International Journal of Approximate Reasoning 141 (2022) 131–145

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Homeostasis phenomenon in conformal prediction and predictive distribution functions ☆

Min-ge Xie *, Zheshi Zheng

Department of Statistics, Rutgers University, United States of America

ARTICLE INFO

Article history:
Received 13 March 2021
Received in revised form 31 August 2021
Accepted 1 September 2021
Available online 22 September 2021

Keywords:
Confidence
Predictive distribution
Robustness
Machine learning
Model mis-specification

ABSTRACT

Conformal prediction is an attractive framework for prediction that is distribution free. In this article, we study in details its *homeostasis property* under a general regression setup and also introduce the concepts of upper and lower predictive distributions and predictive curve to establish connections to left-, right- and two-tailed hypothesis testing problems as well as the developments in confidence distributions. The homeostasis property is very attractive, since it states that under some conditions the prediction results remain valid even if the model used for learning is completely wrong. We show explicitly why the property holds in a model-based setup and also explore the boundary when the property breaks down. Beside the typical assumption used in conformal prediction that the response and covariate pairs $(y, \mathbf{x})$ of all subjects are iid distributed, we also study the classical regression setting in which the design is fixed with given (non-random) covariates $\mathbf{x}$. The trade-offs among learning model accuracy, prediction valid and prediction efficiency are discussed, leading to an emphasis of more efforts on developing better learning models.

1. Introduction

Suppose we have $n+1$ subjects $s_1, \ldots, s_n$ and s_{new}. For the first n subjects, we have an observed data set $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. The new subject s_{new} has values $(\mathbf{x}_{new}, y_{new})$, where we are given only $\mathbf{x}_{new}$ and need to predict the unknown y_{new}. Denote by $\mathcal{X}$ the sample space of $\mathbf{x}$ and $\mathcal{Y}$ the sample space of y. A typical "exchangeable" condition in conformal prediction is that, if a randomly selected pair $(\mathbf{x}_i, y_i)$ from $\mathcal{D}$ is replaced by $(\mathbf{x}_{new}, y_{new})$, the joint distribution of $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} \setminus \{(\mathbf{x}_i, y_i)\}$ remains the same as the distribution of $\mathcal{D}$ [1,23]. For simplicity, we assume that the samples of the $n+1$ subjects are independently identically distributed (iid) random samples from an unknown population $\mathcal{F}$, i.e.,

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new}) \overset{iid}{\sim} \mathcal{F}, \tag{1}$$

which is the simplest and also most commonly used special case in conformal prediction. Later in Section 3 to further study the impact of a learning model in a model-based setup, we also relax the iid requirement in (1) to only assume that

☆ The research is supported in part by US NSF grants DMS-1737857, 1812048, 2015373 and 2027855. We wish to contribute this article to the special issue in honor of Professor Glenn Shafer for his pioneer contributions (together with Professor V. Vovk) on conformal prediction. The authors also wish to thank the editors and the two reviewers for their constructive suggestions and comments. Their suggestions have helped greatly improve the quality of the paper.

* Corresponding author.
E-mail address: mxie@stat.rutgers.edu (M. Xie).

https://doi.org/10.1016/j.ijar.2021.09.001
0888-613X/

$y_{new}|\mathbf{x}_{new}$ relates to $\mathbf{x}_{new}$ the same way as $y_i|\mathbf{x}_i$ relates to $\mathbf{x}_i$, but $\mathbf{x}_{new}$ may be fixed or follow a marginal distribution that is different from that of $\mathbf{x}_i$, i.e.,

$$y_{new}|\mathbf{x}_{new} \text{ relates to } \mathbf{x}_{new} \text{ the same way as } y_i|\mathbf{x}_i \text{ relates to } \mathbf{x}_i \text{ but } \mathbf{x}_{new} \nsim \mathbf{x}_i. \tag{2}$$

The second condition includes the typical assumption used in the classical regression analysis in statistics where the design covariates $\mathbf{x}$ are fixed (cf., e.g., [5,17]). Under condition (2), we will typically consider conditional predictive inference using conformal prediction procedure. Although the article focuses on studies in the two special cases as expressed in (1) and (2), the conclusions reached also have implications to more general exchangeable cases.

The idea behind the conformal prediction development has an intuitive interpretation. In order to make a prediction of the unknown y_{new}, we examine a potential value y^* of the true y_{new} and see how "conformal" the pair $(\mathbf{x}_{n+1}, y^*)$ is among the observed n pairs of iid data $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. The higher the "conformity", the more likely y_{new} takes the potential value y^*. Frequently, a learning model, say $y_i \sim \mu(\mathbf{x}_i)$ for $i = 1 \ldots, n$ and $i = new$, is used to assist prediction. However, it is well-known in the literature that the learning model is not essential; cf., [23]. As we will see later in this article, even if $\mu(\cdot)$ is totally wrong, a conformal prediction can still provide us valid prediction, as long as assumption (1) holds or when (2) holds but with additional strict conditions. This robustness against wrong learning model is referred to as the *homeostasis* property in this article, since it has a "self-rebalance" phenomenon to correct the predictive bias caused by the wrong learning model. An explicit formula of this self-rebalancing correction in regression setting is provided in Section 3.

The homeostasis property is attractive, since it provides an assurance of the outcome even if the learning model used is wrong. It may reduce the burden of model building, a "more difficult task" than prediction [9]. However, there is a trade-off between the use of a wrong learning model and the prediction efficiency, even when the validity is preserved. If the learning model is poorly fit, the predictive result, even if valid, comes with a large uncertainty. Sometime, the large uncertainty can render the prediction result useless in practice. Furthermore, the conditions such as that in (1) play a key role in preserving the homeostasis property. Under a general regression setting, including the case when the covariate variables $\mathbf{x}$ are fixed, we explore the boundary beyond which the homeostasis property breaks down.

Most publications in conformal prediction so far report prediction intervals (or sets) with a pre-specified confidence level. To provide a fuller picture of the predictive inference, we elevate the interval estimation to a predictive function on the space of y_{new}. Vovk et al. [24] introduced the concept of *conformal predictive distribution* function with the frequentist (non-Bayesian) interpretation. In this article, to deal with the discrete nature of conformal prediction, we define three further predictive inference functions, leading to a set of finer concepts of upper- and lower-conformal predictive distributions and predictive curve. We investigate the connections of these functions to hypothesis testing problems and also to the developments in confidence distributions.

The remaining of the article is arranged as follows. Section 2 reviews a general conformal prediction procedure, defines upper- and lower-conformal predictive distributions and predictive curve, and establishes connections to confidence distributions and hypothesis testing problems. Section 3 investigates the model-based conformal prediction and provides a detailed study of the homeostasis phenomenon. Section 4 contains a numerical study example, and Section 5 provides further remarks and discussions. Throughout the paper two commonly-used conformal prediction procedures, known as split-conformal and jackknife-plus approaches, are used as examples to illustrate the investigation.

2. Conformal prediction, hypothesis testing and confidence distribution

2.1. Conformal prediction and three relevant predictive functions on $\mathcal{Y}$

Given $\mathbf{x}_{new}$ and for a potential value of the unknown y_{new}, say y^*, we would like to know how "conformal" the pair $(\mathbf{x}_{new}, y^*)$ is among the observed n pairs of iid data $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. To quantitatively measure the "conformity", we make the following (minimal) assumption:

(A1) For each $i = 1, \ldots, m$, $m \le n$, we can compute a statistic $R_i = R_i(\mathcal{D})$ (referred to as a *conformity score* of object s_i) based on the observed data $\mathcal{D}$. If $(\mathbf{x}_i, y_i)$ is replaced by $(\mathbf{x}_{new}, y^*)$, we can use the same algorithm to calculate the corresponding statistic $R_i^* = R_i(\mathcal{D}_i^*)$ based on the data set $\mathcal{D}_i^* = \mathcal{D} \cup \{(\mathbf{x}_{new}, y^*)\} \setminus \{(\mathbf{x}_i, y_i)\}$.

Different approaches have different ways to compute the statistics R_i. Two concrete examples are provided at the end of this subsection.

Under (1), the pair $(\mathbf{x}_{new}, y_{new})$ has the same distribution as any pair $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, m$. Here, $m \le n$ is any given integer number less than or equal to n. Thus, R_i and R_i^* have the same marginal distribution, if $y^* = y_{new}$. We expect that the pair of values R_i and R_i^* are similar, if y^* is close to y_{new}. This consideration leads us to define the following function from $\mathcal{Y} \mapsto [0, 1]$,

$$Q(y^*) = \frac{1}{m}\sum^{m} \mathbf{1}(R_i^* > R_i) + \frac{1}{2m}\sum^{m} \mathbf{1}(R_i^* = R_i), \text{ for } y^* \in \mathcal{Y},$$

$t=1$ $t=1$

where the second summation term is for potential tie cases. The function $Q(y^*)$ provides an assessment for the degree of "conformality" for the potential value y^*: too large or too small a value of $Q(y^*)$ (i.e., $Q(y^*) \approx 0$ or ≈ 1) indicates that y^* is less likely "conformal" with the observed values in $\mathcal{D}$. Often, $Q(y^*)$ is a monotonically increasing function in y^*, as seen the examples in Section 3. In this case, it can be shown that the $Q(\cdot)$ function is a frequentist asymptotic predictive function in the sense of Shen et al. [20], when $m \to \infty$.

However, for finite sample data and to handle the discreteness in conformal prediction, we consider the following two functions from $\mathcal{Y} \mapsto [0, 1]$,

$$Q^-(y^*) = \frac{\sum_{i=1}^m \mathbf{1}(R_i^* \geq R_i) + 1}{m+1} \quad \text{and} \quad Q^+(y^*) = \frac{\sum_{i=1}^m \mathbf{1}(R_i^* > R_i)}{m+1}. \tag{3}$$

Note that $Q^-(y^*) \geq Q(y^*) \geq Q^+(y^*)$, with the point-wide maximum difference between any pair of functions at the order of $O(\frac{1}{m+1})$. When there are no ties, the maximum difference is bounded by $\frac{1}{m+1}$. The two functions in (3) provide a pointwise envelope box for $Q(y^*)$, similar to the so-called *probability-box* (p-box) development in uncertainty quantification and imprecise probability (cf., e.g., [12]). The box formed by $Q^-(y^*)$ and $Q^+(y^*)$ can help handle inference uncertainty and the discreteness in a conformal prediction procedure. Here, we shift a step of size $\frac{1}{m+1}$, instead of $\frac{1}{m}$, to reflect the nature of discreteness in conformal prediction; cf., [24]. The development of the two inference functions is also similar to that of the so-called lower and upper confidence distributions in statistics for making exact inference of parameters (cf., e.g., [15,7]).

We further define the function

$$PV(y^*) = 2\min\{Q^-(y^*), 1 - Q^+(y^*)\}, \tag{4}$$

for $y^* \in \mathcal{Y}$. Based on the function $PV(y^*)$, we can define an interval (or set) on $\mathcal{Y}$:

$$C_\alpha = \{y^* : PV(y^*) \geq \alpha\}, \tag{5}$$

for a given $\alpha \in (0, 1)$. Under an appropriate choice of algorithm $R_i(\cdot)$, the set C_α is typically a frequentist prediction interval of y_{new} with a confidence level that is equal (or related) to $1 - \alpha$. In fact, all the three functions $Q^-(\cdot)$, $Q^+(\cdot)$ and $PV(\cdot)$ have meaningful interpretations in predictive inference. Further details are provided in Section 2.2.

Throughout the article we elaborate our discussions using two versions of $R_i(\cdot)$, both of which are commonly used conformal prediction procedures in the literature.

