Chapter 11
Kernel Ridge Regression

Vladimir Vovk

Abstract This chapter discusses the method of Kernel Ridge Regression, which is a very simple special case of Support Vector Regression. The main formula of the method is identical to a formula in Bayesian statistics, but Kernel Ridge Regression has performance guarantees that have nothing to do with Bayesian assumptions. I will discuss two kinds of such performance guarantees: those not requiring any assumptions whatsoever, and those depending on the assumption of randomness.

11.1 Introduction

This chapter is based on my talk at the Empirical Inference Symposium (see p. x). It describes some developments influenced by Vladimir Vapnik, which are related to, but much less well known than, the Support Vector Machine. The Support Vector Machine is a powerful combination of the idea of generalized portrait (1962; see Chap. 3) and the kernel methods, and from the very beginning the performance guarantees for it were non-Bayesian, depending only on the *assumption of randomness*: the data are generated independently from the same probability distribution. An example of such a performance guarantee is (3.2); numerous other examples are given in Vapnik [15]. Kernel Ridge Regression (KRR) is a special case of Support Vector Regression, which has been known in Bayesian statistics for a long time. However, the advent of the Support Vector Machine encouraged non-Bayesian analyses of KRR, and this chapter presents two examples of such analyses. The first example is in the tradition of prediction with expert advice

V. Vovk (✉)
Dept. of Computer Science, Royal Holloway, University of London, Egham, Surrey, United Kingdom
e-mail: v.vovk@rhul.ac.uk

B. Schölkopf et al. (eds.), *Empirical Inference*, DOI 10.1007/978-3-642-41136-6_11,

[3] and involves no statistical assumptions whatsoever. The second example belongs to the area of conformal prediction [17] and only depends on the assumption of randomness.

11.2 Kernel Ridge Regression

It appears that the term "Kernel Ridge Regression" was coined in 2000 by Cristianini and Shawe-Taylor [5] to refer to a simplified version of Support Vector Regression; this was an adaptation of the earlier "ridge regression in dual variables" [12]. Take the usual Support Vector Regression in primal variables

$$\begin{array}{ll} \text{minimize} & \|w\|^2 + C \sum_{t=1}^{T} \left((\xi_t)^k + (\xi_t')^k \right) \\ \text{subject to} & (w \cdot x_t + b) - y_t \le \epsilon + \xi_t, \quad t = 1, \dots, T, \\ & y_t - (w \cdot x_t + b) \le \epsilon + \xi_t', \quad t = 1, \dots, T, \\ & \xi_t, \xi_t' \ge 0, \quad t = 1, \dots, T, \end{array}$$

where $(x_t, y_t) \in \mathbb{R}^n \times \mathbb{R}$ are the training examples, w is the weight vector, b is the bias term, ξ_t, ξ_t' are the slack variables, and T is the size of the training set; $\epsilon, C > 0$ and $k \in \{1, 2\}$ are the parameters. Simplify the problem by ignoring the bias term b (it can be partially recovered by adding a dummy attribute 1 to all x_t), setting $\epsilon := 0$, and setting $k := 2$. The optimization problem becomes

$$\text{minimize} \qquad a \|w\|^2 + \sum_{t=1}^{T} (y_t - w \cdot x_t)^2$$

(where $a := 1/C$), the usual Ridge Regression problem. And Vapnik's usual method ([15], Sect. 11.3.2) then gives the prediction

$$\hat{y} = \hat{w} \cdot x = Y'(K + aI)^{-1} k \tag{11.1}$$

for the label of a new object x, where Y is the vector of labels (with components $Y^t := y_t$), K is the Gram matrix $K_{s,t} := x_s \cdot x_t$, and k is the vector with components $k^t := x_t \cdot x$. The kernel trick replaces x_t by $F(x_t)$, and so K by the kernel matrix $K_{s,t} := \mathcal{K}(x_s, x_t)$ and k by the vector $k^t := \mathcal{K}(x_t, x)$, where $\mathcal{K}$ is the kernel $\mathcal{K}(x, x') := F(x) \cdot F(x')$.

This simple observation was made in [12], where this simplified SVR method was called "ridge regression in dual variables". There is no doubt that this calculation has been done earlier as well, but the result does not appear useful. First, compared to the "full" SVM, there is no sparsity of examples (and there is

no sparsity in attributes, as in the case of the Lasso). Having an explicit formula is an advantage, but the formula is not new: mathematically, the formula for KRR coincides with one of the formulas in kriging [4], an old method in geostatistics for predicting values of a Gaussian random field; this formula had been widely used in Bayesian statistics.

