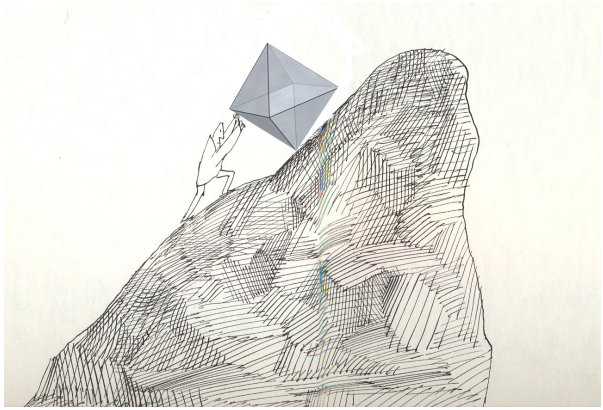
Quantile Regression: An Introductory Overview

Roger Koenker
U. of Illinois

Lan Wang
U. of Minnesota

Xuming He
U. of Michigan

JSM Baltimore: 2 August 2017



Outline of the Overview

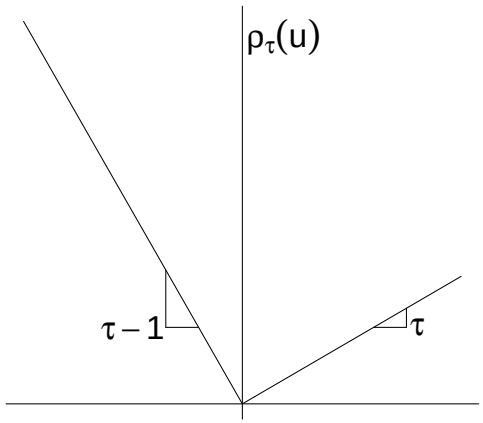
- A (Gentle) Introduction to Quantile Regression Methods – Roger Koenker
- Bootstrap and Other Resampling Methods for Quantile Regression – Xuming He
- Quantile Regression Methods for High Dimensional Data – Lan Wang

Quantile Regression: What is it?

Quantile regression is an evolving set of tools for estimation and statistical inference about models for conditional quantile functions:

$$\hat{g} = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i))$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$. Sorting is replaced by optimization.



Quantile Regression: How does it work?

In the simplest univariate setting asymmetric linear loss requires that,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\rho'_\tau(y_i - g(x_i)) &\equiv n^{-1}\sum_{i=1}^{n}\psi_\tau(y_i - g(x_i)) \\ &= \tau\#\{y_i > \hat{\alpha}\}/n + (\tau-1)\#\{y_i \leqslant \hat{\alpha}\}/n \\ &\approx 0 \end{aligned}$$

so $\hat{\alpha}$ must be chosen so that the proportion of $\{y_i > \hat{\alpha}\}$ is $(1-\tau)$ and the proportion of $\{y_i \leqslant \hat{\alpha}\}$ is τ. i.e. $\hat{\alpha}$ is chosen to balance (counteract) the asymmetry of the loss, Edgeworth (1888).

Quantile Regression: For the Linear Model

When we restrict the class, $\mathcal{G}$, of conditional quantile functions to affine functions we have,

$$\hat{\beta}(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta).$$

Solutions can be efficiently computed by linear programming methods, and are characterized by exact fits to p-element subsets of the data. These fits can be viewed as p-dimensional analogues of the order statistics for the linear model.
As in other forms of regression the covariates, x_i may be expressed as basis expansions in terms of lower dimensional covariates, e.g. $x_i = \varphi_i(z)$.

Quantile Regression Inference

Because $\hat{\beta}(\tau)$ targets the τth conditional quantile only locally, its precision depends crucially only on the conditional density of the response at the τth quantile. Asymptotically,

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{N}(0, H_n^{-1} J_n H_n^{-1}),$$

where

$$J_n = \tau(1-\tau) n^{-1} \sum x_i x_i^\top$$

and

$$H_n = n^{-1} \sum f_i(x_i^\top \beta(\tau)) x_i x_i^\top$$

The latter quantity can be directly estimated, or this can be circumvented by various forms of the bootstrap.

Quantile Regression and Rank Statistics

The linear quantile regression problem has formal dual problem:

$$\hat{a}(\tau) = \max\{y^\top a \mid X^\top a = (1-\tau)X^\top 1,\ a \in [0,1]^n\}$$

- These functions act somewhat like residuals in the quantile regression setting.
- For each observation they indicate the range of $\tau \in [0,1]$ for which y_i lies above or below the fitted quantile regression hyperplane.
- They generalize the rank generating functions of Hájek(1968), and can be used to construct a wide variety of extended rank tests for the linear model as first shown by Gutenbrunner and Jurečková(1992).

