APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind. 2008; **24**:171–183
Published online 8 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.700

Consistency of kernel-based quantile regression

Andreas Christmann[1,*,†] and Ingo Steinwart[2]

[1]*Department of Mathematics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussel, Belgium*
[2]*CCS-3, Mail Stop B256, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.*

SUMMARY

Quantile regression is used in many areas of applied research and business. Examples are actuarial, financial or biometrical applications. We show that a non-parametric generalization of quantile regression based on kernels shares with support vector machines the property of consistency to the Bayes risk. We further use this consistency to prove that the non-parametric generalization approximates the conditional quantile function which gives the mathematical justification for kernel-based quantile regression.

Received 30 November 2006; Revised 19 July 2007; Accepted 25 September 2007

KEY WORDS: consistency; convex risk minimization; empirical risk minimization; kernel; non-parametric; quantile regression

1. INTRODUCTION

Consider a random sample (x_i, y_i) from independent and identically distributed random variables (X_i, Y_i) each with unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$, $1 \leqslant i \leqslant n$. For technical reasons, we assume throughout this work that $\mathcal{X}$ and $\mathcal{Y}$ are closed subsets of $\mathbb{R}^m$ and $\mathbb{R}$, respectively. Recall that in this case P can be split up into the marginal distribution P_X and the regular conditional probability $P(\cdot\,|\,x)$, $x \in \mathcal{X}$, on $\mathcal{Y}$.

The goal of quantile regression is to estimate the conditional quantile, i.e. the set-valued function

$$F^*_{\tau,P}(x) := \{q \in \mathbb{R} : P(Y \leqslant q\,|\,x) \geqslant \tau \text{ and } P(Y \geqslant q\,|\,x) \geqslant 1-\tau\}, \quad x \in \mathcal{X}$$

where $\tau \in (0, 1)$ is a fixed constant. For conceptual simplicity (though mathematically this is not necessary) we assume throughout this paper that $F^*_{\tau,P}(x)$ consists of singletons, so that there exists

*Correspondence to: Andreas Christmann, Department of Mathematics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussel, Belgium.
†E-mail: andreas.christmann@vub.ac.be

a unique conditional quantile function $f^*_{\tau,P}: \mathcal{X} \to \mathbb{R}$ defined by $F^*_{\tau,P}(x) = \{f^*_{\tau,P}(x)\}$, $x \in \mathcal{X}$. Now recall that the so-called pinball loss function

$$L_\tau: \mathbb{R} \to [0, \infty), \quad L_\tau(r) := r(\tau - \mathbf{1}_{\{r<0\}}) = \begin{cases} (\tau-1)r & \text{if } r<0 \\ \tau r & \text{if } r \geqslant 0 \end{cases}$$

has the property that $q^* = f^*_{\tau,P}(x)$ if and only if q^* minimizes the conditional L_τ risk, i.e.

$$\mathbb{E}_{P_{Y|X=x}} L_\tau(Y - q^*) = \inf_{q \in \mathbb{R}} \mathbb{E}_{P_{Y|X=x}} L_\tau(Y - q) \tag{1}$$

Based on this fact Koenker and Bassett [1] proposed the estimator

$$\hat{f}_\tau = \arg \inf_{\theta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n L_\tau(y_i - \langle x_i, \theta \rangle)$$

for cases in which $f^*_{\tau,P}$ is a linear function. In this paper we consider a kernel-based generalization, of $\hat{f}_\tau$ which does not require this linearity assumption on $f^*_{\tau,P}$. In order to introduce this generalization, let $\lambda > 0$ be a regularization parameter and H a reproducing kernel Hilbert space (RKHS) of a kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Recall that the reproducing property gives $f(x) = \langle f, \Phi(x) \rangle$ for all $f \in H$ and $x \in \mathcal{X}$, where $\Phi: \mathcal{X} \to H$ is the canonical feature map defined by $\Phi(x) := k(\cdot, x)$, $x \in \mathcal{X}$. Throughout this paper we additionally assume that k is measurable and H is separable, so that Φ becomes Borel measurable by Petti's measurability theorem (see Diestel and Uhl [2]). Schölkopf *et al*. [3, p. 1216] and Takeuchi *et al*. [4] proposed for $\tau \in (0, 1)$ the kernel-based quantile regression (KBQR) estimator which is defined by

$$f_{P,\lambda} := \arg \min_{f \in H} \mathbb{E}_P L_\tau(Y - f(X)) + \lambda \| f \|^2_H \tag{2}$$

For any fixed data set $D_n = \{(x_i, y_i), 1 \leqslant i \leqslant n\} \subset \mathcal{X} \times \mathcal{Y}$ we obtain the estimator

$$f_{D_n,\lambda} := \arg \min_{f \in H} \frac{1}{n} \sum_{i=1}^n L_\tau(y_i - f(x_i)) + \lambda \| f \|^2_H \tag{3}$$

where $D_n = (1/n) \sum_{i=1}^n \delta_{(x_i, y_i)}$ denotes the empirical distribution. We obtain $f_{D_n,\lambda} = \hat{f}_\tau$, if we choose the linear kernel $k(x, x') := \langle x, x' \rangle$ and $\lambda := 0$.

Our first main result is Theorem 5 which shows that KBQR is risk consistent to the Bayes risk under rather weak assumptions, i.e.

$$\mathbb{E}_P L_\tau(Y - f_{D_n,\lambda_n}(X)) \to \inf\{\mathbb{E}_P L_\tau(Y - f(X)) \,|\, f: X \to \mathbb{R} \text{ measurable}\} \tag{4}$$

holds in probability for $n \to \infty$ for suitable sequences of positive regularization parameters (λ_n). Note that the infimum on the right-hand side of (4) is with respect to *all* measurable functions and not only with respect to all functions in the RKHS H. Our second main result which is Theorem 6 shows that whenever KBQR is Bayes risk consistent it also satisfies

$$\| f_{D_n,\lambda_n} - f^*_{\tau,P} \|_0 \to 0$$

where $\| \cdot \|_0$ denotes a translation invariant metric describing the convergence in probability. Together both results give a mathematical justification for using KBQR in non-parametric quantile regression problems.

