Random Forest Prediction Intervals

Haozhe Zhang[†], Joshua Zimmerman[†], Dan Nettleton[†,*], and Daniel J. Nordman[†,*]
Department of Statistics, Iowa State University

February 16, 2019

Abstract

Random forests are among the most popular machine learning techniques for prediction problems. When using random forests to predict a quantitative response, an important but often overlooked challenge is the determination of prediction intervals that will contain an unobserved response value with a specified probability. We propose new random forest prediction intervals that are based on the empirical distribution of out-of-bag prediction errors. These intervals can be obtained as a by-product of a single random forest. Under regularity conditions, we prove that the proposed intervals have asymptotically correct coverage rates. Simulation studies and analysis of 60 real datasets are used to compare the finite-sample properties of the proposed intervals with quantile regression forests and recently proposed split conformal intervals. The results indicate that intervals constructed with our proposed method tend to be narrower than those of competing methods while still maintaining marginal coverage rates approximately equal to nominal levels.

Keywords: conformal inference; coverage rate; interval width; out-of-bag prediction errors; quantile regression forests.

[*] Corresponding author emails: dnett@iastate.edu, dnordman@iastate.edu.

[†] The authors gratefully acknowledge *The Iowa State University Plant Sciences Institute Scholars Program.*

1 Introduction

The seminal paper on random forests (Breiman, 2001) has nearly 42,000 citations as of December, 2018, according to Google Scholar. The impact of Breiman's random forests on machine learning, predictive analytics, data science, and science in general is difficult to measure but unquestionably substantial. The virtues of random forest methodology, summarized nicely in the recent review article by Biau and Scornet (2016), include no need to specify functional forms relating predictors to a response variable, capable performance for low-sample-size high-dimensional data, general prediction accuracy, easy parallelization, few tuning parameters, and applicability to a wide range of prediction problems with categorical or continuous responses.

Like many algorithmic approaches to prediction, random forests are typically used to produce point predictions that are not accompanied by information about how far those predictions may be from true response values. From the statistical point of view, this is unacceptable; a key characteristic that distinguishes statistically rigorous approaches to prediction from others is the ability to provide quantifiably accurate assessments of prediction error from the same data used to generate point predictions. Thus, our goal here is to develop a prediction interval, based on a random forest prediction, that gives a range of values that will contain an unknown continuous univariate response with any specified level of confidence.

Formally, suppose $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ is a random predictor-response pair distributed according to some unknown distribution $\mathbb{G}$, where Y represents a continuous univariate response that we wish to predict using its predictor information $\boldsymbol{X}$. Suppose $(\boldsymbol{X}, Y)$ is independent of a training set $\boldsymbol{\mathcal{C}}_n$ consisting of observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} \mathbb{G}$. We seek a prediction interval $\mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)$ that will cover the response value Y with probability $1 - \alpha$.

One existing approach for obtaining forest-based prediction intervals involves estimating the conditional distribution of the response variable Y given the predictor vector $\boldsymbol{X} = \boldsymbol{x}$ via quantile regression forests (Meinshausen, 2006). Lower and upper quantiles of an estimated conditional distribution naturally provide a prediction interval for the response at any point $\boldsymbol{x}$ in the predictor space. Prediction intervals produced with quantile regression forests (QRFs) often perform well in terms of conditional coverage at or above nominal levels (i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{X} = \boldsymbol{x}] \geq 1 - \alpha$). QRFs are also very versatile because they do not require the scale or even the shape of the conditional response distribution to be constant across predictor values. However, this versatility comes at a cost. Without stronger assumptions about shared features of the conditional response distributions, each conditional response distribution must be separately estimated using a relatively small amount of data local to the point $\boldsymbol{x}$ in the predictor space at which a prediction interval is desired. This can lead to highly variable estimators of conditional response distributions and QRF intervals that are often quite wide, which diminishes their informativeness and usefulness in some applications. There are, of course, some challenging prediction problems where the flexibility of QRFs is needed, but there are many other problems where common features of conditional response distributions can be exploited to produce more informative prediction intervals.

In contrast to QRF intervals, our approach to interval construction borrows information across the entire training dataset $\boldsymbol{\mathcal{C}}_n$ by assuming that the distribution of a random forest prediction error (response value less the random forest prediction) can be well approximated by the empirical distribution of out-of-bag (OOB) prediction errors obtained from all training observations. Fortunately, the empirical distribution of OOB prediction errors can be obtained with no additional resampling beyond the resampling used to construct a single random forest. Once the empirical distribution of the OOB prediction errors has been obtained, it is straightforward to combine this estimated prediction error distribution with the random forest prediction of the response value for a new case to obtain a prediction interval. By working with a de-trended version of the response, we can focus on estimating one prediction error distribution and use this distribution to obtain all prediction intervals rather than estimating separate conditional response distributions for all new cases as in QRFs.

Our approach is similar to the general technique of prediction interval construction via split conformal (SC) inference (Lei et al., 2018). Prediction intervals with guaranteed finite-sample marginal coverage probability (i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)] \geq 1 - \alpha$) can be generated using SC inference in conjunction with any method for estimating $\mathbb{E}(Y | \boldsymbol{X} = \boldsymbol{x})$, the conditional mean of a response given the predictor variable values in a vector $\boldsymbol{x}$. Our work differs from the random forest interval approach presented as a special case of SC inference by Lei et al. (2018). Rather than relying on a single random partitioning of the training set $\boldsymbol{\mathcal{C}}_n$ into two subsets to obtain cross-validated prediction errors as in SC inference, we use OOB prediction errors that can be naturally obtained from a single random forest constructed from all training observations. Just as SC inference can serve as a general method for interval construction, our OOB-based approach could also be applied with conditional mean estimation techniques other than random forests. We leave investigation of such generalizations to future work and maintain the focus of this paper on random forests.

The rest of this paper is organized as follows. In Section 2, we provide some basic background on the mechanics of random forests, explain some by-products of random forests, and define our approach to random forest prediction interval construction. Section 3 introduces four coverage probability types and explains the asymptotic properties of the proposed out-of-bag random forest prediction intervals. In Section 4, we describe competing approaches for constructing random forest prediction intervals. In Section 5, we compare the finite-sample performance of our prediction intervals to other methods in a simulation study, in terms of four types of coverage rates and interval widths. In Section 6, we evaluate the performance of our approach and others on 60 real datasets. The R code and datasets used in Section 5 and Section 6 are publicly available at `https://github.com/haozhestat/RFIntervals`. The paper concludes with a discussion in Section 7. Proofs of main results and some additional figures are included in the Supplemental Materials.

2 Constructing Random Forest Prediction Intervals

Our proposed OOB prediction interval, defined in Section 2.3, is based on a single random forest and its by-products. We use the random forest algorithm implemented in the R package *randomForest* (Liaw et al., 2002) and summarized in Section 2.1.

2.1 The Random Forest Algorithm

Based on Fortran code originally provided by Leo Breiman and Adele Cutler, the *randomForest* R package (Liaw et al., 2002) provides a convenient tool for generating a random forest. The algorithm has two tuning parameters, referred to as *mtry* and *nodesize* in the *randomForest* R package and in the description of the algorithm below. These tuning parameters are discussed more fully after our formal definition of the algorithm.

1. Draw an equal-probability, with-replacement sample of size n from $\boldsymbol{\mathcal{C}}_n$ to create a bootstrap training dataset $\boldsymbol{\mathcal{C}}_n^* = \{(\boldsymbol{X}_i^*, Y_i^*) : i = 1, \ldots, n\}$.

2. Use $\boldsymbol{\mathcal{C}}_n^*$ to grow a regression tree T^*.

   (a) Start with all the cases in $\boldsymbol{\mathcal{C}}_n^*$ in a single *root node* $\mathcal{N}$.

   (b) Draw a simple random sample $\mathcal{S}$ of *mtry* predictor variables from the set of all p predictor variables.

   (c) Consider partitions of the cases in $\mathcal{N}$ into subnodes $\mathcal{N}_1$ and $\mathcal{N}_2$ that can be defined by considering the values of a predictor variable $x \in \mathcal{S}$ as follows. If x is a quantitative variable, consider all possible partitions where cases in $\mathcal{N}_1$ satisfy $x \le c$ and the cases in $\mathcal{N}_2$ satisfy $x > c$ for some value $c \in \mathbb{R}$. For a categorical predictor variable x, let $\mathcal{A}$ be the set of all the categories of x, and consider all possible partitions where $\mathcal{N}_k$ is set of cases with x in $\mathcal{A}_k$ $(k = 1, 2)$ for some disjoint partition of $\mathcal{A}$ into nonempty subsets $\mathcal{A}_1$ and $\mathcal{A}_2$. From the allowable set of partitions of the cases in $\mathcal{N}$ into subnodes $\mathcal{N}_1$ and $\mathcal{N}_2$ (each defined by a choice of variable x in $\mathcal{S}$ and either a value of $c \in \mathbb{R}$ or a disjoint partition of the categories of x), choose the partition that minimizes

   $$\sum_{k=1}^{2} \sum_{i \in \mathcal{N}_k} \left(Y_i^* - \bar{Y}_k^*\right)^2,$$

   where, for $k = 1, 2$, $\bar{Y}_k^*$ is the average response value for cases in subnode k.

   (d) For each newly created subnode $\widetilde{\mathcal{N}}$ with more than *nodesize* cases, that has variation in the values of the response and in the values of at least one predictor, repeat steps (a) through (d) with $\widetilde{\mathcal{N}}$ in place of $\mathcal{N}$. Any newly created subnode with no more than *nodesize* cases or no variation in either response or predictor vector values is split no further and is known as a *terminal node* of the tree T^*.

3. Independently repeat steps 1 and 2 a total of B times to produce trees $T_1^*, \ldots, T_B^*$ that constitute a random forest denoted as RF. (Note B may be chosen as a function of the training dataset $\boldsymbol{\mathcal{C}}_n$ [i.e., $B \equiv B(\boldsymbol{\mathcal{C}}_n)$] so that Monte Carlo variation in the random forest construction process is not an important source of variation in RF predictions. Put simply, $B \equiv B(\boldsymbol{\mathcal{C}}_n)$ should be large enough so that two random forests constructed from the same training dataset $\boldsymbol{\mathcal{C}}_n$ do not yield practically important differences in predictions for any target $\boldsymbol{x}$ vectors. See Section 2.4 of Biau and Scornet (2016) for a summary of past work on the choice of B.)

The RF point prediction of the response Y for any specified value of the predictor $\boldsymbol{X}$ is $\widehat{Y} = \frac{1}{B}\sum_{b=1}^{B} \widehat{Y}_b^*$, where $\widehat{Y}_b^*$ is the prediction of Y provided by tree T_b^* ($b = 1, \ldots, B$) in RF. Thus, the RF prediction is simply an average of the predictions for Y provided by the trees in RF. For each $b = 1, \ldots, B$, the prediction of Y by tree T_b^* (i.e., $\widehat{Y}_b^*$) is determined as follows. Tree T_b^* is defined by the splitting rules selected for each split in step 2(c) of tree construction and by the collection of cases that reside in each terminal node of the tree. By examining the values of the predictor variables in $\boldsymbol{X}$ and applying the splitting rules to those values, exactly one terminal node of tree T_b^* is identified. (Breiman et al. (2001) referred to the process of identifying the terminal node associated with $\boldsymbol{X}$ as "dropping an $\boldsymbol{X}$ down a tree," a phrase that evokes a useful conceptualization when the root node of the tree is pictured at the top of a tree diagram with the bifurcations associated with splitting rules flowing down to terminal nodes at the bottom of the tree diagram.) Once the terminal node associated with $\boldsymbol{X}$ is identified, the average of the responses for cases in that terminal node provide $\widehat{Y}_b^*$.

In the construction of each regression tree (step 2), there are two important tuning parameters that can impact performance. First, *mtry* determines how many variables are considered when defining the splitting rule at each node in a tree. Second, *nodesize* controls the termination of the tree construction process by defining the maximum terminal node size. If the number of cases in a tree node is greater than *nodesize* (and variation among the response values and predictor values for cases in the node remains), the tree-growing algorithm will split the node by drawing a simple random sample of *mtry* predictor variables and searching for the one variable among those selected that yields the best partition of the node into two subnodes. To evaluate a candidate partition of a node into two subnodes, each response value is centered on its subnode's average response value and then squared and summed across all node observations. The partition that minimizes this sum of squares is considered best. Once every node in a tree is no longer eligible for splitting due to its size or lack of within-node variation, the tree construction process terminates. Both *mtry* and *nodesize* can be tuned to strike an effective balance between variance and bias in predictions, with larger values of *mtry* and smaller values of *nodesize* tending to reduce bias at the cost of greater variance. We will later show that our prediction intervals perform well across a range of typical choices for the tuning parameters *mtry* and *nodesize*.

2.2 Random Forest Weights

For all $b = 1, \ldots, B$, the tree prediction of the bth tree, $\widehat{Y}_b^*$, is determined by finding the terminal node of T_b^* that corresponds to $\boldsymbol{X}$ and then computing the average of the response values for that terminal node. Because the ith training case may be present multiple times in a single terminal node due to bootstrap resampling with replacement, $\widehat{Y}_b^*$ is a weighted average of the original training response values given by

$$\widehat{Y}_b^* = \sum_{i=1}^{n} v_{bi}^* Y_i,$$

for some non-negative weights $v_{b1}^*, \ldots, v_{bn}^*$ that sum to 1 for each $b \in \{1, \ldots, B\}$. Thus, the random forest prediction of Y is an average of weighted averages that may be written as a weighted average of the training response values; i.e.,

$$\widehat{Y} = \frac{1}{B}\sum_{b=1}^{B} \widehat{Y}_b^* = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} v_{bi}^* Y_i = \sum_{i=1}^{n}\left(\frac{1}{B}\sum_{b=1}^{B} v_{bi}^*\right) Y_i = \boldsymbol{w}'\boldsymbol{Y}, \tag{1}$$

where $\boldsymbol{w} = [w_1, \ldots, w_n]' \equiv \left[\frac{1}{B}\sum_{b=1}^{B} v_{b1}^*, \ldots, \frac{1}{B}\sum_{b=1}^{B} v_{bn}^*\right]'$ is a vector of non-negative weights that sum to 1 and $\boldsymbol{Y} = [Y_1, \ldots, Y_n]'$. Due to the algorithm for tree construction and aggregation described in Section 2.1, the weight w_i on training response Y_i will tend to be large when $\boldsymbol{X}_i$ is *close* to $\boldsymbol{X}$, where the notion of closeness is determined in an automated way (via the tree construction process) to account for the relative importance of each component of the predictor vector. In this sense, random forests can be viewed as an adaptive nearest-neighbors prediction method (Lin and Jeon, 2006; Scornet, 2016; Wager and Athey, 2017). Aside from providing this useful interpretation of random forest predictions, random forest weights have been utilized extensively in the development of new methodologies by treating random forests as adaptive weight generators at a high level. For instance, random forest weights play a crucial role in the quantile regression forests of Meinshausen (2006), a point we explain more thoroughly in upcoming Section 4.2. Xu et al. (2016) proposed a case-specific random forest that replaces the uniform bootstrap resampling of training cases in Step 1 of the RF algorithm by a weighted bootstrap, where an initial random forest is used to generate weights specific to a predictor vector of interest. Friedberg et al. (2018) proposed a new approach to high-dimensional nonparametric regression estimation by using random forest weights to define a kernel function for local linear regression.

2.3 Out-of-bag Prediction Intervals

To establish prediction intervals for response Y based on its RF point predictor $\widehat{Y}$ from Section 2.1, we wish to learn about the distribution of the RF prediction error $D \equiv Y - \widehat{Y}$; i.e., we seek the distribution of prediction error that results when predicting a (currently unavailable) response value Y using random forest RF constructed, by necessity, without the use of $(\boldsymbol{X}, Y)$. To gain information about the prediction error distribution, we examine, for each $i = 1, \ldots, n$, the error that results when predicting the ith training response Y_i using a random forest $RF_{(i)}$ constructed without use of case $(\boldsymbol{X}_i, Y_i)$. Such a random forest is readily available for each training case i as a subset of trees from our original random forest RF. From the bootstrap sampling in step 1 of the random forest algorithm described in Section 2.1, approximately $\left(\frac{n-1}{n}\right)^n \approx \exp(-1) \approx 0.368$ of the B trees in the original forest are constructed without $(\boldsymbol{X}_i, Y_i)$. Thus, for each $i = 1, \ldots, n$, there is a subforest $RF_{(i)}$ of RF consisting of approximately $B \cdot \exp(-1)$ trees formed without the use of $(\boldsymbol{X}_i, Y_i)$. For each $i = 1, \ldots, n$, we can use $RF_{(i)}$ to obtain a prediction of Y_i, denoted as $\widehat{Y}_{(i)}$. As in equation (1), we can express $\widehat{Y}_{(i)}$ as $\boldsymbol{w}'_{(i)}\boldsymbol{Y}$, where $\boldsymbol{w}_{(i)}$ is a vector of non-negative weights that sum to 1. Following Breiman (2001), we refer to $\widehat{Y}_{(i)}$ as an out-of-bag (OOB) prediction. Likewise, we refer to the weights in $\boldsymbol{w}_{(i)}$ as OOB weights.

Note that by construction, the ith element of $\boldsymbol{w}_{(i)}$ is zero. Thus, importantly, Y_i is not involved in the OOB prediction $\widehat{Y}_{(i)}$ from forest $RF_{(i)}$, just as Y is not involved in the prediction $\widehat{Y}$ from forest RF. Consequently, the OOB prediction errors $\{D_i \equiv Y_i - \widehat{Y}_{(i)}\}_{i=1}^{n}$ provide a faithful representation of the errors incurred when generating a random forest prediction for a case independent of the training data used to construct the forest.

