Year 2005 *Paper* 182

Cross-validated Bagged Learning

Mark J. van der Laan[*] Sandra E. Sinisi[†]
Maya L. Petersen[‡]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley, saucy54@gmail.com

[‡]Division of Biostatistics, School of Public Health, University of California, Berkeley, mayaliv@berkeley.edu

Cross-validated Bagged Learning

Mark J. van der Laan, Sandra E. Sinisi, and Maya L. Petersen

Abstract

Many applications aim to learn a high dimensional parameter of a data generating distribution based on a sample of independent and identically distributed observations. For example, the goal might be to estimate the conditional mean of an outcome given a list of input variables. In this prediction context, Breiman (1996a) introduced bootstrap aggregating (bagging) as a method to reduce the variance of a given estimator at little cost to bias. Bagging involves applying the estimator to multiple bootstrap samples, and averaging the result across bootstrap samples. In order to deal with the curse of dimensionality, typical practice has been to apply bagging to estimators which themselves use cross-validation, thereby using cross-validation within a bootstrap sample to select fine-tuning parameters trading off bias and variance of the bootstrap sample-specific candidate estimators. In this article we point out that in order to achieve the correct bias variance trade-off for the parameter of interest, one should apply the cross-validation selector externally to candidate bagged estimators indexed by these fine-tuning parameters. In addition we define variable importance as a summary measure of the parameter of interest, and present a novel bootstrap method to achieve statistical inference and p-values based on the (externally) cross-validated bagged estimator. We illustrate the new cross-validated bagging method with a data analysis and investigate the performance of the variable importance measures in a small simulation study.

Contents

1 Introduction and motivation.

Many applications aim to use a learning data set from a particular data generating distribution to construct a predictor of an outcome as a function of a collection of input variables. One can define an optimal predictor as a parameter of the data generating distribution by defining it as the function of input variables which minimizes the expectation of a particular loss function (of the experimental unit and the candidate regression) with respect to the true data generating distribution. If one selects the squared error loss function (i.e., the square of the difference between the outcome and predicted value), then this optimal predictor is the conditional mean of the outcome, given the input variables. In the statistical literature such a location parameter of the conditional distribution of the outcome given the input variables is often referred to as a regression.

In many applications the number of input variables is very large. As a consequence, assuming a fully parameterized regression model such as a linear regression model with only main terms, and minimizing the empirical mean of the loss function (e.g., the sum of squared residual errors in the case of the squared error loss function) is likely to yield poor estimators, since the number of main terms will typically be too large (thereby resulting in over-fitting), and other functional forms of the input variables should be considered. In other words, many current applications frequently demand nonparametric regression estimators. Because of the curse of dimensionality, minimizing the empirical mean of the loss function, i.e., the empirical risk, over all allowed regression functions results in an over-fit of the data.

As a consequence, many estimators follow the sieve loss-based estimation strategy. That is, a sequence of subspaces indexed by fine-tuning parameters is selected, the empirical risk over each subspace is minimized or locally minimized to obtain a subspace-specific (minimum empirical risk) estimator, and the fine-tuning parameter (i.e., the subspace) is selected using an appropriate method to trade off between bias and variance. Examples of fine-tuning parameters indexing constraints on the space of regression functions include initial dimension reduction, the number of terms in the regression model, and the complexity of the allowed functional forms (e.g., basis functions). Each specification of the fine-tuning parameters corresponds to a candidate estimator of the true underlying regression. In order to select among these candidate estimators (i.e., to select these fine-tuning parameters), most algorithms either minimize a penalized empirical risk or minimize the cross-validated risk.

Application of such "machine learning" algorithms to a data set commonly results in a very low dimensional fit. For example in a recent HIV-data application involving prediction of viral replication capacity based on the mutation profile of the virus, in spite of the fact that the employed algorithm searched over a high dimensional space of regression functions, a linear regression with two main terms and a single interaction was selected (Birkner et al., 2005). Although such an estimator is based on a sensible trade off between bias and variance, the resulting fit is disappointing from two perspectives. First, in many applications the true regression is believed to be a function of almost all variables, with many variables making small contributions. Second, a practitioner often wishes to obtain a measure of importance for each variable considered, and such a low dimensional fit reflects zero importance for all variables that do not appear in the estimator. It has been common practice to address the second issue by reporting many of the fits the algorithm has searched over, and to summarize these different fits in a particular manner. Initially, we also followed this approach, but came to the conclusion that the statistical interpretation of such a summary measure is unclear.

Based on these concerns, the following statistical challenge can be formulated: construction of nonparametric regression estimators that 1) are high dimensional, so that the majority of variables contribute to the obtained regression, and 2) still correspond with a sensible trade-off between bias and variance (and thereby have good asymptotic convergence properties). In the current article, we address this challenge. In order to construct high dimensional learners (i.e., 1), we employ the existing machine learning method "**b**ootstrap **agg**regat**ing**" or "bagging" (or "aggregate prediction"), as introduced by Breiman (1996a). However, in order to establish 2), we will provide a fundamental improvement to the current practice of bagging by changing the way cross-validation enters into the methodology.

Breiman suggested bagging as a method to stabilize (and thereby improve upon) a given highly variable estimator. Specifically, given an estimator, Breiman defined a corresponding bagged estimator as the average across bootstrap samples of the bootstrap sample-specific estimators. Since different bootstrap samples typically result in different regression fits, the bagged estimator is typically a very high dimensional regression. Two applications of bagging are provided in random forest and linear regression (Breiman, 2001a). In random forests, the bagged regression estimator is defined as an average of bootstrap-specific classification and regression tree (CART) estimators (Breiman et al., 1984), where in each bootstrap sample CART is applied without cross-validation to obtain a fine partitioning. In the linear regression context, Breiman (1996a) proposed a bagged estimator as the average of bootstrap-specific cross-validated regression estimators, such as a linear regression estimator using forward selection and cross-validation to data adaptively select the size of the model. To conclude, in the current literature on bagging one either aggregates over-fitted regression estimators or one aggregates cross-validated regression estimators.

We note that the latter type of cross-validation within a bootstrap sample provides the right trade-off between bias and variance for the single sample estimator applied to the bootstrap sample. However, since the bagging operation reduces the variance and increases bias, it will typically result in the wrong trade-off for the corresponding bootstrap aggregated estimators. In this article we propose a cross-validated bagged estimator which 1) acknowledges that each estimator indexed by fine-tuning parameters corresponds with a bagged estimator, and 2) uses (external) cross-validation to select among these candidate bagged estimators, and possibly between these estimators and additional (e.g.) non-bagged estimators. By including non-bagged estimators in the set of candidate estimators, this procedure data adaptively selects between bagged and non-bagged estimators, which can be useful in cases where it is unclear if bagging actually improves the prediction performance: see our overview of some of the bagging literature below.

In order to assess the performance of the proposed cross-validated bagged estimator, we propose the use of cross-validation again: that is, the cross-validated bagged estimator is applied to a learning sample, and its fit is evaluated on a test sample, across different splits of the data into learning and test samples. The latter type of procedure involves double cross-validation.

The cross-validated bagged regression estimator maps into specific measures of variable importance for each variable. For example, one could define the importance of a variable as the empirical mean of the variable-specific partial derivative of the regression function, and the underlying true variable importance as the expectation of the partial derivative of the true regression. Beyond reporting a list of estimates of the importance of each variable, we also aim to provide an estimate of the variance of each variable importance estimate and a corresponding p-value assessing its significance relative to a null distribution. Since explicit calculations are hard to carry out, it seems appropriate to propose a re-sampling based approach which involves estimating the true data generating distribution, and establishing the Monte-Carlo variance of the variable importance estimates under this estimated data generating distribution: that is, one might simply apply the bootstrap method. Unfortunately, since the bagged estimators already require running a bootstrap simulation, this would involve bootstrapping a bootstrap-based procedure, an extremely computer intensive proposal. For example, if each bootstrap procedure involves 10,000 bootstrap samples, one would need to calculate $10,000^2$ bootstrap-specific estimators for each value of the fine-tuning parameters.

In this article, we propose a bootstrap method which only requires carrying out a single Monte-Carlo simulation procedure for a) estimating the variance of the variable importance for a given bagged estimator indexed by a fixed value of the fine-tuning parameters, and b) obtaining a p-value.

The organization of this article is as follows. In the next section we present our cross-validated bagged estimator, presented in the context of the general unified loss-based estimation framework as introduced in van der Laan and Dudoit (2003). That is, our estimator applies to any parameter which can be represented as the minimizer over the parameter space of an expectation of a loss function of the experimental unit at a parameter value, where we allow the loss function to be indexed by an unknown nuisance parameter. In particular, this general framework allows us to define the cross-validated bagged estimator of a conditional density, conditional hazard, or conditional location parameter (e.g., mean, median), based on censored and uncensored data. For example, our framework includes prediction of a survival time when the survival time is subject to right censoring. In Section 3 we define a variable importance parameter, the corresponding estimate, and we present our bootstrap method for obtaining a p-value and standard error estimate. In Section 4 we apply our cross-validated bagged estimator to the Deletion/Substitution/Addition Polynomial Regression algorithm introduced in Sinisi and van der Laan (2004), and discuss the resulting new machine learning algorithm. We have implemented this algorithm in C, and it will be made publicly available as an R-function in the near future. In Section 5, we apply this algorithm to a HIV data set to build a predictor of change in plasma HIV RNA level (viral load) based on characteristics of the HIV-infected patient, the sequence of the HIV-virus, and a drug regimen. In addition, we map our final fit into corresponding variable importance measures of the mutations of the HIV-virus and the candidate drugs.

