Distribution-Free Predictive Inference For Regression

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman

Department of Statistics, Carnegie Mellon University

April 15, 2016

Abstract

We develop a general framework for distribution-free predictive inference in regression, using conformal inference. The proposed methodology allows construction of prediction bands for the response variable using any estimator of the regression function. The resulting prediction band preserves the consistency properties of the original estimator under standard assumptions, while guaranteeing finite sample marginal coverage even when the assumptions do not hold. We analyze and compare, both empirically and theoretically, two major variants of our conformal procedure: the full conformal inference and split conformal inference, along with a related jackknife method. These methods offer different tradeoffs between statistical accuracy (length of resulting prediction intervals) and computational efficiency. As extensions, we develop a method for constructing valid in-sample prediction intervals called *rank-one-out* conformal inference, which has essentially the same computational efficiency as split conformal inference. We also describe an extension of our procedures for producing prediction bands with varying local width, in order to adapt to heteroskedascity in the data distribution. Lastly, we propose a model-free notion of variable importance, called *leave-one-covariate-out* or LOCO inference. Accompanying our paper is an R package `conformalInference` that implements all of the proposals we have introduced. In the spirit of reproducibility, all empirical results in this paper can be easily (re)generated using this package.

Contents

1 Introduction

Consider i.i.d. regression data

$$Z_1, \ldots, Z_n \sim P,$$

where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y_i and a d-dimensional vector features (or predictors, or covariates) $X_i = (X_i(1), \ldots, X_i(d))$. The feature dimension d may be large relative to the sample size n (in an asymptotic model, d is allowed to increase with n). Let

$$\mu(x) = \mathbb{E}[Y \,|\, X = x], \quad x \in \mathbb{R}^d,$$

denote the regression function. We are interested in predicting a new response Y_{n+1} from a new feature value X_{n+1}, with no assumptions on μ and P. Formally, given a nominal mis-coverage level $\alpha \in (0, 1)$, we seek to constructing a prediction band $C \subset \mathbb{R}^d \times \mathbb{R}$ based on $Z_1, \ldots, Z_n$ with the property that

$$\mathbb{P}\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1 - \alpha, \tag{1}$$

where the probability is taken over the $n + 1$ i.i.d. draws $Z_1, \ldots, Z_n, Z_{n+1} \sim P$, and for a point $x \in \mathbb{R}^d$ we denote $C(x) = \{y \in \mathbb{R} : (x, y) \in C\}$. The main goal of this paper is to construct prediction bands as in (1) that have finite sample (nonasymptotic) validity and without assumptions on P. A second goal is to construct model-free, prediction-based inferential statements about the importance of each covariate in the prediction model for Y_{n+1} given X_{n+1}.

Our leading example is high-dimensional regression, where $d \gg n$ and a linear function is used to approximate μ (but the linear model is not necessarily assumed to be correct). Common approaches in this setting include greedy methods like forward stepwise regression, and ℓ_1-based methods like the lasso. There is an enormous amount of work dedicated to studying various properties of these methods, but to our knowledge, there is very little work on prediction sets. Our framework provides proper prediction sets for these methods, and for essentially any high-dimensional regression method. It also covers classical linear regression and nonparametric regression techniques. The basis of our framework is *conformal prediction*, a method invented by Vovk et al. (2005).

1.1 Related Work

Conformal inference. The conformal prediction framework was originally proposed as a sequential approach for forming prediction intervals, by Vovk et al. (2005, 2009). The basic idea is simple. Keeping the regression setting introduced above and given a new independent

draw (X_{n+1}, Y_{n+1}) from P, in order to decide if a value y is to be included in $C(X_{n+1})$, we consider testing the null hypothesis that $Y_{n+1} = y$ and construct a valid p-value based on the empirical quantiles of the augmented sample $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ with $Y_{n+1} = y$ (see Section 2 below for details). The data augmentation step makes the procedure immune to overfitting, so that the resulting prediction band always has valid average coverage. Conformal inference has also been studied as a batch (rather than sequential) method, in various settings. For example, Burnaev & Vovk (2014) considered low-dimensional least squares and ridge regression models. Lei et al. (2013) used conformal prediction to construct statistically near-optimal tolerance regions. Lei & Wasserman (2014) extended this result to low-dimensional nonparametric regression. Other extensions, such as classification and clustering, are explored in Lei (2014); Lei et al. (2015).

There is very little work on prediction sets in high-dimensional regression. Hebiri (2010) described an approximation of the conformalized lasso estimator. This approximation leads to a big speedup over the original conformal prediction method build on top of the lasso, but loses the key appeal of conformal inference in the first place—it fails to offer finite sample coverage. Recently Steinberger & Leeb (2016) analyzed a jackknife prediction method in the high-dimensional setting, extending results in low-dimensional regression due to Butler & Rothman (1980). However, this jackknife approach is only guaranteed to have asymptotic validity when the base estimator (of the regression parameters) satisfies strong asymptotic mean squared error and stability properties. This is further discussed in Section 2.4. In our view, a simple, computationally efficient, and yet powerful method that seems to have been overlooked is *split conformal inference* (see Lei et al. (2015); Papadopoulos et al. (2002), or Section 2.2). When combined with, for example, the lasso estimator, the total cost of forming split conformal prediction intervals is dominated by the cost of fitting the lasso, and it always provides finite-sample coverage, in any setting—regardless of whether the lasso estimator is consistent.

High-dimensional inference. A very recent and exciting research thread in the field of high-dimensional inference is concerned with developing methods for constructing confidence intervals for (fixed) population-based targets, and separately, (random) post-selection targets. In the first class, population-based approaches, the linear model is assumed to be true and the focus is on providing confidence intervals for the coefficients in this model (see Belloni et al., 2012; Buhlmann, 2013; Zhang & Zhang, 2014; van de Geer et al., 2014; Javanmard & Montanari, 2014, for example). In the second class, post-selection approaches, the focus is on covering coefficients in the best linear approximation to μ given a subset of selected covariates. Some key references are Berk et al. (2013); Lee et al. (2016); Tibshirani et al. (2016). These methods are all interesting, and they serve different purposes (i.e., the purposes behind the two classes are different). One common thread, however, is that all of these methods rely on nontrivial assumptions—even if the linear model need not be

assumed true, conditions are typically placed on the quality of the regression estimator being considered, the error distribution, the knowledge or estimability of error variance, the homoskedasticity of errors, etc. In contrast, we describe two prediction-based methods for variable importance in Section 5, which do not rely on such conditions at all.

1.2 Summary and Outline

In this paper, we make several methodological and theoretical contributions to conformal inference in regression.

- We provide a general introduction to conformal inference (Section 2), a generic tool to construct distribution-free, finite sample prediction sets. We specifically consider the context of high-dimensional regression, arguably the scenario where conformal inference is most useful, due to the strong assumptions required by (and the fragility of) existing inference methods.

- We conduct extensive simulation studies (Section 3) to assess the two major conformal variants: full and split conformal methods, along with a related jackknife method. These simulations are reproducible using our accompanying R package `conformalInference` (`https://github.com/ryantibs/conformal`), which efficiently implements all methods studied in this paper (including the extensions and variable importance measures described below).

- We develop two extensions of conformal inference (Section 4), allowing for more informative and flexible inference: prediction intervals with in-sample coverage, and prediction intervals with varying local width.

- We provide new theoretical insights for conformal inference: accuracy guarantees for its finite-sample coverage (Theorems 2.1, 2.2), and distribution-free asymptotic, in-sample coverage guarantees (Theorems 2.3, 4.1). We also show that versions of conformal inference approximate certain oracle methods (Section 6). In doing so, we provide near-optimal bounds on the length of the prediction interval under standard assumptions.

- We propose two new, model-free, prediction-based approaches for inferring variable importance based on *leave-one-covariate-out* or LOCO inference (Section 5).

2 Conformal Inference

The basic idea behind conformal prediction theory is related to a simple result about sample quantiles. Let $U_1, \ldots, U_n$ be i.i.d. samples of a scalar random variable (in fact, the arguments that follow hold with the i.i.d. assumption replaced by the weaker assumption of exchangeability). For a given miscoverage level $\alpha \in (0, 1)$, and another i.i.d. sample U_{n+1}, note that

$$\mathbb{P}\Big(U_{n+1} \leq \widehat{q}_{1-\alpha}\Big) \geq 1 - \alpha, \tag{2}$$

where we define the sample quantile $\widehat{q}_{1-\alpha}$ based on $U_1, \ldots, U_n$ by

$$\widehat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{if } \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty & \text{otherwise,} \end{cases}$$

and $U_{(1)} \leq \ldots \leq U_{(n)}$ denote the order statistics of $U_1, \ldots, U_n$. The finite sample coverage property in (2) is easy to verify: by exchangeability, the rank of U_{n+1} among $U_1, \ldots, U_n, U_{n+1}$ is uniformly distributed over the set $\{1, \ldots, n+1\}$.

In our regression problem, where we observe i.i.d. samples $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} \sim P$, $i = 1, \ldots, n$, we might consider the following näive method for constructing a prediction interval for Y_{n+1} at the new feature value X_{n+1}, where (X_{n+1}, Y_{n+1}) is an independent draw from P. Following the idea described above, we can form the prediction interval defined by

$$C_{\text{naive}}(X_{n+1}) = \Big[\widehat{\mu}(X_{n+1}) - \widehat{F}_n^{-1}(1-\alpha),\ \widehat{\mu}(X_{n+1}) + \widehat{F}_n^{-1}(1-\alpha)\Big], \tag{3}$$

where $\widehat{\mu}$ is an estimator of the underlying regression function and $\widehat{F}_n$ the empirical distribution of the fitted residuals $|Y_i - \widehat{\mu}(X_i)|$, $i = 1, \ldots, n$, and $\widehat{F}_n^{-1}(1-\alpha)$ the $(1-\alpha)$-quantile of $\widehat{F}_n$. This is approximately valid for large samples, provided that the estimated regression function $\widehat{\mu}$ is accurate (i.e., enough for the estimated $(1-\alpha)$-quantile $\widehat{F}_n^{-1}(1-\alpha)$ of the fitted residual distribution to be close the $(1-\alpha)$-quantile of the population residuals $|Y_i - \mu(X_i)|$, $i = 1, \ldots, n$). Guaranteeing such an accuracy for $\widehat{\mu}$ generally requires appropriate regularity conditions, both on the underlying data distribution P, and on the estimator $\widehat{\mu}$ itself, such as a correctly specified model and/or an appropriate choice of tuning parameter.

2.1 Conformal Prediction Sets

In general, the naive method (3) can grossly undercover since the fitted residual distribution can often be biased downwards. Conformal prediction intervals (Vovk et al., 2005, 2009; Lei et al., 2013; Lei & Wasserman, 2014) overcome the deficiencies of the naive intervals, and, somewhat remarkably, are guaranteed to deliver proper coverage in finite sample without any assumptions on P or on $\widehat{\mu}$.

Consider the following strategy: for each value $y \in \mathbb{R}$, we construct an augmented regression estimate $\widehat{\mu}_y$, which is trained on the augmented data set $Z_1, \ldots, Z_n, (X_{n+1}, y)$. Now we define

$$R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|, \; i = 1, \ldots, n \quad \text{and} \quad R_{y,n+1} = |y - \widehat{\mu}_y(X_{n+1})|, \tag{4}$$

and we rank $R_{y,n+1}$ among the remaining fitted residuals $R_{y,1}, \ldots, R_{y,n}$, computing

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}(R_{y,i} \leq R_{y,n+1}) = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(R_{y,i} \leq R_{y,n+1}), \tag{5}$$

the proportion of points in the augmented sample whose fitted residual is smaller than the last one, $R_{y,n+1}$. Here $\mathbb{1}(\cdot)$ is the indicator function. By exchangeability, when evaluated at $y = Y_{n+1}$, we see that the constructed statistic $\pi(Y_{n+1})$ is uniformly distributed over the set $\{1/(n+1), 2/(n+1), \ldots, 1\}$, which implies

$$\mathbb{P}\Big((n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil\Big) \geq 1 - \alpha. \tag{6}$$

We may interpret the above display as saying that $1 - \pi(Y_{n+1})$ provides a valid (conservative) p-value for testing the null hypothesis that $H_0 : Y_{n+1} = y$.

By inverting such a test over all possibly values of $y \in \mathbb{R}$, the property (6) immediately leads to our conformal prediction interval at X_{n+1}, namely

$$C_{\text{conf}}(X_{n+1}) = \Big\{ y \in \mathbb{R} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil \Big\}. \tag{7}$$

The steps in (4), (5), (7) must be repeated each time we want to produce a prediction interval (at a new feature value). In practice, we must also restrict our attention in (7) to a discrete grid of trial values y. For completeness, this is summarized in Algorithm 1.

By construction, the conformal prediction interval in (7) has valid finite sample coverage; this interval is also accurate, meaning that it does not substantially over-cover. These are summarized in the following theorem, whose proof is in Appendix A.1.

Theorem 2.1. *If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. pair (X_{n+1}, Y_{n+1}),*

$$\mathbb{P}\Big(Y_{n+1} \in C_{\text{conf}}(X_{n+1})\Big) \geq 1 - \alpha,$$

for the conformal prediction band C_{conf} constructed in (7) (i.e., Algorithm 1). If we assume additionally that for all $y \in \mathbb{R}$, the fitted absolute residuals $R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \ldots, n$ have a continuous joint distribution, then it also holds that

$$\mathbb{P}\Big(Y_{n+1} \in C_{\text{conf}}(X_{n+1})\Big) \leq 1 - \alpha + \frac{1}{n+1}.$$

Algorithm 1 Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, fitting algorithm $\mathcal{A}$, points $\mathcal{X}_{\text{new}} = \{X_{n+1}, X_{n+2}, \ldots\}$ at which to construct prediction intervals, and values $\mathcal{Y}_{\text{trial}} = \{y_1, y_2, \ldots\}$ to act as trial values

Output: Predictions intervals, at each element of $\mathcal{X}_{\text{new}}$

for $x \in \mathcal{X}_{\text{new}}$ **do**

  for $y \in \mathcal{Y}_{\text{trial}}$ **do**

    $\widehat{\mu}_y = \mathcal{A}\big(\{(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)\}\big)$

    $R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \ldots, n$, and $R_{y,n+1} = |y - \widehat{\mu}_y(x)|$

    $\pi(y) = (1 + \sum_{i=1}^{n} \mathbb{1}(R_{y,i} \leq R_{y,n+1}))/(n+1)$

  end for

  $C_{\text{conf}}(x) = \{y \in \mathcal{Y}_{\text{trial}} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\}$

end for

Return $C_{\text{conf}}(x)$, for each $x \in \mathcal{X}_{\text{new}}$

Remark 2.1. *The first part of the theorem on the finite sample validity of conformal intervals in regression is a standard property of all conformal inference procedures and is due to Vovk. The second part—on the anti-conservativeness of conformal intervals—is new. For the second part only, we require that the residuals have a continuous distribution, which is quite a weak assumption, and is used to avoid ties when ranking the (absolute) residuals. By using a random tie-breaking rule, this assumption could be avoided entirely. In practice, the coverage of conformal intervals is highly concentrated around* $1 - \alpha$*, as confirmed by the experiments in Section 3. Other than the continuity assumption, no assumptions are needed in Theorem 2.1 about the the regression estimator* $\widehat{\mu}$ *or the data generating distributions* P*. This is a somewhat remarkable and unique property of conformal inference, and is not true for the jackknife method, as discussed in Section 2.4 (or, say, for the methods used to produce confidence intervals for the coefficients in high-dimensional linear model).*

Remark 2.2. *Generally speaking, as we improve our estimate* $\widehat{\mu}$ *of the underlying regression function* μ*, the resulting conformal prediction interval decreases in length. Intuitively, this happens because a more accurate* $\widehat{\mu}$ *leads to smaller residuals, and conformal intervals are essentially defined by the quantiles of the (augmented) residual distribution. Section 3 gives empirical examples that support this intuition.*

Remark 2.3. *The probability statements in Theorem 2.1 are taken over the i.i.d. pairs* (X_i, Y_i)*,* $i = 1, \ldots, n, n+1$*, and thus they assert average (or marginal) coverage guarantees. This should not be confused with* $\mathbb{P}(Y_{n+1} \in C(x) \,|\, X_{n+1} = x) \geq 1 - \alpha$ *for all* $x \in \mathbb{R}^d$*, i.e., conditional coverage, which is much stronger, and cannot be achieved by finite-length prediction intervals without regularity and consistency assumptions on the model and the estimator (Lei & Wasserman, 2014). Conditional coverage does hold asymptotically under*

certain conditions; see Theorem 6.5 in Section 6.

