CrossMark

Nonparametric predictive distributions based on conformal prediction

Vladimir Vovk[1] · Jieli Shen[2] · Valery Manokhin[1] · Min-ge Xie[2]

Abstract
This paper applies conformal prediction to derive predictive distributions that are valid under a nonparametric assumption. Namely, we introduce and explore predictive distribution functions that always satisfy a natural property of validity in terms of guaranteed coverage for IID observations. The focus is on a prediction algorithm that we call the Least Squares Prediction Machine (LSPM). The LSPM generalizes the classical Dempster–Hill predictive distributions to nonparametric regression problems. If the standard parametric assumptions for Least Squares linear regression hold, the LSPM is as efficient as the Dempster–Hill procedure, in a natural sense. And if those parametric assumptions fail, the LSPM is still valid, provided the observations are IID.

Keywords Conformal prediction · Least Squares · Nonparametric regression · Predictive distributions · Regression

1 Introduction

This paper applies conformal prediction to derive predictive distribution functions that are valid under a nonparametric assumption. In our definition of predictive distribution functions and their property of validity we follow Shen et al. (2018, Section 1), whose terminology we adopt, and Schweder and Hjort (2016, Chapter 12), who use the term "prediction confidence

---

✉ Vladimir Vovk
v.vovk@rhul.ac.uk

Jieli Shen
jlshen2011@gmail.com

Valery Manokhin
valery.manokhin.2015@live.rhul.ac.uk

Min-ge Xie
mxie@stat.rutgers.edu

1   Royal Holloway, University of London, Egham, Surrey, UK

2   Rutgers University, New Brunswick, NJ, USA

distributions". The theory of predictive distributions as developed by Schweder and Hjort (2016) and Shen et al. (2018) assumes that the observations are generated from a parametric statistical model. We extend the theory to the case of regression under the general IID model (the observations are generated independently from the same distribution), where the distribution form does not need to be specified; however, our exposition is self-contained. Our predictive distributions generalize the classical Dempster–Hill procedure (to be formally defined in Sect. 5), which these authors referred to as direct probabilities (Dempster) and $A_{(n)}/H_{(n)}$ (Hill). For a well-known review of predictive distributions, see Lawless and Fredette (2005). The more recent review by Gneiting and Katzfuss (2014) refers to the notion of validity used in this paper as probabilistic calibration and describes it as critical in forecasting; Gneiting and Katzfuss (2014, Section 2.2.3) also give further references.

We start our formal exposition by defining conformal predictive distributions (CPDs), nonparametric predictive distributions based on conformal prediction, and algorithms producing CPDs (conformal predictive systems, CPSs) in Sect. 2; we are only interested in (nonparametric) regression problems in this paper. An unusual feature of CPDs is that they are randomized, although they are typically affected by randomness very little. The starting point for conformal prediction is the choice of a conformity measure; not all conformity measures produce CPDs, but we give simple sufficient conditions. In Sect. 3 we apply the method to the classical Least Squares procedure obtaining what we call the Least Squares Prediction Machine (LSPM). The LSPM is defined in terms of regression residuals; accordingly, it has three main versions: ordinary, deleted, and studentized. The most useful version appears to be studentized, which does not require any assumptions on how influential any of the individual observations is. We state the studentized version (and, more briefly, the ordinary version) as an explicit algorithm. The next two sections, Sects. 4 and 5, are devoted to the validity and efficiency of the LSPM. Whereas the LSPM, as any CPS, is valid under the general IID model, for investigating its efficiency we assume a parametric model, namely the standard Gaussian linear model. The question that we try to answer in Sect. 5 is how much we should pay (in terms of efficiency) for the validity under the general IID model enjoyed by the LSPM. We compare the LSPM with three kinds of oracles under the parametric model; the oracles are adapted to the parametric model and are only required to be valid under it. The weakest oracle (Oracle I) only knows the parametric model, and the strongest one (Oracle III) also knows the parameters of the model. In important cases the LSPM turns out to be as efficient as the Dempster–Hill procedure. All proofs are postponed to Sect. 6, which also contains further discussions. Section 7 is devoted to experimental results demonstrating the validity and, to some degree, efficiency of our methods. Finally, Sect. 8 concludes and lists three directions of further research.

Another method of generating predictive distributions that are valid under the IID model is Venn prediction (Vovk et al. 2005, Chapter 6). An advantage of the method proposed in this paper is that it works in the case of regression, whereas Venn prediction, at the time of writing of this paper, was only known to work in the case of classification (see, however, the recent paper by Nouretdinov et al. 2018, discussed in Sect. 8).

The conference version of this paper (Vovk et al. 2017b), announcing the main results, was published in the Proceedings of COPA 2017. This expanded journal version includes detailed proofs, further discussion of the intuition behind the proposed algorithms, and topics for further research, in addition to improved exposition.

A significant advantage of conformal predictive distributions over traditional conformal prediction is that the former can be combined with a utility function to arrive at optimal decisions. A first step in this direction has been made in Vovk and Bendtsen (2018) (developing ideas of the conference version of this paper).

2 Randomized and conformal predictive distributions

We consider the regression problem with p attributes. Correspondingly, the *observation space* is defined to be $\mathbb{R}^{p+1} = \mathbb{R}^p \times \mathbb{R}$; its element $z = (x, y)$, where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$, is interpreted as an *observation* consisting of an *object* $x \in \mathbb{R}^p$ and its *label* $y \in \mathbb{R}$. Our task is, given a *training sequence* of observations $z_i = (z_i, y_i)$, $i = 1, \ldots, n$, and a new test object $x_{n+1} \in \mathbb{R}^p$, to predict the label y_{n+1} of the $(n+1)$st observation. Our statistical model is the general IID model: the observations $z_1, z_2, \ldots$, where $z_i = (x_i, y_i)$, are generated independently from the same unknown probability measure P on $\mathbb{R}^{p+1}$.

We start from defining predictive distribution functions following Shen et al. (2018, Definition 1), except that we relax the definition of a distribution function and allow randomization. Let U be the uniform probability measure on the interval $[0, 1]$.

Definition 1 A function $Q : (\mathbb{R}^{p+1})^{n+1} \times [0, 1] \to [0, 1]$ is called a *randomized predictive system* (RPS) if it satisfies the following three requirements:

R1a For each training sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and each test object $x_{n+1} \in \mathbb{R}^p$, the function $Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau)$ is monotonically increasing both in $y \in \mathbb{R}$ and in $\tau \in [0, 1]$ (where "monotonically increasing" is understood in the wide sense allowing intervals of constancy). In other words, for each $\tau \in [0, 1]$, the function

$$y \in \mathbb{R} \mapsto Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau)$$

is monotonically increasing, and for each $y \in \mathbb{R}$, the function

$$\tau \in [0, 1] \mapsto Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau)$$

is monotonically increasing.

R1b For each training sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and each test object $x_{n+1} \in \mathbb{R}^p$,

$$\lim_{y \to -\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y), 0) = 0 \tag{1}$$

and

$$\lim_{y \to \infty} Q(z_1, \ldots, z_n, (x_{n+1}, y), 1) = 1.$$

R2 As function of random training observations $z_1 \sim P, \ldots, z_n \sim P$, a random test observation $z_{n+1} \sim P$, and a random number $\tau \sim U$, all assumed independent, the distribution of Q is uniform:

$$\forall \alpha \in [0, 1] : \mathbb{P}\{Q(z_1, \ldots, z_n, z_{n+1}, \tau) \leq \alpha\} = \alpha.$$

The output of the randomized predictive system Q on a training sequence $z_1, \ldots, z_n$ and a test object x_{n+1} is the function

$$Q_n : (y, \tau) \in \mathbb{R} \times [0, 1] \mapsto Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau), \tag{2}$$

which will be called the *randomized predictive distribution (function)* (RPD) output by Q. The *thickness* of an RPD Q_n is the infimum of the numbers $\epsilon \geq 0$ such that the diameter

$$Q_n(y, 1) - Q_n(y, 0) \tag{3}$$

of the set

$$\{Q_n(y, \tau) \mid \tau \in [0, 1]\} \tag{4}$$

is at most ϵ for all $y \in \mathbb{R}$ except for finitely many values. The *exception size* of Q_n is the cardinality of the set of y for which the diameter (3) exceeds the thickness of Q_n. Notice that *a priori* the exception size can be infinite.

In this paper we will be interested in RPDs of thickness $\frac{1}{n+1}$ with exception size at most n, for typical training sequences of length n [cf. (17) below]. In all our examples, $Q(z_1, \ldots, z_n, z_{n+1}, \tau)$ will be a continuous function of τ. Therefore, the set (4) will be a closed interval in $[0, 1]$. However, we do not include these requirements in our official definition.

Four examples of predictive distributions are shown in Fig. 5 below as shaded areas; let us concentrate, for concreteness, on the top left one. The length of the training sequence for that plot (and the other three plots) is $n = 10$; see Sect. 7 for details. Therefore, we are discussing an instance of Q_{10}, of width $1/11$ with exception size 10. The shaded area is $\{(y, Q_{10}(y, \tau)) \mid y \in \mathbb{R}, \tau \in [0, 1]\}$. We can regard (y, τ) as a coordinate system for the shaded area. The cut of the shaded area by the vertical line passing through a point y of the horizontal axis is the closed interval $[Q(y, 0), Q(y, 1)]$, where $Q := Q_{10}$. The notation $Q(y)$ for the vertical axis is slightly informal.

Next we give basic definitions of conformal prediction adapted to producing predictive distributions (there are several equivalent definitions; the one we give here is closer to Vovk et al. 2005, Section 2.2, than to Balasubramanian et al. 2014, Section 1.3). A *conformity measure* is a measurable function $A : (\mathbb{R}^{p+1})^{n+1} \rightarrow \mathbb{R}$ that is invariant with respect to permutations of the first n observations: for any sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$, any $z_{n+1} \in \mathbb{R}^{p+1}$, and any permutation π of $\{1, \ldots, n\}$,

$$A(z_1, \ldots, z_n, z_{n+1}) = A\left(z_{\pi(1)}, \ldots, z_{\pi(n)}, z_{n+1}\right). \tag{5}$$

Intuitively, A measures how large the label y_{n+1} in z_{n+1} is, based on seeing the observations $z_1, \ldots, z_n$ and the object x_{n+1} of z_{n+1}. A simple example is

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \hat{y}_{n+1}, \tag{6}$$

$\hat{y}_{n+1}$ being the prediction for y_{n+1} computed from x_{n+1} and $z_1, \ldots, z_n$ as training sequence (more generally, we could use the whole of $z_1, \ldots, z_{n+1}$ as the training sequence).

The *conformal transducer* determined by a conformity measure A is defined as

$$
\begin{aligned}
Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) &:= \frac{1}{n+1} \left| \{ i = 1, \ldots, n+1 \mid \alpha_i^y < \alpha_{n+1}^y \} \right| \\
&+ \frac{\tau}{n+1} \left| \{ i = 1, \ldots, n+1 \mid \alpha_i^y = \alpha_{n+1}^y \} \right|,
\end{aligned} \tag{7}
$$

where $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ is a training sequence, $x_{n+1} \in \mathbb{R}^p$ is a test object, and for each $y \in \mathbb{R}$ the corresponding *conformity score* α_i^y is defined by

$$
\begin{aligned}
\alpha_i^y &:= A(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, (x_{n+1}, y), z_i), \quad i = 1, \ldots, n, \\
\alpha_{n+1}^y &:= A(z_1, \ldots, z_n, (x_{n+1}, y)).
\end{aligned} \tag{8}
$$

A function is a *conformal transducer* if it is the conformal transducer determined by some conformity measure. A *conformal predictive system* (CPS) is a function which is both a conformal transducer and a randomized predictive system. A *conformal predictive distribution* (CPD) is a function Q_n defined by (2) for a conformal predictive system Q.

Any conformal transducer Q and Borel set $A \subseteq [0, 1]$ define the *conformal predictor*

$$\Gamma^A(z_1, \ldots, z_n, x_{n+1}, \tau) := \{ y \in \mathbb{R} \mid Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) \in A \}. \tag{9}$$

The standard property of validity for conformal transducers is that the values (also called p values) $Q(z_1, \ldots, z_{n+1}, \tau)$ are distributed uniformly on $[0, 1]$ when $z_1, \ldots, z_{n+1}$ are IID and

τ is generated independently of $z_1, \ldots, z_{n+1}$ from the uniform probability distribution U on $[0, 1]$ (see, e.g., Vovk et al. 2005, Proposition 2.8). This property coincides with requirement R2 in the definition of an RPS and implies that the coverage probability, i.e., the probability of $y_{n+1} \in \Gamma^A(z_1, \ldots, z_n, x_{n+1})$, for the conformal predictor (9) is $U(A)$.