1.1 Brief Review of Bagging.

Bagging, or bootstrap aggregating, was introduced by Breiman (1996a) as a tool for reducing the variance of a predictor. The general idea is to generate multiple versions of a predictor and then use these to get an aggregated predictor. The multiple predictors are obtained by using bootstrap replicates of the data, and bagging is meant to yield gains in accuracy. Whether or not bagging will improve accuracy is related to the stability of the procedure that constructs each predictor (Breiman, 1996a). Breiman (1996b) studied instability and stated that k-nearest neighbor methods are *stable*, but that neural networks, classification and regression trees, and subset selection in linear regression were *unstable* methods. Breiman (1996a) found that bagging works well for unstable methods.

Several approaches have been offered to combine different classifiers (LeBlanc and Tibshirani, 1996; Breiman, 1996c; Hothorn and Lausen, 2003). In addition, the following modifications of bagging have been proposed: "nice" bagging (Skurichina and Duin, 1998), sub-bagging or sub-sample aggregating (Buhlmann and Yu, 2002), and iterated bagging or de-biasing (Breiman, 2001b). We provide a brief overview of the various properties of bagging that have been studied and the application of bagging to available algorithms in the literature.

Friedman and Hall (2000) show that bagging reduces variability when applied to highly nonlinear estimators such as decision trees and neural networks, and can also reduce bias for certain types of estimators. Breiman (2001b) show that iterated bagging is effective in reducing both bias and variance. Buja and Stuetzle (2002) look at bagging statistical functionals and U-statistics and apply bagging to CART (Breiman et al., 1984). They find that in the case of bagging CART, both variance and bias can be reduced. Buhlmann and Yu (2002) define the notion of instability and analyze how bagging reduces variance in hard decision problems. Because hard decisions create instability, bagging is helpful to smooth these out, yielding smaller variance and mean squared error. They also look at the bagging effect on piecewise linear spline functions in multivariate adaptive regression splines (MARS) (Friedman, 1991) and find that bagging is unnecessary for MARS. Borra and Ciaccio (2002) also apply bagging to MARS, as well as to project pursuit regression (PPR) and local learning based on recursive covering (DART), and note that in most cases bagging reduces the variability of these methods. Bagging has been viewed from its ability to reduce instability

(Buhlmann and Yu, 2002), its success with nonlinear features of statistical methods (Friedman and Hall, 2000; Buja and Stuetzle, 2002), and Hall and Samworth (2005) address how performance depends on re-sample size.

Skurichina and Duin (1998) offers several conclusions about bagging for linear classifiers; these include that bagging is not necessarily a stabilizing technique where stabilization is defined for linear classifiers, the number of bootstrap replicates should be limited (Breiman, 1996a; Skurichina and Duin, 1998), the usefulness of bagging can be determined by the instability of the classifier, and that bagging improves the performance of a classifier when the classifier is unstable. For computational considerations, it is helpful to have a sense of how many bootstrap replicates are adequate. Breiman (1996a) looked at as few as 10 replicates up to 100 and suggested that fewer replicates are required when the outcome is numerical.

2 The general cross-validated bagged learner.

Suppose that one observes a sample of n i.i.d. observations on a random variable O with data generating distribution P_0, which is known to be an element of a model $\mathcal{M}$. Let $\psi_0 = \Psi(P_0)$ be the parameter of interest of the data generating distribution P_0. We assume that the true parameter (value) ψ_0 can be defined in terms of a *loss function*, $(O, \psi) \rightarrow L(O, \psi)$, as the minimizer of the expected loss, or *risk*. That is,

$$\psi_0 = \Psi(P_0) = \arg\min_{\psi \in \mathbf{\Psi}} \int L(o, \psi) dP_0(o),$$

where the minimum is taken over the parameter space $\mathbf{\Psi} \equiv \{\Psi(P) : P \in \mathcal{M}\}$. In regression with a continuous outcome, a common loss function is the squared error loss, $L(O = (Y, W), \psi) = (Y - \psi(W))^2$, corresponding to the conditional mean $\psi_0(W) = E_0[Y \mid W]$, and if $\psi_0 = dP_0/d\mu$ is the actual density of O or a sub-vector of O, then $L(O, \psi) = -\log \psi(O)$. As in the unified loss based estimation approach presented in van der Laan and Dudoit (2003), it is allowed that the loss function depends on a nuisance parameter $\Upsilon(P_0)$: that is $L(O, \psi) = L(O, \psi \mid \Upsilon(P_0))$. By allowing such unknown loss functions, this framework includes most parameters. In particular, van der Laan and Dudoit (2003) show that for estimation of regressions and densities based on censored data one can use the Inverse Probability of Censoring Weighted (IPCW) Mappings or Double Robust IPCW mapping as presented for general censored data structures in van der Laan and Robins (2003) to map the full data loss function into an observed data loss function indexed by nuisance parameters. Various applications of this unified loss-based cross-validation methodology for estimator selection include selection among regression estimators (Dudoit and van der Laan, 2003), estimator selection with right censored data (Keleş et al., 2003), likelihood-based cross-validation (van der Laan et al., 2003), tree-based estimation and cross-validation with censored data (Molinaro et al., 2003).

For the sake of notational convenience, we suppress the possible dependence of the loss function on a nuisance parameter in the notation, but we will point out at the appropriate places how this affects the proposed estimation procedure.

Let P_n denote the empirical probability distribution of the sample $O_1, \ldots, O_n$, which puts mass $1/n$ on each observation. Consider now a collection of candidate estimators $P_n \rightarrow \hat{\Psi}_s(P_n)$ indexed by a fine tuning parameter s ranging over a set $\mathcal{A}_n$. For example, if $\psi_0(W) = E(Y \mid W)$, then $\hat{\Psi}_s(P_n)$ might represent a particular learning algorithm for estimation of $E(Y \mid W)$ indexed by fine tuning parameters s which are user supplied, such as a support-vector machine algorithm, a forward step-wise algorithm, logic regression (Ruczinski et al., 2003), the D/S/A-polynomial regression algorithm (Sinisi and van der Laan, 2004), and so on. Another class of general examples is obtained by defining

$$\hat{\Psi}_s(P_n) \equiv \arg \min_{\psi \in \mathbf{\Psi}_s} \sum_{i=1}^{n} L(O_i, \psi)$$

as the minimizer of the empirical risk $\sum_i L(O_i, \psi)$ over a sub-parameter space $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ indexed by s, given a collection of subspaces $\mathbf{\Psi}_s$, $s \in \mathcal{A}_n$.

In the case that the loss function depends on a nuisance parameter $\Upsilon(P_0)$, one would estimate the nuisance parameter with an estimator $\hat{\Upsilon}(P_n)$, and minimize the empirical risk corresponding with the estimated loss function. Most estimators can be considered as approximate minimizers of the empirical risk, indexed by parameters defining the search algorithm, such as the space the algorithm searches over and the depth to which it searches the space. We note that we view the estimators as mappings $\hat{\Psi}_s$ from data, P_n, to the parameter space.

Given the empirical distribution P_n, let $P_n^{\#}$ denote the empirical distribution of a sample of n i.i.d. observations $O_1^{\#}, \ldots, O_n^{\#}$ from P_n. Given the s-specific estimator $\hat{\Psi}_s$ we can define a corresponding s-specific bagged estimator as

$$\tilde{\Psi}_s(P_n) \equiv E(\hat{\Psi}_s(P_n^{\#}) \mid P_n).$$

To evaluate the conditional expectation, given the data P_n, one needs to draw many bootstrap samples $P_n^{\#}$ from the empirical probability distribution P_n. For each of these draws, $P_{n1}^{\#}$, …, $P_{nB}^{\#}$, (of size n), the estimators $\hat{\Psi}_s(P_{nb}^{\#})$ based on the bootstrap sample $P_{nb}^{\#}$, $b = 1, \ldots, B$, are calculated and averaged:

$$\tilde{\Psi}_s(P_n) = \lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} \hat{\Psi}_s(P_{nb}^{\#}).$$

This results in a set of candidate bagged estimators $\tilde{\Psi}_s(P_n)$ indexed by s. Our goal is to data adaptively select the s which minimizes the risk of $\tilde{\Psi}_s(P_n)$ over $\mathcal{A}_n$. As such, we propose the *cross-validated bagged estimator* defined as:

$$\hat{\Psi}(P_n) = \hat{\Psi}_{\hat{S}(P_n)}(P_n),$$

where $\hat{S}(P_n)$ is the cross-validation selector based on the loss function $L(\cdot,\cdot)$ corresponding to a cross-validation scheme defined by a random n dimensional vector $B_n \in \{0,1\}^n$. A realization of $B_n = (B_{n,1}, \ldots, B_{n,n})$ defines a particular split of the learning sample of n observations into a training set, $\{i \in \{1, \ldots, n\} : B_{n,i} = 0\}$, and a validation set, $\{i \in \{1, \ldots, n\} : B_{n,i} = 1\}$. We will denote the proportion of observations in the validation set with p. The empirical distributions of the training and validation sets are denoted by P_{n,B_n}^0 and P_{n,B_n}^1, respectively. Formally, the cross-validation selector $\hat{S}(P_n)$ is defined as:

$$\begin{aligned} \hat{S}(P_n) &= \arg\min_{s\in\mathcal{A}_n} E_{B_n} P_{n,B_n}^1 L(\cdot, \tilde{\Psi}_s(P_{n,B_n}^0)) \qquad (1) \\ &= \arg\min_{s\in\mathcal{A}_n} E_{B_n} \sum_{i, B_n(i)=1} L(O_i, \tilde{\Psi}_s(P_{n,B_n}^0)). \end{aligned}$$

At the first equality we used the notation $Pf \equiv \int f(o)dP(o)$. If the loss function depends on an unknown nuisance parameter, then one estimates the unknown loss function on the training sample: that is, one replaces in (1) the loss function by $L(\cdot, \tilde{\Psi}_s(P_{n,B_n}^0) \mid \hat{\Upsilon}(P_{n,B_n}^0))$.

