Flexible distribution-free conditional predictive bands using density estimators

Rafael Izbicki, Gilson Y. Shimizu, Rafael B. Stern
Federal University of São Carlos

Abstract

Conformal methods create prediction bands that control average coverage assuming solely i.i.d. data. Besides average coverage, one might also desire to control conditional coverage, that is, coverage for every new testing point. However, without strong assumptions, conditional coverage is unachievable. Given this limitation, the literature has focused on methods with asymptotical conditional coverage. In order to obtain this property, these methods require strong conditions on the dependence between the target variable and the features. We introduce two conformal methods based on conditional density estimators that do not depend on this type of assumption to obtain asymptotic conditional coverage: Dist-split and CD-split. While Dist-split asymptotically obtains optimal intervals, which are easier to interpret than general regions, CD-split obtains optimal size regions, which are smaller than intervals. CD-split also obtains local coverage by creating prediction bands locally on a partition of the features space. This partition is data-driven and scales to high-dimensional settings. In a wide variety of simulated scenarios, our methods have a better control of conditional coverage and have smaller length than previously proposed methods.

1 Introduction

Supervised machine learning methods predict a response variable, $Y \in \mathscr{Y}$, based on features, $\mathbf{X} \in \mathscr{X}$, using an i.i.d. sample, $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. While most methods yield point estimates, it is often more informative to present prediction bands, that is, a subset of $\mathscr{Y}$ with plausible values for Y (Neter et al., 1996).

A particular way of constructing prediction bands is through *conformal predictions* (Vovk et al., 2005, 2009). This methodology is appealing because it controls the *marginal coverage* of the prediction bands assuming solely i.i.d. data. Specifically, given a new instance, $(\mathbf{X}_{n+1}, Y_{n+1})$, a conformal prediction, $C(\mathbf{X}_{n+1})$, satisfies

$$\mathbb{P}\left(Y_{n+1} \in C(\mathbf{X}_{n+1})\right) \geq 1-\alpha,$$

where $0 < 1-\alpha < 1$ is a desired coverage level. Besides marginal validity one might also wish for stronger guarantees. For instance, *conditional validity* holds when, for every $\mathbf{x}_{n+1} \in \mathscr{X}$,

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \geq 1-\alpha.$$

That is, conditional validity guarantees adequate coverage for each new instance and not solely on average across instances.

Unfortunately, conditional validity can be obtained only under strong assumptions about the the distribution of $(\mathbf{X}, Y)$ (Vovk, 2012; Lei and Wasserman, 2014; Barber et al., 2019). Given this result, effort has been focused on obtaining intermediate conditions. For instance, many conformal methods control *local coverage*:

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A) \geq 1-\alpha,$$

where A is a subset of $\mathscr{X}$ (Lei and Wasserman, 2014; Barber et al., 2019; Guan, 2019). These methods are based on computing conformal bands using only training instances that fall in A. However, to date, these methods do not scale to high-dimensional settings because it is challenging to create A that is large enough so that many training instances fall in A, and yet small enough so that

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A) \approx \mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}),$$

that is, local validity is close to conditional validity.

Another alternative to conditional validity is *asymptotic* conditional coverage (Lei et al., 2018). Under this property, conditional coverage converges to the specified level

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

as the sample size increases. That is, there exist random sets, Λ_n, such that $\mathbb{P}(X_{n+1} \in \Lambda_n | \Lambda_n) = 1 - o_{\mathbb{P}}(1)$ and

$$\sup_{\mathbf{x}_{n+1} \in \Lambda_n} \left| \mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) - 1 - \alpha \right| = o_{\mathbb{P}}(1).$$

In a regression context in which $\mathcal{Y} = \mathbb{R}$, Lei et al. (2018) obtains asymptotic conditional coverage under assumptions such as $Y = \mu(\mathbf{X}) + \epsilon$, where ϵ is independent of $\mathbf{X}$ and has density symmetric around 0. Furthermore, the proposed prediction band converges to the interval with the smallest interval among the ones with adequate conditional coverage.

Despite the success of these methods, there exists space for improvement. In many problems the assumption that ϵ is independent of $\mathbf{X}$ and has a density symmetric around 0 is unrealistic. For instance, in heteroscedastic settings (Neter et al., 1996) ϵ depends on $\mathbf{X}$. It is also common for ϵ to have an asymmetric or even multimodal distribution (Freeman et al., 2017). Furthermore, in these general settings, the smallest region with adequate conditional coverage might not be an interval, which is the outcome of most current methods.

1.1 Contribution

We propose new methods and show that they obtain asymptotic conditional coverage without assuming a particular type of dependence between the target and the features. Specifically, we propose two methods: `Dist-split` and `CD-split`. While `Dist-split` produces prediction bands that are intervals and easier to interpret, `CD-split` yields arbitrary regions, which are generally smaller and appealing for multimodal data. While `Dist-split` converges to an oracle interval, `CD-split` converges to an oracle region. Furthermore, since `CD-split` is based on a novel data-driven way of partitioning the feature space, it also controls local coverage even in high-dimensional settings. Table 1 summarizes the properties of these methods.

The proposed methods also have desirable computational properties. They are based on fast-to-compute split (inductive)-conformal bands (Papadopoulos, 2008; Vovk, 2012; Lei et al., 2018) and on novel conditional density estimation methods that scale to high-dimensional datasets (Lueckmann et al., 2017; Papamakarios et al., 2017; Izbicki and Lee, 2016, 2017; Dalmasso et al., 2019; Pospisil and Lee, 2019) Both methods are easy to compute and scale to large sample sizes as long as the conditional density estimator also does.

