On-line predictive linear regression

Vladimir Vovk Ilia Nouretdinov Alex Gammerman

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 #1

First posted May 3, 2005. Last revised March 21, 2020.

Project web site:
http://alrw.net

Abstract

We consider the on-line predictive version of the standard problem of linear regression; the goal is to predict each consecutive response given the corresponding explanatory variables and all the previous observations. We are mainly interested in prediction intervals rather than point predictions. The standard treatment of prediction intervals in linear regression analysis has two drawbacks: (1) the classical prediction intervals guarantee that the probability of error is equal to the nominal significance level ϵ, but this property per se does not imply that the long-run frequency of error is close to ϵ; (2) it is not suitable for prediction of complex systems as it assumes that the number of observations exceeds the number of parameters. We state a general result showing that in the on-line protocol the frequency of error for the classical prediction intervals does equal the nominal significance level, up to statistical fluctuations. We also describe alternative regression models in which informative prediction intervals can be found before the number of observations exceeds the number of parameters. One of these models, which only assumes that the observations are independent and identically distributed, is popular in machine learning but greatly underused in the statistical theory of regression.

Contents

1 Introduction **1**

2 On-line protocol, part I **4**

3 Conformal prediction **6**

4 Data set **9**

5 The IID model **10**

6 The Gauss linear model **14**

7 The MVA model **16**

8 The IID–Gauss model **19**

9 On-line protocol, part II **20**

10 Conclusion **23**

Appendix A: Proofs **24**

Appendix B: Explicit algorithms **26**

Appendix C: Instruction manual for an R package **30**

Appendix D: Goodness of definitions **32**

References **33**

1 Introduction

Let y_n, $n = 1, 2, \ldots$, be the sequence of response variables to be predicted, and let $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,K})$, $n = 1, 2, \ldots$, be the corresponding vectors of explanatory variables. The standard assumption of linear regression analysis is that the explanatory vectors $\mathbf{x}_n$ are deterministic and

$$y_n = \alpha + \boldsymbol{\beta} \cdot \mathbf{x}_n + \xi_n, \tag{1}$$

where α is an unknown coefficient, $\boldsymbol{\beta} \in \mathbb{R}^K$ is an unknown vector of coefficients, and ξ_n, $n = 1, 2, \ldots$, are IID (independent and identically distributed) Gaussian random variables with mean 0 and unknown variance $\sigma^2 > 0$ (we will write $\xi_n \sim N(0, \sigma^2)$). The model (1) will be called the *Gauss linear model*. It is the standard textbook model.

The standard classes of problems associated with the Gauss linear model are parameter estimation, testing hypotheses about parameters, and prediction. In this paper we will be concerned only with prediction, mainly in the form of prediction intervals rather than point predictions. (It is natural to concentrate on prediction as one of the models that we consider, the IID model, is non-parametric.)

A major drawback of the Gauss linear model is that the corresponding prediction intervals are uninformative (i.e., coincide with the whole real line) unless the number of observations exceeds the number of parameters. The responses of a complex system cannot be realistically expected to be modelled using a small number of parameters, whereas the number of observations can be very limited. This motivates consideration of three other models in this paper, none of which requires that the number of observations should exceed the number of parameters.

Perhaps the most important of these models is what we call the *IID model*: it is only assumed that the sequence of pairs $(\mathbf{x}_n, y_n)$ is IID. This model is non-parametric, effectively involving infinitely many parameters. Despite this, the model does allow one to obtain informative prediction intervals. The IID model, however, also has a fundamental limitation: informative prediction intervals become possible only when the number of observations reaches $1/\epsilon$, where ϵ is the chosen significance level.

Our third regression model combines the assumption (1) with the assumption that $\mathbf{x}_n$ are independent (between themselves and of $\xi_1, \xi_2, \ldots$) and identically distributed Gaussian random vectors. We call it the *MVA model*, with MVA referring to "multivariate analysis". It has also been widely discussed in the statistical literature; e.g., Sampson's [16] "two regressions" refers to the Gauss linear model and the MVA model. This model is narrower than both Gauss linear and IID models, and its strong assumptions ensure that informative prediction intervals can be produced almost right away.

Finally, we consider the combination of the Gauss linear and IID models, which we call the *IID–Gauss model*: in addition to (1) we assume that the explanatory vectors $\mathbf{x}_n$, $n = 1, 2, \ldots$, are random and IID (not necessarily Gaussian,

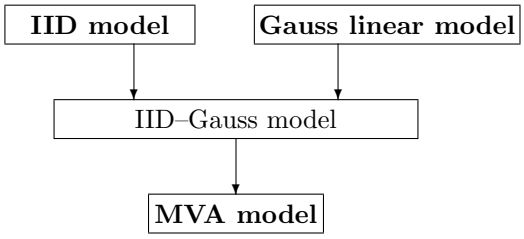

Figure 1: The four models considered in this paper (the three main models are given in boldface).

as in the MVA model) and that the sequence $\xi_1, \xi_2, \ldots$ is independent of the explanatory vectors. This model, however, appears to be of secondary importance. Empirically, it allows informative prediction intervals at significance level ϵ soon after the number of observations exceeds the minimum of $1/\epsilon$ and the number of parameters.

All the models considered in this paper are shown in Figure 1. In this paper we begin (in Section 5) with the IID model. This is the most common model used in modern day statistics and it does not involve the often unrealistic assumption that the noise variables ξ_n are Gaussian or that the explanatory vectors $\mathbf{x}_n$ are Gaussian. An important advantage of the classical Gauss linear model, considered in Section 6, is that the explanatory vectors are not assumed to be IID (in other words, no "random design" is assumed). This model is essentially equivalent to making no assumptions whatsoever about the distribution of $\mathbf{x}_n$ and assuming that the ξ_n in (1) are IID and distributed as $N(0, \sigma^2)$ conditional on $\mathbf{x}_1, \mathbf{x}_2, \ldots$. The Gauss linear model (understood in this way) and the IID model are not comparable between themselves, but both contain the other two models: the IID–Gauss model (Section 8), which is the intersection of the IID and Gauss linear models, and the MVA model (Section 7), which makes the further assumption that the explanatory vectors are Gaussian. These relationships are shown in Figure 1 with the arrows leading from more general to more specific models.

Fisher ([9], Section IV.3) emphatically defended the use of the Gauss linear model even in the case where the distribution of the explanatory vectors is known (with or without parameters). There is also a view in the literature that the Gauss linear model and the MVA model are "essentially equivalent" (for a review of some results in this direction, see [16]). Our conclusion, however, is similar to Brown's [2]: when the MVA model is true, it can be far more useful for prediction; in particular, it can start giving informative prediction intervals long before the number of observations reaches the number of parameters K (or the inverse significance level $1/\epsilon$).

This paper uses a general method of prediction called conformal prediction. The method is reviewed in detail in the monograph [21] and introduced in the

work leading up to that monograph. For each of the four models in Figure 1 we define a suitable confidence predictor, i.e., a strategy for producing prediction intervals or, more generally, prediction regions. For the IID model we follow [21] and for the Gauss linear model we use Fisher's classical confidence predictor. The confidence predictors for the MVA and IID–Gauss models are new.

We are interested in two criteria of quality of confidence predictors, which we call "validity" and "accuracy". For valid confidence predictors, the probability of error equals the nominal significance level ϵ (or at least never exceeds ϵ, in which case we will refer to them as "conservatively valid", or just "conservative", confidence predictors). (We will not be interested in approximate versions of validity and conservative validity; the main problem with such approximate versions is that it is typically difficult to describe conditions under which the approximation is good or tolerable.) The second criterion is applied only to valid confidence predictors: we want the prediction intervals to be as narrow as possible; in this paper we, somewhat arbitrarily, measure the narrowness of a prediction interval $[a, b]$ by its length $b-a$. In particular, we want the prediction intervals to become bounded as soon as possible.

Correspondingly, this paper uses two kinds of entities that one might want to call "models". The first kind is "hard models", such as the four models in Figure 1. These are the usual statistical models: our working hypothesis is that the data set was generated by one of the probability distributions in the model. In particular, the validity of our confidence predictors is allowed to depend on the hard model. By default, the word "model" means "hard model".

In addition to the accepted hard model, one often has other *a priori* information about the data-generating distribution: e.g., only a few parameters might provide the bulk of the information relevant to prediction. Whereas we might hesitate to include such *a priori* information in the hard model explicitly, since it might destroy the validity of our confidence predictor if this information happened to be far from the truth, we might still be able to use such information in designing accurate confidence predictors provided our model is flexible enough. A running example in this paper, introduced in Section 4, will be a linear system with 100 parameters ten of which are felt to be especially important. This will be our "soft model" (not defined formally); whether it is true or not affects only the accuracy, but not validity, of our confidence predictors.

Separation of the available information about the data-generating distribution into the hard model and soft model increases robustness of confidence predictors with respect to modelling errors. If such an error occurs in the soft model, the validity of predictions is not affected. At worst the predictions will become useless, but they will not become misleading (with high probability under any distribution in the hard model). For a further discussion and empirical study, see [11], Section 4.

The property of validity of conformal predictors can be stated in an especially strong form in the on-line prediction protocol. It turns out that the true responses fall outside the corresponding prediction regions independently for different observations. In combination with the law of large numbers this implies that, with high probability, the frequency of error is approximately equal

to the nominal significance level. For the classical prediction intervals in the Gauss linear model this property had been known (as we will see in Section 6) but its importance had not been emphasized prior to the work leading up to [21].

For reviews of the theory of conformal prediction see [11] and [18]. Parts of these two papers are devoted to regression problems.

Section 2 formally introduces the on-line prediction protocol, with a more detailed discussion postponed until Section 9. In Section 3 we describe the method of conformal prediction and state two key results (proved in Appendix A): one asserts the strong validity and the other universality of conformal predictors. Section 4 describes an artificial data set used in later sections for illustrating the performance of various conformal predictors. The following four sections, 5–8, apply the method of conformal prediction to the IID, Gauss linear, MVA, and IID–Gauss models, in this order. Section 10 concludes.

This version of the paper is an extension of the journal version [22]. In addition to the journal version, it contains a discussion of Fisher's fiducial prediction (the first subsection of Section 9) and three extra appendices, B (containing explicit algorithms), C (describing an R package implementing those algorithms), and D.

2 On-line protocol, part I

In our prediction protocol, the task is to sequentially predict y_n, $n = 1, 2, \ldots$, from $\mathbf{x}_n$ and $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n-1$. This on-line protocol is popular in machine learning (see, e.g., [3] and references therein), but most statistical research (except some work on sequential analysis) is still done in the "off-line", or "batch", framework, where one starts from a complete sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. One of the few statisticians advocating the on-line protocol (under the name "prequential", or predictive sequential) has been Philip Dawid [6].

Weak and strong validity and median accuracy

To explain what precisely we mean by validity and accuracy, the two criteria of predictive performance mentioned in Section 1, we will need the notation introduced in the following description of the on-line prediction protocol.

On-line prediction protocol

FOR $n = 1, 2, \ldots$:
  Predictor observes $\mathbf{x}_n \in \mathbb{R}^K$;
  Predictor outputs $\Gamma_n^\epsilon \subseteq \mathbb{R}$ for all $\epsilon \in (0, 1)$;
  Predictor observes $y_n \in \mathbb{R}$;
  $\mathrm{err}_n^\epsilon := \mathbb{I}_{y_n \notin \Gamma_n^\epsilon}$ for all $\epsilon \in (0, 1)$;
  $L_n^\epsilon := \sup \Gamma_n^\epsilon - \inf \Gamma_n^\epsilon$ for all $\epsilon \in (0, 1)$
END FOR.

