Random Forest for Paired Data

M.W. Mitchell[1], J.E. Wulff[1], P.R. Gunst[1]
[1] Metabolon, 617 Davis Drive, Durham, NC 27713

Abstract

Random forest classification is a supervised method that has many advantages over other multivariate methods: it is non-parametric, it is invariant to transformation, and it does not overfit the data, requires no variable selection, and it is fairly easy to implement in ***R***. In particular, it works well with data from the –*omics* sciences such as genomics and metabolomics where the number of variables (p) is much greater than the number of subjects (n), i.e., where "$p >> n$." The out-of-bag error (OOB error) is a good estimate of future performance. However, when the data consists of matched-pairs, such as cancerous and benign tissue from the same subject or time course data, the OOB-error can be severely pessimistic, especially when the intra-subject correlation is very high. In some cases the OOB-error is 100%, indicating perfect misclassification, when the true misclassification is much lower. Additionally, with the computations of variable importance, noise variables with high intra-subject correlation rank lower than those with low intra-subject correlation. We perform an extensive simulation study in order to compare cross-validation techniques for improving the estimate of the error; and we compare different sampling techniques when building the forest to improve the estimate of the error, as well as improve the predictive ability. We also compare the methods on a human metabolomics study. Computing the residuals for each subject performed the best, but has problems with practical application. Sampling by subject performed well, but was comparable to the standard random forest. Leaving one-subject-out cross-validation corrects the bias of the out-of-bag error.

Key Words: random forest, matched pairs, -omics sciences, metabolomics

1. Introduction

Random forest classification (Breiman, 2001) is an ensemble method based on combining a large number of classification trees. An example of such a decision tree for the variable x is the following: if $x > 1$, then it is classified as "A," and classified as "B" otherwise. For each tree, a subset of the observations is used to build the decision rule (the "in-bag" samples), while the remaining samples are used to determine the final prediction (the "out-of-bag" samples). The final classification is based on the majority of the "votes," e.g., if an observation is classified as "A" in 55% of the trees for which it was an out-of-bag sample and classified as "B" in the remaining 45%, its final classification is "A." Thus, the final classification for an observation is based only on decision trees where it was not one of the samples that created the decision rule. The out-of-bag error (OOB error) is computed as the classification error. Also for each tree, only a sample of the variables is considered at each split in order to decrease the correlation between trees in the forest.

Random forest has many advantages: it does not overfit the data, it is invariant to transformation, it is non-parametric, it requires no variable selection, is easy to implement in ***R*** (R Core Team, 2014), and works well even when the number of variables, *p*, is much greater than the number of observations, *n*. This latter property makes it especially useful in the *–omics* sciences such as genomics (Amaratunga et al., 2008; Chen and Ishwaran, 2012) and metabolomics (Gall et al., 2010; Schuler et al., 2014; Innopolito et al., 2014). In these applications, a random forest may be fitted to assess the ability to classify observations in the data set, as well as predict the class statuses of new observations. Additionally, random forest may be used as part of a biomarker selection process by using one of its importance measures.

When there are technical replicates for each subject, i.e., two or more observations from each subject with the same phenotype, the out-of-bag error (OOB-error) is often an underestimate, i.e., the error rate is too optimistic. This is true especially when there is high intra-subject variability relative to the total variability, which results in high intra-subject correlation (ISC). For example, if subject 1 is “A” and has two measurements that are very similar, every time one sample is an in-bag sample it will correctly predict the other sample if it is in the out-of-bag portion. Methods for adapting the random forest or other classification techniques have been proposed for this case. Brenning and Lausen (2008) propose using all of the observations for each subject with subject-level cross-validation. This was applied to a glaucoma data set where there were observations from both eyes. The subjects with the same phenotype for both eyes were used in the final analysis. Adler, Brenning et al. (2011) compare various bootstrapping techniques for several multivariate classification techniques with application to the aforementioned glaucoma data set. Adler, Popatov, and Lausen (2011) extend these methods to an arbitrary number of observations per subject and apply the method to the glaucoma data set. In that paper, the aim is to predict future observations from the same patient, rather than classification of observations from new subjects. The glaucoma data set consisted of 61 variables and 372 subjects. Karpievitch et al. (2009) compare random forest classification for the following cases: (1) unmodified, (2) subject-level averages, and (3) subject-level bootstrapping (“RF++”) with the RF++ method performing the best overall. The method was then applied to a human serum cancer data set where the variables are MALDI-TOF ms data. As with the previous proposals, the methods were applied to data sets where the phenotype is the same for all measurements from the same subject; but unlike the glaucoma data set, the method was applied to a data set with a large number of variables and a small number of subjects (38 subjects, 507 spectra). The subjects had varying numbers of technical replicates (7 to 24).