Because $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n), (\boldsymbol{X}, Y)$ are independent and identically distributed, the OOB prediction errors $D_1, \ldots, D_n$ are identically distributed and have approximately the same distribution as D. The distribution of D differs from the distribution of each OOB prediction error only in that $\widehat{Y}$ is based on the forest RF that involves n training observations and B trees, while each OOB prediction error is based on a forest constructed from $n-1$ observations and comprised of a random number of trees varying around the expected number $B \cdot \exp(-1)$. As n and B grow large, the difference between the distribution of D and the empirical distribution of the OOB prediction errors $D_1, \ldots, D_n$ becomes negligible, and it is reasonable to assume

$$1 - \alpha \approx \mathbb{P}\left[D_{[n,\alpha/2]} \leq D \leq D_{[n,1-\alpha/2]}\right] = \mathbb{P}\left[\widehat{Y} + D_{[n,\alpha/2]} \leq Y \leq \widehat{Y} + D_{[n,1-\alpha/2]}\right], \qquad (2)$$

where $D_{[n,\gamma]}$ is the γ quantile of the empirical distribution of $D_1, \ldots, D_n$. Expression (2) suggests $\left[\widehat{Y} + D_{[n,\alpha/2]}, \widehat{Y} + D_{[n,1-\alpha/2]}\right]$ as a prediction interval for Y with approximate coverage probability $1 - \alpha$. Section 3 provides a formal description of some asymptotic properties of this proposed OOB prediction interval.

When the distribution of D is symmetric, we recommend a slightly modified OOB prediction interval given by $\hat{Y} \pm |D|_{[n,\alpha]}$, where $|D|_{[n,\alpha]}$ is the $1-\alpha$ quantile of the empirical distribution of $|D_1|, \ldots, |D_n|$. In practice, we recommend this symmetric OOB interval unless asymmetry in the empirical distribution of $D_1, \ldots, D_n$ makes the assumption of symmetry for the distribution of D untenable. We use the symmetric version of the OOB interval throughout all the simulations and data analyses presented in this paper.

3 Asymptotic Properties of OOB Prediction Intervals

We assume the following four regularity conditions for asymptotic validity of OOB prediction intervals:

(c.1) $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} \mathbb{G}$.

(c.2) The response variable follows an additive error model; i.e., $Y = m(\boldsymbol{X}) + e$, where $m(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is an unknown mean function and e is a mean-zero error term independent of $\boldsymbol{X}$.

(c.3) The cumulative distribution function (cdf) $F(\cdot)$ of $e = Y - m(\boldsymbol{X})$ is a continuous function over $\mathbb{R}$.

(c.4) The RF prediction $\widehat{Y} \equiv \widehat{m}_n(\boldsymbol{X})$ and associated $RF_{(1)}$ OOB prediction $\widehat{Y}_{(1)} \equiv \widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ are consistent mean estimators; i.e., $\widehat{m}_n(\boldsymbol{X}) \xrightarrow{P} m(\boldsymbol{X})$ and $\widehat{m}_{n,(1)}(\boldsymbol{X}_1) \xrightarrow{P} m(\boldsymbol{X}_1)$ as $n \to \infty$.

Assumptions (c.1)–(c.3) can be viewed as a relaxation of assumptions typically made for multiple linear regression, where $m(\boldsymbol{x})$ is a linear function $\boldsymbol{x}'\boldsymbol{\beta}$ for some unknown $\boldsymbol{\beta} \in \mathbb{R}^p$ and $F(\cdot)$ is the cdf of a normal distribution with mean 0 and some unknown variance $\sigma^2 \in \mathbb{R}^+$. The assumption of consistency of the OOB estimator $\widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ in (c.4) implies consistency of the OOB estimator for any $i = 1, \ldots, n$ because $\widehat{m}_{n,(1)}(\boldsymbol{X}_1), \ldots, \widehat{m}_{n,(n)}(\boldsymbol{X}_n)$ are identically distributed by (c.1). Furthermore, consistency of $\widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ essentially entails the consistency of $\widehat{m}_n(\boldsymbol{X})$ (as the former involves a smaller forest than the latter), but these consistency conditions are each explicitly stated in (c.4) for clarity.

The study of consistency of random forests and other ensemble methods is an active area of research. Because of the complexity of the random forest algorithm described in Section 2.1, proofs of random forest consistency have been established for simplified versions of the algorithm that are more amenable to theoretical study. A history of relevant theoretical developments is outlined by Biau and Scornet (2016). In the remainder of this section, we focus on stating the properties of our OOB intervals that hold when random forests are consistent.

In this paper, the theoretical and numerical properties of prediction intervals are studied with respect to the following four coverage probability types:

- Type I: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)]$ (marginal coverage);
- Type II: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n]$ (conditional coverage given $\boldsymbol{\mathcal{C}}_n$);
- Type III: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{X} = \boldsymbol{x}]$ (conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$); and
- Type IV: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x}]$ (conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{\mathcal{C}}_n$).

The following theorems and their corollaries address these four coverage probability types that can be asymptotically guaranteed for OOB intervals. Proofs of all results are provided in the Supplemental Materials.

Theorem 1 *Under conditions (c.1) – (c.4), the* $100(1-\alpha)\%$ *out-of-bag prediction interval has asymptotically correct conditional coverage rate given* $\boldsymbol{\mathcal{C}}_n$ *for any* $\alpha \in (0,1)$*; that is,*

$$\mathbb{P}\left\{ Y \in \left[\widehat{m}_n(\boldsymbol{X}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{X}) + D_{[n,1-\alpha/2]} \right] \middle| \boldsymbol{\mathcal{C}}_n \right\} \xrightarrow{P} 1 - \alpha \tag{3}$$

as $n \to \infty$ *for any* $\alpha \in (0,1)$.

Theorem 1 is concerned with Type II coverage, i.e., conditional coverage probability given a large training dataset. This conditional coverage probability is relevant when a training dataset is in hand and interest lies in knowing the chance that an OOB prediction interval produced with this training set for a randomly drawn $\boldsymbol{X}$ will cover the random response value Y corresponding to $\boldsymbol{X}$. While Theorem 1 provides an asymptotic result, we study finite-sample properties of the OOB prediction interval for this type of conditional coverage in Section 5 by drawing a single training dataset and empirically approximating the conditional coverage probability for that training dataset. The empirical approximation is obtained by examining the proportion of OOB intervals constructed from the given training dataset that cover Y across a large number of independent $(\boldsymbol{X}, Y)$ draws from $\mathbb{G}$. The process is repeated for many training datasets to learn how conditional coverage probability varies as a function of $\boldsymbol{\mathcal{C}}_n$.

Corollary 1 *Under the conditions for Theorem 1,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{X}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{X}) + D_{[n,1-\alpha/2]}\right]\right\} \to 1-\alpha \tag{4}$$

as $n \to \infty$ for any $\alpha \in (0,1)$.

Corollary 1 is concerned with Type I coverage, i.e., the marginal coverage probability considered by Lei et al. (2018), which is the chance of drawing both training data $\boldsymbol{\mathcal{C}}_n$ and $(\boldsymbol{X}, Y) \sim \mathbb{G}$ so that the resulting prediction interval constructed from $\boldsymbol{\mathcal{C}}_n$ and $\boldsymbol{X}$ covers Y. This marginal coverage probability can be viewed as the conditional probability in Theorem 1 averaged over the distribution of $\boldsymbol{\mathcal{C}}_n$. We investigate the finite-sample properties of our OOB interval's marginal coverage in Section 5 by averaging empirical estimates of conditional coverage over a large number of training dataset drawn from the distribution of $\boldsymbol{\mathcal{C}}_n$.

Theorem 2 *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a fixed vector such that $\widehat{m}_n(\boldsymbol{x}) \overset{P}{\to} m(\boldsymbol{x})$ as $n \to \infty$, and suppose that conditions (c.1) – (c.4) hold. Then, the $100(1-\alpha)\%$ out-of-bag prediction interval has asymptotically correct conditional coverage rate given $\boldsymbol{\mathcal{C}}_n$ and $\boldsymbol{X} = \boldsymbol{x}$ for any $\alpha \in (0,1)$; that is,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{x}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{x}) + D_{[n,1-\alpha/2]}\right] \middle| \boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x}\right\} \xrightarrow{P} 1-\alpha \tag{5}$$

as $n \to \infty$ for any $\alpha \in (0,1)$.

Theorem 2 extends the conditioning on $\boldsymbol{\mathcal{C}}_n$ in Theorem 1 to conditioning on both $\boldsymbol{\mathcal{C}}_n$ and $\boldsymbol{X} = \boldsymbol{x}$. This Type IV coverage probability is relevant for a researcher who has a large training dataset in hand and a particular target value of $\boldsymbol{x}$ for which prediction of the corresponding Y (drawn from the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$) is desired. Finite-sample coverage properties for this type of conditional coverage are studied in Section 5 for selected values of $\boldsymbol{x}$.

Corollary 2 *Under the conditions for Theorem 2,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{x}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{x}) + D_{[n,1-\alpha/2]}\right] \middle| \boldsymbol{X} = \boldsymbol{x}\right\} \to 1-\alpha \tag{6}$$

as $n \to \infty$ *for any* $\alpha \in (0, 1)$.

Corollary 2 provides a relevant result for Type III coverage, i.e., conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$, which is the type of conditional coverage established by Meinshausen (2006) for quantile regression forests (see Section 4.2). The conditional coverage probability in Corollary 2 can be obtained as the expectation of the conditional coverage probability considered in Theorem 2, where the expectation is taken with respect to the distribution of the training dataset $\boldsymbol{\mathcal{C}}_n$. The finite-sample performance of OOB prediction intervals is studied for this type of conditional coverage in Section 5.

4 Alternative Random Forest Intervals

In this section, we describe two existing approaches for generating random forest prediction intervals. These methods are compared with the proposed OOB intervals in simulation and data analysis in Sections 5 and 6, respectively. To our knowledge, our comparison of these methods is the first to appear in the literature. We also mention, in Section 4.3, two recent methods for using random forests to produce a confidence interval for the conditional mean of Y given $\boldsymbol{X} = \boldsymbol{x}$.

4.1 Split Conformal Prediction Intervals

The conformal prediction interval framework originally proposed by Vovk et al. (2005, 2009) is an effective general method for generating reliable prediction intervals. However, the original conformal prediction method is computationally intensive. Lei et al. (2018) proposed a new method, called split conformal (SC) prediction, that is completely general and whose computational cost is a small fraction of the full conformal method. The algorithm for constructing a SC prediction interval using a random forest prediction is as follows:

1. Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{L}_1, \mathcal{L}_2$.

2. Build a random forest from $\{(\boldsymbol{X}_i, Y_i) : i \in \mathcal{L}_1\}$ (a subset of the full training dataset $\boldsymbol{\mathcal{C}}_n$) to obtain an estimate of the mean function $m(\cdot)$ denoted as $\widehat{m}_{n/2}(\boldsymbol{X})$.

3. For each $i \in \mathcal{L}_2$, compute the absolute residual $R_i = |Y_i - \widehat{m}_{n/2}(\boldsymbol{X})|$. Let d be the kth smallest value in $\{R_i : i \in \mathcal{L}_2\}$, where $k = \lceil (n/2+1)(1-\alpha) \rceil$.

4. The split conformal $100(1-\alpha)\%$ prediction interval for Y is $\left[\widehat{m}_{n/2}(\boldsymbol{X}) - d, \widehat{m}_{n/2}(\boldsymbol{X}) + d\right]$.

Under the assumption that $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n), (\boldsymbol{X}, Y) \overset{iid}{\sim} \mathbb{G}$ and that the residuals $\{R_i : i \in \mathcal{L}_2\}$ have a continuous joint distribution, Lei et al. (2018) prove that

$$1 - \alpha \leq \mathbb{P}\left\{Y \in \left[\widehat{m}_{n/2}(\boldsymbol{X}) - d, \widehat{m}_{n/2}(\boldsymbol{X}) + d\right]\right\} \leq 1 - \alpha + \frac{2}{n+2}. \tag{7}$$

Note that this is a very useful result because it guarantees finite-sample marginal coverage at level no less than $1-\alpha$. One potential drawback to the intervals, however, is that they are calibrated for gauging the uncertainty of prediction errors from random forests constructed from $n/2$ rather than n observations. We find that this sample splitting can result in slightly conservative finite-sample performance with regard to interval width. Nonetheless, the SC intervals do work well in our simulations and data analyses presented in Sections 5 and 6.

From a computational standpoint, SC intervals are extremely efficient compared to the original conformal method. Compared to our proposed approach, which requires the construction of only one random forest for both point prediction and interval estimation, SC intervals involve the construction of a random forest from a randomly selected half of the original training dataset. We expect that most users of random forest methodology will desire a random forest point prediction based on the *full* training dataset as well as a prediction interval. Thus, the SC approach for random forests can be viewed as requiring the construction of two forests rather than just the one needed for our random forest point prediction and OOB interval. Of course, this extra cost of a second forest can be avoided altogether for users who are satisfied with the point prediction provided by $\widehat{m}_{n/2}(\boldsymbol{X})$ in step 2 of the SC interval method that is based on a randomly selected half of the training dataset.

4.2 Quantile Regression Forest

As discussed in Section 1, a QRF (Meinshausen, 2006) can be used to estimate the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$, and quantiles from this estimated distribution can be used to form a prediction interval for Y. To understand in more detail how a QRF works, it is useful to revisit the *RF* weights $w_1, \ldots, w_n$ defined in Section 2.2. Based on the algorithm for random forest construction and the method for predicting a response value via a random forest described in Section 2.1, each *RF* weight depends on both the training dataset $\boldsymbol{\mathcal{C}}_n$ and the value of $\boldsymbol{X}$. To emphasize conditioning on $\boldsymbol{X} = \boldsymbol{x}$, we will write, throughout this section, weight w_i as $w_i(\boldsymbol{x})$ for all $i = 1, \ldots, n$.

Equation (1) from Section 2.2 shows that the *RF* prediction of Y can be viewed as the mean of a discrete distribution that places probability $w_i(\boldsymbol{x})$ on Y_i for all $i = 1, \ldots, n$. A QRF uses this discrete distribution as an estimate of the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$. Specifically, write $I(\cdot)$ to denote an indicator function and let $\widehat{H}_n(y|\boldsymbol{x}) = \sum_{i=1}^{n} w_i(\boldsymbol{x}) I(Y_i \leq y)$ serve as an estimator of $H(y|\boldsymbol{x}) \equiv \mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x})$, the conditional cdf of Y given $\boldsymbol{X} = \boldsymbol{x}$. For $\alpha \in (0, 1)$, let $\widehat{Q}_\alpha(\boldsymbol{x}) \equiv \inf\{y \in \mathbb{R} : \widehat{H}_n(y|\boldsymbol{x}) \geq \alpha\}$ denote the α-quantile of the estimated conditional distribution Y given $\boldsymbol{X} = \boldsymbol{x}$. Then, a QRF-based $100(1 - \alpha)\%$ prediction interval for $\boldsymbol{Y}$ is given by $[\widehat{Q}_{\alpha/2}(\boldsymbol{x}), \widehat{Q}_{1-\alpha/2}(\boldsymbol{x})]$. Under regularity conditions and a few simplifying assumptions, Meinshausen (2006) showed that, for any given $\boldsymbol{x}$, the absolute error of the QRF conditional cdf approximation converges uniformly in probability to 0 as $n \to \infty$. Furthermore, an analysis of five datasets in Meinshausen (2006) shows average coverage rates for 95% QRF intervals ranging from 90.2% to 98.6% in five-fold cross-validation analysis. We investigate the performance of QRF prediction intervals relative to SC intervals and our proposed OOB intervals in Sections 5 and 6.

4.3 Confidence Intervals

Wager et al. (2014) use ideas from Efron (1992) and Efron (2014) to develop bias-corrected versions of *Infinitesimal Jackknife* and *Jackknife-after-Bootstrap* estimates of $\text{Var}[\widehat{m}_n(\boldsymbol{x})]$, the variance of the random forest estimator of $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. Because the jackknife-after-bootstrap estimator makes explicit use of OOB tree predictions, there are similarities with our proposed procedure. Although Wager et al. (2014) primarily focus on how well proposed estimators approximate $\mathrm{Var}[\widehat{m}_n(\boldsymbol{x})]$, a footnote regarding intervals displayed in Figure 1 of Wager et al. (2014) proposes a confidence interval of the form $\widehat{m}_n(\boldsymbol{x}) \pm z_\alpha \widehat{\sigma}(\boldsymbol{x})$, where z_α is the $1-\alpha$ quantile of the standard normal distribution and $\widehat{\sigma}(\boldsymbol{x})$ is a standard error computed by taking the square root of the average of jackknife and infinitesimal jackknife estimators of $\mathrm{Var}[\widehat{m}_n(\boldsymbol{x})]$. This interval could be expected to provide coverage of $\mathbb{E}[\widehat{m}_n(\boldsymbol{x})]$ with confidence level approximately equal to $100(1-\alpha)\%$ under the assumption that $\widehat{m}_n(\boldsymbol{x})$ is approximately normal with variance $\widehat{\sigma}^2(\boldsymbol{x})$.

Another approach for constructing confidence intervals from a procedure similar to random forests is proposed in Mentch and Hooker (2016). Instead of aggregating over trees built from full bootstrap samples of size n, Mentch and Hooker (2016) average over trees built on random subsamples of the training dataset and demonstrate that the resulting estimator takes the form of an asymptotically normal incomplete U-statistic. Furthermore, Mentch and Hooker (2016) develop a consistent estimator for the variance of the relevant limiting normal distribution that naturally leads to a confidence interval for the mean of their estimator.

