Conditional variable importance for random forests

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis

BMC Bioinformatics 2008, 9:307
https://doi.org/10.1186/1471-2105-9-307
© Strobl et al; licensee BioMed Central Ltd. 2008

Received: 01 April 2008
Accepted: 11 July 2008
Published: 11 July 2008

Abstract

Background

Random forests are becoming increasingly popular in many scientific fields because they can cope with "small n large p" problems, complex interactions and even highly correlated predictor variables. Their variable importance measures have recently been suggested as screening tools for, e.g., gene expression studies. However, these variable importance measures show a bias towards correlated predictor variables.

Results

We identify two mechanisms responsible for this finding: (i) a preference for the selection of correlated predictors in the tree building process and (ii) an additional advantage for correlated predictor variables induced by the unconditional permutation scheme that is employed in the computation of the variable importance measure. Based on these considerations we develop a new, conditional permutation scheme for the computation of the variable importance measure.

Conclusion

The resulting conditional variable importance reflects the true impact of each predictor variable more reliably than the original marginal approach.

Keywords

Random Forest
Importance Measure
Importance Score
Amino Acid Property
Permutation Scheme

1 Background

Within the past few years, random forests have become a popular and widely used tool for non-parametric regression in many scientific areas. They show high predictive accuracy and are applicable even in high-dimensional problems with highly correlated variables, a situation which often occurs in bioinformatics. Recently, the variable importance measures yielded by random forests have also been suggested for the selection of relevant predictor variables in the analysis of microarray data, DNA sequencing and other applications [2–5].

Identifying relevant predictor variables, rather than only predicting the response by means of some "black-box" model, is of interest in many applications. By means of variable importance measures the candidate predictor variables can be compared with respect to their impact in predicting the response or even their causal effect (see, e.g., [6] for assumptions necessary for interpreting the importance of a variable as a causal effect). In this case a key advantage of random forest variable importance measures, as compared to univariate screening methods, is that they cover the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables. For example, Lunetta et al. [2] find that genetic markers relevant in interactions with other markers or environmental variables can be detected more efficiently by means of random forests than by means of univariate screening methods like Fisher's exact test. In the analysis of amino acid sequence data Segal et al. [7] also point out the necessity to consider interactions between sequence positions. Tree-based methods like random forests can help identify relevant predictor variables even in such high dimensional settings involving complex interactions. Therefore, the impact of different amino acid properties, some of which have been
shown to be relevant in DNA and protein evolution [18], for predicting peptide binding is investigated in our application example in
Section 4. However, we will find in this application example, as often in practical problems, that many predictor variables are highly
correlated.

The issue of correlated predictor variables is prominent in, but not limited to, applications in genomics and other high dimensional
problems. Therefore, it is important to note that in any non-experimental scientific study, where the population of predictor settings cannot be
manipulated independently by the investigator, the distinction between the marginal and the conditional effect of a variable is crucial.

Consider, for example, the apparent correlation between rates of complication after surgery and mortality in hospitals, that was
investigated by Silber and Rosenbaum [19]. It is plausible to believe that the mortality rate of a hospital depends on the rate of
complications – or even that the mortalities are caused by the complications. However, when severity of illness is taken into account, the
correlation dissipates [20].

This phenomenon is known as a spurious correlation (see also Stigler [21] for a historical example). In the hospital mortality example, the
spurious correlation is caused by the fact that hospitals that treat many serious cases have both higher complication and mortality rates.
However, when conditioning on severity of illness (i.e. comparing only patients with similar severity of illness), mortality is no longer
associated with complications.

If you consider this as a prediction problem, once the truly influential background variable (severity of illness) is known, it is clear that the
remaining covariate (complication rate) provides no or little additional information for predicting the response (mortality rate). From a
statistical point of view, however, this distinction can only be made by a conditional importance measure.