To calculate this selector of s, for each possible realization of the sample split B_n and index $s \in \mathcal{A}_n$, B bootstrap samples $P_{n,B_n,b}^{0\#}$ of size $n(1-p)$ are drawn from the training sample P_{n,B_n}^0, $b = 1, \ldots, B$. For each of these

B bootstrap samples we compute the corresponding s-specific estimators $\hat{\Psi}_s(P^{0\#}_{n,B_n,b})$ and average them to obtain:

$$\tilde{\Psi}_s(P^0_{n,B_n}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Psi}_s(P^{0\#}_{n,B_n,b}).$$

The empirical risk of this estimator over the validation sample can now be computed, and averaged over the different splits B_n, as in (1), which results in the so called cross-validated risk of the estimator $\tilde{\Psi}_s(P_n)$. The cross-validation selector is defined as the one which minimizes this cross-validated risk over $s \in \mathcal{A}_n$.

2.1 Contrasting cross-validated bagged learning with bagged cross-validated learning.

It is of interest to contrast this estimator to the bagged cross-validated estimator as used in Breiman (1996a), and followed by other authors. In the current approach, the selection of s via cross-validation is performed within each bootstrap sample. Subsequently the B bootstrap-specific estimators are averaged to arrive at the final estimator. Formally, within a bootstrap sample $P_n^{\#}$ the cross-validation selector of s can be defined as:

$$\hat{S}_{br}(P_n^{\#}) = \arg\min_{s \in \mathcal{A}_n} E_{B_n} P^{1,\#}_{n,B_n} L(\cdot, \hat{\Psi}_s(P^{0,\#}_{n,B_n})).$$

The corresponding estimator based on a single bootstrap sample $P_n^{\#}$ is thus defined as:

$$\hat{\Psi}_{CV}(P_n^{\#}) = \hat{\Psi}_{\hat{S}_{br}(P_n^{\#})}(P_n^{\#}).$$

Finally, the corresponding bagged estimator is the average over a large collection of bootstrap-specific estimators:

$$\tilde{\Psi}_{br}(P_n) = E(\hat{\Psi}_{CV}(P_n^{\#}) \mid P_n).$$

Using cross-validation within a bootstrap sample provides the right selection among the estimators $\hat{\Psi}_s(P_n^{\#})$, $s \in \mathcal{A}_n$ regarding the trade off between bias and variance. However, one would expect it not to perform the right trade-off between bias and variance for the actual bagged estimators $\tilde{\Psi}_s(P_n)$. The reason for this is that the bagged estimator should be less variable as a result of the averaging, and might be more biased due to the double sampling. An increase in bias is due to two (probably cumulative) sources: first, the bias introduced by applying an estimator to a bootstrap sample relative to the empirical sample; and second, the bias introduced by applying the estimator to the empirical sample relative to the truth. In general, the estimators $\hat{\Psi}_s$, $s \in \mathcal{A}_n$, and $\tilde{\Psi}_s$, $s \in \mathcal{A}_n$, are very different classes of estimators, so that a good selector among the un-bagged estimators is not necessarily a good selector among the corresponding bagged estimators.

2.2 Performance of the cross-validation selector.

Let $d(\psi, \psi_0) = \int L(o, \psi) dP_0(o) - \int L(o, \psi_0) dP_0(o)$ denote the risk dissimilarity between a candidate ψ and the true ψ_0 implied by the loss function $L(\cdot, \cdot)$. The results on the cross-validation selector (see van der Laan et al. (2003, 2004)) state that if the loss function is uniformly bounded in its arguments, then the difference of the risk dissimilarity of the cross-validated selected bagged estimator and the risk dissimilarity of the oracle selected bagged estimator is of the order $\log K(n)/np$ plus possibly a term due to estimation of the nuisance parameter in the loss function. The oracle selected bagged estimator is defined as $\hat{\Psi}_{\tilde{S}_{n(1-p)}(P_n)}(P_n)$, where

$$\tilde{S}_{n(1-p)}(P_n) = \arg\min_s E_{B_n} \int L(o, \tilde{\Psi}_s(P^0_{n,B_n})) dP_0(o).$$

Thus for a given data set the oracle selector $\tilde{S}_{n(1-p)}(P_n)$ selects the bagged estimator (based on $n(1-p)$ observations) closest to the truth w.r.t. to the risk dissimilarity.

These results only rely on the loss function to be uniformly bounded in the support of O and the parameter space. They imply that if the number of candidate estimators is polynomial in sample size (and, in the case that the loss function is unknown, that it can be estimated at a better rate than the convergence rate of the oracle selected estimator), then either the cross-validated selected estimator is asymptotically equivalent (up to the constant) to the oracle selected estimator, or it achieves the essentially parametric rate of convergence $\log n/n$.

For most risk dissimilarities one can bound the risk dissimilarity between the bagged estimator $\tilde{\Psi}_s(P_n) = 1/B \sum_b \hat{\Psi}_s(P^{\#}_{n,b})$ and ψ_0 in terms of risk dissimilarity of the un-bagged estimator. In particular, if $d(\psi, \psi_0) = ||\psi - \psi_0||$

for some norm $||\cdot||$, then it follows from the triangle inequality that

$$\begin{aligned} d(\tilde{\Psi}_s(P_n),\psi_0) &\leq \frac{1}{B}\sum_{b=1}^{B} ||\hat{\Psi}_s(P^{\#}_{n,b}) - \hat{\Psi}_s(P_n)|| + ||\hat{\Psi}_s(P_n) - \psi_0|| \\ &= \frac{1}{B}\sum_{b=1}^{B} d\left(\hat{\Psi}_s(P^{\#}_{n,b}), \hat{\Psi}_s(P_n)\right) + d(\hat{\Psi}_s(P_n),\psi_0). \end{aligned}$$

Since the first term on the right-hand side is only affected by the variance of the estimator $\hat{\Psi}_s$ at data generating distribution P_n, one would expect that the second term $d(\hat{\Psi}_s(P_n),\psi_0)$ dominates, and thereby that the rate of convergence of $\hat{\Psi}_s(P_n)$ to ψ_0 w.r.t. the norm $||\cdot||$ implies the same rate of convergence for the corresponding bagged estimator. Thus asymptotic consistency rate results in terms of risk dissimilarity for our proposed cross-validated bagged estimator are implied by asymptotic consistency rate results for the original estimator $\hat{\Psi}_s$. Of course, these calculations tell us little about relative finite sample and asymptotic efficiency between an un-bagged and corresponding bagged estimator.

2.3 Cross-validation selection of the degree of bagging.

In cases where there is a concern that the bagging operation might actually worsen the estimator it is a good idea to let cross-validation select between the original un-bagged estimator and the bagged estimator. In general, the following method might be of interest. Define $\tilde{\Psi}_{s,\alpha} = \alpha\tilde{\Psi}_s + (1-\alpha)\hat{\Psi}_s$ as the weighted average between the bagged and un-bagged estimator, $\alpha \in [0,1]$, and use cross-validation to select (s,α). In this manner, the data are used decide to what degree α the bagging operation should be used, and, by our results establishing asymptotic equivalence with the oracle selector of (α, s), our cross-validated selected estimator will perform asymptotically at least as well as the non-bagged estimator and bagged estimator.

2.4 Assessing the performance of the cross-validated bagged estimator.

One can estimate the risk $\int L(o,\tilde{\Psi}(P_n))dP_0(o)$ of the cross-validated bagged estimator $\tilde{\Psi}(P_n)$ with the cross-validated risk of the estimator $\tilde{\Psi}(P_n) = \tilde{\Psi}_{\hat{S}(P_n)}(P_n)$ using a cross-validation scheme defined by a random vector $B_n^* \in$

$\{0,1\}^n$:

$$E_{B_n^*} \sum_{i, B_n^*(i)=1} L(O_i, \tilde{\Psi}(P^0_{n,B_n^*})).$$

This procedure would require carrying out a B_n-specific cross-validation scheme within each learning sample P^0_{n,B_n^*}, which is often referred to as double cross-validation.

3 Variable importance, inference, and p-values.

Suppose that the parameter space $\mathbf{\Psi}$ consists of d-variate real valued functions $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$. We refer to these functions as functions of a d-dimensional vector W, and we wish to define a measure of variable importance for each variable W_j, $j = 1, \ldots, d$. We define the following function of W and w, j:

$$W_{-j}(w) \equiv (W_1, \ldots, W_{j-1}, W_j = w, W_{j+1}, \ldots, W_d).$$

That is, $W_{-j}(w)$ equals W with its j-th component $W(j)$ set equal to w. In addition, let $W_{-j} = (W_1, \ldots, W_{j-1}, W_{j+1}, \ldots, W_d)$ be the $(d-1)$-dimensional vector with $W(j)$ deleted. Given the true function ψ_0, consider the following variable importance function for variable j:

$$\nu_j(P_0)(w) \equiv \frac{d}{dw} E_{W_{-j}} \psi_0(W_{-j}(w)).$$

That is, $\nu_j(w \mid P_0)$ is defined as the j-th partial derivative of ψ_0 at w averaged over the remaining variables W_{-j} w.r.t. to some specified (e.g., user supplied) probability distribution for W_{-j}.

