Nonparametric predictive distributions based on conformal prediction

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #17

April 5, 2017

Project web site:
http://alrw.net

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 Fisher–Dempster–Hill (FDH) predictive distributions to regression problems. If the standard parametric assumptions for Least Squares linear regression hold, the LSPM is as efficient as the FDH procedure, in a natural sense. And if those parametric assumptions fail, the LSPM is still valid, provided the observations are IID.

Contents

> Let me conclude by observing that $A_{(n)}$ is supported by all of the serious approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It is simple, coherent, and plausible. It can even be argued, I believe, that $A_{(n)}$, along with $H_{(n)}$, constitutes the fundamental solution to the problem of induction.
>
> Bruce M. Hill, 1988

> To be truly useful, however, the methods need extension to regression models with unknown regression parameters.
>
> Christian Genest and Jack Kalbfleisch, 1988

1 Introduction

This paper applies conformal prediction to derive predictive distribution functions (Shen et al., 2017) that are valid under a nonparametric assumption. The theory of predictive distributions as developed by Shen et al. (2017) 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. Our predictive distributions generalize the classical Fisher–Dempster–Hill (FDH) procedure, which these authors referred to as fiducial probabilities, direct probabilities, and $A_{(n)}/H_{(n)}$, respectively. For a relatively recent review of predictive distributions, see Lawless and Fredette (2005).

We start our formal exposition from defining conformal predictive distributions (CPDs), nonparametric predictive distributions based on conformal prediction, in Section 2; we are only interested in 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 Section 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, 4 and 5, are devoted to the validity and efficiency of the LSPM. Whereas the LSPM, as any CPD, 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 Section 5 is how much we should pay 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 FDH procedure. Section 6 is devoted to experimental results.

2 Randomized and conformal predictive distributions

We consider the regression problem with p attributes. 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 $z_i = (z_i, y_i)$, $i = 1, \dots, n$, and a new test object $x_{n+1} \in \mathbb{R}^p$, to predict the label y_{n+1} of the $(n+1)$th observation. Our statistical model is the general IID model: the observations $z_1, z_2, \dots$, where $z_i := (x_i, y_i)$, are generated independently from the same unknown probability measure P on $\mathbb{R}^p \times \mathbb{R}$.

We start from defining predictive distribution functions following Shen et al. (2017, 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 distribution (function)* (RPD) if it satisfies the following three requirements:

R1 i For each training sequence $(z_1, \dots, z_n) \in (\mathbb{R}^{p+1})^n$ and each test object $x_{n+1} \in \mathbb{R}^p$, the function $Q(z_1, \dots, z_n, (x_{n+1}, y), \tau)$ is monotonically increasing both in y and τ (where "monotonically increasing" is understood in the wide sense allowing intervals of constancy). In other words, for each $\tau \in [0, 1]$, the function

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

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

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

is monotonically increasing.

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

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

and

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

R2 As function of random training observations $z_1 \sim P,\dots, z_n \sim P$, 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:

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

The *thickness* of an RPD Q for a training sequence $(z_1,\dots,z_n) \in (\mathbb{R}^{p+1})^n$ is the smallest number $\epsilon \ge 0$ such that, for each test object $x \in \mathbb{R}^p$, the diameter

$$Q(z_1,\dots,z_n,(x,y),1) - Q(z_1,\dots,z_n,(x,y),0) \tag{1}$$

of the set

$$\{Q(z_1,\dots,z_n,(x,y),\tau) \mid \tau \in [0,1]\} \tag{2}$$

is at most ϵ for all $y \in \mathbb{R}$ except for finitely many values. The *exception size* of Q for $(z_1,\dots,z_n)$ is the maximum (over all test objects $x \in \mathbb{R}^p$) cardinality of the set of y for which the diameter (1) exceeds the thickness of Q. 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. In all our examples, $Q(z_1,\dots,z_n,z_{n+1},\tau)$ will be a continuous function of τ. Therefore, the set (2) will be a closed interval in $[0,1]$. However, we do not include these requirements in our official definition.