We will point out throughout this chapter that correlations between predictor variables – regardless of whether they arise from small-
scale characteristics, such as proximities between genetic loci in organisms, or large-scale characteristics, such as similarities in the
clinets of hospitals – severely affect the original random forest variable importance measures, because they can be considered as
measures of marginal importance, even though what is of interest in most applications is the conditional effect of each variable. To
make this distinction more clear, we shortly review previous suggestions from the literature for introducing or illustrating variable
importance in classification and regression trees (termed “classification trees” in the following for brevity, while all results apply to both
classification and regression trees) and random forests. Breiman [22] displays the change in the response variable over the range of one
predictor variable in “partial dependence plots” (see also [10] for a related approach). This method is helpful for revealing important
model coefficients in linear models. However, whether the effect of a variable is interpretable as conditional on all other variables, as in
linear models, may not be guaranteed in other models – and we will point out explicitly below that this is not the case in classification
trees or random forests.

The permutation accuracy importance, that is described in more detail in Section 2.3, follows the rationale that a random permutation of
the values of the predictor variable is supposed to hinder the identification of the predictor variable, thus reducing the prediction
accuracy before and after permuting the predictor variable, i.e. with and without the help of this predictor variable, is used as an
importance measure. The actual permutation accuracy importance measure will be termed “permutation importance” in the following,
while the general concept of the impact of a predictor variable in predicting the response is termed “variable importance”.

The alternative variable importance measure used in random forests, the Gini importance, is based on the principle of impurity reduction that is
followed in most traditional classification tree algorithms. However, it has been shown to be biased when predictor variables vary in
scale characteristics, such as proximities between genetic loci in organisms, or large-scale characteristics, such as similarities in the
clinets of hospitals – severely affect the original random forest variable importance measures, because they can be considered as
measures of marginal importance, even though what is of interest in most applications is the conditional effect of each variable. To
make this distinction more clear, we shortly review previous suggestions from the literature for introducing or illustrating variable
importance in classification and regression trees (termed “classification trees” in the following for brevity, while all results apply to both
classification and regression trees) and random forests. Breiman [22] displays the change in the response variable over the range of one
predictor variable in “partial dependence plots” (see also [10] for a related approach). This method is helpful for revealing important
model coefficients in linear models. However, whether the effect of a variable is interpretable as conditional on all other variables, as in
linear models, may not be guaranteed in other models – and we will point out explicitly below that this is not the case in classification
trees or random forests.

Based on the permutation importance, schemes for variable selection and for providing statements of the "significance of" a predictor
variable (instead of a merely descriptive ranking of the variables w.r.t. their importance scores) have been derived: Breiman and Cutler
[15] suggest a simple significance test that, however, shows poor statistical properties [14]. An approach that, however, shows poor statistical
properties in large scale screening studies is introduced by Diaz-Uriarte and Alvarez de Andres [17], who suggest a backward elimination strategy. This
approach has been shown to provide a reasonable selection of genes in many situations and is freely available in an R package [18],
that also provides different plots for comparing the performance on the original data set to those on a data set with randomly permuted
values. However, the overall null hypothesis that none of the predictor variables contributes anything to the model is not rejected and may
serve as a baseline for significance statements. A similar approach is followed by Rodenburg et al. [19]. However, some recent simulation
studies indicate that the performance of the variable importance measures may not be reliable when predictor variables are correlated:
Even though Archer and Kimes [20] show in their extensive simulation study that the Gini importance can identify influential
predictor variables out of sets of correlated covariates in many settings, the preliminary results of the simulation study of Nicoodemus
and Shugart [21] indicate that the ability of the permutation importance to detect influential predictor variables in sets of correlated
covariates is less reliable than that of alternative machine learning methods and highly depends on the number of previously selected
splitting variables mtry. These studies, as well as our simulation results, indicate that random forests show a preference for correlated
predictor variables, that is also carried forward to any significance test or variable selection scheme constructed from the importance
measures.

In this work we are aiming at providing a deeper understanding of the underlying mechanisms responsible for the observations of [11] and [12].
In addition to this, we want to broaden the scope of consideration and focus on correlated and uncorrelated predictor variables. For this type of
problem we introduce a new, conditional permutation importance for random forests, that better reflects the true influence of predictor variables. Our approach is motivated by the visual means of illustration introduced by Nason et al. [22]. In their "CARTForest" plots they not only display the marginal influence of a predictor variable, like the partial
dependence plots of Breiman [11], but also the influence of random forests and the related method bagging, an ensemble of classification trees is created by means of drawing several bootstrap
samples or subsamples from the original training data and fitting a single classification tree to each sample. Due to the random
variation in the samples and the instability of the single classification trees, the ensemble will consist of a diverse set of trees. For
decision making, a vote (or average) over the predictions of the single trees is used and has been shown to highly outperform the single
trees or random forests.

