



Fuzzy Forests: Extending Random Forests Algorithm for Correlated, High-Dimensional Data

Daniel Conn

UCLA SPH, Biostatistics

Tuck Ngun

Department of Human Genetics David Geffen School of Medicine, UCLA

Gang Li

UCLA SPH, Biostatistics

Christina M. Ramirez

UCLA SPH, Biostatistics

Abstract

In this paper we introduce fuzzy forests, a novel machine learning algorithm for ranking the importance of features in high-dimensional classification and regression problems. Fuzzy forests is specifically designed to provide relatively unbiased rankings of variable importance in the presence of highly correlated features. We introduce our implementation of fuzzy forests in the R package, **fuzzyforest**. Fuzzy forests works by taking advantage of the network structure between features. First, the features are partitioned into separate modules such that the correlation within modules is high and the correlation between modules is low. Then recursive feature elimination random forests (RFE-RFs) are used on each module, separately. From the surviving features, a final group is selected and ranked using one last round of RFE-RFs. The package **fuzzyforest** allows for easy use of Weighted Gene Coexpression Network Analysis (WGCNA) to form modules of features such that the modules are roughly uncorrelated.

Keywords: Random Forests, WGCNA, machine learning, R, networks, $p \gg n$, big data, variable selection, variable importance, variable ranking.

1. Introduction

In the era of high-throughput technologies such as multi-color flow cytometry and next gen-

eration sequencing, high dimensional data has become increasingly common in biomedical research. As the ability to generate data has vastly outpaced our ability to analyze it. In the biomedical sciences as well as the 'Omics fields it is very common for the number of parameters to be much larger than the number of observations, the so-called $p \gg n$ problem. This problem is exacerbated by the fact that the features are often highly correlated and the correlation structure is often unknown a priori.

Identifying important features in this situation has been an area of intense research within the statistics and machine learning community. While model based feature selection algorithms such as the LASSO or SCAD may detect important features in presence of correlation (Raskutti *et al.* 2010), this comes at the cost of making parametric assumptions that may not hold in practice. If $p \gg n$, LASSO can select at most n features. In the setting where there are groups of highly correlated features, the LASSO has a tendency to arbitrarily pick one and shrink the others towards 0 (Bondell and Reich 2008).

Random forests are a popular ensemble machine learning algorithm. Random forests are non-parametric, non-linear, embarrassingly parallelizable, easy to implement, and have been described as one of the best “off-the-shelf” classifiers (Dua *et al.* 2014). Random forest variable importance measures (VIMs) offer a nonparametric alternative to model based feature selection algorithms (Breiman 2001). While random forest VIMs have demonstrated the ability to accurately capture the true importance of features in settings where the features are independent, it is well-known that random forests VIMs are biased when features are correlated with one another ((Strobl *et al.* 2007); (Strobl *et al.* 2008); (Nicodemus and Malley 2009)).

Fuzzy forests handle correlated features by taking a piecewise approach. We first estimate the network structure of the data and partition the set of features into distinct modules such that the correlation within each module is high and the correlation between modules is low. In this regard, we utilize the functionality of Weighted Gene Coexpression Network Analysis, a rigorous framework for detecting correlation networks (Zhang and Horvath 2005). We then use recursive feature elimination random forests (RFE-RFs) as in (Díaz-Uriarte and De Andres 2006) to select the most important features from each module. A final RFE-RF is then applied to the remaining features, from all modules, and selects the top k features.

The article is organized as follows. In section 2 of this article, we briefly review the random forests, WGCNA, and introduce the fuzzy forests algorithm. In section 3, we introduce the R package **fuzzyforest**. In section 4, we conduct simulations comparing fuzzy forests to random forests. In section 5, we use fuzzy forests to determine which immunologic factors are important in 1) determining which patients will be able to control HIV replication without use of anti-retrovirals and 2) determining which patients will have immune systems that won't respond to anti-retrovirals. Section 6 ends the article with a discussion and summary of our results.

2. Variable Importance Measures and the Fuzzy Forests Algorithm

2.1. Variable Importance Measures

In this section, we introduce basic notation and define variable importance measures. We assume that our data comes in the form of n independently and identically distributed (iid) pairs $(X, Y) \sim G_{(X, Y)}$. Here X is a p dimensional feature vector and Y is a scalar outcome.

The value of the v th feature for the i th subject will be denoted by $X_i^{(v)}$. The feature vector for the i th subject is denoted by $X_i = (X_i^{(1)}, \dots, X_i^{(p)})$. Let $X^{(v)} = (X_1^{(v)}, \dots, X_n^{(v)})$ be the set of values for feature v across all n subjects. Finally, we denote the marginal distribution of X by G_X and the marginal distribution of feature v , by $G_{X^{(v)}}$.

