Supplementary Material: A Characterization of Mean Squared Error for Estimator with Bagging

1 Proof of Theorem 2.1

Recall that the estimator with bagging is given by:

$$\tilde{\theta}(L, B) = \frac{\sum_{k=1}^{N} \hat{\theta}(L_{U^k})}{N},$$

where $B = (U^1, .., U^N)$ is a set of N sampling with replacement of the dataset $L = (X_1, ..., X_n)$, taken independently and uniformly from $\mathcal{U} = \{1, .., n\}^{\{1,..,m\}} = (u_k)_{k=1..n^m}$. Since we assumed that $\forall i \in [|1, n^m|]$, $\mathbb{E}_L(\hat{\theta}(L_{u^i})^2) < \infty$ and B living in a finite set, $\tilde{\theta}$ is well defined and the Fubini's theorem is always available for interchange between different expectations.

Lemma 1.1 *Let U be a random variable uniformly distributed over $\mathcal{U}$. Then the expected value of the bagged estimator $\tilde{\theta}$ does not depend on N and is equal to $\mathbb{E}_L\mathbb{E}_U(\hat{\theta}(L_U))$.*

Proof. For N independent identically distributed bagged samples represented by $U^1, ...U^N \in \{1, .., n\}^m$, it follows by linearity of the expectation operator and Fubini's theorem that:

$$\begin{aligned}
\mathbb{E}_{(L,B)}(\tilde{\theta}(L, B)) &= \mathbb{E}_L\left(\mathbb{E}_B\left(\frac{1}{N}\sum_{i=1}^{N} \hat{\theta}(L_{U^i})\right)\right) \\
&= \mathbb{E}_L\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{U^i}\left(\hat{\theta}(L_{U^i})\right)\right) \\
&= \mathbb{E}_L\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{U}\left(\hat{\theta}(L_{U})\right)\right) \\
&= \mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))).
\end{aligned}$$

The proof is completed.

Now we turn to the variance. Previous lemma 1.1 computes the expectations, therefore we only need to compute the second moment.

Lemma 1.2 *The expected value of the bagged estimator satisfies:*

$$\mathbb{E}_{(L,B)}(\tilde{\theta}^2) = \mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))^2) + \frac{1}{N}\mathbb{E}_L(\mathrm{Var}_U(\hat{\theta}(L_U))),$$

where U is a random variable uniformly distributed over $\mathcal{U}$ and

$$\mathrm{Var}_U(\hat{\theta}(L_U))) = \mathbb{E}_U(\hat{\theta}(L_U)^2) - \mathbb{E}_U(\hat{\theta}(L_U))^2.$$

Proof. Remark that:

$$\mathbb{E}_{(L,B)}(\tilde{\theta}^2) = \mathbb{E}_L[\mathbb{E}_B(\tilde{\theta}^2)],$$

and

$$\mathbb{E}_B(\tilde{\theta}^2) = \mathbb{E}_B\left(\frac{1}{N}\sum_{i=1}^{N} \hat{\theta}(L_{U^i})\right)^2.$$

Since U^i is uniformly distributed on $\mathcal{U}$, $\mathbb{E}_B(\tilde{\theta}^2)$ is a symmetric polynomial with respect to $\{\hat{\theta}(L_{u_i})\}_{u_i \in \mathcal{U}}$. Thus there exists two constants C_1 and C_2 such that for any function $\hat{\theta}$,

$$\mathbb{E}_B(\tilde{\theta}^2) = C_1\left(\sum_{i=1}^{n^m} \hat{\theta}(L_{u_i})^2\right) + C_2\left(\sum_{(i,j)\ i \neq j} \hat{\theta}(L_{u_i})\hat{\theta}(L_{u_j})\right). \tag{1}$$

