conquer: Convolution Smoothed Quantile Regression

Wenxin Zhou

UC San Diego

Joint work with Xuming He, Lan Wang,
Kean Ming Tan & Xiaoou Pan

2021 WNAR Conference

- The idea of median regression predates the least squares method by about 50 years.

- Roger Joseph Boscovich (1760), Pierre-Simon Laplace (1789), F. Y. Edgeworth (1887).

- The method of least squares (ℓ_2-method) has significant computational advantages over ℓ_1-method—minimization of absolute errors advocated by Boscovich, Laplace and others, even though the latter is more robust.

- Regression quantiles: Koenker and Bassett, Jr. (1978).

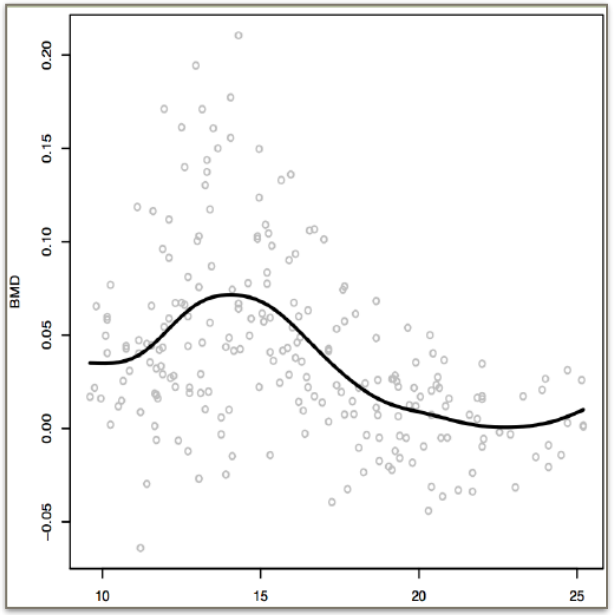
Computational Development

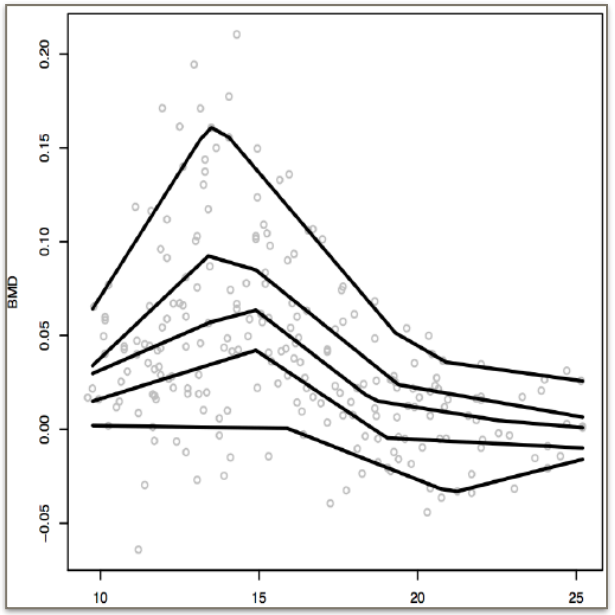
- Simplex-based algorithms: Barrodale & Roberts (1974), Koenker & d'Orey (1987).
  - Slow in larger samples, worst-case analysis suggests the number of iterations may increase exponentially with the sample size.
- Interior point method: Newton-Frisch algorithm (Portnoy & Koenker, 1997) has computational complexity $\mathcal{O}(n^{1+a}p^3 \log n)(0 < a < 1/2)$, conjectured to be improvable to $\mathcal{O}(np^3 \log^2(n))$ (Mizuno-Todd-Ye conjecture).
- Preprocessing (Portnoy & Koenker, 1997): improved complexity $\mathcal{O}((np)^{2(1+a)/3}p^3 \log n + np)$, improvable to $\mathcal{O}(n^{2/3}p^3 \log^2(n) + np)$.
- R "quantreg", MATLAB, SAS, etc.

Quantile regression versus mean regression

- Mean regression: $y_i = m(\mathbf{x}_i) + \varepsilon_i$, $\mathbb{E}(\varepsilon_i \,|\, \mathbf{x}_i) = 0$.

- Quantile regression: $\mathbb{P}\{y \leq Q(\tau, \mathbf{x}) \,|\, \mathbf{x}\} = \tau$
  - No strict distinction between 'signal' and 'noise'.
  - Object of interest: the conditional distribution of $y \,|\, \mathbf{x}$.
  - Contains richer information than conditional mean.