Remark 1 The usual interpretation of (7) is that it is a randomized p value for testing the null hypothesis of the observations being IID. In the case of CPDs, the informal alternative hypothesis is that $y_{n+1} = y$ is smaller than expected under the IID model. Then (6) can be interpreted as a degree of conformity of the observation (x_{n+1}, y_{n+1}) to the remaining observations. Notice the one-sided nature of this notion of conformity: a label can only be strange (non-conforming) if it is too small; large is never strange. This notion of conformity is somewhat counterintuitive, and we use it only as a technical tool.

2.1 Defining properties of distribution functions

Next we discuss why Definition 1 (essentially taken from Shen et al. 2018) is natural. The key elements of this definition are that (1) the distribution function Q is monotonically increasing, and (2) its value is uniformly distributed. The following two lemmas show that these are defining properties of distribution functions of probability measures on the real line. All proofs are postponed to Sect. 6.

First we consider the case of a continuous distribution function; the justification for this case, given in the next lemma, is simpler.

Lemma 1 *Suppose F is a continuous distribution function on $\mathbb{R}$ and Y is a random variable distributed as F. If $Q : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing function such that the distribution of $Q(Y)$ is uniform on $[0, 1]$, then $Q = F$.*

In the general case we need randomization. Remember the definition of the lexicographic order on $\mathbb{R} \times [0, 1]$: $(y, \tau) \leq (y', \tau')$ is defined to mean that $y < y'$ or both $y = y'$ and $\tau \leq \tau'$.

Lemma 2 *Let P be a probability measure on $\mathbb{R}$, F be its distribution function, and Y be a random variable distributed as P. If $Q : \mathbb{R} \times [0, 1] \to \mathbb{R}$ is a function that is monotonically increasing (in the lexicographic order on its domain) and such that the image $(P \times U)Q^{-1}$ of the product $P \times U$, where U is the uniform distribution on $[0, 1]$, under the mapping Q is uniform on $[0, 1]$, then, for all y and τ,*

$$Q(y, \tau) = (1 - \tau)F(y-) + \tau F(y). \tag{10}$$

Equality (10) says that Q is essentially F; in particular, $Q(y, \tau) = F(y)$ at each point y of F's continuity. It is a known fact that if we define Q by (10) for the distribution function F of a probability measure P, the distribution of Q will be uniform when its domain $\mathbb{R} \times [0, 1]$ is equipped with the probability measure $P \times U$.

The previous two lemmas suggest that properties R1a and R2 in the definition of RPSs are the important ones. However, property R1b is formally independent of R1a and R2 in our case of the general IID model (rather than a single probability measure on $\mathbb{R}$): consider, e.g., a conformity measure A that depends only on the objects x_i but does not depend on their labels y_i; e.g., the left-hand side of (1) will be close to 1 for large n and highly conforming x_{n+1}.

2.2 Simplest example: monotonic conformity measures

We start from a simple but very restrictive condition on a conformity measure making the corresponding conformal transducer satisfy R1a. A conformity measure A is *monotonic* if $A(z_1, \ldots, z_{n+1})$ is:

– monotonically increasing in y_{n+1},

$$y_{n+1} \leq y'_{n+1} \implies A(z_1, \ldots, z_n, (x_{n+1}, y_{n+1})) \leq A(z_1, \ldots, z_n, (x_{n+1}, y'_{n+1}));$$

– monotonically decreasing in y_1,

$$y_1 \leq y'_1 \implies A((x_1, y_1), z_2, \ldots, z_n, z_{n+1}) \geq A((x_1, y'_1), z_2, \ldots, z_n, z_{n+1}).$$

(Because of the requirement of invariance (5), being decreasing in y_1 is equivalent to being decreasing in y_i for any $i = 2, \ldots, n$.)

This condition implies that the corresponding conformal transducer (7) satisfies R1a by Lemma 3 below.

An example of a monotonic conformity measure is (6), where $\hat{y}_{n+1}$ is produced by the K-nearest neighbours regression algorithm:

$$\hat{y}_{n+1} := \frac{1}{K} \sum_{k=1}^{K} y_{(k)}$$

is the average label of the K nearest neighbours of x_{n+1}, where $y_{(1)}, \ldots, y_{(n)}$ is the sequence $y_1, \ldots, y_n$ sorted in the order of increasing distances between x_i and x_{n+1} (we assume $n \geq K$ and in the case of ties replace each $y_{(i)}$ by the average of y_j over all j such that the distance between x_j and x_{n+1} is equal to the distance between x_i and x_{n+1}). This conformity measure satisfies, additionally,

$$\lim_{y \to \pm\infty} A(z_1, \ldots, z_n, (x_n, y)) = \pm\infty$$

and, therefore, the corresponding conformal transducer also satisfies R1b and so is an RPS and a CPS.

2.3 Criterion of being a CPS

Unfortunately, many important conformity measures are not monotonic, and the next lemma introduces a weaker sufficient condition for a conformal transducer to be an RPS.

Lemma 3 *The conformal transducer determined by a conformity measure A satisfies condition R1a if, for each $i \in \{1, \ldots, n\}$, each training sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$, and each test object $x_{n+1} \in \mathbb{R}^p$, $\alpha_{n+1}^y - \alpha_i^y$ is a monotonically increasing function of $y \in \mathbb{R}$ (in the notation of (8)).*

Of course, we can fix i to, say, $i := 1$ in Lemma 3. We can strengthen the conclusion of the lemma to the conformal transducer determined by A being an RPS (and, therefore, a CPS) if, e.g.,

$$\lim_{y \to \pm\infty} \left(\alpha_{n+1}^y - \alpha_i^y\right) = \pm\infty.$$

3 Least squares prediction machine

In this section we will introduce three versions of what we call the Least Squares Prediction Machine (LSPM). They are analogous to the Ridge Regression Confidence Machine (RRCM), as described in Vovk et al (2005, Section 2.3) (and called the IID predictor in Vovk et al. 2009), but produce (at least usually) distribution functions rather than prediction intervals.

The *ordinary LSPM* is defined to be the conformal transducer determined by the conformity measure

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \widehat{\widehat{y}}_{n+1} \tag{11}$$

[cf. (6)], where y_{n+1} is the label in z_{n+1} and $\widehat{\widehat{y}}_{n+1}$ is the prediction for y_{n+1} computed using Least Squares from x_{n+1} (the object in z_{n+1}) and $z_1, \ldots, z_{n+1}$ (including z_{n+1}) as training sequence. The right-hand side of (11) is the ordinary residual. However, two more kinds of residuals are common in statistics, and so overall we will discuss three kinds of LSPM. The *deleted LSPM* is determined by the conformity measure

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \hat{y}_{n+1}, \tag{12}$$

whose difference from (11) is that $\widehat{\widehat{y}}_{n+1}$ is replaced by the prediction $\hat{y}_{n+1}$ for y_{n+1} computed using Least Squares from x_{n+1} and $z_1, \ldots, z_n$ as training sequence (so that the training sequence does not include z_{n+1}). The version that will be most useful in this paper will be the "studentized LSPM", which is midway between ordinary and deleted LSPM; we will define it formally later.

Unfortunately, the ordinary and deleted LSPM are not RPS, because their output Q_n [see (2)] is not necessarily monotonically increasing in y (remember that, for conformal transducers, $Q_n(y, \tau)$ is monotonically increasing in τ automatically). However, we will see that this can happen only in the presence of high-leverage points.

Let $\bar{X}$ stand for the $(n + 1) \times p$ data matrix, whose ith row is the transpose x_i' to the ith object (training object for $i = 1, \ldots, n$ and test object for $i = n + 1$). The hat matrix for the $n + 1$ observations $z_1, \ldots, z_{n+1}$ is

$$\bar{H} = \bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'. \tag{13}$$

Our notation for the elements of this matrix will be $\bar{h}_{i,j}$, i standing for the row and j for the column. For the diagonal elements $\bar{h}_{i,i}$ we will use the shorthand $\bar{h}_i$.

The following proposition can be deduced from Lemma 3 and the explicit form [analogous to Algorithm 1 below but using (22)] of the ordinary LSPM. The details of the proofs for all results of this section will be spelled out in Sect. 6.

Proposition 1 *The function Q_n output by the ordinary LSPM [see (2)] is monotonically increasing in y provided $\bar{h}_{n+1} < 0.5$.*

The condition needed for Q_n to be monotonically increasing, $\bar{h}_{n+1} < 0.5$, means that the test object x_{n+1} is not a very influential point. An overview of high-leverage points is given by Chatterjee and Hadi (1988, Section 4.2.3.1), where they start from Huber's 1981 proposal to regard points x_i with $\bar{h}_i > 0.2$ as influential.

The assumption $\bar{h}_{n+1} < 0.5$ in Proposition 1 is essential:

Proposition 2 *Proposition 1 ceases to be true if the constant 0.5 in it is replaced by a larger constant.*

The next two propositions show that for the deleted LSPM, determined by (12), the situation is even worse than for the ordinary LSPM: we have to require $\bar{h}_i < 0.5$ for all $i = 1, \ldots, n$.

Proposition 3 *The function Q_n output by the deleted LSPM according to* (2) *is monotonically increasing in y provided $\max_{i=1,\ldots,n} \bar{h}_i < 0.5$.*

We have the following analogue of Proposition 2 for the deleted LSPM.

Proposition 4 *Proposition 3 ceases to be true if the constant* 0.5 *in it is replaced by a larger constant.*

The best choice, from the point of view of predictive distributions, seems to be the *studentized LSPM* determined by the conformity measure

$$A(z_1, \ldots, z_{n+1}) := \frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} \tag{14}$$

(intermediate between those for the ordinary and deleted LSPM: a standard representation for the deleted residuals $y_i - \hat{y}_i$, where $\hat{y}_i$ is the prediction for y_i computed using $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{n+1}$ as training sequence, is $(y_i - \widehat{y}_i)/(1 - \bar{h}_i)$, $i = 1, \ldots, n + 1$; we ignore a factor independent of i in the definition of internally studentized residuals in, e.g., Seber and Lee 2003, Section 10.2).

An important advantage of studentized LSPM is that to get predictive distributions we do not need any assumptions of low leverage.

Proposition 5 *The studentized LSPM is an RPS and, therefore, a CPS.*

3.1 The studentized LSPM in an explicit form

We will give two explicit forms for the studentized LSPM (Algorithms 1 and 2); the versions for the ordinary and deleted LSPM are similar (we will give an explicit form only for the former, which is particularly intuitive). Predictive distributions (2) will be represented in the form

$$Q_n(y) := [Q_n(y, 0), Q_n(y, 1)]$$

(in the spirit of abstract randomized p values of Geyer and Meeden 2005); the function Q_n maps each potential label $y \in \mathbb{R}$ to a closed interval of $\mathbb{R}$. It is clear that in the case of conformal transducers this interval-valued version of Q_n carries the same information as the original one: each original value $Q_n(y, \tau)$ can be restored as a convex mixture of the end-points of $Q_n(y)$; namely, $Q_n(y, \tau) = (1 - \tau)a + \tau b$ if $Q_n(y) = [a, b]$.

Remember that the vector $(\widehat{y}_1, \ldots, \widehat{y}_{n+1})'$ of ordinary Least Squares predictions is the product of the hat matrix $\bar{H}$ and the vector $(y_1, \ldots, y_{n+1})'$ of labels. For the studentized residuals (14), we can easily obtain

$$\alpha_{n+1}^y - \alpha_i^y = B_i y - A_i, \quad i = 1, \ldots, n,$$

in the notation of (8), where y is the label of the $(n + 1)$st object x_{n+1} and

$$B_i := \sqrt{1 - \bar{h}_{n+1}} + \frac{\bar{h}_{i,n+1}}{\sqrt{1 - \bar{h}_i}}, \tag{15}$$

Algorithm 1 Least Squares Prediction Machine

Require: A training sequence $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$.
Require: A test object $x_{n+1} \in \mathbb{R}^p$.
1: Set $\bar{X}$ to the data matrix for the given $n + 1$ objects.
2: Define the hat matrix $\bar{H}$ by (13).
3: **for** $i \in \{1, 2, \ldots, n\}$ **do**
4:   Define A_i and B_i by (16) and (15), respectively.
5:   Set $C_i := A_i / B_i$.
6: **end for**
7: Sort $C_1, \ldots, C_n$ in the increasing order obtaining $C_{(1)} \leq \cdots \leq C_{(n)}$.
8: Return the predictive distribution (17) for y_{n+1}.

$$A_i = \frac{\sum_{j=1}^n \bar{h}_{j,n+1} y_j}{\sqrt{1 - \bar{h}_{n+1}}} + \frac{y_i - \sum_{j=1}^n \bar{h}_{i,j} y_j}{\sqrt{1 - \bar{h}_i}} \tag{16}$$

[see also (40) below]. We will assume that all B_i are defined and positive; this assumption will be discussed further at the end of this subsection.

