables to enter the model without necessarily impacting the model fit, making variable selection very challenging. Our spike-and-slab wrapper is devised to get around this problem.

The problem of variable selection is fundamentally challenged by the sheer size of possible variable subsets. For linear regression, (a) MCMC implementations exist that capitalize on the availability of marginal likelihood (Guan & Stephens, 2011; Narisetty & He, 2014), (b) optimization strategies exist for both continuous (Ročková, 2017; Ročková & George, 2018) and point-mass spike-and slab priors (Carbonetto & Stephens, 2012). These techniques do not directly translate to tree models, for which tractable marginal likelihoods $\pi(\boldsymbol{Y}^{(n)} \mid \mathcal{S})$ are unavailable. To address this computational challenge, we explore ABC techniques as a new promising avenue for non-parametric spike-and-slab methods.

3 | ABC FOR VARIABLE SELECTION

Performing (approximate) posterior inference in complex models is often complicated by the analytical intractability of the marginal likelihood. ABC is a simulation-based inference framework that obviates the need to compute the likelihood directly by evaluating the proximity of (sufficient statistics of) observed data and pseudo-data simulated from the likelihood. Simon Tavaré first proposed the ABC algorithm for posterior inference (Tavaré et al., 1997) in the 1990's and since then it has widely been used in population genetics, systems biology, epidemiology and phylogeography[1].

Combined with a probabilistic structure over models, marginal likelihoods give rise to posterior model probabilities, a standard tool for Bayesian model choice. When the marginal likelihood is unavailable (our case here), ABC offers a unique computational solution. However, as pointed out by Robert et al. (2011), ABC cannot be trusted for model comparisons when model-wise sufficient summary statistics are not sufficient across models. The ABC approximation to Bayes factors then does not converge to exact Bayes factors, rendering ABC model choice fundamentally untrustworthy. A fresh new perspective to ABC model choice was offered in Pudlo et al. (2015), who rephrase model selection as a classification problem that can be tackled with machine learning tools. Their idea is to treat the ABC reference table (consisting of samples from a prior model distribution and high-dimensional vectors of summary statistics of pseudo-data obtained from the prior predictive distribution) as an actual data set, and to train a RF classifier that predicts a model label using the summary statistics as predictors. Their goal is to produce a stable model decision based on a classifier rather than on an estimate of posterior model probabilities. Our approach has a similar flavour in the sense that it combines machine learning with ABC, but the concept is fundamentally very different. Here, the fusion of Bayesian forests and ABC is tailored to non-parametric variable selection towards obtaining posterior variable inclusion probabilities. Our model selection approach does not suffer from the difficulty of ABC model choice as we *do not* commit to any summary statistics and use random subsets of observations to generate the ABC reference table.

[1] The study of how human beings migrated throughout the world in the past.

3.1 | Naive ABC implementation

For its practical implementation, our Bayesian variable selection method requires sampling from the analytically intractable posterior distribution over subsets $\Pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$ under the *spike-and-forest* prior (4), (3) and (2). Given a single tree partition $\mathcal{T}$, the (conditional) marginal likelihood $\pi(\boldsymbol{Y}^{(n)} \mid \mathcal{T}, \mathcal{S})$ is available in closed form, facilitating implementations of Metropolis-Hastings algorithms (Chipman et al., 1998; Denison et al., 1998) (see Section S.3). However, such MCMC schemes can suffer from poor mixing. Taking advantage of the fact that, despite being intractable, one can *simulate from* the marginal likelihood $\pi(\boldsymbol{Y}^{(n)} \mid \mathcal{S})$, we will explore the potential of ABC as a complementary development to MCMC implementations.

The principle at the core of ABC is to perform approximate posterior inference from a given dataset by simulating from a prior distribution and by comparisons with numerous synthetic datasets. In its standard form, an ABC implementation of model choice creates a reference table, recording a large number of datasets simulated from the model prior and the prior predictive distribution under each model. Here, the table consists of M pairs $(\mathcal{S}_m, \boldsymbol{Y}_m^\star)$ of model indices $\mathcal{S}_m$, simulated from the prior $\pi(\mathcal{S})$, and pseudo-data $\boldsymbol{Y}_m^\star \in \mathbb{R}^n$, simulated from the marginal likelihood $\pi(\boldsymbol{Y}^{(n)} \mid \mathcal{S}_m)$. To generate $\boldsymbol{Y}_m^\star$ in our setup, one can hierarchically decompose the marginal likelihood

$$\pi(\boldsymbol{Y}^{(n)} \mid \mathcal{S}) = \int_{(f_{\mathcal{E},\boldsymbol{B}},\sigma^2)} \pi(\boldsymbol{Y}^{(n)} \mid f_{\mathcal{E},\boldsymbol{B}}, \sigma^2) \mathrm{d}\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}) \tag{6}$$

and first draw $(f^m_{\mathcal{E},\boldsymbol{B}}, \sigma^2_m)$ from the prior $\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S})$ and obtain $\boldsymbol{Y}_m^\star$ from (1), given $(f^m_{\mathcal{E},\boldsymbol{B}}, \sigma^2_m)$. ABC sampling is then followed by an ABC rejection step, which extracts pairs $(\mathcal{S}_m, \boldsymbol{Y}_m^\star)$ such that $\boldsymbol{Y}_m^\star$ is close enough to the actual observed data. In other words, one trims the reference table by keeping only model indices $\mathcal{S}_m$ paired with pseudo-observations that are at most ε-away from the observed data, i.e. $\| \boldsymbol{Y}^{obs} - \boldsymbol{Y}_m^\star \|_2 \leq \epsilon$ for some tolerance level ε. These extracted values comprise an approximate ABC sample from the posterior $\pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$, which should be informative for the relative ordering of the competing models, and thus variable selection (Grelaud et al., 2009). Note that this particular ABC implementation does not require any use of low-dimensional summary statistics, where rejection is based solely on $\boldsymbol{Y}^{obs}$. While theoretically justified, this ABC variant has two main drawbacks.

First, with very many predictors, it will be virtually impossible to sample from all 2^p model combinations at least once, unless the reference table is huge. Consequently, relative frequencies of occurrence of a model $\mathcal{S}_m$ in the trimmed ABC reference table *may not* be a good estimate of the posterior model probability $\pi(\mathcal{S}_m \mid \boldsymbol{Y}^{(n)})$. While the model with the highest posterior probability $\pi(\mathcal{S}_m \mid \boldsymbol{Y}^{(n)})$ is commonly conceived as the right model choice, it may not be the optimal model for prediction. Indeed, in nested correlated designs and orthogonal designs, it is the median probability model that is predictive optimal (Barbieri & Berger, 2004). The median probability model (MPM) consists of those variables whose *marginal* inclusion probabilities $\mathbb{P}(j \in \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)})$ are at least 0.5. While simulation-based estimates of posterior model probabilities $\mathbb{P}(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$ can be imprecise, we argue (and show) that ABC estimates of marginal inclusion probabilities $\mathbb{P}(j \in \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)})$ are far more robust and stable.

The second difficulty is purely computational and relates to the issue of coming up with good proposals $f^m_{\mathcal{E},\boldsymbol{B}}$ such that the pseudo-data are sufficiently close to $\boldsymbol{Y}^{obs}$. Due to the vastness of the tree ensemble space, it would be naive to think that one can obtain solid guesses of f_0 purely by sampling from non-informative priors. This is why we call this ABC implementation naive. These considerations lead us to a new data-splitting ABC modification that uses a random portion of the data to train the prior and to generate pseudo-data with more affinity to the left-out observations.

3.2 | ABC Bayesian forests

By sampling directly from non-informative priors over tree ensembles $\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S})$, the acceptance rate of the naive ABC can be prohibitively small where huge reference tables would be required to obtain only a few approximate samples from the posterior.

To address this problem, we suggest a sample-splitting approach to come up with draws that are less likely to be rejected by the ABC method. At each ABC iteration, we first draw a random subsample $\mathcal{I} \subset \{1, \ldots, n\}$ of size $|\mathcal{I}| = s$ with no replacement. Then we split the observed data $\boldsymbol{Y}^{(n)}$ into two groups, denoted with $\boldsymbol{Y}^{(n)}_{\mathcal{I}}$ and $\boldsymbol{Y}^{(n)}_{\mathcal{I}^c}$, and instead of (6) we consider the marginal likelihood conditionally on $\boldsymbol{Y}^{(n)}_{\mathcal{I}}$

$$\pi(\boldsymbol{Y}^{(n)} \mid \boldsymbol{Y}^{(n)}_{\mathcal{I}}, \mathcal{S}) = \int_{(f_{\mathcal{E},\boldsymbol{B}},\sigma^2)} \pi(\boldsymbol{Y}^{(n)}_{\mathcal{I}^c} | f_{\mathcal{E},\boldsymbol{B}}, \sigma^2) \mathrm{d}\pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}) \tag{7}$$

where

$$\pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}) = \pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \boldsymbol{Y}^{(n)}_{\mathcal{I}}, \mathcal{S}). \tag{8}$$

This simple decomposition unfolds new directions for ABC sampling based on data splitting. Instead of using all observations $\boldsymbol{Y}^{obs}$ to accept/reject each draw, we set aside a random subset of data $\boldsymbol{Y}^{obs}_{\mathcal{I}^c}$ for ABC rejection and use $\boldsymbol{Y}^{obs}_{\mathcal{I}}$ to 'train the prior'. The key observation is that the samples from the prior $\pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S})$, that is the *posterior*$\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \boldsymbol{Y}^{(n)}_{\mathcal{I}}, \mathcal{S})$, will have seen a part of the data and will produce more realistic guesses of f_0. Such guesses are more likely to yield pseudo-data that match $\boldsymbol{Y}^{obs}_{\mathcal{I}^c}$ more closely, thereby increasing the acceptance rate of ABC sampling. Note that the acceptance step is based solely on the left-out sample $\boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}$, not the entire data. Similarly as the naive ABC outlined in the previous section, we first sample the subset $\mathcal{S}$ from the prior $\pi(\mathcal{S})$ and then obtain draws from the conditional marginal likelihood under an updated prior $\pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S})$. This corresponds to an ABC strategy for sampling from $\pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)}_{\mathcal{I}^c})$ under the priors (2) and (8). As will be seen later, this posterior is effective for assessing variable importance. Moreover, if $\pi(\mathcal{S})$ is a good proxy for $\pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)}_{\mathcal{I}})$ (when the training set is small relative to the ABC rejection set), this ABC will produce approximate samples from the original target $\pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$.

The idea of using a portion of the data for training the prior and the rest for model selection goes back to at least Good (1950). The most common prescription for choosing training samples in Bayesian analysis is to convert improper priors into propers ones for meaningful model selection with Bayes factors (Lempers, 1971; O'Hagan, 1995). Berger and Pericchi (1996) advocated choosing the training set as small as possible subject to yielding proper posteriors (so called minimal training samples). Berger and Pericchi (2004) argue that data can vary widely in terms of their information content and the use of single minimal training samples can be inadequate/suboptimal. Since there are many possible training samples, it is natural to average the resulting Bayes factors over the training samples in some fashion. While intrinsic Bayes factors (Berger & Pericchi, 1996) average Bayes factors over all possible minimal training samples, expected posterior priors (Pérez & Berger, 2002) average the prior first. In particular, the empirical expected-posterior prior for model $\mathcal{S}$ (Ghosh & Samanta, 2002; Pérez & Berger, 2002) writes as

$$\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}) = \frac{1}{L} \sum_{l=1}^{L} \pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}), \tag{9}$$

where $\pi_{\mathcal{I}}(f_{\mathcal{E},\boldsymbol{B}},\sigma^2\,|\,\mathcal{S})$ was defined in (8) and where L is the number of all minimal training samples $\mathcal{I}_l$. The marginal likelihood under this prior can be then written as (equation (3.5) in Pérez and Berger 2002) $m(\boldsymbol{Y}^{(n)}\,|\,\mathcal{S})=\frac{1}{L}\sum_{l=1}^{L}\pi(\boldsymbol{Y}^{(n)}\,|\,\boldsymbol{Y}^{(n)}_{\mathcal{I}},\mathcal{S})$, where $\pi(\boldsymbol{Y}^{(n)}\,|\,\boldsymbol{Y}^{(n)}_{\mathcal{I}},\mathcal{S})$ was defined in (7). Our ABC analysis with internal data splitting can be thus regarded as arising from the empirical expected posterior prior (9). While the motivation for using training samples in Bayesian analysis has been largely to make improper priors proper, here we use this idea in a different context to increase ABC acceptance rate.