Split conformal prediction (cf., [14]). In a split conformal prediction procedure, the observed data set $\mathcal{D}$ is randomly split into two subsets $\mathcal{D}_1 = \{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{I}_1}$ and $\mathcal{D}_2 = \{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{I}_2}$ where $\mathcal{I}_1 \cup \mathcal{I}_2 = \{1, 2 \cdots, n\}$ and $|\mathcal{I}_1| = m < n$. Without loss of generality, let us assume that $\mathcal{I}_1 = \{1, \ldots, m\}$. Then, we train a model $\widehat{\mu}(\cdot \mid \mathcal{D}_2)$ based on the subset $\mathcal{D}_2$. For $i = 1, \ldots, m$, we calculate

$$R_i = y_i - \widehat{\mu}(\mathbf{x}_i \mid \mathcal{D}_2) \quad \textit{and} \quad R_i^* = y^* - \widehat{\mu}(\mathbf{x}_{new} \mid \mathcal{D}_2). \tag{6}$$

Here, R_i^* is the same for all i. A prediction interval by (5) typically has a level-$(1 - \alpha)$ coverage.

The split conformal prediction is also known as *inductive conformal prediction* in the machine learning literature; cf., e.g., [22].

Jackknife plus conformal prediction (cf., [3]). In the jackknife plus conformal prediction procedure, we do not split the observed dataset $\mathcal{D}$ but utilize the idea of the so-called deleted residuals. Specifically, let $\widehat{\mu}^{(-i)}(\cdot \mid \mathcal{D}_i)$ be the fitted model based on the data set $\mathcal{D}_i = \mathcal{D} \setminus \{(\mathbf{x}_i, y_i)\}$, for $i = 1, \ldots, m$. Here, $m = n$. Then, we calculate

$$R_i = y_i - \widehat{\mu}^{(-i)}(\mathbf{x}_i \mid \mathcal{D}_i) \quad \textit{and} \quad R_i^* = y^* - \widehat{\mu}^{(-i)}(\mathbf{x}_{new} \mid \mathcal{D}_i), \tag{7}$$

for $i = 1, \ldots, m$. A prediction interval by (5) typically has a level-$(1 - 2\alpha)$ coverage.

2.2. Conformal predictive distributions and predictive curve

The three functions defined in (3) and (4) are associated with the notion of *frequentist predictive distributions*. A predictive distribution in Bayesian inference is well known, but the development of a predictive distribution with confidence interpretation is relatively new; cf., [13,20]. Vovk et al. [24] used conformal prediction to derive predictive distributions and defined the so called *conformal predictive distribution* using randomization. The randomization allows to have an exact confidence statement, i.e., coverage interpretation at the exact $1 - \alpha$ level. Although mathematically clean, the randomization however introduces an additional artificial uncertainty (randomness) into inference statements. Following the development of imprecise probability (cf., e.g., [19,12,16]) and recent work on upper and lower confidence distributions (cf., e.g. [25,7,15]), we define the concepts of upper- and lower- predictive distributions and the associated predictive curve function. Together, the three newly defined predictive distributions can handle the discrete nature of the conformal prediction regardless of the size and the type of data. They provide full information for predictive inference.

A prediction interval obtained by a conformal prediction procedure has the same frequency interpretation as a confidence interval, except that it is developed for a random y_{new} instead of a parameter of interest. That is, if we repeatedly use

a level-$(1-\alpha)$ conformal prediction interval 100 times, the intervals are expected to cover y_{new} $(1-\alpha)100$ times or more. Similarly, a frequentist predictive distribution (with a confidence interpretation) can be viewed as an extension of a confidence distribution but developed for the random y_{new} instead for a parameter of interest. Cox [6] suggested that a confidence distribution be introduced "in terms of the set of confidence intervals of all levels". To better understand the concept of predictive distributions and predictive curve, especially how to relate them to prediction intervals of all levels and hypothesis testing problems, it is prudent to briefly take a look at confidence distribution and confidence curve, and then move on to prediction. We consider a toy example below.

Example 1. We assume in this toy example that $y_1, \ldots, y_n \overset{iid}{\sim} N(\theta, 1)$. Instead of using a point $(\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i)$ or a level-$(1-\alpha)$ interval $(\bar{y} \pm \frac{1}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2}))$ to estimate the unknown parameter θ, a confidence distribution suggests to use a sample-dependent function $N(\bar{y}, \frac{1}{n})$, or more formally in the cumulative distribution function form $H_n(\theta) = \Phi(\sqrt{n}(\theta - \bar{y}))$, to estimate the unknown parameter θ; cf., e.g., [25,18]. A nice feature of a confidence distribution is that it can represent confidence intervals of all levels. For example, the level-$(1-\alpha)$ one-tailed interval $(-\infty, \bar{y} + \frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha)) = (-\infty, H_n^{-1}(1-\alpha))$ and the level-$(1-\alpha)$ two-tailed interval $(\bar{y} + \frac{1}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}), \bar{y} + \frac{1}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})) = (H_n^{-1}(\frac{\alpha}{2}), H_n^{-1}(1-\frac{\alpha}{2}))$. Here, $H_n^{-1}(\cdot)$ is the inverse function of $H_n(\cdot)$.

A closely related concept is confidence curve

$$CV_n(\theta) = 2\min\{H_n(\theta), 1 - H_n(\theta)\},$$

which was first introduced by Birnbaum [4] as an "omnibus form of estimation" that "incorporates confidence limits and intervals at all levels." For any $\alpha \in (0,1)$, $\{\theta : CV_n(\theta) \geq \alpha\}$ is a level-$(1-\alpha)$ two-tailed confidence interval. We could view the function $CV_n(\theta)$ as a result of stacking up two-tailed confidence intervals of all levels $1-\alpha$ for α going from 0 to 1; cf., Fig. 1 (a). The plot of confidence curve function $CV_n(\theta) = 2\min\{\Phi(\sqrt{n}(\theta - \bar{y})), 1 - \Phi(\sqrt{n}(\theta - \bar{y}))\}$ provides a full picture of confidence intervals of all levels $1-\alpha \in (0,1)$, with a peak point corresponding to a median unbiased estimator $\hat{\theta}_M = \bar{y}$ with $\mathbb{P}(\hat{\theta}_M \leq \theta) \geq \frac{1}{2}$ and $\mathbb{P}(\hat{\theta}_M \geq \theta) \geq \frac{1}{2}$. Note that, here, the probability $\mathbb{P}$ and the coverage statements on θ are with regard to the joint distribution of $(y_1, \ldots, y_n)$.

For an unobserved new sample $y_{new} \sim N(\theta, 1)$, the task of prediction focuses on y_{new} instead of $\theta = E(y_{new})$. A predictive distribution is $N(\bar{y}, 1 + \frac{1}{n})$, or in its cumulative distribution function form $Q_n(y) = \Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big)$. Parallel to confidence curve, we can define a predictive curve

$$PV_n(y) = 2\min\{Q_n(y), 1 - Q_n(y)\} = 2\min\left\{\Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big), 1 - \Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big)\right\}. \tag{8}$$

Fig. 1 (b) is a plot of the predictive curve in (8). Again, we can view the function $PV_n(y)$ as a result of stacking up two-tailed prediction intervals of all levels $1-\alpha$ for α going from 0 to 1. The plot of the predictive curve $PV_n(y) = 2\min\big\{\Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big), 1 - \Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big)\big\}$ provides a full picture of prediction intervals of all levels $1-\alpha \in (0,1)$. The peak point in Fig. 1(b) corresponds to a median unbiased point predictor $\hat{y}_M = \bar{y}$ with $\mathbb{P}(\hat{y}_M \leq y_{new}) \geq \frac{1}{2}$ and $\mathbb{P}(\hat{y}_M \geq y_{new}) \geq \frac{1}{2}$. Here, the probability $\mathbb{P}$ and the coverage statements on y_{new} are with regard to the joint distribution of $(y_1, \ldots, y_n, y_{new})$, including y_{new}.

Note that, in Example 1 and for a fixed $t \in \Theta = (-\infty, \infty)$, $H_n(t)$ is the p-value for the one-tailed test $H_o : \theta \leq t$ versus $H_a : \theta > t$ and $CV_n(t)$ is the p-value for the two-tailed test $H_o : \theta = t$ versus $H_a : \theta \neq t$; cf., e.g., [21,25,18]. Thus, $H_n(\theta)$ and $CV_n(\theta)$ can be interpreted as the same quantities of p-value functions of one-tailed and two-tailed tests, respectively. Similarly, the predictive function $Q_n(y)$ and predictive curve $PV_n(y)$ also have the corresponding interpretation of p-value functions of right-tailed test $H_o : y_{new} \leq y$ versus $H_a : y_{new} > y$ and two-tailed test $H_o : y_{new} = y$ versus $H_a : y_{new} \neq y$, respectively.

The example illustrates how these relevant concepts provide inferential information in both estimation and prediction under a continuous and parametric distribution model. In the case of conformal prediction (often with unknown underlying distribution $\mathcal{F}$) and with a finite sample size, we need to deal with discrete functions. To handle the discreteness and following the developments in uncertainty quantification with imprecise probability, we define upper- and lower-predictive distribution functions and, additionally, predictive curve function. These definitions are analogs of the upper- and lower-confidence distributions and confidence curve of parameter estimation, respectively (cf., e.g., [15,7]).

Definition 1. A function $Q^+(\cdot) = Q^+(\mathcal{D}, \mathbf{x}_{new}, \cdot)$ on $(\mathcal{X} \times \mathcal{Y})^{n+1} \to [0,1]$ is said to be an ***upper-predictive distribution function*** for y_{new}, if
(i) For observed $\mathcal{D}$ and given $\mathbf{x}_{new}$, $Q^+(\cdot)$ is a monotonic increasing function on $\mathcal{Y}$ with values ranging within $(0,1)$;
(ii) As a function of the random sample $\mathcal{D}$ and random $(\mathbf{x}_{new}, y_{new})$, $Q^+(y_{new})$, is stochastically less than or equal to a uniformly distributed random variable $U \sim U(0,1)$, i.e.,

$$Pr\{Q^+(y_{new}) < t\} \geq t, \text{ for all } t \in (0,1). \tag{9}$$

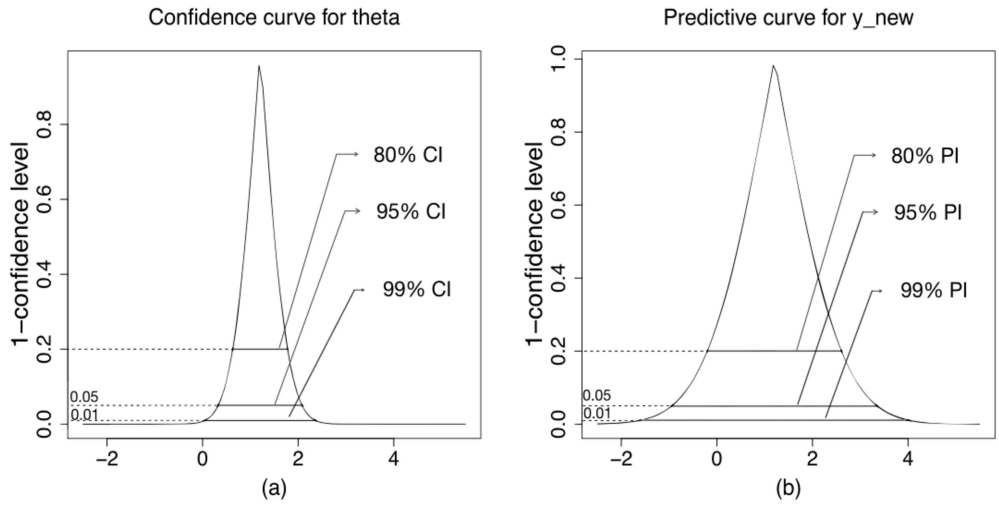

Fig. 1. Plot of (a) confidence curve function $CV_n(\theta) = 2\min\{\Phi(\frac{\theta-\bar{y}}{\sqrt{n}}), 1-\Phi(\frac{\theta-\bar{y}}{\sqrt{n}})\}$; (b) predictive curve function $PV_n(y) = 2\min\{\Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big), 1-\Phi\big(\frac{y-\bar{y}}{\sqrt{1+1/n}}\big)\}$. The plots provide a full picture of (a) confidence intervals and (b) prediction intervals of all levels. In particular, the curves can be formed as stacking up the endpoints of (a) the confidence intervals or (b) the prediction intervals at all levels of $1-\alpha$ for α from 0 to 1. The peak point corresponds to the median unbiased (a) point estimator $\hat{\theta}_M$ of θ and (b) point prediction $\hat{y}_M$ of y_{new}, respectively. The sample data used to generate the plots are from $N(1.35, 1)$ with $n = 5$.

where the probability Pr calculation is for random quantities $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} \in (\mathcal{X} \times \mathcal{Y})^{n+1}$.

