Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

UNCERTAINTY QUANTIFICATION ON SIMULATION ANALYSIS DRIVEN BY RANDOM FORESTS

Amirhossein Meisami
Mark P. Van Oyen

Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave
Ann Arbor, MI 48109, USA

Henry Lam

Department of Industrial Engineering and Operations Research
Columbia University
500 W. 120th Street
New York, NY 1002, USA

ABSTRACT

We consider simulation-based estimation where the inputs are calibrated from predictive models generated by random forests (RFs). RF is a common technique to produce ensemble predictions by aggregating many individual decision trees. This problem arises in data-driven applications such as individualized surgery operations scheduling and supply chain management. We investigate the estimation of the output variance contributed from the noises in the input prediction, which can be used to construct output confidence intervals. In particular, we study the integration of simulation runs with a recently proposed infinitesimal jackknife estimator that can reduce the computational burden from double layers of bootstrapping. We demonstrate our scheme with an elementary numerical example.

1 INTRODUCTION

Modern simulation analysis has been increasingly linked with learning-based predictive modeling; see, e.g., the recent literature in simulation analytics (Jiang, Hong, and Nelson 2016). This paper presents a study along this line, and considers simulation analysis driven by input distributions that are generated from predictive models. More precisely, consider the estimation of $\psi(F_1,\ldots,F_m)$ where $\psi(\cdot)$ is an output summary of interest that can be simulated from the input distributions $F_1,\ldots,F_m$. Here, F_i's are distributions given individual sets of covariates that are calibrated by learning from historical data. Our interest is uncertainty quantification schemes for the simulation output $\psi(F_1,\ldots,F_m)$ under this setting, in particular the estimation of the variance stemming from the noises in building the predictive model, and subsequently the construction of an output confidence interval.

This particular problem arises from the authors' own encounter in individualized surgery operations scheduling (Meisami et al. 2017). In this context, $\psi(F_1,\ldots,F_m)$ is the expected daily total waiting time in a surgery room. $F_1,\ldots,F_m$ denote the distributions of the surgery times for m scheduled patients. These distributions come from a predictive model given each patient's characteristics, such as age, gender, past medical history etc. Clearly, an informed decision-making based on the performance measure requires proper quantification of the variability due to the errors coming from the input prediction.

Another natural instance of this problem, in the area of supply chain and inventory management, is the common newsvendor problem. For instance, in an m-product multi-stage newsvendor problem, the decision maker has to choose the order quantity of each product at different times based on the predicted demand such that cumulative back-order and inventory cost is minimized. The decision maker often has access to information such as customer demographics, seasonality, economic conditions, etc. and use such data to predict the product demand distributions $F_1,\ldots,F_m$. The quantity $\psi(F_1,\ldots,F_m)$ here refers to the

expected cumulative cost under a given ordering policy. Much like the surgery scheduling scenario, it is operationally important to capture the variation and uncertainty in deciding a desirable policy.

One common method in building predictive models for F_i's with covariates, which we focus on in this paper, is the quantile regression forest (QRF) (Meinshausen 2006). This is a generalization of random forest (RF) that produces ensemble prediction by aggregating many decision trees together, where each tree is grown by random splitting and a bootstrap replication of the data set (known as bootstrap aggregating, or bagging in short; Breiman 1996, Breiman 2001). RF and bagging have been shown, both theoretically and empirically, to have desirable properties (e.g., variance reduction, stability; Büchlmann and Yu 2002, Efron 2014), and have become one of the most popular off-the-shelf learning methods. QRF differs from the standard RF in that the output of each tree is a probability distribution on the response variable, so that their aggregation give rise to an overall predicted distribution. Along a similar vein as RF, QRF has been shown to demonstrate superior predictive power than standard tools like quantile regression in certain scenarios (Meinshausen 2006).