The ABC Bayesian forests algorithm is formally summarized in Table 1. It starts by splitting the dataset into two subsets at each (m^{th}) iteration: $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$ for fitting and $\boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}$ for ABC rejection. The algorithm then proceeds by sampling an active set $\mathcal{S}$ from $\pi(\mathcal{S})$. Using the spike-and-slab construction, one can draw Bernoulli indicators $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_p)'$ where $\mathbb{P}(\gamma_j=1\,|\,\theta)=\theta$ for some prior inclusion probability $\theta\in(0,1)$ and set $\mathcal{S}_m=\{j:\gamma_j=1\}$. When sparsity is anticipated, one can choose θ to be small or to arise from a beta prior $\mathcal{B}(a,b)$ for some $a>0$ and $b>0$ (yielding the beta-binomial prior). We discuss other suitable prior model choices in Section 4.

In the *(c) step* of ABC Bayesian forests, one obtains a sample from the posterior of $(f_{\mathcal{E},\boldsymbol{B}},\sigma^2)$, given $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$. For this step, one can leverage existing implementations of Bayesian CART and BART (e.g. the `BART` R package of McCulloch et al. (2018)). A single draw from the posterior is obtained after a sufficient burn-in. In this vein, one can view ABC Bayesian forests as a computational envelope around BART to restrict the pool of available variables. The *(d) step* then consists of predicting the outcome $\boldsymbol{Y}^{\star}_{\mathcal{I}^c_m}$ for left-out observations $\boldsymbol{x}_i$ using (1) for each $i\in\mathcal{I}^c_m$. The last step is ABC rejection based on the discrepancy between $\boldsymbol{Y}^{\star}_{\mathcal{I}^c_m}$ and $\boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}$.

```
Algorithm 1 : ABC Bayesian Forests
Data: Data $(Y_i^{obs}, \boldsymbol{x}_i)$ for $1 \le i \le n$
Result: $\pi_j(\epsilon)$ for $1 \le j \le p$ where $\pi_j(\epsilon) = \widehat{\mathbb{P}}(j \in \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)})$
Set $M$: the number of ABC simulations; $s$: the subsample size; $\epsilon$: the tolerance threshold; $m = 0$
  the counter
  while $m \le M$ do
    (a) Split data $\boldsymbol{Y}^{obs}$ into $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$ and $\boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}$, where $\mathcal{I}_m \subset \{1,\ldots,n\}$ of size $|\mathcal{I}_m| = s$ is obtained
    by sampling with no replacement.
    (b) Pick a subset $\mathcal{S}_m$ from $\pi(\mathcal{S})$.
    (c) Sample $(f^m_{\mathcal{E},\boldsymbol{B}}, \sigma^2_m)$ from $\pi_{\mathcal{I}_m}(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \mathcal{S}_m) = \pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \mid \boldsymbol{Y}^{obs}_{\mathcal{I}_m}, \mathcal{S}_m)$.
    (d) Generate pseudo-data $\boldsymbol{Y}^{\star}_{\mathcal{I}^c_m}$ by sampling white noise $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_m)$ and setting
    $Y^{\star}_i = f^m_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}_i) + \varepsilon_i$ for each $i \notin I_m$.
    (e) Compute discrepancy $\epsilon_m = \|\boldsymbol{Y}^{\star}_{\mathcal{I}^c_m} - \boldsymbol{Y}^{obs}_{\mathcal{I}^c_m}\|_2$.
    if $\epsilon_m < \epsilon$ then
      Accept $(\mathcal{S}_m, f^m_{\mathcal{E},\boldsymbol{B}})$ and set $m = m + 1$
    else
      Reject $(\mathcal{S}_m, f^m_{\mathcal{E},\boldsymbol{B}})$ and set $m = m + 1$
    end
end
Compute $\pi_j(\epsilon)$ as the proportion of times $j^{th}$ variable is used in the accepted $f^m_{\mathcal{E},\boldsymbol{B}}$'s.
```

For the computation of marginal inclusion probabilities $\pi_j(\epsilon)$, one could conceivably report the proportion of ABC accepted samples $\mathcal{S}_m$ that contain the j^{th} variable. However, $\mathcal{S}_m$ is a pool of *available* predictors and not all of them are necessarily used in $f^m_{\mathcal{E},\boldsymbol{B}}$. Thereby, we report the proportion of ABC accepted samples $f^m_{\mathcal{E},\boldsymbol{B}}$ that use the j^{th} variable at least once, that is

TABLE 1 Average out-of-sample mean squared prediction error over 20 independent validation datasets. ABC1 denotes predictions using ABC samples $f^m_{\mathcal{S},\boldsymbol{B}}$ and ABC2 uses ABC variable selection and runs BART ($T = 200$) on the selected subset. T designates the number of trees and c is the selection threshold. The best performing method for each row is denoted in bold

	ABC2	**ABC1**	**ABC1**	**ABC2**	**ABC1**	**ABC1**	**RF**	**RLT**	**DT**	**BART**	**DART**
	$\boldsymbol{T = 20}$	$\boldsymbol{T = 20,}$ $\boldsymbol{c = 0.5}$	$\boldsymbol{T = 20,}$ $\boldsymbol{c = 0.25}$	$\boldsymbol{T = 10}$	$\boldsymbol{T = 10,}$ $\boldsymbol{c = 0.5}$	$\boldsymbol{T = 10,}$ $\boldsymbol{c = 0.25}$					
	Equi-correlation$\rho_{ij} = 0.5$ **for** $i \neq j$										
	Linear										
$p = 100$	5.56	5.58	5.84	5.60	5.84	5.55	5.63	5.45	5.92	5.49	**5.40**
$p = 1000$	5.79	6.15	5.73	5.86	6.28	5.95	5.83	5.70	6.04	5.82	**5.62**
	CART										
$p = 100$	34.21	34.63	37.19	**34.00**	36.10	35.81	34.21	34.64	34.61	35.48	35.57
$p = 1000$	32.00	34.27	35.72	**31.99**	33.93	33.17	32.30	32.40	33.08	33.77	34.04
	Friedman										
$p = 100$	30.32	29.28	31.59	30.52	30.30	**29.03**	31.84	30.17	41.41	31.31	**29.03**
$p = 1000$	33.14	35.97	31.54	33.54	38.42	32.71	34.35	32.22	45.69	32.99	**29.42**
	LLS										
$p = 100$	**26.23**	27.00	28.70	26.25	26.90	27.36	26.80	26.46	28.51	27.42	27.42
$p = 1000$	27.37	26.98	**26.94**	27.38	27.07	27.02	27.18	26.68	30.66	28.21	27.49
	Auto-correlation$\rho_{ij} = 0.9^{\lvert i-j \rvert}$										
	Linear										
$p = 100$	6.17	6.29	6.37	6.20	6.25	6.18	6.37	6.09	6.77	6.17	**5.91**
$p = 1000$	6.39	6.44	**6.00**	6.47	6.21	6.13	6.55	6.20	7.06	6.53	6.42
	CART										
$p = 100$	33.80	37.72	37.28	33.83	36.78	36.61	**33.57**	34.40	35.05	35.61	35.81

(Continues)

TABLE 1 (Continued)

	ABC2	**ABC1**	**ABC1**	**ABC2**	**ABC1**	**ABC1**	**RF**	**RLT**	**DT**	**BART**	**DART**
	$\boldsymbol{T = 20}$	$\boldsymbol{T = 20, c = 0.5}$	$\boldsymbol{T = 20, c = 0.25}$	$\boldsymbol{T = 10}$	$\boldsymbol{T = 10, c = 0.5}$	$\boldsymbol{T = 10, c = 0.25}$					
$p = 1000$	31.57	33.55	37.21	**31.52**	33.52	37.43	31.63	31.88	32.22	33.11	33.43
	Friedman										
$p = 100$	34.09	32.51	34.65	34.27	34.97	32.77	36.88	33.83	48.64	34.21	**30.36**
$p = 1000$	39.09	39.57	32.58	40.58	43.05	33.46	41.80	37.38	49.51	35.96	**30.81**
	LLS										
$p = 100$	28.57	**27.94**	30.71	28.45	28.03	29.12	28.88	27.87	30.69	28.83	28.81
$p = 1000$	29.98	**28.25**	28.96	30.14	28.40	28.38	30.19	28.56	32.29	31.76	29.28

$$\pi_j(\epsilon) = \frac{1}{M(\epsilon)} \sum_{m:\epsilon_m < \epsilon} \mathbb{I}(j \text{ used in } f^m_{\mathcal{E},\boldsymbol{B}}), \tag{10}$$

where $M(\epsilon)$ is the number of accepted ABC samples at ϵ. Each tree ensemble $f^m_{\mathcal{E},\boldsymbol{B}}$ thus performs its own variable selection by picking variables from $\mathcal{S}_m$ rather than from $\{1, \dots, p\}$. Limiting the pool of predictors prevents from too many false positives. In addition, the inclusion probabilities (10) do use the training data $\boldsymbol{Y}^{(n)}_{\mathcal{I}}$ to shrink and update the subset $\mathcal{S}$ by leaving out covariates not picked by $f^m_{\mathcal{E},\boldsymbol{B}}$. In this way, the mechanism for selecting the subsets $\mathcal{S}$ is not strictly sampling from the prior $\pi(\mathcal{S})$ but it seizes the information in the training set $\mathcal{I}$. In this way, $\mathcal{S}_m$'s can be regarded as approximate samples from $\pi(\mathcal{S} \mid \boldsymbol{Y}^{obs})$. When $\mathcal{I} = \emptyset$, we recover the naive ABC as a special case.

3.2.1 | Dynamic ABC

The estimates of marginal inclusion probabilities $\pi_j(\epsilon)$ obtained with ABC Bayesian forests unavoidably depend on the level of approximation accuracy ε. The acceptance threshold ε can be difficult to determine in practice, because it has to accommodate random variation of data around f_0 as well as the error when approximating smooth surfaces f_0 with trees. As $\varepsilon \to 0$, the approximations $\pi_j(\epsilon)$ will be more accurate, but the acceptance rate will be smaller. It is customary to pick ε as an empirical quantile of ϵ_m (Grelaud et al., 2009), keeping only the top few closest samples. Rather than choosing one value ε, we suggest a dynamic strategy by considering a sequence of decreasing values $\epsilon_N > \epsilon_{N-1} > \dots > \epsilon_1 > 0$. By filtering out the ABC samples with stricter thresholds, we track the evolution of each $\pi_j(\epsilon)$ as ε gets smaller and smaller. This gives us a dynamic plot that is similar in spirit to the Spike-and-Slab LASSO (Ročková & George, 2018) or EMVS (Ročková & George, 2014) coefficient evolution plots. However, our plots depict approximations to posterior inclusion probabilities rather than coefficient magnitudes. Other strategies for selecting the threshold ε are discussed in (Csillery et al., 2010; Marin et al., 2012; Sunnaaker et al., 2013).