(As usual, $\mathbb{I}_F$ is defined to be 1 if the condition F holds and 0 if not.) At each step and for each significance level ϵ, Predictor outputs a *prediction region* (usually, although not necessarily, an interval) $\Gamma_n^\epsilon \subseteq \mathbb{R}$. We require that, for all n, the family Γ_n^ϵ of prediction regions should be nested: $\Gamma_n^{\epsilon_1} \subseteq \Gamma_n^{\epsilon_2}$ whenever $\epsilon_1 > \epsilon_2$. An error is registered, $\mathrm{err}_n^\epsilon = 1$, if the prediction region fails to contain the true response y_n, and the accuracy of this particular prediction is measured by the length L_n^ϵ of the corresponding *prediction interval* $\mathrm{co}\,\Gamma_n^\epsilon$ ($\mathrm{co}\,E$ standing for the convex hull of the set E).

Let $\mathrm{Err}_n^\epsilon := \mathrm{err}_1^\epsilon + \cdots + \mathrm{err}_n^\epsilon$ be the cumulative number of errors made up to, and including, step n. In the following sections, we will find it convenient to distinguish between two notions of validity, "weak validity" and "strong validity".

Definition 1. A *confidence predictor* is defined to be a measurable prediction strategy $\Gamma_n^\epsilon = \Gamma^\epsilon(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n)$ in the on-line prediction protocol.

Definition 2. A confidence predictor is *weakly valid* in some statistical model if the probability that $\mathrm{err}_n^\epsilon = 1$ is ϵ, for each $\epsilon \in (0,1)$ and each n under any probability distribution in the model.

The definition of weak validity is standard: cf. [5], (75) on p. 243. Weak validity by itself does not imply that Err_n / n is likely to be close to ϵ for large n.

Definition 3. A confidence predictor is *strongly valid* if it is weakly valid and, for each $\epsilon \in (0,1)$, the events $\mathrm{err}_n^\epsilon = 1$, $n = 1, 2, \ldots$, are independent.

Figure 3 below shows the plot of Err_n^ϵ against n for a specific confidence predictor considered in this paper; it is typical of our predictors that the slopes of the plots of Err_n^ϵ are close to the corresponding significance levels ϵ (we use the significance levels 5%, 1% and 0.5% in all our figures, represented by the corresponding *confidence levels* $1 - \epsilon$ in the legends). This is the only figure in this paper illustrating the validity of our confidence predictors: such figures, in view of the mathematical results guaranteeing validity, tend to be uninformative.

We will measure the accuracy of the predictions made for the first n observations by the median M_n^ϵ of the sequence $L_1^\epsilon, \ldots, L_n^\epsilon$; again, this measure is arbitrary, to a large degree. A plot of M_n^ϵ against n will be called the *median-accuracy plot*; examples of such plots are given in Figures 2 and 4–6.

Unfortunately, the simple notions of validity introduced earlier have to be extended to become useful for our purpose. This is needed because, e.g., the classical prediction intervals are uninformative before the number of observations reaches the number of parameters, and so for small n the error probability is zero rather than ϵ. Let $\mathcal{N}$ be a set of positive integer numbers (we are mainly interested in the case where $\mathcal{N}$ has the form $\{m, m+1, \ldots\}$).

Definition 4. We say that a confidence predictor is *weakly valid for* $n \in \mathcal{N}$ in a statistical model if the probability is ϵ that it makes an error, $\mathrm{err}_n^\epsilon = 1$, at step n under any probability distribution in the model and for all $n \in \mathcal{N}$ and $\epsilon \in (0,1)$. It is *strongly valid for* $n \in \mathcal{N}$ if, in addition, err_n^ϵ, $n \in \mathcal{N}$, are independent for any fixed ϵ.

The role of the on-line protocol

The exposition of this paper is based on the on-line protocol, but the majority of our findings are not constrained to this specific protocol. For example, the fact that valid and informative prediction intervals can become feasible in the MVA model before the number of observations exceeds the number of parameters does not depend on the prediction protocol. In the absence of the on-line protocol, however, "validity" should be understood in the standard sense of weak validity.

3 Conformal prediction

In this section we define a class of confidence predictors, called conformal predictors, and state results about their validity and universality, in a certain sense.

Notions of sufficiency

Fix some *observation space* Z. We will be interested in the space $Z = \mathbb{R}^K \times \mathbb{R}$ of pairs $(\mathbf{x}, y)$; in general, Z is a measurable space assumed to be Luzin, to ensure the existence of regular conditional probabilities. To define conformal predictors, we will need not only a statistical model on Z^∞ but also a sequence of sufficient statistics $S_n : Z^n \to \Sigma_n$, $n = 1, 2, \ldots$; we will always assume that $\Sigma_n = S_n(Z^n)$. We will need a strengthened form of sufficiency; in our definitions we mainly follow Lauritzen [12], Section II.2.

The sequence (S_n) is *algebraically transitive* if there exists a sequence of measurable functions $F_n : \Sigma_{n-1} \times Z \to \Sigma_n$, $n = 2, 3, \ldots$, such that

$$S_n(\zeta_1, \ldots, \zeta_{n-1}, \zeta_n) = F_n(S_{n-1}(\zeta_1, \ldots, \zeta_{n-1}), \zeta_n)$$

for all $(\zeta_1, \ldots, \zeta_{n-1}, \zeta_n) \in Z^n$. Intuitively, $S_n(\zeta_1, \ldots, \zeta_n)$ is the summary of the first n observations, and the condition of algebraic transitivity means that the summary can be updated on-line.

The sequence (S_n) is *totally sufficient* for a statistical model $\mathcal{P}$ on Z^∞ if, for each $n = 1, 2, \ldots$:

- S_n is sufficient for $\mathcal{P}$;
- $\zeta_1, \ldots, \zeta_n$ and $\zeta_{n+1}, \zeta_{n+2}, \ldots$ are conditionally independent given $S_n(\zeta_1, \ldots, \zeta_n)$, where $(\zeta_1, \zeta_2, \ldots) \sim P$, for any $P \in \mathcal{P}$.

The second condition ensures that $S_n(\zeta_1, \ldots, \zeta_n)$ carries all information in $\zeta_1, \ldots, \zeta_n$ that can be used for predicting the future observations $\zeta_{n+1}, \zeta_{n+2}, \ldots$.

A sequence of statistics that is both algebraically transitive and totally sufficient will be called an *ATTS sequence*. In the rest of this paper we will often say "model" to mean a statistical model $\mathcal{P}$ equipped with an ATTS sequence (S_n). This makes the word "model" ambiguous as we often omit "statistical" in "statistical model", but this should not lead to misunderstandings.

Each of the four statistical models considered in this paper (see Figure 1) will be complemented with an ATTS sequence; in all four cases the observation space Z will be $\mathbb{R}^K \times \mathbb{R}$.

Testing conformity

The main ingredient of conformal prediction is statistical testing of conformity of a new observation ζ_n to the old observations $\zeta_1, \dots, \zeta_{n-1}$. In general, our statistical tests will be randomized.

Fix a statistical model $\mathcal{P}$ with an ATTS sequence $S_n : Z^n \to \Sigma_n$. Define Σ_0 to be a fixed one-element set. Any sequence of measurable functions $A_n : \Sigma_{n-1} \times Z \to \mathbb{R}$, $n = 1, 2, \dots$, is called a *nonconformity measure*; A_n will be our test statistics. Given a nonconformity measure (A_n), for each sequence $\zeta_1, \zeta_2, \dots$ of observations and each sequence $\tau_1, \tau_2, \dots \in [0,1]^\infty$ we define the *p-values*

$$p_n = p_n(\zeta_1, \dots, \zeta_n, \tau_n) := \\ \mathbb{P}\left(A_n^{\mathrm{rnd}} > A_n^{\mathrm{obs}} \mid S_n^{\mathrm{rnd}} = S_n^{\mathrm{obs}}\right) + \tau_n \, \mathbb{P}\left(A_n^{\mathrm{rnd}} = A_n^{\mathrm{obs}} \mid S_n^{\mathrm{rnd}} = S_n^{\mathrm{obs}}\right), \\ n = 1, 2, \dots, \quad (2)$$

where $A_n^{\mathrm{rnd}} := A_n(S_{n-1}(\xi_1, \dots, \xi_{n-1}), \xi_n)$ and $S_n^{\mathrm{rnd}} := S_n(\xi_1, \dots, \xi_n)$ are the "random" values, $A_n^{\mathrm{obs}} := A_n(S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n)$ and $S_n^{\mathrm{obs}} := S_n(\zeta_1, \dots, \zeta_n)$ are the "observed" values, and the probabilities are taken with respect to $(\xi_1, \xi_2, \dots) \sim P$ for some $P \in \mathcal{P}$. Since S_n are sufficient statistics, p_n do not depend on $P \in \mathcal{P}$ (at least for a suitable choice of regular conditional probabilities). We will be interested in two cases: *deterministic*, where $\tau_n = 1$ for all n, and *randomized*, where $\tau_1, \tau_2, \dots$ are generated independently from the uniform distribution U on $[0,1]$ (such $\tau_1, \tau_2, \dots$ model the output of a random numbers generator).

Theorem 1. *Suppose that the sequence of observations $(\zeta_1, \zeta_2, \dots) \in Z^\infty$ is generated from a probability distribution $P \in \mathcal{P}$ and that the random numbers $(\tau_1, \tau_2, \dots) \sim U^\infty$ are independent of the observations. The p-values (2) are then independent and distributed uniformly on $[0,1]$:*

$$(p_1, p_2, \dots) \sim U^\infty.$$

For a proof of this theorem, see Appendix A. The fact that $p_n \sim U$ is well known, at least in the continuous case (see, e.g., [5], p. 66; (2) is a version of (1) in [5]).

Conformal prediction

We start by extending, and spelling out in a greater detail, the notion of a confidence predictor: in the general theory of this section and in its application to the IID model in Section 5 we will need an element (typically quite small) of randomization in confidence predictors.

Definition 5. A *randomized confidence predictor* is a measurable function which maps every significance level $\epsilon \in (0,1)$, every data sequence $\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}$, every vector $\mathbf{x}_n$ of explanatory variables, and every number $\tau \in [0,1]$ to a set $\Gamma_n^\epsilon = \Gamma^\epsilon(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n, \tau) \subseteq \mathbb{R}$. We will use the notation Γ_n^ϵ when

the data sequence, the vector of explanatory variables, and the number τ are clear from the context.

Let the observation space be $Z = \mathbb{R}^K \times \mathbb{R}$. Once the p-values (2) are defined, we can use them for confidence prediction (this is a standard procedure; cf. [5], (76) on p. 243): we set

$$\Gamma^\epsilon(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n, \tau_n) := \{y \in \mathbb{R} : p_n((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n, y), \tau_n) > \epsilon\} . \quad (3)$$

Definition 6. The randomized confidence predictor defined by (3) is called the *smoothed conformal predictor determined by* the nonconformity measure (A_n). A *smoothed conformal predictor* is a smoothed conformal predictor determined by some nonconformity measure.

The following statement immediately follows from Theorem 1 and asserts that smoothed conformal predictors are strongly valid.

Corollary 1. *If the sequence of observations $(\mathbf{x}_n, y_n)$, $n = 1, 2, \dots$, is generated by a probability distribution $P \in \mathcal{P}$ and a smoothed conformal predictor is fed with random numbers $(\tau_1, \tau_2, \dots) \sim U^\infty$ independent of the observations, the error sequence $\mathrm{err}_1^\epsilon, \mathrm{err}_2^\epsilon, \dots$ at any significance level ϵ is a sequence of IID Bernoulli random variables with parameter ϵ.*

The adjective "smoothed" refers to using random numbers; if we take $\tau_n = 1$ for all $n = 1, 2, \dots$, we will obtain the definition of a "deterministic conformal predictor", or just "conformal predictor", and in this case we omit τ_n from our notation.

Definition 7. A *conformal predictor* is the confidence predictor defined by

$$\Gamma^\epsilon(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n) := \{y \in \mathbb{R} : p_n((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n, y), 1) > \epsilon\} ,$$

where the p-values p_n are defined by (2).