Let $\hat{\theta} = \hat{\theta}_1 \equiv 1$, it holds

$$1 = C_1 n^m + C_2 n^m(n^m - 1). \tag{2}$$

Now define $\hat{\theta} = \hat{\theta}_2$, such that $\hat{\theta}_2(L_{u_1}) = 1$ and $\hat{\theta}_2(L_{u_i}) = 0$ for all $i \neq 1$. The right hand side of equation (1) equals to C_1 and we need to compute the value of left hand side. Observe that

$$\tilde{\theta}_2 = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_2(L_{U^i}),$$

where U^i is uniformly and independently distribute on $\mathcal{U}$. Thus $\{\hat{\theta}_2(L_{U^i})\}_i$ is a family of independent Bernoulli with parameter $1/n^m$. The second moment of $N\tilde{\theta}_2$ is thus binomial distributed with parameter $(N, 1/n^m)$. Recall that the second moment of a binomial distribution with parameter (N, p) equals to $N(N-1)p^2 + Np$. Therefore:

$$\begin{aligned}
\mathbb{E}_B(\tilde{\theta}_2^2) =& \frac{1}{N^2}\left(\mathbb{E}_B(N\tilde{\theta}_2)^2\right) \\
=& \frac{1}{N^2}\left(N(N-1)n^{-2m} + Nn^{-m}\right).
\end{aligned}$$

Simplifying equation (1) we have:

$$C_1 = \frac{N-1}{N}n^{-2m} + \frac{1}{N}n^{-m}.$$

Together with (2), we deduce :

$$C_2 = \frac{N-1}{N} n^{-2m}.$$

Hence,

$$\begin{aligned}
&\mathbb{E}_B(\tilde{\theta}^2) \\
=&\frac{N-1}{N} n^{-2m} \left(\sum_i \hat{\theta}(L_{u_i})^2 + \sum_{i \neq j} \hat{\theta}(L_{u_i})\hat{\theta}(L_{u_j}) \right) \\
&+ \frac{1}{N} n^{-m} \sum_i \hat{\theta}(L_{u_i})^2 \\
=&\frac{N-1}{N} \mathbb{E}_U(\hat{\theta}(L_U))^2 + \frac{1}{N}\mathbb{E}_U(\hat{\theta}(L_U)^2) \\
=&\mathbb{E}_U(\hat{\theta}(L_U))^2 + \frac{1}{N}\mathrm{Var}_U(\hat{\theta}(L_U)).
\end{aligned}$$

The proof is completed.

Now we are in position to prove theorem:

Proof of theorem. According to lemma 1.1 and 1.2, the variance of the bagging estimator is:

$$\begin{aligned}
&\mathrm{Var}_{(L,B)}(\tilde{\theta}) \\
=&\mathbb{E}_{(L,B)}(\tilde{\theta}^2) - \mathbb{E}_{(L,B)}(\tilde{\theta})^2 \\
=&\mathbb{E}_L \left(\mathbb{E}_U(\hat{\theta}(L_U))^2 \right) \\
&+ \frac{1}{N}\mathbb{E}_L(\mathrm{Var}_U(\hat{\theta}(L_U))) - \mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))^2 \\
=&\frac{1}{N}\mathbb{E}_L(\mathrm{Var}_U(\hat{\theta}(L_U))) + \mathrm{Var}_L(\mathbb{E}_U(\hat{\theta}(L_U))).
\end{aligned}$$

Hence, the MSE of $\tilde{\theta}$ is :

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}) &= \mathrm{Var}(\tilde{\theta}) + \mathrm{Bias}(\tilde{\theta}) \\
&= \frac{1}{N}\mathbb{E}_L\big(\mathrm{Var}_U(\hat{\theta}(L_U))\big) + \mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}(L_U))\big) \\
&\qquad + \big(\mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))) - \theta\big)^2.
\end{aligned}$$

2 Proof of Theorem 3.1

We recall that $\hat{v}(L)$ can be rewritten as:

$$\hat{v}(L) = \frac{1}{n(n-1)} \sum_{i<j} (X_i - X_j)^2. \tag{3}$$

We begin with some lemmas.