However, there is a philosophical and practical difference:

- In kriging, the kernel is estimated from the results of observations and in Bayesian statistics it is supposed to reflect the statistician's beliefs;
- In KRR, as in Support Vector Regression in general, the kernel is not supposed to reflect any knowledge or beliefs about reality, and the usual approach is pragmatic: one consults standard libraries of kernels and uses whatever works.

In the remaining sections of this chapter we will explore KRR in the SVM style, without making Bayesian assumptions. The practical side of this non-Bayesian aspect of KRR is that it often gives good results on real-world data, despite the Bayesian assumptions being manifestly wrong. We will, however, concentrate on its theoretical side: non-Bayesian performance guarantees for KRR.

An important special case of KRR is (ordinary) Ridge Regression (RR): it is a special case (as far as the output is concerned) of KRR for $\mathcal{K}$ as the dot product. However, in the case of RR the usual representation of the prediction is

$$\hat{y} = \hat{w} \cdot x = x'(X'X + aI)^{-1}X'Y \tag{11.2}$$

rather than (11.1), where X is the matrix whose rows are $x_1', \dots, x_T'$; there are many ways to show that (11.2) and (11.1) indeed coincide when $\mathcal{K}$ is the dot product.

Under a standard Bayesian assumption (which we do not state explicitly in general; see, e.g., [17], Sect. 10.3), the conditional distribution of the label y of a new example (x, y) given $x_1, \dots, x_T, x$ and $y_1, \dots, y_T$ is

$$N\left(Y'(K + aI)^{-1}k, \sigma^2 + \frac{\sigma^2}{a}\mathcal{K}(x, x) - \frac{\sigma^2}{a}k'(K + aI)^{-1}k\right), \tag{11.3}$$

where K and k are as before (the postulated probability distribution generating the examples depends on $\mathcal{K}$ and a, and we parameterize a normal probability distribution $N(\mu, \sigma^2)$ by its mean μ and standard deviation σ^2). The mean of the distribution (11.3) is the KRR prediction, but now we have not only a point prediction but also an estimate of its accuracy.

When $\mathcal{K}$ is the dot product, (11.3) can be rewritten as

$$N\left(x'(X'X + aI)^{-1}X'Y, \sigma^2 x'(X'X + aI)^{-1}x + \sigma^2\right). \tag{11.4}$$

In this case the Bayesian assumption can be stated as follows: $x_1, x_2, \dots$ are fixed vectors in $\mathbb{R}^n$ (alternatively, we can make our analysis conditional on their values) and

$$y_t = \theta \cdot x_t + \xi_t, \tag{11.5}$$

where $\theta \sim N(0, (\sigma^2/a)I)$ and $\xi_t \sim N(0, \sigma^2)$ are all independent.

Equations (11.3) and (11.4) give exhaustive information about the next observation; the Bayesian assumption, however, is rarely satisfied.

11.3 Kernel Ridge Regression Without Probability

It turns out that KRR has interesting performance guarantees even if we do not make any stochastic assumptions whatsoever. Due to lack of space no proofs will be given; they can be found in the technical report [21].

In this section we consider the following perfect-information protocol of on-line regression:

Protocol 1 On-line regression protocol

for $t := 1, 2, \ldots$ **do**
  Reality announces $x_t \in \mathbf{X}$
  Learner predicts $\hat{y}_t \in \mathbb{R}$
  Reality announces $y_t \in \mathbb{R}$
end for

First we consider the case where the space $\mathbf{X}$ from which the objects x_t are drawn is a Euclidean space, $\mathbf{X} := \mathbb{R}^n$, and our goal is to compete with linear functions; in this case ordinary Ridge Regression is a suitable strategy for Learner. Then we move on to the case of an arbitrary $\mathbf{X}$ and replace RR by KRR.