Remark 2.4. *Theorem 2.1 still holds if we replace each $R_{y,i}$ by*

$$f\big((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_{n+1}, y);\ (X_i, Y_i)\big), \tag{8}$$

where f is any function that is symmetric in its first n arguments. Such a function f is called the conformity score*, in the context of conformal inference. For example, the value in (8) can be an estimated joint density function evaluated at (X_i, Y_i), or conditional density function at (X_i, Y_i) (the latter is equivalent to the absolute residual $R_{y,i}$ when $Y - \mathbb{E}(Y|X)$ is independent of X, and has a symmetric distribution with decreasing density on $[0, \infty)$.) We will discuss a special locally-weighted conformity score in Section 4.2.*

2.2 Split Conformal Prediction Sets

The original conformal prediction method studied in the last subsection is computationally intensive. For any X_{n+1} and y, in order to tell if y is to be included in $C_{\text{conf}}(X_{n+1})$, we must retrain the model on the augmented data set (which includes the new point (X_{n+1}, y)), and recompute and re-sort the absolute residuals. In some applications, where X_{n+1} is not necessarily observed, prediction intervals are build by evaluating $\mathbb{1}\{y \in C_{\text{conf}}(x)\}$ over all pairs of (x, y) on a fine grid, as in Algorithm 1. In the special cases of kernel density estimation and kernel regression, simple and accurate approximations to the full conformal prediction sets are described in Lei et al. (2013); Lei & Wasserman (2014). In low-dimensional linear regression, the Sherman-Morrison updating scheme can be used to reduce the complexity of the full conformal method, by saving on the cost of solving a full linear system each time the query point (x, y) is changed. But in high-dimensional regression, with relatively sophisticated estimators such as the lasso, performing efficient full conformal inference is still an open problem.

Fortunately, there is an alternative approach, which we call *split conformal prediction*, that is completely general, and whose computational cost is a small fraction of the full conformal method. The split conformal method separates the fitting and ranking steps using sample splitting, and its computational cost is simply that of the fitting step. Similar ideas have appeared in the online prediction literature known under the name *inductive conformal inference* (Papadopoulos et al., 2002; Vovk et al., 2005). The split conformal algorithm summarized in Algorithm 2 is adapted from Lei et al. (2015). Its key coverage properties are given in Theorem 2.2. (Here, and henceforth when discussing split conformal inference, we assume that the sample size n is even, for simplicity, as only very minor changes are needed when n is odd.) The following theorem, proved in Appendix A.1, summarizes the validity of split conformal.

Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, fitting algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$, $\mathcal{I}_2$
$\widehat{\mu} = \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\big)$
$R_i = |Y_i - \widehat{\mu}(X_i)|$, $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$
Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

Theorem 2.2. *If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. draw (X_{n+1}, Y_{n+1}),*

$$\mathbb{P}\Big(Y_{n+1} \in C_{\text{split}}(X_{n+1})\Big) \geq 1 - \alpha,$$

for the split conformal prediction band C_{split} constructed in Algorithm 2. Moreover, if we assume additionally that the residuals R_i, $i \in \mathcal{I}_2$ have a continuous joint distribution, then

$$\mathbb{P}\Big(Y_{n+1} \in C_{\text{split}}(X_{n+1})\Big) \leq 1 - \alpha + \frac{2}{n+2}.$$

In addition to being extremely efficient, compared to the original conformal method, split conformal inference can also hold an advantage in terms of memory requirements. For example, if the fitting procedure $\mathcal{A}$ (in the notation of Algorithm 2) involves variable selection, like the lasso or forward stepwise regression, then we only need to store the selected variables when we evaluate the fit at new points X_i, $i \in \mathcal{I}_2$, and compute residuals, for the ranking step. This can be a big savings in memory when the original variable set is very large, and the selected set is much smaller.

Split conformal prediction intervals also provide an approximate in-sample coverage guarantee, making them easier to illustrate and interpret using the given sample (X_i, Y_i), $i = 1, \ldots, n$, without need to obtain future draws. This is described next.

Theorem 2.3. *Under the conditions of Theorem 2.2, there exists an absolute constant $c > 0$ such that, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{2}{n}\sum_{i \in \mathcal{I}_2} \mathbb{1}\{Y_i \in C_{\text{split}}(X_i)\} - (1 - \alpha)\right| \geq \epsilon\right) \leq 2\exp\Big(cn^2(\epsilon - 4/n)_+^2\Big).$$

Remark 2.5. *Theorem 2.3 implies "half sample" in-sample coverage. It is straightforward to extend this result to the whole sample, by constructing another split conformal prediction band, but with the roles of $\mathcal{I}_1, \mathcal{I}_2$ reversed. This idea is further explored and extended in Section Section 4.1, where we derive Theorem 2.3 as a corollary of another more general result.*

2.3 Multiple Splits

Splitting improves dramatically on the speed of conformal inference, but introduces extra randomness into the procedure. One way to reduce this extra randomness is to combine inferences from several splits. Suppose that we split the training data N times, yielding split conformal prediction intervals $C_{\text{split},1}, \ldots, C_{\text{split},N}$ where each interval is constructed at level $1 - \alpha/N$. Then we define,

$$C_{\text{split}}^{(N)}(x) = \bigcap_{j=1}^{N} C_{\text{split},j}(x), \quad \text{over } x \in \mathbb{R}^d. \tag{9}$$

It follows, using a simple Bonferroni-type argument, that the prediction band $C_{\text{split}}^{(N)}$ has marginal coverage level at least $1 - \alpha$.

Multi-splitting as described above decreases the variability from splitting. But this may come at a price: it is possible that the width of $C_{\text{split}}^{(N)}$ grows with N, though this is not immediately obvious. Replacing α by α/N certainly makes the individual split conformal intervals larger. However, taking an intersection reduces the size of the final interval. Thus there is a "Bonferroni-intersection tradeoff."

The next result shows that, under rather general conditions as detailed in Section 6, the Bonferroni effect dominates and we hence get larger intervals as N increases. For this reason, we suggest using a single split. The proof is given in Appendix A.3.

Theorem 2.4. *Under Assumptions A0, A1, A2 with $\rho_n = o(n^{-1})$ (details given in Section 6), if $|Y - \mu_0(X)|$ has continuous distribution, then with probability tending to 1, $C_{\text{split}}^{(N)}(X)$ is wider than $C_{\text{split}}(X)$.*

Remark 2.6. *Multiple splits were also used in Meinshausen & Buhlmann (2010). However, the situation there is rather different, where the linear model is assumed correct and inference is performed on the coefficients in the true population linear model.*

2.4 Jackknife Prediction Intervals

Lying between the computational complexities of the full and split conformal methods is *jackknife prediction*. This method uses the quantiles of leave-one-out residuals to define prediction intervals, and is summarized in Algorithm 3.

An advantage of the jackknife method over the split conformal method is that it utilizes more of the training data when constructing the absolute residuals, and subsequently, the quantiles. This means that it can often produce intervals of shorter length. A clear

Algorithm 3 Jackknife Prediction Band

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, fitting algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
for $i \in \{1, \ldots, n\}$ **do**
  $\widehat{\mu}^{(-i)} = \mathcal{A}\big(\{(X_\ell, Y_\ell) : \ell \neq i\}\big)$
  $R_i = |Y_i - \widehat{\mu}^{(-i)}(X_i)|$
end for
$d =$ the kth smallest value in $\{R_i : i \in \{1, \ldots, n\}\}$, where $k = \lceil n(1-\alpha) \rceil$
Return $C_{\text{jack}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

disadvantage, however, is that its prediction intervals are not guaranteed to have valid coverage in finite samples. In fact, even asymptotically, its coverage properties do not hold without requiring nontrivial conditions on the base estimator. We note that, by symmetry, the jackknife method has the finite sample in-sample coverage property

$$\mathbb{P}\left(Y_i \in C_{\text{jack}}(X_i)\right) \geq 1 - \alpha, \quad \text{for all } i = 1, \ldots, n.$$

But in terms of out-of-sample coverage (true predictive inference), its properties are much more fragile. Butler & Rothman (1980) show that in a low-dimensional linear regression setting, the jackknife method produces asymptotic valid intervals under regularity conditions strong enough that they also imply consistency of the linear regression estimator. More recently, Steinberger & Leeb (2016) establish asymptotic validity of the jackknife intervals in a high-dimensional regression setting; they do not require consistency of the base estimator $\widehat{\mu}$ per say, but they do require a uniform asymptotic mean squared error bound (and an asymptotic stability condition) on $\widehat{\mu}$. The conformal method requires no such conditions. Even without these regularity conditions, the analyses in Butler & Rothman (1980); Steinberger & Leeb (2016) assume a standard linear model setup, where the regression function is itself a linear function of the features, the features are independent of the errors, and the errors are homoskedastic; none of these conditions are needed in order for the split conformal method (and full conformal method) to have finite simple validity.

3 Empirical Study

Now we examine empirical properties of the conformal prediction intervals under three simulated data settings. In each setting, the samples (X_i, Y_i), $i = 1, \ldots, n$ are generated in an i.i.d. fashion, by first specifying $\mu(x) = \mathbb{E}(Y_i \,|\, X_i = x)$, then specifying a distribution for $X_i = (X_i(1), \ldots, X_i(d))$, and lastly specifying a distribution for $\epsilon_i = Y_i - \mu(X_i)$ (from which we can form $Y_i = \mu(X_i) + \epsilon_i$). These specifications are described below. We write $N(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2, $SN(\mu, \sigma^2, \alpha)$ for the

skewed normal with skewness parameter α, $t(k)$ for the t-distribution with k degrees of freedom, and $\text{Bern}(p)$ for the Bernoulli distribution with success probability p.

Throughout, we will consider the following three experimental setups.

Setting A *(linear, classical)*: the mean $\mu(x)$ is linear in x; the features $X_i(1), \ldots, X_i(d)$ are i.i.d. $N(0,1)$; and the error ϵ_i is $N(0,1)$, independent of the features.

Setting B *(nonlinear, heavy-tailed)*: like Setting A, but where $\mu(x)$ is nonlinear in x, an additive function of B-splines of $x(1), \ldots, x(d)$; and the error ϵ_i is $t(2)$ (thus, without a finite variance), independent of the features.

Setting C *(linear, heteroskedastic, heavy-tailed, correlated features)*: the mean $\mu(x)$ is linear in x; the features $X_i(1), \ldots, X_i(d)$ are first independently drawn from a mixture distribution, with equal probability on the components $N(0,1)$, $SN(0,1,5)$, $\text{Bern}(0.5)$, and then given autocorrelation by redefining in a sequential fashion each $X_i(j)$ to be a convex combination of its current value and $X_i(j-1), \ldots, X_i((j-3) \wedge 1)$, for $j = 1, \ldots, d$; the error ϵ_i is $t(2)$, with standard deviation proportional to $|\mu(X_i)|^3$ (hence, clearly not independent of the features).

Setting A is a simple setup where classical methods are expected to perform well. Setting B explores the performance when the mean is nonlinear and the errors are heavy-tailed. Setting C provides a particularly difficult linear setting for estimation, with heavy-tailed, heteroskedastic errors and highly correlated features. All simulation results in the following subsections are averages over 50 repetitions. Additionally, all intervals are computed at the 90% nominal coverage level. The results can be directly reproduced using the code provided at `https://github.com/ryantibs/conformal`.

3.1 Comparisons to Parametric Intervals from Linear Regression

Here we compare the conformal prediction intervals based on the ordinary linear regression estimator to the classical parametric prediction intervals for linear models. The classical intervals are valid when the true mean is linear and the errors are both normal and homoskedastic, or are asymptotically valid if the errors have finite variance. Recall that the full and split conformal intervals are valid under essentially no assumptions, whereas the jackknife method requires at least a uniform mean squared error bound on the linear regression estimator in order to achieve asymptotic validity (Butler & Rothman, 1980; Steinberger & Leeb, 2016). We empirically compare the classical and conformal intervals across Settings A-C, in both low-dimensional ($n = 100$, $d = 10$) and high-dimensional ($n = 500$, $d = 490$) problems. Note that $d < n$ in this high-dimensional setting, so that the linear regression estimator and the corresponding intervals are well-defined.

In the low-dimensional problem with a linear mean function and normal, homoskedastic errors (Setting A, Table 1), all four methods give reasonable coverage. The parametric intervals are shorter than the conformal intervals, as the parametric assumptions are satisfied and d is small enough for the model to be estimated well. The full conformal interval is shorter than the split conformal interval, but comes at a higher computational cost.

In the other two low-dimensional problems (Settings B and C, Table 1), the assumptions supporting the classical prediction intervals break down. This drives the parametric intervals to over-cover, thus yielding much wider intervals than those from the conformal methods. Somewhat surprisingly (as the linear regression estimator in Settings B and C is far from accurate), the jackknife intervals maintain reasonable coverage at a reasonable length. The full conformal intervals continue to be somewhat shorter than the split conformal intervals, again at a computational cost. Note that the conformal intervals are also using a linear regression estimate here, yet their coverage is still right around the nominal 90% level; the coverage provided by the conformal approach is robust to the model misspecification.

