Unsupervised Learning with Random Forest Predictors

Tao Shi[1, 2] and Steve Horvath[1, 2,]

[1] Department of Human Genetics, David Geffen School of Medicine, UCLA

[2] Department of Biostatistics, School of Public Health, UCLA

E-mail: shorvath@mednet.ucla.edu

February 28, 2005

Authors' Footnote:

Tao Shi (Email: shidaxia@yahoo.com) graduated with a Ph.D. degree in Human Genetics from UCLA. Steve Horvath (Email: shorvath@mednet.ucla.edu) is an Assistant Professor in the Departments of Biostatistics and Human Genetics at the University of California, Los Angeles, CA 90095-7008.

ABSTRACT

A random forest (RF) predictor (Breiman 2001) is an ensemble of individual tree predictors. As part of their construction, RF predictors naturally lead to a dissimilarity measure between the observations. One can also define an RF dissimilarity measure between unlabelled data: the idea is to construct an RF predictor that distinguishes the 'observed' data from suitably generated synthetic data (Breiman 2003). The observed data are the original unlabelled data while the synthetic data are drawn from a reference distribution. Recently, RF dissimilarities have been used successfully in several unsupervised learning tasks involving genomic data. Unlike standard dissimilarities, the relationship between the RF dissimilarity and the variables can be difficult to disentangle. Here we describe the properties of the RF dissimilarity and make recommendations on how to use it in practice.

An RF dissimilarity can be attractive because it handles mixed variable types well, is invariant to monotonic transformations of the input variables, is robust to outlying observations, and accommodates several strategies for dealing with missing data. The RF dissimilarity easily deals with large number of variables due to its intrinsic variable selection, e.g. the *Addcl*1 RF dissimilarity weighs the contribution of each variable on the dissimilarity according to how dependent it is on other variables.

We find that the RF dissimilarity is useful for detecting tumor sample clusters on the basis of tumor marker expressions. In this application, biologically meaningful clusters can often be described with simple thresholding rules.

KEY WORDS: Unsupervised learning; random forests; ensemble predictors; cluster analysis; tumor marker

1. INTRODUCTION

Machine learning methods are often categorized into supervised (outcome labels are used) and unsupervised (outcome label are not used) learning methods. Interestingly, many supervised methods can be turned into unsupervised methods using the following idea: one creates an artificial class label that distinguishes the 'observed' data from suitably generated 'synthetic' data. The observed data are the original unlabelled data while the synthetic data are drawn from a reference distribution. Some supervised learning methods distinguishing observed from synthetic data yield a dissimilarity measure that can be used as input in subsequent unsupervised learning methods (Liu *et al.* 2000; Hastie *et al.* 2001; Breiman 2003). Breiman (2003) proposed to use random forest (RF) predictors to distinguish observed from synthetic data. When the resulting RF dissimilarity is used as input in unsupervised learning methods (e.g. clustering), patterns can be found which may or may not correspond to clusters in the Euclidean sense of the world. The RF dissimilarity has been successfully used in several unsupervised learning tasks involving genomic data: Breiman (2003) applied RF clustering to DNA microarray data, Allen *et al.* (2003) applied it to genomic sequence data, and Shi *et al.* (2004) applied it to tumor marker data. In these real data applications, the resulting clusters often made biological sense, which provides indirect empirical evidence that the method works well in practice.

Many unsupervised learning methods require as input a dissimilarity measure between the observations. Here we describe some important properties of the RF dissimilarity so that potential users can decide when and how to use these dissimilarities in practice. We focus on the properties of the RF dissimilarity when used in partitioning around medoid clustering (Kaufman and Rousseeuw 1990) and in multi-dimensional scaling plots.

After a motivational example (section 1.1), we review the random forest dissimilarity in section 2. We describe how to approximate the relationship between the RF dissimilarity and its variables in section 3. We contrast two classes of RF dissimilarities in section 4. In section 5, we describe how to use the RF dissimilarity in practice. In the discussion, we review the properties of the RF dissimilarity. In the appendix, we provide simple geometric examples.

1.1 MOTIVATION

The RF dissimilarity has been found to be useful for tumor class discovery on the basis of immunohistochemical tumor marker expressions (Shi *et al.* 2004). Tumor marker expressions (protein abundances) are often scored by the percentage of cells staining. These tumor marker expressions can have semi-continuous, highly skewed distributions since many observations may take on the value 0 or 100 percent (Figure 1a).

One (but not the only) reason why the random forest dissimilarity is attractive for tumor marker expressions is that it handles highly skewed variables well: it is invariant with respect to monotonic transformations of the variables, obviating the need for symmetrizing skewed variable distributions.

As an example, we describe the analysis of 307 cancer patients for which 8 tumor marker measurements were available (Shi *et al.* 2004). The goal was to determine whether the cancer patients fall into distinct clusters (unsupervised analysis) and if so, whether the clusters were related to tumor recurrence? The RF dissimilarity was used as input in partitioning around medoid (PAM) clustering (Kaufman and Rousseeuw 1990) to group the patients (tumor samples) into 2 clusters (referred to as RF clusters). For comparison, we also used the Euclidean distance in PAM clustering and refer to the result as Euclidean distance clusters. When cross-tabulating the patients according to their RF and Euclidean distance cluster memberships, we find significant ($p = 1.2e$-15) agreement.

There is indirect empirical evidence that the RF clusters are clinically more meaningful than the Euclidean distance based clusters with regard to post-operative patient survival. In Figures 1b − 1d, we color the 223 tumor samples that fall into RF cluster 1 and Euclidean distance cluster 1 in black, the 36 samples that fall into RF cluster 2 and Euclidean distance cluster 2 in blue, the 25 samples in RF cluster 1 and Euclidean distance cluster 2 in red, and the 23 samples in RF cluster 2 and Euclidean distance cluster 1 in green. The Kaplan Meier plots (Kaplan and Meier 1958) in Figure 1b visualize the survival time distributions: the two pairs of curves, black/red and green/blue, are closer to each other than the black/green and red/blue pairs, i.e. the RF clusters are clinically more meaningful than the Euclidean distance based clusters. Using the logrank test (Cox and Oakes 2001), we find that the RF dissimilarity based clusters have more distinct survival time distributions ($p = 4e$-9) than the Euclidean distance based clusters ($p = 0.019$).

Figures 1c and 1d highlight the differences between the RF and the Euclidean distance clusters in terms of the underlying tumor marker expressions. Figure 1c is a color-coded depiction of the standardized tumor marker expressions (columns) across patients (rows). The patients have been sorted according to RF and Euclidean cluster membership. PAM clustering with the Euclidean distance groups the black and green samples into one cluster and the blue and red samples into the other cluster. This Euclidean distance based grouping makes sense when considering the expression pattern of marker 4, which has the highest variance. However, it makes no sense when considering the expression patterns of markers 1 – 3.

PAM clustering with the RF dissimilarity groups red and black patient samples into one cluster and the green and blue samples into the other cluster. Figure 1c shows that this grouping makes sense, especially when considering markers (columns) 1, 2, and 3 since markers 1 and 2 tend to be under expressed while marker 3 tends to be over-expressed in the green and blue samples. Similarly, the scatterplot in Figure 1d shows that marker expressions 1 and 2 tend to be low for the green and blue samples.

Thus, Figure 1c provides visual evidence that the 2 dissimilarities focus on different markers. The RF dissimilarity focuses on markers that are dependent. Dependent markers may correspond to disease pathways, which drive the clinical outcomes of interest. Thus, another reason why the RF dissimilarity is attractive for tumor marker expressions is that it weighs the contributions of each covariate on the dissimilarity in a natural way: the more related the covariate is to other covariates the more it will affect the definition of the RF dissimilarity.

Another reason why the RF dissimilarity is attractive for tumor marker expression is that it intrinsically dichotomizes the tumor marker expressions. While the information from all 8 tumor markers contributed to the clustering of the samples, RF cluster membership can be predicted with the following rule: tumor samples with $> 65\%$ staining for tumor marker 1 and $> 80\%$ staining for marker 2 are classified into RF cluster 1 (only 9.4% misclassifications), see Figure 1d. For clinical practice, it may be important to find threshold rules for defining cluster membership in terms of a few tumor marker expressions. To arrive at such rules, one can use a supervised learning method, e.g. we used a classification tree predictor (Breiman *et al.* 1984) to predict cluster membership on the basis of the underlying variables. In several tumor marker applications, we have found that RF clusters can be described by cuts along dependent variables. These cuts naturally lead to thresholding rules for describing the resulting clusters. Currently, it is standard practice to dichotomize tumor marker expressions for ease of interpretation and reproducibility in the *supervised* analysis of tumor marker data. But we caution against dichotomizing expressions in an *unsupervised* learning analysis since ad-hoc external threshold values may reduce information or even bias the results. As detailed below, the random forest dissimilarity is based on individual tree predictors, which intrinsically dichotomizes the tumor marker expressions in a principled, data-driven way.

In section 4.1.1, we provide a simulated example that further illustrates why random forest clustering can be particularly useful for detecting clusters, which can be described with thresholding rules.