DOI: 10.1002/asmb

The rest of the paper is organized as follows. Section 2 presents conditions that ensure the existence of $f_{P,\lambda}$. These results will be used to prove our main theorems which are presented in Section 3. Section 4 contains some numerical results. All proofs are given in the Appendix.

2. EXISTENCE AND UNIQUENESS OF INFINITE-SAMPLE KBQR

For any distribution P on $\mathcal{X}\times\mathcal{Y}$ and any measurable map $f:\mathcal{X}\to\mathbb{R}$, we define the L_τ-risk of f with respect to P by

$$\mathcal{R}_{L_\tau,P}(f):=\mathbb{E}_P L_\tau\big(Y-f(X)\big)=\int_{\mathcal{X}}\int_{\mathcal{Y}} L_\tau(y-f(x))\,\mathrm{d}P(y|x)\,\mathrm{d}P_X(x)$$

where we recall that the regular conditional probability $P(\cdot|x)$ exists because $\mathcal{Y}$ is closed (and thus a Polish space). Moreover, note that the above integral is always defined since L_τ is non-negative and continuous, but in general it is not finite. In order to find a condition that ensures $\mathcal{R}_{L_\tau,P}(f)<\infty$ we define

$$|P|_1:=\mathbb{E}_P|Y|=\int_{\mathcal{X}\times\mathcal{Y}}|y|\,\mathrm{d}P(x,y)$$

Now we can formulate a sufficient condition ensuring $\mathcal{R}_{L_\tau,P}(f)<\infty$.

Proposition 1
Let P be a distribution on $\mathcal{X}\times\mathcal{Y}$ with $|P|_1<\infty$ and $f:\mathcal{X}\to\mathbb{R}$ be a function with $f\in L_1(P)$. Then we have $\mathcal{R}_{L_\tau,P}(f)<\infty$.

The following lemma presents in some sense an inverse statement of the above proposition.

Lemma 2
Let $f:\mathcal{X}\to\mathbb{R}$ be a measurable function and P be a distribution on $\mathcal{X}\times\mathcal{Y}$ with $\mathcal{R}_{L_\tau,P}(f)<\infty$. Then we have $|P|_1<\infty$ if and only if $f\in L_1(P)$.

The next result ensures the existence of a solution $f_{P,\lambda}$. In order to formulate it recall that a kernel $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ of an RKHS H is called *bounded* if

$$\|k\|_\infty:=\sup_{x\in\mathcal{X}}\sqrt{k(x,x)}<\infty$$

For such kernels it is well known that the reproducing property yields $\|f\|_\infty\leqslant\|k\|_\infty\cdot\|f\|_H$ for all $f\in H$. In particular, if P is a distribution on $\mathcal{X}\times\mathcal{Y}$ with $|P|_1<\infty$ then the objective function in (2) is always finite by Proposition 1, i.e. we have

$$R^{\mathrm{reg}}_{L_\tau,P,\lambda}(f):=\mathcal{R}_{L_\tau,P}(f)+\lambda\|f\|_H^2<\infty$$

for all $f\in H$. With these preparations we can now establish the existence and uniqueness of $f_{P,\lambda}$.

DOI: 10.1002/asmb

Proposition 3
Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_1 < \infty$, H be an RKHS of a bounded kernel k and $\lambda > 0$. Then there exists a unique minimizer $f_{P,\lambda} \in H$ of

$$f \mapsto \mathcal{R}^{\text{reg}}_{L_\tau, P, \lambda}(f)$$

and we have $\|f_{P,\lambda}\|_H \leqslant \sqrt{|P|_1/\lambda}$.

3. MAIN RESULTS

Our first goal in this section is to present a result that establishes risk consistency of KBQR, i.e. we will show that (4) holds in probability for $n \to \infty$ and suitable sequences of positive regularization parameters (λ_n). Of course, for such convergence to hold it is necessary that the used RKHS H is rich enough in the sense of

$$\mathcal{R}^*_{L_\tau, P, H} := \inf_{f \in H} \mathcal{R}_{L_\tau, P}(f) = \inf\{\mathcal{R}_{L_\tau, P}(f) \mid f : \mathcal{X} \to \mathbb{R} \text{ measurable}\} =: \mathcal{R}^*_{L_\tau, P} \tag{5}$$

The following proposition which is essentially taken from Steinwart *et al*. [5] translates this richness into an easier-to-handle denseness assumption.

Proposition 4
Let H be the RKHS of a bounded kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and μ be a distribution on $\mathcal{X}$. Then the following statements are equivalent:

(i) H is dense in $L_1(\mu)$.
(ii) Equation (5) holds for all distributions P on $\mathcal{X} \times \mathcal{Y}$ with $P_X = \mu$ and $|P|_1 < \infty$.