In the high-dimensional problems (Table 2), the full conformal interval outperforms the parametric interval in terms of both length and coverage across all settings, even in Setting A where the true model is linear. This is due to poor accuracy of linear regression estimation when d is large. The jackknife interval also struggles, again because the linear regression estimate itself is so poor. The split conformal method must be omitted here, since linear regression is not well-defined once the sample is split ($n/2 = 250$, $d = 490$).

Because of the problems that high dimensions pose for linear regression, we also explore the use of ridge regression (Table 3). The parametric intervals here are derived in a similar fashion to those for ordinary linear regression (Burnaev & Vovk, 2014). For all methods except the split conformal method, we use a small fixed value of the ridge regression tuning parameter, $\lambda = 0.01$. For the split conformal method, the base estimator is taken to be ridge regression with cross-validated choice of λ, which is feasible because of the computational advantages of split conformal. The results show that the ridge penalty improves the performance of all methods, but that the conformal methods continue to outperform the parametric one. Moreover, the cross-validated tuning in the split conformal method leads it to dramatically outperform all others in both test error and interval length. The (rough) link between test error and interval length will be further observed in the next subsection.

3.2 Comparisons of Conformal Intervals Across Base Estimators

We explore the behavior of the conformal intervals across a variety of base estimators. We simulate data from Settings A-C, in both low ($n = 200$, $d = 20$) and high ($n = 200$,

Setting A

	Conformal	Jackknife	Split	Parametric
Coverage	0.897 (0.006)	0.885 (0.007)	0.904 (0.006)	0.888 (0.007)
Length	3.561 (0.047)	3.434 (0.046)	3.832 (0.063)	3.433 (0.038)
Time	1.064 (0.002)	0.001 (0)	0.001 (0)	0 (0)

Setting B

	Conformal	Jackknife	Split	Parametric
Coverage	0.895 (0.006)	0.886 (0.006)	0.895 (0.006)	0.923 (0.006)
Length	6.532 (0.494)	6.248 (0.492)	6.707 (0.186)	8.854 (1.771)
Time	1.084 (0.004)	0.001 (0)	0.001 (0)	0 (0)

Setting C

	Conformal	Jackknife	Split	Parametric
Coverage	0.906 (0.006)	0.894 (0.006)	0.899 (0.007)	0.937 (0.005)
Length	19.332 (0.885)	18.021 (0.832)	22.586 (1.539)	27.564 (2.159)
Time	1.116 (0.002)	0.001 (0)	0 (0)	0.001 (0)

Table 1: *Comparison of prediction intervals in low-dimensional problems with* $n = 100$, $d = 10$. *All quantities have been averaged over 50 repetitions, and the numbers in parantheses denote standard errors.*

$d = 2000$) dimensions, and in each case we apply forward stepwise regression (Efroymson, 1960), the lasso (Tibshirani, 1996), the elastic net (Zou & Hastie, 2005), sparse additive models (SPAM, Ravikumar et al., 2009), and random forests (Breiman, 2001). Figures 1 (low-dimensional) and 2 (high-dimensional) show the results of these experiments.

Each method is applied for a range of tuning parameter choices. For the sake of defining a common ground for comparisons, all values are plotted against the relative optimism[1] , defined to be

$$\text{(relative optimism)} = \frac{\text{(test error)} - \text{(train error)}}{\text{(test error)}}.$$

The only exception is the random forest, which gave stable errors over a variety of tuning choices; hence it is represented by a single point in each plot (corresponding to 500 trees in the low-dimensional problems, and 1000 trees in the high-dimensional problems). All curves in the figures represent an average over 50 repetitions, and error bars indicating

[1] In challenging settings (e.g., Setting C), the relative optimism can be negative. This is not an error, but occurs naturally for inflexible estimators and very difficult settings. This is unrelated to conformal inference and to observations about the plot shapes.

Setting A

	Conformal	Jackknife	Parametric
Coverage	0.896 (0.011)	0.886 (0.013)	0.862 (0.015)
Length	7.951 (0.136)	25.995 (0.895)	24.078 (0.834)
Time	109.095 (0.141)	0.546 (0)	0.209 (0)

Setting B

	Conformal	Jackknife	Parametric
Coverage	0.891 (0.016)	0.892 (0.017)	0.866 (0.019)
Length	44.606 (3.964)	60.363 (3.573)	56.099 (3.627)
Time	110.048 (0.056)	0.547 (0)	0.209 (0)

Setting C

	Conformal	Jackknife	Parametric
Coverage	0.902 (0.014)	0.895 (0.013)	0.869 (0.017)
Length	197.754 (16.956)	256.299 (18.811)	234.243 (17.156)
Time	112.417 (0.492)	0.551 (0.001)	0.211 (0)

Table 2: *Comparison of prediction intervals in high-dimensional problems with* $n = 500$, $d = 490$.

Setting A

	Conformal	Jackknife	Split+CV	Parametric
Coverage	0.896 (0.005)	0.891 (0.005)	0.902 (0.006)	1 (0)
Length	4.646 (0.038)	4.595 (0.039)	3.348 (0.026)	18.87 (0.159)
Test error	2.017 (0.046)	2.017 (0.046)	1.017 (0.021)	2.017 (0.046)
Time	109.095 (0.141)	0.546 (0)	1.107 (0.003)	0.209 (0)

Setting B

	Conformal	Jackknife	Split+CV	Parametric
Coverage	0.894 (0.004)	0.891 (0.004)	0.9 (0.005)	0.998 (0.001)
Length	10.186 (0.786)	9.993 (0.76)	5.831 (0.078)	46.873 (4.522)
Test error	20.827 (5.994)	20.827 (5.994)	12.085 (4.741)	20.827 (5.994)
Time	110.048 (0.056)	0.547 (0)	1.097 (0.003)	0.209 (0)

Setting C

	Conformal	Jackknife	Split+CV	Parametric
Coverage	0.902 (0.005)	0.898 (0.006)	0.897 (0.005)	0.997 (0.001)
Length	36.775 (2.095)	36.061 (2.042)	15.04 (0.321)	186.161 (12.389)
Test error	386.608 (100.768)	386.608 (100.768)	276.307 (100.623)	386.608 (100.768)
Time	112.417 (0.492)	0.551 (0.001)	1.152 (0.004)	0.211 (0)

Table 3: *Comparison of prediction intervals in high-dimensional problems with* $n = 500$, $d = 490$, *using ridge regularization.*

the standard errors. In all cases, we used the split conformal method for computational efficiency.

In the low-dimensional problems (Figure 1), the best test error are obtained by the linear methods (lasso, elastic net, stepwise) in the linear Setting A, and by SPAM in the nonlinear (additive) Setting B. In Setting C, all estimators perform quite poorly. We note that across all settings and estimators, no matter the test error performance, (1) the coverage of the conformal intervals is almost exactly 90%, the nominal level, and (2) the interval lengths seem to be highly correlated with test errors.

In the high-dimensional problems (Figure 2), the results are similar. The regularized linear estimators perform best in the linear Setting A, while SPAM dominates in the nonlinear (additive) Setting B and performs slightly better in Setting C. All estimators do reasonably well in Setting A and quite terribly in Setting C, according to test error. Nevertheless, across this range of settings and difficulties, the coverage of the conformal prediction intervals is again almost exactly 90%, and the lengths are highly correlated with test errors.

Close inspection shows that the correspondence between length and test error is not perfect. It can be seen in some settings that the shapes and minima of the length curves tend to favor slightly less optimistic parameter values than the test error curves. In other cases, comparison of methods by relative interval length gives different results to comparison on relative test error. For example, compare stepwise regression and lasso or elastic net in the high-dimensional version of Setting B.

4 Extensions of Conformal Inference

The conformal and split conformal methods, combined with basically any fitting procedure in regression, provide finite sample distribution-free predictive inferences. We describe some extensions of this framework to improve the interpretability and applicability of conformal inference.

4.1 In-Sample Split Conformal Inference

Given samples (X_i, Y_i), $i = 1, \ldots, n$, and a method that outputs a prediction band, we would often like to evaluate this band at some or all of the observed points X_i, $i = 1, \ldots, n$. This is perhaps the most natural way to visualize any prediction band. However, the conformal prediction methods from Section 2 are designed to give a valid prediction interval at a future point X_{n+1}, from the same distribution as $\{X_i, i = 1, \ldots, n\}$, but not yet observed.

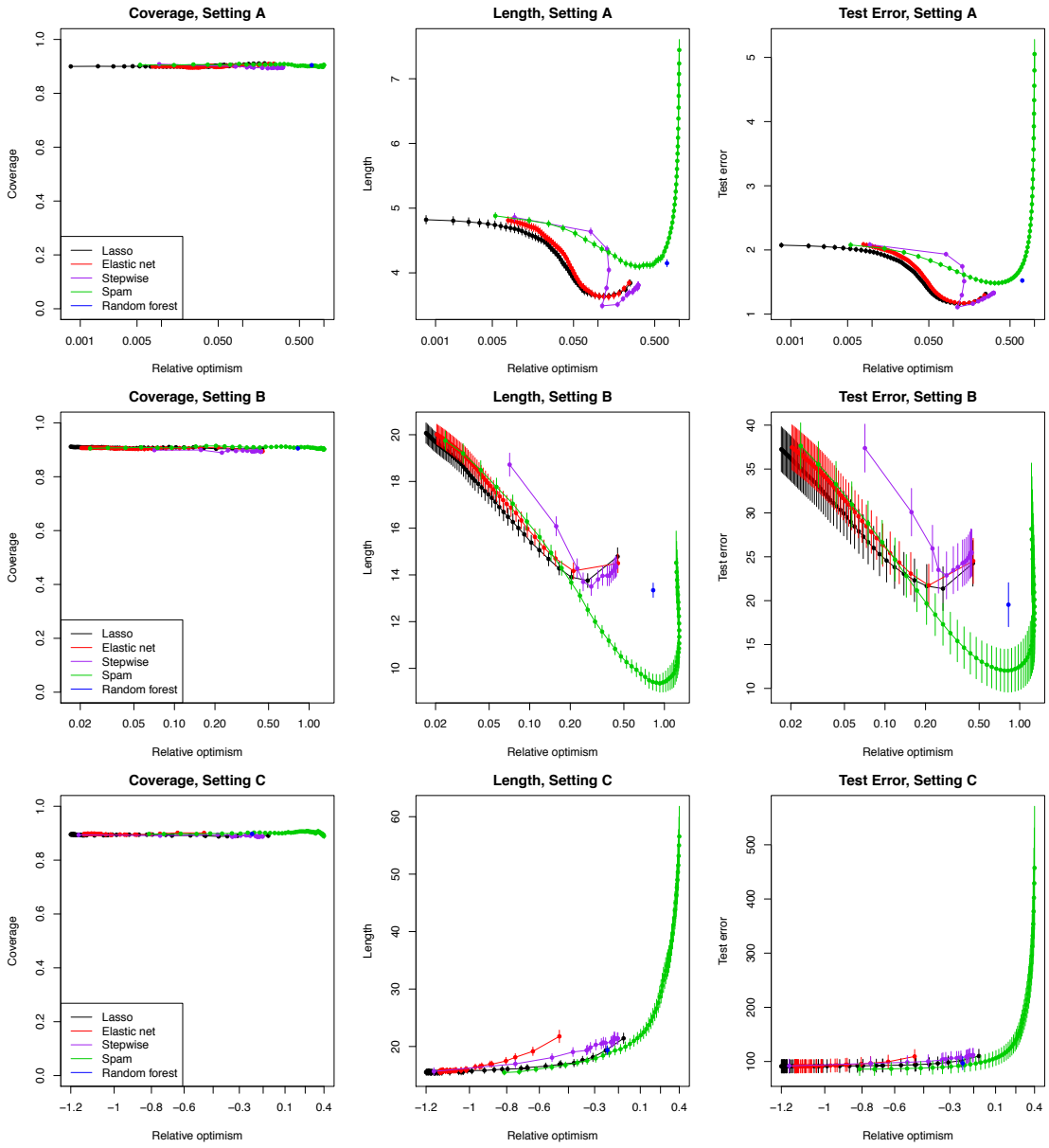

Figure 1: *Comparison of conformal prediction intervals in low-dimensional problems with* $n = 200$, $d = 20$, *across a variety of base estimators.*

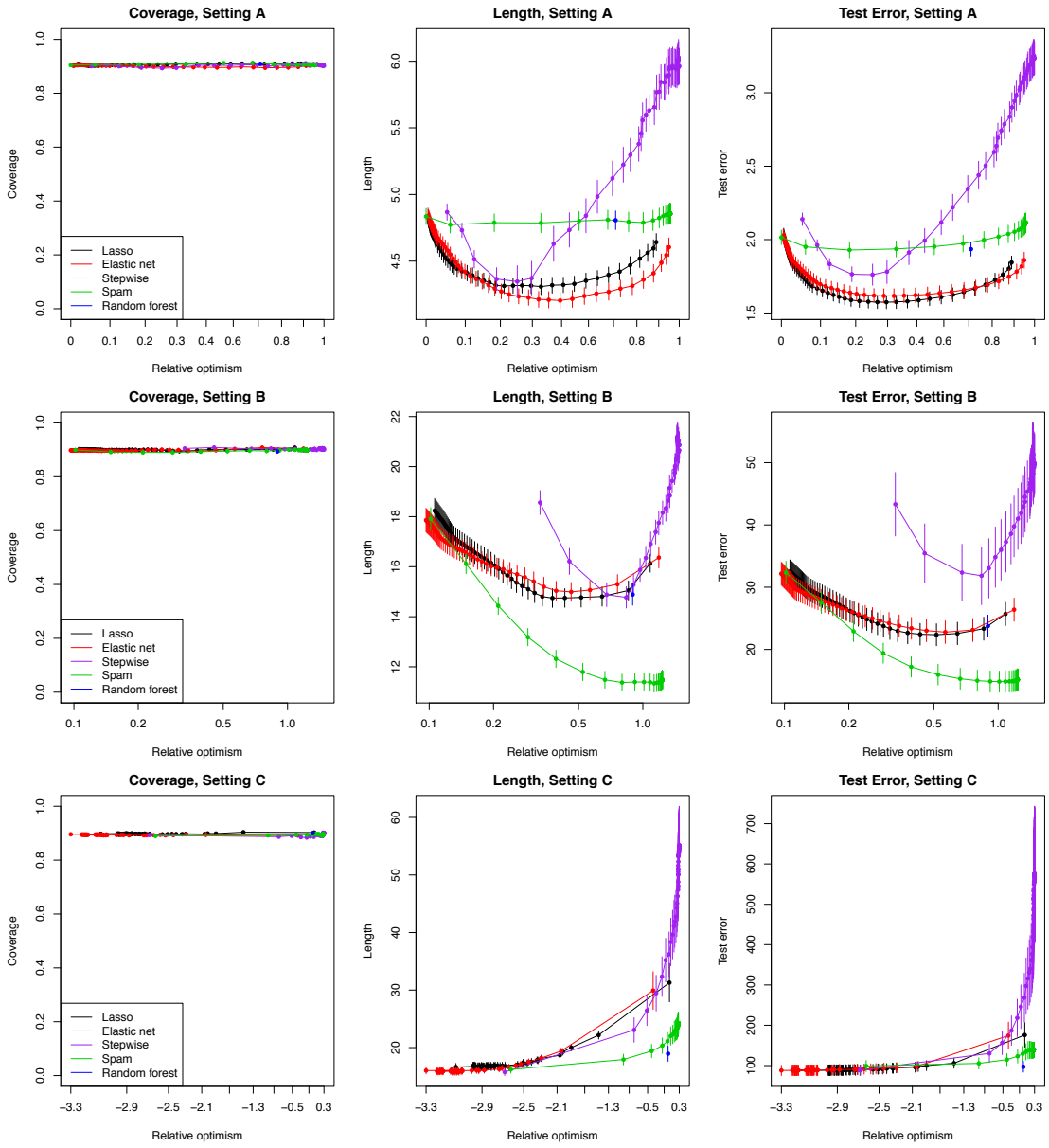

Figure 2: *Comparison of conformal prediction intervals in high-dimensional problems with* $n = 200$, $d = 2000$, *across a variety of base estimators.*

If we apply the full or split conformal prediction methods at an observed feature value, then it is not easy to establish finite sample validity of these methods.