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. 2005a, Section 2.2, than to Balasubramanian et al. 2014, Section 1.3). A *conformity measure* is a measurable function $A : (\mathbb{R}^{p+1})^{n+1} \to \mathbb{R}$ that is invariant with respect to permutations of the first n observations: for any sequence $(z_1,\dots,z_n) \in (\mathbb{R}^{p+1})^n$, any $z_{n+1} \in \mathbb{R}^{p+1}$, and any permutation π of $\{1,\dots,n\}$,

$$A(z_1,\dots,z_n,z_{n+1}) = A\left(z_{\pi(1)},\dots,z_{\pi(n)},z_{n+1}\right). \tag{3}$$

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

$$A(z_1,\dots,z_{n+1}) := y_{n+1} - \hat{y}_{n+1}, \tag{4}$$

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

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

$$Q(z_1,\dots,z_n,(x_{n+1},y),\tau) := \frac{1}{n+1}\left|\left\{i=1,\dots,n+1 \mid \alpha_i^y < \alpha_{n+1}^y\right\}\right| + \frac{\tau}{n+1}\left|\left\{i=1,\dots,n+1 \mid \alpha_i^y = \alpha_{n+1}^y\right\}\right|, \quad (5)$$

where $(z_1,\dots,z_n) \in (\mathbb{R}^{p+1})^n$ 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 scores* α_i^y are defined by

$$\begin{aligned} \alpha_i^y &:= A(z_1,\dots,z_{i-1},z_{i+1},\dots,z_n,(x_{n+1},y),z_i), \qquad i=1,\dots,n, \\ \alpha_{n+1}^y &:= A(z_1,\dots,z_n,(x_{n+1},y)). \end{aligned} \quad (6)$$

A function is a *conformal transducer* if it is the conformal transducer determined by some conformity measure. A *conformal predictive distribution* (CPD) is a function which is both a conformal transducer and a randomized predictive distribution.

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

$$\Gamma^A(z_1,\dots,z_n,x_{n+1},\tau) := \{y \in \mathbb{R} \mid Q(z_1,\dots,z_n,(x_{n+1},y),\tau) \in A\}. \quad (7)$$

The standard property of validity for conformal transducers is that the values (also called p-values) $Q(z_1,\dots,z_{n+1},\tau)$ are distributed uniformly on $[0,1]$ when $z_1,\dots,z_{n+1}$ are IID and τ is generated independently of $z_1,\dots,z_{n+1}$ from the uniform probability distribution U on $[0,1]$ (see, e.g., Vovk et al. 2005a, Proposition 2.8). This property coincides with requirement R2 in the definition of an RPD and implies that the coverage probability, i.e., the probability of $y_{n+1} \in \Gamma^A(z_1,\dots,z_n,x_{n+1})$, for the conformal predictor (7) is $U(A)$.

Remark 1. The usual interpretation of (5) 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} is too small. Then (4) can be interpreted as a degree of conformity of the observation (x_{n+1},y_{n+1}) to the remaining observations.

Defining properties of distribution functions

Next we discuss why Definition 1 (essentially taken from Shen et al. 2017) 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.

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 $\mathbb{R} \times [0,1]$: $(y,\tau) \le (y',\tau')$ is defined to mean that $y < y'$ or both $y = y'$ and $\tau \le \tau'$.

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,*

$$Q(y,\tau) = (1-\tau)F(y-) + \tau F(y). \tag{8}$$

Equality (8) 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 (8) for the distribution function F of a probability measure P, the distribution of Q will be uniform when its domain is equipped with the probability measure $P \times U$: see the literature on randomized p-values, such as the recent review Gurevich and Vovk (2017).

The previous two lemmas suggest that properties R1(i) and R2 in the definition of RPDs are the important ones. However, property R1(ii) is formally independent of R1(i) 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 the labels y_i.

