TEST (2016) 25:254–260
DOI 10.1007/s11749-016-0485-3

DISCUSSION

Comments on: A random forest guided tour

Giles Hooker[1] · **Lucas Mentch**[2]

Published online: 19 April 2016

Abstract We discuss future challenges in developing statistical theory for Random Forests. In particular, we suggest that an analysis of bias and extrapolation is vital to understanding the statistical properties of variable importance measures. We further point to the incorporation of random forests within larger statistical models as an important tool for high-dimensional statistical inference.

Keywords Random forests · Machine learning · Extrapolation · Variable importance

Mathematics Subject Classification 62G09

We would like to congratulate Gérard Biau and Erwan Scornet on their timely, lucid and comprehensive overview of the established theoretical properties of Random Forests (RFs) and related estimators. We believe that the developments they outline represent the basis for the use of ensemble methods within formal statistical inference procedures.

Here we would like to point out what we see as the most important barriers to the widespread use of RF-like methods within statistical analysis; these particularly being an analysis of bias and its implications for measures of variable importance. Such an analysis has important real-world implications as a number of highly cited applied papers in a variety of fields have utilized importance scores for variable selection; see, for example, results in image classification (Bosch et al. 2007), ecology (Cutler

This comment refers to the invited paper available at: doi:10.1007/s11749-016-0481-7.

✉ Giles Hooker
gjh27@cornell.edu

[1] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA

[2] Department of Statistics , University of Pittsburgh, Pittsburgh, USA

et al. 2007), and land cover classification (Gislason et al. 2006) to name just a few. While RFs have performed very well at practical prediction tasks, much more theoretical development is needed to translate this powerful non-parametric tool into valid scientific understanding of the domains where it is applied.

1 Analysis of bias

As the authors outline, much of the recent advances in our understanding of RFs has been in the form of consistency, with more detailed results available for some simplified algorithms. These results are not easy and have relied on either modifying the RF procedure or restricting the class of models examined. While continuing to fill out this territory is important, we also want to advocate for studies into convergence rates and specifically into a characterization of bias for models closer to the original RF algorithm.

The central limit theorems derived in Mentch and Hooker (2015) and Wager and Athey (2015) center on the expected value of the RF, rather than on a generating function; that is, if $m_n(x)$ is an RF built with M_n proper subsamples of n data points, each of size a_n then

$$\sqrt{n}\frac{m_n(\mathbf{x}) - Em_n(\mathbf{x})}{\sqrt{\frac{K_n^2 M_n}{n}\zeta_{1,a_n}(\mathbf{x}) + \frac{1}{M_n}\zeta_{a_n,a_n}(\mathbf{x})}} \xrightarrow{d} N(0,1)$$

where the second term in the denominator provides for a correction that does not require $\lim_{n\to\infty} n/M_n = 0$. Equivalent multivariate limits can be produced for a vector of values $(m_n(\mathbf{x}_1), \ldots, m_n(\mathbf{x}_g))$. Crucially, these statements allow an assessment of variance, but not accuracy, and inference must be interpreted as being focused on the expectation of m_n.

For example, Mentch and Hooker (2014) use these result to test whether the function produced by RFs can be represented in terms of an additive model. In its most general form, this can be expressed as testing the hypothesis that

$$H_0 : Em_n(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_n(\mathbf{x}, \mathbf{z}) + g_n(\mathbf{y}, \mathbf{z}).$$

Such procedures include tests of variable importance (predictions do not change with respect to a variable of interest), strict additivity (by removing the variable $\mathbf{z}$) as well as more complex structures. Tests of this form are practically relevant both because they imply independence of an effect: *the effect of changing* $\mathbf{x}$ *does not depend on* $\mathbf{y}$, and because they provide an indication of which components to jointly examine when trying to interpret m_n in terms of visualisable effects (e.g., Hooker 2004).

These inferential questions are necessarily about $Em_n(\mathbf{x})$ unless the bias reduces at a faster than $\sqrt{n}$ rate. While these tests are still valuable, particularly for interpretation, the removal of this qualification would substantially broaden their applicability. We cannot expect convergence this fast (Stone 1980, provides a limiting result; Biau and Scornet also quote rates for some simplified algorithms) and thus ways to characterize, and potentially correct for, bias are particularly important. A heuristic correction

based on a residual bootstrap was proposed in Hooker and Mentch (2015) which also showed significant improvements in predictive accuracy. However, we do not expect the procedure to provide an improvement in rates.

