A note on constrained nonparametric quantile regression

Yen-Chi Chen
University of Washington
April 1, 2020

Here we review the ideas of nonparametric quantile regression via using the idea of reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. One particular reason of using the RKHS is that it includes the popular spline family. We will explain the quantile regression approach and study two special constraints: non-cross quantile constraint and monotonicity constraint. At the end of this note, we will introduce another idea using rearranging. Most of the contents are based on the following paper:

Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. Journal of machine learning research, 7(Jul), 1231-1264.

Let $Y \in \mathbb{R}$ be the response variable and $X \in \mathbb{R}$ be the (univariate) covariate. The quantile regression aims at finding the conditional quantiles of $Y|X = x$. Formally, given a quantile level $\tau \in [0,1]$, the quantile regression aims at finding

$$m(x;\tau) = F^{-1}(\tau|X = x),$$

where $F(y|x) = P(Y \leq y|X = x)$ is the cumulative distribution function.

The problem of quantile regression is: suppose we observe IID observations

$$(X_1,Y_1),\cdots,(X_n,Y_n),$$

how can we estimate $m(x;\tau)$?

Loss function of quantile regression. The quantile regression can be written as a risk minimization problem via a particular loss function. For a random variable Y, its τ-quantile can be defined as

$$m(\tau) = \mathsf{argmin}_y \mathbb{E}(L_\tau(Y-y)), \quad L_\tau(y) = \begin{cases} \tau y, & y \geq 0, \\ (\tau-1)y, & y < 0. \end{cases} \tag{1}$$

Namely, $L_\tau(y)$ is an asymmetric loss function and the quantile τ controls the slope of this loss. When $\tau = 0.5$, this reduces to the L_1 loss function and the minimizer is the median. As a result, you can easily verify that under suitable conditions (no probability mass at the quantile),

$$m(x;\tau) = \mathsf{argmin}_y \mathbb{E}(L_\tau(Y - f(X))|X = x).$$

Thus, many quantile regression approaches attempt to find the quantile curve by solving

$$\widehat{m}(x;\tau) = \mathsf{argmin}_{m\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L_\tau(Y_i - f(X_i)) + \lambda \cdot \mathsf{pen}(f),$$

where $\mathsf{pen}(f)$ is a penalty term on the (smoothness of) function f and $\mathcal{F}$ is some class of functions (such as splines, RKHS, linear functions, etc) and $\lambda > 0$ is the tuning parameter on the penalty.

1 Quantile regression with RKHS

We first consider the problem where τ is a given quantile. In the next section we will generalize it into the case of multiple quantiles. The RKHS approach for quantile regression starts with decomposing m into $m = g + b$, where $b \in \mathbb{R}$ is the intercept and $g \in \mathcal{H}$ is a function in the RKHS space.

When using RKHS, a natural choice of penalty is the penalty in RKHS, so we rewrite our estimator as

$$\begin{aligned} \widehat{m}(x;\tau) &= \widehat{g}(x) + \widehat{b} \\ (\widehat{g}, \widehat{b}) &= \mathsf{argmin}_{g \in \mathcal{H}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L_\tau(Y_i - g(X_i) - b) + \lambda \|g\|_{\mathcal{H}}, \end{aligned} \tag{2}$$

where $\|g\|_{\mathcal{H}}$ is the RKHS norm of g.

Recall that for any function $g \in \mathcal{H}$, it can always be represented as

$$g(x) = \langle \omega, \phi(x) \rangle,$$

where $\phi(x)$ is the basis function in $\mathcal{H}$ for each x and ω is the coefficient (function) such that the kernel function $K(x,y) = \langle \phi(x), \phi(y) \rangle$.

As a result, we can rewrite the minimization problem of equation (2) as

$$\mathsf{min}_{b \in \mathbb{R}, \omega} \quad \frac{1}{n} \sum_{i=1}^{n} L_\tau(Y_i - \langle \omega, \phi(X_i) \rangle - b) + \frac{\lambda}{2} \|g\|^2_{\mathcal{H}}. \tag{3}$$

Note that we set the tuning parameter to be $\frac{\lambda}{2}$ to simplify derivations later.

The fact that the loss function L_τ has two parts makes the analysis not simple. However, we can simplify the problem by introducing two sets of slack variables $\xi, \xi^* \in \mathbb{R}^n$ and rewrite the minimization problem in equation (3) as

$$\begin{aligned} &\mathsf{min}_{\omega, \xi, \xi^*} \quad \frac{1}{2}\|\omega\|^2 + \frac{1}{\lambda n} \left(\sum_{i=1}^{n} \tau \xi_i + (1-\tau)\xi_i^* \right) \\ &\text{s.t.} \quad (Y_i - \langle \omega, \phi(X_i) \rangle - b) \leq \xi_i, \quad -(Y_i - \langle \omega, \phi(X_i) \rangle - b) \leq \xi_i^*, \quad \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{4}$$

for each $i = 1, \cdots, n$. You can easily verify that the two minimization problems are the same. The slack variable can be interpreted as an activation version of the loss function $L_\tau(y)$. If the part $y \geq 0$ in $L_\tau(y)$, then ξ_i will represents the loss and $\xi_i^* = 0$ (and vice versa for $y < 0$).

To solve the problem of equation (4), we use Lagrangian multipliers. Let $\mu, \mu^*, \eta, \eta^* \in \mathbb{R}^n$ be the Lagrangian multipliers. Then we can rewrite the above minimization problem as the Lagrangian

$$\begin{aligned} \mathcal{L} = &-\frac{1}{2}\|\omega\|^2 - \frac{1}{\lambda n} \left(\sum_{i=1}^{n} \tau \xi_i + (1-\tau)\xi_i^* \right) \\ &+ \sum_{i=1}^{n} \mu_i (\xi_i - Y_i + \langle \omega, \phi(X_i) \rangle + b) + \sum_{i=1}^{n} \mu_i^* (\xi_i^* + Y_i - \langle \omega, \phi(X_i) \rangle - b) \\ &+ \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*). \end{aligned} \tag{5}$$

with constraints that the multipliers are non-negative (KKT conditions). To find the minimal, we have to take the derivative with respect to ω, ξ, ξ^* and set them to be 0. This gives us a few useful equality constraints

$$\begin{aligned}
\frac{\partial}{\partial \xi_i}\mathcal{L} &= 0 = \frac{\tau}{\lambda n} - \mu_i - \eta_i \\
\frac{\partial}{\partial \xi_i^*}\mathcal{L} &= 0 = \frac{1-\tau}{\lambda n} - \mu_i^* - \eta_i^* \\
\frac{\partial}{\partial \omega}\mathcal{L} &= 0 = \omega - \sum_{i=1}^n (\mu_i - \mu_i^*)\phi(X_i) \\
\frac{\partial}{\partial b}\mathcal{L} &= 0 = \sum_{i=1}^n \mu_i - \mu_i^*.
\end{aligned}$$

The third equation gives us a closed-form expression of ω:

$$\omega = \sum_{i=1}^n (\mu_i - \mu_i^*)\phi(X_i) = \sum_{i=1}^n \alpha_i \phi(X_i) = \alpha^T \phi_n$$

and the first two equalities (with the multipliers being non-negative) show constraints on α_i:

$$\frac{\tau - 1}{\lambda n} \leq \alpha_i \leq \frac{\tau}{\lambda n}$$

and the last equality gives the constraint $\sum_i \alpha_i = 0$. Note that $\phi_n = (\phi(X_1), \cdots, \phi(X_n))^T \in \mathbb{R}^n$.

Thus, at the stationary point, the Lagrangian can be written as

$$\mathcal{L}^* = \frac{1}{2}\alpha^T \phi_n \phi_n^T \alpha - \mathbb{Y}_n^T \alpha$$

subject to $1_n^T \alpha = 0$ and $\frac{\tau-1}{\lambda n} \leq \alpha_i \leq \frac{\tau}{\lambda n}$ and $\mathbb{Y}_n = (Y_1, \cdots, Y_n)^T$. The solution is $\widehat{\alpha}$ that minimizes $\mathcal{L}^*$ and we can write $\phi_n \phi_n^T = \mathbf{K}$ as the Gram matrix so the solution is

$$\begin{aligned}
\widehat{\alpha} = \mathsf{argmin}_{\alpha \in \mathbb{R}^n} &\quad \frac{1}{2}\alpha^T \mathbf{K} \alpha - \mathbb{Y}_n^T \alpha \\
\text{s.t.} \quad 1_n^T \alpha = 0, &\quad \frac{\tau - 1}{\lambda n} \leq \alpha_i \leq \frac{\tau}{\lambda n}.
\end{aligned}$$

When we have $\widehat{\alpha}$, we can compute $\widehat{g}(X_i)$ using

$$\widehat{g}(X_i) = \langle \widehat{\omega}, \phi(X_i) \rangle = \langle \widehat{\alpha}^T \phi_n, \phi(X_i) \rangle = \sum_{j=1}^n \widehat{\alpha}_j \langle \phi(X_j), \phi(X_i) \rangle = \sum_{j=1}^n \widehat{\alpha}_j K(X_j, X_i).$$

Similar for the vector $\widehat{g}_n = (\widehat{g}(X_1), \cdots, \widehat{g}(X_n))$, it can be written as

$$\widehat{g}_n = \mathbf{K}\widehat{\alpha}.$$

Note that for any arbitrary point x, $\widehat{g}(x) = \sum_{i=1}^n \widehat{\alpha}_i K(X_i, x)$. The intercept can be estimated by plug-in $\widehat{g}_n$ and minimizes the remaining empirical risk.

2 Non-crossing constraint

In many scenarios, we will not just interested in a particular quantile. Instead, we may want to estimate several quantile curves $\tau_1 < \tau_2 < \cdots < \tau_K$. If we apply the above procedure individually to each quantile curve, estimated quantile curves may cross with each other, leading to an undesirable scenario. Here we discuss how to incorporate the non-crossing constraint into the quantile regression problem.

In this case, each estimated quantile curve would be written as

$$\widehat{m}_k(x) = \widehat{m}(x;\tau_k) = \widehat{g}_k(x) + \widehat{b}_k$$

for each $k = 1, \cdots, K$. The non-crossing constraint can be expressed as the constraint that $\widehat{m}_k(x) \geq \widehat{m}_{k-1}(x)$ which is equivalent to

$$\langle \widehat{\omega}_k, \phi(x) \rangle + \widehat{b}_k \geq \langle \widehat{\omega}_{k-1}, \phi(x) \rangle + \widehat{b}_{k-1}.$$

In RKHS space, it is not easy to impose such constraint for all point x. A relaxed version of the above constraint is

$$\langle \widehat{\omega}_k, \phi(X_i) \rangle + \widehat{b}_k \geq \langle \widehat{\omega}_{k-1}, \phi(X_i) \rangle + \widehat{b}_{k-1} \tag{6}$$

for each $i = 1, \cdots, n$ and $k = 2, \cdots, n$. Namely, we only place the non-crossing constraint at every observed point.

We can easily incorporate equation (6) into equation (4). A complication here is that the optimization problem of different quantiles are no longer separable. So we have to solve a joint optimization problem:

$$\begin{aligned} &\min_{\omega,\xi,\xi^*} \quad \sum_{k=1}^{K} \frac{1}{2}\|\omega_k\|^2 + \frac{1}{\lambda n}\left(\sum_{i=1}^{n} \tau_k \xi_{k,i} + (1-\tau_k)\xi^*_{k,i}\right) \\ &\text{s.t.} \quad (Y_i - \langle \omega_k, \phi(X_i)\rangle - b_k) \leq \xi_{k,i}, \quad -(Y_i - \langle \omega_k, \phi(X_i)\rangle - b_k) \leq \xi^*_{k,i}, \quad \xi_{k,i}, \xi^*_{k,i} \geq 0, \\ &\qquad \langle \omega_k, \phi(X_i)\rangle + b_k \geq \langle \omega_{k-1}, \phi(X_i)\rangle + b_{k-1}. \end{aligned} \tag{7}$$

The additional constraint will introduce an additional Lagrangian multiplier θ. The Lagrangian form of the above problem is

$$\begin{aligned} \mathcal{L} &= \sum_{k=1}^{K} \mathcal{L}_k \\ \mathcal{L}_k &= -\frac{1}{2}\|\omega_k\|^2 - \frac{1}{\lambda n}\left(\sum_{i=1}^{n} \tau_k \xi_{k,i} + (1-\tau_k)\xi^*_{k,i}\right) \\ &\quad + \sum_{i=1}^{n} \mu_{k,i}(\xi_{k,i} - Y_i + \langle \omega_k, \phi(X_i)\rangle + b_k) + \sum_{i=1}^{n} \mu^*_{k,i}(\xi^*_{k,i} + Y_i - \langle \omega_k, \phi(X_i)\rangle - b_k) \\ &\quad + \sum_{i=1}^{n} (\eta_{k,i}\xi_{k,i} + \eta^*_{k,i}\xi^*_{k,i}) \\ &\quad + \sum_{i=1}^{n} \theta_{k,i}(\langle \omega_k, \phi(X_i)\rangle - \langle \omega_{k-1}, \phi(X_i)\rangle + b_k - b_{k-1}) \end{aligned} \tag{8}$$

for $k = 1, 2, \cdots, K$ and $\mathcal{L}_1$ takes the same form as equation (5) and we set $\theta_0 = 0$.

The derivative with respect to ξ and ξ^* are the same–they provide constraints on the range of $\alpha_{k,i} = \mu_{k,i} - \mu^*_{k,i}$. The additional inequality changes the form of ω_k:

$$\frac{\partial}{\partial \omega_k}\mathcal{L} = 0 = \omega_k - \sum_{i=1}^{n}(\mu_{k,i} - \mu^*_{k,i} + \theta_{k,i} - \theta_{k+1,i})\phi(X_i)$$
$$\frac{\partial}{\partial b_k}\mathcal{L} = 0 = \sum_{i=1}^{n}\mu_{k,i} - \mu^*_{k,i} + \theta_{k,i} - \theta_{k+1,i}.$$

Using the notation $\alpha_k = (\alpha_{k,1}, \cdots, \alpha_{k,n})^T$ and $\theta_k = (\theta_{k,1}, \cdots, \theta_{k,n})$, we can write

$$\omega_k = (\alpha_k + \theta_k - \theta_{k+1})^T \phi_n.$$

Using the above stationary points conditions into the Lagrangian, we obtain the following criterion of finding α and θ:

$$\begin{aligned} \mathcal{L}^* &= \sum_k \mathcal{L}^*_k \\ \mathcal{L}^*_k &= \frac{1}{2}\|(\alpha_k + \theta_k - \theta_{k+1})^T\phi_n\|^2 - \mathbb{Y}^T\alpha_k \\ &= \frac{1}{2}\alpha_k^T\mathbf{K}\alpha_k + \alpha_k^T\mathbf{K}(\theta_k - \theta_{k+1}) + (\theta_k - \theta_{k+1})^T\mathbf{K}(\theta_k - \theta_{k+1}) - \mathbb{Y}^T\alpha_k. \end{aligned} \tag{9}$$

We solve the above minimization problem with constraints

$$1_n^T(\alpha_k + \theta_k - \theta_{k+1}) = 0, \quad \frac{\tau_k - 1}{\lambda n} \leq \alpha_{k,i} \leq \frac{\tau_k}{\lambda n}$$

to obtain $\widehat{\alpha}_k$ and $\widehat{\theta}_k$ for each $k = 1, \cdots, K-1$, which also gives us

$$\widehat{g}_k(x) = \sum_{i=1}^{n}(\widehat{\alpha}_{k,i} + \widehat{\theta}_{k,i} - \widehat{\theta}_{k+1,i})K(X_i, x)$$

or

$$\widehat{g}_{k,n} = \mathbf{K}(\widehat{\alpha}_k + \widehat{\theta}_k - \widehat{\theta}_{k+1}) \in \mathbf{R}^n.$$

Note that we can also requires the non-crossing constraint to any set of points $x_1, \cdots, x_L$ that are not necessarily the same as the observed covariates. The kernel matrix in equation (9) will change accordingly.

3 Monotonicity constraint

In addition to the non-crossing constraint, we can also incorporate the monotonicity constraint easily. To simplify the problem, we consider a single quantile again. The same idea can also be applied to multiple quantiles (even with non-crossing constraints). Suppose we want to constraint that quantile curve to be non-decreasing, which translates into

$$g'(x) \geq 0$$

for all x. Again, enforcing this constraint for all points is not easy in the RKHS space. So we relax that constraint by requiring it only on the observed data points, i.e.,

$$g'(X_i) \geq 0$$

for all $i = 1, \cdots, n$.

Here is an interesting property of the derivative of $g(x)$:

$$g'(x) = \frac{d}{dx}\langle \omega, \phi(x) \rangle = \langle \omega, \frac{d}{dx}\phi(x) \rangle = \langle \omega, \phi'(x) \rangle.$$

Thus, the monotonicity constraints becomes

$$\langle \omega, \phi'(X_i) \rangle \geq 0$$

for each $i = 1, \cdots, n$. With this, we can rewrite equation (4) as

$$\begin{aligned} \min_{\omega,\xi,\xi^*} \quad & \frac{1}{2}\|\omega\|^2 + \frac{1}{\lambda n}\left(\sum_{i=1}^{n} \tau\xi_i + (1-\tau)\xi_i^*\right) \\ \text{s.t.} \quad & (Y_i - \langle \omega, \phi(X_i) \rangle - b) \leq \xi_i, \quad -(Y_i - \langle \omega, \phi(X_i) \rangle - b) \leq \xi_i^*, \quad \xi_i, \xi_i^* \geq 0 \\ & \langle \omega, \phi'(X_i) \rangle \geq 0. \end{aligned} \tag{10}$$

The additional constraint introduces a new Lagrangian multiplier ζ and the Lagrangian will be

$$\begin{aligned} \mathcal{L} = & -\frac{1}{2}\|\omega\|^2 - \frac{1}{\lambda n}\left(\sum_{i=1}^{n} \tau\xi_i + (1-\tau)\xi_i^*\right) \\ & + \sum_{i=1}^{n} \mu_i(\xi_i - Y_i + \langle \omega, \phi(X_i) \rangle + b) + \sum_{i=1}^{n} \mu_i^*(\xi_i^* + Y_i - \langle \omega, \phi(X_i) \rangle - b) \\ & + \sum_{i=1}^{n} (\eta_i\xi_i + \eta_i^*\xi_i^*) + \sum_{i=1}^{n} \zeta_i \langle \omega, \phi'(X_i) \rangle. \end{aligned} \tag{11}$$

Again, taking derivatives with respect to μ, η does not change here so they place constraints on $\alpha_i = \mu_i - \mu_i^*$. Also, the derivative of b is the same so we have constraint $\alpha^T 1_n = 0$. What changes is the derivative with respect to ω :

$$\frac{\partial}{\partial \omega}\mathcal{L} = 0 = \omega - \sum_{i=1}^{n} \alpha_i \phi(X_i) + \zeta_i \phi'(X_i).$$

Recall that $\phi_n = (\phi(X_1), \cdots, \phi(X_n))$ and $\psi_n = (\phi'(X_1), \cdots, \phi'(X_n))$. Then we can write

$$\omega = \alpha^T \phi_n + \zeta^T \psi_n.$$

Note that we use notation ψ_n instead of ϕ_n' because ϕ_n itself is a function so people may thought that the derivative ϕ_n' is referring to the derivative of ϕ_n.

Putting the stationary points into equation (11), there are only 3 terms left:

$$\begin{aligned} \mathcal{L}^* &= -\frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{n} \underbrace{(\mu_i - \mu_i^*)}_{=\alpha_i} \langle \omega, \phi(X_i) \rangle + \sum_{i=1}^{n} \zeta_i \langle \omega, \phi'(X_i) \rangle \\ &= -\frac{1}{2}\|\omega\|^2 + \langle \omega, \alpha^T \phi_n \rangle + \langle \omega, \zeta^T \psi_n \rangle. \end{aligned}$$

Inserting $\omega = \alpha^T \phi_n + \zeta^T \psi_n$ into the above equation, then we obtain

$$
\begin{aligned}
\mathcal{L}^* &= \frac{1}{2}\left(\alpha^T \mathbf{K}\alpha + \alpha^T D_2\mathbf{K}\zeta + \zeta^T D_1\mathbf{K}\alpha + \zeta^T D_1 D_2\mathbf{K}\zeta\right) - \mathbb{Y}^T\alpha \\
&= \frac{1}{2}\begin{bmatrix}\alpha \\ \zeta\end{bmatrix}^T \begin{bmatrix}\mathbf{K} & D_2\mathbf{K} \\ D_1\mathbf{K} & D_1D_2\mathbf{K}\end{bmatrix}\begin{bmatrix}\alpha \\ \zeta\end{bmatrix} - \mathbb{Y}^T\alpha,
\end{aligned} \tag{12}
$$

where $D_1\mathbf{K} = \psi_n^T \phi_n$ is the Gram matrix obtained by the partial derivative kernel $\frac{d}{dx_1}K(x_1,x_2)$ and similarly for $D_2\mathbf{K} = \phi_n^T\psi_n$ and $D_1D_2\mathbf{K} = \psi_n^T\psi_n$.

The estimators $\widehat{\alpha}$ and $\widehat{\zeta}$ is obtained by minimizing equation (11) subject to the constraints on α and $\zeta_i \geq 0$. With the estimators $\widehat{\alpha}$ and $\widehat{\zeta}$, we obtain

$$
\widehat{g}(x) = \sum_{i=1}^n \widehat{\alpha}_i K(X_i,x) + \widehat{\zeta}_i K_1(X_i,x),
$$

where $K_1(x_1,x_2) = \frac{\partial}{\partial x_1}K(x_1,x_2)$.

4 Monotonicity constraint: rearrangement approach

Rearrangement provides an alternative approach to quantile regression (and other regression estimator) that encourages monotonicity constraint. The idea of rearrangement comes from the following paper:

- Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. Biometrika, 96(3), 559-575.

It is a method that converts an estimated function to a function with a monotonicity constraint. For simplicity, we assume that we want to fit a quantile regression function that is non-decreasing.

Suppose that the covariates are supported on the interval $[0,1]$ and we have an initial quantile regression estimator $\widehat{m}_\tau(x)$ that is not necessarily non-decreasing. This estimator may be the one from Section 1.

For any function $f:[0,1]\mapsto\mathbb{R}$, its rearrangement is the function

$$
\begin{aligned}
f^\dagger(x) &= \inf\left\{y : \int I(f(u)\leq y)du \geq x\right\} \\
&= \inf\left\{y : \mathsf{Vol}(L_y) \geq x\right\},
\end{aligned}
$$

where $L_y = \{z : f(z)\leq y\}$ is the lower-level set of f at the threshold y. It can be easily seen that $f^\dagger(x)$ is a non-decreasing function and has the same range as f. Moreover, for any threshold $y>0$, the following level sets

$$
L_y = \{z : f(z)\leq\lambda\}, \quad L_y^\dagger = \{z : f^\dagger(z)\leq\lambda\}
$$

have the same volume

$$
\mathsf{Vol}(L_y) = \mathsf{Vol}(L_y^\dagger)
$$

if $f'(x) > 0$ and is differentiable everywhere on $[0,1]$. Namely, $f^\dagger$ and f are similar in terms of the size of regions below any threshold.

In Chernozhukov et al. (2009), the authors proved that if the true function f_0 is non-decreasing, then for any function f, its rearrangement satisfies

$$\left(\int |f^\dagger(x) - f_0(x)|^p dx\right)^{1/p} \leq \left(\int |f(x) - f_0(x)|^p dx\right)^{1/p}$$

for all $p \in [1,\infty]$. Namely, the rearrangement may improve the accuracy of an initial estimator if the truth is non-decreasing. Note that the authors also provided a strict inequality under additional conditions.

Thus, if we know that the true quantile regression $m_\tau(x)$ is non-decreasing, we can always use rearrangement to improve the accuracy.