

Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the setting of a Cancer Prevention Trial

Grant Izmirlian*

August 3, 2004

Abstract: A thorough discussion of the random forest (RF) algorithm as it relates to a SELDI-TOF proteomics study is presented, with special emphasis on its application for cancer prevention: specifically, what makes it an efficient, yet reliable classifier, and what makes it optimal among the many available approaches. The main body of the paper treats the particulars of how to successfully apply the RF algorithm in a proteomics profiling study to construct a classifier and discover peak intensities lending the most strength to the constructed classifier. Via Monte Carlo study, it is shown that (1) under the null hypothesis, the normalized importance measures display non-normal fat tailed asymptotics, so that a step down procedure such as the Benjamini Hochberg False discovery rate results in observed false discovery rates that are highly inflated and (2) some implications about power and sample size are hinted at using a second Monte Carlo study generated under an alternative hypothesis containing an important peak having simple odds ratio 10 for the affected class in a balanced design. This paper appeared in a slightly different and longer format in[1].

*National Cancer Institute; Executive Plaza North, Suite 3131; 6130 Executive Blvd, MSC 7354; Bethesda, MD 20852; izmirlian@nih.gov

1 Introduction

1.1 Overview

One of the most promising developments in the field of biomarkers and early detection has been the advent of genomic expression profiling using microarray technology, as it has provided the ability to profile the expression of an entire genome on a single chip. Thus researchers can determine the expression level of thousands of genes simultaneously and thereby “hunt” for differentially expressed genes across a variety of mRNA samples. For example, an experiment designed to study a potential cancer prevention agent could be designed by treating several samples of LNCaP cells with the agent, while designating another set of samples of LNCaP cells as controls. Then, mRNA collected from each cell population is hybridized to a gene microarray in order to search for differentially expressed genes among them. However, microarray technology has inherent limitations because the actual biological effectors are usually the resulting protein molecules. Thus any study of the genomic expression levels is blind to post-translational modifications and because of this, levels of mRNA expression often correlate poorly with the actual in vivo protein concentration due to differential rates of mRNA translation and varying protein half-lives[2]. For several decades, the identification of serum proteins and peptides has been conducted using mass spectrometry and electrophoresis gels. The limitations of these techniques are similar to those that existed in genomics before the advent of microarray technology. Until recently, only a relatively small number of proteins could be studied simultaneously. Now, however, as in genomics, recent high throughput technology coupled with the analytic tools of bioinformatics has accelerated the rate of discovery within the realm of serum protein chemistry, giving birth to the field of proteomics. Specifically, the first widely used such mass spectrometric technique is known as surface enhanced laser desorption ionization (SELDI) coupled with time of flight (TOF) mass spectrometric detection[3]. The principle behind this is that in the presence of an energy-absorbing matrix such as sinapinic acid (SPH), large molecules such as peptides ionize instead of decomposing when subjected to a nitrogen UV laser. Thus partially purified serum is crystallized with a SPH matrix and placed on a metal slide. Depending upon the range of masses the investigator wishes to study, there are a variety of possible slide surfaces. For example, among hydrophilic exchange media one choice is the strong anion exchange (SAX) surface, which has range from 2 kilo Daltons (kDa) to 50 kDa. This can be used for a “first pass”, while the weak cation exchange (WCX) surface has greater precision over a more narrow range (2 kDa to 20 kDa). The peptides are ionized by the pulsed laser beam and then traverse a magnetic field containing column. Masses are separated according to their times of flight as the latter are proportional to the square of the mass to charge (M/Z) ratio. Since nearly all of the resulting ions have unit charge, the mass to charge ratio is in most cases a mass. Therefore, the terms “ M/Z ratio” and “mass” will be used interchangeably in the following. The spectrum (intensity level as a function of mass) is actually recorded digitally, so that the resulting data obtained on each serum sample (hereafter, subject) is a series of intensity levels at each mass value on a common grid of masses (hereafter, peaks, peak intensities or features). A typical machine has a digitized spectrum of length in the tens of thousands with masses ranging from 2kDa to 50kDa. Notice that, at this level of resolution, that which is conceptualized as “a protein” translates into at least several consecutive peaks. For instance, it is known there are variations isotopic ratios from person to person so that at the level of resolution attainable by

SELDI-TOF, the identity of a given protein in an aggregate of human subjects is most likely a range of similar masses differing by a fraction of a Dalton.

The organization of the remainder of this paper is as follows. First, it is pointed out that, from the point of view of classical statistical methodology, most if not all datasets arising in a proteomic profiling study are under-specified problems so that no unique classification rule can be assigned to a given dataset using any of these methods of classical statistics. Consequently, the analyst must turn to machine-learning for answers. This will open a short synopsis of some of the available machine-learning tools that have enjoyed popular use, especially in the substantive literature (such as a single classification tree, neural network, or genetic algorithm). The following discussion will attempt to give one appreciation for the reasons why many of these algorithms tend to over fit the training data giving results that are not reproducible. Following will be a thorough discussion of the random forest (RF) algorithm[4] and [5]. Specifically, what makes it an efficient yet reliable classifier, and what makes it optimal among the many available approaches. The main body of the paper treats the particulars of how to successfully apply the RF algorithm in a proteomics profiling study to construct a classifier and discover peak intensities most likely responsible for the separation between the classes. Towards that end a nominal t-statistic for the important peak discovery component of analysis is investigated. This has recently become available in the current release of the software[5]. This discussion is illustrated via a Monte Carlo study using “realistic” simulated data generated from characteristics of a proteomic study from the author’s previous work. The case of null relationship between spectra and outcome and the case of a specific relationship of a given magnitude between spectra and outcome will be considered at a range of sample sizes. The newest (version 5) of RF was posted on 11/12 and is available as FORTRAN[5]. There is a version in R originally translated into C from the FORTRAN version 3 which has recently been updated to closely resemble the newest FORTRAN version 5 of the original author’s.[6]

