Supplementary materials for this article are available at http://pubs.amstat.org/toc/tas/63/4.

Variable Importance Assessment in Regression: Linear Regression versus Random Forest

Ulrike GRÖMPING

Relative importance of regressor variables is an old topic that still awaits a satisfactory solution. When interest is in attributing importance in linear regression, averaging over orderings methods for decomposing R^2 are among the state-of-the-art methods, although the mechanism behind their behavior is not (yet) completely understood. Random forests—a machine-learning tool for classification and regression proposed a few years ago—have an inherent procedure of producing variable importances. This article compares the two approaches (linear model on the one hand and two versions of random forests on the other hand) and finds both striking similarities and differences, some of which can be explained whereas others remain a challenge. The investigation improves understanding of the nature of variable importance in random forests. This article has supplementary material online.

KEY WORDS: Linear model; Random forest; Variable importance.

1. INTRODUCTION

Variable importance in regression is an important topic in applied statistics that keeps coming up in spite of critics who basically claim that the question should not have been asked in the first place (cf., e.g., Ehrenberg 1990; Christensen 1992; Stufken 1992). Grömping (2007) recently provided an overview and a detailed discussion of the properties of two methods that derive variable importance in linear regression based on variance decomposition. Chevan and Sutherland (1991) proposed "Hierarchical Partitioning" for more general univariate regression situations; Theil and Chung (1988) discussed information-based approaches including multivariate linear regression.

Recently, random forests have received a lot of attention in biostatistics and other fields. They are popular, because they can handle large numbers of variables with relatively small numbers of observations and in addition provide an assessment of variable importance (cf., e.g., Breiman 2001; Ishwaran 2007; Strobl et al. 2007). There are several recent articles on variable importance in random forests: van der Laan (2006) introduced a general concept based on causal effects. Ishwaran (2007) attempted to get a theoretical handle on MSE reduction in Random Forests based on Breiman et al.'s (1984) Classification And Regression Trees (called RF-CART in the sequel) by investigating the behavior of closed-form expressions of a modified version; of course, a modification for tractability comes with the risk of sacrificing generalizability. For uncorrelated regressors, Strobl et al. (2007) demonstrated that variable importance metrics in RF-CART are biased under relevant circumstances and introduced a different type of forest which does not exhibit this bias in the uncorrelated regressor situations they simulated. For correlated regressors, Strobl et al. (2008) found that the proposed solution of Strobl et al. (2007) does not solve all issues. Thus, they proposed "conditional variable importance" as a modification to the algorithm for determining variable importance in random forests.

Random forests can be used for classification and regression, as was already suggested by Breiman 2001; random survival forests have also been proposed (cf., e.g., Hothorn et al. 2004; Ishwaran et al. 2008). This article investigates the regression situation and compares the newly proposed variable importance measures from two specific types of random forests to the more classical tools for linear regression models. The focus is on inter-regressor correlation as an important determinant of the behavior of variable importance metrics. Here, the random forest variable importance approach can benefit from the somewhat more advanced understanding of what happens in linear models.

Linear regression is a classical parametric method which requires explicit modeling of nonlinearities and interactions, if necessary. It is known to be reasonably robust, if the number of observations n is distinctly larger than the number of variables p ($n \gg p$). With more variables than observations ($p > n$ or even $p \gg n$), linear regression breaks down, unless shrinkage methods are used like ridge regression (Hoerl and Kennard 1970), the lasso (Tibshirani 1996), or the elastic net as a combination of both (Zou and Hastie 2005). Random forests, on the other hand, are nonparametric and allow nonlinearities and interactions to be learned from the data without any need to explicitly model them. Also, they have been reported to work well not only for the $n \gg p$ setting but also for data mining in the $p \gg n$ setting. Reasons for usage of variable importances also differ in the two scenarios (cf. Section 7 for a detailed discussion). This article concentrates on the $n \gg p$ situation. Nevertheless, the findings will also have implications for $p \gg n$ variable selection applications and will shed some light into the black box of random forests.

The next section briefly introduces the example dataset which will be used in Sections 4 and 5 for illustration and method comparison. Section 3 presents the linear model along

Ulrike Grömping is Professor, Department II—Mathematics, Physics, Chemistry, BHT Berlin—University of Applied Sciences, Luxemburger Str. 10, D-13353 Berlin, Germany (E-mail: *groemping@bht-berlin.de*).

with its relative importance metrics, whereas Section 4 introduces two variants of random regression forests and their associated variable importance metric. Section 5 uses the example data to compare the different methods. Section 6 presents a simulation study under systematically varied correlation structures and coefficient vectors like those in the article by Grömping (2007) that compares (i) variance decomposition based on averaging over orderings (Lindeman, Merenda, and Gold 1980; Kruskal 1987a, 1987b; Feldman 2005) and (ii) random forest variable importance metrics for two types of forests. Interpretation of results is followed by pointing out areas of interest for further research. The final Section 7 discusses in detail the purpose-specific conceptual needs for variable importance metrics in both linear model and random forest.

2. SWISS FERTILITY EXAMPLE

The R software (R Development Core Team 2008) contains a small socio-demographic dataset ("swiss") on Fertility in 47 Swiss provinces in 1888 that is suitable for demonstrating the varying behaviors of different approaches. A larger dataset including the same variables for 182 provinces is available online (Switzerland Socio-economic variables 1870 to 1930, *http://opr.princeton.edu/archive/pefp/switz.asp*) and has been used here for better stability of results. The set of variables has been kept the same: Fertility rate in the married population (Fertility), percentage of male population in agriculture jobs (Agriculture), percentage of draftees with highest grade in an army exam (Examination), percentage of draftees with more than primary school education (Education), percentage of catholics (Catholic), and percentage of children who did not survive their first year (Infant.Mortality). It is not entirely clear how the smaller and the larger dataset are related (the maximum of some variables is slightly larger in the smaller dataset). Because the analyses are only intended as examples, this has not been further pursued. Figure 1 provides an overview of the bivariate relations for the data.

3. LINEAR REGRESSION AND IMPORTANCE METRICS

The linear regression model is considered in its usual form

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon,$$
$$\beta_0, \beta_1, \ldots, \beta_p \text{ fixed and unknown,} \qquad (1)$$

where the random variables X_j, $j = 1, \ldots, p$, denote p regressor variables and the random variable ε denotes an error term, which is uncorrelated to the regressors and has expectation 0 and variance $\sigma^2 > 0$. The regressor variances are denoted as v_j, $j = 1, \ldots, p$, the inter-regressor correlations as ρ_{jk}, and the $p \times p$ covariance matrix between regressors is assumed to be positive definite so that any sample regressor matrix with $n > p$ rows is of full column rank with probability 1. Model (1) implies the conditional moments $\mathrm{E}(Y|X_1, \ldots, X_p) = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p$ and $\mathrm{var}(Y|X_1, \ldots, X_p) = \mathrm{var}(\varepsilon|X_1, \ldots, X_p) = \sigma^2$ and the marginal variance model

$$\mathrm{var}(Y) = \sum_{j=1}^{p} \beta_j^2 v_j + 2\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_j\beta_k\sqrt{v_j v_k}\rho_{jk} + \sigma^2. \qquad (2)$$

Formula (2) depends on β_j and v_j through $\beta_j\sqrt{v_j}$ only, the estimated version of which is equivalent to using the standardized coefficient, because division by the standard deviation of the response is not relevant when looking at one response only.