By combining the prediction of a diverse set of trees, bagging utilizes the fact that classification trees are instable but on average
produce the right prediction. This understanding has been supported by several empirical studies (see, e.g., [23–25]) and especially the
theoretical results of Bühmann and Yu [26], who could show that this is achieved by means of smoothing the hard cut decision boundaries created by splitting in single classification trees, which in return reduces the
variance of the prediction.

In random forests, another source of diversity is introduced when the set of predictor variables to select from is randomly restricted
in each split, producing even more diverse trees. In addition to the smoothing of hard decision boundaries, the random selection of
splitting variables in random forests allows predictor variables that are not correlated.

Even though these variables may not be optimal with respect to the current split, their selection may reveal interaction effects with other
variables that otherwise would have been missed and thus work towards the global optimality of the ensemble.

The classification trees, from which the random forests are built, are built recursively in that the next splitting variable is selected by
means of locally optimizing a criterion (such as the Gini gain in the traditional CART algorithm [27] within the current node. This
current node is determined by a configuration of predictor values, that is determined by all previously splits in the same branch of the tree (see, e.g.,
[28] for illustrations). In this respect the evaluation of the next splitting variable can be considered conditional on the previously selected
predictor variables, but regardless of any other predictor variable. In particular, the selection of the first splitting variable involves only
the marginal, univariate association between that predictor variable and the response, regardless of all other predictor variables.

In random forests and the related method bagging, an ensemble of classification trees is created by means of drawing several bootstrap
samples or subsamples from the original training data and fitting a single classification tree to each sample. Due to the random
variation in the samples and the instability of the single classification trees, the ensemble will consist of a diverse set of trees. For
decision making, a vote (or average) over the predictions of the single trees is used and has been shown to highly outperform the single
trees or random forests.

By combining the prediction of a diverse set of trees, bagging utilizes the fact that classification trees are instable but on average
produce the right prediction. This understanding has been supported by several empirical studies (see, e.g., [23–25]) and especially the
theoretical results of Bühmann and Yu [26], who could show that this is achieved by means of smoothing the hard cut decision boundaries created by splitting in single classification trees, which in return reduces the
variance of the prediction.

In random forests, another source of diversity is introduced when the set of predictor variables to select from is randomly restricted
in each split, producing even more diverse trees. In addition to the smoothing of hard decision boundaries, the random selection of
splitting variables in random forests allows predictor variables that are not correlated.

Even though these variables may not be optimal with respect to the current split, their selection may reveal interaction effects with other
variables that otherwise would have been missed and thus work towards the global optimality of the ensemble.

The classification trees, from which the random forests are built, are built recursively in that the next splitting variable is selected by
means of locally optimizing a criterion (such as the Gini gain in the traditional CART algorithm [27] within the current node. This
current node is determined by a configuration of predictor values, that is determined by all previously splits in the same branch of the tree (see, e.g.,
[28] for illustrations). In this respect the evaluation of the next splitting variable can be considered conditional on the previously selected
predictor variables, but regardless of any other predictor variable. In particular, the selection of the first splitting variable involves only
the marginal, univariate association between that predictor variable and the response, regardless of all other predictor variables.
importance will be (a random value from a distribution with mean) zero, while any deviation from the null hypothesis will lead to a

With this framework in mind, we can now take a second look at the random forest permutation importance and ask: Under which null

hypothesis.

into account the permutation scheme that is employed in the computation of the permutation importance. In the following we will first

achieved by dividing the raw importance by its standard error, is provided. However, since recent results [15, 17], the correlated predictors loose most of their advantage because variable selection is now conditional on the previously

chosen variables in the same branch of the tree, that may include the truly influential correlated predictors. However, since variable selection is not conditional on all (or at least all correlated) variables, there is still a preference for the correlated variables with low and zero coefficients $X_j$ and $X_k$ over $X_l$ and $X_m$, with a similar dependency on mtry.