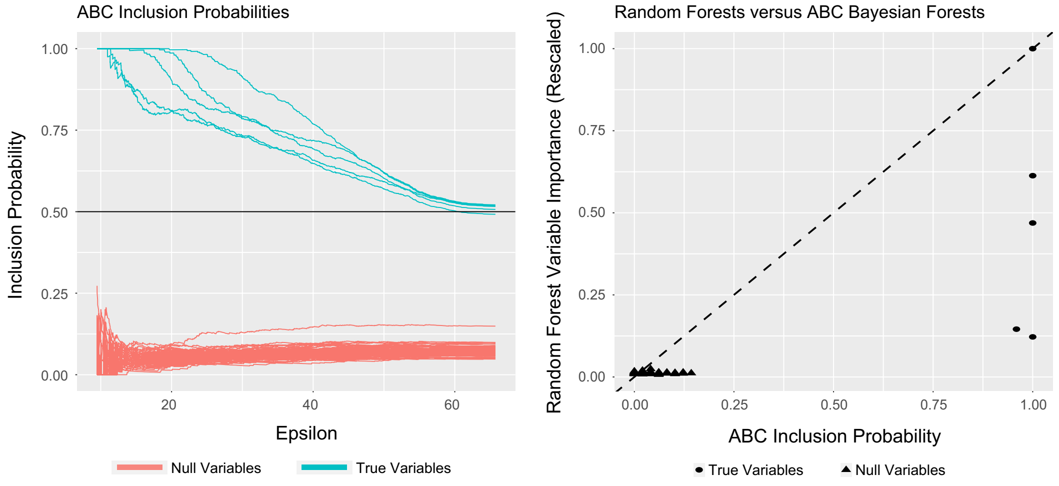
3.3 | ABC Bayesian forests in action

We demonstrate the usefulness of ABC Bayesian forests on the benchmark Friedman dataset (Friedman, 1991), where the observations are generated from (1) with $\sigma = 1$ and

$$f_0(\boldsymbol{x}_i) = 10 \sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5}, \tag{11}$$

where $x_i \in [0, 1]^p$ are *iid* from a uniform distribution on a unit cube. Because the outcome depends on $x_1, \dots, x_p$, the predictors $x_6, \dots, x_p$ are irrelevant, making it more challenging to find $f_0(\boldsymbol{x})$. We begin by illustrating the basic features of ABC Bayesian forests with $p = 100$ and $n = 500$, assuming the beta-binomial prior $\pi(\mathcal{S} \mid \theta)$ with $\theta \sim \mathcal{B}(1, 1)$ (see Section 3.2). At the m^{th} ABC iteration, we draw one posterior sample $f^m_{\mathcal{E},\boldsymbol{B}}$ after 100 burn-in iterations using the BART MCMC algorithm (Chipman et al., 2001) with $T = 10$ trees. We generate $M = 1000$ ABC samples (with $s = n/2$) and we keep track of variables used in $f^m_{\mathcal{E},\boldsymbol{B}}$'s to estimate the marginal posterior inclusion probabilities $\pi_j(\epsilon)$. It is worth pointing out that unlike MCMC, ABC Bayesian forests are embarrassingly parallel, making distributed implementations readily available.

Following the dynamic ABC strategy, we plot the estimates of posterior inclusion indicators $\pi_j(\epsilon)$ as a function of ε (Figure 1). The true signals are depicted in blue, while the noise covariates are in red. The estimated inclusion probabilities clearly segregate the active and non-active variables, even



(a) $T = 10$

(b) Random Forests versus Bayesian Forests

FIGURE 1 (Left) Dynamic ABC plots for evolving inclusion probabilities as ε gets smaller. (Right) Plot of $\pi_j(\epsilon)$ obtained with ABC Bayesian forests (ε is the 5% quantile of ϵ_m's) and the variable importance measure from random forests (rescaled to have a maximum at 1)

for large ε values. This is because BART itself performs variable selection to some degree, where not all variables in $\mathcal{S}_m$ end up contributing to $f^m_{\varepsilon,\boldsymbol{B}}$. For small enough ε, the inclusion probabilities of true signals eventually cross the 0.5 threshold. Based on the median probability model rule (Barbieri & Berger, 2004), one thereby selects the true model when ε is sufficiently small. Because the inclusion probabilities get a bit unstable as ε gets smaller (they are obtained from smaller reference tables), we excluded the 10 smallest ε values from the plot.

We repeated the experiment with more trees ($T = 50$) and a single tree ($T = 1$). Using more trees, one still gets the separation between signal and noise. However, many more noisy covariates would be included by the MPM rule. This is in accordance with Chipman et al. (2001) who state that BART can over-select with many trees. With a single tree, on the other hand, one may miss some of the low-signal predictors, where deeper trees and more ABC iterations would be needed to obtain a clearer separation.

In this simulation, we observe a curious empirical connection between $\pi_j(\epsilon)$, obtained with ABC Bayesian forests (taking top 5% ABC samples), and rescaled variable importance obtained with RF. From Figure 1b, we see that the two measures largely agree, separating the signal coefficients (triangles) from the noise coefficients (dots). However, the RF measure is a bit more conservative, yielding smaller normalized importance scores for true signals. While variable importance for RF is yet not understood theoretically, in the next section we provide conditions under which the posterior distribution is consistent for variable selection.

4 | MODEL-FREE VARIABLE SELECTION CONSISTENCY

In this section, we develop large sample model selection theory for SF priors. As a jumping-off point, we first assume that α (the regularity of f_0) is known, where model selection essentially boils down to finding the active set $\mathcal{S}_0$. Later in this section, we investigate *joint* model selection consistency, acknowledging uncertainty about $\mathcal{S}_0$ and, at the same time, the regularity α.

Several consistency results for non-parametric regression already exist (Yang & Pati, 2017; Zhu et al., 2015). Comminges and Dalalyan (2012) characterized tight conditions on (n, p, q_0), under which it is possible to consistently estimate the sparsity pattern in two regimes. For fixed q_0, consistency is attainable when $(\log p)/n \le c$ for some $c > 0$. When q_0 tends to infinity as $n \to \infty$, consistency is achievable when $c_1 q_0 + \log\log(p/q_0) - \log n \le c_2$ for some $c_1, c_2 > 0$. Throughout this section, we will treat q_0 as fixed and show variable selection consistency when $q_0 \log p \le n^{q_0/(2\alpha+q_0)}$. As an overture to our main result, we start with a simpler case when $T = 1$ (a single tree) and when α is known. The full-fledged result for Bayesian forests and unknown α is presented in Section 4.3. Throughout this section, we will assume $\sigma^2 = 1$.

4.1 | The case of known α

Spike-and-forest mixture priors are constructed in two steps by (1) first specifying a conditional prior $\Pi_{\mathcal{S}}(f)$ on tree (ensemble) functions expressing a qualitative guess on f_0, and then (2) attaching a prior weight $\pi(\mathcal{S})$ to each 'model' (i.e. subset) $\mathcal{S}$. The posterior distribution $\Pi(f \mid \boldsymbol{Y}^{(n)})$ can be viewed as a mixture of individual posteriors for various models $\mathcal{S}$ with weights given by posterior model probabilities $\Pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)})$, that is

$$\Pi(f \mid \boldsymbol{Y}^{(n)}) = \sum_{\mathcal{S}} \Pi(\mathcal{S} \mid \boldsymbol{Y}^{(n)}) \Pi_{\mathcal{S}}(f \mid \boldsymbol{Y}^{(n)}).$$

Our aim is to establish 'model-free variable selection consistency in the sense that

$$\Pi(\mathcal{S} = \mathcal{S}_0 \mid \boldsymbol{Y}^{(n)}) \to 1 \quad \text{in } \mathbb{P}_{f_0}^{(n)} - \text{probability} \quad \text{as} \quad n \to \infty,$$

where $\mathbb{P}_{f_0}^{(n)}$ is the distribution of $\boldsymbol{Y}^{(n)}$ under (1). The adjective 'model-free' merely refers to the fact that we are selecting subsets in a non-parametric regression environment without necessarily committing to a linear model. We start by defining the model index set $\Gamma = \{\mathcal{S} : \mathcal{S} \subseteq \{1, \ldots, p\}\}$, consisting of all 2^p variable subsets, and we partition it into (a) the true model $\mathcal{S}_0$, (b) models that *overfit* $\Gamma_{\mathcal{S} \supset \mathcal{S}_0}$ (i.e. supersets of the true subset $\mathcal{S}_0$) and (c) models that *underfit* $\Gamma_{\mathcal{S} \not\supset \mathcal{S}_0}$ (i.e. models that miss at least one active covariate). Each model $\mathcal{S} \in \Gamma$ is accompanied by a convergence rate $\varepsilon_{n,\mathcal{S}}$ that reflects the inherent difficulty of the estimation problem. For each model $\mathcal{S}$ of size $|\mathcal{S}|$, we define

$$\varepsilon_{n,\mathcal{S}} = C_\varepsilon n^{-\alpha/(2\alpha + |\mathcal{S}|)} \sqrt{\log n} \quad \text{for some} \quad C_\varepsilon > 0, \tag{12}$$

the $\|\cdot\|_n$-near-minimax rate of estimation of a $|\mathcal{S}|$-dimensional α-smooth function.

4.1.1 | Prior specification

Prior distribution on the model index $\Pi(\mathcal{S})$ has to be chosen carefully for model selection consistency to hold when $p > n$ (Moreno et al., 2015). Traditional spike-and-slab priors introduce $\Pi(\mathcal{S})$ through a prior inclusion probability $\theta = \Pi(i \in \mathcal{S}_0 \mid \theta)$, independently for each $i = 1, \ldots, p$. This prior mixing weight is often endowed with a prior, such as the uniform prior $\pi(\theta) = \mathcal{B}(1, 1)$ (Scott & Berger, 2010), yielding a uniform prior on the model size, or the 'complexity prior $\pi(\theta) = \mathcal{B}(1, p^c)$ for $c > 2$ (Castillo & van der Vaart, 2012), yielding an exponentially decaying prior on the model size. We propose a different approach, directly assigning a prior on model weights through

$$\pi(\mathcal{S}) \propto \mathrm{e}^{-C\left(n^{|\mathcal{S}|/(2\alpha+|\mathcal{S}|)}\log n \vee |\mathcal{S}|\log p\right)} \tag{13}$$

where $C > 0$ is a suitably large constant. When $|\mathcal{S}|\log p \leq n^{|\mathcal{S}|/(2\alpha+|\mathcal{S}|)}$, this prior is proportional to $\mathrm{e}^{-C/C_\varepsilon^2 n\varepsilon_{n,\mathcal{S}}^2}$ and, as such, it puts more mass on models that yield faster rates convergence (similarly as in Lember and van der Vaart (2007)). When $|\mathcal{S}|\log p > n^{|\mathcal{S}|/(2\alpha+|\mathcal{S}|)}\log n$, the implied prior on the effective dimensionality $\pi(|\mathcal{S}|) = \binom{p}{|\mathcal{S}|}\pi(\mathcal{S})$ will be exponentially decaying in the sense that $\pi(|\mathcal{S}|) \lesssim \mathrm{e}^{-(C-1)|\mathcal{S}|\log p}$ for $C > 1$. It was recently noted by Castillo and Mismer (2018) that the complexity prior 'penalizes slightly more than necessary'. With our prior specification (13), however, the exponential decay kicks in *only* when $|\mathcal{S}|$ is sufficiently large.