Set $C_i := A_i / B_i$ for all $i = 1, \ldots, n$. Sort all C_i in the increasing order and let the resulting sequence be $C_{(1)} \leq \cdots \leq C_{(n)}$. Set $C_{(0)} := -\infty$ and $C_{(n+1)} := \infty$. The predictive distribution is:

$$Q_n(y) := \begin{cases} [\frac{i}{n+1}, \frac{i+1}{n+1}] & \text{if } y \in (C_{(i)}, C_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n\} \\ [\frac{i'-1}{n+1}, \frac{i''+1}{n+1}] & \text{if } y = C_{(i)} \text{ for } i \in \{1, \ldots, n\}, \end{cases} \tag{17}$$

where $i' := \min\{j \mid C_{(j)} = C_{(i)}\}$ and $i'' := \max\{j \mid C_{(j)} = C_{(i)}\}$. We can see that the thickness of this CPD is $\frac{1}{n+1}$ with the exception size equal to the number of distinct C_i, at most n.

The overall algorithm is summarized as Algorithm 1. Remember that the data matrix $\bar{X}$ has x_i', $i = 1, \ldots, n + 1$, as its ith row; its size is $(n + 1) \times p$.

Finally, let us discuss the condition that all B_i are defined and positive, $i = 1, \ldots, n$. By Chatterjee and Hadi (1988, Property 2.6(b)), $\bar{h}_{n+1} = 1$ implies $\bar{h}_{i,n+1} = 0$ for $i = 1, \ldots, n$; therefore, the condition is equivalent to $\bar{h}_i < 1$ for all $i = 1, \ldots, n + 1$. By Mohammadi (2016, Lemma 2.1(iii)), this means that the rank of the extended data matrix $\bar{X}$ is p and it remains p after removal of any one of its $n + 1$ rows. If this condition is not satisfied, we define $Q_n(y) := [0, 1]$ for all y. This ensures that the studentized LSPM is a CPS.

3.2 The batch version of the studentized LSPM

There is a much more efficient implementation of the LSPM in situations where we have a large test sequence of objects $x_{n+1}, \ldots, x_{n+m}$ instead of just one test object x_{n+1}. In this case we can precompute the hat matrix for the training objects $x_1, \ldots, x_n$, and then, when processing each test object x_{n+j}, use the standard updating formulas based on the Sherman–Morrison–Woodbury theorem: see, e.g., Chatterjee and Hadi (1988, p. 23, (2.18)–(2.18c)). For the reader's convenience we will spell out the formulas. Let X be the $n \times p$ data matrix for the first n observations: its ith row is x_i', $i = 1, \ldots, n$. Set

$$g_i := x_i'(X'X)^{-1} x_{n+1}, \quad i = 1, \ldots, n + 1. \tag{18}$$

Finally, let H be the $n \times n$ hat matrix

$$H := X(X'X)^{-1} X' \tag{19}$$

Algorithm 2 Least Squares Prediction Machine (batch version)

Require: A training sequence $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$.
Require: A test sequence $x_{n+j} \in \mathbb{R}^p$, $j = 1, \ldots, m$.
1: Set X to the data matrix for the n training objects.
2: Set $H = (h_{i,j})$ to the hat matrix (19).
3: **for** $j \in \{1, 2, \ldots, m\}$ **do**
4:     Set $x_{n+1} := x_{n+j}$.
5:     Define an $(n+1) \times (n+1)$ matrix $\bar{H} = (\bar{h}_{i,j})$ by (20) and (21).
6:     **for** $i \in \{1, 2, \ldots, n\}$ **do**
7:         Define A_i and B_i by (16) and (15), respectively.
8:         Set $C_i := A_i / B_i$.
9:     **end for**
10:     Sort $C_1, \ldots, C_n$ in the increasing order obtaining $C_{(1)} \leq \cdots \leq C_{(n)}$.
11:     Return the predictive distribution (17) for the label of x_{n+j}.
12: **end for**

---

for the first n objects; its entries will be denoted $h_{i,j}$, with $h_{i,i}$ sometimes abbreviated to h_i. The full hat matrix $\bar{H}$ is larger than H, with the extra entries

$$\bar{h}_{i,n+1} = \bar{h}_{n+1,i} = \frac{g_i}{1 + g_{n+1}}, \quad i = 1, \ldots, n+1. \tag{20}$$

The other entries of $\bar{H}$ are

$$\bar{h}_{i,j} = h_{i,j} - \frac{g_i g_j}{1 + g_{n+1}}, \quad i, j = 1, \ldots, n. \tag{21}$$

The overall algorithm is summarized as Algorithm 2. The two steps before the outer **for** loop are preprocessing; they do not depend on the test sequence.

3.3 The ordinary LSPM

A straightforward calculation shows that the ordinary LSPM has a particularly efficient and intuitive representation (Burnaev and Vovk 2014, Appendix A):

$$C_i = \frac{A_i}{B_i} = \hat{y}_{n+1} + (y_i - \hat{y}_i) \frac{1 + g_{n+1}}{1 + g_i}, \tag{22}$$

where $\hat{y}_{n+1}$ and $\hat{y}_i$ are the Least Squares predictions for y_{n+1} and y_i, respectively, computed from the test objects x_{n+1} and x_i, respectively, and the observations $z_1, \ldots, z_n$ as the training sequence. The representation (22) is stated and proved in Sect. 6 as Lemma 4. The predictive distribution is defined by (17). The fraction $\frac{1+g_{n+1}}{1+g_i}$ in (22) is typically and asymptotically (at least under the assumptions A1–A4 stated in the next section) close to 1, and can usually be ignored. The two other versions of the LSPM also typically have

$$C_i \approx \hat{y}_{n+1} + (y_i - \hat{y}_i). \tag{23}$$

4 A property of validity of the LSPM in the online mode

In the previous section (cf. Algorithm 1) we defined a procedure producing a "fuzzy" distribution function Q_n given a training sequence $z_i = (x_i, y_i)$, $i = 1, \ldots, n$, and a test object x_{n+1}. In this and following sections we will use both notation $Q_n(y)$ (for an interval) and $Q_n(y, \tau)$

(for a point inside that interval, as above). Remember that U is the uniform distribution on $[0, 1]$.

Prediction in the online mode proceeds as follows:

Protocol 1 ONLINE MODE OF PREDICTION

> Nature generates an observation $z_1 = (x_1, y_1)$
> > from a probability distribution P;
>
> **for** $n = 1, 2, \ldots$ **do**
> > Nature independently generates a new observation
> > > $z_{n+1} = (x_{n+1}, y_{n+1})$ from P;
> >
> > Forecaster announces Q_n, a predictive distribution
> > > based on $(z_1, \ldots, z_n)$ and x_{n+1};
> >
> > set $p_n := Q_n(y_{n+1}, \tau_n)$, where $\tau_n \sim U$ independently
>
> **end for**

Of course, Forecaster does not know P and y_{n+1} when computing Q_n.

In the online mode we can strengthen condition R2 as follows:

Theorem 1 (Vovk et al. 2005, Theorem 8.1) *In the online mode of prediction (in which $(z_i, \tau_i) \sim P \times U$ are IID), the sequence $(p_1, p_2, \ldots)$ is IID and $(p_1, p_2, \ldots) \sim U^{\infty}$, provided that Forecaster uses the studentized LSPM (or any other conformal transducer).*

The property of validity asserted in Theorem 1 is marginal, in that we do not assert that the distribution of p_n is uniform conditionally on x_{n+1}. Conditional validity is attained by the LSPM only asymptotically and under additional assumptions, as we will see in the next section.

5 Asymptotic efficiency

In this section we obtain some basic results about the LSPM's efficiency. The LSPM has a property of validity under the general IID model, but a natural question is how much we should pay for it in terms of efficiency in situations where narrow parametric or even Bayesian assumptions are also satisfied. This question was asked independently by Evgeny Burnaev (in September 2013) and Larry Wasserman. It has an analogue in nonparametric hypothesis testing: e.g., a major impetus for the widespread use of the Wilcoxon rank-sum test was Pitman's discovery in 1949 that even in the situation where the Gaussian assumptions of Student's t-test are satisfied the efficiency ("Pitman's efficiency") of the Wilcoxon test is still 0.95.

In fact the assumptions that we use in our theoretical study of efficiency are not comparable with the general IID model used so far: we will add strong parametric assumptions on the way labels y_i are generated given the corresponding objects x_i but will remove the assumption that the objects are generated randomly in the IID fashion; in this section $x_1, x_2, \ldots$ are fixed vectors. (The reason being that the two main results of this section, Theorems 2 and 3, do not require the assumption that the objects are random and IID.) Suppose that, given the objects $x_1, x_2, \ldots$, the labels $y_1, y_2, \ldots$ are generated by the rule

$$y_i = w'x_i + \xi_i, \tag{24}$$

where w is a vector in $\mathbb{R}^p$ and ξ_i are independent random variables distributed as $N(0, \sigma^2)$ (the Gaussian distribution being parameterized by its mean and variance). There are two

parameters: vector w and positive number σ. We assume an infinite sequence of observations $(x_1, y_1), (x_2, y_2), \ldots$ but take only the first n of them as our training sequence and let $n \to \infty$. These are all the assumptions used in our efficiency results:

A1 The sequence $x_1, x_2, \ldots$ is bounded: $\sup_i \|x_i\| < \infty$.
A2 The first component of each vector x_i is 1.
A3 The empirical second-moment matrix has its smallest eigenvalue eventually bounded away from 0:
$$\liminf_{n \to \infty} \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right) > 0,$$
where $\lambda_{\min}$ stands for the smallest eigenvalue.
A4 The labels $y_1, y_2, \ldots$ are generated according to (24): $y_i = w' x_i + \xi_i$, where ξ_i are independent Gaussian noise random variables distributed as $N(0, \sigma^2)$.

Alongside the three versions of the LSPM, we will consider three "oracles" (at first concentrating on the first two). Intuitively, all three oracles know that the data is generated from the model (24). Oracle I knows neither w nor σ (and has to estimate them from the data or somehow manage without them). Oracle II does not know w but knows σ. Finally, Oracle III knows both w and σ.

Formally, *proper Oracle I* outputs the standard predictive distribution for the label y_{n+1} of the test object x_{n+1} given the training sequence of the first n observations and x_{n+1}, namely it predicts with

$$\hat{y}_{n+1} + \sqrt{1 + g_{n+1}} \hat{\sigma}_n t_{n-p}, \tag{25}$$

where g_{n+1} is defined in (18),

$$\hat{y}_{n+1} := x_{n+1}' (X'X)^{-1} X'Y,$$

$$\hat{\sigma}_n := \sqrt{\frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \quad \hat{y}_i := x_i' (X'X)^{-1} X'Y,$$

X is the data matrix for the training sequence (the $n \times p$ matrix whose ith row is x_i', $i = 1, \ldots, n$), Y is the vector $(y_1, \ldots, y_n)'$ of the training labels, and t_{n-p} is Student's t-distribution with $n - p$ degrees of freedom; see, e.g., Seber and Lee (2003, Section 5.3.1) or Wang et al (2012, Example 3.3). (By condition A3, $(X'X)^{-1}$ exists from some n on.) The version that is more popular in the literature on empirical processes for residuals is *simplified Oracle I* outputting

$$N\left(\hat{y}_{n+1}, \hat{\sigma}_n^2\right). \tag{26}$$

The difference between the two versions, however, is asymptotically negligible (Pinelis 2015), and the results stated below will be applicable to both versions.

