Journal of Machine Learning Research 7 (2006) 1231–1264 Submitted 10/05; Revised 5/06; Published 7/06

Nonparametric Quantile Estimation

Ichiro Takeuchi TAKEUCHI@PA.INFO.MIE-U.AC.JP
Division of Computer Science
Graduate School of Engineering, Mie University
1577, Kurimamachiya-cho, Tsu 514-8507, Japan

Quoc V. Le QUOC.LE@ANU.EDU.AU
Timothy D. Sears TIM.SEARS@ANU.EDU.AU
Alexander J. Smola ALEX.SMOLA@NICTA.COM.AU
RSISE, Australian National University and
Statistical Machine Learning Program, National ICT Australia
0200, ACT, Australia

Editor: Chris Williams

Abstract

In regression, the desired estimate of $y|x$ is not always given by a conditional mean, although this is most common. Sometimes one wants to obtain a good estimate that satisfies the property that a proportion, τ, of $y|x$, will be below the estimate. For $\tau = 0.5$ this is an estimate of the *median*. What might be called median regression, is subsumed under the term *quantile regression*. We present a nonparametric version of a quantile estimator, which can be obtained by solving a simple quadratic programming problem and provide uniform convergence statements and bounds on the quantile property of our estimator. Experimental results show the feasibility of the approach and competitiveness of our method with existing ones. We discuss several types of extensions including an approach to solve the *quantile crossing* problems, as well as a method to incorporate prior qualitative knowledge such as monotonicity constraints.

Keywords: support vector machines, kernel methods, quantile estimation, nonparametric techniques, estimation with constraints

1. Introduction

Regression estimation is typically concerned with finding a real-valued function f such that its values $f(x)$ correspond to the conditional mean of y, or closely related quantities. Many methods have been developed for this purpose, e.g. least mean square (LMS) regression, robust regression (Huber, 1981), or ε-insensitive regression (Vapnik, 1995; Vapnik et al., 1997). Regularized variants include Wahba (1990), penalized by a Reproducing Kernel Hilbert Space (RKHS) norm, and Hoerl and Kennard (1970), regularized via ridge regression.

1.1 Motivation

While these estimates of the mean serve their purpose, there exists a large area of problems where we are more interested in estimating a quantile. That is, we might wish to know other features of the the distribution of the random variable $y|x$:

- A device manufacturer may wish to know what are the 10% and 90% quantiles for some feature of the production process, so as to tailor the process to cover 80% of the devices produced.

- For risk management and regulatory reporting purposes, a bank may need to estimate a lower bound on the changes in the value of its portfolio which will hold with high probability.

- A pediatrician requires a growth chart for children given their age and perhaps even medical background, to help determine whether medical interventions are required, e.g. while monitoring the progress of a premature infant.

These problems are addressed by a technique called Quantile Regression (QR) or Quantile Estimation championed by Koenker (see Koenker, 2005, for a description, practical guide, and extensive list of references). These methods have been deployed in econometrics, social sciences, ecology, etc. The purpose of our paper is:

- To bring the technique of quantile regression to the attention of the machine learning community and show its relation to ν-Support Vector Regression (Schölkopf et al., 2000).

- To demonstrate a nonparametric version of QR which outperforms the currently available nonlinear QR regression formations (Koenker, 2005). See Section 5 for details.

- To derive small sample size results for the algorithms. Most statements in the statistical literature for QR methods are of asymptotic nature (Koenker, 2005). Empirical process results permit us to define two quality criteria and show tail bounds for both of them in the finite-sample-size case.

- To extend the technique to permit commonly desired constraints to be incorporated. As examples we show how to enforce non-crossing constraints and a monotonicity constraint. These constraints allow us to incorporate prior knowlege on the data.

1.2 Notation and Basic Definitions

In the following we denote by $\mathcal{X}, \mathcal{Y}$ the domains of x and y respectively. $X = \{x_1, \ldots, x_m\}$ denotes the training set with corresponding targets $Y = \{y_1, \ldots, y_m\}$, both drawn independently and identically distributed (iid) from some distribution $p(x,y)$. With some abuse of notation y also denotes the vector of all y_i in matrix and vector expressions, whenever the distinction is obvious.

Unless specified otherwise $\mathcal{H}$ denotes a Reproducing Kernel Hilbert Space (RKHS) on $\mathcal{X}$, k is the corresponding kernel function, and $K \in \mathbb{R}^{m \times m}$ is the kernel matrix obtained via $K_{ij} = k(x_i, x_j)$. θ denotes a vector in *feature space* and $\phi(x)$ is the corresponding feature map of x. That is, $k(x,x') = \langle \phi(x), \phi(x') \rangle$. Finally, $\alpha \in \mathbb{R}^m$ is the vector of Lagrange multipliers.

Definition 1 (Quantile) *Denote by $y \in \mathbb{R}$ a random variable and let $\tau \in (0,1)$. Then the τ-quantile of y, denoted by μ_τ is given by the infimum over μ for which $\Pr\{y \leq \mu\} = \tau$. Likewise, the conditional quantile $\mu_\tau(x)$ for a pair of random variables $(x,y) \in \mathcal{X} \times \mathbb{R}$ is defined as the function $\mu_\tau : \mathcal{X} \to \mathbb{R}$ for which pointwise μ_τ is the infimum over μ for which $\Pr\{y \leq \mu | x\} = \tau$.*

1.3 Examples

To illustrate regression analyses with conditional quantile functions, we provide two simple examples here.

1.3.1 ARTIFICIAL DATA

The above definition of conditional quantiles may be best illustrated by a simple example. Consider a situation where the relationship between x and y is represented as

$$y(x) = f(x) + \xi, \text{ where } \xi \sim \mathcal{N}\left(0, \sigma(x)^2\right). \tag{1}$$

Here, note that, the amount of noise ξ is a function of x. Since ξ is symmetric with mean and median 0 we have $\mu_{0.5}(x) = f(x)$. Moreover, we can compute the τ-th quantiles by solving $\Pr\{y \leq \mu | x\} = \tau$ explicitly. Since ξ is normally distributed, we know that the τ-th quantile of ξ is given by $\sigma(x)\Phi^{-1}(\tau)$, where Φ is the cumulative distribution function of the normal distribution with unit variance. This means that

$$\mu_\tau(x) = f(x) + \sigma(x)\Phi^{-1}(\tau).$$

Figure 1 shows the case where x is uniformly drawn from $[-1,1]$ and y is obtained based on (1) with $f(x) = \text{sinc}(x)$ and $\sigma(x) = 0.1\exp(1-x)$. The black circles are 500 data examples and the five curves are $\tau = 0.10, 0.25, 0.50, 0.75$ and 0.90 conditional quantile functions. The probability densities $p(y|x=-0.5)$ and $p(y|x=+0.5)$ are superimposed. The τ-th conditional quantile function is obtained by connecting the τ-th quantile of the conditional distribution $p(y|x)$ for all $x \in \mathcal{X}$. We see that $\tau = 0.5$ case provides the central tendency of the data distribution and $\tau = 0.1$ and 0.9 cases track the lower and upper envelope of the data points, respectively. The error bars of many regression estimates can be viewed as crude quantile regressions. Quantile regression on the other hand tries to estimate such quantities directly.

1.3.2 REAL DATA

The next example is based on actual measurements of bone density (BMD) in adolescents. The data was originally reported in Bachrach et al. (1999) and is also analyzed in Hastie et al. (2001).[1] Figure 2 (a) shows a regression analysis with conditional mean and figure 2 (b) shows that with a set of conditional quantiles for the variable BMD. The response in the vertical axis is relative change in spinal BMD and the covariate in the horizontal axis is the age of the adolescents. The conditional mean analysis (a) provides only the central tendency of the conditional distribution, while apparently the entire distribution of BMD changes according to age. The conditional quantile analysis (b) gives us more detailed description of these changes. For example, we can see that the variance of the BMD changes with the age (heteroscedastic) and that the conditional distribution is slightly positively skewed.

2. Quantile Estimation

Given the definition of $\mu_\tau(x)$ and knowledge of support vector machines we might be tempted to use version of the ε-insensitive tube regression to estimate $\mu_\tau(x)$. More specifically one might try to

1. The data is also available from the website http://www-stat.stanford.edu/ElemStatlearn.

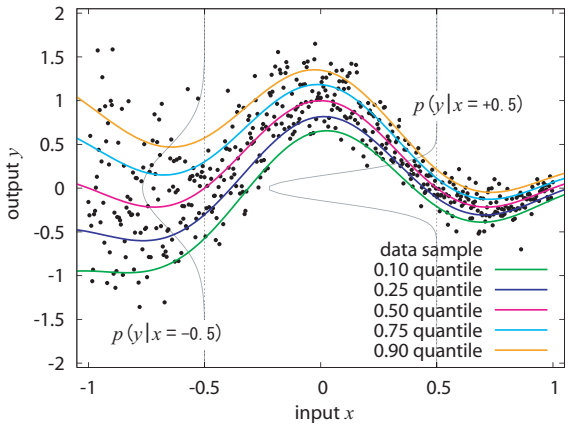


Figure 1: Illustration of conditional quantile functions of a simple artificial system in (1) with $f(x) = \mathrm{sinc}(x)$ and $\sigma(x) = 0.1\exp(1-x)$. The black circles are 500 data examples and the five curves are $\tau = 0.10, 0.25, 0.50, 0.75$ and 0.90 conditional quantile functions. The probability densities $p(y|x=-0.5)$ and $p(y|x=+0.5)$ are superimposed. In this paper, we are concerned with the problem of estimating these conditional quantile functions from training data.

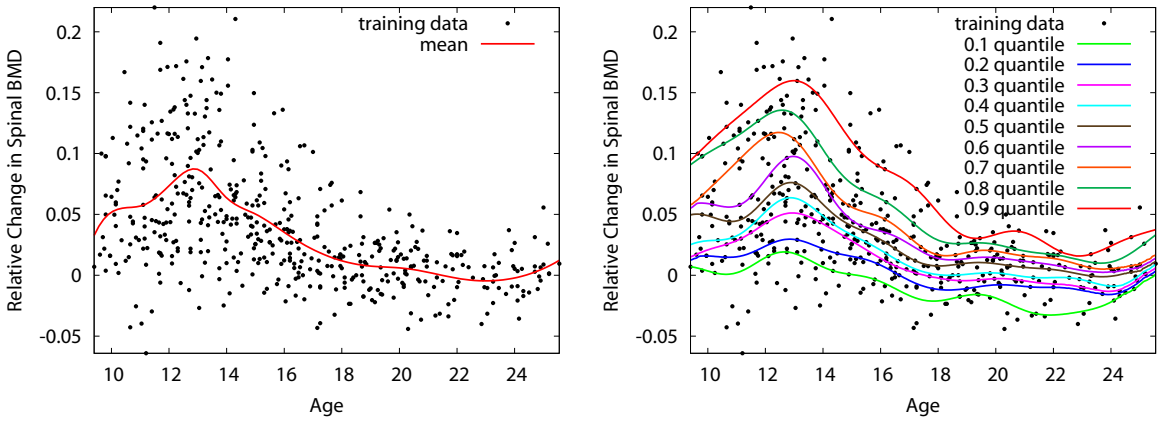
estimate quantiles nonparametrically using an extension of the ν-trick, as outlined in Schölkopf et al. (2000). However this approach carries the disadvantage of requiring us to estimate both an upper and lower quantile *simultaneously*.[2] While this can be achieved by quadratic programming, in doing so we estimate "too many" parameters simultaneously. More to the point, if we are interested in finding an upper bound on y which holds with 0.95 probability we may not want to use information about the 0.05 probability bound in the estimation. Following Vapnik's paradigm of estimating only the relevant parameters directly (Vapnik, 1982) we attack the problem by estimating each quantile separately. For completeness and comparison, we provide a detailed description of a symmetric quantile regression in Appendix A.

2.1 Loss Function

The basic strategy behind quantile estimation arises from the observation that minimizing the ℓ_1-loss function for a location estimator yields the median. Observe that to minimize $\sum_{i=1}^m |y_i - \mu|$ by choice of μ, an equal number of terms $y_i - \mu$ have to lie on either side of zero in order for the derivative wrt. μ to vanish. Koenker and Bassett (1978) generalizes this idea to obtain a regression estimate for any quantile by tilting the loss function in a suitable fashion. More specifically one may show that the following "pinball" loss leads to estimates of the τ-quantile:

Lemma 2 (Quantile Estimator) *Let $Y = \{y_1, \ldots, y_m\} \subset \mathbb{R}$ and let $\tau \in (0,1)$ then the minimizer μ_τ of $\sum_{i=1}^m l_\tau(y_i - \mu)$ with respect to μ satisfies:*

2. Schölkopf et al. (2000) does, in fact, suggests that a choice of different upper bounds on the dual problem would lead to estimators which weigh errors for positive and negative excess differently, that is, which would lead to quantile regression estimators.



(a) Conditional mean analysis (b) Conditional quantile analysis

Figure 2: An illustration of (a) conditional mean analysis and (b) conditional quantile analysis for a data set on bone mineral density (BMD) in adolescents. In (a) the conditional mean curve is estimated by regression spline with least square criterion. In (b) the nine curves are the estimated conditional quantile curves at orders $0.1, 0.2, \ldots, 0.9$. The set of conditional quantile curves provides more informative description of the relationship among variables such as non-constant variance or non-normality of the noise (error) distribution. In this paper, we are concerned with the problem of estimating these conditional quantiles.

$$l_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau-1)\xi & \text{if } \xi < 0 \end{cases} \qquad (2)$$

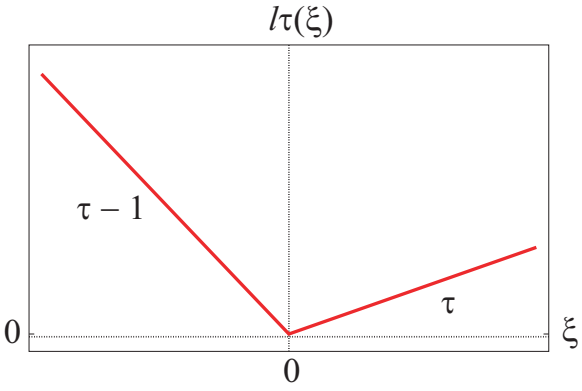


Figure 3: Pinball loss function for quantile estimation.