When the observations from the same subjects have different phenotypes; for example, when collecting tissue samples, there may be one tissue sample collected from a cancerous portion of the tissue and the other sample is taken from a noncancerous portion of the tissue. In this case, the OOB-error can severely overestimate the true error, i.e., it is too *pessimistic*. If there is no mean change and there is high ISC, then it can be the case that the out-bag observations are all misclassified for a given tree. An example is shown in Table 1. Here the rule: $x < 1.6$, then “B,” else “A” correctly classifies all the in-bag samples, but each out-of-bag sample is incorrectly classified. This is a result of the strong similarity between two measurements from the same subject. Furthermore, when random forest randomly selects the variables to consider for each tree, a variable such as this one would be chosen in building the forest because of its strong performance on the in-bag samples.

The focus of this manuscript will be for the case when there are two observations per subject where each observation has a different phenotype and then the number of variables is much larger than the number of observations. In contrast to some of the previously proposed methods, the accuracy of the methods will be assessed on the ability to predict new observations from different subjects, rather than the ability to assess new observations from the same subjects (although the methods still apply). In addition to the accuracy in predicting the status of new observations, the estimated accuracy from the OOB error and cross-validation techniques will be compared. Furthermore, the importance of the variables will be compared for each method. The permutation-based importance measure, rather than the GINI index will be used. The simulated data sets reflect dimensions seen in metabolomics studies. Finally, the methods are applied to a human metabolomics data set.

2. Methods of Random Forest Classification of Paired Data

The methods we compare are listed below.

(1) Standard random forest, sampling with replacement.

This is the same as the standard random forest except the same numbers of observations per group (recommendation of Mitchell (2011)) are chosen for the in-bag samples for each tree.

(2) Standard random forest, sampling without replacement

For the case where "*p* >> *n*," it has been shown the OOB-error can be severely pessimistic (i.e., overestimates the error in predicting new samples) when sampling with replacement (Mitchell, 2011), so here, sampling is performed without replacement, and the same number from each group is sampled for each tree.

a. Because of the issue shown in Table 1, leave-one-out cross-validation (LOO-CV) is performed. This is performed as follows: fit the random forest for all observations except observation *j*, and then predict this observation. Repeat for all *j*. We expect this to produce identical results to the OOB error.

b. Leave-one-subject-out cross-validation (LOSO-CV) is performed. This is performed as follows: fit the random forest using all observations except those from subject *k*. Then predict all of the observations for subject *k*. Repeat for all *k*.

(3) Random Forest on subject-adjusted residuals

For each subject, subtract its mean for each variable. Then fit the random forest on the residuals using sampling without replacement and sampling the same number from each group. This method will have limited utility in predicting the status for observations from new subjects, as the same type of adjustment would need to be performed for the unknown observations (i.e., we would have to know that we have one observation per phenotype for a subject, but it is unknown which is which). However, this method could be used for monitoring new observations from the same subject, and this method can still be used for biomarker selection. Additionally, mislabeled samples or outliers could still be detected.

a. LOO-CV is performed

b. LOSO-CV is also performed

(4) Split

This is similar to the method that was discussed in Addler, Brenning, et al. (2011), but it was not applied to random forest because it is not available in the R-implementation, the "randomForest" package (Liaw and Wiener, 2002). Here, we perform the method as follows: split the data in half, so that there is one observation per subject in each half. Fit a random forest on one of these halves, and then predict these values. Repeat this process *k* times and use a majority rule across all splits to get the final classification. The importance is assessed by averaging the importance across all *k* splits. Alternatively, for the same computation time, one could fit one forest on one split and a separate forest on the other split, and then run *k*/2 splits. The final classification could also be determined by averaging the votes for each split.