2.3 The permutation importance

The rationale of the original random forest permutation importance is the following: By randomly permuting the predictor variables $X_j$, is original association with the response $Y$ is broken. When the permuted variable $X_j$, together with the remaining non-permuted predictor variables, is used to predict the response for the out-of-bag observations, the prediction accuracy (i.e. the number of observations

where $y_i^*(t=1) = f(t|x_i)$

2.4 Types of dependence

We know that the original permutation importance overestimates the importance of correlated predictor variables. Part of this artefact may be due to the preference of correlated predictor variables in early splits as illustrated in Section 2.2. However, we also have to take into account the permuted scheme that is employed in the computation of the permutation importance. In the following we will first outline what notion of independence corresponds to the current permutation scheme of the random forest permutation importance.

Then we will introduce a more sensible permutation scheme that better reflects the true impact of predictor variables.

It can help our understanding to consider the permutation scheme in the context of permutation tests [14]: Usually a null hypothesis is considered that implies the independence of particular (sets of) variables. Under this null hypothesis some permutations of the data are permitted because they preserve the structure determined by the null hypothesis. If, for example, the response variable $Y$ is independent of all predictor variables (global null hypothesis) a permutation of the observed values of $Y$ affects neither the marginal distribution of $Y$ nor the joint distribution of $X_1, X_2, Y$ and $X_3$. Both the joint distribution can be factorized as $P(Y, X_1, X_2, X_3) = P(Y) P(X_1) P(X_2) P(X_3)$ under the null hypothesis. If, however, the null hypothesis is not true, the joint permutation will lead to a deviation in the joint distribution or some test statistic computed from it. Therefore, a change in the distribution or test statistic caused by the permutation can serve as an indicator that the data do not follow the independence structure we would expect under the null hypothesis.

With this framework in mind, we can now take a second look at the random forest permutation importance and ask: Under which null hypothesis would this permutation scheme be permitted? If the data are actually generated under this null hypothesis the permutation importance will be (a random value from a distribution with mean) zero, while any deviation from the null hypothesis will lead to a
In summary, the conditional permutation importance is derived as follows:

\[ I_{\text{conditional}}(X_j | \mathcal{X}_{-j} \cup Z) = \frac{1}{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z)} \left( \frac{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z) \hat{P}(Y \mid \mathcal{X}_{-j} \cup Z_{\pi})}{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z_{\pi})} - 1 \right) \]

where \( I_{\text{conditional}}(X_j | \mathcal{X}_{-j} \cup Z) \) is the conditional permutation importance of \( X_j \), \( \hat{P}(Y \mid \mathcal{X}_{-j} \cup Z) \) is the OOB prediction accuracy before the permutation, and \( \hat{P}(Y \mid \mathcal{X}_{-j} \cup Z_{\pi}) \) is the OOB prediction accuracy after the permutation.

This permutation scheme corresponds to the following null hypothesis:

\[ H_0: (X_j \perp Y | Z) \]

where the conditional distribution can be factorized as:

\[ P(Y, X_j | Z) = P(Y | Z) P(X_j | Z) \quad \text{or} \quad P(Y, X_j | Z) = P(Y | X_j, Z) \]

which is the definition of conditional independence.

In the special case where \( X_j \) and \( Z \) are independent, both permutation schemes will give the same result, as illustrated by our simulation results below. When \( X_j \) and \( Z \) are correlated, however, the original permutation scheme will lead to an apparent increase in the importance of correlated predictor variables, that is due to deviations from the uninteresting null hypothesis of independence between \( X_j \) and \( Z \).

2.5 A new, conditional permutation scheme

Technically, any kind of conditional assessment of the importance of one variable conditional on another is straightforward whenever the variables to be conditioned on, \( Z \), are categorical as in Eq. However, for our aim to conditionally permute the values of \( Y \) within groups of \( Z = z \), where \( Z \) can contain potentially large sets of covariates of different scales of measurement, we want to supply a grid that (i) is applicable to variables of different types, (ii) is parsimonious as possible, but (iii) is also computationally feasible. Our suggestion is to define the grid within which the values of \( X_j \) are permuted for each tree by means of the partition of the feature space induced by that tree. The main advantages of this approach are that this partition was already learned from the data during model fitting, contains splits in categorical, ordered and continuous predictor variables and can thus serve as an internally available means for discretizing the feature space.