Our current work on uncertainty quantification, through estimating the variance contributed from input noises and its subsequent use in confidence interval construction, follows the stream of literature on input uncertainty (e.g., Barton (2012), Song, Nelson, and Pegden (2014), Lam (2016)). In particular, we utilize the general technique of the bootstrap and the delta method. Cheng and Holland (1997) considers the estimation of input-induced variance via both the delta method and the bootstrap, the latter used together with an analysis-of-variance adjustment to reduce bias. Cheng and Holland (1998) and Cheng and Holland (2004) suggest a more efficient refinement of the delta method called the two-point method. Song and Nelson (2015) studies the bootstrap on a mean-variance model in capturing input uncertainty.

To estimate the sampling variance from QRF (or RF) for a prediction problem, standard bootstrapping would require running two layers of resampling, one on the original data and one on the resampled data used to construct the QRFs. In simulation driven by these QRFs, there is an additional effort of simulation runs, thus posing enormous computational burden. As the main focus of this paper, we investigate the integration of simulation with an infinitesimal jackknife (IJ) estimator, proposed by the recent work Efron (2014) and Wager, Hastie, and Efron (2014) that avoids double-layer bootstrapping for RFs in prediction problems, in estimating the output variance contributed from the input noises of the QRF. Throughout this paper we will focus on the essential ideas and omit the rigorous mathematical developments.

In the rest of this paper, Section 2 will first describe the basics of QRF and how to use it in simulation analysis. Section 3 then introduces the IJ estimator and discusses how to blend it efficiently in output uncertainty quantification. Section 4 shows a numerical example.

2 SIMULATION ANALYSIS DRIVEN BY QUANTILE REGRESSION FORESTS

We first describe some basics of QRF, followed by a discussion on how it is used for input modeling in simulation analysis.

2.1 Basics of Quantile Regression Forest

We follow the discussion in Meinshausen (2006). Given a collection of i.i.d. data $(X_j, Y_j), j = 1, \ldots, n$, where $X_j \in \mathscr{B}$ for some appropriate space $\mathscr{B}$ is the covariate or feature vector, and $Y_j \in \mathbb{R}$ is the response, the goal is to estimate $F(y|X = x) = P(Y \leq y|X = x)$ for all $y \in \mathbb{R}$, the conditional probability distribution of the response Y given the feature $X = x$.

To start, we first grow a tree by splitting the feature space $\mathscr{B}$ into non-overlapping rectangular subspaces, say $R_l, l = 1, \ldots, L$, each representing a leaf. This tree is generated randomly, by using a bootstrap resample $\{(X_i^*, Y_i^*)\}_{i=1,\ldots,n}$, i.e., a sample of size n drawn from $\{(X_j, Y_j)\}_{j=1,\ldots,n}$ under sampling-with-replacement, and also randomly splitting the feature space according to some independent random source denoted by θ. For each feature value x, we denote $R_{l(x,\theta)}$ as the leaf that x lies on. For convenience, we denote $(X^*, Y^*) = ((X_j^*, Y_j^*) : j = 1, \ldots, n)$ as the bootstrapped data set. The output of a tree, given x, is the

empirical distribution of Y^* inside the same leaf, i.e.,

$$t(y;x,(X^*,Y^*),\theta) = \frac{\sum_{j:X_j^* \in R_{l(x,\theta)}} I(Y_j^* \le y)}{\#\{j : X_j^* \in R_{l(x,\theta)}\}}$$

We grow these random trees many times, say B times, where each tree contains the bootstrap sample $(X^b,Y^b) = ((X_j^b,Y_j^b) : j = 1,\ldots,n)$ and random realization θ^b. Then the QRF outputs the average of all trees, i.e., given x, it outputs

$$\hat{F}(y|X=x) = \frac{1}{B}\sum_{b=1}^{B} t(y;x,(X^b,Y^b),\theta^b) \tag{1}$$

Besides its superior empirical performance (e.g., compared to quantile regression, especially for extremal quantiles), QRF has been shown to be consistent, nonparametrically, when the leave sizes of the trees (measured by the proportions of data) are grown at a suitable rate relative to the sample size and the splitting mechanism are properly taken (Meinshausen 2006). However, convergence rate results appear only available in the case of subsampled trees (i.e., the bootstrap resample size grows at a smaller rate than n); see, e.g., Mentch and Hooker (2016), Wager and Athey (2017).