Nonparametric Quantile Regression

There are several approaches to the treatment of nonparametric covariate effects,

- Local polynomials (Chaudhuri (1991))

$$\hat{g}(\tau|x) = \text{argmin}_\beta \sum_{i=1}^{n} \rho_\tau(y_i - \sum_{j=0}^{p-1} \beta_j (x_i - x)^j)$$

- Penalization

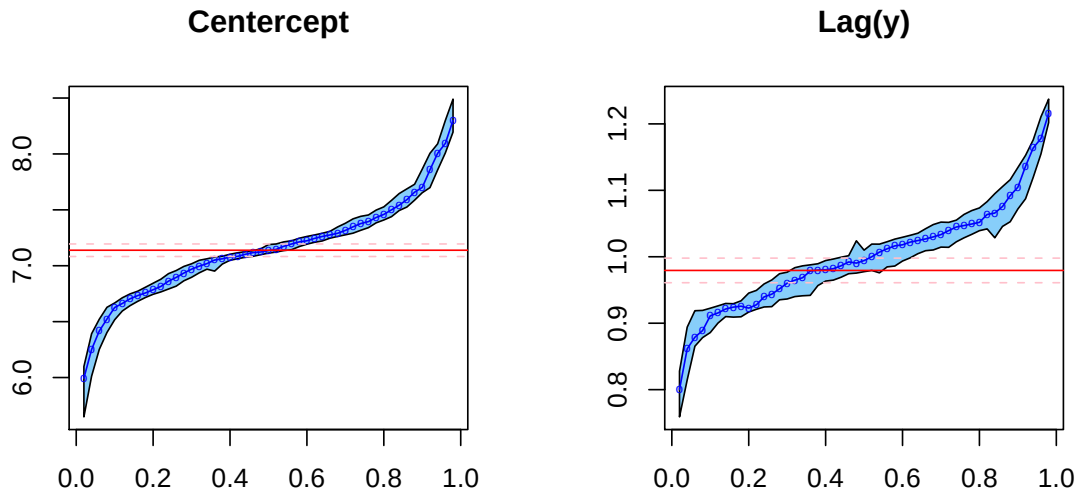
$$\hat{g}_\lambda(\tau) = \text{argmin}_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda P(g).$$

Various penalties are possible. Additive models with total variation roughness penalties have been implemented in the R package **quantreg** employing the function `rqss`.

Quantile Autoregression

Simple autoregressive time-series models can be estimated,

$$\hat{\theta}(\tau) = \text{argmin} \sum \rho_\tau(y_t - \sum \varphi(y_{t-j}, \theta)))$$



Even in the linear φ case there are intriguing stationarity properties of these models. Recently, there has been considerable interest in related frequency domain methods with work by Li, Hagemann, Kley, Hallin and Volgushev, among others.

Quantile Regression for Longitudinal Data

Models for longitudinal data with incidental parameters pose some challenges for quantile regression applications. Initial work on methods for estimating models of the form,

$$Q_{Y_{it}}(\tau|x_{it}) = \alpha_i + x_{it}^\top \beta(\tau),$$

has been done by several authors including Galvao, Lamarche and Kato. Recently, Arellano and Bonhomme have proposed a promising variant on these methods in which incidental parameters are explicitly modeled as functions of covariates.

Quantile Regression Survival Models

Many survival models take the form of transformation models,

$$h(T_i|x_i) = x_i^\top \beta + u_i, \quad \text{with } u_i \sim \text{iid } F.$$

However, the iid error assumption is often questionable. It can be relaxed as,

$$Q_{h(T_i|x_i)}(\tau|x_i) = x_i^\top \beta(\tau).$$

Censoring can be accommodated as in Portnoy (2003) or Peng and Huang (2008). More general censoring schemes can be accommodated employing recent work of Yang, Narisetty and He.

Portfolio Optimization

It is well known that classical Markowitz (mean-variance) portfolio optimization can be reduced to least squares regression. Similarly, optimization of the class of "coherent" measures of risk, like expected shortfall, or lower tail expectation subject to a mean return constraint can be reduced to quantile regression. Note that

$$\begin{aligned}\min_{\alpha} \mathcal{E}_Y \rho_\tau(Y-\alpha) &= \int \rho_\tau(y-\hat{\alpha})\,dF(y) \\ &= \tau\mu(F) - \int_{-\infty}^{\hat{\alpha}} (y-\hat{\alpha})\,dF(x) \\ &= \tau\mu(F) - \int_0^{\tau} F_Y^{-1}(t)\,dt.\end{aligned}$$

Portfolio Optimization II

Now if we set $Y = X\pi$ where X denotes a vector of asset returns and π a vector of portfolio weights, solving,

$$\min_{\pi,\alpha}\{\sum_{t=1}^{T} \rho_\tau(x_t^\top \pi - \alpha) \mid \bar{x}^\top \pi = \mu_0\},$$

minimizes lower tail risk subject to a mean return constraint. More generally, we can use a weighted formulation like this,

$$\min_{\pi,\alpha}\{\sum_{j=1}^{J}\sum_{t=1}^{T} w_j \rho_{\tau_j}(x_t^\top \pi - \alpha_j) \mid \bar{x}^\top \pi = \mu_0\},$$

This is a form of composite quantile regression as in Zou and Yuan (2008), and produces a process π_T indexed by the class of concave functions $\mathcal{C} : [0,1] \mapsto [0,1]$ that may prove to be interesting from a multivariate analysis viewpoint.

Borrowing Strength and the QRious Likelihood

The local nature of quantile regression fitting is usually viewed as a feature, but it can be a bug when the data is sparse; then some form of composite QR method can improve efficiency by borrowing strength across adjacent quantiles. In the most extreme form we can view the model,

$$Q_{Y_i|x_i}(\tau|x_i) - x_i^\top \beta(\tau),$$

as a global model that delivers a global likelihood with associated opportunities to impose additional prior information across quantiles. This approach is is well illustrated in recent work by Wang, Li and He (2012), Carroll and Wei (2009) and Arellano and Bonhomme (2016).