A ***lower-predictive distribution function*** $Q^-(\cdot) = Q^-(\mathcal{D}, \mathbf{x}_{new}, \cdot)$ for y_{new} can be defined similarly, but with (9) replaced by $Pr\left\{Q^-(y_{new}) < t\right\} \le t$ for all $t \in (0,1)$.

A ***predictive curve*** for y_{new} is defined by the upper and lower predictive functions as $PV(y) = 2\min\{Q^-(y), 1-Q^+(y)\}$ for any $y \in \mathcal{Y}$.

Remark. In condition (i) above, we do not require $Q^+(\cdot)$ or $Q^-(\cdot)$ to be a surjective function onto $[0,1]$. So, strictly speaking, $Q^+(\cdot)$ and $Q^-(\cdot)$ may not need to be a cumulative distribution function on $\mathcal{Y}$. However, we still refer to $Q^+(\cdot)$ and $Q^-(\cdot)$ as a upper- and lower-predictive distribution, respectively. The reason is that, because of the stochastic dominance inequalities in the definition of upper- and lower-predictive distributions, we have

$$Pr\left(y_{new} \in \{y : 1 - Q^+(y) \ge \alpha\}\right) \ge 1-\alpha \quad \text{and} \quad Pr\left(y_{new} \in \{y : Q^-(y) \ge \alpha\}\right) \ge 1-\alpha,$$

for any $\alpha \in (0,1)$. Thus, a level-$(1-\alpha)$ upper prediction interval $\{y : Q^+(y) \le 1-\alpha\}$, or lower prediction interval $\{y : Q^-(y) \ge \alpha\}$, or two-sided prediction interval $\{y : PV(y) \ge \alpha\}$ has guaranteed coverage rate of at least $(1-\alpha)100\%$.

We have the following two propositions for the split and jackknife-plus conformal prediction procedures. Their proofs are provided in Appendix. Based on the propositions, the three functions defined in (3) and (4) are lower-, upper-predictive distributions and predictive curve, respectively. These functions can be used to construct one-sided and two-sided prediction intervals (sets) at any levels.

Proposition 1 *(Split conformal prediction procedure). Assume condition (1) holds. In a split conformal prediction procedure defined in (6), if the conformity score R_i^* is increasing in y^*, then $Q^+(y)$, $Q^-(y)$ are increasing in y, and we have*

$$Pr\left(y_{new} \in \{y : Q^+(y) \le \alpha\}\right) \ge \alpha \quad \textit{and} \quad Pr\left(y_{new} \in \{y : Q^-(y) < \alpha\}\right) \le \alpha.$$

In addition,

$$Pr\left(y_{new} \in \{y : PV(y) \ge \alpha\}\right) \ge 1-\alpha.$$

Proposition 2 *(Jackknife-plus conformal prediction procedure). Assume condition (1) holds. In a jackknife plus conformal prediction procedure defined in (7), if the conformity score R_i^* is increasing in y^*, then $Q^+(y)$, $Q^-(y)$ are increasing in y, and we have*

$$Pr\left(y_{new} \in \{y : Q^+(y) \le \alpha/2\}\right) \ge \alpha \quad \textit{and} \quad Pr\left(y_{new} \in \{y : Q^-(y) < \alpha/2\}\right) \le \alpha.$$

In addition,

$$Pr\left(y_{new} \in \{y : PV(y) \ge \alpha/2\}\right) \ge 1-\alpha.$$

The upper- and lower-predictive distributions as well as the predictive curve are associated with hypothesis testing problems. Note that the so-called *p-value functions* are a special type of confidence distributions and confidence distributions can also in turn be utilized to obtain p-values for testing hypotheses [25]. In Example 1, for instance, the confidence distribution $H(t)$ and the confidence curve $CV(t)$ are the p-value functions of right-tailed and two-tailed tests, respectively. More generally, it can be shown that the lower-CD function $H^-(\theta)$ and one minus the upper-CD function $1-H^+(\theta)$ can be treated as a p-value functions of the right-tailed test $H_o:\theta\le t$ versus $H_a:\theta>t$ and the left-tailed test $H_o:\theta\ge t$ versus $H_a:\theta<t$, respectively; the associated confidence curve can be treated as a p-value function of the two-tailed test $H_o:\theta=t$ versus $H_a:\theta\neq t$ (cf., [7,15]). For prediction, the testing problem concerns about the unknown random quantity y_{new} instead of an unknown parameter θ. Similar to the confidence distribution developments, the lower predictive distribution function $Q^-(y)$, one minus the upper predictive distribution function $1-Q^+(y)$ and the predictive curve $PV(y)$ also have the corresponding interpretation of p-value functions of the right-tailed test

$$H_o: y_{new}\le y \quad \text{versus} \quad H_a: y_{new}>y,$$

the left-tailed test

$$H_o: y_{new}\ge y \quad \text{versus} \quad H_a: y_{new}<y$$

and two-tailed test

$$H_o: y_{new}= y \quad \text{versus} \quad H_a: y_{new}\neq y,$$

respectively. By Proposition 1, the tests derived based on $Q^-(y)$, $1-Q^+(y)$ and $PV(y)$ are guaranteed to control the Type-I error less than α, if the split conformal procedure is used. Similarly, by Proposition 2, the tests derived based on $Q^-(y)$, $Q^+(y)$ and $PV(y)$ are guaranteed to have the Type-I error less than 2α, if the jackknife plus conformal procedure is used.

3. Model-based prediction and homeostasis phenomenon

In order to have a in-depth understanding of the homeostasis phenomenon and its boundary, we consider in this section a model-based setup. Specifically, we assume that the observed data $\mathcal{D}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$ and $(\mathbf{x}_{new},y_{new})$ are from the following model:

$$y_i=\mu_0(\mathbf{x}_i)+\epsilon_i, \quad \text{for } i=1,\ldots,n \text{ and } new, \tag{10}$$

where $\mu_0(\cdot)$ is the unknown true model. In addition, we assume the error terms $\epsilon_i=y_i-\mu_0(\mathbf{x}_i)$ are independent draws from an unknown distribution with mean 0. Here, ϵ_i may or may not depend on $\mathbf{x}_i$. For instance, in a regular Gaussian regression model, $\epsilon_i\sim N(0,\sigma^2)$ are iid copies that are free of $\mathbf{x}_i$; however, in a Poisson regression model, the ϵ_i terms depend on $\mathbf{x}_i$.

Since $\mu_0(\cdot)$ is unknown, a learning model is often used, say,

$$y=\mu_1(\mathbf{x})+e. \tag{11}$$

Note that the true model can be re-written as $y=\mu_0(\mathbf{x})+\epsilon=\mu_1(\mathbf{x})+\{\mu_0(\mathbf{x})-\mu_1(\mathbf{x})\}+\epsilon$. It follows that $e=\{\mu_0(\mathbf{x})-\mu_1(\mathbf{x})\}+\epsilon$.

Often, the error term e under the training model has a larger variance than that of the error term ϵ under the true model, i.e., $\text{var}(e)\ge\text{var}(\epsilon)$. For example, when ϵ is independent of $\mathbf{x}$, $\text{var}(e)=\text{var}(\{\mu_0(\mathbf{x})-\mu_1(\mathbf{x})\})+\text{var}(\epsilon)\ge\text{var}(\epsilon)$ and the larger $\text{var}(\{\mu_0(\mathbf{x})-\mu_1(\mathbf{x})\})$ is the larger $\text{var}(e)$ is. A larger error variance typically translates to less accurate inference in estimation and prediction.

In conformal prediction, the discrepant $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ appears only affect efficiency but not validity of the prediction under the iid setup. We have an intuitive explanation why the prediction is still valid even when a totally wrong learning model is used. In particular, when we use a wrong model $\mu_1(\mathbf{x})$, the corresponding point predictor will be biased by the magnitude of $\mu_1(\mathbf{x}_{new})-\mu_0(\mathbf{x}_{new})$, but at the same time the error term e absorbs the *bias*, thus producing residuals with a *shift* by the magnitude of $\mu_0(\mathbf{x}_i)-\mu_1(\mathbf{x}_i)=-\{\mu_1(\mathbf{x}_i)-\mu_0(\mathbf{x}_i)\}$. In a conformal prediction algorithm, the residuals are typically added back to the point prediction to form an overall predictive inference. If $\mathbf{x}_i$ and $\mathbf{x}_{new}$ are iid, then the bias is offset by the shift. Along with the greater residual variance associated with $\text{var}(e)$, the offsetting helps ensure the validity of the conformal prediction. The self-balancing offset to maintain validity is the key of the homeostasis phenomenon.

However in the non-iid case, the self-balancing offset is no longer valid, as the bias $\mu_1(\mathbf{x}_{new})-\mu_0(\mathbf{x}_{new})$ can be quite different from the negative shift $\mu_1(\mathbf{x}_i)-\mu_0(\mathbf{x}_i)$ for some $\mathbf{x}_{new}$. In this case, the conformal prediction algorithm does not produce a valid predictive inference in general. However, it can still be valid in a special case when the bias and shift are relatively small compared to the error term ϵ. For instance, when we use correct (or asymptotically correct) model, we can correctly estimate the distribution of error term. If the error distribution is independent of $\mathbf{x}$, the estimated error terms are conformal and thus we can still get valid predictive interval for a new subject. More details are provided later in the section.

We can express the above discussion in explicit and rigorous mathematical forms under a linear regression model setup: Suppose $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, and $(\mathbf{x}_{new}, y_{new})$ are from

$$y_i = \mu_0(\mathbf{x}_i) + \epsilon_i = \mathbf{x}_i^T \beta + \epsilon_i, \tag{12}$$

where β is the unknown regression coefficient and ϵ_i are iid random errors with mean 0. We would like to compare the performances of conformal prediction procedure under the true model (12) versus under the wrong learning model

$$y_i = \mu_1(\mathbf{x}_i) + e_i = \mathbf{z}_i^T \gamma + e_i, \tag{13}$$

where $\mathbf{z}_i$ is the first q elements of the p covariates of $\mathbf{x}_i$, $q < p$, and γ is the corresponding $q \times 1$ unknown regression coefficient. In the following, we consider two cases: 1) under the typical assumption used in conformal prediction that the response and covariates pairs $(y, \mathbf{x})$ of all subjects are iid distributed; 2) under the classical regression setup of fixed design with given (non-random) covariates $\mathbf{x}$ or the classical regression setup of stochastic $\mathbf{x}$'s. In case 2), we also consider two scenarios that (i) ϵ_i is free of the covariate $\mathbf{x}_i$ and (ii) ϵ_i depends on $\mathbf{x}_i$.

3.1. Split conformal procedure

In a split conformal procedure, the dataset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ is randomly split into two subsets $\mathcal{D}_1$ of size m and $\mathcal{D}_2$ of size $n - m$. In each subset $\mathcal{D}_k$, $k = 1, 2$, we define $\mathbf{Y}_{(k)}$ the response vector. We further define $\mathbf{X}_{(k)}$ the design matrix of p columns and $\mathbf{Z}_{(k)}$ the design matrix of q columns, corresponding to model (12) and (13), respectively. We also have a matrix partition $\mathbf{X}_{(k)} = (\mathbf{Z}_{(k)}, \mathbf{W}_{(k)})$, where $\mathbf{W}_{(k)}$ is a matrix of $(p - q)$ columns.