Lemma 2.1 *Let X be a random variable with finite forth moment satisfying $\mathbb{E}(X) = 0$. Let $X_1, ..., X_n$ be independent copies of X and $L = (X_1, ..., X_n)$. Denote i, j, k, l four distinct index of $\{1, ..., n\}$. Then the following statements hold.*

1. $\mathbb{E}\left((X_i - X_j)^4\right) = 6\mu_2^2 + 2\mu_4$,
2. $\mathbb{E}\left((X_i - X_j)^2(X_i - X_k)^2\right) = 3\mu_2^2 + \mu_4$,
3. $\mathbb{E}\left((X_i - X_j)^2(X_k - X_l)^2\right) = 4\mu_2^2$,

where $\mu_4 := \mathbb{E}(X^4)$ is the forth moment of X. Now denote

$$P = \sum_{i,j} (X_i - X_j)^4,$$

$$Q = \sum_{i,j,k} (X_i - X_j)^2 (X_i - X_k)^2,$$

and

$$R = \sum_{i,j,k,l} (X_i - X_j)^2 (X_k - X_l)^2.$$

Then the following equations hold:

4. $\mathbb{E}_L(P) = 3n(n-1)\mu_2^2 + n(n-1)\mu_4$,
5. $\mathbb{E}_L(Q) = \frac{3}{2}n(n-1)(n-2)\mu_2^2 + \frac{1}{2}n(n-1)(n-2)\mu_4$,
6. $\mathbb{E}_L(R) = \frac{1}{2}n(n-1)(n-2)(n-3)\mu_2^2$.

Proof. The proof of the three first items is immediate by independence of X_i, X_j, X_k, X_l.

Proof of item 4. $P = \sum_{i,j}(X_i - X_j)^4$ contains $n(n-1)/2$ terms. Combining the item 1, the equation holds.

Item 5 and item 6 can be deduce by similar arguments.

According to theorem 2.1 , we need to compute $\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U)))$, $\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big)$ and $\mathrm{Var}_L\big(\mathbb{E}_U(\hat{v}(L_U))\big)$. We begin with $\mathbb{E}_U(\hat{v}(L_U))$ and $\mathbb{E}_U(\hat{v}^2(L_U))$.

Lemma 2.2 *Let X be a random variable with finite forth moment satisfying $\mathbb{E}(X) = 0$. Let $X_1, ..., X_n$ be independent copies of X and $L = (X_1, ..., X_n)$. Let U be a random variable uniformly distributed over $\mathcal{U}$. Then the following equations hold:*

$$\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U))) = \frac{n-1}{n}\mu_2, \tag{4}$$

$$\begin{aligned}
&\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U))^2) \qquad (5) \\
=&\frac{(n-1)(n^2-2n+3)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U)^2)) \qquad (6) \\
=&\frac{n-1}{nm(m-1)}\left(3m - 3 + \frac{n^2-2n+3}{n^2}(m-2)(m-3)\right)\mu_2^2 \\
&+ \frac{n-1}{nm(m-1)}\left(m - 1 + \frac{n-1}{n^2}(m-2)(m-3)\right)\mu_4. \qquad (7)
\end{aligned}$$

Proof. Item 1. Since U is uniformly distributed on $\mathcal{U}$, it holds

$$\mathbb{E}_U(\hat{v}(L_U)) = \frac{1}{n^m}\sum_{i=1}^{n^m}\hat{v}(L_{u_i}) = \frac{1}{n^m}\sum_{i=1}^{n^m}\frac{1}{m(m-1)}\sum_{j<k}(X_{u_i(j)} - X_{u_i(k)})^2. \quad (8)$$

Remark that $\sum_{i=1}^{n^m}\sum_{j<k,j,k\leq m}(X_{L_{u_i}(j)} - X_{L_{u_i}(k)})^2$ is a symmetric polynomial of $(X_j - X_k)^2$, there exists a constant A such that

$$\sum_{i=1}^{n^m}\sum_{j<k,j,k\leq m}(X_{u_i(j)} - X_{u_i(k)})^2 = A\sum_{j<k,j,k\leq n}(X_j - X_k)^2.$$

