Conditional prediction intervals for linear regression

Peter McCullagh
Department of Statistics, University of Chicago
5734 University Avenue, Chicago, IL 60637-1514, USA
pmcc@galton.uchicago.edu

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

Abstract

We construct prediction intervals for the linear regression model with IID errors with a known distribution, not necessarily Gaussian. The coverage probability of our prediction intervals is equal to the nominal confidence level not only unconditionally but also conditionally given a natural σ-algebra of invariant events. This implies, in particular, the perfect calibration of our prediction intervals in the on-line mode of prediction.

1. Introduction

We are interested in procedures for finding a prediction interval $[L, U]$ for the label y of a new (*test*) instance $\mathbf{x}$ given a training set of labelled instances (*examples*) of a fixed size N. The classical notion of prediction intervals only requires that the unconditional coverage probability $\mathbb{P}(y \in [L, U])$ should be equal to the chosen confidence level $1 - \epsilon$, $\epsilon \in (0, 1)$, when both the training set and test example are generated from a given statistical model. This requirement of "unconditional validity" for classical prediction intervals, which they share with confidence intervals originated by Neyman, has often been criticised, starting from Fisher's work ([4] being particularly important).

The other extreme is provided by the case where the labels are generated by a known probability measure (such situations often arise in Bayesian statistics: see, e.g., [9], Section 5.2.4). We can then choose prediction intervals (say, the narrowest or symmetric ones) whose conditional coverage probability is $1 - \epsilon$, $\mathbb{P}(y \in [L, U] \mid \mathcal{F}_N) = 1 - \epsilon$, where " $\mid \mathcal{F}_N$" stands for conditioning on the training set (precise definitions will be given later) and the conditional distribution is assumed continuous. Unfortunately, such guarantees are impossible for the statistical models that we are interested in in this paper (see Appendix A).

In Section 3 we use the pivotal method to define prediction intervals that satisfy $\mathbb{P}(y \in [L, U] \mid \mathcal{K}_N) = 1 - \epsilon$, where " $\mid \mathcal{K}_N$" stands, intuitively, for "conditioning on the relevant aspects of the data set". Our approach is similar to that of [7], but the pivotal method itself goes back to [4]. When applied in the on-line mode our prediction intervals lead to the frequency of error close, with high probability, to the chosen significance level (Section 5). Our empirical studies (Section 4) show that this is approximately true in the batch mode of prediction as well.

2. The Gosset measure

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be a fixed arbitrary sequence of vectors in $\mathbb{R}^K$, for some $K \in \mathbb{N}$; they will play the role of instances. Our statistical model $(P_{\beta,\sigma})$, parameterized by $(\beta, \sigma) \in \mathbb{R}^K \times (0, \infty)$, is that the sequence of labels $y_1, y_2, \ldots$ is generated by

$$y_n = \beta' \mathbf{x}_n + \sigma \xi_n, \tag{1}$$

where $\xi_1, \xi_2, \ldots$ is a sequence of IID noise variables with a known distribution P. Each $P_{\beta,\sigma}$ is a probability measure on $\mathbb{R}^\infty$ (the distribution of the sequence of labels, with the instances fixed). We will assume that P is a continuous probability measure on $\mathbb{R}$ with density p.

Remark 1. The model (1) may appear narrower than the IID model, standard in machine learning, but its advantage is that the instances $\mathbf{x}_n$ can be controlled rather than chosen independently from the same distribution: instead of saying that $\mathbf{x}_n$ are fixed we could say that they are generated by an arbitrary stochastic process with (1) holding conditionally on the realized values of $\mathbf{x}_n$.

Let $\mathcal{G}$ be the group of all transformations

$$g_{a,b} : (y_1, y_2, \ldots) \mapsto (a'\mathbf{x}_1 + by_1, a'\mathbf{x}_2 + by_2, \ldots),$$

where $a \in \mathbb{R}^K$ and $b > 0$, acting on $\mathbb{R}^\infty$. We consider two fundamental σ-algebras on the set $\mathbb{R}^\infty$ of all infinite sequences $(y_1, y_2, \ldots)$ of real numbers. The σ-algebra $\mathcal{F}$ consists of all Borel sets in $\mathbb{R}^\infty$ (it will be our default σ-algebra on $\mathbb{R}^\infty$); by *events* we mean elements of $\mathcal{F}$. For $n = 1, 2, \ldots$, let $Y_n : \mathbb{R}^\infty \to \mathbb{R}$ be the projection onto the nth component: $Y_n(y_1, y_2, \ldots) := y_n$. For each $N = 0, 1, \ldots$, the σ-algebra $\mathcal{F}_N$ on $\mathbb{R}^\infty$ is defined as $\mathcal{F}_N := \sigma(Y_1, \ldots, Y_N)$. The σ-algebra $\mathcal{K}$ on $\mathbb{R}^\infty$ consists of all *invariant* events in $\mathbb{R}^\infty$, i.e., all events that are invariant with respect to the group $\mathcal{G}$. The σ-algebra $\mathcal{K}_N$ on $\mathbb{R}^\infty$ is defined as $\mathcal{K}_N := \mathcal{F}_N \cap \mathcal{K}$.

Lemma 1. *All measures $P_{\beta,\sigma}$ coincide on $\mathcal{K}$.*

Proof. This follows from the fact that all $P_{\beta,\sigma}$ are images of each other under the transformations in $\mathcal{G}$. □

Fix a positive integer number N, the size of the training set (at the beginning of the next section we will impose a mild restriction on N: it should not be too small). Let $(\tilde\beta, \tilde\sigma)$ be an estimator of the parameter (β, σ) that is *equivariant*: if

$$(z_1, z_2, \ldots) = (a'\mathbf{x}_1 + by_1, a'\mathbf{x}_2 + by_2, \ldots), \qquad (2)$$

where $a \in \mathbb{R}^K$, $b > 0$, and $y_n, z_n \in \mathbb{R}$ for all n, then

$$\tilde\beta(z_1, \ldots, z_N) = a + b\tilde\beta(y_1, \ldots, y_N),$$
$$\tilde\sigma(z_1, \ldots, z_N) = b\tilde\sigma(y_1, \ldots, y_N).$$