In the case of both classification and regression we are interested in modeling the conditional mean of Y given a feature vector X . We denote this conditional mean alternatively as $E[Y|X]$ or $f(X)$. In the case of regression, we assume that $Y|X$ has distribution $f(X) + \epsilon$, where the ϵ are iid with variance σ^2 . In the regression setting, Y is unrestricted. In binary classification, Y is restricted to take the value 0 or 1 and $Y|X$ is a Bernoulli trial with mean $E[Y|X = x] = P(Y = 1|X = x)$. Note that random forests can be easily generalized to handle the case of multi-class classification.

If the goal is to predict a new outcome Y based off of measurements, X , a good estimate of $f(X)$ is all that is required. Here, we are interested in more than prediction, in particular we wish to in to understand how $f(X)$ changes as function of particular features. If the value of $f(X)$ varies widely according to a particular value of the v th feature, $X^{(v)}$ is, in some sense, an “important” determinant of the outcome, Y .

If p is low dimensional ($p = 1, 2$), we can simply plot our estimate of $f(X)$ to understand how it varies as function of X . If p is moderate or large, $f(X)$ is difficult to interpret. It is most common in this case, to assume $f(X)$ has a specific parametric form such that $f_\beta(X)$ is known up to a finite dimensional parameter β . In the case of linear regression, β is a vector of regression coefficients and we can measure the importance of one feature versus another feature by examining the absolute magnitude of their corresponding coefficients (assuming the features have also been standardized).

If the parametric model $f_\beta(X)$ is a close approximation to $f(X)$, it is possible that interpretations based off of $\hat{f}_\beta(X)$ will not be misleading. Likewise, if $f_\beta(X)$ is a poor approximation of $f(X)$, the resulting interpretation will be misleading. The parametric approximation $f_\beta(X)$ may be inadequate for a variety of reasons. This may occur if important features are not included. Even if all appropriate features are included, $f_\beta(X)$ may fail to capture important interactions between features. If $f_\beta(X)$ is a linear regression model, $\sum_{v=1}^p \beta_v X^{(v)}$, the true $f(X)$ may be nonlinear in such a way that this best linear approximation fails to capture.

Permutation VIMs provide a means of summarizing the importance of individual features without making parametric assumptions. In the case of regression, the random forest permutation VIM of feature v estimates the following parameter:

$$VIM(v) = E(f(X_i^{(1)}, \dots, X_i^{(v)}, \dots, X_i^{(p)}) - f(X_i^{(1)}, \dots, \tilde{X}_i^{(v)}, \dots, X_i^{(p)}))^2. \quad (1)$$

The above expression deserves further explanation. First, note that the expression is the same for all choices of index, i , because the (X_i, Y_i) are iid with distribution $G_{(X,Y)}$. Next note that f is fixed and the expectation is with respect to the random variables $X_i = (X_i^{(1)}, \dots, X_i^{(v)}, \dots, X_i^{(p)})$ and $\tilde{X}_i^{(v)}$. The random vector X_i has distribution G_X and $\tilde{X}_i^{(v)}$, generated independently of X_i , has distribution $G_{X^{(v)}}$. Here, $\tilde{X}_i^{(v)}$ can be thought of as an independently generated deviation from $X_i^{(v)}$. If the value of $f(X_i)$ changes greatly when $X_i^{(v)}$ is replaced by $\tilde{X}_i^{(v)}$, it implies that the v th feature is highly important. In the case where $f_\beta(X) = \sum_{v=1}^p \beta_v X^{(v)}$ is a linear model, with standardized features ($var(X_i^{(v)}) = 1$), $VIM(v) = \beta_v^2$.

Here, $X_i^{(v)}$ and $\tilde{X}_i^{(v)}$ are iid with distribution $G_{X^{(v)}}$ where $G_{X^{(v)}}$ is the marginal distribution of $X_i^{(v)}$. This form of the VIM is given in a slightly different form in [Gregorutti et al. \(2013\)](#) and [Zhu et al. \(2012\)](#). [Gregorutti et al. \(2013\)](#) present a similar expression for the case of classification. These authors also discuss conditions under which the estimate of the permutation VIM derived from random forests is consistent.