A simple way to obtain valid in-sample predictive inference from the conformal methods is to treat each X_i as a new feature value and use the other $n-1$ points as the original features (running either the full or split conformal methods on these $n-1$ points). This approach has two drawbacks. First, it seriously degrades the computational efficiency of the conformal methods—for full conformal, it multiplies the cost of the (already expensive) Algorithm 1 by n, making it perhaps intractable for even moderately large data sets; for split conformal, it multiplies the cost of Algorithm 2 by n, making it as expensive as the jackknife method in Algorithm 3. Second, if we denote by $C(X_i)$ the prediction interval that results from this method at X_i, for $i = 1, \ldots, n$, then one might expect the empirical coverage $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \in C(X_i)\}$ to be at least $1-\alpha$, but this is not easy to show due to the complex dependence between the indicators.

Our proposed technique overcomes both of these drawbacks, and is a variant of the split conformal method that we call *rank-one-out* or ROO split conformal inference. The basic idea is quite similar to split conformal, but the ranking is conducted in a leave-one-out manner. The method is presented in Algorithm 4. For simplicity (as with our presentation of the split conformal method in Algorithm 2), we assume that n is even, and only minor modifications are needed for n odd. Computationally, ROO split conformal is very attractive. First, the fitting algorithm $\mathcal{A}$ (in the notation of Algorithm 4) only needs to be run twice. Second, for each split, the ranking of absolute residuals needs to be calculated just once; with careful updating, it can be reused in order to calculate the prediction interval for X_i in $O(1)$ additional operations, for each $i = 1, \ldots, n$.

By symmetry in their construction, the ROO split conformal intervals have the in-sample finite sample coverage property

$$\mathbb{P}\left(Y_i \in C_{\text{roo}}(X_i)\right) \geq 1-\alpha, \quad \text{for all } i = 1, \ldots, n.$$

A practically interesting performance measure is the empirical in-sample average coverage $n^{-1}\sum_{i=1}^n \mathbb{1}\{Y_i \in C_{\text{roo}}(X_i)\}$. Our construction in Algorithm 4 indeed implies a weak dependence among the random indicators in this average, which leads to a slightly worse coverage guarantee for the empirical in-sample average coverage, with the difference from the nominal $1-\alpha$ level being of order $\sqrt{\log n/n}$, with high probability.

Theorem 4.1. *If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for the ROO split conformal prediction band C_{roo} constructed in Algorithm 4, there exists an absolute constant $c > 0$, such that for all $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \in C_{\text{roo}}(X_i)\} \geq 1-\alpha-\epsilon\right) \geq 1 - 2\exp(-cn\epsilon^2).$$

Algorithm 4 Rank-One-Out Split Conformal

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, fitting algorithm $\mathcal{A}$
Output: Prediction intervals at each X_i, $i = 1, \ldots, n$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$, $\mathcal{I}_2$
for $k \in \{1, 2\}$ **do**
  $\widehat{\mu}_k = \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_k\}\big)$
  for $i \in I_{3-k}$ **do**
    $R_i = |Y_i - \widehat{\mu}_k(X_i)|$
  end for
  for $i \in I_{3-k}$ **do**
    $d_i =$ the mth smallest value in $\{R_i : i \in \mathcal{I}_{3-k} \setminus \{i\}\}$, where $m = \lceil n/2(1-\alpha) \rceil$
    $C_{\text{roo}}(X_i) = [\widehat{\mu}_k(X_i) - d_i, \widehat{\mu}_k(X_i) + d_i]$
  end for
end for
Return intervals $C_{\text{roo}}(X_i)$, $i = 1, \ldots, n$

Furthermore, if we assume additionally that the residuals $(R_i, i = 1, ..., n)$, have a continuous joint distribution, then for all $\epsilon > 0$,

$$\mathbb{P}\bigg(1 - \alpha - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \in C_{\text{roo}}(X_i)\} \leq 1 - \alpha + \frac{2}{n} + \epsilon\bigg) \geq 1 - 2\exp(-cn\epsilon^2).$$

The proof of Theorem 4.1 relies on McDiarmid's Inequality. It is conceptually straightforward but requires a careful tracking of dependencies and is deferred until Appendix A.2.

Remark 4.1. *An even simpler, and conservative approximation to each in-sample prediction interval $C_{\text{roo}}(X_i)$ is*

$$\widetilde{C}_{\text{roo}}(X_i) = [\widehat{\mu}_k(X_i) - \widetilde{d}_k, \widehat{\mu}_k(X_i) + \widetilde{d}_k], \tag{10}$$

where, using the notation of Algorithm 4, we define $\widetilde{d}_k$ to be the mth smallest of $\{R_i : i \in \mathcal{I}_{3-k}\}$, for $m = \lceil (1-\alpha)n/2 \rceil + 1$. Hence, now only a single sample quantile from the fitted residuals is needed for each split. As a price, each interval in (10) is wider than its counterpart from Algorithm 4 by at most one inter-quantile difference. Moreover, the results of Theorem 4.1 carry over to the prediction band $\widetilde{C}_{\text{roo}}$: in the second probability statement (trapping the empirical in-sample average coverage from below and above), we need only change the $2/n$ term to $6/n$.

In Appendix A.1, we prove Theorem 2.3 as a modification of Theorem 4.1.

4.2 Locally-Weighted Conformal Inference

The full conformal and split conformal methods both tend to produce prediction bands $C(x)$ whose width (i.e., length, we will use these two terms interchangeably) is roughly constant over $x \in \mathbb{R}^d$. In fact, for split conformal, the width is exactly constant over x. For full conformal, the width can vary slightly as x varies, but the difference is often negligible as long as the fitting method is moderately stable. This property—the width of $C(x)$ being roughly immune to x—is desirable if the spread of the residual $Y - \mu(X)$ does not vary substantially as X varies. However, in some scenarios this will not be true, i.e., the residual variance will vary nontrivially with X, and in such a case we want the conformal band to adapt correspondingly.

We now introduce an extension to the conformal method that can account for nonconstant residual variance. Recall that, in order for the conformal inference method to have valid coverage, we can actually use any conformity score function to generalize the definition of (absolute) residuals, as in (8) of Remark 2.4. For the present extension, we modify the definition of residuals in Algorithm 1 by inversely scaling the fitted residuals by an estimated error spread. Formally

$$R_{y,i} = \frac{|Y_i - \widehat{\mu}_y(X_i)|}{\widehat{\rho}_y(X_i)}, \; i = 1, \ldots, n, \quad \text{and} \quad R_{y,n+1} = \frac{|y - \widehat{\mu}_y(x)|}{\widehat{\rho}_y(x)}, \tag{11}$$

where now $\widehat{\rho}_y(x)$ denotes an estimate of the conditional mean absolute deviation (MAD) of $(Y - \mu(X))|X = x$, as a function of $x \in \mathbb{R}^d$. We choose to estimate the error spread by the mean absolute deviation of the fitted residual rather than the standard deviation, since the former exists in some cases in which the latter does not. Here, the conditional mean $\widehat{\mu}_y$ and conditional MAD $\widehat{\rho}_y$ can either be estimated jointly, or more simply, the conditional mean $\widehat{\mu}_y$ can be estimated first, and then the conditional MAD $\widehat{\rho}_y$ can be estimated using the collection of fitted absolute residuals $|Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \ldots, n$ and $|y - \widehat{\mu}_y(X_{n+1})|$. With the locally-weighted residuals in (11), the validity and accuracy properties of the full conformal inference method carry over.

For the split conformal and the ROO split conformal methods, the extension is similar. In Algorithm 2, we instead use locally-weighted residuals

$$R_i = \frac{|Y_i - \widehat{\mu}(X_i)|}{\widehat{\rho}(X_i)}, \; i \in \mathcal{I}_2, \tag{12}$$

where the conditional mean $\widehat{\mu}$ and conditional MAD $\widehat{\rho}$ are fit on the samples in $\mathcal{I}_1$, either jointly or in a two-step fashion, as explained above. The output prediction interval at a point x must also be modified, now being $[\widehat{\mu}(x) - \widehat{\rho}(x)d, \; \widehat{\mu}(x) + \widehat{\rho}(x)d]$. In Algorithm 4, analogous modifications are performed. Using locally-weighted residuals, as in (12), the validity and accuracy properties of the split methods, both finite sample and asymptotic, in

Theorems 2.2, 2.3 and 4.1, again carry over. The jackknife interval can also be extended in a similar fashion.

Figure 3 displays a simple example of the split conformal method using locally-weighted residuals. We let $n = 1000$, drew i.i.d. copies $X_i \sim \text{Unif}(0, 2\pi)$, $i = 1, \ldots, n$, and let

$$Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i, \ i = 1, \ldots, n,$$

for i.i.d. copies $\epsilon_i \sim N(0, 1)$, $i = 1, \ldots, n$. We divided the data set randomly into two halves $\mathcal{I}_1, \mathcal{I}_2$, and fit the conditional mean estimator $\widehat{\mu}$ on the samples from the first half $\mathcal{I}_1$ using a smoothing spline, whose tuning parameter was chosen by cross-validation. This was then used to produce a 90% prediction band, according to the usual (unweighted) split conformal strategy, that has constant width by design; it is plotted, as a function of $x \in \mathbb{R}$, in the top left panel of Figure 3. For our locally-weighted version, we then fit a conditional MAD estimator $\widehat{\rho}$ on $|Y_i - \widehat{\mu}(X_i)|$, $i \in \mathcal{I}_1$, again using a smoothing spline, whose tuning parameter was chosen by cross-validation. Locally-weighted residuals were used to produce a 90% prediction band, with locally-varying width, plotted in the top right panel of the figure. Visually, the locally-weighted band adapts better to the heteroskedastic nature of the data. This is confirmed by looking at the length of the locally-weighted band as a function of x in the bottom right panel. It is also supported by the improved empirical average length offered by the locally-weighted prediction band, computed over 5000 new draws from $\text{Unif}(0, 2\pi)$, which is 1.105 versus 1.247 for the unweighted band. Finally, in terms of average coverage, again computed empirically over the same 5000 new draws, both methods are very close to the nominal 90% level, with the unweighted version at 89.7% and the weighted version at 89.9%. Most importantly, the locally-weighted version does a better job here of maintaining a conditional coverage level of around 90% across all x, as shown in the bottom left panel, as compared to the unweighted split conformal method, which over-covers for smaller x and under-covers for larger x.

5 Model-Free Variable Importance: LOCO

In this section, we discuss the problem of estimating the importance of each variable in a prediction model. A key question is: how do we assess variable importance when we are treating the working model as incorrect? One possibility, if we are fitting a linear model with variable selection, is to interpret the coefficients as estimates of the parameters in the best linear approximation to the mean function μ. This has been studied in, e.g., Wasserman (2014); Buja et al. (2014); Tibshirani et al. (2016). However, we take a different approach for two reasons. First, our method is not limited to linear regression. Second, the spirit of our approach is to focus on predictive quantities and we want to measure variable

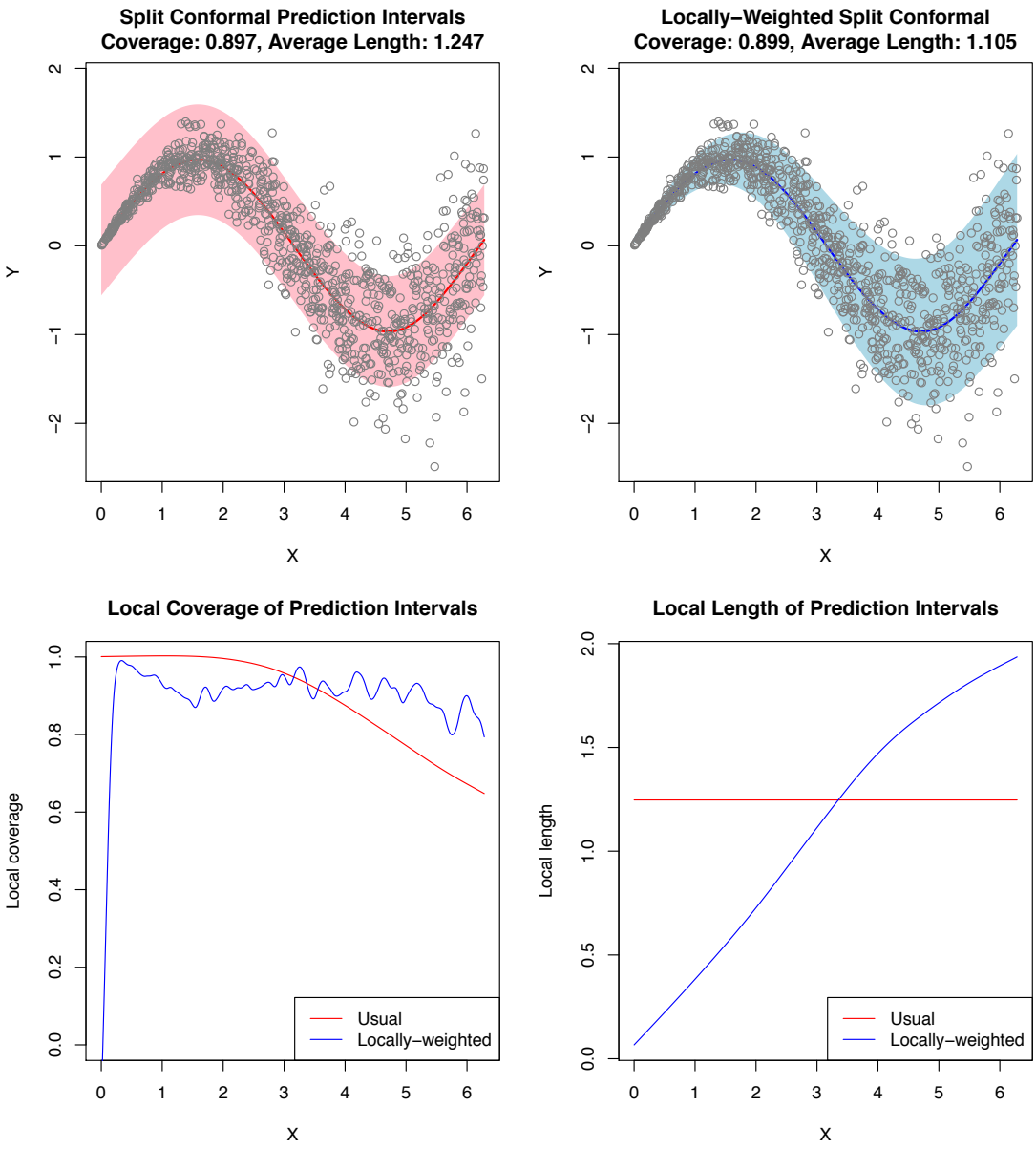

Figure 3: *A simple univariate example of the usual (unweighted) split conformal and locally-weighted split conformal prediction bands. The top left panel shows the split conformal band, and the top right shows the locally-weighted split conformal band; we can see that the latter properly adapts to the heteroskedastic nature of the data, whereas the former has constant length over all x (by construction). The bottom left and right panels plot the empirical local coverage and local length measures (which have been mildly smoothed as functions of x for visibility). The locally-weighted split conformal method maintains a roughly constant level of local coverage across x, but its band has a varying local length; the usual split conformal method exhibits precisely the opposite trends.*

importance directly in terms of prediction. Our approach is similar in spirit to the variable importance measure used in random forests (Breiman, 2001).