If W_j is a continuous variable, then the partial derivative is the usual derivative, and if $W_j \in \{1, \ldots, K_j\}$ is an ordered categorical variable, then the derivative at $w = k$ just refers to the difference between ψ_0 at $W_j = k+1$ and ψ_0 at $W_j = k$. The variable importance $\nu_j(w \mid P_0)$ is a function in w. It can be summarized in a real valued number by averaging its absolute values w.r.t. to some specified probability distribution G_0 on the set of possible values for the j-th variable W_j:

$$\bar{\nu}_j(P_0) \equiv \int \mid \frac{d}{dw} E_{W_{-j}} \psi_0(W_{-j}(w)) \mid dG_0(w).$$

These variable importance measures were proposed in Sinisi and van der Laan (2004) in the context of prediction.

These variable importance parameters can be straightforwardly generalized to importance parameters for a joint subset of variables by replacing the single partial derivatives by joint first order partial derivatives.

The variable importance measure $\nu_j(P_0)(w)$ measures the effect of a change of W_j at w on ψ_0. In particular, if $\psi_0(W) = E(Y \mid W)$, then this variable importance corresponds to the so called G-computation formula for the marginal causal effect of an intervention $W_j = w$ to $W_j = w + 1$ (e.g., W_j is categorical) on the outcome Y, assuming a causal counterfactual model: see Sinisi and van der Laan (2004). The function $\nu_j(P_0)(\cdot)$ can be plotted to observe any fluctuations in importance of variable W_j over its range. For illustrations of this variable importance function in the context of classification and regression trees we refer to Molinaro and van der Laan (2005).

The cross-validated bagged estimator $\tilde{\Psi}(P_n)$ can be mapped into estimates of these variable importance measures, by substitution:

$$\hat{\nu}_j(P_n)(w) \equiv \frac{d}{dw} E_{W_{-j}} \tilde{\Psi}(P_n)(W_{-j}(w)).$$

Since the cross-validated bagged estimator $\tilde{\Psi}(P_n)$ will typically depend on most of the d variables, this collection of estimators of variable importance functions $\hat{\nu}_j(P_n)$, $j = 1, \ldots, d$, provides an interesting list of output for a data analysis. Alternative variable importance measures were originally introduced in (Breiman, 2001a) for bagged prediction based on CART.

3.1 Bootstrap based inference for variable importance measures.

In this subsection we will present a bootstrap method for assessing the standard error of the estimator

$$\hat{\nu}_{j,s}(P_n)(w) \equiv \frac{d}{dw} E_{W_{-j}} \tilde{\Psi}_s(P_n)(W_{-j}(w)),$$

based on the s-specific bagged estimator $\tilde{\Psi}_s(P_n) \equiv E(\hat{\Psi}_s(P_n^{\#}) \mid P_n)$ for a given s. We suggest reporting these inference results with $s = \hat{S}(P_n)$, where we have to acknowledge that these results do not take into account that s is selected data adaptively. Since variable importance is typically a much smoother functional of the data generating distribution than ψ_0, the variability due to data adaptive selection of s might be small relative to the actual variance of the variable importance measure.

Real valued linear summary measures of our parameter.

In order to simplify notation and to illustrate the generality, let $\tilde{V}_s(P_n)$ represent a real valued linear summary measure of the bagged estimator $\tilde{\Psi}_s(P_n)$, such as an evaluation or difference at a value w of $\hat{\nu}_{j,s}(P_n)$. Let $\hat{V}_s(P_n)$ represent the same real valued linear summary measure of the non-bagged estimator $\hat{\Psi}_s(P_n)$. We consider $\hat{V}_s(P_n)$ as an estimator of the underlying real valued linear summary measure $V(P_0)$ of the true ψ_0. Due to the fact that $\tilde{V}_s(P_n)$ is a linear function of the bagged estimator $\tilde{\Psi}_s(P_n)$, we can represent $\tilde{V}_s(P_n)$ as an s-specific bagged estimator corresponding with the un-bagged estimator $\hat{V}_s(P_n)$:

$$\tilde{V}_s(P_n) = E(\hat{V}_s(P_n^{\#}) \mid P_n).$$

Estimation of the variance.

We are concerned with estimating the variance of $\tilde{V}_s(P_n)$ with a particular re-sampling method. Firstly, we note that bootstrapping the estimate $\tilde{V}_s(P_n)$ would require bootstrapping a procedure which itself requires bootstrapping. Therefore, we are concerned with developing a bootstrap method which circumvents the use of this double bootstrap.

Our proposal is based on the Theorem of Pythagoras in the Hilbert space of functions of $(P_n, P_n^{\#})$ endowed with inner product being the covariance operator. In this Hilbert space the conditional expectation $E(Y \mid X)$ represents the projection of Y onto the subspace of functions of X, and the Theorem of Pythagoras states

$$\text{VAR}(E(Y \mid X)) = \text{VAR}(Y) - \text{EVAR}(Y \mid X).$$

Application of this result with $Y = \hat{V}_s(P_n^{\#})$, and $X = P_n$, teaches us that

$$\sigma^2 \equiv \text{VAR}\tilde{V}_s(P_n) = \text{VAR}\hat{V}_s(P_n^{\#}) - E\text{VAR}(\hat{V}_s(P_n^{\#}) \mid P_n).$$

We will use this representation of the variance of our bagged estimator $\tilde{V}_s(P_n)$ to propose an estimator of this variance. First, we estimate $E\text{VAR}(\hat{V}_s(P_n^{\#}) \mid P_n)$ with $\text{VAR}(\hat{V}_s(P_n^{\#}) \mid P_n)$, which can be directly computed as a by-product of our computation of $\tilde{V}_s(P_n)$, since the latter already involved sampling many realizations $\hat{V}_s(P_n^{\#})$, conditional on the data P_n. It remains to present an estimator of the marginal variance of $\hat{V}_s(P_n^{\#})$. We note that the experiment defining the latter random variable is given by 1) sampling a sample P_n from the true data generating distribution P, 2) taking a sample $P_n^{\#}$ from P_n, and 3) evaluating $\hat{V}_s(P_n^{\#})$. Therefore, given an estimate $\tilde{P}_n$ of the true data generating distribution P, the bootstrap version of this experiment is defined by 1) sampling an empirical distribution $\tilde{P}_n^*$ of n observations from $\tilde{P}_n$, 2) sampling an empirical distribution $\tilde{P}_n^{\#}$ from $\tilde{P}_n^*$, and 3) evaluating $\hat{V}_s(\tilde{P}_n^{\#})$. In order to avoid a very large number of ties in the double bootstrapped empirical distribution $\tilde{P}_n^{\#}$, we propose to sample $\tilde{P}_n^*$ from a smooth $\tilde{P}_n$, or use a weighted (Bayesian) bootstrap in which one samples for each observation O_i a weight from (e.g.) an exponential distribution with mean $1/n$ (Rubin, 1981). In this manner one guarantees that $\tilde{P}_n^*$ is a probability distribution with support being the whole sample $\{O_1, \ldots, O_n\}$.

Given this bootstrap experiment, one can now estimate the variance of $\hat{V}_s(P_n^{\#})$ by simply simulating a large number B of replicates $\hat{V}_s(\tilde{P}_{n,b}^{\#})$, $j = 1, \ldots, B$. We note that the latter bootstrap method for estimating the variance of $\hat{V}_s(P_n^{\#})$ is not more computer intensive than a single bootstrap since it only involves calculating the estimator $\hat{V}_s$ once for each draw of $\tilde{P}_n^{\#}$, and drawing $\tilde{P}_n^{\#}$ is not more computer intensive than drawing $P_n^{\#}$.

We conclude that we can estimate σ^2 with the following bootstrap-based variance estimate:

$$\sigma_n^2 \equiv \frac{1}{B} \sum_{b=1}^{B} \left(\hat{V}_s(\tilde{P}_{n,b}^{\#}) - 1/B \sum_{b=1}^{B} \hat{V}_s(\tilde{P}_{n,b}^{\#}) \right)^2 - \frac{1}{B} \sum_{b=1}^{B} \left(\hat{V}_s(P_{n,b}^{\#}) - \tilde{V}_s(P_n) \right)^2 .$$

Recall that $\tilde{V}_s(P_n) = 1/B \sum_{b=1}^{B} \hat{V}_s(P_{n,b}^{\#})$, $P_{n,b}^{\#}$ is a bootstrap sample from our data P_n, and $\tilde{P}_{n,b}^{\#}$ represents a draw from the randomly drawn $\tilde{P}_n^*$.

To summarize, our variance estimator is calculated as follows:

- Sample B bootstrap samples $P_{n,b}^{\#}$ from our data P_n, $b = 1, \ldots, B$. Calculate the corresponding $x_1 = (\hat{V}_s(P_{n,b}^{\#}), b = 1, \ldots, B)$.

- Sample a weighted empirical distribution $\tilde{P}_{n,b}^*$ corresponding with P_n (Bayesian bootstrap) and, given $\tilde{P}_{n,b}^*$, sample a standard empirical distribution $\tilde{P}_{n,b}^{\#}$ from $\tilde{P}_{n,b}^*$, $b = 1, \ldots, B$. Calculate the corresponding $x_2 = (\hat{V}_s(\tilde{P}_{n,b}^{\#}) : b = 1, \ldots, B)$.

- Given these two B dimensional vectors x_1, x_2, compute

$$\sigma_n^2 = \frac{1}{B} \sum_{b=1}^{B} (x_1(b) - \bar{x}_1)^2 - \frac{1}{B} \sum_{b=1}^{B} (x_2(b) - \bar{x}_2)^2 .$$

3.2 Obtaining a P-value.

In this subsection we present the analogue of our bootstrap method described above for obtaining a p-value for testing the null hypothesis $H_0 : V_0 = 0$.