Under the true learning model (12) and from the least squares estimation, we have

$$R_i = y_i - \widehat{\mu}_0(\mathbf{x}_i; \mathcal{D}_2) = y_i - \mathbf{x}_i^T \hat{\beta}_{(2)} \quad \text{and} \quad R_i^* = y^* - \widehat{\mu}(\mathbf{x}_{new} \mid \mathcal{D}_2) = y^* - \mathbf{x}_{new}^T \hat{\beta}_{(2)}, \tag{14}$$

for $i \in \mathcal{D}_1 = \{1, \ldots, m\}$, where $\hat{\beta}_{(2)} = (\mathbf{X}_{(2)}^T \mathbf{X}_{(2)})^{-1} \mathbf{X}_{(2)}^T \mathbf{Y}_{(2)}$ is the least squares estimator based on $\mathcal{D}_2$. It follows that

$$Q^-(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* \geq \mathbf{x}_{new}^T \hat{\beta}_{(2)} + (y_i - \mathbf{x}_i^T \hat{\beta}_{(2)})\} + 1}{m+1} \text{ and } Q^+(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* > \mathbf{x}_{new}^T \hat{\beta}_{(2)} + (y_i - \mathbf{x}_i^T \hat{\beta}_{(2)})\}}{m+1}.$$

Based on (5), the level-$(1 - \alpha)$ prediction interval of y_{new}:

$$C_\alpha = \left[\mathbf{x}_{new}^T \hat{\beta}_{(2)} + \{y_i - \mathbf{x}_i^T \hat{\beta}_{(2)}\}_{[\frac{\alpha(m+1)}{2}]-1}, \mathbf{x}_{new}^T \hat{\beta}_{(2)} + \{y_i - \mathbf{x}_i^T \hat{\beta}_{(2)}\}_{[(1-\frac{\alpha}{2})(m+1)]}\right), \tag{15}$$

where $\{y_i - \mathbf{x}_i^T \hat{\beta}_{(2)}\}_K$ is the Kth sample quantile of $y_1 - \mathbf{x}_1^T \hat{\beta}_{(2)}, \ldots, y_m - \mathbf{x}_m^T \hat{\beta}_{(2)}$ and $[a]$ is the largest integer that does not exceed a. Note that, given $\mathbf{x}_{new}$, the point predictor $\mathbf{x}_{new}^T \hat{\beta}_{(2)}$ is an unbiased estimator of $\mathrm{E}(y_{new} | \mathbf{x}_{new}) = \mathbf{x}_{new}^T \beta$. Also, $\mathrm{E}(y_i - \mathbf{x}_i^T \hat{\beta}_{(2)}) = 0$, for $i = 1, \ldots, m$. This prediction interval (18) is "centered" at the unbiased point predictor $\mathbf{x}_{new}^T \hat{\beta}_{(2)}$ and its width is determined by the "spread" of the mean-zero "noises" $\{y_i - \mathbf{x}_i^T \hat{\beta}_{(2)}, i = 1, \ldots, m\}$.

Under the wrong learning model (13),

$$R_i = y_i - \widehat{\mu}_1(\mathbf{z}_i; \mathcal{D}_2) = y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)} \text{ and } R_i^* = y^* - \widehat{\mu}(\mathbf{z}_{new} \mid \mathcal{D}_2) = y^* - \mathbf{z}_{new}^T \hat{\gamma}_{(2)}, \tag{16}$$

for $i = 1, \ldots, m$, where $\hat{\gamma}_{(2)} = (\mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)})^{-1} \mathbf{Z}_{(2)}^T \mathbf{Y}_{(2)}$. It follows that

$$Q^-(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* \geq \mathbf{z}_{new}^T \hat{\gamma}_{(2)} + (y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)})\} + 1}{m+1} \text{ and } Q^+(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* > \mathbf{z}_{new}^T \hat{\gamma}_{(2)} + (y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)})\}}{m+1},$$

and the corresponding prediction interval of y_{new} is

$$\widetilde{C}_\alpha = \left[\mathbf{z}_{new}^T \hat{\gamma}_{(2)} + \{y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)}\}_{[\frac{\alpha(m+1)}{2}]-1}, \mathbf{z}_{new}^T \hat{\gamma}_{(2)} + \{y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)}\}_{[(1-\frac{\alpha}{2})(m+1)]}\right). \tag{17}$$

Given $\mathbf{x}_{new}$, the point predictor $\mathbf{z}_{new}^T \hat{\gamma}_{(2)}$ when using model (13) is biased due to missing the covariates $\mathbf{w}_i$'s in model (13), i.e.,

$$\begin{aligned} bias &= \mathrm{E}\left[\mathbf{z}_{new}^T \hat{\gamma}_{(2)} | \mathbf{X}, \mathbf{x}_{new}\right] - \mathbf{x}_{new}^T \beta = \mathbf{z}_{new}^T (\mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)})^{-1} \mathbf{Z}_{(2)}^T (\mathbf{Z}_{(2)} \beta_1 + \mathbf{W}_{(2)} \beta_2) - \mathbf{x}_{new}^T \beta \\ &= -\mathbf{w}_{new}^T \beta_2 + \mathbf{z}_{new}^T (\mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)})^{-1} \mathbf{Z}_{(2)}^T \mathbf{W}_{(2)} \beta_2, \end{aligned}$$

where β_1 and β_2 are the first p and the last $(p - q)$ elements of β, respectively. Meanwhile, there is a non-zero shift in the "residual" $R_i = y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)}$:

$$\begin{aligned} shift(i) &= E\left[y_i - \mathbf{z}_i^T \hat{\gamma}_{(2)} | \mathbf{X}, \mathbf{x}_{new}\right] = \mathbf{x}_i^T \beta - \mathbf{z}_i^T (\mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)})^{-1} \mathbf{Z}_{(2)}^T (\mathbf{Z}_{(2)} \beta_1 + \mathbf{W}_{(2)} \beta_2) \\ &= \mathbf{w}_i^T \beta_2 - \mathbf{z}_i^T (\mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)})^{-1} \mathbf{Z}_{(2)}^T \mathbf{W}_{(2)} \beta_2, \end{aligned}$$

for $i = 1, \ldots, m$. The $shift(i)$ and *bias* often have the opposite signs and thus, when added together as in (17), they cancel each other at some amount. The amount of cancellation depends on whether or not $\mathbf{x}_{new} = (\mathbf{z}_{new}, \mathbf{w}_{new})$ resembles $\mathbf{x}_i = (\mathbf{z}_i, \mathbf{w}_i)$.

3.1.1. The iid case where $\mathbf{x}_{new} \sim \mathbf{x}_i$

In this iid case, by Proposition 1, both predictive intervals in (15) and (17) are valid. Here, we would like to explain why (17) still has the valid coverage even though it is derived based on the wrong model (13). Let us start with a hypothetical case that the new individual is the "average individual" of the observed data with $\mathbf{x}_{new} = \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i = (\bar{\mathbf{w}}, \bar{\mathbf{z}})$. Then, the bias of the point predictor is $bias = -\bar{\mathbf{w}}^T\beta_2 + \bar{\mathbf{z}}^T(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)})^{-1}\mathbf{Z}_{(2)}^T\mathbf{W}_{(2)}\beta_2$ and the average shift of the residual terms is $\frac{1}{m}\sum_{i=1}^{m} shift(i) = -bias$. They exactly cancel each other out in the prediction interval (19). This cancellation explains why the prediction interval (19) is still on target, even if the learning model is wrong. The cancellation is not as complete, when new $\mathbf{x}_{new} \sim \mathbf{x}_i$ (not the "average individual" $\mathbf{x}_{new} = \bar{\mathbf{x}}$). It appears that the combination of an enlarged interval and the cancellation of the *bias* and *shift* helps ensure the validity of conformal prediction.

The wrong learning model (13) has an impact on the length of the prediction interval. The proposition below states that the width of the prediction interval based on the wrong model $\mu_1(\cdot)$ is expected to be wider than that based on the correct model $\mu_0(\cdot)$. A proof can be found in Appendix.

Proposition 3. *Under model (12), assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $\mathbf{x}_i$'s are iid from a normal distribution and $\beta_2^T\Sigma_{\mathrm{w|z}}\beta_2 > 0$, where $\Sigma_{\mathrm{w|z}} = \mathrm{var}(\mathbf{w}_1|\mathbf{z}_1)$, then*

$$\lim_{m,n-m\to\infty} \mathbb{P}\Big(\{y_i - \mathbf{z}_i^T\widehat{\gamma}_{(2)}\}_{[(1-\frac{\alpha}{2})(m+1)]} - \{y_i - \mathbf{z}_i^T\widehat{\gamma}_{(2)}\}_{[\frac{\alpha(m+1)}{2}]-1} > \{y_i - \mathbf{x}_i^T\widehat{\beta}_{(2)}\}_{[(1-\frac{\alpha}{2})(m+1)]} - \{y_i - \mathbf{x}_i^T\widehat{\beta}_{(2)}\}_{[\frac{\alpha(m+1)}{2}]-1}\Big) = 1.$$

That is, with probability tending to 1, the length of prediction interval (15) is greater than that of prediction interval (17).

3.1.2. The non-iid case where $\mathbf{x}_{new} \nsim \mathbf{x}_i$

The iid assumption is crucial to ensure the validity of a prediction when using a wrong learning model. If $\mathbf{x}_{new} \nsim \mathbf{x}_i$ or $\mathbf{x}_{new}$ and $\mathbf{x}_i$ are fixed, Proposition 1 does not apply and the prediction performance of (15) and (17) is quite different than before. Our discussion is around two specific scenarios: (i) ϵ_i and ϵ_{new} are iid and independent of $\mathbf{x}$'s; and (ii) ϵ_i and ϵ_{new} are independent but ϵ_i depends on $\mathbf{x}_i$ (for example, in the Poisson model ϵ_i depends on $\mu_0(\mathbf{x}_i)$).

Under scenario (i) and if we use the true model $\mu_0(\cdot)$ in the training, we have based on (14) that $R_i = \epsilon_i + \mathbf{x}_i^T(\hat{\beta}_{(2)} - \beta) = \epsilon_i + O_p(\frac{1}{\sqrt{n-m}})$. Similarly, when $y^* = y_{new}$, we have $R_i^* = \epsilon_{new} + \mathbf{x}_{new}^T(\hat{\beta}_{(2)} - \beta) = \epsilon_{new} + O_p(\frac{1}{\sqrt{n-m}})$. Since ϵ_i and ϵ_{new} are iid under scenario (i), the R_i and R_i^* are approximate conformal (up to $O_p(\frac{1}{\sqrt{n-m}})$). The statement of approximate conformal holds as long as $\mathbf{x}_k^T(\hat{\beta}_{(2)} - \beta) = O_p(\frac{1}{\sqrt{n-m}})$, for $k \in \mathcal{I}_1 \cup \{new\}$, a condition that typically holds under the standard design conditions imposed in the classical regression models. In this case, the prediction is still valid with a correct confidence statement. The result is summarized in the following proposition and a proof is given in the Appendix. In the proposition, the standard conditions of classical regression are used to ensure that $\widehat{\beta}_{(2)}$ is consistent, two of which considered in the proof are: (i) for fixed covariates $\mathbf{x}_i$, we assume $\lambda_{min}(\mathbf{X}_{(2)}^T\mathbf{X}_{(2)}) \to \infty$ and $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of the target matrix; (ii) for random covariates $\mathbf{x}_i$, we assume $\mathbf{x}_i$ is iid, compact and $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)$ is invertible, for $i = 1, \ldots, n$. See, e.g., [8], [10] and [11] for further details of the standard conditions.