Software

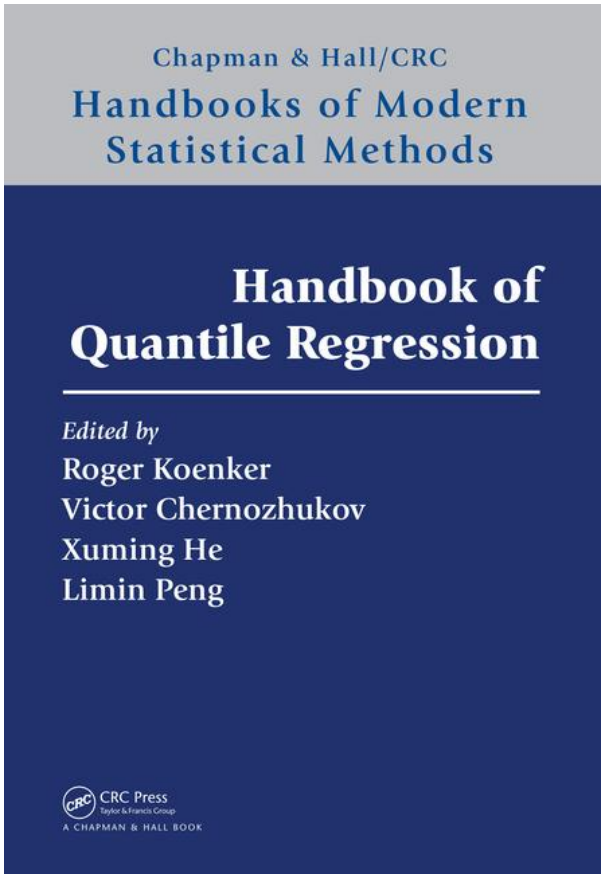
Early development of computational methods for quantile regression was carried out at Bell Labs in the S language, and has continued to be developed in R in my package `quantreg` available from CRAN. Included are functions:

- rq Basic linear model fitting and inference,
- nlrq Nonlinear model fitting and inference,
- crq Censored linear model fitting and inference
- rqss Nonparametric model fitting via total variation penalties
- dynrq Time-series model fitting
- qrisk Portfolio optimization via QR methods

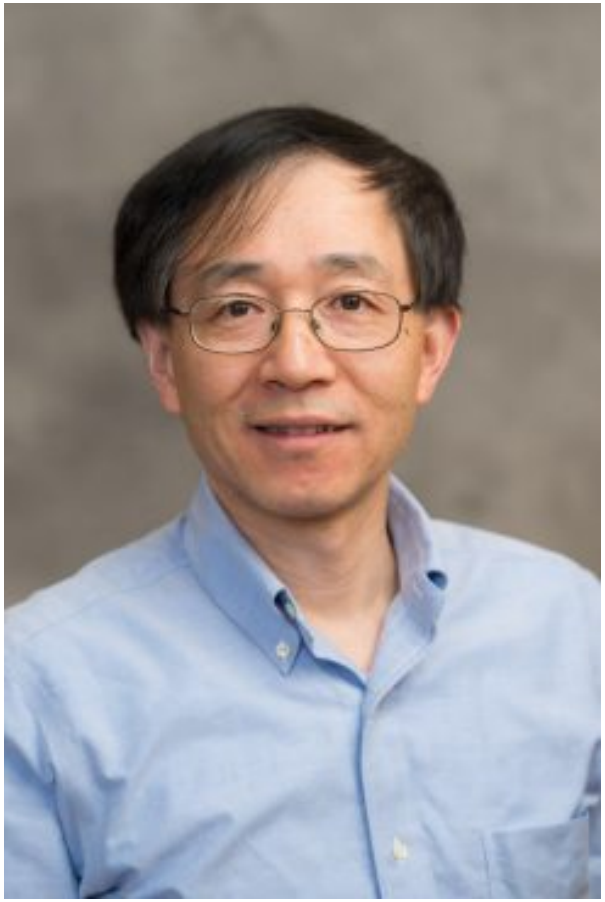
Some of this functionality is also available in SAS Proc Quantreg, and to a lesser extent in Stata.

Further Developments

More about recent developments will be provided by my colleagues Xuming He and Lan Wang, but I can't resist a brief advertisement:



Resampling Methods for Quantile Regression Inference



Why resampling methods?

The asymptoptic variance-covariance of quantile regression estimator involves

$$H_n = n^{-1} \sum f_i(x_i^\top \beta(\tau)) x_i x_i^\top$$

and a direct estimation is challenging.

Resamping methods have advantages:

- take advantage of computer power to replace analytic derivations;
- tend to have good finite-sample approximation accuracy;
- have the flexibility to work under less stringent model assumptions.

Two settings

- Correlation model: (x_i, y_i) treated as a random sample.
- Regression model: x_i treated as fixed.

Correlation Model

Paired bootstrap: (x_i^*, y_i^*) sampled with replacement from the original sample

- For each bootstrap sample $\{(x_i, y_i), i = 1, \cdots, n\}$, compute the quantile estimate $\beta^*(\tau)$,

$$\beta^*(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i^* - x_i^{*\top}\beta).$$

- Repeat this B times to get $\beta_1^*(\tau), \cdots, \beta_B^*(\tau)$

The distribution of $\beta^*(\tau) - \hat{\beta}(\tau)$ is approximately the same as the sampling distribution of $\hat{\beta}(\tau) - \beta(\tau)$.