Notice that when a conformal predictor makes an error, the corresponding smoothed conformal predictor also makes an error. In combination with Corollary 1, we can see that conformal predictors are *conservative*, in the sense that, for each ϵ, their error sequence $\mathrm{err}_1^\epsilon, \mathrm{err}_2^\epsilon, \dots$ is dominated by a sequence of IID Bernoulli random variables with parameter ϵ. In particular, whereas we have $\lim_{n\to\infty}(\mathrm{Err}_n^\epsilon / n) = \epsilon$ a.s. for smoothed conformal predictors, we only have $\limsup_{n\to\infty}(\mathrm{Err}_n^\epsilon / n) \le \epsilon$ a.s. for conformal predictors.

We will see that there is no difference between conformal predictors and the corresponding smoothed conformal predictors for the Gauss linear model and $n \ge K + 3$ since the second addend on the right-hand side of (2) is then zero. There is also no difference for the MVA model and $n \ge 3$; however, the difference

is important (although usually barely noticeable on error and accuracy plots) for the IID model.

A natural question is whether there are other ways to achieve validity, except conformal prediction. The following theorem will give a negative answer to a version of this question.

Definition 8. A confidence predictor Γ is *invariant* if Γ_n^ϵ, $n > 1$, depends on the first $n-1$ observations only through the value of S_{n-1} on those observations.

The use of invariant confidence predictors is natural in view of the sufficiency principle; see, e.g., [5], Section 2.3 (ii). Let $\mathcal{N}$ be a set of positive integers. We say that a confidence predictor $\Gamma^\dagger$ is *at least as accurate as* another confidence predictor Γ for $n \in \mathcal{N}$ if

$$(\Gamma^\dagger)^\epsilon(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n) \subseteq \Gamma^\epsilon(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n)$$

for all ϵ, all $n \in \mathcal{N}$, and P-almost all $\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n$, under any probability distribution $P \in \mathcal{P}$.

Recall that a statistic S taking values in a measurable space Σ is said to be *boundedly complete* (with respect to the statistical model $\mathcal{P}$) if, for any bounded measurable function $f : \Sigma \to \mathbb{R}$, the following condition is satisfied: the expected value $\mathbb{E}_P(f(S))$ of $f(S)$ is zero under all $P \in \mathcal{P}$ only if $f(S) = 0$ P-almost surely for all $P \in \mathcal{P}$.

Theorem 2. *Let $\mathcal{N}$ be a set of positive integers. Suppose the ATTS statistics S_n are boundedly complete for $n \in \mathcal{N}$. If a confidence predictor Γ is invariant and weakly valid for $n \in \mathcal{N}$, then there is a conformal predictor that is at least as accurate as Γ for $n \in \mathcal{N}$.*

This theorem is also proved in Appendix A. An important step towards its proof was made by Takeuchi ([19], p. 31).

The condition of bounded completeness holds for the Gauss linear model and the MVA model by the standard completeness result for exponential statistical models (see, e.g., Theorem 4.1 in [13]), and it is also known to hold for the IID model (see the theorem on p. 797 in [1] or Theorem 1 in [14]).

4 Data set

We will illustrate the accuracy of various confidence predictors using the following artificially generated data set with 600 observations and $K = 100$ explanatory variables. The components $x_{n,k}$ of $\mathbf{x}_n$ are independently generated from $N(0,1)$, and the responses y_n are generated according to (1) with $\xi_n \sim N(0,1)$ independent between themselves and of all $x_{n,k}$, with $\alpha = 100$ and with the following components β_k of $\boldsymbol{\beta}$:

$$\beta_k := \begin{cases} (-1)^{k-1}10 & k = 1, \dots, 10 \\ (-1)^{k-1} & k = 11, \dots, 100. \end{cases}$$

Table 1: Steps at which informative prediction becomes possible for the four models; ϵ is the significance level ($\epsilon < 1/2$ is assumed) and K is the number of parameters.

Model	The first step at which prediction intervals can become informative
IID model	$\lceil 1/\epsilon \rceil$
Gauss linear model	$K+3$
MVA model	3
IID–Gauss model	$\min(\lceil 1/\epsilon \rceil, K+3)$

The probability distribution generating this data set belongs to all four models considered in this paper (Figure 1). It is natural to expect that more specific models, when true, will lead to better predictions. In one respect this is true: more general models allow informative predictions later, as shown in Table 1 (to be explained in later sections). However, soon after the threshold given in the table is reached, the quality of prediction becomes very similar on our data set.

The (informal) soft model guiding the choice of the nonconformity measure will include the assumption of linearity (1) and the knowledge, or guess, that the first 10 explanatory variables are much more important than the rest.

Relationship (1) between the response and explanatory variables can be written as

$$y_n = \boldsymbol{\gamma} \cdot \mathbf{z}_n + \xi_n, \tag{4}$$

where

$$\boldsymbol{\gamma} := \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} \in \mathbb{R}^{K+1} \text{ and } \mathbf{z}_n := \begin{pmatrix} 1 \\ \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{K+1}.$$

For $l = 1, 2, \ldots$, let $\boldsymbol{Z}_l$ be the $l \times (K+1)$ matrix whose rows are $\mathbf{z}_i'$, $i = 1, \ldots, l$, and $\mathbf{y}_l$ be the vector whose ith element is y_i, $i = 1, \ldots, l$. We will sometimes refer to the first column of $\boldsymbol{Z}_l$ as the *dummy* column.

5 The IID model

The statistical model considered in this section is non-parametric: we simply assume that the observations $(\mathbf{x}_n, y_n)$ are IID. Notice that this does not involve the assumption of linearity of the "true" regression function or the assumption of a Gaussian noise. Linearity is, however, an important component of the soft model used for choosing a suitable nonconformity measure.

The ATTS statistics are

$$S_n := \wr(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\int,$$

where we use $\wr a_1, \dots, a_n \int$ to denote the bag, or multiset, consisting of $a_1, \dots, a_n$ (some of these elements may coincide). For each n, the conditional distribution of $(\xi_1, \dots, \xi_n)$ given that

$$\wr \xi_1, \dots, \xi_n \int = \wr (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \int,$$

where ξ_i are IID random elements taking values in $\mathbb{R}^K \times \mathbb{R}$, coincides (with probability one) with the probability distribution on the orderings $(\mathbf{x}_{\pi(1)}, y_{\pi(1)}), \dots, (\mathbf{x}_{\pi(n)}, y_{\pi(n)})$ of the bag $\wr (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \int$ that arises when different permutations π are chosen with the same probability $1/n!$.

The IID model is typical in that there is a great flexibility in choosing a nonconformity measure for use in conformal prediction. Suppose, e.g., that the number of explanatory variables K is too large for us to estimate all the β_k and α in the soft model (1). We believe, however, that the first $K_n^\dagger \ll K$ of the explanatory variables are especially important, and it is feasible to estimate the corresponding β_k, $k = 1, \dots, K_n^\dagger$, and α.

Fix temporarily a positive integer number n. We will write $\mathbf{y}$ for $\mathbf{y}_n$, $\boldsymbol{Z}$ for $\boldsymbol{Z}_n$ and $K^\dagger$ for $K_n^\dagger$. Let $\boldsymbol{U}$ be the submatrix of $\boldsymbol{Z}$ consisting of the first $K^\dagger + 1$ columns of $\boldsymbol{Z}$: those that correspond to the explanatory variables deemed to be useful at this stage plus the dummy column $\mathbf{1}$. To test the conformity of the nth observation to the first $n-1$ observations, we will first fit a hyperplane to all n observations using the relevant explanatory variables. Applying a small "ridge coefficient" $a > 0$ to avoid the need to invert singular matrices, we obtain the vector of residuals

$$\mathbf{e} := \mathbf{y} - \boldsymbol{U} \left(\boldsymbol{U}' \boldsymbol{U} + a \boldsymbol{I} \right)^{-1} \boldsymbol{U}' \mathbf{y}, \tag{5}$$

whose components will be denoted $e_1, \dots, e_n$.

We will be interested in the conformal predictor determined by the nonconformity measure

$$A_n \left(S_{n-1} \left(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1} \right), (\mathbf{x}_n, y_n) \right) := |e_n|. \tag{6}$$

Deleted and, especially, studentized residuals would also be a natural choice (see, e.g., [21], pp. 34–35). In our experience, however, the difference is not significant, and we stick to the simplest choice. The confidence predictor obtained from this conformal predictor by replacing the prediction regions Γ_n^ϵ with the prediction intervals $\operatorname{co} \Gamma_n^\epsilon$ will be called the *IID predictor* (cf. the comments at the end of this section).

The IID predictor can be implemented fairly efficiently. First notice that for the IID model the formula (2) for p-values can be simplified to

$$p_n = \frac{|\{i : \alpha_i > \alpha_n\}| + \tau_n \, |\{i : \alpha_i = \alpha_n\}|}{n}, \tag{7}$$

where $\alpha_i := A_n(\wr \zeta_1, \dots, \zeta_{i-1}, \zeta_{i+1}, \dots, \zeta_n \int, \zeta_i)$, i ranges over $\{1, \dots, n\}$, and $|E|$ stands for the size of the set E. In the case of the nonconformity measure (6), $\alpha_i = |e_i|$. The residuals (5) can be written in the form

$$\mathbf{e} = \mathbf{y} - \boldsymbol{U} \left(\boldsymbol{U}' \boldsymbol{U} + a \boldsymbol{I} \right)^{-1} \boldsymbol{U}' \mathbf{y} = \boldsymbol{C} \mathbf{y},$$

where $\boldsymbol{C}$ is the matrix $\boldsymbol{I} - \boldsymbol{U}\left(\boldsymbol{U}'\boldsymbol{U} + a\boldsymbol{I}\right)^{-1}\boldsymbol{U}'$, not depending on the response variables. If we fix the first $n-1$ response variables y_i and vary the last one, y, the residuals $e_i = e_i(y)$, $i = 1, \dots, n$, become linear functions of y (this fact will also be used in Section 7). By (7) with $\tau_n := 1$, the p-value is the fraction of $i = 1, \dots, n$ satisfying $|e_i(y)| \geq |e_n(y)|$; therefore, as y varies from $-\infty$ to ∞, the p-value can change only at the at most $2n-2$ points (called *critical* points) which are solutions to the linear equations $e_i(y) = e_n(y)$ and $e_i(y) = -e_n(y)$. This divides the real line into at most $4n-3$ intervals: the critical points, considered as degenerate closed intervals, the open intervals bounded on both sides by adjacent critical points, and the two unbounded open intervals to the left of the leftmost critical point and to the right of the rightmost critical point; if there are no critical points, this collapses into one unbounded open interval $\mathbb{R}$. We can compute the p-value for one point in each of these intervals and then compute Γ_n^ϵ as the union of the intervals with p-values exceeding ϵ. The computation of the IID prediction interval $\operatorname{co}\Gamma_n^\epsilon$ can be simplified if we notice that the set Γ_n^ϵ is closed (which is opposite to what we will have for the Gauss linear and MVA models): assuming that the set of critical points is non-empty, $\operatorname{co}\Gamma_n^\epsilon$ is bounded if and only if the two unbounded intervals have p-values at most ϵ, in which case the end-points of $\operatorname{co}\Gamma_n^\epsilon$ can be found as the leftmost and rightmost critical points with p-values exceeding ϵ. Computing Γ_n^ϵ and $\operatorname{co}\Gamma_n^\epsilon$ from scratch (e.g., without using the results of computations from the previous steps of the on-line protocol) takes time $O(n \log n)$ (see [21], p. 33).

For use in our experiments with the artificial data set described in Section 4, we take

$$K_n^\dagger := \begin{cases} 10 & \text{if } n < 103 \\ 100 & \text{otherwise,} \end{cases} \tag{8}$$

and so define $\boldsymbol{U}$ as the first 11 columns of $\boldsymbol{Z}$ if $n < 103$ and as the full $\boldsymbol{Z}$ otherwise. Our chosen value for the threshold, 103, appeared to us slightly less arbitrary than other choices, since it is the first step when the classical prediction intervals (see (10) below) become bounded. However, the quality of the estimates of α and the 100 components of $\boldsymbol{\beta}$ is still poor when n is close to 103. This affects the quality of our prediction intervals but does not show on the median-accuracy plots. The value of the ridge coefficient is always $a = 0.01$.