Proposition 4. *Under model (12), assume $\epsilon_1, \ldots, \epsilon_n, \epsilon_{new}$ are iid from some continuous distribution $G(\cdot)$ and the design matrix of $\mathbf{x}_i$'s follows the aforementioned standard conditions imposed in the classical regression. For C_α defined in (15), we have*

$$\lim_{(n-m)\to\infty} \mathbb{P}\big(y_{new} \in C_\alpha \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new}\big) \geq 1 - \alpha.$$

In Proposition 4, the probability statement $\mathbb{P}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$ is with regard to the joint distribution of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$ at the fixed covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new}$. That is to say, for any given new target, the predictive interval obtained by the split conformal procedure using the correct training model $\mu_0(\cdot)$ has asymptotic right coverage, if $(n-m) \to \infty$. Note that, if the covariates $\mathbf{x}$'s are random, this probability statement $\mathbb{P}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$ is for the conditional probability of $(y_1, \ldots, y_n, y_{new})$, given $(\mathbf{x}_1 \ldots, \mathbf{x}_n, \mathbf{x}_{new})$, or equivalently, the conditional probability of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$, given $(\mathbf{x}_1 \ldots, \mathbf{x}_n, \mathbf{x}_{new})$. Otherwise, if $\mathbf{x}$'s from a fixed design with non-random covariates, the probability statement $\mathbb{P}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$ is just with regard to the joint distribution of $(y_1, \ldots, y_n, y_{new})$, or equivalently, the joint distribution of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$.

The above conditional probability statement (when the covariates $\mathbf{x}$'s are random) is different from that discussed in [2]. In [2], the authors assumed that the observed samples $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ and the testing data $(\mathbf{x}_{new}, y_{new})$

are all iid random samples from the same unknown distribution P; i.e., $(\mathbf{x}_i, y_i), (\mathbf{x}_{new}, y_{new}) \overset{iid}{\sim} P$. They studied prediction intervals with "distribution-free conditional coverage" in the sense that $\mathbb{P}(y_{new} \in \widehat{C}_n(\mathbf{x}_{new}) \mid \mathbf{x}_{new} = x) \geq 1 - \alpha$ for all P and almost all x. In particular, the probability $\mathbb{P}(\cdot \mid \mathbf{x}_{new} = x)$ in their article is with regard to the joint probability of $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$ conditioned only on $\mathbf{x}_{new} = x$. On the contrary, in our random design case with random covariates $\mathbf{x}$'s, we only need the assumption (2) specified on page 2 and the distribution of $\mathbf{x}_{new}$ does not need to be the same as $\mathbf{x}_i$. This is weaker than the assumption that $(\mathbf{x}_{new}, y_{new})$ has the same joint distribution as $(\mathbf{x}_i, y_i)$ adopted by Barber et al. [2]. More importantly, our conditional probability statement is for $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$ conditioned on all covariates $(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$, which is stronger than that discussed in [2]. Our conditional coverage implies theirs. The same discussion holds for the Jackknife plus conformal procedure in the next subsection.

Under scenario (i) but the wrong training model $\mu_1(\cdot)$ is used, we have based on (16) that $R_i = \epsilon_i + (\mathbf{w}_i^T - \mathbf{z}_i^T(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)})^{-1}\mathbf{Z}_{(2)}^T\mathbf{W}_{(2)})\beta_2$ and, with $y^* = y_{new}$, $R_i^* = \epsilon_{new} + (\mathbf{w}_{new}^T - \mathbf{z}_{new}^T(\mathbf{Z}_{(2)}^T\mathbf{Z}_{(2)})^{-1}\mathbf{Z}_{(2)}^T\mathbf{W}_{(2)})\beta_2$. When $\mathbf{x}_{new} \approx \mathbf{x}_i$, the difference between R_i and R_{new} can be very large and they are not conformal. The prediction does not provide us a valid inference with a correct confidence statement, which is confirmed in the numerical study in Section 4.

Under scenario (ii) and since $\mathbf{x}_{new} \approx \mathbf{x}_i$, we have $\epsilon_{new} \approx \epsilon_i$. Regardless which training model is used, the difference between R_i and R_{new} can be very large and they are not conformal. The prediction no longer provides us a valid inference with a correct confidence statement.

3.2. Jackknife plus conformal procedure

The jackknife plus conformal procedure does not need to split the dataset $\mathcal{D}$ and $m = n$. We define notations: $\mathbf{Y}$ is the $m \times 1$ response vector of the training (observed) data, $\mathbf{X}$ and $\mathbf{Z}$ are the $m \times p$ and $m \times q$ design matrices, respectively, and we have a matrix partition $\mathbf{X} = (\mathbf{Z}, \mathbf{W})$.

Under the true learning model $\mu_0(\mathbf{x}_i)$ and from the least squares estimation, we have,

$$R_i = y_i - \hat{\mu}_0^{(-i)}(\mathbf{x}_i|\mathcal{D}_i) = y_i - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T)^{-1}(\mathbf{X}^T\mathbf{Y} - \mathbf{x}_i y_i) = \frac{y_i - \mathbf{x}_i^T\hat{\beta}}{1 - h_{ii}} = u_i$$

for $i = 1, \ldots, m = n$, where $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the least squares estimator using entire set of observed data $\mathcal{D}$, $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ and $u_i = \frac{y_i - \mathbf{x}_i^T\hat{\beta}}{1 - h_{ii}}$ is the deleted residual. Similarly, we have

$$R_i^* = y^* - \hat{\mu}_0^{(-i)}(\mathbf{x}_{new}|\mathcal{D}_i) = y^* - \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T)^{-1}(\mathbf{X}^T\mathbf{Y} - \mathbf{x}_i y_i) = y^* - \mathbf{x}_{new}^T\hat{\beta} + h_{i,new}u_i,$$

where $h_{i,new} = \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$. It follows that

$$Q^-(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* \geq \mathbf{x}_{new}^T\hat{\beta} + (1 - h_{i,new})u_i\} + 1}{m + 1} \text{ and } Q^+(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* > \mathbf{x}_{new}^T\hat{\beta} + (1 - h_{i,new})u_i\}}{m + 1}.$$

By (5), the prediction interval of y_{new} is:

$$C_\alpha = \left[\mathbf{x}_{new}^T\hat{\beta} + \{(1 - h_{i,new})u_i\}_{[\frac{\alpha(m+1)}{2}]-1}, \mathbf{x}_{new}^T\hat{\beta} + \{(1 - h_{i,new})u_i\}_{[(1-\frac{\alpha}{2})(m+1)]}\right]. \tag{18}$$

Note that, given $\mathbf{x}_{new}$, the point predictor $\mathbf{x}_{new}^T\hat{\beta}$ is an unbiased estimator of $E(y_{new}|\ \mathbf{x}_{new}) = \mathbf{x}_{new}^T\beta$ and $E\{(1 - h_{i,new})u_i|\mathbf{x}_{new}\} = 0$, for $i = 1, \ldots, n$. The prediction interval (18) is "centered" at the unbiased predictor $\mathbf{x}_{new}^T\hat{\beta}$ and its width is determined by the "spread" of the mean-zero "noises" $\{(1 - h_{i,new})u_i, i = 1, \ldots, m = n\}$.

When the wrong training model $\mu_1(\mathbf{z})$ is used, we can use a similar derivation to get

$$Q^-(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* < \mathbf{z}_{new}^T\hat{\gamma} + (1 - g_{i,new})v_i\} + 1}{m + 1} \text{ and } Q^+(y^*) = \frac{\sum_{i=1}^m \mathbf{1}\{y^* \leq \mathbf{z}_{new}^T\hat{\gamma} + (1 - g_{i,new})v_i\}}{m + 1},$$

where $\hat{\gamma} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$ is the least squares estimator using the wrong model, $g_{ii} = \mathbf{z}_i^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}_i$, $g_{i,new} = \mathbf{z}_{new}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}_i$ and $v_i = \frac{y_i - \mathbf{z}_i^T\hat{\gamma}}{1 - g_{ii}}$. In this case and by (5), a prediction interval of y_{new} is

$$\widetilde{C}_\alpha = \left[\mathbf{z}_{new}^T\hat{\gamma} + \{(1 - g_{i,new})v_i\}_{[\frac{\alpha(m+1)}{2}]-1}, \mathbf{z}_{new}^T\hat{\gamma} + \{(1 - g_{i,new})v_i\}_{[(1-\frac{\alpha}{2})(m+1)]}\right]. \tag{19}$$

As in the split conformal case, due to missing the covariates $\mathbf{w}_i$, the point predictor $\mathbf{z}_{new}^T\hat{\gamma}$ under the wrong training model (13) is biased:

$$bias = \mathbb{E}\left[\mathbf{z}_{new}^T\hat{\gamma}|\mathbf{X}, \mathbf{x}_{new}\right] - \mathbf{x}_{new}^T\beta = -\mathbf{w}_{new}^T\beta_2 + \mathbf{z}_{new}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}\beta_2,$$

where β_2 is the last $(p - q)$ elements of β. Furthermore, the expectations of the residual terms have also nonzero shifts:

$$Shift(i) = \mathbb{E}\{(1-g_{i,new})v_i|\mathbf{X}, \mathbf{x}_{new}\} = \frac{1-g_{i,new}}{1-g_{ii}}\left(\mathbf{w}_i^T - \mathbf{z}_i^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}\right)\beta_2 = \frac{1-g_{i,new}}{1-g_{ii}}(\mathbf{w}_i^{\perp})^T\beta_2$$

where $\mathbf{w}_i^{\perp}$ is the ith row of the matrix $\mathbf{W}^{\perp} = \{I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\}\mathbf{W}$. Again, the $shift(i)$ and *bias* often have the opposite signs and thus, when added together in (19), they cancel each other to some extent, but the amount of cancellation depends on $\mathbf{x}_{new} = (\mathbf{z}_{new}, \mathbf{w}_{new})$ and $\mathbf{x}_i = (\mathbf{z}_i, \mathbf{w}_i)$.

3.2.1. The iid case where $\mathbf{x}_{new} \sim \mathbf{x}_i$

In this iid case, by Proposition 2, both predictive intervals in (18) and (19) are valid with a guaranteed coverage of $(1-2\alpha)100\%$ or more. We would like to explain why (19) still has the valid coverage even though it is derived based on the wrong model (13). Again, we start with the hypothetical case that the new individual is the "average individual" with $\mathbf{x}_{new} = \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i = (\bar{\mathbf{w}}, \bar{\mathbf{z}})$. In this case, the bias of the point predictor and the average shift of the residual terms are $bias = -(\bar{\mathbf{w}}^{\perp})^T\beta_2$ and $average\ shift = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1-1/n}{1-g_{ii}}(\mathbf{w}_i^{\perp})^T\beta_2\right\}$, respectively. Since $\frac{1-1/n}{1-g_{ii}} \approx 1$ when $\mathbf{z}_i$'s are iid (cf., Lemma A1 in Supplementary), the *average shift* $\approx (\bar{\mathbf{w}}^{\perp})^T\beta_2 = -bias$, thus they are approximately canceled out in the prediction interval (19). This cancellation explains in part why the prediction interval (19) is still roughly on target, even if the training model is wrong. The cancellation is not as complete, when the testing data $\mathbf{x}_{new}$ is just an iid sample and not the "average" $\bar{\mathbf{x}}$. Again, as in the split conformal prediction procedure discussed before, the combination of an enlarged interval and the cancellation of the *bias* and *shift* helps ensure the validity of conformal prediction under a wrong model in the iid case.

A wrong learning model also impacts the lengths of the prediction intervals. The proposition below is for a jackknife plus conformal procedure. It states that the width of the prediction interval using the wrong training model $\mu_1(\cdot)$ is expected to be wider than that based on the correct model $\mu_0(\cdot)$, if $\mathbf{x}_{new} = \bar{\mathbf{x}}$. A proof can be found in the Supplementary.

Proposition 5. *Under model (12), assume $\epsilon_1, \ldots, \epsilon_n, \epsilon_{new} \overset{iid}{\sim} N(0, \sigma^2)$, $\mathbf{x}_i$'s and $\mathbf{x}_{new}$ are iid from a normal distribution and $\beta_2^T\Sigma_{w|z}\beta_2 > 0$, where $\Sigma_{w|z} = \mathrm{var}(\mathbf{w}_1|\mathbf{z}_1)$. Suppose $\mathbf{x}_{new} = \bar{\mathbf{x}}$, then*

$$\lim_{n\to\infty}\mathbb{P}\Bigg(\{(1-g_{i,new})v_i\}_{[(1-\frac{\alpha}{2})(m+1)]} - \{(1-g_{i,new})v_i\}_{[\frac{\alpha(m+1)}{2}]-1} > \{(1-h_{i,new})u_i\}_{[(1-\frac{\alpha}{2})(m+1)]} - \{(1-h_{i,new})u_i\}_{[\frac{\alpha(m+1)}{2}]-1}\Bigg) = 1.$$

That is, with probability tending to 1, the length of prediction interval (18) is greater than that of prediction interval (19).

