On-line predictive linear regression

Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman
Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
{vovk,ilia,alex}@cs.rhul.ac.uk

February 2, 2008

Abstract

Gauss linear model; independent identically distributed observations; multivariate analysis; on-line protocol; prequential statistics; regression 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. The standard treatment of prediction in linear regression analysis has two drawbacks: (1) the usual 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 does equal the nominal significance level, up to statistical fluctuations, and we 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.

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) normal random variables with mean 0 and variance $\sigma^2 > 0$ (we will write $\xi_n \sim N(0, \sigma^2)$). The model (1) will be called the *Gauss linear model* (following Seal's 1967 suggestion).

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. In §2 we formally introduce the on-line prediction protocol, with a more detailed discussion postponed to §7. In §3 we note an important advantage of the on-line protocol: the true responses fall outside the standard prediction intervals independently for different observations; in combination with the law of large numbers this implies that their frequency of error is approximately equal to the nominal significance level. In §4 this result is stated for a wide class of models and a wide class of prediction strategies.

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. Sometimes realistic models will be non-parametric, effectively involving infinitely many parameters (as in §6). In §4 we state a result (theorem 2) suggesting that the Gauss linear model is too restrictive to permit informative prediction intervals in such cases.

In §§5–6 we consider three alternatives to the Gauss linear model, none of which require that the number of observations should exceed the number of parameters. We start from a regression model that has also been widely discussed in the statistical literature (the other of Sampson's 1974 two regressions); we call it the *MA model* (with MA referring to "multivariate analysis"). This 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 normal random vectors. Fisher (1973, §IV.3) emphatically defended the use of the Gauss linear model even in the case where the distribution of the explanatory vector is known (with or without parameters). There is also a view in the literature that the Gauss linear model and the MA model are "essentially equivalent" (for a review of some results in this direction, see Sampson 1974). Our conclusion, however, is similar to Brown's (1990): when the MA 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.

In §6 we explore regression in what we call the *de Finetti model*: it is only assumed that the sequence of pairs $(\mathbf{x}_n, y_n)$ is IID. Despite the non-parametric nature of this model, it also allows one to obtain informative prediction intervals before the number of observations reaches the number of parameters. The de Finetti 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. At the end of §6 we consider the com-

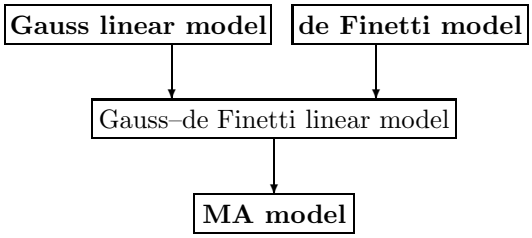

Figure 1: The four models considered in this paper (the three main models are given in boldface).

bination of the Gauss linear and de Finetti models, which we call the *Gauss–de Finetti linear model*: in addition to (1) we assume that the explanatory vectors $\mathbf{x}_n$, $n = 1, 2, \ldots$, are random and IID and that the sequence $\xi_1, \xi_2, \ldots$ is independent of the explanatory vectors. This model, however, appears to be of secondary importance.

The models considered in this paper are shown in figure 1, with arrows leading from more general to more specific models (formally, a statistical model is more general than another statistical model if the convex hull of the second model is a subset of the convex hull of the first model). For each model we will define a suitable prediction strategy; it is natural to expect that more specific models, when true, will lead to better predictions.

We will be interested in two criteria of quality of prediction strategies, which we call "validity" and "accuracy". For valid prediction strategies, 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", prediction strategies). The second criterion is applied only to valid prediction strategies: we want the prediction intervals to be as narrow as possible; in this paper we, somewhat arbitrarily, measure the narrowness of a prediction interval by its Euclidean length. In particular, we want the prediction intervals to become bounded as soon as possible.

The idea of learning complex systems from a small number of observations is familiar in machine learning and has also become popular in statistics (see, e.g., Lindsay *et al.* 2004, §3.3.4). In the context of this paper, this is a feasible goal. First, such learning has a limited purpose: prediction of the future responses. Many aspects of the system are irrelevant or not very important for prediction. Second, one often has *a priori* information about the system: 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 model explicitly, since it would destroy the validity of our prediction strategy if this information happened to be far from the truth, we might still be able to use such information in designing the prediction strategy provided our model is flexible enough. A running example in this paper, introduced in the next section, will be a linear system with 100 parameters ten of which are felt to be especially

important.

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, 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 Dawid (1984).

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 §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 $\epsilon \in (0, 1)$;
  $\mathrm{lth}_n^\epsilon := \mathrm{length}(\mathrm{co}\,\Gamma_n^\epsilon)$ for all $\epsilon \in (0, 1)$
END FOR.

(As usual, $\mathrm{co}\,E$ stands for the convex hull of the set E in a linear space and $\mathbb{I}_F$ 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* (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 lth_n^ϵ of the corresponding *prediction interval* $\mathrm{co}\,\Gamma_n^\epsilon$ (as usual, the length of an interval with end-points a and b is defined to be $|a - b|$).

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". A measurable prediction strategy in the on-line protocol (or, as we will say, *confidence predictor*) is *weakly valid* in some statistical model (such as (1)) 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. (Cf. Cox & Hinkley 1974, (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. A *strongly valid* confidence predictor is one for which, in addition, the events $\mathrm{err}_n^\epsilon = 1$, $n = 1, 2, \ldots$, are independent.