As Figure 2 shows, the IID predictor works well for our data set if the significance level is not too demanding: it can be seen from (7) (with $\tau_n := 1$) that for the IID prediction interval $\operatorname{co}\Gamma_n^\epsilon$ to be bounded the number of observations n has to be at least $1/\epsilon$ (as Table 1 says). For example, for the significance level $\epsilon = 0.5\%$, the IID predictor requires 200 observations to produce bounded predictions, and this shows on the median-accuracy plot at $n = 399$ (since for $n < 399$ at least half of the observed prediction intervals are infinitely wide).

The IID model is non-parametric but we can see that it still admits valid confidence predictors (or conservative confidence predictors if one insists on using deterministic predictors). The threshold $1/\epsilon$ can be said to play the role of the number of parameters, and the non-parametric nature of the model is reflected in the fact that $1/\epsilon \to \infty$ as $\epsilon \to 0$. Since $1/\epsilon$ tends to ∞ relatively

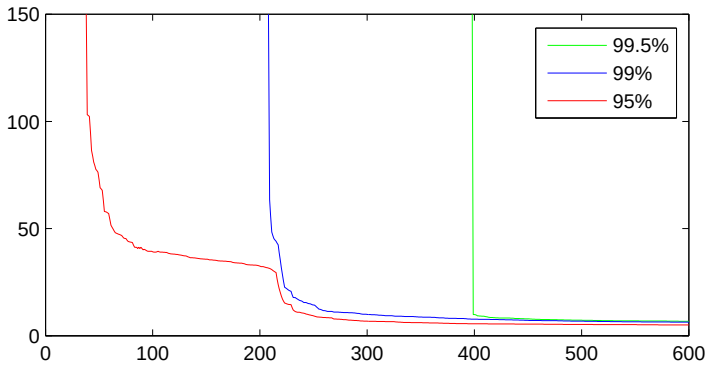

Figure 2: The median-accuracy plot for the IID predictor. The three significance levels used in this and all the following figures are $\epsilon = 0.05, 0.01, 0.005$, shown in the form $100(1-\epsilon)\%$ (the corresponding confidence levels) in the legends.

slowly, such an infinite-dimensional model may be better for the purpose of prediction than a K-dimensional model with a very large K.

Theorem 2 is not directly applicable to the IID model, since only smoothed conformal predictors are valid, as the latter term is used in this paper. Two results of the same nature about the IID model are stated in [21], Section 2.4.

There are two sources of conservativeness for the IID predictor as described above (and used for producing Figure 2). First, we used a deterministic predictor (taking $\tau_n = 1$ for all n), and second, we replaced each prediction region by its convex hull. Our experiments (see, e.g., Figure 3) show that we still have approximate validity.

For each model considered in this paper except the Gauss linear model we define a nonconformity measure involving the matrix $\boldsymbol{U}$ defined earlier in this section. In the case of the IID model, we have used the nonconformity measure (6) and called the corresponding conformal predictor with Γ_n^ϵ replaced by $\operatorname{co}\Gamma_n^\epsilon$ the IID predictor (it was called "Ridge Regression Confidence Machine" in [21]). Of course, our brief term is somewhat misleading: it should always be borne in mind that the conformal predictor leading to the IID predictor is only one of many conformal predictors that can be defined in the IID model. Similarly, in the following three sections we will introduce the Gauss predictor, the MVA predictor, and the IID–Gauss predictor, which will also correspond to specific nonconformity measures.

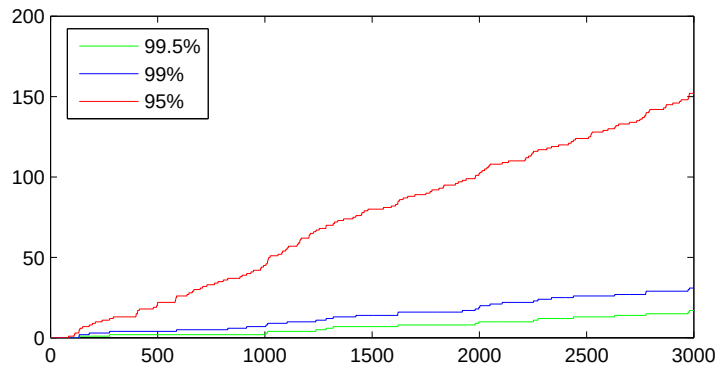

Figure 3: The cumulative numbers of errors made by the IID predictor: Err_n^ϵ is plotted against n.

6 The Gauss linear model

Let $\hat{\boldsymbol{\gamma}}_l := (\boldsymbol{Z}_l'\boldsymbol{Z}_l)^{-1}\boldsymbol{Z}_l'\mathbf{y}_l$ be the least-squares estimate of the parameter vector $\boldsymbol{\gamma}$ in (4) from the first l observations. For simplicity, we will assume that the matrix $\boldsymbol{Z}_l$ has full rank (i.e., rank $\boldsymbol{Z}_l = \min(l, K+1)$) for all l; this implies that $\hat{\boldsymbol{\gamma}}_l$ is well defined for $l \geq K+1$.

Let $\hat{y}_n$ be the least-squares prediction $\hat{\boldsymbol{\gamma}}_{n-1} \cdot \mathbf{z}_n$ for y_n and

$$\hat{\sigma}_l^2 := \frac{1}{l-K-1}(\mathbf{y}_l - \boldsymbol{Z}_l\hat{\boldsymbol{\gamma}}_l)'(\mathbf{y}_l - \boldsymbol{Z}_l\hat{\boldsymbol{\gamma}}_l)$$

be the standard estimate of σ^2 from $\boldsymbol{Z}_l$ and $\mathbf{y}_l$. It is well known that in the Gauss linear model the ratio

$$T_n := \frac{y_n - \hat{y}_n}{\sqrt{1 + \mathbf{z}_n'(\boldsymbol{Z}_{n-1}'\boldsymbol{Z}_{n-1})^{-1}\mathbf{z}_n}\hat{\sigma}_{n-1}}, \quad n = K+3, K+4, \ldots, \tag{9}$$

has the t-distribution with $n-K-2$ degrees of freedom. This gives the classical weakly valid prediction interval for the nth response,

$$\Gamma_n^\epsilon := \left\{ y \in \mathbb{R} : |y - \hat{y}_n| < t_{n-K-2}^{\epsilon/2}\sqrt{1 + \mathbf{z}_n'(\boldsymbol{Z}_{n-1}'\boldsymbol{Z}_{n-1})^{-1}\mathbf{z}_n}\hat{\sigma}_{n-1} \right\}, \quad n \geq K+3, \tag{10}$$

where t_m^δ is the upper δ point of the t-distribution with m degrees of freedom. (See, e.g., [17], (5.27).) We set Γ_n^ϵ to $\mathbb{R}$ when $n < K+3$.

Later in this section we will see that Corollary 1 implies the following property of the classical prediction intervals for the Gauss linear model.

Corollary 2. *Let $\epsilon \in (0,1)$. The events $y_n \notin \Gamma_n^\epsilon$, $n = K+3, K+4, \dots$, are independent. In particular, the confidence predictor (10) is strongly valid for $n \geq K+3$.*

Remark 1. Corollary 2 and, more generally, the fact that the statistics (9) are independent was established in [7] following [15]. It is interesting that both papers use the independence of (9) for testing rather than for prediction.

Let us now see that some conformal predictor outputs the classical prediction intervals (10). This will demonstrate that Corollary 2 is indeed a special case of Corollary 1.

The ATTS statistics for the Gauss linear model are

$$S_n(\mathbf{x}_1, y_1, \dots, \mathbf{x}_n, y_n) := \left(\mathbf{x}_1, \dots, \mathbf{x}_n, \sum_{i=1}^n y_i, \sum_{i=1}^n y_i \mathbf{x}_i, \sum_{i=1}^n y_i^2 \right).$$

(It is natural to have $\mathbf{x}_1, \dots, \mathbf{x}_n$ as components of S_n, although they are superfluous under our original definition, in which $\mathbf{x}_1, \mathbf{x}_2, \dots$ are deterministic.) The prediction intervals (10) are precisely the prediction regions output by the conformal predictor corresponding to the nonconformity measure

$$A_n\left(S_{n-1}\left(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}\right), (\mathbf{x}_n, y_n)\right) := \frac{|y_n - \hat{y}_n|}{\sqrt{1 + \mathbf{z}_n' (\boldsymbol{Z}_{n-1}' \boldsymbol{Z}_{n-1})^{-1} \mathbf{z}_n} \hat{\sigma}_{n-1}} \tag{11}$$

(cf. (9); the goodness of the definition follows from the formulas given at the beginning of this section). The expression on the right-hand side of (11) can be replaced by other natural expressions, such as $|y_n - \hat{y}_n|$. See [21], Section 8.5, for further details.

According to our general convention, the conformal predictor (10) is called the *Gauss predictor* (although its discoverer was Fisher rather than Gauss).

We have already mentioned that the classical confidence predictor, Γ_n^ϵ given by (10), does not work when there are many parameters; in particular, it is required that $n \geq K+3$. Theorem 2 shows that there is hardly any way to use the knowledge that the first 10 explanatory variables are the important ones without abandoning the Gauss linear model: no weakly valid confidence predictor in a very wide and natural class can produce informative prediction intervals unless $n \geq K+3$. Indeed, since the conditional distribution of the first n observations given S_n is concentrated at one point for $n \leq K+1$ and at two points for $n = K+2$ with probability one, no conformal predictor and, therefore, no weakly valid invariant confidence predictor can give a bounded prediction region Γ_n^ϵ for $\epsilon < 0.5$ and $n \leq K+2$.

Remark 2. A common reaction to the importance of the condition $n \geq K+3$ is that one can use only a subset of explanatory variables when $n < K+3$. A simple answer is that we are interested in confidence predictors that are valid under the Gauss linear model (1), not under some other model that is only

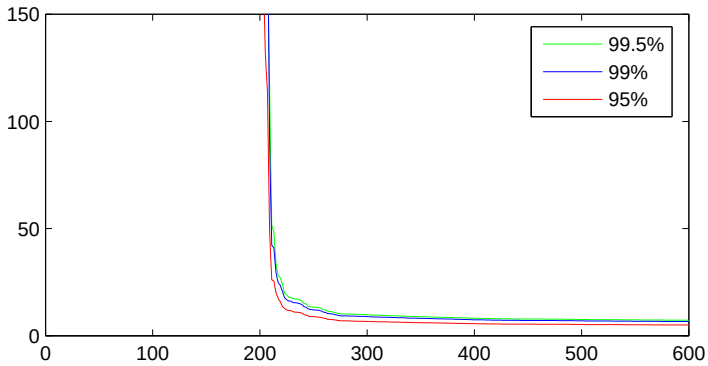

Figure 4: The median-accuracy plot for the classical prediction intervals.

"approximately true", in some ill-defined sense. However, the reader might wonder how the case of the Gauss linear model is different from the case of the IID model, where we are allowed to (and in Section 5 do) take into account only a subset of explanatory variables before n reaches $K+3$. The crucial difference is that if our data $(\mathbf{x}_n, y_n)$ conform to the IID model, discarding a subset of the explanatory variables will give us data $(\mathbf{x}'_n, y_n)$ that still conform to the IID model; on the other hand, if our data $(\mathbf{x}_n, y_n)$ conform to the Gauss linear model, discarding a subset of the explanatory variables will give us data $(\mathbf{x}'_n, y_n)$ that do not conform to the IID model, unless the coefficients β_k in front of the discarded explanatory variables $x_{n,k}$ (see (1)) happen to be precisely zero. More generally, applying any transformation $\mathbf{x}'_n := \phi(\mathbf{x}_n)$ to the explanatory vectors does not lead us outside the IID model. The MVA model is similar to the IID model in this respect, except that the transformation ϕ has to be linear.