Consider a null distribution P_n^* for which it is known that the parameter $V(P_n^*) = 0$. For example, if $\psi_0(W) = E(Y|W)$, and $V(P_0)$ denotes a variable importance parameter defined above, then one can set P_n^* equal to the empirical distribution defined by 1) Y and W being independent, and 2) the marginal distributions equal to the marginal empirical distributions of Y and W, respectively. In this case, sampling n draws from P_n^* corresponds closely to a sample from the permutation distribution, where a draw from the permutation distribution corresponds to a sample $(X_i, Y_{\pi(i)})$, $i = 1, \ldots, n$, where $(\pi(1), \ldots, \pi(n))$ is a permutation of $(1, \ldots, n)$. Therefore, in this regression example one can view the permutation distribution as the null distribution for the sample of n observations.

Let P_n^0 be an empirical distribution of a sample of n i.i.d. observations from P_n^*. We are concerned with evaluating the variance $\sigma^2(P_n^*)$ of $\tilde{V}_s(P_n^0)$, given P_n^*, with a particular re-sampling method. One can compute a p-value for an observed value v of $\tilde{V}_s(P_n)$ under the normal distribution $N(0, \hat{\sigma}^2(P_n^*))$ centered at zero as follows:

$$\text{P-value}(v) = \bar{\Phi}(v/\sigma(P_n^*)).$$

Here $\bar{\Phi}$ denotes the survivor function of the standard normal.

As in the previous subsection, application of the variance formula for conditional expectations tells us that

$$\sigma^2(P_n^*) \equiv \text{VAR}_{P_n^*}\tilde{V}_s(P_n^0)) = \text{VAR}_{P_n^*}\hat{V}_s(P_n^{0\#}) - E_{P_n^*}\text{VAR}(\hat{V}_s(P_n^{0\#}) \mid P_n^0),$$

where now the variance and expectation are w.r.t. to our null distribution P_n^*.

As in the previous subsection, we can use this representation of the variance of our bagged estimator $\tilde{V}_s(P_n^0)$ at data generating distribution P_n^* to propose an evaluation of this variance $\sigma^2(P_n^*)$. First, $\text{VAR}(\hat{V}_s(P_n^{0\#}) \mid P_n^0)$, can be directly computed as a by-product of our computation of $\tilde{V}_s(P_n^0)$, since the latter already involved sampling many realizations $\hat{V}_s(P_n^{0\#})$, conditional on the data P_n^0. In order to approximate $E\text{VAR}(\hat{V}_s(P_n^{0\#}) \mid P_n^0)$ we could average this P_n^0-specific conditional variance across a number of draws P_n^0 from P_n^*. The presentation of the evaluation of the marginal variance of $\hat{V}_s(P_n^{0\#})$ is the complete analogue of the previous subsection and therefore is not repeated here.

Confidence interval.

Given an estimate of the bias (as presented below) and variance of our estimator $\tilde{V}_s(P_n)$ of the true real valued parameter of interest V_0, we could obtain inference for V_0 based on the normal working model

$$\tilde{V}_s(P_n) \sim N(V_0 + \text{Bias}_n, \sigma_n^2).$$

Under this working model, we have that

$$\tilde{V}_s(P_n) + \frac{\text{Bias}_n}{\sigma_n} \pm Z_{1-\alpha/2}\sigma_n$$

is a $1-\alpha$ confidence interval for V_0, where $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The appropriateness of assuming a normal distribution for the estimates of variable importance needs to be theoretically investigated. However, we note that the bagging provides a smoothness of the distribution of our estimator of variable importance, which thereby makes the normal model more appropriate than it would be without bagging.

Estimation of the bias.

It is of additional interest to estimate the bias of $E\tilde{V}_s(P_n) - V_0$, where $V_0 = V(\psi_0)$ represents the true real valued linear summary measure of ψ_0. We note that this bias can also be represented as

$$\text{Bias} \equiv E\hat{V}_s(P_n^{\#}) - V_0.$$

Therefore, we could estimate this bias with the following estimator:

$$\text{Bias}_n = \frac{1}{B}\sum_{b=1}^{B} \hat{V}_s(\tilde{P}_{n,b}^{\#}) - V(\Psi(\tilde{P}_n)).$$

In order to estimate the true parameter value $V(\Psi(\tilde{P}_n))$ in the world with the true data generating distribution being $\tilde{P}_n$, one could simply generate a very large sample from $\tilde{P}_n$ and fit with our cross-validated bagged estimator the true parameter $\Psi(\tilde{P}_n)$. In other words, Bias_n simply denotes the actual bias of $\hat{V}_s(P_n^{\#})$ (and thereby $\tilde{V}_s(P_n)$) at an estimated data generating distribution $\tilde{P}_n$.

4 Cross-validated bagged regression based on the DSA: Simulation and Data Analysis.

In this section we illustrate our proposed cross-validated bagged estimation methodology to estimate a regression function $E(Y \mid W)$. In Figure 1 we provide a graphical overview of the methodology.

Our candidate estimators $\hat{\Psi}_s(P_n)$ are chosen to be the Deletion/Substitution/Addition (D/S/A) polynomial regression estimators indexed by a three dimensional integer vector $s = (k_0, k_1, k_2)$, as introduced and implemented in Sinisi and van der Laan (2004). Given the three integer values (k_0, k_1, k_2), this estimator 1) computes for each of the d variables in $W \in \mathbb{R}^d$ the t-statistic for the marginal regression on Y, 2) selects the top k_0 variables ranked by this t-statistic, 3) aims to minimize, using the D/S/A algorithm, the empirical mean of squared errors over all linear regressions in maximally k_1 tensor products of polynomial powers in these selected k_0 variables, where these tensor products involve a product of maximally k_2 terms. That is, this estimator maps into linear regressions in k_1 polynomial basis functions, such as $w_3 w_4^2$ and $w_1 w_5 w_7$, where k_0 indexes a data adaptive dimension reduction and k_2 indexes a bound on the allowed complexity of the basis functions. For details about the actual minimization strategy the D/S/A algorithm follows, we refer to Sinisi and van der Laan (2004).

4.1 Simulated Example.

A motivation for obtaining a bagged estimator is to gain information about each variable in order to form a more suitable measure of variable importance, but one cost of this is computing time. To look at the relative gain of estimating variable importance measures from a bagged fit, we performed a brief simulation. The following illustrates the D/S/A un-bagged and bagged estimators on three simulated data sets. In each of the three settings, the true model is: $y = \sum_j \frac{1}{j} x_j + \varepsilon$ where $x_j \sim U(0,1)$, $\varepsilon \sim N(0, \sigma^2)$, $j = 1, \ldots, 20$, and $n = 250$. The fine-tuning parameters, k_1 and k_2, were chosen using 5-fold cross-validation, the bagged estimate is based on 200 bootstrap replications, and the simulations were repeated 10 times in each case. The variations of the three simulations are:

1. $\sigma^2 = 1$

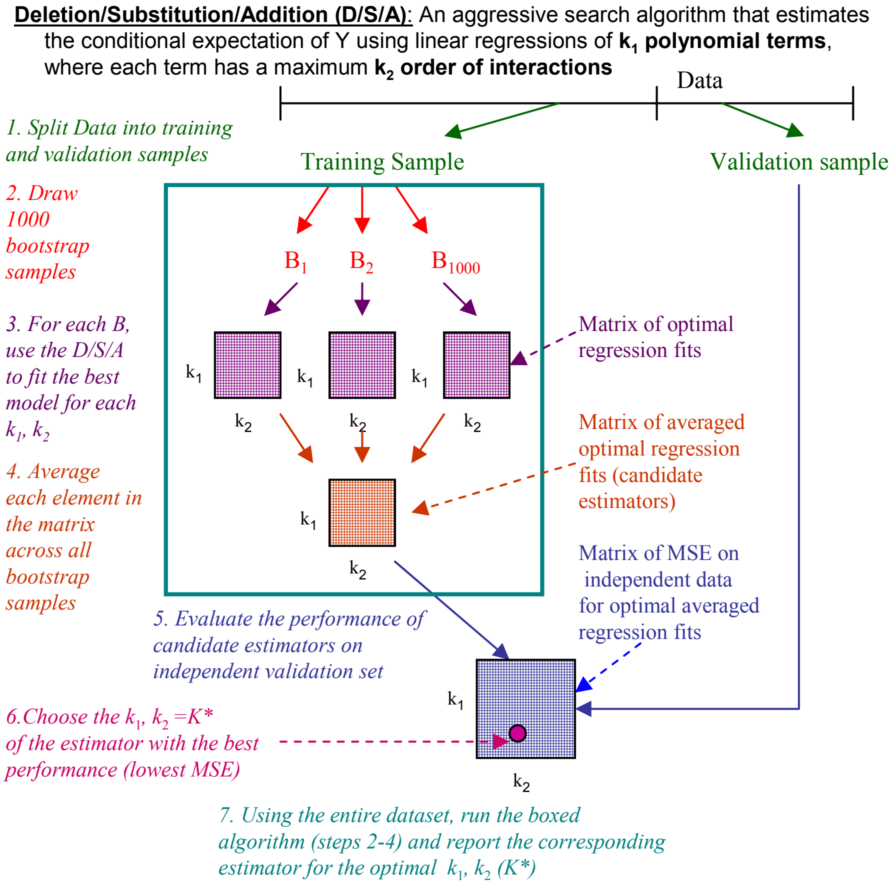
Deletion/Substitution/Addition (D/S/A): An aggressive search algorithm that estimates the conditional expectation of Y using linear regressions of k1 polynomial terms, where each term has a maximum k2 order of interactions
Data
1. Split Data into training and validation samples
Training Sample
Validation sample
2. Draw 1000 bootstrap samples
B1
B2
B1000
3. For each B, use the D/S/A to fit the best model for each k1, k2
k1
k2
Matrix of optimal regression fits
4. Average each element in the matrix across all bootstrap samples
Matrix of averaged optimal regression fits (candidate estimators)
Matrix of MSE on independent data for optimal averaged regression fits
5. Evaluate the performance of candidate estimators on independent validation set
6.Choose the k1, k2 =K* of the estimator with the best performance (lowest MSE)
7. Using the entire dataset, run the boxed algorithm (steps 2-4) and report the corresponding estimator for the optimal k1, k2 (K*)

Figure 1: Graphical representation of the cross-validated bagged D/S/A learner

2. $\sigma^2 = 0.25$

3. $\sigma^2 = 0.25$, and in addition to the 20 uniform variables that form the true model, there are 5 additional noise variables, $x \sim N(1, 1)$

In each setting, the maximal allowed order of interaction is one, meaning that only main terms were allowed to enter the fits ($k_2 = 1$). k_1 ranges from 1 to 10 in the first two settings and from 1 to 15 in the third setting. Changing k_1 from 10 to 20 did not affect the results.