Assuming that the level of smoothness α is known, the optimal number of steps (i.e. tree bottom leaves K) needed to achieve the rate-optimal performance for estimating f_0 should be of the order $n^{q_0/(2\alpha+q_0)} = 1/C_\varepsilon^2 n\varepsilon_{n,\mathcal{S}_0}^2/\log n$ (Ročková & van der Pas, 2020). For our toy setup with a known α, we thus assume a point-mass prior on K with an atom near the optimal number of steps for each given $\mathcal{S}$, i.e.

$$\pi(K \mid \mathcal{S}) = \mathbb{I}[K = K_{\mathcal{S}}], \qquad \text{where} \quad K_{\mathcal{S}} = \lfloor C_K/C_\varepsilon^2 n\varepsilon_{n,\mathcal{S}}^2/\log n \rfloor \tag{14}$$

for some $C_K > 0$ such that $K_{\mathcal{S}_0} = 2^{q_0 s}$ for some $s \in \mathbb{N}$. In Section 4.2, we allow for more flexible trees with variable sizes.

4.1.2 | Identifiability

The active variables ought to be sufficiently relevant in order to make their identification possible. To this end, we introduce a non-parametric signal strength assumption, making sure that f_0 is not too flat in active directions (Comminges & Dalalyan, 2012; Yang & Pati, 2017).

We first introduce the notion of an approximation gap. For any given model $\mathcal{S}$, we denote with $\mathcal{F}_{\mathcal{S}}$ a set of approximating functions (only single trees $f_{\mathcal{T},\boldsymbol{\beta}}$ with $K_{\mathcal{S}}$ leaves for now) and define the approximation gap as follows:

$$\delta_n^{\mathcal{S}} \equiv \inf_{f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{F}_{\mathcal{S}}} \|f_0 - f_{\mathcal{T},\boldsymbol{\beta}}\|_n = \|f_0 - f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}\|_n, \tag{15}$$

where $f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}$ is the $\|\cdot\|_n$-projection of f_0 onto $\mathcal{F}_{\mathcal{S}}$. For identifiability of $\mathcal{S}_0$, we require that those models that miss one of the active covariates have a large separation gap.

Definition 4 (Identifiability)We say that $\mathcal{S}_0$ is (f_0, ε)-*identifiable* if, for some $M > 0$,

$$\inf_{i \in \mathcal{S}_0} \delta_n^{\mathcal{S}_0 \setminus i} > 2M\varepsilon. \tag{16}$$

We provide a more intuitive explanation of (16) in terms of directional variability of f_0. The best approximating tree $f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}$ can be written as

$$f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}(\boldsymbol{x}) = \sum_{k=1}^{K_{\mathcal{S}}} \mathbb{I}(\boldsymbol{x} \in \hat{\Omega}_k^{\mathcal{S}})\hat{\beta}_k \text{ with } \hat{\beta}_k = \bar{f}_0(\hat{\Omega}_k^{\mathcal{S}}) \equiv \frac{1}{n(\hat{\Omega}_k^{\mathcal{S}})} \sum_{\boldsymbol{x}_i \in \hat{\Omega}_k^{\mathcal{S}}} f_0(\boldsymbol{x}_i),$$

where $\widehat{\mathcal{T}} = \{\widehat{\Omega}_{k_S}^{S}\}_{k=1}^{K_S}$ is the tree-shaped partition of the $\|\cdot\|_n$-projection of f_0 defined in (15) with K_S leaves and where $n(\widehat{\Omega}_k^S) = \sum_{i=1}^{n}\mathbb{I}(\boldsymbol{x}_i \in \widehat{\Omega}_k^S) \equiv n\mu(\widehat{\Omega}_k^S)$. The separation gap in (15) can be then re-written as

$$\delta_n^S = \sqrt{\sum_{k=1}^{K_S} \mu(\widehat{\Omega}_k^S) V[f_0 \,|\, \widehat{\Omega}_k^S]},$$

where

$$V[f_0 \,|\, \widehat{\Omega}_k^S] \equiv \frac{1}{n(\widehat{\Omega}_k^S)} \sum_{\boldsymbol{x}_i \in \widehat{\Omega}_k^S} \left(f_0(\boldsymbol{x}_i) - \bar{f}_0(\widehat{\Omega}_k^S) \right)^2$$

is the local variability of f_0 inside $\widehat{\Omega}_k^S$. Given this characterization, (16) will be satisfied, for instance, when variability of f_0 inside best approximating cells that miss an active direction is too large, that is $\inf_{i\in\mathcal{S}_0} \inf_k V[f_0 \,|\, \widehat{\Omega}_k^{\mathcal{S}_0\setminus i}] > 4M^2\varepsilon^2$.

Our identifiability condition is a theoretical assumption on f_0 which indicates how large signal in each direction should be in order to be capturable. It generalizes the more traditional sufficient 'beta-min conditions' (Castillo et al., 2015; Zhao & Yu, 2006) for variable selection consistency (see Remark 4.1). Here, we gauge the amount of signal in terms of local variation in cells that *do not split* on an active covariate. Intuitively, if we do not split on $i \in \mathcal{S}_0$, the 'variation of f_0 inside the cells of the best tree we can get without i will be too large. The following example links our identifiability assumption with beta-min conditions.

Example 4 Assume for now that $p = 2$ and that f_0 is linear, i.e.

$$f_0(\boldsymbol{x}_i) = a + bx_{i1} + cx_{i2}.$$

Moreover, assume that $n = 16$ predictor observations are located on a regular grid $\mathcal{X} = \{k/4\colon 1 \le k \le 4\} \times \{j/4\colon 1 \le j \le 4\}$, where × denotes the Cartesian product. Suppose $\mathcal{S}_0 = \{1,2\}$ and set $\mathcal{S} = \mathcal{S}_0 \setminus \{2\} = \{1\}$ and $K_S = 2$. It can be verified that the partition $\widehat{\mathcal{T}}$ of the best approximating tree that *does not* split on the covariate x_2 consists of two rectangles $\widehat{\Omega}_1^S = [0, 1/2) \times [0, 1]$ and $\widehat{\Omega}_2^S = [1/2, 1] \times [0, 1]$. Then we have

$$\bar{f}_0(\widehat{\Omega}_1^S) = a + \frac{3}{2}\left(\frac{b}{4}\right) + \frac{5}{2}\left(\frac{c}{4}\right) \quad \text{and} \quad \bar{f}_0(\widehat{\Omega}_2^S) = a + \frac{7}{2}\left(\frac{b}{4}\right) + \frac{5}{2}\left(\frac{c}{4}\right)$$

and thereby

$$(\delta_m^S)^2 = V(f_0 \,|\, \widehat{\Omega}_1^S) = V(f_0 \,|\, \widehat{\Omega}_2^S) = \frac{1}{4}\frac{b^2}{16} + \frac{5}{4}\frac{c^2}{16}. \tag{17}$$

From the expression (17) we can immediately see the connection to the beta-min conditions. When the signal in the direction of x_2 is large enough, i.e. $c > 16/\sqrt{5}M\varepsilon$, our identifiability condition will be satisfied.

The second sufficient condition needed for methods such as the LASSO to fully recover $\mathcal{S}_0$ is 'irrepresentability' (Van De Geer & Bühlmann, 2009; Zhao & Yu, 2006). This condition restricts the amount of correlation between (active and non-active) covariates by imposing a regularization constraint on the magnitudes of regression coefficients of the inactive predictors onto the active ones. Here, we generalize the notion of irrepresentability to the non-parametric setup. Consider an underfitting model $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \not\supset \mathcal{S}_0$, where $\mathcal{S}_1 \subset \mathcal{S}_0$ are true positives and $\mathcal{S}_2$ is a possibly empty set of false positives, that is $\mathcal{S}_2 \cap \mathcal{S}_0 = \emptyset$. We define

$$\rho_n^{\mathcal{S}} \equiv \frac{1}{n}\sum_{i=1}^{n}[f_0(\boldsymbol{x}_i) - f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}_1}(\boldsymbol{x}_i)][f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}(\boldsymbol{x}_i) - f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}_1}(\boldsymbol{x}_i)], \tag{18}$$

the sample covariance between the surplus signals in f_0 and $f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}}$ obtained by removing the effect of $f_{\hat{\mathcal{T}},\hat{\boldsymbol{\beta}}}^{\mathcal{S}_1}$. This quantity will be large if noise covariates inside $\mathcal{S}_2$ can compensate for the missed true covariates in $\mathcal{S}_0 \setminus \mathcal{S}_1$, i.e. when the true and fake covariates are strongly correlated. To obviate this substitution effect, we introduce the following nonparametric ‘irrepresentability’ condition. Similarly as in Zhao and Yu (2006), we require that ‘the total amount of an irrelevant covariate represented by the covariates in the true model’ is small.

Definition 4 (Irrepresentability)We say that *ε-irrepresentability* holds for f_0 and $\mathcal{S}_0$ if, for some $M > 0$, we have $\sup_{\mathcal{S}\not\supset\mathcal{S}_0} |\rho_n^{\mathcal{S}}| < \frac{M}{2}\varepsilon$, where $\rho_n^{\mathcal{S}}$ was defined in (18).

It follows from Lemma S.1.2 (Appendix) that under the irrepresentability and identifiability conditions (Definitions 4.1 and 4.2), we obtain

$$\inf_{\mathcal{S}\not\supset\mathcal{S}_0} \inf_{f_{\mathcal{T},\boldsymbol{\beta}}\in\mathcal{F}_{\mathcal{S}}} \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n > M\varepsilon. \tag{19}$$

This condition essentially states that *all* models that miss *at least one* active covariate (i.e. not only subsets of the true model) have a large separation gap.

The following theorem characterizes variable selection consistency of spike-and-tree posterior distributions. Namely, the posterior distribution over the model index is shown to concentrate on the true model $\mathcal{S}_0$. One additional assumption is needed to make sure that the (fixed) design $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is sufficiently regular. Ročková and van der Pas (2020) define the notion of a fixed $\mathcal{S}_0$-regular design in terms of cell diameters of a *k-d* tree partition (Definition 3.3). This assumption essentially excludes outliers, making sure that the data cloud is spread evenly in active directions (while permitting correlation between covariates).