Simplest example: monotonic conformity measures

We start from a simple but very restrictive condition on a conformity measure making the corresponding conformal transducer satisfy R1(i). A conformity measure A is *monotonic* if $A(z_1,\dots,z_{n+1})$ is:

- monotonically increasing in y_{n+1},

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

- monotonically decreasing in y_1,

  $$y_1 \le y'_1 \implies A((x_1,y_1),z_2,\dots,z_n,z_{n+1}) \ge A((x_1,y'_1),z_2,\dots,z_n,z_{n+1}).$$

  (Because of the requirement of invariance (3), being decreasing in y_1 is equivalent to being decreasing in y_i for any $i = 2,\dots,n$.)

This condition implies that the corresponding conformal transducer (5) satisfies R1(i) by Lemma 3 below.

An example of a monotonic conformity measure is (4), where $\hat{y}$ is produced by the K-nearest neighbours regression algorithm:

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

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 R1(ii) and so is an RPD and CPD.

Criterion of being a CPD

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

Lemma 3. *The conformal transducer determined by a conformity measure A satisfies condition R1(i) 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 $x_{n+1} \in \mathbb{R}^p$, $\alpha^y_{n+1} - \alpha^y_i$ is a monotonically increasing function of $y \in \mathbb{R}$ (in the notation of (6)).*

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 RPD (and, therefore, a CPD) if, e.g.,

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

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. (2005a, Section 2.3) (and called the IID predictor in Vovk et al. 2005b), 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

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \widehat{y}_{n+1} \tag{9}$$

(cf. (4)), where y_{n+1} is the label in z_{n+1} and $\widehat{y}_{n+1}$ 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 (9) 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

$$A(z_1, \ldots, z_{n+1}) := y_{n+1} - \hat{y}_{n+1}, \tag{10}$$

whose difference from (9) is that $\widehat{y}_{n+1}$ 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 RPD, because their output $Q_n : \mathbb{R} \times [0,1] \to [0,1]$ for a test object $x_{n+1} \in \mathbb{R}^p$,

$$Q_n(y, \tau) := Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) \tag{11}$$

(cf. (5)) is not necessarily monotonically increasing in y (remember that $Q_n(y, \tau)$ 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

$$\bar{H} = \bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'. \tag{12}$$

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) of the ordinary LSPM.

Proposition 1. *The function Q_n output by the ordinary LSPM (see (11)) 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 proposition shows that for the deleted LSPM, determined by (10), 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 (11) 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

$$A(z_1, \dots, z_{n+1}) := \frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{1 - \bar{h}_{n+1}}} \tag{13}$$

(intermediate between those for the ordinary and deleted LSPM: a standard representation for the deleted residuals is $(y_i - \widehat{y}_i)/(1 - \bar{h}_i)$, $i = 1, \dots, n+1$; 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 RPD and, therefore, a CPD.*

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 (11) will be represented in the form

$$Q_n(y) := [Q_n(y, 0), Q_n(y, 1)] \tag{14}$$

(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 $Q_n(y) = [a, b]$.

For the studentized residuals (13), we can easily obtain

$$\alpha^y_{n+1} - \alpha^y_i = B_i y - A_i, i = 1, \dots, n,$$

in the notation of (6), where y is the label of the $(n+1)$th object x_{n+1} and

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

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

We will assume that all B_i are defined and positive; this assumption will be further discussed at the end of this subsection.

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

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

Algorithm 1 Least Squares Prediction Machine

Require: A training sequence $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$.
Require: A test object $x_{n+1} \in \mathbb{R}^p$.
Set $\bar{X}$ to the data matrix for the given $n+1$ objects.
Define the hat matrix $\bar{H}$ by (12).
for $i \in \{1, 2, \dots, n\}$ **do**
    Define A_i and B_i by (16) and (15), respectively.
    Set $C_i := A_i / B_i$.
end for
Sort $C_1, \dots, C_n$ in the increasing order obtaining $C_{(1)} \le \dots \le C_{(n)}$.
Return the predictive distribution (17) for y_{n+1}.