Figure 4 gives the median-accuracy plot for the confidence predictor (10); the predictor works very well soon after the number of observations reaches $K+3=103$. Since the median is plotted, the good quality of the prediction intervals shows only from $n=205$: indeed, for $n<205$ at least half of the observed prediction intervals are infinitely wide.

7 The MVA model

Remember that the MVA model assumes, besides (1), that $\mathbf{x}_n$ are generated independently from the same unknown multivariate Gaussian distribution on $\mathbb{R}^K$, with the noise random variables $\xi_1, \xi_2, \dots$ independent of $\mathbf{x}_1, \mathbf{x}_2, \dots$. The

ATTS statistics in the MVA model are

$$S_n := \left(\sum_{i=1}^n \mathbf{x}_i, \sum_{i=1}^n y_i, \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i', \sum_{i=1}^n y_i \mathbf{x}_i, \sum_{i=1}^n y_i^2 \right);$$

equivalently, the statistics can be defined to be the empirical means and covariances of all variables, i.e., the response and the explanatory variables.

Let $\mathbf{y} := \mathbf{y}_n$, $\boldsymbol{Z} := \boldsymbol{Z}_n$, $K^\dagger := K_n^\dagger$ and $\boldsymbol{U}$ be as in Section 5. Suppose the value of the statistic S_n is known. The vector of residuals (5) can now be written as

$$\mathbf{e} := \mathbf{y} - \boldsymbol{U} \left(\boldsymbol{U}' \boldsymbol{U} + a \boldsymbol{I} \right)^{-1} \boldsymbol{U}' \mathbf{y} = \mathbf{y} - \boldsymbol{U} \mathbf{c}, \tag{12}$$

where $\mathbf{c} := \left(\boldsymbol{U}' \boldsymbol{U} + a \boldsymbol{I} \right)^{-1} \boldsymbol{U}' \mathbf{y}$ is a known vector. Since the joint distribution of $\mathbf{y}$ and the non-dummy columns of $\boldsymbol{U}$ is invariant with respect to rotations around the vector $\mathbf{1}$, the distribution of $\mathbf{e}$ will also be invariant with respect to such rotations. It might help the reader's intuition to notice that knowing the value of S_n is equivalent to knowing the lengths of and the angles between the following $K+2$ vectors: the $K+1$ columns of $\boldsymbol{Z}$ and $\mathbf{y}$.

In the rest of this section we will assume $n \geq 3$ (with arbitrary conventions for $n = 1, 2$). Let $e_1, \dots, e_n$ be the components of the vector (12) of residuals and $\overline{e}_{n-1}$ be the average of $e_1, \dots, e_{n-1}$. A standard statistical result (stated in Section 9; see (15)) allows us to conclude that

$$\sqrt{\frac{n-1}{n}} \frac{e_n - \overline{e}_{n-1}}{\sqrt{\frac{1}{n-2} \sum_{i=1}^{n-1} (e_i - \overline{e}_{n-1})^2}} \tag{13}$$

has the t-distribution with $n-2$ degrees of freedom.

Let us see how to implement the conformal predictor corresponding to the nonconformity measure

$$A_n \left(S_{n-1}(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n, y_n) \right) := \frac{e_n - \overline{e}_{n-1}}{\sqrt{\sum_{i=1}^{n-1} (e_i - \overline{e}_{n-1})^2}}, \tag{14}$$

which is proportional to (13); the fact that the right-hand side of (14) depends on the first $n-1$ observations only through the value of S_{n-1} can be seen from the representation (12), where $\mathbf{c}$ is a known vector. First we replace the true value y_n by variable y ranging over $\mathbb{R}$. Each residual e_i becomes a linear (according to (12), where $\mathbf{c}$ also depends on y) function $e_i(y)$ of y, and the prediction region can be written as

$$\Gamma_n^\epsilon := \left\{ y \in \mathbb{R} : \sqrt{\frac{n-1}{n}} \frac{|e_n(y) - \overline{e}_{n-1}(y)|}{\sqrt{\frac{1}{n-2} \sum_{i=1}^{n-1} (e_i(y) - \overline{e}_{n-1}(y))^2}} < t_{n-2}^{\epsilon/2} \right\}.$$

The inequality in this formula is quadratic in y, so Γ_n^ϵ is easy to find. We can see that the prediction region for y_n is an interval (empirically, this is the typical case), the union of two rays, the empty set, or the whole real line.

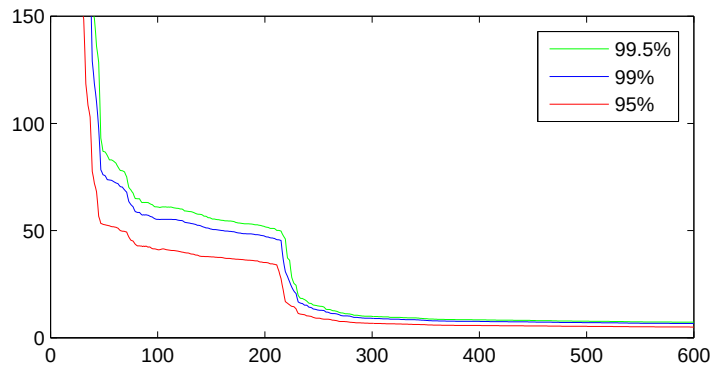

Figure 5: The median-accuracy plot for the MVA predictor.

Replacing Γ_n^ϵ by $\operatorname{co}\Gamma_n^\epsilon$ in the conformal predictor we have just defined gives the *MVA predictor*. Our experiments with the artificial data set of Section 4 are carried out as before (cf. (8)): $\boldsymbol{U}$ is defined as the first 11 columns of $\boldsymbol{Z}$ if $n < 103$ and as the full $\boldsymbol{Z}$ otherwise.

The median-accuracy plot for the MVA predictor and our artificial data set is shown in Figure 5. Before the threshold 103 the predictor quickly learns α and the first 10 parameters β_k, and its performance more or less stabilizes before quickly improving again when it starts learning the other parameters from $n = 103$ onwards; the second improvement in the performance shows on the median-accuracy plot from $n = 205$.

The performance of the MVA predictor is better than the performance of any other confidence predictor considered in this paper. Of course, this should not be taken to mean that the other predictors are worse. Different predictors are based on different information about the data set. None of the predictors "knows" that the components of $\mathbf{x}_n$ are realizations of independent standard Gaussian random variables; even the MVA model, the narrowest model considered in this paper, allows arbitrary means of and arbitrary correlations between different explanatory variables for the same observation. The Gauss predictor does not know that the $\mathbf{x}_n$ are IID and Gaussian. The IID predictor only knows that the observations $(\mathbf{x}_n, y_n)$ are IID, and the IID–Gauss predictor, introduced in the next section, knows, in addition, that the y_n are generated by (1).

The median-accuracy plot for each of the four predictors is essentially determined by that for the MVA predictor and the threshold for the corresponding model as shown in Table 1. It is convenient to represent each line on a median-accuracy plot as the function that maps each value for the accuracy in the interval $[0, 150]$ to the first step at which that accuracy is achieved (so the graph of this function is obtained by rotating the page by 90° countercloc-

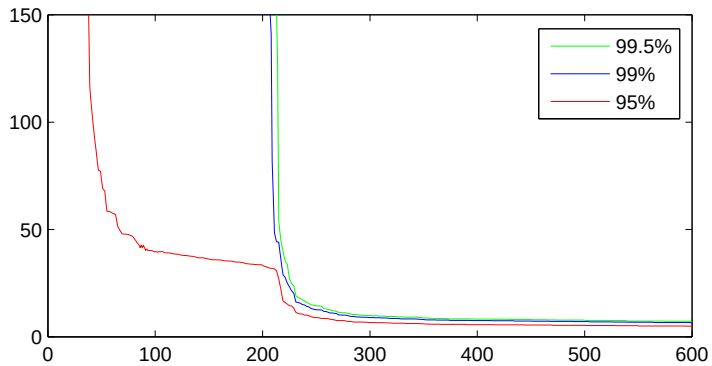

Figure 6: The median-accuracy plot for the IID–Gauss predictor.

kwise). Each of the three functions in Figure 2 is, approximately, the maximum of $2\lceil 1/\epsilon \rceil$ and the corresponding function in Figure 5. Similarly, each of the three functions in Figure 4 is, approximately, the maximum of $2(K+3) = 206$ and the corresponding function in Figure 5. As usual, the factor of 2 appears because of the use of median in our accuracy plots.

8 The IID–Gauss model

As defined in Section 1, the IID–Gauss model is the combination of the Gauss linear and IID models: we assume both that the observations are IID and that the responses are generated by (1) with $\xi_1, \xi_2, \ldots$ independent of $\mathbf{x}_1, \mathbf{x}_2, \ldots$. Correspondingly, the ATTS statistics are

$$S_n := \left(\wr \mathbf{x}_1, \ldots, \mathbf{x}_n \int, \sum_{i=1}^n y_i, \sum_{i=1}^n y_i \mathbf{x}_i, \sum_{i=1}^n y_i^2 \right).$$

Using the nonconformity measure (6) and replacing the prediction regions output by the corresponding conformal predictor with their convex hulls, we obtain the *IID–Gauss predictor*. Its performance on our usual data set is shown in Figure 6. We do not know whether the IID–Gauss predictor can be implemented efficiently, and Figure 6 was produced using Monte-Carlo sampling from the conditional distributions given S_n. However, comparing Figure 6 to Figures 2 (to the left of $n = 205$) and 4 (to the right of $n = 205$), we can see that the following simple confidence predictor will work almost as well as the IID–Gauss predictor on our data set: predict using the IID predictor if $n < 103$ and predict using the Gauss predictor if $n \geq 103$. As in all other cases in this paper where the threshold $n = K + 3 = 103$ appears, the best switch-over point will

be slightly greater than $K+3$, but the question of when exactly to switch is outside the scope of this paper.

Remark 3. The IID predictor and the IID–Gauss predictor use the same non-conformity measure, (6), but still produce very different median-accuracy plots at confidence level 99.5%. This happens because of the conditioning on the event $S_n^{\mathrm{rnd}} = S_n^{\mathrm{obs}}$ in the definition (2). Since the ATTS statistics perform more radical data compression in the case of the IID–Gauss model, the achievable values of $\mathbb{P}(A_n^{\mathrm{rnd}} \geq A_n^{\mathrm{obs}} \mid S_n^{\mathrm{rnd}} = S_n^{\mathrm{obs}})$ (corresponding to (2) with $\tau_n := 1$) are much smaller than the $1/n$ achievable under the IID model.

As in the previous section, there is a close connection between Figures 5 and 6: each of the three functions in Figure 6 is, approximately, the maximum of $2\min(\lceil 1/\epsilon \rceil, K+3)$ and the corresponding function in Figure 5. The distributive law of max over min now implies that each of the three functions in Figure 6 is the minimum of the corresponding functions in Figures 2 and 4.

9 On-line protocol, part II

Theorem 1 sheds new light not only on the main topic of this paper, predictive linear regression, but also on some more classical corners of statistics. In this section we will discuss, in particular, Fisher's fiducial prediction and Wilks's non-parametric prediction intervals. At the end of the section we discuss relaxations of the on-line protocol.

The Gaussian model

Let us consider the model (1) with the $\mathbf{x}_n$ absent (i.e., $K=0$); in other words, y_n is an IID sequence with $y_n \sim N(\alpha, \sigma^2)$ and unknown α and $\sigma^2 > 0$. This model will be called the *Gaussian model*. Notice that the MVA model and the IID–Gauss model also reduce to the Gaussian model when $K=0$.