2. RANDOM FOREST DISSIMILARITIES

An RF predictor is an ensemble of individual classification tree predictors (Breiman 2001). For each observation, each individual tree votes for one class and the forest predicts the class that has the plurality of votes. The user has to specify the number of randomly selected variables (*mtry*) to be searched through for the best split at each node. The Gini index (Breiman *et al.* 1984) is used as the splitting criterion. The largest tree possible is grown and is not pruned. The root node of each tree in the forest contains a bootstrap sample from the original data as the training set. The observations that are not in the training set, roughly 1/3 of the original data set, are referred to as out-of-bag (OOB) observations. One can arrive at OOB predictions as follows: for a case in the original data, predict the outcome by plurality vote involving only those trees that did not contain the case in their corresponding bootstrap sample. By contrasting these OOB predictions with the training set outcomes, one can arrive at an estimate of the prediction error rate, which is referred to as the OOB error rate. The RF construction allows one to define several measures of variable importance. In this article, we use the node purity based variable importance measure. A discussion of the different importance measures is beyond the scope of this article. Another by-product of the RF construction is the RF dissimilarity measure, which is the focus of this article.

2.1 The RF Dissimilarity for Labelled Data

We will briefly review how to use random forests to arrive at a dissimilarity measure for *labelled* data, i.e., an outcome is available (Breiman 2003). Since an individual tree is unpruned, the terminal nodes will contain only a small number of observations. The training data are run down each tree. If observations i and j both land in the same terminal node, the similarity between i and j is increased by one. At the end of the forest construction, the similarities are symmetrized and divided by the number of trees. The similarity between an observation and itself is set equal to one. The similarities between objects form a matrix, SIM, which is symmetric, positive definite, and each entry lies in the unit interval $[0, 1]$. The RF dissimilarity is defined as $(DISSIM_{ij}) = (\sqrt{1 - SIM_{ij}})$. The RF dissimilarity can be used as input of multi-dimensional scaling (MDS), which yields a set of points in an Euclidean space such that the Euclidean distances between these points are approximately equal to the dissimilarities (Cox and Cox 2001). The aim of MDS is to choose a low dimensional configuration of points which minimizes a 'stress' function. Different stress functions lead to different MDS procedures. In this paper, we will use classical (cMDS) and isotonic (isoMDS) (Kruskal and Wish 1978; Shepard, Romney, Nerlove and Board 1972) multidimensional scaling as implemented in the R (R Development Core Team 2004) functions *cmdscale* and *isoMDS*, respectively. The function *cmdscale* is in the standard distribution of R. The function *isoMDS* is implemented in the contributed package *MASS* (Venables and Ripley 2002).

2.2 The RF Dissimilarity for Unlabelled Data

We will now review how to use random forests to arrive at a dissimilarity measure for *unlabelled* data (Breiman 2003). The idea is to use the similarity matrix generated from a RF predictor that distinguishes observed from 'synthetic' data. The observed data are the original, unlabelled data while the synthetic data are drawn from a reference distribution. A synthetic class outcome is defined by labelling the observed data by class 1 and the synthetic data by class 2. By restricting the resulting labelled similarity measure to the observed data, one can define a similarity measure between unlabelled observations. The similarity measure strongly depends on how the synthetic observations are generated. We focus on two methods that have been implemented in the *randomForest* function of the contributed R package *randomForest* (Liaw and Wiener 2002). The function provides an R interface for Breiman's original FORTRAN implementation of RF.

In *Addcl*1 sampling, synthetic data are added by randomly sampling from the product of empirical marginal distributions of the variables. The tree predictors of the random forest aim to separate synthetic from observed data. Hence each tree will be enriched with splitting variables that are dependent on other variables. Thus the resulting RF dissimilarity measure will be built on the basis of dependent variables. In general, we find that this sampling option works best in practice. For example, it was used the motivating example described above.

In *Addcl*2 sampling, synthetic data are added by randomly sampling from the hyperrectangle that contains the observed data, i.e. the variables of synthetic observations have a uniform distribution with range determined by the minimum and maximum of the corresponding observed variable. The *Addcl*2 sampling option has been removed from the recent versions of Breiman's FORTRAN implementation, but it is still implemented in the R package.

We use the RF dissimilarity as input for partitioning around medoids (PAM) clustering which is implemented in the R function *pam* in the contributed package *cluster*. But many other clustering procedures (e.g. hierarchical clustering) could be used as well. Since we have found that the *Addcl*1 dissimilarity often works best in practice (see for example the wine data example below), we refer to the combination of RF dissimilarity and PAM clustering as RF clustering. But to highlight some of the theoretical properties of RF dissimilarities, we will also consider *Addcl*2 sampling. Another artificial sampling scheme that leads to a simple geometric interpretation, is described in example $ExNULL$ in the appendix.

3. APPROXIMATING THE RF DISSIMILARITY

There are several approaches for disentangling the relationship between an RF dissimilarity and its variables. The RF variable importance measure can be used to determine which variables are important for defining the RF dissimilarity. By construction, the variable importance is determined by how the distribution of the observed (class 1) data differs from that of the synthetic (class 2) data. It is often useful to model the RF clusters in terms of their underlying variables using graphs or easily interpretable supervised learning methods. For example, one can study how important variables vary across the clusters (e.g., parallel coordinate plots, boxplots) or use the cluster label as outcome in a classification tree.

When dealing with quantitative (interval scale) variables, one can *sometimes* find a Euclidean distance-based approximation of the *Addcl*1 RF dissimilarity *if each variable is equally important for distinguishing observed from synthetic observations.* This is done by i) replacing the variable values by their ranks, ii) standardizing the result to mean 0 and variance 1, and iii) using the resulting variables in a Euclidean distance. To motivate i), note that the RF dissimilarity depends only on variable ranks (scale invariance) since the underlying tree node splitting criterion (Gini index) considers only variable ranks. To motivate ii) note that standardization puts the variables on a more equal footing in a Euclidean distance.

In several real data applications, e.g. the wine example below, we have found that this approximation works quite well. But in general, the two dissimilarity measures can differ substantially, see example *ExAddcl*1, table 1, and example *ExX* in the *Appendix* section, which shows that the RF dissimilarity is in general not invariant to rotations in the variable space.

3.1 Real Data Example: The Wine Data

The wine data set (in the R library *mlbench*) contains 13 variables that describe the chemical analysis results of wines from three different cultivars in Italy. We interpret the cultivars as external, true clusters and compare the RF clustering results with it using the adjusted Rand index (Rand 1971; Hubert and Arabie 1985), which is a commonly used measure of agreement between two partitions. The Rand index assumes its maximum of 1 in case of perfect agreement while its expected value in the case of random partitions is 0. We find that *Addcl*1 RF clustering coupled with classical MDS leads to the highest clustering agreement (adj. Rand=0.93), while *Addcl*2 clustering with either cMDS (adj. Rand = 0.19) or isoMDS (adj. Rand = 0.21) does not perform well (the first and second column of Figure 2). The RF variable importance measure shows that variables 6 and 7 are important (third column in Figure 2). These two variables also have the highest correlation coefficient (0.86) among all variable pairs, which is consistent with our explanation of *Addcl*1 clustering: the trees preferentially use variables that are highly dependent.

We find that *Addcl*1 RF clustering (adjusted Rand = 0.93) is far superior to PAM clustering with a Euclidean distance (adj. Rand = 0.37). There are two possible reasons for this superior performance: First, the random forest dissimilarity is based on the ranks of the variables while the Euclidean distance is scale dependent. Standardizing each variable ($mean = 0$, $variance = 1$) improves the performance of the Euclidean distance based PAM clustering considerably (adj. Rand = 0.74). Second, the RF dissimilarity may focus on those variables that contain information that relate to the cultivars, i.e. its intrinsic feature selection selects the variables that matter for the external cluster label.

To show that the RF dissimilarity can be approximated with the Euclidean distance based procedure outlined above, consider the 4^{th} column of Figure 2: the *Addcl*1 RF dissimilarities are strongly correlated (Spearman correlation = 0.90) with the Euclidean distance approximations. the *Addcl*2 RF dissimilarities are much less correlated (correlation = 0.48) with the Euclidean distances in this example.

4. THE *Addcl*1 AND *Addcl*2 RF DISSIMILARITIES

In the following, we describe several simulated examples that highlight properties of the *Addcl*1 and *Addcl*2 RF dissimilarity.

4.1 Simulated Examples

4.1.1 Example *ExRule*

This example is meant to illustrate why the RF dissimilarity works better in the motivational example involving tumor marker expressions. The underlying cluster structure can be described using a simple thresholding rule. There are 2 signal variables. For observations in cluster 1 and cluster 2, the 2 signal variables $X1$ and $X2$ have random uniform distributions on the intervals $U[0.8, 1.0]$ and $U[0, 0.8]$, respectively. Thus cluster 1 observations can be predicted using the threshold rule $X1 > 0.8$ and $X2 > 0.8$. We simulate 150 cluster 1 observations and 150 cluster 2 observations. Noise variable X3 is simulated to follow a binary (Bernoulli) distribution with hit probability 0.5, which is independent of all other variables. Noise variables $X4, \ldots, X10$ are simulated by randomly permuting variable X1, i.e. they follow the same distribution of X1 but are independent of all other variables.

In Figure 3 the observations have been colored as follows: black if $X1 > 0.8$ and $X3 = 1$, red if $X1 > 0.8$ and $X3 = 0$, blue if $X1 \leq 0.8$ and $X3 = 0$, and green if $X1 \leq 0.8$ and $X3 = 1$.

The *Addcl*1 RF dissimilarity focuses on variables $X1$ and $X2$ in its construction since these are the only variable that are dependent. This can can be seen from the RF variable importance measures depicted in Figure 3c. Figures 3a and b show that when the *Addcl*1 RF dissimilarity is used as input of classical or isotonic multidimensional scaling, it results in distinct point clusters that correspond to whether or not $X1 > 0.8$.