Figure 5 below shows the plot of Err_n^ϵ against n for a specific confidence predictor constructed 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). This is the only figure in this paper illustrating the validity of our prediction strategies: 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 Lth_n^ϵ of the sequence $\mathrm{lth}_1^\epsilon, \ldots, \mathrm{lth}_n^\epsilon$; again, this measure is arbitrary, to a large degree. A plot of Lth_n^ϵ against n will be called the *median-accuracy plot*; examples of such plots are given in figures 2–4 and 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 standard 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 N be a set of positive integer numbers (we are mainly interested in the case where N has the form $\{m, m+1, \ldots\}$). We say that a confidence predictor is *weakly valid for* $n \in N$ 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 $n \in N$ and $\epsilon \in (0, 1)$. It is *strongly valid for* $n \in N$ if, in addition, err_n^ϵ, $n \in N$, 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 MA 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 The Gauss linear model

The Gauss linear model (1) can be written as

$$y_n = \boldsymbol{\gamma} \cdot \mathbf{z}_n + \xi_n, \tag{2}$$

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 $\mathbf{z}_i'$, $i = 1, \ldots, l$, $\mathbf{y}_l$ be the vector whose ith element is y_i, $i = 1, \ldots, l$, and $\hat{\boldsymbol{\gamma}}_l := (\boldsymbol{Z}_l' \boldsymbol{Z}_l)^{-1} \boldsymbol{Z}_l' \mathbf{y}_l$ be the least squares estimate of the parameter vector $\boldsymbol{\gamma}$ in (2) from the first l

observations. We will sometimes refer to the first column of $\boldsymbol{Z}_l$ as the *dummy* column. 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 $l \geq K+1$.

It is well known that in the Gauss linear model the ratio

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

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

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

is the standard estimate of σ^2 from $\boldsymbol{Z}_l$ and $\mathbf{y}_l$, has the t-distribution with $n-K-2$ degrees of freedom. This gives the standard weakly valid prediction interval for the nth response,

$$\Gamma_n^\epsilon := \begin{cases} \left\{ y \in \mathbb{R} : |y - \hat{y}_n| < t_{n-K-2}^{\epsilon/2}\sqrt{1 + \mathbf{z}_n'(\boldsymbol{Z}_{n-1}'\boldsymbol{Z}_{n-1})^{-1}\mathbf{z}_n}\hat{\sigma}_{n-1} \right\} & n \geq K+3 \\ \mathbb{R} & \text{otherwise,} \end{cases} \tag{4}$$

where t_m^δ is the upper δ point of the t-distribution with m degrees of freedom. (See, e.g., Seber & Lee 2003, (5.27).)

Proposition 1 *The events $y_n \notin \Gamma_n^\epsilon$, $n = K+3, K+4, \ldots$, are independent. In particular, the confidence predictor (4) is strongly valid for $n \geq K+3$.*

Remark We have not seen proposition 1 stated explicitly in the literature, but it and, more generally, the fact that the statistics (3) are independent, can be regarded as known. Lemma 1 in Brown *et al.* (1975) asserts that (3) with $\hat{\sigma}_{n-1}$ removed are independent $N(0, \sigma^2)$ random variables. (This can be used for prediction when the standard deviation σ is known.) Seillier-Moiseiwitsch (1993, Example 1) proves that (3) are independent when $K = 0$. It is interesting that both papers use the independence of (3) for testing rather than for prediction.

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 $\boldsymbol{\beta}$:

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

We will suppose the statistician analyzing these data knows, or suspects, that the first 10 explanatory variables are much more important than the rest.

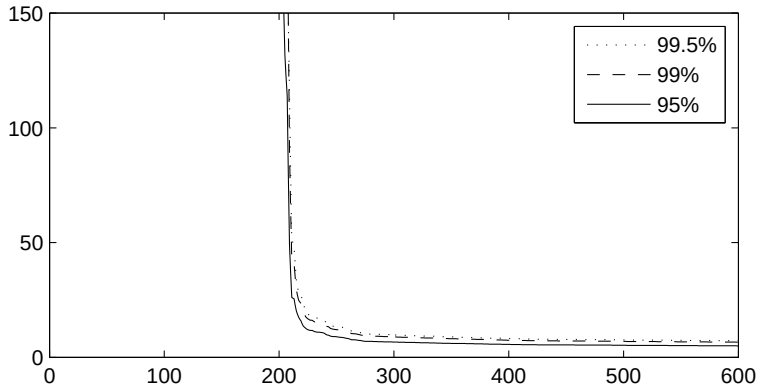

Figure 2: The median-accuracy plot for the standard prediction intervals. The three significance levels used in this and all the following figures are $\epsilon = 0.05, 0.01, 0.005$, shown in the form $100(1-\epsilon)\%$ (the corresponding confidence levels) in the legends.

We have already mentioned that the standard confidence predictor, (4), does not work when there are many parameters; in particular, it is required that $n \geq K + 3$. In the next section we will see 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 $n \geq K + 3$. Figure 2 gives the median-accuracy plot for the confidence predictor (4); 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 after $n = 205$.