Correlation Model

Generalized bootstrap: each re-sample is generated through random weights $(w_1, \cdots w_n)$ with mean 1;

$$\beta^*(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w_i \rho_\tau(y_i - x_i^\top \beta).$$

Reference: Chatterjee and Bose (2005)

Estimating equation bootstrap

Let

$$S_n(\beta) = n^{-1/2} \sum_{i=1}^{n} x_i \big(I(y_i - x_i^\top \beta < 0) - \tau \big).$$

Given x_i, the distribution of $S_n(\beta)$ is pivotal when β takes the true quantile coefficient β_τ. Suppose U has the same distribution, then

$$S_n(\beta_U) = U$$

provides a resampling distribution for the quantile estimate.

Reference: Parzen, Wei and Ying (1994)

Markov chain marginal bootstrap

MCMB (He and Hu, 2002): iteratively solve the marginal equations of $S_n(\beta) = U$.

For the iteration from $\beta^{(k)}$ to $\beta^{(k+1)}$, with $p = 2$ for illustration,

- Solve $\beta_1^{(k+1)}$ from

$$\sum_{i=1}^{n} x_{i,1}\left(I(y_i - x_{i,1}\beta_1 - x_{i,2}\beta_2^{(k)}) - \tau\right) = U_1.$$

Markov chain marginal bootstrap

MCMB (He and Hu, 2002): iteratively solve the marginal equations of $S_n(\beta) = U$.

For the iteration from $\beta^{(k)}$ to $\beta^{(k+1)}$, with $p = 2$ for illustration,

- Solve $\beta_1^{(k+1)}$ from

$$\sum_{i=1}^{n} x_{i,1}\big(I(y_i - x_{i,1}\beta_1 - x_{i,2}\beta_2^{(k)}) - \tau\big) = U_1.$$

- Solve $\beta_2^{(k+1)}$ from

$$\sum_{i=1}^{n} x_{i,2}\big(I(y_i - x_{i,1}\beta_1^{(k+1)} - x_{i,2}\beta_2) - \tau\big) = U_2.$$

Markov chain marginal bootstrap

MCMB (He and Hu, 2002): iteratively solve the marginal equations of $S_n(\beta) = U$.

For the iteration from $\beta^{(k)}$ to $\beta^{(k+1)}$, with $p = 2$ for illustration,

- Solve $\beta_1^{(k+1)}$ from

$$\sum_{i=1}^{n} x_{i,1}\big(I(y_i - x_{i,1}\beta_1 - x_{i,2}\beta_2^{(k)}) - \tau\big) = U_1.$$

- Solve $\beta_2^{(k+1)}$ from

$$\sum_{i=1}^{n} x_{i,2}\big(I(y_i - x_{i,1}\beta_1^{(k+1)} - x_{i,2}\beta_2) - \tau\big) = U_2.$$

- Keep iterating with independent draws of U_1 and U_2 each time.

MCMB

Properties of MCMB:

- Computational complexity is linear in p.
- At each step, the solution is a weighted univariate quantile calculation.
- The MCMB chain is approximately an AR series whose stationary distribution is the same as the sampling distribution of the quantile regression estimate.

Regression Model

How to keep $x_i^* = x_i$ fixed?

- **Residual bootstrap:** $y_i^* = x_i^\top \hat{\beta}_\tau + e_i^*$, where e_i^* is a bootstrap sample of the residuals $\hat{e}_i$.

- **Wild bootstrap:**

$$y_i^* = x_i^\top \hat{\beta}_\tau + w_i |\hat{e}_i|,$$

where w_i are drawn from an appropriate distribution. One example: two-point mass distribution with probabilities $1-\tau$ and τ at $w = 2(1-\tau)$ and -2τ, respectively.

Reference: Feng, He, and Hu (2011)

Bayesian Inference

Ingredients: working likelihood for the data $\mathcal{D}$ + prior distribution

- **Asymmetric Laplacian likelihood:**

$$L(\beta; \mathcal{D}) = \frac{\tau^n (1-\tau)^n}{\sigma^n} \exp\left\{ - \frac{\sum_{i=1}^n \rho_\tau (y_i - x_i^\top \beta)}{\sigma} \right\}$$

  Reference: Yu and Moyeed (2001)

- **Empirical likelihood:**
  Reference: Lancaster and Sun (2009), Yang and He (2012)

Asymmetric Laplacian likelihood:

Pros:

- The quantile regression estimate $\hat{\beta}(\tau)$ is the maximum likelihood estimate under the asymmetric Lapacian.
- Efficient MCMC algorithms are available (Kozumi and Kobayashi, 2011; Yue and Rue, 2011).
- The posterior is proper even with flat priors on $\beta(\tau)$.

Cons:

- The working likelihoods at two different τ's might not be compatible.
- The posterior variance is not approximating the sampling variance of the quantile regression estimator.

Posterior Inference from Asymmetric Laplacian likelihood:

The posterior variance from the asymmetric Laplacian likelihood can be adjusted to provide an asymptotically valid estimate of the sampling variance.

$$\hat{\Sigma}_{\text{adj}} = \frac{n}{\sigma^2}\hat{\Sigma}_P \hat{J}_n \hat{\Sigma}_P \approx \text{Var}(\hat{\beta}(\tau))$$

where $\hat{\Sigma}_P \approx (\sigma/n)H_n^{-1}$ is the posterior variance, and

$$J_n = \tau(1-\tau)n^{-1}\sum x_i x_i^\top$$

$$H_n = n^{-1}\sum f_i(x_i^\top \beta(\tau))x_i x_i^\top .$$

Posterior Inference (continued):

$$\hat{\Sigma}_{\text{adj}} = \frac{n}{\sigma^2}\hat{\Sigma}_P \hat{J}_n \hat{\Sigma}_P \approx \text{Var}(\hat{\beta}(\tau))$$

- Asymptotically, the adjusted variance is invariant in the choice of σ in the asymmetric likelihood specification.
- The asymmetric Laplacian likelihood generalizes to censored quantile regression.
- R package: *bayesQR*.