In contrast, the *Addcl*2 RF dissimilarity focuses on variable $X3$ in its construction since its binary distribution in the observed data is quite different from its uniform distribution in the synthetic data. This can be seen from the RF variable importance measures depicted in Figure 3f. Since the RF *Addcl*2 definition is driven by variable $X3$, the point clouds that result in classical and isotonic MDS plots, are defined by the values of $X3$, see Figure 3d and e. To a lesser extent the *Addcl*2 dissimilarity is also defined by variables 1 and 2 since their dependence can also be used to distinguish observed from synthetic observations.

Similar to the *Addcl*2 RF dissimilarity, the Euclidean distance is determined by variable $X3$ as can be seen from the coloring of the point clouds in classical and isotonic MDS plots (Figures 3g and h. The reason for this is that $X3$ is the variable with the highest variance (Figure 3i).

Example *ExRule* illustrates why the RF clusters in the motivational example were different from those of the Euclidean distance. In this tumor expression data example, the Euclidean distance also focused on tumor marker with the highest variance while the RF distance focused on the most dependent tumor markers. In both examples, the clusters could be described using a simple thresholding rule.

4.1.2 Example *ExDep*

This example shows the effects of highly collinear variables. In this example, all variables have independent, random uniform distributions except that variable 1 equals variable 2. More specifically, the data consists of 150 observations with 5 random uniform variables $U[0, 1]$. Variables 2 through 5 are independent while variable 1 equals variable 2. The synthetic data that results from *Addcl*1 sampling have 5 independent, random uniform variables. Since variables 1 and 2 are highly dependent in the observed but not in the synthetic data, they are most important for distinguishing observed from synthetic data. Thus the *Addcl*1 RF dissimilarity, which is identical to the *Addcl*2 RF dissimilarity in this case, focuses on variables 1 and 2 in its construction. This can be seen from the multidimensional scaling plot in Figure 4. The points fall along a U-shape. The color coding of the points reveals that the U-shape corresponds to high, medium, and low values of variable 1. Similarly, when the RF dissimilarity is used as input of PAM clustering (k = 3 clusters), the resulting 3 clusters correspond to low, medium, and high values of variable 1. Depending on the application this may or may not be an interesting result. If the 2 variables measure the expression levels of distinct tumor markers, tight co-expression may reflect the disease pathway. In this case, it may be important to define clusters on the basis of the expression levels of the tumor markers in the pathway. However, if the 2 variables simply represent redundant variables (e.g. weight and body mass index) the corresponding clusters may be trivial. This potential problem is not particular to the RF dissimilarity but plagues other dissimilarities as well. As a data preparation step, it is often advisable to remove redundant variables.

4.1.3 Example *ExAddcl*1

This example is used to describe a situation where the *Addcl*1 RF dissimilarity leads to clusters while the *Addcl*2 RF dissimilarity does not. Further, it shows that the *Addcl*2 RF dissimilarity may give rise to spuriously distinct point clouds in classical multi-dimensional scaling plots. The example was originally motivated from the study of tumor marker data (Shi *et al.* 2004). As discussed in the motivational example, tumor marker expressions are often scored by the percentage of cells staining. These tumor marker scores (percentages) can have semi-continuous, highly skewed distributions since many observations may take on the value 0 and/or 100 percent. The data set *ExAddcl*1, contains 120 observations with 18 independent binary noise variables and 2 signal variables defined as follows. The first signal variable contains random uniform values for the first 60 observations but is set to 0 for the remaining 60 observations. The second signal variable is set to 0 for the first 60 variables but has random uniform values in the remaining 60 observations. Thus a scatter plot of the data along the 2 signal variables would exhibit an 'L' shape. Figure 5 shows that when the *Addcl*1 dissimilarity is used as input of classical MDS, one finds 2 distinct point clouds. These 2 distinct point clouds correspond to the arms of the 'L' in the aforementioned scatter plot of the signal variables. The variable importance measure shows that the first 2 (semi-continuous) variables are most important for defining these point clouds. In contrast, the *Addcl*2 dissimilarity leads to 4 distinct point clouds in a classical MDS plot. The coloring reveals that the point clouds do not correspond to the underlying true clustering structure determined by variables 1 and 2 (see also the variable importance plot). In additional (unreported) simulation examples, we have found that the *Addcl*2 RF dissimilarity may give rise to spuriously distinct point clouds when used as input of classical multi-dimensional scaling. Therefore, we advise to use isotonic MDS for the *Addcl*2 RF dissimilarity.

4.1.4 Example *ExAddcl*2

This example is used to describe a situation where the *Addcl*2 RF dissimilarity leads to clusters while the *Addcl*1 RF dissimilarity does not. The data consisted of 100 simulated observations with 20 uncorrelated variables. The first variable was simulated to have a Bernoulli (binary) distribution with hit probability 0.5 while the remaining 19 variables were drawn from a random uniform distribution. Figure 6 shows that only the *Addcl*2 dissimilarity in conjunction with isoMDS scaling leads to distinct point clouds corresponding to the values of the binary variable. In contrast, an MDS plot involving the *Addcl*1 RF dissimilarity does not exhibit any distinct point clouds. This result can be explained as follows. For the *Addcl*2 RF dissimilarity, synthetic are generated by sampling from the hyper-rectangle that contains the observed data, i.e. the variables have independent uniform distributions. If the observed data have several uniformly distributed variables and few binary variables, synthetic and observed data can only be distinguished by the binary variables, i.e. these variables will be important for the definition of the RF dissimilarity. This can be seen from the variable importance plot in Figure 6: when using *Addcl*2 samples, the binary variable is recognized to be important while the 19 random uniform variables are unimportant. When using *Addcl*1 sampling, the variable importance measure tells the opposite story. Here the synthetic data have the same distribution as the observed data, i.e. the synthetic class outcome has no relation to the variables. In this case, the Gini criterion will favor splitting variables that assume many values. Thus the resulting RF dissimilarity will be defined by variables with many values.

4.2 Which RF Dissimilarity Should Be Used?

The examples above illustrate which patterns and clusters can be detected as a result of using the *Addcl*1 or the *Addcl*2 RF dissimilarity as input in PAM clustering or multidimensional scaling plots. These patterns may or may not correspond to clusters in the Euclidean distance sense of the word. Depending on the research goal, the user should decide before the analysis whether the patterns found using a particular RF dissimilarity would be of interest.

To be more specific, consider example *Addcl*2 for which one variables is binary (an extreme case of a mixture distribution with well separated components) and the remaining variables follow random uniform distributions. All variables are independent. A Euclidean distance and the *Addcl*2 RF dissimilarity lead to distinct point clouds (clusters) in a multi-dimensional scaling plot. However, the *Addcl*1 RF dissimilarity does not lead to distinct point clouds in an MDS plot. How one judges the performance of the *Addcl*1 dissimilarity in this case depends on the research goal. In many applications, one would want that the MDS plot leads to two distinct point clouds corresponding to the values of the binary variable. Then the *Addcl*1 dissimilarity would be unsuitable. However, if the binary variable encodes patient gender, the resulting point clusters would probably be uninteresting. Then the *Addcl*1 RF dissimilarity would be suitable.

5. USING THE RF DISSIMILARITY IN PRACTICE

We find that the RF clustering results are robust with respect to the choice of the RF parameter mtry (the number of variables considered at each node split). This can be illustrated by studying noised up versions of example $ExAddcl1$, see Table 1. To be specific, noise was simulated to have a uniform distribution between $U[0, L]$ for different noise levels $L = 0.1, 0.2, 0.3$. The noise was added to the signal variables of those observations that had a signal variable value of 0. As is to be expected, the $Addcl2$ RF dissimilarity fails completely in this example and we discuss the performance of the $Addcl1$ RF dissimilarity in the following. Table 1 records how the Rand index changes as a function of $mtry$. Reasonably large values of $mtry$ (default is the square root of the number of variables) lead to good agreement (high Rand index) between RF clusters and the underlying true cluster label. Relatively large values of $mtry$ ensure that the signal variables have an increased chance of being chosen as node splitters. Very low values of $mtry$ lead to inferior performance. Other (unreported) simulations show that very high values of $mtry$ can lead to poor clustering performance if the cluster structure is determined by many variables.

We find that the value of mtry that minimizes the out-of-bag error rate does not necessarily lead to the best clustering performance. This can again be illustrated using Table 1: the values of $mtry$ that lead to the smallest error rate, do not necessarily lead to the highest Rand index. The OOB error rate simply measures how different the observed data are from the synthetic data according to the RF predictor.

Roughly speaking, there is an inverse relationship between OOB error rate and the cluster signal in the data. This can be seen in Figure 7, which depicts the relationship between the adjusted Rand index and the out-of-bag error rate for different noised up versions of examples $ExAddcl1$ and $ExAddcl2$. Low noise levels (high signal) lead to a low OOB error rate, which in turn correspond to high values of the Rand index. Figure 7 also shows that an OOB error rate of 50% may correspond to high signal data in $Addcl1$ sampling.

5.1 Computational Considerations

When dealing with labelled data, RF predictors do not overfit the data: the more trees a forest contains, the more accurate it is (Breiman 2001). When using RF to arrive at a dissimilarity measure, a similar property holds: the more trees are grown, the better the dissimilarity measure will be able to separate observations. In the examples of this paper, we typically used 2000 trees per forest.