Proper Oracle II outputs the predictive distribution

$$N\left(\hat{y}_{n+1}, (1 + g_{n+1})\sigma^2\right). \tag{27}$$

Correspondingly, *simplified Oracle II* outputs the predictive distribution

$$N\left(\hat{y}_{n+1}, \sigma^2\right); \tag{28}$$

the difference between the two versions of Oracle II is again asymptotically negligible under our assumptions. For future reference, *Oracle III* outputs the predictive distribution

$$N\left(w' x_{n+1}, \sigma^2\right).$$

Our notation is Q_n for the conformal predictive distribution (2), as before, Q_n^{I} for simplified or proper Oracle I's predictive distribution, (26) or (25) (Theorem 2 will hold for both), and Q_n^{II} for simplified or proper Oracle II's predictive distribution, (28) or (27) (Theorem 3 will hold for both). Theorems 2 and 3 are applicable to all three versions of the LSPM.

Theorem 2 *The random functions $G_n : \mathbb{R} \to \mathbb{R}$ defined by*

$$G_n(t) := \sqrt{n} \left(Q_n(\hat{y}_{n+1} + \hat{\sigma}_n t, \tau) - Q_n^{\mathrm{I}}(\hat{y}_{n+1} + \hat{\sigma}_n t) \right)$$

weakly converge to a Gaussian process Z with mean zero and covariance function

$$\mathrm{cov}(Z(s), Z(t)) = \Phi(s)\left(1 - \Phi(t)\right) - \phi(s)\phi(t) - \frac{1}{2} s t \phi(s)\phi(t), \quad s \leq t.$$

Theorem 3 *The random functions $G_n : \mathbb{R} \to \mathbb{R}$ defined by*

$$G_n(t) := \sqrt{n} \left(Q_n(\hat{y}_{n+1} + \sigma t, \tau) - Q_n^{\mathrm{II}}(\hat{y}_{n+1} + \sigma t) \right)$$

weakly converge to a Gaussian process Z with mean zero and covariance function

$$\mathrm{cov}(Z(s), Z(t)) = \Phi(s)\left(1 - \Phi(t)\right) - \phi(s)\phi(t), \quad s \leq t. \tag{29}$$

In Theorems 2 and 3, we have $\tau \sim U$; alternatively, they will remain true if we fix τ to any value in [0, 1]. For simplified oracles, we have $Q_n^{\mathrm{I}}(\hat{y}_{n+1} + \hat{\sigma}_n t) = \Phi(t)$ in Theorem 2 and $Q_n^{\mathrm{II}}(\hat{y}_{n+1} + \sigma t) = \Phi(t)$ in Theorem 3. Our proofs of these theorems (given in Sect. 6) are based on the representation (22) and the results of Mugantseva (1977) (see also Chen 1991, Chapter 2).

Applying Theorems 2 and 3 to a fixed argument t, we obtain (dropping τ altogether):

Corollary 1 *For a fixed $t \in \mathbb{R}$,*

$$\sqrt{n} \left(Q_n(\hat{y}_{n+1} + \hat{\sigma}_n t) - Q_n^{\mathrm{I}}(\hat{y}_{n+1} + \hat{\sigma}_n t) \right)$$
$$\Rightarrow N \left(0, \Phi(t)(1 - \Phi(t)) - \phi(t)^2 - \frac{1}{2} t^2 \phi(t)^2 \right)$$

and

$$\sqrt{n} \left(Q_n(\hat{y}_{n+1} + \sigma t) - Q_n^{\mathrm{II}}(\hat{y}_{n+1} + \sigma t) \right) \Rightarrow N \left(0, \Phi(t)(1 - \Phi(t)) - \phi(t)^2 \right).$$

Figure 1 presents plots for the asymptotic variances, given in Corollary 1, for the two oracular predictive distributions: black for Oracle I ($\Phi(t)(1 - \Phi(t)) - \phi(t)^2 - \frac{1}{2} t^2 \phi(t)^2$ vs t) and blue for Oracle II ($\Phi(t)(1 - \Phi(t)) - \phi(t)^2$ vs t); the red plot will be discussed later in this section. The two asymptotic variances coincide at $t = 0$, where they attain their maximum of between 0.0908 and 0.0909.

We can see that under the Gaussian model (24) complemented by other natural assumptions, the LSPM is asymptotically close to the oracular predictive distributions for Oracles I and II, and therefore is approximately conditionally valid and efficient (i.e., valid and efficient given $x_1, x_2, \ldots$). On the other hand, Theorem 1 guarantees the marginal validity of the LSPM under the general IID model, regardless of whether (24) holds.
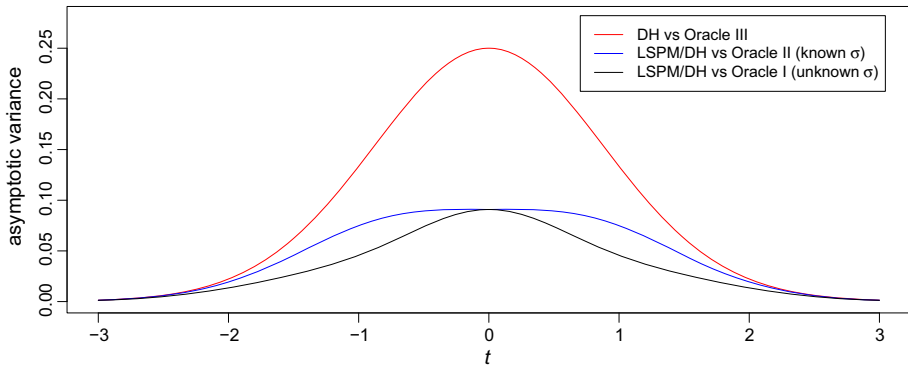
Fig. 1 The asymptotic variances for the Dempster–Hill (DH) procedure as compared with the truth (Oracle III, red) and for the LSPM and DH procedure as compared with the oracular procedures for known σ (Oracle II, blue) and unknown σ (Oracle I, black); in black and white, red is highest, blue is intermediate, and black is lowest

5.1 Comparison with the Dempster–Hill procedure

In this subsection we discuss a classical procedure that was most clearly articulated by Dempster (1963, p. 110) and Hill (1968, 1988); therefore, in this paper we refer to it as the *Dempster–Hill procedure*. Both Dempster and Hill trace their ideas to Fisher's (1939; 1948) nonparametric version of his fiducial method, but Fisher was interested in confidence distributions for quantiles rather than predictive distributions. Hill (1988) also referred to his procedure as Bayesian nonparametric predictive inference, which was abbreviated to nonparametric predictive inference (NPI) by Frank Coolen (Augustin and Coolen 2004). We are not using the last term since it seems that all of this paper (and the whole area of conformal prediction) falls under the rubric of "nonparametric predictive inference". An important predecessor of Dempster and Hill was Jeffreys (1932), who postulated what Hill later denoted as $A_{(2)}$ (see Lane 1980 and Seidenfeld 1995 for discussions of Jeffreys's paper and Fisher's reaction).

The Dempster–Hill procedure is the conformal predictive system determined by the conformity measure

$$A(z_1, \ldots, z_{n+1}) = A(y_1, \ldots, y_{n+1}) = y_{n+1}; \tag{30}$$

it is used when the objects x_i are absent. (Both Dempster and Hill consider this case.) It can be regarded as the special case of the LSPM for the number of attributes $p = 0$; alternatively, we can take $p = 1$ but assume that all objects are $x_i = 0$. The predictions $\hat{y}$ are always 0 and the hat matrices are $\bar{H} = 0$ and $H = 0$ (although the expressions (13) and (19) are not formally applicable), which means that (11), (12), and (14) all reduce to (30). It is easy to see that the predictive distribution becomes, in the absence of ties (Dempster's and Hill's usual assumption),

$$Q_n(y) := \begin{cases} [\frac{i}{n+1}, \frac{i+1}{n+1}] & \text{if } y \in (y_{(i)}, y_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n\} \\ [\frac{i-1}{n+1}, \frac{i+1}{n+1}] & \text{if } y = y_{(i)} \text{ for } i \in \{1, \ldots, n\} \end{cases} \tag{31}$$

(cf. (17)), where $y_{(1)} \leq \cdots \leq y_{(n)}$ are the y_i sorted in the increasing order, $y_{(0)} := -\infty$, and $y_{(n+1)} := \infty$. This is essentially Hill's assumption $A_{(n)}$ (which he also denoted A_n); in his words: "A_n asserts that conditional upon the observations $X_1, \ldots, X_n$, the next observation

X_{n+1} is equally likely to fall in any of the open intervals between successive order statistics of the given sample" (Hill 1968, Section 1). The set of all continuous distribution functions F compatible with Hill's $A_{(n)}$ coincides with the set of all continuous distribution functions F satisfying $F(y) \in Q_n(y)$ for all $y \in \mathbb{R}$, where Q_n is defined by (31).

Notice that the LSPM, as presented in (23), is a very natural adaptation of $A_{(n)}$ to the Least Squares regression.

Since (31) is a conformal transducer (provided a point from an interval in (31) is chosen randomly from the uniform distribution on that interval), we have the same guarantees of validity as those given above: the distribution of (31) is uniform over the interval [0, 1].

As for efficiency, it is interesting that, in the most standard case of IID Gaussian observations, our predictive distributions for linear regression are as precise as the Dempster–Hill ones asymptotically when compared with Oracles I and II. Let us apply the Dempster–Hill procedure to the location/scale model $y_i = w + \xi_i$, $i = 1, 2, \ldots$, where $\xi_i \sim N(0, \sigma^2)$ are independent. As in the case of the LSPM, we can compare the Dempster–Hill procedure with three oracles (we consider only simplified versions): Oracle I knows neither w nor σ, Oracle II knows σ, and Oracle III knows both w and σ.

It is interesting that Theorems 2 and 3 (and therefore the blue and black plots in Fig. 1) are applicable to both the LSPM and Dempster–Hill predictive distributions. (The fact that the analogous asymptotic variances for standard linear regression are as good as those for the location/scale model was emphasized in the pioneering paper by Pierce and Kopecky 1979.) The situation with Oracle III is different. Donsker's (1952) classical result implies the following simplification of Theorems 2 and 3, where Q^{III} stands for Oracle III's predictive distribution (independent of n).

Theorem 4 *In the case of the Dempster–Hill procedure, the random function $G_n : \mathbb{R} \to \mathbb{R}$ defined by*

$$G_n(t) := \sqrt{n} \left(Q_n(w + \sigma t, \tau) - Q^{III}(w + \sigma t) \right) = \sqrt{n} \left(Q_n(w + \sigma t, \tau) - \Phi(t) \right) \quad (32)$$

weakly converges to a Brownian bridge, i.e., a Gaussian process Z with mean zero and covariance function

$$\mathrm{cov}(Z(s), Z(t)) = \Phi(s)(1 - \Phi(t)), \quad s \le t.$$

The variance $\Phi(t)(1 - \Phi(t))$ of the Brownian bridge is shown as the red line in Fig. 1. However, the analogue of the process (32) does not converge in general for the LSPM (under this section's assumption of fixed objects).

6 Proofs, calculations, and additional observations

In this section we give all proofs and calculations for the results of the previous sections and provide some additional comments.

6.1 Proofs for Sect. 2

Here we prove Lemmas 1–3.

Proof of Lemma 1

Suppose there is $y \in \mathbb{R}$ such that $Q(y) \neq F(y)$. Fix such a y. The probability that $Q(Y) \leq Q(y)$ is, on the one hand, $Q(y)$ and, on the other hand, $F(y')$, where

$$y' := \sup\{y'' \mid Q(y'') = Q(y)\}.$$

(The first statement follows from the distribution of Y being uniform and the second from the definition of F in conjunction with its continuity.) Since $Q(y) \neq F(y)$, we have $y' > y$, and we know that $Q(y) = Q(y'-) = F(y') > F(y)$. We can see that Q maps the whole interval $[y, y')$ of positive probability $F(y') - F(y)$ to one point, which contradicts its distribution being uniform.