A full study of bias should include not just rates, but a characterization of the dependence of bias on the covariate distribution. In particular, as we discuss in more detail below, the extrapolation behavior of RFs should be examined. Here we provide a heuristic claim that trees extrapolate as additive models constructed from marginal components—that is by averaging regressions made using subsets of the covariates. Specifically, we imagine a data set that lies close to a d-dimensional manifold $\mathcal{M}$ within $\mathbb{R}^p$. At any point far from $\mathcal{M}$, predictions will be obtained from leaves which all have to intersect $\mathcal{M}$. We expect leaves to involve splits of at most d variables, since splitting more that d variables will yield marginal information. The prediction in such a leaf is then a local average over the manifold, whichever d covariates are employed. The overall prediction is then given in terms of averages of each part of the manifold that can be reached from our prediction point by moves along d axes.

This heuristic is demonstrated in Fig. 1. Here we have simulated a bivariate set of covariate values of size 200 with uniform marginals and a Gaussian copula with correlation parameter 0.9. We then defined a noiseless response to be $Y = X_1$ and obtained an RF using these data. In Fig. 1 we examine the match between these data and the true model along with a heuristic model given by the average of marginal linear regressions on X_1 and X_2, estimated separately. Here, our RF model gives responses close to the generating model near the data, but increasingly mimics the average of marginals model in regions far from the observations. The lower left panel overlays the splits from the first six trees in the RF. We see splits extending from the data to the edge of covariate space, but using one or other axes depending on the tree, giving credence to our heuristic expectation. Close to the data, splits are much shorter, appropriately mimicking the response.

This explanation is very much heuristic and must be modified to account for several factors: this RF splits more frequently on X_1 than X_2, for example, meaning the approximation away from the data is best described by a weighted average of marginal regressions, and we do not have a description of behavior in areas with low (as opposed to no) data density. Nonetheless, we believe this simple example highlights an important need for further study; as we will show below, these behaviors also have consequences for using RF models for understanding and inference.

2 Implications for variable importance

A crucial implication of the observations above is in understanding the effect of the covariate distribution on various measures of variable importance. While diagnostic tools have generally been considered secondary to predictive performance, measures like variable importance have been a significant factor in the popularity of RFs and have been used in a diverse array of fields such as image classification (Bosch et al. 2007), ecology (Cutler et al. 2007), and drug efficacy (Gunther et al. 2003). Nonetheless, a body of literature has demonstrated that these measures can be misleading in the presence of correlated covariates.