In principle, any partition derived from a classification tree can be used to define the permutation grid. Here we used partitions produced by unbiased conditional inference trees [33], that employ binary splitting as in the standard CART algorithm [33]. This means that, if \( \mathcal{X} \) is the number of categories of an unordered or ordered categorical variable, up to \( \mathcal{X} \), but potentially less than \( \mathcal{X} \), subsets of the data are separated.

Continuous variables are treated in the same way: Every binary split in a variable provides one or more cutpoints, that can induce a more or less fine graded grid on this variable. By using the grid resulting from the current tree we are able to condition in a straightforward way not only on categorical, but also on continuous variables and create a grid that may be more parsimonious than the more or less fine graded grid on this variable. By using the grid resulting from the current tree we are able to condition in a

Within this grid permute the values of \( X_j \) and compute the ob-prediction accuracy after permutation:

\[ I_{\text{conditional}}(X_j | \mathcal{X}_{-j} \cup Z) = \frac{1}{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z)} \left( \frac{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z) \hat{P}(Y \mid \mathcal{X}_{-j} \cup Z_{\pi})}{\hat{P}(Y \mid \mathcal{X}_{-j} \cup Z_{\pi})} - 1 \right) \]

For all variables \( Z \) to be conditioned on: Extract the cutpoints that split this variable in the current tree and create a grid by means of bisecting the sample space in each cutpoint.

The difference between the prediction accuracy before and after the permutation accuracy again gives the importance of \( X_j \) for one tree (see Equation 1). The importance of \( X_j \) for the forest is again computed as an average over all trees.

To determine the variables \( Z \) to be conditioned on, the most conservative – or rather overcautious – strategy would be to include all other variables as conditioning variables, as was indicated by our initial notation. A more intuitive choice is to include only those variables whose empirical correlation with the variable of interest \( X_j \) exceeds a certain moderate threshold, as we do with the Pearson correlation coefficient for continuous variables in the following simulation study and application example. For the more general case of predictor variables of different scales of measurement the framework promoted by Hothorn et al. [31] provides p-values of conditional independence tests as measures of association. The p-values have the advantage that they are comparable for variables of all types and can serve as an intuitive and objective means for selecting the variables \( Z \) to be conditioned on in any problem. Another option is to let the user himself select certain variables to condition on, if, e.g., a hypothesis of interest includes certain independences.
The conditional permutation importance will be freely available in the next release of the party package for recursive partitioning results produced with different random seeds do not vary systematically. Only then it is assured that the differences between, e.g., can help identify the truly relevant predictor variables. Our simulation results also illustrate the impact of the choice of the random forest entirely eliminate the preference for correlated predictor variables, it has been shown to provide a more fair means of comparison that the true impact of each predictor variable better than the original, marginal approach. Even though the conditional permutation cannot conditional permutation scheme uses the partition that is automatically provided by the fitted model as a conditioning grid and reflects variables and suggested a new, conditional permutation scheme for the computation of the variable importance measure. This new, conditional permutation scheme is that the variability of the conditional importance is lower than that of the unconditional importance within each level of mtry.

With respect to the identifiability of few influential predictors from a set of correlated and other noise variables (which was the task in [32] and [35]), we see from the importance scores for \( X_2 \) that this predictor has a similar importance regardless of whether we use the original or the conditional permutation scheme. This shows that importance rankings in data sets that contain complex correlations between predictor variables can be severely more pronounced for large mtry as expected, but similar effects can be observed for smaller values. For piecewise-constant functions, that can be more easily much better suited as a means of comparison than the original importance. For piecewise constant functions, that can be more easily conditional importance reflects the variability of the conditional importance is lower than that of the unconditional importance within each level of mtry. With respect to the identifiability of few influential predictors from a set of correlated and other noise variables (which was the task in [32] and [35]), we can see from the importance scores for \( X_2 \) that this predictor has a similar importance regardless of whether we use the original or the conditional permutation scheme. This shows that importance rankings in data sets that contain complex correlations between predictor variables can be severely more pronounced for large mtry as expected, but similar effects can be observed for smaller values. For piecewise-constant functions, that can be more easily piecewise-constant functions, that can be more easily compared