1.2 Classifiers-An Overview

The goal of a proteomic profiling study is to try to relate these proteomic profiles to the clinical population of origin (hereafter, class membership, or class). In general this relationship could involve an arbitrary degree of complexity. Since a variety of methods for elucidating this relationship will be discussed, they will in general be referred to as classifiers. A classifier is an algorithm that uses training data, containing both the proteomic profile and the class membership on each element in a sample of a reasonable size, and uses it to define a classification algorithm whereby all future proteomic profiles with unknown class membership can be assigned a predicted class. After the training step, a validation step is necessary in order to estimate the generalization error in the trained classifier, and a variety of strategies exist for accomplishing this task. The one point in common among such strategies is that all require a validation set made up of samples not in the training set. The overall generalization error rate is one minus the fraction of predictions concordant with the truth. In two class problems (“1” = “affected” or “responds”, “0” = “healthy” or “doesn’t respond”), the sensitivity and specificity are defined as the fraction of predictions that are concordant with the truth among those in the “1” and “0” classes, respectively. Below we will briefly outline some of the alternate strategies for the validation step. At that point we will see that the way in which the RF algorithm does this is optimal among all such strategies in the literature[4] and [5]. An important point can be made here. Suppose that a choice of classifier

has been made and imagine an idealized situation in which such an abundance of data exists that, for each unique proteomic spectrum, there is a reasonable sample size sharing, nearly identically, that proteomic spectrum. Given a particular spectrum, X , then among the sample sharing the spectrum X , the average proportion of predictions for each class, j , obtained in the validation phase, is an estimate of the true underlying conditional probability: $\max_{j=0,1} \mathbb{P}\{Y = j | \text{spectrum is } X\}$, where Y denotes the class membership of a randomly drawn subject from the universe of individuals having spectrum X . In the remainder we refer to this as the science underlying the problem, since the presence or absence of true dependence of the outcome upon the spectrum that is the target of any such investigation. Furthermore, one should keep in mind that the complement of the true underlying conditional probability given above provides a lower limit to the expected error rate attainable by any classifier (the Bayes risk). Another important point to be made is that the credibility of the estimated error rate (including estimated sensitivity and specificity) depends entirely upon the robustness of the classifier to slight perturbations in the training data, and the manner in which it is validated.

1.3 Generalization Error-Some Caveats

Perhaps the first notable work in the cancer detection area was a case/control study of sera collected from ovarian cancer patients and controls, which used proteomics (SELDI-TOF) spectra, to train a classification algorithm to distinguish sera of ovarian cancer patients from that of controls[7]. Interestingly, that study very quickly drew a lot of attention, as sensitivity, specificity and positive predictive value (PPV) were reported to be 100%, 96% and 94%, respectively. However, since the appearance of that work, it has been the topic of much controversy. First, the 94% positive predictive value is misleading, as it is based upon the nearly 50% prevalence of ovarian cancer in the case/control study. When based instead upon the fraction of a percent prevalence that is known to exist in the population at large, this translates to a PPV of roughly 9% [8] and [9]. Second, it has been observed that the use of these findings to screen a healthy population would require a much higher specificity[10]. Finally, the difficulty in reproducing these results has called into question the validity of the findings[11].

1.4 Under-specified Statistical Problems

In order to better appreciate the origin of the above-mentioned caveats, an intuitive discussion of an under-specified statistical problem in classical statistics is now presented. In other words, why isn't linear or quadratic discriminant analysis (LDA, QDA) or even (polytomous) logistic regression appropriate for most proteomics profiling studies? The reason is that there is too much data on too few subjects. This has been called "the curse of dimensionality" by some authors[12]. Consider that a typical proteomics spectrum, even after peak detection, has on the order of 100 peaks, while typical sample sizes are less than 100. To compound the problem, since it is believed that the true mechanism at the level of the proteome that distinguishes one class from another is highly complex, all possible interactions among the peaks must be considered. In the case of the two-class problem and logistic regression, one is faced with 2100 regression coefficients with a sample size of less than 100. To say that the system does not have a unique solution is putting it mildly! Consider that one can assign arbitrary values to an arbitrary choice of all but 100 of the 2100 regression coefficients, and still there will exist a solution for the remaining

100. Consequently, nothing can be said about the values of any of the regression coefficients with any degree of certainty at all, so that it is impossible to construct a classifier based upon maximum likelihood. This includes linear and quadratic discriminant analysis and classification using categorical regression, unless one is willing to a priori throw away a large portion of the available features. As an aside, we remark that penalized maximum likelihood is a viable alternative and in fact, support vector machines, one of the popular methods for analyzing proteomic profiling studies, is equivalent to ridge stabilized maximum likelihood[13]. The disadvantage of this family of approaches is the large degree of tuning required as well as its vulnerability to the curse of dimensionality[12].

1.5 Classification Trees

For these reasons, machine-learning techniques must be used instead of the tools of classical statistical inference in proteomic profiling studies. One of the simplest machine-learning classifiers is the classification tree (CT)[14]. It functions by using features from the spectra to successively split the training set into two portions or nodes until all subjects belonging to the same node share the same class. The training phase begins with all subjects in the root node. The root node is split into two child nodes by selecting the M/Z value which separates the sample pool according to a threshold intensity level so that the split results in the greatest decrease in node class heterogeneity between parent and child nodes. Heterogeneity is measured in the CT using the Gini index, which is one minus the sum of squared class proportions. The decrease in heterogeneity is measured as the difference between the Gini index at the parent node and the weighted average Gini indices at the child nodes. In this manner, all resulting nodes are split until either the within node class purity is 100%, or further splitting is impossible due to identical feature profiles (not likely to occur in proteomics studies). Such nodes are called terminal nodes. Notice that the 100% node purity requirement sometimes results in a node containing a single subject. Thus, the algorithm just described is a complete decision rule and class membership prediction is done for spectra of unknown origin by placing it into the trained tree and assigning the class label corresponding to the terminal node into which it falls.

1.6 Bias and Variance

Because the “best variable” selection is repeated each time a node is split until the training set is fit perfectly, it is easy to see why a single CT represents an extreme case of over fit in the training data. Recall the language of the introductory paragraphs above, i.e. at the validation step and within a group of similar spectrum profiles, the average proportion of predictions for a given class, j , is an estimate of the true underlying conditional probability: $\max_{j=0,1} \mathbb{P}\{Y = j \mid \text{spectrum is } X\}$, where X and Y denote the spectrum of peak intensities and class membership of a randomly drawn subject from our source population. Classifiers such as a single CT, that fit the training data perfectly, can be shown to have high variance but low bias[15]. This is to say that in an idealized series of individual studies, attempts to replicate the same conclusions, using separate (but equally distributed) datasets, will result in highly variable error rates obtained in the validation step of each study. However if the classifier contains enough potential for complexity, i.e. a suitably rich span of possible associations between spectrum and class membership (hereafter, complexity span of the classifier), as does the classification tree, then the

average estimate among those individual studies would be an accurate estimate of the true underlying conditional probability of misclassification. Table 1 below lists various base classifiers by the magnitude of bias and according to the magnitude of variance. The term “base classifier” is used here to distinguish a classifier that is a single instance of itself from aggregate classifiers, discussed shortly, that are constructed by aggregating over the predictions made by multiple instances of a base classifier. Notice first, in table 1, that a single classification tree (CT), genetic algorithms (GA) and artificial neural nets (ANN) are all algorithms that have low bias but high variance. Recall that linear-, and quadratic- discriminant analyses (LDA, QDA) are unsuitable for proteomic studies because they require sample size to be larger than the number of features plus interactions. It is still worth mentioning that LDA and QDA have low variance, but their complexity span is limited by the fact that they function by partitioning the space of spectrum intensities into regions of homogenous predicted class and these regions are constrained to have linear (or quadratic) boundaries. This can result in bias. While Bayesian discriminant analysis can handle under-specified problems and has a richer range of boundary types, the investigator is required to choose a particular type of richness at the outset, and the algorithm is highly sensitive to this choice in under-specified problems, so that this too results in bias. The k-nearest neighbors (kNN) classifier is low in variance because it doesn’t necessarily fit the training data perfectly. However, its ability to detect trend between a few peaks and class membership deteriorates as the spectrum size increases, and thus has potentially a high bias as dimensionality increases and a strong signal among just a few peaks becomes overwhelmed by noise. Incidentally, support vector machines share this caveat-i.e. become overwhelmed by noise[12]. Clearly, an ideal classifier belongs to the low variance, low bias cell, which is empty in the diagram. However, construction of an aggregate classifier from a base classifier such as the CT, ANN or GA having low bias but high variance has the effect of variance reduction, producing a classifier of low bias and low variance.