Note that it was shown by Steinwart *et al.* [5] that e.g. the popular Gaussian radial basis function (RBF) kernel defined by $k(x, x') = \exp(-\gamma^{-2}\|x - x'\|^2)$ for $\gamma > 0$ fixed and $x, x' \in \mathbb{R}^m$ satisfies condition (i) of Proposition 4 for *all* distributions μ on $\mathbb{R}^m$. Obviously, this kernel is also bounded, because $|k(x, x')| \leqslant 1$ for all $x, x' \in \mathbb{R}^m$. Moreover, for *compact* $\mathcal{X}$, and continuous kernels k on $\mathcal{X}$, condition (i) of Proposition 4 is satisfied for all distributions μ on $\mathcal{X}$ if k is *universal* in the sense of Steinwart [6], i.e. if its RKHS is dense in the space $C(\mathcal{X})$ of continuous functions mapping $\mathcal{X}$ to $\mathbb{R}$. Examples of such kernels including the Gaussian RBF kernel are described by Steinwart [6]. Finally, note that polynomial kernels $k(x, x') = (c + \langle x, x' \rangle)^m$, $m \geqslant 1$, $c \geqslant 0$, $x, x' \in \mathbb{R}^m$, are also popular in practice, but they are neither bounded nor dense for general measures μ.

We can now formulate our first main result.

Theorem 5
Let $\mathcal{X} \subset \mathbb{R}^m$ be a closed subset and H be the RKHS of a bounded measurable kernel k on $\mathcal{X}$ such that H is dense in $L_1(\mu)$ for all distributions μ on $\mathcal{X}$. Furthermore, let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$. Then the KBQR estimator defined by (3) using λ_n for sample sets of length n is risk consistent in the sense of (4) for all distributions P with $|P|_1 < \infty$. Moreover, if (λ_n) actually satisfies $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$ then (4) holds even almost surely.

DOI: 10.1002/asmb

In order to formulate our second main result let us introduce some more notations. To this end let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ and $f, g : \mathcal{X} \to \mathbb{R}$ be measurable functions. We write

$$\|f\|_{L_0(P_X)} := \|f\|_0 := \int_{\mathcal{X}} \min\{1, |f|\} \, \mathrm{d}P_X$$

and $d(f, g) := \|f - g\|_0$. It is elementary to check that d is a *translation invariant* metric on the space of all measurable functions defined on $\mathcal{X}$, and furthermore a simple application of Chebyshev's inequality shows that d describes the convergence in probability P_X.

The following result shows that under the assumptions of Theorem 5 the KBQR estimator approximates the conditional quantile function in terms of $\|\cdot\|_{L_0(P_X)}$.

Theorem 6
Let $\mathcal{X} \subset \mathbb{R}^m$ be a closed subset and H be the RKHS of a bounded measurable kernel k on $\mathcal{X}$ such that H is dense in $L_1(\mu)$ for all distributions μ on $\mathcal{X}$. Furthermore, let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$. Then the KBQR estimator defined by (3) satisfies

$$\|f_{D_n, \lambda_n} - f^*_{\tau, P}\|_{L_0(P_X)} \to 0 \tag{6}$$

in probability for $n \to \infty$ and all distributions P on $\mathcal{X} \times \mathcal{Y}$ with $|P|_1 < \infty$. Moreover, if (λ_n) actually satisfies $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then (6) holds even almost surely.

It is interesting to note that the assumption $F^*_{\tau, P}(x) = \{f^*_{\tau, P}(x)\}$ is only needed to formulate Theorem 6 in terms of $\|\cdot\|_0$. However, Theorem 3.16 of Steinwart [7] which is used in the proof of Theorem 6 actually provides a framework to replace $\|\cdot\|_0$ by a more general notion of closedness if the assumption $F^*_{\tau, P}(x) = \{f^*_{\tau, P}(x)\}$ is violated.

In some sense the convergence with respect to $\|\cdot\|_0$ is rather weak and one may wonder whether it can be replaced by some stronger notion of convergence. For example, note that for $\tau = 0.5$ Theorem 5 established the convergence

$$\mathbb{E}_P |Y - f_{D_n, \lambda_n}(X)| - \mathbb{E}_P |Y - f^*_{\tau, P}(X)| \to 0, \quad n \to \infty \tag{7}$$

which naturally raises the question whether we actually have

$$\mathbb{E}_P |f_{D_n, \lambda_n}(X) - f^*_{\tau, P}(X)| \to 0 \tag{8}$$

Of course, the inverse triangle inequality $\|a| - |b\| \leqslant |a - b|$ immediately shows that (8) implies (7), but since for general $a, b, c \in \mathbb{R}$ the inequality $|a - c| - |b - c| \geqslant |a - c|$ is false, we conjecture that without additional assumptions on P the convergence in (8) does not follow the one in (7). In this direction it is also interesting to note that the framework developed by Steinwart [7] suggests that for certain classes of distributions P we can actually replace $\|\cdot\|_{L_0(P_X)}$ by some (quasi)-norm $\|\cdot\|_{L_p(P_X)}$. However, such considerations are out of the scope of this paper.

Another interesting question is whether we can establish convergence rates in Theorem 5 or Theorem 6. Of course, it is well known in learning theory that such convergence rates require additional assumptions on the distribution P, e.g. in terms of the approximation properties of H with respect to $f^*_{\tau, P}$. Moreover, the techniques used in the proof of Theorem 5 or Theorem 6 are tuned to provide consistency under rather minimal assumptions on $\mathcal{X}$, $\mathcal{Y}$, P and H, but in general these techniques are too weak to obtain good convergence rates in the statistical analysis. Because of these reasons, convergence rates are also out of the scope of this paper.