Reference: Yang, Wang and He (2016) (with discussions)

Summary of Computer-intensive Methods

With the bootstrap (and its variants) or Bayesian computation, we can carry out approximate inference on quantile regression without direct estimation of the (troublesome?) asymptotic variance.

Computers are our friends for quantile regression, especially with large n and/or large p problems.

Quantile Regression Methods for High Dimensional Data

Motivating example: birth weight data (Votavova et al., 2011) with genetic information

- **Response**: Birth weight of baby (in kilograms).
- **Covariates**: Age of mother, gestational age, parity, measurement of the amount of cotinine, a chemical found in tobacco, in the blood and mother's BMI, genetic data from the peripheral blood sample (24,539 probes).
- Lower quantiles of infant birth weight are of particular interest.

Motivating example (cont'd)

Q-SCAD .1		Q-SCAD .3		Q-SCAD .5	
Covariate	Frequency	Covariate	Frequency	Covariate	Frequency
Gestational Age	82	Gestational Age	86	Gestational Age	69
1687073 (SOGA1)	24	1804451 (LEO1)	33	2334204 (ERCC6L)	57
		1755657 (RASIP1)	27	1732467 (OR2AG1)	52
		1658821 (SAMD1)	23	1656361 (LOC201175)	31
		2059464 (OR5P2)	14	1747184 (PUS7L)	5
		2148497 (C20orf107)	6		
		2280960 (DEPDC7)	3		

Frequency of covariates selected at three quantiles among 100 random partitions

Motivating example (cont'd)

- Gestational age is identified to be important with high frequency at all three quantiles under consideration.
- The gene SOGA1 is a suppressor of glucose, which is interesting because maternal gestational diabetes is known to have a significant effect on birth weight [Gilliam et al. (2003)].
- The genes OR2AG1, OR5P2 and DEPDC7 are all located on chromosome 11, the chromosome with the most selected genes. Chromosome 11 also contains PHLDA2, a gene that has been reported to be highly expressed in mothers that have children with lower birth weight [Ishida et al. (2012)].
- **The genes selected at the three different quantiles are not overlapping.** This is an indication of the heterogeneity in the data. The variation in frequency is likely due to the relatively small sample size.

Advantages of quantile approach in high dimension ($p \gg n$)

- **Quantile-adaptive sparsity**:
  A small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may be different when we consider different conditional quantiles.
- **Weaker conditions on random error distribution**:
  No need to impose restrictive distributional or moment conditions on the random errors and allow their distributions to depend on the covariates.
- **Robustness** with respect to outliers in Y.

Quantile-adaptive nonlinear variable screening

- **Quantile-based approach** (He, Wang and Hong (2013, AOS)):
  - It allows the sets of active variables to vary across quantiles, thus making it more flexible to accommodate heterogeneity.
  - It is model-free and avoids the difficult task of specifying the form of a statistical model in a high dimensional space.
- The R codes can be found at:
  https://www.stt.msu.edu/users/hhong/example1b.txt
- The set of active variables at quantile level α is define as

$$M_\alpha = \{j : Q_\alpha(Y|\mathbf{X}) \text{ functionally depends on } X_j\},$$

  where $Q_\alpha(Y|\mathbf{X})$ is the αth conditional quantile of Y given $\mathbf{X} = (X_1, \ldots, X_p)^\top$.
- **Example**:

$$Y = m(\boldsymbol{\alpha}_1^\top \mathbf{X}_{A_1}) + \sigma(\boldsymbol{\alpha}_2^\top \mathbf{X}_{A_2})\epsilon,$$

  where m and σ are known or unknown functions, $\boldsymbol{\alpha}_i$ $(i = 1, 2)$ are vectors of nonzero coefficients, A_i are subsets of $\{1, 2, \ldots, p\}$, ϵ has a standard normal distribution.

Ranking by marginal quantile utility

- Y and X_j are independent $\iff Q_\alpha(Y|X_j) - Q_\alpha(Y) = 0, \ \forall \ \alpha \in (0,1)$,

  where $Q_\alpha(Y|X_j)$ is the αth conditional quantile of Y given X_j and $Q_\alpha(Y)$ is the αth unconditional quantile of Y.
- Let $\widehat{\beta}_j = \text{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \rho_\alpha(Y_i - \pi(X_{ij})^T \beta)$, and define

$$\widehat{f}_{nj}(t) = \pi(t)^T \widehat{\beta}_j - F^{-1}_{Y,n}(\alpha)$$

  where $F^{-1}_{Y,n}(\alpha)$ is the α-th sample quantile function.
- We will select the subset of variables

$$\widehat{M}_\alpha = \{1 \leqslant j \leqslant p : \|\widehat{f}_{nj}\|_n^2 \geqslant \nu_n\}$$

  where $\|\widehat{f}_{nj}\|_n^2 = n^{-1} \sum_{i=1}^n \widehat{f}^2_{nj}(X_{ij})$.