2.2 Input Modeling

Consider the performance measure described in the introduction $\psi(F_1,\ldots,F_m)$, where given F_i's, $\psi(\cdot)$ can be estimated by running simulation. An example would be the expectation-type measure

$$\psi(F_1,\ldots,F_m) = E_{F_1,\ldots,F_m}[h(\mathbf{Y}_1,\ldots,\mathbf{Y}_m)] \tag{2}$$

where $\mathbf{Y}_i$ denotes a sequence of i.i.d. variates generated from F_i, and h is some cost function that can be evaluated given the $\mathbf{Y}_i$'s.

We consider the situation where each input distribution F_i is a conditional distribution given a feature value x_i, i.e., $F_i(y) = P(Y \le y|X = x_i)$. As mentioned in the introduction, these F_i's could represent the distributions of the random surgery times given patients' observed characteristics. Using the QRF, a natural estimate for $\psi(F_1,\ldots,F_m)$ will be $\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)$, where $\hat{F}_i(y) = \hat{F}(y|X = x_i)$ is given by (1), and $\hat{\psi}$ is a simulation approximation of ψ. In the context of (2), this means

$$\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m) = \frac{1}{R}\sum_{r=1}^{R} h(\mathbf{Y}_1^r,\ldots,\mathbf{Y}_m^r)$$

where each $\mathbf{Y}_i^r$ consists of independent realizations of variates generated from input distribution $\hat{F}_i$. Under the assumptions that guarantee nonparametric consistency of QRF and mild regularity conditions on $\psi(\cdot)$, $\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)$ is consistent for estimating $\psi(F_1,\ldots,F_m)$ when the bootstrap size B in the QRF and the simulation size R grow to ∞, and hence $\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)$ is a justified estimate of the performance measure.

For reliable decision-making, we need a measure of the variability on the point estimate of $\psi(F_1,\ldots,F_m)$. This variability comes from both the noises of the simulation replications (called stochastic or simulation uncertainty) and the input data (called input uncertainty); see, e.g., Barton (2012), Song, Nelson, and Pegden (2014), Lam (2016). This variability can be captured by a confidence interval (CI), with a confidence level say $1-\alpha$, given by

$$\left[\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m) - z_{1-\alpha/2}\sqrt{\sigma^2 + \frac{\tau^2}{R}}, \hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m) + z_{1-\alpha/2}\sqrt{\sigma^2 + \frac{\tau^2}{R}}\right] \tag{3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The variance component σ^2 is the input-induced variance, whereas τ^2/R is the simulation-induced variance. Interval (3) has assumed that the

bootstrap size B is large enough to make the resampling variance negligible, and a Gaussian approximation is asymptotically valid. This assumption is not formally justified here. In Section 3, we will consider a modified version of (3) and provide further discussion.

The quantity τ^2 is relatively easy to estimate, for instance by taking the sample variance of all the simulation replications. The quantity σ^2, in the nonparametric setting, typically involves estimating the variance of the so-called influence function of $\psi(\cdot)$, which can be viewed as a functional derivative with respect to the input distributions (Hampel et al. 2011, Huber 2011). It could be difficult to efficiently estimate σ^2 by directly using the form of the influence function (e.g., in a long-horizon problem). Therefore, bootstrapping has been used as a major tool for estimating σ^2 (e.g., Cheng and Holland 1997, Song and Nelson 2015). In the current context, this is done by repeatedly drawing bootstrap resamples from $\{(X_i,Y_i)\}_{i=1,\ldots,n}$. For each resample, fit the QRF pretending the resample is the original data, and calculate $\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)$. The variance of these outcomes gives the bootstrap variance that approximates σ^2.

Note that the above scheme for estimating σ^2 requires two layers of bootstrapping, the first corresponding to the generation of resamples on which QRFs are trained on, and the second to the generation of trees in constructing the QRFs and subsequently the drive of simulation runs. The overall computational effort is on the order of the product of the outer bootstrap size and the sum of the tree and the simulation replication sizes.