The first two summands of (2) constitute the part of the variance that is explained by the regressors, and R^2 from a linear

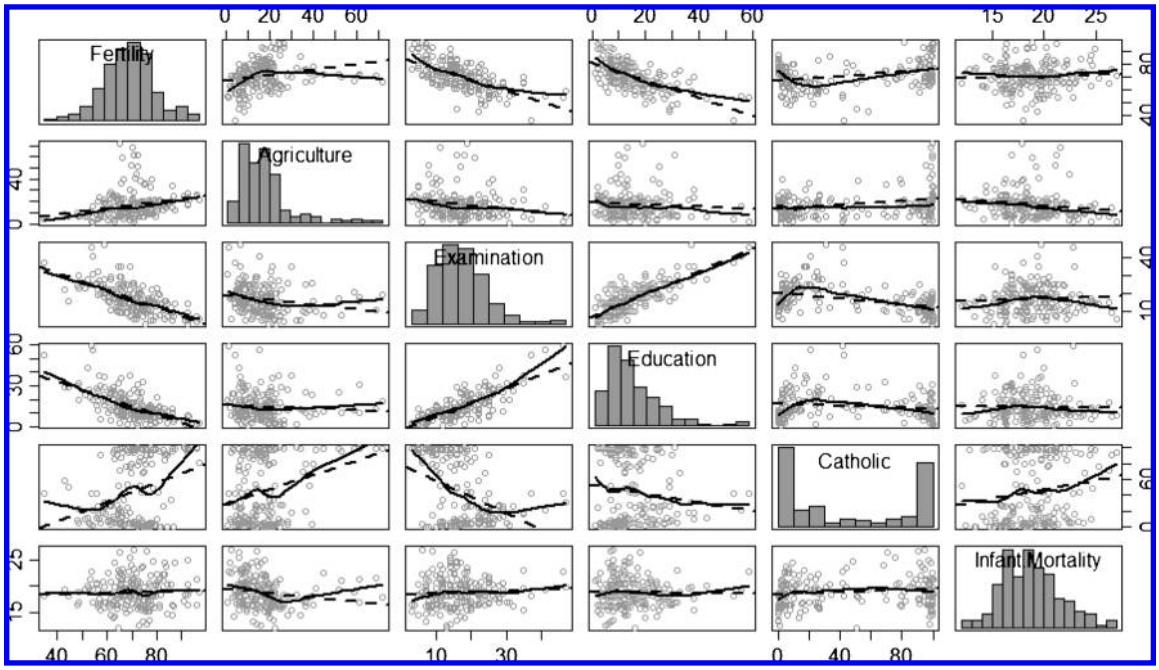

Figure 1. Scatterplot matrix of Swiss Fertility data with linear and loess lines.

model with n independent observations is consistent for the proportion of the first two summands of (2) in the total var(Y). Variable importance methods that decompose R^2 thus have to decompose the first two summands of (2). It is well known that this variance—or equivalently the model sum of squares or R^2—can be uniquely decomposed in case of uncorrelated regressors, but that different methods lead to different results for correlated regressors. In particular, the increase in R^2 allocated to a certain regressor X_k depends on which regressors are already present in the model before adding X_k. For any particular order of regressors, a unique allocation can be determined by allocating each regressor the increase in R^2 (or explained variance) when it is added to the model. Lindeman, Merenda, and Gold (1980, henceforth, LMG) proposed to average such order-dependent allocations over all $p!$ orderings for a fair unique assessment. This approach (independently proposed also by Kruskal (1987a, 1987b)) will be termed LMG throughout this article. Feldman (2005) proposed a modified version with data-dependent weights that favor strong predictors (called PMVD for Proportional Marginal Variance Decomposition following Feldman). For mathematical detail on LMG and PMVD, see, for example, the article by Grömping 2007. Note that LMG has been proposed under various different names, for example, dominance analysis (Budescu 1993) or Shapley value regression (Lipovetsky and Conklin 2001). In this article, LMG and PMVD are compared to random forest variable importance assessments.

4. RANDOM FORESTS FOR REGRESSION

A forest is an ensemble of trees—like in real life. Breiman (2001) introduced the general concept of random forests and proposed one specific instance of this concept, which we will consider as RF-CART in the following. A further instance proposed by Strobl et al. (2007), RF-CI, will also be introduced below. The types of trees used in these two concepts will be presented in the following subsection, before introducing their integration into a forest approach in Section 4.2.

4.1 Regression Trees

A regression tree (cf. Figure 2 for an example) is built by recursively partitioning the sample (= the "root node") into more and more homogeneous groups, so-called nodes, down to the "terminal nodes." Each split is based on the values of one variable and is selected according to a splitting criterion. Once a tree has been built, the response for any observation can be predicted by following the path from the root node down to the appropriate terminal node of the tree, based on the observed values for the splitting variables, and the predicted response value simply is the average response in that terminal node. For example, the left tree shown in Figure 2, modeling Fertility based on five candidate regressors, would predict the Fertility for a Swiss province with Education 20% and Agriculture 5% as 55.69% (go left on both splits). The regression function thus estimated from a tree is a multidimensional step function. This article considers binary trees only, that is, trees that split a parent node into two children at any step.

4.1.1 CART Trees

The CART algorithm proposed by Breiman et al. (1984) chooses the split for each node such that maximum reduction in overall node impurity is achieved, where impurity is measured as the total sum of squared deviations from node centers. CART first grows a tree very large and subsequently "prunes" it, that is, cuts off branches that do not add to predictive performance according to a pruning criterion that can differ from the splitting criterion. The reason for pruning a large tree instead of growing a small tree only in the first place is an improvement in the predictive performance of the tree (stopping too early might miss out on later improvements). If only one tree is built, care is needed to make sure the tree does not overfit the data. For this purpose, the degree of pruning is typically decided based on cross-validation. If CART trees are used in random forests, they are typically grown quite large, and no pruning is done (cf. right tree in Figure 2); for details, see Section 4.2.

4.1.2 Conditional Inference Trees

The splitting approach in CART trees has been known for a long time to be unfair in the presence of regressor variables of different types, categorical variables with different numbers of categories, or differing numbers of missing values (cf., e.g., Breiman 1984; Shih and Tsai 2004). To avoid this variable selection bias, Hothorn, Hornik, and Zeileis (2006b) proposed to use multiplicity-adjusted conditional tests (cf. Hothorn et al.

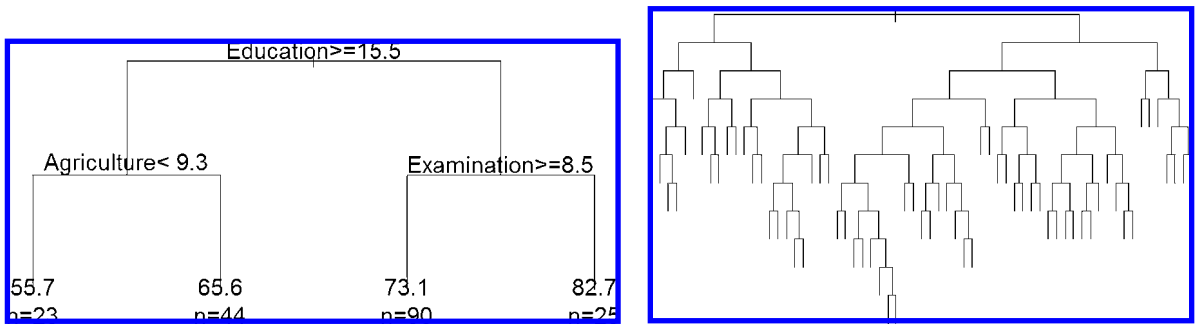

Figure 2. Individual tree and tree used in a forest (computed with R-package rpart; Therneau and Atkinson 1997). *Left*: A CART tree for Fertility for the Swiss data. Splits are labeled with the splitting criterion. Split condition true goes to the left, false to the right. *Right*: An unpruned CART tree for the same data that might be used in a forest (nodes of size 5 or less are terminal).

2006a) rather than maximum impurity reduction as the splitting criterion. The key idea of their approach is to untangle variable selection and split selection, similarly to the well-known CHAID approach (Kass 1980) for classification. Thus, for any node to be split, the procedure conducts a global permutation test of the null hypothesis of no association between any of the regressors and the response within the node. If this global null hypothesis is not rejected, the node is not split and becomes a terminal node. Otherwise, individual null hypotheses of no association to the response are tested for each variable, and the variable with the smallest p-value is selected for splitting. Subsequently, the splitting rule is determined based on the selected variable. With this method, pruning is not needed, because trees are not grown any further, once there is no further statistically significant split. For the Swiss Fertility data, the conditional tree is similar to the left CART tree in Figure 2. Analogously to CART, where pruning is usually omitted when growing forests, conditional inference trees for forests are also per default grown larger than those used in single tree analyses (cf. next section).