(5) Sub-sampling by Subject

Here the goal is to fit a random forest by randomly sampling the individual similar to "Paired (both)" from Adler, Brenning, et al. (2011) or the R++ method of Karpievitch et al. (2009). This method was not applied to the random forest in Adler, Brenning, et al. (2011) because it is not available in the standard R-package. Although it is possible to modify the Fortran or C-versions, we prefer the *R*-package ("randomForest") as is, so we program this in *R* by fitting a random forest with one tree on half the individuals (thus all are the in-bag samples), then predicting the remaining individuals (the out-of-bag). This process is repeated *NT* times, where *NT* is the number of trees. The "majority rules" method is applied to the votes in order to achieve the final classification. The permutation importance is also programmed in *R* to produce the same measure produced by the standard package. In other words, to compute the importance for a given variable, *k*, permute its values for the out-of-bag data for each tree. Then for each tree, *j*, compute the difference in accuracy, $e_{kj} = a_{0kj} – a_{kj}*$ where a_{0kj} is the original accuracy and $a_{kj}*$ is the accuracy after the values of variable *k* are permuted. Let m_k represent the mean e_{kj}, across all *NT* trees, and let s_k represent the standard deviation of e_{kj} across all *NT* trees. Then the importance of variable *k* is given by $m_k/(s_k/\text{sqrt}(NT))$, where "sqrt" represents the square root function.

3. Simulation Study

For the simulation study, 400 normal random variables were modeled where there are two groups, "A" and "B." The variables for each group were modeled with the same covariance matrix. The numbers of subjects simulated were the following: 6, 10, and 20 (12, 20, and 40 total observations). These dimensions were chosen to represent metabolomic data, as animal studies typically have fewer than 10 animals per group, and smaller human studies may have groups of size 20.

Let σ_0^2 represent the intra-subject variance, and let σ^2 represent the total variance. Using basic statistical theory (or as shown in Karpievitch et al., 2009), it is can be shown that the intra-subject correlation ISC = σ_0^2/σ^2. Let x_{ijk} represent the values of a given variable for subject *i*, observation *j*, and group *k* (*k* = 1 for group A and *k* = 2 for group B). The means for groups A and B, respectively, were set to {E(x_{ij1}) = 0 and E(x_{ij2}) = 0}, {E(x_{ij1}) = 0.24 and E(x_{ij2}) = 0}, or {E(x_{ij1}) = 0 and E(x_{ij2}) = 0.24} depending on the variable and the run. For each variable Var(x_{ij1}) = Var(x_{ij2}) = σ^2 = 0.3. Independent variables were used, as well as clusters of correlated variables.

For the case of random noise, i.e., E(x_{ij1}) = 0 and E(x_{ij2}) = 0 for all 400 variables, the following sets of ISCs are considered: (a) all 0 (two independent measurements per subject), (b) all 0.5, (c) all 0.9, (d) mixed: 40 variables each with intra-class correlations of 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, i.e., variables 1-40 have intra-class correlation of 0, variables 41-80 have intra-class correlation of 0.1, variables 81-120 have intra-class correlation of 0.2, …, variables 361:400 have intra-class correlation of 0.9 . For (a) – (d) all 400 variables were independent. Since ISC = σ_0^2/σ^2, σ_0^2 = ISC*σ^2. Thus, Cov(x_{ijk}, $x_{ij'k}$) = σ_0^2 = ISC*σ^2, and Cov(x_{ijk}, $x_{i'j'k}$) = 0 for a given variable.

For the case with differential variables, 10 variables were simulated with means with E(x_{ij1}) = 0.24 and E(x_{ij2}) = 0; and 10 variables were simulated to have means with E(x_{ij1}) = 0 and E(x_{ij2}) = 0.24. The remaining 380 variables have no mean differences between the two groups, i.e., E(x_{ij1}) = 0 and E(x_{ij2}) = 0. The same intra-correlations as (a) – (d) were considered. With (d) there was one variable with E(x_{ij1}) = 0.24 and E(x_{ij2}) = 0 and one variable with E(x_{ij1}) = 0 and E(x_{ij2}) = 0.24 for each of the ten sets of 40 variables. For (a) – (d) all variables were independent. Then case (d) was modified for correlated variables. With each set of 40, there were five by five correlated blocks of 0.9, 0.8, 0.7, 0.5, and 0.25, respectively, with the other 15 variables being independent. More specifically, for variables 1-40, all had ISCs of 0. Variables 1-5 had pair-wise correlations of 0.9, variables 6-10 had pair-wise correlations of 0.8, variables 11-15 had pair-wise correlations of 0.7, variables 16-20 had pair-wise correlations of 0.5, variables 21-25 had pair-wise correlations of 0.25, and variables 26-40 had pair-wise correlations of 0. For variables 41-80, each has ISC of 0.1, variables 41-45 had pair-wise correlations of 0.9, variables 46-50 had pair-wise correlations of 0.8, etc. For the correlated variable case, there were 17 variables with E(x_{ij1}) = 0.24 and E(x_{ij2}) = 0: variables 46-50, 115, 221-225, 355, 361-365, and there were 13 variables with E(x_{ij1}) = 0 and E(x_{ij2}) = 0.24: variables 30, 176:180, 192, 279, 291-295.