### 3 Results

For the simulation design introduced in Section 2.1, Figure 3 shows the median and interquartile range (over 500 iterations) of the importance scores of each variable for the different permutation schemes: the original marginal permutation and the newly suggested conditional permutation scheme. The set of variables \( Z \) to be conditioned on was chosen here to include all variables with an empirical correlation \( r \geq 0.2 \). The simulation design introduced in Section 2.1, Figure 3 shows the median and interquartile range (over 500 iterations) of the importance scores of each variable for the different permutation schemes: the original marginal permutation and the newly suggested conditional permutation scheme. The set of variables \( Z \) to be conditioned on was chosen here to include all variables with an empirical correlation \( r \geq 0.2 \).

#### 4 Example: Relating amino acid sequence to phenotype in peptide-binding data

As an example we consider peptide-binding data that were previously analysed with recursive partitioning techniques by Segal et al. [7]. The data set includes 105 variables for a total of \( n = 310 \) amino acid sequences. The response to be predicted is a binding property that can be coded as a binary variable (binding/no binding). The remaining variables in this data set contains 13 amino acids encoded for each of the eight positions. These properties include, e.g., volume, polarity, bulkiness, flexibility, aromaticity, and charge, yielding in total 104 continuous predictor variables. A random forest with 1000 trees and \( m \text{try} = 104 \) (which corresponds to bagging [22, 24] as a special case of a random forest where \( m \text{try} \) is equal to the number of candidate predictors and variable selection is not randomly restricted) was fit to the data set. The permutation importance was computed both with the unconditional or the conditional permutation scheme. The resulting importance scores are displayed in Figure 4 (note that the absolute values of the scores should not be interpreted). The few predictor variables whose importance scores reach highest or even exceed the plotting area would be selected for further analysis by any means. However, for some of the variables with the next smaller importance scores the ranking strongly depends on the permutation scheme. We will focus our illustration on the ranking of three exemplary predictor variables, "h2y8", "flex8" and "pol3", that are highlighted in Figure 4. We find in the unconditional view in the top panel of Figure 4 that the variable "h2y8" appears to be of higher importance than "pol3". Interestingly, if we replace the conditional importance with the original importance the ranking strongly depends on the permutation scheme. We will focus our illustration on the ranking of three exemplary predictor variables, "h2y8", "flex8" and "pol3", that are highlighted in Figure 4. We find in the unconditional view in the top panel of Figure 4 that the variable "h2y8" appears to be of higher importance than "pol3" (rank of "h2y8": 8, "flex8": 9, "pol3": 11). However, in the conditional view in the bottom panel of Figure 4 their order is reversed and it turns out that "pol3" is really more important than "h2y8" and "flex8" (ranks "h2y8": 9, "flex8": 8, "pol3": 7). This change in the ranks of the predictor variables is most pronounced for large mtry as expected, but similar effects can be observed for smaller values.