4.2 Random Forests

A random forest consists of a large number, *ntree*, of trees, for example 1000. Theoretical results on random forests (Breiman 2001; Ishwaran 2007) are asymptotic in the number of trees, and a large number of trees has been reported to be particularly important when interest is in diagnostic quantities like variable importance (cf., e.g., Breiman 2002). A random forest is random in two ways: (i) each tree is based on a random subset of the observations, and (ii) each split within each tree is created based on a random subset of *mtry* candidate variables. Trees are quite unstable, so that this randomness creates differences in individual trees' predictions. The overall prediction of the forest is the average of predictions from the individual trees—because individual trees produce multidimensional step functions, their average is again a multidimensional step function that can nevertheless predict smooth functions because it aggregates a large number of different trees. For visualization of forest results, main effects and interaction plots similar to what has been proposed by Friedman (1991) for MARS can be used; cf. also Section 5 below.

4.2.1 RF-CART and RF-CI

RF-CART, that is, random forests based on CART trees, are most well known, because they have already been proposed in the fundamental article by Breiman (2001). Recently, Strobl et al. (2007) proposed to base random forests on the conditional inference trees discussed in Section 4.1.2. To highlight that the key difference lies in the different type of underlying trees (CART versus CI = conditional inference), this approach is called RF-CI in the following. Here, the two forest types have been applied as implemented in the R-packages randomForest (Liaw and Wiener 2002 based on Breiman 2001, 2002) and party (function cforest; Strobl et al. 2007). The default number of trees ($ntree = 500$) is identical for both forest variants. The number *mtry* of variables to be considered for each split is a tuning parameter, which has the default floor($p/3$) for RF-CART. With $mtry = 1$, the splitting variable would be determined completely at random, whereas $mtry = p$ would eliminate one aspect of randomness for the forest, and it has been recommended to try half and twice the default as well (Liaw and Wiener 2002). For RF-CI, *mtry* has no meaningful default. Choice of *mtry* is further discussed in connection with simulation results (cf. Section 6.3).

RF-CART and RF-CI use different default sampling approaches: RF-CART uses a with-replacement sample of size n, RF-CI a without-replacement sample of size $0.632 * n$. According to Strobl et al. (2007), this difference is inconsequential in the setting investigated here with continuous regressors only. The defaults of RF-CART and RF-CI also result in very different tree sizes: RF-CART uses large unpruned trees that are grown with a lower limit for the size of nodes to be considered for splitting (nodes of size 5 or less are not split), for example, like the one on the right in Figure 2. RF-CI trees for forests are per default built without significance testing (tests without multiplicity adjustment are conducted at a default significance level of 100%), limiting tree size by restricting minimum split size for a node to 20 and minimum node size to 7. This stricter criterion, together with the sampling approach, implies that trees in RF-CI are per default substantially smaller than those in RF-CART (e.g., 8 to 11 terminal nodes for a typical tree within a RF-CI forest of the example data, in comparison to about 50 to 70 terminal nodes for RF-CART trees). Settings for both RF-CART and RF-CI could be adjusted such that the trees become larger or smaller by adjusting node splitting criteria, or—in case of RF-CI—by introducing a significance level smaller than 1. Increasing the minimum node size for RF-CART has been proposed by Segal, Barbour, and Grant (2004) for improving prediction accuracy, and selected simulation results with this modification will be discussed in Section 6.3.

4.2.2 Mean Squared Error

According to random sampling of observations, regardless whether with or without replacement, (an average of) 36.8% of the observations are not used for any individual tree—that is, are "out of the bag" = OOB for that tree. The accuracy of a random forest's prediction can be estimated from these OOB data as

$$\text{OOB-MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{\hat{y}_{i\,\text{OOB}}})^2,$$

where $\overline{\hat{y}_{i\,\text{OOB}}}$ denotes the average prediction for the ith observation from all trees for which this observation has been OOB. Analogously to linear regression, with the overall sum of squares $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ defined in the usual way, OOB-R^2 can be obtained as $1 - \text{OOB-MSE}/\text{SST}$.

4.2.3 Variable Importance

A very well-known variable importance metric in CART trees and random forests is the so-called Gini importance for classification and its analogue, average impurity reduction, for regression forests. However, because of impurity's bias for selecting split variables, the resulting variable importance metrics are of course also biased (cf., e.g., Shih and Tsai 2004;

Strobl et al. 2007). Breiman (2002) suggested reduction in MSE when permuting a variable (Method 1), called "MSE reduction" in the following, as the method of choice. Although he discarded this metric for excessive variability in situations with many regressors in a later version of the Random Forests manual (Breiman 2003), permutation-based MSE reduction has been adopted as the state-of-the-art approach by various authors (Diaz-Uriarte and Alvarez de Andrés 2006; Ishwaran 2007; Genuer, Poggi, and Tuleau 2008; Strobl et al. 2008). Therefore, this permutation-based "MSE reduction" is also used as the random forest importance criterion in this article. It is determined as follows: For tree t, the OOB mean squared error is calculated as the average of the squared deviations of OOB-responses from their respective predictions:

$$\text{OOBMSE}_t = \frac{1}{n_{\text{OOB},t}} \sum_{\substack{i=1:\\ i\in\text{OOB}_t}}^{n} (y_i - \hat{y}_{i,t})^2,$$

where the $\hat{}$ indicates predictions, $\text{OOB}_t = \{i : \text{observation } i \text{ is OOB for tree } t\}$, that is, summation is over OOB observations only, and $n_{\text{OOB},t}$ is the number of OOB observations in tree t. If regressor X_j does not have predictive value for the response, it should not make a difference if the values for X_j are randomly permuted in the OOB data before the predictions are generated. Thus,

$$\text{OOBMSE}_t(X_j \text{ permuted}) = \frac{1}{n_{\text{OOB},t}} \sum_{\substack{i=1:\\ i\in\text{OOB}_t}}^{n} (y_i - \hat{y}_{i,t}(X_j \text{ permuted}))^2 \qquad (3)$$

should not be substantially larger (or might by chance even be smaller) than OOBMSE_t. For each variable X_j in each tree t, the difference $\text{OOBMSE}_t(X_j \text{ permuted}) - \text{OOBMSE}_t$ is calculated based on one permutation of the variable's out-of-bag data for the tree (this difference is of course 0 for a variable that happens to be not involved in any split of tree t). The MSE reduction according to regressor X_j for the complete forest is obtained as the average over all *ntree* trees of these differences. It is worthwhile to realize that the thus-calculated MSE reduction is NOT the same as the reduction in the forest's MSE by having variable X_j available (versus not having X_j in the set of explanatory variables). This was also pointed out by Ishwaran et al. (2008).

Whereas LMG and PMVD naturally decompose R^2, such natural decomposition does not occur for forests' MSE reduction. Thus, for comparison purposes, all variable importance metrics in this article have been normalized to sum to 100%.

5. ALL METRICS APPLIED TO THE SWISS FERTILITY DATA

For the Swiss Fertility data, a linear model has been fitted using quadratic effects for Agriculture and Catholic and linear effects for the other three regressors, based on the impressions from Figure 1. The random forests have been fitted with default settings, apart from an increased *ntree* to ensure stability of the variable importance assessment (cf. Breiman 2002 or Genuer, Poggi, and Tuleau 2008 for recommended *ntree* values). In addition to the default $mtry = 1$, $mtry = 2$ has been run for comparison.