4 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* Ω (we will be interested in the space $\Omega = \mathbb{R}^K \times \mathbb{R}$ of pairs $(\mathbf{x}, y)$; in general, Ω is a measurable space assumed to be Borel, to ensure the existence of regular conditional probabilities). To define conformal predictors, we will need not only a statistical model on Ω^∞ but also a sequence of sufficient statistics $S_n : \Omega^n \to \Sigma_n$; we will always assume that $\Sigma_n = S_n(\Omega^n)$. We will need a strengthened form of sufficiency; in our definitions we mainly follow Lauritzen (1988), §II.2.

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

$$S_n(\omega_1, \ldots, \omega_{n-1}, \omega_n) = F_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n)$$

for all $(\omega_1, \ldots, \omega_{n-1}, \omega_n) \in \Omega^n$. Intuitively, $S_n(\omega_1, \ldots, \omega_n)$ 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 Ω^∞ if, for each $n = 1, 2, \ldots$:

- S_n is sufficient for $\mathcal{P}$;
- $\omega_1, \ldots, \omega_n$ and $\omega_{n+1}, \omega_{n+2}, \ldots$ are conditionally independent given $S_n(\omega_1, \ldots, \omega_n)$, where $(\omega_1, \omega_2, \ldots) \sim P$, for any $P \in \mathcal{P}$.

The second condition ensures that $S_n(\omega_1, \ldots, \omega_n)$ carries all information in $\omega_1, \ldots, \omega_n$ that can be used for predicting the future observations $\omega_{n+1}, \omega_{n+2}, \ldots$.

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 a sequence (S_n) of ATTS statistics (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 Ω will be $\mathbb{R}^K \times \mathbb{R}$. In particular, the ATTS statistics for the Gauss linear model are

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

(It is natural to have $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as components of S_n, although in principle they are superfluous.)

Testing conformity

The main ingredient of conformal prediction is statistical testing of conformity of a new observation ω_n to the old observations $\omega_1, \ldots, \omega_{n-1}$. In general, our statistical tests will be randomized.

Fix a statistical model $\mathcal{P}$ with an ATTS sequence $S_n : \Omega^n \to \Sigma_n$. Any sequence of measurable functions $A_n : \Sigma_{n-1} \times \Omega \to \mathbb{R}$, $n = 1, 2, \ldots$, is called a *nonconformity measure*; A_n will be our test statistics. (We define Σ_0 to be a fixed one-element set.) Given such an (A_n), for each sequence $\omega_1, \omega_2, \ldots$ of

observations and each sequence $\tau_1, \tau_2, \ldots \in [0,1]^\infty$ we define the *p-values*

$$\begin{aligned} p_n = p_n(\omega_1, \ldots, \omega_n, \tau_n) &:= \mathbb{P}\Big(A_n(S_{n-1}(\xi_1, \ldots, \xi_{n-1}), \xi_n) \\ &> A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \mid S_n(\xi_1, \ldots, \xi_n) = S_n(\omega_1, \ldots, \omega_n)\Big) \\ &+ \tau_n \, \mathbb{P}\Big(A_n(S_{n-1}(\xi_1, \ldots, \xi_{n-1}), \xi_n) \\ &= A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \mid S_n(\xi_1, \ldots, \xi_n) = S_n(\omega_1, \ldots, \omega_n)\Big), \quad n = 1, 2, \ldots, \end{aligned} \tag{5}$$

where $(\xi_1, \xi_2, \ldots) \sim P$ for some $P \in \mathcal{P}$. (This definition uses fixed versions of regular conditional probabilities that do not depend on $P \in \mathcal{P}$.) We will be interested in two cases: *deterministic*, where $\tau_n = 1$ for all n, and *randomized*, where $\tau_1, \tau_2, \ldots$ are generated independently from the uniform distribution U on $[0,1]$ (such $\tau_1, \tau_2, \ldots$ model the output of a random numbers generator).

Theorem 1 *Suppose that the observations $\omega_n \in \Omega$, $n = 1, 2, \ldots$, are generated from a probability distribution $P \in \mathcal{P}$ and that the random numbers $(\tau_1, \tau_2, \ldots) \sim U^\infty$ are independent of the observations. The p-values (5) are then independent and distributed uniformly on $[0,1]$:*

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

For a proof of this theorem, see the appendix. The fact that $p_n \sim U$ is well known, at least in the continuous case (see, e.g., Cox & Hinkley 1974, p. 66; (5) is a version of Cox & Hinkley's (1)).

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 de Finetti model in §6 we will need an element (typically quite small) of randomization in confidence predictors. A *randomized confidence predictor* is a measurable function which maps every significance level $\epsilon \in (0,1)$, every data sequence $\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}$, every vector $\mathbf{x}_n$ of explanatory variables and every number $\tau \in [0,1]$ to a set $\Gamma^\epsilon_n = \Gamma^\epsilon(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n, \tau) \subseteq \mathbb{R}$; we will use the notation Γ^ϵ_n when the data sequence, the vector of explanatory variables and the number τ are clear from the context.