A is thus the coefficient of $(X_1 - X_2)^2$ in both side. For $U \in \mathcal{U}$, denote $n_1(U)$ and $n_2(U)$ the number of X_1 and the number of X_2 in the sample of bagging $\{X_{U(j)}\}_{j=1,...,m}$. Then the coefficient of $(X_1 - X_2)^2$ in $\sum_{j<k,j,k\leq m}(X_{U(j)} - X_{U(k)})^2$ is $n_1(U)n_2(U)$. Therefore,

$$A = \sum_{j=1}^{n^m} n_1(u_j)n_2(u_j). \quad (9)$$

Consider the function

$$G(X_1, ..., X_n) = \mathbb{E}_U\left(\exp\left(\sum_{i=1}^{m} X_{U(i)}\right)\right).$$

Since G can also be rewritten as

$$G = \frac{1}{n^m}\sum_{j=1}^{n^m}\prod_{i=1}^{n}\exp(n_i(u_j)X_i),$$

we deduce that

$$\sum_{i=1}^{n^m} n_1(u_i)n_2(u_i) = n^m\partial_1\partial_2 G(0, ..., 0).$$

On the other hand, uniformly pick a U out of $\mathcal{U}$ is equivalent to uniformly pick the value of $U(1), U(2), \cdots, U(m)$ out of $\{1, \cdots, n\}$ independently. Hence,

$$G(X_1, ..., X_n) = \mathbb{E}_U\left(\exp\left(\sum_{i=1}^{m} X_{U(i)}\right)\right) = \prod_{i=1}^{m}\mathbb{E}_U\left(\exp\left(X_{U(i)}\right)\right) = \mathbb{E}_B\left(\exp\left(X_u\right)\right)^m,$$

where u is uniformly distributed on $\{1, \cdots, n\}$. Therefore

$$G(X_1, ..., X_n) = \mathbb{E}_U\left(\exp\left(X_u\right)\right)^m = \frac{1}{n^m}\left(\sum_{i=1}^{n} exp(X_i)\right)^m, \quad (10)$$

and

$$n^m\partial_1\partial_2 G(0, ..., 0) = m(m-1)n^{m-2}.$$

Together with equation (9), it holds:

$$A = m(m-1)n^{m-2}.$$

Plugin into equation (8), we have

$$\mathbb{E}_U(\hat{v}(L_U)) = \frac{1}{n^2}\sum_{j<k}(X_j - X_k)^2. \quad (11)$$

Recall that the standard variance estimator is an unbias estimator and can be written as:

$$\hat{v}(L) = \frac{1}{n(n-1)}\sum_{j<k}(X_j - X_k)^2,$$

we deduce that

$$\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U))) = \frac{n-1}{n}\mu_2.$$

Item 2. By equation (11) and a simple development,

$$\mathbb{E}_U(\hat{v}(L_U))^2 = \frac{1}{n^4}\left(\sum_{j<k}(X_j - X_k)^2\right)^2 = \frac{1}{n^4}(P + 2Q + 2R). \quad (12)$$

Combining with lemma 2.1, it holds:

$$\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U))^2) = \frac{1}{n^4}\left(\mathbb{E}(P) + 2\mathbb{E}(Q) + 2\mathbb{E}(R)\right) = \frac{(n-1)(n^2-2n+3)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4. \quad (13)$$

Item 3. We remark again $\mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U)^2))$ is a symmetric polynomial function of $(X_i - X_j)^2$ and the degree of this polynomial is 4. Thus there exists constants α, β and γ such that:

$$\mathbb{E}_U(\hat{v}^2(L_U)) = \frac{1}{n^m}\frac{1}{m^2(m-1)^2}\sum_{i=1}^{n^m}\hat{v}^2(L_{u_i}) = \frac{1}{n^m}\frac{1}{m^2(m-1)^2}(\alpha P + \beta Q + \gamma R). \quad (14)$$