Table 1 shows normalized variable importance metrics from all approaches. Figure 3 shows effects plots for the linear model and for RF-CART. Within the linear model, LMG and PMVD allocations are almost identical, apart from the split within the only pair of strongly correlated regressors, Examination and Education (correlation: 0.79): Here, LMG gives Examination the benefit of the doubt, whereas PMVD allocates almost no contribution to Examination. PMVD's bootstrap confidence interval for the share of Examination (not shown) includes the full LMG interval for the same share in this example (cf. also Grömping 2007 for a discussion of the variance properties of PMVD and LMG); that is, PMVD is extremely variable here. Importances from the two types of forest behave somewhat differently both from each other and from the linear model assessments. None of the forest metrics shows any similarity to PMVD regarding Examination's share, whereas the RF-CI allocation for Education is almost as extreme as PMVD's with $mtry = 2$. Table 1 also shows that the RF-CI assessment strongly depends on *mtry*, whereas the RF-CART assessment is more stable over *mtry*. This behavior will also be confirmed by the simulation study of Section 6. Results on *mtry* from the literature will be discussed in that context (Section 6.3).

Table 1. Relative importance of the five effects for "Fertility" normalized to sum 100%* in linear regression and random forest models.

			Random Forest MSE reduction, *ntree* = 2000			
Response:	Linear model** ($R^2 = 61.3\%$)		*mtry* = 1***		*mtry* = 2***	
Fertility	PMVD	LMG	RF-CART	RF-CI	RF-CART	RF-CI
Agriculture	21.3	22.0	26.1	20.7	31.2	14.4
Examination	1.0	25.6	22.9	28.8	20.2	31.8
Education	56.3	31.5	28.5	35.9	30.1	44.9
Catholic	18.2	18.3	17.6	12.9	13.3	8.6
Infant.Mortality	3.3	2.7	4.9	1.7	5.2	0.3
Total	100.0	100.0	100.0	100.0	100.0	100.0

*Normalization to sum 100% is not recommended for data analysis purposes, but is helpful for making metrics' relative assessments comparable.

**Agriculture and Catholic quadratic, the other variables linear; calculation with R-package relaimpo (cf. Grömping 2006).

***$mtry$ is the number of candidate variables randomly selected for each split in each tree.

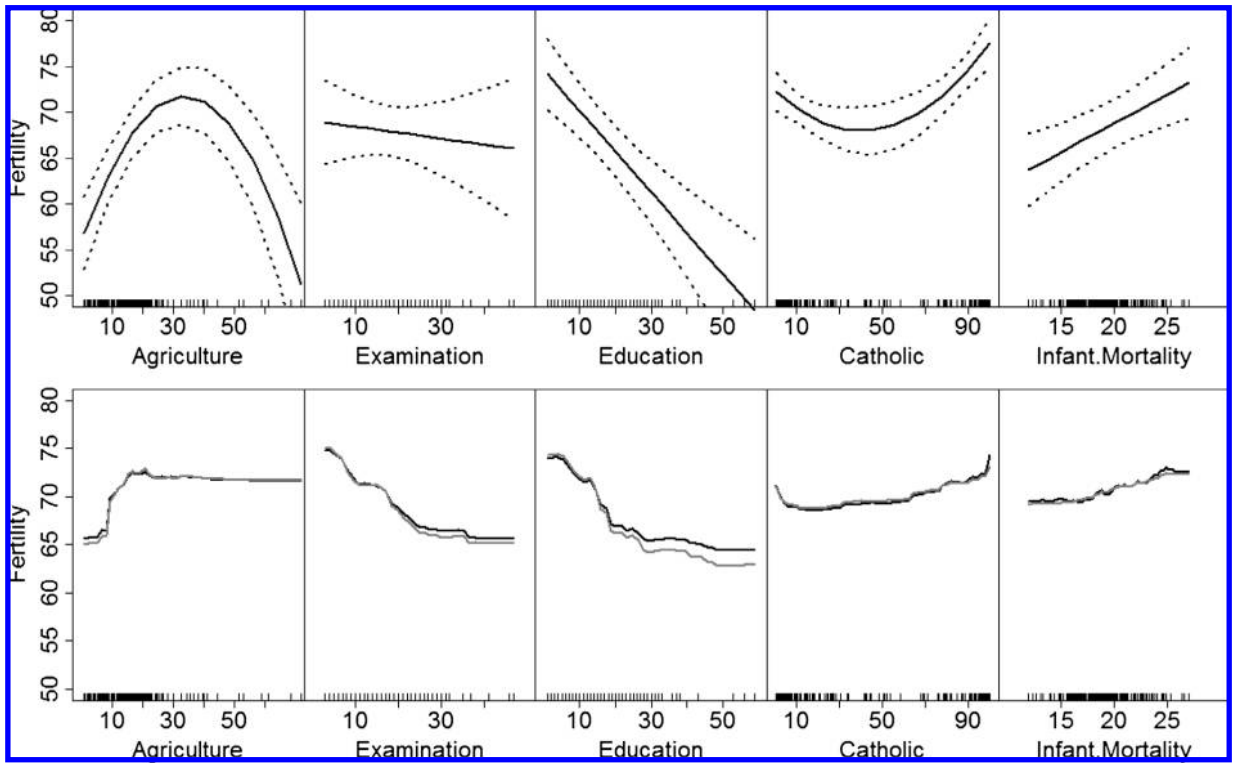

Figure 3. Main effects plot for the linear model (top, with 95% bands) and RF-CART ($mtry = 1$ (black) and $mtry = 2$ (gray)). Rugs at the bottom represent individual data values for the 182 Swiss provinces.

Of course, with this real data example, there is no guarantee that the data are adequately described by the chosen linear model with three linear and two quadratic effects and without interactions. Differences between allocations from the linear model and the two forest approaches in the example analysis may be due both to method differences and to deviations between the postulated linear model structure and the true model that the nonparametric forest approaches have attempted to estimate.

6. SIMULATION STUDY

To focus on the method-related differences only, this section reports a simulation study for truly linear models. Note that forests can only approximate linear models, because they always fit step functions. However, it has been proven that they closely approximate any smooth expectation model for large samples and large numbers of trees (Ishwaran 2007) over a finite closed rectangular regressor space. Of course, the linear model is on its home turf, whereas the forests are far less parsimonious for fitting a truly linear model. This is reflected by much larger dispersion of forest-based variable importance metrics, which is, however, not in the focus of this study. It can be conjectured that, under the default settings used here, RF-CART is more efficient in approximating a continuous model than RF-CI because of the larger number of nodes in each tree.

6.1 Simulation Scenarios

The simulation scenarios have been chosen as a subset from those in Grömping (2007); for these scenarios, LMG and PMVD have already been investigated, and their true values (in terms of limits when increasing the sample size to infinity) can be calculated. Here, 100 samples of size $n = 1000$ each have been simulated in the following way: Four regressors $X_1, \ldots, X_4$ have been generated as n independent observation vectors from four-variate normal distributions with expectations 0, variances 1, and correlations $\text{corr}(X_j, X_k) = \rho^{|j-k|}$ with $\rho = -0.9$ to 0.9 in steps of 0.1, $j, k = 1, \ldots, 4$. Note that negative values of ρ provide mixed-sign correlation structures among regressors, that is, positive and negative values of ρ yield structurally different correlation matrices. Responses have been created from the regressors using seven different vectors of true model coefficients: $\boldsymbol{\beta}_1 = (4, 1, 1, 0.3)^T$, $\boldsymbol{\beta}_2 = (1, 1, 1, 0.3)^T$, $\boldsymbol{\beta}_3 = (4, 1, 0, 0)^T$, $\boldsymbol{\beta}_4 = (1, 1, 1, 1)^T$, $\boldsymbol{\beta}_5 = (1.2, 1, 1, 0.3)^T$, $\boldsymbol{\beta}_6 = (1, 1, 1, 0)^T$, $\boldsymbol{\beta}_7 = (4, 3.5, 3, 2.5)^T$), adding normal random error with σ^2 such that $R^2 = 50\%$. The two forest methods have been implemented with 500 trees for each forest using their previously given default settings (cf. Section 4.2). The aforementioned instability of MSE reductions (Breiman 2002; Genuer, Poggi, and Tuleau 2008) is not an issue here, because variation is not the topic of investigation and averages over 100 simulation runs are reasonably stable. Simulation code for implementing similar studies—adjusted to run with the current version of R-package party for RF-CI—can be found online as supplemental material.