11.3.1 Ordinary Ridge Regression

In this section, $\mathbf{X} = \mathbb{R}^n$. The RR strategy for Learner is given by the formula $\hat{y}_t := b'_{t-1} A_{t-1}^{-1} x_t$, where $b_0, b_1, \ldots$ is the sequence of vectors and $A_0, A_1, \ldots$ is the sequence of matrices defined by

$$b_T := \sum_{t=1}^{T} y_t x_t, \qquad A_T := aI + \sum_{t=1}^{T} x_t x'_t$$

(cf. (11.2)), where $a > 0$ is a parameter. The incremental update of the matrix A_t^{-1} can be done effectively by the Sherman–Morrison formula. The following performance guarantee is proved in [21], Sect. 2.

Theorem 11.1. *The Ridge Regression strategy for Learner with parameter $a > 0$ satisfies, at any step T,*

$$\sum_{t=1}^{T} \frac{(y_t - \hat{y}_t)^2}{1 + x_t' A_{t-1}^{-1} x_t} = \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right). \tag{11.6}$$

The part $x_t' A_{t-1}^{-1} x_t$ in the denominator of (11.6) is usually close to 0 for large t.

Theorem 11.1 has been adapted to the Bayesian setting by Zhdanov and Kalnishkan [20], who also notice that it can be extracted from [1] (by summing their (4.21) in an exact rather than an estimated form).

Theorem 11.1 and its kernel version (Theorem 11.2 below) imply surprisingly many well-known inequalities.

Corollary 11.1. *Assume $|y_t| \le \mathbf{y}$ for all t, clip the predictions of the Ridge Regression strategy to $[-\mathbf{y}, \mathbf{y}]$, and denote them by $\hat{y}_t^{\mathbf{y}}$. Then*

$$\sum_{t=1}^{T} (y_t - \hat{y}_t^{\mathbf{y}})^2 \le \min_{\theta} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + 4\mathbf{y}^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right). \tag{11.7}$$

The bound (11.7) is exactly the bound obtained in [16] (Theorem 4) for the algorithm merging linear experts with predictions clipped to $[-\mathbf{y}, \mathbf{y}]$, which does not have a closed-form description and so is less interesting than clipped RR. The bound for the strategy called the AAR in [16] has $\mathbf{y}^2$ in place of $4\mathbf{y}^2$ ([16], Theorem 1). (The AAR is very similar to RR: its predictions are $b_{t-1}' A_t^{-1} x_t$ rather than $b_{t-1}' A_{t-1}^{-1} x_t$; it is called the Vovk–Azoury–Warmuth algorithm in [3].) The regret term in (11.7) has the logarithmic order in T if $\|x_t\|_\infty \le X$ for all t, because

$$\ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) \le n \ln \left(1 + \frac{TX^2}{a} \right) \tag{11.8}$$

(the determinant of a positive definite matrix is bounded by the product of its diagonal elements; see [2], Chap. 2, Theorem 7). From Theorem 11.1 we can also deduce Theorem 11.7 in [3], which is somewhat similar to Corollary 11.1. That theorem implies (11.7) when RR's predictions happen to be in $[-\mathbf{y}, \mathbf{y}]$ without clipping (but this is not what Corollary 11.1 asserts).

RR is not as good as the AAR in the setting where $\sup_t |y_t| \le \mathbf{y}$ and the goal is to obtain bounds of the form (11.7) (since the AAR is to some degree optimized for this setting), but is still very good; and we can achieve an interesting equality (rather than inequality) for it.

The upper bound (11.7) does not hold for the RR strategy if the coefficient 4 is replaced by any number less than $\frac{3}{2\ln 2} \approx 2.164$, as can be seen from an example given in Theorem 3 [16], where the left-hand side of (11.7) is $4T + o(T)$, the minimum on the right-hand side is at most T, $\mathbf{y} = 1$, and the logarithm is $2T \ln 2 + O(1)$. It is also known that there is no strategy achieving (11.7) with the coefficient less than 1 instead of 4, even in the case where $\|x_t\|_\infty \le X$ for all t: see Theorem 2 in [16].

There is also an upper bound on the cumulative square loss of the RR strategy without a logarithmic part and without assuming that the labels are bounded.

Corollary 11.2. *If $\|x_t\|_2 \le Z$ for all t then the Ridge Regression strategy for Learner with parameter $a > 0$ satisfies, at any step T,*

$$\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \le \left(1 + \frac{Z^2}{a}\right) \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right).$$

This bound is better than the bound in Corollary 3.1 of [8], which has an additional regret term of the logarithmic order in time.