Theorem 4.1 *Assume* $f_0 \in \mathcal{H}_p^\alpha \cap C(\mathcal{S}_0)$ *for some* α ∈ *(0,1] and* $\mathcal{S}_0 \subset \{1, \ldots, p\}$ *with* $q_0 = |\mathcal{S}_0|$ *and* $\|f_0\|_\infty \lesssim B$. *Denote with* $\tilde{\varepsilon}_n = C_\varepsilon n^{-\alpha/(2\alpha+q_n)}\sqrt{\log n}$, *where* $q_n = C_q\lceil n\varepsilon_{n,\mathcal{S}_0}^2/\log p\rceil$ *for some* $C_q > 0$, *and assume* $q_0 \log p \le n^{q_0/(2\alpha+q_0)}$ *with* $2 \le q_0 = \mathcal{O}(1)$ *as* n → ∞. *Assume that (a)* $\mathcal{S}_0$ *is* $(f_0, \tilde{\varepsilon}_n)$-identifiable*, (b)* $\tilde{\varepsilon}_n$*-irrepresentability holds and that (c) the design* $\mathcal{X}$ *is* $\mathcal{S}_0$*-regular. Under the* spike-and-tree *prior comprising (with T = 1) (4),(5),(13) with C > 2 and (14), we have*

$$\Pi[\mathcal{S} = \mathcal{S}_0 \,|\, \boldsymbol{Y}^{(n)}] \to 1 \quad \text{in } \mathbb{P}_{f_0}^{(n)} - \text{probability as } n \to \infty.$$

Proof Section S.1.1

Remark 4.1 The assumption of $(f_0, \tilde{\varepsilon}_n)$-identifiability pertains to the more traditional sufficient beta-min conditions for variable selection consistency in sparse high-dimensional models. For example, Castillo et al. (2015) in their Corollary 1 require that $\min_{i\in S_0}|\beta_i^0| \ge M\sqrt{\frac{q_0 \log p}{n}}$, for some ‘large enough constant’ M>0 that depends on the compatibility number (see, e.g. Definition 2.1 in Castillo et al. (2015)) of the design matrix X (rescaled to have an $\|\cdot\|_2$ norm $\sqrt{n}$). Our identifiability threshold also depends on the rate of convergence ε_n (similarly as in Castillo et al. 2015). However, unlike in the linear models we measure the signal strength in a non-parametric way. Lastly, note that the identifiability gap $\tilde{\varepsilon}_n$ in Theorem 4.1 is a bit larger than the near-minimax rate $\varepsilon_{n,\mathcal{S}_0}$. This requirement will be relaxed in the next section, where α will be treated as unknown.

For *iid* models, Ghosal et al. (2008) considered the problem of nonparametric Bayesian model selection and averaging and characterized conditions under which the posterior achieves adaptive rates of convergence. The authors also study the posterior distribution of the model index, showing that it puts a negligible weight on models that are bigger than the optimal one. Yang and Pati (2017) characterized similar conditions for the non-*iid* case, see Section S.1.1 for more details.

Remark 4.2 (Theory for ABC)It is worth pointing out that Theorem 4.1 is obtained for the *actual* posterior $\pi(\mathcal{S}\,|\,\boldsymbol{Y}^{(n)})$, not the ABC posterior. Theory for ABC recently started emerging with the first results focussing on ABC bias (Barber et al., 2015), consistency and asymptotic normality (Frazier et al., 2018, 2020; Martin et al., 2014) and on convergence of the posterior mean (Li & Fearnhead, 2018). For our non-parametric regression scenario, we can conclude (variable selection) consistency for ABC Bayesian forests under the assumption that the residual variance σ^2 decreases with the sample size (as is typical in the Gaussian sequence model). In particular, Theorem S.1.1 in Supplemental Materials (Section S.1.4) shows that the ABC posterior concentrates at the rate $\lambda_n = 4\epsilon_n^T/3 + 1/\sqrt{n}$, where $\epsilon_n^T = \sqrt{2\log n/n}$ is the ABC tolerance level. This result implies that the ABC posterior will *not* reward underfitting model as long as our identifiability and irrepresentability conditions are satisfied with $\varepsilon = \lambda_n$. Regarding over-fitting models, an ABC analogue of Lemma 1.1 (Section 1.1.2 in Supplemental Materials) implies that the ABC posterior probability of over-fitting models goes to zero, which concludes variable selection consistency of a (naive) ABC method. These considerations can be extended to ABC Bayesian forests with data splitting using the empirical expected posterior prior justification in (9). More details are in Supplemental Materials (Section S.1.4).

Remark 4.3 (Consistency of the Median Probability Model)In Section 3.3, we used the median probability model rule which may not the same as the highest-posterior model whose consistency we have shown in Theorem 4.1. However, even when $p \to \infty$ it can be verified (as in Corollary 4.1 in Narisetty and He 2014) that the median probability model is *also* consistent under the same assumptions as Theorem 4. 1. In particular, $\mathbb{P}_{f_0}^{(n)}[\cap_{i=1}^{p} E_i] \to 1$ as $n \to \infty$ where $E_i = \{\Pi(\gamma_i = \gamma_i^0\,|\,\boldsymbol{Y}^{(n)}) > 0.5\}$ and where $\gamma_i = \mathbb{I}(i \in \mathcal{S})$ are binary inclusion indicators and $\gamma_i^0 = \mathbb{I}(i \in \mathcal{S}_0)$.

4.2 | The case of unknown α

The fact that the level α has to be known for the consistency to hold makes the result in Theorem 4.1 somewhat theoretical. In this section, we provide a joint consistency result for the unknown regularity level K and, at the same time, the unknown subset $\mathcal{S}_0$. Finding the optimal regularity level K, given $\mathcal{S}_0$, is a model selection problem of independent interest (Lafferty & Wasserman, 2001). Here, we acknowledge uncertainty about *both* K and $\mathcal{S}_0$ by assigning a joint prior distribution on $(K, \mathcal{S})$. Namely, we consider an analogue of (13), where $n^{|\mathcal{S}|/(2\alpha+|\mathcal{S}|)}$ is now replaced with K log n (according to (14)), i.e.

$$\pi(K, \mathcal{S}) \propto \mathrm{e}^{-C(K\log n \vee |\mathcal{S}|\log p)} \quad \text{for} \quad 1 \le K \le n \quad \text{and} \quad \mathcal{S} \subseteq \{1, \ldots, p\}. \tag{20}$$

This prior penalizes models with too many splits or too many covariates. We now regard each model as a *pair of indices*$(K, \mathcal{S})$, where the 'true' model is characterized by $\Gamma_0 = (K_{\mathcal{S}_0}, \mathcal{S}_0)$ with $K_{\mathcal{S}_0}$ defined in (14). Again, we partition the model index set $\Gamma = \{(K, \mathcal{S}): \mathcal{S} \subseteq \{1, \dots, p\}, 1 \leq K \leq n\}$ into (a) the true model Γ_0, (b) models that underfit $\Gamma_{\{\mathcal{S} \not\supset \mathcal{S}_0\} \cup \{K < K_{\mathcal{S}_0}\}}$ (i.e. miss at least one covariate *or* use less than the optimal number of splits), and (c) models that overfit $\Gamma_{\{\mathcal{S} \supset \mathcal{S}_0\} \cap \{K \geq K_{\mathcal{S}_0}\}}$ (i.e. use too many variables and splits).

We combine the identifiability and irrepresentability conditions into one as follows:

$$\inf_{\{\mathcal{S} \not\supset \mathcal{S}_0\} \cup \{K < K_{\mathcal{S}_0}\}} \inf_{f_{\mathcal{T},\beta} \in \mathcal{F}_{\mathcal{S}}(K)} \|f_{\mathcal{T},\beta} - f_0\|_n > M\varepsilon_{n,\mathcal{S}_0} \tag{21}$$

for some $M > 1$, where $\mathcal{F}_{\mathcal{S}}(K)$ consists of all trees with K bottom leaves and splitting variables $\mathcal{S}$. This condition is an analogue of (19), essentially stating that one cannot approximate f_0 with an error smaller than a multiple of the near-minimax rate using underfitting models.

Theorem 4.2 *Assume $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$ for some α ∈ (0,1] and $\mathcal{S}_0 \subset \{1, \dots, p\}$ such that $|\mathcal{S}_0| = q_0$ and $\|f_0\|_\infty \lesssim B$. Assume $q_0 \log p \leq n^{q_0/(2\alpha + q_0)}$ and $2 \leq q_0 = \mathcal{O}(1)$ as $n \to \infty$. Furthermore, assume that the design $\mathcal{X}$ is $\mathcal{S}_0$-regular and that (21) holds. Under the spike-and-tree prior comprising (with $T = 1$) (4), (5) and (20) for $C > 3$, we have*

$$\Pi\left[\{\mathcal{S} = \mathcal{S}_0\} \cap \{K_{\mathcal{S}_0} \leq K \leq K_n\} \,|\, \boldsymbol{Y}^{(n)}\right] \to 1 \quad \text{in } \mathbb{P}_{f_0}^{(n)} - \text{probability as } n \to \infty,$$

where $K_{\mathcal{S}_0}$ was defined in (14) and $K_n = \lceil \overline{C} n \varepsilon_{n,\mathcal{S}_0}^2 / \log n \rceil$ for some $\overline{C} > C_K / C_\varepsilon^2$.

Proof Section S.1.2

Note that both $K_{\mathcal{S}_0}$ and K_n are of the same (optimal) order, where the marginal posterior distribution $\Pi(K \,|\, \boldsymbol{Y}^{(n)})$ squeezes inside these two quantities as $n \to \infty$. Lafferty and Wasserman (2001) provide a similar result for their RODEO method, without the variable selection consistency part. Yang and Pati (2017) also provide a similar result for Gaussian processes, without the regularity selection consistency part. Here, we characterize *joint* consistency for both subset and regularity model selection.

4.3 | Variable selection consistency with Bayesian forests

Finally, we provide a variant of Theorem 4.2 for tree ensembles. Each Bayesian forest (i.e. additive regression tree) model is characterized by a triplet $(\mathcal{S}, T, \boldsymbol{K})$, where $\mathcal{S}$ is the active variable subset, $T \in \mathbb{N}$ is the number of trees and $\boldsymbol{K} = (K^1, \dots, K^T)' \in \mathbb{N}^T$ is a vector of the bottom leave counts for the T trees. Rate-optimality of Bayesian forests can be achieved for a wide variety of priors, ranging from many weak learners (large T and small K^t's) to a few strong learners (small T and large K^t's) (Ročková & van der Pas, 2020). The optimality requirement is that the *total* number of leaves in the ensemble $\sum_{t=1}^T K^t$ behaves like $K_{\mathcal{S}_0}$, defined earlier in (14).

We thereby define models in terms of equivalence classes rather than individual triplets $(\mathcal{S}, T, \boldsymbol{K})$. We construct each equivalence class $E(Z)$ by combining ensembles with the same number Z of total leaves that is

$$E(Z) = \bigcup_{T=1}^{\min\{Z,n\}} \left\{ \boldsymbol{K} \in \mathbb{N}^T : \sum_{t=1}^T K^t = Z \right\}. \tag{22}$$

The cardinality of $E(Z)$, denoted with $\Delta(E(Z))$, satisfies $\Delta(E(Z)) \leq Z!p(Z)$, where $p(Z)$ is the partitioning number (i.e. the number of ways one can write Z as a sum of positive integers). The 'true' model $\Gamma_0 = (\mathcal{S}_0, E(K_{\mathcal{S}_0}))$ consists of an equivalence class of forests that split on variables inside $\mathcal{S}_0$ with a total number of $K_{\mathcal{S}_0}$ leaves. Similarly as before, we define underfitting model classes $\Gamma_{\{\mathcal{S} \not\supset \mathcal{S}_0\} \cup \{E(Z): Z < K_{\mathcal{S}_0}\}}$ and overfitting model classes $\Gamma_{\{\mathcal{S} \supset \mathcal{S}_0\} \cap \{E(Z): Z \geq K_{\mathcal{S}_0}\}}$. Regarding the prior on T, similarly as Ročková and van der Pas (2020), we consider

$$\pi(T) \propto \mathrm{e}^{-C_T T}, \quad T = 1, \ldots, n, \quad \text{for} \quad C_T > 0. \tag{23}$$

Given T, we assign a joint prior over $\mathcal{S}_0$ and $\boldsymbol{K} \in \mathbb{N}^T$ as follows:

$$\pi(\mathcal{S}, \boldsymbol{K} \mid T) \propto \mathrm{e}^{-C\max\{|\mathcal{S}|\log p;\ \sum_{t=1}^{T} K^t \log n\}} \quad \text{for} \quad C > 1. \tag{24}$$

We conclude this section with a model selection consistency result for Bayesian forests under the following identifiability condition

$$\inf_{\{\mathcal{S} \not\supset \mathcal{S}_0\} \cup \{E(Z): Z < K_{\mathcal{S}_0}\}} \inf_{f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F}_{\mathcal{S}}(\boldsymbol{K})} \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n > M\varepsilon_{n,\mathcal{S}_0}, \tag{25}$$

where $\mathcal{F}_{\mathcal{S}}(\boldsymbol{K})$ denotes all forests $f_{\mathcal{E},\boldsymbol{B}}$ that split on variables $\mathcal{S}$ and consist of T trees with $\boldsymbol{K} = (K^1, \ldots, K^T)'$ bottom leaves.