6.2 Descriptive Method Comparisons

This section describes simulation results in terms of average normalized variable importances from random forests and true normalized values for LMG and PMVD—against which their estimates are consistent. LMG and PMVD values are shown as reference curves to indicate that they are the theoretical limits as opposed to averages over simulation runs. This is not meant to indicate that they represent an overall gold standard. A key difference between LMG and PMVD can be observed in Figure 4: For LMG, the importance allocated to the regressor with the largest coefficient decreases substantially in favor of the importances allocated to the other regressors with increasing degree of correlation. This is called "equalizing behavior" in the following.

Average variable importance from the forests with $mtry = 1$ (Figure 4) is found to be quite similar to LMG. However, the similarities are far from perfect, and average variable importance from RF-CI shows a slight tendency toward PMVD. With increasing *mtry* (cf. Figures 5 and 6), variable importance for RF-CI becomes more and more similar to PMVD, whereas variable importance for RF-CART is *relatively* stable over *mtry* and remains similar to LMG.

Whereas RF-CI shows a much stronger dependence on *mtry* than RF-CART, dependence of allocations on the correlation parameter depends on the situation: LMG and RF-CART show a much stronger dependence, for example, in Figure 5, whereas the dependence is stronger for PMVD and particularly for RF-CIs in Figure 6. The most striking feature for Figure 6 is the strong correlation-dependence of allocations to the first three X's for RF-CI. This is much stronger than for PMVD and increases with increasing *mtry*, so that the regressor with the largest coefficient loses its first rank to the second regressor for a high positive correlation parameter and high *mtry*, whereas the regressor with the smallest coefficient overtakes the two medium regressors for a strongly negative correlation parameter (and thus a mixed-sign correlation pattern) with increasing *mtry*. This behavior is not yet understood.

Apart from the correlation pattern and the tuning parameter *mtry*, the sample size is also relevant for variable importance allocations: For simulations with smaller sample size ($n = 100$, not shown)—for which good approximation of a linear model by a random forest is not guaranteed—RF-CI allocations were far less similar to PMVD even in situations where agreement is almost perfect for $n = 1000$ (e.g., $\boldsymbol{\beta}_1$ with $mtry = 4$). In case of RF-CART, a closer similarity to LMG has been observed for smaller samples, whereas larger samples tend to be more equalizing than LMG between weaker and stronger regressors (but not for regressors with no influence) for low degrees of correlation (cf., e.g., Figure 5).

6.3 Discussion of the Dependence on *mtry*

The tuning parameter *mtry* deserves special attention. Breiman (2001) recommended $mtry = \sqrt{p}$ for classification forests and observed that larger numbers would be needed for situations for which many irrelevant inputs confuse the picture. He indicated that $mtry \ll p$ should improve predictive performance

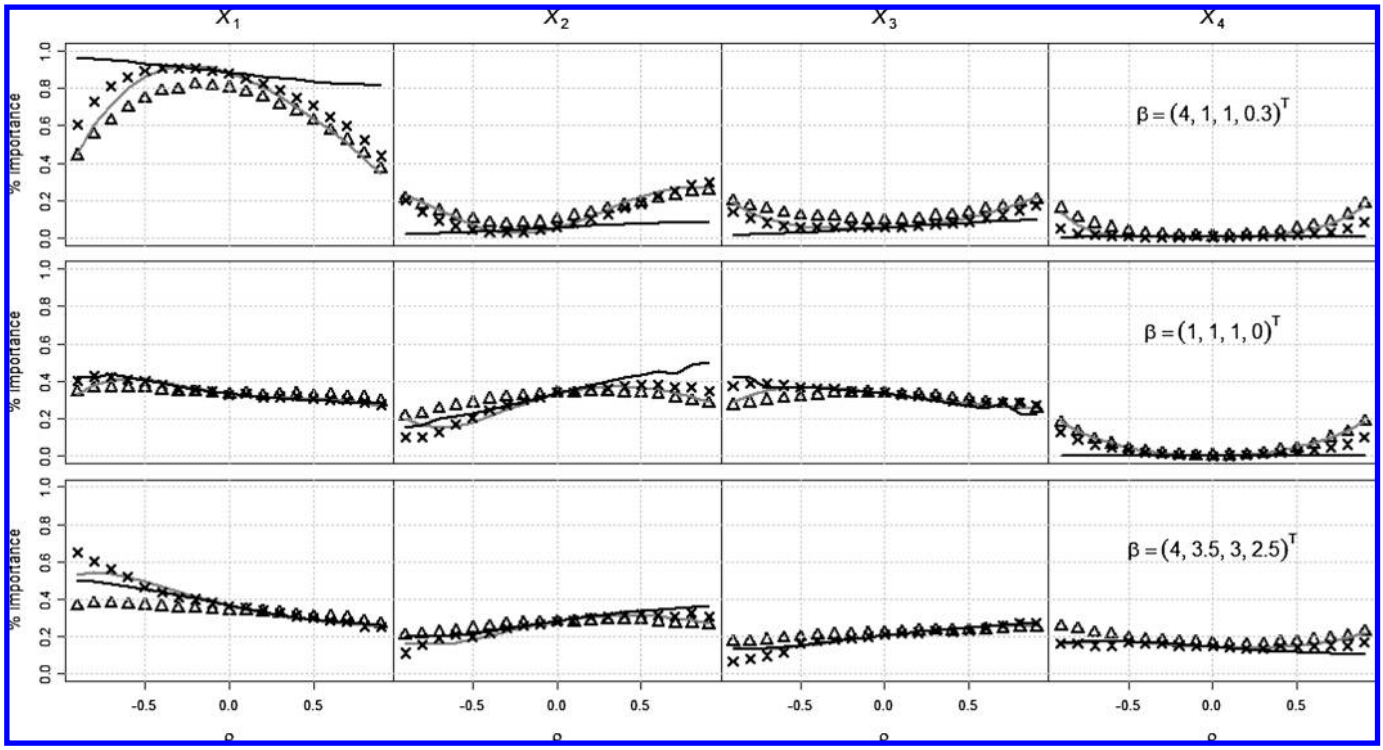

Figure 4. Average normalized importances for the four X's from 100 simulated datasets based on $mtry = 1$ variables per split, $\boldsymbol{\beta}_1 = (4, 1, 1, 0.3)^T$ (top), $\boldsymbol{\beta}_6 = (1, 1, 1, 0)^T$ (middle), and $\boldsymbol{\beta}_7 = (4, 3.5, 3, 2.5)^T$ (bottom), $\text{corr}(X_j, X_k) = \rho^{|j-k|}$ with $\rho = -0.9$ to 0.9 in steps of 0.1. Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation. △: variable importance (% MSE reduction) from RF-CART, ×: variable importance (% MSE reduction) from RF-CI.