Methods (1) – (5) were applied to sets of 6, 10, and 20 subjects and for each simulation run, the random forest was then applied to a "test" set with the same data structure (so each subject has one observation from "A" and one observation of "B"). A total of 250 simulation runs were performed for each combination and 1,000 trees used for each method. For the "split" method (4), k = 100 splits were taken for each run. All simulations were performed in ***R*** version 3.0.3 (R Core Team, 2014) with the randomForest package (Liaw and Weiner, 2002).

4. Simulation Study Results

We assess and compare the following for the methods under consideration: (1) the estimate of the error for new data, as given by the OOB error or one of the cross-validation methods, (2) a comparison of the actual error for new data as determined by the average error for the test sets, and (3) a comparison of the importance measures. For the no mean case, the results were similar for all samples sizes, so just the results for n = 6 are shown in Table 2. All methods had predictive accuracies of 50% on the test sets, as desired. However, the OOB errors were severe over-estimates for some methods. For the standard random forest with either with or without replacement, the errors were often over 95% and in some cases were 100% (perfect misclassification). These were errors increased with the increasing ISC. This was expected as discussed previously (example from Table 1). The bias was the same for LOO-CV, which was expected, but LOSO-CV corrects this bias.

The method with the residuals produced OOB errors that were extreme underestimates of the error for predicting new observations. This occurs because subtracting the mean forces results in ISC = -1 for the residuals, regardless of the initial correlation. Let $\mathbf{x_A}$ and $\mathbf{x_B}$ represent the values of a variable for the two groups ordered by subject. Then residuals are $(\mathbf{x_A} – (\mathbf{x_A} + \mathbf{x_B})/2) = (\mathbf{x_A} - \mathbf{x_B})/2$ and $(\mathbf{x_B} – (\mathbf{x_A} + \mathbf{x_B})/2) = -(\mathbf{x_A} - \mathbf{x_B})/2$. The correlation of $(\mathbf{x_A} - \mathbf{x_B})/2$ and $-(\mathbf{x_A} - \mathbf{x_B})/2$ is equal to -1. Hence, trees of noise variables could appear to separate depending on the samples chosen. This bias was corrected by applying LOSO-CV, but not LOO-CV.

In terms of the variable importance we can see the effect of various ISCs by comparing the average importance for the “mixed” case (40 variables each with ISCs of 0, 0.1, …, 0.9). Figure 1 shows these for the no mean case for n = 20 subjects. We can see that the residual and split method have similar importance for all variables regardless of ISC, while the standard random forest has higher importance for lower ISCs, while the subject-sampled random forest has higher importance for higher ISCs. This follows for the reasons illustrated in the example illustrated in Table 1.

For the cases with differential variables, the same biases were seen for the OOB-error for the standard methods and the residual method, and as with the no mean case, LOSO-CV performs well in correcting this. The results are shown in Figure 2. There is no improvement in accuracy when sampling with replacement compared to sampling without replacement. The split method and the sampling by subject methods performed similarly to the standard random forest with the sampling by subject performing slightly better for the case where all ISCs were 0.9. Even though the OOB error was severely optimistic for the residual method, its actual accuracy on the test sets was much better than the other methods. This bias in the OOB error is a result of the induced correlation discussed previously. Furthermore, to apply to the test set, the subjects had to be known in order to compute the residuals, while for the other methods which observations belonging to which subjects did not need to be specified. This effect can also be seen in the average importance for each variable (see Figure 3), as here the variables with the highest ISCs were more important, as this large source of variation has been removed.