Theorem 4.3 *Assume $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$ for some α ∈ (0,1] and $\mathcal{S}_0 \subset \{1, \ldots, p\}$ such that $|\mathcal{S}_0| = q_0$ and $\|f_0\|_\infty \lesssim B$. Assume $q_0 \log p \leq n^{q_0/(2\alpha + q_0)}$, where $2 \leq q_0 = \mathcal{O}(1)$ as $n \to \infty$. Furthermore, assume that the design is $\mathcal{S}_0$-regular and that (25) holds. Under the spike-and-forest prior comprising (4), (5), (23) and (24), we have*

$$\Pi\left[\{\mathcal{S} = \mathcal{S}_0\} \cap \left\{K_{\mathcal{S}_0} \leq \sum_{t=1}^{T} K^t \leq K_n\right\} \mid \boldsymbol{Y}^{(n)}\right] \to 1 \quad \text{in } \mathbb{P}_{f_0}^{(n)}\text{-probability as } n \to \infty,$$

where $K_{\mathcal{S}_0}$ was defined in (14) and $K_n = \lceil \overline{C} n\varepsilon_{n,\mathcal{S}}^2 / \log n \rceil$ for some $\overline{C} > C_K / C_\varepsilon^2$.
Proof Section S.1.3

5 | SIMULATION STUDY

We evaluate the performance of ABC Bayesian forests on simulated data. We consider the following performance criteria: Precision $= 1 - \text{FDP} = \frac{\text{TP}}{\text{TP}+\text{FP}}$, Power $= \frac{\text{TP}}{\text{TP}+\text{FN}}$(defined as the proportion of true signals discovered as such), Hamming Distance (HD)= FP+FN (where FP and FN denotes the number of false positives and false negatives, respectively) and the area under the ROC curve (AUC). Traditionally, AUC assesses how well a classification method can differentiate between two classes in the absence of a clear decision boundary. We use this criterion to assess variable importance since many of the considered selection methods are based on an importance measure and, as such, do not have a clear decision boundary.

The synthetic data are generated from the model (1), where $\boldsymbol{x}_i$'s for $i = 1, \ldots, n$ are drawn independently from $N_p(0, \Sigma)$ with $\Sigma = (\rho_{ij})_{i,j=1}^{p,p}$. We make our comparisons under different combinations of f_0, σ and Σ. In particular, we consider a relatively large noise level with $\sigma = 5$ ($\sigma = \sqrt{5}$ for the linear setup) and

1. medium equi-correlation $\rho_{ij} = 0.5$ for $i \neq j$ with $\rho_{ii} = 1$,
2. high auto-correlation $\rho_{ij} = 0.\,9^{|i-j|}$.

Regarding the mean function f_0, we consider four choices: (1) a linear setup with $f_0(\boldsymbol{x}_i) = x_{i1} + 2x_{i2} + 3x_{i3} - 2x_{i4} - x_{i5}$; (2) the Friedman setup as described in (11); (3) a CART (tree-based) function $f_0(\boldsymbol{x}_i)$ generated from the first 5 covariates using the `rpart` function in R; (4) a simulated example from Liang et al. (2018) (denoted with LLS hereafter) with $f_0(\boldsymbol{x}_i) = \frac{10x_{i2}}{1+x_{i1}^2} + 5\sin(x_{i3}x_{i4} + 2x_{i5})$. For the auto-correlation case, we permuted the covariates so that signals are not next to each other.

For each combination of settings, we repeat our simulation over 20 different datasets assuming $n = 500$ and $p \in \{100{,}1000\}$. We compare ABC Bayesian forests with RF, dynamic trees (DT) of Taddy et al. (2011b), BART (Chipman et al., 2010), DART of Linero (2018), LASSO and SFs (the MCMC counterpart of ABC Bayesian forests outlined in Section S.3 of the Supplemental Materials). ABC Bayesian forests are trained with $M = 1000$ ABC samples, where only a fraction of ABC samples (top 10%) are kept in the reference table. The prior $\pi(\mathcal{S})$ is the usual beta-binomial prior with $\theta \sim \mathcal{B}(1, 1)$. Inside each ABC step, we sample a subset of size $s = n/2$ and draw a tree ensemble using the default Bayesian CART prior (Chipman et al., 1998) and $T \in \{10, 20\}$ trees. For each ABC sample, we draw the last BART sample after $B = 200$ burn-in MCMC iterations. A sensitivity analysis to the choice s, T, B and M is reported in the Supplemental Materials (Section 4). Two versions of BART (without ABC) were deployed using the `R` package `BART`: (1) the standard BART from Chipman et al. (2010) with $T = 20$ (as recommended in Bleich et al. 2014), and (2) the sparse version DART of Linero (2018) with a Dirichlet prior (`sparse = TRUE, a = 0.5, b = 1`) with $T = 200$. Both versions are run with 10,000 MCMC samples after 10,000 burn-in. For LASSO, we use the `glmnet` package in `R` (Friedman et al., 2010) using the 1-se rule to select the penalty λ. For RF, we deploy the `randomForest` package in `R` (Liaw & Wiener, 2002) using the default number of 500 trees where variable importance is based on the difference in predictions (with and without each covariate) in out-of-bag samples.

To select variables with RF, there are at least three commonly used strategies: (1) recursive feature elimination (RFE) implemented in the `caret` package with 5-fold cross-validation (as suggested in Linero 2018); (2) truncating importance at the $1-\alpha$ quantile of a standard normal distribution (as suggested by Breiman and Cutler 2013); (3) truncating importance at the Bonferroni-corrected $(1-\alpha/p)$ quantile of a standard normal distribution (Bleich et al., 2014). We report the third method, which was seen to perform the best. For BART and DART, we select those variables which have been split on inside a forest at least once on average. Alternative strategies based on truncating inclusion probabilities (Linero, 2018) using data-adaptive thresholds (Bleich et al., 2014) did not perform better, in general. For ABC, we report results for two selection thresholds 0.5 and 0.25. For SF, we report the median probability model.

The performance comparisons for variable selection are summarized in Figure 2 (equi-correlation $\rho_{ij} = 0.5$) and Figure 3 (autocorrelation $\rho_{ij} = 0.\,9^{\,|i-j|}$). These figures show that ABC has an advantage in terms of AUC, suggesting that ABC can rank variables more efficiently. While RF tend to have a higher power, they are plagued with false discoveries (i.e. smaller precision). ABC Bayesian forests, on the other hand, are seen to yield fewer false discoveries (i.e. higher precision) relative to the other procedures. The ABC threshold 0.5 yields higher precision, whereas 0.25 yields higher power.

While ABC Bayesian forests were designed to explore the posterior distribution over models, it is natural to ask whether they also yield reasonable prediction. There are various ways to perform prediction with our ABC method. One natural strategy is to save each draw $f_{\boldsymbol{S},\boldsymbol{B}}^{m}$ at the m^{th} ABC iteration when $\epsilon_m < \epsilon$ and average out individual predictions obtained from these single draws. Alternatively,

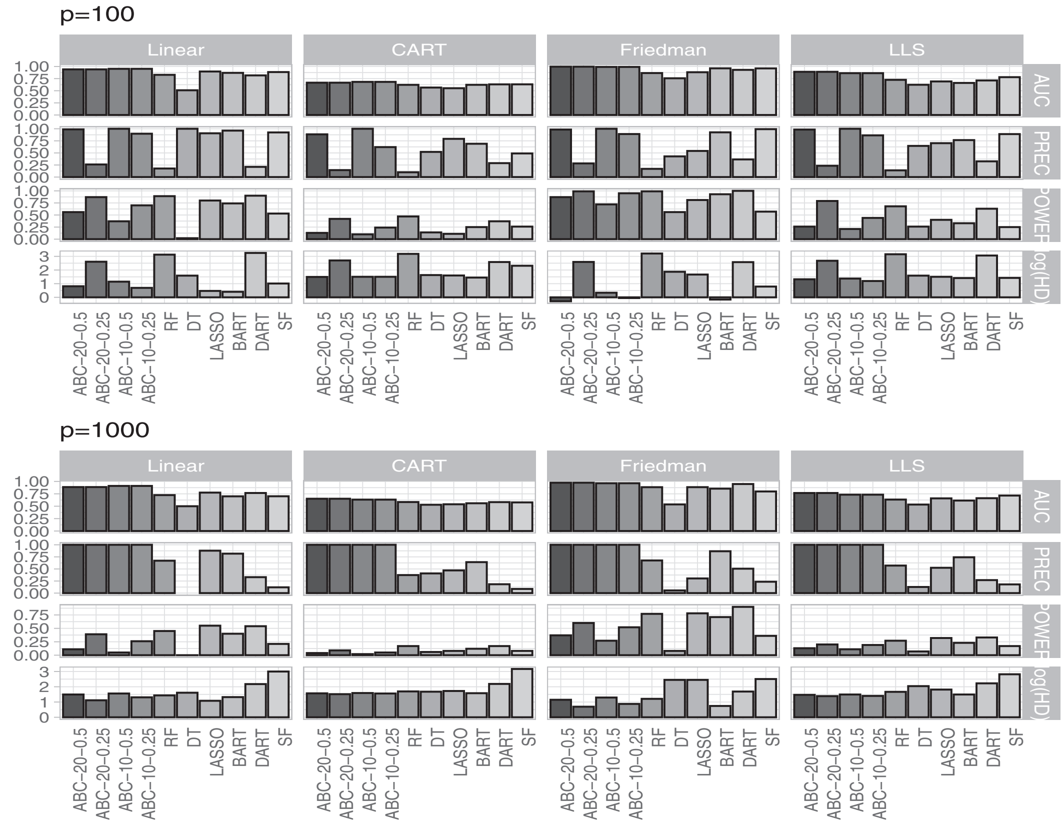


FIGURE 2 Average variable selection performance under equicorrelation $\rho_{ij} = 0.5$ over 20 simulations. Each panel corresponds to a different dimension $p \in \{100,1000\}$. Each row reports a different statistic: AUC is the area under the ROC curve, PREC $= 1 - \text{FDP} = \frac{\text{TP}}{\text{TP}+\text{FP}}$, POWER $= \frac{\text{TP}}{\text{TP}+\text{FN}}$, log (HD)= log (FP+FN). ABC is run for $T \in \{10,20\}$ and cutoff $\in \{0.5, 0.25\}$. Each column indicates a different data generating process

one could first select variables based on ABC Bayesian forests and then run a separate BART method (using the default number of $T = 200$ trees which is recommended for prediction) with the selected variables. Using both strategies, we report average out-of-sample mean squared prediction error, where the average is taken over 20 independent validation samples generated from the same data generating process (Table 1). We include both ABC predictions described above and denote them as ABC1 and ABC2, respectively, for the two different thresholds ($c \in \{0.5, 0.25\}$) and for the two choices of the number of trees ($T \in \{10, 20\}$).