1.7 Bootstrap Aggregation or Bagging

This can be explained as follows. Ideally, if every proteomics study could be replicated under identical circumstances hundreds of, each producing datasets statistically independent from one another, but drawn from the same underlying distribution, then a separate classifier could be trained on data from each study. Next, given a spectrum, x , the ensemble could be used to predict class membership corresponding to x by giving each of the classifiers a vote, counting votes for each of the classes, and assigning, as the predicted class, the one with most votes. If the number of independent studies is m , and the variance of any given statistic derived on a single base classifier in a single study is v , then the same statistic derived on the aggregate classifier has variance v/m . The summary statistic(s) could be for example, the generalization error and a list of peak importances, as done in the Monte Carlo study which follows. Realistically, however, one never has a series of independent and identically distributed (i.i.d.) studies, just the outcome of a single study. However, a series of bootstrap replicates of the dataset can be formed by drawing (with replacement) a sample of the same size. Note that drawing with replacement allows the possibility that a single element in the original sample is multiply represented in the bootstrap sample. It is well known that drawing a sample of size n with replacement from an original sample of the same size tends to include roughly two thirds of the original sample with duplication in the remaining third. These bootstrap replicates approximate a series of i.i.d. studies[16]. In this manner, a

low bias high variance base classifier such as CT can be aggregated resulting in a classifier having both low bias and variance. In the machine-learning literature this is called bagging (an acronym for bootstrap aggregating) the base classifier[17]. To reiterate, the training phase proceeds by training multiple base classifiers on each of a series of bootstrap replicates of the training set. This produces an ensemble of trained classifiers, each returning a predicted class when presented with a new spectrum. The aggregate classifier assigns predicted classes using majority vote. Another advantage of aggregate classifiers is that they offer a validation scheme whereby all of the available data can be used for training and for validation, while maintaining separation between training and validation, ensuring reliable estimates of error rates and related statistics. This validation scheme works as follows. Consider a dataset consisting of a sample of size n . Each base classifier in the ensemble is trained on a bootstrap replicate from the entirety of available data. However, each of these bootstrap replicates (being samples of size n drawn with replacement from the original sample of size n) tends to leave out roughly one third of the sample. Thus each classifier in the ensemble is trained on roughly two thirds of the original data. Consequently, each element in the sample of size n trains roughly of all classifiers in the ensemble so that it can be used to validate the remaining classifiers. The use of bagging and the related out of bag cross-validation method has been called “632 cross-validation” by some authors[18], because the “roughly” above is in actuality which is approximately 0.632. Notice how this differs from “leave one (or more) out cross-validation”. In the latter validation scheme, the training set size is $n - 1$ (or less) instead of n and more importantly, due of the high degree of overlap among the ensemble of resulting datasets, this ensemble of datasets do not approximate an i.i.d. series of studies so that there is neither reduction in variance nor increase in accuracy of the error rate estimate. Next, it is easy to see why 632 cross-validation is superior to split sample cross-validation because in the latter method one is forced to train on only a portion (half, two thirds, etc.) of the data. Finally, none of these other schemes has the effect of increasing the reliability of the error rate estimates, as does the combination of bagging and 632 cross-validation.

2 The Random Forest Algorithm

2.1 Random Forests is Bagging coupled with Random Feature Selection

The RF algorithm is conducted by bagging a classification tree, with 632 cross-validation, with an additional modification. Random feature selection in the construction of each tree and at each node is done in order to enhance the degree of independence among trees in the ensemble. This means that during the training phase, within each classification tree, each time a node needs to be split, the search for the best feature to split on is limited to a subset of all features, and that subset is randomly drawn from the entirety of features. This random draw is made separately for all nodes in each tree in the ensemble, and is of fixed size, m , a parameter set by the analyst. The default value of m is the square root of the total number of features, but results are fairly insensitive to this choice. Note that the subset selection process will tend to include each of the available features roughly at an equal number of nodes among all trees in the ensemble. Secondly, random feature selection still results in a low bias, high variance base classifier, since each CT is “grown” (continues splitting) until 100% node purity is reached. Consequently,

random feature selection has the effect of reducing correlation between individual classifiers in the ensemble, while maintaining strength of the aggregate i.e. its sensitivity to the true structure underlying our data³. To summarize, bagging always outperforms the base classifier both from the standpoint of reliability and from the standpoint of strength. The level of reliability and strength attained by bagging is enhanced by random feature subset selection. See figure 1.

2.2 Important Peak Discovery-The MDM measure

The next topic of discussion is the manner by which the RF algorithm detects important peaks once the classifier has been trained. The principle behind the quantification of the importance of a given peak to the classification algorithm is intuitively clear. The investigator is interested in identifying peaks that differentiate the classes from one another. Thus, if a particular peak lends discriminatory ability to the classification algorithm, then replacing its values by noise should attenuate the discriminatory ability of the algorithm as a whole. This attenuation is measured via the mean decrease in margin (MDM) measure, which is computed at each peak, j , as follows. First during the out-of-bag (o.o.b.) validation stage, each subject's amenability to classification is quantitated by the margin, which (in the two class situation) is the amount by which o.o.b votes for the correct class exceed those for the incorrect class. Once the margin for each subject is computed, each margin is recomputed using the same spectrum, only with intensity values at the peak, j , "noised" by selecting at random from the intensity values at peak j from among the other spectra in the dataset. If the peak, j , is important to the classifier, then the margin should have decreased. The MDM measure is the average over subjects of this decrease in margin at the subject level (true margin minus margin of noised spectra), and larger values of MDM are indicative of greater importance. The newly released version 5 of RF (available as FORTRAN) computes an estimated variance in the MDM importance measure corresponding to each peak in the spectrum, and resulting z-scores and corresponding p-values are returned⁴. Thus it is possible do important peak selection within the realm of statistical inference.