In a wide variety of simulation studies, we show that our proposed methods obtain better conditional coverage and smaller band length than alternatives in the literature. For example, Figure 1 illustrates `CD-split`, `Dist-split` and the reg-split method from Lei et al. (2018) on the toy example from Lei and Wasserman (2014). The bottom right plot shows that both `CD-split` and `Dist-split` get close to controlling conditional coverage. Since `Dist-split` can yield only intervals, `CD-split` obtains smaller bands in the region in which $\mathbf{Y}$ is bimodal. In this region `CD-split` yields a collection of intervals around each of the modes.

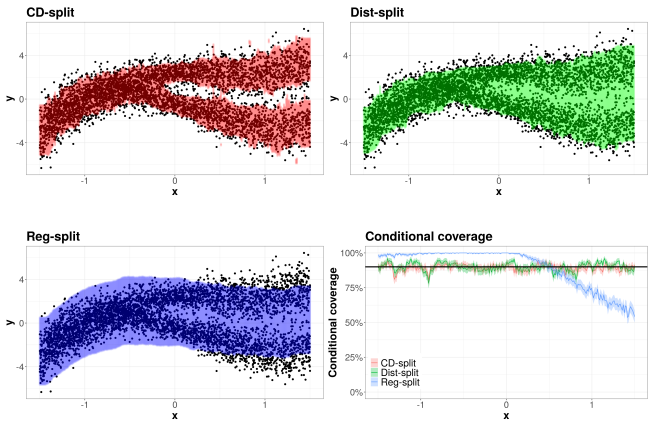

Figure 1: Comparison between `CD-split`, `Dist-split` and the reg-split method from Lei et al. (2018).

This paper is organized as follows: Section 2 and Section 3 introduce, respectively, `Dist-split`and `CD-split`. Experiments are shown in Section 4. All proofs can be found in the Appendix.

Notation. Unless stated otherwise, we study a univariate regression setting such that $\mathcal{Y} = \mathbb{R}$. Data from an i.i.d. sequence is split into two parts, $\mathbb{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ and $\mathbb{D}' = \{(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)\}$. Both datasets have the same size solely to simplify notation. Also, the new instance, $(\mathbf{X}_{n+1}, Y_{n+1})$, has the same distribution as the other sample units. Finally, $q(\alpha; \{t_1, \ldots, t_m\})$ is the α empirical quantile of $\{t_1, \ldots, t_m\}$.

2 `Dist-split`

The `Dist-split` method is based on the fact that, if $F(y|\mathbf{x})$ is the conditional distribution of Y_{n+1} given $\mathbf{X}_{n+1}$, then $F(Y_{n+1}|\mathbf{X}_{n+1})$ has uniform distribution. Therefore, if $\widehat{F}$ is close to F, then $\widehat{F}(Y_{n+1}|\mathbf{X}_{n+1})$ approximately uniform, and does not depend on $\mathbf{X}_{n+1}$. That is, obtaining marginal coverage for $\widehat{F}(Y_{n+1}|\mathbf{X}_{n+1})$ is close to obtaining conditional coverage.

Definition 2.1 (`Dist-split` prediction band)**.** Let $\widehat{F}(y|\mathbf{x}_{n+1})$ be an estimate based on $\mathbb{D}'$ of the conditional distribution of Y_{n+1} given $\mathbf{x}_{n+1}$. The `Dist-split` prediction band, $C(\mathbf{x}_{n+1})$, is

$$C(\mathbf{x}_{n+1}) := \left\{ y : q(.5\alpha; \mathcal{T}(\mathbb{D})) \leq \widehat{F}(y|\mathbf{x}_{n+1}) \leq q(1 - .5\alpha; \mathcal{T}(\mathbb{D})) \right\}$$
$$= \left[\widehat{F}^{-1}\left(q(.5\alpha; \mathcal{T}(\mathbb{D})) | \mathbf{x}_{n+1} \right); \widehat{F}^{-1}\left(q(1 - .5\alpha; \mathcal{T}(\mathbb{D})) | \mathbf{x}_{n+1} \right) \right]$$

Table 1: Properties of `Dist-split` and `CD-split`.

Method	Marginal coverage	Asymptotic conditional coverage	Local coverage	Prediction bands are intervals	Can be used for classification?
`Dist-split`	✓	✓	✗	✓	✗
`CD-split`	✓	✓	✓	✗	✓

where $\mathscr{T}(\mathbb{D}) = \{\widehat{F}(Y_i|\mathbf{X}_i), i = 1,\dots,n\}$ and $\widehat{F}^{-1}$ is the generalized inverse of a cdf.

Algorithm 1 shows an implementation of `Dist-split`.