When deriving our prediction algorithm in the next section, we will choose a specific equivariant estimator, but all results stated before we do so hold for any equivariant estimator. Set

$$\nu(y_1, y_2, \ldots) := \left(\frac{y_1 - \tilde\beta'(y_1, \ldots, y_N)\mathbf{x}_1}{\tilde\sigma(y_1, \ldots, y_N)}, \frac{y_2 - \tilde\beta'(y_1, \ldots, y_N)\mathbf{x}_2}{\tilde\sigma(y_1, \ldots, y_N)}, \ldots \right);$$

we will sometimes call ν the *normalizing transformation*. By $\nu_n(y_1, y_2, \ldots)$ we will mean the nth element of the sequence $\nu(y_1, y_2, \ldots)$, and by $\nu(y_1, \ldots, y_N)$ we will mean the first N elements of the sequence $\nu(y_1, y_2, \ldots)$ (there is no dependence on y_n, $n > N$). In statistics, $\nu(y_1, \ldots, y_N)$ is known as the configuration statistic for the training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$; the function ν is maximal invariant with respect to the group $\mathcal{G}$.

Lemma 2. *The mapping ν is $\mathcal{K}$-measurable.*

Proof. Let us abbreviate $\tilde\beta(y_1, \ldots, y_N)$ to $\tilde\beta_y$, $\tilde\sigma(y_1, \ldots, y_N)$ to $\tilde\sigma_y$, $\tilde\beta(z_1, \ldots, z_N)$ to $\tilde\beta_z$, and $\tilde\sigma(z_1, \ldots, z_N)$ to $\tilde\sigma_z$. Since ν is $\mathcal{F}$-measurable, it suffices to check that ν is *invariant*, i.e., constant on each orbit. This follows immediately from the equivariance of the estimator $(\tilde\beta, \tilde\sigma)$: Equation (2) implies

$$\begin{aligned}\nu_n(z_1, z_2, \ldots) &= \frac{z_n - \tilde\beta_z'\mathbf{x}_n}{\tilde\sigma_z} \\ &= \frac{a'\mathbf{x}_n + by_n - (a + b\tilde\beta_y)'\mathbf{x}_n}{b\tilde\sigma_y} \\ &= \frac{y_n - \beta_y'\mathbf{x}_n}{\tilde\sigma_y} = \nu_n(y_1, y_2, \ldots). \qquad \square\end{aligned}$$

Corollary 1. *The distribution of $\nu(Y_1, Y_2, \ldots)$ under $P_{\beta,\sigma}$ does not depend on (β, σ).*

Proof. This follows immediately from Lemmas 1 and 2. □

By the *Gosset measure* G (with respect to the density p, equivariant estimator $(\tilde\beta, \tilde\sigma)$, and sample size N) we will mean the image $P_{\beta,\sigma}\nu^{-1}$ of any measure $P_{\beta,\sigma}$ under the normalizing transformation ν. Corollary 1 says that it does not matter which β and σ we take.

The following corollary strengthens Lemma 2.

Corollary 2. $\mathcal{K} = \sigma(\nu)$.

Proof. By Lemma 2, $\mathcal{K} \supseteq \sigma(\nu)$. On the other hand, since $\nu(y_1, y_2, \ldots)$ always belongs to the same orbit as $(y_1, y_2, \ldots)$, we have $E \in \mathcal{K} \implies E = \nu^{-1}(E) \in \sigma(\nu)$, i.e., $\mathcal{K} \subseteq \sigma(\nu)$. □

3. Prediction intervals for the linear regression model

In this section a training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ of the fixed size N will usually be represented as the $N \times K$ matrix $\mathbf{X}$ whose rows are the vectors $\mathbf{x}_n'$, $n = 1, \ldots, N$, and the $N \times 1$ vector $\mathbf{y}$ of all y_ns. We will always assume that $N > K + 1$ and that $\mathbf{X}$ is a full rank matrix. Our goal is to predict the label of a new instance $\mathbf{x}_{N+1}$. As in the previous section, the instances $\mathbf{x}_n$ are regarded as fixed (in the next section they will be generated stochastically, but our conclusions will be still valid, since our analysis can be applied conditionally on knowing the instances: cf. Remark 1 above).

Fix a significance level $\epsilon \in (0, 1)$. An *interval predictor* is a pair of measurable functions $L : \mathbb{R}^N \to \mathbb{R}$ and $U : \mathbb{R}^N \to \mathbb{R}$ such that $L \le U$. Another representation of the interval predictor is as the function $\Gamma(\mathbf{y}) := [L(\mathbf{y}), U(\mathbf{y})]$ mapping the labels to the corresponding prediction interval. The interval predictor is *unconditionally*

valid (for a given statistical model) if its coverage probability is $1-\epsilon$: $\mathbb{P}(\mathrm{err}) = \epsilon$ under each probability measure in the given model, where $\mathrm{err} = \mathrm{err}^\Gamma(Y_1, Y_2, \ldots)$ is the event $\{Y_{N+1} \notin \Gamma(Y_1, \ldots, Y_N)\}$ (sometimes called an *error*). It is $\mathcal{K}_N$*-valid* if $\mathbb{P}(\mathrm{err} \mid \mathcal{K}_N) = \epsilon$ a.s.

There exist many interval predictors Γ such that

$$G(\mathrm{err} \mid \mathcal{F}_N) = \epsilon \quad \text{a.s.} \tag{3}$$

(assuming that the conditional distribution of Y_{N+1} given $\mathcal{F}_N$ with respect to the Gosset measure is continuous; this assumption is satisfied for the noise distributions and equivariant estimator $(\tilde\beta, \tilde\sigma)$ used in our empirical studies in Section 4).