DOI: 10.1002/asmb

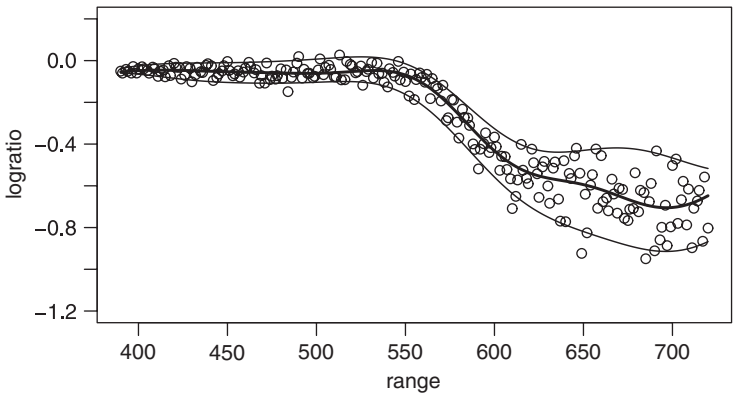

Figure 1. LIDAR data set ($n=221$). Kernel-based quantile regression based on the RBF kernel with $\gamma^2=0.5$ and $\lambda=\frac{1}{700}n^{-1/3}$. Considered quantile levels: $\tau=0.05, 0.50$, and 0.95.

4. NUMERICAL RESULTS

4.1. Example: LIDAR data set

We now analyze data concerning the so-called LIDAR technique. LIDAR is the abbreviation of LIght Detection And Ranging. This technique uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere. We consider the logarithm of the ratio of received light from two laser sources as the response variable $Y=$ `logratio`, whereas the single explanatory variable $X=$ `range` is the distance traveled before the light is reflected back to its source. A scatterplot of the data set, consisting of $n=221$ observations, is shown in Figure 1, together with the fitted curves based on KBQR using the standard RBF kernel for the median and the lower and upper 5% quantiles. See Ruppert *et al*. [8] for more details on this data set. KBQR clearly shows that the relationship between both variables is non-linear, almost constant for values of `range` below 550 and decreasing for higher values of `range`. However, KBQR also shows that the variability of `logratio` is non-constant and much greater for values of `range` say above 600 than for values below this value.

4.2. Simulation results

Now we shall describe a small simulation and its results to investigate how well the asymptotical results derived in Section 3 on the consistency of KBQR work for small to moderate sized sample sizes. We consider $n\in\{221, 1000, 4000\}$ and use the same parameter settings than in the previous subsection, i.e. $\lambda_n=\frac{1}{700}n^{-1/3}$, and $\gamma^2=0.5$ for the RBF kernel. The number of replications in the simulation was set to 1000. For each replication $\ell\in\{1,\ldots,1000\}$, we independently generated n data points $x_i^{(\ell)}$ for `range` from a continuous uniform distribution with support $(390, 720)$. Furthermore, for each replication we generated n data points $y_i^{(\ell)}$ for `logratio` according to independent normal distributions with conditional expectations and variances given by

$$\mu(x_i^{(\ell)}) := -0.05 - \frac{0.7}{1+\exp[-(x_i^{(\ell)}-600)/10]}$$

DOI: 10.1002/asmb

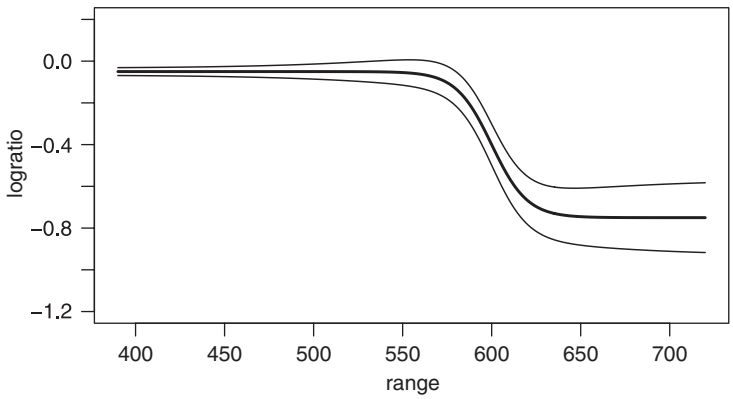

Figure 2. True quantile regression curves for the simulation. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

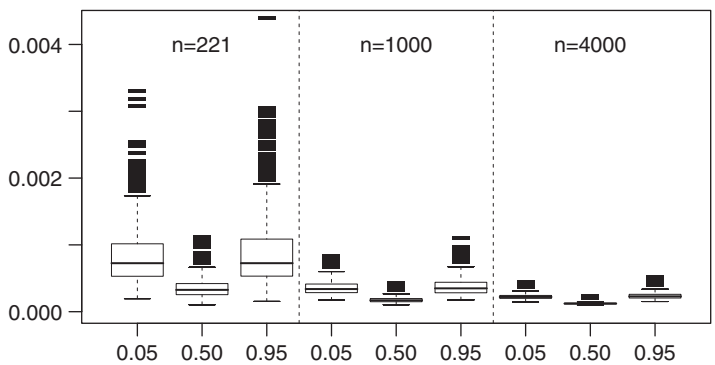

Figure 3. Simulation results for the criterion IMSE_τ for kernel-based quantile regression based on the RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

and

$$\sigma^2(x_i^{(\ell)}) := \left(0.01 + \frac{0.1}{1+\exp[-(x_i^{(\ell)} - 600)/50]}\right)^2$$

respectively. The true conditional τ-quantile curves are thus given by $f^*_{\tau,P}(x) = \mu(x) + u_\tau \sigma(x)$, $\tau \in (0, 1)$, where u_τ defines the τ-quantile of a normal distribution with mean 0 and variance 1. The curves for the conditional medians and the conditional lower and upper 5% quantiles are shown in Figure 2 to illustrate that this model generates data sets similar to the LIDAR data set, see Figure 1.