Let the observation space be $\Omega = \mathbb{R}^K \times \mathbb{R}$. Once the p-values (5) are defined, we can use them for confidence prediction (this is a standard procedure; cf. Cox & Hinkley 1974, (76) on p. 243): we set

$$\begin{aligned} &\Gamma^\epsilon(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n, \tau_n) \\ &\qquad := \{y \in \mathbb{R} : p_n((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n, y), \tau_n) > \epsilon\}. \end{aligned} \tag{6}$$

This randomized confidence predictor 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.

Corollary 1 *If the observations* $(\mathbf{x}_n, y_n)$ *are generated by a probability distribution* $P \in \mathcal{P}$ *and a smoothed conformal predictor is fed with random numbers* $(\tau_1, \tau_2, \ldots) \sim U^\infty$ *independent of the observations, the error sequence* $\mathrm{err}^\epsilon_1, \mathrm{err}^\epsilon_2, \ldots$ *at any significance level* ϵ *is Bernoulli with parameter* ϵ.

This immediately follows from theorem 1 and asserts that smoothed conformal predictors are strongly valid.

The adjective "smoothed" refers to using random numbers; if we take $\tau_n = 1$ for all $n = 1, 2, \ldots$, we will obtain the definition of a *deterministic conformal predictor*, or just *conformal predictor* (in this case we omit τ_n from our notation). 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 $\mathrm{err}^\epsilon_1, \mathrm{err}^\epsilon_2, \ldots$ is dominated by a Bernoulli sequence with parameter ϵ. In particular, whereas we have $\lim_{n\to\infty}(\mathrm{Err}^\epsilon_n / n) = \epsilon$ a.s. for smoothed conformal predictors, we only have $\limsup_{n\to\infty}(\mathrm{Err}^\epsilon_n / n) \le \epsilon$ a.s. for conformal predictors.

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 (5) is then zero. There is also no difference for the MA model and $n \ge 3$; however, the difference is important (although usually barely noticeable on error and accuracy plots) for the de Finetti model.

Proposition 1 is a special case of corollary 1 corresponding to the nonconformity measure

$$A_n\left(S_{n-1}\left(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}\right), (\mathbf{x}_n, y_n)\right) := \frac{|y_n - \hat{y}_n|}{\sqrt{1 + \mathbf{z}_n'(\boldsymbol{Z}_{n-1}'\boldsymbol{Z}_{n-1})^{-1}\mathbf{z}_n}\hat{\sigma}_{n-1}} \tag{7}$$

(cf. (3); the goodness of the definition follows from the formulas given at the beginning of §3). The expression on the right-hand side of (7) can be replaced by other natural expressions, such as $|y_n - \hat{y}_n|$—see Vovk *et al.* (2005), §8.5.

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.

We say that a confidence predictor is *invariant* if Γ^ϵ_n, $n > 1$, depends on the first $n - 1$ observations only through the value of S_{n-1}. (The use of invariant confidence predictors is natural in view of the sufficiency principle; see, e.g., Cox & Hinkley 1974, §2.3 (iii).) Let 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 $n \in N$ if

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

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

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

This theorem is also proved in the appendix. In some form it was known already in the late 1970s to Kei Takeuchi.

The condition of bounded completeness holds for the Gauss linear model and the MA model by the standard completeness result for exponential statistical models (see, e.g., theorem 4.1 in Lehmann 1986), and it is also known to hold for the de Finetti model (see the theorem on p. 797 in Bell *et al.* 1960 or theorem 1 in Mattner 1996).

Therefore, it is not a coincidence that the standard confidence predictor (4) does not work until n exceeds $K+2$: since the conditional distributions P_n are concentrated at one point for $n \le K+1$ 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^ϵ_n for $\epsilon < 0.5$ and $n \le K+2$.

5 The MA model

Remember that the MA model assumes, besides (1), that $\mathbf{x}_n$ are generated independently from the same multivariate normal distribution on $\mathbb{R}^K$, with the noise random variables $\xi_1, \xi_2, \ldots$ independent of $\mathbf{x}_1, \mathbf{x}_2, \ldots$. The ATTS statistics in the MA 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 ATTS statistics can be defined to be the empirical means and covariances of all variables, i.e., the response and the explanatory variables).

In the MA model, 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. We believe, however, that the first $K^\dagger_n \ll K$ of the explanatory variables are especially important, and it is feasible to estimate the corresponding β_k, $k = 1, \ldots, K^\dagger_n$, and α.

Fix a positive integer number n. We will write $\mathbf{y}$ for $\mathbf{y}_n$, $\boldsymbol{Z}$ for $\boldsymbol{Z}_n$ and $K^\dagger$ for $K^\dagger_n$. 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 to avoid the need to invert singular matrices, we obtain the vector of residuals

$$\mathbf{e} := \mathbf{y} - \boldsymbol{U} \left(\boldsymbol{U}' \boldsymbol{U} + a \boldsymbol{I} \right)^{-1} \boldsymbol{U}' \mathbf{y} = \mathbf{y} - \boldsymbol{U} \mathbf{c}; \tag{8}$$

notice that $\mathbf{c} := (\boldsymbol{U}'\boldsymbol{U} + a\boldsymbol{I})^{-1}\boldsymbol{U}'\mathbf{y}$ is a known vector when the value of the statistic S_n is known. Since the joint distribution of $\mathbf{y}$ and the non-dummy columns of $\boldsymbol{U}$ is invariant w.r. to rotations around the vector $\mathbf{1}$, the distribution of $\mathbf{e}$ will also be invariant w.r. to such rotations. (It might help the 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 $n = 1, 2$). A standard statistical result (stated in §7; see (14)) allows us to conclude that

$$\sqrt{\frac{n-1}{n}} \frac{e_n - \overline{e}_{n-1}}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}}, \tag{9}$$

where $e_1, \ldots, e_n$ are the components of the vector (8) of residuals and $\overline{e}_{n-1}$ is the average of $e_1, \ldots, e_{n-1}$, has the t-distribution with $n-2$ degrees of freedom.