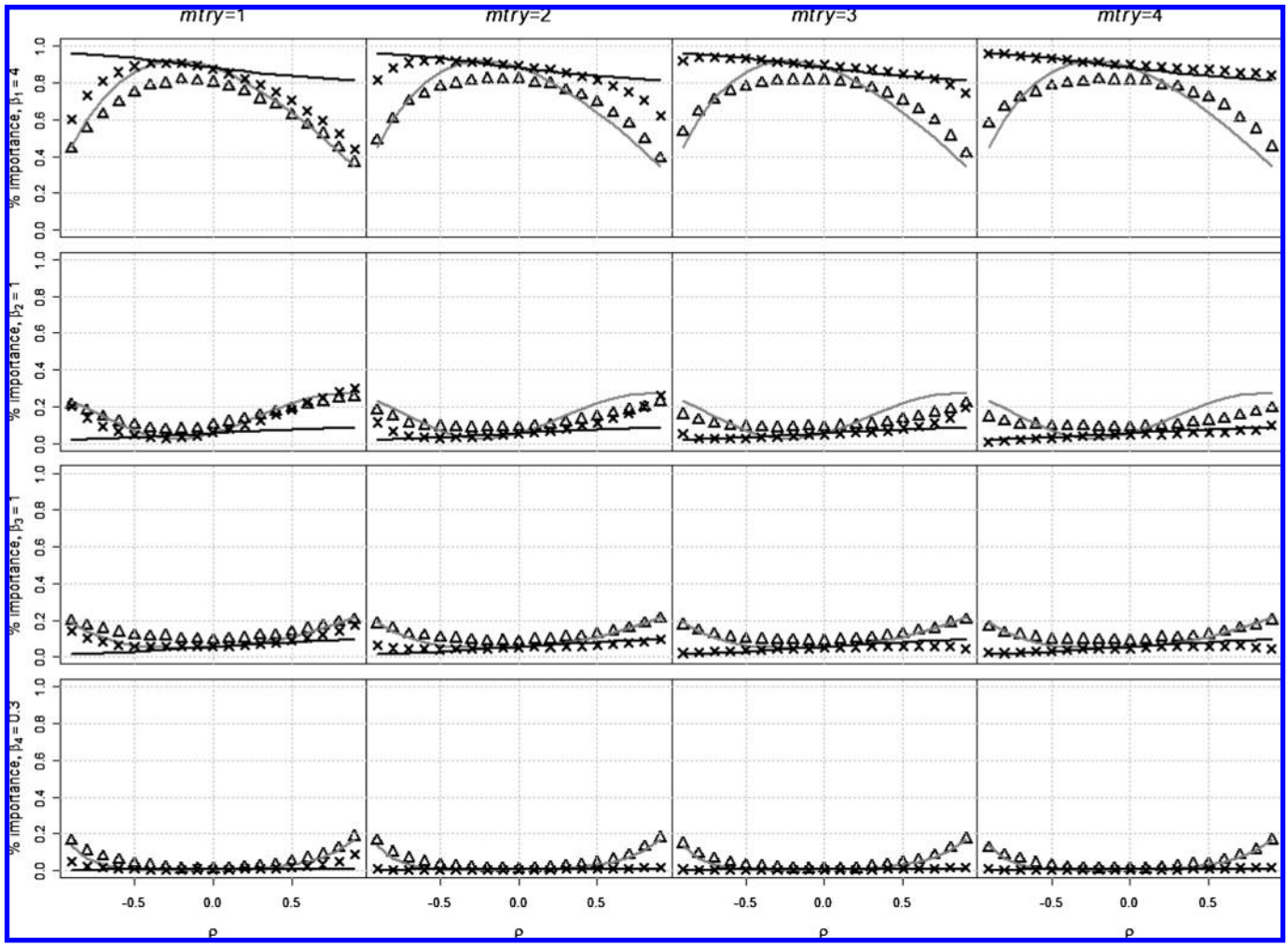

Figure 5. Average normalized importances for the four X's (top to bottom: X_1 to X_4) from 100 simulated datasets for $mtry = 1, 2, 3, 4$ (left to right) with $\boldsymbol{\beta}_1 = (4, 1, 1, 0.3)^T$, $\text{corr}(X_j, X_k) = \rho^{|j-k|}$ with $\rho = -0.9$ to 0.9 in steps of 0.1. Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation. △: variable importance (% MSE reduction) from RF-CART, ×: variable importance (% MSE reduction) from RF-CI.

versus using all variables, because lower correlation between individual trees improves prediction accuracy. He also observed this effect to be much weaker in regression forests, and consequently the proposed default of $p/3$ for regression forests (Liaw and Wiener 2002) is much larger than $\sqrt{p}$ for large numbers of variables. Recently Genuer, Poggi, and Tuleau (2008) presented simulation results that indicate better prediction accuracy with $mtry = p$ for the artificial data from the article by Friedman (1991). Within our simulation study, the *mtry* value that minimized OOB-MSE depended on both the model and the correlation between regressors (but was generally $mtry = 1$ or $mtry = 2$, and in most cases performance differences were small).

Diaz-Uriarte and Alvarez de Andrés (2006) investigated *mtry*'s impact on variable importance for classification forests and found the default values to work satisfactorily and independently of the settings of other parameters like *ntree*. The empirical investigations in this article have shown that the choice of *mtry* can substantially affect the allocated importance in random forests for regression. The reasons for this behavior are most obvious when considering $mtry = 1$: By random choice, a regressor with no relation to the response, which would have never been selected for a split given any competition, will by chance sometimes become the basis of splitting. If the regressor is unrelated to both the response and the other regressors, it will only get a weak share allocated both in RF-CART and in RF-CIs (e.g., Figure 4, X_4 in the middle scenario for $\rho = 0$). This is because even if such a regressor has by chance been made the basis of a split, the % MSE reduction will be on average zero. Now, consider a variable with no conditional influence of its own (coefficient 0 in the model) but a strong genuine correlation to one or more of the regressors that do have a nonzero coefficient. With $mtry = 1$, such a variable will sometimes be the only candidate for a split and will as such—because of its covariance with the response—create splits that do reduce impurity and will also decrease % MSE for OOB cases; that is, the permuted variable will perform worse than the original variable for the OOB cases because the split based on this variable picks up a real influence of the correlated regressor(s) that is lost otherwise for the particular tree. With $mtry = 2$ or larger, one single