DOI: 10.1002/asmb

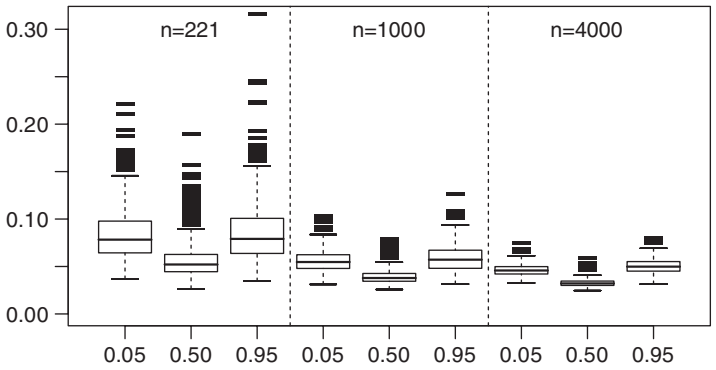

Figure 4. Simulation results for the criterion mBias_τ for kernel-based quantile regression based on the RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

We use two criteria to measure how well the KBQR estimates approximate the true conditional quantiles. Our first criterion is

$$\mathrm{IMSE}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \left[\frac{1}{n} \sum_{i=1}^{n} (f_{D_n,\lambda_n}(x_i^{(\ell)}) - f^*_{\tau,P}(x_i^{(\ell)}))^2 \right]$$

which is an empirical version of the integrated mean-squared error. To measure the worst case behavior of the KBQR estimates we use the criterion

$$\mathrm{mBias}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \max_{1 \leqslant i \leqslant n} |f_{D_n,\lambda_n}(x_i^{(\ell)}) - f^*_{\tau,P}(x_i^{(\ell)})|$$

which is an empirical version of the maximum bias. The random number generation and the plots were made with the statistical software `R` [9]. The program `mySVM` [10] was used for the computation of the KBQR estimates.

The boxplots given in the Figures 3 and 4 show that the KBQR estimates perform quite well with respect to both criteria under the considered circumstances, because both criteria have relatively small values and their values decrease with increasing sample sizes. A considerable improvement is obtained by increasing the sample size by a factor of around 4. The simulations indicate that the consistency results derived in Section 3 can be useful even for small to moderate sample sizes.

5. CONCLUSION

In this paper we proved that the KBQR proposed by Schölkopf *et al*. [3, p. 1216] and Takeuchi *et al*. [4] is risk consistent, i.e. the L_τ-risk of the KBQR estimator converges in probability to the Bayes risk which is defined as the smallest L_τ-risk for all measurable functions. A similar result was recently obtained by Christmann and Steinwart [11] for support vector regression (see Schölkopf and Smola [12] for an introduction). Moreover, we have shown that the KBQR estimator converges in probability to the conditional quantile function, which provides a mathematical justification of this method. We mention that KBQR also has good robustness properties due to the Lipschitz

DOI: 10.1002/asmb

continuity of the pinball loss function if a bounded and continuous kernel is used, see Christmann and Van Messem [13].

It might be possible to get rid of the assumption $|P|_1<\infty$ when considering KBQR if one changes the empirical regularized minimization problem (3) to

$$f_{P,\lambda} := \arg\inf_{f\in H} \mathbb{E}_P L_\tau^*(Y-f(X))+\lambda\|f\|_H^2$$

where $L_\tau^*(y,t) := L_\tau(y-t)-L_\tau(y)$ for $y,t\in\mathbb{R}$. However, loss functions that can take on negative values are beyond the scope of this paper.

For another non-parametric generalization of $\hat{f}_\tau$ based on splines, we refer to Koenker *et al.* [14] and He and Ng [15].

APPENDIX

The appendix contains the proofs of our results. We begin by summarizing some properties of the pinball loss function. For each $\tau\in(0,1)$, the pinball loss function L_τ is convex, satisfies $L_\tau(0)=0$ and $\lim_{|r|\to\infty} L_\tau(r)=\infty$, and is Lipschitz continuous with Lipschitz constant $|L_\tau|_1=\max\{\tau,1-\tau\}$. Furthermore,

$$\min\{\tau,1-\tau\}\,|r| \leqslant L_\tau(r) \leqslant |L_\tau|_1\,|r|, \quad r\in\mathbb{R} \tag{A1}$$

Proof of Proposition 1
By (A1) we have

$$\mathscr{R}_{L_\tau,P}(f)=\mathbb{E}_P L_\tau(Y-f(X)) \leqslant |L_\tau|_1 \mathbb{E}_P(|Y|+|f(X)|) \leqslant |P|_1+\|f\|_{L_1(P)}<\infty \qquad \square$$

Proof of Lemma 2
For all $a,b\in\mathbb{R}$ we have $|a-b|\geqslant|a|-|b|$. Now let us assume that we know $f\in L_1(P)$. Using (A1) we obtain

$$\infty>\mathscr{R}_{L_\tau,P}(f)\geqslant\min\{\tau,1-\tau\}\mathbb{E}_P(|Y-f(X)|)\geqslant\min\{\tau,1-\tau\}\mathbb{E}_P(|Y|-|f(X)|)$$

From this we obtain $|P|_1<\infty$. The converse implication can be shown analogously. $\square$

Proof of Proposition 3
Our proof follows DeVito *et al.* [16] in a streamlined fashion. Combining (A1) with Lemma 2 of [5], we see that $\mathscr{R}_{L_\tau,P}: L_1(P_X)\to\mathbb{R}$ is continuous. Furthermore, $\mathrm{id}: H\to L_1(P_X)$ is continuous since k is bounded and hence $\mathscr{R}^{\mathrm{reg}}_{L_\tau,P,\lambda}: H\to\mathbb{R}$ is continuous. This map is also convex, and the set $\{f\in H:\mathscr{R}^{\mathrm{reg}}_{L_\tau,P,\lambda}(f)\leqslant\delta_{P,\lambda}\}$ is bounded and non-empty, because it contains $0\in H$. Therefore, Ekeland and Turnbull [17, Proposition II.4.6] ensure the existence of $f_{P,\lambda}$. The uniqueness follows from the strict convexity of $\mathscr{R}^{\mathrm{reg}}_{L_\tau,P,\lambda}$. The last assertion is trivial. $\square$

Our next goal is to obtain a representation of $f_{P,\lambda}$. To this end we need the notion of subdifferentials which is recalled in the following definition.