The intervals of Wager et al. (2014) and Mentch and Hooker (2016) are confidence intervals for the expected value of estimators of $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$. When the estimators they consider are unbiased (or at least $\sqrt{n}$-consistent) for $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$, their proposed intervals serve as confidence intervals for $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$. Because our focus is on prediction intervals for Y (conditional mean plus random error) that are necessarily wider than confidence intervals for $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$, we do not consider these confidence intervals further in the current paper.

5 Simulation Study

In this section, we use simulated examples to illustrate the finite-sample performance of our proposed OOB prediction intervals. We compare OOB, SC and QRF interval widths and their Type I through IV coverage rates introduced in Section 3. The R package *conformalInference* is used to construct split conformal prediction intervals, and the R package *quantregForest* is used to build quantile regression forests.

We simulate data from an additive error model: $Y = m(\boldsymbol{X}) + \epsilon$, where the predictor

$\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ with $p = 10$ and ϵ is the error term. The distribution of predictor vector $\boldsymbol{X}$, the distribution of error term ϵ, the mean function $m(\cdot)$, and the training sample size n may all affect the performance of prediction intervals. In our simulation study, a factorial design is considered for these four factors:

- Mean functions : $m(\boldsymbol{x}) = x_1 + x_2$ (*linear*), $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|)$ (*nonlinear*), and $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|) + x_1 x_2$ (*nonlinear with interaction*).

- Distributions of errors: $\epsilon \sim N(0, 1)$ (*homoscedastic*), $\epsilon \sim t_3/\sqrt{3}$ (*heavy-tailed*), $\epsilon \sim \mathrm{N}\left(0, \frac{1}{2} + \frac{1}{2}\frac{|m(\boldsymbol{X})|}{E|m(\boldsymbol{X})|}\right)$ (*heteroscedastic*).

- Distributions of predictors: $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (*uncorrelated*), and $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$ (*correlated*), where $\boldsymbol{\Sigma}_p$ is an AR(1) covariance matrix with $\rho = 0.6$ and diagonal values equal to 1.

- Training sample sizes: $n = 200, 500, 1000, 2000$, and 5000.

The full-factorial design results in 90 different simulation scenarios. For each of the 90 scenarios, the random forest tuning parameters are selected from $mtry \in \{1, \ldots, 10\}$ and $nodesize \in \{1, \ldots, 5\}$ to minimize average cross-validated mean squared prediction error over five-fold cross-validation for 10 randomly generated datasets. The selected tuning parameters for any given scenario are then used for construction of all random forests and intervals for each dataset simulated according to that scenario. Dataset-specific adaptive tuning and performance for different choices of *mtry* and *nodesize* is studied in Section 6. The number of trees is 2000 for all random forests built in the simulation study (Oshiro et al., 2012). Following Lei et al. (2018), we set the nominal level at 0.9 for all prediction intervals constructed in this section.

5.1 Evaluating Type I and II coverage rates

To evaluate the Type I and II coverage rates, we simulate 200 datasets for each of our 90 simulation scenarios. Each dataset consists of training cases (n = 200, 500, 1000, 2000, or 5000) and 500 test cases randomly and independently generated from the joint distribution of $(\boldsymbol{X}, Y)$. For each interval method and each simulated dataset, Type II coverage is estimated by calculating the percentage of 500 test case response values contained in their prediction intervals. Type I coverage for each simulation scenario and interval method is estimated by averaging over the 200 Type II coverage estimates obtained from the 200 simulated datasets for each simulation scenario. Because results for scenarios involving uncorrelated predictors lead to the same conclusions as results for correlated predictors, figures for the former are displayed in the Supplemental Materials.

Figure 1 and Figure S.1 summarize the Type I and II coverage rate estimates for OOB, SC and QRF intervals for all training sample sizes and data-generating models. Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot. This average represents the empirical Type I coverage rate for any given scenario. Estimates of the Type I coverage rates of OOB and SC prediction intervals are very close to 0.9 (the nominal level). In contrast, QRF prediction intervals are more likely to over-cover or under-cover target response in terms of Type I coverage. As the sample size n increases, the OOB and SC Type II coverage rate estimates show decreased variation and become more concentrated around 0.9. Additionally, the coverage rates of OOB and SC prediction intervals are stable across the mean functions, predictor correlations, and measurement error distributions in our simulation study.

Given the random forest for any simulated dataset, OOB interval width is the same for all test cases. Similarly, the SC method produces intervals of constant width across test cases. On the other hand, the width of QRF intervals varies across test cases. Thus, for each simulated dataset, we record one OOB interval width, one SC interval width, and 500 QRF interval widths. To compare the interval widths of these three methods, we average the 500 QRF interval widths for each simulated dataset. Boxplots summarizing the distributions of interval widths are provided in Figure S.2 and Figure S.3. To provide a clearer comparison of interval widths, we compute the ratio of the SC interval width relative to the OOB interval width and the ratio of the average QRF interval width to the OOB interval width for each simulated dataset. Boxplots of the $\log_2$ transformation of the ratios are presented in Figure 2 and Figure S.4. Figure S.2 and Figure S.3 show that the interval widths shrink as sample size increases. Figure 2 and Figure S.4 indicate that OOB prediction intervals tend to be narrower than intervals produced by competing methods.

The only exceptions occur when QRF intervals have coverage rates substantially below the nominal level.

5.2 Evaluating Type III and IV coverage rates

The simulation settings for evaluating the Type III and IV coverage probabilities are the same as in Section 5.1 except that no test cases are simulated. Instead, for each simulated training dataset, OOB, SC and QRF prediction intervals are generated for $\boldsymbol{X} = \boldsymbol{x}$, where $\boldsymbol{x}$ is a specified 10-dimensional predictor vector. Using the known conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ for the given simulation scenario, we compute the exact Type IV coverage probability for each interval. The Type III coverage rate for any interval method and simulation scenario is then estimated by averaging over the 200 Type IV coverage rate estimates computed from the 200 training datasets simulated for that scenario.

Figures 3, 4, S.5, and S.6 show the boxplots of Type IV coverage rate estimates, i.e., estimates of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x}]$ for OOB, SC and QRF prediction intervals and $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{1}$ (10-dimensional vectors of zeros and ones, respectively). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot. This average represents the empirical Type III coverage rate for any given scenario. As in the Type I and II coverage results presented in Section 5.1, we see that OOB and SC intervals perform similarly across all scenarios with respect to Type III and IV coverage. In contrast, QRF intervals tend to be more variable within scenarios than OOB and SC intervals in terms of Type IV coverage and display Type III coverage values that often differ from the corresponding values for OOB and SC intervals. QRF intervals clearly perform better for some scenarios (*Linear*×*Heteroscedastic* scenarios, for example) and worse for others (e.g., seven of the nine panels in Figure 3).

Aside from the size of the training dataset n, major factors that affect finite-sample Type III and IV coverage include the shape of the mean function $m(\cdot)$ in a neighborhood of $\boldsymbol{x}$ and $\text{Var}(\epsilon|\boldsymbol{X} = \boldsymbol{x})$ relative to $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$ when error variance is heteroscedastic. To understand the impact of these factors, consider simulation scenarios involving the nonlinear mean function $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|)$. This nonlinear function achieves a global maximum at $\boldsymbol{x} = \boldsymbol{0}$. Because $\mathbb{P}\{m(\boldsymbol{X}) < m(\boldsymbol{0})\} = 1$, each training case has a conditional

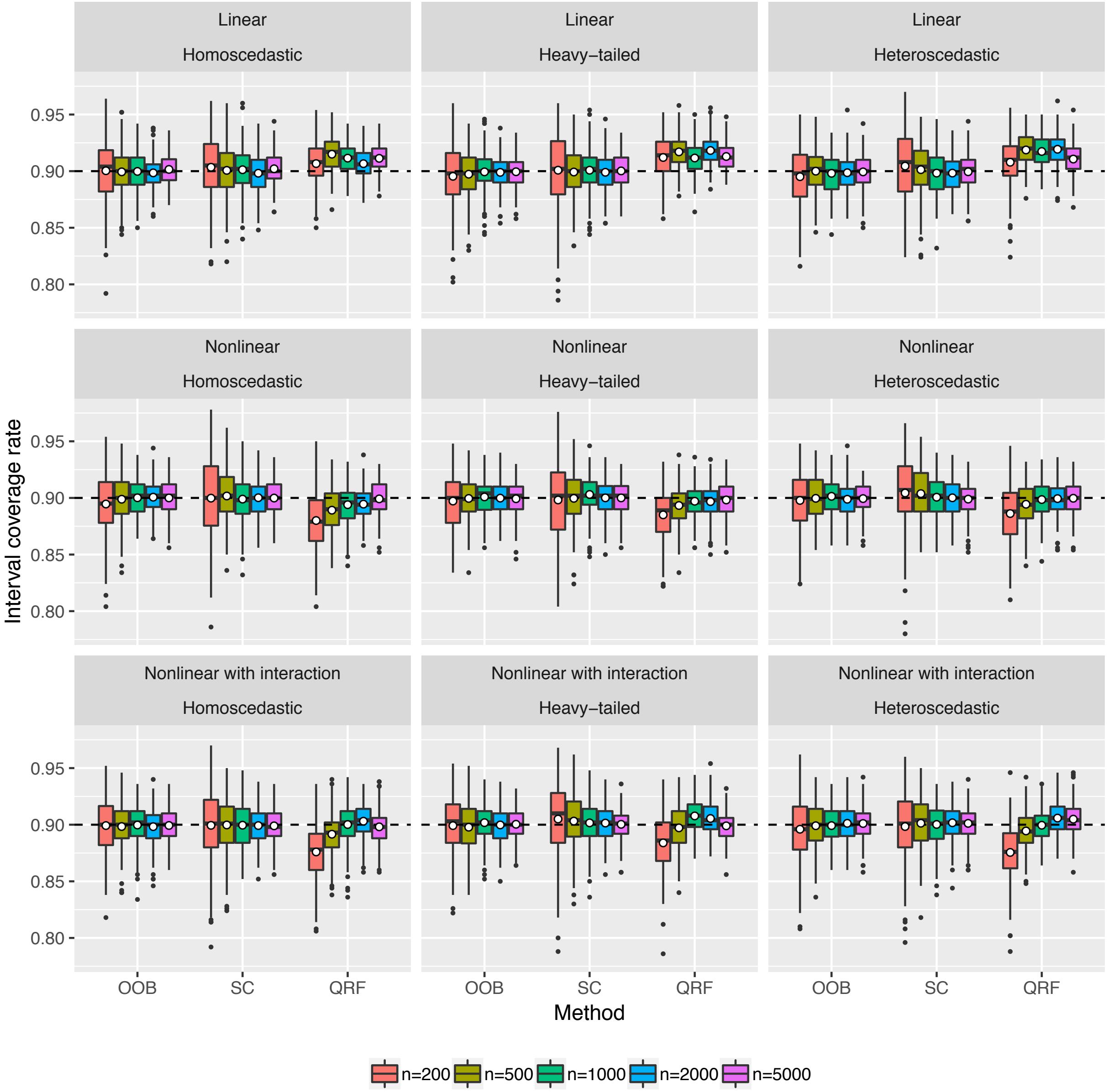


Figure 1: Boxplots of the Type II coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n]$, of out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot, and represents an estimate of Type I coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)]$.

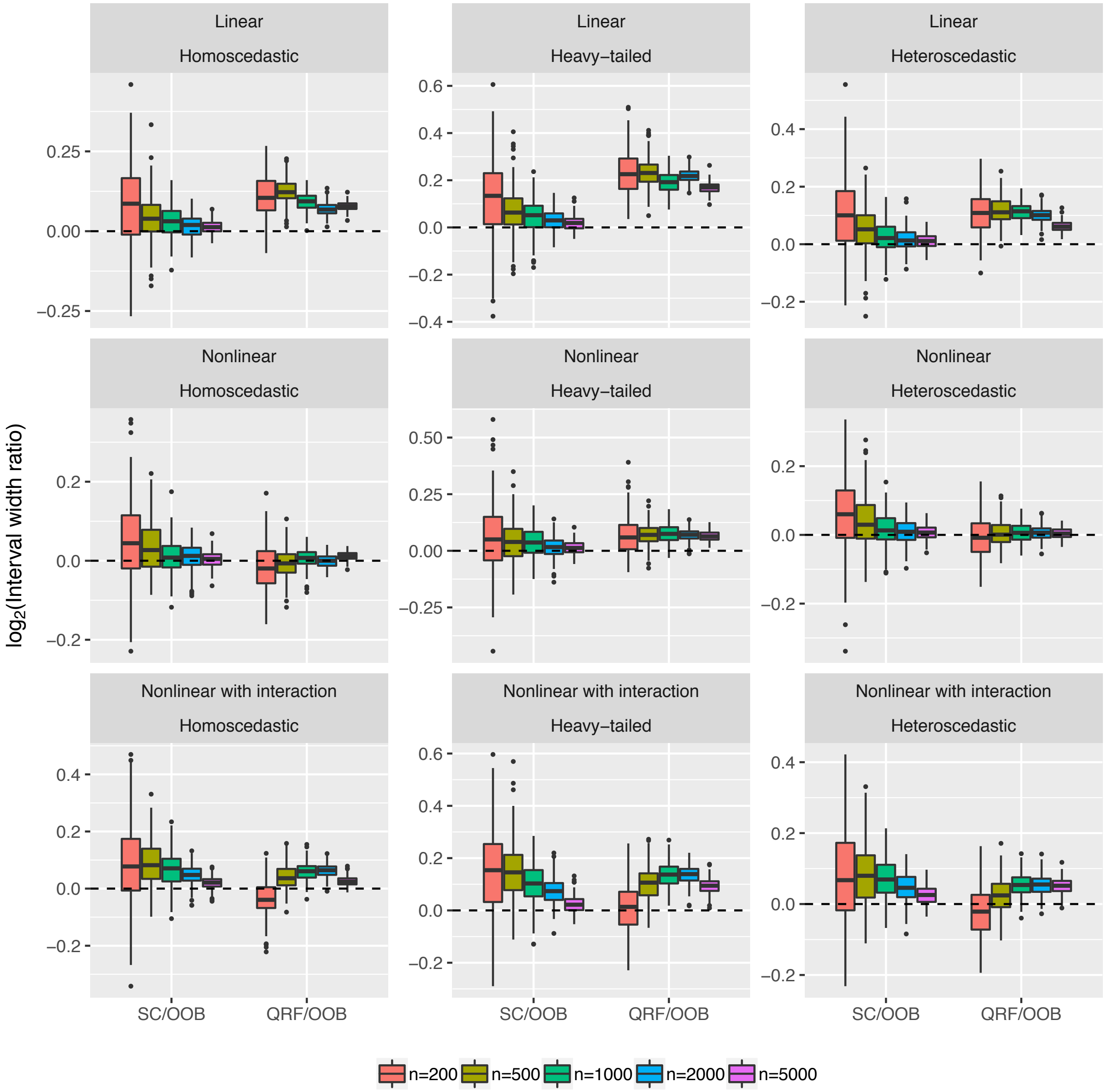


Figure 2: Boxplots of the $\log_2$ ratios of split conformal (SC) interval widths to out-of-bag (OOB) interval widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval widths when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors).

mean response strictly less than $m(\mathbf{0})$ with probability one (i.e., $\mathbb{P}_{\boldsymbol{X}_i}\{\mathbb{E}(Y_i|\boldsymbol{X}_i) < m(\mathbf{0})\} = 1$ for all $i = 1, \ldots, n$). Because a random forest prediction is simply a weighted average of training responses (as discussed in Section 2.2), the random forest estimator of $m(\mathbf{0})$ has expectation less than $m(\mathbf{0})$. This bias at $\boldsymbol{x} = \mathbf{0}$ leads to larger prediction errors at $\boldsymbol{x} = \mathbf{0}$ than for other points in the predictor domain and under-coverage for OOB, SC and QRF intervals visible in the middle row of Figure 3.

The under-coverage problem at $\boldsymbol{x} = \mathbf{0}$ in the nonlinear case is exacerbated for OOB and SC intervals for the heteroscedastic case. The OOB and SC intervals rely on a single distribution of prediction errors estimated by combining information from prediction errors made throughout the training dataset rather than the prediction errors made at any specified $\boldsymbol{x}$ vector. Thus, all else equal, an OOB or SC prediction interval will tend to over-cover response values at a value $\boldsymbol{x}$ for which the error variance is relatively low and under-cover response values at a value $\boldsymbol{x}$ for which the error variance is relatively high. For the *Nonlinear*×*Heteroscedastic* case with $\boldsymbol{x} = \mathbf{0}$, $\text{Var}(\epsilon|\boldsymbol{X} = \mathbf{0})$ is more than twice $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$, the mean error variance over the predictor space. Thus, the severe under-coverage of OOB and SC intervals in the second row and third column of Figure 3 is as expected due to both underestimation of the mean function and relatively large error variance at $\boldsymbol{x} = \mathbf{0}$. Although QRF intervals suffer from the same random forest bias problem that plagues OOB and SC intervals, the adaptive width of QRF intervals typically provides improved Type III and IV coverage results for QRF intervals relative to OOB and SC intervals in heteroscedastic scenarios.