3 SIMULATION-BASED INFINITESIMAL JACKKNIFE ESTIMATOR

Given the difficulties in approximating σ^2 in (3), we consider an alternate approach that requires looking at a slightly different point estimator. For convenience, let $\mathbf{F} = (F_1,\ldots,F_m)$ and $\mathbf{F}^*(y) = (t(y;x_i,(X^*,Y^*),\theta^*) : i = 1,\ldots,m)$ as the vector of input distributions predicted by each tree. Our point estimator depicted in Section 2.2 can be expressed as $\hat{\psi}(\hat{E}[\mathbf{F}^*])$ where $\hat{E}[\cdot]$ denotes the sample average of B independent realizations of trees. As the numbers of trees and simulation replications grow to ∞, this approximates $\psi(E_*[\mathbf{F}^*])$ where $E_*[\cdot]$ denotes the expectation generated by the bootstrapped trees. CI (3) is derived from the conjectured central limit theorem $\psi(E_*[\mathbf{F}^*]) \sim N(\psi(\mathbf{F}),\sigma^2)$.

Our considered alternate point estimator takes the form

$$\hat{E}[\hat{\psi}(\mathbf{F}^*)] \tag{4}$$

that interchanges $\hat{E}$ and $\hat{\psi}$. As the numbers of trees and simulation replications grow to ∞, the estimator (4) approximates

$$E_*[\psi(\mathbf{F}^*)] \tag{5}$$

where $E_*[\cdot]$ denotes the expectation with respect to the bootstrap. We note that (5) is equivalent to

$$E_*[\psi^*(\mathbf{F}^*)], \tag{6}$$

where $\psi^*(\mathbf{F}^*)$ denotes an unbiased simulation replication in estimating $\psi(\mathbf{F}^*)$ given $\mathbf{F}^*$ (e.g., $h(\mathbf{Y}_1,\ldots,\mathbf{Y}_m)$ with $\mathbf{Y}_i$ being i.i.d. sequence generated from F_i^*), and with slight abuse of notation, $E_*[\cdot]$ in (6) denotes the expectation generated from the combination of a bootstrapped tree and the simulation replication. Note that (6) can be viewed as the limiting output of a bagging scheme, where each "base learner" is now a realization of $\psi^*(\mathbf{F}^*)$. From this view, an implementable bagging scheme to approximate (6) is

$$\hat{E}[\psi^*(\mathbf{F}^*)] = \frac{1}{B}\sum_{b=1}^{B} h(\mathbf{Y}_1^b,\ldots,\mathbf{Y}_m^b) \tag{7}$$

where $(\mathbf{Y}_i^b)_{i=1,\ldots,m}$ is generated from $\mathbf{F}^b = (t(y;x_i,(X^b,Y^b),\theta^b) : i = 1,\ldots,m)$. This is similar to using the established QRF, but instead of outputting $t(y;x,(X^b,Y^b),\theta^b)$ for each tree, we output a copy of the simulation output generated by the input models $\mathbf{F}^b$.

Our subsequent analysis assumes that when n is large and under smoothness conditions on $\psi(\cdot)$, $\mathbf{F}^* \approx E_*[\mathbf{F}^*]$ and $E_*[\psi^*(\mathbf{F}^*)] = E_*[\psi(\mathbf{F}^*)] \approx \psi(E_*[\mathbf{F}^*])$ well enough so that a central limit theorem holds for $E_*[\psi^*(\mathbf{F}^*)]$ and implies $E_*[\psi^*(\mathbf{F}^*)] \overset{approx.}{\sim} N(\psi(\mathbf{F}), \sigma^2)$. Thus, together with the simulation variability, a $(1-\alpha)$-level CI for $\psi(\mathbf{F})$ is

$$\left[\hat{E}[\psi^*(\mathbf{F}^*)] - z_{1-\alpha/2}\sqrt{\sigma^2 + \frac{\tau^2}{R}}, \hat{E}[\psi^*(\mathbf{F}^*)] + z_{1-\alpha/2}\sqrt{\sigma^2 + \frac{\tau^2}{R}}\right] \tag{8}$$