Our proposal, *leave-one-covariate-out* or LOCO inference, proceeds as follows. Denote by $\widehat{\mu}$ our estimate of the mean function, fit on data (X_i, Y_i), $i \in \mathcal{I}_1$ for some $\mathcal{I}_1 \subseteq \{1, \ldots, n\}$. To investigate the importance of the jth covariate, we refit our estimate of the mean function on the data set $\{(X_i(-j), Y_i) : i \in \mathcal{I}_1\}$, where in each $X_i(-j) = (X_i(1), \ldots, X_i(j-1), X_i(j+1), \ldots, X_i(d)) \in \mathbb{R}^{d-1}$, we have removed the jth covariate. Denote by $\widehat{\mu}_{(-j)}$ this refitted mean function, and denote the excess prediction error of covariate j, at a new i.i.d. draw (X_{n+1}, Y_{n+1}), by

$$\Delta_j(X_{n+1}, Y_{n+1}) = |Y_{n+1} - \widehat{\mu}_{(-j)}(X_{n+1})| - |Y_{n+1} - \widehat{\mu}(X_{n+1})|.$$

The random variable $\Delta_j(X_{n+1}, Y_{n+1})$ measures the increase in prediction error due to not having access to covariate j in our data set, and will be the basis for inferential statements about variable importance. There are two ways to look at $\Delta_j(X_{n+1}, Y_{n+1})$.

5.1 Local Measure of Variable Importance

Using the conformal prediction bands, we can construct a valid prediction interval for the random variable $\Delta_j(X_{n+1}, Y_{n+1})$, as follows. Let C denote a conformal prediction set for Y_{n+1} given X_{n+1}, having coverage $1-\alpha$, constructed from either the full or split methods—in the former, the index set used for the fitting of $\widehat{\mu}$ and $\widehat{\mu}_{(-j)}$ is $\mathcal{I}_1 = \{1, \ldots, n\}$, and in the latter, it is $\mathcal{I}_1 \subsetneq \{1, \ldots, n\}$, a proper subset (its complement $\mathcal{I}_2$ is used for computing the appropriate sample quantile of residuals). Now define

$$W_j(x) = \left\{ |y - \widehat{\mu}_{(-j)}(x)| - |y - \widehat{\mu}(x)| \ : \ y \in C(x) \right\}.$$

From the finite-sample validity of C, we immediately have

$$\mathbb{P}\left(\Delta_j(X_{n+1}, Y_{n+1}) \in W_j(X_{n+1}), \text{ for all } j = 1, \ldots, d\right) \geq 1 - \alpha. \tag{13}$$

It is important to emphasize that the prediction sets $W_1, \ldots, W_d$ are valid in finite sample, without distributional assumptions. Furthermore, they are uniformly valid over j, and there is no need to do any multiplicity adjustment. One can decide to construct $W_j(X_{n+1})$ at a single fixed j, at all $j = 1, \ldots, d$, or at a randomly chosen j (say, the result of a variable selection procedure on the given data (X_i, Y_i), $i = 1, \ldots, n$), and in each case the interval(s) will have proper coverage.

Just like the guarantees from conformal inference, the coverage statement in (13) is marginal over X_{n+1}, and in general, does not hold conditionally at $X_{n+1} = x$. However, to summarize

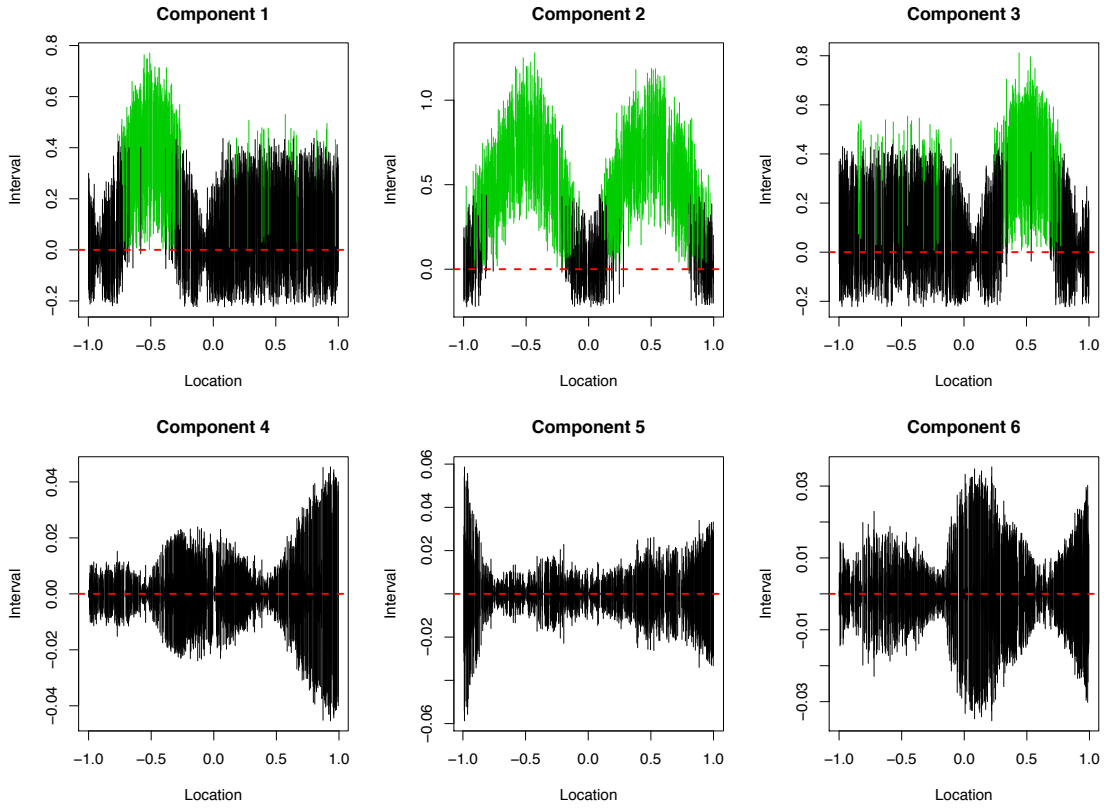

Figure 4: *In-sample prediction intervals for* $\Delta_j(X_i)$ *across all covariates* $j = 1, \ldots, 6$ *and samples* $i = 1, \ldots, 1000$*, in an additive model setting described in the text. Each interval that lies strictly above zero is colored in green.*

the effect of covariate j, we can still plot the intervals $W_j(X_i)$ for $i = 1, \ldots, n$, and loosely interpret these as making local statements about variable importance.

We illustrate this idea in a low-dimensional additive model, where $d = 6$ and the mean function is $\mu(x) = \sum_{j=1}^{6} f_j(x(j))$, with $f_1(t) = \sin(\pi(1+t))\mathbb{1}(t < 0)$, $f_2(t) = \sin(\pi t)$, $f_3(t) = \sin(\pi(1+t))\mathbb{1}(t > 0)$, and $f_4 = f_5 = f_6 = 0$. We generated $n = 1000$ i.i.d pairs (X_i, Y_i), $i = 1, \ldots, 1000$, where each $X_i \sim \text{Unif}[-1, 1]^d$ and $Y_i = \mu(X_i) + \epsilon_i$ for $\epsilon_i \sim N(0, 1)$. We then computed each interval $W_j(X_i)$ using the ROO split conformal technique at the miscoverage level $\alpha = 0.1$, using an additive model as the base estimator (each component modeled by a spline with 5 degrees of freedom). The intervals are plotted in Figure 4. We can see that many intervals for components $j = 1, 2, 3$ lie strictly above zero, indicating that leaving out such covariates is damaging to the predictive accuracy of the estimator. Furthermore, the locations at which these intervals lie above zero are precisely locations at

which the underlying components f_1, f_2, f_3 deviate significantly from zero. On the other hand, the intervals for components $j = 4, 5, 6$ all contain zero, as expected.

5.2 Global Measures of Variable Importance

For a more global measure of variable importance, we can focus on the distribution of $\Delta_j(X_{n+1}, Y_{n+1})$, marginally over (X_{n+1}, Y_{n+1}). We rely on a splitting approach, where the index set used for the training of $\widehat{\mu}$ and $\widehat{\mu}_{(-j)}$ is $\mathcal{I}_1 \subsetneq \{1, \ldots, n\}$, a proper subset. Denote by $\mathcal{I}_2$ its complement, and by $\mathcal{D}_k = \{(X_i, Y_i) : i \in \mathcal{I}_k\}$, $k = 1, 2$ the data samples in each index set. Define

$$G_j(t) = \mathbb{P}\Big(\Delta_j(X_{n+1}, Y_{n+1}) \leq t \;\Big|\; \mathcal{D}_1\Big),\ t \in \mathbb{R},$$

the distribution function of $\Delta_j(X_{n+1}, Y_{n+1})$ conditional on the data $\mathcal{D}_1$ in the first half of the data-split. We will now infer parameters of G_j such as its mean θ_j or median m_j. For the former parameter,

$$\theta_j = \mathbb{E}\Big[\Delta_j(X_{n+1}, Y_{n+1}) \;\Big|\; \mathcal{D}_1\Big],$$

we can obtain the asymptotic $1 - \alpha$ confidence interval

$$\left[\widehat{\theta}_j - \frac{z_{\alpha/2} s_j}{\sqrt{n/2}},\ \widehat{\theta}_j + \frac{z_{\alpha/2} s_j}{\sqrt{n/2}}\right],$$

where $\widehat{\theta}_j = (n/2)^{-1} \sum_{i \in \mathcal{I}_2} \Delta_j(X_i, Y_i)$ is the sample mean, s_j^2 is the analogous sample variance, measured on $\mathcal{D}_2$, and $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Similarly, we can perform a one-sided hypothesis test of

$$H_0 : \theta_j \leq 0 \quad \text{versus} \quad H_1 : \theta_j > 0$$

by rejecting when $\sqrt{n/2} \cdot \widehat{\theta}_j / s_j > z_\alpha$. Although these inferences are asymptotic, the convergence to its asymptotic limit is uniform (say, as governed by the Berry-Esseen Theorem) and independent of the feature dimension d (since $\Delta_j(X_{n+1}, Y_{n+1})$ is always univariate). To control for multiplicity, we suggest replacing α in the above with $\alpha/|S|$ where S is the set of variables whose importance is to be tested.

Inference for parameter θ_j requires existence of first and second moments for the error term. In practice it may be more stable to consider the median parameter

$$m_j = \text{median}\Big[\Delta_j(X_{n+1}, Y_{n+1}) \;\Big|\; \mathcal{D}_1\Big],$$

we can get nonasymptotic inferences using standard, nonparametric tests such as the sign test or the Wilcoxon signed-rank test, applied to $\Delta_j(X_i, Y_i)$, $i \in \mathcal{I}_2$. This allows us to test

$$H_0 : m_j \leq 0 \quad \text{versus} \quad H_1 : m_j > 0$$

with finite sample validity under essentially no assumptions on the distribution G_j (the sign test only requires continuity, and the Wilcoxon test requires continuity and symmetry). Confidence intervals for m_j can be obtained by inverting the (two-sided) versions of the sign and Wilcoxon tests, as well. Again, we suggest replacing α with $\alpha/|S|$ to adjust for multiplicity, where S is the set of variables to be tested.

We finish with an example of a high-dimensional linear regression problem with $n = 200$ observations and $d = 500$ variables. The mean function $\mu(x)$ was defined to be a linear function of $x(1), \ldots, x(5)$ only, with coefficients drawn i.i.d. from $N(0, 4)$. We drew $X_i(j) \sim N(0, 1)$ independently across all $i = 1, \ldots, 200$ and $j = 1, \ldots, 500$, and then defined the responses by $Y_i = \mu(X_i) + \epsilon_i$, for $\epsilon_i \sim N(0, 1)$, $i = 1, \ldots, 200$. A single data-split was applied, and the on the first half we fit the lasso estimator $\widehat{\mu}$ with the tuning parameter λ chosen by 10-fold cross-validation. The set of active predictors S was collected, which had size $|S| = 17$; the set S included the 5 truly relevant variables, but also 12 irrelevant ones. We then refit the lasso estimator $\widehat{\mu}_{(-j)}$ using the same cross-validation, with covariate j excluded, for each $j \in S$. On the second half of the data, we applied the Wilcoxon rank sum test to compute confidence intervals for the median excess test error due to variable dropping, m_j, for each $j \in S$. These intervals were properly corrected for multiplicity: each was computed at the level $1 - 0.1/17$ in order to obtain a simultaneous level $1 - 0.1 = 0.9$ of coverage. Figure 5 displays the results. We can see that the intervals for the first 5 variables are well above zero, and those for the next 12 all hover around zero, as desired. We have a number of other examples that show similar behavior but, due to space limitations, we are exlcuding these examples.

The problem of inference after model selection is a large topic and we are only dealing with the issue briefly here. In a future paper we thoroughly compare several approaches including LOCO.

6 Statistical Accuracy

Conformal inference offers reliable coverage under no assumptions other than i.i.d. In this section we investigate the statistical accuracy of conformal prediction intervals by bounding the length of the resulting intervals $C(X)$. Unlike coverage guarantee, such statistical accuracy must be established under appropriate regularity conditions on both the model and the fitting method. Our analysis starts from a very mild set of conditions, and moves toward the standard assumptions typically made in high dimensional regression, where we show that conformal methods achieve near-optimal statistical efficiency.