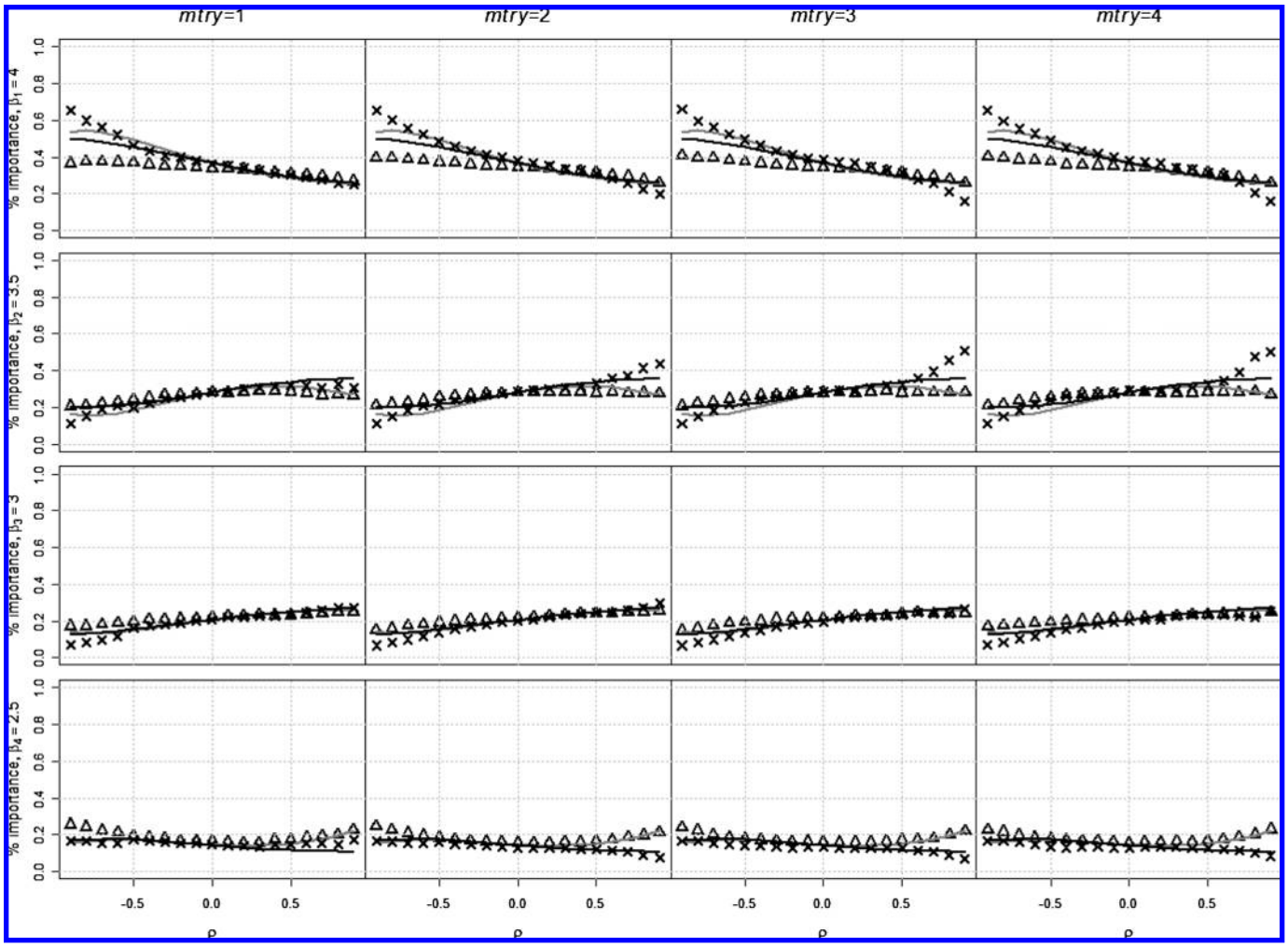

Figure 6. Average normalized importances for the four X's (top to bottom: X_1 to X_4) from 100 simulated datasets for $mtry = 1, 2, 3, 4$ (left to right) with $\boldsymbol{\beta}_7 = (4, 3.5, 3, 2.5)^T$, $\text{corr}(X_j, X_k) = \rho^{|j-k|}$ with $\rho = -0.9$ to 0.9 in steps of 0.1. Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation. △: variable importance (% MSE reduction) from RF-CART, ×: variable importance (% MSE reduction) from RF-CI.

weak regressor is usually in competition with another stronger regressor and will not win often, unless it is related to a stronger regressor: A weak regressor will stand in for a related stronger regressor, if the stronger regressor is neither sufficiently accommodated for in previous splits of the tree nor present in the current candidate set. With *mtry* increasing, it becomes less and less likely that a weak regressor will become the splitter. RF-CI will likely have stopped splitting before the weak regressors come into play with larger *mtry*. Therefore, it is plausible that the weaker effects will get lower allocations for larger values of *mtry* in RF-CI. This phenomenon has also been observed by Strobl et al. (2008, sec. 2.2) in the classification context. These considerations do not, however, capture the full nature of RF-CI allocations; especially, the behavior in Figure 6 that has been described above still awaits explanation.

Allocations from RF-CART were far less dependent on *mtry*. Conjecturing this behavior to be due to the much larger size of individual trees, a few simulations with minimum node sizes for splits of 2, 21, and 51 instead of the default 6 were run. These show that RF-CART with limited tree size (= larger minimum node size for splits) exhibits a much stronger dependence of allocations on *mtry*. The effect of reducing tree size was a strong increase in the equalizing behavior for small values of *mtry*, even for uncorrelated regressors. Note that—as previously mentioned for the standard tree setup—this equalizing behavior concerns strong and weak regressors but not regressors with *no* real effect. Also note that limiting tree size does not cause RF-CART to behave like RF-CI for $mtry = p$; RF-CART remains more equalizing than RF-CI.

6.4 Aspects for Further Research

As pointed out in the previous subsection, whereas it is reasonably well understood why strong variables substantially gain importance relative to weaker variables with increasing *mtry* in RF-CIs, other aspects of how variable importances depend on *mtry* should be subject to further investigation (e.g., behavior of RF-CI allocations in Figure 6). Also, it would be interesting to separate the effects of tree size from the effects of using p-values rather than maximum impurity reductions in comparing RF-CART to RF-CI. Furthermore, whereas it is convincing

that p-values are good at fairly treating unimportant variables of all types (i.e., null variables), it is not so clear that they are unbiased at representing the relative merits of several important variables of different strengths. The difficulty for investigating this lies in the fact that there is no easily identifiable gold-standard importance assessment of the strength for non-null variables (cf. also the following section).

The shape of the regressor space might be another interesting topic for further research. Contrary to linear models, forests (of course) do not only react to the mean- and covariance-structure of response and regressors but also to further aspects of the data. For example, variable importance in a forest has been found to depend on the shape of the regressor space: in a simulation (not shown) for the scenario of Section 6.1 with $\boldsymbol{\beta}_1$ and uncorrelated regressors distributed uniformly over a four-dimensional cube, variable importances differ depending on whether the cube sides are parallel to the coordinate axes or rotated (cf. graph and program in the supplemental material). It would be interesting to investigate this further.

7. THE CONCEPT OF VARIABLE IMPORTANCE

Variable importance is not very well defined as a concept. Even in the well-developed linear model and the standard $n \gg p$ situation, there is no theoretically defined variable importance metric in the sense of a parametric quantity that a variable importance estimator *should* try to estimate. In the absence of a clearly agreed true value, ad hoc proposals for empirical assessment of variable importance have been made, and desirability criteria for these have been formulated, for example, "decomposition" of R^2 into "nonnegative contributions attributable to each regressor" has been postulated (cf., e.g., Grömping 2007 for more detail). Popular approaches for empirically assessing a variable's importance include squared correlations (a completely marginal approach) and squared standardized coefficients (an approach conditional on all other variables in the model), as critically discussed, for example, by Darlington (1968). In line with Johnson and Lebreton's (2004) (vague) definition of relative importance, LMG and PMVD account for both marginal and conditional aspects of importance by averaging over R^2 contributions from models with different variables preceding the respective variable of interest (and thus conditioned upon). LMG has been heuristically introduced and found a justification as a Shapley value later on (e.g., Stufken 1992) whereas PMVD was introduced by Feldman (2005) to satisfy the "exclusion" property, which requests that a variable with coefficient 0 should be allocated zero importance. Thus, PMVD is closer to a conditional perspective than LMG, in that it honors conditional unimportance of a variable given all others.

With their request that both conditional and marginal aspects need to be reflected in measuring relative importance, Johnson and Lebreton (2004) aimed for explanatory importance. Depending on the research question at hand, the focus of the variable importance assessment in regression can be explanatory or predictive importance or a mixture of both; cf. the article by Grömping (2007) for a more detailed discussion. To clearly differentiate between these two foci, consider the following very simple examples of a causal chain:

$$X_2 \to X_1 \to Y, \tag{4}$$

$$X_2 \leftarrow X_1 \to Y. \tag{5}$$

In causal chain (4), X_2 indirectly influences the response; in causal chain (5), X_2 is correlated to the response but does not influence it. If all relations are linear, a linear model for Y with regressors X_1 and X_2 will have a zero coefficient for X_2 in both (4) and (5), that is, conditional on X_1, X_2 does not contribute anything to the prediction. Thus, in the presence of sufficient data, both coefficient-based approaches and PMVD would allocate importance zero to X_2. This makes perfect sense for prediction purposes, and in causal chain (5) also for explanatory importance; however, most people would disagree that X_2 is unimportant for Y in an explanatory or causal sense in causal chain (4).

In random forest applications, variable importance is typically used for variable selection; random forests are especially popular for $p \gg n$ scenarios. In parallel to the distinction between explanatory/causal importance and predictive importance in conventional regression models, variable selection can serve two different objectives (cf., e.g., Diaz-Uriarte and Alvarez de Andrés 2006; Genuer, Poggi, and Tuleau 2008), namely (a) to identify important variables highly related to the response variable for explanatory and interpretation purposes (parallel to explanatory/causal) or (b) to identify a small number of variables sufficient for a good prediction of the response variable (parallel to prediction). In prediction or variable selection with purpose (b), one would strive to avoid redundancy and obtain a parsimonious prediction model. It is not so important that the model contains all relevant variables, as long as prediction works well. For example, ideally one would only select X_1 in causal chains (4) or (5), but selection of X_2 alone would also be acceptable, if the relation between X_1 and X_2 were sufficiently strong. (Because the linear model or the forest itself does not contain information about structural relationships between variables, it can be very difficult or even impossible to distinguish between the causal chains shown in (4) and (5) and the respective causal chains with the two X's in swapped roles.) On the other hand, for explanatory/causal importance or identification of potentially important variables in variable selection, it would be considered detrimental if an important variable is missed because of a low variable importance allocation, even though another highly correlated variable might well stand in for this variable in terms of prediction. Thus, one would certainly want the variable selection method to find both X_1 and X_2 in causal chain (4), whereas one would prefer not to consider X_2 as important in causal chain (5). However, because the two causal chains are indistinguishable for a linear model or a random forest, one would have to accept identifying X_2 as important in (5) as the price for being able to find it in (4).