Algorithm 1 `Dist-split`

Input: Data $(\mathbf{X}_i, Y_i)$, $i = 1,...,n$, miscoverage level $\alpha \in (0,1)$, algorithm $\mathscr{B}$ for fitting conditional cumulative distribution function

Output: Prediction band for $\mathbf{x}_{n+1} \in \mathbb{R}^d$

1: Randomly split $\{1,2,...,n\}$ into two subsets $\mathbb{D}$ e $\mathbb{D}'$
2: Fit $\hat{F} = \mathscr{B}(\{(\mathbf{X}_i, Y_i) : i \in \mathbb{D}'\})$ **// Estimate cumulative distribution function**
3: Let $\mathscr{T}(\mathbb{D}) = \{\hat{F}(y_i|\mathbf{x}_i), i \in \mathbb{D}\}$
4: Let $t_1 = q(\alpha/2; \mathscr{T}(\mathbb{D}))$ and $t_2 = q(1-\alpha/2; \mathscr{T}(\mathbb{D}))$ **// Compute the quantiles of the set $\mathscr{T}(\mathbb{D})$**
5: **return** $\left\{y : t_2 \geq \widehat{F}(y|\mathbf{x}_{n+1}) \geq t_1\right\}$

`Dist-split` adequately controls the marginal coverage. Furthermore, it exceeds the specified $1-\alpha$ coverage by at most $(n+1)^{-1}$. These results are presented in Theorem 2.2.

Theorem 2.2 (Marginal coverage)**.** *Let $C(\boldsymbol{X}_{n+1})$ be such as in Definition 2.1. If both $F(y|\mathbf{x})$ and $\widehat{F}(y|\mathbf{x})$ are continuous for every $\mathbf{x} \in \mathscr{X}$, then*

$$1-\alpha \leq \mathbb{P}(Y_{n+1} \in C(\boldsymbol{X}_{n+1})) \leq 1-\alpha+\frac{1}{n+1}.$$

Under additional assumptions `Dist-split` also obtains asymptotic conditional coverage and converges to an optimal oracle band. Two types of assumptions are required. First, that the conditional density estimator, $\widehat{F}$ is consistent. This assumption is an adaptation to density estimators of the consistency assumption for regression estimators in Lei et al. (2018). Also, we require that $F(y|\mathbf{x})$ is differentiable and $F^{-1}(\alpha^*|\mathbf{x})$ is uniformly smooth in a neighborhood of $.5\alpha$ and $1-.5\alpha$. These assumptions are formalized below.

Assumption 2.3 (Consistency of density estimator)**.** There exist $\eta_n = o(1)$ and $\rho_n = o(1)$ such that

$$\mathbb{P}\left(\mathbb{E}\left[\sup_{y\in\mathscr{Y}}\left(\widehat{F}(y|\mathbf{X}) - F(y|\mathbf{X})\right)^2 \middle| \widehat{F}\right] \geq \eta_n\right) \leq \rho_n$$

Assumption 2.4. For every $\mathbf{x} \in \mathscr{X}$, $F(y|\mathbf{x})$ is differentiable. Also, if $q_\alpha = F^{-1}(\alpha)$, then there exists $M^{-1} > 0$ such that $\inf_{\mathbf{x}} \frac{dF(y|\mathbf{x})}{dy} \geq M^{-1}$ in a neighborhood of $q_{0.5\alpha}$ and of $q_{1-0.5\alpha}$.

Given the above assumptions, `Dist-split` satisfies desirable theoretical properties. First, it obtains asymptotic conditional coverage. Also, `Dist-split` converges to the optimal *interval* according to the commonly used (Parmigiani and Inoue, 2009) loss function

$$L((a,b), Y_{n+1}) = \alpha(b-a) + (a - Y_{n+1})_+ + (Y_{n+1} - b)_+,$$

that is, `Dist-split` satisfies

$$C(\mathbf{X}_{n+1}) \approx \left[F^{-1}(.5\alpha|\mathbf{X}_{n+1}); F^{-1}(1-.5\alpha|\mathbf{X}_{n+1})\right]$$

These results are formalized in Theorem 2.5.

Theorem 2.5. *Let $C_n(\mathbf{X}_{n+1})$ be the prediction band in Definition 2.1 and $C^*(\mathbf{X}_{n+1})$ be the optimal prediction interval according to*

$$L((a,b), Y_{n+1}) = \alpha(b-a) + (a - Y_{n+1})_+ + (Y_{n+1} - b)_+.$$

Under Assumptions 2.3 and 2.4,

$$\lambda(C_n(\mathbf{X}_{n+1}) \Delta C^*(\mathbf{X}_{n+1})) = o_{\mathbb{P}}(1),$$

where λ is the Lebesgue measure.

Corollary 2.6. *`Dist-split` achieves asymptotic conditional coverage under Assumptions 2.3 and 2.4.*

`Dist-split` converges to the same oracle as recently proposed conformal quantile regression methods (Romano et al., 2019; Sesia and Candès, 2019). However, the experiments in Section 4 show that `Dist-split` usually outperforms these methods.

If the distribution of $Y|\mathbf{x}$ is not symmetric and unimodal, `Dist-split` may obtain larger regions than necessary. For example, a union of two intervals better represents a bimodal distribution than a single interval. The next section introduces `CD-split` which obtains prediction bands that are more general than intervals.

3 `CD-split`

The intervals output by `Dist-split` are wider than necessary when the target distribution is multimodal, such as in fig. 1. In order to overcome this issue, `CD-split` yields prediction bands that approximate $\{y : f(y|\mathbf{x}_{n+1}) > t\}$, the highest posterior region.

A possible candidate for this approximation is $\{y : \widehat{f}(y|\mathbf{x}_{n+1}) > t\}$, where $\widehat{f}$ is a conditional density estimator. However, the value of t that guarantees conditional coverage varies according to $\mathbf{x}$. Thus, in order to obtain conditional validity, it is necessary to choose t adaptively. This adaptive choice for t is obtained by making $C(\mathbf{x}_{n+1})$ depend only on samples close to $\mathbf{x}_{n+1}$, similarly as in Lei and Wasserman (2014); Barber et al. (2019); Guan (2019).