2.2. An Introduction to Random Forests

Random forests is a popular ensemble method that has been applied in the setting of both classification and regression. The random forests algorithm works by combining the predictions of an ensemble of classification and regression trees. Each tree is grown on a separate bootstrap sample of the data. The number of trees grown in this manner is denoted as $ntree$. The subjects that are not selected in a particular bootstrap sample are said to be “out of bag.” Note that roughly one third of observations will be out of bag for each tree. These samples play the important role of serving as a test set for each tree.

Call the k th tree $\hat{f}_k(X)$. In the case of regression trees, $\hat{f}(X) = \frac{1}{ntree} \sum_{k=1}^{ntree} \hat{f}_k(X)$. In the case of classification, $\hat{f}(X)$ is the majority vote of the $ntree$ predictions given by $\hat{f}_k(X)$. Each regression tree is highly unstable and gives highly variable predictions. Averaging multiple trees over many bootstrap samples leads to more stable estimates of $f(X)$. The algorithm described thus far is known as bagging (bootstrap-aggregating). This algorithm is a special case of random forests.

A further element of randomness is introduced by random forests. Before a node in a particular tree is split, a subset of features is chosen at random. The best splitting rule involving only these randomly selected features is then used to split the node. The number of randomly selected features at each stage is commonly denoted as $mtry$. High values of $mtry$ tend to lead to just a few important features getting selected at the majority of nodes. If $mtry = p$, then random forests are equivalent to bagging. Lower values of $mtry$ allow more features to play a role in the estimation $f(X)$. In the case of regression, a common default value of $mtry$ is \sqrt{p} . In the case of classification $\lfloor p/3 \rfloor$ is common choice.

Random forest VIMs are obtained by testing how predictive accuracy suffers when the values of individual predictors are randomly permuted. If a feature is important for prediction, then permuting the values of that feature destroys the relationship with the outcome. Thus, for important features we should subsequently observe significant deterioration in predictive performance. VIMs measure the average decline in predictive performance across multiple trees.

The VIM for the v th feature is calculated as follows. Let $OOB_k \subset \{1, \dots, n\}$ be the indices for the out of bag sample from the k th tree. Let $\pi_k^{(v)} = (\pi_{k1}^{(v)}, \dots, \pi_{kn}^{(v)})$ be a random permutation of OOB_k and let $\tilde{X}_i^{(v)} = (X_i^{(1)}, \dots, X_{\pi_{ki}^{(v)}}^{(v)}, \dots, X_i^{(p)})$, where $i \in OOB_k$. In the case of regression, the variable importance of the i th feature from the k th tree is defined as

$$\widehat{VIM}_k(v) = \frac{\sum_{i \in OOB_k} (y_i - \hat{f}_k(\tilde{X}_i^{(v)}))^2 - (y_i - \hat{f}_k(X_i))^2}{|OOB_k|} \quad (2)$$

The variable importance for the entire random forest is defined as

$$\widehat{VIM}(v) = \frac{\sum_{k=1}^{ntree} \widehat{VIM}_k(v)}{ntree} \quad (3)$$

2.3. A Brief Review of WGCNA

There are a variety of algorithms for partitioning features into distinct, high correlation modules. In this regard, WGCNA is our method of choice. WGCNA is a rigorous framework for detecting correlation networks (Zhang and Horvath 2005). Although it was first developed to detect modules of highly correlated genes, it has found application in a variety of biological contexts. The R package **WGCNA** is a robust, computationally efficient, and well-documented implementation of the WGCNA framework. We expect that researchers already familiar with **WGCNA** will easily adopt the fuzzy forests algorithm and we expect that newcomers to WGCNA will be able to make good use of **WGCNA**'s fine documentation and tutorials.

WGCNA is a rigorous framework for constructing a network of features. This network is then used to determine clusters or modules of inter-related features. We briefly review the steps of a WGCNA network analysis. Our review closely follows that of Zhang and Horvath (2005). The user first specifies a similarity function $s_{ij} = S(X^{(i)}, X^{(j)})$ taking values between 0 and 1. Both unsigned and signed networks are possible. If the features are continuous, the most common choice of similarity function is $|Corr(X^{(i)}, X^{(j)})|$ or $\frac{1+Corr(X^{(i)}, X^{(j)})}{2}$ according to whether the network is unsigned or signed (Zhang and Horvath 2005).

This similarity matrix is then transformed into an adjacency matrix $A = [a_{ij}]$. The adjacency function determine how similarities translate into properties of the network. For example, in a hard-thresholded network, where nodes are either connected or unconnected, the adjacency function determines how high the correlation has to be in order for two nodes to be connected.