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', $i = 1, \dots, n+1$, 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, $i = 1, \dots, n$. By Chatterjee and Hadi (1988, Property 2.6(b)), $\bar{h}_{n+1} = 1$ implies $\bar{h}_{i,n+1} = 0$ for $i = 1, \dots, n$; therefore, the condition is equivalent to $\bar{h}_i < 1$ for all $i = 1, \dots, n+1$. By Mohammadi (2016, Lemma 2.1(iii)), this means that the rank of the extended hat matrix $\bar{H}$ 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 CPD.

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 $x_{n+1}, \dots, x_{n+m}$ instead of just one test object x_{n+1}. In this case we can precompute the hat matrix for the training objects $x_1, \dots, x_n$, 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', $i = 1, \dots, n$. Set

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

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

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

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

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

Algorithm 2 Least Squares Prediction Machine (batch version)

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

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

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

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

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 $z_1, \dots, z_n$ as the training sequence. The predictive distribution is defined by (17). The fraction 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 version 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 on-line mode

In the previous section (see 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, the 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. *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 wide-spread 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 and 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 There is a ball in $\mathbb{R}^p$ containing all objects $x_i \in \mathbb{R}^p$, $i = 1, 2, \ldots$.

A2 The first component of each 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, we define *proper Oracle I* as 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 as

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

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

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

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). However, the version

that is more popular in the literature on empirical processes for residuals is the *simplified Oracle I*

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

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

Proper Oracle II is defined as the predictive distribution

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

Correspondingly, the *simplified Oracle II* is 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* is defined as the predictive distribution

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

Our notation is Q_n for the conformal predictive distribution (11), as before, Q_n^{I} for the simplified or proper Oracle I predictive distribution, (26) or (25) (Theorem 2 will hold for both), and Q_n^{II} for the simplified or proper Oracle II 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 function $G_n : \mathbb{R} \to [0,1]$ 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 converges to a Gaussian process Z with mean zero and covariance function

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

Theorem 3. *The random function $G_n : \mathbb{R} \to [0,1]$ 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 converges to a Gaussian process Z with mean zero and covariance function

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

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 (omitted in this version of the paper) 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):

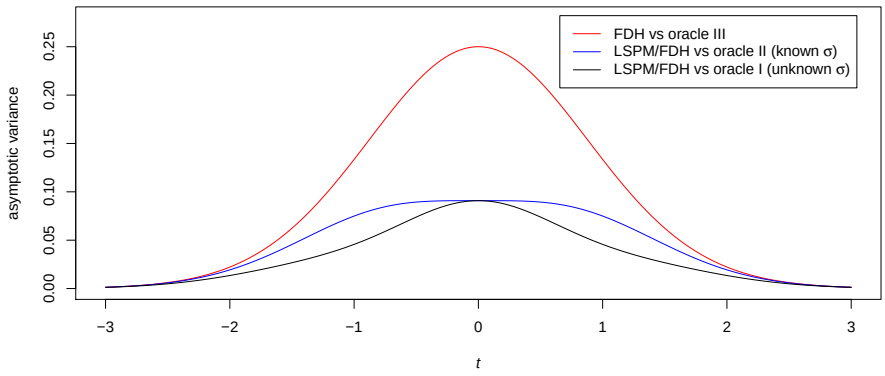

Figure 1: The asymptotic variances for the FDH procedure as compared with the truth (Oracle III, red) and for the LSPM and FDH as compared with the oracle procedures for known σ (Oracle II, blue) and unknown σ (Oracle I, black)

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

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

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

Figure 1 presents plots for the asymptotic variances, given in Corollary 1, for the two oracular predictive distributions: black for Oracle I and blue for Oracle II (the red plot will be discussed later in this section). The maximum of both asymptotic variances, attained at $y = 0$, is 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 oracle predictive distributions, and therefore is approximately conditionally valid and efficient. On the other hand, Theorem 1 guarantees the marginal validity of the LSPM under the general IID model, regardless of whether (24) holds.

Comparison with the FDH procedure

In this subsection we discuss a classical procedure that was most clearly articulated by Dempster (1963, p. 110) and Hill (1968, 1988). Both Dempster and Hill trace their ideas to Fisher's (1939, 1948) nonparametric version of his fiducial method. In this paper we refer to this procedure as the *Fisher–Dempster–Hill (FDH) procedure*, although Fisher (1939, p. 4) himself traces this idea back to Student's 1908 paper where he introduced his famous t-test. 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 falls under the rubric of "nonparametric predictive inference".