We find that the RF dissimilarity can vary considerably as a function of the particular realization of the synthetic data. When it comes to the RF dissimilarity, we strongly recommend to average the results of multiple forest constructions. The RF dissimilarity is subject to Monte Carlo fluctuations if few forests are grown or if the total number of trees is low. To study how the cluster performance depends on the number of forests and the number of trees per forest, we averaged the results of several of simulated examples described above. We used the Rand index to measure the agreement between the RF cluster label and the true (simulated) cluster label. Figure 8 shows boxplots that visualize the distribution of the Rand index for different numbers of forests. The total number of trees was fixed to 5000 and the number of trees per forest was chosen accordingly. *Overall, we find that the results are fairly robust with respect to the number of forests as long as the proximities of at least 5 forests are averaged.* On the other end of the spectrum (e.g. 5000 forests with 1 tree each), the clustering performance is diminished. Incidentally, we averaged the RF similarities of at least 8 different forests in the examples of this paper. As pointed out by a reviewer: instead of averaging the proximities of several forests, it may be easier to run a single forest with a very large second class (suitably re-weighted).

The computation of the RF dissimilarity can be parallelized in 2 ways: the tree computations can be parallelized and the forest computations can be parallelized since the final RF dissimilarity is the average over the forest dissimilarities.

6. DISCUSSION

Our motivational example provides evidence that the RF dissimilarity can be particularly useful for detecting clusters that are described by thresholding rules. It has been found that such clusters are clinically meaningful in tumor marker expression studies (Shi *et al.* 2004; Seligson *et al.* 2005).

In general, random forest dissimilarities are a class of dissimilarities that are highly dependent on how the synthetic data are generated. Example $ExNull$ in the appendix provides an extreme example for showing that the RF dissimilarity may be very different from the Euclidean distance when synthetic observations are chosen in an artificial way. Further, the PAM clustering results reported in table 1 show that the RF dissimilarity is in general very different from the Mahalanobis distance.

We describe the patterns and clusters that can be detected as a result of using different random forests dissimilarities in PAM clustering or multidimensional scaling plots. Depending on the research goal, the user should decide before the analysis whether the patterns found using a particular RF dissimilarity would be of interest. For the motivational example (tumor marker data), the clusters have a rectangular shape along highly dependent variables, which is attractive in this application.

It is relatively difficult to provide a geometric interpretation of RF clustering, which is why we have discussed several strategies for disentangling the relationship between variables and the RF dissimilarity (a Euclidean distance approximation, tree predictors, boxplots, etc). In the appendix, we list several examples that allow for a simple geometric interpretation.

Important properties of the RF dissimilarity derive from the underlying tree predictors and the forest construction.

The RF dissimilarity handles mixed variable types well, i.e. it can handle both categorical and ordered variables in a simple and natural way. It inherits this property from the underlying tree predictors (Breiman *et al.* 1984).

The RF dissimilarity is invariant with respect to monotonic transformations of the variable values. This is because the node splits (based on the Gini index) only depend on the variable ranks (Breiman *et al.* 1984). But it is not invariant to orthogonal transformations (rotations) in the original Euclidean space as demonstrated by example ExX in the appendix.

The RF dissimilarity is robust to outlying observations because it is based on feature ranks and bagging is used in the random forest construction.

The RF dissimilarity allows for several straightforward approaches of dealing with missing variable values. On the one hand are approaches that have been suggested for tree predictors, e.g. the use of surrogate splits (Breiman *et al.* 1984) or the 'missings together'(MT) approach (Zhang and Bracken 1996). On the other hand, is an iterative imputation approach suggested by Breiman (2003). A study of these missing data approaches is beyond the scope of this manuscript.

The RF dissimilarity easily deals with a large number of variables due to its intrinsic variable selection. The *Addcl*1 RF dissimilarity weighs the contribution of each variable on the dissimilarity according to how dependent it is on other variables. In general, the RF dissimilarity will use those variables in its construction that distinguish observed from synthetic data according to the underlying tree predictors.

The *Addcl*1 RF dissimilarity can be used as input of classical or isotonic (non-parametric) MDS plots. Usually, we use classical MDS for the *Addcl*1 dissimilarity. But we strongly recommend to use isoMDS for the *Addcl*2 RF dissimilarity to avoid the detection of spurious patterns (see example *ExAddcl*1).

The *Addcl*2 RF dissimilarity favors discrete variables over continuous variables as illustrated in example *ExAddcl*2. This may or may not be a desirable property of the *Addcl*2 RF dissimilarity as discussed in section 4.2. In general, we advocate the use of the *Addcl*1 RF dissimilarity over the use of the *Addcl*2 RF dissimilarity. Breiman's recent Fortran program only implements the *Addcl*1 dissimilarity.

There is empirical evidence that the *Addcl*1 RF dissimilarity can be superior to standard distance measures in several applications (Allen *et al.* 2003; Shi *et al.* 2004). But it is clear that other dissimilarity measures will be superior in other applications. For example, when it is desirable to weigh the variable contributions according to their scale, the RF dissimilarity may be a bad choice since it only considers feature ranks. If possible, the choice of the dissimilarity should be guided by the research goal and by domain knowledge.

ACKNOWLEDGMENT

We thank the reviewers and Andy Yip (UCLA) for very valuable comments. T.S. was a doctoral trainee supported by the UCLA IGERT Bioinformatics Program funded by NSF DGE 9987641. We thank David Seligson, UCLA, for the tumor marker data.

APPENDIX: SIMPLE GEOMETRIC EXAMPLES

Here we provide rather artificial examples that allow for a simple geometric interpretation of RF clustering and that highlight some important properties. But these examples are not meant to show the advantages of the RF dissimilarity directly.

Example $ExNULL$**:** This example allows for a simple geometric interpretation of RF clustering and illustrates that the RF dissimilarity may be very different from the Euclidean distance when synthetic observations are chosen in an artificial way. The data consisted of 100 observations with 20 independent, random uniform variables, i.e. there are no clusters in the Euclidean distance sense of the word (Figure 9a). However, by generating synthetic observations in an artificial way, one can arrive at an RF dissimilarity that leads to 4 distinct point clouds in a multi-dimensional scaling plot. Specifically, synthetic data are simulated such that the first 2 variables are constant with value equal to 0.5 while the remaining 18 variables have independent random uniform distributions. Clearly, variables 1 and 2 distinguish observed from synthetic data. Note that both MDS plots show 4 distinct point clouds in the observed data (Figures 9c and d). The four point clouds correspond to the four quadrants in Figures 9b and the boundaries of the quadrants correspond to hyperplanes that separate synthetic from observed data. In this example, the RF clusters have a simple geometrical interpretation: *RF clusters are delineated by hyper-rectangles along (important) variable axes that isolate observed data from synthetic data.* In general, the RF construction can lead to much more flexible boundaries between observed and synthetic data. But in this particular example, we find that hyper-rectangles are a good approximation. It is clear that the hyper-rectangles originate from the underlying tree predictors which recursively partition the feature space along the variable axes.

Example ExX**:** This example is used to illustrate that RF clustering is in general not rotation invariant in a Euclidean space. Further, it allows for a simple geometric interpretation of RF clustering. The data consisted of 300 points that formed a cross in the 2-dimensional Euclidean plane. There are two versions of the data: one where the cross lines up with the coordinate axes (Figure 10b) and one where the cross is rotated by 45 degrees (Figure 10a). The RF construction aims to separate observed from synthetic data by cuts along variable axes. When the synthetic data are generated by $Addcl2$ sampling, this is not possible and RF clustering fails to detect the 4 cross arms (Figure 10a). However, when the data are rotated by 45 degrees, cuts along variable axes succeed at separating observed from synthetic data (Figure 10b).

REFERENCES

Allen, E., Horvath, S., Kraft, P., Tong, F., Spiteri, E., Riggs, A., & Marahrens, Y. (2003), "High Concentrations of LINE Sequence Distinguish Monoallelically-Expressed Genes", *Proceedings of the National Academy of Sciences*, 100(17), 9940-9945.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall, New York.

Breiman, L. (2001), "Random forests", *Machine Learning*, 45(1), 5–32.

Breiman, L. (2003), "Random Forests Manual v4.0", Technical report, UC Berkeley, ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf.

Cox, T. F., and Cox, M. A. A. (2001), *Multidimensional scaling*, : Chapman and Hall/CRC.

Cox, D. R., Oakes D. (1990) Analysis of survival data. New York, NY: Chapman and Hall.

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., Mischel, P., and Nelson, S. F. (2004), "Gene expression profiling of gliomas strongly predicts survival." *Cancer Research*, 64(18), 6503-6510.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The elements of statistical learning: data mining, inference, and prediction*, Springer series in statistics, New York: Springer.

Hubert, L. and Arabie, P. (1985), "Comparing partitions." *J. Classification*, 2, 193–218.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J.of the American Statistical Association*, 53, 457–48.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding groups in data : an introduction to cluster analysis*, Wiley series in probability and mathematical statistics. Applied probability and statistics,, New York: Wiley.

Kruskal, J. B., and Wish, M. (1978), *Multidimensional scaling*, Beverly Hills, Calif.: Sage Publications.

Liaw, A. and Wiener, M. (2002), "Classification and regression by *randomForest*." *R News: The Newsletter of the R Project* (http://CRAN.R-project.org/doc/Rnews/), 2(3), 18–22.

Liu, B., Xia, Y., & Yu, P. (2000), "CLTree-Clustering through decision tree construction", Technical report, IBM Research.