We first collect some common assumptions and notation that will be used throughout this section. Further assumptions will be stated when they are needed.

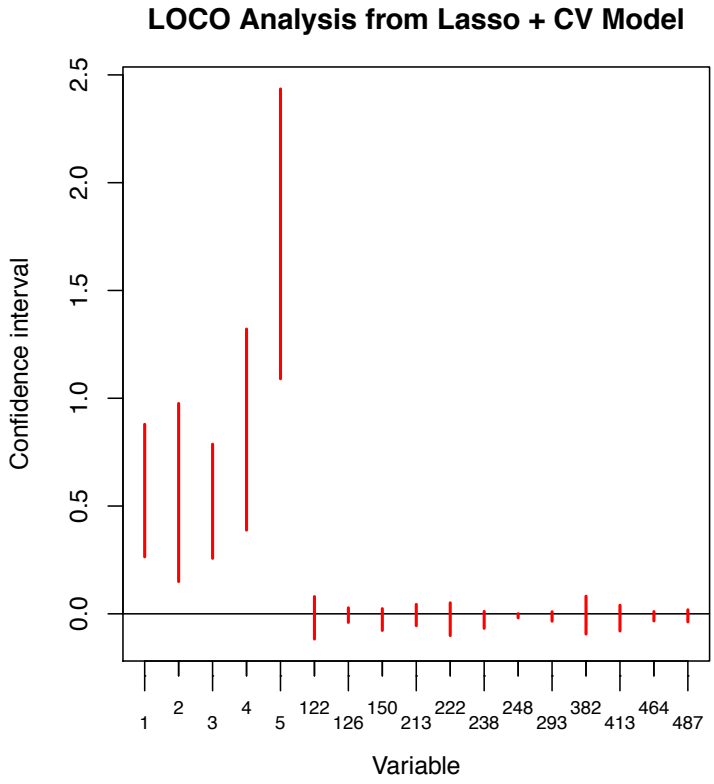

Figure 5: *Wilcoxon-based confidence intervals for the median excess test error due to variable dropping, for all selected variables in a high-dimensional linear regression example with* $n = 200$ *and* $d = 500$ *described in the text.*

> **Assumption A0.** (iid data) The data $(X_1, Y_1), ..., (X_n, Y_n)$ are iid from a common distribution P on $\mathbb{R}^d \times \mathbb{R}$, with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), \;\; x \in \mathbb{R}$.

Assumption A0 is the very basic assumption throughout this paper.

> **Assumption A1.** (independent and symmetric noise) The noise $\epsilon = Y - \mu(X)$ is independent of X, and the density function of ϵ is symmetric about 0 and non-increasing on $[0, \infty)$. We denote the distribution function of $|\epsilon|$ as F.

Assumption A1 is weaker than the assumptions usually made in the regression literature. In particular, we do not even require ϵ to have finite first moment. The symmetry and monotonicity requirement is for presentation convenience, and can be dropped by considering appropriate quantiles or density level sets of ϵ. The continuity of the distribution of ϵ also ensures that with probability 1 the fitted residuals will all be distinct, making inversion of

empirical distribution function easily tractable. However, our other assumptions, such as the stability or consistency of the base estimator (given below), may implicitly impose some further moment conditions on ϵ, making our assumption on ϵ comparable to the standard ones.

Two oracle bands. To quantify the accuracy of the prediction bands constructed with full and split conformal, we will compare their lengths to the length of the idealized prediction bands obtained by two oracles: the "super-oracle" and a regular oracle. The "super orcale" has complete knowledge of the regression function $\mu(x)$ and the error distribution, while a regular oracle has knowledge only of the residual distribution, i.e. of the distribution of $Y - \widehat{\mu}_n(X)$, where (X, Y) is a new draw from from P, independent of the sample used to compute the regression estimator $\widehat{\mu}_n$.

Assumptions A0 and A1 imply that the "super-oracle" prediction band is

$$C_s^*(x) = [\mu(x) \pm q_\alpha], \text{ where } q_\alpha \text{ is the upper } \alpha \text{ quantile of } |\epsilon|.$$

The band $C_s^*(x)$ is optimal in the sense that (i) it is has conditional coverage: $P(Y \in C(x) \mid X = x) \geq 1 - \alpha$, (ii) it has the shortest length among all bands with conditional coverage, (iii) it has the shortest average length among all bands with marginal coverage (Lei & Wasserman, 2014).

With a base fitting method $\mathcal{A}_n$ and a sample of size n, we can mimic the super oracle by substituting μ with $\widehat{\mu}_n$. In order to have valid prediction, the band needs to accommodate randomness of $\widehat{\mu}_n$ and the new independent sample (X, Y). Thus it is natural to consider the "oracle"

$$C_o^*(x) = [\widehat{\mu}_n(x) \pm q_{n,\alpha}], \text{ where } q_{n,\alpha} \text{ is the upper } \alpha \text{ quantile of } |Y - \widehat{\mu}_n(X)|.$$

We note that the definition of $q_{n,\alpha}$ is unconditional, so the randomness is regarding the $(n+1)$-tuple $((X_1, Y_1), ..., (X_n, Y_n), (X, Y))$. The band $C_o^*(x)$ is still impractical because the distribution of $|Y - \widehat{\mu}_n(X)|$ is unknown but its quantiles can be estimated. Unlike the super oracle band, in general the oracle band only offers marginal coverage: $\mathbb{P}(Y \in C_o^*(X)) \geq 1 - \alpha$ over the randomness of the $(n+1)$-tuple.

Our main theoretical results in this section can be summarized as follows.

1. If the base estimator is consistent, then the two oracle bands have similar length (Section 6.1).

2. If the base estimator is stable under re-sampling and small perturbation, then the conformal prediction bands are close to the oracle band (Section 6.2).

3. If the base estimator is consistent, then the conformal prediction bands are close to the super-oracle (Section 6.3).

All proofs of this section are deferred to Appendix A.3.

6.1 Comparing the Oracles

Intuitively, if $\widehat{\mu}_n$ is close to μ, then the two oracle bands should be close. Denote the estimation error by Δ_n:

$$\Delta_n(x) = \widehat{\mu}_n(x) - \mu(x)\,.$$

Theorem 6.1 (Comparing oracle bands)**.** *Under Assumptions A0, A1, let f be the density function of $|\epsilon|$. Assume further that f has continuous derivative that is uniformly bounded by a positive constant M. Let F_n be the CDF of $|Y - \widehat{\mu}_n(X)|$ and f_n the corresponding density. Then we have*

$$\sup_t |F_n(t) - F(t)| \leq (M/2)\mathbb{E}\Delta_n^2(X) \tag{14}$$

where the expectation is taken over the randomness of $\widehat{\mu}_n$ and X.

Furthermore, if f is lower bounded by $r > 0$ on $(q_\alpha - \eta, q_\alpha + \eta)$ for some $\eta > (M/2r)\mathbb{E}(\Delta_n^2(X))$, then

$$|q_{n,\alpha} - q_\alpha| \leq (M/2r)\mathbb{E}[\Delta_n^2(X)]\,. \tag{15}$$

In the definition of the oracle bands, the width is $2q_\alpha$ for the super oracle and $2q_{n,\alpha}$ for the oracle. Theorem 6.1 indicates that the oracle bands have similar width, with a difference proportional to $\mathbb{E}(\widehat{\mu}_n(X) - \mu(X))^2$. It is worth mentioning that we do not require the estimate $\widehat{\mu}_n$ to be consistent. Instead, Theorem 6.1 applies whenever $\mathbb{E}(\widehat{\mu}_n(X) - \mu(X))^2$ is smaller than some constant, as specified by the triplet (M, r, η) in the theorem. Moreover, it is also remarkable that the estimation error $\widehat{\mu}_n - \mu$ has only second order impact on the oracle prediction band. This is due to the assumption that ϵ has symmetric density.

6.2 Oracle Approximation Under Stability Assumptions

Now we provide sufficient conditions under which the split conformal and full conformal intervals approximate the regular oracle.

Case I: split conformal. In the case of split conformal, our main assumption is the *sampling stability*.

Assumption A2. (sampling stability) The estimated regression function satisfies

$$\mathbb{P}(\|\widehat{\mu}_n - \mu_0\|_\infty \geq \eta_n) \leq \rho_n,$$

for some $\eta_n = o(1)$ and $\rho_n = o(1)$ as $n \to \infty$, and some function μ_0.

We do not need to assume that μ_0 is anywhere close to the true regression function μ. We only need the estimate $\widehat{\mu}_n$ to concentrate around μ_0. So this is just a stability assumption rather than consistency assumption. For example, this is satisfied in nonparametric regression with over-smoothing. In case of $\mu_0 = \mu$, this becomes a sup-norm consistency assumption, and is satisfied, for example, for lasso estimators under standard assumptions, fixed-dimension ordinary least square with bounded predictor, and standard nonparametric regression on compact domain. Usually η_n has the form of $c(\log n/n)^{-r}$, and ρ_n is of order n^{-c}, for some fixed $c > 0$ (the choice of the constant c is arbitrary and will only impact the constant term in front of η_n).

When the sampling stability fails to hold, then conditioning on $\widehat{\mu}_n$ the residual $Y - \widehat{\mu}_n(X)$ may have a substantially different distribution as F_n. Therefore the split conformal interval can be substantially different from the oracle.

Remark 6.1. *The sup-norm bound required in Assumption A2 can be weakened to an $\ell_{p,X}$ norm bound with $p > 0$ where $\ell_{p,X}(g) = [\mathbb{E}_X|g(X)|^p]^{1/p}$ for any function g. The idea is that when $\ell_{p,X}$ norm bound holds, by Markov inequality the ℓ_∞ norm bound holds (with another vanishing sequence η_n) outside a small set whose probability is vanishing. Such a small set will have negligible impact on the quantiles. An example of this argument is given in the proof of Theorem 6.4.*

Theorem 6.2 (Split conformal and oracle)**.** *Let $\alpha \in (0, 1)$ be fixed and $C_{n,\text{spit}}$ be the split conformal interval fitted from $2n$ samples. Under Assumptions A0, A1, A2, assume further that $\widetilde{f}$, the density of $|Y - \mu_0(X)|$, is lower bounded by $r > 0$ in an open neighborhood of its upper α quantile. Let $\nu_{n,\text{split}}$ be the width of the split conformal band, then*

$$\nu_{n,\text{split}} - 2q_{n,\alpha} = O_P(\rho_n + \eta_n + n^{-1/2}).$$

Case II: full conformal. By definition, the full conformal band has varying width at different location of x. In practice the variation is negligible, but needs to be accounted for in our theoretical investigation. As a consequence, the probabilistic statement of our result on the full conformal method will cover the randomness of both $\widehat{\mu}_n$ as well as X, the point at which prediction band is to be constructed. Here X is a fresh random draw from the marginal distribution of the predictor.

Recall that for any candidate value y, we will fit the regression function with augmented data, where the $(n+1)$th data point is (X, y). We denote this fitted regression function by

$\widehat{\mu}_{n,(X,y)}$. Due to the arbitrariness of y, we must limit the range of y under consideration. Here we consider all $y \in \mathcal{Y} \subset \mathbb{R}^1$. Typically, $\mathcal{Y}$ is a compact interval of fixed length.

Like the split conformal analysis, our sufficient conditions for the full conformal band to approximate the oracle also requires the sampling stability (Assumption A2). But it also requires the following perturb-one sensitivity condition.

Assumption A3. (perturb-one sensitivity) There exist $\eta_n = o(1)$ and $\rho_n = o(1)$ such that

$$\mathbb{P}\left(\sup_{y\in\mathcal{Y}} \|\widehat{\mu}_n - \widehat{\mu}_{n,(X,y)}\|_\infty \geq \eta_n\right) \leq \rho_n\,.$$

The perturb-one sensitivity requires that the fitted regression function does not change much if we only perturb the y value of the last data entry. It is satisfied, for example, by kernel and local polynomial regression with non-overfitting bandwidth, least square regression with well-conditioned design, ridge regression, and even the Lasso under certain conditions (Thakurta & Smith, 2013).

For a similar reason as in Remark 6.1, we can weaken the ℓ_∞ norm requirement to an $\ell_{p,X}$ norm bound for any $p > 0$.

Theorem 6.3 (Oracle approximation of full Conformal)**.** *Under the same assumptions as in Theorem 6.2, assume in addition that Y is supported on $\mathcal{Y}$ such that Assumption 3 holds. Let $\nu_{n,\mathrm{conf}}(X)$ be the length of the full conformal prediction interval at X. Then*

$$\nu_{n,\mathrm{conf}}(X) - 2q_{n,\alpha} = O_P(\eta_n + \rho_n/r + n^{-1/2}) = o_P(1).$$

6.3 Super-oracle Approximation and Asymptotic Conditional Coverage

Combining the results in Sections 6.1 and 6.2, we immediately get $\nu_{n,\mathrm{split}} - 2q_\alpha = o_P(1)$ and $\nu_{n,\mathrm{conf}} - 2q_\alpha = o_P(1)$ when $\mathbb{E}\Delta_n^2(X) = o(1)$. In fact, when the estimate $\widehat{\mu}_n$ is consistent, we can further establish conditional coverage results for conformal prediction bands. That is, they have not only near-optimal length, but also near-optimal location.

The only additional assumption we need here is consistency of $\widehat{\mu}_n$. A natural condition would be $\mathbb{E}\Delta_n^2(X) = o(1)$. Our analysis uses an even weaker assumption.

Assumption A4. (Consistency of base estimator) The estimated regression function $\widehat{\mu}_n$ satisfies

$$\mathbb{P}\left\{\mathbb{E}_X\left[(\widehat{\mu}_n(X) - \mu(X))^2 \mid \widehat{\mu}_n\right] \geq \eta_n\right\} \leq \rho_n$$

for some $\eta_n = o(1)$ and $\rho_n = o(1)$.

It is easy to verify that Assumption A4 is implied by $\mathbb{E}\Delta_n^2(X) = o(1)$ using Markov's inequality. Many consistent estimators in the literature have this property, such as the Lasso estimator under sparse eigenvalue condition with appropriate tail behavior of X, and the kernel and local polynomial estimator for nonparametric regression on a compact domain.

We will show that conformal bands are close to the super oracle, and hence have approximately correct conditional coverage, which is formally defined as follows.