5. Data application

This data set was originally used for internal validation experiments where EDTA-plasma samples were collected from 42 in-house volunteers. Samples were taken under both fed and fasted conditions. There were three subjects with only one measurement, and these were removed from the analysis. Mass spectrometry (LC/ms) was used to measure the metabolites – more detail on this mass spectrometry platform is described in Evans et al. (2014). The samples were all run in one batch on the instruments, and thus no normalization was performed. Any metabolite not present in at least 50% of the samples was removed from analysis. Of the remaining metabolites, any missing values were assumed to be below the limit of detection and were imputed with the observed minimum (on a per metabolite basis). Although the study was not run in order to test fed vs. fasted, we are performing this analysis to illustrate the aforementioned methods.

The following methods were applied: (1) standard random forest, sampling without replacement, LOSO-CV to estimate the error, (2) split method, (3) sampling by subject, (4) residual with LOSO-CV. The results are shown in Table 3. The standard, split, and subject-sampling methods produced the same error rates. Although the residual method performed the best, it is not applicable to a new a data set unless we know that each subject had two measurements, one fed and one fasted. The variables ranked as important also differ for the methods.

The variable importance values for the various methods are shown in Figure 4. Each method is compared to the standard method, so values below the line $y = x$ indicate importance values lower for the same variable, and those above the line indicate values that are higher. From Figure 4, it can be seen that the method where the subjects are sub-sampled produces very similar results to the standard random forest. The split method produces consistently lower values, but the ranking is similar. The residual method differed the most from the others. From the simulation study, we expect the biggest differences for those with the highest ISC, with the residual method having higher values for strong variables with high ISC. One example of this is paraxanthine, which is a caffeine metabolite. Its ISC = 0.86. The importance measures for the standard, residual, split, and subject-sampling methods are 3.8, 10.9, 0.9, and 4.6, respectively. We see that the importance for the residual method is much higher than the standard for this variable, while the split method is much lower, as expected

6. Discussion

Through an extensive simulation study and application to real data, we see that OOB error for the standard random forest is an overestimate of its predictive error in future data, i.e., the error rate is too pessimistic. However, performing LOSO-CV eliminates this bias. The standard random forest does not take into account the paired structure of the data. Thus, one solution would be to create subject-adjusted data, such as computing the simple residuals. This method produces OOB errors that are underestimates of its true predictive errors; however, this bias is eliminated by performing LOSO-CV. This method has stronger predictive ability over all the other methods, but suffers from a lack of applicability to new data sets in many situations. However, it can still be useful for biomarker selection and determining outliers. Another solution is to sample one observation from each subject, fit the random forest, repeat, and then aggregate the results. Here the data are now independent, but only half the data is being used for each split. However, its predictive ability is comparable to the standard random forest. Furthermore for the split method, the importance is not affected by the ISC, while the standard forest gives higher rankings to those with low ISCs and the residual method gives higher rankings to those with higher ISCs. Another natural solution to this issue is to randomly sample the subjects. This option is not currently available in the standard *R*-package, but we have developed code in *R* for this. Its error estimate needs no further adjusting. Its true predictive ability is comparable to the standard random forest. Because the code was written in *R*, rather than modifying the C or Fortran, this may take substantively longer to run, depending on the size of the data set and the number of trees. Thus, it may be preferable to simply run the standard random forest and estimate the error with LOSO-CV or run the split method.

Acknowledgements

We would like to acknowledgment Adam Kennedy from Metabolon who designed the original study used in the data application and set up the sample collection protocol.

References

Adler, W., Brenning, A., Potapov, S., Schmid, M., Lausen, B., 2011. Ensemble classification of paired data. Computational Statistics and Data Analysis 55, 1933-1941.

Adler, W., Potapov, S., Lausen, B., 2011. Classification of repeated measurements data using tree-based ensemble methods. Comput Stat 26:355-369; doi: 10.1007/s00180-011-0249-1.

Amaratunga, D., Cabrera, J., Lee, Y-S., 2008. Enriched random forests. Bioinformatics 24 (18), 2010-2014.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5-32.

Brenning, A., Lausen, B., 2008. Estimating error rates in the classification of paired organs. Statistics in Medicine 27, 4515-4531.

Chen, X., Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics 99(6), 323-329.

Evans, A., Bridgewater, B., Liu, Q., Mitchell, M., Robinson, R., Dai, H., Stewart, S., DeHaven, C., Miller, L., (2014), High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. Metabolomics, 4:132. doi: 10.4172/2153-0769.1000132.

Gall W., Beebe, K., Lawton K., Adam K-P., Mitchell, M., Nakhle, P., Ryals, J., Milburn, M., Nannipieri, M., Camastra, S., Natali, A., Ferrannini, E., 2010. Alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. PLoS One, 5:e10883; doi: 10.1371/journal.pone.0010883.