1. *The number of terms, m_-, with $y_i < \mu_\tau$ is bounded from above by τm.*
2. *The number of terms, m_+, with $y_i > \mu_\tau$ is bounded from above by $(1-\tau)m$.*
3. *For $m \to \infty$, the fraction $\frac{m_-}{m}$, converges to τ if $\Pr(y)$ does not contain discrete components.*

Proof Assume that we are at an optimal solution. Then, increasing the minimizer μ by $\delta\mu$ changes the objective by $[(1-m_+)(1-\tau) - m_+\tau]\,\delta\mu$. Likewise, decreasing the minimizer μ by $\delta\mu$ changes the objective by $[-m_-(1-\tau) + (1-m_-)\tau]\,\delta\mu$. Requiring that both terms are nonnegative at opti-

mality in conjunction with the fact that $m_- + m_+ \leq m$ proves the first two claims. To see the last claim, simply note that the event $y_i = y_j$ for $i \neq j$ has probability measure zero for distributions not containing discrete components. Taking the limit $m \to \infty$ shows the claim. ■

The idea is to use the same loss function for functions, $f(x)$, rather than just constants in order to obtain quantile estimates conditional on x. Koenker (2005) uses this approach to obtain linear estimates and certain nonlinear spline models. In the following we will use kernels for the same purpose.

2.2 Optimization Problem

Based on $l_\tau(\xi)$ we define the expected quantile risk as

$$R[f] := \mathbf{E}_{p(x,y)}\left[l_\tau(y - f(x))\right]. \tag{3}$$

By the same reasoning as in Lemma 2 it follows that for $f : \mathcal{X} \to \mathbb{R}$ the minimizer of $R[f]$ is the quantile $\mu_\tau(x)$. Since $p(x,y)$ is unknown and we only have X, Y at our disposal we resort to minimizing the empirical risk plus a regularizer:

$$R_{\mathrm{reg}}[f] := \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}} \text{ where } f = g + b \text{ and } b \in \mathbb{R}. \tag{4}$$

Here $\|\cdot\|_{\mathcal{H}}$ is RKHS norm and we require $g \in \mathcal{H}$. Notice that we do not regularize the constant offset, b, in the optimization problem. This ensures that the minimizer of (4) will satisfy the quantile property:

Lemma 3 (Empirical Conditional Quantile Estimator) *Assuming that f contains a scalar unregularized term, the minimizer of (4) satisfies:*

1. *The number of terms m_- with $y_i < f(x_i)$ is bounded from above by τm.*
2. *The number of terms m_+ with $y_i > f(x_i)$ is bounded from above by $(1-\tau)m$.*
3. *If (x,y) is drawn iid from a distribution $\Pr(x,y)$, with $\Pr(y|x)$ continuous and the expectation of the modulus of absolute continuity of its density satisfying $\lim_{\delta\to 0}\mathbf{E}\left[\varepsilon(\delta)\right] = 0$. With probability 1, asymptotically, $\frac{m_-}{m}$ equals τ.*

Proof For the two claims, denote by f^* the minimum of $R_{\mathrm{reg}}[f]$ with $f^* = g^* + b^*$. Then $R_{\mathrm{reg}}[g^* + b]$ has to be minimal for $b = b^*$. With respect to b, however, minimizing R_{reg} amounts to finding the τ quantile in terms of $y_i - g(x_i)$. Application of Lemma 2 proves the first two parts of the claim.

For the second part, an analogous reasoning to Schölkopf et al. (2000, Proposition 1) applies. In a nutshell, one uses the fact that the measure of the δ-neighborhood of $f(x)$ converges to 0 for $\delta \to 0$. Moreover, for kernel functions the entropy numbers are well behaved (Williamson et al., 2001). The application of the union bound over a cover of such function classes completes the proof. Details are omitted, as the proof is identical to that of Schölkopf et al. (2000). ■

Later, in Section 4 we discuss finite sample size results regarding the convergence of $\frac{m_-}{m} \to \tau$ and related quantities. These statements will make use of scale sensitive loss functions. Before we do that, let us consider the practical problem of minimizing the regularized risk functional.

2.3 Dual Optimization Problem

Here we compute the dual optimization problem to (4) for efficient numerical implementation. Using the connection between RKHS and feature spaces we write $f(x) = \langle \phi(x), w \rangle + b$ and we obtain the following equivalent to minimizing $R_{\text{reg}}[f]$.

$$\underset{w,b,\xi_i^{(*)}}{\text{minimize}} \quad C\sum_{i=1}^{m} \tau\xi_i + (1-\tau)\xi_i^* + \frac{1}{2}\|w\|^2 \tag{5a}$$

$$\text{subject to} \quad y_i - \langle \phi(x_i), w \rangle - b \leq \xi_i \text{ and } \langle \phi(x_i), w \rangle + b - y_i \leq \xi_i^* \text{ where } \xi_i, \xi_i^* \geq 0 \tag{5b}$$

Here we used $C := 1/(\lambda m)$. The dual of this problem can be computed straightforwardly using Lagrange multipliers. The dual constraints for ξ and ξ^* can be combined into one variable. This yields the following dual optimization problem

$$\underset{\alpha}{\text{minimize}} \quad \frac{1}{2}\alpha^\top K\alpha - \alpha^\top \vec{y} \text{ subject to } C(\tau-1) \leq \alpha_i \leq C\tau \text{ for all } 1 \leq i \leq m \text{ and } \vec{1}^\top\alpha = 0. \tag{6}$$

We recover f via the familiar kernel expansion

$$w = \sum_i \alpha_i \phi(x_i) \text{ or equivalently } f(x) = \sum_i \alpha_i k(x_i, x) + b. \tag{7}$$

Note that the constant b is the dual variable to the constraint $\mathbf{1}^\top\alpha = \mathbf{0}$. Alternatively, b can be obtained by using the fact that $f(x_i) = y_i$ for $\alpha_i \notin \{C(\tau-1), C\tau\}$. The latter holds as a consequence of the KKT-conditions on the primal optimization problem of minimizing $R_{\text{reg}}[f]$.

Note that the optimization problem is very similar to that of an ε-SV regression estimator (Vapnik et al., 1997). The key difference between the two estimation problems is that in ε-SVR we have an additional $\varepsilon\|\alpha\|_1$ penalty in the objective function. This ensures that observations with deviations from the estimate, i.e. with $|y_i - f(x_i)| < \varepsilon$ do not appear in the support vector expansion. Moreover the upper and lower constraints on the Lagrange multipliers α_i are matched. This means that we balance excess in both directions. The latter is useful for a regression estimator. In our case, however, we obtain an estimate which penalizes loss unevenly, depending on whether $f(x)$ exceeds y or vice versa. This is exactly what we want from a quantile estimator: by this procedure errors in one direction have a larger influence than those in the converse direction, which leads to the shifted estimate we expect from QR. A practical advantage of (6) is that it can be solved directly with standard quadratic programming code rather than using pivoting, as is needed in SVM regression (Vapnik et al., 1997).

A practical estimate does require a procedure for setting the regularization parameter. Figure 4 shows how QR responds to changing the regularization parameter. All three estimates in Figure 4 attempt to compute the median, subject to different smoothness constraints. While they all satisfy the quantile property having half the points on either side of the regression, some estimates appear track the observations better. This issue is addressed in Section 5 where we compute quantile regression estimates on a range of data sets.

3. Extensions and Modifications

Our optimization framework lends itself naturally to a series of extensions and modifications of the regularized risk minimization framework for quantile regression. In the following we discuss some extensions and modifications.

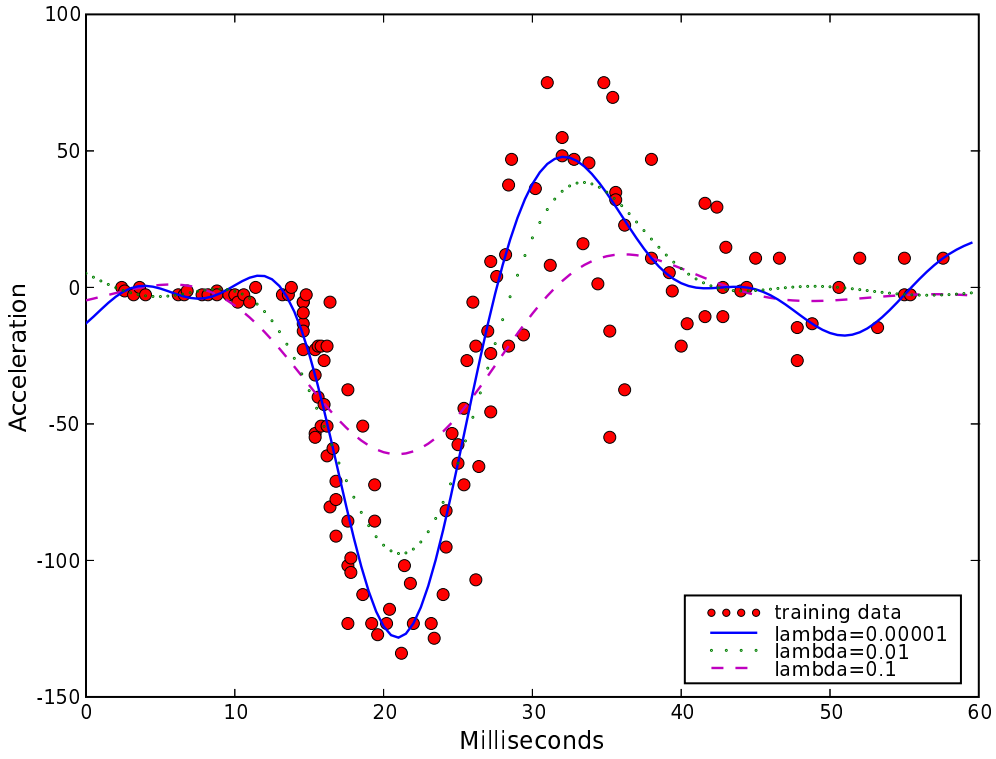


Figure 4: The data set measures acceleration in the head of a crash test dummy v. time in tests of motorcycle crashes. Three regularized versions of the median regression estimate ($\tau = 0.5$). While all three variants satisfy the quantile property, the degree of smoothness is controlled by the regularization constant λ. All three estimates compare favorably to a similar graph of nonlinear QR estimates reported by Koenker (2005).

3.1 Non-Crossing Constraints

When we want to estimate several conditional quantiles (e.g. $\tau = 0.1, 0.2, \ldots, 0.9$), two or more estimated conditional quantile functions can cross or overlap. This embarrassing phenomenon called *quantile crossings* occurs because each conditional quantile function is independently estimated (Koenker, 2005; He, 1997). Figure 5(a) shows BMD data presented in 1.3.2 and $\tau = 0.1, 0, 2, \ldots, 0.9$ conditional quantile functions estimated by the kernel-based estimator described in the previous section. Both of the input and the output variables are standardized in $[0,1]$. We note quantile crossings at several places, especially at the outside of the training data range ($x < 0$ and $1 < x$). In this subsection, we address this problem by introducing *non-crossing constraints*.[3] Figure 5(b) shows a family of conditional quantile functions estimated with the non-crossing constraints.

Suppose that we want to estimate n conditional quantiles at $0 < \tau_1 < \tau_2 < \ldots < \tau_n < 1$. We enforce *non-crossing* constraints at l points $\{x_j\}_{j=1}^{l}$ in the input domain $\mathcal{X}$. Let us write the model for the τ_h-th conditional quantile function as $f_h(x) = \langle \phi(x), w_h \rangle + b_h$ for $h = 1, 2, \ldots, n$. In $\mathcal{H}$ the non-crossing constraints are represented as linear constraints

$$\left\langle \phi(x_j), \omega_h \right\rangle + b_h \leq \left\langle \phi(x_j), \omega_{h+1} \right\rangle + b_{h+1}, \text{ for all } 1 \leq h \leq n-1,\ 1 \leq j \leq l. \tag{8}$$

Solving (5) or (6) for $1 \leq h \leq n$ with non-crossing constraints (8) allows us to estimate n conditional quantile functions not crossing at l points $x_1, \ldots, x_l \in \mathcal{X}$. The primal optimization problem is given by

$$\underset{w_h, b_h, \xi_{hi}^{(*)}}{\text{minimize}} \quad \sum_{h=1}^{n} \left[C \sum_{i=1}^{m} \tau_h \xi_{hi} + (1-\tau_h)\xi_{hi}^{*} + \frac{1}{2}\|w_h\|^2 \right] \tag{9a}$$

$$\text{subject to } \ y_i - \langle \phi(x_i), w_h \rangle - b_h = \xi_{hi} - \xi_{hi}^{*} \text{ where } \xi_{hi}, \xi_{hi}^{*} \geq 0, \quad \text{for all } 1 \leq h \leq n,\ 1 \leq i \leq m. \tag{9b}$$

$$\left\{ \left\langle \phi(x_j), \omega_{h+1} \right\rangle + b_{h+1} \right\} - \left\{ \left\langle \phi(x_j), \omega_h \right\rangle + b_h \right\} \geq 0, \quad \text{for all } 1 \leq h \leq n-1,\ 1 \leq j \leq l. \tag{9c}$$

Using Lagrange multipliers, we can obtain the dual optimization problem:

$$\underset{\alpha_h, \theta_h}{\text{minimize}} \quad \sum_{h=1}^{n} \left[\frac{1}{2}\alpha_h^{\top} K \alpha_h + \alpha_h^{\top} \tilde{K}(\theta_{h-1} - \theta_h) + \frac{1}{2}(\theta_{h-1} - \theta_h)^T \bar{K}(\theta_{h-1} - \theta_h) - \alpha_h^{\top} \vec{y} \right] \tag{10a}$$

$$\text{subject to } \ C(\tau_h - 1) \leq \alpha_{hi} \leq C\tau_h, \text{ for all } 1 \leq h \leq n, 1 \leq i \leq m, \tag{10b}$$

$$\theta_{hj} \geq 0, \text{ for all } 1 \leq h \leq n, 1 \leq j \leq l, \ \vec{1}^{\top} \alpha_h = 0, \text{ for all } 1 \leq h \leq n, \tag{10c}$$