Proof of Lemma 2

First we prove that $Q(y, 1) = F(y)$ for all $y \in \mathbb{R}$. Fix a $y \in \mathbb{R}$ such that $Q(y, 1) \neq F(y)$, assuming it exists. Set

$$y' := \sup\{y'' \mid Q(y'', 1) = Q(y, 1)\}. \tag{33}$$

Since $Q(y, 1) \neq F(y)$ and, for $(Y, \tau) \sim P \times U$,

$$Q(y, 1) = \mathbb{P}(Q(Y, \tau) \leq Q(y, 1)) \geq \mathbb{P}(Q(Y, 1) \leq Q(y, 1))$$
$$\geq \mathbb{P}((Y, 1) \leq (y, 1)) = \mathbb{P}(Y \leq y) = F(y),$$

we have $Q(y, 1) > F(y)$. Next we consider two cases:

– if the supremum in (33) is attained,

$$F(y) < Q(y, 1) = \mathbb{P}(Q(Y, 1) \leq Q(y, 1)) = \mathbb{P}((Y, 1) \leq (y', 1)) = F(y'),$$

and so Q maps the lexicographic interval $((y, 1), (y', 1)]$ of positive probability $F(y') - F(y)$ into one point;

– if the supremum in (33) is not attained,

$$F(y) < Q(y, 1) = \mathbb{P}(Q(Y, 1) \leq Q(y, 1)) = \mathbb{P}((Y, 1) < (y', 1)) = F(y'-),$$

and so Q maps the lexicographic interval $((y, 1), (y', 0))$ of positive probability $F(y'-) - F(y)$ into one point.

In both cases we get a contradiction with the distribution of Q being uniform, which completes the proof that $Q(y, 1) = F(y)$ for all $y \in \mathbb{R}$.

In the same way we prove that $Q(y, 0) = F(y-)$ for all $y \in \mathbb{R}$.

Now (10) holds trivially when F is continuous at y. If it is not, $Q^{-1}((F(y-), F(y))$ will only contain points (y, τ) for various τ, and so (10) is the only way to ensure that the distribution of Q is uniform.

Proof of Lemma 3

Let us split all numbers $i \in \{1, \ldots, n+1\}$ into three classes: i of class I are those satisfying $\alpha_i^y > \alpha_{n+1}^y$, i of class II are those satisfying $\alpha_i^y = \alpha_{n+1}^y$, and i of class III are those satisfying $\alpha_i^y < \alpha_{n+1}^y$. Each of those numbers is assigned a *weight*: 0 for i of class I, $\tau/(n+1)$ for i of class II, and $1/(n+1)$ for i of class III; notice that the weights are larger for higher-numbered classes. According to (7), $Q_n(y, \tau)$ is the sum of the weights of all $i \in \{1, \ldots, n+1\}$. As y increases, each individual weight can only increase (as i can move only to a higher-numbered class), and so the total weight $Q_n(y, \tau)$ can also only increase.

6.2 Comments and proofs for Sect. 3

After a brief discussion of Ridge Regression Prediction Machines (analogous to Ridge Regression Confidence Machines, mentioned at the beginning of Sect. 3), we prove Propositions 1–5 and find the explicit forms for the studentized, ordinary, and deleted LSPM.

Ridge regression prediction machines

We can generalize LSPM to the *Ridge Regression Prediction Machine* (RRPM) by replacing the Least Squares predictions in (11), (12), and (14) by Ridge Regression predictions (see Vovk et al. 2017a for details). In this paper we are interested in the case $p \ll n$, and so Least Squares often provide a reasonable result as compared with Ridge Regression. When we move on to the kernel case (and Kernel Ridge Regression), the Least Squares method ceases to be competitive. Vovk et al. (2017a) extend some results of this paper to the kernel case replacing the LSPM by the RRPM.

Remark 2 The early versions of the Ridge Regression Confidence Machines used $\left| y_{n+1} - \widehat{y}_{n+1} \right|$ in place of the right-hand side of (11) (see, e.g., Vovk et al. 2005, Section 2.3). For the first time the operation $|\cdots|$ of taking the absolute value was dropped in Burnaev and Vovk (2014) to facilitate theoretical analysis.

Proof of Proposition 1

According to Lemma 3, $Q_n(y, \tau)$ will be monotonically increasing in y if $\alpha_{n+1}^y - \alpha_i^y$ is a monotonically increasing function of y. We will use the notation $e_i := y_i - \hat{y}_i$ (suppressing the dependence on y) for the ith residual, $i = 1, \ldots, n+1$, in the data sequence $z_1, \ldots, z_n, (x_{n+1}, y)$; y_{n+1} is understood to be y. In terms of the hat matrix $\bar{H}$ (which does not depend on the labels), the difference $e_{n+1} - e_i$ can be written as

$$
\begin{aligned}
\alpha_{n+1}^y - \alpha_i^y &= e_{n+1} - e_i \\
&= (y_{n+1} - \hat{y}_{n+1}) - (y_i - \hat{y}_i) \\
&= y - \hat{y}_{n+1} + \hat{y}_i + c \\
&= y - (\bar{h}_{n+1,1} y_1 + \cdots + \bar{h}_{n+1,n} y_n + \bar{h}_{n+1} y) \\
&\quad + (\bar{h}_{i,1} y_1 + \cdots + \bar{h}_{i,n} y_n + \bar{h}_{i,n+1} y) + c \\
&= (1 - \bar{h}_{n+1} + \bar{h}_{i,n+1}) y + c,
\end{aligned}
\tag{34}
$$

where c stands for a constant (in the sense of not depending on y), and different entries of c may stand for different constants. We can see that Q_n will be a nontrivial monotonically increasing function of y whenever

$$
1 - \bar{h}_{n+1} + \bar{h}_{i,n+1} > 0
\tag{35}
$$

for all $i = 1, \ldots, n$. Since $\bar{h}_{i,n+1} \in [-0.5, 0.5]$ (see Chatterjee and Hadi 1988, Property 2.5(b) on p. 17), we can see that it indeed suffices to assume $\bar{h}_{n+1} < 0.5$.

Proof of Proposition 2

We are required to show that our $c = 0.5$ is the largest c for which the assumption $\bar{h}_{n+1} < c$ is still sufficient for $Q_n(y, \tau)$ to be a monotonically increasing function of y. For $\epsilon \in (0, 1)$,

consider the data set

$$\bar{X} = \begin{pmatrix} -1 + \epsilon \\ 1 \end{pmatrix} \tag{36}$$

(so that $n = 1$; we have two observations: one training observation and one test observation). The hat matrix is

$$\bar{H} = \frac{1}{2 - 2\epsilon + \epsilon^2} \begin{pmatrix} (1-\epsilon)^2 & -1+\epsilon \\ -1+\epsilon & 1 \end{pmatrix}.$$

The coefficient in front of y in the last line of (34) [i.e., the left-hand side of (35)] now becomes

$$1 - \frac{1}{2 - 2\epsilon + \epsilon^2} + \frac{-1+\epsilon}{2 - 2\epsilon + \epsilon^2} = \frac{\epsilon^2 - \epsilon}{2 - 2\epsilon + \epsilon^2} < 0.$$

Therefore, $Q_n(\cdot, \tau)$ is monotonically decreasing and not monotonically increasing. On the other hand,

$$\bar{h}_{n+1} = \bar{h}_2 = \frac{1}{2 - 2\epsilon + \epsilon^2}$$

can be made as close to 0.5 as we wish by making ϵ sufficiently small.

Proof of Proposition 3

Let $e_{(i)}$ be the *deleted residual*: $e_{(i)} := y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is computed using Least Squares from the data set $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{n+1}$ (so that z_i is deleted from $z_1, \ldots, z_{n+1}$, where we set temporarily $z_{n+1} := (x_n, y)$). It is well known that

$$e_{(i)} = \frac{e_i}{1 - \bar{h}_i},$$

where e_i is the ordinary residual, as used in the proof of Proposition 1 (for a proof, see, e.g., Montgomery et al. 2012, Appendix C.7). Let us check when the difference $e_{(n+1)} - e_{(i)}$ is a monotonically increasing function of $y = y_{n+1}$. Analogously to (34), we have, for any $i = 1, \ldots, n$:

$$\begin{aligned} e_{(n+1)} - e_{(i)} &= \frac{e_{n+1}}{1 - \bar{h}_{n+1}} - \frac{e_i}{1 - \bar{h}_i} \\ &= \frac{y_{n+1} - \hat{y}_{n+1}}{1 - \bar{h}_{n+1}} - \frac{y_i - \hat{y}_i}{1 - \bar{h}_i} \\ &= \frac{y - \bar{h}_{n+1}y}{1 - \bar{h}_{n+1}} - \frac{y_i - \bar{h}_{i,n+1}y}{1 - \bar{h}_i} + c \\ &= y - \frac{y_i - \bar{h}_{i,n+1}y}{1 - \bar{h}_i} + c \\ &= y \frac{1 - \bar{h}_i + \bar{h}_{i,n+1}}{1 - \bar{h}_i} + c. \end{aligned} \tag{37}$$

Therefore, it suffices to require

$$1 - \bar{h}_i + \bar{h}_{i,n+1} > 0, \tag{38}$$

which is the same condition as for the ordinary LSPM [see (35)] but with i and $n+1$ swapped. Therefore, it suffices to assume $\bar{h}_i < 0.5$.

Proof of Proposition 4

The statement of the proposition is obvious from the proofs of Propositions 2 and 3: motivated by the conditions (35) and (38) being obtainable from each other by swapping i and $n + 1$, we can apply the argument in the proof of Proposition 2 to the data set

$$\bar{X} = \begin{pmatrix} 1 \\ -1 + \epsilon \end{pmatrix}$$

(which is (36) with its rows swapped).

Proof of Proposition 5

Similarly to (34) and (37), we obtain:

$$
\begin{aligned}
\alpha^y_{n+1} - \alpha^y_i &= \frac{e_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{e_i}{\sqrt{1 - \bar{h}_i}} \\
&= \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{y_i - \hat{y}_i}{\sqrt{1 - \bar{h}_i}} \\
&= \frac{y - \bar{h}_{n+1} y}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{y_i - \bar{h}_{i,n+1} y}{\sqrt{1 - \bar{h}_i}} + c \\
&= \sqrt{1 - \bar{h}_{n+1}}\, y + \frac{\bar{h}_{i,n+1}}{\sqrt{1 - \bar{h}_i}}\, y + c.
\end{aligned}
\tag{39}
$$

Therefore, we need to check the inequality

$$\sqrt{1 - \bar{h}_{n+1}} + \frac{\bar{h}_{i,n+1}}{\sqrt{1 - \bar{h}_i}} \geq 0.$$

This inequality can be rewritten as

$$\bar{h}_{i,n+1} \geq -\sqrt{(1 - \bar{h}_{n+1})(1 - \bar{h}_i)}$$

and follows from Chatterjee and Hadi (1988), Property 2.6(b) on p. 19.

Computations for the studentized LSPM

Now we need the chain (39) with a more careful treatment of the unspecified constants c:

$$
\begin{aligned}
\alpha_{n+1} - \alpha_i &= \frac{e_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{e_i}{\sqrt{1 - \bar{h}_i}} \\
&= \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{y_i - \hat{y}_i}{\sqrt{1 - \bar{h}_i}} \\
&= \frac{y - \sum_{j=1}^{n} \bar{h}_{j,n+1} y_j - \bar{h}_{n+1} y}{\sqrt{1 - \bar{h}_{n+1}}} - \frac{y_i - \sum_{j=1}^{n} \bar{h}_{i,j} y_j - \bar{h}_{i,n+1} y}{\sqrt{1 - \bar{h}_i}} \\
&= \left(\sqrt{1 - \bar{h}_{n+1}} + \frac{\bar{h}_{i,n+1}}{\sqrt{1 - \bar{h}_i}} \right) y
\end{aligned}
$$

$$-\left(\frac{\sum_{j=1}^{n}\bar{h}_{j,n+1}y_j}{\sqrt{1-\bar{h}_{n+1}}}+\frac{y_i-\sum_{j=1}^{n}\bar{h}_{i,j}y_j}{\sqrt{1-\bar{h}_i}}\right)$$

$$=B_i y - A_i,\tag{40}$$

where the last equality is just the definition of B_i and A_i, also given by (15) and (16) above.

