Noname manuscript No.
(will be inserted by the editor)

Rejoinder on: A Random Forest Guided Tour

Gérard Biau · Erwan Scornet

Received: date / Accepted: date

First of all, we would like to thank S. Arlot, P. Bühlmann, R. Genuer, P. Geurts, G. Hooker, F. Leonardi, L. Mentch, S. Wager and L. Wehenkel for their insightful and stimulating comments on our review paper, as well as for their thorough investigation. We also thank the Editors-in-Chief for letting us the opportunity to comment the issues raised in the discussions.

The comments all underline the importance of random forests and of the connected topic of variable selection. Of special interest is the diversity of perspectives, which include theoretical, practical, and computational issues. To summarize, there are five main points in the discussions that are quite recurrent:

(*i*) How can the results on "simplified" random forests be used to gain access to the complex machinery of Breiman's forests?
(*ii*) Do the existing results on Breiman's forests extend to the non-i.i.d. setting?

Gérard Biau
Sorbonne Universités, UPMC Univ Paris 06, CNRS, Laboratoire de Statistique Théorique et Appliquée (LSTA), boîte 158, 4 place Jussieu, 75005 Paris, France
Institut universitaire de France
Tel.: +331 44 27 85 63
Fax: +331 44 27 33 42
E-mail: gerard.biau@upmc.fr

Erwan Scornet
Sorbonne Universités, UPMC Univ Paris 06, CNRS, Laboratoire de Statistique Théorique et Appliquée (LSTA), boîte 158, 4 place Jussieu, 75005 Paris, France
Tel.: +331 44 27 85 62
Fax: +331 44 27 46 70
E-mail: erwan.scornet@upmc.fr

(*iii*) What is the best randomization scheme? (feature selection at each node? at the beginning of tree construction?)
(*iv*) How does the correlation between features impact the forest procedure and the variable importance?
(*v*) Which splitting criterion is the most adapted to a given learning task?

It is unfortunately not possible to address all these exciting issues within the confines of this rejoinder. In effect, each of them is a research area in its own, and they all together define an ambitious multi-year research program. We would like instead to add a sixth item to the list above, regarding the out-of-bag (OOB) error estimate properties (which is defined on page 10 of the manuscript).

Consider a forest in the classification regime, where each pair $(\mathbf{X}_i, Y_i)$ takes its values in, say, $[0,1]^d \times \{0,1\}$ and $n \geq 2$. Assuming that the resampling prior to the j-th tree construction is done with bootstrap (so, $a_n = n$ and replacement is allowed), we end up with two data sets: the original observations $\mathscr{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$, and the bootstrapped data set $\mathscr{D}_n^{(j)} = ((\mathbf{X}_1^{(j)}, Y_1^{(j)}), \ldots, (\mathbf{X}_n^{(j)}, Y_n^{(j)}))$, with possible repetitions. In the notation of the article, the j-th tree classifier is $m_n(\mathbf{X}; \Theta_j, \mathscr{D}_n)$. In the sequel we set $\Theta^{(j)} \equiv \Theta_j$ and use the more explicit notation $m_n(\mathbf{X}; \Theta^{(j)}, \mathscr{D}_n^{(j)})$, which highlights the fact that the tree is grown with the resampled data $\mathscr{D}_n^{(j)}$.

The OOB error estimate is defined as follows. For any observation $\mathbf{X}_i$, let

$$B_n^{(i)} = \left\{ j \in \{1, \ldots, M\} : (\mathbf{X}_i, Y_i) \notin \mathscr{D}_n^{(j)} \right\}$$

be the set of indices j such that the j-th tree does not use $\mathbf{X}_i$ in its construction (i.e., $\mathbf{X}_i$ is not selected in the j-th bootstrap step). Accordingly, let $m_{M,n}^{\mathrm{OOB}}(\mathbf{X}_i; \mathscr{D}_n)$ be the majority vote among trees that do not use $\mathbf{X}_i$ in their construction, that is

$$m_{M,n}^{\mathrm{OOB}}(\mathbf{X}_i; \mathscr{D}_n) = \begin{cases} 1 & \text{if } \frac{1}{|B_n^{(i)}|} \sum_{j \in B_n^{(i)}} m_n(\mathbf{X}_i; \Theta^{(j)}, \mathscr{D}_n^{(j)}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Then, the OOB error estimate is but the error of the $m_{M,n}^{\mathrm{OOB}}(\mathbf{X}_i; \mathscr{D}_n)$ averaged over all $\mathbf{X}_i$:

$$\hat{L}_{M,n}^{\mathrm{OOB}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{m_{M,n}^{\mathrm{OOB}}(\mathbf{X}_i; \mathscr{D}_n) \neq Y_i}.$$

Then, if $m_{M,n}(\mathbf{X}; \Theta_1, \ldots, \Theta_M, \mathscr{D}_n)$ denotes Breiman's random forest, an interesting open problem is to compare $\hat{L}_{M,n}^{\mathrm{OOB}}$ with the natural target

$$L_{M,n} = \mathbb{P}[m_{M,n}(\mathbf{X}; \Theta_1, \ldots, \Theta_M, \mathscr{D}_n) \neq Y \mid \mathscr{D}_n],$$

where the probability is taken with respect to both $(\mathbf{X}, Y)$ and $\Theta_1, \ldots, \Theta_M$. The random quantity $L_{M,n}$ measures the effectiveness of the forest, and since it cannot be computed, the immediate need of the statistician is to estimate it by $\hat{L}_{M,n}^{\mathrm{OOB}}$ as accurately as possible. Therefore, the challenge that we put on the sixth position of the list is the following one:

(vi) Derive an exponential inequality for $\mathbb{P}[|\hat{L}^{\text{OOB}}_{M,n} - L_{M,n}| \geq \varepsilon]$.