Asymptotic properties of the RR strategy can be further studied using Corollary A.1 of Kumon et al. [9]. Kumon et al.'s result states that when $\|x_t\|_2 \le 1$ for all t, then $x_t' A_{t-1}^{-1} x_t \to 0$ as $t \to \infty$. It is clear that we can replace $\|x_t\|_2 \le 1$ for all t by $\sup_t \|x_t\|_2 < \infty$. This gives the following corollary, which can be summarized as follows. If there exists a very good expert (asymptotically), then RR also predicts very well. If there is no such very good expert, RR performs asymptotically as well as the best regularized expert.

Corollary 11.3. *Let $a > 0$ and $\hat{y}_t$ be the predictions output by the Ridge Regression strategy with parameter a. Suppose $\sup_t \|x_t\|_2 < \infty$. Then*

$$\left(\exists \theta \in \mathbb{R}^n : \sum_{t=1}^{\infty}(y_t - \theta' x_t)^2 < \infty\right) \Longrightarrow \sum_{t=1}^{\infty}(y_t - \hat{y}_t)^2 < \infty$$

and

$$\left(\forall \theta \in \mathbb{R}^n : \sum_{t=1}^{\infty}(y_t - \theta' x_t)^2 = \infty\right)$$

$$\Longrightarrow \lim_{T\to\infty} \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\min_{\theta \in \mathbb{R}^n}\left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right)} = 1.$$

11.3.2 Kernel Ridge Regression

In this section, $\mathbf{X}$ is an arbitrary set. Let $\mathcal{F}$ be the RKHS with kernel $\mathcal{K}$ of functions on $\mathbf{X}$. The KRR strategy for Learner with parameter $a > 0$ is defined by the formula (11.1) applied to the past examples.

The following version of Theorem 11.1 for KRR can be derived from Theorem 11.1 itself (see [21], Sect. 6, for details).

Theorem 11.2. *The KRR strategy with parameter $a > 0$ for Learner satisfies, at any step T,*

$$\sum_{t=1}^{T} \frac{(y_t - \hat{y}_t)^2}{1 + \frac{1}{a}\mathcal{K}(x_t, x_t) - \frac{1}{a}k_t'(K_{t-1} + aI)^{-1}k_t} = \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2 \right).$$

The denominator on the left-hand side tends to 1 under some regularity conditions:

Lemma 11.1 ([20], Lemma 2). *Let $\mathcal{K}$ be a continuous kernel on a compact metric space. Then*

$$\mathcal{K}(x_t, x_t) - k_t'(K_{t-1} + aI)^{-1}k_t \to 0 \text{ as } t \to \infty.$$

Again, we can derive several interesting corollaries from Theorem 11.2.

Corollary 11.4. *Assume $|y_t| \le \mathbf{y}$ for all t and let $\hat{y}_t^{\mathbf{y}}$ be the predictions of the KRR strategy clipped to $[-\mathbf{y}, \mathbf{y}]$. Then*

$$\sum_{t=1}^{T} (y_t - \hat{y}_t^{\mathbf{y}})^2 \le \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2 \right) + 4\mathbf{y}^2 \ln \det \left(I + \frac{1}{a} K_T \right). \quad (11.9)$$

But now we have a problem: in general, the ln det term is not small compared to T. However, we still have the analogue of Corollary 11.3 (for a detailed derivation, see [20]).

Corollary 11.5 ([20], Corollary 4). *Let $\mathbf{X}$ be a compact metric space and $\mathcal{K}$ be a continuous kernel on $\mathbf{X}$. Then*

$$\left(\exists f \in \mathcal{F} : \sum_{t=1}^{\infty} (y_t - f(x_t))^2 < \infty \right) \Longrightarrow \sum_{t=1}^{\infty} (y_t - \hat{y}_t)^2 < \infty$$

and

$$\left(\forall f \in \mathcal{F} : \sum_{t=1}^{\infty}(y_t - f(x_t))^2 = \infty\right) \Implies \lim_{T\to\infty} \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2\right)} = 1.$$

To obtain a non-asymptotic result of this kind under the assumption $\sup_t |y_t| \le \mathbf{y}$, let us first assume that the number of steps T is known in advance. We will need the notation $c_{\mathcal{F}} := \sqrt{\sup_{x\in\mathbf{X}} \mathcal{K}(x,x)}$. Bounding the logarithm of the determinant in (11.9) we have

$$\ln\det\left(I + \frac{1}{a}\mathbf{K}_T\right) \le T\ln\left(1 + \frac{c_{\mathcal{F}}^2}{a}\right)$$