For prediction at $\boldsymbol{x} = \mathbf{1}$, the second row of Figure 4 shows improved performance for all intervals relative to the $\boldsymbol{x} = \mathbf{0}$ case. Random forest bias at $\boldsymbol{x} = \mathbf{1}$ is relatively minimal because the average value of $m(\boldsymbol{x})$ for $\boldsymbol{x}$ near $\mathbf{1}$ is relatively close to $m(\mathbf{1})$. This leads to Type III and IV coverages near the nominal 0.90 level for the homoscedastic and heavy-tailed scenarios. Over-coverage for OOB and SC intervals results for the *Nonlinear*×*Heteroscedastic* case in Figure 4 because the error variance at $\boldsymbol{x} = \mathbf{1}$ is less than 75% of the mean error variance $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$. The Type III and IV coverage results for OOB intervals presented in Figures 3, 4, S.5, and S.6 are as expected when considering the shape of the mean function near $\boldsymbol{x}$ and the value of $\text{Var}(\epsilon|\boldsymbol{X} = \boldsymbol{x})$ relative to

$\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$ in each scenario.

In response to a referee's comment, we have generated Figures S.7 and S.8 that evaluate Type III and IV coverage at $\boldsymbol{x} = \boldsymbol{x}_3 \equiv (3, -3, 3, \ldots, 3)'$. Whether predictor variables are correlated or uncorrelated, the multivariate normal distribution of $\boldsymbol{X}$ in our simulation study assigns very low probability to neighborhoods containing $\boldsymbol{x}_3$. Thus, most simulated training datasets will contain no observations in close proximity to $\boldsymbol{x}_3$. Nonetheless, a random forest predictor will find "nearest neighbors" in the training dataset as those with the highest weights in (1). The resulting extrapolation may or may not work well, depending on the true mean function $m(\cdot)$. Figures S.7 and S.8 show that OOB and SC intervals have highly variable Type IV coverage and Type III coverage near (but often below) the nominal level for linear and nonlinear scenarios. For the scenarios involving the nonlinear mean function with interaction, the Type III and IV coverage levels for OOB and SC intervals are estimated to be zero or near zero. This is not surprising considering that $m(\boldsymbol{X})$ tends to be much greater than $m(\boldsymbol{x}_3)$ with probability near one when $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|) + x_1x_2$. Thus, regardless of the training observations that receive the greatest weight in (1), the random forest prediction is likely to be substantially greater than $m(\boldsymbol{x}_3)$ so that large prediction errors are likely. QRF intervals are wide and over-cover for our linear and nonlinear scenarios and show severe under-coverage for the nonlinear scenarios with interaction. None of the prediction interval approaches we have studied can be recommended for prediction in a region of the predictor space where no training data are available, but we know of no approach that can be generally trusted for such extrapolation.

6 Data Analysis

In this section, we compare the performance of OOB, SC and QRF prediction intervals on 60 actual datasets, summarized in Table 1. The majority of the datasets (40 out of 60) were analyzed by Chipman et al. (2010). The other 20 datasets come from the UC Irvine Machine Learning Repository website. These datasets span various application areas, including biological science, physical science, social science, engineering, and business. Sample sizes range from 96 to 45730, and the number of predictors ranges from 3 to 100. Prior to analysis, we standardize the response variable for each dataset to make the interval widths

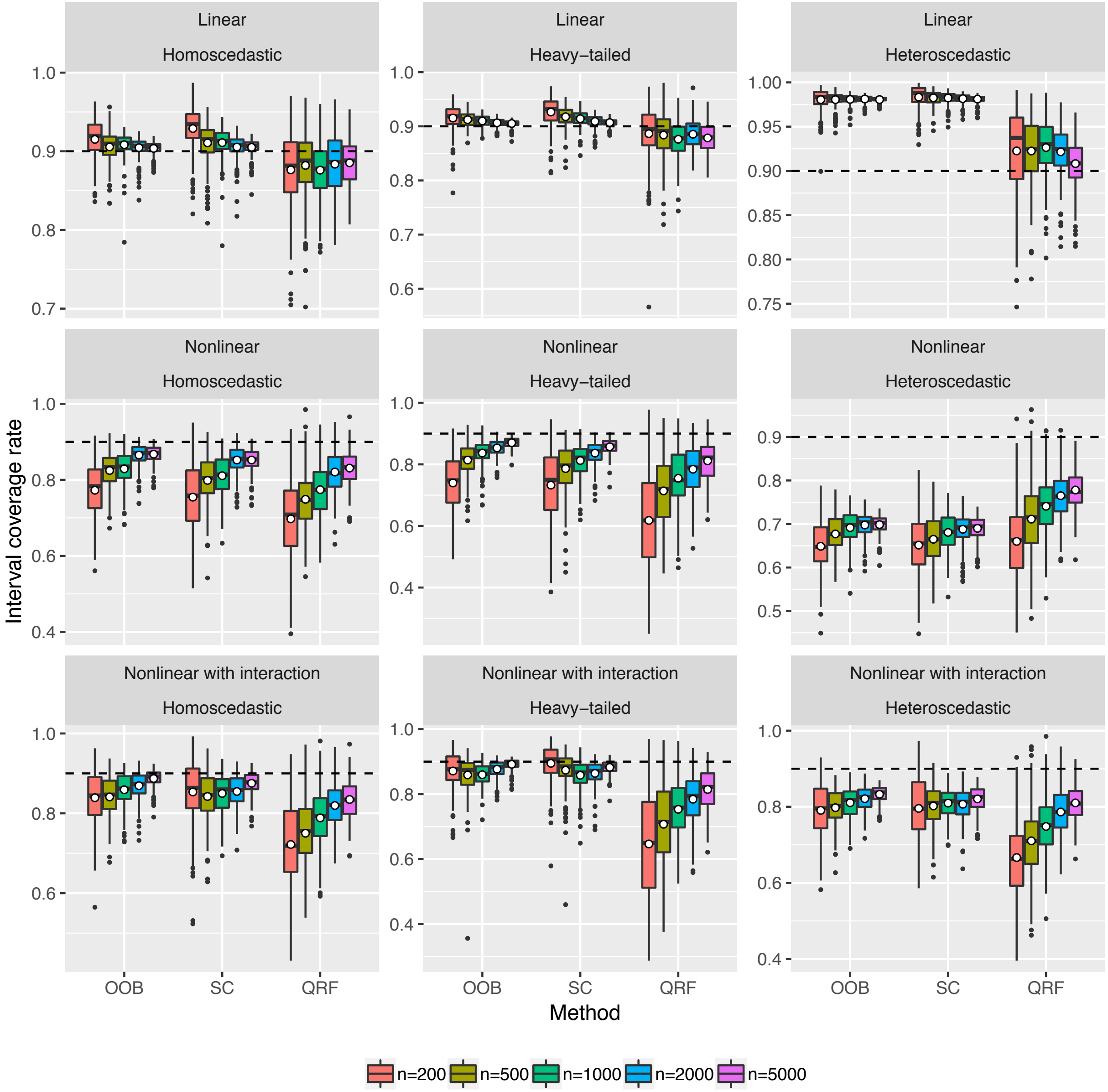


Figure 3: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \mathbf{0}]$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{X} = \mathbf{0}]$.

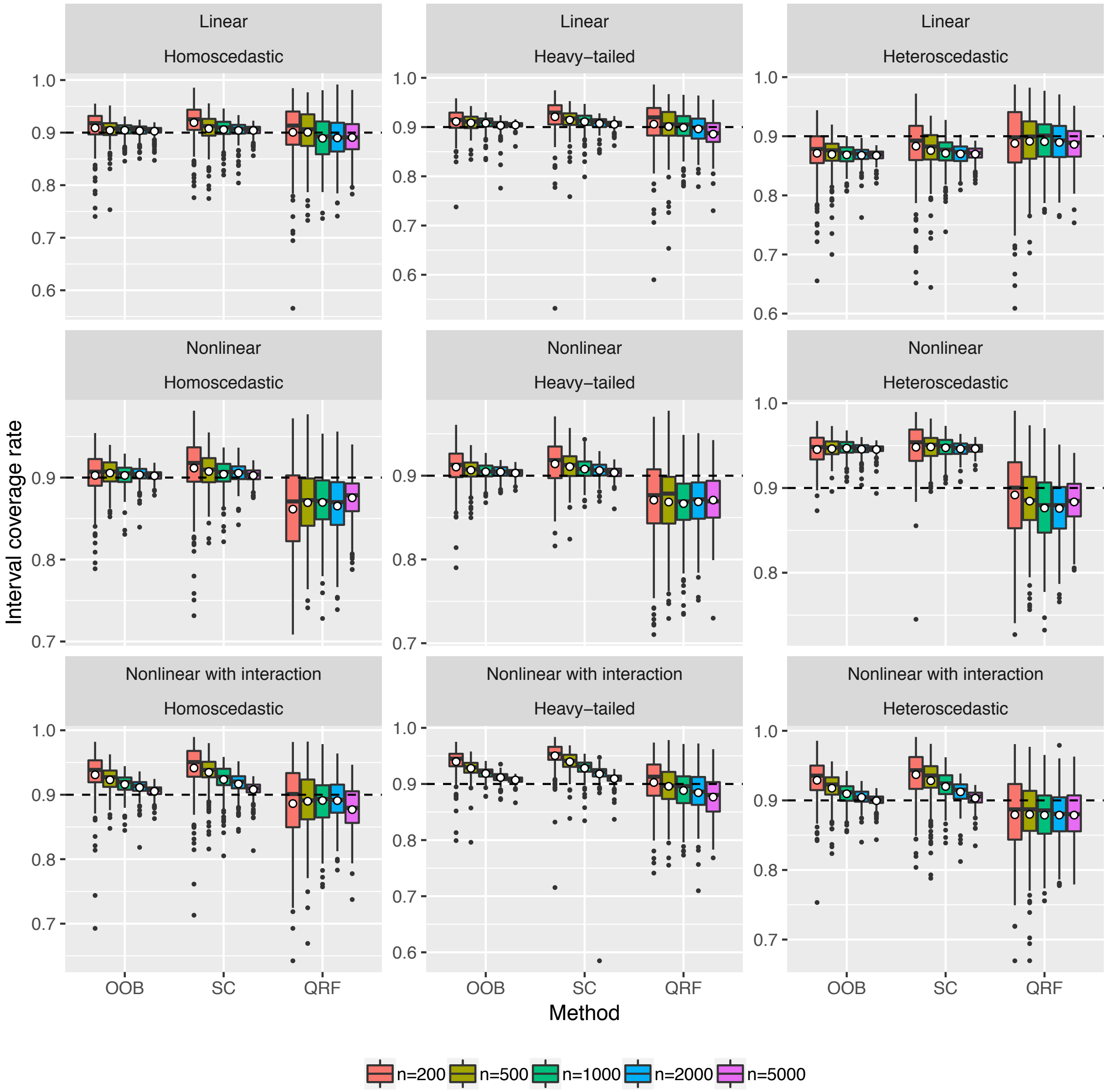


Figure 4: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \mathbf{1}]$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{X} = \mathbf{1}]$.

for different datasets more comparable. Cases with one or more missing values are omitted. The number of trees is 2000 for all random forests built in this section, and the nominal coverage rate is set at 0.9.

Because the repeated measures of the response variable given a fixed predictor vector $\boldsymbol{X} = \boldsymbol{x}$ are not common in these datasets, Type III and IV coverage probabilities are difficult to evaluate. Thus, only Type I and II coverage probabilities are considered in this section. Our approach to empirically assess Type I and II coverage probabilities is through five-fold cross validation. For each run of five-fold cross validation, we randomly partition the whole dataset into five non-overlapping parts. Four parts are combined to form a training set that is used to compute prediction intervals for the response values of cases in the fifth part. Then we calculate the percentages of response values in the fifth part contained by their intervals to approximate Type II coverage rate. All $\binom{5}{4}$ training/test sets are analyzed for each partition, and a total of 20 random partitions are analyzed for each dataset. For each dataset and method, this process yields 100 empirical Type II coverage rates, which can be averaged to obtain an empirical Type I coverage rate.

The empirical coverage rates (Type I: circles, Type II: boxplots) for all three methods for all 60 datasets are presented in Figure S.9 - S.11. Figure 5 shows a summary of all the Type II coverage rate estimates with datasets on the horizontal axis in ascending order by the average value of the OOB, SC and QRF Type I coverage rate estimates. Relative interval widths are summarized in Figure 6, where we present the $\log_2$ ratio of the average width of SC intervals to the average width of OOB intervals, and the average width of QRF intervals to the average width of OOB intervals. The order of datasets in Figure 6 is the same as the order in Figure 5.

The findings from real data analysis are consistent with the conclusions made in the simulation study. Both the OOB prediction intervals and the SC prediction intervals have good Type I coverage rates centered at 0.9, but the Type I coverage rate of QRF intervals deviate substantially from 0.9 for many datasets. Furthermore, OOB prediction intervals are narrower than SC prediction intervals for almost all 60 datasets, and the widths of OOB prediction intervals tend to be similar to or narrower than QRF interval widths. The few exceptions occur for datasets with QRF coverage rate estimates well below 0.9.

For the data analysis results presented so far in this section, the *mtry* and *nodesize* tuning parameters of random forests are selected for each dataset by five-fold cross validation to minimize cross-validated mean squared prediction error over (*mtry*, *nodesize*) $\in \left\{ \left\lceil \frac{1}{2} \left\lfloor \frac{p}{3} \right\rfloor \right\rceil, \left\lfloor \frac{p}{3} \right\rfloor, 2 \left\lfloor \frac{p}{3} \right\rfloor \right\} \times \{1, 5\} = \{2, 3, 6\} \times \{1, 5\}$, following the advice of Breiman as recounted by Liaw et al. (2002). The tuning parameters are then fixed at the selected values during the subsequent OOB, SC and QRF interval evaluation (which also involves five-fold cross-validation, although five-fold cross-validation is repeated 20 times for coverage probability estimation). To show how the three prediction intervals adapt to other choices of the random forest tuning parameters, we evaluate the performance of the prediction intervals on one real data example, the Concrete Strength dataset from UCI, for each combination of *nodesize* $\in \{1, 5\}$ and *mtry* $\in \{2, 4, 6, 8\}$. The results are illustrated in Figure 7. As in our other analyses, OOB and SC prediction intervals tend to cover close to 90% of the test case response values on average, and OOB intervals are narrower than both SC and QRF intervals regardless of the *mtry* and *nodesize* values. The QRF intervals have estimated Type I coverage rates sometimes above and sometimes below the nominal level depending on the tuning parameter values. Both the OOB and SC intervals show stable performance across tuning parameter values, while QRF intervals are sensitive to the choice of tuning parameters in terms of coverage and width. Overall, the OOB intervals perform uniformly best across the investigated tuning parameter values for this dataset.

7 Concluding Remarks

We propose OOB prediction intervals as a straightforward technique for constructing prediction intervals from a single random forest and its by-products. We have provided theory that guarantees asymptotic coverage (of various types) for OOB intervals under regularity conditions. Our simulation analysis in Section 5 and our analysis of 60 datasets in Section 6 provide evidence for reliability and efficiency of OOB intervals across a wide range of sample sizes and scenarios that do not necessarily conform to the assumptions required for our theorems. Thus, the performance record for OOB intervals established in this paper indicates that OOB prediction intervals can be used with confidence for a wide array of practical problems.

Table 1: Name, n = total number of observations (excluding observations with missing values), and p = number of predictor variables for 60 datasets.