Given an interval predictor Γ for the Gosset measure G, we can define an interval predictor Γ' for the original linear regression model $(P_{\beta,\sigma})$ as

$$\Gamma'(y_1, \ldots, y_N) := \\ \tilde\sigma_y \Gamma\left(\frac{y_1 - \tilde\beta_y' \mathbf{x}_1}{\tilde\sigma_y}, \ldots, \frac{y_N - \tilde\beta_y' \mathbf{x}_N}{\tilde\sigma_y}\right) + \tilde\beta_y' \mathbf{x}_{N+1}, \tag{4}$$

where we again use the notation $\tilde\beta_y := \tilde\beta(y_1, \ldots, y_N)$ and $\tilde\sigma_y := \tilde\sigma(y_1, \ldots, y_N)$. In words, to obtain $\Gamma'(y_1, \ldots, y_N)$ we first normalize $(y_1, \ldots, y_N)$, then apply Γ to obtain a prediction interval, and finally apply the inverse transformation to that prediction interval. Notice that the interval predictor Γ' is *equivariant* in the sense that $\{Y_{N+1} \in \Gamma'(Y_1, \ldots, Y_N)\} \in \mathcal{K}$.

Proposition 1. *If Γ is an interval predictor satisfying (3), the interval predictor Γ' defined by (4) is $\mathcal{K}_N$-valid.*

Proof. Using Corollary 2 and Equations (3) and (4), we obtain:

$$\begin{aligned}
&P_{\beta,\sigma}(\mathrm{err}^{\Gamma'} \mid \mathcal{K}_N) \\
&= P_{\beta,\sigma}(Y_{N+1} \notin \Gamma'(Y_1, \ldots, Y_N) \mid \mathcal{K}_N) \\
&= P_{\beta,\sigma}\left(\frac{Y_{N+1} - \tilde\beta_y' \mathbf{x}_{N+1}}{\tilde\sigma_y} \notin \right. \\
&\qquad \left. \Gamma\left(\frac{Y_1 - \tilde\beta_y' \mathbf{x}_1}{\tilde\sigma_y}, \ldots, \frac{Y_N - \tilde\beta_y' \mathbf{x}_N}{\tilde\sigma_y}\right) \mid \mathcal{K}_N\right) \\
&= P_{\beta,\sigma}(\mathrm{err}^\Gamma(\nu_1, \nu_2, \ldots) \mid \mathcal{K}_N) \\
&= P_{\beta,\sigma}(\mathrm{err}^\Gamma(\nu_1, \nu_2, \ldots) \mid \nu_1, \ldots, \nu_N) \\
&= G(\mathrm{err}^\Gamma(Y_1, Y_2, \ldots) \mid Y_1, \ldots, Y_N) \\
&= G(\mathrm{err}^\Gamma \mid \mathcal{F}_N) = \epsilon \quad \text{a.s.}
\end{aligned}$$

□

Now we can define our interval predictor. Among the interval predictors Γ satisfying (3) we choose the *symmetric* one, i.e., the interval predictor $\Gamma = [L, U]$ such that $G(Y_{N+1} < L \mid \mathcal{F}_N) = G(Y_{N+1} > U \mid \mathcal{F}_N) = \epsilon/2$ a.s. (Such an interval predictor is essentially unique.) The predictor Γ' defined by (4) will be called the *symmetric pivotal interval predictor* (abbreviated to *SPIP*). Proposition 1 says that it is $\mathcal{K}_N$-valid (and so, in particular, unconditionally valid).

MCMC implementation of the SPIP

Suppose $(Y_1, Y_2, \ldots)$ are distributed according to $P_{0,1} = P^\infty$. Let $(Z_1, Z_2, \ldots) := \nu(Y_1, Y_2, \ldots)$, and let B and Σ be the random vector $\tilde\beta(Y_1, \ldots, Y_N)$ and random variable $\tilde\sigma(Y_1, \ldots, Y_N)$, respectively. If A and B are random elements, we let $f_{A|B}(a \mid b)$ stand for the conditional density of A at point a given $B = b$. We will also use similar notation for unconditional distributions: $f_A(a)$ stands for the density of A at a. We will be interested in regular (in particular, continuous) versions of conditional distributions, and so will omit the qualification "a.s."

In this subsection we choose a specific equivariant estimator $(\tilde\beta, \tilde\sigma)$; as we explain in the following subsection, we will obtain the same SPIP for any other equivariant estimator. As $\tilde\beta$ we take the least squares estimate $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ of β, and as $\tilde\sigma$ we take $\|\mathbf{y} - \mathbf{X}\tilde\beta\|/\sqrt{N}$, one of the standard estimates of σ. These are the maximum likelihood estimates when the noise distribution is Gaussian, but we will use them for other noise distributions as well. (The other standard estimate of σ, $\|\mathbf{y} - \mathbf{X}\tilde\beta\|/\sqrt{N-K}$, is the maximum likelihood estimate of σ based on the residuals $\mathbf{y} - \mathbf{X}\tilde\beta$. We could also use it as $\tilde\sigma$: the divisor is irrelevant for our construction.)

The following lemma (generalizing Fisher's [4] Equation (4), corresponding to $K = 1$ and $x_1 = x_2 = \cdots = 1$) will be the main ingredient of our prediction algorithm.

Lemma 3. *Suppose $\beta = 0$ and $\sigma = 1$. The conditional density of (B, Σ) given $Z_1 = z_1, \ldots, Z_N = z_N$ is proportional to $\tilde\sigma^{N-K-1} p(\tilde\beta'\mathbf{x}_1 + \tilde\sigma z_1) \cdots p(\tilde\beta'\mathbf{x}_N + \tilde\sigma z_N)$ (with the coefficient of proportionality a function of $z_1, \ldots, z_N$).*

Proof. To each $\mathbf{y} \in \mathbb{R}^N$ there corresponds a vector of residuals $\mathbf{r} := \mathbf{y} - \mathbf{X}\tilde\beta$, where $\tilde\beta$ is chosen to minimize the length of $\mathbf{r}$. The mapping $\mathbf{y} \mapsto \mathbf{r}$ is the projection onto the $(N-K)$-dimensional subspace S_1 of residuals, defined as the orthogonal complement of the column space of the matrix $\mathbf{X}$. Dividing $\mathbf{r}$ by $\tilde\sigma$ is essentially the same as the transformation $\mathbf{r} \mapsto \mathbf{r}/|\mathbf{r}|$ for $\mathbf{r} \neq 0$. It further transforms the residuals $\mathbf{r}$ into an element $\mathbf{z} = (z_1, \ldots, z_N)$ of a sphere S_2 in S_1 of a fixed radius, and the mapping $\mathbf{r} \mapsto \mathbf{z}$ is the projection of S_1 onto S_2 (each point $\mathbf{r}$ of S_1 different from the origin is mapped to the intersection of S_2 and the ray emanating from the origin in the direction of $\mathbf{r}$). The topological dimension of S_2 is $N - K - 1$, one fewer than that of S_1.