R Development Core Team (2004), "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association*, 66(336), 846–850.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., & Brown, P. O. (2000), "Systematic variation in gene expression patterns in human cancer cell lines." *Nature Genetics*, 24(3), 227-235.

Seligson, D. B., Horvath, S., Shi, T., Yu, H., Tze, S., Grunstein, M., Kurdistani, S. (2005), "Global histone modification patterns predict risk of prostate cancer recurrence." *Conditionally Accepted*

Shepard, R. N., Romney, A. K., Nerlove, S. B., & Board., M. S. S. (1972), *Multidimensional scaling; theory and applications in the behavioral sciences*, New York: Seminar Press.

Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., Horvath, S. (2004), "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma." *Modern*

Pathology, Oct 29

Venables, W. N. and Ripley, B. D. (2002), *Modern applied statistics with S-PLUS. Fourth Edition.* Springer-Verlag, New York.

Zhang, H. P., Bracken, M. B. (1996), "Tree-based, two-stage risk factor analysis for spontaneous abortion." *American Journal of Epidemiology*, 144:989-996.

Table 1: Simulation study $ExAddcl1$. Different levels of noise are added to the signal variables. The RF clustering performance (measured by the adj. Rand index) is recorded for different ways of generating synthetic data($Addcl1$ and $Addcl2$), different MDS procedures (classical and isotonic MDS), and different numbers of random variables ($mtry$). We also report the Rand indices when a Euclidean or a Mahalanobis distance is used on the same data. Note that the $Addcl1$ dissimilarity leads to the highest Rand indices while the other dissimilarities fail to detect the clusters.

			Addcl1			**Addcl2**		**Euclidean**	**Mahalanobis**
noise	*mtry*	*Error*	*Rand*	*Rand*	*Error*	*Rand*	*Rand*	*Rand*	*Rand*
			($cMDS$)	($isoMDS$)		($cMDS$)	($isoMDS$)		
0.1	1	0.6	0	0	0	0.014	0	0.011	0.002
0.1	3	0.61	0.24	0.27	0	0	0	0.011	0.002
0.1	6	0.66	0.75	0.81	0	0.01	0	0.011	0.002
0.1	12	0.55	0.81	0.69	0	0	0	0.011	0.002
0.1	15	0.57	0.63	0.75	0	0	0	0.011	0.002
0.2	1	0.55	0.003	0.001	0	0.02	0	0.011	0.003
0.2	3	0.62	0.004	0.12	0	0.02	0	0.011	0.003
0.2	6	0.61	0.48	0.43	0	0	0.002	0.011	0.003
0.2	12	0.58	0.53	0.44	0	0	0	0.011	0.003
0.2	15	0.56	0.53	0.48	0	0	0	0.011	0.003
0.3	1	0.66	0.012	0.038	0	0	0	0.055	0.005
0.3	3	0.6	0	0.074	0	0	0	0.055	0.005
0.3	6	0.61	0.44	0.17	0	0	0	0.055	0.005
0.3	12	0.61	0.39	0.39	0	0	0	0.055	0.005
0.3	15	0.61	0.48	0.44	0	0	0	0.055	0.005

FIGURE CAPTIONS

Figure 1. a) The histogram of the first tumor marker expression illustrates that tumor marker expressions tend to be highly skewed. b) Kaplan-Meier plots visualize the survival time distributions of tumor sample groups defined by cross-tabulating random forest cluster membership with the Euclidean distance based cluster membership. Clearly, the RF dissimilarity leads to clusters that are more meaningful with respect to post-operative survival time. c) A heatmap depiction of the tumor marker expressions which are standardized to mean 0 and variance 1 for each marker. Rows correspond to tumor samples and columns to tumor markers. The samples are sorted according to the colors defined in b). The column-side color bar represents the different group memberships as shown in a). Clearly, samples in RF cluster 2 (blue and green side bar colors) show low expressions in tumor markers 1 and 2. d) A scatter plot involving tumor markers 1 and 2, which have been colored according to RF and Euclidean cluster membership defined in b). The horizontal and vertical lines correspond to threshold values found when using a tree predictor for predicting cluster membership on the basis of the 2 tumor markers. The upper right hand corner (tumor marker 1 expression $> 65\%$ and marker 2 expression $> 80\%$) is enriched with red and black points, i.e. with RF cluster 1 samples.

Figure 2. The wine data are depicted in two dimensional scaling plots. The observations are labelled by the cultivar (external label) and colored by RF cluster membership. The figures in the first and second row correspond to the $Addcl1$ and the $Addcl2$ RF dissimilarity, respectively. The figures in the first and second column correspond to classical and isotonic MDS plots. The last column contains scatter plots of the RF dissimilarities versus the Euclidean distance involving the standardized variable ranks.

Figure 3. The *ExRule* data are depicted in 2 dimensional MDS plots. Plots in the first and second columns use classical and isotonic MDS, respectively. The MDS plots in the upper, middle and lower rows are based on the $Addcl1$ RF, $Addcl2$ RF, and the Euclidean dissimilarity, respectively. The coloring of the points is explained in the text. Note that the $Addcl1$ RF dissimilarity leads to 2 point clouds that are determined by variables 1 and 2, see also the RF variable importance plot in c). In contrast, the $Addcl2$ RF dissimilarity and the Euclidean distance lead to point clouds that are mainly distinguished by variable 3. The RF variable importance plot in f) shows that the $Addcl2$ RF dissimilarity focuses on variable 3 in its construction. The variance

plot in i) shows that variable 3 has the highest variance, which explains why it dominates the definition of the Euclidean distance.

Figure 4. The *ExDep* data (5 random uniform variables) are depicted using the first two (highly dependent) variables, $var1$ and $var2$. The observed values lie on the diagonal ($var1 = var2$) and are colored and labelled by the Addcl1 RF clustering label, i.e. the result of using the RF dissimilarity as input of PAM clustering (k = 3 clusters). The synthetic data are colored in turquoise and labelled by '5'. b) Classical multi-dimensional scaling plot of the Addcl1 RF dissimilarity. The order of the colors in the 2 figures reveals that the RF clusters correspond to low, medium, and high values of the variable $var1$.

Figure 5. The $ExAddcl1$ data are depicted in 2 dimensional scaling plots. The observations are labelled and colored by the true underlying cluster structure (the arms of the 'L' shape defined by signal variables 1 and 2). The MDS plots in the top and bottom row depict the $Addcl1$ and $Addcl2$ dissimilarities, respectively. Plots in the first and second column use classical and isotonic MDS, respectively. The RF variable importance measure is plotted in the third column. The variables are enumerated on the x-axis and the corresponding variable importance on the y-axis.

Figure 6. The $ExAddcl2$ data are depicted in 2 dimensional scaling plots. The observations are labelled and colored by the true underlying cluster structure (the values of the binary variable). The MDS plots in the top and bottom row depict the $Addcl1$ and $Addcl2$ dissimilarities, respectively. Plots in the first and second column use classical and isotonic MDS, respectively. The RF variable importance measure (based on node purity) is plotted in the third column: variables are listed on the x-axis, variable importance is on the y-axis.

Figure 7. The relationship between the adjusted Rand index and the out-of-bag error rate. a) The results for the $Addcl1$ RF dissimilarity for noised up versions of Example $ExAddcl1$, see also Table 1. High dependency between variables 1 and 2 (low noise levels) lead to low OOB error rates, which in turn correspond to high values of the Rand index. b) The analogous results of the $Addcl2$ RF dissimilarity for noised up versions of Example $ExAddcl2$.

Figure 8. Boxplots of the Rand index of agreement (between cluster labels and simulated

underlying true dependence structure) versus the number of forests used. The number of trees per forest was chosen such that the total number of trees was 5000. The middle line in each boxplot represents the median; the upper and lower hinges of the box show the median of the upper and lower halves of the data. The ends of the line segments attached to the box extend to the smallest data value and the largest data value.

Figure 9. a) The $ExNULL$ data are depicted using only the first two variables, $var1$ and $var2$. No dependence structure is present. b) Same plot as in a) but 100 synthetic points have been inserted in the center of the plot (colored in turquoise). The original observations are labelled and colored by the same color and label as in plot c). c) An isotonic MDS plot that uses the RF dissimilarity as input. Clearly, 4 distinct point clouds emerge. The coloring and the labelling shows that the 4 point clouds correspond to the 4 quadrants around the point $(0.5, 0.5)$ in plot b). d) Classical MDS plot that uses RF dissimilarity as input.

Figure 10. Scatterplot of the ExX data along the two variable axes. Points are labelled by the arm of the cross ($k = 1, \ldots, 4$) and colored by their RF cluster label. The synthetic data are generated by $Addcl2$ sampling and are represented by solid, turquoise points. The RF construction aims to separate observed from synthetic data by cuts along variable axes. In figure a) this is difficult and RF clustering assigns the points along intervals of the x-axis. In figure b) cuts along variable axes succeed at separating synthetic from observed observations and the resulting RF clusters correspond to the 4 arms of the cross. Incidentally, when a Euclidean distance is used as input of $k = 4$ medoid (PAM) clustering, the resulting clusters correspond to the arms of the cross irrespective of the orientation of the cross.