The coefficients α, β and γ are the coefficients of $(X_1 - X_2)^4$, $(X_1 - X_2)^2(X_1 - X_3)^2$ and $(X_1 - X_2)^2(X_3 - X_4)^2$ respectively. We use similar arguments as in the proof of item 1. Denote $n_i(U)$ the number of X_i in the sample of bagging $\{X_{U(j)}\}_{j=1,...,m}$. Again according to property of G, then it holds:

$$\alpha = \sum_{i=1}^{n^m} n_1(u_i)^2 n_2(u_i)^2 = n^m\partial_1^2\partial_2^2 G(0, \cdots, 0),$$

$$\beta = 2\sum_{i=1}^{n^m} n_1(u_i)^2 n_2(u_i) n_3(u_i) \\ = 2n^m \partial_1^2 \partial_2 \partial_3 G(0, \cdots, 0),$$

and

$$\gamma = 2\sum_{i=1}^{n^m} n_1(u_i) n_2(u_i) n_3(u_i) n_4(u_i) \\ = 2n^m \partial_1 \partial_2 \partial_3 \partial_4 G(0, \cdots, 0).$$

Therefore, with equation (10) and (14), we have

$$\begin{aligned} &\mathbb{E}_U(\hat{v}^2(L_U)) \\ =&\left(\frac{1}{n^2 m(m-1)} + \frac{2(m-2)}{n^3 m(m-1)} + \frac{(m-2)(m-3)}{n^4 m(m-1)}\right) P \\ &+ \left(\frac{2(m-2)}{n^3 m(m-1)} + \frac{2(m-2)(m-3)}{n^4 m(m-1)}\right) Q \\ &+ \frac{2(m-2)(m-3)}{n^4 m(m-1)} R. \end{aligned} \tag{15}$$

Item 3 follows with equation (15) and lemma 2.1.

Now we are ready to prove theorem 3.1.

Proof of theorem 3.1 According to theorem 2.1 and lemma 2.2, it holds

$$\mathbb{E}_{(L,B)}(\tilde{v}) = \mathbb{E}_L(\mathbb{E}_U(\hat{v}(L_U))) = \frac{n-1}{n}\mu_2. \tag{16}$$

On the other hand, applying lemma 2.2 successively, we have:

$$\begin{aligned} &\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big) \\ =&\mathbb{E}_L\left(\mathbb{E}_U\left(\hat{v}(L_U)^2\right)\right) - \mathbb{E}_L\left(\mathbb{E}_U(\hat{v}(L_U))^2\right) \\ =&\frac{n-1}{nm(m-1)}\left(3m - 3 + \frac{n^2-2n+3}{n^2}(6-4m)\right)\mu_2^2 \\ &+ \frac{n-1}{nm(m-1)}\left(m - 1 + \frac{n-1}{n^2}(6-4m)\right)\mu_4, \end{aligned}$$

and:

$$\begin{aligned} &\mathrm{Var}_L\big(\mathbb{E}_U(\hat{v}(L_U))\big) \\ =&\mathbb{E}_L\left(\mathbb{E}_U(\hat{v}(L_U))^2\right) - \mathbb{E}_L\left(\mathbb{E}_U(\hat{v}(L_U))\right)^2 \\ =&\frac{(n-1)(n^2-2n+3)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4 - \frac{(n-1)^2}{n^2}\mu_2^2 \\ =&\frac{(3-n)(n-1)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4. \end{aligned} \tag{17}$$

The proof is completed.