Definition 3.1 (`CD-split` prediction band)**.** Let $\widehat{f}(y|\mathbf{x}_{n+1})$ be a conditional density estimate obtained from data $\mathbb{D}'$ and $0 < 1-\alpha < 1$ be a coverage level. Let d be a distance on the feature space and $\mathbf{x}_1^c, \ldots, \mathbf{x}_J^c \in \mathscr{X}$ be centroids chosen so that $d(\mathbf{x}_i^c, \mathbf{x}_j^c) > 0$. Consider the partition of the feature space that associates each $\mathbf{x} \in \mathscr{X}$ to the closest $\mathbf{x}_j^c$, i.e., $\mathscr{A} = \{A_j : j = 1, \ldots, J\}$, where $A_j = \left\{\mathbf{x} \in \mathscr{X} : d(\mathbf{x}, \mathbf{x}_j^c) < d(\mathbf{x}, \mathbf{x}_k^c) \text{ for every } k \neq j\right\}$. The `CD-split` prediction band for Y_{n+1} is:

$$C(\mathbf{x}_{n+1}) = \left\{y : \widehat{f}(y|\mathbf{x}_{n+1}) \geq q(\alpha; \mathscr{T}(\mathbf{x}_{n+1}, \mathbb{D}))\right\},$$

where $\mathscr{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \{\widehat{f}(y_i|\mathbf{x}_i) : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\}$, where $A(\mathbf{x}_{n+1})$ is the element of $\mathscr{A}$ to which $\mathbf{x}_{n+1}$ belongs to.

Remark 1 (Multivariate responses)**.** *Although we focus on univariate targets, `CD-split` can be extended to the case in which $\mathbf{Y} \in \mathbb{R}^p$. As long as an estimate of $f(\mathbf{y}|\mathbf{x})$ is available, the same construction can be applied.*

The bands given by `CD-split` control local coverage in the sense proposed by Lei and Wasserman (2014).

Definition 3.2 (Local validity; Definition 1 of Lei and Wasserman (2014))**.** Let $\mathscr{A} = \{A_j : j \geq 1\}$ be a partition of $\mathscr{X}$. A prediction band C is locally valid with respect to $\mathscr{A}$ if, for every j and $\mathbb{P}$,

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A_j) \geq 1 - \alpha$$

Theorem 3.3 (Local and marginal validity)**.** *The `CD-split` band is locally valid with respect to $\mathscr{A}$. It follows from Lei and Wasserman (2014) that the `CD-split` band is also marginally valid.*

Although `CD-split` controls local coverage, its performance drastically depends on the chosen partition of the feature space. If the partition is not chosen well, local coverage may be far from conditional coverage. For instance, if the partition is defined according to the Euclidean distance (Lei and Wasserman, 2014; Barber et al., 2019), then the method will not scale to high-dimensional feature spaces. In these settings small Euclidean neighborhoods have few data points and, therefore, large neighborhoods must be taken. As a result, local coverage is far from conditional coverage. We overcome this drawback by using a specific data-driven partition. In order to build this metric, we start by defining the profile of a density, which is illustrated in fig. 2.

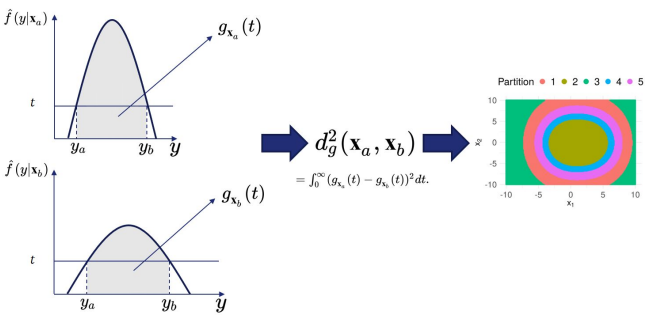

Figure 2: Illustration of the profile distance, which is used in `CD-split` for partitioning the feature space.

Definition 3.4 (Profile of a density)**.** For every $\mathbf{x} \in \mathbb{R}^d$ and $t \geq 0$, the profile of $\widehat{f}(y|\mathbf{x})$, $g_{\mathbf{x}}(t)$, is

$$g_{\mathbf{x}}(t) := \int_{\{y : \widehat{f}(y|\mathbf{x}) \geq t\}} \widehat{f}(y|\mathbf{x}) \, dy.$$

The profile of a density is the cumulative distribution function associated to its level sets. It is used to define the profile distance in the feature space.

Definition 3.5 (Profile distance)**.** The profile distance[1] between $\mathbf{x}_a, \mathbf{x}_b \in \mathscr{X}$ is

$$d_g^2(\mathbf{x}_a, \mathbf{x}_b) := \int_0^\infty \left(g_{\mathbf{x}_a}(t) - g_{\mathbf{x}_b}(t)\right)^2 dt,$$

Contrary to the Euclidean distance, the profile distance is appropriate even for high-dimensional data. For instance, two points might be far in Euclidean distance and still have similar conditional densities. In this case one would like these points to be on the same partition element. The profile obtains this result by measuring the distance between instances based on the distance between their conditional densities. By grouping points with similar conditional densities, the profile distance allows partition elements to be larger without compromising too much the approximation of local validity to conditional validity. This property is illustrated in the following examples.