(cf. (11.8)). Since we know the number T of steps in advance, we can choose a specific value for a; let $a := c_{\mathcal{F}}\sqrt{T}$. This gives us an upper bound with the regret term of the order $O(\sqrt{T})$ for any $f \in \mathcal{F}$:

$$\sum_{t=1}^{T}(y_t - \hat{y}_t^{\mathbf{y}})^2 \le \sum_{t=1}^{T}(y_t - f(x_t))^2 + c_{\mathcal{F}}(\|f\|_{\mathcal{F}}^2 + 4\mathbf{y}^2)\sqrt{T}.$$

If we do not know the number of steps in advance, it is possible to achieve a similar bound using a mixture of KRR over the parameter a with a suitable prior over a:

$$\sum_{t=1}^{T}(y_t - \hat{y}_t^{\mathbf{y}})^2 \le \sum_{t=1}^{T}(y_t - f(x_t))^2 + 8\mathbf{y}\max\left(c_{\mathcal{F}}\|f\|_{\mathcal{F}}, \mathbf{y}\delta T^{-1/2+\delta}\right)\sqrt{T+2} + 6\mathbf{y}^2\ln T + c_{\mathcal{F}}^2\|f\|_{\mathcal{F}}^2 + O(\mathbf{y}^2) \quad (11.10)$$

for any arbitrarily small $\delta > 0$, where the constant implicit in $O(\mathbf{y}^2)$ depends only on δ. (No proof of this result has been published.) The inequality (11.10) still looks asymptotic in that it contains an O term; however, it is easy to obtain an explicit (but slightly messier) non-asymptotic inequality.

In particular, (11.10) shows that if $\mathcal{K}$ is a universal kernel [14] on a topological space $\mathbf{X}$, KRR is competitive with all continuous functions on $\mathbf{X}$: for any continuous $f : \mathbf{X} \to \mathbb{R}$,

$$\limsup_{T\to\infty}\frac{1}{T}\left(\sum_{t=1}^{T}(y_t - \hat{y}_t^{\mathbf{y}})^2 - \sum_{t=1}^{T}(y_t - f(x_t))^2\right) \le 0 \quad (11.11)$$

(assuming $|y_t| \le \mathbf{y}$ for all t). For example, (11.11) holds for $\mathbf{X}$ a compact set in $\mathbb{R}^n$, $\mathcal{K}$ an RBF kernel, and $f : \mathbf{X} \to \mathbb{R}$ any continuous function (see Example 1 in [14]). For continuous universal kernels on compact spaces, (11.11) also follows from Corollary 11.5.

11.4 Kernel Ridge Regression in Conformal Prediction

Suppose we would like to have prediction intervals rather than point predictions, and we would like them to have guaranteed coverage probabilities. It is clear that to achieve this we need a stochastic assumption; it turns out that the randomness assumption is often sufficient to obtain informative prediction intervals. In general, our algorithms will output prediction sets (usually intervals, but not always); to obtain prediction intervals we will apply convex closure (which can only improve coverage probability).

The special case of conformal prediction discussed in this section works as follows. Suppose we have an "underlying algorithm" (such as KRR) producing point predictions in $\mathbb{R}$. Let $(x_1, y_1), \dots, (x_T, y_T)$ be a training set and x_{T+1} be a new object. To find the prediction set for y_{T+1} at a significance level $\epsilon \in (0, 1)$:

- For each possible label $z \in \mathbb{R}$:
  - Set $y_{T+1} := z$;
  - For each $t \in \{1, \dots, T+1\}$ compute the *nonconformity score* $\alpha_t^z := |y_t - \hat{y}_t^z|$, where $\hat{y}_t^z$ is the point prediction for the label of x_t computed by the underlying algorithm from the extended training set $(x_1, y_1), \dots, (x_{T+1}, y_{T+1})$;
  - Compute the p-value

$$p(z) := \frac{1}{T+1} \sum_{t=1}^{T+1} \mathbf{1}_{\{\alpha_t^z \ge \alpha_{T+1}^z\}},$$

    where $\mathbf{1}_{\{\cdot\}}$ is the indicator function;
- Output the prediction set $\{z \in \mathbb{R} \mid p(z) > \epsilon\}$, where ϵ is the given significance level.