No.	Name	n	p	No.	Name	n	p
1	Abalone	4177	8	31	Facebook Metrics	495	17
2	Air Quality	9357	12	32	Fame	1318	22
3	Airfoil Self-Noise	1503	5	33	Fat	252	14
4	Ais	202	12	34	Fishery	6806	14
5	Alcohol	2462	18	35	Hatco	100	13
6	Amenity	3044	21	36	Hydrodynamics	308	6
7	Attend	838	9	37	Insur	2182	6
8	Auto MPG	392	7	38	Istanbul Stock	536	6
9	Automobile	159	18	39	Laheart	200	16
10	Baseball	263	20	40	Medicare	4406	21
11	Basketball	96	4	41	Mumps	1523	3
12	Beijing PM2.5	41757	11	42	Mussels	201	4
13	Boston	506	13	43	Naval Propulsion Plants	11934	16
14	Budget	1729	10	44	Optical Network	630	9
15	Cane	3775	9	45	Ozone	330	8
16	Cardio	375	9	46	Parkinsons	5875	21
17	College	694	24	47	PM2.5 of Five Cities	21436	9
18	Community Crime	1994	100	48	Price	159	15
19	Computer Hardware	209	6	49	Protein Structure	45730	9
20	Concrete Strength	1030	8	50	Rate	144	9
21	Concrete Slump Test	103	9	51	Rice	171	15
22	Cps	534	10	52	Scenic	113	10
23	CPU	209	7	53	Servo	167	4
24	Cycle Power Plant	9568	4	54	SML2010	4137	21
25	Deer	654	13	55	Smsa	141	10
26	Diabetes	375	15	56	Strike	625	5
27	Diamond	308	4	57	Tecator	215	10
28	Edu	1400	5	58	Tree	100	8
29	Energy Efficiency	768	8	59	Triazine	186	28
30	Enroll	258	6	60	Wage	3380	13

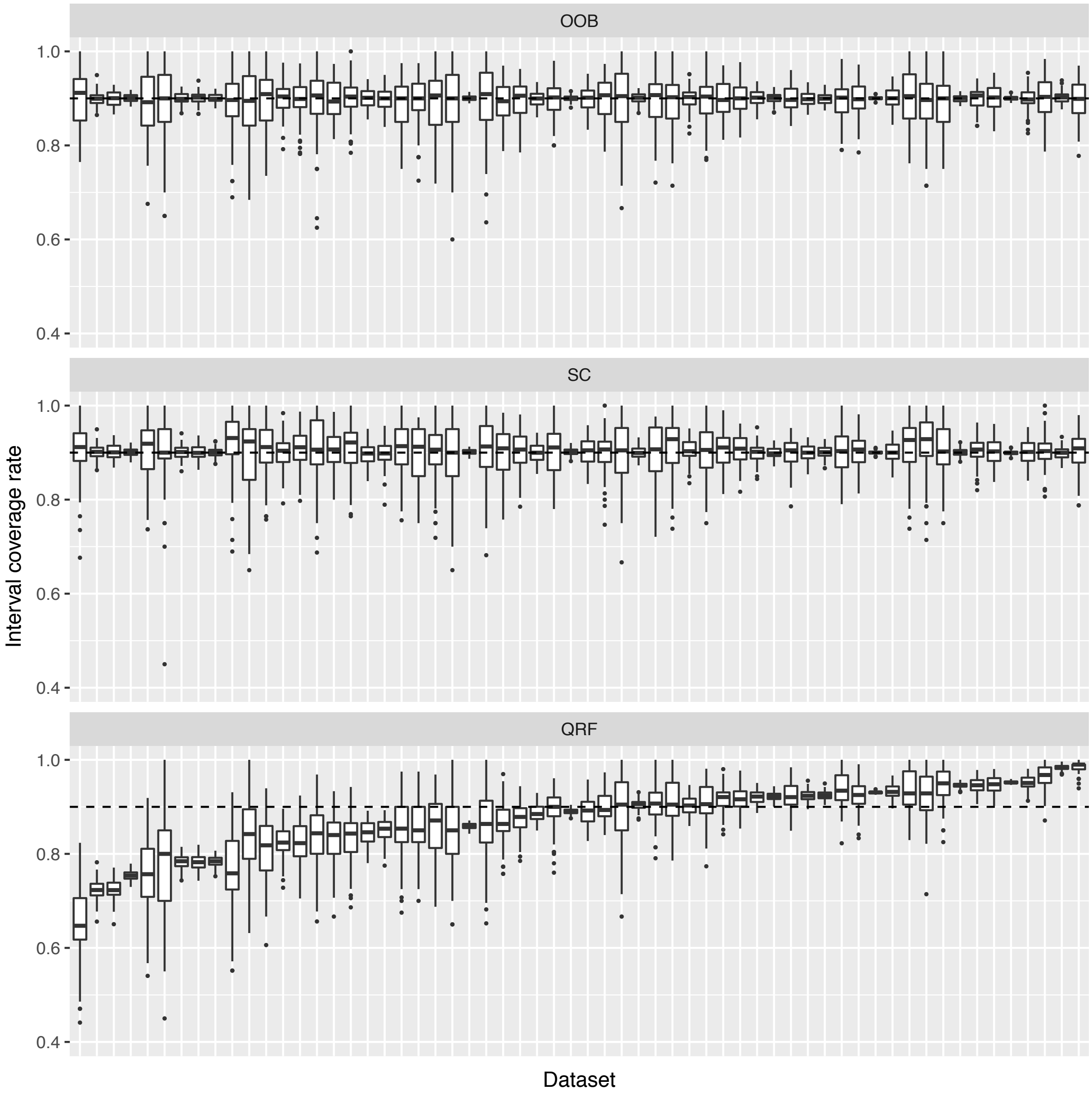


Figure 5: Boxplots of Type II coverage rates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 60 datasets. The ordering of the datasets on the horizontal axis is the same for all three panels and is determined by the average Type I coverage rates of OOB, SC and QRF prediction intervals.

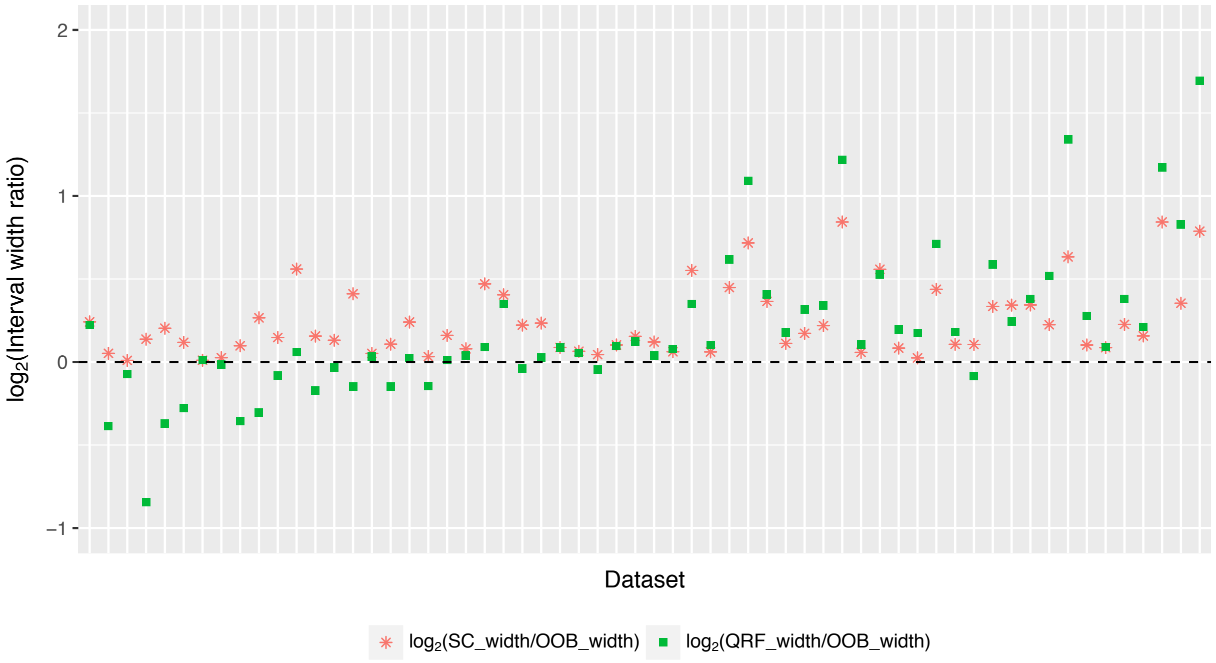
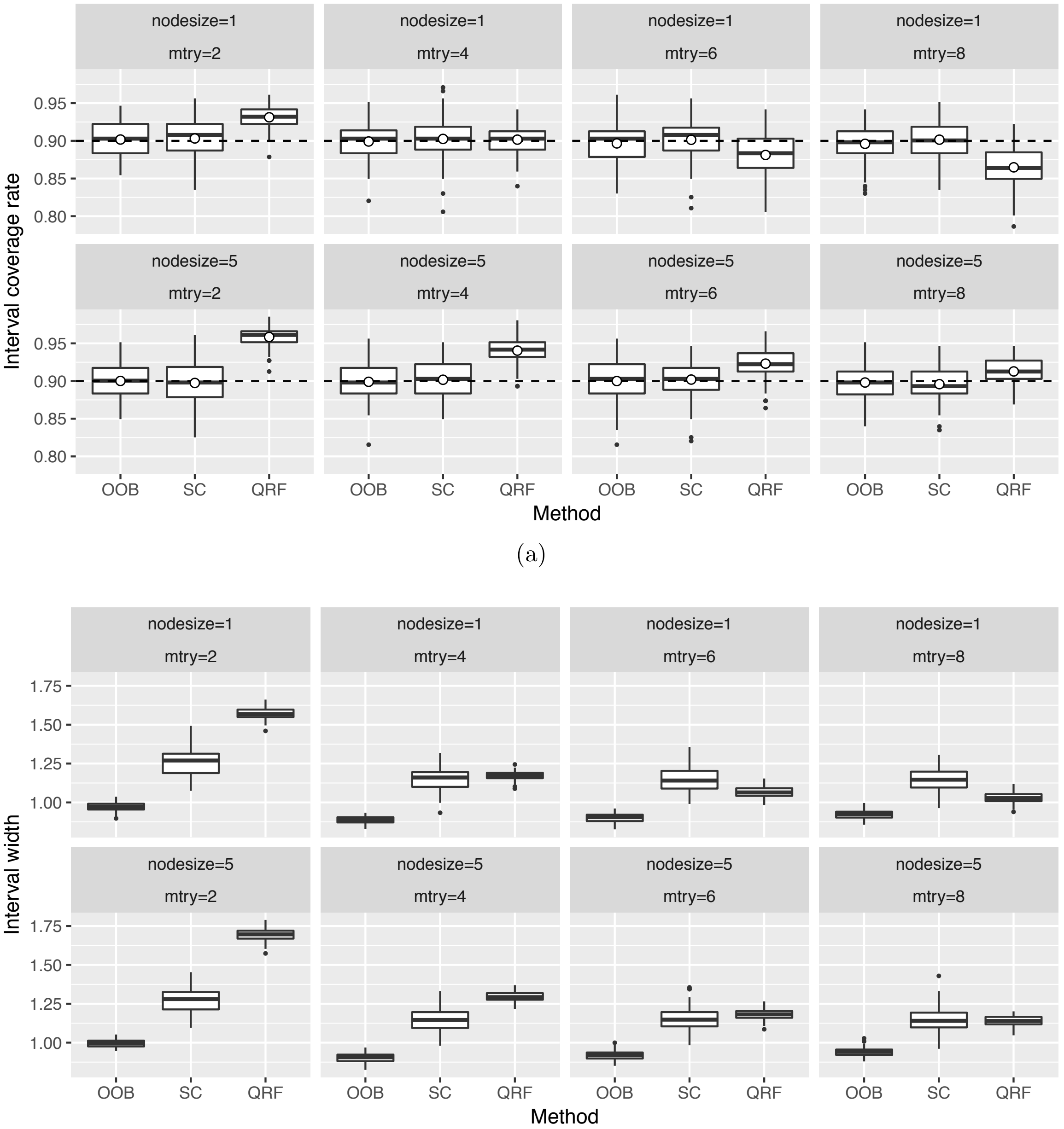


Figure 6: A plot of the $\log_2$ ratios of split conformal (SC) interval width averages to out-of-bag (OOB) interval width averages, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval width averages for 60 datasets.

Our numerical results show that QRF prediction intervals tend to have Type I and Type II coverage rates that deviate from the nominal level, sometimes over-covering and sometimes under-covering target response values, more often than the other methods we studied. Furthermore, when QRF intervals do cover at the nominal Type I or Type II rate, they tend to be wider than OOB intervals. In most of our simulation scenarios involving heteroscedastic errors, QRF prediction intervals outperformed OOB and SC intervals with respect to Type III and Type IV coverage. This is not surprising because QRF intervals are designed to provide Type III coverage, while SC intervals are only guaranteed to provide marginal (Type I) coverage. Furthermore, the theorems presented in this paper – that guarantee asymptotically correct coverage rates for OOB intervals – rely on an assumption of homoscedasticity. Nonetheless, OOB and SC intervals outperform QRF intervals with respect to Type III and IV coverage in some of our simulation scenarios involving heteroscedasticity (and in most scenarios involving homoscedasticity).

To assess the validity of the homoscedasticity assumption for any particular dataset, we suggest examining a residual plot of OOB prediction errors against estimated mean values.



(a)

(b)

Figure 7: The effect of tuning parameters on prediction intervals for the example of Concrete Strength dataset: (a) boxplots of Type II coverage rates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals under different combinations of *mtry* and *nodesize*; (b) boxplots of interval widths for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals under different combinations of *mtry* and *nodesize*.

Other variations on residual plots – e.g., plots of OOB prediction errors vs. important predictors, plots of absolute OOB prediction errors vs. estimated mean values, etc. – may also be used to identify discrepancies between assumptions and data. As in traditional multivariate linear regression, a transformation of the response variable may be useful for variance stabilization. In some cases, such transformations may be unavailable or undesirable. In these situations, simple modifications to our approach as in Lei et al. (2018) can be made to account for nonconstant error variance. More specifically, Lei et al. (2018) provide an extension to SC inference, known as Locally Weighted Conformal Inference, that yields prediction intervals with good empirical coverage properties when the error variance is a function of the predictor vector. A completely analogous technique can be used to improve the performance of OOB intervals when error variance changes across the predictor space.

Our comparison of OOB and SC inference shows that these methods produce intervals that behave similarly with respect to coverage probability. However, OOB intervals tend to be narrower, and thus more informative, than SC intervals. The SC intervals come with a guarantee of finite-sample Type I coverage probability at or above any specified level of confidence under very general conditions. Although this marginal coverage guarantee is very appealing, our numerical results in simulations and in the analysis of 60 real datasets provide compelling evidence in favor of OOB intervals. We recommend that an OOB interval be used alongside a random forest point prediction to provide a range of plausible response values for those drawing conclusions from data.

References

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–127.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2018). Local linear forests. *arXiv preprint arXiv:1807.11408*.

Lei, J., GSell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 154–168. Springer.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V., Nouretdinov, I., Gammerman, A., et al. (2009). On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651.

Xu, R., Nettleton, D., and Nordman, D. J. (2016). Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65.

Supplemental Materials to *Random Forest Prediction Intervals*

Haozhe Zhang, Joshua Zimmerman, Dan Nettleton*, and Daniel J. Nordman*

Department of Statistics, Iowa State University, Ames IA 50011

*Corresponding author emails: dnett@iastate.edu, dnordman@iastate.edu

The Supplemental Materials provided here consist of two parts. Proofs of the distributional results, regarding the coverage properties of out-of-bag prediction intervals (Section 3 of the main manuscript), are detailed in Section S.1. In Section S.2 and Section S.3, we then provide some additional figures to further support the numerical studies given in Sections 5-6 of the main manuscript.

S.1 Proofs of Main Results

Proofs of Theorem 1 and Corollary 1. Corollary 1 follows from the convergence of the conditional probability in Theorem 1 combined with the boundedness of the conditional probability by 1; consequently, the expected value of the conditional probability in Theorem 1 (or, equivalently, the unconditional probability in Corollary 1) converges to $1-\alpha$.

For the proof of Theorem 1, we require some notation as well as statements of Lemmas 1-2 to follow; proofs of these technical lemmas appear after that of Theorem 1. Let $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ be iid random vectors where $Y - m(\boldsymbol{X})$ has continuous cdf F under condition (c.3), i.e., $F(t) = \mathbb{P}\{Y - m(\boldsymbol{X}) \leq t\}$, $t \in \mathbb{R}$. Based on $\boldsymbol{\mathcal{C}}_n \equiv \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$, let $\widehat{Y} \equiv \widehat{m}_n(\boldsymbol{X})$ denote the *RF* estimator of $m(\boldsymbol{X})$ and, for $i = 1, \ldots, n$, let $\widehat{Y}_{(i)} = \widehat{m}_{n,(i)}(\boldsymbol{X}_i)$ denote the associated oob estimator of $m(\boldsymbol{X}_i)$ (i.e., based on the subforest $RF_{(i)}$ involving observations $\boldsymbol{\mathcal{C}}_n \setminus \{(\boldsymbol{X}_i, Y_i)\}$), where condition (c.4) entails

$$|\widehat{m}_n(\boldsymbol{X}) - m(\boldsymbol{X})| \xrightarrow{P} 0 \quad \text{and} \quad |\widehat{m}_{n,(1)}(\boldsymbol{X}_1) - m(\boldsymbol{X}_1)| \xrightarrow{P} 0 \quad \text{as } n \to \infty. \tag{S.1}$$

From the prediction differences $D_{n,i} \equiv D_i \equiv Y_i - \widehat{m}_{n,(i)}(\boldsymbol{X}_i)$, $i = 1, \ldots, n$, let $D_{[n,\gamma]} \equiv \inf\{t \in \mathbb{R} : \widehat{F}_n(t) \geq \gamma\}$ denote the $\gamma \in (0,1)$ empirical quantile based on the empirical distribution

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(D_{n,i} \leq t), \quad t \in \mathbb{R},$$

as an estimator of F, where $I(\cdot)$ denotes the indicator function above.

Lemma 1 *Under conditions (c.1)-(c.4), as $n \to \infty$,*

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| \xrightarrow{P} 0$$

and $F(D_{[n,\gamma_1]}) - F(D_{[n,\gamma_2]}) \xrightarrow{P} 1-\alpha$ for any $\gamma_1, \gamma_2, \alpha \in (0,1)$ with $\gamma_1 - \gamma_2 = 1-\alpha$.

Lemma 2 *Under conditions (c.1)-(c.4), as $n \to \infty$,*

$$\Delta_n \equiv \sup_{t\in\mathbb{R}} |P_*\{Y - \widehat{m}_n(\boldsymbol{X}) < t\} - F(t)| = \sup_{t\in\mathbb{R}} |P_*\{Y - \widehat{m}_n(\boldsymbol{X}) \leq t\} - F(t)| \xrightarrow{P} 0,$$

where $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n)$ denotes conditional probability given $\boldsymbol{\mathcal{C}}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$.