The FDH procedure 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 $x_n = 0$, $n = 1, 2, \ldots$; in any case, we can ignore the objects. The predictions $\hat{y}$ are always 0, and the hat matrices are $\bar{H} = 0$ and $H = 0$. The LSPM conformity measures become

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

instead of (9), (10), or (13). It is easy to see that the predictive distribution becomes, in the absence of ties (Dempster's and Hill's usual assumption),

$$Q_n(y) := \begin{cases} [\frac{i}{n+1}, \frac{i+1}{n+1}] & \text{if } y \in (y_{(i)}, y_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n\} \\ [\frac{i-1}{n+1}, \frac{i+1}{n+1}] & \text{if } y = y_{(i)} \text{ for } i \in \{1, \ldots, n\} \end{cases} \tag{30}$$

(cf. (17)), where $y_{(1)} \le \cdots \le y_{(n)}$ are the y_i sorted in the increasing order. 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 (30).

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

Since (30) is a conformal transducer (provided a point from an interval in (30) is chosen randomly from the uniform distribution on that interval), we have the same guarantees of validity as those given above: the distribution of (30) 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 FDH ones asymptotically when compared with Oracles I and II. Let us apply the FDH 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 FDH 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.

It is interesting that Theorems 2 and 3 (and therefore the blue and black plots in Figure 1) are applicable both to the LSPM and FDH 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).

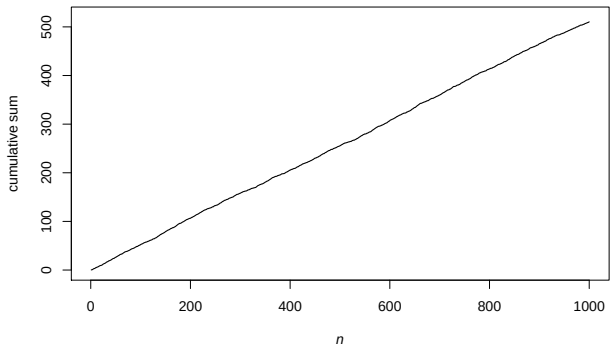

Figure 2: The cumulative sums S_n of the p-values vs $n = 1, \dots, 1000$

Theorem 4. *In the case of the FDH procedure, the random function* $G_n : \mathbb{R} \to [0,1]$ *defined by*

$$G_n(t) := \sqrt{n}\left(Q_n(w + \sigma t, \tau) - Q^{\mathrm{III}}(w + \sigma t)\right) = \sqrt{n}\left(Q_n(w + \sigma t, \tau) - \Phi(t)\right) \tag{31}$$

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

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

The variance $\Phi(t)(1 - \Phi(t))$ of the Brownian bridge is shown as the red line in Figure 1. However, the analogue of the process (31) does not converge in general for the LSPM (under this section's assumption of fixed objects).

6 Experimental results

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

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 $N := 1000$ of IID observations $z_1, \dots, z_N$ and the corresponding p-values $p_n := Q_n(y_{n+1}, \tau_n)$, $n = 1, \dots, N$, 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 $n = 1, \dots, N$; as expected, it

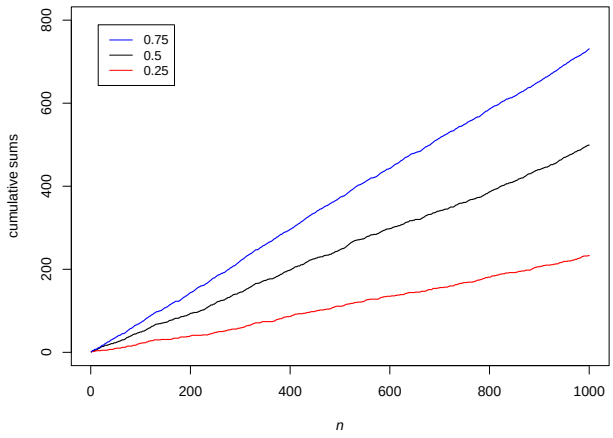

Figure 3: The cumulative sums S_n^α vs $n = 1, \ldots, 1000$ for $\alpha \in \{0.25, 0.5.0.75\}$