The fact that

$$T_n := \sqrt{\frac{n-1}{n}} \frac{y_n - \overline{y}_{n-1}}{\hat{\sigma}_{n-1}}, \tag{15}$$

where

$$\overline{y}_l := \frac{1}{l} \sum_{i=1}^{l} y_i \quad \text{and} \quad \hat{\sigma}_l^2 := \frac{1}{l-1} \sum_{i=1}^{l} (y_i - \overline{y}_l)^2,$$

has the t-distribution with $n-2$ degrees of freedom [8] allows us to conclude that $y_n \in \Gamma_n^\epsilon$ with probability $1-\epsilon$, where the prediction interval Γ_n^ϵ for y_n is defined by

$$\Gamma_n^\epsilon := \left\{ y \in \mathbb{R} : \left| y - \overline{y}_{n-1} \right| < t_{n-2}^{\epsilon/2} \sqrt{\frac{n}{n-1}} \hat{\sigma}_{n-1} \right\}, \quad n = 3, 4, \ldots, \tag{16}$$

and $\epsilon \in (0,1)$ is the chosen significance level. This prediction interval is a special case of (10).

Fisher discussed (16) and related confidence predictors in his last book ([9], Sections V.3–4) under the rubric of "fiducial prediction". It appears that the idea of fiducial prediction is less controversial (and less often discussed) than the related idea of fiducial inference for parameter values; besides, we will be interested in the least controversial aspects of fiducial prediction. Fisher's comments about fiducial prediction in Sections V.3–4 are all applicable to the predictor (16), although in Section V.3 he discusses prediction of exponentially distributed rather than Gaussian random variables.

To some extent answering his critics ("some teachers assert that statements of fiducial probability cannot be tested by observations"), he writes that "fiducial statements about future observations" (such as (16), although this passage is about exponentially distributed responses) "are verifiable by subsequent observations to any degree of precision required". The following is our reconstruction (we believe the only possible reconstruction) of *Fisher's verification protocol*, as applied to the prediction intervals (16). Fix a significance level $\epsilon \in (0,1)$ and $l \in \{2,3,\dots\}$ (the *sample size*; we might consider samples of different sizes, but we will stick to the simplest case). For $m = 1,2,\dots$, generate the *mth sample*

$$y_{(m-1)(l+1)+1}, y_{(m-1)(l+1)+2}, \dots, y_{m(l+1)-1}$$

and the *mth test observation* $y_{m(l+1)}$. Register an error if the mth prediction interval computed from the mth sample according to (16) fails to contain the mth test observation:

$$\mathrm{err}_m^\dagger := \begin{cases} 0 & \text{if } |y_{m(l+1)} - \overline{y}| < t_{l-1}^{\epsilon/2} \sqrt{\frac{l+1}{l}} \sqrt{\frac{1}{l-1} \sum_{i=(m-1)(l+1)+1}^{m(l+1)-1} (y_i - \overline{y})^2} \\ 1 & \text{otherwise,} \end{cases}$$

where

$$\overline{y} := \frac{1}{l} \sum_{i=(m-1)(l+1)+1}^{m(l+1)-1} y_i.$$

As in the on-line protocol, the errors $\mathrm{err}_m^\dagger$, $m = 1,2,\dots$, are independent. The frequency of error gets arbitrarily close to ϵ with an arbitrarily high probability as the number of observations increases.

Fisher's verification protocol has a serious drawback: as Fisher puts it,

> In carrying out such a verification [...], it is to be supposed that the investigator is not deflected from his purpose by the fact that new data are becoming available from which predictions, better than the one he is testing, could at any time be made. For verification, the original prediction must be held firmly in view. This, of course, is a somewhat unnatural attitude for a worker whose main preoccupation is to improve his ideas.

Indeed, when making his prediction for the mth test observation, the "investigator" is asked to ignore the first $m-1$ samples. The protocol seems to be an artificial device rather than a description of what "a worker whose main

preoccupation is to improve his ideas" might do in reality. Let us see, however, what happens if all the previous observations *are* used when making the mth prediction; in this case, the sequence of errors becomes

$$\mathrm{err}_m^\ddagger := \begin{cases} 0 & \text{if } |y_{m(l+1)} - \overline{y}| < t^{\epsilon/2}_{m(l+1)-2}\sqrt{\frac{m(l+1)}{m(l+1)-1}}\sqrt{\frac{1}{m(l+1)-2}\sum_{i=1}^{m(l+1)-1}(y_i-\overline{y})^2} \\ 1 & \text{otherwise,} \end{cases}$$

where

$$\overline{y} := \frac{1}{m(l+1)-1}\sum_{i=1}^{m(l+1)-1} y_i.$$

As $\mathrm{err}_m^\ddagger$, $m = 1, 2, \ldots$, is a subsequence of the sequence of errors err_n^ϵ, $n = 1, 2, \ldots$, in the on-line protocol, the errors are still independent. Theorem 1 cures the drawback.

In Sections V.3–4 of his book, Fisher is also interested in the problem of predicting some characteristic of a future sample from a given sample (our problem of predicting one future observation is a special case, corresponding to a second sample of size one). Fisher's verification protocol is defined in essentially the same way and leads to independent errors. It can be shown that, in all cases considered by Fisher, the independence persists when the predictions for test samples are based on all available information. (We assume that Fisher's fiducial distributions for the characteristics of the second sample have been replaced by prediction intervals, maybe unbounded on one side; he does this explicitly only for his exponential example in Section V.3.)

Fisher's theory of fiducial prediction is based on the fact that a value such as (15) has a known distribution for each n; therefore, it can be used as a "pivot" to project this known distribution onto the future observation y_n. This idea may be difficult to formalize, but Fisher's observation that (15) has a known distribution can be strengthened: Theorem 1 (applied to the nonconformity measure (15)) implies that the random variables T_n, $n = 3, 4, \ldots$, have the t-distribution with $n-2$ degrees of freedom and are independent in the on-line protocol. Therefore, not only the individual T_n have known distributions, but also the whole sequence $(T_1, T_2, \ldots)$ has a known distribution (the product of t-distributions).

The univariate IID model

The IID model is different from all the other models in this paper (see Figure 1) in that it gives a univariate model different from the Gaussian model in the case where the explanatory variables are absent. The construction of prediction and tolerance intervals in the univariate IID model, which says that $y_1, y_2, \ldots$ form an IID sequence, was undertaken by many authors following the pioneering paper by Wilks [23]. Wilks's work was later extended to the multivariate case: see, e.g., Fraser [10]; this extension, however, is not directly related to our IID predictors. For simplicity, let us assume in this subsection, as is customary in literature,

that the distribution of one observation is continuous. Correspondingly, we will assume that the realized values of y_n, $n = 1, 2, \ldots$, are all different.

For each $n = 1, 2, \ldots$, define $T_n \in \{1, 2, \ldots, n\}$ as the smallest i such that $y_n < y_{(n-1,i)}$, where $y_{(n-1,1)}, \ldots, y_{(n-1,n-1)}$ is the sequence of the first $n-1$ observations $y_1, \ldots, y_{n-1}$ sorted in the ascending order; if $y_n > y_{(n-1,n-1)}$, set $T_n := n$. Each T_n is a "pivot", being distributed uniformly on the set $\{1, \ldots, n\}$. Wilks suggested the following prediction intervals based on this fact: fix a number $r \in \{1, 2, \ldots\}$ and define $\Gamma_n^{2r/n}$, $n = 2r+1, 2r+2, \ldots$, to be the interval $(y_{(n-1,r)}, y_{(n-1,n-r)})$; the probability of error, $y_n \notin \Gamma_n^{2r/n}$, is then $2r/n$. Now Theorem 1 implies that the whole random sequence $(T_1, T_2, \ldots)$ has a known distribution: namely, it is distributed according to the product $U_1 \times U_2 \times \cdots$ of the uniform distributions U_n on $\{1, \ldots, n\}$. In particular, Wilks's prediction intervals $\Gamma_n^{2r/n}$, $n = 2r+1, 2r+2, \ldots$, lead to independent errors.

Relaxations of the on-line protocol

This paper concentrates on the on-line prediction protocol. Smoothed conformal predictors lead to independent errors in the on-line protocol, and Theorem 2 suggests that conformal predictors are the most natural weakly valid confidence predictors. This is why we included the requirement of independence in the definition of strong validity, despite the fact that the error frequency can be shown to approach the error probability ϵ with probability approaching one even when the requirement of independence is relaxed in certain ways.

The situation changes when we move outside the on-line protocol. The on-line protocol is natural, but in one respect it is overly restrictive: the true response y_n becomes known before the prediction for the next response y_{n+1} is made. It can be shown that the error frequency will still converge to ϵ if the true response is only given for a small fraction of observations, and even for those observations it can be given with a delay ([21], Section 4.3; see also [20] for an empirical study). The independence of errors, however, will be lost (we can still have "approximate independence", but this is a much more elusive notion than ordinary independence).

10 Conclusion

In this paper we considered the problem of prediction in three main regression models. One of these models, the Gauss linear model, is the standard textbook one. The MVA model seems to have been somewhat neglected, partly because of philosophical reasons: according to the conditionality principle ([5], Section 2.3(iii)) one should condition on the observed values of the explanatory variables to make the prediction (or estimate, etc.) more relevant to the data at hand. In most of this paper we took a pragmatic approach, studying which models permit one to produce informative prediction intervals in different circumstances without being restricted *a priori* by general principles. We did use

the sufficiency principle in our interpretation of Theorem 2, but we admit that this makes the theorem less convincing. Surprisingly, the IID model appears to have been neglected in the field of regression, even in non-parametric statistics, where the value of this model is in principle well understood.

Acknowledgments

This paper has greatly benefited from Glenn Shafer's advice, from a discussion with Steffen Lauritzen, and from correspondence with Philip Dawid. Comments by the anonymous referees of the journal version and by Prof. Susan Murphy helped us improve the presentation. This work was partially supported by EPSRC (grant EP/F002998/1), MRC (grant S505/65), and the Royal Society.

Appendix A: Proofs of the theorems

In this appendix we will prove the two main results stated in this paper, Theorems 1 and 2. A version of Theorem 1 was proved in Section 8.7 of [21], but we reproduce the principal points of the proof to make our exposition self-contained. A special case of Theorem 2 (namely, for the IID model) was proved in Section 2.6 of [21].

Proof of Theorem 1

In this proof, $\zeta_1, \zeta_2, \dots$ will be random observations generated by $P \in \mathcal{P}$, $(\zeta_1, \zeta_2, \dots) \sim P$, and $\tau_1, \tau_2, \dots$ will be random numbers, $(\tau_1, \tau_2, \dots) \sim U^\infty$. For each $n = 0, 1, \dots$ let $\mathcal{G}_n$ be the σ-algebra generated by the random elements

$$S_n(\zeta_1, \dots, \zeta_n), \zeta_{n+1}, \tau_{n+1}, \zeta_{n+2}, \tau_{n+2}, \dots .$$

So $\mathcal{G}_0$ is the most informative σ-algebra and $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \mathcal{G}_2 \supseteq \cdots$. It will be convenient to write $\mathbb{P}_{\mathcal{G}}(E)$ and $\mathbb{E}_{\mathcal{G}}(\xi)$ for the conditional probability $\mathbb{P}(E \mid \mathcal{G})$ and expectation $\mathbb{E}(\xi \mid \mathcal{G})$, respectively, given a σ-algebra $\mathcal{G}$.