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

$$A_n\left(S_{n-1}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n, y_n)\right) := \frac{e_n - \overline{e}_{n-1}}{\sqrt{\sum_{i=1}^{n-1}(e_i - \overline{e}_{n-1})^2}} \tag{10}$$

(proportional to (9); the fact that the right-hand side of (10) depends on the first $n-1$ observations only through the value of S_{n-1} can be seen from the representation (8), 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 (8), 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.

For use in our experiments with the artificial data set described in §3, we 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 (it is the first step when the standard prediction intervals (4) become bounded), but 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 $a = 0.01$.

For each model considered in this paper except the Gauss linear model we define a nonconformity measure involving the matrix $\boldsymbol{U}$ defined in the previous paragraph. In the case of the MA model, we use the nonconformity measure

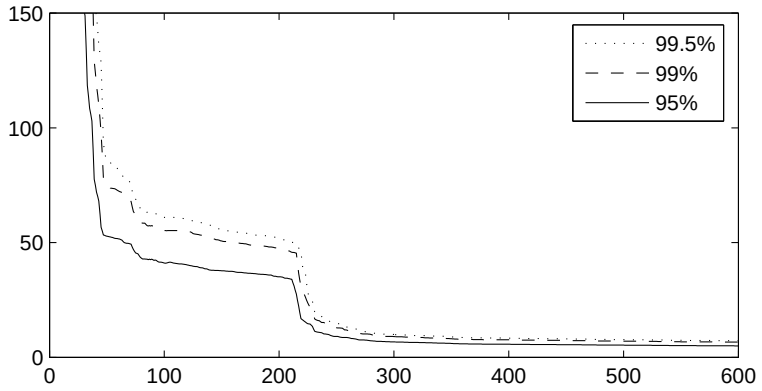

Figure 3: The median-accuracy plot for the MA predictor.

(10) and call the corresponding conformal predictor with Γ^ϵ_n replaced by $\operatorname{co}\Gamma^\epsilon_n$ the *MA predictor*. Of course, this brief term is somewhat misleading: it should always be borne in mind that the conformal predictor leading to the MA predictor is only one of many conformal predictors that can be defined in the MA model. Similarly, in the next section we will introduce the de Finetti predictor (called "Ridge Regression Confidence Machine" in Vovk *et al.* 2005) and the Gauss–de Finetti predictor, which will also correspond to specific nonconformity measures. In the same spirit, the confidence predictor (4) will be called the *Gauss predictor*.

The median-accuracy plot for the MA predictor and our artificial data set is shown in figure 3. 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 MA predictor is better than the performance of any other confidence predictor considered in this paper, but this, of course, 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 normal variables; even the MA 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 normal. In the following section we will introduce the de Finetti predictor, which only knows that the observations $(\mathbf{x}_n, y_n)$ are IID, and the Gauss–de Finetti predictor, which knows, in addition, that the y_n are generated by (1).

6 The de Finetti 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 normal noise. The ATTS statistics are

$$S_n := \{\!|(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\!\}, \tag{11}$$

where we use $\{\!|a_1, \dots, a_n|\!\}$ 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

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

where ξ_i are IID random elements taking values in $\mathbb{R}^K \times \mathbb{R}$, assigns (with probability one) the same probability, $1/n!$, to every ordering $(\mathbf{x}_{\pi(1)}, y_{\pi(1)}), \dots, (\mathbf{x}_{\pi(n)}, y_{\pi(n)})$ of the bag $\{\!|(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\!\}$.

We attach de Finetti's name to this model since de Finetti, in his study of exchangeability, was the first to understand the role of the statistics (11).

In the case of the de Finetti model, we will be interested in the conformal predictor determined by the nonconformity measure

$$A_n\left(S_{n-1}\left(\mathbf{x}_1, y_1, \dots, \mathbf{x}_{n-1}, y_{n-1}\right), (\mathbf{x}_n, y_n)\right) := |e_n|, \tag{12}$$

where we continue to use $e_1, \dots, e_n$ for denoting the components of the vector of residuals (8). (Deleted and, especially, studentized residuals would also be a natural choice—see, e.g., Vovk *et al.* 2005, pp. 34–35; in our experience, however, the difference is not significant, and we stick to the simplest choice.) As usual, we call the confidence predictor obtained from this conformal predictor by replacing the prediction regions Γ^ϵ_n with the prediction intervals $\operatorname{co}\Gamma^\epsilon_n$ simply the *de Finetti predictor*.