#### 4 Example: Relating amino acid sequence to phenotype in peptide-binding data

As an example we consider peptide-binding data that were previously analysed with recursive partitioning techniques by Segal et al. [7]. The data set includes 105 variables for a total of \( n = 310 \) amino acid sequences. The response to be predicted is a binding property that can be coded as a binary variable (binding/no binding). The remaining variables in this data set contains 13 amino acids encoded for each of the eight positions. These properties include, e.g., volume, polarity, bulkiness, flexibility, aromaticity, and charge, yielding in total 104 continuous predictor variables. A random forest with 1000 trees and \( m \text{try} = 104 \) (which corresponds to bagging [22, 24] as a special case of a random forest where \( m \text{try} \) is equal to the number of candidate predictors and variable selection is not randomly restricted) was fit to the data set. The permutation importance was computed both with the unconditional or the conditional permutation scheme. The resulting importance scores are displayed in Figure 4 (note that the absolute values of the scores should not be interpreted). The few predictor variables whose importance scores reach highest or even exceed the plotting area would be selected for further analysis by any means. However, for some of the variables with the next smaller importance scores the ranking strongly depends on the permutation scheme. We will focus our illustration on the ranking of three exemplary predictor variables, "h2y8", "flex8" and "pol3", that are highlighted in Figure 4. We find in the unconditional view in the top panel of Figure 4 that the variable "h2y8" appears to be of higher importance than "pol3". Interestingly, if we replace the conditional importance with the original importance the ranking strongly depends on the permutation scheme. We will focus our illustration on the ranking of three exemplary predictor variables, "h2y8", "flex8" and "pol3", that are highlighted in Figure 4. We find in the unconditional view in the top panel of Figure 4 that the variable "h2y8" appears to be of higher importance than "pol3" (rank of "h2y8": 8, "flex8": 9, "pol3": 11). However, in the conditional view in the bottom panel of Figure 4 their order is reversed and it turns out that "pol3" is really more important than "h2y8" and "flex8" (ranks "h2y8": 9, "flex8": 8, "pol3": 7). This change in the ranks of the predictor variables is most pronounced for large mtry as expected, but similar effects can be observed for smaller values.

#### 5 Discussion and conclusion

We have investigated the sources of preferences in the variable importance measures of random forests in favor of correlated predictor variables and suggested a new, conditional permutation scheme for the computation of the variable importance measure. This new, conditional permutation scheme uses the partition that is automatically provided by the fitted model as a conditioning grid and reflects the true impact of each predictor variable better than the original, marginal approach. Even though the conditional permutation cannot entirely eliminate a preference for correlated predictor variables, it has been shown to provide a more fair means of comparison that can help identify the truly relevant predictor variables. Our simulation results also illustrate the impact of the choice of the random forest tuning parameter mtry. While the default value mtry = p = Math phenomeMTDGVsDB@v=faeagat1tmrz9s2avn6vV5IVKts0learuWH9MDH5MBPija92xxaXtAldB4l4bbdbbhffHDYgaaracPG6WnIi=XHvBvGVg1G=+eEesex0xXdbab91fY5dbRqq is often found to be optimal with respect to prediction accuracy in empirical studies [33], our findings indicate that in case of correlated predictor variables different values of mtry should be considered. It is however, it should also be noted that any interpretation of random forest variable importance scores can only be sensible when the number of trees is chosen sufficiently large such that the results produced with different random seeds do not vary systematically. Only then it is assured that the differences between, e.g., unconditional and conditional importance are not only due to random variation. The conditional permutation importance will be freely available in the next release of the party package for recursive partitioning [21, 31] in the R system for statistical computing [22].
Authors’ original submitted files for images
Below are the links to the authors’ original submitted files for images:
12859_2008_2292_MOESM4_ESM.pdf Authors’ original file for image 1
12859_2008_2292_MOESM5_ESM.pdf Authors’ original file for image 2
12859_2008_2292_MOESM6_ESM.pdf Authors’ original file for image 3
12859_2008_2292_MOESM7_ESM.pdf Authors’ original file for image 4

Authors’ contributions
CS defined the research question, suggested the conditional variable importance, set up and performed the simulation experiments and drafted the manuscript. A-LB analyzed the peptide-binding data. TK, TA and AZ contributed to the theoretical understanding and presentation of the problem. All authors contributed to and approved the final version of the manuscript.

Authors’ Affiliations
(1) Department of Statistics, Ludwig Maximilians-Universität München, Ludwigsstraße 33, D-80539 München, Germany
(2) Sylva Lawry Centre for Multiple Sclerosis Research, Hohenfelder Straße 1, D-81677 München, Germany
(3) Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Augasse 2 – 6, A-1090 Wien, Austria

References
Unfortunately, variable-importance measures in random forests are subject to the same bias in favor of variables with many categories and continuous variables that affects variable selection in single trees, and also to a new source of bias induced by the resampling scheme (Strobl et al., 2007b). Both problems can be addressed in party to guarantee unbiased variable selection and variable importance for predictor variables of different types. Even though this refined approach can provide reliable variable-importance measures in many applications, the original permutation importance can be m... Conditional variable importance for random forests. BMC Bioinformatics, 9:307, 2008. C. Strobl, J. Malley, and G. Tutz.