Let us now describe another stimulating challenge. Letting

$$A_n^{(j)} = \big\{ i \in \{1, \ldots, n\} : (\mathbf{X}_i, Y_i) \notin \mathscr{D}_n^{(j)} \big\}$$

and denoting by $\alpha_n^{(j)}$ its cardinality, we may define the j-th *individual* OOB error estimate as

$$\hat{L}_n^{(j)} = \mathbb{1}_{\alpha_n^{(j)} \neq 0} \times \frac{1}{\alpha_n^{(j)}} \sum_{i \in A_n^{(j)}} \mathbb{1}_{m_n(\mathbf{X}_i; \Theta^{(j)}, \mathscr{D}_n^{(j)}) \neq Y_i}.$$

$\hat{L}_n^{(j)}$ is the estimation of the j-th tree error evaluated over the data that are left out by the j-th bootstrap. We note that the event $[\alpha_n^{(j)} = 0]$ has probability $n!/n^n$, which by Stirling's approximation behaves as $\sqrt{2\pi n} e^{-n}$ as $n \to \infty$.

Logically, the *global* OOB error estimate is the average of the individual error estimates. Thus, for a forest with M trees, we have

$$\hat{L}_{M,n} = \frac{1}{M} \sum_{j=1}^{M} \hat{L}_n^{(j)}.$$

Let

$$L_n = \mathbb{P}[m_n(\mathbf{X}; \Theta, \mathscr{D}_n) \neq Y \,|\, \mathscr{D}_n]$$

be the error of a random tree. The following lemma is proved at the end of the discussion.

Lemma 1 *For all $\varepsilon > 0$ and $n \geq 2$,*

$$\mathbb{P}\big[|\hat{L}_{M,n} - L_n| \geq \varepsilon\big] \leq 2Me^{-n \min\left(\frac{\varepsilon^2}{160}, \frac{2}{25}\right)} + Me\sqrt{n}e^{-n} + 2e^{-M\varepsilon^2/2}.$$

In particular, with the choice $M = \lceil n/80 \rceil$, regardless of the distribution of $(\mathbf{X}, Y)$,

$$\mathbb{P}\big[|\hat{L}_{M,n} - L_n| \geq \varepsilon\big] \leq \left(\frac{n}{40} + 4 + e\right) e^{-n \min\left(\frac{\varepsilon^2}{160}, \frac{2}{25}\right)}.$$

Lemma 1 shows that $\hat{L}_{M,n}$ and L_n are asymptotically exponentially close, provided $M = \lceil n/80 \rceil$. This distribution-free result is not surprising since, given the data set $\mathscr{D}_n$, L_n is but the error of a single random tree averaged over the randomization parameter Θ.

Observe that Θ is of the form $\Theta = (\Theta_1, \Theta_2)$, where Θ_1 describes the bootstrap subset selection prior to the tree growing, and Θ_2 encapsulates the random feature selection in action at the nodes of the tree. So, for each tree of the forest, $\Theta_1^{(j)}$ chooses with replacement n items within the list $\{1, \ldots, n\}$ and the j-th tree is grown with the bootstrapped data subset $\mathscr{D}_n^{(j)}$. On the other hand, if this tree were to be grown with the original data set $\mathscr{D}_n$ instead of $\mathscr{D}_n^{(j)}$—that is, if we used *all* available data—then we would measure its prediction performance via the criterion

$$\bar{L}_n = \mathbb{P}[m_n(\mathbf{X};\Theta_2,\mathscr{D}_n) \neq Y \,|\, \mathscr{D}_n].$$

The second challenge that we pose is as follows:

$(vi)'$ Derive an exponential inequality for $\mathbb{P}[|\hat{L}_{M,n} - \bar{L}_n| \geq \varepsilon]$.

Put differently, we would like to know under which conditions on the distribution of $(\mathbf{X}, Y)$ the global OOB error estimation process is smart enough to accurately estimate the average error of a tree grown with the whole data set, without any prior bootstrap randomization. A possible route to follow is to note that each tree of the forest is a Layered Nearest Neighbor estimate (LNN, see page 13 of the article) and adapt stability arguments given by Devroye and Wagner (1969) for the holdout estimate of the classification error of k-local rules. We believe however that the analysis is more involved in the case of forests, since the tree rests upon the highly nonlocal CART program.