The best method under each simulation setting is marked in bold. When the data becomes more non-linear (CART and LLS setups) and the correlation among variables gets stronger, ABC tends to outperform the other methods. DART, on the other hand, works better for more linear datasets. Note that our default ABC implementation internally uses only a *small* number of $B = 200$ burn-in iterations and a small number of trees. For prediction, it has been recommended that BART is deployed with a larger number of trees (Chipman et al., 2010). In addition, the ABC computation produces forest samples $f^{m}_{S,B}$ which are from an *approximate* posterior. These two facts may affect resulting predictions which may not necessarily outperform BART (DART) across-the-board.

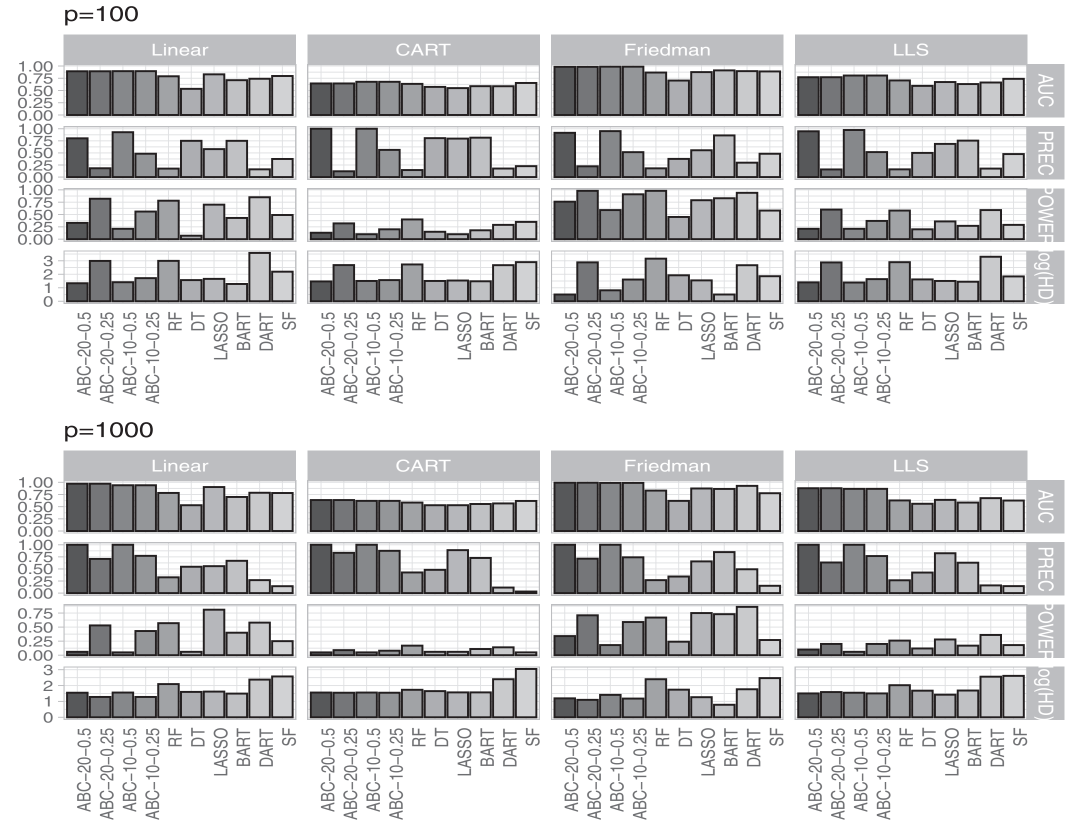


FIGURE 3 Average variable selection performance under autocorrelation $\rho_{ij} = 0.9^{|i-j|}$ over 10 simulations. Each panel corresponds to a different dimension $p \in \{100, 1000\}$. Each row reports a different statistics: AUC is the area under the ROC curve, PREC = 1 − FDP = $\frac{\text{TP}}{\text{TP+FP}}$, POWER = $\frac{\text{TP}}{\text{TP+FN}}$, log (HD)= log (FP+FN). ABC is run for $T \in \{10,20\}$ and cutoff $\in \{0.5,0.25\}$. Each column indicates a different data generating process

6 | HIV DATA

To further illustrate the usefulness of our approach, we consider a dataset described and analyzed in Rhee et al. (2006) and Barber and Candès (2015). The data consists of genotype and resistance measurements (log-decrease in susceptibility) for three drug classes, that is protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). The data are publicly available from the Stanford HIV Drug Resistance Database.[2]

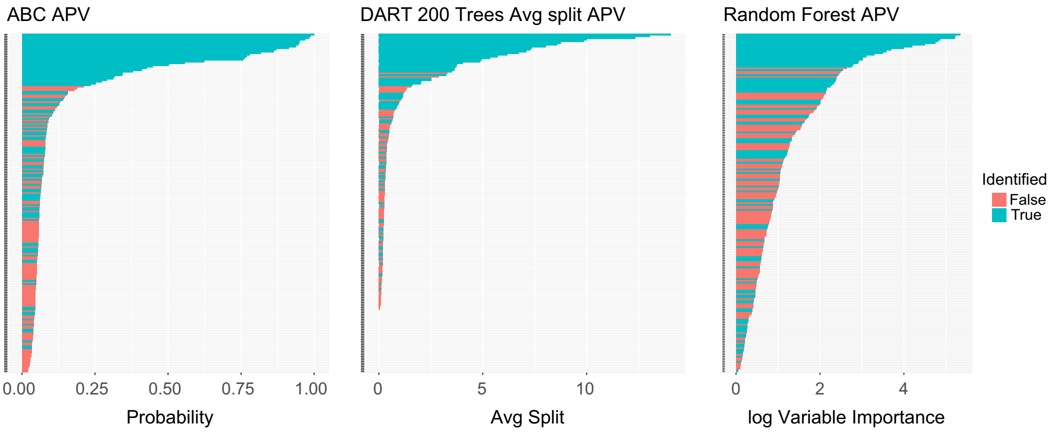
The goal of this analysis is to identify possible non-polymorphic mutation positions which result in a log-fold increase of laboratory-tested drug resistance. The design matrix $X = (x_{ij})_{i,j=1}^{n,p}$ consists of binary indicators $x_{ij} \in \{0, 1\}$ for whether or not the j^{th} mutation occurred in the i^{th} sample. As in Barber and Candès (2015), only mutations that appear at least 3 times are taken into consideration. One appealing feature of this dataset is the availability of a proxy to the 'ground truth'. Indeed, in an independent experimental study, Rhee et al. (2005) identified mutations that are present at a significantly higher frequency in patients who have been treated with each drug. Similarly as Barber and

[2] https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/

Candès (2015), we treat this experimental data as an approximation to the truth for comparisons and for validation of our findings.

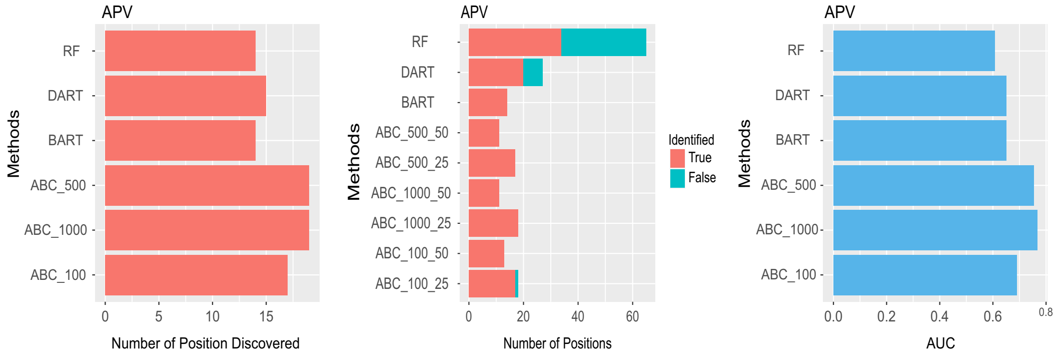
We run ABC with $M = 10{,}000$ iterations, where each internal BART sample is obtained after 200 burnin iterations with 20 trees. The top 1000 ABC samples with the smallest ϵ_m are kept and used to compute inclusion probabilities for each mutation. For illustration, we visualize results for one of the PI drugs (APV) and report the results for all the drugs in the Supplemental Material (Section S.5). The inclusion probabilities have been ordered and plotted in Figure 4, where the mutations experimentally validated by Rhee et al. (2005) (a proxy for true signals) are denoted in blue and the rest is in red. For comparisons, we also included the importance measure (the average number of splits on each variable) from DART run with 20,000 MCMC iterations and $T = 200$ trees as well as the importance measure (on a log scale) from RF run with 500 trees.

Figure 4 reveals that ABC Bayesian forests have a strong separation power, where experimentally validated mutations generally have a higher inclusion probability. Compared to DART and RF, ABC clearly stands out as being more effective in weeding out 'noise'. We gauge the strength of the signal/noise separation using several descriptive statistics. In these comparisons, we also consider plain BART method (using $T = 20$ trees and 20,000 MCMC iterations) and ABC using the top 100 and 500 samples with the smallest tolerance level ϵ_m. Since the selection of the cut-off point is not obvious for BART and RF, we first select variables based on an adaptive cut-off point so that there are no false discoveries (i.e. the cut-off is the largest importance weight of a *not* experimentally validated mutation). From the plot of the number of 'True' locations selected (displayed in Figure 5a) we can see that all three ABC implementations find more signal variables. Next, we choose the cut-off point in an automated way, where ABC importance probabilities are truncated at 0.5 and 0.25, BART and DART measures are truncated at one (i.e. the variable has been used on average at least once), and RF select variables using recursive feature elimination as explain in the previous section. Similarly to Barber and Candès (2015), we report the number of 'True' locations and 'False' locations (Figure 5b). RF selection is plagued with false discoveries and DART is not free from false identifications either. The ABC selection cutoff 0.5 results in a more conservative selection, where lowering the cutoff point



(a) ABC (b) DART (c) RF

FIGURE 4 A barplot of ordered importance measures (inclusion probabilities for ABC, importance measures for DART and RF) for each of the $p = 201$ mutations for the drug APV, where blue represents mutations found in Rhee et al. (2005). (a) Inclusion probabilities are computed using the top 1000 out of $M = 10{,}000$ ABC samples; (b) Average split of DART with 20,000 MCMC iterations; (c) log variable importance of random forest with 500 trees



(a) Adaptive Cut-off (b) Automated Cut-off (c) AUC

FIGURE 5 (a) The number of true discoveries using an adaptive cut-off; (b) The number of true (red) and false (blue) discoveries using an automated cut-off; (c) The AUC of each method

to 0.25 yields more discoveries. Finally, from the plot of the AUC values for all considered methods (Figure 5c), we conclude that ABC is better at separating the experimentally validated mutations from the rest even using a very few filtered ABC samples.

7 | DISCUSSION

This paper makes advancements at two fronts. One is the proposal of ABC Bayesian forests for variable selection based on a new idea of data splitting, where a fraction of data is first used for ABC proposal and the rest for ABC rejection. This new strategy increases ABC acceptance rate. We have shown that ABC Bayesian forests are highly competitive with (and often better than) other tree-based variable selection procedures. The second development is theoretical and concerns consistency for variable and regularity selection. Continuing the theoretical investigation of BART by Ročková and van der Pas (2020), we proposed new complexity priors which jointly penalize model dimensionality and tree size. We have shown joint consistency for variable *and* regularity selection when the level of smoothness is unknown and no greater than 1. Our results are the first model selection consistency results for BART priors.