The main purpose of using $\hat{E}[\psi^*(\mathbf{F}^*)]$ instead of $\hat{\psi}(\hat{\mathbf{F}})$ in Section 2.2 is the availability of an estimate for σ^2 without resorting to a double-layer bootstrapping, given by the infinitesimal jackknife (IJ) estimator proposed by Efron (2014) and Wager, Hastie, and Efron (2014), in the form

$$\hat{V}_{IJ} = \sum_{i=1}^{n} Cov_*(N_i^*, \psi^*(\mathbf{F}^*))^2 \tag{9}$$

where N_i^* is the number of counts of the observation (X_i, Y_i) appearing in the bootstrap resample for building F^*, and $Cov_*(\cdot,\cdot)$ denotes the covariance generated by the boostrapping. Under some conditions using subsampled trees, it is shown that $\hat{V}_{IJ}/\sigma^2 \xrightarrow{p} 1$ (Wager and Athey 2017). However, as far as we know, there have been no such convergence results available for QRFs with full-size trees, and we may need to resort to the use of subsampled trees for a rigorous analysis leading to (8).

A bias-reduced estimator for (9) using a finite tree size B (Wager, Hastie, and Efron 2014) is given by

$$\hat{V}_{IJ}^B = \sum_{i=1}^{n} \widehat{Cov}_i^2 - \frac{n}{B^2}\sum_{b=1}^{B}(\psi^b(\mathbf{F}^b) - \overline{\psi^*(\mathbf{F}^*)})^2 \tag{10}$$

where

$$\widehat{Cov}_i = \frac{1}{B}\sum_{b=1}^{B}(N_i^b - 1)(\psi^b(\mathbf{F}^b) - \overline{\psi^*(\mathbf{F}^*)})$$

N_i^b and $\psi^b(\mathbf{F}^b)$ denote the realizations of N_i^* and $\psi^*(\mathbf{F}^*)$ in the b-th tree, and $\overline{\psi^*(\mathbf{F}^*)} = (1/B)\sum_{b=1}^{B}\psi^b(\mathbf{F}^b)$. As explained in Wager, Hastie, and Efron (2014), the first term in (10) is the plug-in estimator of (9), and the second term corrects the bias of the plug-in estimator from order $\sigma^2/(nB)$ to order $n\sigma^2/B$. We note that Wager, Hastie, and Efron (2014) focuses on the mean prediction, but their analysis clearly applies to any type of function on the response, including $\psi^*(\mathbf{F}^*)$.

We have used one intra-tree simulation replication in the illustration above. Note that (6) remains unchanged if one increases the intra-tree simulation replication size from one to any number ((5) is an extreme that uses an infinite number of simulation replications). The conjectured asymptotic analysis remains the same except that the intra-tree variance for $\psi^*(\mathbf{F}^*)$ decreases in the usual Monte Carlo rate. Thus, one can use $\hat{V}_{IJ}^B$ with $\psi^b(\cdot)$ redefined accordingly. Our numerical experiment in the next section indicates that under a fixed budget, a single intra-tree replication does appear to perform superior to other choices.

4 NUMERICAL EXAMPLE

In this section we provide numerical results on the method discussed in Section 3 using the well-known "Auto MPG" data set that was analyzed in Wager, Hastie, and Efron (2014). We use 7 features (number of cylinders, displacement, horsepower, weight, acceleration, year, and origin) over a training set of 312 observations to build our QRF. For this proof-of-concept example, we fix four feature values in a separate test set, and consider a performance measure

$$\psi(F_1, \ldots, F_4) = E_{F_1,\ldots,F_4}[h(Y_1, \ldots, Y_4)] \tag{11}$$

where h is a simple sum function

$$h(y_1,\ldots,y_4) = y_1 + y_2 + y_3 + y_4$$

and Y_i's are four independent miles-per-gallon (MPG) responses generated from the QRF given the four feature values. We use simulation to estimate $\psi(F_1,\ldots,F_4)$ (even though it can be computed exactly).