is an approximately straight line with slope 0.5. Figure 3 presents three plots: the cumulative sums $S_n^\alpha := \sum_{i=1}^n \mathbf{1}_{\{p_i \le \alpha\}}$, 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, Figure 4 plots A_N^α against $\alpha \in [0, 1]$, where $A_N^\alpha := \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{p_i \le \alpha\}}$. The result is, approximately, the main diagonal of the square $[0, 1]^2$, as it should be.

Efficiency

Next we explore empirically the efficiency of the studentized LSPM. Figure 5 compares the conformal predictive distribution with the true (Oracle III) 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 Figure 6 we show the first plot of Figure 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 Figure 6 is similar except that the training sequence is of length 100 and there are 20 attributes generated independently from the stan-

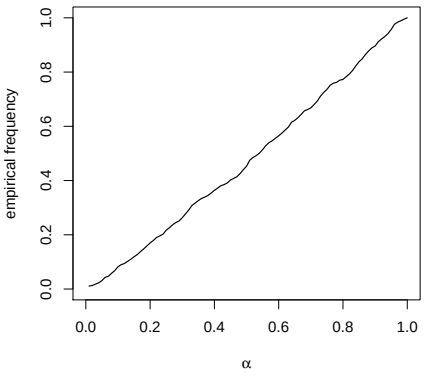

Figure 4: The calibration curve: A_N^α vs $\alpha \in [0,1]$ for $N = 1000$

dard Gaussian distribution except for the first one, which is the dummy all-1 attribute; the labels are generated as before, $y_n \sim 2x_{n,2} + N(0,1)$.

Acknowledgments

This work has been supported by the EPSRC (grant EP/K033344/1), EU Horizon 2020 Research and Innovation programme (grant 671555), US NSF (grant DMS1513483), and Leverhulme Magna Carta Doctoral Centre.

References

Thomas Augustin and Frank P. A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124: 251–272, 2004.

Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Amsterdam, 2014.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 10, February 2014.

Samprit Chatterjee and Ali S. Hadi. *Sensitivity Analysis in Linear Regression*. Wiley, New York, 1988.