Definition (Asymptotic conditional coverage)**.** *Let $\{C_n(x) : x \in \mathbb{R}^d\}$ be a sequence of possibly random prediction bands. We say C_n has asymptotic conditional coverage at level $(1-\alpha)$ if there exists a sequence of possibly random sets $\Lambda_n \subset \mathbb{R}^d$ such that $\mathbb{P}(X \in \Lambda_n \mid \Lambda_n) = 1 - o_P(1)$ and*

$$\inf_{x \in \Lambda_n} \left| \mathbb{P}(Y \in C_n(x) \mid X = x) - (1-\alpha) \right| = o_P(1)\,.$$

Theorem 6.4 (Super-oracle approximation for split conformal)**.** *Under Assumptions A0, A1, A4, assuming in addition that $|Y - \mu(X)|$ has density bounded away from zero in an open neighborhood of its upper α quantile, the split conformal interval satisfies*

$$L(C_{\text{split}}(X) \triangle C_s^*(X)) = o_P(1)\,,$$

where $L(\cdot)$ denotes the Lebesgue measure and $\triangle$ denotes the symmetric difference between two sets. As a consequence, $C_{\text{split}}(\cdot)$ has asymptotic conditional coverage at level $1-\alpha$.

Remark 6.2. *The proof of Theorem 6.4 also works, with simple adaptation, even if η_n does not vanish. In this case we do not have consistency but the error will contain a term involving η_n.*

The super oracle approximation for full conformal prediction bands is similar, provided that the perturb-one sensitivity condition (A3) holds.

Theorem 6.5 (Super-oracle approximation for full conformal)**.** *Under same conditions as in Theorem 6.4, assuming in addition that A3 holds, we have*

$$L(C_{\text{conf}}(X) \triangle C_s^*) = o_P(1),$$

and $C_{\text{conf}}(\cdot)$ has asymptotic conditional coverage at level $1-\alpha$.

7 Conclusion

Current high-dimensional inference methods make strong assumptions while little is known about their robustness against model misspecification. We showed that if we focus on

prediction bands, almost all existing point estimators can be used to build valid prediction bands, even when the model is grossly misspecified, as long as the data are i.i.d. Conformal inference is similar to the jackknife, bootstrap and cross-validation in the use of symmetry of data. A remarkable difference in conformal inference is its "out-of-sample fitting". That is, unlike most existing prediction methods which fit a model using the training sample and then apply the fitted model to any new data points for prediction, the full conformal method refits the model each time when a new prediction is requested at a new value of X. An important and distinct consequence of such an "out-of-sample fitting" is the guaranteed finite sample coverage property.

The distribution-free coverage offered by conformal inference is in a marginal sense. The conditional coverage may be larger than $1-\alpha$ at some values of x and smaller than $1-\alpha$ at other values of x. This should not be viewed as a disadvantage of conformal inference, because the statistical accuracy of the conformal prediction bands crucially depends on the base estimator. In a sense, conformal inference broadens the scope and value of any point estimator with nearly no costs: if the estimator is accurate (which usually requires an approximately correctly specified model and a good regularization parameter), then the conformal prediction band is near-optimal; if the estimator is bad, we still have valid marginal coverage. As a result, it makes sense to use conformal prediction band as a diagnostic and comparison tool for regression function estimators.

There are many directions in conformal inference that are worth exploring. Here we give a short list. First, we need to better understand the trade-off between full and split conformal methods. The split conformal is fast but at the cost of less accurate inference. Also in practice it would be desirable to reduce the additional randomness caused by splitting the data. In this paper we show that simply taking intersections of bands from multiple splits leads to less accurate results. It would be practically appealing to develop novel methods that more efficiently combine results from multiple splits. Second, it would be interesting to see how conformal inference can help with model-free variable selection. Our leave-one-covariate-out (LOCO) method is a first step in this direction. However, the current version of LOCO based on excess prediction error can only be implemented with the split conformal method due to computational reasons. When split conformal is used, the inference is then conditional on the model fitted in the first half of the data. The effect of random splitting inevitably raises an issue of selective inference, which needs to be appropriately addressed. In a future paper, we will report on detailed comparisons of LOCO with other approaches to high-dimensional inference.

References

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, *41*(2), 802–837.

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

Buhlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, *19*(4), 1212–1242.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2014). Models as approximations: How random predictors and model violations invalidate classical inference in regression. ArXiv: 1404.1578.

Burnaev, E., & Vovk, V. (2014). Efficiency of conformalized ridge regression. *Proceedings of the Annual Conference on Learning Theory*, *25*, 605–622.

Butler, R., & Rothman, E. (1980). Predictive intervals based on reuse of the sample. *Journal of the American Statistical Association*, *75*(372), 881–889.

Efroymson, M. A. (1960). Multiple regression analysis. In *Mathematical Methods for Digital Computers*, vol. 1, (pp. 191–203). Wiley.

Hebiri, M. (2010). Sparse conformal predictors. *Statistics and Computing*, *20*(2), 253–266.

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*, 2869–2909.

Lee, J., Sun, D., Sun, Y., & Taylor, J. (2016). Exact post-selection inference with application to the lasso. *Annals of Statistics, to appear*.

Lei, J. (2014). Classification with confidence. *Biometrika*, *101*(4), 755–769.

Lei, J., Rinaldo, A., & Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, *74*(1), 29–43.

Lei, J., Robins, J., & Wasserman, L. (2013). Distribution free prediction sets. *Journal of the American Statistical Association*, *108*, 278–287.

Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, *76*(1), 71–96.

Meinshausen, N., & Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, *72*(4), 417–473.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, (pp. 345–356). Springer.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B*, *71*(5), 1009–1030.

Steinberger, L., & Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. ArXiv: 1602.05801.

Thakurta, A. G., & Smith, A. (2013). Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, (pp. 819–850).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association, to appear*.

van de Geer, S., Buhlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, *42*(3), 1166–1201.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

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

Wasserman, L. (2014). Discussion: A significance test for the lasso. *Annals of Statistics*, *42*(2), 501–508.

Zhang, C.-H., & Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, *76*(1), 217–242.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.

A Technical Proofs

A.1 Proofs of Section 2

Proof of Theorem 2.1. The first part (the lower bound) comes directly from the definition of the conformal interval in (7), and the (discrete) p-value property in (6). We focus on the second part (upper bound). Define $\alpha' = \alpha - 1/(n+1)$. By assuming a continuous joint distribution of the fitted residuals, we know that the values $R_{y,1}, \ldots, R_{y,n+1}$ are all distinct with probability one. The set C_{conf} in (7) is equivalent to the set of all points y such that $R_{y,n+1}$ ranks among the $\lceil (n+1)(1-\alpha) \rceil$ smallest of all $R_{y,1}, \ldots, R_{y,n+1}$. Consider now the set $D(X_{n+1})$ consisting of points y such that $R_{y,n+1}$ is among the $\lceil (n+1)\alpha' \rceil$ largest. Then by construction

$$\mathbb{P}\Big(Y_{n+1} \in D(X_{n+1})\Big) \geq \alpha',$$

and yet $C_{\text{conf}}(X_{n+1}) \cap D(X_{n+1}) = \emptyset$, which implies the result. ∎

Proof of Theorem 2.2. The first part (lower bound) follows directly by symmetry between the residual at (X_{n+1}, Y_{n+1}) and those at (X_i, Y_i), $i \in \mathcal{I}_2$. We prove the upper bound in the second part. Assuming a continuous joint distribution of residual, and hence no ties, the set $C_{\text{split}}(X_{n+1})$ excludes the values of y such that $|y - \widehat{\mu}(X_{n+1})|$ is among the $(n/2) - \lceil (n/2+1)(1-\alpha) \rceil$ largest in $\{R_i : i \in \mathcal{I}_2\}$. Denote the set of these excluded points as $D(X_{n+1})$. Then again by symmetry,

$$\mathbb{P}\Big(Y_{n+1} \in D(X_{n+1})\Big) \geq \frac{(n/2) - \lceil (n/2+1)(1-\alpha) \rceil}{n/2+1} \geq \alpha - 2/(n+2),$$

which completes the proof. ∎

Proof of Theorem 2.4. Without loss of generality, we assume that the sample size is $2n$. The individual split conformal interval has length infinity if $\alpha/N < 1/n$. Therefore, we only need to consider $2 \leq N \leq \alpha n \leq n$. Also in this proof we will ignore all the rounding issues by directly working with the empirical quantiles. The differences caused by rounding is negligible.

For each $1 \leq j \leq N$, $C_{\text{split},j}(X)$, the jth split conformal prediction band at X, is an interval with radius $\widehat{F}_{n,j}^{-1}(1-\alpha/N)$, where $\widehat{F}_{n,j}$ is the empirical CDF of fitted absolute residuals in the ranking subsample in the jth split.

We focus on the event $\{\sup_{1 \leq j \leq N} \|\widehat{\mu}_j - \mu_0\|_\infty < \eta_n\}$, which has probability at least $1 - N\rho_n \geq 1 - n\rho_n \to 1$. In this event the length of $C_{\text{split}}^{(N)}(X)$ is at least

$$2 \min_{1 \leq j \leq N} \widetilde{F}_{n,j}^{-1}(1-\alpha/N) - 2\eta_n \,,$$

where $\widetilde{F}_{n,j}$ is the empirical CDF of the absolute residuals in the ranking subsample in the jth split.

Without loss of generality, let $C_{\text{split}}(X) = C_{\text{split},1}(X)$, with length no more than $2\widetilde{F}_{n,1}^{-1}(1-\alpha) + 2\eta_n$ on the events we focus on. Therefore, it suffices to show that

$$P\left[\widetilde{F}_{n,1}^{-1}(1-\alpha) < \widetilde{F}_{n,j}^{-1}(1-\alpha/N) - 2\eta_n, \quad \forall\, 1 \le j \le N\right] \to 1\,. \tag{16}$$

Let $\widetilde{F}$ be the CDF of $|Y - \mu_0(X)|$. Note that it is $\widetilde{F}_j$, instead of $\widehat{F}_j$, that corresponds to F. By Dvoretzky-Kiefer-Wolfowitz inequality we have,

$$\begin{aligned} P\left[\widetilde{F}_{n,j}^{-1}(1-\alpha/N) \le \widetilde{F}^{-1}(1-\alpha/1.6)\right] &\le P\left[\|\widetilde{F}_{n,j} - \widetilde{F}\|_\infty \ge \alpha(1/1.6 - 1/N)\right] \\ &\le P\left[\|\widetilde{F}_{n,j} - \widetilde{F}\|_\infty \ge \alpha/8\right] \\ &\le 2\exp\left(-n\alpha^2/32\right) \end{aligned}$$

Taking union bound

$$P\left[\inf_{1\le j\le N} \widetilde{F}_{n,j}^{-1}(1-\alpha/N) \le \widetilde{F}^{-1}(1-\alpha/1.6)\right] \le 2N\exp(-n\alpha^2/32)\,.$$

On the other hand

$$P\left[\widetilde{F}_{n,1}^{-1}(1-\alpha) \ge \widetilde{F}^{-1}(1-\alpha/1.4)\right] \le P\left[\|\widetilde{F}_{n,1} - \widetilde{F}\|_\infty \ge \alpha(1-1/1.4)\right] \le 2\exp(-n\alpha^2/8)\,.$$

So with probability at least $1 - 2\exp(-n\alpha^2/8) - 2N\exp(-n\alpha^2/32)$ we have

$$\inf_{1\le j\le N} \widetilde{F}_{n,j}^{-1}(1-\alpha/N) - \widetilde{F}_{n,1}^{-1}(1-\alpha) \ge \widetilde{F}^{-1}(1-\alpha/1.6) - \widetilde{F}^{-1}(1-\alpha/1.4) > 0\,.$$

Therefore we conclude (16) because $\eta_n = o(1)$. ∎

Proof of Theorem 2.3. Comparing the close similarity of $\widetilde{d}_1$ in (10) and d in Algorithm 2, we see that $\widetilde{d}_1 = d$ if we choose the target coverage levels to be $1-\alpha$ for the regular split conformal band C_{split}, and $1-(\alpha + 2\alpha/n)$ for the modified ROO split conformal band $\widetilde{C}_{\text{roo}}$. The desired result follows immediately by replacing α by $\alpha + 2\alpha/n$ in Theorem 4.1, as it applies to $\widetilde{C}_{\text{roo}}$ (explained in the above remark). ∎

A.2 Proofs of Section 4

Proof of Theorem 4.1. For notation simplicity we assume $I_1 = \{1, ..., n/2\}$, and r_i's are in increasing order for $i = 1, ..., n/2$. Let $m = \lceil (1-\alpha)n/2 \rceil$. Then $\mathbb{1}(Y_i \in C_{\text{roo}}(X_i)) =$

$\mathbf{1}(r_i \le d_i)$ where d_i is the mth smallest value in $r_1, ..., r_{i-1}, r_{i+1}, ..., r_{n/2}$. Now we consider changing a sample point, say, (X_j, Y_j), in $\mathcal{I}_1$ and denote the resulting possibly disordered residuals by $r'_1, ..., r'_{n/2}$, and define d'_i correspondingly. We now consider the question "for which values of $i \in \mathcal{I}_1 \backslash \{j\}$ can we have $\mathbb{1}(r_i \le d_i) \neq \mathbb{1}(r'_i \le d'_i)$?" Recall that we assume $r_1 \le r_2 \le ... \le r_{n/2}$. If $i \le m-1$ and $i \neq j$, then $d_i \ge r_m$, $d'_i \ge r_{m-1}$, $r_i = r'_i$, and hence $\mathbb{1}(r_i \le d_i) = \mathbb{1}(r'_i \le d'_i) = 1$. If $i \ge m+2$ and $i \neq j$, then using similar reasoning we have $\mathbb{1}(r_i \le d_i) = \mathbb{1}(r'_i \le d'_i) = 0$. Therefore, changing a single data point can change $\mathbb{1}(Y_i \in C_{\text{roo}}(X_i))$ for at most three values of i ($i = m, m+1, j$). Because the input sample points are independent, using McDiarmid's inequality we have

$$\mathbb{P}\left((2/n) \sum_{i \in \mathcal{I}_1} \mathbb{1}(Y_i \in C_{\text{roo}}(X_i)) \le 1 - \alpha - \epsilon \right) \le \exp(-cn\epsilon^2) .$$

The claim follows by switching $\mathcal{I}_1$ and $\mathcal{I}_2$ and adding the two inequalities up.

Now we consider the other direction. We only need to show that $\mathbb{P}(Y_j \notin C_{\text{roo}}(X_j)) \ge \alpha - 2/n$. Under the continuity assumption, with probability one the residuals r_j are all distinct. Let $j \in \mathcal{I}_k$ for $k = 1$ or 2. By construction, $C_{\text{roo}}(X_j)$ does not contain the y values such that $|y - \widehat{\mu}_{3-k}(X_j)|$ is among the $n/2 - \lceil (n/2)(1-\alpha) \rceil$ largest of $\{r_i : i \in \mathcal{I}_k \backslash \{j\}\}$. Denote this set by $D_{\text{roo}}(X_j)$. Then the standard conformal argument implies that

$$\mathbb{P}(Y_i \in D_{\text{roo}}(X_i)) \ge \frac{n/2 - \lceil (n/2)(1-\alpha) \rceil}{n/2} \ge \alpha - \frac{2}{n} .$$

And we can establish the corresponding exponential deviation inequality using the same reasoning above.