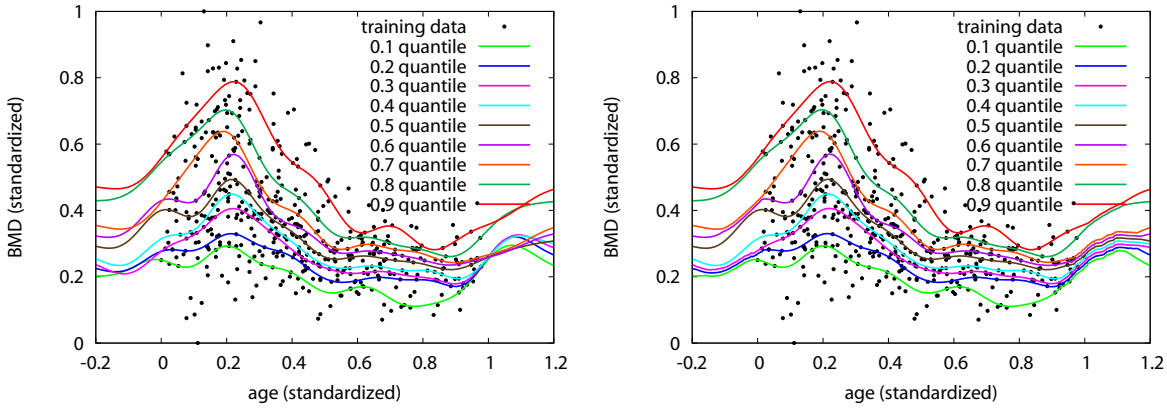
where θ_{hj} is the Lagrange multiplier of (9c) for all $1 \leq h \leq n$, $1 \leq j \leq l$, $\tilde{K}$ is $m \times l$ matrix with its (i, j)-th entry $k(x_i, x_j)$, $\bar{K}$ is $l \times l$ matrix with its (j_1, j_2)-th entry $k(x_{j1}, x_{j2})$ and θ_h is l-vector with its j-th entry θ_{hj} for all $1 \leq h \leq n$. For notational convenience we define $\theta_{0j} = \theta_{nj} = 0$ for all $1 \leq j \leq l$. The model for conditional quantile τ_h-th quantile function is now represented as

$$f_h(x) = \sum_{i=1}^{m} \alpha_{hi} k(x, x_i) + \sum_{j=1}^{l} (\theta_{h-1i} - \theta_{hi}) k(x, x_j) + b_h. \tag{11}$$

3. A part of the contents in this subsection was presented by one of the authors (Takeuchi and Furuhashi, 2004).

In section 5.2.1 we empirically investigate the effect of non-crossing constraints on the generalization performances.

It is worth noting that, after enforcing the non-crossing constraints, the quantile property as in Lemma 3 may not be guaranteed. This is because the method both tries to optimize for the quantile property and the non-crossing property (in relation to other quantiles). Hence, the final outcome may not empirically satisfy the quantile property. Yet, the non-crossing constraints are very nice because they ensure the semantics of the quantile definition: lower quantile level should not cross the higher quantile level.



(a) Without *non-crossing* constraints

(b) With *non-crossing* constraints

Figure 5: An example of *quantile crossing* problem in BMD data set presented in Section 1. Both of the input and the output variable are standardized in $[0,1]$. In (a) the set of conditional quantiles at $0.1, 0.2, \ldots, 0.9$ are estimated by the kernel-based estimator presented in the previous section. Quantile crossings are found at several points, especially at the outside of the training data range ($x < 0$ and $1 < x$). The plotted curves in (b) are the conditional quantile functions obtained with *non-crossing* constraints explained in Section 3.1. There are no *quantile crossing* even at the outside of the training data range.

3.2 Monotonicity and Growth Curves

Consider the situation of a health statistics office which wants to produce growth curves. That is, it wants to generate estimates of y being the height of a child given parameters x such as age, ethnic background, gender, parent's height, etc. Such curves can be used to assess whether a child's growth is abnormal.

A naive approach is to apply QR directly to the problem of estimating $y|x$. Note, however, that we have additional information about the biological process at hand: the height of every individual child is a *monotonically increasing* function of age. Without observing large amounts of data, there is no guarantee that the estimates $f(x)$, will also be monotonic functions of age. Figure 6 is an example of quantile regression with monotonicity constraints. The data set is taken from Mammen et al. (2001). Fuel efficiency (in miles per gallon) is studied as a function of engine output.

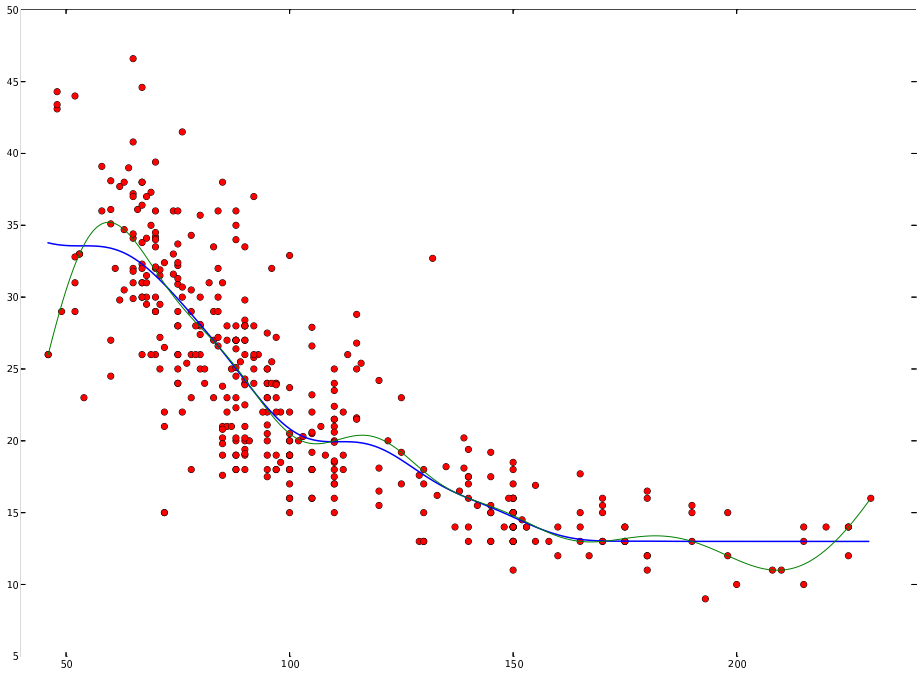


Figure 6: Example plots from quantile regression with and without monotonicity constraints. The thin line represents the nonparametric quantile regression without monotonicity constraints whereas the thick line represents the nonparamtric quantile regression with monotonicity constraints.

To address this problem we adopt an approach similar to (Vapnik et al., 1997; Smola and Schölkopf, 1998) and impose constraints on the derivatives of f directly. While this only ensures that f is monotonic on the observed data X, we could always add more locations x_i' for the express purpose of enforcing monotonicity.

Formally, we require that for a differential operator D, such as $D = \partial_{x_{\text{age}}}$ the estimate $Df(x) \geq 0$ for all $x \in X$. Using the linearity of inner products we have

$$Df(x) = D(\langle \phi(x), w \rangle + b) = \langle D\phi(x), w \rangle = \langle \psi(x), w \rangle \text{ where } \psi(x) := D\phi(x). \tag{12}$$

Note that accordingly inner products between ψ and ϕ can be obtained via $\langle \psi(x), \phi(x') \rangle = D_1 k(x,x')$ and $\langle \psi(x), \psi(x') \rangle = D_1 D_2 k(x,x')$, where D_1 and D_2 denote the action of D on the first and second argument of k respectively. Consequently the optimization problem (5) acquires an additional set of

constraints and we need to solve

$$\begin{aligned}\underset{w,b,\xi_i}{\text{minimize}} \;\; & C\sum_{i=1}^{m}\tau\xi_i + (1-\tau)\xi_i^* + \frac{1}{2}\|w\|^2 \\ \text{subject to} \;\; & y_i - \langle\phi(x_i), w\rangle - b \leq \xi_i, \;\; \langle\phi(x_i), w\rangle + b - y_i \leq \xi_i^*, \\ & \langle\psi(x_i), w\rangle \geq 0, \;\; \xi_i, \xi_i^* \geq 0.\end{aligned}$$

Since the additional constraint does not depend on b it is easy to see that the quantile property still holds. The dual optimization problem yields

$$\underset{\alpha,\beta}{\text{minimize}} \;\; \frac{1}{2}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}^\top \begin{bmatrix}K & D_1K\\ D_2K & D_1D_2K\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} - \alpha^\top\vec{y} \tag{13a}$$

$$\text{subject to} \;\; C(\tau-1) \leq \alpha_i \leq C\tau \text{ and } 0 \leq \beta_i \text{ for all } 1 \leq i \leq m \text{ and } \vec{1}^\top\alpha = 0. \tag{13b}$$

Here D_1K is a shorthand for the matrix of entries $D_1k(x_i, x_j)$ and D_2K, D_1D_2K are defined analogously. Here $w = \sum_i \alpha_i\phi(x_i) + \beta_i\psi(x_i)$ or equivalently $f(x) = \sum_i \alpha_i k(x_i, x) + \beta_i D_1k(x_i, x) + b$.

Example Assume that $x \in \mathbb{R}^n$ and that x_1 is the coordinate with respect to which we wish to enforce monotonicity. Moreover, assume that we use a Gaussian RBF kernel, that is

$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\left\|x - x'\right\|^2\right). \tag{14}$$

In this case $D_1 = \partial_1$ with respect to x and $D_2 = \partial_1$ with respect to x'. Consequently we have

$$D_1k(x, x') = \frac{x_1' - x_1}{\sigma^2}k(x, x'); D_2k(x, x') = \frac{x_1 - x_1'}{\sigma^2}k(x, x') \tag{15a}$$

$$D_1D_2k(x, x') = \left[\sigma^{-2} - \frac{(x_1 - x_1')^2}{\sigma^4}\right]k(x, x'). \tag{15b}$$

Plugging the values of (15) into (13) yields the quadratic program. Note also that both $k(x, x')$ and $D_1k(x, x')$ in (15a), are used in the function expansion.

If x_1 were drawn from a discrete (yet ordered) domain we could replace D_1, D_2 with a finite difference operator. This is still a linear operation on k and consequently the optimization problem remains unchanged besides a different functional form for D_1k.

An alternative to the above approach is not to modify the optimization problem but to ensure the constraints by modifying the function in the hypothesis space which is much simpler to implement as in Le et al. (2006).

3.3 Other Function Classes

Semiparametric Estimates RKHS expansions may not be the only function classes desired for quantile regression. For instance, in the social sciences a semiparametric model may be more desirable, as it allows for interpretation of the linear coefficients (Gu and Wahba, 1993; Smola et al., 1999; Bickel et al., 1994). In this case we add a set of parametric functions f_i and solve

$$\text{minimize } \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \text{ where } f(x) = g(x) + \sum_{i=1}^{n}\beta_i f_i(x) + b. \tag{16}$$

For instance, the function class f_i could be linear coordinate functions, that is, $f_i(x) = x_i$. The main difference to (6) is that the resulting optimization problem exhibits a larger number of equality constraint. We obtain (6) with the additional constraints

$$\sum_{j=1}^{m} \alpha_j f_i(x_j) = 0 \text{ for all } i. \tag{17}$$

Linear Programming Regularization Convex function classes with ℓ_1 penalties can be obtained by imposing an $\|\alpha\|_1$ penalty instead of the $\|g\|_{\mathcal{H}}^2$ penalty in the optimization problem. The advantage of this setting is that minimizing

$$\text{minimize } \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \lambda \sum_{j=1}^{n} |\alpha_i| \text{ where } f(x) = \sum_{i=1}^{n} \alpha_i f_i(x) + b. \tag{18}$$

is a *linear program* which can be solved efficiently by existing codes for large scale problems. In the context of (18) the functions f_i constitute the generators of the convex function class. This approach is similar to Koenker et al. (1994) and Bosch et al. (1995). The former discuss ℓ_1 regularization of expansion coefficients whereas the latter discuss an explicit second order smoothing spline method for the purpose of quantile regression. Most of the discussion in the present paper can be adapted to this case without much modification. For details on how to achieve this see Schölkopf and Smola (2002). Note that smoothing splines are a special instance of kernel expansions where one assumes explicit knowledge of the basis functions.

Relevance Vector Regularization and Sparse Coding Finally, for sparse expansions one can use more aggressive penalties on linear function expansions than those given in (18). For instance, we could use a staged regularization as in the RVM (Tipping, 2001), where a quadratic penalty on each coefficient is exerted with a secondary regularization on the penalty itself. This corresponds to a Student-t penalty on α.

Likewise we could use a mix between an ℓ_1 and ℓ_0 regularizer as used in Fung et al. (2002) and apply successive linear approximation. In short, there exists a large number of regularizers, and (non)parametric families which can be used. In this sense the RKHS parameterization is but one possible choice. Even so, we show in Section 5 that QR using the RKHS penalty yields excellent performance in experiments.

Neural Networks, Generalized Models Our method does not depend on the how the function class is represented (not only the Kernelized version), in fact, one can use Neural Networks or Generalized Models for estimation as long as the loss function is kept the same. This is the main reason why this paper is called *Non-parametric quantile estimation*.

4. Theoretical Analysis

In this section we state some performance bounds for our estimator.