2.3 Adjustments for Multiple Testing-the Benjamini Hochberg FDR

As is the case in the analysis of gene expression micro-arrays, multiple hypothesis testing is an issue here as well. We recommend use of the Benjamini-Hochberg (BH) step-down procedure to control the false discovery rate or FDR[19]. This can be described as follows. If one does nothing about multiple hypothesis testing and applies the naive "nominal p-value is less than 0.05" filter to a list of more than one hundred hypotheses, i.e. the per comparison error rate, (PCER), procedure, then it is expected that roughly 5 peaks having purely happenstance relationship with the outcome will be determined to be as significant. This may be alright if the resulting list is much longer than 5, but if the important list is of length 5 or so, then decisions based upon this analysis will probably end up wasting someone's time and money in the lab. The strictest way to adjust for multiple testing is the Bonferoni procedure. It works by comparing nominal p-values corresponding to each test to 0.05 divided by the total number of hypothesis tests (peaks in this case). This controls the global type I error, which means that the chance of falsely identifying at most one or more peaks is less than 0.05. While this is appropriate in the analysis of multi-arm clinical trials, in which a single false positive finding

is clearly undesirable, it is clearly too conservative for filtering a list of candidate peaks. The BH step down procedure is somewhere in between the Bonferoni and PCER procedures, and works by controlling the false discovery rate, and this is more in keeping with the philosophy of this type of biologic investigation. The BH procedure controls the proportion of falsely identified peaks among the number of peaks identified as significant. While Bonferoni controls the chance of falsely identifying one or more, BH controls expected proportion falsely identified among those identified. To illustrate the BH step-down procedure, consider the analysis of a hypothetical proteomics profiling study, here, using simulated data. The manner in which this hypothetical dataset was simulated is described below. Imagine that the peak detection phase of the pre-processing step produced 138 peaks. Table 2 below lists the top 25 peaks, sorted by nominal p-values based upon the MDM importance measure nominal t-statistics. Column 1 lists the “mass name”; column 2, the MDM measure (difference in percent of o.o.b. votes times 100); column 3, the nominal t-statistics; column 4, nominal p-values; and column 5, the BH step-down values (so called because the direction is from higher to lower p-values, even though one starts at the end of the list and works upward). The latter are calculated as the peak rank (row number) times the desired false discovery rate (FDR), in this case 0.10, divided by the total number of hypothesis tests, in this case 138. The procedure works by starting at the bottom of the list and comparing nominal p with BH step-down value. As we move from row 138 upward, the first row in which the nominal p is less than the corresponding BH step-down value becomes the dividing line—all rows including this one are considered to be “discoveries”. Assuming that the hypothesis tests in question are statistically independent, this procedure controls the expected false discovery rate¹⁸. Thus, of the discoveries made according to this procedure, the expected proportion that occurred purely by chance is guaranteed to be less than or equal to the stipulated FDR value if the total number of discoveries is “large”. Under violations to the independence assumption, the procedure is supposed to be conservative. Notice that in table 2, the first 5 lines are considered to be discoveries. If in reality, #22 is the only true discovery then the observed proportion of false discoveries would be 80%. However, in the above-described procedure, replacing the nominal p-values with the true p-values (discussed below) results in no peaks identified as significant i.e. observed false discovery rate of 0%. To summarize, the results of a RF analysis consist of the overall o.o.b. error rate, sensitivity and specificity, and additionally, a list of peak importances. Here we use the MDM importance measures sorted by p-values combined with the BH step down procedure.

3 Monte Carlo Simulation Study

3.1 Simulation of “Realistic Spectra”

Next, we turn the discussion to benchmarking the algorithm via Monte Carlo study. This is done using simulated data. An attempt was made to simulate “realistic looking” proteomic spectra by using moments derived from a proteomics study in which the author was involved. These profiles will be generated first according to the peak intensities distribution described below. Secondly, the “outcome” variable, class membership, will be simulated conditional upon each spectrum. Section 3.2 describes results of a Monte Carlo investigation in which there is no association between the spectra and the class membership (i.e. under the global null hypothesis), while section 3.3 describes results of a second Monte Carlo investigation using

data simulated under a specific type of “alternative hypothesis”. Specifically, each dataset consisted of a sample of 100 spectra, while each spectrum contained 138 peaks. Each spectrum was generated according to a 138 dimensional correlated log-normal distribution (i.e. logged values are multivariate normal) with mean vector and marginal variances taken from the author’s previous work. There was an indication of a high degree of correlation in the proteomics study, but the sample size available was not sufficient for a stable estimate. However, correlations over 0.9 were considered as high while others were considered low. Thus the correlation matrix used in the simulation consisted of zeros and 0.9’s off the diagonal, with 78 of the 138 peak intensities belonging to a correlation pair of 0.9. This type of correlation produces realistic looking spectra sharing many gross visual features with the real proteomics study from the author’s previous work, and is in accordance with biological intuition, which states that practically all of what one sees in a human serum spectrum are the proteins of normal cellular processes. For example, consider a group of proteins that are involved with a specific process part of normal metabolism. If that process is running faster in a given subject then one expects concentrations of all proteins in that group to be affected

3.2 Data under the “Global Null”

Having a reasonable mechanism in place for simulating spectra, the next topic of discussion is the conditional distribution of class membership given the spectrum distribution. Of interest is the behavior of the RF classification algorithm, and more specifically, its criteria for selecting important peaks and also for quantifying the level of confidence in this selection. Moreover, of particular interest is this behavior both in the presence of true effect and in the absence of any effect. In order to do this systematically, the properties of the algorithm are first investigated via Monte Carlo (MC) study using a series of 1000 replicated i.i.d. datasets under the global null hypothesis (no association between spectra and outcome). Each of these datasets was given a sample size of 100 and an overall prevalence of 50%. The RF algorithm was fit to each MC replicate dataset using 1000 trees. The median, 5th and 95th MC percentiles for the sensitivity (Se) and specificity (Sp) were 48.1% (11.4%, 81.7%) and 49.0% (12.2%, 82.1%), respectively. While it appears that high values of each had non-negligible probability of occurring, it is important to note that such high values never occurred together, as the Youden index[20], $(Se + Sp - 1)$, and overall error rate had medians, 5th and 95th MC percentiles of -3.8% (-23.5%, 15.9%) and 50% (40%, 62%), respectively. Notice that the latter of these, the overall o.o.b. error rate and its 95% MC confidence interval are precisely that expected from 100 flips of a fair coin, the mechanism that generated the outcomes. Next, Monte Carlo variance estimates for the per peak MDM importance measures were derived and compared these with the variance estimate used in the FORTRAN version 5, and discovered that the latter are up to 10-fold larger than those estimated via Monte Carlo. Next, nominal t statistics based upon the correct standard error were obtained by dividing each MDM measure by its MC standard error estimate. A kernel density estimate was formed from the pooled sample of nominal t-statistics across peaks and MC replicates. This is displayed in figure 2. Quite puzzling is the fact that it appears to be skewed and largely kurtotic, with values 2.72 and 18.0, respectively. In passing we remark that the individual peak wise distributions of these nominal t-statistics behaved similarly to the distribution of the aggregate mentioned here. The areas to the right of the point of intersection, 2.29, under the true distribution and under the standard normal distribution are 0.023 and 0.0085 respectively. In the