Statistical properties

- **Sure screening property**:

$$P\left(M_\alpha \subset \widehat{M}_\alpha\right) \to 1, \text{as } n \to \infty.$$

- **Controlling false discovery**: Under certain regularity conditions, there exist some positive constants δ_1 and δ_2 such that for all n sufficiently large,

$$\begin{aligned} & P\left(|\widehat{M}_\alpha| \leqslant 2d^2 n^\tau \lambda_{max}(\boldsymbol{\Sigma})/\delta^*\right) \\ \geqslant \; & 1 - p\left\{11\exp\left(-\delta_1 n^{1-4\tau}\right) + 12d^2\exp\left(-\delta_2 d^{-3} n^{1-2\tau}\right)\right\}. \end{aligned}$$

Especially, $P\left(|\widehat{M}_\alpha| \leqslant 2d^2 n^\tau \lambda_{max}(\boldsymbol{\Sigma})/\delta^*\right) \to 1$ as $n \to \infty$.

Penalized/regularized quantile regression in high dimension

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}_\tau + \epsilon_i, \ i = 1, \ldots, n,$$

where ϵ_i are independent random errors such that $P(\epsilon_i \leqslant 0 | \mathbf{x}_i) = \tau$, $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ with $x_{i0} = 1$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of unknown parameters.

- **Sparsity**: Let $A_0 = \{j : \beta_j^* \neq 0\}$ and $|A_0| = q$. Assume that $q \ll n$.
- **Penalized linear quantile regression**:

$$Q(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ.

Difference choices of penalty function

High-dimensional quantile regression $p \gg n$ with

- L_1 **(or Lasso) penalty** (Tibshirani, 1996) was studied by Belloni and Chernozhukov (2011), Bradic, Fan and Wang (2011), Kato (2011), Wang (2013), among others
- **Nonconvex penalty** function (e.g., SCAD (Fan and Li, 2001) and MCP (Zhang, 2010)) was studied by Wang, Wu and Li (2012), Sherwood and Wang (2016), among others.
- Two-stage **adaptive penalty**: van de Geer (2003), Fan, Fan and Barut (2014), Fan Xue and Zou (2014) (adaptive L_1 penalty); Zheng, Peng and He (2015, 2017, adaptive regional penalty), among others.

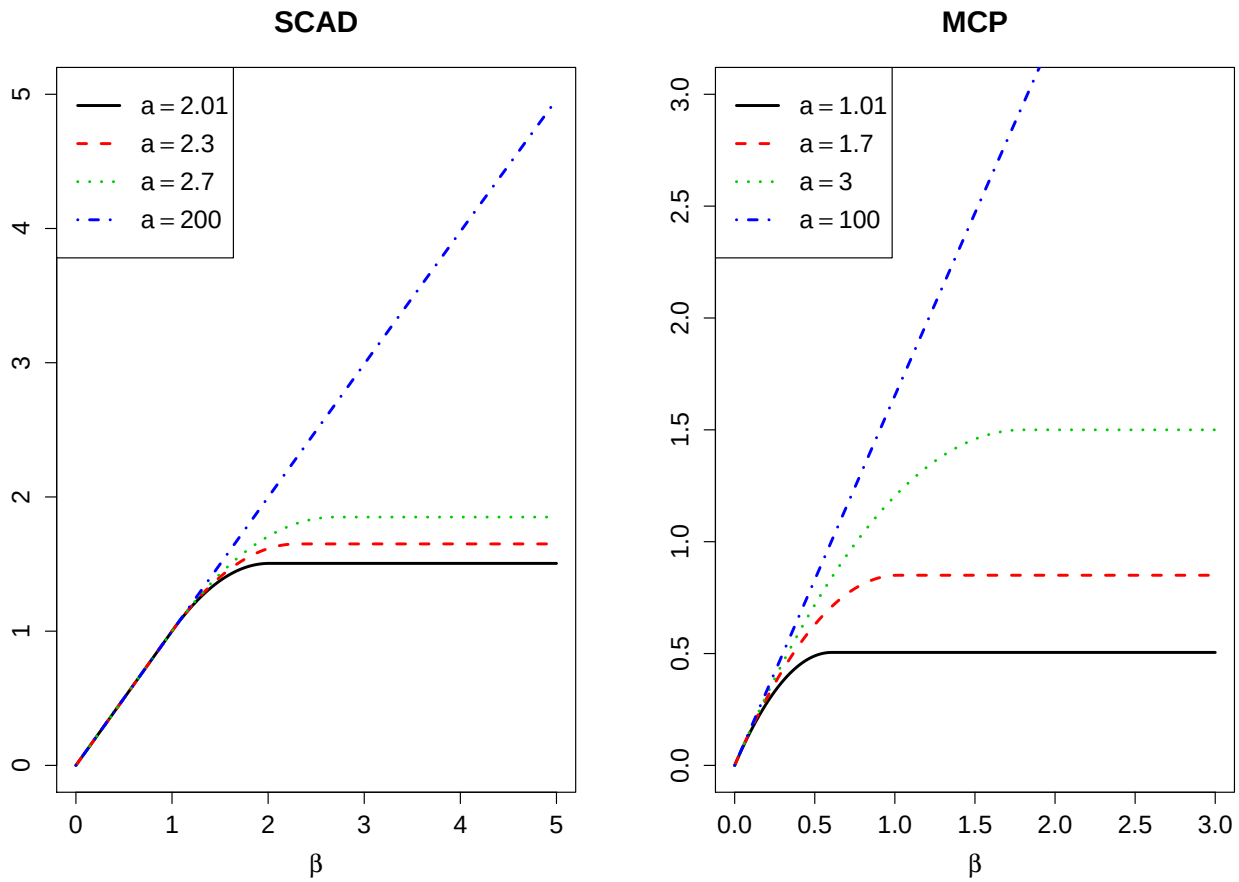
L_1 penalized quantile regression

- L_1 penalty: $p_\lambda(|\beta_j|) = |\beta_j|$, $j = 1, \ldots, p$.
- Computationally convenient due to the convex structure.
- The use of L_1 penalty achieves accurate prediction under relaxed conditions.
- **Near-oracle rate of estimation** (Belloni and Chernozhukov, 2011): under regularity conditions:

$$\|\widehat{\beta}^{L1} - \beta_0\|_2 = O_p\left(\sqrt{\frac{q \log p}{n}}\right),$$

where q is the unknown sparsity level, p is the number of candidate covariates; and n is the sample size.