4.1 Performance Indicators

We first need to discuss how to evaluate the performance of the estimate f versus the true conditional quantile $\mu_\tau(x)$. Two criteria are important for a good quantile estimator f_τ:

- f_τ needs to satisfy the quantile property as well as possible. That is, we want that

$$\Pr_{X,Y}\{|\Pr\{y < f_\tau(x)\} - \tau| \geq \varepsilon\} \leq \delta. \tag{19}$$

  In other words, we want that the probability that $y < f_\tau(x)$ does not deviate from τ by more than ε with high probability, when viewed over all draws (X,Y) of training data. Note however, that (19) does not imply having a conditional quantile estimator at all. For instance, the constant function based on the unconditional quantile estimator with respect to Y performs extremely well under this criterion. Hence we need a second quantity to assess how closely $f_\tau(x)$ tracks $\mu_\tau(x)$.

- Since μ_τ itself is not available, we take recourse to (3) and the fact that μ_τ is the minimizer of the expected risk $R[f]$. While this will not allow us to compare μ_τ and f_τ directly, we can at least compare it by assessing how close to the minimum $R[f_\tau^*]$ the estimate $R[f_\tau]$ is. Here f_τ^* is the minimizer of $R[f]$ with respect to the chosen function class. Hence we will strive to bound

$$\Pr_{X,Y}\{R[f_\tau] - R[f_\tau^*] > \varepsilon\} \leq \delta. \tag{20}$$

These statements will be given in terms of the Rademacher complexity of the function class of the estimator as well as some properties of the loss function used in select it. The technique itself is standard and we believe that the bounds can be tightened considerably by the use of *localized* Rademacher averages (Mendelson, 2003), or similar tools for empirical processes. However, for the sake of simplicity, we use the tools from Bartlett and Mendelson (2002), as the key point of the derivation is to describe a new setting rather than a new technique.

4.2 Bounding $R[f_\tau^*]$

Definition 4 (Rademacher Complexity) *Let $X := \{x_1,\ldots,x_m\}$ be drawn iid from $p(x)$ and let $\mathcal{F}$ be a class of functions mapping from (X) to $\mathbb{R}$. Let σ_i be independent uniform $\{\pm 1\}$-valued random variables. Then the Rademacher complexity $\mathcal{R}_m$ and its empirical variant $\hat{\mathcal{R}}_m$ are defined as follows:*

$$\hat{\mathcal{R}}_m(\mathcal{F}) := \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{m}\sum_1^n \sigma_i f(x_i)\right| \quad \middle| X\right] \text{ and } \mathcal{R}_m(\mathcal{F}) := \mathbf{E}_X\left[\hat{\mathcal{R}}_m(\mathcal{F})\right]. \tag{21}$$

Conveniently, if Φ is a Lipschitz continuous function with Lipschitz constant L, one can show (Bartlett and Mendelson, 2002) that

$$\mathcal{R}_m(\Phi\circ\mathcal{F}) \leq 2L\mathcal{R}_m(\mathcal{F}) \text{ where } \Phi\circ\mathcal{F} := \{g|g = \phi\circ f \text{ and } f\in\mathcal{F}\}. \tag{22}$$

An analogous result exists for empirical quantities bounding $\hat{\mathcal{R}}_m(\Phi\circ\mathcal{F}) \leq 2L\hat{\mathcal{R}}_m(\mathcal{F})$. The combination of (22) with Bartlett and Mendelson (2002, Theorem 8) yields:

Theorem 5 (Concentration for Lipschitz Continuous Functions) *For any Lipschitz continuous function Φ with Lipschitz constant L and a function class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ and probability measure on $\mathcal{X}$ the following bound holds with probability $1-\delta$ for all draws of X from $\mathcal{X}$:*

$$\sup_{f\in\mathcal{F}}\left|\mathbf{E}_x[\Phi(f(x))] - \frac{1}{m}\sum_{i=1}^m \Phi(f(x_i))\right| \leq 2L\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{8\log 2/\delta}{m}}. \tag{23}$$

We can immediately specialize the theorem to the following statement about the loss for QR:

Theorem 6 *Denote by f_τ^* the minimizer of the $R[f]$ with respect to $f \in \mathcal{F}$. Moreover assume that all $f \in \mathcal{F}$ are uniformly bounded by some constant B. With the conditions listed above for any sample size m and $0 < \delta < 1$, every quantile regression estimate f_τ satisfies with probability at least $(1-\delta)$*

$$R[f_\tau] - R[f_\tau^*] \leq 2\max L\mathcal{R}_m(\mathcal{F}) + (4+LB)\sqrt{\frac{\log 2/\delta}{2m}} \text{ where } L = \{\tau, 1-\tau\}. \tag{24}$$

Proof We use the standard bounding trick that

$$\begin{aligned} R[f_\tau] - R[f_\tau^*] &\leq \left|R[f_\tau] - R_{\text{emp}}[f_\tau]\right| + R_{\text{emp}}[f_\tau^*] - R[f_\tau^*] && (25)\\ &\leq \sup_{f\in\mathcal{F}} \left|R[f] - R_{\text{emp}}[f]\right| + R_{\text{emp}}[f_\tau^*] - R[f_\tau^*] && (26)\end{aligned}$$

where (25) follows from $R_{\text{emp}}[f_\tau] \leq R_{\text{emp}}[f_\tau^*]$. The first term can be bounded directly by Theorem 5. For the second part we use Hoeffding's bound (Hoeffding, 1963) which states that the deviation between a bounded random variable and its expectation is bounded by $B\sqrt{\frac{\log 1/\delta}{2m}}$ with probability δ. Applying a union bound argument for the two terms with probabilities $2\delta/3$ and $\delta/3$ yields the confidence-dependent term. Finally, using the fact that l_τ is Lipschitz continuous with $L = \max(\tau, 1-\tau)$ completes the proof. ■

Example Assume that $\mathcal{H}$ is an RKHS with radial basis function kernel k for which $k(x,x) = 1$. Moreover assume that for all $f \in \mathcal{F}$ we have $\|f\|_{\mathcal{H}} \leq C$. In this case it follows from Mendelson (2003) that $\mathcal{R}_m(\mathcal{F}) \leq \frac{2C}{\sqrt{m}}$. This means that the bounds of Theorem 6 translate into a rate of convergence of

$$R[f_\tau] - R[f_\tau^*] = O(m^{-\frac{1}{2}}). \tag{27}$$

This is as good as it gets for nonlocalized estimates. Since we do not expect $R[f]$ to vanish except for pathological applications where quantile regression is inappropriate (that is, cases where we have a deterministic dependency between y and x), the use of localized estimates (Bartlett et al., 2002) provides only limited returns. We believe, however, that the constants in the bounds could benefit from considerable improvement.

4.3 Bounds on the Quantile Property

The theorem of the previous section gave us some idea about how far the sample average quantile loss is from its true value under p. We now proceed to stating bounds to which degree f_τ satisfies the quantile property, i.e. (19).

In this view (19) is concerned with the deviation $\mathbf{E}\left[\chi_{(-\infty,0]}(y - f_\tau(x))\right] - \tau$. Unfortunately $\chi_{(-\infty,0]} \circ \mathcal{F}$ is not scale dependent. In other words, small changes in $f_\tau(x)$ around the point $y = f_\tau(x)$ can have large impact on (19). One solution for this problem is to use an artificial margin ε and ramp functions $r_\varepsilon^+, r_\varepsilon^-$ as defined in (28) and Figure 7. These functions are Lipschitz continuous with constant $L = 1/\varepsilon$. This leads to:

$$r_\varepsilon^+(\xi) := \min\{1, \max\{0, 1-\xi/\varepsilon\}\} \quad (28a)$$
$$r_\varepsilon^-(\xi) := \min\{1, \max\{0, -\xi/\varepsilon\}\} \quad (28b)$$

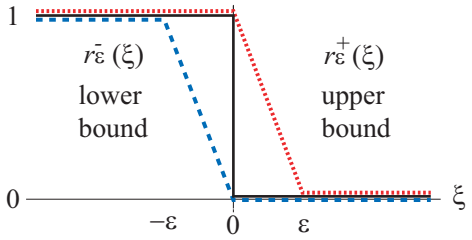


Figure 7: Ramp functions bracketing the characteristic function via $r_\varepsilon^+ \geq \chi_{(-\infty,0]} \geq r_\varepsilon^-$.

Theorem 7 *Under the assumptions of Theorem 6 the expected quantile is bounded with probability* $1-\delta$ *each from above and below by*

$$\frac{1}{m}\sum_{i=1}^m r_\varepsilon^-(y_i - f(x_i)) - \Delta \leq \mathbf{E}\left[\chi_{(-\infty,0]}(y - f_\tau(x))\right] \leq \frac{1}{m}\sum_{i=1}^m r_\varepsilon^+(y_i - f(x_i)) + \Delta, \quad (29)$$

where the statistical confidence term is given by $\Delta = \frac{2}{\varepsilon}\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{-8\log\delta}{m}}$.

Proof The claim follows directly from Theorem 5 and the Lipschitz continuity of r_ε^+ and r_ε^-. Note that r_ε^+ and r_ε^- minorize and majorize $\xi_{(-\infty,0]}$, which bounds the expectations. Next use a Rademacher bound on the class of loss functions induced by $r_\varepsilon^+ \circ \mathcal{F}$ and $r_\varepsilon^- \circ \mathcal{F}$ and note that the ramp loss has Lipschitz constant $L = 1/\varepsilon$. Finally apply the union bound on upper and lower deviations. ■

Note that Theorem 7 allows for some flexibility: we can decide to use a very conservative bound in terms of ε, i.e. a large value of ε to reap the benefits of having a ramp function with small L. This leads to a lower bound on the Rademacher average of the induced function class. Likewise, a small ε amounts to a potentially tight approximation of the empirical quantile, while risking loose statistical confidence terms.

5. Experiments

The present section mirrors the theoretical analysis of the previous section.

5.1 Experiments with Standard Nonparametric Quantile Regression

We check the performance of various quntile estimators with respect to two criteria:

- Expected risk with respect to the ℓ_τ loss function. Since computing the true conditional quantile is impossible and all approximations of the latter rely on intermediate density estimation, this is the only objective criterion we could find. We denote this loss measure as *pinball loss*.

- Simultaneously we need to ensure that the estimate satisfies the quantile property, that is, we want to ensure that the estimator we obtained does indeed produce numbers $f_\tau(x)$ which exceed y with probability close to τ. The quantile property was measured by *ramp loss*.[4]

5.1.1 MODELS

We compare the following four models:

- An unconditional quantile estimator. Given the simplicity of the function class (constants!) this model should tend to underperform all other estimates in terms of minimizing the empirical risk. By the same token, it should perform best in terms of preserving the quantile property. This appears as `uncond`.

- Linear QR as described in Koenker and Bassett (1978). This uses a linear unregularized model to minimize l_τ. In experiments, we used the `rq` routine available in the *R* package[5] called `quantreg`. This appears as `linear`.

- Nonparametric QR as described by Koenker et al. (1994). This uses a spline model for each coordinate individually, with linear effect. The fitting routine used was `rqss`, also available in `quantreg`.[6] The regularization parameter in this model was chosen by 10-fold cross-validation within the training sample. This appears as `rqss`.

- Nonparametric quantile regression as described in Section 2. We used Gaussian RBF kernels with automatic kernel width (ω^2) and regularization (C) adjustment by 10-fold cross-validation within training sample.[7] This appears as `npqr`.

As we increase the complexity of the function class (from constant to linear to nonparametric) we expect that (subject to good capacity control) the expected risk will decrease. Simultaneously we expect that the quantile property becomes less and less maintained, as the function class grows. This is exactly what one would expect from Theorems 6 and 7. As the experiments show, performance of the `npqr` method is comparable or significantly better than other models. In particular it preserves the quantile property well.

Notes on Gaussian RBF kernel parameter selection trick The parameter σ in the Gaussian kernel could be chosen by the following trick. We fist subsample the training data (if the training data set is not large, use the whole training data), then compute the distance between the points and find the distances at 0.9 and 0.1 quantile of all the distances, the average distance of these two distances is set to be the initial σ_0. This is to guarantee that the kernel parameter is neither too big or too small. Other values of σ to be selected in the experiments (via cross-validation) are $[10^{-4}\sigma_0, \ldots, \sigma_0, \ldots, 10^3\sigma_0, 10^4\sigma_0]$. In general, depending on the problems, one may set the search space to be finer (the distance between two consecutive items in the list is smaller) or coarser (the distance between two consecutive items in the list is larger), or even a higher value for maximum item in the list, and a smaller value for minimum item in the list, etc.

4. In the experiments we set $\varepsilon = 0$ in (28) for simplicity. Thus, it might be appropriate to call it as *step loss* rather than ramp loss. However, we keep to use the term "ramp loss" throughout this paper.
5. See http://cran.r-project.org/.
6. Additional code containing bugfixes and other operations necessary to carry out our experiments is available at http://users.rsise.anu.edu.au/∼timsears.
7. Code will be available as part of the CREST toolbox for research purposes.

5.1.2 DATA SETS

We chose 20 regression data sets from the following R packages: `mlbench, quantreg, alr3` and `MASS`. The first library contains data sets from the UCI repository. The last two were made available as illustrations for regression textbooks. The data sets are all documented and available in *R*. Data sets were chosen not to have any missing variables, to have suitable datatypes, and to be of a size where all models would run on them.[8] In most cases either there was an obvious variable of interest, which was selected as the *y*-variable, or else we chose a continuous variable arbitrarily. The sample sizes vary from $m = 38$ (CobarOre) to $m = 1375$ (heights), and the number of regressors vary from $d = 1$ (5 sets) and $d = 12$ (BostonHousing). Some of the data sets contain categorical variables. We omitted variables which were effectively record identifiers, or obviously produced very small groupings of records. Finally, we *standardized* all data sets coordinatwise to have zero mean and unit variance before running the algorithms. This had a side benefit of putting the pinball loss on similar scale for comparison purposes.