The underlying measure on the $(N-K-1)$-dimensional sphere S_2 is the standard surface measure. The underly-

ing measure on $(\tilde\beta, \tilde\sigma)$ is the restriction of the $(K+1)$-dimensional Lebesgue measure to $\mathbb{R}^K \times (0,\infty)$. Finally, the underlying measure on $(y_1, \dots, y_N)$ is the N-dimensional Lebesgue measure.

Since

$$f_{\mathrm{B},\Sigma|Z_1,\dots,Z_N}(\tilde\beta, \tilde\sigma \mid z_1, \dots, z_N) = \frac{f_{\mathrm{B},\Sigma,Z_1,\dots,Z_N}(\tilde\beta, \tilde\sigma, z_1, \dots, z_N)}{f_{Z_1,\dots,Z_N}(z_1, \dots, z_N)}$$

([9], Proposition B.45) and the denominator is independent of $(\tilde\beta, \tilde\sigma)$, we only need to prove that the density $f_{\mathrm{B},\Sigma,Z_1,\dots,Z_N}(\tilde\beta, \tilde\sigma, z_1, \dots, z_N)$ is proportional to

$$\tilde\sigma^{N-K-1} f_{Y_1,\dots,Y_N}(\tilde\beta'\mathbf{x}_1 + \tilde\sigma z_1, \dots, \tilde\beta'\mathbf{x}_N + \tilde\sigma z_N) = \tilde\sigma^{N-K-1} f_{Y_1,\dots,Y_N}(y_1, \dots, y_N),$$

where $(\tilde\beta, \tilde\sigma, z_1, \dots, z_N)$ are connected with $(y_1, \dots, y_N)$ as described above.

In other words, we are required to check that the Jacobian of the mapping

$$(\tilde\beta, \tilde\sigma, z_1, \dots, z_N) \mapsto (y_1, \dots, y_N)$$

is proportional to $\tilde\sigma^{N-K-1}$. (This mapping is somewhat unusual in that it is defined on the half-cylinder $\mathbb{R}^K \times (0,\infty) \times S_2$, but the notion of Jacobian for it is similar to the standard one: it is the factor by which the mapping expands a small volume in its domain.) The mapping from $(\tilde\beta, \tilde\sigma, z_1, \dots, z_N)$ to $(\tilde\beta, r_1, \dots, r_N)$ has Jacobian proportional to the surface area of a sphere of radius $\tilde\sigma$, i.e., to $\tilde\sigma^{N-K-1}$. The linear mapping from $(\tilde\beta, \mathbf{r})$ to $\mathbf{y} = \mathbf{X}\tilde\beta + \mathbf{r}$ has a constant Jacobian that is different from zero since the mapping is bijective and both $(\tilde\beta, \mathbf{r}) \in \mathbb{R}^K \times S_1$ and $\mathbf{y} \in \mathbb{R}^N$ range over spaces of dimension N. □

Our prediction algorithm, given as Algorithm 1, uses MCMC sampling (see, e.g., [3]) from the conditional distribution of $(\tilde\beta, \tilde\sigma)$ given $\nu(y_1, \dots, y_N)$. Its inputs are the training set $\mathbf{x}_1, y_1, \dots, \mathbf{x}_N, y_N$ and a new instance $\mathbf{x}_{N+1}$. The duration of the burn-in period is B, and $B+M$ is the overall duration of the random walk. Let s_β and s_σ be small positive constants (the standard deviations of the Gaussian proposal distributions for B and $\log\Sigma$, respectively) and I be the identity matrix (whose size will be clear from the context). Fix a significance level $\epsilon \in (0, 1/2)$; for simplicity we assume that $M\epsilon/2$ is an integer.

The correctness of the algorithm (to any accuracy with probability approaching one as $M \to \infty$) follows from the facts that the stationary distribution of the ergodic Markov chain (β_m, σ_m), $m = 0, 1, \dots$, is $f_{\mathrm{B},\Sigma|Z_1,\dots,Z_N}(\cdot,\cdot \mid z_1, \dots, z_N)$ and that

$$f_{Z_{N+1}|Z_1,\dots,Z_N}(z_{n+1} \mid z_1, \dots, z_N)$$

Algorithm 1 MCMC SPIP based on noise distribution with density p

Compute $\tilde\beta$, $\tilde\sigma$ from the training set;
for $n = 1, \dots, N$ **do**
  $z_n := (y_n - \tilde\beta'\mathbf{x}_n)/\tilde\sigma$;
end for
set $\beta_0 := 0$, $\sigma_0 := 1$, and $p_0 := p(z_1) \cdots p(z_N)$;
for $m = 1, \dots, B+M$ **do**
  set $\beta_m := \beta_{m-1} + N(0, s_\beta^2 I)$ and $\log\sigma_m := \log\sigma_{m-1} + N(0, s_\sigma^2)$;
  set $p_m := \sigma_m^{N-K-1} \times p(\beta_m'\mathbf{x}_1 + \sigma_m z_1) \cdots p(\beta_m'\mathbf{x}_N + \sigma_m z_N)$;
  if $p_m < p_{m-1}$ **then**
    with probability $1 - p_m/p_{m-1}$,
    redefine $\beta_m := \beta_{m-1}$, $\sigma_m := \sigma_{m-1}$, and $p_m := p_{m-1}$;
  end if
  sample ξ_m from p and set $\zeta_m := (\xi_m - \beta_m'\mathbf{x}_{N+1})/\sigma_m$;
end for
order $\zeta_{B+1}, \dots, \zeta_{B+M}$ into an increasing sequence $\zeta_{(1)}, \dots, \zeta_{(M)}$;
output the prediction interval $[\tilde\beta'\mathbf{x}_{N+1} + \tilde\sigma\zeta_{(M\epsilon/2)}, \tilde\beta'\mathbf{x}_{N+1} + \tilde\sigma\zeta_{(M(1-\epsilon/2))}]$.