The results, across the 10 simulations, are summarized in Tables 1-6. In Tables 1,3,5, the following quantities are reported: β represents the true importance measure for the corresponding variable, the mean estimated importance measure (averaged across the 10 simulations), the variance and bias of the estimated importance measures, the mean squared error (MSE) of the estimated importance measure, and the ratio of the bagged MSE to the un-bagged MSE. The RSS, R^2, and estimate of the true risk is computed for each un-bagged and bagged fit and averaged across the 10 repeats and reported in Tables 2,4,6.

It is clear that the un-bagged and bagged estimators are comparable in terms of prediction, as in each case the bagged estimator has a slight improvement in RSS and true risk over the un-bagged estimator. Our simulations results for the MSE suggest that bagging provides a better estimate than the un-bagged estimator of the variable importance measures (VIM) for variables which have a high probability of not being included in an un-bagged fit. It is also of interest to note that even for the pure noise variables $X_{21}, \ldots, X_{25}$ in simulation three, which have a true VIM of zero, the bagged estimator seems to perform well relative to the un-bagged estimator. These results are based on only 10 simulations, but we expect that these findings will hold under more simulations. In the future, we will consider a more in-depth simulation study where the number of repetitions is increased and the other fine-tuning parameters (e.g., k_2) are selected with cross-validation.

4.2 Data analysis example.

The data example is drawn from the Stanford University HIV Drug Resistance Database (http://hivdb.stanford.edu/), which contains observational clinical data from patients infected with HIV. The application has two goals: 1) **Prediction:** Construct a predictor of a patient's change in plasma HIV

RNA level (viral load) as a function of his/her treatment history, current drug regimen, and mutations in the HIV virus infecting him/her. This predictor would then allow identification of a treatment regimen likely to result in the biggest decrease in a given patient's viral load (best treatment response). 2) **Variable Importance:** Estimate variable importance measures for viral mutations in the HIV strains infecting a patient. Under assumptions, these variable importance measures can be interpreted as summaries of the effect of mutations on clinical virologic response to specific antiretroviral drugs and drug combinations.

Data structure.

The data has the following structure:

1. viral genotype, summarized as the presence or absence of each viral mutation considered by the Stanford scoring system to have some effect on virologic response to antiretroviral therapy (see http://hivdb.stanford.edu/)

2. treatment regimen initiated following assessment of viral genotype, which might involve changing some or all of the drugs in a patient's previous regimen

3. change in viral load over baseline, measured at 11-36 weeks after beginning the current treatment regimen and while still on this regimen

4. treatment history prior to initiating the current regimen

The observed data structure on a subject is written as (W, A, Z, Y) where W represents the treatment history and baseline viral load, $A \in \{0,1\}^{71}$ represents the binary mutation profile of the virus (genotype), $Z \in \{0,1\}^{22}$ is the treatment regimen assigned to the patient, and Y is the patient's change in log viral load.

The data were stratified based on individual drugs or common treatment regimens. In particular, for this illustration we focus on a group of 295 patients who received the following three antiretroviral drugs exclusively: zidovudine (AZT), lamivudine (3TC), and indinavir (IDV). The data structure now reduces to (W, A, Y) because Z is the same for these 295 patients. We consider 16 baseline variables (W) that are considered potential confounders of the two drug classes in this particular regimen, Protease Inhibitors (PI) and Nucleoside Reverse Transcriptase Inhibitors (NRTI), and 48 mutations

(A) that are thought to affect viral resistance to at least one drug in either of these two classes.

As described above, the D/S/A algorithm was used (Sinisi and van der Laan, 2004) to estimate $E(Y|W, A)$. The fine-tuning parameters (k_0, k_1, k_2) were set at (64,5,3), where k_1 and k_2 were selected via cross-validation, allowing k_1 to range between 1 and 5 and k_2 to range between 1 and 3. The bagged estimator was based on 1000 bootstrap replications.

Prediction.

In order to select the best model for the purpose of prediction, the bagged and un-bagged estimators indexed by the choice of (k_1, k_2) corresponding to the estimator with the lowest cross-validated risk were selected. The following un-bagged estimator was selected: $-2.6106 + 1.9015w_6$ ($k_0 = 64, \hat{k}_1 = 1, \hat{k}_2 = 1$) where w_6 denotes "PI.fail". This fit had a low $R^2 = 0.043$ and a cross-validated risk of 176.7.

The cross-validated risks for the bagged estimators corresponding to different choices of k_1 and k_2 are given in Table 7. The cross validation-selected bagged estimator corresponding to the choice of $\hat{k}_1 = 1, \hat{k}_2 = 2$ consisted of the average across bootstrap replications of the *best* predictor of size one (and maximum order of interactions two) in each bootstrap replication. This produced a rather low-dimensional aggregated predictor with 114 terms, an R^2 of 0.056, and a cross validated risk of 176.9. Of note, the bagging procedure did not result in an improvement in cross-validated risk.

The cross-validated risk can be used to provide a rough estimate of the standard error of the predictor:

$$\begin{aligned} s.\hat{e}. = \sqrt{\text{cross-validated risk/size of validation set}} \\ &= \sqrt{176.9/59} \\ &= 1.7 \end{aligned}$$

Considering that the outcome (change in log viral load) ranges from -4.2 to 2.1, the cross-validated risk suggests that even the true optimal predictor, $E(Y|W, A)$, has little predictive power, due to large variance residuals. The high variability of the residuals explains the selection of a low dimensional k_1, as the estimator is forced to reduce the large variance at a high cost to bias. Given the cross validated risk, as well as the low R^2, we conclude that prediction may not represent a reasonable goal to pursue with this data set.

Variable Importance.

An additional goal of this application is to identify mutations with effects on virologic response and to estimate their variable importance. As discussed above, cross validated risks were similarly poor across all combinations of k_1 and k_2. As a result, in choosing k_1 and k_2 to obtain estimates of variable importance, we chose to use a more over-fitted bagged estimator corresponding with $k_1 = 5, k_2 = 3$ for the following reasons: 1) As variable importance is a much smoother parameter than $E(Y|W, A)$, in estimation of variable importance the variance of the predictor becomes less of a concern, while bias becomes a greater concern. 2) Treatment history variables are included as variables in the regression for this application not to improve prediction, but because they have the potential to confound the effect of mutations on response; past treatments can affect which mutations are observed at baseline, and also have independent effects on virologic response (for example, as a result of archived mutations which are not measured but nonetheless contribute to resistance). Control of this confounding requires that the multi-variable regression fits contain more than one term each. 3) Use of a richer bagged estimator allowed us to identify as many mutations with non-zero variable importance as possible.

The final bagged estimator corresponding with $k_1 = 5, k_2 = 3$ has 1266 terms. The R^2 for this predictor is 0.14. Several points regarding interpretation of this estimator and the corresponding variable importance measures deserve discussion.

Before considering the mutation effects on virologic response identified by our estimator, it is worth noting the obvious, but nonetheless important, fact that mutations must occur with reasonable frequency in the sample for their effects on virologic response to be identified. Several major mutations known to confer significant resistance to drugs in the regimen (for example, the known AZT-resistance mutations Q151L/M and Y115F and IDV-resistance mutations I84C and V82A/F/S/T) occur only once or not at all in the sample. Consequently, several major resistance mutations do not appear in the final bagged estimator and have variable importance measures of zero. Table 8 shows the frequency of PI- and NRTI-associated mutations in the sample.

Table 9 shows the mutations and mutation interactions with the largest effect on virologic response, as identified from the results of the final bagged estimator after integrating out treatment history variables $(\widehat{E_W}(E(Y|A = a, W) - E(Y|A = 0, W)))$. Mutations with coefficients that have an absolute value greater than 0.05 are shown. The coefficients can be interpreted as the expected change in log viral load associated with the presence of the corresponding mutation or mutation interaction. Note that a positive coefficient implies that the corresponding mutation results in an increase in viral load over its baseline level, or in other words, contributes to a poor treatment response. The corresponding variable importance measures for mutations identified by the bagged estimator are reported in Tables 10 and 11.