The de Finetti predictor can be implemented fairly efficiently. First notice that for the de Finetti model the formula (5) for p-values can be simplified to

$$p_n = \frac{|\{i : \alpha_i > \alpha_n\}| + \tau_n |\{i : \alpha_i = \alpha_n\}|}{n}, \tag{13}$$

where $\alpha_i := A_n(\{\!|\omega_1, \dots, \omega_{i-1}, \omega_{i+1}, \dots, \omega_n|\!\}, \omega_i)$, i ranges over $\{1, \dots, n\}$, and $|E|$ stands for the size of the set E. In the case of the nonconformity measure (12), $\alpha_i = |e_i|$. The residuals (8) 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 $e_i = e_i(y)$ become linear functions of y (this fact was already used in the previous section). By (13) with $\tau_n := 1$, the p-value is the fraction

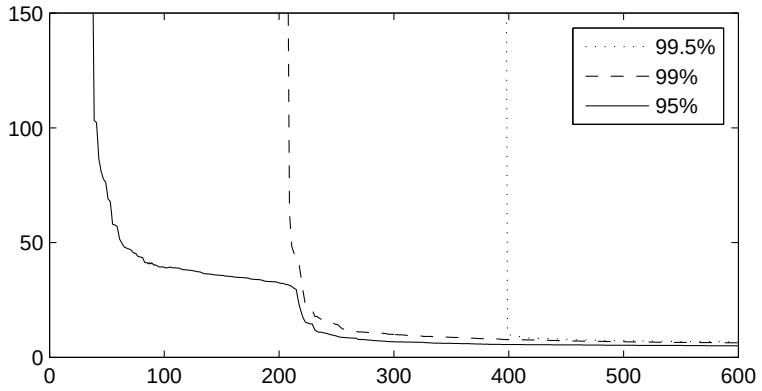

Figure 4: The median-accuracy plot for the de Finetti predictor.

of $i = 1, \ldots, n$ 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 $2n$ 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 $4n + 1$ 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^ϵ_n as the union of the intervals with p-values exceeding ϵ. The computation of the de Finetti prediction interval $\operatorname{co}\Gamma^\epsilon_n$ can be simplified if we notice that the set Γ^ϵ_n is closed (which is opposite to what we have for the Gauss linear and MA models): assuming that the set of critical points is non-empty, $\operatorname{co}\Gamma^\epsilon_n$ is bounded if and only if the two unbounded intervals have p-values at most ϵ, in which case the end-points of $\operatorname{co}\Gamma^\epsilon_n$ can be found as the leftmost and rightmost critical points with p-values exceeding ϵ. Computing Γ^ϵ_n and $\operatorname{co}\Gamma^\epsilon_n$ 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 Vovk *et al.* 2005, p. 33).

As figure 4 shows, the de Finetti predictor works well for our data set if the significance level is not too demanding: it is clear that for the de Finetti prediction interval $\operatorname{co}\Gamma^\epsilon_n$ to be bounded the number of observations n has to be at least $1/\epsilon$. The median-accuracy plot for $\epsilon = 5\%$ is almost as good as the corresponding plot for the MA predictor. For the significance level $\epsilon = 0.5\%$, the de Finetti predictor requires 200 observations to produce bounded predictions, and this shows on the median-accuracy plot at $n = 399$. At the significance level $\epsilon = 1\%$ the de Finetti predictor performs about the same as the Gauss predictor, but for a different reason: $1/\epsilon$ just happens to coincide with K.

The de Finetti model is non-parametric but we can see that it still admits valid predictors (or conservative predictors if one insists on using deterministic

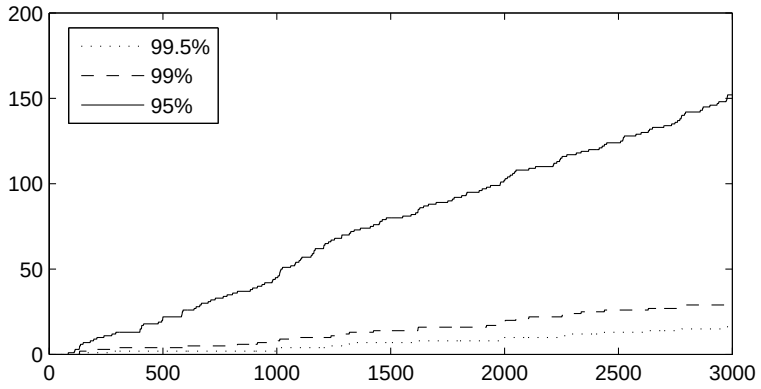

Figure 5: The cumulative numbers of errors made by the de Finetti predictor: Err_n^ϵ is plotted against n.

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

Theorem 2 is not directly applicable to the de Finetti model, since only smoothed conformal predictors are valid, as the latter term is used in this paper. In Vovk *et al.* (2005), §2.4, we state two results of the same nature about the de Finetti model.

There are two sources of conservativeness for the de Finetti predictor as described above (and used for producing figures 4). 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 5) show that we still have approximate validity.