3.2.2. The non-iid case where $\mathbf{x}_{new} \nsim \mathbf{x}_i$

The discussion is very similar to that under the split conformal prediction procedure. We again consider two specific scenarios: (i) ϵ_i and ϵ_{new} are iid and independent of $\mathbf{x}$'s; and (ii) ϵ_i and ϵ_{new} are independent but ϵ_i depends on $\mathbf{x}_i$.

Under scenario (i) and if we use the true model $\mu_0(\cdot)$ in training, we have $R_i = y_i - \mathbf{x}_i^T\widehat{\beta}^{(-i)} = \epsilon_i + \mathbf{x}_i^T(\beta - \widehat{\beta}^{(-i)}) = \epsilon_i + O_p(\frac{1}{\sqrt{n-1}})$. Similarly, with $y^* = y_{new}$, $R_i^* = \epsilon_{new} + \mathbf{x}_{new}(\beta - \widehat{\beta}^{(-i)}) = \epsilon_{new} + O_p(\frac{1}{\sqrt{n-1}})$. Here, $\widehat{\beta}^{(-i)}$ is the least squared estimator of β based on the data set $\mathcal{D}^{(-i)} = \mathcal{D}\setminus\{(\mathbf{x}_i, y_i)\}$. Since ϵ_i and ϵ_{new} are iid, the R_i and R_i^* are approximate conformal (up to $O_p(\frac{1}{\sqrt{n-1}})$). Thus, the prediction is still valid with a correct confidence statement. This result is summarized in the following proposition and a proof is given in the Appendix. In the proposition, to ensure that $\widehat{\beta}^{(-i)}$ is consistent, the standard conditions imposed in classical regression are the same as those in Proposition 4 except that the design matrix $\mathbf{X}_{(2)}$ for the split conformal method is replaced by the entire design matrix $\mathbf{X}$ for the jackknife plus conformal method.

Proposition 6. *Under model (12), assume $\epsilon_1, \ldots, \epsilon_n, \epsilon_{new}$ are iid from some continuous distribution $G(\cdot)$ and the design matrix of $\mathbf{x}_i$'s follows the aforementioned standard conditions imposed in the classical regression. For C_α defined in (18), we have*

$$\lim_{n\to\infty}\mathbb{P}\left(y_{new} \in C_\alpha \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new}\right) \geq 1 - 2\alpha.$$

Similar to Proposition 4, the probability statement $\mathbb{P}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$ in Proposition 6 is with regard to the joint distribution of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$ at the fixed covariates $\mathbf{x}_{new}$ (and also fixed $\mathbf{x}_i$'s). That is to say, for any given new target, the predictive interval obtained by the Jackknife plus procedure using the correct training model $\mu_0(\cdot)$ has asymptotic right coverage, as $n \to \infty$. Note that, this probability statement $\mathbb{P}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$ is a conditional probability of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$, given $(\mathbf{x}_1 \ldots, \mathbf{x}_n, \mathbf{x}_{new})$, if the covariates $\mathbf{x}$'s are random; otherwise it is just the joint probability of ϵ's.

Under scenario (i) but a wrong training model $\mu_1(\cdot)$ is used, we prove in the appendix that we can express $R_i = \epsilon_i + \frac{1}{1-g_{ii}}\left\{\mathbf{w}_i^T - \mathbf{z}_i^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}\right\}\beta_2 + O_p(\frac{1}{\sqrt{n-1}})$. Similarly, with $y^* = y_{new}$, we can show that $R_i^* = \epsilon_{new} + \frac{1}{1-g_{ii}}\left\{\mathbf{w}_{new}^T - \mathbf{z}_{new}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}\right\}\beta_2 + O_p(\frac{1}{\sqrt{n-1}})$. When $\mathbf{x}_{new} = (\mathbf{z}_{new}^T, \mathbf{w}_{new}^T)^T \nsim \mathbf{x}_i = (\mathbf{z}_i^T, \mathbf{w}_i^T)^T$, the scores R_i and R_i^* are typically not conformal with each other, thus the prediction does not provide us a valid inference with a correct confidence statement.

Table 1
A summary on whether a conformal prediction procedure can provide us with a valid confidence statement under different assumption and scenarios. Yes[1] is only for the joint distribution of $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$. If $\mathbf{x}$'s are random, Yes[2] covers both the joint distribution of $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$ and the conditional distribution of $(y_1, \ldots, y_n, y_{new})|(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$. If the covariates $\mathbf{x}$'s are fixed and non-random, Yes[2] is only for the joint distribution of $(y_1, \ldots, y_n, y_{new})$. For the conditional statement or in the case that the covariates $\mathbf{x}$'s are from a fixed design, the YES statement is for an asymptotic coverage requiring a large enough sample size.

		Training model	
		True model	Wrong model
1) The iid case with $\mathbf{x}_{new} \sim \mathbf{x}_i$	Both scenarios (i) & (ii)	Yes[1]	Yes[1]
2) The non-iid case with $\mathbf{x}_{new} \nsim \mathbf{x}_i$	Scenario (i): ϵ is independent of $\mathbf{x}$	Yes[2]	No
	Scenario (ii): ϵ is not independent of $\mathbf{x}$	No	No

Under scenario (ii) with $\mathbf{x}_{new} \nsim \mathbf{x}_i$, we have $\epsilon_{new} \nsim \epsilon_i$. Again, regardless which training model that we use, the difference between R_i and R_{new} can be very large and they are not conformal. The prediction no longer provides us a valid inference with a correct confidence statement.

To end this section, we summarize our overall findings in Table 1. Note that in the iid case with $\mathbf{x}_{new} \sim \mathbf{x}_i$, the valid statement is regarding to the coverage of the joint distribution of $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$. In scenario (i) of the non iid case with $\mathbf{x}_{new} \nsim \mathbf{x}_i$, if $\mathbf{x}$'s are random, the valid statement is regarding to the coverage of both the joint distribution of $\mathcal{D} \cup \{(\mathbf{x}_{new}, y_{new})\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_{new}, y_{new})\}$ and the conditional distribution of $(y_1, \ldots, y_n, y_{new})|(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{new})$. For the conditional case, the statement refers to an asymptotic coverage for a large sample size. If $\mathbf{x}$'s are from a classical design with fixed non-random covariates, the valid statement is only regarding to the joint distribution of $(y_1, \ldots, y_n, y_{new})$, or equivalently, the joint distribution of $(\epsilon_1, \ldots, \epsilon_n, \epsilon_{new})$.

4. Numerical studies

We conduct a numerical study in two examples, one is under the regular linear regression setting with a continuous response variable y and the other under a Poisson model with a categorical response variable y. The linear regression example covers the scenarios in the first two rows of Table 1 and the Poisson example covers the scenarios in the first and the third rows of Table 1. Both examples together cover all cases and scenarios in Table 1 and provide empirical evidence that supports our discussions.

Example 2 *(Linear regression).* Consider a linear model of two covariates with the true model

$$y_i = \mu_0(\mathbf{x}_i) + \epsilon_i = \beta_0 + \beta_1 z_i + \beta_2 w_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{20}$$

$i = 1, \ldots n$, where ϵ_i and $\mathbf{x}_i$ are independent. In our numerical study, $(\beta_0, \beta_1, \beta_2) = (-1, 2, 2)$ and $\mathbf{x}_i = (z_i, w_i)^T \overset{iid}{\sim} N(\mu_x, \Sigma_x)$ with $\mu_x = (0, 0)^T$, $\sigma^2 = 1$ and the (k, k')-element of Σ_x equal to $0.5^{|k-k'|}/2$, $k, k' \in \{1, 2\}$ and $n = 500$.

For the new data, we consider two cases. Case 1: under the iid assumption that $\mathbf{x}_{new} \sim \mathbf{x}_i$ and, given $\mathbf{x}_{new}$, $(\mathbf{x}_{new}, y_{new})$ follows (20); Case 2: the marginal distribution of $\mathbf{x}_{new} \nsim \mathbf{x}_i$ with $\mathbf{x}_{new} \sim N(\tilde{\mu}_x, \tilde{\Sigma}_x)$ and, given $\mathbf{x}_{new}$, $(\mathbf{x}_{new}, y_{new})$ follows (20). Here, $\tilde{\mu}_x = \mu_x + (2, 2)^T$ and the (k, k')-element of $\tilde{\Sigma}_x$ is $0.8^{|k-k'|}/2$, $k, k' \in \{1, 2\}$.

In addition to the correct model (a) $\mu_0(\mathbf{x}_i) = \beta_0 + \beta_1 z_i + \beta_2 w_i$, three wrong learning models are considered:

(b) $\mu_1(\mathbf{x}_i) = \gamma_0 + \gamma_1 z_i$ (partially correct, without covariate w_i);

(c) $\mu_2(\mathbf{x}_i) = \xi_0 + \xi_1 z_i^2$ (a wrong regression form).

(d) $\mu_3(\mathbf{x}_i) = \eta_0$ (without any covariates);

For model fitting, we use the least squares method in all three cases.

Reported in each cell of Table 2 are the coverage rate and average length (inside brackets) of 95% conformal prediction intervals for y_{new}, computed based on 500 repetitions. As expected, in the iid case, all learning models can provide valid prediction results, with the smallest interval length obtained under the true model. In the non-iid case, only the true model can provide a valid prediction. The other three learning models do not provide valid predictive inference in terms of a correct coverage rate, even though their prediction intervals are wider. The results in both cases underscore the importance of using a correct learning model for prediction.

In order to get the full picture of the prediction intervals of all confidence levels under different scenarios and different learning models, we plot in Fig. 2 the predictive curves obtained using the split and jackknife plus conformal procedures. Each of the six plots (a1)-(c1) and (a2)-(c2) are based on one simulated data set from 500 repetitions (other simulation data sets provide similar plots with the same messages). Plots (a1)-(a2) are for $\mathbf{x}_{new} = (-0.009, 0.006) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, (b1)-(b2) are for $\mathbf{x}_{new} = (0.111, -0.637) \overset{iid}{\sim} \mathbf{x}_i$ and (c1)-(c2) $\mathbf{x}_{new} = (1.948, 0.592) \nsim \mathbf{x}_i$. In each plot, we have four predictive curves

Table 2
Performance of 95% prediction intervals under four learning models and in two scenarios (coverage rates (before brackets) and average interval lengths (inside brackets)). Model $\mu_1(\cdot)$ is a partially wrong model, $\mu_2(\cdot)$ is a completely wrong model and $\mu_3(\cdot)$ does not use any covariates. Training data size = 500; Testing data size = 1; Repetition = 500.