Data Set	Sample Size	No. Regressors (x)	Y Var.	Dropped Vars.
caution	100	2	y	-
ftcollinssnow	93	1	Late	YR1
highway	39	11	Rate	-
heights	1375	1	Dheight	-
sniffer	125	4	Y	-
snowgeese	45	4	photo	-
ufc	372	4	Height	-
birthwt	189	7	bwt	ftv, low
crabs	200	6	CW	index
GAGurine	314	1	GAG	-
geyser	299	1	waiting	-
gilgais	365	8	e80	-
topo	52	2	z	-
BostonHousing	506	13	medv	-
CobarOre	38	2	z	-
engel	235	1	y	-
mcycle	133	1	accel	-
BigMac2003	69	9	BigMac	City
UN3	125	6	Purban	Locality
cpus	209	7	estperf	name

Table 1: Data Set facts

8. The last requirement, using `rqss` proved to be challenging. The underlying spline routines do not allow extrapolation beyond the previously seen range of a coordinate, only permitting interpolation. This does not prevent fitting, but does randomly prevent forecasting on unseen examples, which was part of our performance metric.

5.1.3 RESULTS

We tested the performance of the **4** models. For each model we used 10-fold cross-validation to assess the confidence of our results. As mentioned above, a regularization parameter in `rqss` and ω^2 and C in `npqr` were automatically chosen by 10-fold cross-validation **within** the training sample, i.e. we used *nested* cross-validation. To compare across all four models we measured both *pinball loss* and *ramp loss*. The 20 data sets and three different quantile levels ($\tau \in \{0.1, 0.5, 0.9\}$) yield 60 trials for each model. The full results are shown in Appendix B. In summary, we conclude as follows:

- In terms of *pinball loss*, the performance of our `npqr` were comparable or better than other three models.

  `npqr` performed significantly better than other three models in 14 of the 60 trials, while `rqss` performed significantly better than other three models in only one of the 60 trials. In the rest of 45 trials, no single model performed significantly better then the others. All these statements are based on the two-sided paired-sample t-test with significance level 0.05. We got similar but a bit less conservative results by (nonparametric) Wilcoxon signed rank test.

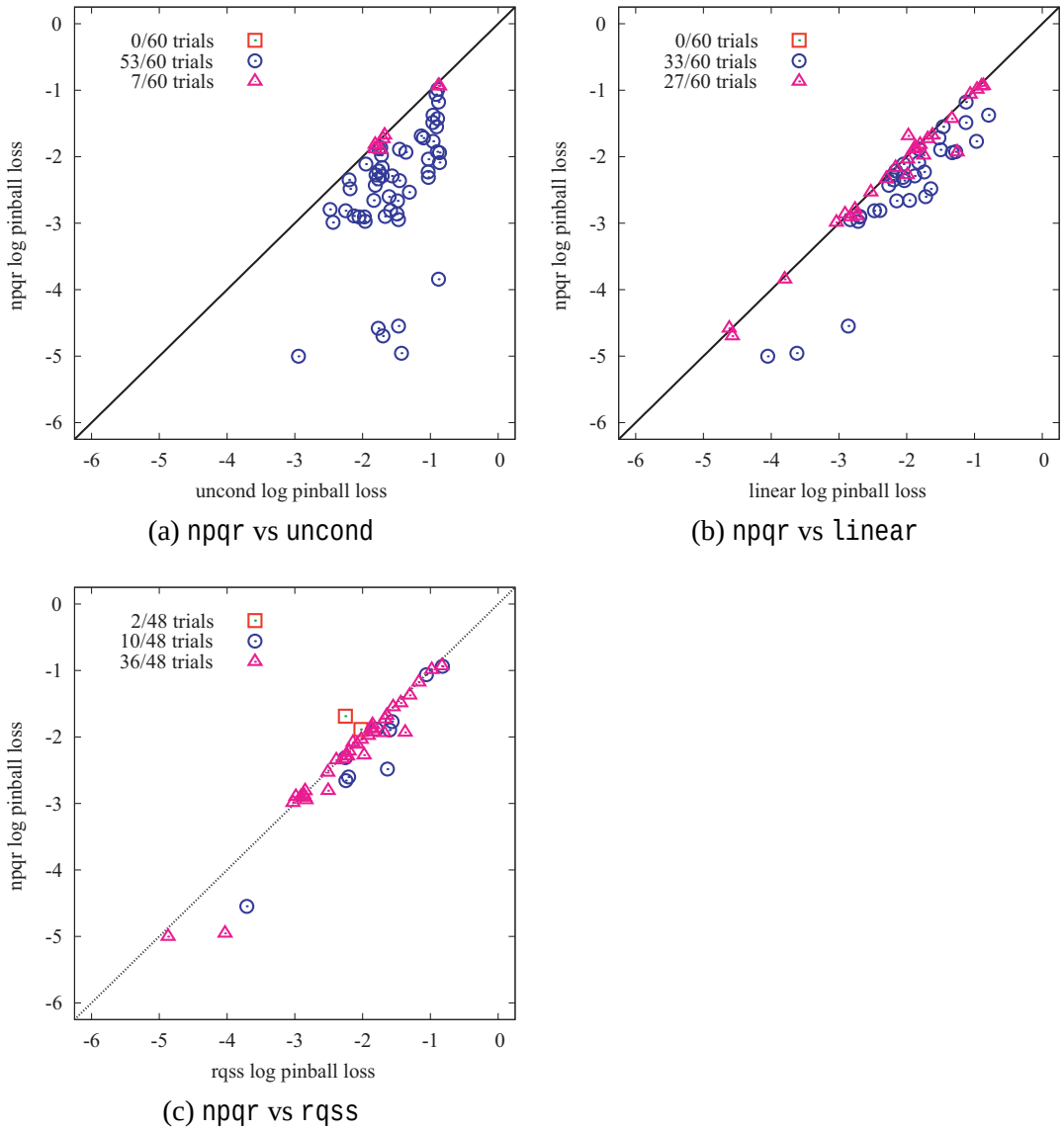
  Figure 8 depicts the comparison of `npqr` performance with each of `uncond`, `linear` and `rqss` models. Each of three plots contain 60 points corresponding to 60 trials (3 different τs times 20 data sets).[9] The vertical axis indicates the log pinball losses of `npqr` and the horizontal axis indicates those of the alternative. The points under (over) the 45 degree line means that the `npqr` was better (worse) than the alternative. Circles (squares) indicate that `npqr` was significantly better (worse) than the alternative at 0.05 significance level in paired-sample t-test, while triangles indicate no significant difference.

- In terms of *ramp loss* (quantile property), the performance of our `npqr` were comparable to other three models for intermediate quantile ($\tau = 0.5$). All four models produced ramp losses close to the desired quantile, although flexible nonparametric models `rqss` and `npqr` were noisier in this regard. When $\tau = 0.5$, the number of $f_\tau(x)$ which exceed y did NOT deviate significantly from the binomial distribution $B(\text{ sample size }, \tau)$ in all 20 data sets.

  On the other hand, for extreme quantiles ($\tau = 0.1$ and 0.9), `rqss` and `npqr` showed a small but significant bias towards the median in a few trials. We conjecture that this bias is related to the problem of *data piling* (Hall et al., 2005). See section 6 for the discussion.

  Note that the quantile property, as such, is not informative measure for *conditional* quantile estimation. It merely measures *unconditional* quantile estimation performances. For example, `uncond`, the constant function based on the unconditional quantile estimator with respect to Y (straightforwardly obtained by sorting $\{y_i\}_{i=1}^m$ without using $\{x_i\}_{i=1}^m$ at all), performed best under this criterion. It is clear that the less flexible model would have the better quantile property, but it does not necessarily mean that those less flexible ones are better for conditional quantile functions.

9. In the comparison between `npqr` and `rqss`, 48 trials were examined since in the other 12 trials `rqss` was unable to produce estimates, due to its construction of the function system.



(a) `npqr` vs `uncond` (b) `npqr` vs `linear`

(c) `npqr` vs `rqss`

Figure 8: Log-log plots of out-of-sample performances. The plots show `npqr` versus (a) `uncond`, (b) `linear` and (c) `rqss`; combining the average pinball losses of all 60 trials (3 quantiles times 20 data sets). The points under (over) the 45 degree line means that the `npqr` was better (worse) than the alternative. Circle (squares) indicate that `npqr` was significantly better (worse) than the alternative at 0.05 significance level in paired-sample t-test, while triangles indicate no significant difference.

5.2 Experiments on Nonparametric Quantile Regression with Additional Constraints

We empirically investigate the performances of nonparametric quantile regression estimator with the additional constraints described in section 3. Imposing constraints is one way to introduce the prior knowledge on the data set being analyzed. Although additional constraints always increase training errors, we will see that these constraints can sometimes reduce test errors. The full results are shown in Appendix B.

5.2.1 NON-CROSSING CONSTRAINTS

First we look at the effect of non-crossing constraints on the generalization performances. We used the same 20 data sets mentioned in the previous subsection. We denote the `npqrs` trained with non-crossing constraints as `noncross` and `npqr` indicates standard one here. We made comparisons between `npqr` and `noncross` with $\tau \in \{0.1, 0.5, 0.9\}$. The results for `noncross` with $\tau = 0.1$ were obtained by training a pair of non-crossing models with $\tau = 0.1$ and 0.2. The results with $\tau = 0.5$ were obtained by training three non-crossing models with $\tau = 0.4$, 0.5 and 0.6. The results with $\tau = 0.9$ were obtained by training a pair of non-crossing models with $\tau = 0.8$ and 0.9. In this experiment, we simply impose non-crossing constraints only at a single test point to be evaluated. The kernel width and smoothing parameter were always set to be the selected ones in the above standard `npqr` experiments. The confidences were assessed by 10-fold cross-validation in the same way as the previous section. The complete results are found in the tables in Appendix B. The performances of `npqr` and `noncross` are quite similar since `npqr` itself could produce *almost* non-crossing estimates and the constraints only make a *small* adjustments only when there happen to be the violations.

5.2.2 MONOTONICITY CONSTRAINTS

We compare two models:

- Nonparametric QR as described in Section 2 (`npqr`).
- Nonparametric QR with monotonicity constraints as described in Section 3.2 (`npqrm`).

We use two data sets:

- The *cars* data set as described in Mammen et al. (2001). Fuel efficiency (in miles per gallon) is studied as a function of engine output.
- The *onions* data set as described in Ruppert and Carroll (2003). log(Yield) is studied as a function of density, we use only the measurements taken at Purnong Landing.

We tested the performance of the two methods on 3 different quantiles ($\tau \in \{0.1, 0.5, 0.9\}$). In the experiments with *cars*, we noticed that the data is not truly monotonic. This is because, smaller engines may correspond to cheap cars and thus may not be very efficient. Monotonic models (`npqrm`) tend to do worse than standard models (`npqr`) for lower quantiles. With higher quantiles, `npqrm` tends to do better than the standard `npqr`. For the `onions` data set, as the data is truly monotonic, the `npqrm` does better than the standard `npqr` in terms of the pinball loss.

6. Discussion and Extensions

Frequently in the literature of regression, including quantile regression, we encounter the term "exploratory data analysis". This is meant to describe a phase before the user has settled on a "model", after which some statistical tests are performed, justifying the choice of the model. Quantile regression, which allows the user to highlight many aspects of the distribution, is indeed a useful tool for this type of analysis. We also note that no attempts at statistical modeling beyond automatic parameter choice via cross-validation, were made to tune the results. So the effort here stays true to that spirit, yet may provide useful estimates immediately.

In the Machine Learning literature the emphasis is more on short circuiting the modeling process. Here the two approaches are complementary. While not completely model-free, the experience of building the models in this paper shows how easy it is to estimate the quantities of interest in QR, with little of the angst of model selection, thanks to regularization. It is interesting to consider whether kernel methods, with regularization, can blur the distinction between model building and data exploration in statistics.

In summary, we have presented a Quadratic Programming method for estimating quantiles which bests the state of the art in statistics. It is easy to implement, comes with uniform convergence results and experimental evidence for its soundness. We also introduce non-crossing and monotonicity constraints as extensions to avoid some undesirable behaviors in some circumstances.

Overly Optimistic Estimates for Ramp Loss The experiments show us that the there is a bias towards the median in terms of the ramp loss. For example, if we run a quantile estimator with $\tau = 0.05$, then we will not necessarily get the empirical quantile is also at 0.05 but more likely to be at 0.08 or higher. Likewise, the empirical quantile will be 0.93 or lower if the estimator is run at 0.9. This affects all estimators, using the pinball loss as the loss function, not just the kernel version.

This is because the algorithm tends to aggressively push a number of points to the kink in the training set, these points may then be miscounted (see Lemma 3). The main reason behind it is that the extreme quantiles tend to be less smooth, the regularizer will therefore makes sure we get a simpler model by biasing towards the median (which is usually simpler). However, in the test set it is very unlikely to get the points lying exactly at the kink. Figure 9 shows us there is a linear relationship between the fraction of points at and below the kink (for low quantiles) and below the kink (for higher quantiles) with the empirical ramp loss.

Accordingly, in order to get a better performance in terms of the ramp loss, we just estimate the quantiles, and if they turn out to be too optimistic on the training set, we use a slightly lower (for $\tau < 0.5$) or higher (for $\tau > 0.5$) value of τ until we have exactly the right quantity.

The fact that there is a number of points sitting exactly on the kink (quantile regression - this paper), the edge of the tube (ν-SVR - see Schölkopf et al., 2000), or the supporting hyperplane (single-class problems and novelty detection - see Schölkopf et al., 1999) might affect the overall accuracy control in the test set. This issue deserves further scrutiny.

Estimation with constraints We introduce non-crossing and monotonicity constraints in the context of nonparametric quantile regression. However, as discussed in Mammen et al. (2001), other constraints can also be applied very similiarly to the constraints described in this paper but might be in different estimation contexts. Here are some variations (we just give directions for the first two, the rest can be applied in the same manner)

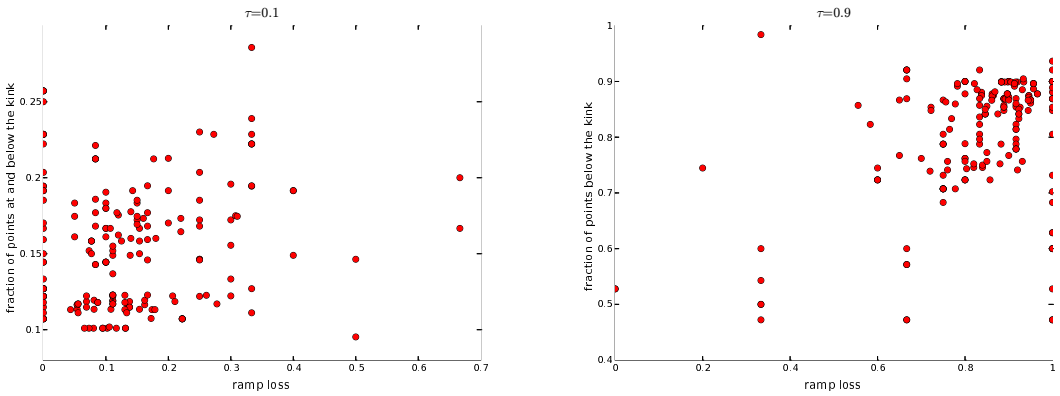


Figure 9: Illustration of the relationship between quantile in training and ramp loss.

- *Bivariate extreme-value distributions*. Hall and Tajvidi (2000) propose methods to estimate the dependence function of a bivariate extreme-value distribution. They require to estimate a **convex** function f such that $f(0) = f(1) = 1$ and $f(x) \geq \max(x, 1-x)$ for $x \in [0,1]$. We can also apply this approach to our method as to the monotonicity constraint, all we have to do is to ensure $\langle \phi(0), w \rangle + b = \langle \phi(1), w \rangle + b = 1$, $\langle \phi''(x), w \rangle \geq 0$ and $\langle \phi(x), w \rangle + b \geq \max(x, 1-x)$ for $x \in [0,1]$.

- *Positivity constraints*. The regression function is positive. In this case, we must ensure $\langle \phi(x), w \rangle + b > 0$, $\forall x$.

- *Boundary conditions*. The regression function is defined in $[a,b]$ and assumed to be v at the boundary point a or b.

- *Additive models with monotone components*. The regression function $f : \mathbb{R}^n \to \mathbb{R}$ is of additive form $f(x_1, ..., x_n) = f_1(x_1) + ... + f_n(x_n)$ where each additive component f_i is monotonic.

- *Observed deriatives*. Assume that m samples are observed corresponding with m regression functions. Now, the constraint is that f_j coincides with the derivative of f_{j-1} (same notation with last point) (Cox, 1988).

Future Work Quantile regression has been mainly used as a data analysis tool to assess the influence of individual variables. This is an area where we expect that nonparametric estimates will lead to better performance.

Being able to estimate an upper bound on a random variable $y|x$ which hold with probability τ is useful when it comes to determining the so-called Value at Risk of a portfolio. Note, however, that in this situation we want to be able to estimate the regression quantile for a large set of different portfolios. For example, an investor may try to optimize their portfolio allocation to maximize return while keeping risk within a constant bound. Such uniform statements will need further analysis if we are to perform nonparametric estimates. We need more efficient optimization algorithm for non-crossing constraints since we have to work with $o(nm)$ dual variables. Simple SVM (Vishwanathan et al., 2003) would be a promising candidate for this purpose.

Acknowledgments

National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported by grants of the ARC, by the Pascal Network of Excellence and by Japanese Grants-in-Aid for Scientific Research 16700258. We thank Roger Koenker for providing us with the latest version of the *R* package `quantreg`, and for technical advice. We thank Shahar Mendelson and Bob Williamson for useful discussions and suggestions. We also thank the anonymous reviewers for valuable feedback.

Appendix A. Nonparametric ν-Support Vector Regression

In this section we explore an alternative to the quantile regression framework proposed in Section 2. It derives from Schölkopf et al. (2000). There the authors suggest a method for adapting SV regression and classification estimates such that automatically only a quantile ν lies beyond the desired confidence region. In particular, if $p(y|x)$ can be modeled by additive noise of equal degree (i.e. $y = f(x) + \xi$ where ξ is a random variable independent of x) Schölkopf et al. (2000) show that the ν-SV regression estimate does converge to a quantile estimate.

A.1 Heteroscedastic Regression

Whenever the above assumption on $p(y|x)$ is violated ν-SVR will not perform as desired. This problem can be amended as follows: one needs to turn the margin $\varepsilon(x)$ into a nonparametric estimate itself. This means that we solve the following optimization problem.

$$\begin{aligned}
\underset{\theta_1,\theta_2,b,\varepsilon}{\text{minimize}} \quad & \frac{\lambda_1}{2}\|\theta_1\|^2 + \frac{\lambda_2}{2}\|\theta_2\|^2 + \sum_{i=1}^{m}(\xi_i + \xi_i^*) - \nu m\varepsilon && \text{(30a)}\\
\text{subject to} \quad & \langle\phi_1(x_i),\theta_1\rangle + b - y_i \le \varepsilon + \langle\phi_2(x_i),\theta_2\rangle + \xi_i && \text{(30b)}\\
& y_i - \langle\phi_1(x_i),\theta_1\rangle - b \le \varepsilon + \langle\phi_2(x_i),\theta_2\rangle + \xi_i^* && \text{(30c)}\\
& \xi_i, \xi_i^* \ge 0 && \text{(30d)}
\end{aligned}$$

Here ϕ_1, ϕ_2 are feature maps, θ_1, θ_2 are corresponding parameters, ξ_i, ξ_i^* are slack variables and b, ε are scalars. The key difference to the heteroscedastic estimation problem described in Schölkopf et al. (2000) is that in the latter the authors assume that the specific form of the noise is *known*. In (30) instead, we make no such assumption and instead we estimate $\varepsilon(x)$ as $\langle\phi_2(x),\theta_2\rangle + \varepsilon$.

One may check that the dual of (30) is obtained by

$$\begin{aligned}
\underset{\alpha,\alpha^*}{\text{minimize}} \quad & \frac{1}{2\lambda_1}(\alpha-\alpha^*)^\top K_1(\alpha-\alpha^*) + \frac{1}{2\lambda_2}(\alpha+\alpha^*)^\top K_1(\alpha+\alpha^*) + (\alpha-\alpha^*)^\top y && \text{(31a)}\\
\text{subject to} \quad & \vec{1}^\top(\alpha-\alpha^*) = 0 && \text{(31b)}\\
& \vec{1}^\top(\alpha+\alpha^*) = Cm\nu && \text{(31c)}\\
& 0 \le \alpha_i, \alpha_i^* \le 1 \text{ for all } 1 \le i \le m && \text{(31d)}
\end{aligned}$$

Here K_1, K_2 are kernel matrices where $[K_i]_{jl} = k_i(x_j, x_l)$ and $\vec{1}$ denotes the vector of ones. Moreover, we have the usual kernel expansion, this time for the regression $f(x)$ and the margin $\varepsilon(x)$ via

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\,k_1(x_i,x) + b \text{ and } \varepsilon(x) = \sum_{i=1}^{m}(\alpha_i + \alpha_i^*)\,k_2(x_i,x) + \varepsilon. \tag{32}$$

The scalars b and ε can be computed conveniently as dual variables of (31) when solving the problem with an interior point code (see Schölkopf and Smola, 2002, for more details).

A.2 The ν-Property

As in the parametric case also (30) has the ν-property. However, it is worth noting that the solution $\varepsilon(x)$ need not be positive throughout unless we change the optimization problem slightly by imposing a nonnegativity constraint on ε. The following theorem makes this reasoning more precise:

Theorem 8 *The minimizer of (30) satisfies*

1. *The fraction of points for which $|y_i - f(x_i)| < \varepsilon(x_i)$ is bounded by $1-\nu$.*
2. *The fraction of constraints (30b) and (30c) with $\xi_i > 0$ or $\xi_i^* > 0$ is bounded from above by ν.*
3. *If (x,y) is drawn iid from a distribution $\Pr(x,y)$, with $\Pr(y|x)$ continuous and the expectation of the modulus of absolute continuity of its density satisfying $\lim_{\delta\to 0}\mathbf{E}[\varepsilon(\delta)] = 0$. With probability 1, asymptotically, the fraction of points satisfying $|y_i - f(x_i)| = \varepsilon(x_i)$ converges to 0.*

Moreover, imposing $\varepsilon \geq 0$ is equivalent to relaxing (31c) to $\vec{1}^\top(\alpha - \alpha^*) \leq Cm\nu$. If in addition K_2 has only nonnegative entries then also $\varepsilon(x) \geq 0$ for all x_i.

Proof The proof is essentially similar to that of Lemma 3 and Schölkopf et al. (2000). However note that the flexibility in ε and potential $\varepsilon(x) < 0$ lead to additional complications. However, if both f and $\varepsilon(x)$ have well behaved entropy numbers, then also $f \pm \varepsilon$ are well behaved.

To see the last set of claims note that the constraint $\vec{1}^\top(\alpha - \alpha^*) \leq Cm\nu$ is obtained again directly from dualization via the condition $\varepsilon \geq 0$. Since $\alpha_i, \alpha_i^* \geq 0$ for all i it follows that $\varepsilon(x)$ contains only nonnegative coefficients, which proves the last part of the claim. ■

Note that in principle we could enforce $\varepsilon(x_i) \geq 0$ for all x_i. This way, however, we would lose the ν-property and add even more complication to the optimization problem. A third set of Lagrange multipliers would have to be added to the optimization problem.

A.3 An Example

The above derivation begs the question why one should not use (31) instead of (6) for the purpose of quantile regression. After all, both estimators yield an estimate for the upper and lower quantiles.

Firstly, the combined approach is numerically more costly as it requires optimization over twice the number of parameters, albeit at the distinct advantage of a sparse solution, whereas (6) always leads to a dense solution.

The key difference, however, is that (31) is prone to producing estimates where the margin $\varepsilon(x) < 0$. While such a solution is clearly unreasonable, it occurs whenever the margin is rather small and the overall tradeoff of simple f vs. simple ε yields an advantage by keeping f simple. With enough data this effect vanishes, however, it occurs quite frequently, even with supposedly distant quantiles, as can be seen in Figure 10.

In addition, the latter suffers from the assumption that the error be symmetrically distributed. In other words, if we are just interested in obtaining the 0.95 quantile estimate we end up estimating

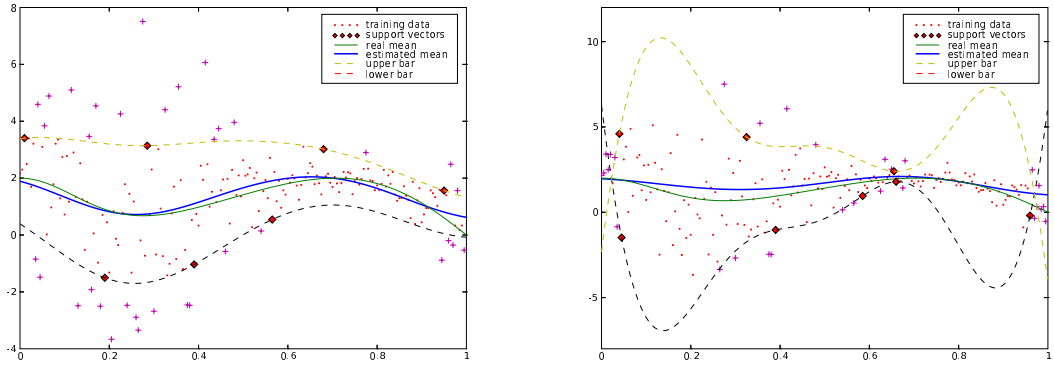


Figure 10: Illustration of the heteroscedastic SVM regression on artificial data set generated from (1) with $f(x) = \sin \pi x$ and $\sigma(x) = \exp(\sin 2\pi x)$. On the left, $\lambda_1 = 1$, $\lambda_2 = 10$ and $\nu = 0.2$, the algorithm successfully regresses the data. On the right, $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\nu = 0.2$, the algorithm fails to regress the data as ε becomes negative.

the 0.05 quantile on the way. In addition to that, we make the assumption that the additive noise is symmetric.

We produced this derivation and experiments mainly to make the point that the adaptive margin approach of Schölkopf et al. (2000) is insufficient to address the problems posed by quantile regression. We found empirically that it is much easier to adjust QR instead of the symmetric variant.

In summary, the symmetric approach is probably useful only for parametric estimates where the number of parameters is small and where the expansion coefficients ensure that $\varepsilon(x) \geq 0$ for all x.

Appendix B. Experimental Results

In this appendix, we show the detail results on the experiments.

B.1 Standard Nonparametric Quantile Regression

Here we assemble six tables to display the comparisons among four models, `uncond`, `linear`, `rqss` and `npqr`. Each table represents *pinball loss* or *ramp loss* for each of $\tau = 0.1$, 0.5 and 0.9 cases.

	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Pinball Loss	Table 2	Table 4	Table 6
Ramp Loss	Table 3	Table 5	Table 7

Tables 2, 4, and 6 show the average pinball loss for each data set. A lower figure is preferred in each case. The bold figures indicate the best (smallest) performances. The circles '$\circ$' indicate that the difference from the second best model were statistically significant at 0.05 level with two-sided paired-sample t-test. NA denotes cases where rqss (Koenker et al., 1994) was unable to produce estimates, due to its construction of the function system.

Tables 3, 5 and 7, show the ramp loss, a measure for quantile property. In each table a figure close to the intended quantile (10, 50 or 90) is preferred. The figures in round brackets denote the

p-values under the null-hypothesis that the ramp loss, i.e. the number of test points (x,y) such that $y < f_\tau(x)$, is a sample from binomial distribution $B(\text{ sample size}, \tau)$. The bold figures indicate the best (closest to the intended quantile τ) performances. The bullets '$\bullet$' indicate that the ramp loss deviate significantly from binomial distribution $B(\text{ sample size}, \tau)$.

B.2 Nonparametric Quantile Regression with Constraints

Next, we show the results on constrained nonparameteric quantile regression.

B.2.1 NON-CROSSING CONSTRAINTS

Table 8 shows the average pinball loss comparison between the nonparametric quantile regression without (`npqr`) and with (`noncross`) non-crossing constraints. The bold figures indicate the better (smaller) performances The circles '$\circ$' indicate that the difference were statistically significant at 0.05 level with two-sided paired-sample t-test.

Table 9 shows the ramp loss, a measure for quantile property, of `npqr` and `noncross`. The figures in round brackets denote the p-values under the null-hypothesis that the ramp loss, i.e. the number of test points (x,y) such that $y < f_\tau(x)$, is a sample from binomial distribution $B(\text{ sample size}, \tau)$. The bold figures indicate the better (closeer to the intended quantile τ) performances. The bullets '$\bullet$' indicate that the ramp loss deviated significantly from binomial distribution $B(\text{ sample size}, \tau)$.

B.2.2 MONOTONICITY CONSTRAINTS

We tested on the cars and the onions data set for monotonicity with respect to engine size and diameter respectively. Note that on the engines data set the monotonicity constraint is not perfectly satisfied. Table 10 shows the average pinball loss comparison between the nonparametric quantile regression without (`npqr`) and with (`npqrm`) monotonicity constraints. See above for the notation of the table. Table 11 shows the ramp loss, a measure for quantile property, of `npqr` and `npqrm`. See above for the notation of the table.

data set	uncond	linear	rqss	npqr
caution	11.09 ± 0.95	11.18 ± 1.04	**9.18 ± 0.93**	9.56 ± 0.92
ftcollinssnow	16.28 ± 1.18	16.48 ± 1.19	**15.68 ± 1.33**	16.24 ± 1.17
highway	11.27 ± 1.48	19.32 ± 5.11	19.51 ± 4.44	∘ **8.34 ± 1.18**
heights	17.20 ± 0.44	15.28 ± 0.39	15.27 ± 0.40	**15.26 ± 0.39**
sniffer	13.92 ± 0.99	6.78 ± 0.68	**5.44 ± 0.58**	5.48 ± 0.64
snowgeese	8.74 ± 1.44	**4.79 ± 0.89**	4.85 ± 0.90	5.03 ± 0.87
ufc	17.06 ± 0.72	10.02 ± 0.42	10.11 ± 0.44	**9.70 ± 0.42**
birthwt	18.29 ± 1.39	18.44 ± 1.24	18.85 ± 1.28	**17.68 ± 1.16**
crabs	18.27 ± 0.97	1.03 ± 0.08	NA	**0.91 ± 0.07**
GAGurine	10.53 ± 0.55	8.39 ± 0.41	**5.79 ± 0.43**	6.00 ± 0.63
geyser	17.15 ± 0.52	11.50 ± 0.49	11.10 ± 0.49	**10.91 ± 0.49**
gilgais	12.84 ± 0.49	5.93 ± 0.40	5.75 ± 0.44	**5.46 ± 0.35**
topo	20.41 ± 2.45	9.12 ± 1.32	8.15 ± 1.30	**6.03 ± 0.91**
BostonHousing	14.05 ± 0.56	6.60 ± 0.34	NA	∘ **5.10 ± 0.42**
CobarOre	17.88 ± 2.28	17.36 ± 1.97	14.71 ± 2.20	**13.80 ± 2.70**
engel	11.92 ± 0.65	6.49 ± 0.79	5.68 ± 0.45	**5.55 ± 0.37**
mcycle	19.99 ± 0.86	17.87 ± 0.98	10.98 ± 0.66	∘ **7.39 ± 0.90**
BigMac2003	8.37 ± 1.17	6.31 ± 0.95	NA	**6.13 ± 0.96**
UN3	18.02 ± 1.06	**11.47 ± 0.97**	NA	**11.47 ± 1.02**
cpus	5.25 ± 0.69	1.74 ± 0.34	0.77 ± 0.18	**0.67 ± 0.23**

Table 2: Method Comparison: Pinball Loss ($\times 100$, $\tau = 0.1$)

data set	uncond	linear	rqss	npqr
caution	**11.00** (**0.59**)	12.00 (0.40)	• 16.00 (0.04)	12.00 (0.40)
ftcollinssnow	**10.00** (**0.91**)	11.10 (0.65)	12.20 (0.44)	12.20 (0.44)
highway	**10.80** (**0.70**)	• 20.00 (0.03)	• 26.70 (0.00)	• 20.00 (0.03)
heights	9.60 (0.66)	**10.00** (**0.92**)	**10.00** (**0.92**)	**10.00** (**0.92**)
sniffer	7.80 (0.57)	13.70 (0.15)	**12.00** (**0.37**)	• 15.90 (0.02)
snowgeese	12.50 (0.32)	**9.70** (**0.95**)	**9.70** (**0.95**)	13.60 (0.32)
ufc	9.70 (0.92)	**9.90** (**0.94**)	11.80 (0.21)	10.50 (0.68)
birthwt	**10.00** (**0.86**)	12.00 (0.27)	12.60 (0.18)	11.60 (0.38)
crabs	**10.00** (**0.88**)	12.00 (0.29)	NA	13.30 (0.09)
GAGurine	10.40 (0.68)	**9.90** (**0.96**)	10.70 (0.55)	12.10 (0.19)
geyser	**9.70** (**0.96**)	11.20 (0.48)	10.70 (0.60)	12.20 (0.21)
gilgais	9.50 (0.88)	**10.40** (**0.71**)	• 13.50 (0.03)	12.40 (0.12)
topo	**8.90** (**0.84**)	13.40 (0.29)	16.00 (0.14)	• 19.40 (0.03)
BostonHousing	**9.70** (**0.89**)	11.50 (0.24)	NA	• 15.00 (0.00)
CobarOre	**8.50** (**0.93**)	12.70 (0.35)	16.10 (0.16)	16.10 (0.16)
engel	**10.20** (**0.81**)	9.40 (0.85)	**10.20** (**0.81**)	12.20 (0.20)
mcycle	**10.00** (**0.92**)	11.50 (0.51)	11.40 (0.51)	12.00 (0.35)
BigMac2003	**9.00** (**0.92**)	• 18.00 (0.04)	NA	14.30 (0.16)
UN3	9.50 (0.97)	12.00 (0.37)	NA	**10.30** (**0.74**)
cpus	**9.40** (**0.95**)	12.20 (0.29)	• 15.30 (0.01)	• 19.10 (0.00)

Table 3: Method Comparison: Ramp Loss ($\times 100, \tau = 0.1$)

data set	uncond	linear	rqss	npqr
caution	38.13 ± 3.44	32.40 ± 2.91	23.76 ± 2.74	**22.56 ± 2.68**
ftcollinssnow	42.10 ± 2.95	40.82 ± 2.95	44.07 ± 3.24	**39.08 ± 3.09**
highway	38.35 ± 6.34	45.39 ± 7.04	27.17 ± 3.26	**25.33 ± 3.62**
heights	40.08 ± 0.81	**34.50 ± 0.72**	34.66 ± 0.72	34.53 ± 0.72
sniffer	35.74 ± 3.13	12.78 ± 1.11	10.50 ± 0.98	∘ **9.92 ± 0.94**
snowgeese	32.08 ± 6.33	13.85 ± 3.46	**10.49 ± 2.53**	18.50 ± 4.96
ufc	40.21 ± 1.55	23.20 ± 0.95	21.23 ± 0.90	**21.22 ± 0.90**
birthwt	41.05 ± 2.14	38.15 ± 1.96	37.55 ± 2.08	**37.19 ± 1.96**
crabs	41.52 ± 1.99	2.24 ± 0.13	NA	**2.14 ± 0.12**
GAGurine	40.75 ± 1.81	27.87 ± 1.46	16.02 ± 1.20	**14.57 ± 1.11**
geyser	41.57 ± 1.84	32.50 ± 1.23	31.03 ± 1.36	**30.75 ± 1.40**
gilgais	42.10 ± 1.51	16.12 ± 1.01	**11.72 ± 0.69**	12.40 ± 0.66
topo	42.17 ± 3.86	26.51 ± 2.71	18.58 ± 2.65	**14.39 ± 1.65**
BostonHousing	35.57 ± 1.60	17.50 ± 0.95	NA	∘ **10.76 ± 0.61**
CobarOre	41.37 ± 4.97	41.93 ± 5.20	43.61 ± 4.59	**39.29 ± 6.69**
engel	35.75 ± 2.33	13.72 ± 1.14	13.25 ± 0.92	**13.01 ± 0.85**
mcycle	38.38 ± 3.04	37.88 ± 2.76	20.87 ± 1.52	∘ **17.06 ± 1.42**
BigMac2003	33.24 ± 5.12	21.75 ± 2.85	NA	∘ **17.89 ± 3.05**
UN3	40.79 ± 2.61	26.32 ± 1.70	NA	**23.96 ± 1.84**
cpus	23.00 ± 3.30	5.73 ± 1.04	2.45 ± 0.61	∘ **1.06 ± 0.17**

Table 4: Method Comparison: Pinball Loss ($\times 100$, $\tau = 0.5$)

data set	uncond	linear	rqss	npqr
caution	52.00 (0.62)	**49.00** (**0.92**)	**51.00** (**0.76**)	**49.00** (**0.92**)
ftcollinssnow	50.60 (0.84)	**49.70** (**1.00**)	48.60 (0.84)	51.40 (0.68)
highway	**48.30** (**1.00**)	44.20 (0.52)	45.00 (0.75)	41.70 (0.34)
heights	49.30 (0.63)	**50.10** (**0.91**)	49.80 (0.91)	50.30 (0.79)
sniffer	47.80 (0.72)	**51.00** (**0.72**)	**51.00** (**0.72**)	51.30 (0.72)
snowgeese	48.10 (1.00)	49.20 (1.00)	51.70 (0.77)	**50.60** (**0.77**)
ufc	49.20 (0.80)	**50.00** (**0.96**)	51.60 (0.50)	50.60 (0.80)
birthwt	48.90 (0.77)	**50.00** (**0.88**)	47.80 (0.56)	50.30 (0.88)
crabs	49.50 (0.94)	50.50 (0.83)	NA	**50.00** (**0.94**)
GAGurine	49.20 (0.78)	50.90 (0.69)	51.40 (0.61)	**49.80** (**0.96**)
geyser	48.60 (0.64)	**49.80** (**1.00**)	49.50 (0.91)	49.20 (0.82)
gilgais	48.70 (0.68)	**50.00** (**0.92**)	49.70 (0.92)	50.70 (0.75)
topo	**47.70** (**0.89**)	**47.70** (**0.89**)	**47.70** (**0.89**)	54.80 (0.49)
BostonHousing	**49.70** (**0.89**)	49.60 (0.89)	NA	51.70 (0.40)
CobarOre	46.40 (0.87)	44.50 (0.63)	**47.90** (**0.87**)	59.40 (0.14)
engel	50.90 (0.70)	49.70 (1.00)	49.60 (1.00)	**50.00** (**0.90**)
mcycle	**49.10** (**0.86**)	51.30 (0.73)	51.40 (0.73)	48.80 (0.86)
BigMac2003	49.30 (1.00)	**50.00** (**0.81**)	NA	44.20 (0.34)
UN3	**49.40** (**1.00**)	**50.60** (**0.86**)	NA	48.60 (0.86)
cpus	49.20 (0.89)	51.30 (0.68)	**49.70** (**1.00**)	51.80 (0.58)

Table 5: Method Comparison: Ramp Loss ($\times 100, \tau = 0.5$)

data set	uncond	linear	rqss	npqr
caution	23.35 ± 3.19	15.04 ± 1.54	∘ **13.19** ± **1.57**	15.16 ± 1.76
ftcollinssnow	18.71 ± 1.21	19.77 ± 1.76	19.35 ± 1.90	**18.67** ± **1.74**
highway	25.67 ± 3.71	28.49 ± 6.75	25.34 ± 6.09	**14.48** ± **3.53**
heights	17.63 ± 0.47	**15.47** ± **0.39**	15.50 ± 0.39	**15.47** ± **0.39**
sniffer	23.01 ± 3.62	5.87 ± 0.43	5.88 ± 0.44	∘ **5.25** ± **0.40**
snowgeese	26.94 ± 6.93	7.97 ± 2.67	8.09 ± 3.52	**7.94** ± **2.61**
ufc	18.05 ± 0.96	10.94 ± 0.45	10.84 ± 0.56	**10.15** ± **0.53**
birthwt	16.21 ± 1.03	16.17 ± 1.03	16.53 ± 1.19	∘ **15.20** ± **0.91**
crabs	17.09 ± 0.90	**0.99** ± **0.07**	NA	1.02 ± 0.08
GAGurine	20.86 ± 0.67	15.22 ± 0.83	10.51 ± 1.17	**10.13** ± **1.05**
geyser	14.21 ± 0.72	12.92 ± 0.67	12.48 ± 0.63	**12.10** ± **0.61**
gilgais	18.83 ± 0.72	6.74 ± 0.49	**5.06** ± **0.37**	5.51 ± 0.37
topo	16.50 ± 2.40	13.67 ± 2.80	13.84 ± 3.04	**10.30** ± **2.17**
BostonHousing	22.68 ± 1.28	11.67 ± 0.95	NA	∘ **6.96** ± **0.63**
CobarOre	17.63 ± 2.06	22.28 ± 3.43	20.16 ± 2.92	**15.01** ± **2.12**
engel	22.44 ± 2.57	**5.44** ± **0.43**	5.64 ± 0.65	5.70 ± 0.57
mcycle	15.97 ± 1.21	14.06 ± 1.00	10.58 ± 0.89	∘ **7.02** ± **0.56**
BigMac2003	23.29 ± 4.97	13.06 ± 2.20	NA	∘ **9.45** ± **2.85**
UN3	16.36 ± 1.00	10.37 ± 0.73	NA	∘ **8.81** ± **0.61**
cpus	24.01 ± 4.26	2.67 ± 0.26	1.78 ± 0.72	**0.71** ± **0.17**

Table 6: Method Comparison: Pinball Loss ($\times 100$, $\tau = 0.9$)

data set	uncond	linear	rqss	npqr
caution	**90.00** (**0.90**)	**90.00** (**0.90**)	89.00 (0.83)	89.00 (0.83)
ftcollinssnow	**90.30** (**0.82**)	89.20 (0.91)	88.30 (0.65)	89.20 (0.91)
highway	**89.20** (**0.89**)	• 64.20 (0.00)	• 61.70 (0.00)	• 70.00 (0.00)
heights	89.50 (0.58)	**90.00** (**0.94**)	89.80 (0.85)	90.10 (0.87)
sniffer	**89.40** (**0.97**)	87.60 (0.53)	86.80 (0.37)	84.60 (0.09)
snowgeese	**88.90** (**0.95**)	85.00 (0.32)	85.00 (0.32)	83.90 (0.32)
ufc	**89.80** (**0.94**)	90.30 (0.79)	88.50 (0.36)	88.30 (0.28)
birthwt	88.70 (0.68)	87.60 (0.38)	88.00 (0.38)	**88.90** (**0.68**)
crabs	**89.00** (**0.70**)	87.00 (0.20)	NA	87.10 (0.20)
GAGurine	89.50 (0.82)	**89.80** (**0.96**)	89.40 (0.82)	87.80 (0.25)
geyser	88.50 (0.48)	89.40 (0.74)	**90.40** (**0.81**)	89.10 (0.60)
gilgais	**89.10** (**0.59**)	88.30 (0.30)	87.10 (0.09)	• 83.90 (0.00)
topo	**89.10** (**0.84**)	87.10 (0.52)	85.70 (0.52)	• 77.70 (0.01)
BostonHousing	**90.10** (**0.89**)	88.80 (0.38)	NA	• 80.30 (0.00)
CobarOre	**89.10** (**0.93**)	85.80 (0.66)	79.10 (0.06)	85.80 (0.66)
engel	88.90 (0.65)	**90.00** (**0.85**)	89.10 (0.65)	89.40 (0.81)
mcycle	88.60 (0.70)	**88.80** (**0.70**)	87.70 (0.51)	86.20 (0.23)
BigMac2003	**89.30** (**0.92**)	84.30 (0.16)	NA	• 77.70 (0.01)
UN3	**88.00** (**0.53**)	86.70 (0.24)	NA	85.80 (0.15)
cpus	**89.30** (**0.87**)	87.80 (0.40)	• 82.60 (0.00)	• 82.10 (0.00)

Table 7: Method Comparison: Ramp Loss ($\times 100, \tau = 0.9$)

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
data set	npqr	noncross	npqr	noncross	npqr	noncross
caution	9.56 ± 0.92	**9.55 ± 0.92**	22.56 ± 2.68	**22.51 ± 2.68**	15.16 ± 1.76	**15.15 ± 1.76**
ftcollinssnow	**16.24 ± 1.17**	**16.24 ± 1.17**	39.08 ± 3.09	**38.81 ± 3.09**	**18.67 ± 1.74**	**18.67 ± 1.74**
highway	8.34 ± 1.18	**8.20 ± 1.20**	25.33 ± 3.62	**25.30 ± 3.57**	14.48 ± 3.53	**14.41 ± 3.53**
heights	**15.26 ± 0.39**	15.27 ± 0.39	**34.53 ± 0.72**	34.54 ± 0.72	**15.47 ± 0.39**	15.48 ± 0.39
sniffer	5.48 ± 0.64	**5.43 ± 0.64**	9.92 ± 0.94	**9.91 ± 0.94**	5.25 ± 0.40	**5.19 ± 0.40**
snowgeese	**5.03 ± 0.87**	**5.03 ± 0.87**	**18.50 ± 4.96**	18.59 ± 4.98	7.94 ± 2.61	**7.88 ± 2.62**
ufc	**9.70 ± 0.42**	**9.70 ± 0.39**	**21.22 ± 0.90**	21.23 ± 0.90	10.15 ± 0.53	**9.92 ± 0.49**
birthwt	**17.68 ± 1.16**	17.69 ± 1.16	**37.19 ± 1.96**	37.21 ± 1.96	**15.20 ± 0.91**	**15.20 ± 0.91**
crabs	**0.91 ± 0.07**	**0.91 ± 0.07**	**2.14 ± 0.12**	**2.14 ± 0.12**	1.02 ± 0.08	**1.01 ± 0.08**
GAGurine	6.00 ± 0.63	**5.99 ± 0.63**	**14.57 ± 1.11**	**14.57 ± 1.11**	**10.13 ± 1.05**	**10.13 ± 1.05**
geyser	**10.91 ± 0.49**	**10.91 ± 0.49**	30.75 ± 1.40	**30.71 ± 1.40**	**12.10 ± 0.61**	12.11 ± 0.61
gilgais	**5.46 ± 0.35**	**5.46 ± 0.35**	12.40 ± 0.66	**12.37 ± 0.66**	**5.51 ± 0.37**	**5.51 ± 0.37**
topo	**6.03 ± 0.91**	6.04 ± 0.91	∘ **14.39 ± 1.65**	15.54 ± 1.62	10.30 ± 2.17	**10.21 ± 2.16**
BostonHousing	5.10 ± 0.42	**5.04 ± 0.42**	10.76 ± 0.61	**10.73 ± 0.61**	6.96 ± 0.63	∘ **6.85 ± 0.62**
CobarOre	13.80 ± 2.70	**13.66 ± 2.63**	**39.29 ± 6.69**	40.00 ± 6.61	∘ **15.01 ± 2.12**	15.13 ± 2.12
engel	**5.55 ± 0.37**	**5.55 ± 0.37**	13.01 ± 0.85	**12.96 ± 0.85**	**5.70 ± 0.57**	**5.70 ± 0.57**
mcycle	**7.39 ± 0.90**	**7.39 ± 0.90**	17.06 ± 1.42	**17.03 ± 1.42**	7.02 ± 0.56	**7.00 ± 0.55**
BigMac2003	**6.13 ± 0.96**	6.36 ± 1.02	17.89 ± 3.05	∘ **17.72 ± 3.05**	**9.45 ± 2.85**	9.48 ± 2.84
UN3	**11.47 ± 1.02**	11.52 ± 1.04	23.96 ± 1.84	**23.81 ± 1.81**	**8.81 ± 0.61**	8.82 ± 0.61
cpus	∘ **0.67 ± 0.23**	1.30 ± 0.18	∘ **1.06 ± 0.17**	1.35 ± 0.17	∘ **0.71 ± 0.17**	0.87 ± 0.18

Table 8: Pinball loss comparison between the nonparametric quantile regression without (npqr) and with (noncross) non-crossing constraints.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
data set	npqr	noncross	npqr	noncross	npqr	noncross
caution	**12.00 (0.40)**	**12.00 (0.40)**	**49.00 (0.92)**	**49.00 (0.92)**	**89.00 (0.83)**	**89.00 (0.83)**
ftcollinssnow	**12.20 (0.44)**	**12.20 (0.44)**	**51.40 (0.68)**	**51.40 (0.68)**	**89.20 (0.91)**	**89.20 (0.91)**
highway	• 20.00 (0.03)	• **13.30 (0.03)**	41.70 (0.34)	**45.00 (0.34)**	• **70.00 (0.00)**	• 56.70 (0.00)
heights	**10.00 (0.92)**	9.90 (0.92)	**50.30 (0.79)**	**50.30 (0.79)**	**90.10 (0.87)**	**90.10 (0.87)**
sniffer	• **15.90 (0.02)**	• **15.90 (0.02)**	**51.30 (0.72)**	**51.30 (0.72)**	84.60 (0.09)	**85.40 (0.09)**
snowgeese	**13.60 (0.32)**	**13.60 (0.32)**	**50.60 (0.77)**	**50.60 (0.77)**	**83.90 (0.32)**	**83.90 (0.32)**
ufc	**10.50 (0.68)**	10.70 (0.68)	**50.60 (0.80)**	**50.60 (0.80)**	**88.30 (0.28)**	88.20 (0.28)
birthwt	11.60 (0.38)	**10.00 (0.38)**	50.30 (0.88)	**50.20 (0.88)**	**88.90 (0.68)**	**88.90 (0.68)**
crabs	13.30 (0.09)	**13.00 (0.09)**	**50.00 (0.94)**	49.50 (0.94)	**87.10 (0.20)**	87.00 (0.20)
GAGurine	12.10 (0.19)	**11.60 (0.19)**	49.80 (0.96)	**49.90 (0.96)**	87.80 (0.25)	**88.10 (0.25)**
geyser	12.20 (0.21)	**12.10 (0.21)**	49.20 (0.82)	**49.60 (0.82)**	**89.10 (0.60)**	89.00 (0.60)
gilgais	**12.40 (0.12)**	**12.40 (0.12)**	**50.70 (0.75)**	50.80 (0.75)	• 83.90 (0.00)	• **84.20 (0.00)**
topo	• **19.40 (0.03)**	• **19.40 (0.03)**	**54.80 (0.49)**	56.30 (0.49)	• **77.70 (0.01)**	• **77.70 (0.01)**
BostonHousing	• **15.00 (0.00)**	• 15.10 (0.00)	51.70 (0.40)	**51.50 (0.40)**	• 80.30 (0.00)	• **80.80 (0.00)**
CobarOre	**16.10 (0.16)**	**16.10 (0.16)**	**59.40 (0.14)**	**59.40 (0.14)**	**85.80 (0.66)**	**85.80 (0.66)**
engel	**12.20 (0.20)**	**12.20 (0.20)**	**50.00 (0.90)**	50.10 (0.90)	**89.40 (0.81)**	**89.40 (0.81)**
mcycle	**12.00 (0.35)**	**12.00 (0.35)**	**48.80 (0.86)**	48.10 (0.86)	86.20 (0.23)	**87.40 (0.23)**
BigMac2003	**14.30 (0.16)**	16.00 (0.16)	**44.20 (0.34)**	43.70 (0.34)	• 77.70 (0.01)	• **79.30 (0.01)**
UN3	**10.30 (0.74)**	**10.30 (0.74)**	**48.60 (0.86)**	47.80 (0.86)	85.80 (0.15)	**86.70 (0.15)**
cpus	• **19.10 (0.00)**	• 20.60 (0.00)	**51.80 (0.58)**	46.90 (0.58)	• 82.10 (0.00)	• **82.50 (0.00)**

Table 9: Ramp loss (quantile property) comparison between the nonparametric quantile regression without (npqr) and with (noncross) non-crossing constraints.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
data set	npqr	npqrm	npqr	npqrm	npqr	npqrm
cars	**0.65** ± **0.15**	0.66 ± 0.16	**1.59** ± **0.32**	1.61 ± 0.23	0.79 ± 0.16	**0.77** ± **0.16**
onions	2.68 ± 1.21	**2.27** ± **0.71**	4.93 ± 1.58	**4.89** ± **1.47**	1.86 ± 0.73	**1.84** ± **0.37**

Table 10: Pinball loss comparison between the nonparametric quantile regression without (npqr) and with (npqrm) monotonicity constraints.

	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
data set	npqr	monotonic	npqr	monotonic	npqr	monotonic
cars	12.00 (0.24)	**11.00** (**0.24**)	**51.00** (**0.88**)	**51.00** (**0.88**)	**89.00** (**0.82**)	**89.00** (**0.82**)
onions	• 18.00 (0.00)	• **17.00** (**0.00**)	**48.00** (**0.44**)	**48.00** (**0.44**)	• **86.00** (**0.01**)	• 80.00 (0.00)

Table 11: Ramp loss (quantile property) comparison between the nonparametric quantile regression without (npqr) and with (npqrm) monotonicity constraints.

References

L. K. Bachrach, T. Hastie, M. C. Wang, B. Narashimhan, and R. Marcus. Bone mineral acquisition in healthy asian, hispanic, black and caucasian youth, a longitudinal study. *Journal of Clinical Endocrinal Metabolism*, 84:4702– 4712, 1999.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher averages. In *Proc. Annual Conf. Computational Learning Theory*, pages 44–58, 2002.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. J. Hopkins Press, Baltimore, ML, 1994.

R. J. Bosch, Y. Ye, and G. G.Woodworth. A convergent algorithm for quantile regression with smoothing splines. *Computational Statistics and Data Analysis*, 19:613–630, 1995.

D. D. Cox. Approximation of method of regularization estimators. *Annals of Statistics*, 16:694–713, 1988.

G. Fung, O. L. Mangasarian, and A. J. Smola. Minimal kernel classifiers. *Journal of Machine Learning Research*, 3:303–321, 2002.

C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society B*, 55:353–368, 1993.

P. Hall and N. Tajvidi. Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, 2000.

P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society - Series B*, 2005. forthcoming.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

X. He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, may 1997.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–680, 1994.

Q. V. Le, A. J. Smola, and T. Gärtner. Simpler knowledge-based support vector machines. In *Proc. Intl. Conf. Machine Learning*, 2006.

E. Mammen, J. S. Marron, B. A. Turlach, and M. P. Wand. A general projection framework for constrained smoothing. *Statistical Science*, 16(3):232–248, August 2001.

S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures on Machine Learning*, number 2600 in LNAI, pages 1–40. Springer, 2003.

D. Ruppert and R. J. Carroll. *Semiparametric Regression*. Wiley, 2003.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, R. C. Williamson, A. J. Smola, and J. Shawe-Taylor. Single-class support vector machines. In J. Buhmann, W. Maass, H. Ritter, and N. Tishby, editors, *Unsupervised Learning*, Dagstuhl-Seminar-Report 235, pages 19–20, 1999.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

A. J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

A. J. Smola, T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 585–591, Cambridge, MA, 1999. MIT Press.

I. Takeuchi and T. Furuhashi. Non-crossing quantile regressions by SVM. In *Proc. International Joint Conference on Neural Networks*, 2004.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

V. Vapnik, S. Golowich, and A. J. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Berlin, 1982.

S. V. N. Vishwanathan, A. J. Smola, and M. N. Murty. SimpleSVM. In Tom Fawcett and Nina Mishra, editors, *Proc. Intl. Conf. Machine Learning*, Washington DC, 2003. AAAI press.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization bounds for regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transaction on Information Theory*, 47(6):2516–2532, 2001.