The ordinary and deleted LSPM

Here we will do the analogues of the calculation (40) for the ordinary and deleted LSPM. For the ordinary LSPM we obtain

$$\alpha_{n+1} - \alpha_i = e_{n+1} - e_i$$
$$= (y_{n+1} - \hat{y}_{n+1}) - (y_i - \hat{y}_i)$$
$$= \left(y - \sum_{j=1}^{n}\bar{h}_{j,n+1}y_j - \bar{h}_{n+1}y\right) - \left(y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j - \bar{h}_{i,n+1}y\right)$$
$$= \left(1 - \bar{h}_{n+1} + \bar{h}_{i,n+1}\right)y$$
$$\quad - \left(\sum_{j=1}^{n}\bar{h}_{j,n+1}y_j + y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j\right)$$
$$= B_i y - A_i,$$

with the notation

$$B_i := 1 - \bar{h}_{n+1} + \bar{h}_{i,n+1},\tag{41}$$

$$A_i = \sum_{j=1}^{n}\bar{h}_{j,n+1}y_j + y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j.\tag{42}$$

For the deleted LSPM the calculation (40) becomes:

$$\alpha_{n+1} - \alpha_i = \frac{e_{n+1}}{1 - \bar{h}_{n+1}} - \frac{e_i}{1 - \bar{h}_i}$$
$$= \frac{y_{n+1} - \hat{y}_{n+1}}{1 - \bar{h}_{n+1}} - \frac{y_i - \hat{y}_i}{1 - \bar{h}_i}$$
$$= \frac{y - \sum_{j=1}^{n}\bar{h}_{j,n+1}y_j - \bar{h}_{n+1}y}{1 - \bar{h}_{n+1}} - \frac{y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j - \bar{h}_{i,n+1}y}{1 - \bar{h}_i}$$
$$= \left(1 + \frac{\bar{h}_{i,n+1}}{1 - \bar{h}_i}\right)y - \left(\frac{\sum_{j=1}^{n}\bar{h}_{j,n+1}y_j}{1 - \bar{h}_{n+1}} + \frac{y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j}{1 - \bar{h}_i}\right)$$
$$= B_i y - A_i,$$

with the notation

$$B_i := 1 + \frac{\bar{h}_{i,n+1}}{1 - \bar{h}_i},$$

$$A_i = \frac{\sum_{j=1}^{n}\bar{h}_{j,n+1}y_j}{1 - \bar{h}_{n+1}} + \frac{y_i - \sum_{j=1}^{n}\bar{h}_{i,j}y_j}{1 - \bar{h}_i}.$$

6.3 Comments and proofs for Sect. 5

There are different notions of weak convergence of empirical processes used in literature, but in this paper (in particular, Theorems 2 and 3) we use the old-fashioned one due to Skorokhod: see, e.g., Billingsley (1999, except for Section 15). We will regard empirical distribution functions and empirical processes to be elements of a function space which we will denote $\mathbb{D}$: its elements are càdlàg (i.e., right-continuous with left limits) functions $f : \mathbb{R} \to \mathbb{R}$, and the distance between $f, g \in \mathbb{D}$ will be defined to be the Skorokhod distance (either d or d° in the notation of Billingsley 1999, Theorem 12.1) between the functions $t \in [0, 1] \mapsto f(\Phi^{-1}(t))$ and $t \in [0, 1] \mapsto g(\Phi^{-1}(t))$ in $D[0, 1]$. (Here Φ is the standard Gaussian distribution function; we could have used any other function on the real line that is strictly monotonically increasing from 0 to 1.)

Proofs of Theorems 2 and 3 for the ordinary LSPM

We will start our proof from the ordinary LSPM, in which case the predictive distribution is particularly simple.

Lemma 4 (Burnaev and Vovk 2014) *In the case of the ordinary LSPM, we have* (22).

Proof Remember that, in our notation, X is the data matrix based on the first n observations, $\bar{X}$ is the data matrix based on the first $n + 1$ observations, H is the hat matrix $X(X'X)^{-1}X'$ based on the first n observations, and $\bar{H}$ is the hat matrix $\bar{X}(\bar{X}'\bar{X})^{-1}X'$ based on the first $n+1$ observations; the elements of H are denoted as $h_{i,j}$ and the elements of $\bar{H}$ as $\bar{h}_{i,j}$, except that for the diagonal elements we do not repeat the index. Besides, we let $Y \in \mathbb{R}^n$ stand for the vector of the training labels $(y_1, \ldots, y_n)'$. To compute C_i we will use the formulas (41), (42) and (20), (21):

$$
\begin{aligned}
B_i &= 1 - \bar{h}_{n+1} + \bar{h}_{i,n+1} \\
&= 1 - \frac{x'_{n+1}(X'X)^{-1}x_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} + \frac{x'_i(X'X)^{-1}x_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} \\
&= \frac{1 + x'_i(X'X)^{-1}x_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}.
\end{aligned}
$$

and, letting $\hat{y}$ stand for the predictions computed from the first n observations,

$$
\begin{aligned}
A_i &= y_i - \sum_{j=1}^{n} \bar{h}_{i,j} y_j + \sum_{j=1}^{n} \bar{h}_{j,n+1} y_j \\
&= y_i - \sum_{j=1}^{n} h_{i,j} y_j + \sum_{j=1}^{n} \frac{x'_i(X'X)^{-1}x_{n+1}x'_{n+1}(X'X)^{-1}x_j}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} y_j \\
&\quad + \sum_{j=1}^{n} \frac{x'_j(X'X)^{-1}x_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} y_j \\
&= y_i - \hat{y}_i + \frac{x'_i(X'X)^{-1}x_{n+1}x'_{n+1}(X'X)^{-1}X'Y}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} + \frac{Y'X(X'X)^{-1}x_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} \\
&= y_i - \hat{y}_i + \frac{x'_i(X'X)^{-1}x_{n+1}\hat{y}_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} + \frac{\hat{y}_{n+1}}{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}
\end{aligned}
$$

$$= y_i - \hat{y}_i + \frac{1 + x_i'(X'X)^{-1}x_{n+1}}{1 + x_{n+1}'(X'X)^{-1}x_{n+1}}\hat{y}_{n+1}.$$

This gives

$$C_i = A_i/B_i = (y_i - \hat{y}_i)\frac{1 + x_{n+1}'(X'X)^{-1}x_{n+1}}{1 + x_i'(X'X)^{-1}x_{n+1}} + \hat{y}_{n+1},$$

i.e., (22).   □

Now Theorems 2 and 3 will follow from Mugantseva (1977) and Chen (1991). Mugantseva only treats simple linear regression, and in general we deduce Theorem 2 from Chen (1991, Theorem 2.4.3) and deduce Theorem 3 from Chen's Theorems 2.4.3 and 2.3.2. However, to make those results applicable we need to show that the fraction $\frac{1+g_{n+1}}{1+g_i}$ in (22) can be ignored; the following lemma shows that both g_{n+1} and g_i are sufficiently close to 1.

Lemma 5 *Under our conditions A1–A4, $\max_{i=1,\ldots,n+1} |g_i| = O(n^{-1}) = o(n^{-1/2})$.*

Proof We have, for all allowed sequences $x_1, x_2, \ldots,$

$$\max_{i=1,\ldots,n+1} |g_i| \leq \frac{\|x_{n+1}\| \max_{i=1,\ldots,n+1} \|x_i\|}{\lambda_{\min}(X'X)}$$

$$< \frac{\|x_{n+1}\| \max_{i=1,\ldots,n+1} \|x_i\|}{n\epsilon} = O(n^{-1}),$$

with the inequality holding for some $\epsilon > 0$ from some n on.   □

We will spell out the details of the proof only for Theorem 3. Since Q_n is concentrated at the points $C_1, \ldots, C_n$, its transformation $t \mapsto Q_n(\hat{y}_{n+1} + \sigma t)$ is concentrated at

$$t_i = \frac{C_i - \hat{y}_{n+1}}{\sigma} = \frac{y_i - \hat{y}_i}{\sigma}\frac{1 + g_{n+1}}{1 + g_i} \tag{43}$$

[by (22)]. If we replace (43) by

$$\bar{t}_i = \frac{y_i - \hat{y}_i}{\sigma}$$

the desired result reduces to Mugantseva's (as presented by Chen 1991), so we need to check that this replacement is valid. We will use the fact that, by Lemma 5,

$$\frac{1 + g_{n+1}}{1 + g_i} = 1 + o(n^{-1/2})$$

as $n \to \infty$ uniformly in $i = 1, \ldots, n$. Let F_n be the empirical distribution function determined by the random points $t_1, \ldots, t_n$ and $\bar{F}_n$ be the empirical distribution function determined by the random points $\bar{t}_1, \ldots, \bar{t}_n$. Let

$$G_n(t) := n^{1/2}(F_n(t) - \Phi(t))$$

$$\bar{G}_n(t) := n^{1/2}(\bar{F}_n(t) - \Phi(t))$$

be the corresponding empirical processes. We know that $\bar{G}_n$ weakly converge to the zero-mean Gaussian process Z with the covariance function (29). Our goal is to prove that the same is true about G_n.

The idea is to use Prokhorov's theorem, in the form of Theorem 13.1 in Billingsley (1999), first proving that the finite-dimensional distributions of G_n converge to those of Z

and then that the sequence G_n is tight. The functional space $D(-\infty, \infty)$ is defined and studied in Billingsley (1999, p. 191); we can use it in place of $\mathbb{D}$ if we consider, without loss of generality, the domains of G_n and $\bar{G}_n$ to be bounded. Let $\pi_{t_1^*, \ldots, t_k^*}$ be the projection of $D(-\infty, \infty)$ onto $\mathbb{R}^k$: $\pi_{t_1^*, \ldots, t_k^*}(x) := (x(t_1^*), \ldots, x(t_k^*))$.

Lemma 6 *The finite-dimensional distributions of G_n weakly converge to Z: $\pi_{t_1^*, \ldots, t_k^*}(G_n) \Rightarrow \pi_{t_1^*, \ldots, t_k^*}(Z)$.*

Proof For simplicity, we will only consider two-dimensional distributions. To see that $\pi_{t_1^*, t_2^*}(G_n) \Rightarrow \pi_{t_1^*, t_2^*}(Z)$, notice that, for some $\epsilon_n \to 0$,

$$\mathbb{P}\left(G_n(t_1^*) \le a_1, G_n(t_2^*) \le a_2\right)$$
$$= \mathbb{P}\left(n^{1/2}(F_n(t_1^*) - \Phi(t_1^*)) \le a_1, n^{1/2}(F_n(t_2^*) - \Phi(t_2^*)) \le a_2\right)$$
$$\le \mathbb{P}\Big(n^{1/2}(\bar{F}_n(t_1^* - \epsilon_n n^{-1/2}) - \Phi(t_1^*)) \le a_1,$$
$$\qquad n^{1/2}(\bar{F}_n(t_2^* - \epsilon_n n^{-1/2}) - \Phi(t_2^*)) \le a_2\Big)$$
$$\le \mathbb{P}\Big(n^{1/2}(\bar{F}_n(t_1^* - \epsilon_n n^{-1/2}) - \Phi(t_1^* - \epsilon_n n^{-1/2})) \le a_1 + \epsilon_n,$$
$$\qquad n^{1/2}(\bar{F}_n(t_2^* - \epsilon_n n^{-1/2}) - \Phi(t_2^* - \epsilon_n n^{-1/2})) \le a_2 + \epsilon_n\Big) \tag{44}$$
$$= \mathbb{P}\left(\bar{G}_n(t_1^* - \epsilon_n n^{-1/2}) \le a_1 + \epsilon_n, \bar{G}_n(t_2^* - \epsilon_n n^{-1/2}) \le a_2 + \epsilon_n\right)$$
$$\le \mathbb{P}\left(\bar{G}_n(t_1^* - \epsilon_n n^{-1/2}) \le a_1 + \epsilon, \bar{G}_n(t_2^* - \epsilon_n n^{-1/2}) \le a_2 + \epsilon\right) \tag{45}$$
$$\to \mathbb{P}\left(Z(t_1^*) \le a_1 + \epsilon, Z(t_2^*) \le a_2 + \epsilon\right) \tag{46}$$
$$\le \mathbb{P}\left(Z(t_1^*) \le a_1, Z(t_2^*) \le a_2\right) + \delta. \tag{47}$$

The inequality (44) follows from $\left|\Phi'\right| \le 1$. The inequality (45) holds from some n on for any $\epsilon > 0$. By making ϵ sufficiently small we can make the δ in (47) arbitrarily small. The convergence (46) follows from Lemma 8 below. In the same way we can prove the opposite inequality

$$\mathbb{P}\left(G_n(t_1^*) \le a_1, G_n(t_2^*) \le a_2\right)$$
$$= \mathbb{P}\left(n^{1/2}(F_n(t_1^*) - \Phi(t_1^*)) \le a_1, n^{1/2}(F_n(t_2^*) - \Phi(t_2^*)) \le a_2\right)$$
$$\ge \mathbb{P}\Big(n^{1/2}(\bar{F}_n(t_1^* + \epsilon_n n^{-1/2}) - \Phi(t_1^*)) \le a_1,$$
$$\qquad n^{1/2}(\bar{F}_n(t_2^* + \epsilon_n n^{-1/2}) - \Phi(t_2^*)) \le a_2\Big)$$
$$\ge \mathbb{P}\Big(n^{1/2}(\bar{F}_n(t_1^* + \epsilon_n n^{-1/2}) - \Phi(t_1^* + \epsilon_n n^{-1/2})) \le a_1 - \epsilon_n,$$
$$\qquad n^{1/2}(\bar{F}_n(t_2^* + \epsilon_n n^{-1/2}) - \Phi(t_2^* + \epsilon_n n^{-1/2})) \le a_2 - \epsilon_n\Big)$$
$$= \mathbb{P}\left(\bar{G}_n(t_1^* + \epsilon_n n^{-1/2}) \le a_1 - \epsilon_n, \bar{G}_n(t_2^* + \epsilon_n n^{-1/2}) \le a_2 - \epsilon_n\right)$$
$$\ge \mathbb{P}\left(\bar{G}_n(t_1^* + \epsilon_n n^{-1/2}) \le a_1 - \epsilon, \bar{G}_n(t_2^* + \epsilon_n n^{-1/2}) \le a_2 - \epsilon\right)$$
$$\to \mathbb{P}\left(Z(t_1^*) \le a_1 - \epsilon, Z(t_2^*) \le a_2 - \epsilon\right)$$
$$\ge \mathbb{P}\left(Z(t_1^*) \le a_1, Z(t_2^*) \le a_2\right) - \delta.$$

$\square$

The second step in the proof of Theorem 3 is to prove the tightness of the perturbed empirical distribution functions for the residuals.