Appendix: Proof of Lemma 1. Let, for fixed j,

$$L_n^{(j)} = \mathbb{P}[m_n(\mathbf{X};\Theta^{(j)},\mathscr{D}_n^{(j)}) \neq Y \,|\, \Theta^{(j)},\mathscr{D}_n]$$

and observe that

$$\mathbb{E}[\hat{L}_n^{(j)} \,|\, \Theta^{(j)},\mathscr{D}_n^{(j)}] = \mathbb{1}_{\alpha_n^{(j)}\neq 0} L_n^{(j)}. \tag{1}$$

Also notice that

$$\begin{aligned}
&\mathbb{P}\big[|\hat{L}_n^{(j)} - L_n^{(j)}| \geq \varepsilon\big] \\
&\quad\leq \mathbb{P}\big[|\hat{L}_n^{(j)} - \mathbb{1}_{\alpha_n^{(j)}\neq 0} L_n^{(j)}| \geq \varepsilon/2\big] + \mathbb{P}\big[|\mathbb{1}_{\alpha_n^{(j)}= 0} L_n^{(j)}| \geq \varepsilon/2\big] \\
&\quad\leq \mathbb{P}\big[|\hat{L}_n^{(j)} - \mathbb{1}_{\alpha_n^{(j)}\neq 0} L_n^{(j)}| \geq \varepsilon/2\big] + \mathbb{P}[\alpha_n^{(j)} = 0] \\
&\quad= \mathbb{P}\big[|\hat{L}_n^{(j)} - \mathbb{1}_{\alpha_n^{(j)}\neq 0} L_n^{(j)}| \geq \varepsilon/2\big] + n!/n^n.
\end{aligned}$$

Therefore, using (1) and Hoeffding's inequality (Hoeffding, 1963), we obtain

$$\begin{aligned}
&\mathbb{P}\big[|\hat{L}_n^{(j)} - L_n^{(j)}| \geq \varepsilon\big] \\
&\leq \mathbb{E}\Big[\mathbb{P}\Big[\big|\hat{L}_n^{(j)} - \mathbb{E}[\hat{L}_n^{(j)} \,|\, \Theta^{(j)},\mathscr{D}_n^{(j)}]\big| \geq \varepsilon/2 \,\Big|\, \Theta^{(j)},\mathscr{D}_n^{(j)}\Big]\Big] + n!/n^n \\
&\leq 2\mathbb{E}e^{-\alpha_n^{(j)}\varepsilon^2/2} + n!/n^n.
\end{aligned}$$

A second application of (one-sided) Hoeffding's inequality shows that for $t > 0$, with probability larger than $1 - e^{-2t^2/n}$, $\alpha_n^{(j)} > -t + n(1-1/n)^n$. Thus, for all $t > 0$,

$$\begin{aligned}
\mathbb{E}e^{-\alpha_n^{(j)}\varepsilon^2/2} &\leq e^{(t\varepsilon^2 - n(1-1/n)^n\varepsilon^2)/2} + e^{-2t^2/n} \\
&\leq e^{-n\varepsilon^2(-t/n+1/4)/2} + e^{-2t^2/n},
\end{aligned}$$

since $(1-1/n)^n \geq 1/4$ for all $n \geq 2$. The choice $t = n/5$ yields

$$\mathbb{E}e^{-\alpha_n^{(j)}\varepsilon^2/2} \leq 2e^{-n\min\left(\frac{\varepsilon^2}{40},\frac{2}{25}\right)}.$$

Putting all the pieces together, we conclude that, for all $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\big[|\hat{L}_{M,n} - L_n| \geq \varepsilon\big] \\
&\quad\leq \mathbb{P}\Big[\frac{1}{M}\sum_{j=1}^{M}|\hat{L}_n^{(j)} - L_n^{(j)}| \geq \varepsilon/2\Big] + \mathbb{P}\Big[\Big|\frac{1}{M}\sum_{j=1}^{M}L_n^{(j)} - L_n\Big| \geq \varepsilon/2\Big] \\
&\quad\leq M\Big(2e^{-n\min\left(\frac{\varepsilon^2}{160},\frac{2}{25}\right)} + n!/n^n\Big) + \mathbb{P}\Big[\Big|\frac{1}{M}\sum_{j=1}^{M}L_n^{(j)} - L_n\Big| \geq \varepsilon/2\Big].
\end{aligned}$$

By Hoeffding's inequality, the last term is upper bounded by $2e^{-M\varepsilon^2/2}$. Thus,

$$\mathbb{P}\big[|\hat{L}_{M,n} - L_n| \geq \varepsilon\big] \leq 2Me^{-n\min\left(\frac{\varepsilon^2}{160},\frac{2}{25}\right)} + Me\sqrt{n}e^{-n} + 2e^{-M\varepsilon^2/2}.$$

Finally, letting $M = \lceil n/80 \rceil$, we have

$$\mathbb{P}\big[|\hat{L}_{M,n} - L_n| \geq \varepsilon\big] \leq \left(\frac{n}{40} + 4 + e\right)e^{-n\min\left(\frac{\varepsilon^2}{160},\frac{2}{25}\right)}.$$

References

L. Devroye and T.J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 15: 525–531, 1969.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.