Ippolito D., Lewis, J., Yu, C., Leon, L., Stallings, J., 2014. Alteration in circulating metabolites during and after heat stress in the conscious rat: potential biomarkers of exposure and organ specific injury. BMC Physiology. 14:14; doi: 10.1186/s12899-014-0014-0.

Karpievitch, Y., Hill, E., Leclerc, A., Dabney A., Almeida, J., 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PLoS ONE 4(9), e7087.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2(3), 18-22.

Mitchell, M., 2011. Bias of the random forest out-of-bag (OOB) error for certain input parameters. Open Journal of Statistics 1(3); doi: 10.4236/ojs.2011.13024.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Schuler, K., Rambally, B., DiFurio, M., Sampey, B., Gehrig, P., Makowski, L., Bae-Jump, V., 2015. Antiproliferative and metabolic effects of metformin in a preoperative window clinical trial for endometrial cancer. Cancer Medicine. 4(2), 161-173; doi: 10.1002/cam4.353

Tables and Figures

Table 1: Example of a classification tree with high ISC – all the in-bag samples are correctly classified, but all of the out-of-bag samples are incorrectly classified.

SUBJECT ID	x	G	STATUS	PREDICTED
1	2.0	A	IN	
1	2.1	B	OUT	A
2	0.4	A	OUT	B
2	0.3	B	IN	
3	3.0	A	IN	
3	3.1	B	OUT	A
4	1.0	A	OUT	B
4	1.2	B	IN	

Table 2: Error Rates for the No Mean Case, Simulation Study, n=6

ISC	METHOD	OOB Error	LOO-CV	LOSO-CV	Test Set Error
all 0	subject-sampling	0.49			0.50
all 0	resid	0.00	0.00	0.51	0.52
all 0	split	0.51			0.50
all 0	standard	0.53	0.51	0.47	0.51
all 0	standard_wReplace	0.62			0.50
all 0.5	subject-sampling	0.51			0.52
all 0.5	resid	0.00	0.00	0.48	0.51
all 0.5	split	0.50			0.50
all 0.5	standard	0.97	0.97	0.52	0.52
all 0.5	standard_wReplace	0.98			0.50
all 0.9	subject-sampling	0.50			0.50
all 0.9	resid	0.00	0.00	0.52	0.51
all 0.9	split	0.50			0.50
all 0.9	standard	1.00	1.00	0.50	0.50
all 0.9	standard_wReplace	1.00			0.51
mix	subject-sampling	0.48			0.49
mix	resid	0.00	0.00	0.48	0.51
mix	split	0.51			0.50
mix	standard	0.94	0.95	0.51	0.51
mix	standard_wReplace	0.97			0.50

Table 3: Error Rates for Random Forest Methods on the Metabolomics Data Set

Method	Error
Standard (OOB)	0.155
Standard (LOSO-CV)	0.095
Split	0.095
Subject-Sampling	0.095
Residual (LOSO-CV)	0