Lemma 7 *The sequence G_n, $n = 1, 2, \ldots$, is tight.*

Proof We will use the standard notation for càdlàg functions x on a closed interval of the real line (Billingsley 1999, Section 12): $j(x)$ stands for the size of the largest jump of x, $w_x(T) := \sup_{s,t \in T} |x(s) - x(t)|$ for any subset T of the domain of x, $w_x(\delta) := \sup_t w_x[t, t+\delta]$ for any $\delta > 0$, and $w'_x(\delta) := \inf_{\{t_i\}} \max_{i \in \{1,\ldots,v\}} w_x[t_{i-1}, t_i)$, where $t_0 < t_1 < \cdots < t_v$ range over the partitions of the domain $[t_0, t_v]$ of x that are δ-sparse in the sense of $\min_{i \in \{1,\ldots,v\}}(t_i - t_{i-1}) > \delta$.

We know that $\bar{G}_n \Rightarrow Z$ and, therefore, $\bar{G}_n$ is tight. By Billingsley (1999, Theorem 13.4), the continuity of Z implies that $j(\bar{G}_n) \Rightarrow 0$. This can be written as

$$\forall \epsilon > 0 : \lim_n \mathbb{P}(j(\bar{G}_n) \geq \epsilon) = 0$$

and in combination with

$$\forall \epsilon > 0 : \lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}(w'_{\bar{G}_n}(\delta) \geq \epsilon) = 0$$

(this is Billingsley 1999, Theorem 13.2(ii)) and

$$w_x(\delta) \leq 2w'_x(\delta) + j(x)$$

(this is Billingsley 1999, (12.9)) implies

$$\forall \epsilon > 0 : \lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}(w_{\bar{G}_n}(\delta) \geq \epsilon) = 0. \tag{48}$$

The statement of the lemma will follow from Billingsley (1999, Theorem 13.2, Corollary). We will only check condition (ii) (i.e., (7.7) in Billingsley 1999) for G_n; in other words, we will check (48) with G_n in place of $\bar{G}_n$. It suffices to notice that

$$
\begin{aligned}
w_{G_n}(\delta) &= \sup_{|t_2 - t_1| \leq \delta} |G_n(t_2) - G_n(t_1)| = \sup_{|t_2 - t_1| \leq \delta} (G_n(t_2) - G_n(t_1)) \\
&= \sup_{|t_2 - t_1| \leq \delta} \sqrt{n}\, (F_n(t_2) - \Phi(t_2) - F_n(t_1) + \Phi(t_1)) \\
&\leq \sup_{|t_2 - t_1| \leq \delta} \sqrt{n}\, \left(\bar{F}_n(t_2 + \epsilon_n n^{-1/2}) - \Phi(t_2) - \bar{F}_n(t_1 - \epsilon_n n^{-1/2}) + \Phi(t_1) \right) \\
&\leq \sup_{|t_2 - t_1| \leq \delta} \sqrt{n}\Big(\bar{F}_n(t_2 + \epsilon_n n^{-1/2}) - \Phi(t_2 + \epsilon_n n^{-1/2}) \\
&\qquad - \bar{F}_n(t_1 - \epsilon_n n^{-1/2}) + \Phi(t_1 - \epsilon_n n^{-1/2}) \Big) + \epsilon_n n^{-1/2} \\
&= \sup_{|t_2 - t_1| \leq \delta} \left(\bar{G}_n(t_2 + \epsilon_n n^{-1/2}) - \bar{G}_n(t_1 - \epsilon_n n^{-1/2}) \right) + \epsilon_n n^{-1/2} \\
&\leq \sup_{|t'_2 - t'_1| \leq 2\delta} \left(\bar{G}_n(t'_2) - \bar{G}_n(t'_1) \right) + \epsilon_n n^{-1/2} \tag{49} \\
&= w_{\bar{G}_n}(2\delta) + \epsilon_n n^{-1/2},
\end{aligned}
$$

where the inequality (49) holds from some n on.   □

Now Theorem 3 follows from Lemmas 6 and 7 by Billingsley (1999, Theorem 13.1).
The following lemma was used in the proof of Lemma 6.

Lemma 8 *Suppose a sequence $\bar{G}_n$ of random functions in $D(-\infty, \infty)$ weakly converges to a random function Z in $C(-\infty, \infty)$ and suppose $t_n \to t$ are real numbers (or, more generally, t_n are random variables converging to t in probability). Then $\bar{G}_n(t_n)$ weakly converges to $Z(t)$.*

Proof By Billingsley (1999, Theorem 3.9), $(\bar{G}_n, t_n) \Rightarrow (Z, t)$. By the mapping theorem (Billingsley 1999, Theorem 2.7), $\bar{G}_n(t_n) \Rightarrow Z(t)$. □

Proofs for the studentized LSPM

Let us see that Theorems 2 and 3 still hold for the deleted and studentized LSPM. For concreteness, we will only consider the studentized LSPM. We have the following stronger form of Lemma 5.

Lemma 9 *Under conditions A1–A4,* $\max_{i,j=1,\ldots,n+1} \left| \bar{h}_{i,j} \right| = O(n^{-1})$.

Proof As in the proof of Lemma 5, we have, for all permitted sequences $x_1, x_2, \ldots$,

$$\max_{i,j=1,\ldots,n} \left| h_{i,j} \right| \leq \frac{\max_{i,j=1,\ldots,n} \|x_i\| \, \|x_j\|}{\lambda_{\min}(X'X)} = O(1).$$

It remains to combine this with (20), (21), and the statement of Lemma 5. □

We will use the old notation B and A for the ordinary LSPM, (41) and (42), but will supply B and A with primes, writing B' and A', for the studentized LSPM, (15) and (16). Since B and B' are very close to 1,

$$B_i' = \sqrt{1 - \bar{h}_{n+1}} + \frac{\bar{h}_{i,n+1}}{\sqrt{1 - \bar{h}_i}} = 1 + O(n^{-1}),$$

$$B_i = 1 - \bar{h}_{n+1} + \bar{h}_{i,n+1} = 1 + O(n^{-1}),$$

we only need to check that A and A' are very close between themselves. The difference between them,

$$A_i' - A_i = \left(\frac{\sum_{j=1}^n \bar{h}_{j,n+1} y_j}{\sqrt{1 - \bar{h}_{n+1}}} + \frac{y_i - \sum_{j=1}^n \bar{h}_{i,j} y_j}{\sqrt{1 - \bar{h}_i}} \right)$$
$$- \left(\sum_{j=1}^n \bar{h}_{j,n+1} y_j + y_i - \sum_{j=1}^n \bar{h}_{i,j} y_j \right)$$
$$= \sum_{j=1}^n O(n^{-2}) y_j + O(n^{-1}) y_i + \sum_{j=1}^n O(n^{-2}) y_j,$$

has a Gaussian distribution conditionally on $x_1, x_2, \ldots$, and its variance is $O(n^{-2})$. Now it suffices to apply the method of the previous subsection to

$$t_i' := \frac{A_i'/B_i' - \hat{y}_{n+1}}{\sigma},$$

which can also be regarded as perturbed $\bar{t}_i$.

7 Experimental results

In this section we explore experimentally the validity and efficiency of the studentized LSPM.

7.1 Online validity

First we check experimentally the validity of our methods in the online mode of prediction. It is guaranteed by our theoretical results but provides an opportunity to test the correctness of our implementation.

We generate IID observations $z_n = (x_n, y_n)$, $n = 1, \ldots, 1001$, and the corresponding p values $p_n := Q_n(y_{n+1}, \tau_n)$, $n = 1, \ldots, N$, $N := 1000$, in the online mode. In our experiments, $x_n \sim N(0, 1)$, $y_n \sim 2x_n + N(0, 1)$, and, as usual, $\tau_n \sim U$, all independent. Figure 2 plots $S_n := \sum_{i=1}^{n} p_i$ vs $n = 1, \ldots, N$; as expected, it is an approximately straight line with slope 0.5. Figure 3 presents three plots: the cumulative sums $S_n^\alpha := \sum_{i=1}^{n} \mathbf{1}_{\{p_i \leq \alpha\}}$, where $\mathbf{1}$ is the indicator function, vs $n = 1, \ldots, N$, for three values of α, $\alpha \in \{0.25, 0.5, 0.75\}$. For each of the three αs the result is an approximately straight line with slope α. Finally, Fig. 4 plots A_N^α against $\alpha \in [0, 1]$, where $A_N^\alpha := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{p_i \leq \alpha\}}$. The result is, approximately, the main diagonal of the square $[0, 1]^2$, as it should be.
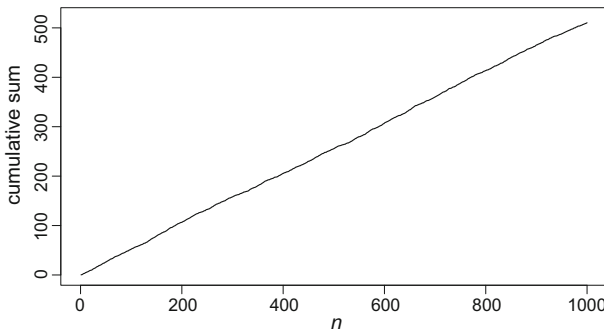
Fig. 2 The cumulative sums S_n of the p values versus $n = 1, \ldots, 1000$
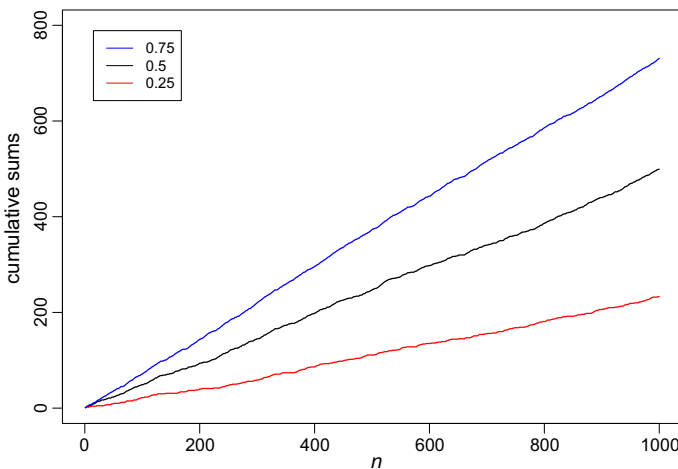
Fig. 3 The cumulative sums S_n^α versus $n = 1, \ldots, 1000$ for $\alpha \in \{0.25, 0.5.0.75\}$

