Conditional variable importance in R package `extendedForest`

Stephen J. Smith, Nick Ellis, C. Roland Pitcher

December 2, 2013

Contents

1 Introduction

The `gradientForest` package was developed to analyse large numbers of potential predictor variables by integrating the individual results from random forest analyses over a number of species. The random forests for each species were produced by the R package `extendedForest` consisting of modifications that we made to the original `randomForest` package [Liaw and Wiener, 2002]. One of the major modifications made to `randomForest` was to the method for calculating variable importance when two or more predictor variables were correlated.

Many of the predictor variables used in ecological studies are either naturally (e.g., decreasing temperatures with water depth) or functionally (e.g., benthic irradiance are calculated as a function of bottom depth and light attenuation) correlated. While some of these predictors may determine species distribution or abundance other collinear predictors may not.

The random subset approach for fitting predictor variables at each node could result in a correlated but less influential predictor standing in for more highly influential predictors in the early splits of an individual tree depending upon which predictor is selected in the subset. This tendency can be lessened by increasing the subsample size of predictors for each node but the trade-off would be an increase in correlation between trees in the forest with concurrent increase in generalization error and a decrease in accuracy [Breiman, 2001; see also Grömping, 2009].

Strobl et˜al. [2008] have also demonstrated that the permutation method for estimating variable importance exhibits a bias towards correlated predictor variables. The underlying reason for this behaviour has to do with the structure of the null hypothesis, i.e., independence between the response Y and the predictor X_j being permuted, implied by the importance measure. A small value for the importance measure would suggest that Y and X_j are independent but also assumes that X_j is independent of the other predictor variables Z in the model that were not permuted ($Z = X, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$). Correlation between X_j and Z will result in an

apparent increase in importance reflecting the lack of independence between X_j and Z instead of only reflecting the lack of independence between X_j and Y.

To remedy this situation, Strobl et˜al. [2008] proposed a conditional permutation approach where the values for X_j in the OOB sample for each tree are permuted within partitions of the values of the predictors in each tree that are correlated with X_j. Permutation importance is calculated by passing the OOB samples reconfigured with this permutation grid through each respective tree in the standard way.

In this document we present the results of a simulation study demonstrating the impact of correlation between predictor variables on determining variable importance and how the conditional permutation method implemented in `extendedForest` reduces this impact.

2 Methods

2.1 Conditional permutation

Our implementation of the method of Strobl et˜al. [2008] in `extendedForest` is based on the following. For predictor X_j determine all predictors, $X_i' (i = 1, \ldots, k; i \neq j \,\text{and}\, k \leq p-1)$ that are correlated with X_j above some threshold ρ^*. For tree t, find the first K split values, $s_1, \ldots, s_K$ and indices $v_i, \ldots, v_K$ on the X_i'. For each observation l in the OOB sample designate a grouping or partitioning label as,

$$g_l = \sum_{i=1}^{K} 2^{i-1} I(X'_{v_i} < s_i) \tag{1}$$

where $I(\cdot)$ is the indicator function taking value 1 if its argument is true and 0 otherwise.

Permutation of X_j in the OOB sample is applied within the above groups and calculation of the permutation importance measure proceeds as before. The grouping labels will take at most 2^K different values, although some combinations may be missing. K should be chosen not so large that there are too few observations per partition. We used a rule-of-thumb $K = \lfloor \log_2(0.368 N_a^{\text{sites}}/2) \rfloor$, which, if the sites were uniformly distributed among partitions, would ensure at least two points per partition.

2.2 Simulation Study

We used the simulation study design in Strobl et˜al. [2008] to demonstrate the difference between determining variable importance by conditional and marginal permutation. The response variable was set as a function of twelve predictor variables, i.e.,

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_{12} x_{i,12} + \varepsilon_i,$$

where $\varepsilon_i \sim \mathrm{N}(0, 0.5)$. The coefficients for the predictor variables were set so that only six of the twelve were influential.

Table 1: Regression coefficients for linear model used in simulation.

	Predictor variables											
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
β_j	5	5	2	0	-5	-5	-2	0	0	0	0	0

The correlation structure was introduced by setting the predictor variables as a sample from a multivariate normal distribution with a zero mean vector and covariance Σ. All predictors were