3 Proof of theorem 3.3

According to theorem 2.1 and proposition 3.2, it holds

$$\begin{aligned} &\mathrm{MSE}(\tilde{v}(L,B)) - \mathrm{MSE}(\hat{v}) \\ =&\frac{1}{N}\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big) + \left(\frac{5n-3}{n^3} - \frac{2}{n(n-1)}\right)\mu_2^2 \\ &+ \frac{-2n+1}{n^3}\mu_4 \\ =&\frac{1}{n^2}\left(-2\mu_4 + 3\mu_2^2\right) + \frac{1}{n^3}\left(\mu_4 - \mu_2^2\right) \\ &+ \frac{1}{N}\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big) + \mathcal{O}\left(\frac{1}{n^4}\right). \end{aligned}$$

On the other hand, simplifying equation (1) from theorem 2.1, we have:

$$\frac{1}{N}\mathbb{E}\big(\mathrm{Var}_U(\hat{v}(L_U))\big) = \frac{1}{Nm}\left(\mu_4 - \mu_2^2\right) + o\left(\frac{1}{Nm}\right).$$

We deduce therefore

$$\begin{aligned} &\mathrm{MSE}(\tilde{v}(L,B)) - \mathrm{MSE}(\hat{v}) \\ =&\frac{1}{Nm}\left(\mu_4 - \mu_2^2\right) + \frac{1}{n^2}\left(-2\mu_4 + 3\mu_2^2\right) + o\left(\frac{1}{Nm} + \frac{1}{n^2}\right). \end{aligned}$$

We deduce theorem 3.3 by comparing the latter equation with 0.

4 Algorithm Complexity

The algorithm complexity decomposes as follows: The estimation of the kurtosis takes $O(n)$, The estimation of each inner variance takes $O(n)$, thus the estimation of the bagged variance takes $O(Nn)$. The total complexity is thus $O(n) + O(Nn) \approx O(Nn)$. Given that the condition $N > \frac{n}{2}$ must hold, we deduce that the complexity is in $O(n^2)$. Compared to traditional variance estimation in $O(n)$, the complexity is thus greater. As a result, this algorithm can only be useful in cases where the sample size n is not too large.

5 Experimental Setup Details

For all experiments, whenever possible we optimized and parallelized our computation using the Python Numba[1] library. All the experiments were run on a single CPU 3,1 GHz Intel Core i5 with 8 GB of RAM.

5.1 First Experiment

To generate the datasets, we used the make_regression function from sklearn[2]. Regarding the first experiment, we used the Numpy library to estimate the

[1] http://numba.pydata.org/

[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_regression.html

parameters of the linear regression (linalg.lstsq function with default hyperparamters[3]). For the Decision Tree regression, we use the scikit-learn implementation[4] with default hyperparameters. To fit the non-linear curve on the estimated data points, we used the optimize.curve_fit function from the Scipy package[5].

5.2 Second Experiment

For the second experiment, we essentially used the random module from the Numpy library[6].

5.3 Third Experiment

For the third experiment, we used the rv_discrete module from the Scipy library to design a custom discrete distribution[7].

6 Comparison with (Buja and Stuetzle, 2006)

The paper of (Buja and Stuetzle, 2006) studied the bagging-statistic and they have shown that, for the empirical variance $\hat{v}_{bs} = \text{mean}(X^2) - \text{mean}(X)^2$, beneficial effects of bagging (in terms of MSE) exist iff $\kappa > 2$. In their paper, the empirical variance is denoted by U and the corresponding bagging statistic is denoted by U^{bag}, in order to avoid ambiguity of notations and keep the simplicity, we denoted it by $\hat{v}_{bs}$ instead of U and kept U^{bag} the bagging statistic describe in (Buja and Stuetzle, 2006). Here we will use our approach to recover this results.

According to our notations, the empirical variance

$$\begin{aligned}\hat{v}_{bs}(L) &= \frac{1}{n}\sum_{i}^{n} X_i^2 - \frac{1}{n^2}\left(\sum_{i}^{n} X_i\right)^2 \\ &= \frac{Card(L)-1}{Card(L)}\hat{v}(L) = \frac{n-1}{n}\hat{v}(L).\end{aligned}$$

The bagging-statistic defined by (Buja and Stuetzle, 2006) is in fact $\mathbb{E}_U(\hat{v}_{bs}(L_U))$. On can easily check it by applying the equation of part 2 in (Buja and Stuetzle, 2006), we found:

$$U^{bag} = \frac{(n-1)(m-1)}{mn^2}\sum_i X_i^2 - \frac{m-1}{mn^2}\sum_{j,k} X_j X_k.$$