Several of the mutations known to confer significant resistance to at least one of the three drugs with which the patients in the sample are treated are identified by the bagged estimator and corresponding variable importance measures. For example, both the L90M and M46I/L/V mutations confer significant resistance to IDV, while others, such as L10F/I/R/V may act as important accessory mutations, improving viral fitness and increasing resistance to IDV in the presence of additional PI-associated mutations. Similarly, the M184I/V mutation is the primary mutation conferring resistance to 3TC, while D67E/G/N, K70R/G/E, L210W, and T215F/Y are thymidine analog mutations, a class of mutations that contribute resistance to AZT.

However, while some of the mutations and interactions identified by the bagged estimator and reflected in the variable importance measures are supported by previous research and mechanistic understanding, the significance of others is less clear. While it is possible that some of these may represent effects on viral resistance to the AZT+3TC+IDV regimen that have not previously been identified in the literature, there are several alternative explanations that reflect the complexity of dealing with observational clinical data.

1) The presence of viral mutations may be acting to some extent as a surrogate marker for the ability of patients to adhere to prescribed regimens. As adherent patients infected with resistant virus generally do better than patients who do not take their drugs at all, resistance mutations in this context may be associated with an improved virologic outcome. Unfortunately, no data on patient adherence are currently available in this data-set.

2) Mutations can have complex effects on the virus; a given mutation may simultaneously increase resistance and decrease viral fitness, contributing to an escape from complete virologic suppression (and thus an increase in viral

load) in some individuals, while resulting in an impaired ability of the virus to replicate (and thus a slight decrease in viral load) in individuals with ongoing viral replication.

3) There is a strong potential for selection bias in the sample. Specifically, in this observational database, clinicians assign patients to drug regimens that they feel are likely to be effective, based on the patient's current viral mutation profile and treatment history. This has resulted in under-representation of mutations known to cause resistance to the current regimen of interest among patients treated with that regimen, limiting the ability of these mutations to be detected, as discussed above. In addition, selective assignment of drug regimen can cause bias in estimates of the effects of well-represented mutations, if, for example, patients with a mutation and expected to respond well are over-represented among the people receiving the regimen, and patients with the mutation expected to respond poorly are under-represented.

4) Finally, although summaries of treatment history were included in the estimator, it is possible that residual confounding remains.

Due to these limitations, we believe the results presented should be interpreted with extreme caution. In the future we plan to perform several additional analyses aimed at addressing some of the issues raised. In addition, we hope to apply the algorithm to laboratory-based data with the aim of predicting viral resistance (measured *in vitro* as drug-specific viral phenotype) based on viral mutations. While retaining the same general structure, this related application would greatly simplify the identification of mutation effects on resistance by eliminating the issues raised above. Nonetheless, the current data example provides an illustration of the power of the algorithm to data adaptively identify a high-dimensional estimator, with minimal *a priori* assumptions about which variables to include and about the appropriate complexity of the component regression models.

References

M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL `http://www.bepress.com/sagmb/vol4/iss1/art8`. Article 8.

S. Borra and A. Di Ciaccio. Improving nonparametric regression methods by bagging and boosting. *Computational Statistics and Data Analysis*, 38: 407–420, 2002.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996a.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383, 1996b.

L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996c.

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

L. Breiman. Using iterated bagging to debias regressions. *Machine Learning*, 45:261–277, 2001b.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. The Wadsworth Statistics/Probability series. Wadsworth International Group, 1984.

P. Buhlmann and B. Yu. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2002.

A. Buja and W. Stuetzle. Observations on bagging, 2002. URL `http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-bag-wxs.pdf`.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Technical Report 126, Division of Biostatistics, University of California, Berkeley, Feb. 2003. URL `www.bepress.com/ucbbiostat/paper126/`.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991. Discussion by A. R. Barron and X. Xiao.

J. H. Friedman and P. Hall. On bagging and nonlinear estimation, 2000. URL http://www-stat.stanford.edu/~jhf/ftp/bag.ps.

P. Hall and R. J. Samworth. Properties of bagged nearest-neighbour classifiers. *Journal of the Royal Statistical Society: Series B*, 2005. To appear.

T. Hothorn and B. Lausen. Bundling classifiers by bagging trees. *Computational Statistics and Data Analysis*, 2003.

S. Keleş, M. J. van der Laan, and S. Dudoit. Asymptotically optimal model selection method with right censored outcomes. Technical Report 124, Division of Biostatistics, University of California, Berkeley, Sept. 2003. URL www.bepress.com/ucbbiostat/paper124/.

M. LeBlanc and R. Tibshirani. Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91:1641–1650, 1996.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. Technical Report 135, Division of Biostatistics, University of California, Berkeley, Sept. 2003. URL www.bepress.com/ucbbiostat/paper135/.

A. M. Molinaro and M. J. van der Laan. Cross-validated bagged partitioning estimators. Technical report, Division of Biostatistics, University of California, Berkeley, 2005. In preparation.

D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003. URL www.biostat.jhsph.edu/~iruczins/publications/publications.html.

S. E. Sinisi and M. J. van der Laan. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL http://www.bepress.com/sagmb/vol3/iss1/art18. Article 18.

M. Skurichina and R. P. W. Duin. Bagging for linear classifiers. *Pattern Recognition*, 31:909–930, 1998.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, Nov. 2003. URL `www.bepress.com/ucbbiostat/paper130/`.

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood based cross-validation. Technical Report 125, Division of Biostatistics, University of California, Berkeley, Feb. 2003. URL `www.bepress.com/ucbbiostat/paper125/`.

M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The cross-validated adaptive epsilon-net estimator. Technical Report 142, Division of Biostatistics, University of California, Berkeley, February 2004. URL `www.bepress.com/ucbbiostat/paper142/`.

M. J. van der Laan and J. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, 2003.

Table 1: Simulation One, un-bagged vs. bagged estimator, $\sigma^2 = 1$

x	β	mean		var		bias		MSE		ratio
1	1	0.9015	0.7893	0.0471	0.0766	-0.0985	-0.2107	0.0521	0.1133	2.175
2	0.5	0.2894	0.3089	0.1492	0.0895	-0.2106	-0.1911	0.1787	0.1170	0.655
3	0.3333	0.0499	0.0771	0.0249	0.0053	-0.2834	-0.2562	0.1027	0.0704	0.685
4	0.25	0.0000	0.0812	0.0000	0.0142	-0.2500	-0.1688	0.0625	0.0413	0.661
5	0.2	0.0698	0.0994	0.0487	0.0332	-0.1302	-0.1006	0.0608	0.0400	0.658
6	0.1667	0.0000	0.0190	0.0000	0.0078	-0.1667	-0.1476	0.0278	0.0288	1.037
7	0.1429	0.0000	-0.0104	0.0000	0.0151	-0.1429	-0.1532	0.0204	0.0371	1.818
8	0.125	0.0000	0.0359	0.0000	0.0039	-0.1250	-0.0891	0.0156	0.0114	0.732
9	0.1111	0.0000	0.0070	0.0000	0.0021	-0.1111	-0.1041	0.0123	0.0127	1.032
10	0.1	0.0538	0.0304	0.0290	0.0113	-0.0462	-0.0696	0.0282	0.0150	0.533
11	0.0909	0.0450	0.0436	0.0202	0.0122	-0.0459	-0.0473	0.0203	0.0132	0.650
12	0.0833	0.0000	0.0489	0.0000	0.0057	-0.0833	-0.0344	0.0069	0.0063	0.908
13	0.0769	0.0000	0.0101	0.0000	0.0134	-0.0769	-0.0668	0.0059	0.0165	2.792
14	0.0714	0.0000	0.0161	0.0000	0.0018	-0.0714	-0.0553	0.0051	0.0047	0.924
15	0.0667	0.0000	0.0273	0.0000	0.0019	-0.0667	-0.0394	0.0044	0.0033	0.732
16	0.0625	0.0545	0.0238	0.0297	0.0078	-0.0080	-0.0387	0.0268	0.0085	0.319
17	0.0588	0.0000	0.0196	0.0000	0.0014	-0.0588	-0.0392	0.0035	0.0028	0.808
18	0.0556	0.0360	0.0152	0.0130	0.0022	-0.0196	-0.0404	0.0120	0.0036	0.299
19	0.0526	0.0000	0.0224	0.0000	0.0011	-0.0526	-0.0302	0.0028	0.0019	0.679
20	0.05	0.0578	0.0323	0.0334	0.0188	0.0078	-0.0177	0.0301	0.0172	0.571

Table 2: Simulation One - Summary Measures

averages	un-bagged	bagged
RSS	254.5	247.1
R^2	0.093	0.120
true risk est	1.059	1.049