Non-convex penalized linear quantile regression ($p \gg n$)



SCAD and MCP penalty functions ($\lambda = 1$)

A numerical example

Simulation results ($n = 300$, $p = 600$, $\tau = 0.7$)

Method	Size	P1	P2	AE
LS-Lasso	24.30 (0.61)	100%	7%	1.40 (0.03)
Q-Lasso ($\tau = 0.5$)	25.76 (0.94)	100%	10%	1.05 (0.03)
Q-Lasso ($\tau = 0.7$)	32.74 (1.22)	90%	90%	1.78 (0.05)
LS-SCAD	6.04 (0.25)	100%	0%	0.38 (0.02)
Q-SCAD ($\tau = 0.5$)	6.14 (0.36)	100%	7%	0.19 (0.01)
Q-SCAD ($\tau = 0.7$)	9.97 (0.54)	100%	100%	0.38 (0.03)
LS-MCP	5.56 (0.19)	100%	0%	0.38 (0.02)
Q-MCP ($\tau = 0.5$)	5.33 (0.23)	100%	3%	0.18 (0.01)
Q-MCP ($\tau = 0.7$)	7.56 (0.32)	98%	98%	0.37 (0.03)

Size denotes the average number of non-zero regression coefficients; P1 denotes the proportion of simulation runs including X_6, X_{12}, X_{15} and X_{20}; P2 denotes the proportion of simulation runs X_1 is selected; and AE denotes the absolute estimation error defined by $\sum_{j=0}^{p} |\widehat{\beta}_j - \beta_j|$.

Non-convex penalized linear quantile regression (cont'd)

- **SCAD penalty**:

$$p_\lambda(|\beta|) = \lambda|\beta| I(0 \leqslant |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \leqslant |\beta| \leqslant a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } a > 2.$$

- **MCP penalty**:

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \leqslant |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \geqslant a\lambda), \quad a > 1.$$

- **Oracle property** (Wang, Wu and Li, 2012): Assume some regularity conditions. The oracle estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^\top, \mathbf{0}^\top)^\top$ satisfies that $P(\widehat{\boldsymbol{\beta}} \in \mathcal{B}_n(\lambda)) \to 1$ as $n \to \infty$.

High-dimensional semiparametric quantile regression

- **Partially linear additive quantile regression**:

$$Q_{Y_i|x_i,z_i}(\tau) = x_i'\beta_0 + g(z_i), \quad i = 1, ..., n,$$

where x_i is a p_n-dimensional vector of covariates, z_i is a d-dimensional vector of covariates, $g(z_i) = g_0 + \sum_{j=1}^{d} g_j(z_{ij})$, $g_0 \in \mathcal{R}$ and g_j satisfies $E(g_j(z_{ij})) = 0$. Assume $\beta_0 = (\beta_{01}', \mathbf{0}')'$ and β_{01} is q_n-dimensional.

- **Penalized semiparametric quantile regression estimator**:

$$(\hat{\beta}, \hat{\xi}) = \underset{(\beta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta - \Pi(z_i)'\xi) + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

- Choose λ that minimizes the following high-dimensional BIC criterion (Lee, Noh and Park, 2013):

$$\mathrm{QBIC}(\lambda) = \log\left(\sum_{i=1}^{n} \rho_\tau\left(Y_i - X_i'\hat{\beta}_\lambda - \Pi(z_i)'\hat{\xi}_\lambda\right)\right) + \nu_\lambda \frac{\log(p)\log(\log(n}{2n}$$

- Oracle property: Sherwood and Wang (2016).

Estimation and variable selection for multiple quantiles

- Let $\tau_1 < \tau_2 < ... < \tau_M$ be the set of quantiles of interest, where $M > 0$ is a positive integer. We assume that

$$Q_{Y_i|\mathbf{x}_i,\mathbf{z}_i}(\tau_m) = \mathbf{x}_i'\boldsymbol{\beta}_0^{(m)} + g_0^{(m)}(\mathbf{z}_i), \quad m = 1, ..., M,$$

  where $g_0^{(m)}(\mathbf{z}_i) = g_{00}^{(m)} + \sum_{j=1}^{d} g_{0j}^{(m)}(z_{ij})$, with $g_{00}^{(m)} \in \mathcal{R}$, $g_{0j}^{(m)}$ satisfies $E\left[g_{0j}^{(m)}(z_{ij})\right] = 0$.
- We are interested in the high-dimensional case where **most of the linear covariates have zero coefficients across all M quantiles**, for which group selection will help us combine information across quantiles.