This hard threshold function, denoted by $signum(s_{ij}, \tau)$, is the simplest choice of adjacency function. If $s_{ij} \geq \tau$, $a_{ij} = signum(s_{ij}, \tau) = 1$ otherwise $a_{ij} = 0$. Nodes are either connected or unconnected if the $signum$ adjacency function is used. In practice, a soft-thresholded network is often more plausible than a hard-thresholded one. The power function $a_{ij} = s_{ij}^q$ is common choice of soft-thresholding adjacency function. Large values of q yield behavior closer to a hard-thresholded network. Setting $q = 1$ is equivalent to using the similarity function. Once an adjacency function is calculated, a hierarchical clustering tree algorithm is used to define clusters of features.

It is common to apply this hierarchical clustering algorithm to the topological overlap matrix rather than the adjacency matrix. The topological overlap between two nodes is defined as

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (4)$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_u a_{iu}$ (Horvath 2011). The topological overlap between two nodes can be high even if a_{ij} is low. This occurs when the two nodes are strongly connected to the same set of nodes. Use of topological overlap rather than the adjacencies may lead to more distinct modules (Zhang and Horvath 2005).

In many biological contexts, it is suspected that only a few features are highly connected. This prior knowledge leads to the scale-free criterion for determining which value of q to select. A network is said to have a generalized scale-free topology if $p(k) \sim k^{-\gamma}$ or $\log_{10}(p(k)) \propto \log_{10}(k)$ (Zhang and Horvath 2005). This suggests that one should select a smaller value of q such that the R^2 between $\log_{10}(p(k))$ and $\log_{10}(k)$ is high.

2.4. The Fuzzy Forests Algorithm

The fuzzy forests algorithm is an extension of random forests designed to obtain less biased variable importance rankings in the presence of correlated features. In following section, we describe the motivation behind fuzzy forests and explain why it provides less biased estimates of VIMs. In this section, we describe the algorithm.

The fuzzy forests algorithm can be subdivided into two steps: a screening step and a selection step. The screening step works in a piecewise fashion to screen out unimportant features. The screening step takes as input a partition of the features such that the correlation within each partition is roughly constant. Our package, **fuzzyforest**, facilitates the use of WGCNA to obtain such a partition of features. However, it is possible to use alternative methods to partition the features. Denote this partitioning of the features by the set $P = \{P_1, \dots, P_m\}$. Let $p_l = |P_l|$ so that $\sum_{l=1}^m p_l = p$.

The screening step operates independently on each partition. For each element of the partition P_l , RFE is used to screen out unimportant features. Starting with all features in the partition P_l , a random forest is fit using all features in P_l . The least important features are then eliminated, call the reduced set of features after the first random forest $P_l^{(1)}$. For example, the features with VIM in bottom 25% might be dropped at each step. A 2nd random forest is fit using features in $P_l^{(1)}$. The least important features from this latest random forest are then eliminated leading a further reduced set of features $P_l^{(2)} \subset P_l^{(1)} \subset P_l$. Call the subset obtained after the k th iteration $P_l^{(k)}$ and let $p_l^{(k)} = |P_l^{(k)}|$. Features are eliminated in this manner until a user-specified stopping criteria is reached. For example, features may be eliminated until 5% of the original P_l features remain.

The user must specify a few tuning parameters at the screening step. First, the user must specify how many features are to be dropped after each step of the RFE. We call this fraction the *drop_fraction*. The user must also specify a stopping criteria. In **fuzzyforest** the user specifies what percentage of the original p_l features to retain. This percentage will be called the *keep_fraction*. The first time the number of features drops below $keep_fraction * p_l$, the RFE will stop and the top $\lfloor keep_fraction * p_l \rfloor$ features will be selected. More precisely, for the first iteration k such that $p_l^{(k)} < keep_fraction * p_l$, we retain the top $\lfloor keep_fraction * p_l \rfloor$ features from $P_l^{(k-1)}$.

For each random forest, *mtry* and *ntree* must be appropriately selected. Since the number of features varies across random forests, *mtry* and *ntree* must be a function of the current number of features. Suppose we are at iteration k and are about fit a random forest to obtain $P_l^{(k+1)} \subset P_l^{(k)}$. In the case of regression, **fuzzyforest** lets $mtry = \sqrt{p_l^{(k)}} mtry_factor$. For classification **fuzzyforest** sets $mtry = \lfloor p_l^{(k)} / 3 \rfloor mtry_factor$. In both cases, *mtry_factor* must be pre-specified by the user with a default of 1. The parameter *ntree* must be set high enough to be able to pick up the effects of important variables, however if *ntree* is set too high, the iterative series of random forests will take a long time to run. The package **fuzzyforest** sets $ntree = \max(min_ntree, p_l^{(k)} * ntree_factor)$.

