NATIONAL
ICT AUSTRALIA
LIMITED

Nonparametric Quantile Estimation

NICTA Technical Report 1000005T.1
Statistical Machine Learning Program, Canberra
June 2005

Quoc V. Le QUOCLE@ANU.EDU.AU
Tim Sears TIM.SEARS@ANU.EDU.AU
Alexander J. Smola ALEX.SMOLA@NICTA.COM.AU
National ICTA Australia and National University, Canberra ACT

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. We provide uniform convergence statements and guarantees on the quality of margins. In addition to that, experimental results show the feasibility and competitiveness of our method.

Keywords

Support Vector Machines, Kernel Methods, Quantile Regression, Nonparametric Techniques

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, amongst them least mean square (LMS) regression [**?**], robust regression [8], or ϵ-insensitive regression [16, 17]. Regularized variants penalize e.g. by a Reproducing Kernel Hilbert Space (RKHS) norm [19] or via Ridge Regression [7].

1.1 Motivation

While these estimates serve their purpose, as can be shown both experimentally and theoretically, there exists a large area of problems where we are more interested in estimating a quantile. That is, we wish to obtain other information about 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) championed by Koenker (see [9] 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 [12].
- To demonstrate a nonparametric version of QR which outperforms the currently available nonlinear QR regression formations [9]. 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 [9]. 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 one example we show how to enforce a monotonicity constraint. This feature could be desirable in a growth chart, and not otherwise guaranteed, if data is scarce or many variables are included in the regression.

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 $\Pr\{y \leq \mu|x\} = \tau$.*

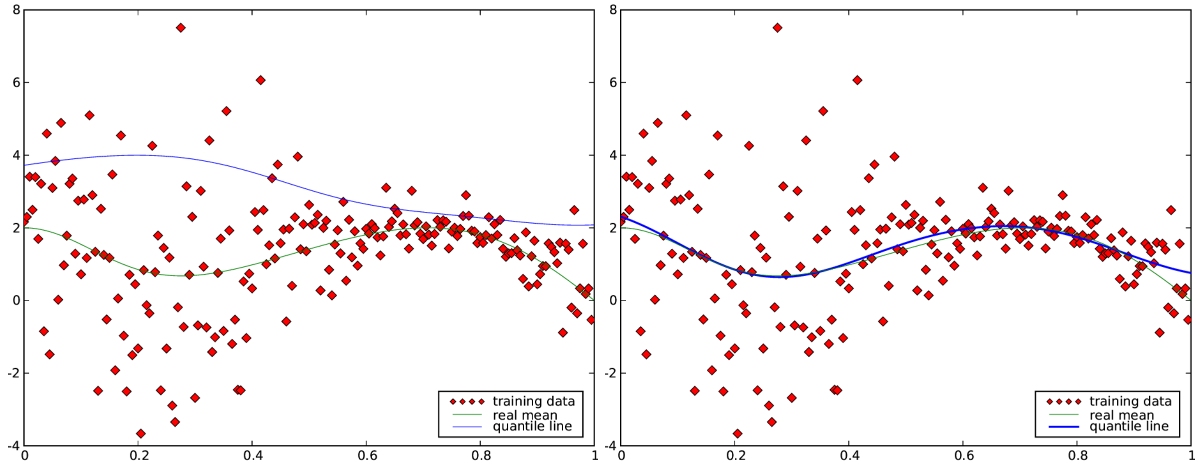


Figure 1: Illustration of the nonparametric quantile regression on toy dataset. On the left, $\tau = 0.9$. On the right, $\tau = 0.5$ the quantile regression line approximates the median of the data very closely (since ξ is normally distributed median and mean are identical).

1.3 An Example

The definition may be best illustrated by simple example. Consider a situation where x is drawn uniformly from $[0, 1]$ and y is given by

$$y(x) = \sin \pi x + \xi \text{ where } \xi \sim \mathcal{N}\left(0, e^{\sin 2\pi x}\right).$$

Here the amount of noise is a function of the location. Since ξ is symmetric with mean and mode 0 we have $\mu_{0.5}(x) = \sin \pi x$. Moreover, we can compute the quantiles by solving for $\Pr\left\{y \le \mu | x\right\} = \tau$ explicitly.

Since ξ is normal we know that the quantiles of ξ are given by $\Phi^{-1}(\tau) \sin 2\pi x$, where Φ is the cumulative distribution function of the normal distribution with unit variance. This means that

$$\mu_\tau(x) = \sin \pi x + \Phi^{-1}(\tau) \sin 2\pi x. \tag{1}$$

Figure 1 shows two quantile *estimates*. We see that depending on the choice of the quantile, we obtain a close approximation of the median ($\tau = 0.5$), or a curve which tracks just inside the upper envelope of the data ($\tau = 0.9$). The error bars of many regression estimates can be viewed as crude quantile regressions: one tries to specify the interval within which, with high probability, the data may lie. Note, however, that the latter does not entirely correspond to a quantile regression: error bars just give an upper bound on the range within which an estimate lies, whereas QR aims to estimate the exact boundary at which a certain quantile is achieved. In other words, it corresponds to tight error bars.

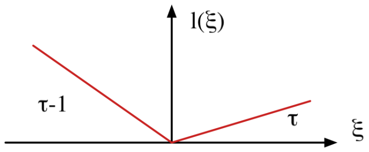
2. Quantile Regression

Given the definition of $q_\tau(x)$ and knowledge of support vector machines we might be tempted to use version of the ϵ-insensitive tube regression to estimate $q_\tau(x)$. More specifically one might want to estimate quantiles nonparametrically using an extension of the ν-trick, as outlined in [12]. The latter, however, has the disadvantage of requiring us to estimate both an upper and lower quantile *simultaneously*. While this can be achieved by quadratic programming, in doing so we would 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 [18] we attack the problem by estimating each quantile separately. That said, we provide a detailed description of a symmetric quantile regression in Appendix A.

2.1 Loss Function

The basic idea of 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. [9] 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:

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



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:*

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 optimality 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. Koencker [9] 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

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

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:

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

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 $\lim_{\delta \to 0} \mathbf{E}\left[\epsilon(\delta)\right] = 0$. With probability 1, asymptotically, $\frac{m_-}{m}$ equals τ.*

Proof For the two claims, denote by f^* the minimum of $R_{\text{reg}}[f]$ with $f^* = g^* + b^*$. Then $R_{\text{reg}}[g^* + b]$ 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 [12, 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 [20]. 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 in [12]. ■

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 $R_{\text{reg}}[f]$.

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

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

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 have the well known 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 $\vec{1}^\top \alpha = 0$. 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 [17]. The key difference between the two estimation problems is that in ϵ-SVR we have an additional $\epsilon\|\alpha\|_1$ penalty in the objective function. This ensures that observations with deviations from the estimate, i.e. with $|y_i - f(x_i)| < \epsilon$ 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.

The 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 [17]. Figure 2 shows how QR behaves subject to changing the model class, that is, subject to changing the regularization parameter. All three estimates in Figure 2 attempt to compute the median subject to different smoothness constraints. While they all satisfy the quantile property of having a fraction of $\tau = 0.5$ points on either side of the regression, they track the observations more or less closely. Section 5 gives an in-depth comparison of quantile regression estimators on a large range of datasets.

3. Extensions and Modifications

The mathematical programming 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 modifications which allow for a richer model class, and others which enforce monotonicity in the estimate.

3.1 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*

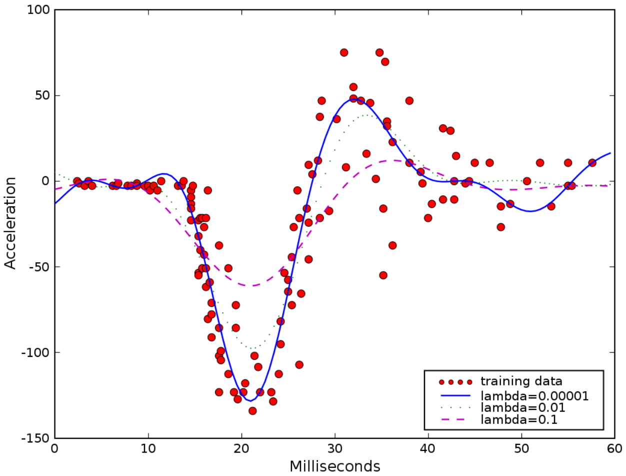


Figure 2: The data set measures acceleration in the head of a crash test dummy v. time in tests of motorcycle crashes. Three regularized versions of the median regression estimate ($\tau = 0.5$). While all three variants satisfy the quantile property, the degree of smoothness is controlled by the regularization constant λ. All three estimates compare favorably to a similar graph of nonlinear QR estimates reported in [9].

increasing function of age. This property needs to translate to the quantile estimates $\mu_\tau(x)$. There is no guarantee, however, that unless we observe large amounts of data, the estimates $f(x)$ will also be monotonic functions of age.

To address this problem we use an approach similar to [17, 14] 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

$$Df(x) = D\left(\langle \phi(x), w\rangle + b\right) = \langle D\phi(x), w\rangle = \langle \psi(x), w\rangle \text{ where } \psi(x) := D\phi(x). \tag{8}$$

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

$$\underset{w,b}{\text{minimize}} \quad C\sum_{i=1}^{m} \tau\xi_i + (1-\tau)\xi_i^* + \frac{1}{2}\|w\|^2 \tag{9}$$
$$\text{subject to } \; y_i - \langle \phi(x_i), w\rangle - b \leq \xi_i, \;\; \langle \phi(x_i), w\rangle + b - y_i \leq \xi_i^* \text{ and } \langle \psi(x_i), w\rangle \geq 0 \text{ where } \xi_i, \xi_i^* \geq 0.$$

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

$$\underset{\alpha,\beta}{\text{minimize}} \quad \frac{1}{2}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}^\top \begin{bmatrix} K & D_1K \\ D_2K & D_1D_2K\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} - \alpha^\top \vec{y} \tag{10a}$$

$$\text{subject to } \; C(\tau-1) \leq \alpha_i \leq C\tau \text{ and } 0 \leq \beta_i \text{ for all } 1 \leq i \leq m \text{ and } \vec{1}^\top\alpha = 0. \tag{10b}$$

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 $f(x) = \sum_i \alpha_i k(x_i, x) + \beta_i D_1 k(x_i, x) + b$.

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

$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\left\|x - x'\right\|^2\right). \tag{11}$$

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

$$D_1 k(x,x') = \frac{x'_1 - x_1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right) \tag{12a}$$

$$D_2 k(x,x') = \frac{x_1 - x'_1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right) \tag{12b}$$

$$D_1 D_2 k(x,x') = \left[\sigma^{-2} - \frac{(x_1 - x'_1)^2}{\sigma^4}\right] \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right). \tag{12c}$$

Plugging the values of (12) into (10) yields the quadratic program. Note also that both $k(x,x')$ and $D_1 k(x,x')$ ((12a)), 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 $D_1 k$.

3.2 Function Classes

Semiparametric Estimates RKHS expansions may not be the only function classes desired for quantile regression. For instance, a semiparametric model may be more desirable as it allows for interpretation of the linear coefficients in as often desired in the social sciences [5, 13, 3]. In this case we add a set of parametric functions f_i and solve

$$\text{minimize } \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \text{ where } f(x) = g(x) + \sum_{i=1}^{n} \beta_i f_i(x) + b. \tag{13}$$

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

$$\sum_{j=1}^{m} \alpha_j f_i(x_j) = 0 \text{ for all } i. \tag{14}$$

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

$$\text{minimize } \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \lambda \sum_{j=1}^{n} |\alpha_i| \text{ where } f(x) = \sum_{i=1}^{n} \alpha_i f_i(x) + b. \tag{15}$$

is a *linear program* which can be solved efficiently by existing codes for large scale problems. In the context of (15) the functions f_i constitute the generators of the convex function class. This approach is similar to [**?**]. 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 [11].

Relevance Vector Regularization and Sparse Coding Finally, for sparse expansions one can use more aggressive penalties on linear function expansions than those given in (15). For instance, we could use a staged regularization as in the RVM [15], 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 [4] 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.

3.3 Transitivity

Let f_τ and $f_{\tau'}$ be estimates obtained by solving (6) for values $\tau \geq \tau'$. It is only natural to require that $f_\tau(x) \geq f_{\tau'}(x)$ for all x, as this property also holds for the quantiles $\mu_\tau(x)$ and $\mu_{\tau'}(x)$. However the property, is by no means guaranteed when simply solving the optimization problem. In other words, the quantile regressions we obtain need not be monotonic as a function of the quantile parameters. This is due to the fact that a higher quantile level need not necessarily correspond to a simpler function. As a result, we may produce non-monotonic quantile estimates $f_\tau(x)$. This problem can be overcome in three different ways:

- Increasing the sample size while keeping the model complexity fixed will lead to monotonicity, as the estimates converge to the best values within the model class.
- Suitable capacity control for fixed sample size has the same effect.
- The constraints can be enforced explicitly by solving several optimization problems jointly. This means that for $\tau \geq \tau'$ one solves the joint optimization problem given by two instances of (4) with the additional set of constraints $f_\tau(x_i) \geq f_{\tau'}(x_i)$.

  The downside of the last approach is that by coupling two quadratic programs the computational complexity is much increased. Secondly, enforcing the constraint on the sample does not guarantee its satisfaction everywhere. Thirdly imposing the constraints may break the quantile property of Lemma 3.

Consequently proper capacity appears to be the best practical strategy to minimize the probability of occurrence of such a problem. This will be addressed in the following section.

4. Theoretical Analysis

4.1 Performance Indicators

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

  $$\Pr_{X,Y}\left\{\left|\Pr\left\{y < f_\tau(x)\right\} - \tau\right| \geq \epsilon\right\} \leq \delta. \tag{16}$$

  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 (16) 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

  $$\Pr_{X,Y}\left\{R[f_\tau] - R[f_\tau^*] > \epsilon\right\} \leq \delta. \tag{17}$$

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 [10], or similar tools for empirical processes. However, for the sake of simplicity, we use the tools from [2], 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 R_m and its empirical variant $\hat{R}_m$ are defined as follows:*

$$\hat{R}_m(\mathcal{F}) := \mathbf{E}_\sigma\left[\sup_{f \in \mathcal{F}}\left|\frac{2}{m}\sum_1^n \sigma_i f(x_i)\right| \, \middle| X\right] \text{ and } R_m(\mathcal{F}) := \mathbf{E}_X\left[\hat{R}_m(\mathcal{F})\right]. \tag{18}$$

Conveniently, if Φ is a Lipschitz continuous function with Lipschitz constant L, one can show [2] that

$$R_m(\Phi \circ \mathcal{F}) \leq 2LR_m(\mathcal{F}) \text{ where } \Phi \circ \mathcal{F} := \{g | g = \phi \circ f \text{ and } f \in \mathcal{F}\}. \tag{19}$$

An analogous result exists for empirical quantities bounding $\hat{R}_m(\Phi \circ \mathcal{F}) \leq 2L\hat{R}_m(\mathcal{F})$. The combination of (19) with [2, Theorem 8] yields:

$$r_\epsilon^+(\xi) := \min\{1, \max\{0, 1 - \xi/\epsilon\}\} \qquad \text{(25a)}$$

$$r_\epsilon^-(\xi) := \min\{1, \max\{0, -\xi/\epsilon\}\} \qquad \text{(25b)}$$

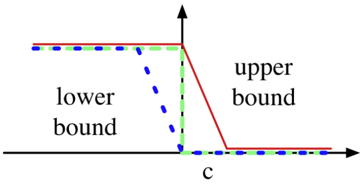


Figure 3: Ramp functions bracketing bracketing the characteristic function via $r_\epsilon^+ \geq \chi_{(-\infty,0]} \geq r_\epsilon^-$.

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 $\mathcal{X}$:*

$$\sup_{f \in \mathcal{F}} \left| \mathbf{E}_x \left[\Phi(f(x))\right] - \frac{1}{m}\sum_{i=1}^m \Phi(f(x_i)) \right| \leq 2LR_m(\mathcal{F}) + \sqrt{\frac{8\log 2/\delta}{m}}. \tag{20}$$

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 $(1-\delta)$*

$$R[f_\tau] - R[f_\tau^*] \leq 2\max LR_m(\mathcal{F}) + (4 + LB)\sqrt{\frac{\log 2/\delta}{2m}} \text{ where } L = \{\tau, 1-\tau\}\,. \tag{21}$$

Proof We use the standard bounding trick that

$$R\left[f_\tau\right] - R\left[f_\tau^*\right] \leq \left|R\left[f_\tau\right] - R_{\text{emp}}\left[f_\tau\right]\right| + R_{\text{emp}}\left[f_\tau^*\right] - R\left[f_\tau^*\right] \tag{22}$$

$$\leq \sup_{f\in\mathcal{F}} \left|R\left[f\right] - R_{\text{emp}}\left[f\right]\right| + R_{\text{emp}}\left[f_\tau^*\right] - R\left[f_\tau^*\right] \tag{23}$$

where (22) follows from $R_{\text{emp}}\left[f_\tau\right] \leq R_{\text{emp}}\left[f_\tau^*\right]$. The first term can be bounded directly by Theorem 5. For the second part we use Hoeffding's bound [6] 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 [10] that $R_m(\mathcal{F}) \leq \frac{2C}{\sqrt{m}}$. This means that the bounds of Theorem 6 translate into a rate of convergence of

$$R\left[f_\tau\right] - R\left[f_\tau^*\right] = O(m^{-\frac{1}{2}}). \tag{24}$$

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 [1] 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. (16).

In this view (16) 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 (16). One solution for this problem is to use an artificial margin ϵ and ramp functions $r_\epsilon^+, r_\epsilon^-$ as defined in Figure 3. These functions are Lipschitz continuous with constant $L = 1/\epsilon$. This leads to:

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

$$\frac{1}{m}\sum_{i=1}^{m} r_\epsilon^-(y_i - f(x_i)) - \Delta \leq \mathbf{E}\left[\chi_{(-\infty,0]}(y - f_\tau(x))\right] \leq \frac{1}{m}\sum_{i=1}^{m} r_\epsilon^+(y_i - f(x_i)) + \Delta, \tag{26}$$

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

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 $r_\epsilon^+ \circ \mathcal{F}$ and $r_\epsilon^- \circ \mathcal{F}$ and note that the ramp loss has Lipschitz constant $L = 1/\epsilon$. 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 above. That is, we check the performance of various quantile 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 [9] this is the only objective criterion we could find.
- 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.

5.1 Models

We compare the following four models:

- An unconditional quantile estimator. This should be the baseline of all other estimates in terms of minimizing the expected risk. Given the simplicity of the function class it should perform best in terms of preserving the quantile property.
- Linear QR as described in [9]. This uses the a linear unregularized model to minimize l_τ. In experiments, we used the `rq` routine available in the R package called `quantreg`.
- Nonparametric QR as described by [9] (Ch. 7). This uses a spline model for each coordinate individually, with linear effect. The fitting routine used was `rqss`, also available in `quantreg`.[1]
- 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. This appears as `nprq`.[2]

As preprocessing all coordinates of x_i were rescaled to zero mean and unit variance coordinate-wise.

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, the `npqr` method outperforms all other estimators significantly in most cases. Moreover, it compares favorably in terms of preserving the quantile property.

[1] Additional code containing bugfixes and other operations necessary to carry out our experiments is available at http://sml.nicta.com.au/∼sears/qr/.

[2] Code will be available as part of the CREST toolbox for research purposes.

5.2 Datasets

We chose 20 regression datasets from the following R packages: `mlbench, quantreg, alr3` and `MASS`. The first library contains datasets 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^3. 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. [4] 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 $m = 39$ (highway) 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. In order to make a fair comparison, we *standardized* all datasets (to have zero mean and unit variance) before running the algorithms.

Data Set	Sample Size	dimension (x)	y variables	dropped variables
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	126	6	Purban	Locality
cpus	209	7	estperf	name

Table 1: Dataset facts

5.3 Results

We tested the performance of the **4** algorithms on **3** different quantiles ($\tau \in \{0.1, 0.5, 0.9\}$). For each model we used 10-fold cross-validation to assess the confidence of our results. For the `npqr` model, kernel width and smoothness parameters were automatically chosen by cross-validation within the training sample. We performed 10 runs on the training set to adjust parameters, then chose the best parameter setting based on the pinball loss averaged over 10 splits. To compare across all four models we measured both pinball loss and quantile performance.

The full results are shown in appendix B. The 20 data sets and three quantile levels yield 60 trials for each model. In terms of pinball loss averaged across 10 tests the `npqr` model performed best or tied on 51 of the 60 trials, showing the clear advantage of the proposed method. The results are consistent across quantile levels. We can get another impression of performance by looking at the loss in each of the 10 test runs that enter each trial. This is depicted in Figure 5.3. In a large majority of test cases the `npqr` model error is smaller than that of the other models, resulting in a "cloud" below the 45 degree line.

Moreover, the quantile properties of all four methods are comparable. All four models produced ramp losses close to the desired quantile, although the `rqss` and `npqr` models were noisier in this regard. The complete results for the ramp loss are presented in last three tables in Appendix B.

[3] See http://cran.r-project.org/

[4] 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 limit forecasting on unseen examples, which was part of our performance metric.

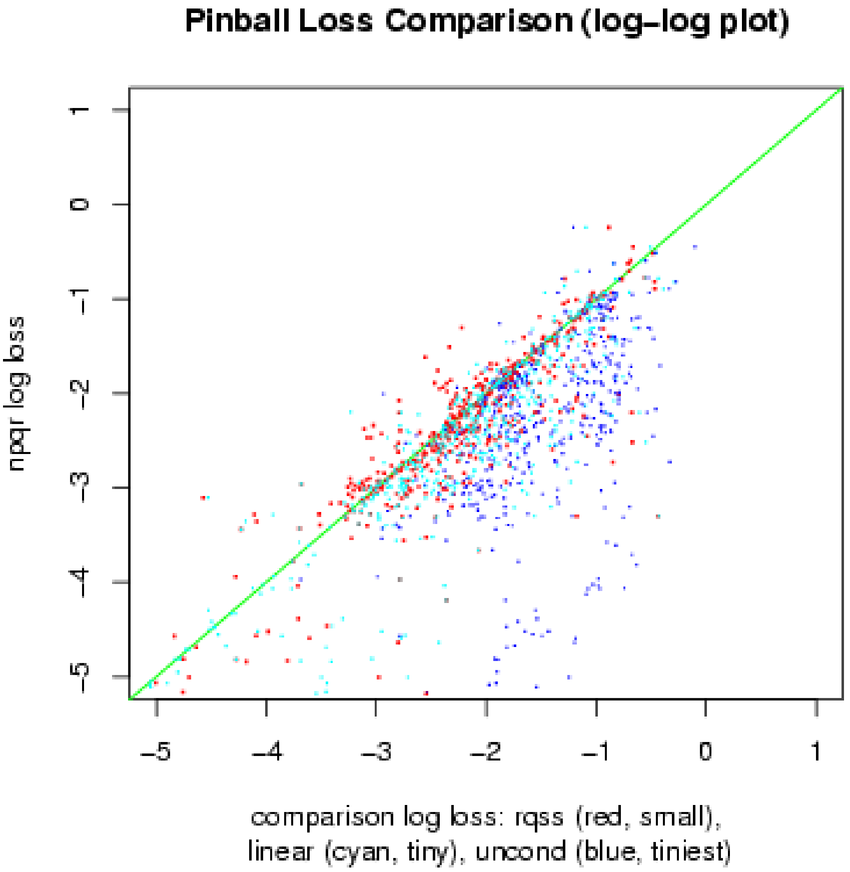


Figure 4: A log-log plot of test results for npqr versus each of the other three models. The relative weight of the "cloud" lying below 45-degree line provides an impression of the performance of the npqr versus the other models.

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 many aspects of the modeling process. While not truly model-free, the experience of comparing the models in this paper shows how easy it is to estimate the quantities of interest in QR, without any of the angst of model selection. It is interesting to consider whether kernel methods, with proper regularization, are a good substitute for some traditional modeling activity. In particular we were able to some simpler traditional statistical estimates significantly, which allows the human modeler to focus on statistical concerns at a higher level.

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, we provided uniform convergence results and experimental evidence for its soundness.

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.

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 and by the Pascal Network of Excellence. 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.

References

[1] P.L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher averages. In *Proceedings of the 15th conference on Computational Learning Theory COLT'02*, pages 44–58, 2002.

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

[3] 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.

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

[5] 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.

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

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

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

[9] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[10] 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.

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

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

[13] 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.

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

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

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

[17] 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.

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

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

[20] 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.

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 [12]. 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) [12] 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 $\epsilon(x)$ into a nonparametric estimate itself. This means that we solve the following optimization problem.

$$\underset{\theta_1,\theta_2,b,\epsilon}{\text{minimize}} \quad \frac{\lambda_1}{2}\|\theta_1\|^2 + \frac{\lambda_2}{2}\|\theta_2\|^2 + \sum_{i=1}^{m}(\xi_i + \xi_i^*) - \nu m \epsilon \tag{27a}$$

$$\text{subject to} \quad \langle \phi_1(x_i), \theta_1 \rangle + b - y_i \leq \epsilon + \langle \phi_2(x_i), \theta_2 \rangle + \xi_i \tag{27b}$$

$$y_i - \langle \phi_1(x_i), \theta_1 \rangle - b \leq \epsilon + \langle \phi_2(x_i), \theta_2 \rangle + \xi_i^* \tag{27c}$$

$$\xi_i, \xi_i^* \geq 0 \tag{27d}$$

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 [12] is that in the latter the authors assume that the specific form of the noise is *known*. In (27) instead, we make no such assumption and instead we estimate $\epsilon(x)$ as $\langle \phi_2(x), \theta_2 \rangle + \epsilon$.

By Lagrange multiplier methods one may check that the dual of (27) is obtained by

$$\underset{\alpha,\alpha^*}{\text{minimize}} \quad \frac{1}{2\lambda_1}(\alpha - \alpha^*)^\top K_1(\alpha - \alpha^*) + \frac{1}{2\lambda_2}(\alpha + \alpha^*)^\top K_1(\alpha + \alpha^*) + (\alpha - \alpha^*)^\top y \tag{28a}$$

$$\text{subject to} \quad \vec{1}^\top(\alpha - \alpha^*) = 0 \tag{28b}$$

$$\vec{1}^\top(\alpha + \alpha^*) = Cm\nu \tag{28c}$$

$$0 \leq \alpha_i, \alpha_i^* \leq 1 \text{ for all } 1 \leq i \leq m \tag{28d}$$

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 $\epsilon(x)$ via

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\, k_1(x_i, x) + b \text{ and } \epsilon(x) = \sum_{i=1}^{m}(\alpha_i + \alpha_i^*)\, k_2(x_i, x) + \epsilon. \tag{29}$$

The scalars b and ϵ can be computed conveniently as dual variables of (28) when solving the problem with an interior point code.

A.2 The ν-Property

As in the parametric case also (27) has the ν-property. However, it is worth noting that the solution $\epsilon(x)$ 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 (27) satisfies*

1. *The fraction of points for which $|y_i - f(x_i)| < \epsilon(x_i)$ is bounded by $1 - \nu$.*
2. *The fraction of constraints (27b) and (27c) 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 $\lim_{\delta \to 0} \mathbf{E}[\epsilon(\delta)] = 0$. With probability 1, asymptotically, the fraction of points satisfying $|y_i - f(x_i)| = \epsilon(x_i)$ converges to 0.*

Moreover, imposing $\epsilon \geq 0$ is equivalent to relaxing (28c) to $\vec{1}^\top(\alpha - \alpha^*) \leq Cm\nu$. If in addition K_2 has only nonnegative entries then also $\epsilon(x) \geq 0$ for all x_i.

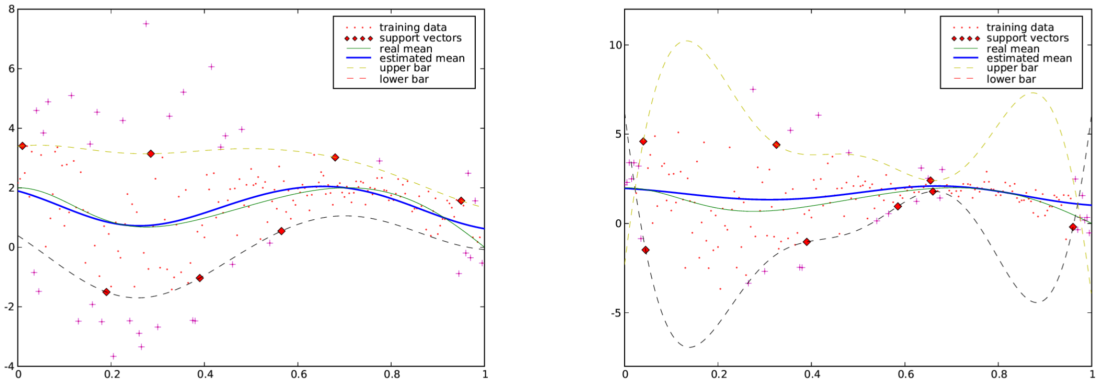


Figure 5: Illustration of the heteroscedastic SVM regression on toy dataset. On the left, $\lambda_1 = 1$, $\lambda_2 = 10$ and $\nu = 0.2$, the algorithm successfully regresses the data. On the right, $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\nu = 0.2$, the algorithm fails to regress the data as ϵ becomes negative.

Proof The proof is essentially identical to that of Lemma 3 and that of [12]. However note that the flexibility in ϵ and potential $\epsilon(x) < 0$ lead to additional complications. However, if both f and $\epsilon(x)$ have well behaved entropy numbers, then also $f \pm \epsilon$ 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 $\epsilon \geq 0$. Since $\alpha_i, \alpha_i^* \geq 0$ for all i it follows that $\epsilon(x)$ contains only nonnegative coefficients, which proves the last part of the claim. ■

Note that in principle we could enforce $\epsilon(x_i) \geq 0$ 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 (28) 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 (28) is prone to producing estimates where the margin $\epsilon(x) < 0$. 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 5.

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 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 [12] 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 $\epsilon(x) \geq 0$ for all x.

B. Experimental Results

Here we assemble six tables to display the results across the four models. The first three tables report the pinball loss for each data set and the standard devation across the 10 test runs. A lower figure is preferred in each case. NA denotes cases where rqss [9] was unable to produce estimates, due to its construction of the function system.

In the next three tables we measure the ramp loss. In each table a figure close the the intended quantile (10, 50 or 90) is preferred. For further discussion see the Results section of the paper.

Data Set	uncond	linear	rqss	npqr
caution	11.09 ± 2.56	11.18 ± 3.37	**9.18 ± 3.09**	9.22 ± 3.54
ftcollinssnow	**16.31 ± 5.31**	16.55 ± 6.00	17.52 ± 5.12	**16.31 ± 5.31**
highway	11.38 ± 5.79	16.36 ± 9.65	20.51 ± 19.52	**9.62 ± 5.04**
heights	17.20 ± 2.23	15.28 ± 2.21	15.28 ± 2.23	**15.26 ± 2.20**
sniffer	13.98 ± 2.63	6.66 ± 1.67	5.29 ± 1.79	**5.09 ± 1.29**
snowgeese	8.71 ± 4.21	4.64 ± 2.40	4.65 ± 2.44	**3.89 ± 2.64**
ufc	17.03 ± 2.86	10.01 ± 1.35	10.11 ± 1.12	**9.73 ± 1.30**
birthwt	18.31 ± 2.59	18.39 ± 2.39	18.73 ± 2.93	**17.56 ± 3.22**
crabs	18.27 ± 3.36	1.03 ± 0.33	NA	**0.94 ± 0.28**
GAGurine	11.08 ± 1.47	7.22 ± 1.30	5.82 ± 1.03	**5.24 ± 1.31**
geyser	17.11 ± 1.97	11.51 ± 1.15	11.10 ± 1.39	**9.90 ± 1.76**
gilgais	12.88 ± 1.51	5.92 ± 1.59	5.75 ± 1.79	**5.61 ± 1.45**
topo	20.38 ± 8.61	9.22 ± 3.68	8.19 ± 3.53	**5.98 ± 2.35**
BostonHousing	14.07 ± 1.77	6.61 ± 1.05	NA	**5.10 ± 1.31**
CobarOre	17.72 ± 8.95	16.55 ± 6.49	**12.83 ± 6.36**	15.15 ± 7.28
engel	11.93 ± 1.82	6.51 ± 2.33	**5.70 ± 1.17**	5.72 ± 1.20
mcycle	20.03 ± 2.38	17.81 ± 3.43	10.98 ± 2.43	**8.06 ± 2.97**
BigMac2003	8.67 ± 2.40	6.46 ± 2.08	NA	**6.04 ± 2.10**
UN3	18.02 ± 4.53	11.57 ± 2.28	NA	**11.07 ± 3.50**
cpus	5.25 ± 1.75	1.73 ± 0.89	0.74 ± 0.37	**0.59 ± 0.53**

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

Data Set	uncond	linear	rqss	npqr
caution	38.16 ± 10.19	32.40 ± 8.39	23.76 ± 8.09	**22.50 ± 8.98**
ftcollinssnow	41.96 ± 11.03	41.00 ± 11.34	42.28 ± 11.21	**39.15 ± 11.58**
highway	41.86 ± 22.46	39.47 ± 19.26	**26.05 ± 12.27**	26.71 ± 18.26
heights	40.09 ± 2.99	**34.50 ± 2.88**	34.66 ± 2.86	**34.50 ± 2.88**
sniffer	35.64 ± 6.12	12.63 ± 3.88	10.23 ± 2.76	**9.52 ± 2.54**
snowgeese	31.31 ± 15.80	13.23 ± 9.00	10.95 ± 8.82	**9.50 ± 4.44**
ufc	40.17 ± 5.26	23.20 ± 2.64	**21.21 ± 2.68**	21.25 ± 2.44
birthwt	41.13 ± 7.31	38.14 ± 6.97	37.28 ± 5.97	**36.95 ± 7.04**
crabs	41.47 ± 7.03	2.24 ± 0.44	NA	**2.06 ± 0.45**
GAGurine	36.60 ± 4.26	23.61 ± 4.19	16.08 ± 3.14	**12.10 ± 2.29**
geyser	41.28 ± 7.17	32.30 ± 4.55	30.79 ± 3.88	**29.90 ± 5.55**
gilgais	42.02 ± 5.30	16.11 ± 3.91	**11.76 ± 2.92**	12.32 ± 2.80
topo	41.23 ± 15.98	26.13 ± 8.79	18.02 ± 9.43	**14.51 ± 5.51**
BostonHousing	35.63 ± 5.28	17.51 ± 3.54	NA	**11.21 ± 2.37**
CobarOre	42.14 ± 19.73	41.65 ± 18.84	44.24 ± 12.18	**36.67 ± 18.64**
engel	35.83 ± 7.13	13.73 ± 3.15	**13.23 ± 2.05**	13.39 ± 2.47
mcycle	38.73 ± 9.72	38.19 ± 9.16	21.02 ± 5.18	**17.81 ± 5.66**
BigMac2003	34.97 ± 10.89	21.99 ± 7.11	NA	**17.72 ± 8.80**
UN3	40.83 ± 8.81	26.45 ± 4.30	NA	**22.77 ± 3.36**
cpus	23.03 ± 8.61	5.69 ± 2.23	2.49 ± 1.79	**0.99 ± 0.36**

Table 3: Method Comparison: Pinball Loss ($\times 100$, $\tau = 0.5$)

Data Set	uncond	linear	rqss	npqr
caution	23.28 ± 9.63	15.04 ± 3.37	13.19 ± 3.36	**11.73 ± 2.70**
ftcollinssnow	**18.80 ± 4.45**	19.73 ± 6.14	20.18 ± 6.41	18.84 ± 6.32
highway	25.89 ± 13.58	21.83 ± 18.57	17.63 ± 14.94	**11.54 ± 6.52**
heights	17.64 ± 1.28	**15.47 ± 0.85**	15.50 ± 0.91	**15.47 ± 0.85**
sniffer	23.38 ± 9.69	5.82 ± 1.63	5.84 ± 1.57	**5.29 ± 0.93**
snowgeese	26.60 ± 18.81	7.79 ± 8.98	8.51 ± 12.54	**3.69 ± 2.48**
ufc	18.03 ± 2.89	10.94 ± 1.31	10.83 ± 1.51	**10.21 ± 1.74**
birthwt	16.18 ± 3.34	16.13 ± 3.22	16.36 ± 3.72	**15.22 ± 3.28**
crabs	17.09 ± 3.08	0.99 ± 0.24	NA	**0.98 ± 0.23**
GAGurine	22.65 ± 5.14	15.72 ± 5.07	10.57 ± 3.27	**6.42 ± 1.18**
geyser	14.12 ± 2.53	12.83 ± 2.34	12.37 ± 2.47	**11.10 ± 1.63**
gilgais	18.91 ± 1.99	6.75 ± 2.07	**5.07 ± 1.68**	5.53 ± 0.98
topo	16.96 ± 7.12	13.46 ± 11.52	13.16 ± 11.01	**8.96 ± 5.90**
BostonHousing	22.62 ± 5.33	11.59 ± 2.94	NA	**7.02 ± 2.53**
CobarOre	17.21 ± 4.31	21.76 ± 6.03	19.38 ± 5.21	**16.33 ± 6.28**
engel	22.59 ± 6.86	5.43 ± 1.08	5.64 ± 1.81	**5.41 ± 1.15**
mcycle	16.10 ± 3.21	14.16 ± 3.44	10.69 ± 3.57	**7.12 ± 1.95**
BigMac2003	24.48 ± 17.33	13.47 ± 6.21	NA	**12.81 ± 11.20**
UN3	16.36 ± 2.97	10.38 ± 2.19	NA	**8.90 ± 2.02**
cpus	23.61 ± 10.46	2.69 ± 0.57	1.83 ± 2.31	**0.57 ± 0.31**

Table 4: Method Comparison: Pinball Loss (×100, $\tau = 0.9$)

Data Set	uncond	linear	rqss	npqr
caution	**11.0 ± 8.8**	12.0 ± 9.2	16.0 ± 10.7	18.0 ± 14.8
ftcollinssnow	**10.0 ± 9.7**	11.1 ± 9.1	12.2 ± 11.0	**10.0 ± 9.7**
highway	**10.8 ± 15.7**	20.0 ± 23.3	26.7 ± 37.8	20.0 ± 23.3
heights	**9.6 ± 2.8**	10.0 ± 2.4	10.0 ± 2.2	9.8 ± 2.4
sniffer	**7.8 ± 10.1**	13.7 ± 9.6	12.0 ± 13.1	11.4 ± 7.2
snowgeese	**12.5 ± 17.7**	9.7 ± 12.6	9.7 ± 12.6	16.1 ± 20.5
ufc	**9.7 ± 3.9**	9.9 ± 5.4	11.8 ± 4.0	9.9 ± 5.2
birthwt	**10.0 ± 7.8**	12.0 ± 6.7	12.6 ± 5.1	11.7 ± 7.1
crabs	**10.0 ± 8.5**	12.0 ± 9.8	NA	13.0 ± 6.7
GAGurine	10.4 ± 5.1	**9.9 ± 4.7**	10.7 ± 6.4	16.5 ± 10.3
geyser	**9.7 ± 8.3**	11.2 ± 6.2	10.7 ± 6.9	10.8 ± 7.3
gilgais	**9.5 ± 6.9**	10.4 ± 4.9	13.5 ± 4.6	11.3 ± 4.6
topo	**8.9 ± 15.0**	13.4 ± 13.3	16.0 ± 24.6	18.0 ± 17.5
BostonHousing	**9.7 ± 4.7**	11.5 ± 4.6	NA	14.9 ± 4.9
CobarOre	**8.5 ± 14.3**	12.7 ± 22.8	16.1 ± 17.0	14.2 ± 22.9
engel	**10.2 ± 7.1**	9.4 ± 6.8	10.2 ± 7.9	10.3 ± 8.5
mcycle	**10.0 ± 9.6**	11.5 ± 9.1	11.4 ± 9.1	14.3 ± 9.9
BigMac2003	**9.0 ± 11.4**	18.0 ± 22.9	NA	14.3 ± 19.4
UN3	**9.5 ± 10.0**	12.0 ± 9.7	NA	14.5 ± 8.6
cpus	**9.4 ± 8.9**	12.2 ± 10.2	15.3 ± 7.9	15.7 ± 10.1

Table 5: Method Comparison: Ramp Loss(×100, $\tau = 0.1$)

Data Set	uncond	linear	rqss	npqr
caution	52.0 ± 22.5	**49.0 ± 13.7**	51.0 ± 14.5	51.0 ± 19.7
ftcollinssnow	50.6 ± 14.0	**49.7 ± 16.9**	48.6 ± 19.8	51.4 ± 24.3
highway	48.3 ± 31.9	**44.2 ± 38.5**	45.0 ± 38.5	54.2 ± 24.0
heights	**49.3 ± 5.7**	50.1 ± 5.1	49.8 ± 4.9	50.1 ± 5.1
sniffer	**47.8 ± 8.1**	51.0 ± 13.0	51.0 ± 11.8	51.3 ± 15.2
snowgeese	48.1 ± 27.6	49.2 ± 32.7	51.7 ± 26.9	**47.2 ± 18.8**
ufc	**49.2 ± 8.6**	50.0 ± 6.8	51.6 ± 6.8	51.6 ± 5.5
birthwt	48.9 ± 14.3	50.0 ± 14.3	**47.8 ± 13.9**	49.4 ± 10.9
crabs	49.5 ± 10.9	50.5 ± 9.8	NA	**49.5 ± 7.6**
GAGurine	49.2 ± 11.8	50.9 ± 8.0	51.4 ± 17.0	**49.0 ± 15.6**
geyser	**48.6 ± 11.2**	49.8 ± 7.8	49.5 ± 6.8	50.9 ± 10.8
gilgais	**48.7 ± 10.5**	50.0 ± 10.6	49.7 ± 10.0	50.6 ± 8.9
topo	47.7 ± 23.3	**47.7 ± 19.1**	47.7 ± 21.3	56.3 ± 22.5
BostonHousing	**49.7 ± 6.0**	49.6 ± 8.4	NA	50.5 ± 9.8
CobarOre	46.4 ± 23.0	**44.5 ± 22.2**	47.9 ± 27.7	57.9 ± 34.6
engel	50.9 ± 9.0	49.7 ± 8.6	**49.6 ± 8.6**	50.5 ± 10.5
mcycle	**49.1 ± 11.7**	51.3 ± 11.6	51.4 ± 13.7	53.1 ± 11.0
BigMac2003	49.3 ± 14.6	50.0 ± 20.8	NA	**44.3 ± 23.0**
UN3	**49.4 ± 9.6**	50.6 ± 11.8	NA	50.3 ± 14.2
cpus	**49.2 ± 13.7**	51.3 ± 18.3	49.7 ± 11.7	49.2 ± 14.9

Table 6: Method Comparison: Ramp Loss (×100, $\tau = 0.5$)

Data Set	uncond	linear	rqss	npqr
caution	90.0 ± 10.5	90.0 ± 10.5	89.0 ± 12.0	**82.0 ± 10.3**
ftcollinssnow	90.3 ± 11.1	89.2 ± 12.9	**88.3 ± 12.9**	89.2 ± 12.9
highway	89.2 ± 22.2	64.2 ± 32.4	**61.7 ± 29.4**	75.8 ± 26.8
heights	**89.5 ± 2.3**	90.0 ± 1.8	89.8 ± 1.8	90.0 ± 1.8
sniffer	89.4 ± 7.0	87.6 ± 12.4	86.8 ± 10.4	**85.1 ± 11.5**
snowgeese	88.9 ± 12.4	**85.0 ± 17.5**	**85.0 ± 17.5**	90.3 ± 17.2
ufc	89.8 ± 5.1	90.3 ± 5.2	88.5 ± 6.3	**87.4 ± 4.4**
birthwt	88.7 ± 9.7	**87.6 ± 10.0**	88.0 ± 9.0	88.9 ± 9.1
crabs	89.0 ± 9.7	**87.0 ± 8.9**	NA	87.0 ± 9.2
GAGurine	89.5 ± 3.8	89.8 ± 6.3	89.4 ± 5.6	**87.0 ± 7.5**
geyser	**88.5 ± 5.6**	89.4 ± 6.4	90.4 ± 6.0	89.3 ± 4.4
gilgais	89.1 ± 6.0	88.3 ± 4.5	87.1 ± 6.7	**86.1 ± 5.7**
topo	89.1 ± 15.0	87.1 ± 14.8	**85.7 ± 19.4**	87.1 ± 14.8
BostonHousing	90.1 ± 4.4	88.8 ± 6.1	NA	**85.4 ± 4.8**
CobarOre	89.1 ± 15.7	85.8 ± 16.7	**79.1 ± 22.6**	86.7 ± 17.2
engel	**88.9 ± 6.3**	90.0 ± 6.6	89.1 ± 6.9	90.0 ± 5.5
mcycle	88.6 ± 7.6	88.8 ± 7.4	87.7 ± 7.4	**86.6 ± 5.5**
BigMac2003	89.3 ± 8.0	84.3 ± 16.0	NA	**72.7 ± 30.1**
UN3	88.0 ± 14.8	86.7 ± 9.8	NA	**85.8 ± 10.4**
cpus	89.3 ± 7.1	87.8 ± 7.7	82.6 ± 6.4	**82.5 ± 11.8**

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