Figure 1: Comparison of Average Importance, No Mean Case, n=20, mixed ISC

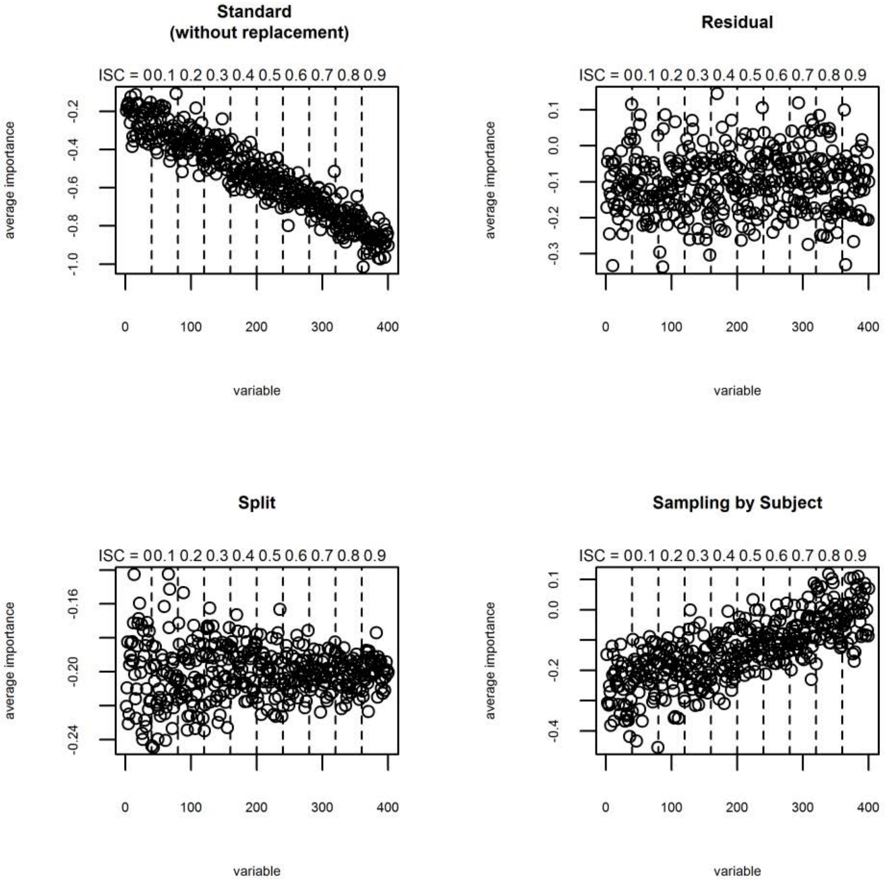


Figure 2: Comparison of error rates on the test sets

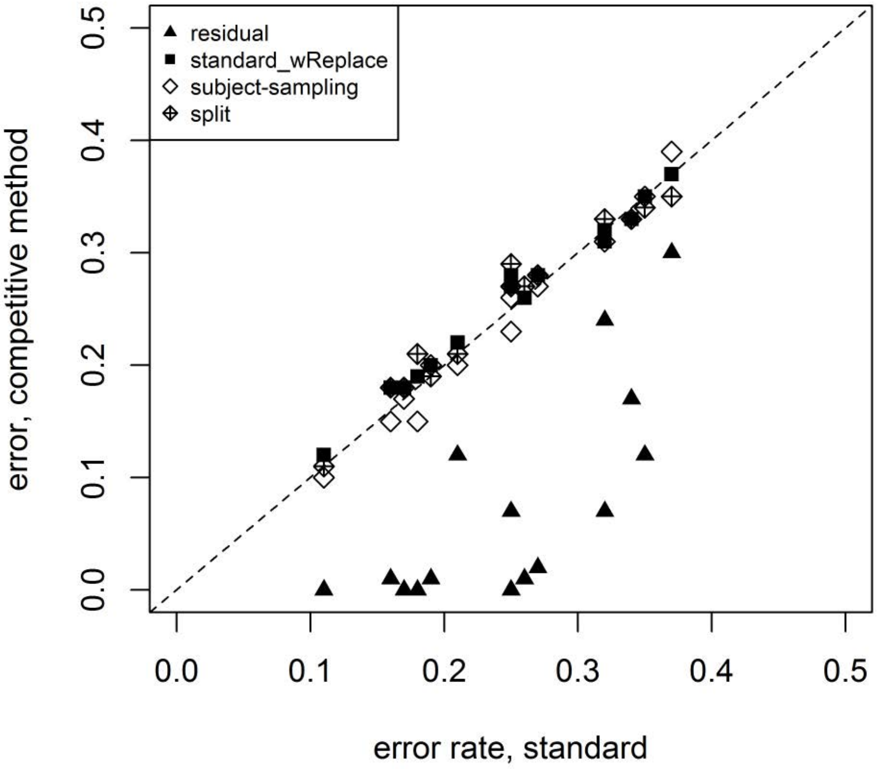


Figure 3: Comparison of Average Importance, n=20; 40 differential independent variables