The selection step consists of one last RFE. This RFE is applied to all features that have been selected at the screening step. Note that a separate choice of *drop_fraction*, *mtry_factor*, *min_tree*, and *ntree_factor* may be used. In the package **fuzzyforest**, *keep_fraction* is implicitly defined by user. The user specifies how many features will selected by the selection step.

2.5. Motivation for Fuzzy Forests Algorithm

The selection step of fuzzy forests is motivated by the following observation concerning the theoretical properties of VIMs. Let $G_{P^{(l)}}$ denote the joint distribution of the features in partition $P^{(l)}$ and let $X^{P^{(l)}} \sim G_{P^{(l)}}$. Suppose that $X^{P^{(i)}} \perp X^{P^{(j)}}$ for all $i \neq j$ and suppose that $f(X) = \sum_{j=1}^m f_j(X^{P^{(j)}})$. If we fit a random forest using only the features in $P^{(l)}$ we are no longer estimating $E[Y|X] = f(X)$, instead we are estimating $E[E[Y|X]|X^{P^{(l)}}] = E[f(X)|X^{P^{(l)}}] = \sum_{j=1}^m E[f_j(X^{P^{(j)}})|X^{P^{(l)}}] = f_l(X^{P^{(l)}}) + \sum_{j \neq l} E[f_j(X^{P^{(j)}})]$

Suppose feature v is in partition $P^{(l)}$. The variable importance obtained by fitting a random forest to only those features in $P^{(l)}$ is estimating the following quantity:

$$VIM^*(v) = E(f_l(X_i^{(l_1)}, \dots, X_i^{(v)}, \dots, X_i^{(l_m)}) - f_l(X_i^{(l_1)}, \dots, \tilde{X}_i^{(v)}, \dots, X_i^{(l_m)}))^2. \quad (5)$$

Here, $X_i^{(k)}$ is the k th element of partition $P^{(l)}$. As in equation (1), $X^{(v)}$ and $\tilde{X}^{(v)}$ are iid from $G_{X^{(v)}}$. We see from this equation that $VIM^*(v) = VIM(v)$ if the true regression function is additive across modules and if the modules are independent of one another. If our assumptions are met, the VIMs obtained by analyzing each module separately are asymptotically the same as those that would have been obtained if VIMs were obtained by analyzing all features at once.

Therefore, the selection step of fuzzy forests achieves two goals. First of all, it reduces the number of features that have to be analyzed at one time. Second, the finite sample bias caused by correlated features is alleviated. In (Nicodemus and Malley 2009), it is observed that insignificant features that are correlated with a significant feature are more likely to be chosen at the root of tree than uncorrelated significant features. The high importance of these insignificant correlated features comes at the cost of the significant uncorrelated features. When we analyze each module separately, features in different groups are no longer competing against one another.

This above observations suggest that we should fit a separate random forest to each module and then report the the resulting estimates of the VIMs. Why does the fuzzy forests algorithm diverge from this simple procedure? We stress that the VIMs obtained by analyzing each module separately are, in general, different than the VIMs obtained by fitting a single random forests. The assumptions of additivity across modules and independence of modules are quite strong.

The fuzzy forests algorithm relaxes these overly restrictive assumptions. Fuzzy forests allows for interactions between features that were found to be important within modules. In biological applications, modules might represent different biological systems. Thus, it is reasonable to allow these systems to interact with one another. Fuzzy forests is implicitly making the assumption that the features that have high VIMs within modules are also most likely to be the features involved in interactions across modules.

It is important that RFE is applied at the selection step. When features from separate modules are combined, the potential for bias due to correlation between features is re-introduced. The iterative random forests, in practice, manage to eliminate unimportant predictors which are correlated with important ones.

We do not recommend interpreting the estimated VIMs too closely. While the ranking of the VIMs obtained from fuzzy forests are less biased than those derived from random forests, the estimated VIMs may still be quite biased.

3. The fuzzyforest package

The package **fuzzyforest** has two functions for fitting the fuzzy forests model. The first is **WGCNA_fuzzyforest**, the second is **fuzzyforest**. The function **WGCNA_fuzzyforest** automatically carries out a WGCNA analysis on the features. Then it uses the newly derived modules as input to the fuzzy forest algorithm. The WGCNA analysis is carried out via the **blockwiseModules** function from the package **WGCNA**.