The desire for an adequate variable selection method for purpose (a) was the starting point for Zou and Hastie's (2005) introduction of the elastic net: They modified the lasso, which is known to have a tendency to select one representative of a group of strongly correlated variables only (i.e., to work well for purpose (b)), into showing a "grouping property," that is,

a tendency to select correlated variables together. Put simply, grouping is achieved by biasing coefficients such that for highly correlated regressors coefficients with higher absolute values are shrunken toward the origin and coefficients with lower absolute values are "shrunken" away from the origin, so that the whole group ends up to be chosen together. "Grouping" of correlated regressors is conceptually close to what has been previously called LMG's "equalizing" behavior. Thus, purpose (a) is served better because of a reduced risk of missing an important variable in the presence of further variables highly correlated with it.

Within the range of squared (standardized) coefficients at the conditional extreme, over PMVD, LMG to squared marginal correlations at the marginal extreme, simulations showed that RF-CART and RF-CI with small *mtry* behave similarly to LMG, that is, balance between conditional and marginal approach leaning toward marginal, whereas RF-CI with large *mtry* behaves more similarly to PMVD, that is, also balances between marginal and conditional approach leaning toward the conditional end. Strobl et al. (2008) positioned themselves at the conditional extreme by considering it "bias" that variables with the same coefficients receive different importances due to inter-regressor correlations; thus, following their logic, one would also have to reject the idea of predicting Y based on X_2 instead of X_1 in causal chains (4) or (5) above, except for the limiting case for which there is a perfectly deterministic relation between the two regressors in the data. Of course, with $p \gg n$ variable selection, unbiased estimation of all coefficients is impossible—additional constraints or penalties (like in the elastic net) reintroduce estimability but also bias into the estimates. Strobl et al. (2008) suggested getting closer to conditional importance by using RF-CI (instead of RF-CART) together with a conditional instead of an unconditional permutation algorithm for the OOB data in (3). However, their approach does not attack a key driver of marginality in the random forest approach: as long as $mtry < p$, correlated variables will more or less frequently stand in for stronger predictors and thus act as splitters for marginal reasons. Results from the simulation study suggest that using RF-CI with $mtry = p$ might already go a long way toward making RF-CI more conditional.

Both random forest variable importances and LMG and PMVD are based on shares of the explained model variance or reductions in error variation by using the true instead of the permuted values for a variable in prediction (again, beware of misinterpreting forest's MSE reduction; cf. also Section 4.2.3). Achen (1982) referred to measures based on the response's dispersion as "dispersion importance" (cf. also Johnson and Lebreton 2004; Grömping 2007; Genuer, Poggi, and Tuleau 2008). It is obvious from (2) that the variance induced by regressors depends on both the coefficients and the correlations between regressors. For example, a group of regressors with positive regression coefficients and large positive inter-regressor correlations (e.g., 0.9) contributes more than twice as much to the variance (2) than another group of uncorrelated regressors with the same absolute magnitude of regression coefficients, according to the impact of the second summand. Thus, an importance metric based on dispersion must be expected to reflect this dependence on inter-regressor correlations, allocating higher importance to correlated regressors. If this is unwanted, one may go to great lengths to eliminate the correlation influences—for example, by introducing intricate data-dependent weights as in PMVD or a conditional permutation scheme as in the work of Strobl et al. 2008—or one might try to find an altogether different basis for variable importance allocation, for example, standardized coefficients in the linear model; unfortunately, there is no unbiased equivalent for these in the $p \gg n$ situation.

SUPPLEMENTAL MATERIAL

Simulation programs: A preparation program (utilityPrograms.R, R program code) loads all required packages and provides calculation routines that are used in the simulation. The simulation function is provided in a separate file (simulationProgram.R, R program code). After running the previous two programs, (callSimulationProgram.R, R program code) can be adapted to do the simulations that the user is interested in. Comments within the programs give further instructions. (programs.zip)

Dependence of variable importance in random forests on the shape of the regressor space: As pointed out in Section 6.4, the simulations have indicated that variable importance in random forests depends on the shape of the regressor space. This is further explained in this supplement. (Supplement_shape.pdf, Acrobat file)

[Received September 2008. Revised August 2009.]

REFERENCES

Achen, C. H. (1982), *Interpreting and Using Regression*, Beverly Hills, CA: Sage.

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

——— (2002), "Manual on Setting Up, Using, and Understanding Random Forests V3.1," unpublished manuscript, available at *http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf*.

——— (2003), "Manual on Setting Up, Using, and Understanding Random Forests V4.0," unpublished manuscript, available at *ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf*.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Budescu, D. V. (1993), "Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression," *Psychological Bulletin*, 114, 542–551.

Chevan, A., and Sutherland, M. (1991), "Hierarchical Partitioning," *The American Statistician*, 45, 90–96.

Christensen, R. (1992), Comment on "Hierarchical Partitioning," by A. Chevan and M. Sutherland, *The American Statistician*, 46, 74.

Darlington, R. B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.

Diaz-Uriarte, R., and Alvarez de Andrés, S. (2006), "Gene Selection and Classification of Microarray Data Using Random Forest," *BMC Bioinformatics*, 7, 3.

Ehrenberg, A. S. C. (1990), "The Unimportance of Relative Importance," *The American Statistician*, 44, 260.

Feldman, B. (2005), "Relative Importance and Value," unpublished manuscript, available at *http://www.prismanalytics.com/docs/RelativeImportance050319.pdf*.

Friedman, J. (1991), "Multivariate Additive Regression Splines," *The Annals of Statistics*, 19, 82–91.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008), "Random Forests: Some Methodological Insights," Research Report 6729, Institut National de Recherche en Informatique et en Automatique. ISSN 0249–6399.

Grömping, U. (2006), "Relative Importance for Linear Regression in R: The Package relaimpo," *Journal of Statistical Software*, 17, 1. Available at *http://www.jstatsoft.org/v17/i01/*.

——— (2007), "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition," *The American Statistician*, 61, 139–147.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2006a), "A Lego System for Conditional Inference," *The American Statistician*, 60, 257–263.

Hothorn, T., Hornik, K., and Zeileis, A. (2006b), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–674.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004), "Bagging Survival Trees," Statistics in Medicine, 23 (1), 77–91.

Ishwaran, H. (2007), "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The Annals of Applied Statistics*, 2, 841–860.

Johnson, J. W., and Lebreton, J. M. (2004), "History and Use of Relative Importance Indices in Organizational Research," *Organizational Research Methods*, 7, 238–257.

Kass, G. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29, 119–127.

Kruskal, W. (1987a), "Relative Importance by Averaging Over Orderings," *The American Statistician*, 41, 6–10.

——— (1987b), Correction to "Relative Importance by Averaging Over Orderings," *The American Statistician*, 41, 341.

Liaw, A., and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis*, Glenview, IL: Scott, Foresman.

Lipovetsky, S., and Conklin, M. (2001), "Analysis of Regression in Game Theory Approach," *Applied Stochastic Models in Business and Industry*, 17, 319–330.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org*, ISBN 3-900051-07-0.

Segal, M. R., Barbour, J. D., and Grant, R. M. (2004), "Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests," *Statistical Applications in Genetics and Molecular Biology*, 3, article 2.

Shih, Y.-S., and Tsai, H.-W. (2004), "Variable Selection Bias in Regression Trees With Constant Fits," *Computational Statistics and Data Analysis*, 45, 595–607.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9, 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25.

Stufken, J. (1992), "On Hierarchical Partitioning," *The American Statistician*, 46, 70–71.

Theil, H., and Chung, C.-F. (1988), "Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions," *The American Statistician*, 42, 249–252.

Therneau, T., and Atkinson, E. (1997), "An Introduction to Recursive Partitioning Using the rpart Routines," Technical Report 61, Mayo Clinic, Section of Statistics.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

van der Laan, M. (2006), "Statistical Inference for Variable Importance," *The International Journal of Biostatistics*, 2, 1008–1008.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320.