$$\begin{aligned}
&= \int_{\mathbb{R}^K\times(0,\infty)} f_{Z_{N+1}|\mathrm{B},\Sigma,Z_1,\dots,Z_N}(z_{N+1} \mid \tilde\beta, \tilde\sigma, z_1, \dots, z_N) \\
&\quad \times f_{\mathrm{B},\Sigma|Z_1,\dots,Z_N}(\tilde\beta, \tilde\sigma \mid z_1, \dots, z_N) d(\tilde\beta, \tilde\sigma) \\
&= \int_{\mathbb{R}^K\times(0,\infty)} f_{Z_{N+1}|\mathrm{B},\Sigma}(z_{N+1} \mid \tilde\beta, \tilde\sigma) \\
&\quad \times f_{\mathrm{B},\Sigma|Z_1,\dots,Z_N}(\tilde\beta, \tilde\sigma \mid z_1, \dots, z_N) d(\tilde\beta, \tilde\sigma) \qquad (5)
\end{aligned}$$

(the second equality follows from the conditional independence of Z_{N+1} and $(Z_1, \dots, Z_N)$ given $(\tilde\beta, \tilde\sigma)$, which in turn follows from the conditional independence of Y_{N+1} and $(Y_1, \dots, Y_N)$ given $(\tilde\beta, \tilde\sigma)$).

In our empirical studies reported in the next section we use $B = M = 100{,}000$ (although already $B = M = 2000$ would have led to very similar results) and $s_\beta = s_\sigma = 0.1$.

The role of the estimator

As defined above, the SPIP potentially depends not only on the noise distribution P but also on the equivariant estimator $(\tilde\beta, \tilde\sigma)$. The dependence on $(\tilde\beta, \tilde\sigma)$ is worrying because it might lead to an inefficient interval predictor when $(\tilde\beta, \tilde\sigma)$ is not an efficient estimator of (β, σ) under P; we have such a situation when Algorithm 1 is applied to a non-Gaussian noise distribution. In this subsection we will see that in fact the SPIP does not depend on the choice of $(\tilde\beta, \tilde\sigma)$.

Let $(\tilde\beta, \tilde\sigma)$ be any equivariant estimator (not necessarily the one defined in the previous subsection). The im-

age $\nu(\mathbb{R}^\infty)$ of the normalizing transformation ν coincides with the set of all sequences $(y_1, y_2, \ldots) \in \mathbb{R}^\infty$ for which $\tilde{\beta}(y_1, \ldots, y_N) = 0$ and $\tilde{\sigma}(y_1, \ldots, y_N) = 1$. We will call $\nu(\mathbb{R}^\infty)$ the *Gosset space* for $(\tilde{\beta}, \tilde{\sigma})$, and we will write $\tilde{\nu}$ for ν to indicate the dependence on $(\tilde{\beta}, \tilde{\sigma})$. The Gosset space determines $\tilde{\nu}$ (indeed, $\tilde{\nu}$ maps each element of $\mathbb{R}^\infty$ to the unique element of its orbit that belongs to the Gosset space) and therefore determines the Gosset measure $P_{0,1}\tilde{\nu}^{-1}$. The Gosset measure is concentrated on the Gosset space, and we will sometimes regard the Gosset space as the measure space equipped with the Gosset measure (more accurately, with the restriction of the Gosset measure to the Gosset space).

It is easy to see that all Gosset measures coincide on $\mathcal{K}$ with $P_{\beta,\sigma}$ and, therefore, with each other. What is even more important for us, we will see that all Gosset spaces are isomorphic as measure spaces.

Let $(\hat{\beta}, \hat{\sigma})$ be another equivariant estimator. The corresponding normalizing transformation will be denoted $\hat{\nu}$. By Lemma 2, each normalizing transformation collapses each orbit into one point, and so both $\tilde{\nu}$ and $\hat{\nu}$ are projections: $\tilde{\nu}\tilde{\nu} = \tilde{\nu}$ and $\hat{\nu}\hat{\nu} = \hat{\nu}$. Moreover, these projections are such that $\hat{\nu}\tilde{\nu} = \hat{\nu}$ and $\tilde{\nu}\hat{\nu} = \tilde{\nu}$. We can see that the restriction $\hat{\nu}|_{\tilde{\nu}(\mathbb{R}^\infty)}$ of $\hat{\nu}$ to the Gosset space $\tilde{\nu}(\mathbb{R}^\infty)$ is a bijection between the Gosset spaces $\tilde{\nu}(\mathbb{R}^\infty)$ and $\hat{\nu}(\mathbb{R}^\infty)$, with the inverse mapping $\tilde{\nu}|_{\hat{\nu}(\mathbb{R}^\infty)}$. The corresponding Gosset measures $\tilde{G}$ and $\hat{G}$ are related by $\tilde{G} = \hat{G}\tilde{\nu}^{-1}$ and $\hat{G} = \tilde{G}\hat{\nu}^{-1}$. In other words, the Gosset spaces $\tilde{\nu}(\mathbb{R}^\infty)$ and $\hat{\nu}(\mathbb{R}^\infty)$ are isomorphic as measure spaces with isomorphism $\iota := \hat{\nu}|_{\tilde{\nu}(\mathbb{R}^\infty)}$.