Model and Assumptions

- Goal: learn the effect of a $p \times 1$ vector of covariates $\mathbf{x} = (x_1, \ldots, x_p)^\intercal$ $(x_1 \equiv 1)$ on the entire distribution of y.
- Conditional quantile model:

$$Q(\tau, \mathbf{x}) = F^{-1}_{y|\mathbf{x}}(\tau) \approx \mathbf{x}^\intercal \beta^*(\tau), \quad \tau \in [\tau_L, \tau_U] \subseteq (0,1).$$

- $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \overset{\text{iid}}{\sim} (y, \mathbf{x})$.

Given $\tau \in [\tau_L, \tau_U]$, $(y, \mathbf{x})$ admits the characterization

$$y \approx \mathbf{x}^\intercal \beta^*(\tau) + \varepsilon(\tau),$$

where $\mathbb{P}\{\varepsilon(\tau) \leq 0 \,|\, \mathbf{x}\} = \tau$. Let $f_{\varepsilon(\tau)}(\,\cdot\,|\,\mathbf{x})$ be the conditional density function of $\varepsilon(\tau)$ given $\mathbf{x}$.

QR-series Approximation for NP Model

- QR-series approximation to $\mathbf{x} \mapsto Q(\tau, \mathbf{x})$: fix τ, let $\mathbf{x} \mapsto Z(\mathbf{x}) = (Z_1(\mathbf{x}), \ldots, Z_m(\mathbf{x}))^\intercal$ be a vector of series approximating functions of dimension $m = m_n$.
  - B-splines (or regression splines)
  - Polynomials
  - Fourier series
  - Compactly supported wavelets

- QR-series approximation error $R(\tau, \mathbf{x}) = Q(\tau, \mathbf{x}) - Z(\mathbf{x})^\intercal \beta^*(\tau)$ vanishes asymptotically under appropriate conditions when $m = m_n \to \infty$ as $n \to \infty$ (Belloni, *et al.*, 2019).

Quantile Regression

Given $\tau \in (0,1)$, the standard QR estimator is

$$\hat{\beta} = \hat{\beta}(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\intercal \beta),$$

where $\rho_\tau(u) = \{\tau - 1(u < 0)\}u$ is the check function.

- robustness against outliers in the response, especially in the case of median regression
- ability to capture heterogeneity in the set of important predictors at different quantile levels of the response distribution caused by, e.g., heteroscedastic variance

Theoretical Development

- Consistency: Bassett & Koenker (1986), Zhao, Rao & Chen (1993), El Bantli & Hallin (1999), etc.
- Rate of convergence: Ruppert & Carroll (1980), Pollard (1991), Hjort & Pollard (1993), Knight (1998), etc.
- Bahadur representation & Normal approximation: Jureckova & Sen (1984), Portnoy (1984), Koenker & Portnoy (1987), Portnoy & Koenker (1989), Gutenbrunner & Jureckova (1992), Hendricks & Koenker (1991), He & Shao (1996, 2000), Arcones (1996), Koenker & Machado (1999), Koenker & Xiao (2002).

Classical Asymptotics: $n \to \infty$ and p is fixed.

Challenges in QR

- Lack of strong convexity: quantile loss is piecewise linear and its "curvature energy" is concentrated in a single point. This is substantially different from other popular loss functions, e.g. ℓ_2, logistic and Huber, or even Tukey and Hampel, which are at least locally strongly convex.
- Lack of smoothness: quantile loss is not everywhere differentiable. Theoretically, it leads to an error term of order $\mathcal{O}_{\mathbb{P}}(n^{-1/4})$ or $\mathcal{O}_{\mathbb{P}}\{(p^3/n)^{1/4}\}$ in the Bahadur representation.

  - Welsh (1989) shows that $p^3 \log^2(n) = o(n)$ suffices for normal approximation (fixed design).
  - Huber regression requires $p^2 = o(n)$.