Next, for $\alpha \in (0,1)$, writing $\mathcal{P}_{*,n} \equiv P_*(D_{[n,\alpha/2]} \leq Y - \widehat{m}_n(\boldsymbol{X}) \leq D_{[n,1-\alpha/2]})$ to denote the target conditional coverage probability given $\boldsymbol{\mathcal{C}}_n$, we have

$$\begin{aligned}\mathcal{P}_{*,n} &= P_*(Y - \widehat{m}_n(\boldsymbol{X}) \leq D_{[n,1-\alpha/2]}) - P_*(Y - \widehat{m}_n(\boldsymbol{X}) < D_{[n,\alpha/2]}) \\ &= F(D_{[n,1-\alpha/2]}) - F(D_{[n,\alpha/2]}) + R_n,\end{aligned}$$

for a remainder R_n defined by subtraction. Then, $\mathcal{P}_{*,n} \xrightarrow{P} (1-\alpha)$ follows as $n \to \infty$ in Theorem 1 by using Lemma 1 along with the bound on the remainder $|R_n| \leq 2\Delta_n \xrightarrow{P} 0$ under Lemma 2. □

Proof of Lemma 1. The second claim of Lemma 1 follows from the first using that F is continuous. To see this, we consider showing $F(D_{[n,\gamma]}) \xrightarrow{P} \gamma$ for a fixed value $\gamma \in (0,1)$. For $a \equiv \inf\{t \in \mathbb{R} : F(t) \geq \gamma\}$ and $b \equiv \sup\{t \in \mathbb{R} : F(t) \leq \gamma\}$, note $a \leq b$ and that $F(a-\epsilon) < \gamma < F(b+\epsilon)$ holds for any $\epsilon > 0$. From this, the first Lemma 1 claim yields that $\mathbb{P}(\widehat{F}_n(a-\epsilon) < \gamma < \widehat{F}_n(b+\epsilon)) \to 1$ as $n \to \infty$ for any given $\epsilon > 0$. The event $\widehat{F}_n(a-\epsilon) < \gamma < \widehat{F}_n(b+\epsilon)$ implies that $D_{[n,\gamma]} \in [a-\epsilon, b+\epsilon]$ so that $|F(D_{[n,\gamma]}) - \gamma| \leq \Lambda(\epsilon) \equiv$

$F(b+\epsilon) - F(a-\epsilon)$ further holds, because F is non-decreasing with $F(a) = F(b) = \gamma$. Now $F(D_{[n,\gamma]}) \xrightarrow{P} \gamma$ follows by $\lim_{n\to\infty} \mathbb{P}\{|F(D_{[n,\gamma]}) - \gamma| \leq \Lambda(\epsilon)\} = 1$ for each $\epsilon > 0$ combined with $\lim_{\epsilon\downarrow 0} \Lambda(\epsilon) = 0$.

To establish the first claim of Lemma 1, it suffices, by Poyla's theorem and the continuity of F, to show that $\widehat{F}_n(t) \xrightarrow{P} F(t)$ for any fixed $t \in \mathbb{R}$. Note that, using $m(\boldsymbol{X}_1) - \widehat{m}_{n,(1)}(\boldsymbol{X}_1) \overset{d}{=} m(\boldsymbol{X}_2) - \widehat{m}_{n,(2)}(\boldsymbol{X}_2) \xrightarrow{P} 0$ in (S.1) along with Slutsky's theorem, we have

$$\begin{pmatrix} D_{n,1} \\ D_{n,2} \end{pmatrix} = \begin{pmatrix} Y_1 - m(\boldsymbol{X}_1) \\ Y_2 - m(\boldsymbol{X}_2) \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{X}_1) - \widehat{m}_{n,(1)}(\boldsymbol{X}_1) \\ m(\boldsymbol{X}_2) - \widehat{m}_{n,(2)}(\boldsymbol{X}_2) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Y_1 - m(\boldsymbol{X}_1) \\ Y_2 - m(\boldsymbol{X}_2) \end{pmatrix} \quad \text{(S.2)}$$

as $n \to \infty$, where $Y_1 - m(\boldsymbol{X}_1)$ and $Y_2 - m(\boldsymbol{X}_2)$ are again iid with continuous cdf F. By the iid properties of the random vectors in $\boldsymbol{\mathcal{C}}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$ along with (S.2), we then have

$$\mathbb{E}\widehat{F}_n(t) = \mathbb{P}(D_{n,1} \leq t) \to F(t) \quad \text{as } n \to \infty$$

for any given $t \in \mathbb{R}$, as well as

$$\begin{aligned} \mathrm{Var}[\widehat{F}_n(t)] &= \frac{1}{n}\mathrm{Var}[I(D_{n,1} \leq t)] + \frac{n(n-1)}{n^2}\mathrm{Cov}\left[I(D_{n,1} \leq t), I(D_{n,2} \leq t)\right] \\ &\leq \frac{1}{n} + \mathbb{P}(D_{n,1} \leq t, D_{n,2} \leq t) - [\mathbb{P}(D_{n,1} \leq t)]^2 \\ &\to [F(t)]^2 - [F(t)]^2 = 0 \end{aligned}$$

as $n \to \infty$. This shows $\widehat{F}_n(t) \xrightarrow{P} F(t)$ and completes the proof of Lemma 1. □

Proof of Lemma 2. The equality of the suprema defining Δ_n follows from one-sided limit behavior of cdfs (e.g., $\lim_{t\uparrow s} P_*(Y - \widehat{m}_n(\boldsymbol{X}) \leq t) = P_*(Y - \widehat{m}_n(\boldsymbol{X}) < s)$ and $\lim_{t\downarrow s} P_*(Y - \widehat{m}_n(\boldsymbol{X}) < t) = P_*(Y - \widehat{m}_n(\boldsymbol{X}) \leq s))$ along with $F(t) = \mathbb{P}(Y - m(\boldsymbol{X}) < t)$, $t \in \mathbb{R}$, by continuity. Writing $Y - \widehat{m}_n(\boldsymbol{X}) = [Y - m(\boldsymbol{X})] + [m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})]$, the conditional cdf of $[Y - m(\boldsymbol{X})]$ given $\boldsymbol{\mathcal{C}}_n$ is F (i.e., the continuous unconditional cdf), as $[Y - m(\boldsymbol{X})]$ is independent of $\boldsymbol{\mathcal{C}}_n$. Hence, to establish Lemma 2, it suffices to prove that the conditional distribution of $[m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})]$ given $\boldsymbol{\mathcal{C}}_n$ converges to a distribution that is degenerate at 0 (in probability). For any integer $\ell \geq 1$, $P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \xrightarrow{P} 0$ follows as

$n \to \infty$ using that

$$\mathbb{E}P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) = \mathbb{P}(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \to 0$$

by (S.1). This implies the desired probabilistic convergence and completes the proof of Lemma 2. [That is, if $P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \xrightarrow{P} 0$ for any integer $\ell \geq 1$, then for any subsequence $\{n_j\} \subset \{n\}$, one may extract a further subsequence $\{n_k\} \subset \{n_j\}$ such that the set of sample points

$$A \equiv \{\omega \in \Omega : P_*(|m(\boldsymbol{X}) - \widehat{m}_{n_k}(\boldsymbol{X})| > \ell^{-1})(\omega) \to 0 \text{ as } n_k \to \infty \text{ for all } \ell \geq 1\}$$

has $\mathbb{P}(A) = 1$ on some probability space $(\Omega, \mathcal{F}, P)$; consequently, along the subsequence $\{n_k\}$ and pointwise on A, the distribution of $|m(\boldsymbol{X}) - \widehat{m}_{n_k}(\boldsymbol{X})|$ under P_* converges weakly to a degenerate distribution at 0 (i.e., with probability 1). As the subsequence $\{n_j\} \subset \{n\}$ was arbitrary, the weak convergence of the distribution of $|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})|$ under P_* must hold in probability.] □

Proofs of Theorem 2 and Corollary 2. By re-defining the conditional probability P_* in the proof of Theorem 1 to denote conditional probability $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x})$ given both $\boldsymbol{\mathcal{C}}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$ and $\boldsymbol{X} = \boldsymbol{x}$ (rather than given $\boldsymbol{\mathcal{C}}_n$ alone), the same proof for Theorem 1 then applies to show Theorem 2. This is because Lemma 1 remains valid along with a version of Lemma 2 with respect to the re-defined conditional probability P_*; namely, under Theorem 2 assumptions, the corresponding Lemma 2 result becomes

$$\Delta_n \equiv \sup_{t \in \mathbb{R}} |P_* \{Y - \widehat{m}_n(\boldsymbol{x}) < t\} - F(t)| = \sup_{t \in \mathbb{R}} |P_* \{Y - \widehat{m}_n(\boldsymbol{x}) \leq t\} - F(t)| \xrightarrow{P} 0,$$

as $n \to \infty$, under the conditional probability $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x})$. This recasting of Lemma 2 can be justified using the same essential argument given in the previous proof of Lemma 2 with two modifications: we use that the conditional distribution of $Y - m(\boldsymbol{X}) \equiv Y - m(\boldsymbol{x})$ given $\boldsymbol{\mathcal{C}}_n$ and $\boldsymbol{X} = \boldsymbol{x}$ has cdf F (because $e = Y - m(\boldsymbol{X})$, with cdf F, is independent of $\boldsymbol{X}$ by condition (c.2) and independent of $\boldsymbol{\mathcal{C}}_n$ by assumption) and we apply $\widehat{m}_n(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ in place of $\widehat{m}_n(\boldsymbol{X}) \xrightarrow{P} m(\boldsymbol{X})$. Theorem 2 then yields Corollary 2 in the same manner as Corollary 1 follows from Theorem 1. □

S.2 Additional Figures for Section 5

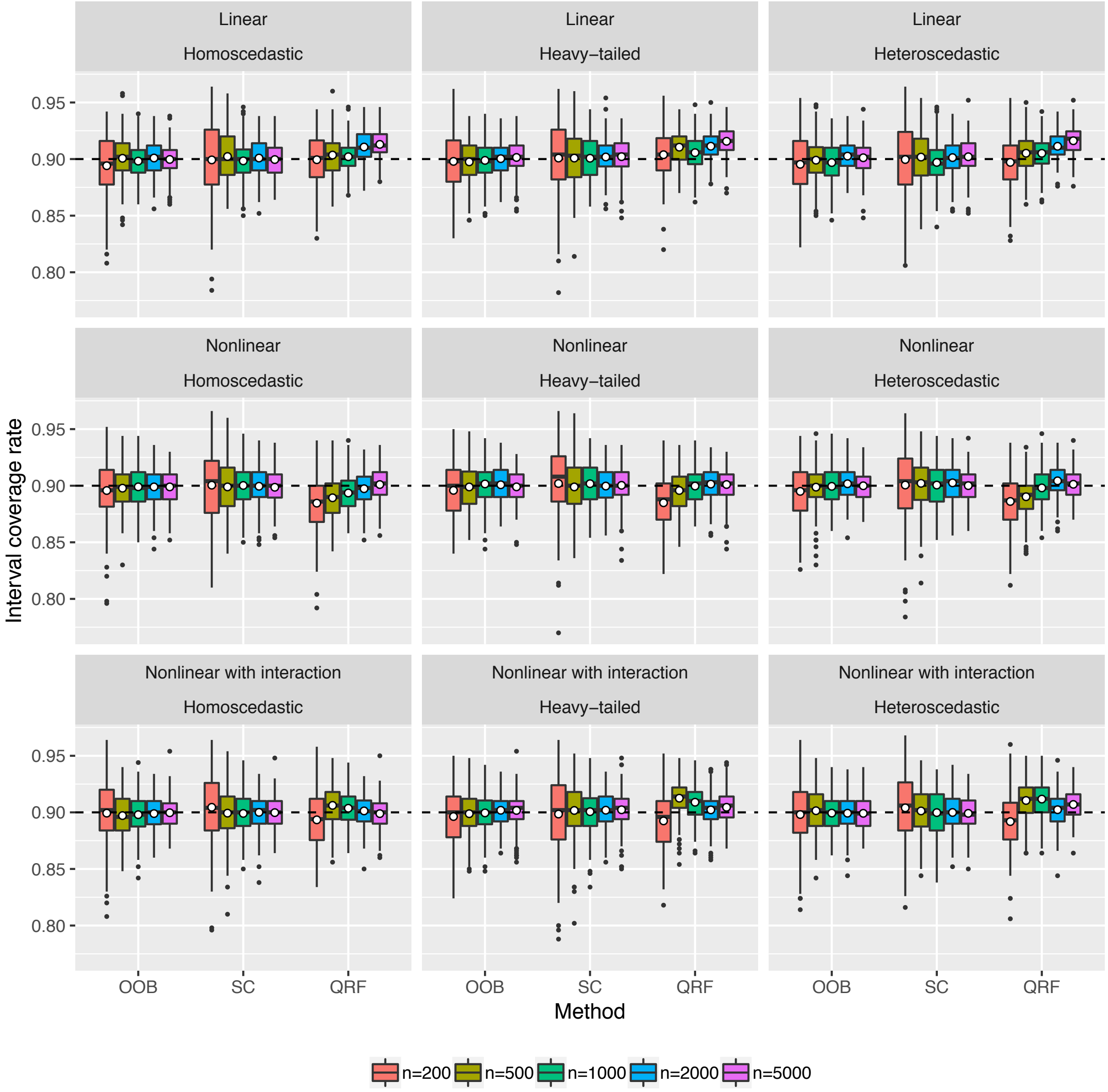


Figure S.1: Boxplots of the Type II coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n]$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot, and represents an estimate of Type I coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)]$.

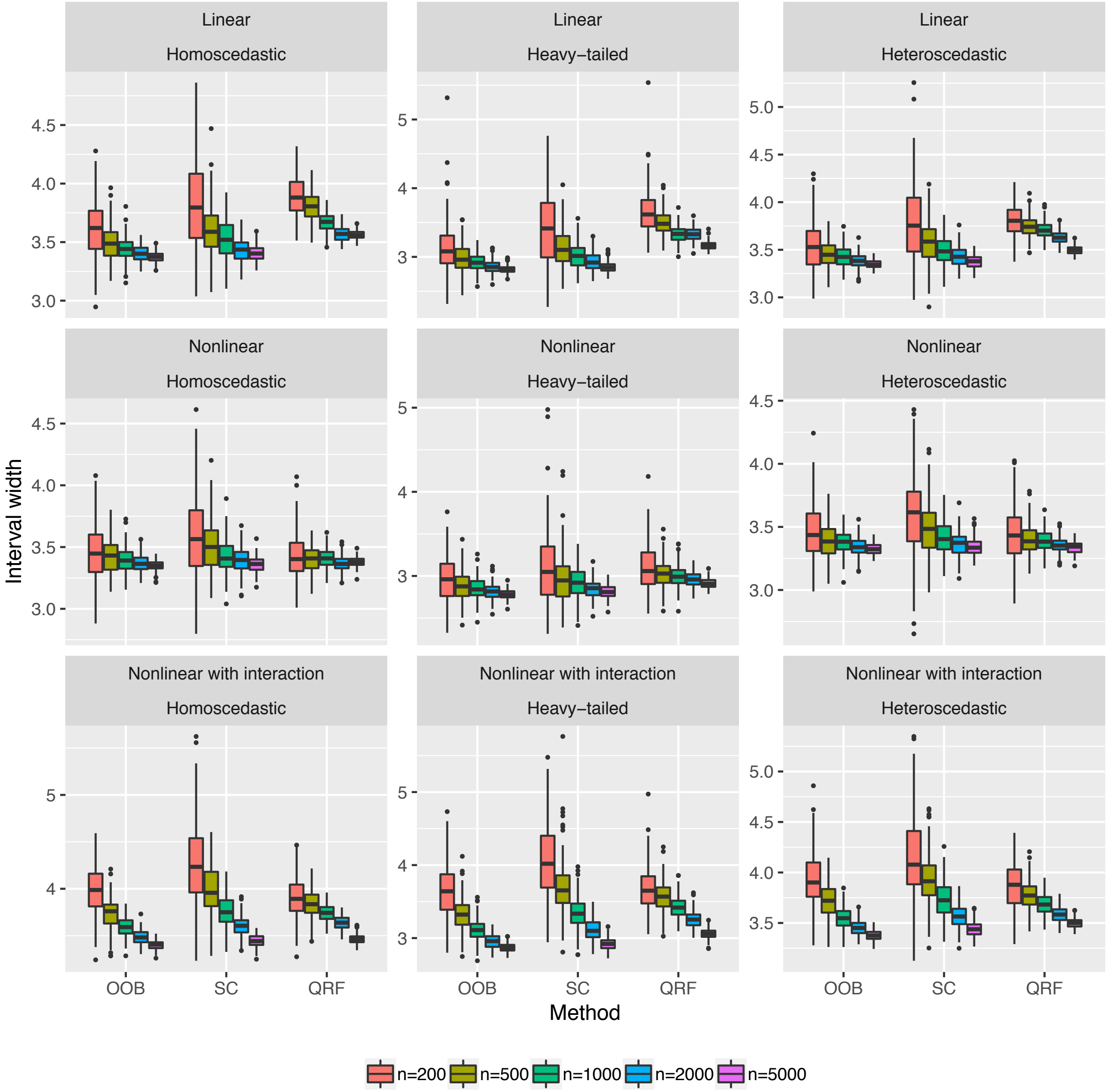


Figure S.2: Boxplots of interval widths for out-of-bag (OOB) prediction intervals and split conformal (SC) prediction intervals, and the average interval widths of quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors).