The Gauss–de Finetti linear model

As defined in §1, the Gauss–de Finetti linear model is the combination of the Gauss linear and de Finetti 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

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

Using the nonconformity measure (12) and replacing the prediction regions output by the corresponding conformal predictor with their convex hulls, we obtain the *Gauss–de Finetti predictor*. Its performance on our usual data set is shown in figure 6. We do not know whether the Gauss–de Finetti predictor

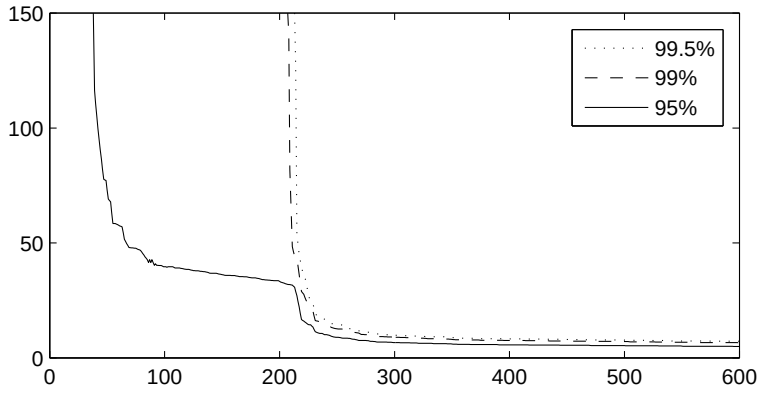

Figure 6: The median-accuracy plot for the Gauss–de Finetti predictor.

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 4 (to the left of $n = 205$) and 2 (to the right of $n = 205$), we can see that the following simple prediction strategy will work almost as well as the Gauss–de Finetti predictor on our data set: predict using the de Finetti 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.)

7 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 MA model and the Gauss–de Finetti model also reduce to the Gaussian model when $K = 0$.

The fact that

$$T_n := \sqrt{\frac{n-1}{n}} \frac{y_n - \overline{y}_{n-1}}{\hat{\sigma}_{n-1}}, \tag{14}$$

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 (Fisher 1925) 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

$$\Gamma_n^\epsilon := \left\{ y \in \mathbb{R} : \left| y - \overline{y}_{n-1} \right| < t_{n-2}^{\epsilon/2} \sqrt{\frac{n}{n-1}} \hat{\sigma}_{n-1} \right\}, \quad n = 3, 4, \ldots, \tag{15}$$

and $\epsilon \in (0,1)$ is the chosen significance level. This prediction interval is a special case of (4).

Fisher discussed (15) and related confidence predictors in his last book (Fisher 1973, §§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 §§V.3–4 are all applicable to the predictor (15), although in §V.3 he discusses prediction of exponentially rather than normally distributed 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 (15), 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 (15). Fix a significance level $\epsilon \in (0,1)$ and $l \in \{2, 3, \ldots\}$ (the *sample size*; we might consider samples of different sizes, but we will stick to the simplest case). For $m = 1, 2, \ldots$, generate the *mth sample*

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

and the *mth test observation* $y_{m(l+1)}$. Register an error if the mth prediction interval computed from the mth sample according to (15) 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 $\mathrm{err}_m^\dagger$, $m = 1, 2, \ldots$, are independent. The frequency of error gets arbitrarily close to ϵ with an arbitrarily high probability as the number of observations increases.

The 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.

Fisher's theory of fiducial prediction is based on the fact that a value such as (14) 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 (14) has a known distribution can be strengthened: theorem 1 (applied to the nonconformity measure (14)) 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 de Finetti model

The de Finetti 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 de Finetti model, which says that $y_1, y_2, \ldots$ form an IID sequence, was undertaken by many authors following the pioneering paper by Wilks (1941). (This work was later extended to the multivariate case: see, e.g., Fraser 1957; this extension, however, is not directly related to our de Finetti 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 (Vovk *et al.* 2005, §4.3). 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).

8 Conclusion

In this paper we considered the problem of prediction in three main models for linear regression. One of these models, the Gauss linear model, is the standard textbook one. The MA model seems to have been somewhat neglected, partly because of philosophical reasons (one conditions on the observed values of the explanatory variables to make the prediction, or estimate, etc., more relevant). In 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 accept that this makes the theorem less convincing.) It remains a mystery to us why the de Finetti model has been completely neglected in the field of regression, even in non-parametric statistics, where the value of the de Finetti model is in principle well understood.

Acknowledgments

This paper has greatly benefited from Glenn Shafer's comments, and we are grateful to Steffen Lauritzen for a useful discussion. This work was partially supported by MRC (grant S505/65) and the Royal Society.

Appendix: 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 §8.7 of Vovk *et al.* (2005), but we reproduce the principal points of the proof to make our exposition self-contained. A special case of theorem 2 (namely, for the de Finetti model) was proved in §2.6 of Vovk *et al.* (2005).

Proof of theorem 1

In this proof, $\omega_1, \omega_2, \dots$ will be random observations generated by $P \in \mathcal{P}$, $(\omega_1, \omega_2, \dots) \sim P$, 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

$$S_n(\omega_1, \dots, \omega_n), \omega_{n+1}, \tau_{n+1}, \omega_{n+2}, \tau_{n+2}, \dots .$$

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 $\epsilon \in (0, 1)$,*

$$\mathbb{P}_{\mathcal{G}_n}(p_n \le \epsilon) = \epsilon. \tag{16}$$

Proof For a given value of the summary $S_n(\omega_1, \dots, \omega_n)$ of the first n observations, consider the conditional distribution function F of the random variable $\eta := A_n(S_{n-1}(\omega_1, \dots, \omega_{n-1}), \omega_n)$ (because of the total sufficiency, it does not matter whether we further condition on $\omega_{n+1}, \tau_{n+1}, \omega_{n+2}, \tau_{n+2}, \dots$). Define $F(x-)$ to be $\sup_{t<x} F(t)$. Therefore, our task reduces to showing 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 (5)). The latter fact is usually stated in statistics textbooks for continuous F (see, e.g., Cox & Hinkley 1974, §3.2), 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 ω_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, $\Omega := \mathbb{R}^K \times \mathbb{R}$ and ω_i stands for $(\mathbf{x}_i, y_i)$. Let $n \in N$.