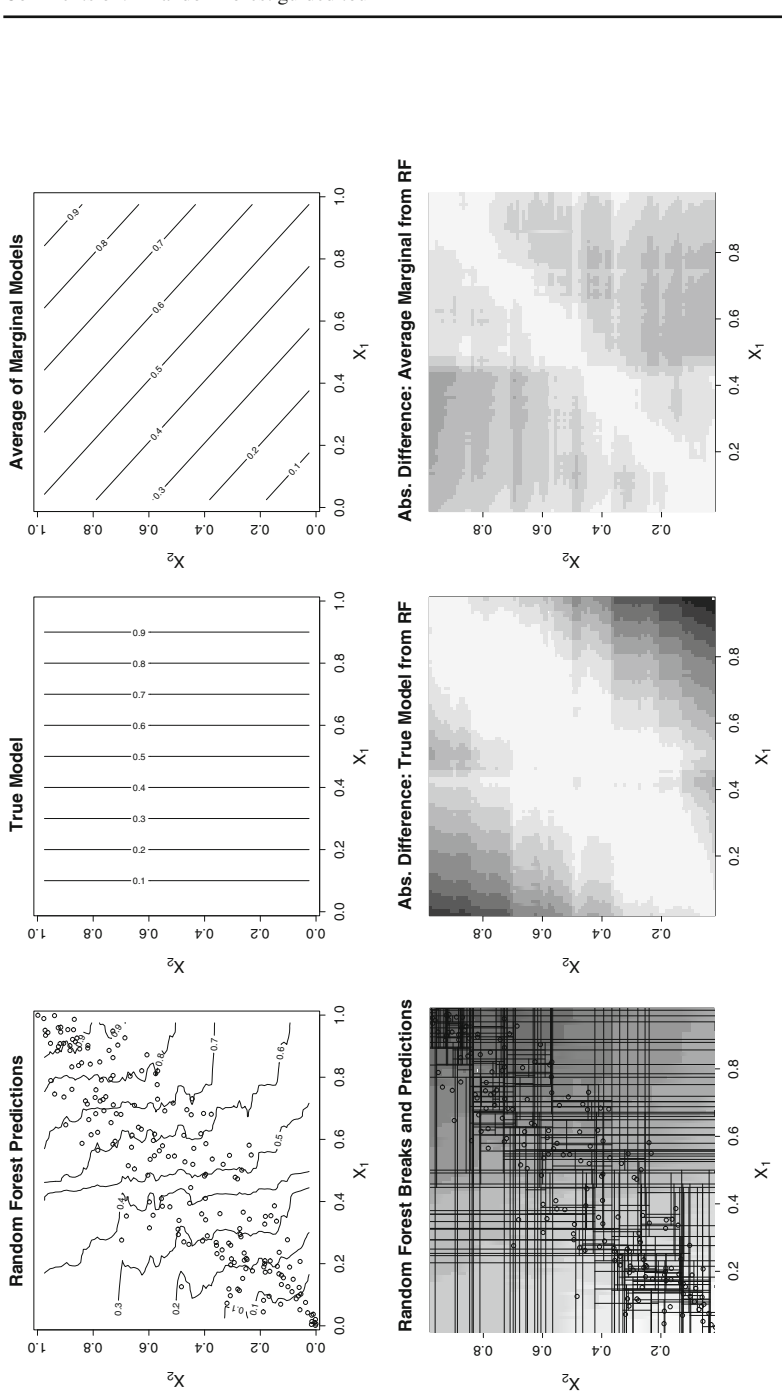

Fig. 1 Results from an experiment in extrapolation for RFs. *Top row* from *left*: contours of RF predictions when learned using covariate values given by *circles*; true values of the response given to the RF; predictions from a model given by averaging marginal regressions or the response on each covariate separately. *Bottom row* from *left*: RF predictions with splits from six example trees overlaid to show extrapolation behavior; absolute difference between RF predictions and the true model; absolute differences between RF predictions and the average of marginals model. *Greyscale* intensities in the two absolute difference plots are scaled to be directly comparable

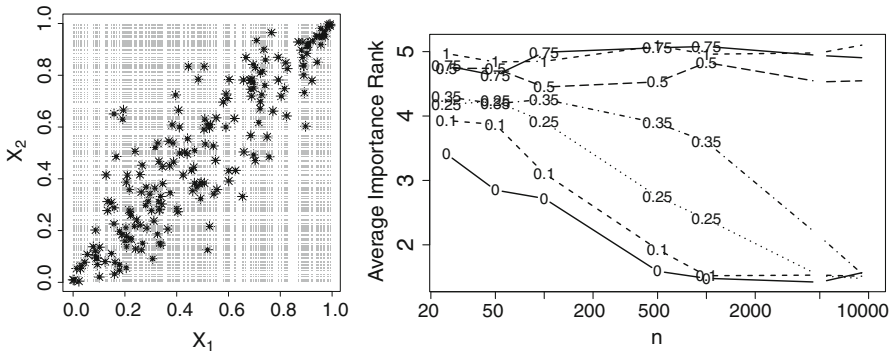

Fig. 2 *Left* a set of bivariate covariates (*asterisk*) and the corresponding points used to calculate MDA. *Right* results of an experiment in variable importance measures. We provide the average rank of the first out of seven independent covariates by sample size (x-axis) and a Gaussian copula correlation parameter (labels) between the first and second covariates

We believe that the correct explanation for this can be found in Hooker (2007); that measures such as Mean Decrease in Accuracy (MDA) unduly rely on the behavior of RFs as they extrapolate. This is because the variable importance of X_1, say, is based on value of the function measured by permuting its values; meaning we average predictive accuracy over a covariate distribution given by the product of the marginal distributions of X_1 and X_{-1}. This is illustrated in the left panel of Fig. 2 where we have plotted the measurement points used for variable importance scores from our example above, along with the original data distribution.