DOI: 10.1002/asmb

Definition A1 (Subdifferential)**.**
Let H be a Hilbert space, $F: H \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in H$ with $F(w) \neq \infty$. Then the subdifferential of F at w is defined by

$$\partial F(w) := \{w^* \in H : \langle w^*, v - w \rangle \leqslant F(v) - F(w) \text{ for all } v \in H\}$$

With the help of the subdifferential ∂L_τ we can now recall a result shown by DeVito *et al*. [16] (in a slightly generalized form) which in turn is a generalization of a representation derived by Steinwart [18].

Proposition A2
Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_1 < \infty$, k be a bounded, measurable kernel k over $\mathcal{X}$ with separable RKHS H, and $\Phi: \mathcal{X} \to H$ be the canonical feature map of k. Then for all $\lambda > 0$ there exists a bounded and measurable function $h_\lambda: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $h_\lambda(x, y) \in \partial L_\tau(y - f_{P,\lambda}(x))$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and

$$f_{P,\lambda} = -\frac{1}{2\lambda} \mathbb{E}_P h_\lambda \Phi \tag{A2}$$

With the help of Proposition A2 we can now state the following stability result.

Theorem A3
Let P, H, Φ, and h_λ be as in Proposition A2. Then we have $\|h_\lambda\|_\infty \leqslant |L_\tau|_1$ and for all distributions Q on $\mathcal{X} \times \mathcal{Y}$ with $|Q|_1 < \infty$ we have

$$\|f_{P,\lambda} - f_{Q,\lambda}\|_H \leqslant \frac{1}{\lambda} \|\mathbb{E}_P h_\lambda \Phi - \mathbb{E}_Q h_\lambda \Phi\|_H \tag{A3}$$

Proof
Let us first show the upper bound for $\|h_\lambda\|_\infty$. To this end we observe

$$|h_\lambda(x, y)| \leqslant |\partial L_\tau(y - f_{P,\lambda}(x))| \leqslant |L_\tau(y - \cdot)_{|[-f_{P,\lambda}(x), f_{P,\lambda}(x)]}|_1 \leqslant |L_\tau|_1$$

and hence we deduce $\|h_\lambda\|_\infty \leqslant |L_\tau|_1$. In order to prove (A3) we first observe that the definition of the subdifferential yields

$$h(x, y)(f_{Q,\lambda}(x) - f_{P,\lambda}(x)) \leqslant L_\tau(y - f_{Q,\lambda}(x)) - L_\tau(y - f_{P,\lambda}(x))$$

and hence

$$\mathbb{E}_Q L_\tau(Y - f_{P,\lambda}(X)) + \langle f_{Q,\lambda} - f_{P,\lambda}, \mathbb{E}_Q h\Phi \rangle \leqslant \mathbb{E}_Q L_\tau(Y - f_{Q,\lambda}(X)) \tag{A4}$$

Moreover, an easy calculation shows

$$\lambda \|f_{P,\lambda}\|_H^2 + 2\lambda \langle f_{Q,\lambda} - f_{P,\lambda}, f_{P,\lambda} \rangle + \lambda \|f_{P,\lambda} - f_{Q,\lambda}\|_H^2 = \lambda \|f_{Q,\lambda}\|_H^2 \tag{A5}$$

Combining (A4) and (A5) it follows

$$\mathcal{R}^{\text{reg}}_{L_\tau, Q, \lambda}(f_{P,\lambda}) + \langle f_{Q,\lambda} - f_{P,\lambda}, \mathbb{E}_Q h\Phi + 2\lambda f_{P,\lambda} \rangle + \lambda \|f_{P,\lambda} - f_{Q,\lambda}\|_H^2 \leqslant \mathcal{R}^{\text{reg}}_{L_\tau, Q, \lambda}(f_{Q,\lambda}) \leqslant \mathcal{R}^{\text{reg}}_{L_\tau, Q, \lambda}(f_{P,\lambda})$$

DOI: 10.1002/asmb

Therefore, by using the representation $f_{P,\lambda} = -(1/2\lambda)\mathbb{E}_P h\Phi$ we obtain

$$\begin{aligned}\lambda \| f_{P,\lambda} - f_{Q,\lambda} \|_H^2 &\leqslant \langle f_{P,\lambda} - f_{Q,\lambda}, \mathbb{E}_Q h\Phi - \mathbb{E}_P h\Phi \rangle \\ &\leqslant \| f_{P,\lambda} - f_{Q,\lambda} \|_H \cdot \| \mathbb{E}_Q h\Phi - \mathbb{E}_P h\Phi \|_H\end{aligned}$$

From this we easily obtain the assertion. □

Proof of Proposition 4
The implication (i) ⇒ (ii) immediately follows from Theorem 3 of Steinwart *et al.* [5] and the converse implication can be easily established by combining Theorem 8 with a straightforward modification of Example 5 of Steinwart *et al.* [5]. □

In order to prove Theorem 5 we need some preliminary results. Our first lemma shows that the influence of the regularization term $\lambda \| f_{P,\lambda} \|_H^2$ used in the definition of KBQR vanishes for $\lambda \to 0$.