Let $\Gamma = \tilde{\Gamma}$ be a symmetric interval predictor satisfying (3) (but remember that now our notation for G is $\tilde{G}$). We are only interested in the values $\tilde{\Gamma}(y_1, \ldots, y_N)$ for $(y_1, y_2, \ldots)$ in the Gosset space $\tilde{\nu}(\mathbb{R}^\infty)$. Let $\hat{\Gamma}$ be the interval predictor corresponding to $\tilde{\Gamma}$ under the isomorphism ι; in other words, $\hat{\Gamma}$ is defined by the right-hand side of (4) for $(y_1, y_2, \ldots)$ in the Gosset space $\hat{\nu}(\mathbb{R}^\infty)$. Because of the isomorphism, $\hat{\Gamma}$ will be a symmetric interval predictor satisfying (3) with G replaced by $\hat{G}$ and err standing for $\mathrm{err}^{\hat{\Gamma}}$. It is clear that we will obtain the same prediction interval $\Gamma'(y_1, \ldots, y_N)$ regardless of whether we apply the recipe (4) to $\tilde{\Gamma}$ and $(\tilde{\beta}, \tilde{\sigma})$ or to $\hat{\Gamma}$ and $(\hat{\beta}, \hat{\sigma})$. In other words, we will obtain the same SPIP from $(\tilde{\beta}, \tilde{\sigma})$ and $(\hat{\beta}, \hat{\sigma})$.

4. Empirical studies

In our empirical studies we apply Algorithm 1 to the following noise distributions: the Gaussian distribution with density $p(y) \propto e^{-y^2/2}$, the Laplace distribution with density $p(y) \propto e^{-|y|}$, and Student's t-distribution on 4 degrees of freedom with density $p(y) \propto (1 + y^2)^{-5/2}$ (up to scaling).

The Gaussian distribution is standard. The prediction intervals produced by the SPIP are identical to the classical prediction intervals (defined in, e.g., [10], Section 5.3.1): indeed, by (5) and Basu's theorem, the conditional density $f_{Z_{N+1}|Z_1,\ldots,Z_N}(\cdot \mid z_1, \ldots, z_N)$ does not depend on $z_1, \ldots, z_N$ and so is identical to the unconditional density $f_{Z_{N+1}}$. The Laplace distribution is used in robust linear regression (see, e.g., [10], Sections 3.13.1 and 11.12.1, or [1]).

Lange et al. [6] suggest using t-distributions in robust linear regression; they report ([6], p. 883) that the t-distribution on 4 degrees of freedom worked well in many of their applications. However, our emphasis will be on the more standard Gaussian and Laplace distributions.

First we investigate the behaviour of our prediction intervals on artificially generated data sets. Figure 1 shows four calibration plots, each of which is a graph of the error rate against the significance level for data generated by p_1 and the algorithm using p_2 (where p_1 and p_2 are to be defined later). For each plot we generated 5,500 examples $(\mathbf{x}_n, y_n)$ from the model (1) with $K = 1$, $\beta = \sigma = 2$, $\mathbf{x}_n$ generated independently from the uniform distribution on $[0, 1]$, and ξ_n generated independently (among themselves and $\mathbf{x}_1, \ldots, \mathbf{x}_{5500}$) from a distribution with density p_1. The first $N = 500$ examples were used as the training set and the remaining 5,000 examples as the test set. We ran Algorithm 1 (with $N + k$ in place of $N + 1$) based on a noise distribution with density p_2 to predict the label y_{N+k} of each test instance $\mathbf{x}_{N+k}$, $k = 1, \ldots, 5000$, for a fine grid of significance levels $\epsilon \in (0, 0.2]$. Each of the four plots shows the percentage of the test examples $(\mathbf{x}_{N+k}, y_{N+k})$, $k = 1, \ldots, 5000$, for which y_{N+k} was not covered by the prediction interval produced for $\mathbf{x}_{N+k}$ by Algorithm 1 as function of ϵ. We call them calibration plots since the function being close to the bisector of the first quadrant means that the prediction intervals are well calibrated: the prediction algorithm's frequency of error is close to the nominal significance level. We concentrate on the most interesting range of small ϵ, which includes, in particular, the standard values of 5% and 1%. Since the results of our experiments are random, each calibration plot is shown for five different initial states of the MATLAB random number generator.

The top left plot has both p_1 and p_2 equal to the Gaussian distribution, and the bottom right plot has both p_1 and p_2 equal to the Laplace distribution. These two plots demonstrate empirically the validity of our prediction algorithm: when it is provided with the correct model, its predictions are well calibrated. The top right plot describes an application of a robust prediction algorithm (based on the Laplace distribution) to benign (Gaussian) data. The algorithm is rather conservative: at significance level 5% it typically makes between 1% and 2% of errors, while at 1% the percentage of errors is typically below 0.05%. The bottom left plot describes an application of an optimistic prediction al-

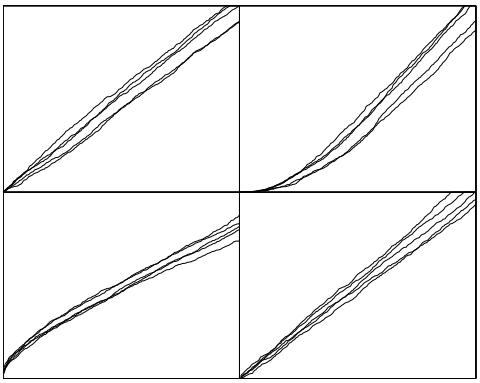

Figure 1. Calibration plots for the Gaussian prediction intervals on Gaussian (top left) and Laplace (bottom left) data and for the Laplace prediction intervals on Gaussian (top right) and Laplace (bottom right) data. The horizontal axis is $\epsilon \in (0, 0.2]$, and the range of the vertical axis is also $[0, 0.2]$.

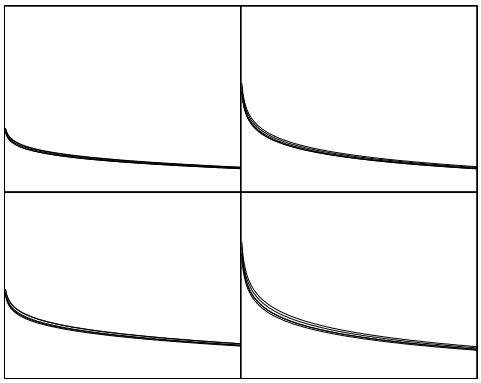

Figure 2. The median widths of prediction intervals for various $\epsilon \in (0, 0.2]$, with the same layout as Figure 1. The range of the vertical axis is always $[0, 40]$.

gorithm (based on the Gaussian distribution) to somewhat unruly (Laplace) data. For interesting values of the significance level, the predictions are not well calibrated: at significance level 5%, the percentage of wrong predictions is around 6–7$\%$, and at 1% it is around 2–3$\%$.