Split conformal prediction				
	True model	Wrong model		
	$\mu_0(\cdot)$	$\mu_1(\cdot)$	$\mu_2(\cdot)$	$\mu_3(\cdot)$
1) The iid case with $\mathbf{x}_{new} \sim \mathbf{x}_i$	.956 (3.96)	.962 (6.25)	.956 (10.52)	.954 (10.39)
2) The non-iid with $\mathbf{x}_{new} \nsim \mathbf{x}_i$	.95 (3.97)	.724 (6.27)	.254 (10.54)	.248 (10.48)
Jackknife+ conformal prediction				
	True model	Wrong model		
	$\mu_0(\cdot)$	$\mu_1(\cdot)$	$\mu_2(\cdot)$	$\mu_3(\cdot)$
1) The iid case with $\mathbf{x}_{new} \sim \mathbf{x}_i$	.96 (3.96)	.966 (6.24)	.944 (10.50)	.958 (10.46)
2) The non-iid with $\mathbf{x}_{new} \nsim \mathbf{x}_i$	.942 (3.96)	.726 (6.25)	.236 (10.28)	.238 (10.42)

corresponding to four working models, plus the target (oracle) predictive curve of $PV(y) = 2\max\{\Phi(y-\mu_{new}), 1-\Phi(y-\mu_{new})\}$ obtained by pretending that we know exactly y_{new}'s distribution: $y_{new} \sim N(\mu_{new}, 0.5)$ with $\mu_{new} = -1 + (2,2)\mathbf{x}_{new}$. In each of the plots, the predictive curves trained with the correct learning model (black solid curves) are very close to the target oracle predictive curves (lightest gray solid curves), indicating that if we use the true model as the learning model, we are able to provide accurate prediction at all confidence levels. Under the wrong models, however, the results are different. In plots (a1)-(a2) with $\mathbf{x}_{new}$ being the "average individual", we see an almost complete cancellation of *bias* and *shift* as described earlier. However, the predictive curves are much wider than those based on the correct model. Plots (b1)-(b2) are for the iid case of $\mathbf{x}_{new} \sim \mathbf{x}_i$. In this case the curves are similar to those in plots (a1)-(a2), although the cancellations are not as complete as for the "average individual." Nevertheless, the enlarged interval widths help maintain the coverage. Plots (c1)-(c2) are for the non-iid case, in which the cancellations of *bias* and *shift* are not effective when wrong learning models are used, leading to wrong predictions. In all plots, we can also see that a partially correct model $\mu_1(\cdot)$ performs better than the other two completely wrong models $\mu_2(\cdot)$ and $\mu_3(\cdot)$.

In summary, when we train prediction algorithms using a wrong model, the iid assumption is essential for the validity of prediction, and using a wrong model often results in wider, sometimes much wider, prediction intervals. When we train the same algorithms using the correct model, the validity and efficiency of the predictions are observed in both the iid and non-iid scenarios conditional on $\mathbf{x}$'s. The results provide numerical support for the first two rows in Table 1.

Example 3 *(Poisson regression).* Suppose that the response y is a Poisson count that follows a generalized linear model: $y_i|\mathbf{x}_i \sim \text{Poisson}(\mu_0(\mathbf{x}_i))$ with $\mu_0(\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = e^{\beta_0+\beta_1 z_i+\beta_2 w_i}$, for $i = 1, \ldots, n$ and new. In the form of (10), we have

$$y_i = \mu_0(\mathbf{x}_i) + \epsilon_i = e^{\beta_0+\beta_1 z_i+\beta_2 w_i} + \epsilon_i, \tag{21}$$

where $\epsilon_i = y_i - \mu_0(\mathbf{x}_i)$ is a mean 0 error term that deponents on $\mathbf{x}_i$. In our numerical study, $(\beta_0, \beta_1, \beta_2) = (-1, 1, 1)$, $\mathbf{x}_i = (z_i, w_i)^T \overset{iid}{\sim} N(\mu_x, \Sigma_x)$ with $\mu_x = (1,1)^T$ and the (k,k')-element of Σ_x equal to $0.5^{|k-k'|}/5$, $k, k' \in \{1,2\}$ and $n = 500$.

For the new data, we consider two cases (Table 3). Case 1: under the iid assumption that $\mathbf{x}_{new} \sim N(\mu_x, \Sigma_x)$ and $y_{new}|\mathbf{x}_{new}$, follows (21); Case 2: the marginal distribution of $\mathbf{x}_{new}$ is instead from $\mathbf{x}_{new} \sim N(\tilde{\mu}_x, \tilde{\Sigma}_x)$ and $y_{new}|\mathbf{x}_{new}$ follows (21). Here, $\tilde{\mu}_x = \mu_x + (1,1)^T$ and $\tilde{\Sigma}_x = 2\Sigma_x$.

In addition to the correct model (a) $\mu_0(\mathbf{x}_i) = e^{\beta_0+\beta_1 z_i+\beta_2 w_i}$, three wrong learning models are considered:

(b) $\mu_1(\mathbf{x}_i) = e^{\gamma_0+\gamma_1 z_i}$ (partially correct, without covariate w_i);

(c) $\mu_2(\mathbf{x}_i) = e^{\xi_0+\xi_1 z_i^2}$ (a wrong regression form).

(d) $\mu_3(\mathbf{x}_i) = e^{\eta_0}$ (without any covariates).

For model fitting, we use the maximum likelihood method in all cases.

Since the Poisson count y_{new} is discrete, we report in each cell of Table 2 the coverage rate and average cardinality (inside brackets) of 95% conformal predictive sets for y_{new}, using the split conformal procedure, computed based on 500 repetitions. As expected, in the iid case, all learning models provide valid prediction coverages and the smallest set is observed under the true model. In the non-iid case, all four models do not provide valid predictive inference in terms of a correct coverage rate, even for the true model.

Same as in Example 2, we plot in Fig. 3 the predictive curve functions obtained using the split conformal prediction procedure and based on a simulated data set. Other simulated data sets (in 500 repetitions) and also the jackknife conformal procedure produce more or less the same plots. Note that $\mathcal{Y}$ is a discrete space containing all non-negative integers, the plots resemble bar charts. Plot (a1)-(a3) is for $\mathbf{x}_{new} = (0.63, 0.79)$ (a realization from $\mathbf{x}_{new} \sim \mathbf{x}_i$) and (b1)-(b3) $\mathbf{x}_{new} = (2.31, 1.84)$

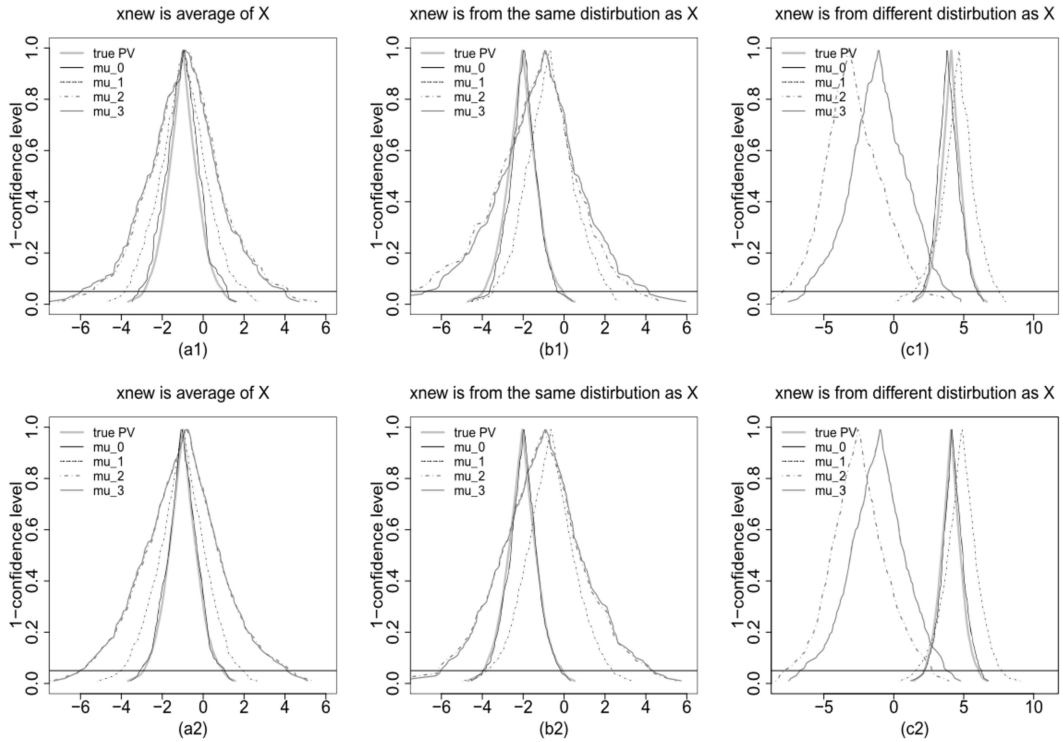

Fig. 2. Plots of predictive curves for (a1) & (a2): $\mathbf{x}_{new} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$; (b1) & (b2): $\mathbf{x}_{new} \overset{iid}{\sim} \mathbf{x}_i$ and (c1) & (c2): $\mathbf{x}_{new} \nsim \mathbf{x}_i$. In each plot, the lightest gray solid curve is the target (oracle) predictive curve $PV_n(y) = 2\max\{\Phi(y-\mu_{new}), 1-\Phi(y-\mu_{new})\}$, obtained assuming that the distribution of y_{new} is completely known. The predictive curves in black and darker gray are obtained using the four working models, respectively. The solid black curve is for learning model $\mu_0(\cdot)$, the dotted black for $\mu_1(\cdot)$, dashed gray for $\mu_2(\cdot)$ and solid gray for $\mu_3(\cdot)$. Each of the six plots (a1)-(c1) and (a2)-(c2) are based on a simulated data set (out of 500 repetitions). Plots (a1)-(c1) are obtained using the split conformal prediction procedure, and plots (a2)-(c2) are obtained using the jackknife plus conformal prediction procedure.

Table 3
Performance of 95% prediction intervals under four learning models and in two scenarios (coverage rates (before brackets) and average cardinality (inside brackets)). Model $\mu_1(\cdot)$ is a partially wrong model, $\mu_2(\cdot)$ is a completely wrong model and $\mu_3(\cdot)$ does not use any covariates. Training data size = 500; Testing data size = 1; Repetition = 500.

Split conformal prediction				
	True model	Wrong model		
	$\mu_0(\cdot)$	$\mu_1(\cdot)$	$\mu_2(\cdot)$	$\mu_3(\cdot)$
1) The iid case with $\mathbf{x}_{new} \sim \mathbf{x}_i$	.946 (7.078)	.942 (9.502)	.94 (9.65)	.978 (14.57)
2) The non-iid with $\mathbf{x}_{new} \nsim \mathbf{x}_i$	.566 (7.81)	.438 (10.52)	.396 (10.79)	.382 (14.6)

(a realization from $\mathbf{x}_{new} \nsim \mathbf{x}_i$). In each plot, we have two predictive curves corresponding to the true working models $\mu_0(\cdot)$ and one of the wrong models $\mu_1(\cdot)$-$\mu_3(\cdot)$, plus the target (oracle) predictive curve of $PV(y) = 2\min\left(Q^-(y), 1-Q^+(y)\right)$ obtained by pretending that we know exactly y_{new}'s distribution: $y_{new} \sim Poi(\mu_{new})$ with $\mu_{new} = \exp\left(-1+(1,1)\mathbf{x}_{new}\right)$.

In plot (a1)-(a3) with $\mathbf{x}_{new} \sim \mathbf{x}_i$, we see that the predictive curves are close to the target oracle curve in each plot, and the curve obtained from the true model is more concentrated than the others using the wrong training models, meaning that the confidence set obtained is smaller.

Plot (b1)-(b3) is for the non-iid case with $\mathbf{x}_{new} \nsim \mathbf{x}_i$. None of the confidence curves is close to the target oracle predictive curve and they do not provide sufficient coverage for y_{new}. In this simulation setup, we find that the prediction intervals obtained using the correct training model are too narrow. Also, the predictions by a wrong model have large biases and, even though their interval lengths are longer, they still can not sufficiently cover y_{new}. The results can also be seen in the Fig. 4, where we plot the corresponding level-95% predictive sets for y_{new} in Fig. 3 using the four different training models and the target prediction intervals obtained using the oracle predictive curve.

In summary, when we train prediction algorithms with the training error depending on the covariates, the iid assumption is essential for the validity of prediction. Even if the training model is true, the coverage is not correct. In addition, using a wrong model often results in wider, sometimes much wider, prediction interval. The messages are consistent with those in Table 1.