Lemma A4
Let H be an RKHS over $\mathcal{X}$ with bounded kernel k and P be a distribution on $\mathcal{X} \times \mathcal{Y}$ such that $|P|_1 < \infty$. Then we have

$$\lim_{\lambda \to 0^+} \mathcal{R}^{\text{reg}}_{L_\tau, P, \lambda}(f_{P,\lambda}) = \mathcal{R}^*_{L_\tau, P, H}$$

Proof
For $\varepsilon > 0$ we fix an $f_\varepsilon \in H$ such that $\mathcal{R}_{L_\tau, P}(f_\varepsilon) \leqslant \mathcal{R}^*_{L_\tau, P, H} + \varepsilon$. Then for all $\lambda < \varepsilon \| f_\varepsilon \|_H^{-2}$ we have

$$\mathcal{R}_{L_\tau, P, H} \leqslant \lambda \| f_{P,\lambda} \|_H^2 + \mathcal{R}_{L_\tau, P}(f_{P,\lambda}) \leqslant \lambda \| f_\varepsilon \|_H^2 + \mathcal{R}_{L_\tau, P}(f_\varepsilon) \leqslant 2\varepsilon + \mathcal{R}^*_{L_\tau, P, H}$$ □

The next lemma gives a simple but useful approximation of $|\mathcal{R}_{L_\tau, P}(f) - \mathcal{R}_{L_\tau, P}(g)|$.

Lemma A5
Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_1 < \infty$. For all bounded measurable functions $f, g : \mathcal{X} \to \mathcal{Y}$, we have

$$\left| \mathcal{R}_{L_\tau, P}(f) - \mathcal{R}_{L_\tau, P}(g) \right| \leqslant |L_\tau|_1 \, \| f - g \|_\infty$$

Proof
The Lipschitz continuity of L_τ immediately gives

$$|\mathcal{R}_{L_\tau, P}(f) - \mathcal{R}_{L_\tau, P}(g)| \leqslant \int |L_\tau(y - f(x)) - L_\tau(y - g(x))| \, \mathrm{d}P(x, y) \leqslant |L_\tau|_1 \, \| f - g \|_\infty$$ □

Under the assumptions of Lemma A4 and Proposition 4 we immediately see that $\mathcal{R}_{L_\tau, P}(f_{P,\lambda_n}) \to \mathcal{R}^*_{L_\tau, P}$ holds for $\lambda_n \to 0$. Therefore, we obtain risk consistency whenever we can show that $|\mathcal{R}_{L_\tau, P}(f_{P,\lambda_n}) - \mathcal{R}_{L_\tau, P}(f_{D_n,\lambda_n})| \to 0$ holds in probability for $n \to \infty$ and suitable null sequences (λ_n). Our main tool for ensuring this convergence will be Theorem A3 which in particular describes the behavior of $\| f_{P,\lambda_n} - f_{D_n,\lambda_n} \|_\infty$ if we let Q be an empirical measure based on a sample set of length n. Lemma A5 showed how the norm of this difference can be used to estimate $|\mathcal{R}_{L_\tau, P}(f_{P,\lambda_n}) - \mathcal{R}_{L_\tau, P}(f_{D_n,\lambda_n})|$.

Let us now deal with the stochastic analysis of $|\mathcal{R}_{L_\tau, P}(f_{P,\lambda}) - \mathcal{R}_{L_\tau, P}(f_{D_n,\lambda})| \to 0$. To this end we need the following theorem which can be found in Chapter 3 of Yurinsky [19].

DOI: 10.1002/asmb

Theorem A6 (Hoeffding's inequality in Hilbert spaces)**.**
Let $(\Omega, \mathscr{A}, P)$ be a probability space, H be a separable Hilbert space and $B>0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent H-valued, bounded random variables with $\|\xi_i\|_\infty \leqslant B$ for all $i=1, \ldots, n$. Then for all $\varepsilon \geqslant n^{-1/2}$ we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}_P\xi_i)\right\|_H \geqslant \varepsilon\right) \leqslant \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon^2 n}{\varepsilon B + 3B^2}\right)$$

Proof of Theorem 5
To avoid handling with too many constants let us assume $\|k\|_\infty = 1$. Obviously, this implies $\|f\|_\infty \leqslant \|k\|_\infty \|f\|_H = \|f\|_H$ for all $f \in H$ and hence Lemma A5 implies

$$\left|\mathscr{R}_{L,P}(f_{P,\lambda_n}) - \mathscr{R}_{L,P}(g)\right| \leqslant |L_\tau|_1 \left\|f_{P,\lambda_n} - g\right\|_\infty \leqslant \|f_{P,\lambda_n} - g\|_H \tag{A6}$$

for all $g \in H$. For $n \in \mathbb{N}$ and $\lambda_n > 0$ we now write $h_n : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ for the function we obtain by Proposition A2 and Theorem A3. Moreover, let $\varepsilon > 0$ and D_n be a training set of length n with empirical distribution D_n such that

$$\|\mathbb{E}_P h_n \Phi - \mathbb{E}_{D_n} h_n \Phi\|_H \leqslant \lambda_n \varepsilon \tag{A7}$$

Then Theorem A3 gives $\|f_{P,\lambda_n} - f_{D_n,\lambda_n}\|_H \leqslant \varepsilon$ and hence (A6) yields

$$|\mathscr{R}_{L,P}(f_{P,\lambda_n}) - \mathscr{R}_{L,P}(f_{D_n,\lambda_n})| \leqslant \|f_{P,\lambda_n} - f_{D_n,\lambda_n}\|_H \leqslant \varepsilon \tag{A8}$$