In Figure 2 we give the median widths of the prediction intervals at significance levels $\epsilon \in (0, 0.2]$, with p_1 being the Gaussian distribution for the two top plots and the Laplace distribution for the two bottom plots, and with p_2 being the Gaussian distribution for the two left-hand plots and the Laplace distribution for the two right-hand plots.

We know that the unconditional, and conditional on $\mathcal{K}_N$, coverage probability of our prediction intervals is equal to the confidence level $1 - \epsilon$; this is illustrated by the top left and bottom right plots of Figure 1. (It should be remembered that there is always implicit conditioning on the observed instances: cf. Remark 1.) An interesting question is how stable the fully conditional, i.e., conditional on $\mathcal{F}_N$, coverage probabilities are. The results for our experimental setup are shown in Figure 3. We generated five training sets, each of size 500; box plots 1 to 5 describe the results for the first training set, 6 to 10 for the second training set, etc. For each training set we generated five test sets of size 5,000 following the same distribution. For each of the test examples $(\mathbf{x}, y)$ we computed the fully conditional coverage probability of the corresponding prediction interval (computed from the instance $\mathbf{x}$ and the corresponding training set, with the label y ignored). Box plot 1 gives some statistics for the first test set generated for the first training set, box plot 2 gives statistics for the second test set generated for the first training set, etc. Namely, each box plot gives the median coverage probability, the quartile coverage probabilities, and the maximum and minimum coverage probabilities for the prediction intervals generated for the test instances. We can see that for the same training set the box plots are very similar (because of the large size of the test sets), but the variation of coverage probabilities between the training sets is significant.

In the next section we will see the strength of the guarantee $\mathbb{P}(\mathrm{err} \mid \mathcal{K}_N) = \epsilon$ compared with the usual Neyman-type guarantee $\mathbb{P}(\mathrm{err}) = \epsilon$. On the other hand, Figure 3 shows the weakness of the guarantee $\mathbb{P}(\mathrm{err} \mid \mathcal{K}_N) = \epsilon$ compared with the ideal (but impossible) guarantee $\mathbb{P}(\mathrm{err} \mid \mathcal{F}_N) = \epsilon$: it is not uncommon for the probability $\mathbb{P}(\mathrm{err} \mid \mathcal{F}_N)$ to be as high as 5.5% for our prediction intervals when the nominal significance level is 5%.

We have also applied three kinds of prediction intervals to the ChickWeight data set ([2], Example 5.3; [5], Table A.2; part of the standard R distribution, package `datasets`). The data set gives weight versus age of chicks on different diets. The body weights of the chicks were measured at birth, every second day thereafter until day 20, and on day 21; some measurements (22 in total) are missing. The range of the body weights is 35 to 373 grams. There are four groups of chicks on different protein diets. Our task was to predict a chick's weight given its age. We used the chicks on diets 1 and 2 as the training set (of size 340) and the chicks on diets 3 and 4 as the test set (of size 238).

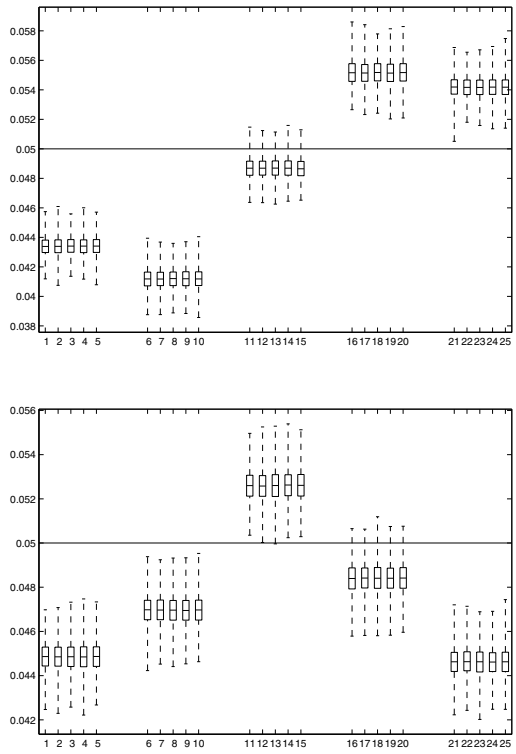

Figure 3. The fully conditional coverage probabilities of Gaussian (top) and Laplace (bottom) prediction intervals for $\epsilon = 5\%$.

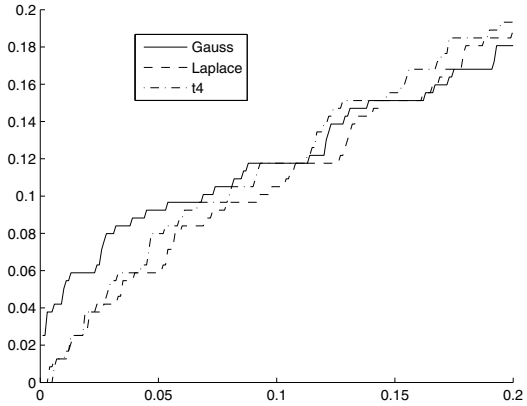

Figure 4. The calibration plots for the Chick-Weight data set, with $\epsilon \in (0, 0.2]$.

It is clear that all three models that we have discussed are wrong for this data set, for a multitude of reasons, and our question is which is more useful for predicting the weights of the chicks in the test set. Figure 4 shows that the Laplace prediction intervals (i.e., those produced by Algorithm 1 based on the Laplace distribution) are fairly well calibrated over our range of significance levels, and that the t4 prediction intervals (i.e., those produced by Algorithm 1 based on the t distribution on 4 degrees of freedom) are not very different. Figure 5 gives the median widths of the prediction intervals at significance levels $\epsilon \in (0, 0.2]$, as in Figure 2.