Smoothed Estimating Equation (SEE) Approach

- Population loss $Q(\beta) = \mathbb{E}_{(y,\mathbf{x})\sim P}\, \rho_\tau(y - \mathbf{x}^\intercal\beta)$, and

$$\beta^* = \beta^*(\tau) = \arg\min_{\beta\in\mathbb{R}^p} Q(\beta).$$

- If $F_{\varepsilon(\tau)|\mathbf{x}}$ is continuously differentiable, Q is twice differentiable and strongly convex at least in a neighborhood of β^*. Moreover, β^* satisfies the first-order condition:

$$\nabla Q(\beta) = \mathbb{E}\{1(y < \mathbf{x}^\intercal\beta) - \tau\}\mathbf{x}\,|_{\beta=\beta^*} = 0.$$

- Sample analog:

$$\frac{1}{n}\sum_{i=1}^{n} \{1(y_i - \mathbf{x}_i^\intercal\beta < 0) - \tau\}\mathbf{x}_i = 0.$$

The QR estimator $\hat\beta$ solves this equation approximately.

- SEE approach (Whang 2006; Kaplan & Sun, 2017):

$$\frac{1}{n}\sum_{i=1}^{n}\{G(-r_i(\beta)/h)-\tau\}\mathbf{x}_i=0,$$

where $r_i(\beta)=y_i-\mathbf{x}_i^{\intercal}\beta$, G is a smooth function and $h>0$ is the bandwidth.

- Horowitz's method (Horowitz, 1998): smooth the criterion function by replacing the indicator in the check function by a kernel counterpart:

$$\ell_h^{\mathrm{H}}(u)=u\{\tau-G(-u/h)\}.$$

- Horowitz's smoothed check function is not convex!

M-estimation Viewpoint

- Smoothed loss function

$$\hat{Q}_h(\beta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\rho_\tau * K_h}_{=:\ell_h}(y_i - \mathbf{x}_i^\intercal \beta),$$

where K is a kernel function, $K_h(u) = (1/h)K(u/h)$ and

$$\ell_h(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty} \rho_\tau(u) K_h(v-u)\,\mathrm{d}v.$$

- Convolution smoothed QR (*conquer*):

$$\hat{\beta}_h = \hat{\beta}_h(\tau) \in \arg\min_{\beta\in\mathbb{R}^p} \hat{Q}_h(\beta).$$

- Any optimum $\hat{\beta}_h$ satisfies the FOC

$$\nabla \hat{Q}_h(\beta) = \frac{1}{n} \sum_{i=1}^{n} \{\bar{K}(-r_i(\beta)/h) - \tau\}\mathbf{x}_i,$$

where $\bar{K}(u) = \int_{-\infty}^{u} K(t)\,\mathrm{d}t$.

- Convexity:

$$\nabla^2 \hat{Q}_h(\beta) = \frac{1}{n} \sum_{i=1}^{n} K_h(y_i - \mathbf{x}_i^\intercal \beta) \cdot \mathbf{x}_i \mathbf{x}_i^\intercal.$$

Provided that K is non-negative, $\hat{Q}_h$ is convex and hence any minimizer satisfies the first-order moment condition.

- Fixed-p asymptotics: Fernandes, Guerre & Horta (2021).
- Growing-p (non)asymptotics: He, *et al.* (2020).

Convolution versus Deconvolution

Adding noise $\{u_i\}$ to the response leads to noisy QR

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i + u_i - \mathbf{x}_i^\intercal \beta).$$

Since the noise distribution can be specified, consider

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_u\{\rho_\tau(y_i + u_i - \mathbf{x}_i^\intercal \beta)\}.$$

- Gaussian kernel/noise: $u_i \sim N(0, h^2)$.
- Uniform kernel/noise: $u_i \sim \text{Unif}(-h, h)$.
- Laplacian kernel/noise: $u_i \sim \text{Laplace}(0, h)$.

Computational Methods

- Gradient descent (GD): starting at iteration 0 with an initial estimate $\hat{\beta}^0$, at iteration $t = 0,1,2,\ldots,$, computes

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \eta_t \cdot \nabla \hat{Q}_h(\hat{\beta}^t),$$

where $\nabla \hat{Q}_h(\beta) = (1/n) \sum_i \left\{ \bar{K}\left(\frac{x_i^{\mathrm{T}} \beta - y_i}{h} \right) - \tau \right\} \mathbf{x}_i.$

- Barzilai-Borwein step size (Barzilai & Borwein, 1988): for $t = 1,2,\ldots,$ define

$$\delta^t = \hat{\beta}^t - \hat{\beta}^{t-1}, \quad g^t = \nabla \hat{Q}_h(\hat{\beta}^t) - \nabla \hat{Q}_h(\hat{\beta}^{t-1}).$$