This explanation is in contrast to that given in Strobl et al. (2008) which identifies variable importance as measuring the marginal importance of a variable. If this were the case, we would expect to see consistently poor variable importance even with large data. A simple experiment suffices to demonstrate the differences between these explanations: we generated seven dimensional covariates from a uniform distribution with correlation induced in the first two dimensions via a Gaussian copula and calculated a response from the model

$$Y = 0.8X_1 + 0.8X_2 + \sum_{j=3}^{7} X_j$$

with no noise process. Here, X_1 and X_2 should be registered as less important than the remaining variables, but correlation between them will induce larger perceived marginal effects and also produce the extrapolation described above. The right panel of Fig. 2 provides the average rank (higher = more important) in MDA of X_1 over 100 simulations at a variety of data sizes and correlation parameters. Here we observe that at high correlations, X_1 tends to be erroneously awarded one of the highest importances of all seven covariates. At lower correlations, however, while the average rank of X_1's MDA starts higher than the average covariate, it decreases as the data size grows. We contend that this behavior is a consequence of larger sample sizes "filling out" the covariate space, leading to lower extrapolation.

Hooker (2007) proposed a weighting scheme to correct for this behavior, based on the density of covariates. However, a fuller analysis of bias will provide better insight into which values of the covariates may be reasonably used to provide inference or insight. We note that formal hypothesis tests proposed in Mentch and Hooker (2014) can readily accommodate weighting query points for RFs. We expect that an improved understanding of extrapolation will play an important role in providing valid statistical procedures. We also note that the conditional variable importance measures in Strobl et al. (2008) would still be expected to be effective here because they avoid extrapolation, but will not allow tests of more complex structure to be carried out.

3 Extensions to boosting and semiparametric regression

A further important direction of theoretical research is the incorporation of RF methods into larger statistical models. This has the potential both to extend the class of problems that can be addressed via RF-type methods as well as to avoid some of the issues of bias discussed above. An initial result in Wager and Athey (2015) shows nicely that RFs, when used to model nuisance components can still yield valid (i.e., asymptotically unbiased) inference about other parts of a model obtained from RF residuals. This is commonly found in partially linear models: including a non-parametric term in a model still allows inference about parametric components of the model; see Li and Racine (2007), for example. We speculate that equivalent results can be obtained for models of the form

$$y = \mathbf{x}\beta + f(\mathbf{z}) + \epsilon$$

where $f(\mathbf{z})$ is modeled by an ensemble of decision trees that use covariates z.

Of course, obtaining these types of model requires more radical changes to RF fitting algorithms. Such models can be fit using boosting methods (see Fahrmeir et al. 2013, for parametric models) or via backfitting schemes (Sorokina et al. 2007; Lou et al. 2013). These, of course, will all require new theoretical development, for consistency, rates, and central limit theorems.

Acknowledgments This work was supported by NSF grants DMS-103252 and DEB-1353039.

References

Bosch A, Zisserman A, Muoz X (2007) Image classification using random forests and ferns. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007. IEEE. pp 1–8

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792

Fahrmeir L, Kneib T, Lang S, Marx B (2013) Regression: models, methods and applications. Springer, Berlin Heidelberg

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pat Recogn Lett 27(4):294–300

Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP (2003) Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. Proc Natl Acad Sci 100(16):9608–9613

Hooker G (2004) Variable interaction networks. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Hooker G (2007) Generalized functional ANOVA diagnostics for high dimensional functions of dependent variables. J Comput Graph Stat 16:709–732

Hooker G, Mentch L (2015) Bootstrap bias corrections for ensemble methods. arXiv preprint arXiv:1506.00553

Li Q, Racine JS (2007) Nonparametric econometrics. Princeton University Press, Princeton

Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. In: Proceedings of the Ninteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Mentch L, Hooker G (2014) Detecting feature interactions in bagged trees and random forests. ArXiv e-prints

Mentch L, Hooker G (2015) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. J Mach Learn Res (In press)

Sorokina D, Caruana R, Riedewald M (2007) Additive groves of regression trees. In: Proceedings of the 18th European Conference on Machine Learning (ECML'07)

Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. Ann Stat 1348–1360

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

Wager S, Athey S (2015) Estimation and inference of heterogeneous treatment effects using random forests. arXiv preprint arXiv:1510.04342