[3] https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.linalg.lstsq.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html

[6] https://docs.scipy.org/doc/numpy/reference/routines.random.html

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.rv_discrete.html

On the other hand, from (11), it holds

$$\mathbb{E}_U(\hat{v}_{bs}(L_U)) = \frac{m-1}{m}\mathbb{E}_U(\hat{v}(L_U)) = \frac{m-1}{mn^2}\sum_{j<k}(X_j - X_k)^2,$$

which is agreed with U^{bag} by expanding the latter expression.

Now we are able to compute the bagging effect for estimator $\hat{v}_{bs}$. Noticing that $\hat{v}_{bs}(L) = \frac{n-1}{n}\hat{v}(L)$ and $\hat{v}_{bs}(L_U) = \frac{m-1}{m}\hat{v}(L_U)$. By (16) and (17),

$$\mathbb{E}(U^{bag}) = \mathbb{E}_L(E_U(\hat{v}_{bs}(L_U))) = \frac{(m-1)(n-1)}{mn}\mu_2,$$

and

$$\begin{aligned}\text{Var}(U^{bag}) &= \text{Var}_L(E_U(\hat{v}_{bs}(L_U))) \\ &= \frac{(m-1)^2}{m^2}\left(\frac{(3-n)(n-1)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4\right).\end{aligned}$$

On the other hand, since $\hat{v}_{bs}(L)$ is the empirical variance,

$$\mathbb{E}_L(\hat{v}_{bs}(L)) = \frac{n-1}{n}\mu_2,$$

and according to Proposition 3.2,

$$\begin{aligned}\text{Var}(\hat{v}_{bs}(L)) &= \text{Var}\left(\frac{n-1}{n}\hat{v}(L)\right) \\ &= \frac{(3-n)(n-1)}{n^2(n-1)}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4.\end{aligned}$$

Now, denote as in (Buja and Stuetzle, 2006) that $g = n/m$ and assuming $0 < g < \infty$, simplifying expressions of $\text{Var}(\hat{v}_{bs}(L))$, $\mathbb{E}_L(\hat{v}_{bs}(L))$, $\text{Var}(U^{bag})$ and $\mathbb{E}(U^{bag})$ we are able to compare the MSE of $\hat{v}_{bs}$ and U^{bag}:

$$\begin{aligned}&\text{MSE}(U^{bag}) - \text{MSE}(\hat{v}_b s) \\ &= \frac{1}{n^2}g(-2\mu_4 + (4+g)\mu_2^2) + \mathcal{O}\left(\frac{1}{n^3}\right).\end{aligned}$$

Therefore, we deduce that the bagging statistic has a smaller MSE iff to $\kappa > 2 + g/2$. In (Buja and Stuetzle, 2006), assuming one can take an optimal g, then bagging statistic has a smaller MSE iff $\kappa > 2$. We recall that the bagging statistic is the quantity requires the number of iterations N goes to ∞. A similar calculation as in the proof of Theorem 3.3, we can reformulate and extend result of (Buja and Stuetzle, 2006) as following, which gives an indication of the number of iteration needed and the optimal sample size.

Theorem 6.1 *As n goes to ∞, bagging reduces on average the MSE of the empirical variance estimator if and only if:*

$$-2\mu_4 + \left(4 + \frac{n}{m}\right)\mu_2^2 < 0,$$

and

$$N > \frac{\mu_4 - \mu_2^2}{2\mu_4 - 4\mu_2^2 - (n/m)\mu_2^2}n.$$

We can see that for distribution with $\kappa > 2$, the sample size m should greater than $\frac{2n}{\kappa-2}$ in order to have beneficial effect for the empirical variance estimator. Compare to theorem 3.3, there is no constraint on sample size for the unbiased variance estimator.

References

Andreas Buja and Werner Stuetzle. Observations on bagging. *Statistica Sinica*, pages 323–351, 2006.