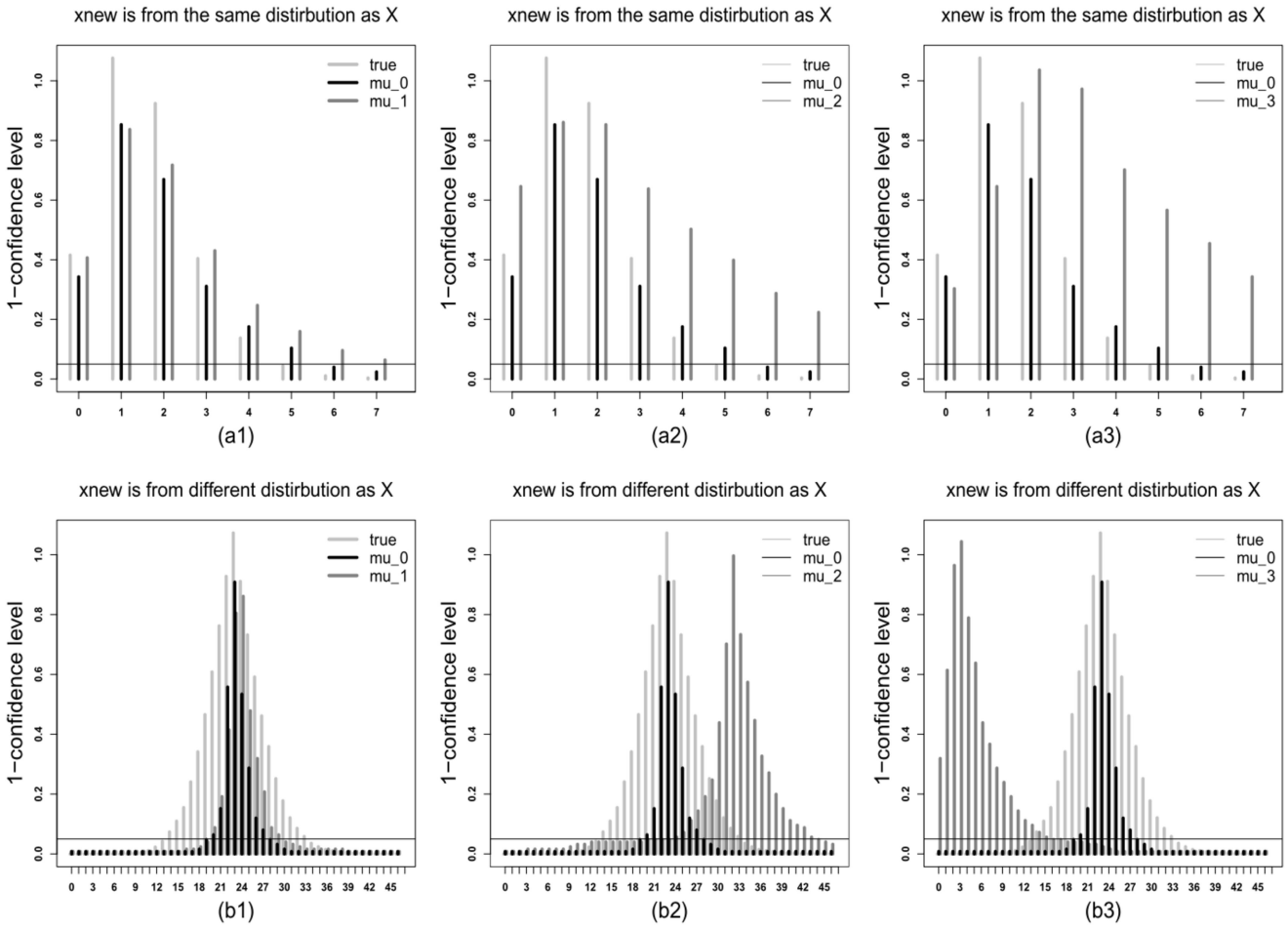

Fig. 3. Plots of predictive curves for (a1)-(a3): $\mathbf{x}_{new} \overset{iid}{\sim} \mathbf{x}_i$ and (b1)-(b3): $\mathbf{x}_{new} \nsim \mathbf{x}_i$. In each plot, the lightest gray line is the target (oracle) predictive curve $PV_n(y) = 2\max\{Q^+(y), 1 - Q^-(y)\}$, obtained assuming that the distribution of y_{new} is completely known. The black curve is for learning model $\mu_0(\cdot)$ and darker gray for $\mu_1(\cdot)$ in (a1),(b1); for $\mu_2(\cdot)$ in (a2),(b2) and for $\mu_3(\cdot)$ in (a3),(b3). The plots in each row are based on the same simulated data set (a data set out of 500 repetitions). This figure is obtained using the split conformal prediction procedure. Similar figures are obtained using the jackknife plus procedure (not included in the paper).

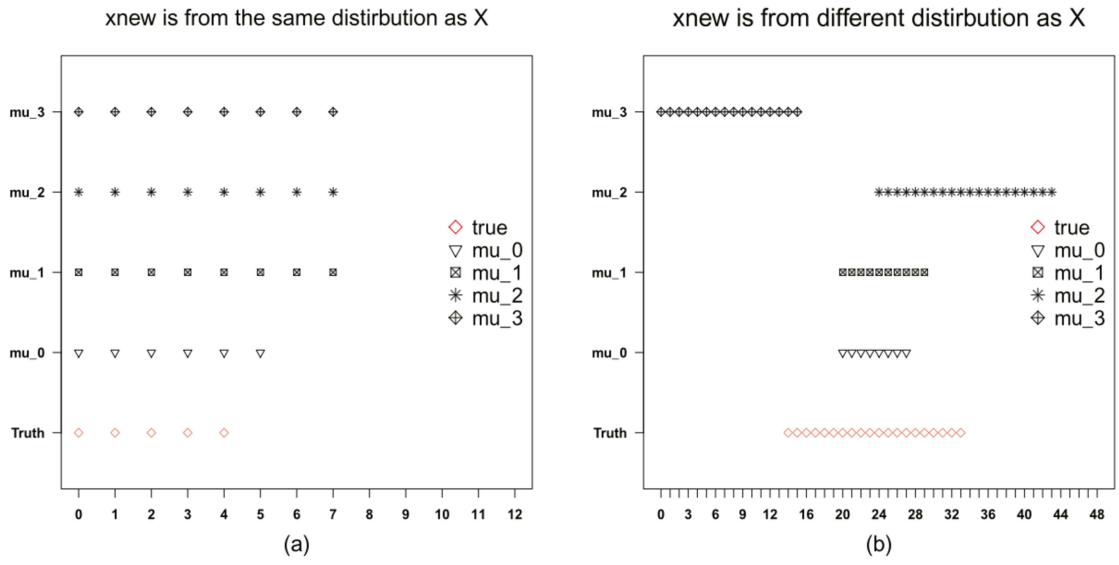

Fig. 4. Plots of level-95% predictive sets under different training models for (a): $\mathbf{x}_{new} \overset{iid}{\sim} \mathbf{x}_i$ and (b): $\mathbf{x}_{new} \nsim \mathbf{x}_i$. They are compared to the oracle level-95% prediction interval assuming we know the actual predictive distribution of y_{new}. These two plots correspond to the two rows of Fig. 3, respectively.

5. Conclusion

"The 21st Century has seen the rise of a new breed" of "stunningly successful prediction algorithms" [9]. The conformal prediction algorithm is one of such successful stories that have been attracting increased interest. Different than a conventional prediction algorithm in computer science, a conformal prediction procedure provides inference conclusions with a quantified uncertainty and a clear frequentist interpretation. Conformal prediction is non-parametric and distribution free. It has a wide range of applications. When the condition is right, an inference conclusion from a conformal prediction procedure is resistant to the use of wrong learning models. This robust homeostasis property provides an assurance for problems whose models are difficult to fit. It opens the door for us to use data-driven black-box approaches to tackle many complex and difficulty problems.

In this article, we have specifically studied in details the homeostasis property under a general regression setup. To deal with the discrete nature of the typically conformal prediction problems, we also introduced the concepts of upper and lower predictive distributions and predictive curve to establish connections to left-, right- and two-tailed hypothesis testing problems as well as the developments of confidence distributions. Our study explores the boundary at which the homeostasis property breaks down. Beside the typical assumption used in conformal prediction that the response and covariate pairs $(y, \mathbf{x})$ of all subjects are iid distributed, we study the classical regression setting that the design is fixed with given (non-random) covariates $\mathbf{x}$. The trade-off among learning model accuracy, prediction valid and prediction efficiency is discussed, leading to an emphasis of more efforts on developing better learning models and also better understanding of the impact of error assumptions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijar.2021.09.001.

References

[1] D.J. Aldous, Exchangeability and related topics, in: École d'Été de Probabilités de Saint-Flour XIII—1983, Springer, 1985, pp. 1–198.
[2] R.F. Barber, E.J. Candes, A. Ramdas, R.J. Tibshirani, The limits of distribution-free conditional predictive inference, arXiv preprint, arXiv:1903.04684, 2019.
[3] R.F. Barber, E.J. Candes, A. Ramdas, R.J. Tibshirani, Predictive inference with the jackknife+, Ann. Stat. 49 (2021) 486–507.
[4] A. Birnbaum, Confidence curves: an omnibus technique for estimation and testing statistical hypotheses, J. Am. Stat. Assoc. 56 (1961) 246–249.
[5] R.D. Cook, S. Weisberg, Residuals and Influence in Regression, Chapman and Hall, 1982.
[6] D.R. Cox, Some problems connected with statistical inference, Ann. Math. Stat. 29 (1958) 357–372.
[7] Y. Cui, M. Xie, Confidence distribution and distribution estimation for modern statistical inference, in: H. Pham (Ed.), Handbook of Engineering Statistics, 2nd edition, Springer, 2021.
[8] H. Drygas, Weak and strong consistency of the least squares estimators in regression models, Z. Wahrscheinlichkeitstheor. Verw. Geb. 34 (1976) 119–127.
[9] B. Efron, Prediction, estimation, and attribution (with discussion), J. Am. Stat. Assoc. 115 (2020) 636–655.
[10] F. Eicker, Asymptotic normality and consistency of the least squares estimators for families of linear regressions, Ann. Math. Stat. (1963) 447–456.
[11] L. Fahrmeir, H. Kaufmann, Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, Ann. Stat. 13 (1985) 342–368.
[12] S. Ferson, V. Kreinovich, L. Ginzburg, D. Myers, K. Sentz, Constructing Probability Boxes and Dempster-Shafer structures, Sandia National Laboratories, Tech. Rep., Technical report SANDD2002-4015, 2003.
[13] F. Lawless, M. Fredette, Frequentist prediction intervals and predictive distributions, Biometrika 92 (2005) 529–542.
[14] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, J. Am. Stat. Assoc. 113 (2018) 1094–1111.
[15] X. Luo, T. Dasgupta, M. Xie, R.Y. Liu, Leveraging the Fisher randomization test using confidence distributions: inference, combination and fusion learning, J. R. Stat. Soc. B 83 (2021) 777–797, https://doi.org/10.1111/rssb.12429.
[16] R. Martin, C. Liu, Inferential Models: Reasoning with Uncertainty, Vol. 145, CRC Press, 2015.
[17] S. Rosset, R.J. Tibshirani, From fixed-X to random-X regression: bias-variance decompositions, covariance penalties, and prediction error estimation, J. Am. Stat. Assoc. 115 (2020) 138–151.
[18] T. Schweder, N. Hjort, Confidence, Likelihood and Probability, Cambridge University Press, Cambridge, U.K., 2016.
[19] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, 1976.
[20] J. Shen, R. Liu, M. Xie, Prediction with confidence—a general framework for predictive inference, J. Stat. Plan. Inference 195 (2018) 126–140.
[21] K. Singh, M. Xie, W.E. Strawderman, Confidence Distribution (CD) – Distribution Estimator of a Parameter, IMS Lecture Notes-Monograph Series, vol. 54, Institute of Mathematical Statistics, 2007, pp. 132–150.
[22] V. Vovk, Conditional validity of inductive conformal predictors, in: Asian Conference on Machine Learning, PMLR, 2012, pp. 475–490.
[23] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer Science & Business Media, 2005.
[24] V. Vovk, J. Shen, V. Manokhin, M. Xie, Nonparametric predictive distributions by conformal prediction, Mach. Learn. 108 (2019) 445–474.
[25] M. Xie, K. Singh, Confidence distribution, the frequentist distribution estimator of a parameter (with discussion), Int. Stat. Rev. 81 (2013) 3–39.