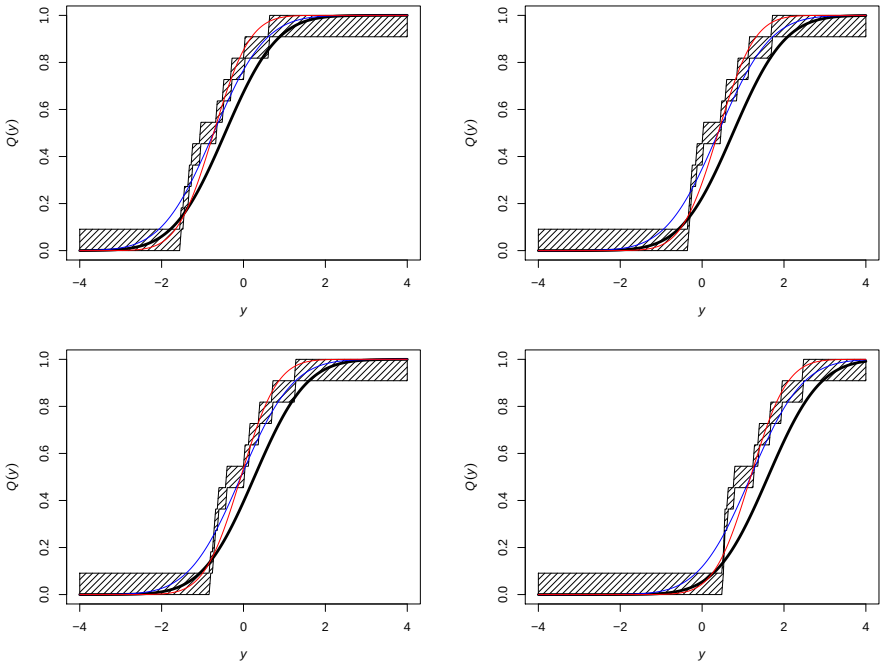

Figure 5: Examples of true predictive distribution functions (black), their conformal estimates (represented by the shaded areas), and the distribution functions output by simplified Oracle I (red) and Oracle II (blue) for a tiny training sequence (of length 10 with two attributes, the first one being the dummy all-1 attribute)

Gemai Chen. *Empirical Processes Based on Regression Residuals: Theory and Applications*. PhD thesis, Department of Mathematics and Statistics, Simon Fraser University, August 1991.

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

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

Ronald A. Fisher. "Student". *Annals of Eugenics*, 9:1–9, 1939.

Ronald A. Fisher. Conclusions fiduciaires. *Annales de l'Institut Henry Poincaré*, 10:191–213, 1948.

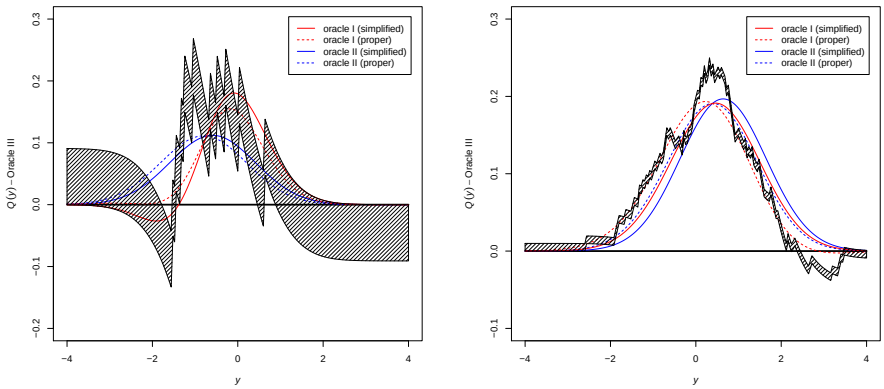

Figure 6: The left-hand plot is the first (upper left) plot of Figure 5 normalized by subtracting the true distribution function (the thick black line in Figure 5, which now coincides with the x-axis) and with the outputs of the proper oracles added; the right-hand plot is an analogous plot for a larger training sequence (of length 100 with 20 attributes, the first one being the dummy attribute)

Christian Genest and Jack Kalbfleisch. Bayesian nonparametric survival analysis: comment. *Journal of the American Statistical Association*, 83:780–781, 1988.

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

Yuri Gurevich and Vladimir Vovk. p-values, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 16, February 2017.

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

Bruce M. Hill. De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In Dennis V. Lindley, José M. Bernardo, Morris H. DeGroot, and Adrian F. M. Smith, editors, *Bayesian Statistics 3*, pages 211–241. Oxford University Press, Oxford, 1988.

Jerald F. Lawless and Marc Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92:529–542, 2005.

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

Lyudmila A. Mugantseva. Testing normality in one-dimensional and multidimensional linear regression. *Theory of Probability and its Applications*, 22: 591–602, 1977. Russian original in: Теория вероятностей и ее применения.

Donald A. Pierce and Kenneth J. Kopecky. Testing goodness of fit for the distribution of errors in regression models. *Biometrika*, 66:1–5, 1979.

Iosif Pinelis. Exact bounds on the closeness between the Student and standard normal distributions. *ESAIM: Probability and Statistics*, 19:24–27, 2015. arXiv:1101.3328v2.

George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley, Hoboken, NJ, second edition, 2003.

Jieli Shen, Regina Liu, and Minge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 2017. To appear.

Student (William S. Gosset). The probable error of a mean. *Biometrika*, 6: 1–25, 1908.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005a.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 1, May 2005b.

C. M. Wang, Jan Hannig, and Hari K. Iyer. Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142:1980–1990, 2012.