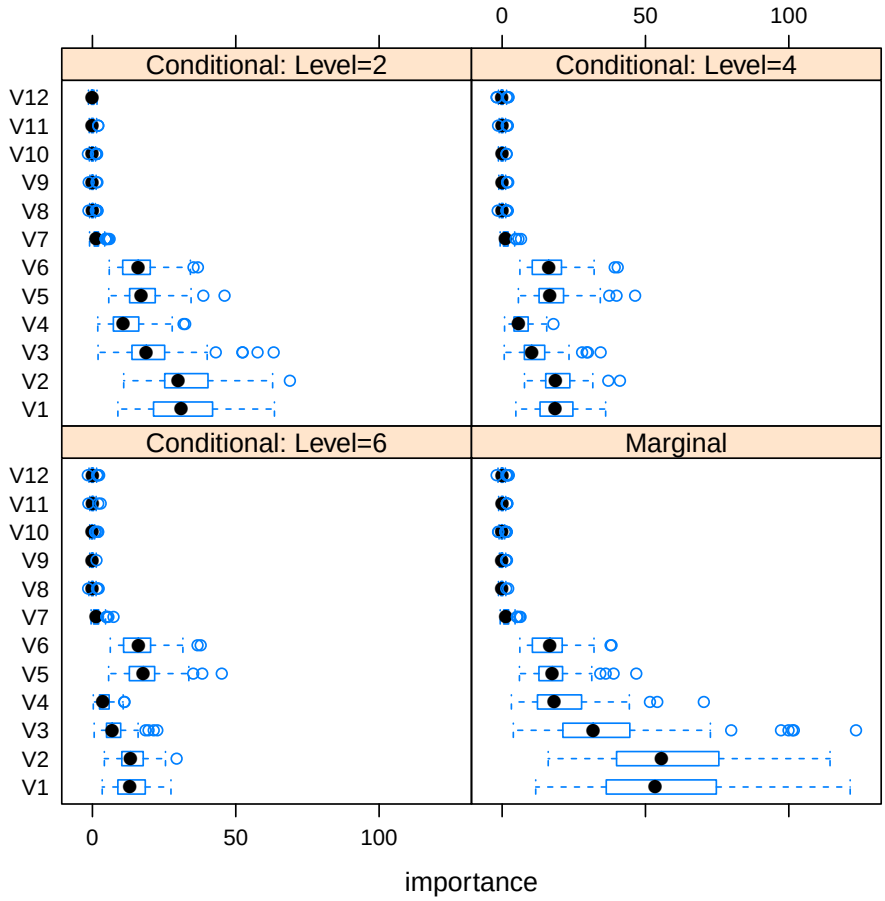
defined to have unit variance, $\sigma_{j,j} = 1$ and only the first four predictors were block-correlated with $\sigma_{j,j'} = 0.9$ ($i \neq j' \leq 4$). Off-diagonal elements were set to zero for the rest of the predictors. The R code used to run the simulation follows.

```
> require(extendedForest)
> require(MASS)
> #Set up covariance matrix
>
> Cov <- matrix(0,12,12)
> Cov[1:4,1:4] <- 0.9
> diag(Cov)[] <- 1
> #Coefficients for linear model
>
> beta <- c(5,5,2,0,-5,-5,-2,0,0,0,0,0)
> # Set the maximum number of partitions to compute the importance
> # from conditional permutation distribution of each variable
> maxK<-c(0,2,4,6)
> # Set the number of records (or sites) and the number of simulations.
> nsites<- 100
> nsim <- 100
> imp <- array(0,dim=c(12,4,nsim))
> #Simulation
>
> set.seed(222)
> for (sim in 1:nsim) {
+   X <- mvrnorm(nsites, rep(0,12), Sigma=Cov)
+   Y <- X%*%beta + rnorm(nsites,0,0.5)
+   df <- cbind(Y=Y,as.data.frame(X))
+   for (lev in 1:4) {
+     RF <- randomForest(Y ~ .,df, maxLevel=maxK[lev], importance=TRUE, ntree=500, corr.thre
+     imp[,lev,sim] <- RF$importance[,1]
+   }
+ }
> dimnames(imp) <- list(rownames(RF$importance), as.character(maxK), NULL)
> imp <- as.data.frame.table(imp)
```

3 Results

Marginal permutation identifies variables 1 and 2 as most important, followed by variables 3 to 7. even though variable 4 had no influence on the response, its correlation to variables 1 to 3 resulted in this variable being ranked as more important than variables 5 to 7.

```
> require(lattice)
> names(imp) <- c("var","maxK","sim","importance")
> print(bwplot(var ~ importance | ifelse(maxK=="0", "Marginal", paste("Conditional: Level",m
```



Variable importance obtained for differing number of partitions for the permutation grid are presented. Based on our rule-of-thumb the number of partitions should be set to 4 and there appears to be little difference between the results for $K = 4$ and $K = 6$. In both these cases the importance of variables 1 and 2 were very similar to those for variables 5 and 6 as expected from the linear model. Further, variable 4 is now just slightly ahead of variable 7 in importance. Apparently, this approach eliminates most but not all of effects of correlation. It is possible that increasing the number of partitions may reduce the importance of variable 4 even more (compare results from $K = 4$ and $K = 6$), however, at some point there will not be enough observations at each partition to calculate importance.

4 Session information

The simulation and output in this document were generated in the following computing environment.

- R version 3.0.2 Patched (2013-11-30 r64358), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils

- Other packages: MASS˜7.3-29, extendedForest˜1.6, lattice˜0.20-24
- Loaded via a namespace (and not attached): grid˜3.0.2, tools˜3.0.2

References

L.˜Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63:308–319, 2009. doi: 10.1198/tast.2009.08199.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

C.˜Strobl, A.L. Boulesteix, T.˜Kneib, T.˜Augustin, and A.˜Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.