following, important peak discovery was conducted using nominal t-statistics based upon MDM importance measures divided by MC estimates of standard error. To get a feel for the chance of committing “type-I” errors at the peak discovery phase, the BH procedure was applied to sorted lists in each of the 1000 MC datasets generated under the global null. The BH criterion for selecting important peaks at FDR of 10% was applied in two ways: by referring t-statistics to the standard normal curve, and by referring them to a kernel density estimate of their true (non-normal) underlying distribution. The latter was obtained via Monte Carlo simulation. When referring to the nominal (but incorrect) null distribution, the empirical false discovery rate (proportion of the Monte Carlo reps in which any peak was identified using the BH criteria) was 79%. This is quite inflated relative to the nominal FDR of 10%! However, when referring to the true null distribution, the empirical FDR was 9.5%, which is very close to and below the stipulated upper bound of 10%. To summarize, the MC study of data simulated under the “null hypothesis” has demonstrated that RF has a very low chance of committing “type-I” errors from the standpoint of the estimated error rate. However, at the stage of important feature selection using MDM nominal t-statistics, it has been demonstrated that in order to maintain control over the expected false discovery rate, p-values should be derived using the true underlying null distribution, which is non-normal. The ramifications for the study of a given dataset when the true null distribution is unknown are to use re-sampling by random allotment of class membership in order to determine percentiles under the null for use in a peak selection procedure such as the BH-FDR procedure.

3.3 Data under the “Alternative”

Having studied the application of the RF algorithm to data under the global null via Monte Carlo simulation, a similar study using data generated under “the alternative” is now presented. The dataset will contain a signal that will hopefully be detected using the RF classifier and peak importance MDM measure. Before describing the manner in which this “data under the alternative” was generated, some clarification is in order. If the analytic tools are well understood in terms of formal statistical inference and an approximation exists for the power function then given a parameter value and sample size, one can derive the approximate power to detect it. Then, the accuracy of the above-mentioned approximation can be tested by simulating data according to parameters specified and simply measuring the proportion of times the null was rejected—simple enough! That being said the reader now can appreciate the caveats faced in the present situation. There is no approximate power function. In the appendix is provided a variance formula for the MDM importance measure, but the null distribution is of unknown form. Consequently, one cannot have any feel for the magnitude in signal that is detectable at a particular sample size. Instead a particular choice of signal is made. This will be followed up with a statement about efficiency and a heuristic argument that can be used to give a rough estimate for sample size will be given. Towards this end, three peaks, #22, #88 and #96, were chosen more or less arbitrarily to be designated as “truly important”. This choice was made arbitrarily, but in accordance with reality, the selection was made from among the peaks present at smaller concentrations. Next, these intensity levels were categorized at their respective medians. Next, eight categories were constructed from all of the possible combinations among the three intensities being above or below their respective medians. The conditional probabilities for class “1” membership that were assigned to each of the eight cells are listed in table 3. Notice that in each of the columns corresponding to one of the

three “important” masses, #22, #88 and #96, a “1” indicates that the particular intensity at that mass is greater than its corresponding median value. The last column lists the probability of membership in the “1” class that was assigned to each cell. Thus the generation of the dataset is accomplished by first simulation of spectra as described in the above and then, to each complete spectrum, generating the appropriate -valued variable having success probability from column 4 in table 3 corresponding to the particular cell among the eight listed in table 3 to which the spectrum belongs. Notice that the “classification space” is the non-negative octant of 3-dimensional space, and that there are eight regions within this space of homogenous class membership distribution, and these eight cells are divided by 3 planes, one through each axis at a median. This fairly simplistic situation has linear boundaries and as such does not test the might of RF to detect oddly shaped boundaries. However, it’s intended purpose is to demonstrate the ability of RF to identify truly important peaks having a sufficient level of strength for classification, even in the presence of an abundance of peaks completely unassociated with class membership. Notice as well that while the cell shapes are simplistic, the model is not linear on the logit scale in the intensity levels categorized at the medians. The above generation of datasets was carried out within a Monte Carlo study, using 1000 such generated datasets each having sample size Each of these datasets was analyzed using the RF algorithm, again using 1000 trees. The medians, 5th and 95th MC percentiles corresponding to the overall o.o.b. error rate, sensitivity and specificity were 32% (24.9%, 40%), 76.9% (64.4%, 87.1%) and 63.8% (54.2%, 73.4%). Once again, the BH step-down procedure was applied in two ways to the sorted lists of MDM importance measure t-statistics that resulted from analyses of each of the 1000 simulated datasets-by referring t-statistics to the nominal (but incorrect) standard normal curve and by referring instead to their true underlying distribution under the global null as described above. Of the three peaks #22, #88 and #96, the strongest of these, mass #22, with a mean value of 4.4, was detected in 39% of the MC datasets when referring to the standard normal curve using the BH step-down procedure with a FDR of 10%, but when referring to the true null distribution, the 39% power dropped to only 1.5%. Thus an empirical estimate of the power to detect an important peak having MDM=4.4 at a sample size of 100 under a FDR of 10% is only 1.5%. Notice that inference on the MDM’s is in reality based only on a sample size of since the asymptotic variance is of order equal to the size of the average o.o.b. sample size. For sake of comparison, consider a power calculation for the two by two table for association between class membership and the intensity at mass #22, categorized at its median. The overall prevalence and true logged relative odds are given by and where the π ’s are the probabilities listed in the rightmost column of table 3, i.e. the conditional probability for class “1” given the spectrum, and the p ’s are the population fractions in each of the cells. The eight cell proportions were estimated from a dataset of size 10,000. Next, note that, given an overall prevalence of 0.34, the sample size required for a power of 90% to detect a logged odds ratio of 2.23 in a simple 1 way association under type I error of 10% is Since, as remarked above, inference on the MDM measures in the RF algorithm is based upon the average o.o.b. sample, then an o.o.b. sample of 70 corresponds to a total sample of size A second Monte Carlo study based upon data generated under the “alternative hypothesis” mentioned above, having sample size was conducted. This resulted in medians, 5th and 95th MC percentiles for the overall o.o.b. error rate, sensitivity and specificity of 31.1% (25.3%, 36.3%), 79.1% (70.1%, 86.4%), and 64.0% (57.0%, 71.2%), respectively. This time, the MDM corresponding to mass #22, with a mean value of 7.0, was detected using the BH procedure with FDR

10% with an empirical power of 93% when referring to the nominal (but incorrect) standard normal curve, but only 2.5% when referring to the true null distribution. The first result, based upon assumed normality of the null distribution suggests, at least under the current set of circumstances, that peak discovery using RF is no less efficient in the use of data in discovering a single peak of a given strength among 137 other non-significant peaks than knowing the correct peak and the correct split in advance and using the two by two table test for association, except for the factor of e . This is not to imply that RF's ability to sort through noise does not result in loss of efficiency as it was shown above that referral to the wrong null distribution is not fixed at the same type I error (in this case FDR). In truth the power to detect the above effect at a sample size of 190, when referring to the true underlying null distribution, is only 2.5%.