Lemma 1. *For any step $n = 1, 2, \dots$ and any $\epsilon \in (0, 1)$,*

$$\mathbb{P}_{\mathcal{G}_n}(p_n \le \epsilon) = \epsilon.$$

Proof. For a given value of the summary $S_n(\zeta_1, \dots, \zeta_n)$ of the first n observations, consider the conditional distribution function F of the random variable $\eta := A_n(S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n)$ (because of the total sufficiency, it does not matter whether we further condition on $\zeta_{n+1}, \tau_{n+1}, \zeta_{n+2}, \tau_{n+2}, \dots$). Define $F(x-)$ to be $\sup_{t<x} F(t)$. Our task is to show that the conditional probability of the event

$$1 - F(\eta) + \tau_n(F(\eta) - F(\eta-)) \le \epsilon \tag{17}$$

is ϵ (since the left-hand side of (17) coincides with the right-hand side of the definition (2)). The latter fact is usually stated in statistics textbooks for continuous F (see, e.g., [5], p. 66), but it is also easy to check in general. □

Lemma 2. *For any step $n = 1, 2, \ldots$, p_n is $\mathcal{G}_{n-1}$-measurable.*

Proof. This follows from the definition: p_n is defined in terms of ζ_n, τ_n and the summary of the first $n-1$ observations. □

Now we can easily prove the theorem. First we demonstrate that, for any $n = 1, 2, \ldots$ and any $\epsilon_1, \ldots, \epsilon_n \in (0, 1)$,

$$\mathbb{P}_{\mathcal{G}_n}(p_n \le \epsilon_n, \ldots, p_1 \le \epsilon_1) = \epsilon_n \cdots \epsilon_1 \quad \text{a.s.} \tag{18}$$

The proof is by induction on n. For $n = 1$, (18) is a special case of Lemma 1. For $n > 1$ we obtain, from Lemmas 1 and 2, standard properties of conditional expectations, and the inductive assumption:

$$\begin{aligned}\mathbb{P}_{\mathcal{G}_n}(p_n \le \epsilon_n, \ldots, p_1 \le \epsilon_1) &= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{p_n \le \epsilon_n} \mathbb{I}_{p_{n-1} \le \epsilon_{n-1}, \ldots, p_1 \le \epsilon_1}\right)\right) \\ &= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{p_n \le \epsilon_n} \mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{p_{n-1} \le \epsilon_{n-1}, \ldots, p_1 \le \epsilon_1}\right)\right) = \mathbb{E}_{\mathcal{G}_n}(\mathbb{I}_{p_n \le \epsilon_n} \epsilon_{n-1} \cdots \epsilon_1) \\ &= \epsilon_n \epsilon_{n-1} \cdots \epsilon_1 \quad \text{a.s.}\end{aligned}$$

The "tower property" of conditional expectations immediately implies

$$\mathbb{P}(p_n \le \epsilon_n, \ldots, p_1 \le \epsilon_1) = \epsilon_n \cdots \epsilon_1.$$

Therefore, the distribution of the first n p-values $p_1, \ldots, p_n$ is U^n, for all $n = 1, 2, \ldots$. This implies that the distribution of the infinite sequence $p_1, p_2, \ldots$ is U^∞.

Proof of Theorem 2

In this proof, $Z := \mathbb{R}^K \times \mathbb{R}$ and ζ_i stands for $(\mathbf{x}_i, y_i)$. Let $n \in \mathcal{N}$.

For each summary $s \in \Sigma_n$ let $f(s)$ be the conditional probability given $S_n(\zeta_1, \ldots, \zeta_n) = s$ that Γ makes an error at a significance level ϵ when predicting y_n from $\zeta_1, \ldots, \zeta_{n-1}$ and $\mathbf{x}_n$, the observations $\zeta_1, \zeta_2, \ldots$ being generated from $P \in \mathcal{P}$. We know that the expected value of $f(S_n(\zeta_1, \ldots, \zeta_n))$ is ϵ under any $P \in \mathcal{P}$, and this, by the bounded completeness of S_n, implies that $f(s) = \epsilon$ for almost all (under PS_n^{-1} for any $P \in \mathcal{P}$) summaries s. Define $E(s, \epsilon)$ to be the set of all pairs $(s', \zeta) = (s', (\mathbf{x}, y)) \in \Sigma_{n-1} \times Z$ such that $F_n(s', \zeta) = s$ (where F_n is the function from the definition of the algebraic transitivity of the S_n) and Γ makes an error at the significance level ϵ when predicting y and fed with $\zeta_1, \ldots, \zeta_{n-1}$ satisfying $S_{n-1}(\zeta_1, \ldots, \zeta_{n-1}) = s'$ and with $\mathbf{x}$ (since Γ is invariant, whether an error is made depends only on s', not on the particular $\zeta_1, \ldots, \zeta_{n-1}$). It is clear that

$$\epsilon_1 \le \epsilon_2 \Longrightarrow E(s, \epsilon_1) \subseteq E(s, \epsilon_2)$$

and

$$\mathbb{P}\left((S_{n-1}(\zeta_1, \ldots, \zeta_{n-1}), \zeta_n) \in E(s, \epsilon) \mid S_n(\zeta_1, \ldots, \zeta_n) = s\right) = \epsilon \quad \text{a.s.},$$

where $(\zeta_1, \zeta_2, \ldots) \sim P \in \mathcal{P}$.

In this proof we say "conformity measure" to mean a nonconformity measure which is used for computing p-values in the opposite way to (2): the ">" in (2) is replaced by "<". Let us check that the conformal predictor $\Gamma^\dagger$ determined by the conformity measure

$$A_n(s', \zeta) := \inf \{\epsilon : (s', \zeta) \in E(F_n(s', \zeta), \epsilon)\}$$

is at least as accurate as Γ. By the monotone convergence theorem for conditional expectations,

$$\begin{aligned}
&\mathbb{P}\left(A_n(S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n) \le \epsilon \mid S_n(\zeta_1, \dots, \zeta_n) = s\right) \\
&\quad = \lim_{\delta \downarrow \epsilon} \mathbb{P}\left(A_n(S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n) < \delta \mid S_n(\zeta_1, \dots, \zeta_n) = s\right) \\
&\quad \le \lim_{\delta \downarrow \epsilon} \mathbb{P}\left((S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n) \in E(s, \delta) \mid S_n(\zeta_1, \dots, \zeta_n) = s\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \lim_{\delta \downarrow \epsilon} \delta = \epsilon \quad \text{a.s.},
\end{aligned}$$

where $(\zeta_1, \zeta_2, \dots) \sim P \in \mathcal{P}$ and δ is constrained to be a rational number. Therefore, at each significance level ϵ and for P-almost all $(\zeta_1, \dots, \zeta_n) \in Z^n$,

$$\begin{aligned}
&y_n \in (\Gamma^\dagger)^\epsilon(\zeta_1, \dots, \zeta_{n-1}, \mathbf{x}_n) \Longleftrightarrow \mathbb{P}\left(A_n^{\text{rnd}} \le A_n^{\text{obs}} \mid S_n^{\text{rnd}} = S_n^{\text{obs}}\right) > \epsilon \\
&\quad \Longrightarrow A_n^{\text{obs}} > \epsilon \Longrightarrow (S_{n-1}(\zeta_1, \dots, \zeta_{n-1}), \zeta_n) \notin E(S_n(\zeta_1, \dots, \zeta_n), \epsilon) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Longleftrightarrow y_n \in \Gamma^\epsilon(\zeta_1, \dots, \zeta_{n-1}, \mathbf{x}_n),
\end{aligned}$$

in the notation of (2).

Appendix B: Explicit algorithms

For all three main models considered in this paper there are relatively efficient algorithms for computing prediction intervals. In this appendix we will describe two of them, one new (the MVA predictor) and the other not known in the statistical community (the IID predictor).

The MVA predictor

The idea of this algorithm was described in Section 7, and here we will give explicit formulas which the reader might find useful in writing computer programs implementing this idea. Let $\boldsymbol{P}$ be the matrix

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{U}\left(\boldsymbol{U}'\boldsymbol{U} + a\boldsymbol{I}\right)^{-1}\boldsymbol{U}'$$

projecting the responses $\mathbf{y}$ onto the residuals $\mathbf{e}$ (see (12)). Introducing the notation

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n-1} \\ y \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n-1} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y \end{pmatrix} = \mathbf{y}_0 + y\mathbf{u}$$

(so that $\mathbf{y}_0$ is the vector of the first $n-1$ responses followed by a 0 and $\mathbf{u}$ is the vector whose only non-zero component is a 1 at the nth position), we can define the vector of residuals as

$$\mathbf{e} = \boldsymbol{P}(\mathbf{y}_0 + y\mathbf{u}) = \boldsymbol{P}\mathbf{y}_0 + y\boldsymbol{P}\mathbf{u}. \tag{19}$$

Subtract from each component of $\boldsymbol{P}\mathbf{y}_0$ the mean of the first $n-1$ components of $\boldsymbol{P}\mathbf{y}_0$ and subtract from each component of $\boldsymbol{P}\mathbf{u}$ the mean of the first $n-1$ components of $\boldsymbol{P}\mathbf{u}$; let the resulting vectors be $\mathbf{a}$ and $\mathbf{b}$, respectively. The prediction region consists of the y satisfying

$$\sqrt{\frac{n-1}{n}}\frac{|a_n + yb_n|}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(a_i + yb_i)^2}} < t, \tag{20}$$

where a_i and b_i are the components of $\mathbf{a}$ and $\mathbf{b}$, respectively, and t stands for $t_{n-2}^{\epsilon/2}$. We can rewrite (20) as

$$(a_n + b_n y)^2 < t^2 \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n-1}(a_i + b_i y)^2,$$

or, explicitly as a quadratic inequality,

$$Ay^2 + 2By + C < 0,$$

where

$$\begin{aligned}
A &= (n-1)(n-2)b_n^2 - t^2 n \sum_{i=1}^{n-1} b_i^2, \\
B &= (n-1)(n-2)a_n b_n - t^2 n \sum_{i=1}^{n-1} a_i b_i, \\
C &= (n-1)(n-2)a_n^2 - t^2 n \sum_{i=1}^{n-1} a_i^2.
\end{aligned}$$

The solutions to the corresponding quadratic equation are

$$\frac{-B \pm \sqrt{D}}{A}, \tag{21}$$

where the discriminant D is

$$D = B^2 - AC.$$

Therefore, the prediction interval (defined, as usual, to be the convex hull of the prediction region) is:

- unbounded if
$$A < 0$$
or
$$A = 0 \text{ and } B \neq 0$$
or
$$A = B = 0 \text{ and } C < 0;$$
- empty if
$$A > 0 \text{ and } D \leq 0$$
or
$$A = B = 0 \text{ and } C \geq 0;$$
- otherwise (i.e., if $A > 0$ and $D > 0$), the open interval with the end-points (21).

The IID predictor

In this subsection we describe explicitly the IID predictor. We will follow mainly [21], Section 2.3, although our current algorithm is simpler, as we are only interested in finding the prediction intervals (the convex hulls of the prediction regions) and not the prediction regions themselves.

Rewrite the vector of residuals (19) in the form $\mathbf{a} + y\mathbf{b}$, where $\mathbf{a} := \boldsymbol{P}\mathbf{y}_0$ and $\mathbf{b} := \boldsymbol{P}\mathbf{u}$. For each $i = 1, \dots, n$, let

$$S_i := \{y : |e_i(y)| \geq |e_n(y)|\} = \{y : |a_i + b_i y| \geq |a_n + b_n y|\},$$

where a_i and b_i are the components of $\mathbf{a}$ and $\mathbf{b}$. Each set S_i (always closed) will be the real line, the union of two rays, a ray, an interval (by an interval we will always mean a bounded interval), a point (which is a special case of an interval), or empty. Indeed, as we are interested in $|a_i + b_i y|$ we can assume $b_i \geq 0$ for $i = 1, \dots, n$ (if necessary, multiply both a_i and b_i by -1). If $b_i \neq b_n$ then $|e_i(y)|$ and $|e_n(y)|$ are equal at two points (which may coincide):

$$-\frac{a_i - a_n}{b_i - b_n} \quad \text{and} \quad -\frac{a_i + a_n}{b_i + b_n}; \tag{22}$$

in this case, S_i is an interval (possibly a point), the union of two rays, or the real line. If $b_i = b_n \neq 0$ then $|e_i(y)| = |e_n(y)|$ at the only point

$$-\frac{a_i + a_n}{2b_i} \tag{23}$$

(and S_i is a ray) unless $a_i = a_n$, in which case $S_i = \mathbb{R}$. If $b_i = b_n = 0$, S_i is either $\emptyset$ or $\mathbb{R}$.

To calculate the p-value $p(y)$ for any potential response y of x_n, we count how many S_i include y and divide by n:

$$p(y) = \frac{|\{i = 1, \dots, n : y \in S_i\}|}{n}.$$

This formula immediately implies that each Γ_n^ϵ, as defined by (3), is a closed set: setting $k := \lfloor \epsilon n \rfloor + 1$,

$$\Gamma_n^\epsilon = \{y : p(y) > \epsilon\} = \{y : |\{i = 1, \dots, n : y \in S_i\}| \geq k\} \\ = \bigcup_{\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}} S_{i_1} \cap \dots \cap S_{i_k};$$

the last expression is a finite union of closed sets.