The second function **fuzzyforest** assumes that the features have already been partitioned into separate modules. In particular a user is able to carefully conduct a separate module detection analysis using **WGCNA**. Such an analysis might find the smallest power q such that the scale-free topology criterion is met. The function **fuzzyforest** is then able to carry out the fuzzy forests algorithm using the output of **WGCNA**.

A number of tuning parameters must be specified before fuzzy forests can be run. These tuning parameters are organized into separate control objects. Tuning parameters related to WGCNA are specified with an S3 object of type **WGCNA_control** with constructor **WGCNA_control()**. Similarly tuning parameters related to the screening step and the select step are specified through objects of type **screen_control** and **select_control**.

We demonstrate the workings of **fuzzyforest** with a small simulation. The features have multivariate normal distribution and were generated using the package **mvtnorm**. There are 5 modules. Within each module the correlation between features is .8 and the variance of each feature is 1. The modules are independent of one another. The true model is linear and two of the 5 modules have 2 significant coefficients set to 5.

```
R> n <- 100
R> p <- 50
R> mod_sigma <- matrix(.8,10,10); diag(mod_sigma)=1
R> sigma <- as.matrix(.bdiag(lapply(1:5, function(j) mod_sigma)))
R> X <- rmvnorm(n, mean=rep(0, 50), sigma=sigma)
R> beta <- c(rep(5,2),rep(0,10-2),rep(5,2),rep(0,50-12))
R> y <- X%%beta + rnorm(n, sd=.1)
R> X <- as.data.frame(X)
R> fit <- WGCNA_fuzzyforest(X,y,WGCNA_params=WGCNA_control(minModuleSize=1)
+ ,screen_params = screen_control(keep_fraction=.5, drop_fraction=.2)
+ ,select_params = select_control(number_selected=4, drop_fraction=.1)
+ ,num_processors=4)
```

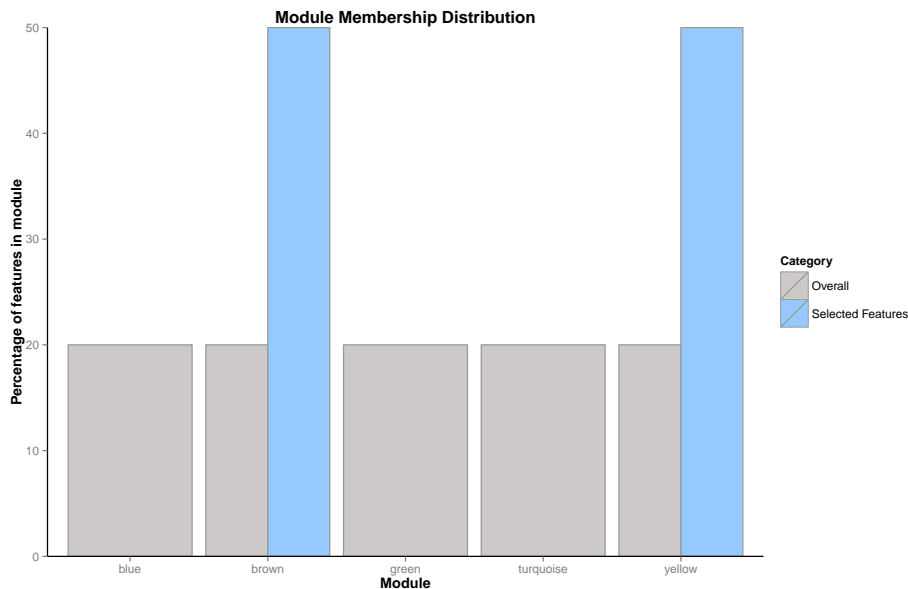
If the outcome, **y**, is a **factor** the random forests will be constructed using classification trees, otherwise regression trees will be used. **WGCNA_fuzzyforest** returns an S3 object of type **fuzzy_forest** consisting of a list. This list contains a **data.frame** with the ranking and VIM of the selected features. It also contains the final random forest that was fit at the selection step, along with the module membership and further information about the weighted co-expression network.

Objects of type **fuzzy_forest** have the usual generic methods. The function **print** returns the list of selected features. The function **predict(fuzzy_forest, new_data)** takes in a **data.frame** or **matrix** and produces predictions via a random forest based on the selected features.

The number of features selected by fuzzy forests must be pre-specified. It is possible to fit separate models selecting a different numbers of features for each model. The predictive accuracy on a test set can then be used to select one of these models. If a test set is unavailable a plot of the test error versus the number of features selected may shed light on when the fuzzy forests algorithm begins over-fitting. Note that these models will not necessarily be nested.