Table 3: Simulation Two, un-bagged vs. bagged estimator, $\sigma^2 = 0.25$

x	β	mean		var		bias		MSE		ratio
1	1	0.9752	0.9708	0.0067	0.0063	-0.0248	-0.0292	0.0066	0.0065	0.985
2	0.5	0.5171	0.5035	0.0137	0.0200	0.0171	0.0035	0.0126	0.0180	1.430
3	0.3333	0.2722	0.1962	0.0310	0.0259	-0.0612	-0.1372	0.0316	0.0422	1.334
4	0.25	0.1981	0.1922	0.0330	0.0247	-0.0519	-0.0578	0.0324	0.0256	0.789
5	0.2	0.1310	0.1171	0.0332	0.0270	-0.0690	-0.0829	0.0346	0.0312	0.901
6	0.1667	0.1531	0.1625	0.0424	0.0355	-0.0135	-0.0042	0.0383	0.0320	0.834
7	0.1429	0.0816	0.0675	0.0176	0.0077	-0.0613	-0.0754	0.0196	0.0126	0.646
8	0.125	0.0644	0.0675	0.0187	0.0124	-0.0606	-0.0575	0.0205	0.0145	0.708
9	0.1111	0.0789	0.0630	0.0165	0.0065	-0.0322	-0.0481	0.0159	0.0081	0.510
10	0.1	0.0547	0.0533	0.0133	0.0074	-0.0453	-0.0467	0.0140	0.0088	0.628
11	0.0909	0.0672	0.0552	0.0129	0.0062	-0.0237	-0.0357	0.0122	0.0068	0.561
12	0.0833	0.0530	0.0352	0.0144	0.0086	-0.0303	-0.0481	0.0139	0.0101	0.725
13	0.0769	0.0223	0.0274	0.0050	0.0024	-0.0546	-0.0495	0.0075	0.0046	0.620
14	0.0714	0.0000	0.0205	0.0000	0.0009	-0.0714	-0.0509	0.0051	0.0034	0.665
15	0.0667	0.0395	0.0333	0.0072	0.0018	-0.0272	-0.0334	0.0072	0.0027	0.375
16	0.0625	-0.0613	-0.0221	0.0191	0.0129	-0.1238	-0.0846	0.0325	0.0188	0.578
17	0.0588	0.0000	0.0071	0.0000	0.0017	-0.0588	-0.0518	0.0035	0.0042	1.216
18	0.0556	0.0261	0.0507	0.0068	0.0085	-0.0294	-0.0048	0.0070	0.0077	1.098
19	0.0526	0.0000	-0.0070	0.0000	0.0009	-0.0526	-0.0596	0.0028	0.0044	1.571
20	0.0500	0.0218	0.0235	0.0047	0.0018	-0.0282	-0.0265	0.0051	0.0023	0.451

Table 4: Simulation Two - Summary Measures

averages	un-bagged	bagged
RSS	63.40	62.82
R^2	0.353	0.358
true risk est	0.278	0.273

Table 5: Simulation Three, un-bagged vs. bagged estimator, $\sigma^2 = 0.25$, added 5 noise variables

x	β	mean		var		bias		MSE		ratio
1	1	1.0093	1.0156	0.0053	0.0056	0.0093	0.0156	0.0048	0.0053	1.088
2	0.5	0.4344	0.4001	0.0115	0.0210	-0.0656	-0.0999	0.0147	0.0289	1.973
3	0.3333	0.2496	0.2416	0.0365	0.0269	-0.0838	-0.0918	0.0398	0.0326	0.820
4	0.25	0.2234	0.2264	0.0432	0.0286	-0.0266	-0.0236	0.0396	0.0263	0.665
5	0.2	0.0736	0.0753	0.0145	0.0059	-0.1264	-0.1247	0.0290	0.0209	0.720
6	0.1667	0.1235	0.0925	0.0178	0.0107	-0.0431	-0.0742	0.0179	0.0151	0.847
7	0.1429	0.0766	0.0754	0.0156	0.0111	-0.0663	-0.0675	0.0184	0.0145	0.786
8	0.125	0.1009	0.0875	0.0179	0.0049	-0.0241	-0.0375	0.0167	0.0058	0.347
9	0.1111	0.0858	0.0606	0.0128	0.0037	-0.0253	-0.0505	0.0122	0.0058	0.481
10	0.1	0.0624	0.0492	0.0176	0.0129	-0.0376	-0.0508	0.0172	0.0142	0.826
11	0.0909	0.0273	0.0230	0.0074	0.0006	-0.0636	-0.0679	0.0107	0.0051	0.479
12	0.0833	0.0363	0.0435	0.0132	0.0096	-0.0470	-0.0398	0.0141	0.0102	0.723
13	0.0769	0.0594	0.0647	0.0159	0.0069	-0.0176	-0.0123	0.0146	0.0064	0.435
14	0.0714	0.0000	0.0068	0.0000	0.0009	-0.0714	-0.0647	0.0051	0.0050	0.973
15	0.0667	0.0448	0.0309	0.0089	0.0084	-0.0219	-0.0358	0.0085	0.0088	1.036
16	0.0625	0.0523	0.0505	0.0129	0.0074	-0.0102	-0.0120	0.0117	0.0068	0.579
17	0.0588	0.0656	0.0518	0.0267	0.0122	0.0068	-0.0070	0.0241	0.0110	0.457
18	0.0556	0.0000	0.0040	0.0000	0.0019	-0.0556	-0.0516	0.0031	0.0044	1.421
19	0.0526	0.0168	0.0142	0.0028	0.0010	-0.0358	-0.0384	0.0038	0.0024	0.616
20	0.05	0.0159	0.0180	0.0025	0.0012	-0.0341	-0.0320	0.0034	0.0021	0.616
21	0	0.0478	0.0363	0.0018	0.0010	0.0478	0.0363	0.0039	0.0022	0.564
22	0	0.0248	0.0209	0.0018	0.0008	0.0248	0.0209	0.0022	0.0011	0.504
23	0	0.0412	0.0376	0.0023	0.0010	0.0412	0.0376	0.0038	0.0023	0.616
24	0	0.0172	0.0239	0.0014	0.0007	0.0172	0.0239	0.0015	0.0012	0.795
25	0	0.0000	0.0074	0.0000	0.0001	0.0000	0.0074	0.0000	0.0002	

Table 6: Simulation Three - Summary Measures

averages	un-bagged	bagged
RSS	58.73	57.45
R^2	0.391	0.404
true risk est	0.336	0.325

Table 7: Data Example - Cross-validated risks for bagged estimators

k_1/k_2	1	2	3
1	178.2	176.9	177.0
2	181.8	180.4	182.6
3	188.9	192.9	194.0
4	189.1	189.3	191.8
5	208.9	196.0	190.4

Table 8: Mutation frequency in sample (patients treated with AZT+3TC+IDV)

PI-associated Mutations	Frequency	NRTI-associated Mutations	Frequency
L63P	193	T215F/Y	200
I93L	72	M41L	120
M36I/L/V	46	D67E/G/N	110
L10F/I/R/V	39	K70R/G/E	106
A71T/V/I	27	K219E/N/Q/R	87
K20I/M/R	13	L210W	86
L90M	6	V118I	51
M46I/L/V	4	T69D/N	42
N83I	4	T215C/D/E/I/V/S	38
G73C/S/T	3	E44A/D	38
F53L/Y	2	L74I/V	28
V82A/F/S/T	1	T69deletion/A/I/S	14
I84A/V	1	V75A/I/M/T/S	8
I54L/M/S/T/V	1	M184I/V	7
L33F	1	F116Y	2
I84C	0	Q151L/M	1
V32I	0	K65R	1
G48V	0	T69insertion	0
N88S	0	D67deletion	0
I54A	0	F77L	0
I47A/V	0	A62V	0
N88D/T	0	Y115F	0
L24I/F	0		
D30N	0		
I50V	0		
I50L	0		

Table 9: Estimated joint effects of mutations and mutation interactions on virologic response

PI- and RT- associated Mutations	Coefficient
L90M	0.329025
I93L * V75A/I/M/T/S	0.270962
L63P * D67E/G/N * V75A/I/M/T/S	0.224218
L10F/I/R/V * L63P	0.209222
I93L * D67E/G/N * V75A/I/M/T/S	0.164361
K20I/M/R * M36I/L/V	0.159395
L10F/I/R/V * M36I/L/V	0.150934
L10F/I/R/V	0.139267
D67E/G/N/ *V75A/I/M/T/S * K219E/N/Q/R	0.131429
K20I/M/R *K70R/G/E	0.118476
M184I/V	0.113037
D67E/G/N/ * V75A/I/M/T/S	0.107799
L63P * K70R/G/E * V75A/I/M/T/S	0.092744
L63P * T69D/N * V75A/I/M/T/S	0.090789
K20I/M/R	0.08876
T69D/N * V75A/I/M/T/S	0.088729
L10F/I/R/V * M41L	0.08825
K70R/G/E * V75A/I/M/T/S	0.084049
L10F/I/R/V * L63P * V118I	0.083536
M41L	0.08236
L10F/I/R/V * M36I/L/V * V118I	0.076455
M41L * D67E/G/N * M184I/V	0.070175
L63P * V75A/I/M/T/S	0.067187
M184I/V * L210W	0.0659
M41L * M184I/V	0.058845
T215F/Y	0.05381
M41L * K70R/G/E	-0.05292
L74I/V	-0.05579
A71T/V/I * I93L	-0.05994
A71T/V/I	-0.10026
T215C/D/E/I/V/S * K219E/N/Q/R	-0.1028
L10F/I/R/V * T69 /A/I/S	-0.1146
T69 /A/I/S	-0.19919

Table 10: Variable Importance Measures for PI-associated mutations

PI-associated Mutations	Variable Importance
L10F/I/R/V	0.331854
L90M	0.329099
K20I/M/R	0.19375
A71T/V/I	0.124193
M36I/L/V	0.049674
L63P	0.045315
I93L	0.012281
L33F	0.009722
M46I/L/V	0.008582
N83I	0.006002
I84A/V	0.004401
G73C/S/T	0.003127
F53L/Y	0.00089

Table 11: Variable Importance Measures for NRTI-associated mutations

NRTI-associated Mutations	Variable Importance
T69 /A/I/S	0.222174
M184I/V	0.186227
V75A/I/M/T/S	0.138039
T215C/D/E/I/V/S	0.091995
K219E/N/Q/R	0.082242
T215F/Y	0.068212
E44A/D	0.060438
L74I/V	0.057606
K70R/G/E	0.026778
T69D/N	0.023309
L210W	0.019655
D67E/G/N	0.017697
V118I	0.007738
F116Y	0.000406
M41L	0.000399