The following algorithm implicitly keeps track of the number $M(y)$ of $i = 1, \dots, n-1$ such that $y \in S_i$ for each $y \in \mathbb{R}$. For each critical point y_j, $j = 1, \dots, m$, let $NM(j)$ be the change in $M(y)$, as y increases, at y_j due to y_j being a critical point. The set of critical points will be represented as a sequence P (initially empty); the critical points will be listed in P and in NM in the same order. The algorithm will also compute the number L of S_i that include points to the left of the left-most critical point and the number R of S_i that include points to the right of the right-most critical point. It is clear that L and NM (or R and NM) determine M. The word "add" in the description of the algorithm means "attach at the end" (unless we say explicitly "add in front") and has nothing to do with arithmetic addition.

The algorithm is given a significance level ϵ and outputs the corresponding prediction interval $\bar{\Gamma}_n^\epsilon := \operatorname{co} \Gamma_n^\epsilon$. We will assume that the set of critical points ((22) and (23)) is not empty and will supply the matrix $\boldsymbol{U}$ with a lower index to explicitly indicate the dependence on n.

IID PREDICTOR

$\boldsymbol{P} := \boldsymbol{I} - \boldsymbol{U}_n(\boldsymbol{U}_n' \boldsymbol{U}_n + a\boldsymbol{I})^{-1} \boldsymbol{U}_n'$;
$\mathbf{a} = (a_1, \dots, a_n)' := \boldsymbol{P}(y_1, \dots, y_{n-1}, 0)'$;
$\mathbf{b} = (b_1, \dots, b_n)' := \boldsymbol{P}(0, \dots, 0, 1)'$;
FOR $i = 1, \dots, n$
    IF $(b_i < 0)$ $a_i := -a_i$; $b_i := -b_i$ END IF
END FOR
$P := NM := ()$; $L := R := 0$;
FOR $i = 1, \dots, n-1$
    IF $(b_i \neq b_n)$
        add the two points (22) in the increasing order to P;
        IF (the corresponding S_i is an interval)
            add 1 and -1 to NM
        END IF
        IF (the corresponding S_i is a union of disjoint rays)
            add -1 and 1 to NM and increase L and R by 1
        END IF
        IF (the corresponding S_i is the real line)
            reverse adding (22) to P and increase L and R by 1
        END IF
    ELSEIF $(a_i = a_n)$

```
        increase L and R by 1
    ELSEIF (b_n ≠ 0)
        add (23) to P;
        IF (the corresponding S_i, which is a ray, is going to the right)
            add 1 to NM and increase R by 1
        ELSE
            add −1 to NM and increase L by 1
        END IF
    ELSEIF (|a_i| ≥ |a_n|)
        increase L and R by 1
    END IF
END FOR
add −∞ in front of P and L + 1 in front of NM
add ∞ at the end of P and −R − 1 at the end of NM
sort P in ascending order breaking ties using −NM;
apply the same permutation to NM;
find the smallest i_1 such that the sum of the first i_1 elements of NM exceeds ε;
find the largest i_2 such that the sum of the first i_2 elements of NM exceeds ε;
Γ̄_n^ε := [P(i_1), P(i_2)].
```

This algorithm is run by Predictor at each step $n = 1, 2, \ldots$ of the on-line prediction protocol. It is important that both points (22) should be added to P, even when they coincide. Finding i_2 can be done by search from the right, since the sum of all elements of NM is 0.

Let us suppose that the number K of explanatory variables is bounded above by a constant. The computation time of our algorithm is $O(n \log n)$. Indeed,

$$\mathbf{a} = (y_1, \ldots, y_{n-1}, 0)' - \boldsymbol{U}_n \left[(\boldsymbol{U}_n' \boldsymbol{U}_n + a\boldsymbol{I})^{-1} \boldsymbol{U}_n' (y_1, \ldots, y_{n-1}, 0)' \right]$$

and

$$\mathbf{b} = (0, \ldots, 0, 1)' - \boldsymbol{U}_n \left[(\boldsymbol{U}_n' \boldsymbol{U}_n + a\boldsymbol{I})^{-1} \boldsymbol{U}_n' (0, \ldots, 0, 1)' \right]$$

can be computed in time $O(n)$, and sorting P can be done in time $O(n \log n)$ (see, e.g., [4], Part II).

Appendix C: Instruction manual for the R package `PredictiveRegression`

The R package `PredictiveRegression` implements the three prediction algorithms (IID predictor, Gauss predictor, and MVA predictor) described in this paper. It is available from CRAN via `http://www.r-project.org/`.

The program implementing the IID predictor is `iidpred`. It is called using the command `iidpred(train,test,epsilons,ridge)` with the default values `c(0.05,0.01)` for `epsilons` and `0` for `ridge`. The arguments are:

`train` The training set as a matrix of size $N \times (K+1)$. Each row describes an observation. Columns 1 to K are the explanatory variables, and column $K+1$ is the response variables.

`test` The test set as a matrix of size $N_2 \times K$. Each row corresponds to an observation (but without the response variable). Columns 1 to K are the explanatory variables.

`epsilons` A vector of several significance levels. Each significance level `epsilons`$[j]$ is a number between 0 and 1. The default value is $(5\%, 1\%)'$.

`ridge` The ridge coefficient, a nonnegative number. The default value is 0; setting it to a small positive constant might lead to more stable results.

(There is no argument $K^\dagger$ and $K^\dagger = K$ is always used.) The output is a list of three elements:

`output[[1]]` This is the matrix of lower bounds of prediction intervals. Its size is $N_2 \times N_\epsilon$, where N_2 is the number of test observations and N_ϵ is the number of significance levels. The element `output[[1]]`$[i,j]$ of `output[[1]]` is the lower bound a of the prediction interval $[a,b]$ for the ith test observation and for the jth significance level `epsilons`$[j]$ in the vector `epsilons`.

`output[[2]]` The matrix of upper bounds b, with the same structure as `output[[1]]`. Typically $a =$ `output[[1]]`$[i,j]$ and $b =$ `output[[2]]`$[i,j]$ are real numbers such that $a \le b$. Exceptions: a is allowed to be $-\infty$ and b is allowed to be ∞; the only case where $a > b$ is $a = \infty$ and $b = -\infty$ (the empty prediction $[a,b]$).

`output[[3]]` The termination code, which is one of:

0 normal termination;

1 illegal parameters (the training and test sets have different numbers of explanatory variables);

2 too few observations for all significance levels.

This program implements the algorithm described on pp. 30–34 of [21] and, in more detail, in the previous section. A simple example of its use is

```
train <- matrix(c(0,10,20,30, 1.01,10.99,21.01,30.99),
   nrow=4, ncol=2);
test <- matrix(c(5,15,25), nrow=3, ncol=1);
output <- iidpred(train,test,c(0.05,0.2),0.01);
print(output[[1]]);
print(output[[2]]);
```

The programs implementing the MVA predictor and the Gauss predictor are `mvapred` and `gausspred`, respectively. They are called in the same way as `iidpred`, with the only exception that `gausspred` does not require the parameter `ridge`. The termination code for `mvapred` and `gausspred` (returned in `output[[3]]`) is:

0 normal termination;

1 illegal parameters (the training and test sets have different numbers of explanatory variables);

2 too few observations.

An example of use of `mvapred` is

```
train <- matrix(c(0,10,20,30, 1.01,10.99,21.01,30.99),
   nrow=4,ncol=2);
test <- matrix(c(5,15,25), nrow=3, ncol=1);
output <- mvapred(train,test,c(0.05,0.2),0.01);
print(output[[1]]);
print(output[[2]]);
```

and an example of use of `gausspred` is

```
train <- matrix(c(1,2,3,4, 2.01,2.99,4.01,4.99),
   nrow=4, ncol=2);
test <- matrix(c(0,10,20), nrow=3, ncol=1);
output <- gausspred(train,test,c(0.05,0.2));
print(output[[1]]);
print(output[[2]]);
```

Appendix D: Goodness of definitions of nonconformity measures

In this appendix we will check in detail that all our nonconformity measures A_n depend on the first $n-1$ observations only through the value of S_{n-1} (briefly this was explained in the main part of the paper). For the IID model and the IID–Gauss model this is obvious.

In the case of the Gauss linear model, to check that the expression on the right-hand side of (11) depends on the first $n-1$ observations only through the value of S_{n-1}, we can first check this for $\boldsymbol{Z}'_{n-1}\boldsymbol{Z}_{n-1}$, then for $\hat{\gamma}_{n-1}$, and finally for $\hat{y}_n$ and $\hat{\sigma}^2_{n-1}$ (the definition of the latter has to be expanded).

In the case of the MVA model, we are required to check that the right-hand side of (14) depends on the first $n-1$ observations only through the value of S_{n-1}. According to (12), $\mathbf{e} = \mathbf{y} - \boldsymbol{U}\mathbf{c}$, where $\mathbf{c}$ is a known vector. First notice

that e_n depends only on the nth observation and that $\overline{e}_{n-1}$ depends only on the average of $y_1, \ldots, y_{n-1}$ and the average of $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$. Therefore, both e_n and $\overline{e}_{n-1}$ (equivalently, $\sum_{i=1}^{n-1} e_i$) depend on the first $n-1$ observations only through the value of S_{n-1}. It remains to show that $\sum_{i=1}^{n-1} e_i^2$ depends on the first $n-1$ observations only through the value of S_{n-1}. This immediately follows from the definition of S_{n-1}.

References

[1] C. B. Bell, David Blackwell, and Leo Breiman. On the completeness of order statistics. *Annals of Mathematical Statistics*, 31:794–797, 1960.

[2] Lawrence D. Brown. An ancillarity paradox which appears in multiple linear regression (with discussion). *Annals of Statistics*, 18:471–538, 1990.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[5] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[6] A. Philip Dawid. Statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society A*, 147:278–292, 1984.

[7] Joaquín Diaz, Federico J. O'Reilly, and Santiago Rincon-Gallardo. A set of independent sequential residuals for the multivariate regression model. *Journal of Statistical Planning and Inference*, 8:21–25, 1983.

[8] Ronald A. Fisher. Applications of "Student's" distribution. *Metron*, 5:90–104, 1925.

[9] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, New York, third edition, 1973.

[10] Donald A. S. Fraser. *Nonparametric Methods in Statistics*. Wiley, New York, 1957.

[11] Alex Gammerman and Vladimir Vovk. Hedging predictions in machine learning (with discussion). *Computer Journal*, 50:151–177, 2007. A version of this paper (without the discussion) is published as `arXiv:cs/0611011` [cs.LG].

[12] Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*, volume 49 of *Lecture Notes in Statistics*. Springer, New York, 1988.

[13] Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer, New York, second edition, 1986.

[14] L. Mattner. Complete order statistics in parametric models. *Annals of Statistics*, 24:1265–1282, 1996.

[15] Federico O'Reilly and C. P. Quesenberry. The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests. *Annals of Statistics*, 1:74–83, 1973.

[16] Allan R. Sampson. A tale of two regressions. *Journal of the American Statistical Association*, 69:682–689, 1974.

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

[18] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[19] Kei Takeuchi. *Statistical Prediction Theory (in Japanese)*. Baihūkan, Tokyo, 1975.

[20] Stijn Vanderlooy, Laurens van der Maaten, and Ida Sprinkhuizen-Kuyper. Off-line learning with Transductive Confidence Machines: an empirical evaluation. In Petra Perner, editor, *Proceedings of the Fifth International Conference on Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Artificial Intelligence*, pages 310–323, Berlin, 2007. Springer.

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

[22] Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37:1566–1590, 2009. (Also published as `arXiv:0906.3123` [math.ST].) This is a short version of this paper.

[23] Samuel S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12:91–96, 1941.