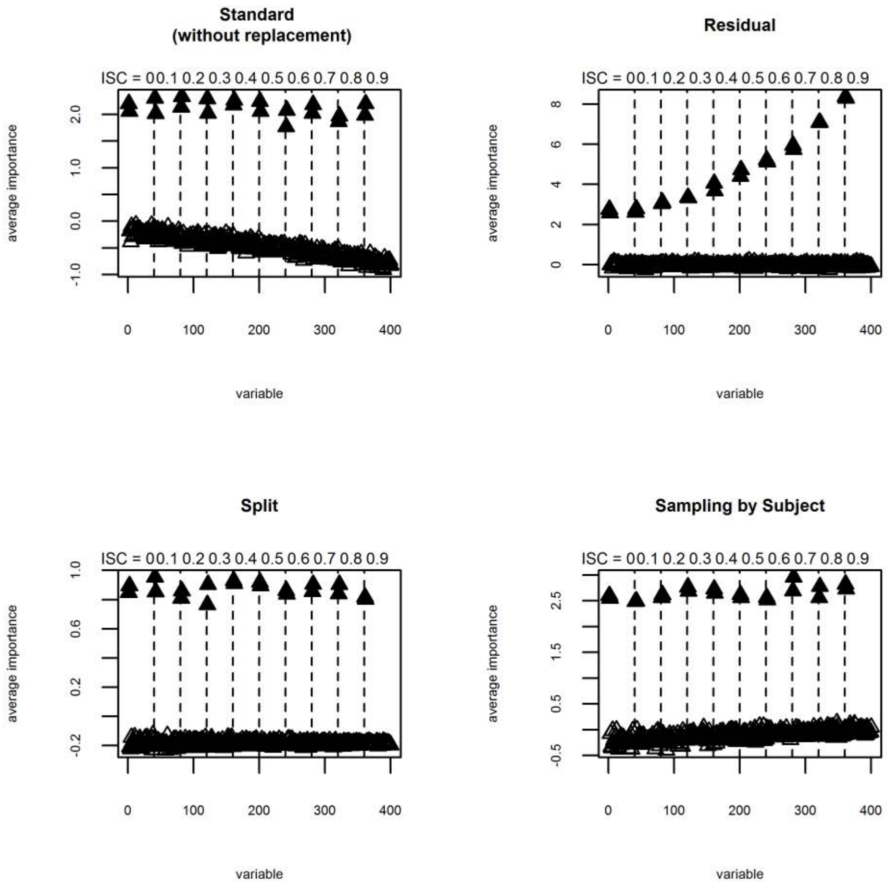
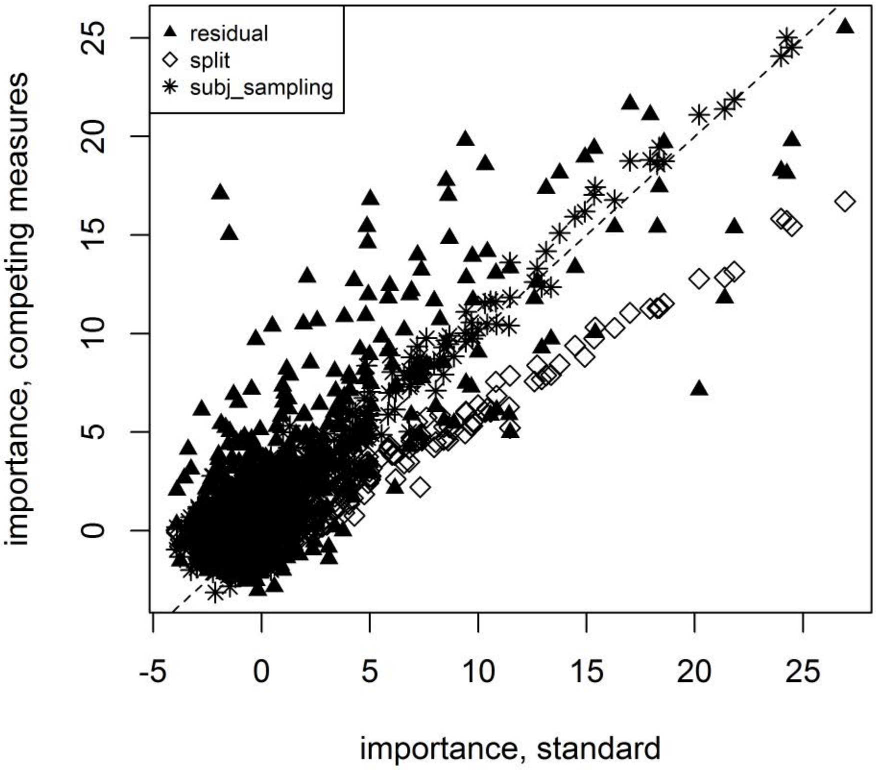


Figure 4: Comparison of Variable Importance, Human Metabolomics Data


▲ residual
◇ split
✱ subj_sampling
importance, competing measures
importance, standard