4 Discussion

The topic of aggregating base classifiers has been one of intense and fruitful activity in the area of machine-learning. Nonetheless, even among the most current substantive work in medical bioinformatics, the analysis tools of popular choice are base classifiers [21]. It cannot be stressed enough that, from the standpoint of reproducibility and validity, no other tool can be expected to match the performance of a bagged classifier. The reason for this is that the bagged classifier gives the closest and most reliable approximation to the true relationship between the spectra and the outcome. Furthermore, 632 cross-validation is made possible, and this is clearly the most efficient use of available data. The reason for this is that bootstrapped replicates are sent to train each classifier so that each element in the sample will have been left out of roughly 1/3 of the training sets of all of the classifiers in the ensemble, so that validation is done by sending each element in the sample to classifiers which it did not train. This is nearly as good as having an independent validation sample in addition to a bagged classifier, depending upon the degree of overlap (statistical dependence) of the classifiers in the ensemble. Bootstrapping ensures a degree of independence, depending upon how large a series of bootstrapped replicates and how large the sample size. The random forests algorithm reduces the level of correlation even further by using random feature selection every time a node is split. The performance of the RF algorithm was investigated via Monte Carlo simulation. This was done both under the assumption that no relationship exists between the spectra and outcome (the global null), and under a stipulated "effect" of a given magnitude at two different sample sizes. The first Monte Carlo investigation using data under the global null demonstrated a non-normal asymptotic distribution for the mean decrease in margin (MDM) t-statistics, having a larger tail to the right of the moderate-to-extreme value, 2.48. It was noted that this has a profound effect on important peak selection using a multiple testing selection criteria such as the Benjamini-Hochberg false discovery rate (FDR). Peaks selected using MDM t-statistics filtered using the BH criteria controlled at FDR=10% resulted in an empirical false discovery rate of 79% when referring to the incorrect (standard normal) null distribution. When referred to the correct null distribution, the empirical false discovery rate was 9.5%. In addition, the relative efficiency of peak selection using importance measures returned by RF filtered using BH at FDR=10% relative to logistic regression endowed with the knowledge of the correct peak and the correct split was hinted at but not resolved. Further work needs to be done to understand the nature of the true null distribution, and in the mean time,

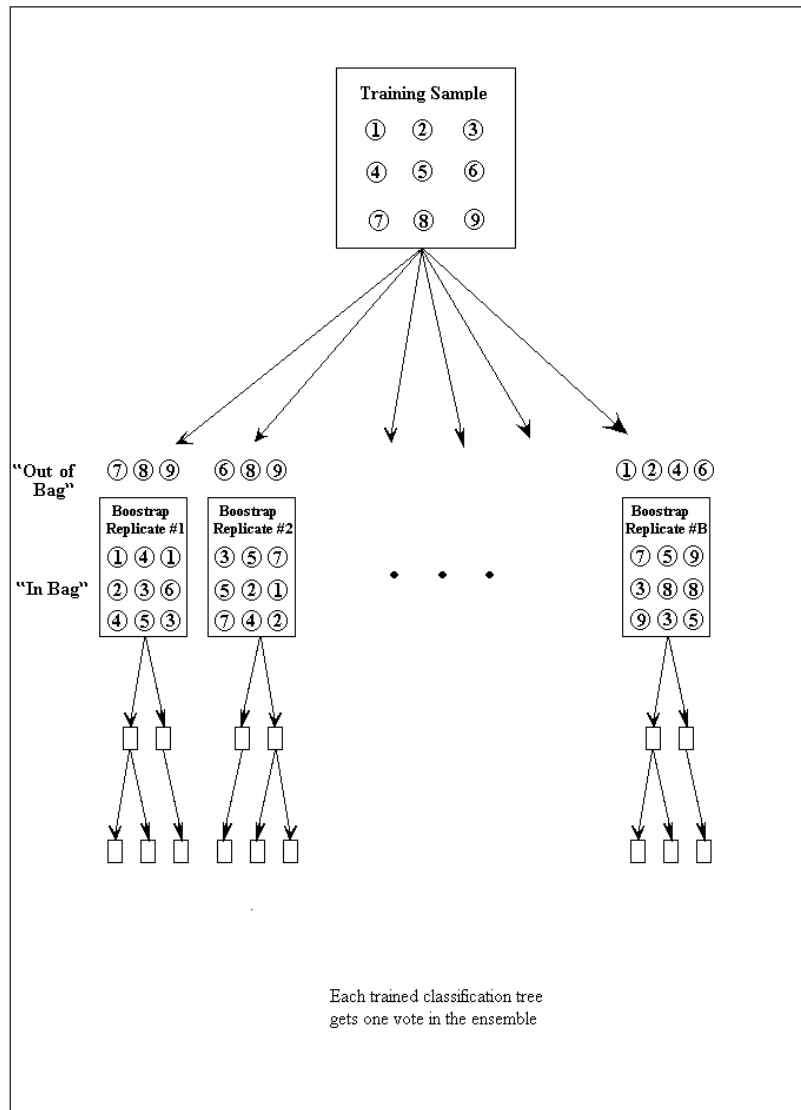
an additional series of simulation studies may shed light upon the efficiency of RF relative to logistic regression in the hands of a deity. For now it can be said that the upper bound is , and most likely, the correct answer is substantially smaller. The implications of this are profound. If an investigator is putting together a proteomics profiling study from scratch (as opposed to adding onto a completed clinical trial), the question is posed “what kind of reasonable effect is expected of the strongest single predictor split at the most informative threshold”? From this one can easily obtain the required sample size from a binomial test of proportions. The required sample to sift through all of the extraneous information (i.e. to do data mining) is going to be substantially larger than three times the above. In conclusion, the most striking advantages of the random forest algorithm are its robustness to noise, its simplicity, lack of dependence upon tuning parameters, and speed in computation. As an additional note we mention that an entirely new set of visual diagnostic tools has been made available⁴. Two such new diagnostic tools are the person-peak specific MDM measure (to assess the influence at the person by peak level) and predicted prototypes (to create meaningful summary displays of a trained classifier). A thorough investigation of these is recommended. Finally, there has recently appeared a plethora of publications promising a comparison of various statistical techniques in the analysis of proteomics profiling data. Usually these apply a range of techniques to one or more datasets taken from profiling studies¹. Presumably, the use of clinically obtained data is intended to attach a greater level of credibility within the audience of bench scientists and medical researchers. However, several clinically obtained datasets cannot begin to illuminate the statistical properties of a analytic tool in the way that can be done using “realistic looking data” generated in a Monte Carlo study. For example, “how does the method fare when there really is no relationship between the spectra and the outcome variable” and “how does the method fare when there really is a relationship between the spectrum and the outcome variable of a given strength” are questions that can only be answered via a thorough simulation study. This paper only begins to provide a thorough study, and as one does not yet exist, this can be taken as an invitation to the reader to write one. One important issue would be to pay better attention than time has permitted here to the generation of even more realistic looking spectra. By “realistic” it is meant that the goal is to create data that share a reasonable amount of statistical characteristics with proteomic spectra of human serum. For this purpose spectra from several studies could be combined to provide stable estimates of the required cross moments, and substantive experts would be required to weigh in on the level of “reality” having been attained. The most thorough kind of “methods bench-marking” work should contain “real data” as well as a thorough Monte Carlo study. While this paper has not quite made the cut stipulated by the above razor, it is hoped that a convincing case has been made.