Example 3.6. [Location family] Let $h(y)$ be a density, $\mu(\mathbf{x})$ a function, and $Y|\mathbf{x} \sim h(y - \mu(\mathbf{x}))$. In this case, $d_g(\mathbf{x}_a, \mathbf{x}_b) = 0$, for every $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^d$. For instance, if $Y|\mathbf{x} \sim N(\beta^t \mathbf{x}, \sigma^2)$, then all instances have the same profile. Indeed, in this special scenario, if `CD-split` uses a unitary partition, then conditional validity is obtained.

Example 3.7. [Irrelevant features] If $\mathbf{x}_S$ is a subset of the features such that $f(y|\mathbf{x}) = f(y|\mathbf{x}_S)$, then $d_g(\mathbf{x}_a, \mathbf{x}_b)$ does not depend on the irrelevant features, S^c. While irrelevant features do not affect the profile distance, they can

[1] The profile distance is a metric on the quotient space $\mathscr{X}/\sim$, where $\sim$ is the equivalence relation $\mathbf{x}_a \sim \mathbf{x}_b \iff g_{\mathbf{x}^a} = g_{\mathbf{x}^b}$ a.e.

have a large impact in the Euclidean distance in high-dimensional settings.

Also, if all samples that fall into the same partition as $\mathbf{x}_{n+1}$ have the same profile as $\mathbf{x}_{n+1}$ according to f, then the statistics used in `CD-split` , $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D})$ are i.i.d. data. Thus, the quantile used in `CD-split` will be the α quantile of $f(Y_{n+1}|\mathbf{x}_{n+1})$. This in turn makes $C(\mathbf{x}_{n+1})$ the smallest prediction band with conditional validity of $1-\alpha$. Theorem 3.8, below, formalizes this statement.

Theorem 3.8. *Assume that all samples that fall into the same partition as $\mathbf{x}_{n+1}$, say $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_m, Y_m)$, are such that $g_{\mathbf{x}_i} = g_{\mathbf{x}_{n+1}}$, and that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ is continuous as a function of y for every $\mathbf{x} \in \mathcal{X}$. Let $T_m := q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))$ be the cutoff used in `CD-split`. For every $\alpha \in (0,1)$*

$$T_m \xrightarrow[a.s.]{m \longrightarrow \infty} t^*$$

where $t^* = t^*(\mathbf{x}_{n+1}, \alpha)$ is the cutoff associated to the oracle band, the smallest predictive region with coverage $1-\alpha$.

Given the above reasons, the profile density captures what is needed of a meaningful neighborhood that contains many samples even in high dimensions. Indeed, consider a partition of the feature space, $\mathscr{A}$, that has the property that all samples that belong to the same element of $\mathscr{A}$ have the same oracle cutoff t^*. Theorem 3.9 shows that the coarsest partition that has this property is the one induced by the profile distance.

Theorem 3.9. *Assume that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ is continuous as a function of y for every $\mathbf{x} \in \mathcal{X}$. For each sample $\mathbf{x} \in \mathcal{X}$ and miscoverage level $\alpha \in (0,1)$, let $t^*(\mathbf{x}, \alpha)$ be the cutoff of the oracle band for $f(y|\mathbf{x})$ with coverage $1-\alpha$. Consider the equivalence relation $\mathbf{x}_a \sim \mathbf{x}_b \iff d_g(\mathbf{x}_a, \mathbf{x}_b) = 0$.*

(i) If $\mathbf{x}_a \sim \mathbf{x}_b$, then $t^*(\mathbf{x}_a, \alpha) = t^*(\mathbf{x}_b, \alpha)$ for every $\alpha \in (0,1)$

(ii) If $\sim'$ is any other equivalence relation such that $\mathbf{x}_a \sim' \mathbf{x}_b$ implies that $t^*(\mathbf{x}_a, \alpha) = t^*(\mathbf{x}_b, \alpha)$ for every $\alpha \in (0,1)$, then $\mathbf{x}_a \sim' \mathbf{x}_b \Rightarrow \mathbf{x}_a \sim \mathbf{x}_b$.

Based on the above motivation, we use `CD-split` with the profile distance. In order to compute the prediction bands, we need to define the centroids $\mathbf{x}_i^c$. Ideally, the partitions should be such that: (i) all sample points inside a given element of the partition have similar profile, and (ii) sample points that belong to different elements of the partition have profiles that are very different from each other. We accomplish this by choosing the partitions by applying a k-means++ clustering algorithm (Arthur and Vassilvitskii, 2007) using the profile distance. This is done by applying the standard (Euclidean) k-means++ algorithm to the data points $\mathbf{w}_i := \tilde{g}(\mathbf{x}_i)$, where $\tilde{g}(\mathbf{x}_i)$ is a discretization of the function $g(\mathbf{x}_i)$, obtained by evaluating $g(\mathbf{x}_i)$ on a grid of values. The points, $\mathbf{w}_1^c, \ldots, \mathbf{w}_J^c$, are the centroids of such clusters. Figure 3 illustrates the partitions that are obtained in one dataset. The profile distance allows samples that are far from each other in the Euclidean sense to fall into the same element of the partition. This is the key reason why our method scales to high-dimensional datasets. Algorithm 2 shows pseudo-code for implementing `CD-split`.