Estimation and variable selection for multiple quantiles (cont'd)

- Write $\boldsymbol{\beta}_0^{(m)} = \left(\beta_{01}^{(m)}, \beta_{02}^{(m)}, ..., \beta_{0p_n}^{(m)}\right)'$, $m = 1, ..., M$. Let $\bar{\boldsymbol{\beta}}^{0j}$ be the M-vector $\left(\beta_{0j}^{(1)}, ..., \beta_{0j}^{(M)}\right)'$, $1 \leqslant j \leqslant p_n$.
- We estimate $\left(\boldsymbol{\beta}_0^{(m)}, \xi_0^{(m)}\right)$, $m = 1, ..., M$, by minimizing

$$n^{-1} \sum_{i=1}^{n} \sum_{m=1}^{M} \rho_{\tau_m}(Y_i - \mathbf{x}_i' \boldsymbol{\beta}^{(m)} - W(z_i)' \boldsymbol{\xi}^{(m)}) + \sum_{j=1}^{p} p_\lambda(\|\bar{\boldsymbol{\beta}}^j\|_1),$$

where $\bar{\boldsymbol{\beta}}^j$ is the M-vector $\left(\beta_j^{(1)}, ..., \beta_j^{(M)}\right)'$, $1 \leqslant j \leqslant p$. The penalty function encourages group-wise sparsity and forces the covariates that have no effect on any of the M quantiles to be excluded together.

New algorithms for large-scale data

- **QICD algorithm**: iterative coordinate-descent algorithm for high-dimensional nonconvex penalized quantile regression (Peng and Wang, 2015, *JCGS*. R package: QICD
- **Parallel-computing based algorithms**:
  Yu, Lin and Wang (2017+) A parallel algorithm for large-scale nonconvex penalized quantile regression. To appear in *JCGS*.
  Gu et al. (2017) ADMM for high-dimensional sparse penalized quantile regression. *Technometrics*.

High-dimensional quantile regression with focus on an interval of quantile levels (Zheng, Peng and He, 2015, AOS)

- **Motivation**: simultaneously estimating and selecting variables over a set of quantile levels $\Delta \in (0, 1)$.
- The active/relevant set of covariates is defined as

$$S_\Delta = \text{support}(\beta_0(\tau), \tau \in \Delta) = \left\{ j \in 2, \ldots, p : \exists\, \tau \in \Delta, |\beta_0^j(\tau)| > 0 | \right\}.$$

- **Adaptively weighted L_1 penalized quantile regression**:

$$Q(\beta, \tau) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i^T \beta) + \lambda_n \sum_{j=2}^{p} w_j(\tau) |\beta^{(j)}|,$$

- $\widehat{S}_\Delta = \left\{ j \in 2, \ldots, p : \exists\, \tau \in \Delta, |\hat{\beta}^{(j)}(\tau)| > 0 | \right\}.$

High-dimensional quantile regression with focus on an interval of quantile levels (cont'd)

- Examples of weights that reflect regional focus:

$$w_j(\tau) = \frac{1}{\sup_{\tau\in\Delta} |\tilde{\beta}^{(j)}(\tau)|}, \text{ and } \quad w_j(\tau) = \frac{1}{\int_\Delta |\tilde{\beta}^{(j)}(\tau)| d\tau},$$

  where $\tilde{\beta}^{(j)}(\tau)$ is an initial estimator of $\beta_0^{(j)}(\tau)$.
- It's recommended to use L_1 penalized quantile regression for the initial estimator.

High-dimensional quantile regression with focus on an interval of quantile levels (cont'd)

- **Oracle property**: with probability approaching one, the proposed estimator can successfully identify the set of relevant covariates, including those having effects on some or all quantile levels in Δ.
- **A GIC criterion** for tuning parameter selection:

$$\mathrm{GIC}(\lambda) = \int_{\Delta} \log \hat{\sigma}_{\lambda}(\tau) d\tau + |\hat{S}_{\lambda}| \phi(n),$$

where $\hat{\sigma}_{\lambda}(\tau) = n^{-1} \sum_{i=1}^{n} \rho_{\tau}(Y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}}_{\lambda}(\tau))$, $\phi(n)$ is a sequence converging to zero with n.

- Extended to censored data in Zheng, Peng and He (2017+, AOS).

Discussions and conclusions

Discussion 1: Could we further refine the theory for high-dimensional quantile regression?

- The classical theory for quantile regression requires regularity conditions comparable with those for LS (e.g., Chapter 4, Koenker, 2005).
- More complex conditions on the design matrix (e.g., the restricted nonlinear impact condition) are required for high-dimensional quantile regression comparing to penalized least squares regression.

Discussions and conclusions (cont'd)

Discussion 2: Could we further bridge the gap between convex and nonconvex penalty approaches?

- Adaptive weighting is a nice idea to use L_1 penalized quantile regression as an initial estimator to further reduce the bias. Minimal signal strength assumption is needed to ensure model selection consistency.
- Loh and Wainwright (2015 JMLR; 2017+, AOS) proved that any local solution of non-convex penalized M-estimator will lie within statistical precision of the underlying parameter vector but requires smooth loss function.

Discussions and conclusions (cont'd)

- **Quantile regression is useful for modeling high-dimensional heterogeneous data**: quantile-specific sparsity, weaker error distribution assumptions.
- **Still an active area for research**:
  - Statistical inference for high-dimensional quantile regression: Belloni, Chernozhukov, Kato (2015, Biometrika), Zhao, Kolar and Liu (2014, on arXiv), Bradic and Kolar (2017+, on arXiv)
  - Lee, Liao, Seo and Shin (2017+, JASA): high-dimensional change-point quantile regression
  - Lv, Lin, Lian and Huang (2017+, AOS): oracle inequality for sparse additive quantile regression in reproducing kernel hilbert space.
  - Many others...

A comprehensive introduction to quantile regression