We should mention that, although we choose to use a very simple h, the same procedure essentially applies when h is the total waiting and overtime cost of a surgery schedule and Y_i's denote the predicted surgery durations. Similarly, h can be the overall back-order and holding cost in a multi-product newsvendor problem and Y_i's denote the predicted product demands.

We first show two results (Figures 1 and 2) following the analysis in Wager, Hastie, and Efron (2014). Figure 1 shows the point estimate of (11), given by (7), and the standard deviation estimate, given by (10), on 20 instances of (11) with different feature values. We use one intra-tree simulation replication in our experiment. The behaviors of these values show similar patterns as Figure 1 in Wager, Hastie, and Efron (2014), except that the variation of the standard deviation estimate appears more uniform regardless of the distance to the actual values. This may be explained by the summation operation h that gives an averaging effect.

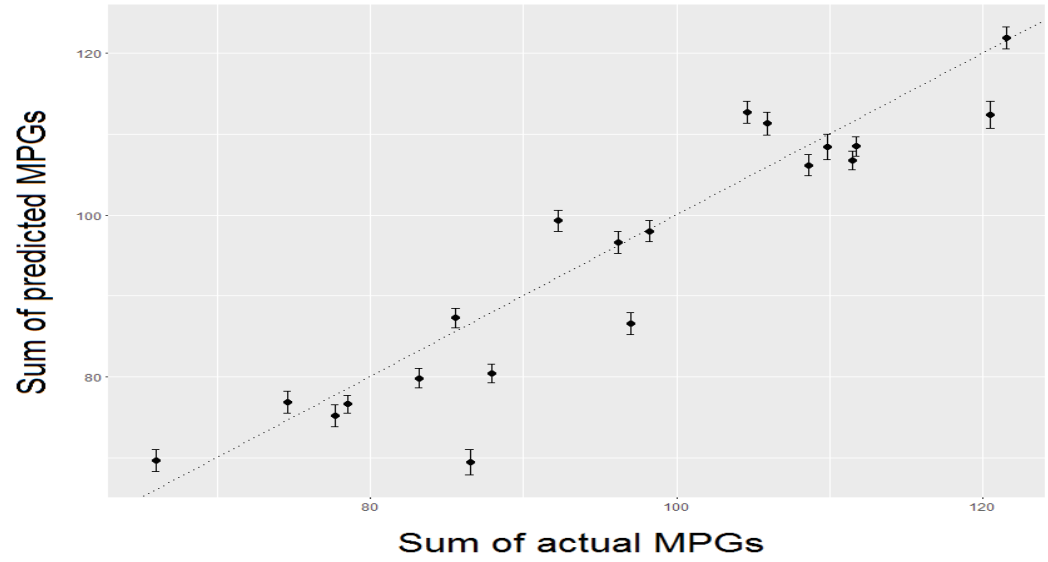

Figure 1: Point and standard deviation estimates on (11) with the "Auto MPG" data set, following the representation of Figure 1 in Wager, Hastie, and Efron 2014. We depict the results on 20 different instances of (11). The dot and the half-length of each interval represent the point and standard deviation estimates respectively.

Next, Figure 2 shows a comparison between the IJ variance estimate and the estimate generated by direct bootstrapping. The latter requires a double-layer bootstrap, first on the generation of the resample used to build QRF, and second on the generation of the trees within each QRF. We use a resample size of $1,000$ in each layer and one intra-tree simulation replication. We use $B = 2000$ and also one intra-tree simulation replication in the IJ estimate. Figure 2 shows that the IJ variance estimate is generally close to the direct bootstrap, suggesting the validity of the former.