For each summary $s \in \Sigma_n$ let $f(s)$ be the conditional probability given $S_n(\omega_1, \ldots, \omega_n) = s$ that Γ makes an error at a significance level ϵ when predicting y_n from $\omega_1, \ldots, \omega_{n-1}$ and $\mathbf{x}_n$, the observations $\omega_1, \omega_2, \ldots$ being generated from $P \in \mathcal{P}$. We know that the expected value of $f(S_n(\omega_1, \ldots, \omega_n))$ 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 $(s', \omega) = (s', (\mathbf{x}, y)) \in \Sigma_{n-1} \times \Omega$ such that $F_n(s', \omega) = s$ (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 $\omega_1, \ldots, \omega_{n-1}$ satisfying $S_n(\omega_1, \ldots, \omega_{n-1}) = s'$ and with $\mathbf{x}$ (since Γ is invariant, whether an error is made depends only on s', not on the particular $\omega_1, \ldots, \omega_{n-1}$). It is clear that

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

and

$$\mathbb{P}\left((S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \in E(s, \epsilon) \mid S_n(\omega_1, \ldots, \omega_n) = s\right) = \epsilon \quad \text{a.s.},$$

where $(\omega_1, \omega_2, \ldots) \sim P \in \mathcal{P}$.

In this proof we say "conformity measure" to mean a nonconformity measure which is used for computing p-values in the opposite way to (5): the "$>$" in (5) is replaced by "$<$". Let us check that the conformal predictor $\Gamma^\dagger$ determined by the conformity measure

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

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

$$\begin{aligned}
&\mathbb{P}\left(A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \le \epsilon \mid S_n(\omega_1, \ldots, \omega_n) = s\right) \\
&\quad = \lim_{\delta \downarrow \epsilon} \mathbb{P}\left(A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) < \delta \mid S_n(\omega_1, \ldots, \omega_n) = s\right) \\
&\le \lim_{\delta \downarrow \epsilon} \mathbb{P}\left((S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \in E(s, \delta) \mid S_n(\omega_1, \ldots, \omega_n) = s\right) = \lim_{\delta \downarrow \epsilon} \delta = \epsilon \quad \text{a.s.},
\end{aligned}$$

where $(\omega_1, \omega_2, \ldots) \sim P \in \mathcal{P}$ and δ is constrained to be a rational number. Therefore, at each significance level ϵ and for all $(\omega_1, \ldots, \omega_n) \in \Omega^n$,

$$\begin{aligned}
y_n \in (\Gamma^\dagger)^\epsilon(\omega_1, \ldots, \omega_{n-1}, \mathbf{x}_n) &\iff \mathbb{P}\big(A_n(S_{n-1}(\xi_1, \ldots, \xi_{n-1}), \xi_n) \\
&\le A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \mid S_n(\xi_1, \ldots, \xi_n) = S_n(\omega_1, \ldots, \omega_n)\big) > \epsilon \\
&\implies A_n(S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) > \epsilon \\
&\implies (S_{n-1}(\omega_1, \ldots, \omega_{n-1}), \omega_n) \notin E(S_n(\omega_1, \ldots, \omega_n), \epsilon) \\
&\iff y_n \in \Gamma^\epsilon(\omega_1, \ldots, \omega_{n-1}, \mathbf{x}_n) \quad \text{a.s.},
\end{aligned}$$

where $(\xi_1, \xi_2, \ldots) \sim P \in \mathcal{P}$.

References

Bell, C. B., D. Blackwell, and L. Breiman (1960). On the completeness of order statistics. *Annals of Mathematical Statistics 31*, 794–797.

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

Brown, R. L., J. Durbin, and J. M. Evans (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *Journal of the Royal Statistical Society* B *37*, 149–192.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society* A *147*, 278–292.

Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron 5*, 90–104.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference* (Third ed.). New York: Hafner.

Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. New York: Wiley.

Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics*, Volume 49 of *Lecture Notes in Statistics*. New York: Springer.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (Second ed.). New York: Springer.

Lindsay, B. G., J. Kettenring, and D. O. Siegmund (2004). A report on the future of statistics (with discussion). *Statistical Science 19*, 387–413.

Mattner, L. (1996). Complete order statistics in parametric models. *Annals of Statistics 24*, 1265–1282.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association 69*, 682–689.

Seal, H. L. (1967). Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. *Biometrika 54*, 1–24.

Seber, G. A. F. and A. J. Lee (2003). *Linear Regression Analysis* (Second ed.). Hoboken, NJ: Wiley.

Seillier-Moiseiwitsch, F. (1993). Sequential probability forecasts and the probability integral transform. *International Statistical Review 61*, 395–408.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. New York: Springer.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics 12*, 91–96.