The generic function `plot` yields a visual display of which modules are important. The grey bars represent what percentage of the total p features fall into a particular module. The blue bars represent the percentage of selected features that fall into each module. Applying the function `plot` to the object `fit` above, we obtain the graph in figure 1.

Figure 1: Result of `plot(fit)`



4. Simulations

In this section we demonstrate the performance of fuzzy forests on a number of simulated data sets. Our first set of simulations assumes a logistic regression model. In all simulations, the sample size was $n = 1000$. We let the number of features equal $p = 100$ and 1000 . For each simulation we carried out 100 simulations. In each simulation, the number of features selected was 10. For each simulation, we also fit random forests 100 times.

In the case, where $p = 100$ there are 3 modules each of size 25. The features were generated from a multivariate normal distribution. Each feature had variance 1. The correlation within each module was 0.8. The modules were independent of one another and there were an additional 25 features that were independent of each other and independent of each module. The true regression function is of the form $\log(p_i/(1 - p_i)) = X_i'\beta$ where p_i is probability that subject i has outcome 1. For the case when $p = 100$, the first 3 features are significant with values 1, .5, and .25. The first 3 independent features are also significant with the same

Table 1: Results of Simulation for $p = 100$

Fuzzy Forests			Random Forests		
Feature	Avg VIM	Probability of Selection	Feature	Avg VIM	Probability of Selection
x_1	.058	.75	x_1	.010	.98
x_2	.023	.32	x_2	.004	.57
x_3	.014	.21	x_3	.002	.26
x_4	.007	.10	x_4	.001	.12
x_{76}	.074	1.0	x_{76}	.009	.94
x_{77}	.062	.83	x_{77}	.005	.64
x_{78}	.001	.14	x_{78}	.002	.26
x_{79}	.000	0.0	x_{79}	.000	0

Table 2: Results of Simulation for $p = 1000$

Fuzzy Forests			Random Forests		
Feature	Avg VIM	Probability of Selection	Feature	Avg VIM	Probability of Selection
x_1	.023	.70	x_1	.006	.99
x_2	.007	.22	x_2	.002	.66
x_3	.004	.11	x_3	.001	.31
x_4	.000	.01	x_4	.000	.08
x_{901}	.032	1.0	x_{901}	.005	.94
x_{902}	.026	.82	x_{902}	.000	.00
x_{903}	.003	.08	x_{903}	.000	.00
x_{904}	.000	0.0	x_{904}	.000	.00

values. We therefore have

$$\log(p_i/(1 - p_i)) = x_1 + .5x_2 + .25x_3 + x_{76} + .5x_{77} + .25x_{78} \quad (6)$$

where x_1, x_2 and x_3 are highly correlated and x_{76}, x_{77} , and x_{78} are independent.

We present below the results of fuzzy forests on 100 simulated data sets. In addition to presenting the VIMs for the significant predictors, we also present the VIMs for x_4 and x_{79} . The simulation demonstrates fuzzy forests still has a tendency to select unimportant features that are correlated with important ones.

In the case where $p = 1000$, there were 9 modules each of size 100. Modules were independent of one another and the correlation structure within modules was once again .8. The last 100 features were independent of one another and of the 9 modules. The first 3 of these independent features were significant yielding the following regression function:

$$\log(p_i/(1 - p_i)) = x_1 + .5x_2 + .25x_3 + x_{901} + .5x_{902} + .25x_{903} \quad (7)$$

We show the average VIMs over 100 simulations and present the VIMs of x_4 and the x_{904} in addition to the significant ones. Note that the VIMs are at an overall lower level. The selection probability of the most significant features remain somewhat similar to the case where $p = 100$. The selection probability of x_4 is lower than it was in the case of $p = 100$, however, the selection probabilities were overall lower in the simulation.

Table 3: Results of Fuzzy Forests Algorithm

Feature	VIM
4pPD1p	.056
4pR5n38nDRnPD1P	.0273
4pR5n38nDRpPD1p	.020
3p4pRAnR7nR5n38nDRpPD1p	.018
4pRApR7n	.018

5. Applications

We demonstrate a typical analysis by using **fuzzyforest** to discover immunologic profiles that predict how an HIV-infected patients responds to antiretroviral therapy. The subjects in the study were enrolled in either the Options Project or the SCOPE project, both longitudinal cohorts based at the University of California, San Francisco (UCSF). In this particular analysis, we wish to identify novel immunologic signatures that are predictive of whether a patient will be an immunologic responder or an immunologic non-responder.