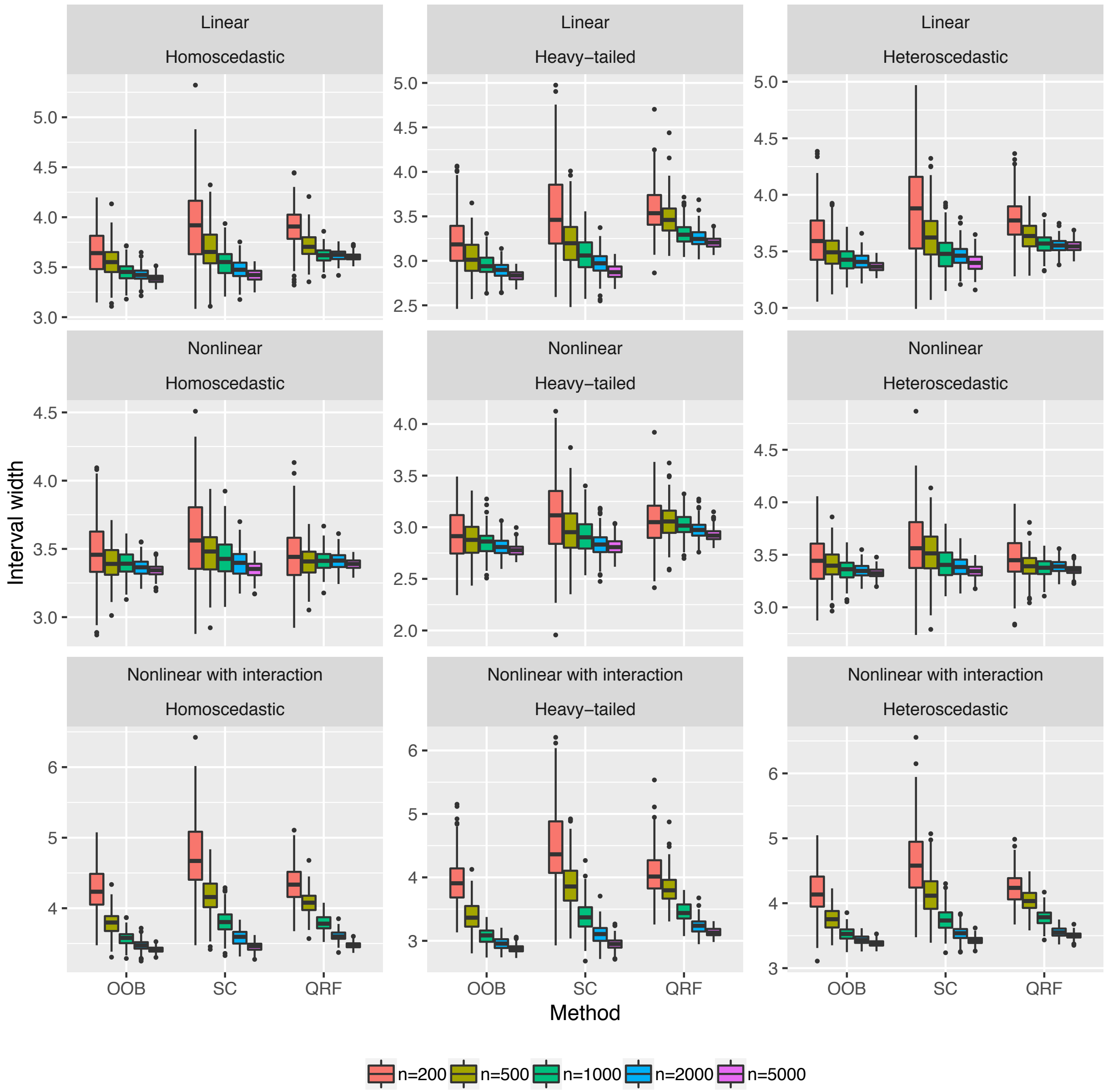


Figure S.3: Boxplots of the $\log_2$ ratios of split conformal (SC) interval widths to out-of-bag (OOB) interval widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval widths when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors).

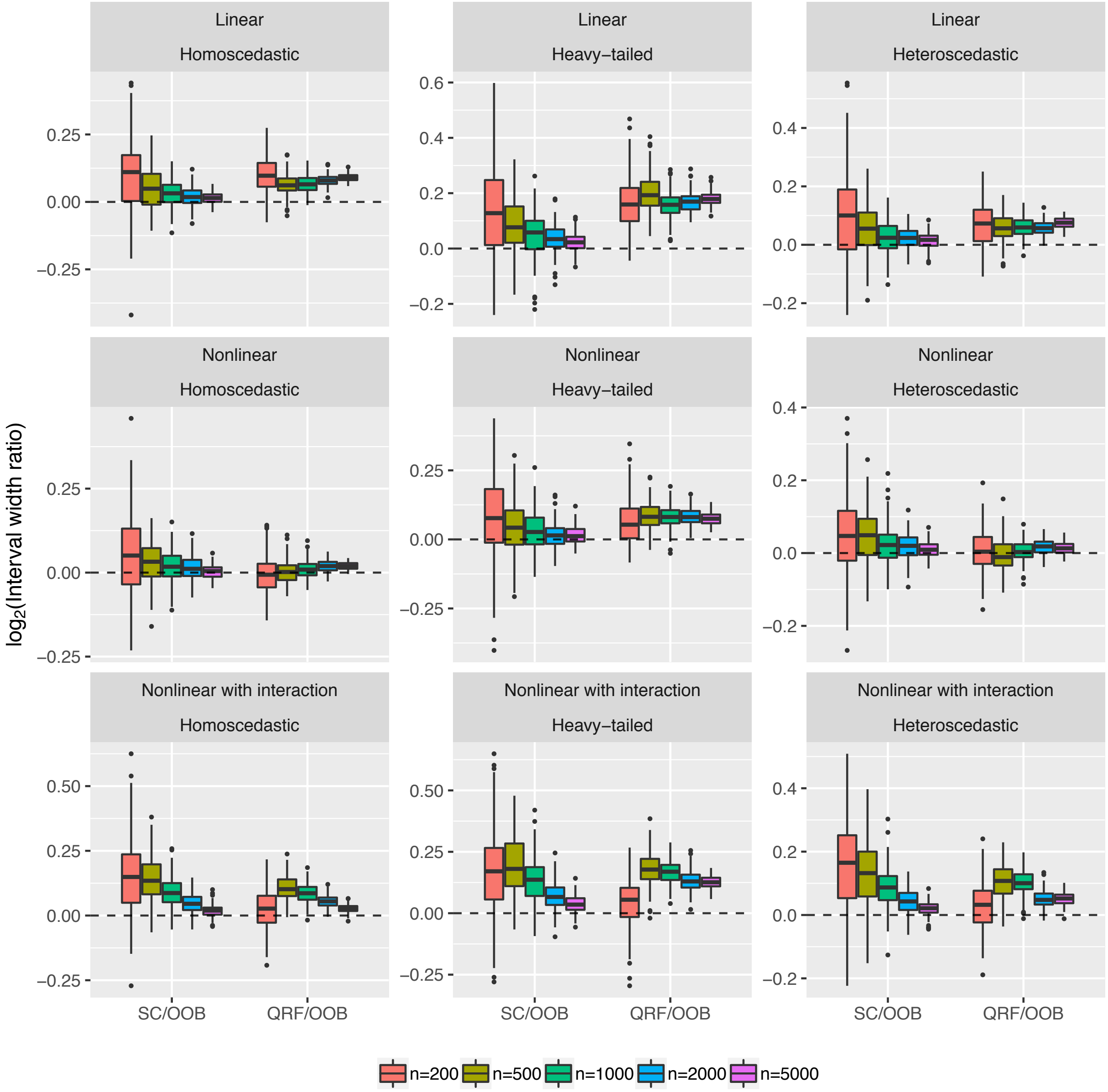


Figure S.4: Boxplots of the $\log_2$ ratios of split conformal interval (CONF) widths to out-of-bag interval (OOB) widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag interval (OOB) widths when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors).

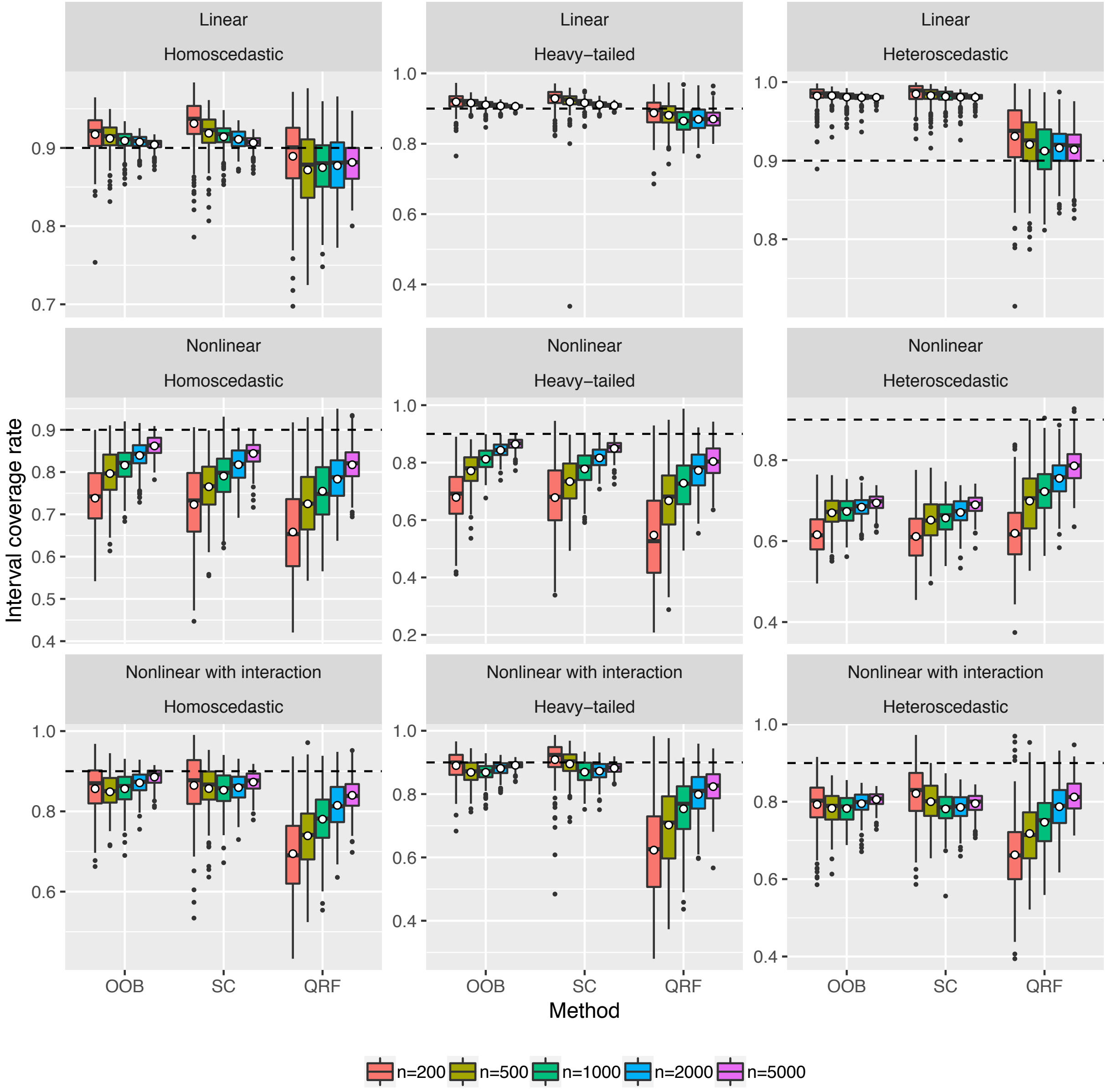


Figure S.5: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \mathbf{0}]$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{X} = \mathbf{0}]$.

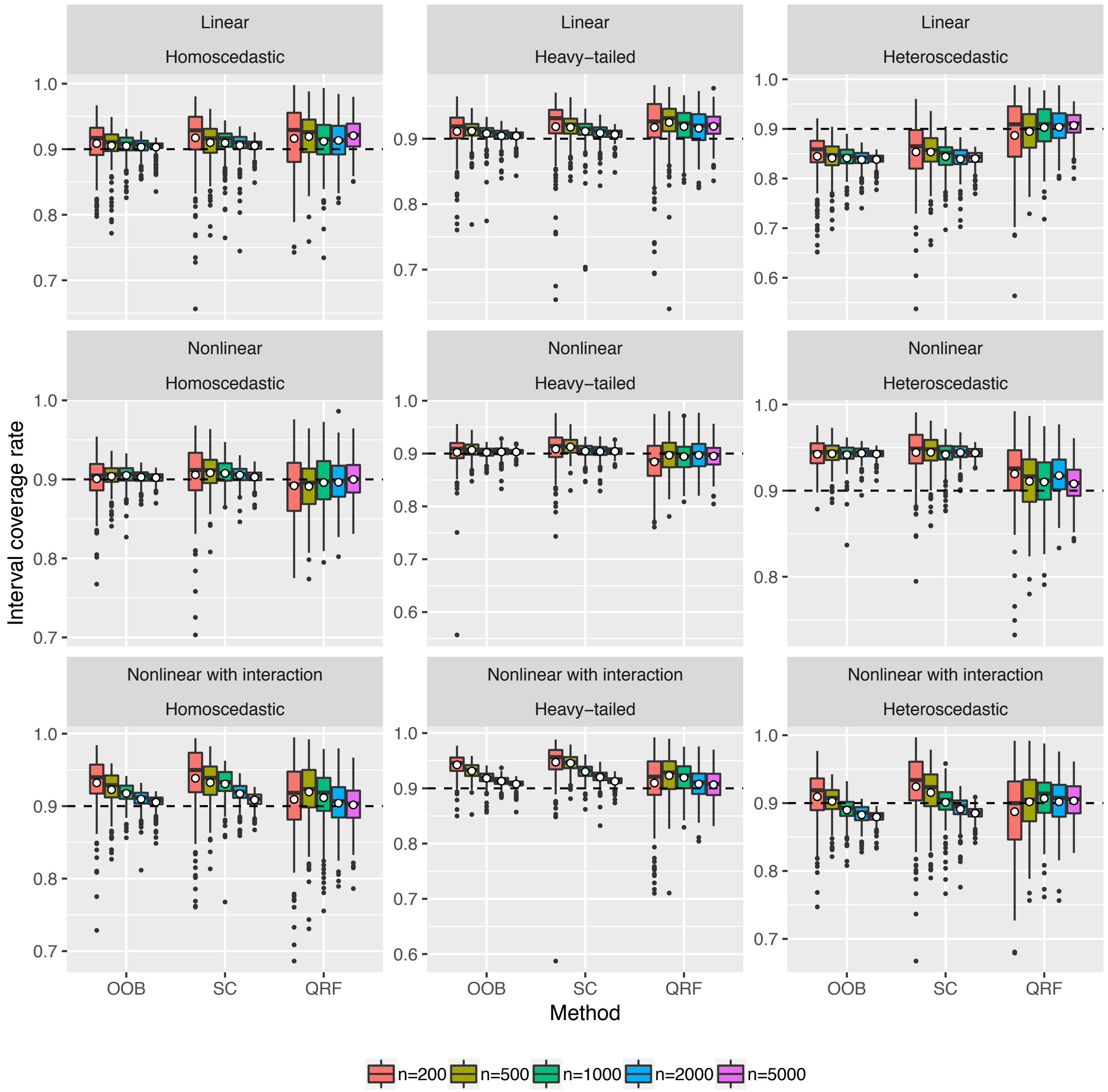


Figure S.6: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \mathbf{1}]$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{X} = \mathbf{1}]$.

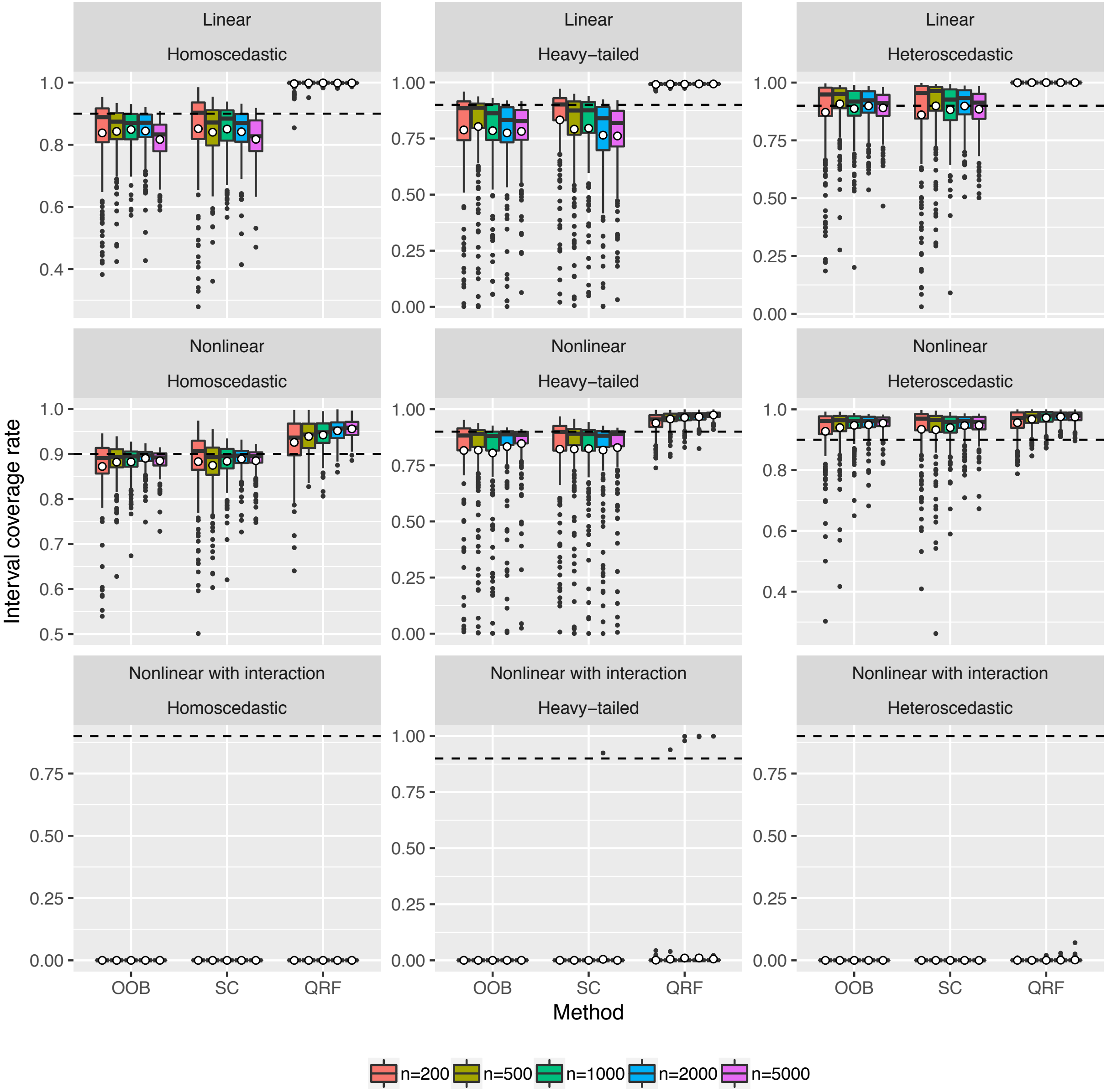


Figure S.7: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n) | \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$.

