Dependence of Variable Importance in Random Forests on the Shape of the Regressor Space

Supplement to “Variable Importance Assessment in Regression: Linear Regression Versus Random Forest”

By Ulrike Grömping
Beuth Hochschule für Technik Berlin – University of Applied Sciences, Germany

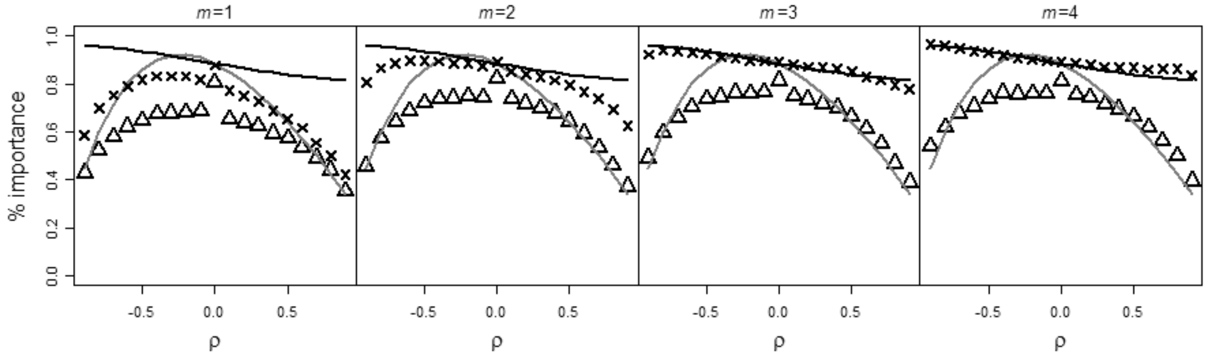

Figure: Averaged normalized importances for X_1 from 100 simulated datasets (simulation process described below) for m=1,2,3,4 (left to right) with $\boldsymbol{\beta}_1$=(4,1,1,0.3)^T, corr(X_j,X_k)=$\rho^{|j-k|}$ with ρ=–0.9 to 0.9 in steps of 0.1
Grey line: true normalized LMG allocation; Black line: true normalized PMVD allocation
△: Variable importance (% MSE Reduction) from RF-CART; ×: Variable importance (% MSE Reduction) from RF-CI

Let **C**(ρ) denote the 4x4 correlation matrix among the regressors, and let **R**(ρ)Λ(ρ)**R**^{T}(ρ) denote its eigen value decomposition, depending on the correlation parameter ρ. For the Figure above, simulated regressor variables with correlation matrix **C**(ρ) have been generated by

- first generating a vector of 4 uncorrelated uniform random variables
- and then pre-multiplying this vector with **R**(ρ)Λ(ρ)$^{1/2}$.

Response data have subsequently been generated according to a linear model with normal random error. Function **`randatunif`** in **`utilityPrograms.R`** does this.

It is striking in the above Figure that the forests’ variable importances show a discontinuous behavior at ρ=0, especially obvious with m=1. This discontinuity gave rise to the observation that variable importances depend on the shape of the regressor space, as will be pointed out below.

C(ρ) for ρ=0 is the 4-dimensional identity matrix; for this matrix, the eigen vectors are not uniquely determined. The depicted average variable importance for ρ=0 is based on unit eigen vectors, which implies that the regressors are a uniform sample from a 4-dimensional cube with sides parallel to the axes of the coordinate system. However, the eigen vectors of **C**(ρ) for $|\rho| \rightarrow 0$ converge against a different set of eigen vectors (direction arbitrary, order depending on the sign of ρ), which can be seen in function **`randatunif.rotcube`** in **`utilityPrograms.R`**. When multiplying the uncorrelated regressors from a unit cube with sides parallel to the axes with the above-mentioned limiting set of eigen vectors (cf. function **`randatunif.rotcube`**), the resulting regressors are still uncorrelated but come from a rotated cube. Simulating uncorrelated regressors according to this procedure leads to average normalized variable importances for ρ=0 that are exactly where expected when smoothly interpolating the triangles for non-zero ρ in the Figure above.

Thus, allocations to X_1 (strongest regressor) in the uncorrelated regressor case are higher, if cube sides are parallel to the axes of the coordinate system (and thus to split directions of trees) than if cube sides are rotated; this effect might be similar in nature to correlation.