We now consider varying the intra-tree simulation replication size, denoted R, as discussed at the end of Section 3. We consider three different scenarios under a total simulation budget of 2000 replications, first with 200 trees and 10 intra-tree replications, second with 400 trees and 5 intra-tree replications, and third with 2000 trees and only one intra-tree replication. Table 1 shows that the variance estimates are

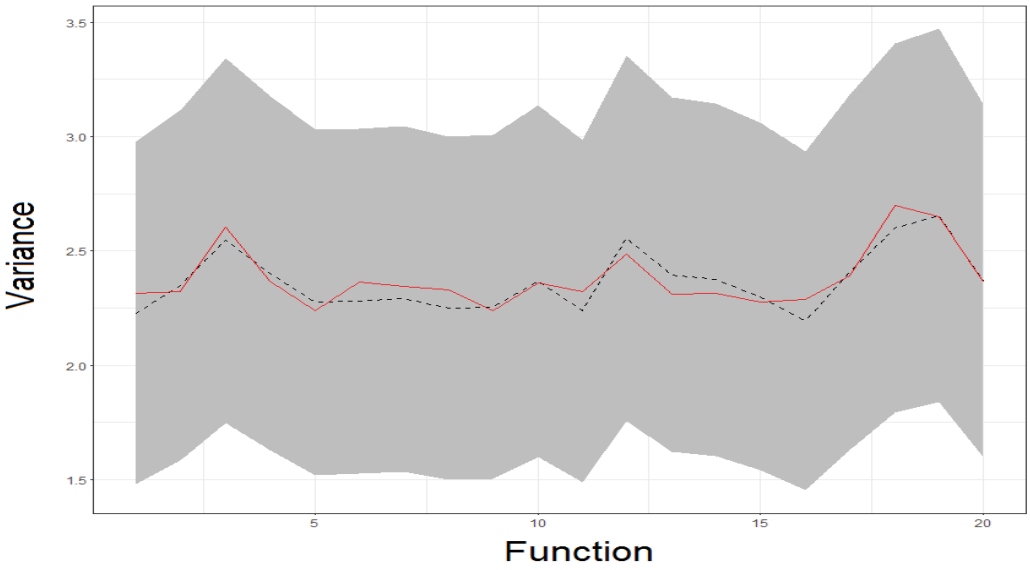

Figure 2: Variance estimates using IJ and direct boostrapping, on 20 different instances of (11). Solid line shows IJ, and dashed line shows direct bootstrapping. The half-height of the gray band is the standard deviation of the bootstrap distribution in using the direct bootstrapping.

most uniform in the $B = 2000$ case, which are also significantly higher than the other two cases. Since Figure 2 demonstrates that the estimates using $B = 2000$ matches closely with direct bootstrapping, they can serve as a reasonable benchmark. This implies that increasing R under a fixed budget in this example may degrade the estimation reliability.

5 CONCLUSION

We have studied the estimation of the variance contributed from input noise, as a way to quantify input uncertainty and construct a confidence interval, in stochastic simulation driven by input models calibrated from QRFs. Standard bootstrapping on this problem requires double layers of resampling and simulation runs. We study the integration of an IJ scheme recently proposed in the context of prediction problems, which avoids the use of double-bootstrapping, into the simulation-based estimation of the input-induced variance. We present a strategy and demonstrate it through an elementary numerical example. Future work includes the mathematical analysis and more extensive testing of our approach.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453, CMMI-1548201 and CAREER CMMI-1653339.

REFERENCES

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, O. R. R. Pasupathy, and A. Uhrmacher, 1–12. Piscataway, New Jersey: IEEE.

Breiman, L. 1996. "Bagging Predictors". *Machine Learning* 24 (2): 123–140.

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

Büchlmann, P., and B. Yu. 2002. "Analyzing Bagging". *Annals of Statistics* 1:927–961.

Table 1: Variance estimates on (11) with different combinations of tree and intra-tree replication sizes under fixed simulation budget.