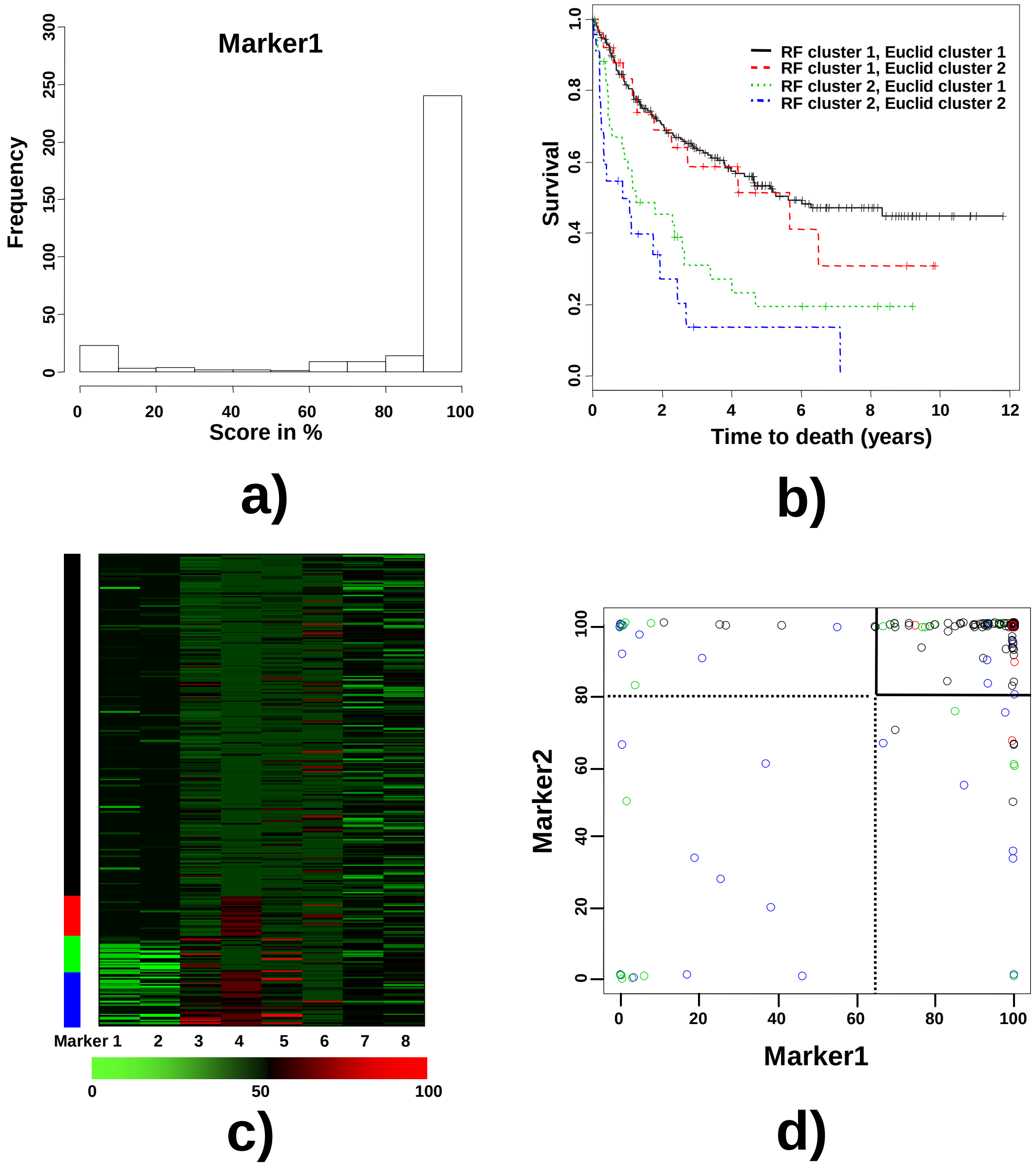
Marker1
Frequency
0
50
100
150
200
250
300
0
20
40
60
80
100
Score in %
a)
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Time to death (years)
RF cluster 1, Euclid cluster 1
RF cluster 1, Euclid cluster 2
RF cluster 2, Euclid cluster 1
RF cluster 2, Euclid cluster 2
b)
Marker 1
2
3
4
5
6
7
8
0
50
100
c)
Marker2
0
20
40
60
80
100
Marker1
0
20
40
60
80
100
d)

Figure 1:

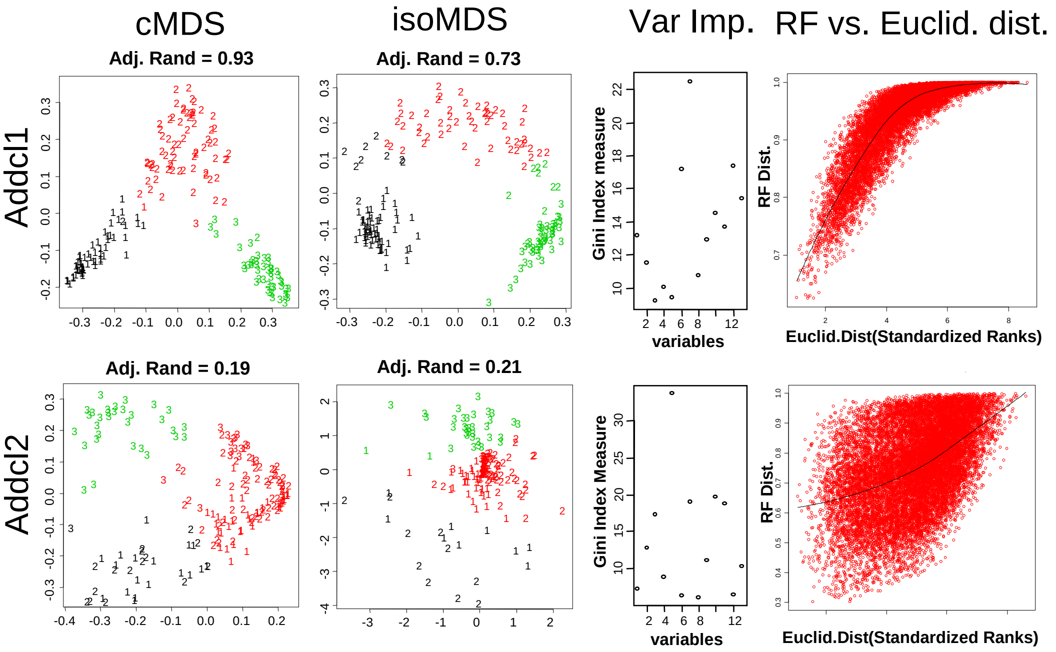
cMDS
isoMDS
Var Imp.
RF vs. Euclid. dist.
Adj. Rand = 0.93
Adj. Rand = 0.73
Addcl1
Gini Index measure
variables
RF Dist.
Euclid.Dist(Standardized Ranks)
Adj. Rand = 0.19
Adj. Rand = 0.21
Addcl2
Gini Index Measure
variables
RF Dist.
Euclid.Dist(Standardized Ranks)

Figure 2:

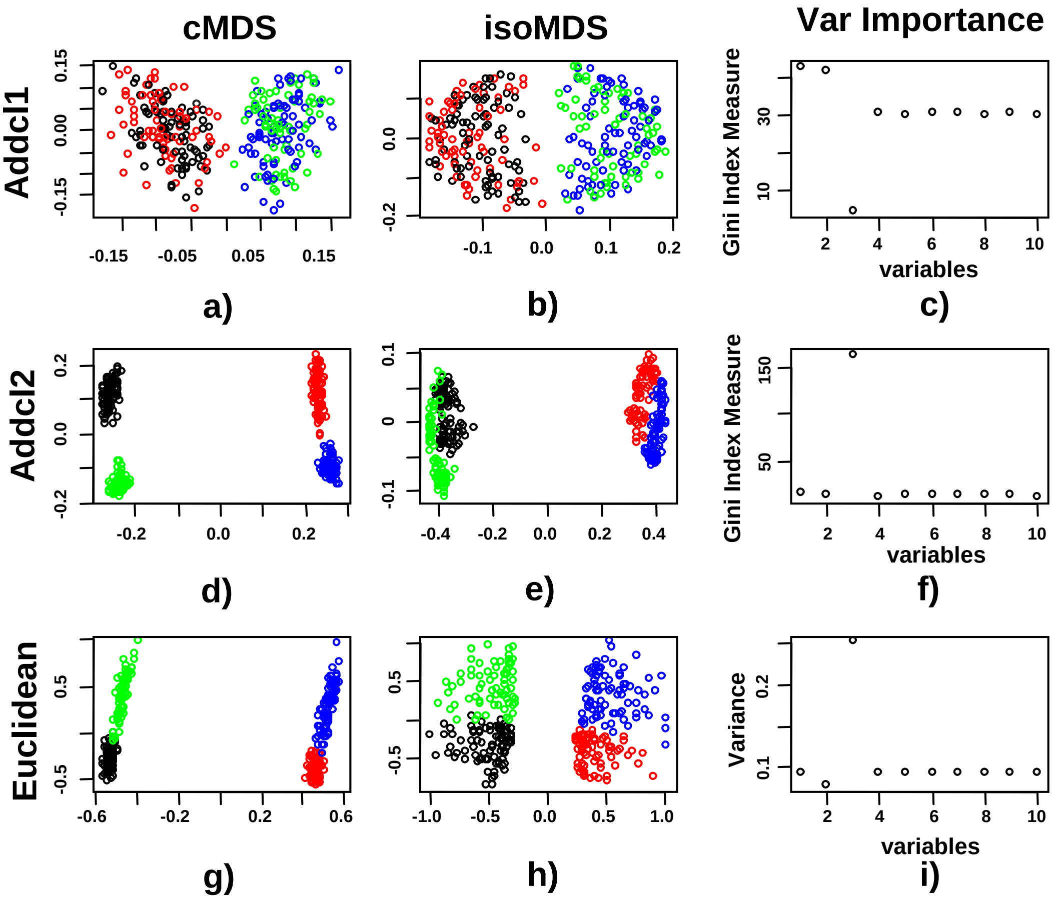
cMDS
isoMDS
Var Importance
Addcl1
Addcl2
Euclidean
Gini Index Measure
Variance
variables
a)
b)
c)
d)
e)
f)
g)
h)
i)