In general, we have found that the best calibration was usually achieved by the Gaussian prediction intervals (with the Laplace and t4 ones somewhat conservative) or by the Laplace and t4 prediction intervals (with the Gaussian ones somewhat miscalibrated). The performance of Laplace and t4 prediction intervals was broadly similar, despite the different nature of the tails of the corresponding noise distributions (decaying exponentially fast in the case of Laplace and polynomially fast in the case of t4).

To be on the safe side, our recommendation would be to use the Laplace or t4 prediction intervals when in doubt. Alternatively, a safe prediction algorithm could output the union of the Gaussian, Laplace, and t4 prediction intervals (as suggested by Morgenthaler [8], Section 3, in the context of confidence estimation; later in the paper Morgenthaler considers potentially more efficient prediction algorithms sacrificing the conditional validity).

5. Implications for on-line prediction

In Sections 2–4 we discussed using our prediction algorithm in the batch mode, when test labels are predicted repeatedly using a fixed training set. In the on-line mode, we start from the empty training set and add each new example $(\mathbf{x}_n, y_n)$, $n = 1, 2, \ldots$, to the training set after predicting its label y_n. An important class of prediction intervals are *strong prediction intervals*, which satisfy $\mathbb{P}(\mathrm{err}_N \mid \mathrm{err}_1, \ldots, \mathrm{err}_{N-1}) = \epsilon$ a.s. for each $N = 1, 2, \ldots$, where err_n is the event of an error being made at the nth step (in other words, which make errors independently at different steps). The random binary sequence $\mathbb{I}_{\mathrm{err}_1}, \mathbb{I}_{\mathrm{err}_2}, \ldots$ of indicators of errors for strong prediction intervals has the same distribution as for fully conditional prediction intervals (i.e., those satisfying $\mathbb{P}(\mathrm{err}_N \mid \mathcal{F}_{N-1}) = \epsilon$ a.s. for all N). In this sense strong prediction intervals are perfectly calibrated. The prediction intervals constructed in this paper can be applied in the on-line mode, and they are automatically strong (cf. [11], Proposition 2): indeed, they are equivariant (in particular, err_n are $\mathcal{K}_n$-measurable) and $\mathbb{P}(\mathrm{err}_N \mid \mathcal{K}_{N-1}) = \epsilon$ a.s. for each N. Notice that the Neyman-type guarantee $\mathbb{P}(\mathrm{err}_N) = \epsilon$ would not have been sufficient for this conclusion. It can be shown ([12], Chapter 4) that strong prediction intervals have certain properties

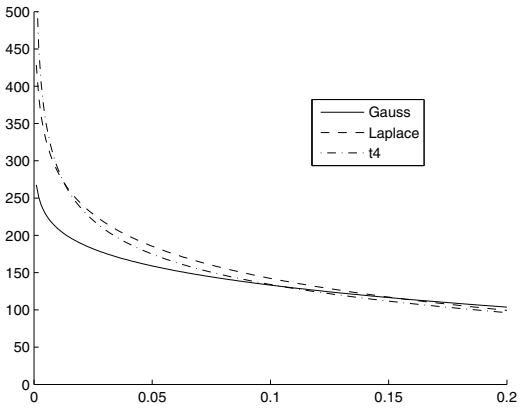

Figure 5. Median widths of the prediction intervals for the ChickWeight data set.

of validity in the batch and "semi-on-line" (intermediate between batch and on-line) modes, and we saw in Section 4 that they are empirically approximately valid in the batch mode.

Acknowledgements

This work was supported in part by the NSF grant DMS 0906592, EPSRC grant EP/F002998/1, MRC grant G0802594, EU FP7 grant 201381, Cyprus Research Promotion Foundation, and VLA. In our empirical studies we used MATLAB® and the R system.

References

[1] P. Bloomfield and W. L. Steiger. *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhäuser, Boston, 1983.

[2] M. J. Crowder and D. J. Hand. *Analysis of Repeated Measures*. Chapman and Hall, London, 1990.

[3] P. Diaconis. The Markov chain Monte Carlo revolution. *Bulletin (New Series) of the American Mathematical Society*, 46:179–205, 2008.

[4] R. A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society A*, 144:285–307, 1934.

[5] D. J. Hand and M. J. Crowder. *Practical Longitudinal Data Analysis*. Chapman and Hall, London, 1996.

[6] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modelling using the t distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

[7] J. F. Lawless. On the estimation of safe life when the underlying life distribution is Weibull. *Technometrics*, 15:857–865, 1973.

[8] S. Morgenthaler. Robust confidence intervals for a location parameter: The configural approach. *Journal of the American Statistical Association*, 81:518–525, 1986.

[9] M. J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

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

[11] F. Seillier-Moiseiwitsch. Sequential probability forecasts and the probability integral transform. *International Statistical Review*, 61:395–408, 1993.

[12] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

A An impossibility result

In Section 1 we said that no interval predictor can be $\mathcal{F}_N$-valid in the models considered in this paper. Now we prove this simple claim.

Proposition 2. *Suppose p is bounded above and $\epsilon < 1$. No prediction interval satisfies $P_{\beta,\sigma}(\mathrm{err} \mid \mathcal{F}_N) \le \epsilon$ a.s. for all β and σ simultaneously.*

Proof. If we consider a fixed probability measure $P_{\beta,\sigma}$ from our model, the conditional distribution of Y_{N+1} given the training set coincides with the distribution of $\beta' \mathbf{x}_{N+1} + \sigma\xi$, $\xi \sim P$, and does not depend on the training set. Set $c := \sup_y p(y)$. The density of the conditional distribution of Y_{N+1} is bounded above by c/σ. This means that a prediction interval satisfying $\mathbb{P}(\mathrm{err} \mid \mathcal{F}_N) \le \epsilon$ must have length at least $(1-\epsilon)\sigma/c$. As σ is unbounded above, the prediction interval has to be of infinite length in order to be suitable for all σ simultaneously. □