Our ABC sampling routine has the potential to be extended in various ways. Sampling from $\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \,|\, \boldsymbol{Y}^{obs}_{\mathcal{I}_m}, \mathcal{S}_m)$ in ABC Bayesian forests is one way of distilling $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$ to propose a candidate ensemble $f^m_{\mathcal{E},\boldsymbol{B}}$. We noticed that the ABC acceptance rate can be further improved by replacing a randomly sampled tree with a fitted tree. Indeed, instead of drawing from $\pi(f_{\mathcal{E},\boldsymbol{B}}, \sigma^2 \,|\, \boldsymbol{Y}^{obs}_{\mathcal{I}_m}, \mathcal{S})$, one can *fit* a tree $\widehat{f}^m_{\mathcal{T},\boldsymbol{\beta}}$ to $\boldsymbol{Y}^{obs}_{\mathcal{I}_m}$ using recursive partitioning algorithms (such as the `rpart` R package of Therneau and Atkinson (2018)) or with BART (by taking the posterior mean estimate $\widehat{f}^m_{\mathcal{E},\boldsymbol{B}} = \mathbb{E}[f_{\mathcal{E},\boldsymbol{B}} \,|\, \boldsymbol{Y}^{obs}_{\mathcal{I}_m}, \mathcal{S}]$). This variant, further referred to as ABC Forest Fit, is indirectly linked to other model-selection methods based on resampling.

Felsenstein (1985) proposed a 'first-order bootstrap' to assess confidence of an estimated tree phylogeny. The idea was to construct a tree from each bootstrap sample and record the proportion of bootstrap trees that have a feature of interest (for us, this would be variables used for splits). Efron and Tibshirani (1998) embedded this approach within a parametric bootstrap framework, linking the bootstrap confidence level to both frequentist p-values and Bayesian a posteriori model probabilities.

The authors proposed a second-order extension by reweighting the first-order resamples according to a simple importance sampling scheme. This second-order variant performs frequentist calibration of the a posteriori probabilities and amounts to performing Bayesian analysis with Welch–Peers uninformative priors. Efron (2012) further develops the connection between parametric Bootstrap and posterior sampling through reweighting in exponential family models. Using non-parametric bootstrap ideas, Newton and Raftery (1994) introduce the weighted likelihood bootstrap (WLB) to sample from approximate posterior distributions. The WLB samples are obtained by maximum reweighted likelihood estimation with random weights. Such posterior sampling can be beneficial when, for instance, maximization is easier than Gibbs sampling from conditionals. In a similar spirit, our ABC Forest Fit variant would perform optimization (instead of sampling) on a random subset of the dataset to obtain a candidate tree/ensemble.

It is worth pointing out that $\widehat{f}^{m}_{\mathcal{E},\boldsymbol{B}}$ does not necessarily have to be a tree/forest. We suggest trees because they are are easily trainable and produce stable results using traditional software packages. In principle, however, this method could be deployed in tandem with other non-parametric methods, such as deep learning, to perform variable selection.

ACKNOWLEDGMENTS

This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

REFERENCES

Barber, R.F. & Candès, E.J. (2015) Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43, 2055–2085.

Barber, S., Voss, J. & Webster, M. (2015) The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9, 80–105.

Barbieri, M.M. & Berger, J.O. (2004) Optimal predictive model selection. *The Annals of Statistics*, 32, 870–897.

Berger, J.O. & Pericchi, L.R. (1996) The intrinsic Bayes factor for linear models. *Bayesian Statistics*, 5, 25–44.

Berger, J.O. & Pericchi, L.R. (2004) Training samples in objective Bayesian model selection. *The Annals of Statistics*, 32, 841–869.

Bleich, J., Kapelner, A., George, E.I. & Jensen, S.T. (2014) Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics*, 1750–1781.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Breiman, L. & Cutler, A. (2013) Online manual for random forests. Available from: http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.html.

Breiman, L., Friedman, J., Olshen, R. & Stone, C.J. (1984) *Classification and regression trees*. New York: Chapman and Hall.

Candes, E., Fan, Y., Janson, L. & Lv, J. (2018) Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 551–577.

Carbonetto, P. & Stephens, M. (2012) Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7, 73–108.

Castillo, I. & Mismer, R. (2018) Empirical Bayes analysis of spike and slab posterior distributions. *arXiv preprint arXiv:1801.01696*.

Castillo, I. & van der Vaart, A. (2012) Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40, 2069–2101.

Castillo, I., Schmidt-Hieber, J. & Van der Vaart, A. (2015) Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43, 1986–2018.

Chipman, H.A., George, E.I. & McCulloch, R.E. (1998) Bayesian CART model search. *Journal of the American Statistical Association*, 93, 935–948.

Chipman, H., George, E.I. & McCulloch, R.E. (2001) The practical implementation of Bayesian model selection. In: *Model Selection*. Institute of Mathematical Statistics, pp. 65–116.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266–298.

Comminges, L. & Dalalyan, A.S. (2012) Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40, 2667–2696.

Csillery, K., Blum, M.G., Gaggiotti, O.E. & Francois, O. (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25, 410–418.

Denison, D.G., Mallick, B.K. & Smith, A.F. (1998) A Bayesian CART algorithm. *Biometrika*, 85, 363–377.

Efron, B. (2012) Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6, 1971.

Efron, B. & Tibshirani, R. (1998) The problem of regions. *The Annals of Statistics*, 26(5), 1687–1718.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96, 1348–1360.

Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39, 783–791.

Frazier, D.T., Martin, G.M., Robert, C.P. & Rousseau, J. (2018) Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105, 593–607.

Frazier, D.T., Robert, C.P. & Rousseau, J. (2020) Model misspecification in approximate Bayesian computation: Consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 421–444.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, 19, 1–67.

Friedman, J., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1.

George, E.I. & McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881–889.

Ghosal, S., Lember, J. & Van Der Vaart, A. (2008) Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2, 63–89.

Ghosh, J.K. & Samanta, T. (2002) Nonsubjective Bayes testing? An overview. *Journal of Statistical Planning and Inference*, 103, 205–223.

Good, I.J. (1950) *Probability and the weighing of evidence*. London: C. Griffin.

Gramacy, R. & Lee, H. (2008) Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103, 1119–11303.

Grelaud, A., Robert, C.P. & Marin, J.-M. (2009) ABC methods for model choice in Gibbs random fields. *Comptes Rendus Mathematique*, 347, 205–210.

Guan, Y. & Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5, 1780–1815.

Hill, J. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20, 217–240.

Ishwaran, H. (2007) Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1, 519–537.

Jiang, B., Wu, T., Zheng, C. & Wong, W. (2017) Learning summary statistic for approximate Bayesian computation via deep neural network. *Statistica Sinica*, 27, 1595–1618.

Johnson, V.E. & Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107, 649–660.

Kazemitabar, J., Amini, A., Bloniarz, A. & Talwalkar, A.S. (2017) Variable importance using decision trees. In: *Advances in Neural Information Processing Systems*, pp. 425–434.

Lafferty, J. & Wasserman, L. (2001) Iterative Markov Chain Monte Carlo computation of reference priors and minimax risk. In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 293–300.

Lafferty, J. & Wasserman, L. (2008) RODEO: Sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1), 28–63.

Lember, J. & van der Vaart, A. (2007) On universal Bayesian adaptation. *Statistics & Decisions*, 25, 127–152.

Lempers, F.B. (1971) Posterior probabilities of alternative linear models.

Li, W. & Fearnhead, P. (2018) Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, 105, 301–318.

Liang, F., Li, Q. & Zhou, L. (2018) Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113, 955–972.

Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. *R News*, 2, 18–22.

Lin, Y. & Zhang, H.H. (2006) Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34, 2272–2297.

Linero, A.R. (2018) Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113, 1–11.

Marin, J.-M., Pudlo, P., Robert, C.P. & Ryder, R.J. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, 22, 1167–1180.

Martin, J.S., Jasra, A., Singh, S.S., Whiteley, N., Del Moral, P. & McCoy, E. (2014) Approximate Bayesian computation for smoothing. *Stochastic Analysis and Applications*, 32, 397–420.

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C. & Pratola, M. (2018) BART: Bayesian Additive Regression Trees. Available from: https://CRAN.R-project.org/package=BART. R package version 1.6.

Moreno, E., Girón, J. & Casella, G. (2015) Posterior model consistency in variable selection as the model dimension grows. *Statistical Science*, 30, 228–241.

Narisetty, N.N. & He, X. (2014) Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42, 789–817.

Newton, M.A. & Raftery, A.E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56, 3–48.

O'Hagan, A. (1995) Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 99–118.

Pérez, J.M. & Berger, J.O. (2002) Expected-posterior prior distributions for model selection. *Biometrika*, 89, 491–512.

Plagnol, V. & Tavaré, S. (2004) Approximate Bayesian Computation and MCMC. In: *Monte Carlo and Quasi-Monte Carlo Methods 2002*. Springer, pp. 99–113.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. & Robert, C.P. (2015) Reliable ABC model choice via random forests. *Bioinformatics*, 32, 859–866.

Radchenko, P. & James, G.M. (2010) Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105, 1541–1553.

Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 1009–1030.

Rhee, S.-Y., Fessel, W.J., Zolopa, A.R., Hurley, L., Liu, T., Taylor, J., et al. (2005) Hiv-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype b isolates and implications for drug-resistance surveillance. *The Journal of infectious diseases*, 192, 456–465.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D.L. & Shafer, R.W. (2006) Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103, 17355–17360.

Robert, C.P., Cornuet, J.-M., Marin, J.-M. & Pillai, N.S. (2011) Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108, 15112–15117.

Ročková, V. (2017) Particle EM for variable selection. *Journal of the American Statistical Association*, 113, 1–30.

Ročková, V. & George, E.I. (2014) EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109, 828–846.

Ročková, V. & George, E.I. (2018) The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113, 431–444.

Ročková, V. & van der Pas, S. (2020) Posterior concentration for Bayesian regression trees and their ensembles. *The Annals of Statistics (in revision)*, 48, 2108–2131.

Ročková, V. & Saha, E. (2019) On theory for BART. In: *Artificial Intelligence and Statistics*.

Savitsky, T., Vannucci, M. & Sha, N. (2011) Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical Science*, 26, 130.

Scheipl, F. (2011) spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in r. *arXiv preprint arXiv:1105.5253*.

Scott, J.G. & Berger, J.O. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 2587–2619.

Sunnaaker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M. & Dessimoz, C. (2013) Approximate Bayesian computation. *PLoS computational biology*, 9, e1002803.

Taddy, M., Gramacy, R. & Polson, N. (2011a) Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106, 409–123.

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

Tavaré, S., Balding, D.J., Griffiths, R.C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, 145, 505–518.

Therneau, T. & Atkinson, B. (2018) rpart: Recursive Partitioning and Regression Trees. Available from: https://CRAN.R-project.org/package=rpart. R package version 4.1-13.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Turlach, B.A. (2004) Discussion on least angle regression. *The Annals of Statistics*, 32, 481–490.

Van De Geer, S.A. & Bühlmann, P. (2009) On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3, 1360–1392.

Yang, Y. & Pati, D. (2017) Bayesian model selection consistency and oracle inequality with intractable marginal likelihood. *arXiv preprint arXiv:1701.00311*.

Zhao, P. & Yu, B. (2006) On model selection consistency of LASSO. *Journal of Machine Learning Research*, 7, 2541–2563.

Zhu, R., Zeng, D. & Kosorok, M.R. (2015) Reinforcement learning trees. *Journal of the American Statistical Association*, 110, 1770–1784.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Liu Y, Ročková V, Wang Y. Variable selection with ABC Bayesian forests. *J R Stat Soc Series B*. 2021;00:1–29. https://doi.org/10.1111/rssb.12423