Fig. 4 The calibration curve: A_N^α versus $\alpha \in [0, 1]$ for $N = 1000$
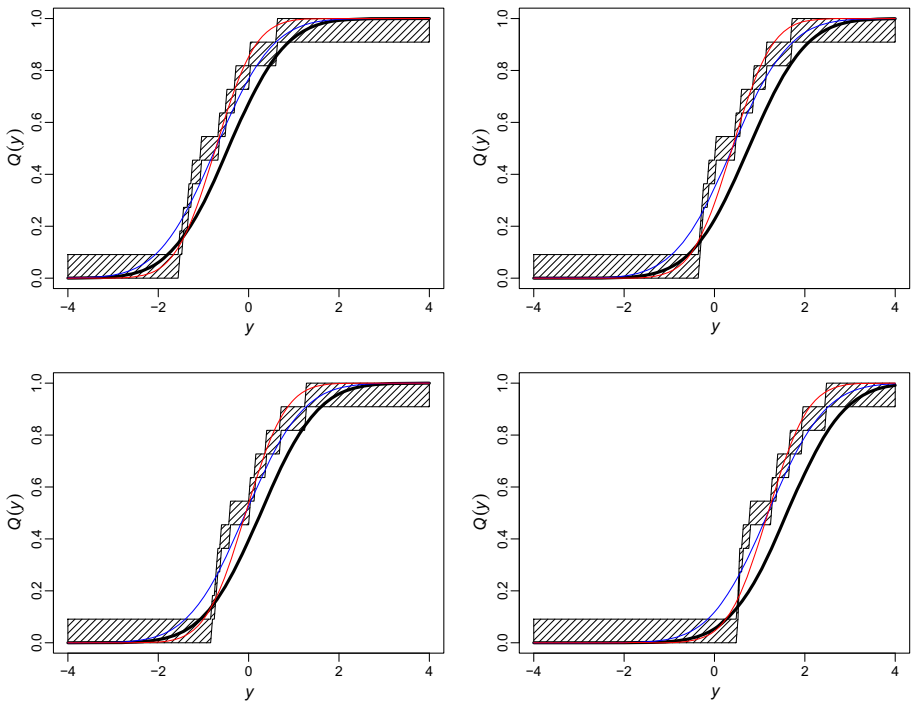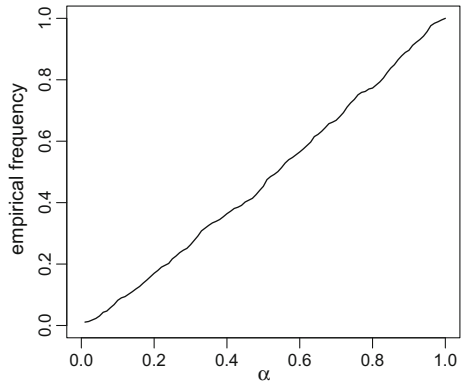

Fig. 5 Examples of true predictive distribution functions (black), their conformal estimates (represented by the shaded areas), and the distribution functions output by simplified Oracle I (red) and Oracle II (blue) for a tiny training sequence (of length 10 with two attributes, the first one being the dummy all-1 attribute); in black and white, the true predictive distribution functions are the thick lines, and Oracle I is always farther from them in the uniform metric than Oracle II is
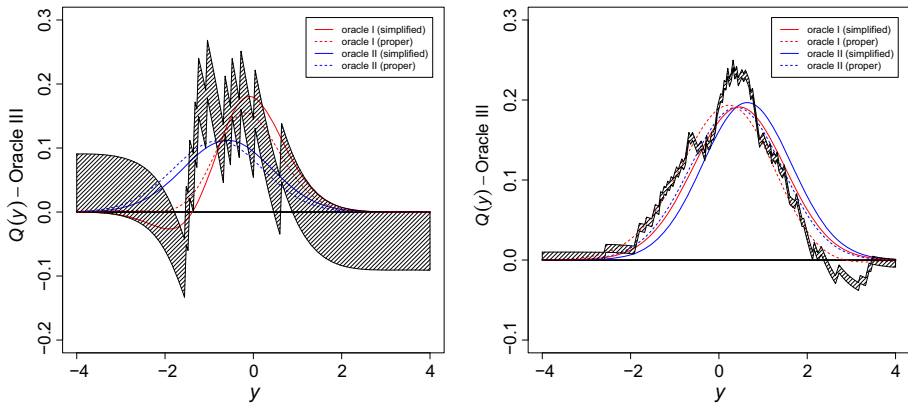
Fig. 6 The left-hand plot is the first (upper left) plot of Fig. 5 normalized by subtracting the true distribution function (the thick black line in Fig. 5, which now coincides with the x-axis) and with the outputs of the proper oracles added; the right-hand plot is an analogous plot for a larger training sequence (of length 100 with 20 attributes, the first one being the dummy attribute)

7.2 Efficiency

Next we explore empirically the efficiency of the studentized LSPM. Figure 5 compares the conformal predictive distribution with the true (Oracle III's) distribution for four randomly generated test objects and a randomly generated training sequence of length 10 with 2 attributes. The first attribute is a dummy all-1 attribute; remember that Theorems 2 and 3 depend on the assumption that one of the attributes is an identical 1 (without it, the plots become qualitatively different: cf. Chen 1991, Corollary 2.4.1). The second attribute is generated from the standard Gaussian distribution, and the labels are generated as $y_n \sim 2x_{n,2} + N(0, 1)$, $x_{n,2}$ being the second attribute. We also show (with thinner lines) the output of Oracle I and Oracle II, but only for the simplified versions, in order not to clutter the plots. Instead, in the left-hand plot of Fig. 6 we show the first plot of Fig. 5 that is normalized by subtracting the true distribution function; this time, we show the output of both simplified and proper Oracles I and II; the difference is not large but noticeable. The right-hand plot of Fig. 6 is similar except that the training sequence is of length 100 and there are 20 attributes generated independently from the standard Gaussian distribution except for the first one, which is the dummy all-1 attribute; the labels are generated as before, $y_n \sim 2x_{n,2} + N(0, 1)$.

Since Oracle III is more powerful than Oracles I and II (it knows the true data-generating distribution), it is more difficult to compete with; therefore, the black line is farther from the shaded area than the blue and red lines for all four plots in Fig. 5. The estimated distribution functions being to the left of the true distribution functions is a coincidence: the four plots correspond to the values 0–3 of the seed for the R pseudorandom number generator, and for other seeds the estimated distribution functions are sometimes to the right and sometimes to the left.

8 Conclusion

This paper introduces conformal predictive distributions in regression problems. Their advantage over the usual conformal prediction intervals is that a conformal predictive distribution Q_n contains more information; in particular, it can produce a plethora of prediction intervals:

e.g., for each $\epsilon > 0$, $\{y \in \mathbb{R} \mid \epsilon/2 \leq Q_n(y, \tau) \leq 1 - \epsilon/2\}$ is a conformal prediction interval at confidence level $1 - \epsilon$.

These are natural possible topics for further research:

– This paper is based on the most traditional approach to weak convergence of empirical processes, originated by Skorokhod and described in detail by Billingsley (1999). This approach encounters severe difficulties in more general situations (such as multi-dimensional labels). Alternative approaches have been proposed by numerous authors, including Dudley (using the uniform topology and ball σ-algebra, Dudley 1966, 1967) and Hoffmann-Jørgensen (dropping measurability and working with outer integrals; see, e.g., van der Vaart and Wellner 1996, Section 1.3 and the references in the historical section). Translating our results into those alternative languages might facilitate various generalizations.

– Another generalization of the traditional notion of weak convergence is Belyaev's notion of weakly approaching sequences of random distributions (Belyaev and Sjöstedt–de Luna 2000). When comparing the LSPM with Oracle III, we limited ourselves to stating the absence of weak convergence and calculating the asymptotics of 1-dimensional distributions; Belyaev's definition is likely to lead to more precise results.

– The recent paper by Nouretdinov et al. (2018) uses inductive Venn–Abers predictors to produce predictive distributions, with very different guarantees of validity. Establishing connections between the approach of this paper and that of Nouretdinov et al. (2018) is an interesting direction of further research.

References

Augustin, T., & Coolen, F. P. A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, *124*, 251–272.

Balasubramanian, V. N., Ho, S. S., & Vovk, V. (Eds.). (2014). *Conformal prediction for reliable machine learning: Theory, adaptations, and applications*. Amsterdam: Elsevier.

Belyaev, Y., & Sjöstedt-de Luna, S. (2000). Weakly approaching sequences of random distributions. *Journal of Applied Probability*, *37*, 807–822.

Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). New York: Wiley.

Burnaev, E., & Vovk, V. (2014). Efficiency of conformalized ridge regression. *JMLR: Workshop and Conference Proceedings*, *35*, 605–622. **(COLT 2014)**.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. New York: Wiley.

Chen, G. (1991). *Empirical processes based on regression residuals: Theory and applications*. PhD thesis, Department of Mathematics and Statistics, Simon Fraser University

Dempster, A. P. (1963). On direct probabilities. *Journal of the Royal Statistical Society B*, *25*, 100–110.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems. *Annals of Mathematical Statistics*, *23*, 277–281.

Dudley, R. M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois Journal of Mathematics*, *10*, 109–126.

Dudley, R. M. (1967). Measures on non-separable metric spaces. *Illinois Journal of Mathematics*, *11*, 449–453.

Fisher, R. A. (1939). Student. *Annals of Eugenics*, *9*, 1–9.

Fisher, R. A. (1948). Conclusions fiduciaires. *Annales de l'Institut Henry Poincaré*, *10*, 191–213.

Genest, C., & Kalbfleisch, J. (1988). Bayesian nonparametric survival analysis: Comment. *Journal of the American Statistical Association*, *83*, 780–781.

Geyer, C. J., & Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values (with discussion). *Statistical Science*, *20*, 358–387.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, *63*, 677–691.

Hill, B. M. (1988). De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In D. V. Lindley, J. M. Bernardo, M. H. DeGroot, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 3, pp. 211–241). Oxford: Oxford University Press.

Jeffreys, H. (1932). On the theory of errors and least squares. *Proceedings of the Royal Society of London A*, *138*, 48–55.

Lane, D. A. (1980). Fisher, Jeffreys, and the nature of probability. In S. E. Fienberg & D. V. Hinkley (Eds.), *R. A. Fisher: An appreciation, lecture notes in statistics* (Vol. 1, pp. 148–160). Berlin: Springer.

Lawless, J. F., & Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, *92*, 529–542.

Mohammadi, M. (2016). On the bounds for diagonal and off-diagonal elements of the hat matrix in the linear regression model. *REVSTAT Statistical Journal*, *14*, 75–87.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Hoboken, NJ: Wiley.

Mugantseva, L. A. (1977). Testing normality in one-dimensional and multi-dimensional linear regression. *Theory of Probability and Its Applications*, *22*, 591–602.

Nouretdinov, I., Volkhonskiy, D., Lim, P., Toccaceli, P., & Gammerman, A. (2018). Inductive Venn-Abers predictive distribution. *Proceedings of Machine Learning Research*, *60*, 15–36. **(COPA 2018)**.

Pierce, D. A., & Kopecky, K. J. (1979). Testing goodness of fit for the distribution of errors in regression models. *Biometrika*, *66*, 1–5.

Pinelis, I. (2015). Exact bounds on the closeness between the Student and standard normal distributions. *ESAIM: Probability and Statistics*, *19*, 24–27.

Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability: Statistical inference with confidence distributions*. Cambridge: Cambridge University Press.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). Hoboken, NJ: Wiley.

Seidenfeld, T. (1995). Jeffreys, Fisher, and Keynes: Predicting the third observation, given the first two. In A. F. Cottrell & M. S. Lawlor (Eds.), *New perspectives on Keynes* (pp. 39–52). Durham, NC: Duke University Press.

Shen, J., Liu, R., & Xie, M. (2018). Prediction with confidence—A general framework for predictive inference. *Journal of Statistical Planning and Inference*, *195*, 126–140.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. New York: Springer.

Vovk, V., & Bendtsen, C. (2018). Conformal predictive decision making. *Proceedings of Machine Learning Research*, *91*, 52–62. **(COPA 2018)**.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.

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

Vovk, V., Nouretdinov, I., Manokhin, V., & Gammerman, A. (2017a). *Conformal predictive distributions with kernels*. On-line compression modelling project (new series). http://alrw.net. Working paper 20

Vovk, V., Shen, J., Manokhin, V., & Xie, M. (2017b). Nonparametric predictive distributions based on conformal prediction. *Proceedings of Machine Learning Research*, *60*, 82–102. **(COPA 2017)**.

Wang, C. M., Hannig, J., & Iyer, H. K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, *142*, 1980–1990.