An immunologic non-responder is a patient on antiretroviral therapy with undetectable levels of the virus (< 50 copies/ml) whose CD4+ T cell counts nevertheless remain low (< 350 cells/mm³). Similarly, an immunologic responder is an aviremic patient with undetectable levels of the virus and CD4+ T cell counts above 350 cells/mm³. There were 59 immunologic non-responders and 142 immunologic responders in the data set.

The features are derived flow cytometry, measuring T cell maturation, activation, dysfunction, senescence, antigen-specificity and proliferation. Because many of the flow cytometry parameters are highly correlated we used WGCNA to identify modules within the feature space. We used the scale-free topology criterion to determine the power of the adjacency function and set $q = 7$. A heatmap displays the module correlation structure of the feature space. We found 10 modules. The largest module was of size 27. The smallest module was of size 9.

We used the resulting modules memberships as input to the function **fuzzyforest**. Because of the small size of the modules we set **keep_fraction** to 0.25. We tested out multiple values of for **number_selected**. The ranking of features was robust to settings of this parameter. We display the results when selecting 5 features.

One of the strongest predictors of immunologic non-response was PD-1 expression on CD4+ T cells. This replicates findings from earlier work [Hunt *et al.* \(2008\)](#). PD-1 expression was evident across many cell types, especially in effector T cells. Non-responders also had higher percentages of cells that expressed HLA-DR.

6. Discussion

In this paper we have presented the fuzzy forests algorithm as an extension of random forests that can provide less biased feature selection in the presence of correlation between features. We discussed conditions under which fuzzy forests is expected to outperform random forests. We carried out simulations comparing random forests to fuzzy forests.

We also introduced an implementation of fuzzy forests in the **fuzzyforest** package. The **fuzzy-**

forest package has two functions for fitting the fuzzy forests algorithm. The first implementation, `WGCNA_fuzzyforest` automatically carries out WGCNA to partition the features into separate modules. These modules are then used by the fuzzy forests algorithm for feature selection. The second implementation, `fuzzyforest` lets the user determine how features should be partitioned before the fuzzy forests algorithm is used for feature selection.

We then used fuzzy forests to investigate how immunologic profiles determine a patient's response to HART. In this analysis, we took a hands-on approach to determining how the set of features should be partitioned. We used the scale free topology to determine the power q of the adjacency function. We also attempted to understand the biological significance of the resulting modules. Then we used the `fuzzyforest` function for feature selection.

Acknowledgements

This work was partially funded by NSF IIS 1251151.

References

- Bondell HD, Reich BJ (2008). "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR." *Biometrics*, **64**(1), 115–123.
- Breiman L (2001). "Random Forests." *Machine learning*, **45**(1), 5–32.
- Díaz-Uriarte R, De Andres SA (2006). "Gene Selection and Classification of Microarray Data Using Random Forest." *BMC bioinformatics*, **7**(1), 3.
- Dua S, Acharya UR, Dua P (2014). *Machine learning in healthcare informatics*. Springer.
- Gregorutti B, Michel B, Saint-Pierre P (2013). "Correlation and Variable Importance in Random Forests." *arXiv preprint arXiv:1310.5726*.
- Horvath S (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer.
- Hunt PW, Brenchley J, Sinclair E, McCune JM, Roland M, Shafer KP, Hsue P, Emu B, Krone M, Lampiris H, *et al.* (2008). "Relationship Between T Cell Activation and CD4+ T Cell Count in HIV-Seropositive Individuals with Undetectable Plasma HIV RNA Levels in the Absence of Therapy." *Journal of Infectious Diseases*, **197**(1), 126–133.
- Nicodemus KK, Malley JD (2009). "Predictor correlation impacts machine learning algorithms: implications for genomic studies." *Bioinformatics*, **25**(15), 1884–1890.
- Raskutti G, Wainwright MJ, Yu B (2010). "Restricted Eigenvalue Properties for Correlated Gaussian Designs." *The Journal of Machine Learning Research*, **11**, 2241–2259.
- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008). "Conditional Variable Importance for Random Forests." *BMC bioinformatics*, **9**(1), 307.

- Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC bioinformatics*, **8**(1), 25.
- Zhang B, Horvath S (2005). “A General Framework for Weighted Gene Co-Expression Network Analysis.” *Statistical applications in genetics and molecular biology*, **4**(1).
- Zhu R, Zeng D, Kosorok MR (2012). “Reinforcement Learning Trees.”

Affiliation:

Daniel Conn
Department of Biostatistics
UCLA School of Public Health
Los Angeles, United States of America
E-mail: djconn17@gmail.com

Figure 2: Module Correlation Structure