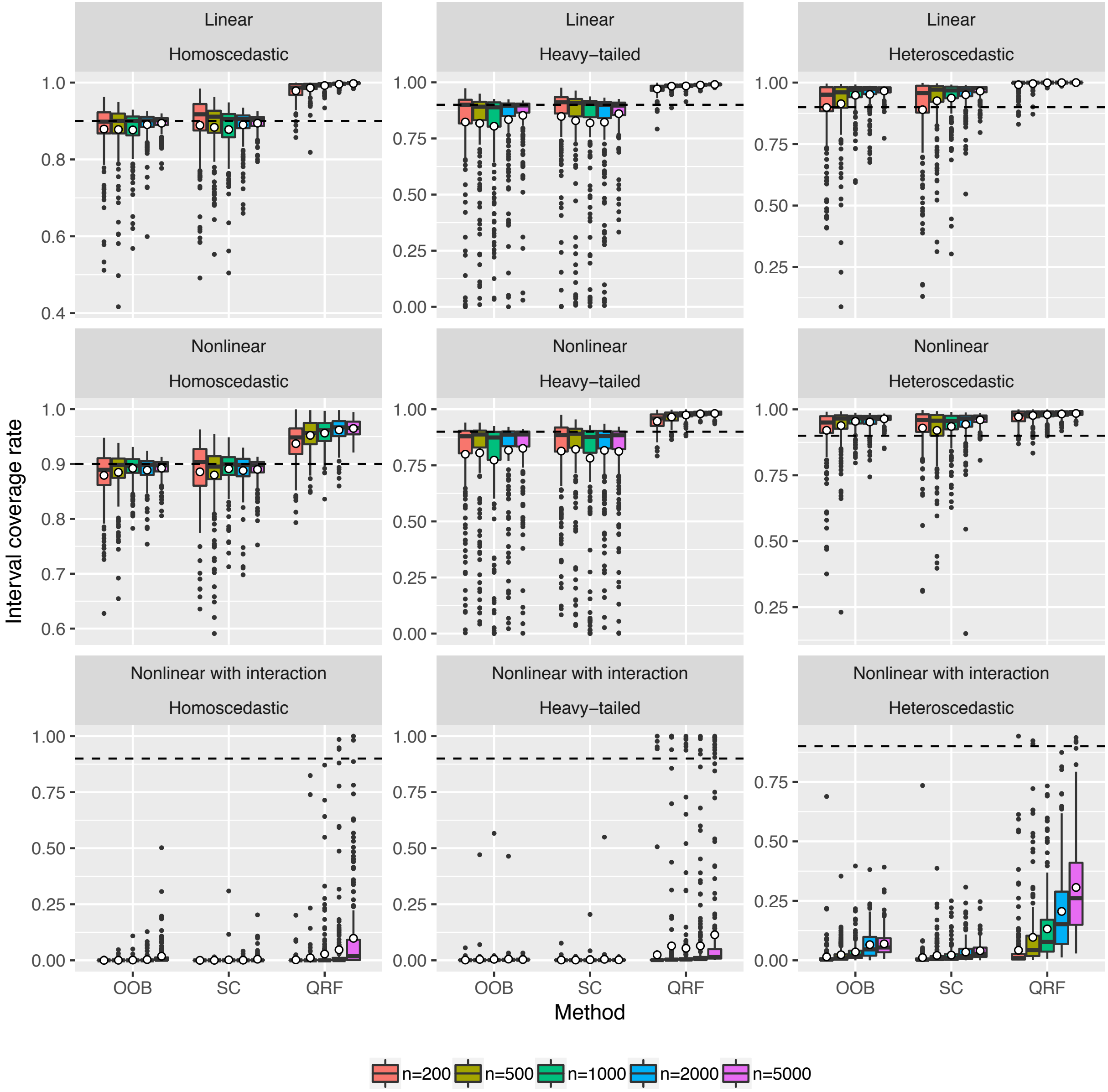


Figure S.8: Boxplots of the Type IV coverage rate estimates, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$, for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \boldsymbol{\mathcal{C}}_n)|\boldsymbol{X} = (3, -3, 3, \cdots, 3)']$.

S.3 Additional Figures for Section 6

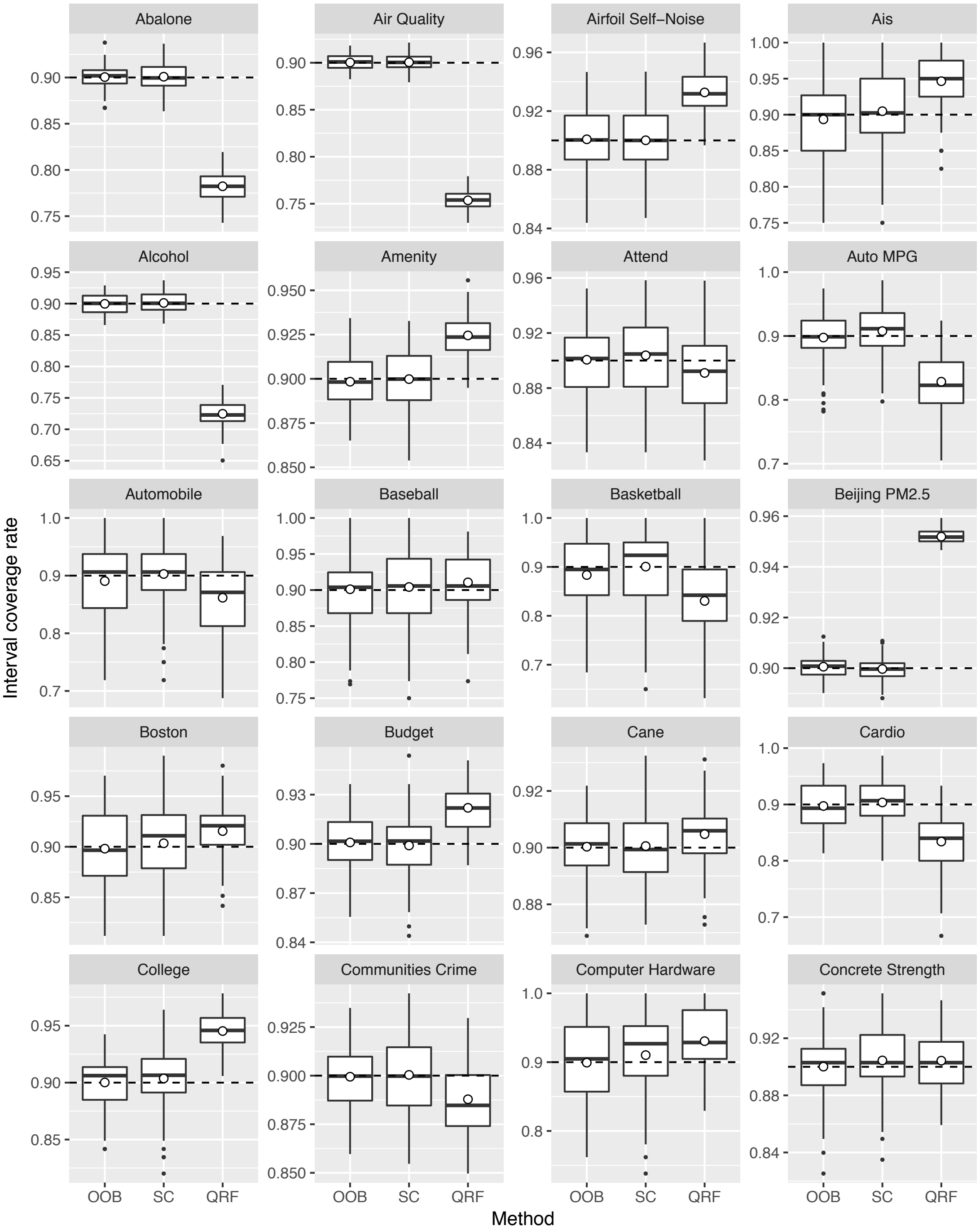


Figure S.9: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Abalone, Air Quality, Airfoil Self-Noise, Ais, Alcohol, Amenity, Attend, Auto MPG, Automobile, Baseball, Basketball, Beijing PM2.5, Boston, Budget, Cane, Cardio, College, Communities Crime*, *Computer Hardware*, and *Concrete Strength*. The circles represent empirical Type I coverage rates.

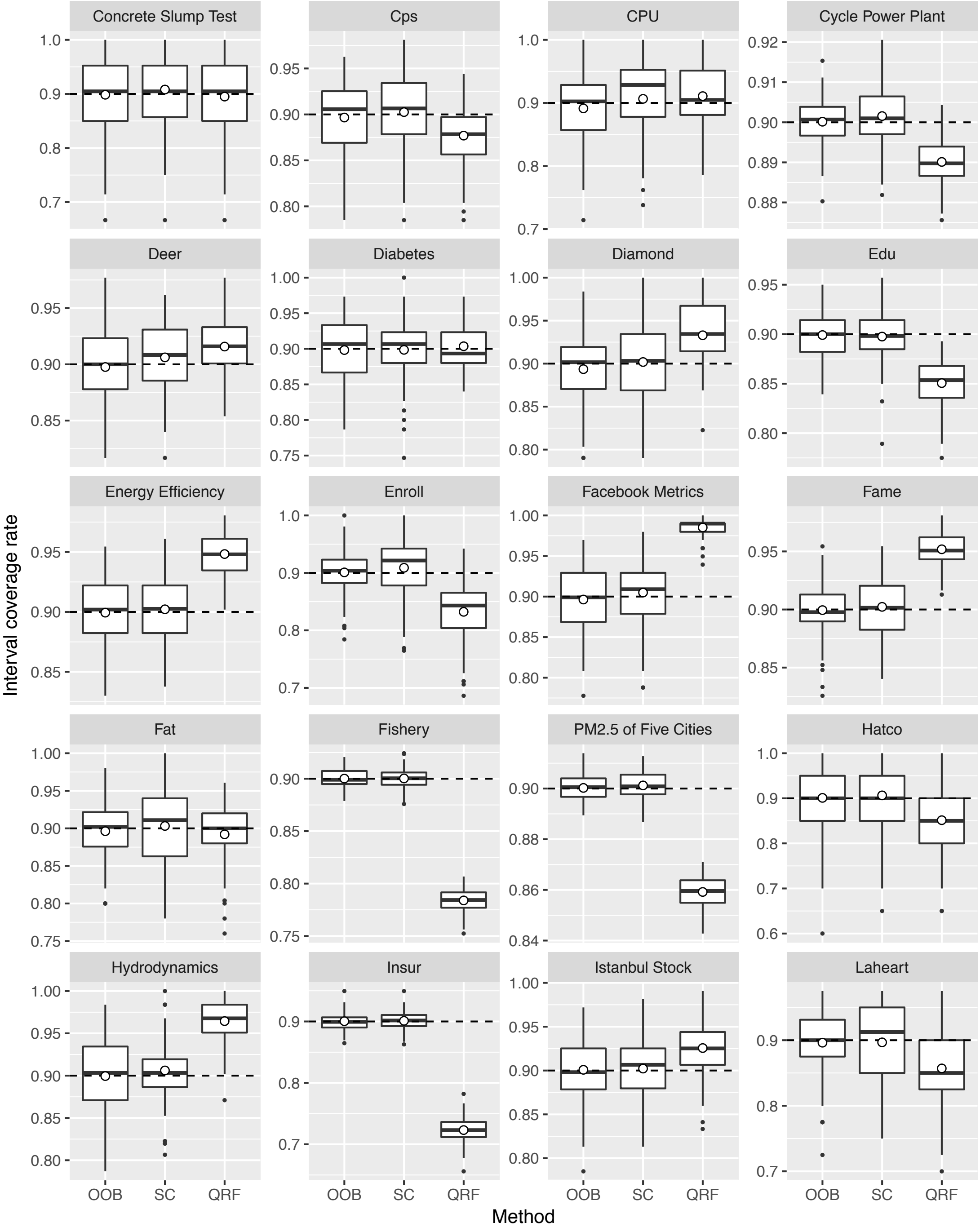


Figure S.10: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Concrete Slump Test, Cps, CPU, Cycle Power Plant, Deer, Diabetes, Diamond, Edu, Energy Efficiency, Enroll, Facebook Metrics, Fame, Fat, Fishery, Hatco, Hydrodynamics, Insur, Istanbul Stock, Laheart*, and *Medicare*. The circles represent empirical Type I coverage rates.

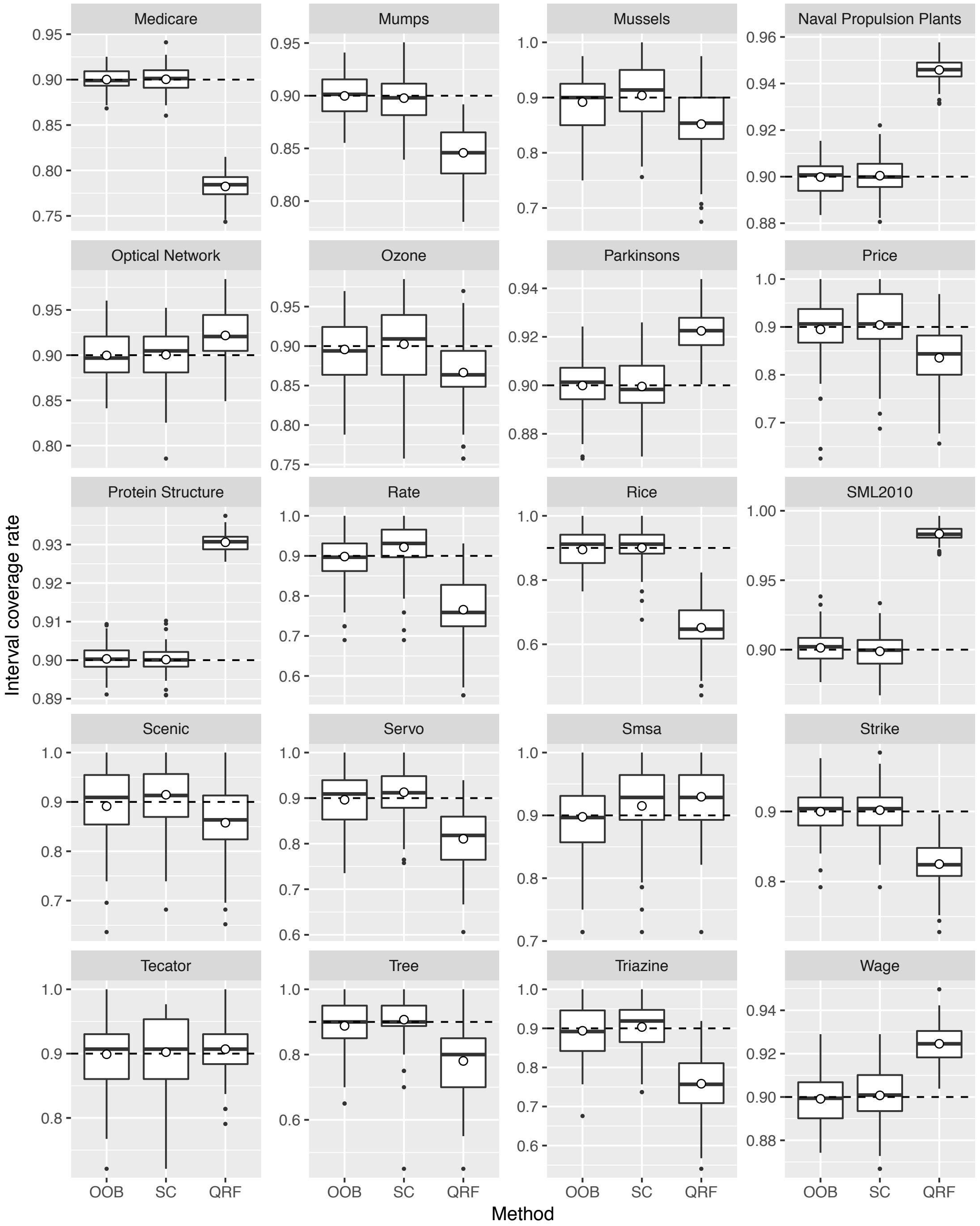


Figure S.11: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Mumps, Mussels, Naval Propulsion Plants, Optical Network, Ozone, Parkinsons, PM2.5 of Five Cities, Price, Protein Structure, Rate, Rice, Scenic, Servo, SML2010, Smsa, Strike, Tecator, Tree, Triazine*, and *Wage*. The circles represent empirical Type I coverage rates.