References

- [1] Izmirlian, G., ©2004. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. *Ann. N.Y. Acad. Sci.* 1020, 154–174.
- [2] Wu, B., Long, A. D., Abbott, D., *et al.*, 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19, 1636–1643.
- [3] Yip, T.-T., and Lomas, L., 2002. Seldi proteinchip array in oncoproteomic research. *Technology in Cancer Research and Treatment* 1, 273–280.
- [4] Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
- [5] Breiman, L., and Cutler, A., 2003. The random forest package, version 5, in FORTRAN. <http://www.math.usu.edu/~adele/forests/index.htm>.
- [6] Liaw, A., Weiner, M., 2003. The random forest package, version 3.91 in R. <http://cran.us.r-project.org/>.
- [7] Petricoin III, E. F., Ardkani, A. M., Hitt, B. A., *et al.*, 2002. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 359, 572–577.
- [8] Elwood, M., 2002. Correspondence: Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 360, 170.
- [9] Rockhill, B., 2002. Correspondence: Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 360, 169.
- [10] Diamandis, E. P., 2002. Correspondence: Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 360, 170.
- [11] Pollack, A., 2004. New Cancer Test Stirs Hope and Concern. *The New York Times*, Section F, page 1.
- [12] Hastie, T., Tibshirani, R., Friedman, J. H., 2001. *The Elements of Statistical Learning*. Springer, New York.
- [13] Svetnik, V., 2003. Personal communication.
- [14] Breiman, L., Friedman, J. H., Olshen, R. A., P., T., 1984. *Classification and Regression Trees*. Chapman and Hall, New York.
- [15] Breiman, L., 1998. Arcing classifiers. *The Annals of Statistics* 26, 801–823.
- [16] Efron, B., 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7, 1–26.
- [17] Breiman, L., 1996. Bagging predictors. *Machine Learning* 26, 123–140.
- [18] Efron, B., Tibshirani, R., 1995. Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Tech. rep., Stanford University.
- [19] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57, 289–300.

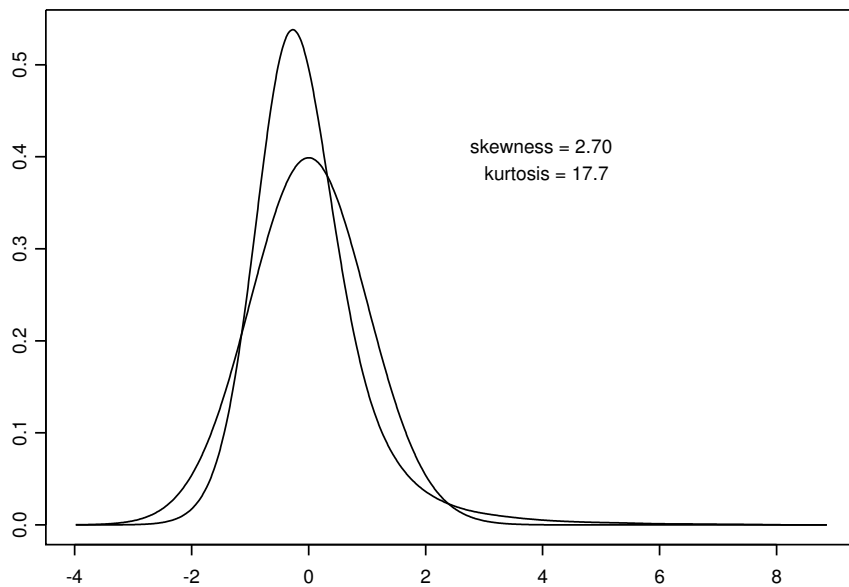
- [20] Youden, W. J., 1950. Index for rating diagnostic tests. *Cancer* 3, 32–35.
- [21] Yanagisawa, K., Shyr, Y., Xu, B.-J., P., T., 2003. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet* 262, 433–439.
- [22] Durrett, R., 1984. *Probability: Theory and Examples*. Wadsworth and Brooks/Cole, Belmont CA.

Figure 1



Schematic of “bagging” or bootstrap aggregation using the classification tree as the base classifier. Note that out-of-bag samples are used to validate all trees for which they are “out-of-bag”. In ordinary bagged classification trees, splitting is done at every node in every tree by selecting the best feature to split on resulting in purest nodes. In the random forest algorithm, the search for best feature is limited to a randomly drawn subset of size around the square root of the total number of features. Random subsets are drawn repeatedly every time a node is split.

Figure 2



Upper curve: Kernel density plot for t-statistics from Monte Carlo (# reps = 1000) study generated from the global null, pooled across 138 peaks with skewness and kurtosis shown; lower curve: standard normal.

Table 1

	Low Variance	High Variance
Low Bias		CT, NN, GA
High Bias	kNN, LDA, QDA, BDA	

Bias/Variance characteristics of several base classifiers BDA=Bayesian discriminant analysis, CT=classification trees, GA=genetic algorithms, kNN=k nearest neighbors, LDA=linear discriminant analysis, NN=neural nets, QDA=quadratic discriminant analysis.

Table 2

Mass #	MDM	t-stat	Nominal p	True p	BH
M.062	1.930	4.200	1.33e-05	0.004	0.000725
M.022	7.520	4.040	2.67e-05	0.005	0.001450
M.037	1.290	3.090	9.89e-04	0.011	0.002170
M.026	1.260	3.060	0.001	0.012	0.002900
M.060	1.270	2.780	0.003	0.015	0.003620
M.044	1.020	2.480	0.006	0.021	0.004350
M.055	0.295	2.200	0.014	0.029	0.005070
M.029	0.291	2.080	0.019	0.033	0.005800
M.036	0.665	1.470	0.070	0.073	0.006520
M.009	0.603	1.470	0.071	0.074	0.007250
M.002	0.226	1.430	0.076	0.078	0.007970
M.020	0.539	1.170	0.121	0.114	0.008700
M.004	0.484	1.150	0.124	0.117	0.009420
M.068	0.495	1.130	0.129	0.121	0.010100
M.069	0.455	1.120	0.132	0.123	0.010900
M.042	0.154	1.100	0.136	0.127	0.011600
M.067	0.474	1.090	0.138	0.129	0.012300
M.053	0.149	1.070	0.143	0.133	0.013000
M.085	0.136	1.040	0.149	0.139	0.013800
M.045	0.396	0.964	0.168	0.156	0.014500
M.028	0.317	0.786	0.216	0.204	0.015200
M.064	0.304	0.771	0.220	0.209	0.015900
M.135	0.320	0.770	0.221	0.209	0.016700
M.019	0.316	0.768	0.221	0.210	0.017400
M.131	0.111	0.703	0.241	0.231	0.018100

The top 25 MDM importance measure nominal t values sorted by p-value. Source-data simulated under “the alternative”.