For $\widetilde{C}_{\text{roo}}(X_j)$, the lower bound follows from that of $C_{\text{roo}}(X_j)$ because $\widetilde{C}_{\text{roo}}(X_j) \supseteq C_{\text{roo}}(X_j)$. To prove the upper bound, realize that $\widetilde{C}_{\text{roo}}(X_j)$ does not contain y such that $|y - \widehat{\mu}_{3-k}(X_j)|$ is among the $(n/2) - \lceil (n/2)(1-\alpha) \rceil - 1$ largest of $\{r_i : i \in \mathcal{I}\}$. Hence it does not contain all y's such that $|y - \widehat{\mu}_{3-k}(X_j)|$ is among the $(n/2) - \lceil (n/2)(1-\alpha) \rceil - 2$ largest of $\{r_i : i \in \mathcal{I} \backslash \{j\}\}$. Comparing with the argument for C_{roo}, the extra -2 in the ranking changes $2/n$ to $6/n$ in the second probability statement in the theorem. ∎

A.3 Proofs of Section 6

Proof of Theorem 6.1. For any $t > 0$, by Fubini theorem and independence between ϵ and (Δ_n, X),

$$\begin{aligned} F_n(t) &= \mathbb{P}(|Y - \widehat{\mu}_n(X)| \le t) \\ &= \mathbb{P}(-t + \Delta_n(X) \le \epsilon \le t + \Delta_n(X)) \\ &= \mathbb{E}_{\widehat{\mu}_n, X}[F_0(t + \Delta_n(X)) - F_0(-t + \Delta_n(X))], \end{aligned} \tag{17}$$

where F_0 is the CDF of ϵ.

Let f_0 be the density function of F_0. We can approximate F_0 at any t using first order Taylor expansion

$$F_0(t+\delta) = F_0(t) + \delta f_0(t) + \delta^2 R(t,\delta)\,,$$

where $R(t,\delta) = 0.5\int_0^1 (1-u) f_0'(t+u\delta)du$ satisfies $\sup_{t,\delta}|R(t,\delta)| \le M/4$.

Next, using symmetry of F_0 we have $f_0(t) = f_0(-t)$ for all t, the RHS of (17) becomes

$$\begin{aligned}
&\mathbb{E}_{\widehat{\mu}_n,X}[F_0(t+\Delta_n(X)) - F_0(-t+\Delta_n(X))]\\
=&\mathbb{E}_{\widehat{\mu}_n,X}\left[F_0(t) + \Delta_n(X) f_0(t) + \Delta_n^2(X) R(t,\Delta_n(X))\right.\\
&\qquad \left. - F_0(-t) - \Delta_n(X) f_0(-t) - \Delta_n^2(X) R(-t,\Delta_n(X))\right]\\
=&F_0(t) - F_0(-t) + \mathbb{E}_{\widehat{\mu}_n,X}[\Delta_n^2(X) W]\\
=&F(t) + \mathbb{E}_{\widehat{\mu}_n,X}[\Delta_n^2(X) W]
\end{aligned}$$

where $W = R(t,\Delta_n(X)) - R(-t,\Delta_n(X))$. Equation (14) follows immediately since $|W| \le M/2$, almost surely.

Next we show equation (15). Because F has density at least $r > 0$ in an open neighborhood of q_α. If $t < q_\alpha - \delta$ for some $\delta > (M/2r)\mathbb{E}(\Delta_n^2(X))$ then

$$\begin{aligned}
F_q(t) \le& F(q_\alpha - \delta) + (M/2)\mathbb{E}(\Delta_n^2(X))\\
\le& F(q_\alpha) - \delta r + (M/2)\mathbb{E}(\Delta_n^2(X))\\
<& 1-\alpha\,.
\end{aligned}$$

Thus $q_{n,\alpha} \ge q_\alpha - (M/2r)\mathbb{E}(\Delta_n^2(X))$. Similarly we can show that $q_{n,\alpha} \le q_\alpha + (M/2r)\mathbb{E}(\Delta_n^2(X))$, and hence establish the claimed result. ∎

Proof of Theorem 6.2. Let $\widetilde{q}_\alpha$ be the upper α quantile of $|Y - \mu_0(X)|$. We first show that

$$|\widetilde{q}_\alpha - q_{n,\alpha}| \le \rho_n/r + \eta_n\,. \tag{18}$$

To see this, note that

$$\begin{aligned}
&P(|Y - \widehat{\mu}_n(X)| \le \widetilde{q}_{\alpha+\rho_n} - \eta_n)\\
\le& P\left[|Y - \widehat{\mu}_n(X)| \le \widetilde{q}_{\alpha+\rho_n} - \eta_n, \|\mu_0 - \widehat{\mu}_n\|_\infty \le \eta_n\right] + \rho_n\\
\le& \mathbb{P}\left[|Y - \mu_0(X)| \le \widetilde{q}_{\alpha+\rho_n}\right] + \rho_n\\
=& 1 - \alpha - \rho_n + \rho_n = 1-\alpha\,.
\end{aligned}$$

Thus $q_{n,\alpha} \ge \widetilde{q}_{\alpha+\rho_n} - \eta_n \ge \widetilde{q}_\alpha - \rho_n/r - \eta_n$. Similarly, we have $q_{n,\alpha} \le \widetilde{q}_{\alpha-\rho_n} + \eta_n \le \widetilde{q}_\alpha + \rho_n/r + \eta_n$.

The width of split conformal band equals $2\widehat{F}_n^{-1}(1-\alpha)$, where $\widehat{F}_n$ is the empirical CDF of $\{|Y_i - \widehat{\mu}_n(X_i)| : 1 \le i \le n\}$, and $\widehat{\mu}_n = \mathcal{A}_n((X_i, Y_i)_{i=n+1}^{2n})$. On the event $\{\|\widehat{\mu}_n - \mu_0\|_\infty \le \eta_n\}$, we have $\big||Y_i - \widehat{\mu}_n(X_i)| - |Y_i - \mu_0(X_i)|\big| \le \eta_n$ for all $1 \le i \le n$. Therefore, let $\widetilde{F}_n$ be the empirical CDF of $\{|Y_i - \mu_0(X_i)| : 1 \le i \le n\}$, we have

$$\mathbb{P}\left(|\widehat{F}_n^{-1}(1-\alpha) - \widetilde{F}_n^{-1}(1-\alpha)| \le \eta_n\right) \ge 1 - \rho_n\,. \tag{19}$$

Using standard empirical quantile theory for i.i.d data and using the assumption that $\widetilde{f}$ is bounded from below by $r > 0$ in a neighborhood of its upper α quantile, we have

$$\widetilde{F}_n^{-1}(1-\alpha) = \widetilde{q}_\alpha + O_P(n^{-1/2}). \tag{20}$$

Combining (18), (19), and (20), we conclude that

$$|\widehat{F}_n^{-1}(1-\alpha) - q_{n,\alpha}| = O_P(\eta_n + \rho_n + n^{-1/2})\,. \qquad \blacksquare$$

Proof of Theorem 6.3. We focus on the event

$$\{\|\widehat{\mu}_n - \mu_0\|_\infty \le \eta_n\} \cap \left\{\sup_{y\in\mathcal{Y}} \|\widehat{\mu}_n - \widehat{\mu}_{n,(X,y)}\|_\infty \le \eta_n\right\},$$

which, by assumption, has probability at least $1 - 2\rho_n \to 1$.

On this event, we have

$$\Big||Y_i - \widehat{\mu}_{n,(X,y)}(X_i)| - |Y_i - \mu_0(X_i)|\Big| \le 2\eta_n\,, \quad i = 1, ..., n\,. \tag{21}$$

$$\Big||y - \widehat{\mu}_{n,(X,y)}(X)| - |y - \mu_0(X)|\Big| \le 2\eta_n\,. \tag{22}$$

With (21) and (22), by the construction of full conformal prediction interval we can directly verify the following two facts.

1. $y \in C_{\text{conf}}(X)$ if $|y - \mu_0(X)| \le \widetilde{F}_n^{-1}(1-\alpha) - 4\eta_n$;
2. $y \notin C_{\text{conf}}(X)$ if $|y - \mu_0(X)| \ge \widetilde{F}_n^{-1}(1-(\alpha - 3/n)) + 4\eta_n$,

where $\widetilde{F}_n$ is the empirical CDF of $\{|Y_i - \mu_0(X_i)| : 1 \le i \le n\}$.

Therefore, the length of $C_{\text{conf}}(X)$ satisfies

$$\nu_{n,\text{conf}}(X) = 2\widetilde{q}_\alpha + O_P(\eta_n + n^{-1/2})\,.$$

The claimed result follows by further combining the above equation with (18). $\blacksquare$

Proof of Theorem 6.4. Without loss of generality, we assume that $C_{\text{split}}(\cdot)$ is obtained using $2n$ data points. The proof consists of two steps. First we show that $\widehat{\mu}_n(X) - \mu(X) = o_P(1)$. Second we show that $\widehat{F}_n^{-1}(1-\alpha) - q_\alpha = o_P(1)$, where $\widehat{F}_n$ is the empirical CDF of $\{|Y_i - \widehat{\mu}_n(X_i)| : 1 \le i \le n\}$ with $\widehat{\mu}_n = \mathcal{A}_n((X_i, Y_i)_{i=n+1}^{2n})$.

We now show the first part. Throughout this proof we focus on the event that $\mathbb{E}_X(\widehat{\mu}_n(X) - \mu(X))^2 \le \eta_n$, which has probability at least $1-\rho_n$ by Assumption A4. On this event, using Markov's inequality, we have that $\mathbb{P}(X \in B_n^c \mid \widehat{\mu}_n) \ge 1 - \eta_n^{1/3}$, where $B_n = \{x : |\widehat{\mu}_n(x) - \mu(x)| \ge \eta_n^{1/3}\}$. Therefore we conclude that $\mathbb{P}_{X,\widehat{\mu}_n}(|\widehat{\mu}_n(X) - \mu(X)| \ge \eta_n^{1/3}) \le \eta_n^{1/3} + \rho_n \to 0$ as $n \to \infty$. Part 1 of the proof is complete.

For the second part, define $\mathcal{I}_1 = \{i : \ 1 \le i \le n, \ X_i \in B_n^c\}$ and $\mathcal{I}_2 = \{1, ..., n\} \backslash \mathcal{I}_1$. Note that B_n is independent of $(X_i, Y_i)_{i=1}^n$. Using Hoeffding's inequality conditionally on $\widehat{\mu}_n$ we have $|\mathcal{I}_2| \le n\eta_n^{1/3} + c\sqrt{n \log n} = o(n)$ with probability tending to 1, for some absolute constant $c > 0$. The probability holds both conditionally or unconditionally on $\widehat{\mu}_n$.

Let $\widehat{G}_{n,1}$ be the empirical CDF of $\{|Y_i - \widehat{\mu}_n(X_i)| : i \in \mathcal{I}_1\}$, and $\widetilde{G}_{n,1}$ be the empirical CDF of $\{|Y_i - \mu(X_i)| : i \in \mathcal{I}_1\}$. By definition of $\mathcal{I}_1$ we know that $\big||Y_i - \widehat{\mu}_n(X_i)| - |Y_i - \mu(X_i)|\big| \le \eta_n^{1/3}$ for all $i \in \mathcal{I}_1$. All empirical quantiles of $\widehat{G}_{n,1}$ and $\widetilde{G}_{n,1}$ are at most $O_P(\sqrt{n})$ apart, because $|\mathcal{I}_1| = n(1 + o_P(1))$.

The half width of $C_{\text{split}}(X)$ is $\widehat{F}_n^{-1}(1-\alpha)$. According to the definition of $\mathcal{I}_1$, we have

$$\widehat{G}_{n,1}^{-1}\left(1 - \frac{n\alpha}{|\mathcal{I}_1|}\right) \le \widehat{F}_n^{-1}(1-\alpha) \le \widehat{G}_{n,1}\left(1 - \frac{n\alpha - |\mathcal{I}_2|}{|\mathcal{I}_1|}\right).$$

Both $\frac{n\alpha}{|\mathcal{I}_1|}$ and $\frac{n\alpha - |\mathcal{I}_2|}{|\mathcal{I}_1|}$ are $\alpha + o_P(1)$. As a result we conclude that

$$\widehat{F}_n^{-1}(1-\alpha) - q_\alpha = o_P(1).$$

The second part of the proof is complete. ∎

Proof of Theorem 6.5. The proof naturally combines those of Theorems 6.3 and 6.4.

Using the same argument as in the proof of Theorem 6.4, we can define the set B_n and index sets $\mathcal{I}_1$, $\mathcal{I}_2$.

Now we assume the event $\{X \in B_n^c\}$, which has probability tending to 1. Then by definition of B_n and the fact that $\eta_n \le \eta_n^{1/3}$ we have

$$\Big||Y_i - \widehat{\mu}_{n,(X,y)}(X_i)| - |Y_i - \mu_0(X_i)|\Big| \le 2\eta_n^{1/3}, \quad \forall\ i \in \mathcal{I}_1. \tag{23}$$

$$\Big||y - \widehat{\mu}_{n,(X,y)}(X)| - |y - \mu_0(X)|\Big| \le 2\eta_n^{1/3}. \tag{24}$$

By definition of $C_{\text{conf}}(X)$ and following the same reasoning as in the proof of Theorem 6.3, we can verify the following facts:

1. $y \in C_{\text{conf}}(X)$ if $|y - \mu_0(X)| \le \widetilde{G}_{n,1}^{-1}\left(1 - \frac{n\alpha}{|\mathcal{I}_1|}\right) - 4\eta_n^{1/3}$;
2. $y \notin C_{\text{conf}}(X)$ if $|y - \mu_0(X)| \ge \widetilde{G}_{n,1}^{-1}\left(1 - \frac{n\alpha - |\mathcal{I}_2| - 3}{|\mathcal{I}_1|}\right) + 4\eta_n^{1/3}$,

where $\widetilde{G}_{n,1}$ is the empirical CDF of $\{|Y_i - \mu_0(X_i)| : i \in \mathcal{I}_1\}$.

Both $\frac{n\alpha}{|\mathcal{I}_1|}$ and $\frac{n\alpha - |\mathcal{I}_2| - 3}{|\mathcal{I}_1|}$ are $\alpha + o_P(1)$, and hence

$$\widetilde{G}_{n,1}^{-1}\left(1 - \frac{n\alpha}{|\mathcal{I}_1|}\right) = q_\alpha + o_P(1)\,, \quad \widetilde{G}_{n,1}^{-1}\left(1 - \frac{n\alpha - |\mathcal{I}_2| - 3}{|\mathcal{I}_1|}\right) = q_\alpha + o_P(1)\,.$$

Thus the lower (upper) end point of $C_{\text{conf}}(X)$ is $q_\alpha + o_P(1)$ below (above) $\mu(X)$. The proof is complete. ■