	Variance Estimates		
Mean Sum	$B = 200$ & $R = 10$	$B = 400$ & $R = 5$	$B = 2000$ & $R = 1$
106.56	0.46	0.95	2.06
96.97	1.85	1.07	2.30
112.54	11.99	0.36	2.28
121.85	0.66	0.48	2.10
106.73	0.63	0.59	2.08
108.49	0.50	0.34	2.17
99.24	0.30	0.50	2.07
97.87	0.71	0.43	2.09
87.07	0.44	0.76	2.04
69.25	1.90	0.48	2.15
76.76	3.52	0.66	2.16
69.55	0.94	0.30	2.16
75.04	0.44	0.64	2.29
80.46	0.54	0.66	2.12
76.76	0.27	1.02	2.10
79.70	0.52	0.43	2.07
86.58	0.25	0.95	2.01
111.49	0.12	0.47	2.24
113.14	0.24	0.37	2.10
108.55	0.33	0.23	2.15

Cheng, R. C., and W. Holland. 1997. "Sensitivity of Computer Simulation Experiments to Errors in Input Data". *Journal of Statistical Computation and Simulation* 57 (1-4): 219–241.

Cheng, R. C., and W. Holland. 1998. "Two-point Methods for Assessing Variability in Simulation Output". *Journal of Statistical Computation Simulation* 60 (3): 183–205.

Cheng, R. C., and W. Holland. 2004. "Calculation of Confidence Intervals for Simulation Output". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 14 (4): 344–362.

Efron, B. 2014. "Estimation and Accuracy After Model Selection". *Journal of the American Statistical Association* 109 (507): 991–1007.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 2011. *Robust Statistics: The Approach Based on Influence Functions*, Volume 114. New York: John Wiley & Sons.

Huber, P. J. 2011. *Robust Statistics*. Berlin Heidelberg: Springer.

Jiang, G., L. J. Hong, and B. L. Nelson. 2016. "A Simulation Analytics Approach to Dynamic Risk Monitoring". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, WSC '16, 437–447. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers, Inc.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Winter Simulation Conference (WSC), 2016*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192. Piscataway, NJ, USA: IEEE.

Meinshausen, N. 2006. "QuantileRegression Forests". *Journal of Machine Learning Research* 7 (Jun): 983–999.

Meisami, A., H. Lam, M. Van Oyen, C. Stromblad, and N. Kastango. 2017. "An Individualized Learning Methodology for the Surgery Scheduling Problem". *Working paper, University of Michigan Industrial and Operations Engineering Dept.*.

Mentch, L., and G. Hooker. 2016. "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests". *Journal of Machine Learning Research* 17 (26): 1–41.

Song, E., and B. L. Nelson. 2015. "Quickly Assessing Contributions to Input Uncertainty". *IIE Transactions* 47 (9): 893–909.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Diallo, I. Ryzhov, L. Yilmaz, S. Buckley, and J. Miller, 162–176. Piscataway, New Jersey: IEEE.

Wager, S., and S. Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests". *Journal of the American Statistical Association* (just-accepted).

Wager, S., T. Hastie, and B. Efron. 2014. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife". *Journal of Machine Learning Research* 15 (1): 1625–1651.

AUTHOR BIOGRAPHIES

AMIRHOSSEIN MEISAMI is a Ph.D. candidate in the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. His research interests include data-driven decision making and integration of machine learning and optimization. His e-mail address is meisami@umich.edu.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics at Harvard University, and was a faculty member in the Department of Mathematics and Statistics at Boston University and the Department of Industrial and Operations Engineering at the University of Michigan before joining Columbia. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khl2114@columbia.edu.

MARK VAN OYEN Mark Van Oyen is a Professor of Industrial and Operations Engr. (IOE) at the University of Michigan. His interests include the analysis, design, and control of stochastic systems (models and applications). His current research emphasizes healthcare operations and medical decision making. He co-authored papers that won the 2016 Manufacturing and Service Operations Management (MSOM) Best Paper Award, 2016 MSOM Service Science SIG Best Paper Award, and the 2010 Pierskalla Award. He has served as Associate Editor for *Operations Research, Naval Research Logistics, IIE Transactions*, and *IIE Transactions on Healthcare Syst. Engr.* and Senior Editor for *Flexible Services & Manufacturing*. He was a faculty member of the Northwestern Univ. Sch. of Engr. (1993-2005) and Loyola Univ. of Chicago's Sch. of Bus. Admin. (1999-2005). His email address is vanoyen@umich.edu.