Table 3

#22	#88	#96	$\mathbb{P}\{Y = 1 \mid X\}$
0	0	0	0.05
1	0	0	0.15
0	1	0	0.70
1	1	0	0.70
0	0	1	0.10
1	0	1	0.30
0	1	1	0.10
1	1	1	0.99

Class “1” membership probabilities assigned to each of the eight cells defined by categorizing 3 arbitrarily chosen peak intensities at their respective medians. In each of the first 3 columns, a “1” indicates “is greater than its corresponding median value”.

5 Appendix: Derivation of the MDM variance

Denote the data by $T_n = \{(X_i, Y_i) : i = 1, \dots, n\}$ where $X_i \in R_d$, a bounded subset of \mathbb{R}^d , has distribution $F(dx) = \mathbb{P}\{X_i \in dx\}$ and $\pi(x) = \mathbb{P}\{Y_i = 1 \mid X_i = x\}$. Attention is restricted here to the two-class situation for ease in exposition, but the ideas presented here generalize easily. Call T_n the training set. For $b = 1, \dots, m$ let $T_n^{(b)} = \{(X_i, Y_i) : i \in \mathcal{I}_{n,b}\}$ denote a bootstrap sample from T_n , which is to say that $\mathcal{I}_{n,b}$ is a random sample with replacement drawn from $\{1, \dots, n\}$. In the language of the text, $\mathcal{I}_{n,b}$ is the portion of the training sample that is “in-bag” for the bootstrap replicate b . Denote its complement as $\mathcal{O}_{n,b}$ which is the portion of the training sample that is “out-of-bag” for the bootstrap replicate b . Next, let $C(\cdot, T_n^{(b)}, \xi_b)$, denote the b -th classifier in the ensemble, which has trained on $T_n^{(b)}$ and is considered a stochastic function from R_d to $\{0, 1\}$. Here, ξ_b is a random vector containing codings for all random feature subset selections at each node within the tree trained on bootstrap replicate b . For a given subject, i , a tree, b , and a feature, j , the raw margin[4] is defined as:

$$\Delta_{i,b,n}^{(j)} = I\left(C(X_i, T_n^{(b)}, \xi_b) = Y_i\right) - I\left(C(X_i^{(j)}, T_n^{(b)}, \xi_b) = Y_i\right)$$

where $X_i^{(j)}$ is equal to X_i at all components except at the j -th, which is replaced with a random draw from the components of the rest of the sample, i.e.

$$X_i^{(j)} = \begin{cases} X_{\eta(i),j} & \text{for } \ell = j \\ X_\ell & \text{otherwise} \end{cases}$$

and η is a uniform random permutation on $\{1, \dots, n\}$. The MDM importance measure can now be defined as:

$$\bar{\Delta}_{j,n,m} = \frac{1}{m} \sum_{b=1}^m \frac{1}{|\mathcal{O}_{n,b}|} \sum_{i \in \mathcal{O}_{n,b}} \Delta_{i,b,n}^{(j)}$$

By the sub-additive ergodic theorem [22], for fixed n ,

$$\bar{\Delta}_{j,n,m} \xrightarrow{\text{a.s.}} \mathbb{E}\left[\Delta_{i,b,n}^{(j)} \mid T_n\right] \text{ as } m \rightarrow \infty$$

Next, the dependence of the variance upon n and m is investigated. The variance is equal to the expectation of the conditional variance given the out-of-bag set in question. The variance of conditional expectation terms are zero since the conditional expectations are of a mean of a random sum of identically distributed terms:

$$\begin{aligned}
\text{var} [\bar{\Delta}_{j,n,m}] &= \frac{1}{m} \mathbb{E} [|\mathcal{O}_{n,b}|^{-1}] \text{var} [\Delta_{i,b,n}^{(j)}] + \frac{1}{m} \mathbb{E} \left[\frac{|\mathcal{O}_{n,b}| - 1}{|\mathcal{O}_{n,b}|} \right] \text{cov} [\Delta_{i,b,n}^{(j)}, \Delta_{i',b,n}^{(j)}] \\
&+ \frac{m-1}{m} \mathbb{E} \left[\frac{|\mathcal{O}_{n,b} \cap \mathcal{O}_{n,b'}|}{|\mathcal{O}_{n,b}| |\mathcal{O}_{n,b'}|} \right] \text{cov} [\Delta_{i,b,n}^{(j)}, \Delta_{i,b',n}^{(j)} \mid \mathcal{O}_{n,b} \cap \mathcal{O}_{n,b'} \neq \emptyset] \mathbb{P}\{\mathcal{O}_{n,b} \cap \mathcal{O}_{n,b'}\} \\
&+ \frac{m-1}{m} \mathbb{E} \left[\frac{|\mathcal{O}_{n,b} \Delta \mathcal{O}_{n,b'}|}{|\mathcal{O}_{n,b}| |\mathcal{O}_{n,b'}|} \right] \text{cov} [\Delta_{i,b,n}^{(j)}, \Delta_{i,b',n}^{(j)} \mid \mathcal{O}_{n,b} \Delta \mathcal{O}_{n,b'} \neq \emptyset] \mathbb{P}\{\mathcal{O}_{n,b} \Delta \mathcal{O}_{n,b'}\} \\
&+ \frac{m-1}{m} \mathbb{E} \left[\frac{|\mathcal{I}_{n,b} \cap \mathcal{I}_{n,b'}|}{|\mathcal{I}_{n,b}| |\mathcal{I}_{n,b'}|} \right] \text{cov} [\Delta_{i,b,n}^{(j)}, \Delta_{i,b',n}^{(j)} \mid \mathcal{I}_{n,b} \cap \mathcal{I}_{n,b'} \neq \emptyset] \mathbb{P}\{\mathcal{I}_{n,b} \cap \mathcal{I}_{n,b'}\} \\
&= O\left(\frac{1}{nm}\right) + O\left(\frac{1}{m}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right)
\end{aligned}$$

where Δ (not to be confused with Δ) is the symmetric set difference. The point to be made here is that since m , the number of bootstrap replicates in the ensemble (trees in the forest), can be made arbitrarily large it follows that the dominating terms in the variance of the importance measure based upon a sample of size n arise from the last three terms above. The source of these correlation terms is due to individuals that are (i) out-of-bag in both (ii) in-bag in one and out-of-bag in the other and (iii) in-bag in two trees, b and b' .