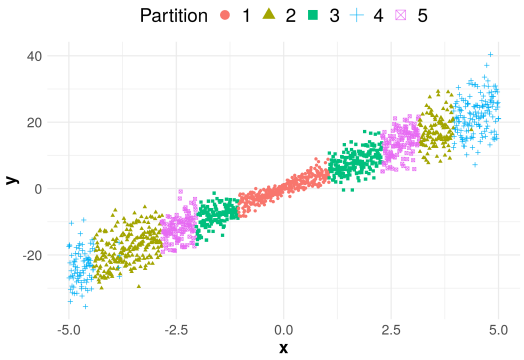

Figure 3: Scatter plot of data generated according to $Y|x \sim N(5x, 1+|x|)$. Colors indicate partitions that were obtained using the profile of the estimated densities. Note that points that are far from each other on the x-axis can have similar densities and belong to the same element of the partition. This allows larger partition elements while preserving the optimal cutoff (Theorem 3.9).

Algorithm 2 `CD-split`

Input: Data $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$, miscoverage level $\alpha \in (0,1)$, algorithm $\mathscr{B}$ for fitting conditional density function, number of elements of the partition J.

Output: Prediction band for $\mathbf{x}_{n+1} \in \mathbb{R}^d$

1: Randomly split $\{1, 2, \ldots, n\}$ into two subsets $\mathbb{D}$ and $\mathbb{D}'$
2: Fit $\hat{f} = \mathscr{B}(\{(\mathbf{x}_i, Y_i) : i \in \mathbb{D}'\})$ **// Estimate cumulative density function**
3: Compute $\mathscr{A}$, the partition of $\mathcal{X}$, by applying k-means++ on the profiles of the samples in $\mathbb{D}$'
4: Compute $g_{\mathbf{x}_{n+1}}(t) = \int_{\{y: \hat{f}(y|\mathbf{x}) \geq t\}} \hat{f}(y|\mathbf{x}) dy$, for all $t > 0$ **// Profile of the density (Definition 3.4)**
5: Find $A(\mathbf{x}_{n+1}) \in \mathscr{A}$, the element of $\mathscr{A}$ such that $\mathbf{x}_{n+1} \in \mathscr{A}$
6: Compute $g_{\mathbf{x}_i}(t) = \int_{\{y: \hat{f}(y|\mathbf{x}) \geq t\}} \hat{f}(y|\mathbf{x}) dy$, for all $t > 0$ and $i \in \mathbb{D}$ **// Profile of the densities (Definition 3.4)**
7: Let $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \{\hat{f}(y_i|\mathbf{x}_i), i \in \mathbb{D} : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\}$
8: Let $t = q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))$ **// Compute the α- quantile of the set $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D})$**
9: **return** $\{y : \hat{f}(y|\mathbf{x}^*) \geq t\}$

3.1 Multiclass classification

If the sample space $\mathscr{Y}$ is discrete, we use a similar construction to that of Definition 3.1. More precisely, the

`CD-split` prediction band is given by

$$C(\mathbf{x}_{n+1}) = \left\{y : \widehat{\mathbb{P}}(Y = y|\mathbf{x}_{n+1}) \geq q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))\right\}, \text{ where}$$
$$\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \left\{\widehat{\mathbb{P}}(Y_i = y_i|\mathbf{x}_i), i = 1, \ldots, n : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\right\},$$

$A(\mathbf{x}_{n+1})$ is the element of $\mathscr{A}$ to which $\mathbf{x}_{n+1}$ belongs to, and

$$d_g^2(\mathbf{x}_a, \mathbf{x}_b) = \sum_{y \in \mathscr{Y}} \left(\widehat{\mathbb{P}}(Y = y|\mathbf{x}_a) - \widehat{\mathbb{P}}(Y = y|\mathbf{x}_b)\right)^2.$$

Theorems analogous to those presented in the last section hold in the classification setting as well.

Remark 2. *While `CD-split` controls the coverage of C conditional on the value* $\mathbf{x}_{n+1}$*, in a classification setting some methods control* class-specific coverage *(Sadinle et al., 2019), defined as*

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1})|Y_{n+1} = y) \geq 1 - \alpha_y.$$

4 Experiments

We consider the following settings with $d = 20$ covariates:

- **[Asymmetric]** $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} = 5x_1 + \epsilon$, where $\epsilon \sim \text{Gamma}(1 + 2|x_1|, 1 + 2|x_1|)$.
- **[Bimodal]** $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-1.5, 1.5)$, and $Y|\mathbf{x} \sim 0.5\text{N}(f(\mathbf{x}) - g(\mathbf{x}), \sigma^2(\mathbf{x})) + 0.5\text{N}(f(\mathbf{x}) + g(\mathbf{x}), \sigma^2(\mathbf{x}))$, with $f(\mathbf{x}) = (x_1 - 1)^2(x_1 + 1)$, $g(\mathbf{x}) = 2\mathbb{I}(x_1 \geq -0.5)\sqrt{x_1 + 0.5}$, and $\sigma^2(\mathbf{x}) = 1/4 + |x_1|$. This is the example from Lei and Wasserman (2014) with $d - 1$ additional irrelevant variables.
- **[Heteroscedastic]** $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} \sim \text{N}(x_1, 1 + |x_1|)$.
- **[Homoscedastic]** $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} \sim \text{N}(x_1, 1)$

We compare the performance of the following methods:

- **[Reg-split]** The regression-split method (Lei et al., 2018), based on the conformal score $|Y_i - \widehat{r}(\mathbf{x}_i)|$, where $\widehat{r}$ is an estimate of the regression function.
- **[Local Reg-split]** The local regression-split method (Lei et al., 2018), based on the conformal score $\frac{|Y_i - \widehat{r}(\mathbf{x}_i)|}{\widehat{\rho}(\mathbf{x}_i)}$, where $\widehat{\rho}(\mathbf{x}_i)$ is an estimate of the conditional mean absolute deviation of $|Y_i - r(\mathbf{x}_i)|\mathbf{x}_i$.
- **[Quantile-split]** The conformal quantile regression method (Romano et al., 2019; Sesia and Candès, 2019), based on conformalized quantile regression.
- **[Dist-split]** From section 2.
- **[CD-split]** From section 3 with partitions of size $\lceil \frac{n}{100} \rceil$.

Each experiment is performed with comparable settings. Each experiment uses a coverage level of $1 - \alpha = 90\%$ and is run 5,000 times. Also, random forests (Breiman, 2001) are used to estimate all quantities needed in each method, namely: the regression function in Reg-split, the conditional mean absolute deviation in Local Reg-split, the conditional quantiles via quantile forests (Meinshausen, 2006) in Quantile-split, and the conditional density via FlexCode (Izbicki and Lee, 2017) in `Dist-split` and `CD-split`. A conditional cumulative distribution estimate, $\widehat{F}(y|\mathbf{x})$ is obtained by integrating the conditional density estimate: $\widehat{F}(y|\mathbf{x}) = \int_{-\infty}^{y} \widehat{f}(y|\mathbf{x})dy$. The tuning parameters of all methods were set to be the default values of the packages that were used.

Figure 4 shows the performance of each method as a function of the sample size. While the left side figures display how well each method controls conditional coverage, the right side displays the average size of the regions that are obtained. The control of the conditional coverage is measured through the conditional coverage absolute deviation, that is, $\mathbb{E}[|\mathbb{P}(Y^* \in C(\mathbf{X}^*)|\mathbf{X}^*) - (1 - \alpha)|]$. Since all of the methods obtain marginal coverage very close to the nominal 90% level, this information is not displayed in the figure. Figure 4 shows that, in all settings, `CD-split` is the method which best controls conditional coverage. Also, in most cases its prediction bands also have the smallest size. Similarly, `Dist-split` frequently is the second method with both highest control of conditional coverage and also smallest prediction bands.

We also apply `CD-split` to a classification setting. We consider $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} N(0, 1)$ and $Y|\mathbf{X}$ follows the logistic model, $\mathbb{P}(Y = i|\mathbf{x}) \propto \exp\{\boldsymbol{\beta} \cdot \mathbf{x}\}$, where $\boldsymbol{\beta} = (-6, -5, -1.5, 0, 1.5, 5, 6)$. We compare `CD-split` to Probability-split, the method described in Sadinle et al. (2019, Sec. 4.3), which has the goal of controlling global coverage. Probability-split is a particular case of `CD-split`: it corresponds to applying `CD-split` with $J = 1$ partitions. Figure 5 shows the results. `CD-split` better controls conditional coverage. On the other hand, its prediction bands are, on average, larger than those of Probability-split.

5 Final remarks

We introduce `Dist-split` and `CD-split`, which obtain asymptotic conditional coverage and converge to optimal oracle bands, even in high-dimensional feature spaces. These results do not require assumptions about the dependence between the target variable and the features. Both methods are based on estimating conditional densities. While `Dist-split` necessarily leads to intervals, which are easier to interpret, `CD-split` leads to smaller prediction regions. A simulation study shows that both methods yield smaller prediction bands and