Figure 3:

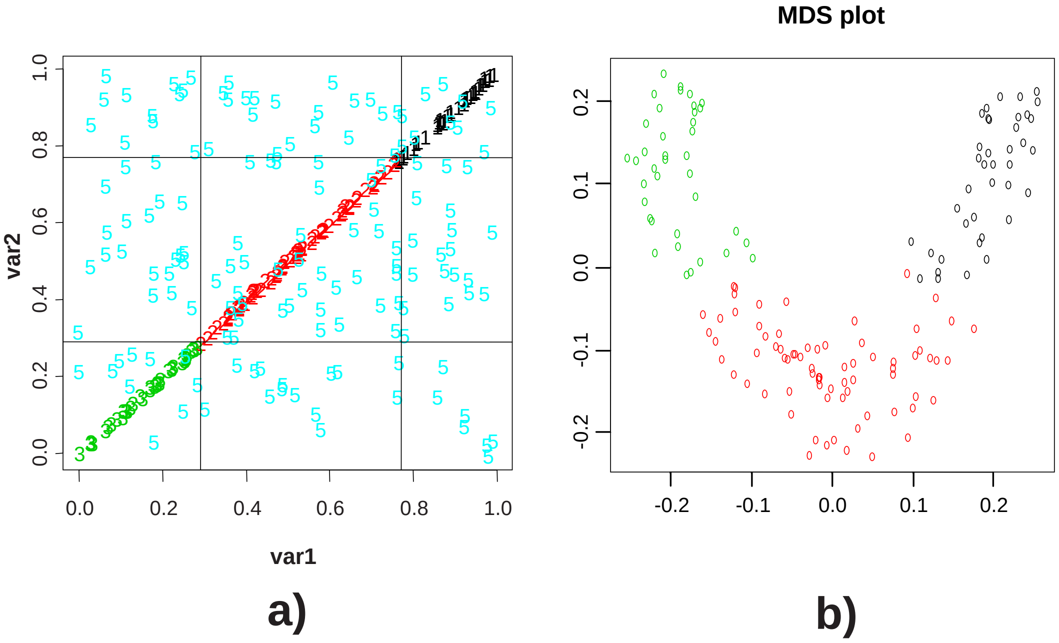

Figure 4:

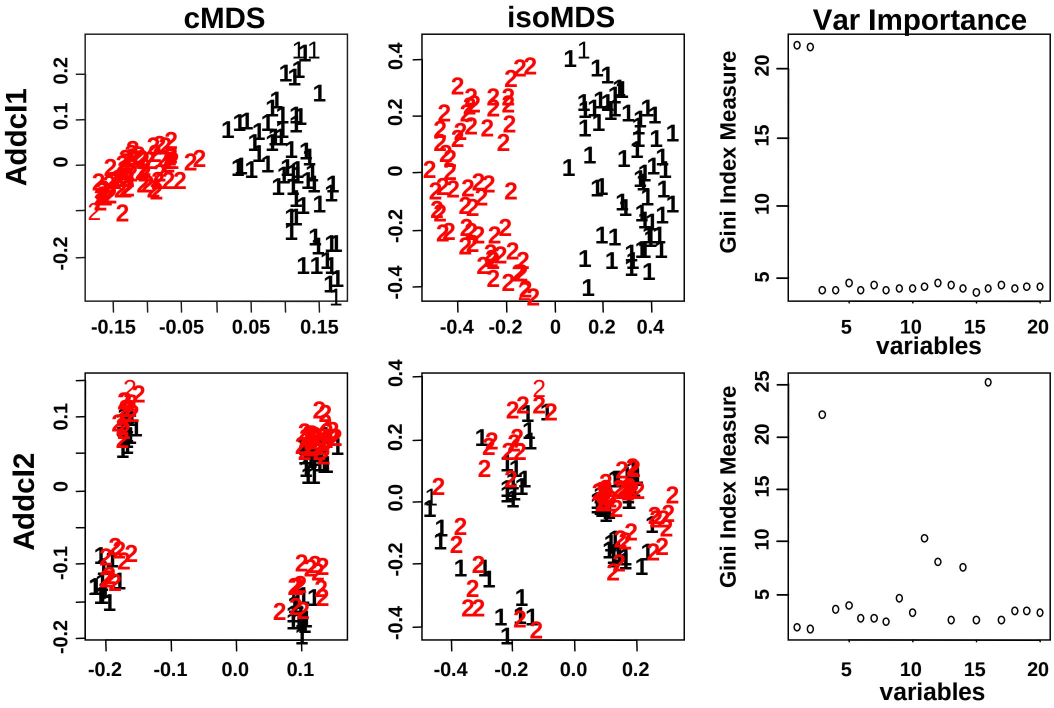
cMDS
isoMDS
Var Importance
Addcl1
Addcl2
Gini Index Measure
variables
-0.15
-0.05
0.05
0.15
-0.2
-0.1
0.0
0.1
-0.4
-0.2
0
0.2
0.4
5
10
15
20
25

Figure 5:

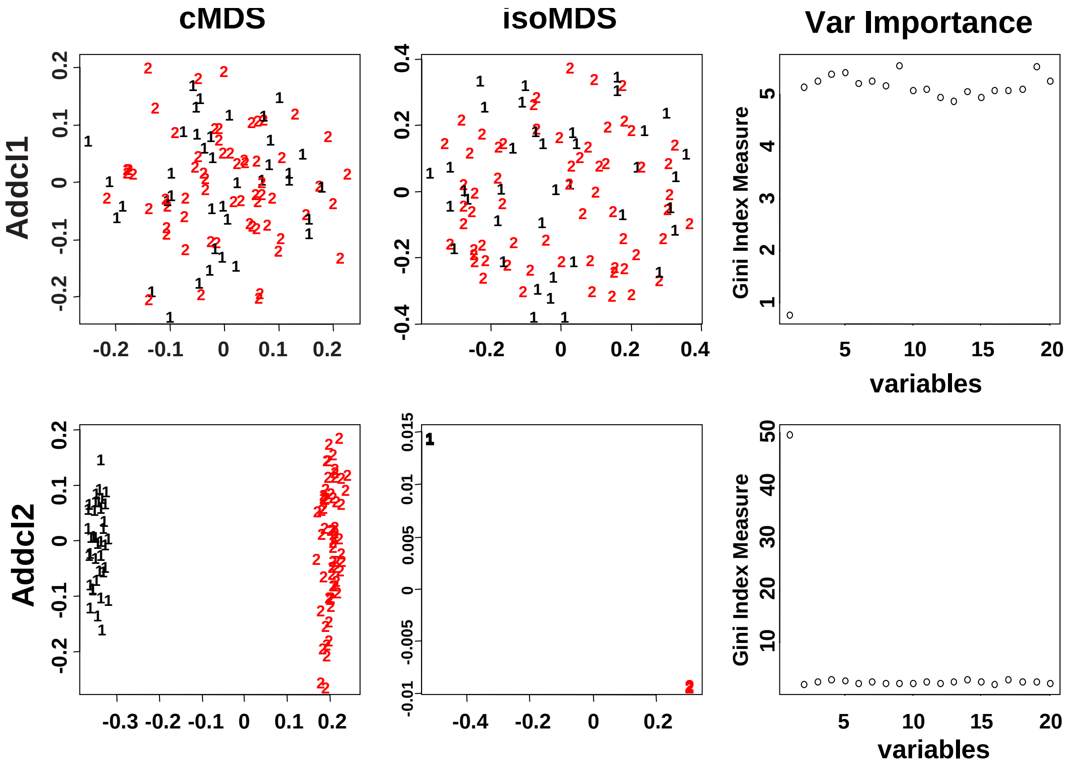
cMDS
isoMDS
Var Importance
Addcl1
Addcl2
Gini Index Measure
variables
-0.2 -0.1 0 0.1 0.2
-0.2 0 0.2 0.4
5 10 15 20
-0.3 -0.2 -0.1 0 0.1 0.2
-0.4 -0.2 0 0.2
0.015 0.01 0.005 0 -0.005 -0.01
10 20 30 40 50
1 2 3 4 5

Figure 6:

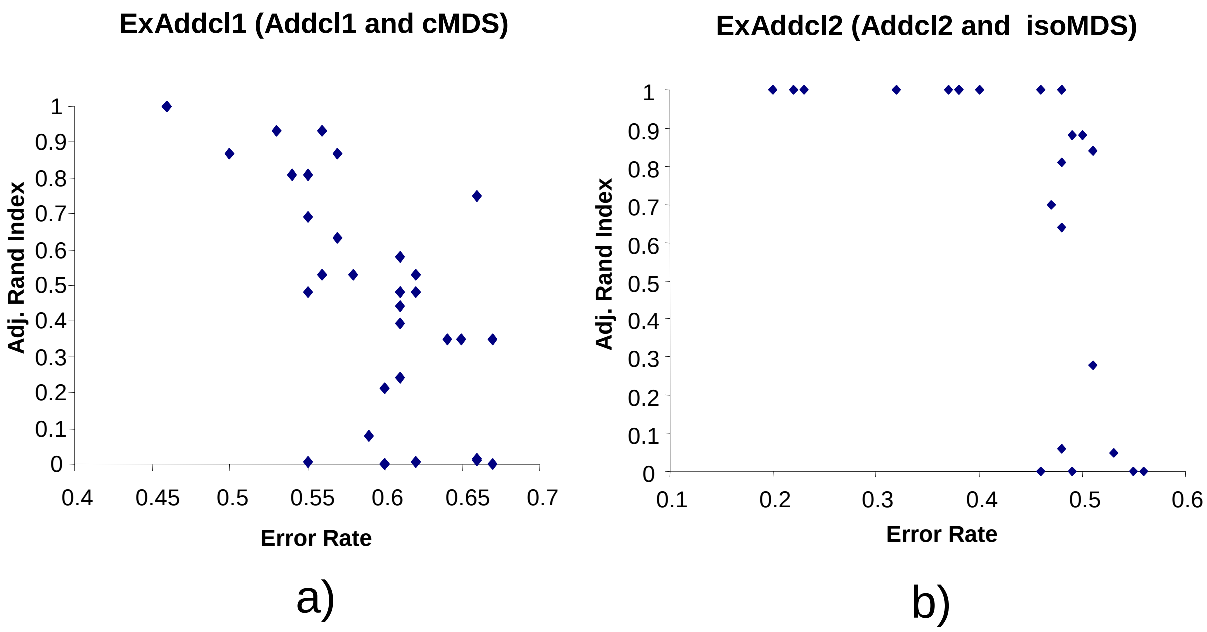
ExAddcl1 (Addcl1 and cMDS)
Adj. Rand Index
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.4
0.45
0.5
0.55
0.6
0.65
0.7
Error Rate
a)
ExAddcl2 (Addcl2 and isoMDS)
Adj. Rand Index
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
Error Rate
b)

Figure 7:

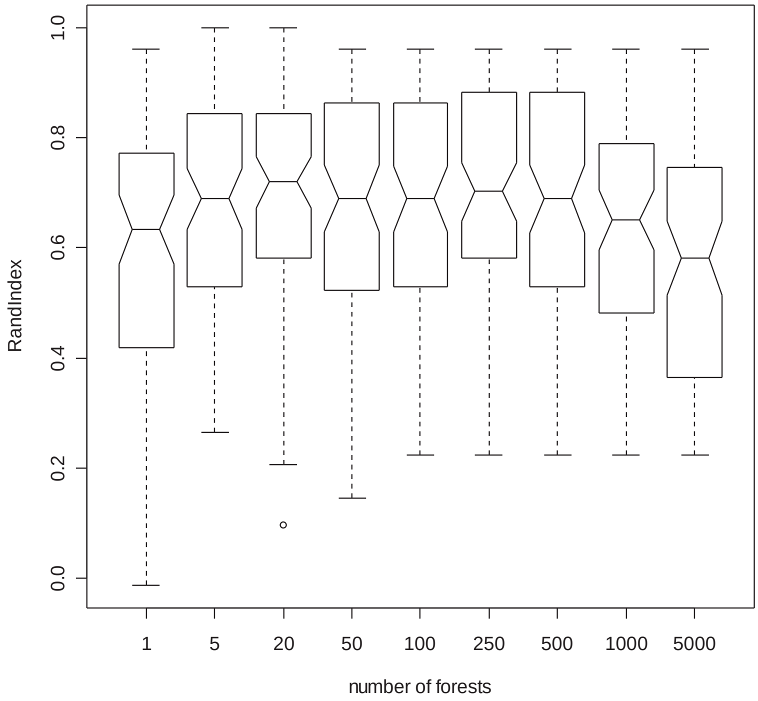
RandIndex
0.0
0.2
0.4
0.6
0.8
1.0
1
5
20
50
100
250
500
1000
5000
number of forests

Figure 8:

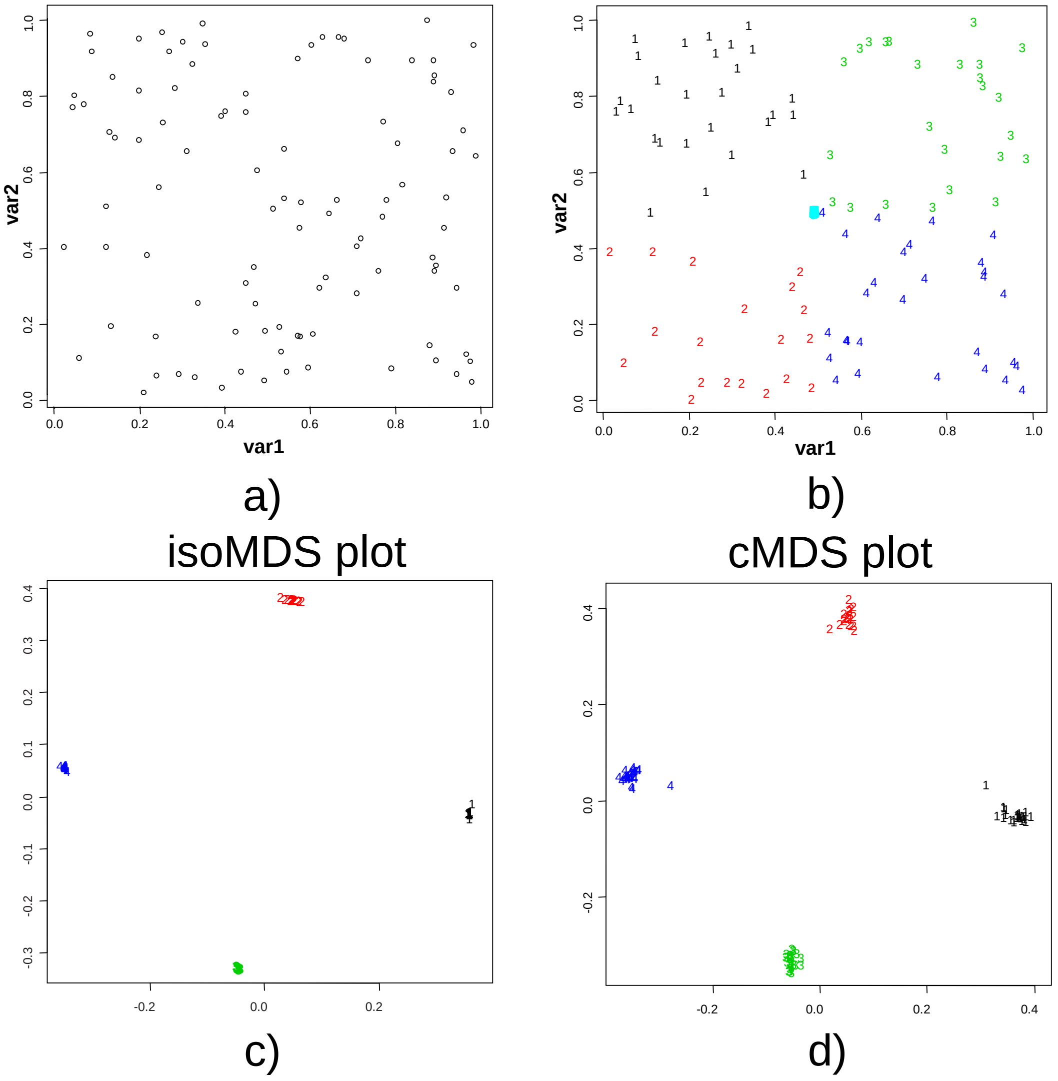

Figure 9:

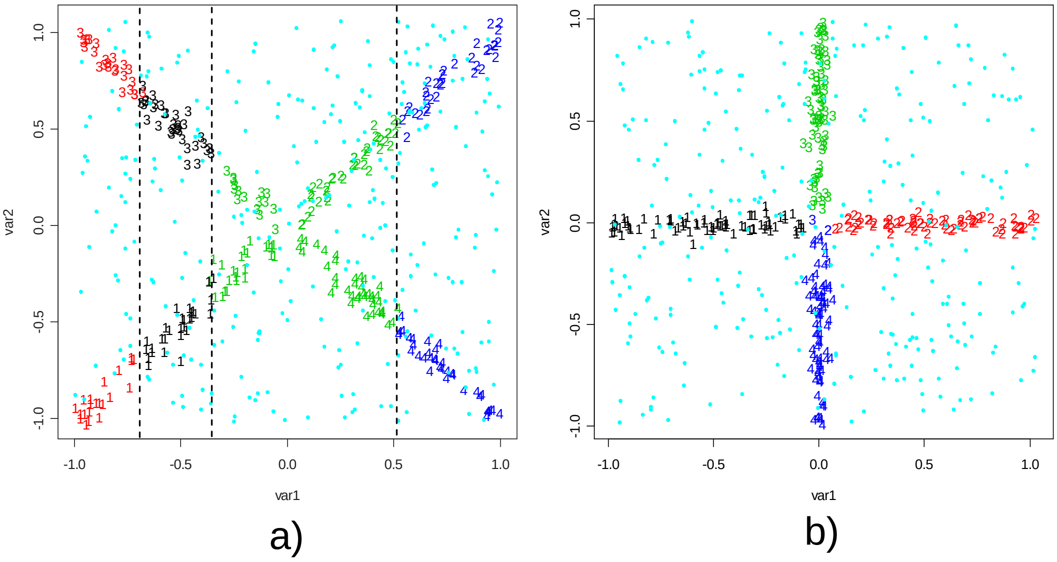

Figure 10: