Optimal Estimation for Quantile Regression with Functional Response

Xiao Wang, Purdue University

Mathematical and Statistical Challenges in Neuroimaging Data Analysis

Collaborators

SAMSI CCNS

Zhengwu Zhang,	SAMSI
Linglong Kong,	University of Alberta
Hongtu Zhu,	UNC Chapel Hill

Functional Regression with Functional Response

- Functional Regression (Morris 2015)
- Functional Response (Hongtu Zhu ...):

$$Y_i(s) = X_i^T \beta(s) + \eta_i(s), i = 1, \ldots, n.$$

- Recover the conditional mean of $Y(s)$ given X and the location s.
- Various imaging segmentation and registration methods end up with preprocessing results non-consistent or with errors.
- The error distributions are unknown, assuming Gaussian for convenience in many applications though.
- The variances of errors are varying spatially within the brain. Quantile regression (QR) is able to give a full picture of the data. These features make QR more appealing than its cousin, the ordinary least squares.
- In this paper, we would like to recover the $100\tau\%$ quantile of the conditional distribution of $Y(s)$ given X and the location s.

Quantile Regression

- Quantile Regression (Koenker and Basset 1978) vs. Mean Regression

$$y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n.$$

- Quadratic function vs. Check function:

$$\rho_\tau(r) = \begin{cases} \tau r & \text{if } r > 0 \\ -(1-\tau)r & \text{otherwise} \end{cases}$$

- Quantile regression provides better estimators than mean regression WHEN
  - Data are skewed
  - Data contain outliers
- Quantile regression does not require specifying any error distribution.
- Many nonparametric and semiparametric quantile regression models ... (Koenker 2005; ...)

ADNI DTI Data

- **Dataset**: 203 subjects from ADNI
- **Response:** mean Fractional Anisotropy (FA) values along midsagittal corpus callosum skeleton (TBSS pipeline).
- **Covariates:** Gender, Age, Alzheimer's Disease Assessment Scale, Mini-Mental State Examination.

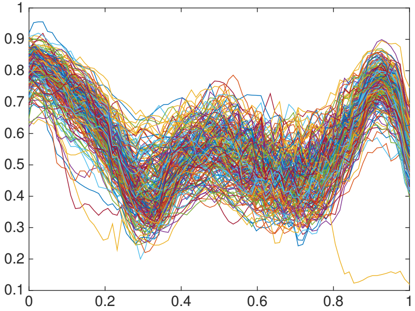


Figure : FA curves along corpus callosum skeleton.

ADNI Hippocampus Image Data

- **Dataset:** 403 subjects from ADNI
- **Response:** Hippocampus images
- **Covariates:** Gender, Age, and Behavior score

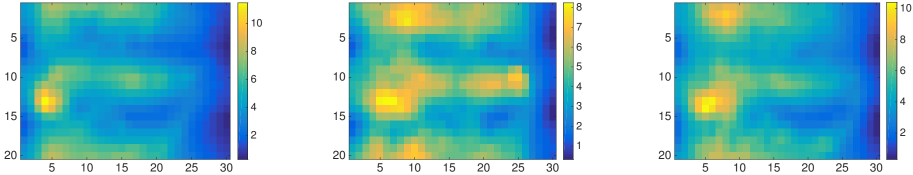


Figure : Observed left hippocampus images.

Quantile Regression with Functional Response

- For a given $\tau \in (0,1)$, consider a quantile regression model with varying-coefficients and functional responses,

$$Y(s) = X^T \beta_\tau(s) + \eta_\tau(s)$$

- $\eta_\tau(\cdot)$ is a stochastic process whose τth quantile is zero for a fixed s given X.
- The conditional quantile function of $Y(s)$ given X for any $\tau \in (0,1)$ can be expressed by

$$Q_{Y(s)}(\tau|X) = X^T \beta_\tau(s)$$

- The unknown parameters $\beta_\tau = (\beta_1, \ldots, \beta_p)$, where $\beta_k \in \mathcal{H}(K)$, a RKHS generated by a pd kernel K.

$$K(s,t) = (1 + \langle s,t \rangle)^d, \quad K(s,t) = \exp(-\|s-t\|^2/2\sigma^2)$$

- Suppose that we observe $(X_i, Y_i(s_{ij}))$ for subjects $i = 1, \ldots, n$ and locations $s_{i1}, \ldots, s_{im_i}$. Our goal is to investigate the estimation of the coefficient functions $\beta_{\tau k}$, $k = 1, \ldots, p$.

Loss Function

- Fixed design: the functional response are observed at the same locations across curves, that is, $m_1 = m_2 = \cdots = m_n := m$ and $s_{1j} = s_{2j} = \cdots = s_{jn} := s_j$ for $j = 1, \ldots, m$.
- Random design: the s_{ij} are independently sampled from a distribution $\pi(s)$.
- L_2-distance: For two function vectors $f_1, f_2 \in \mathcal{F}^p$, define

$$\left\| f_1 - f_2 \right\|_{s,2}^2 = \begin{cases} \dfrac{1}{m} \displaystyle\sum_{j=1}^{m} \sum_{k=1}^{p} (f_{1k}(s_j) - f_{2k}(s_j))^2 & \text{fixed design} \\ \displaystyle\int_{\mathcal{S}} \sum_{k=1}^{p} (f_{1k}(s) - f_{2k}(s))^2 \pi(s) ds & \text{random design} \end{cases}$$

- We measure the accuracy of the estimation of $\hat{\beta}_\tau$ by

$$\mathcal{E}_{n\tau}(\hat{\beta}_\tau, \beta_\tau) = \left\| \hat{\beta}_\tau - \beta_\tau \right\|_{s,2}^2.$$

Rate of Convergence: Lower Bound

- Fix $\tau \in (0,1)$. Suppose the eigenvalues $\{\rho_k : k \geq 1\}$ of the reproducing kernel K satisfies $\rho_k \asymp k^{-2r}$ for some constant $0 < r < \infty$. Then
  a. For the fixed design,
  $$\lim_{a_\tau \to 0} \lim_{n,m\to\infty} \inf_{\tilde{\beta}_\tau} \sup_{\beta_\tau \in \mathcal{F}^p} \mathbb{P}\Big(\mathcal{E}_{n\tau}(\tilde{\beta}_\tau, \beta_\tau) \geq a_\tau (n^{-1} + m^{-2r})\Big) = 1; \tag{1}$$
  b. For the random design,
  $$\lim_{a_\tau \to 0} \lim_{n,m\to\infty} \inf_{\tilde{\beta}_\tau} \sup_{\beta_\tau \in \mathcal{F}^p} \mathbb{P}\Big(\mathcal{E}_{n\tau}(\tilde{\beta}_\tau, \beta_\tau) \geq a_\tau ((nm)^{-\frac{2r}{2r+1}} + n^{-1})\Big) = 1. \tag{2}$$

  The above infimums are taken over all possible estimators $\tilde{\beta}_\tau$ based on the training data.
- If τ belongs to a compact interval of $(0,1)$, a_τ may not depend on τ.

Rate of Convergence: Fixed Design

- Under the common design, the minimax rate is of the order $m^{-2r} + n^{-1}$. This rate is fundamentally different from the usual nonparametric rate of $(nm)^{2r/(2r+1)}$ (Stone 1982).
- The rate is jointly determined by the sampling frequency m and the number of curves n rather than the total number of observations mn.
- When the functionals are sparsely sampled, that is, $m = O(n^{1/2r})$, the optimal rate is of the order m^{-2r} , solely determined by the sampling frequency. On the other hand, when the sampling frequency is high, that is, $m \gg n^{1/2r}$, the optimal rate remains $1/n$ regardless of m.

Rate of Convergence: Random Design

- Similar to the common design, there is a phase transition phenomenon in the optimal rate of convergence with a boundary at $m = n^{1/2r}$.
- When the sampling frequency m is small, that is, $m = O(n^{1/2r})$, the optimal rate is of the order $(nm)^{2r/(2r+1)}$ which depends jointly on the values of both m and n.
- In the case of high sampling frequency with $m \gg n^{1/2r}$, the optimal rate is always $1/n$ and does not depend on m.

Rate of Convergence

- When m is above the boundary, that is, $m \gg n^{1/2r}$, there is no difference between the fixed and random designs. When m is below the boundary, that is, $m \ll n^{1/2r}$, the random design is always superior to the fixed design in that it offers a faster rate of convergence.

Objective Function

- Penalized estimator: Minimize

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau\Big(Y_i(s_{ij}) - X_i^T\beta(s_{ij})\Big) + \lambda\sum_{k=1}^{p}\|\beta_k\|_K^2$$

- Representer Theorem:

$$\hat{\beta}_k(s) = \sum_{i=1}^{\tilde{m}}\theta_i\xi_i(s) + \sum_{j=1}^{m}\beta_j K(s_j, s), \qquad k = 1, \dots, p$$

- Matrix form: Minimize

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau\Big(Y_{ij} - b_{ij}^T\theta - a_{ij}^T\beta\Big) + \lambda\beta^T\Sigma\beta$$

ADMM Algorithm

- Write the optimization into an equivalent form:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - u_{ij}) + \lambda\beta^T\Sigma\beta \\ \text{subject to} \quad & u_{ij} = b_{ij}^T\theta + a_{ij}^T\beta, i = 1, \ldots, n, j = 1, \ldots, m \end{aligned}$$

- Augmented Lagrangian:

$$\begin{aligned} L_\eta(u, \xi, \theta, \beta) = & \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - u_{ij}) + \lambda\beta^T\Sigma\beta + \sum_{i=1}^{n}\sum_{j=1}^{m} \xi_{ij}(u_{ij} - b_{ij}^T\theta - a_{ij}^T\beta) \\ & + \frac{\eta}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(u_{ij} - b_{ij}^T\theta - a_{ij}^T\beta)^2 \end{aligned}$$

- ADMM update:

$$\begin{aligned} u_{ij}^{k+1} = & \text{argmin}_{u_{ij}} \left(\rho_\tau(Y_{ij} - u_{ij}) + \xi_{ij}^k(u_{ij} - b_{ij}^T\theta^k - a_{ij}^T\beta^k) + \frac{\eta}{2}(u_{ij} - b_{ij}^T\theta^k - a_{ij}^T\beta^k)^2\right) \\ (\theta^{k+1}, \beta^{k+1}) = & \text{argmin}_{\theta,\beta} \left(\lambda\beta^T\Sigma\beta + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\xi_{ij}^k a_{ij}^T\beta + \frac{\eta}{2}(u_{ij}^{k+1} - b_{ij}^T\theta - a_{ij}^T\beta)^2\right)\right) \\ \xi_{ij}^{k+1} = & \xi_{ij}^k + \eta(u_{ij}^{k+1} - b_{ij}^T\theta - a_{ij}^T\beta^{k+1}) \end{aligned}$$

ADMM Algorithm

- consider the proximal operator of ρ_τ with parameter μ and λ such that

$$\operatorname{prox}_{\rho_\tau,\mu,\lambda}(v) = \arg\min_x \left(\rho_\tau(x-\mu) + \frac{1}{2\lambda}(x-v)^2\right). \tag{3}$$

- The solution to (3) can be explicitly obtained, and $x^+ = \operatorname{prox}_{\rho_\tau,\mu,\lambda}(v) = S_{\tau,\mu,\lambda}(v)$, where

$$S_{\tau,\mu,\lambda}(v) = \begin{cases} v - \lambda\tau & v > \mu + \lambda\tau \\ 0 & \mu - \lambda(1-\tau) \le v \le \mu + \lambda\tau \\ v + \lambda(1-\tau) & v < \mu - \lambda(1-\tau). \end{cases}$$

- When $\tau = 1/2$ and $\mu = 0$, $S_{\tau,\mu,\lambda}(\cdot)$ is the well-known soft thresholding operator such that

$$S_{1/2,0,\lambda}(v) = \left(1 - \frac{\lambda}{2|v|}\right)_+ v,$$

(for $v \neq 0$) which is a shrinkage operator.

Choice of Smoothing Parameter

- RCV:

$$RCV = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\rho_\tau(Y_{ij} - X_i^T\hat{\beta}^{[-i]}(s_{ij}))$$

- SIC:

$$SIC(\lambda) = \log\left(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau(Y_{ij} - X_i^T\hat{\beta}(s_{ij}))\right) + \frac{\log(mn)}{2nm}df$$

- GACV:

$$GACV(\lambda) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau(Y_{ij} - X_i^T\hat{\beta}(s_{ij}))}{mn - df}$$

Degrees of Freedom

- Let $\hat{Y}_{ij} = X_i^T \hat{\beta}(s_{ij})$.

$$div(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\partial \hat{Y}_{ij}}{\partial Y_{ij}}$$

- This quantity first appeared under SURE formula (Stein 1981). It can be considered an estimate the effective dimension for a general modeling procedure (Efron 1986; Meyer and Woodroofe 2000).
- Define $\mathcal{E} = \{(i,j) : Y_{ij} - X_i^T \hat{\beta}(s_{ij}) = 0\}$. We show that

$$div(\hat{Y}) = |\mathcal{E}|$$

Rate of Convergence: Upper Bound

- Fix $\tau \in (0,1)$. Suppose the eigenvalues $\{\rho_k : k \geq 1\}$ of the reproducing kernel K satisfies $\rho_k \asymp k^{-2r}$ for some constant $0 < r < \infty$. Then
  a. For the fixed design,
  $$\lim_{A_\tau \to \infty} \lim_{n,m\to\infty} \sup_{\beta_\tau \in \mathcal{F}^p} \mathbb{P}\Big(\mathcal{E}_{n\tau}(\hat{\beta}_\tau, \beta_\tau) \geq A_\tau(n^{-1} + m^{-2r})\Big) = 1; \tag{4}$$
  b. For the random design,
  $$\lim_{A_\tau \to 0} \lim_{n,m\to\infty} \sup_{\beta_\tau \in \mathcal{F}^p} \mathbb{P}\Big(\mathcal{E}_{n\tau}(\hat{\beta}_\tau, \beta_\tau) \geq A_\tau((nm)^{-\frac{2r}{2r+1}} + n^{-1})\Big) = 1. \tag{5}$$
- For τ belonging to a compact interval of $(0,1)$, the result holds uniformly for τ.

1D Simulated Data Analysis

- Data are simulated from the model:

$$y_i(s_j) = x_{i1}\beta_1(s_j) + x_{i2}\beta_2(s_j) + x_{i3}\beta_3(s_j) + \eta_i(s_j, \tau), i = 1, .., n, j = 1, ..., m,$$

where

$$\begin{aligned}
[x_{i1}, x_{12}, x_{i3}] &= [1, \sim Bernoulli(0.5), \sim uniform(0, 1)] \\
[\beta_1(s), \beta_2(s), \beta_3(s)] &= [5s^2, 5(1-s)^4, 2s^2 + 5] \\
\eta_i(s_j) &= v_i(s_j) + \epsilon_i(s_j), \epsilon_i(s_j) \sim N(0, 0.1), v_i \sim GP(0, \Sigma) \\
\eta_i(s_j, \tau) &= \eta_i(s_j) - F^{-1}(\tau), F \text{ is marginal density of } \eta_i(s_j)
\end{aligned}$$

- Use root mean integrated squared error (RMISE) to measure the quality of estimated β_i

$$RMISE_\tau = \left(\frac{1}{m} \sum_{j=1}^{m} \|\hat{\beta}_l(s_j, \tau) - \beta_l(s_j, \tau)\|^2 \right)^{1/2} \quad l = 1, 2, 3,$$

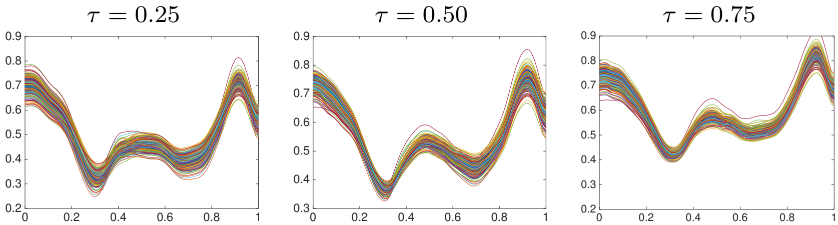
1D Simulated Data Analysis

- Averaged RMISE over 100 simulation runs are reported for $\tau = 0.5$ and $\tau = 0.75$ for sample size $n = 20, 50, 100, 200$

	$\tau = 0.5$			$\tau = 0.75$		
n	$\beta_1(s)$	$\beta_2(s)$	$\beta_3(s)$	$\beta_1(s)$	$\beta_2(s)$	$\beta_3(s)$
20	2.49	2.30	3.82	2.85	2.05	4.36
50	1.55	1.35	2.55	1.43	1.44	2.21
100	1.16	0.91	1.8	1.35	0.95	1.99
200	0.88	0.71	1.36	0.79	0.62	1.30

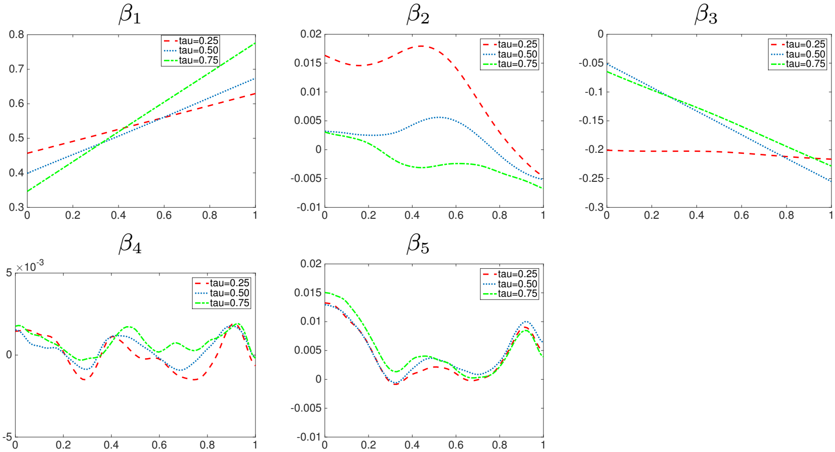
ADNI DTI Data

- Recall:
  - **Response:** y_i=mean Fractional Anisotropy (FA) curves along midsagittal corpus callosum skeleton
  - **Covariates:** x_i= [Gender, Age, Alzheimer's Disease Assessment Scale, Mini-Mental State Examination]
- Predicted τ's quantile for $\tau = 0.25, 0.5$ and 0.75



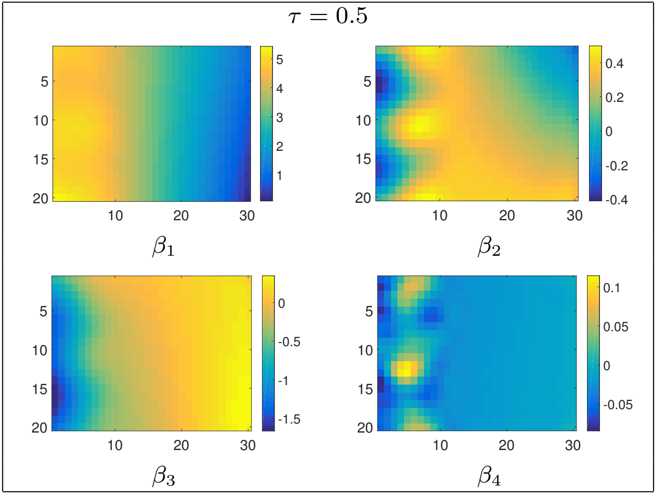
ADNI DTI Data

- Coefficient β_l for $\tau = 0.25, 0.5$ and 0.75



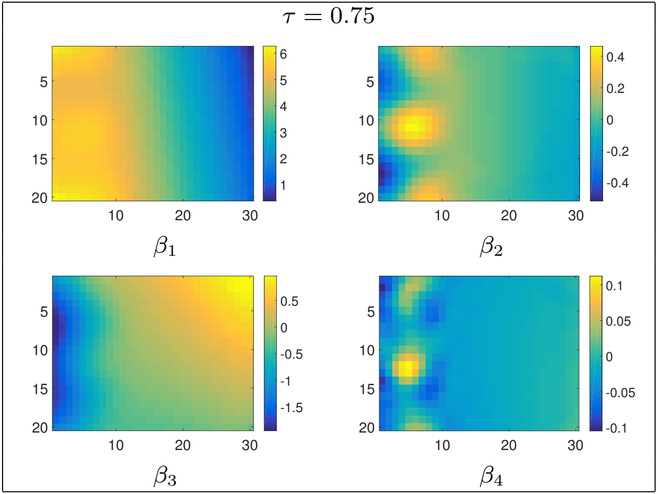
ADNI Hippocampus Image Data

- Coefficient images β_l for $\tau = 0.5$:



ADNI Hippocampus Image Data

- Coefficient images β_l for $\tau = 0.75$:



Conclusion

- Estimation
- Improve the speed of the algorithm
- Inference
- Variable selection: knots selection and variable selection simultaneously