Let us now estimate the probability of D_n satisfying (A7). To this end we first observe that $\lambda_n n^{1/2} \to \infty$ implies that for all sufficiently large n we have $\lambda_n \varepsilon \geqslant n^{-1/2}$. Moreover, Theorem A3 shows $\|h_n\|_\infty \leqslant 1$ and our assumption $\|k\|_\infty = 1$ thus yields $\|h_n \Phi\|_\infty \leqslant 1$. Consequently, Theorem A6 yields

$$P^n(D_n \in (\mathscr{X} \times \mathscr{Y})^n : \|\mathbb{E}_P h_n \Phi - \mathbb{E}_{D_n} h_n \Phi\|_H \leqslant \lambda_n \varepsilon) \geqslant 1 - \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon^2 \lambda_n^2 n}{\varepsilon\lambda_n + 3}\right)$$

for all sufficiently large n. Using $\lambda_n n^{1/2} \to \infty$ and $\lambda_n \to 0$, we thus find that the probability of sample sets D_n satisfying (A7) converges to 1 if $|D_n| = n \to \infty$. As we have seen above this implies that (A8) holds true with probability tending to 1. Now, since $\lambda_n \to 0$ we additionally have $|\mathscr{R}_{L,P}(f_{P,\lambda_n}) - \mathscr{R}_{L,P}| \leqslant \varepsilon$ for all sufficiently large n and hence we finally obtain the first assertion.

In order to show the second assertion we define $\varepsilon_n := (\ln(n+1)))^{-1/2}$, and $\delta_n := \mathscr{R}_{L,P}(f_{P,\lambda_n}) - \mathscr{R}^*_{L,P} + \varepsilon_n$, $n \geqslant 1$. Moreover, for an infinite sample $D_\infty := ((x_1, y_1), \ldots) \in (\mathscr{X} \times \mathscr{Y})^\infty$ we write $D_n := ((x_1, y_1), \ldots, (x_n, y_n))$. With these notations we define

$$A_n := \{D_\infty \in (\mathscr{X} \times \mathscr{Y})^\infty : \mathscr{R}_{L,P}(f_{D_n,\lambda_n}) - \mathscr{R}^*_{L,P} > \delta_n\}, \quad n \geqslant 1$$

Now, our above estimates together with $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$ yields

$$\sum_{i=1}^{\infty} P^\infty(A_n) \leqslant \sum_{i=1}^{\infty} \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon_n^2 \lambda_n^2 n}{\varepsilon_n \lambda_n + 3}\right) < \infty$$

DOI: 10.1002/asmb

and hence we obtain

$$P^{\infty}(\{D_{\infty} \in (\mathcal{X} \times \mathcal{Y})^{\infty} | \exists n_0 \forall n \geqslant n_0 : \mathcal{R}_{L,P}(f_{D_n,\lambda_n}) - \mathcal{R}^*_{L,P} \leqslant \delta_n\}) = 1$$

by the Borel–Cantelli lemma. Since $\lambda_n \to 0$ implies $\delta_n \to 0$ we then find the assertion. □

Proof of Theorem 6
We have already seen in Theorem 5 that the KBQR estimator satisfies

$$\mathcal{R}_{L_\tau,P}(f_{D_n,\lambda_n}) \to \mathcal{R}^*_{L_\tau,P}$$

for $n \to \infty$. Moreover, $(y,t) \mapsto L_\tau(y,t)$ is a supervised convex loss function in the sense of Steinwart [7] whose conditional risks have a unique minimizer, namely $f^*_{\tau,P}$. Consequently, Theorem 3.16 of Steinwart [7] in the form of Remark 3.18 yields the assertion. □

REFERENCES

1. Koenker R, Bassett G. Regression quantiles. *Econometrica* 1978; **46**:33–50.
2. Diestel J, Uhl J. *Vector Measures*. American Mathematical Society: Providence, RI, 1977.
3. Schölkopf B, Smola A, Williamson R, Bartlett P. New support vector algorithms. *Neural Computation* 2000; **12**:1207–1245.
4. Takeuchi I, Le Q, Sears T, Smola A. Nonparametric quantile estimation. *Journal of Machine Learning Research* 2006; **7**:1231–1264.
5. Steinwart I, Hush D, Scovel C. Function classes that approximate the Bayes risk. *Proceedings of the 19th Annual Conference on Learning Theory*, Carnegie Mellon University Pittsburgh, Pennsylvania, *COLT 2006*. Springer: Berlin, 2006; 79–93.
6. Steinwart I. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* 2001; **2**:67–93.
7. Steinwart I. How to compare different loss functions. *Constructive Approximation* 2005; **26**:225–287.
8. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, 2003.
9. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2006 ISBN 3-900051-07-0, http://www.R-project.org.
10. Rüping S. *mySVM-Manual*. University of Dortmund, Department of Computer Science, 2000. http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM.
11. Christmann A, Steinwart I. Consistency and robustness of kernel based regression. *Bernoulli* 2007; **13**:799–819.
12. Schölkopf B, Smola A. *Learning with Kernels. Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press: Cambridge, MA, 2002.
13. Christmann A, Van Messem A. Bouligand derivatives and robustness of support vector machines. *Technical Report*, Vrije Universiteit Brussel, Tentatively accepted by: *Journal of Machine Learning Research*, 2007.
14. Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika* 1994; **81**:673–680.
15. He X, Ng P. COBS: qualitatively constrained smoothing via linear programming. *Computational Statistics* 1999; **14**:315–337.
16. DeVito E, Rosasco L, Caponnetto A, Piana M, Verri A. Some properties of regularized kernel methods. *Journal of Machine Learning Research* 2004; **5**:1363–1390.
17. Ekeland I, Turnbull T. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press: Chicago, 1983.
18. Steinwart I. Sparseness of support vector machines. *Journal of Machine Learning Research* 2003; **4**:1071–1105.
19. Yurinsky V. *Sums and Gaussian Vectors*. Lecture Notes in Mathematics, vol. 1617. Springer: Berlin, 1995.

DOI: 10.1002/asmb