This set predictor is the *conformal predictor* based on the given underlying algorithm. Conformal predictors have a guaranteed coverage probability:

Theorem 11.3. *The probability that the prediction set output by a conformal predictor is an error (i.e., fails to include y_{T+1}) does not exceed the significance level ϵ.*

Moreover, in the on-line prediction protocol (Protocol 1, in which Reality outputs (x_t, y_t) independently from the same probability distribution), the long-run

frequency of errors also does not exceed ϵ almost surely. For a proof, see [17] (Theorem 8.1).

The property of conformal predictors asserted by Theorem 11.3 is their *validity*. Validity being achieved automatically, the remaining desiderata for conformal predictors are their "efficiency" (we want the prediction sets to be small, in a suitable sense) and "conditional validity" (we might want to have prespecified coverage probabilities conditional on the training set or some property of the new example).

The idea of conformal prediction is inspired by the Support Vector Machine (and the notation α for nonconformity scores is adapted from Vapnik's Lagrange multipliers). The immediate precursor of conformal predictors was described in the paper [7] co-authored by Vapnik, which is based on the idea that a small number of support vectors warrants a high level of confidence in the SVM's prediction. This idea was present in Vapnik and Chervonenkis's thinking in the 1960s: see, e.g., (3.2) and [15], Theorem 10.5. The method was further developed in [17]; see [13] for a tutorial.

In the case where the conformal predictor is built on top of RR or KRR, there is no need to go over all potential labels $z \in \mathbb{R}$. The set prediction for the example (x_{T+1}, y_{T+1}) can be computed in time $O(T \log T)$ (in the case of RR) or $O(T^2)$ (in the case of KRR). This involves only solving linear equations and sorting; the simple resulting algorithm is called the Ridge Regression Confidence Machine (RRCM) in [11] and [17]. There is an R package implementing the RRCM (in the case of RR), `PredictiveRegression`, available from `CRAN`.

The Bayes predictions (11.3) and (11.4) can be easily converted into prediction intervals. But they are valid only under the postulated probability model, whereas the prediction intervals output by the RRCM are valid under the randomness assumption (as is common in machine learning). This is illustrated by Fig. 11.1, which is a version of Wasserman's Fig. 1 in [19]. We consider the simplest case, where $x_t = 1$ for all t; therefore, the examples (x_t, y_t) can be identified with their labels $y_t \in \mathbb{R}$, which we will call *observations*. The chosen significance level is 20 % and the kernel $\mathcal{K}$ is the dot product. In the top plot, the four observations are generated from $N(1, 1)$; in the middle plot, from $N(10, 1)$; and in the bottom plot, from $N(100, 1)$. The blue lines are the prediction intervals computed by the RRCM with $a = 1$ and the red lines are the Bayes prediction intervals computed as the shortest intervals containing 80 % of the mass (11.4) with $a = 1$ and $\sigma = 1$.

All observations are generated from $N(\theta, 1)$ for various constants θ. When $\theta = 1$ (and so the Bayesian assumption (11.5) can be regarded as satisfied), the Bayes prediction intervals are on average only slightly shorter than the RRCM's (the Bayes prediction interval happens to be wider in Fig. 11.1; for a random seed of the random number generator, the Bayes prediction intervals are shorter in about 54 % of cases). But as θ grows, the RRCM's prediction intervals also grow (in order to cover the observations), whereas the width of the Bayes prediction intervals remains constant. When $\theta = 100$ (and so (11.5) is clearly violated), the Bayes prediction intervals give very misleading results.

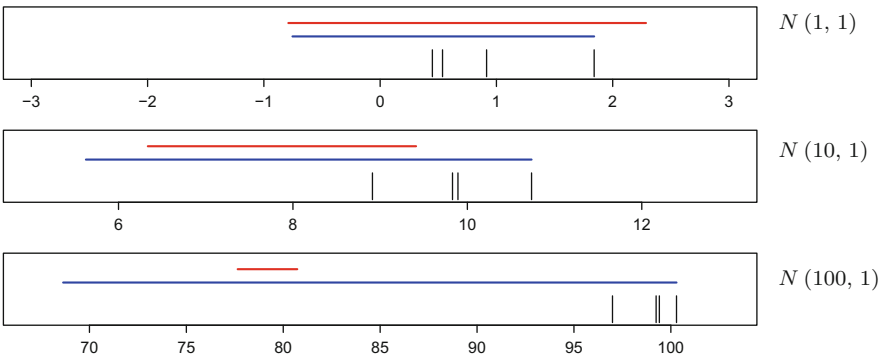

Fig. 11.1 In the *top plot*, the four observations (shown as *short vertical lines*) are generated from $N(1, 1)$; in the *middle plot*, from $N(10, 1)$; and in the *bottom plot*, from $N(100, 1)$. The *blue lines* are prediction intervals computed by a conformal predictor, and the *red lines* are Bayes prediction intervals