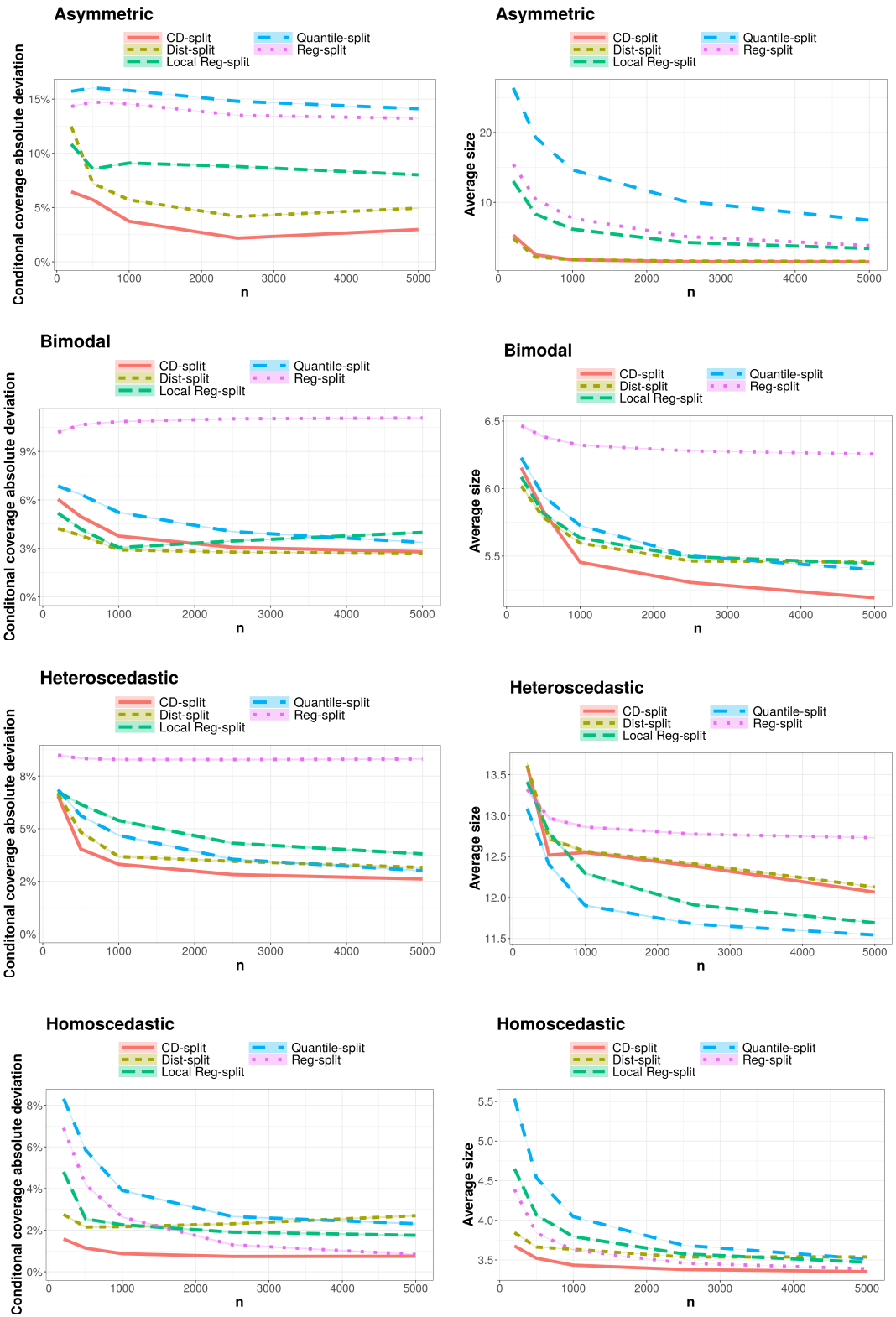

Figure 4: Performance of each conformal method as a function of the sample size. Left panels show how much the conditional coverage varies with $\mathbf{x}$; right panels display the average size of the prediction bands.

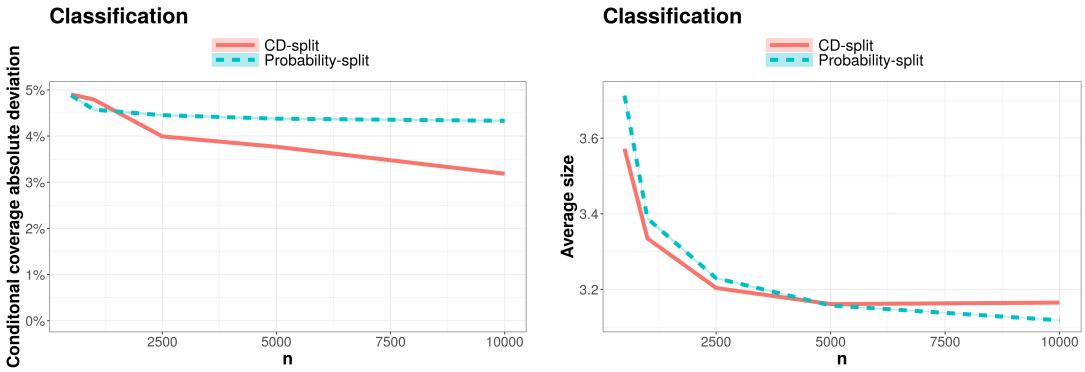

Figure 5: Performance of each conformal method as a function of the sample size. Left panel shows how much the conditional coverage vary with $\mathbf{x}$; right panel displays the average size of the prediction bands.

better control of conditional coverage than other methods in the literature under a variety of settings. We also show that `CD-split` leads to good results in classification problems.

`CD-split` is based on a novel data-driven metric on the feature space that is appropriate for defining neighborhoods for conformal methods, in particular in high-dimensional settings. It might be possible to use this metric with other conformal methods to obtain asymptotic conditional coverage.

R code for implementing `Dist-split` and `CD-split` is available at https://github.com/rizbicki/predictionBands.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Rafael Izbicki is grateful for the financial support of FAPESP (grants 2017/03363-8 and 2019/11321-9) and CNPq (grant 306943/2017-4). The authors are also grateful for the suggestions given by Luís Gustavo Esteves and Jing Lei.

References

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*.

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

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., and Malz, A. I. (2019). Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. *arXiv preprint arXiv:1908.11523*.

Freeman, P. E., Izbicki, R., and Lee, A. B. (2017). A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565.

Guan, L. (2019). Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*.

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831.

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

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

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299.

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

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. IntechOpen.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347.

Parmigiani, G. and Inoue, L. (2009). *Decision theory: Principles and approaches*, volume 812. John Wiley & Sons.

Pospisil, T. and Lee, A. B. (2019). (f) rfcde: Random forests for conditional density estimation and functional data. *arXiv preprint arXiv:1906.07177*.

Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression.

Sadinle, M., Lei, J., and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Sesia, M. and Candès, E. J. (2019). A comparison of some conformal quantile regression methods. *arXiv preprint arXiv:1909.05433*.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490.

Vovk, V. et al. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

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