In parametric statistics, it is widely believed that the choice of the prior does not matter much: the data will eventually swamp the prior. However, even in parametric statistics the model (such as $N(\theta, 1)$) itself may be wrong.

In nonparametric statistics, the situation is much worse:

> the prior can swamp the data, no matter how much data you have

(Diaconis and Freedman [6], Sect. 4). In this case, using Bayes prediction intervals becomes particularly problematic. The RRCM can be interpreted as an example of *renormalized Bayes*, as discussed in [18] and later papers.

As mentioned earlier, the RRCM is valid under the assumption of randomness; no further assumptions are required. However, conditional validity and, especially, efficiency do require extra assumptions. For example, [10] uses standard statistical assumptions used in density estimation to demonstrate the conditional validity and efficiency of a purpose-built conformal predictor. It remains an open problem to establish whether similar results hold for the RRCM.

Acknowledgements I am deeply grateful to Vladimir Vapnik for numerous discussions and support over the years, starting from our first meetings in the summer of 1996. Many thanks to Alexey Chervonenkis, Alex Gammerman, Valya Fedorova, and Ilia Nouretdinov for their advice and help. This work has been supported in part by the Cyprus Research Promotion Foundation (TPE/ORIZO/0609(BIE)/24) and EPSRC (EP/K033344/1).

References

1. Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Mach. Learn. **43**, 211–246 (2001)
2. Beckenbach, E.F., Bellman, R.: Inequalities. Springer, Berlin (1965)

3. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge (2006)
4. Cressie, N.: The origins of kriging. Math. Geol. **22**, 239–252 (1990)
5. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Methods. Cambridge University Press, Cambridge (2000)
6. Diaconis, P., Freedman, D.: On the consistency of Bayes estimates (with discussion). Ann. Stat. **14**, 1–67 (1986)
7. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, pp. 148–155. Morgan Kaufmann, San Francisco (1998)
8. Kakade, S.M., Ng, A.Y.: Online bounds for Bayesian algorithms. In: Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems, Vancouver (2004)
9. Kumon, M., Takemura, A., Takeuchi, K.: Sequential optimizing strategy in multi-dimensional bounded forecasting games. Stoch. Process. Appl. **121**, 155–183 (2011)
10. Lei, J., Wasserman, L.: Distribution free prediction bands. Tech. Rep. `arXiv:1203.5422 [stat.ME]`, `arXiv.org` e-Print archive (2012). To appear in the Journal of the Royal Statistical Society B
11. Nouretdinov, I., Melluish, T., Vovk, V.: Ridge regression confidence machine. In: Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, pp. 385–392. Morgan Kaufmann, San Francisco (2001)
12. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Shavlik, J.W. (ed.) Proceedings of the Fifteenth International Conference on Machine Learning, Madison, pp. 515–521. Morgan Kaufmann, San Francisco (1998)
13. Shafer, G., Vovk, V.: A tutorial on conformal prediction. J. Mach. Learn. Res. **9**, 371–421 (2008)
14. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2**, 67–93 (2001)
15. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
16. Vovk, V.: Competitive on-line statistics. Int. Stat. Rev. **69**, 213–248 (2001)
17. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)
18. Wasserman, L.: Frequentist Bayes is objective (comment on articles by Berger and by Goldstein). Bayesian Anal. **1**, 451–456 (2006)
19. Wasserman, L.: Frasian inference. Stat. Sci. **26**, 322–325 (2011)
20. Zhdanov, F., Kalnishkan, Y.: An identity for kernel ridge regression. Theor. Comput. Sci. **473**, 157–178 (2013)
21. Zhdanov, F., Vovk, V.: Competing with Gaussian linear experts. Tech. Rep. `arXiv:0910.4683` [cs.LG], `arXiv.org` e-Print archive (2009). Revised in May 2010