The BB step sizes are

$$\eta_{1,t} = \langle \delta^t, \delta^t \rangle / \langle \delta^t, g^t \rangle \quad \text{and} \quad \eta_{2,t} = \langle \delta^t, g^t \rangle / \langle g^t, g^t \rangle.$$

- As $\tau \approx 0$ or 1, the Hessian becomes more ill-conditioned:

$$\nabla^2 \hat{Q}_h(\beta) = \frac{1}{n} \sum_{i=1}^{n} K_h(y_i - \mathbf{x}_i^\intercal \beta) \cdot \mathbf{x}_i \mathbf{x}_i^\intercal.$$

The step sizes computed in GD-BB may sometimes vibrate drastically, causing instability of the algorithm. To stabilize the algorithm, we take

$$\eta_t = \min\{\eta_{1,t}, \eta_{2,t}, C\}, \, t = 1,2,\ldots$$

For example, $C = 10$.

- Scale the covariate inputs to have zero mean and/or unit variance before applying the GD-BB method.

Initialization via Expectile Regression

- Asymmetric quadratic loss (Newey & Powell, 1987)

$$e_\tau(u) = |\tau - 1(u < 0)| \cdot u^2/2.$$

- Given a univariate random variable Z with $\mathbb{E}|Z| < \infty$, the scale parameter

$$e_\tau = \arg\min_{u \in \mathbb{R}} \mathbb{E}\{e_\tau(Z - u) - e_\tau(Z)\}$$

is called the τ-expectile (Newey & Powell, 1987) or Efron's ω-mean with $\omega = \tau/(1-\tau)$ (Efron, 1991).

Robustified Expectile Regression (*retire*)

- Robustified asymmetric quadratic loss:

$$r_c(u) = |\tau - 1(u < 0)| \cdot H_c(u),$$

where H_c $(c > 0)$ is the Huber loss

$$H_c(u) = 0.5u^2 1(|u| \le c) + (c|u| - c^2/2)1(|u| > c).$$

- We use *retire* estimator as an initial estimate:

$$\hat{\beta}_c^0 \in \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} r_c(y_i - \mathbf{x}_i^\intercal \beta).$$

- When $\tau = 1/2$, this becomes Huber's M-estimator.

Bootstrap Inference with Conquer

Let $\{w_i\}_{i=1}^n$ be iid with $\mathbb{E}(w_i) = 1$, $\mathrm{Var}(w_i) = 1$.

The bootstrapped conquer is defined as

$$\hat{\beta}^\flat \in \arg\min_\beta \hat{Q}_h^\flat(\beta), \quad \hat{Q}_h^\flat = \frac{1}{n}\sum_{i=1}^n w_i \ell_h(y_i - \mathbf{x}_i^\intercal \beta).$$

(i) $w_i \sim \mathcal{N}(1,1)$ (negative weights breaks the convexity)

(ii) $w_i \sim \mathrm{Exp}(1)$

(iii) $w_i \sim 2\mathrm{Ber}(0.5)$ (recommended for large samples)

Normal Distribution Calibrated Inference

- As shown in He, *et al* (2020), for each $j = 1, \ldots, p$,

$$n^{1/2}\sigma_j^{-1}(\hat{\beta}_h - \beta^*)_j \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

where σ_j^2 is the j-th diagonal entry of

$$\mathbf{H}^{-1}\mathbf{E}[\{\bar{K}(-\varepsilon_i/h) - \tau\}^2\mathbf{x}\mathbf{x}^\intercal]\mathbf{H}^{-1}.$$

and $\mathbf{H} = \mathbb{E}\{f_\varepsilon(0 \mid \mathbf{x})\mathbf{x}\mathbf{x}^\intercal\}$.

- Kernel-type matrix estimators:

$$\hat{\mathbf{H}} = \frac{1}{nh}\sum_{i=1}^n \phi(\hat{\varepsilon}_i/h)\mathbf{x}_i\mathbf{x}_i^\intercal, \quad \hat{\mathbf{\Sigma}}(\tau) = \frac{1}{n}\sum_{i=1}^n \{\bar{K}(-\hat{\varepsilon}_i/h) - \tau\}^2\mathbf{x}_i\mathbf{x}_i^\intercal.$$

Here we use the same bandwidth as for the estimator.

Penalized Conquer in High Dimensions

$\beta^* \in \mathbb{R}^p$ is sparse: $\|\beta^*\|_0 \leq s \ll n \ll p$.

- ℓ_1-penalized QR (Belloni & Chernozhukov, 11):

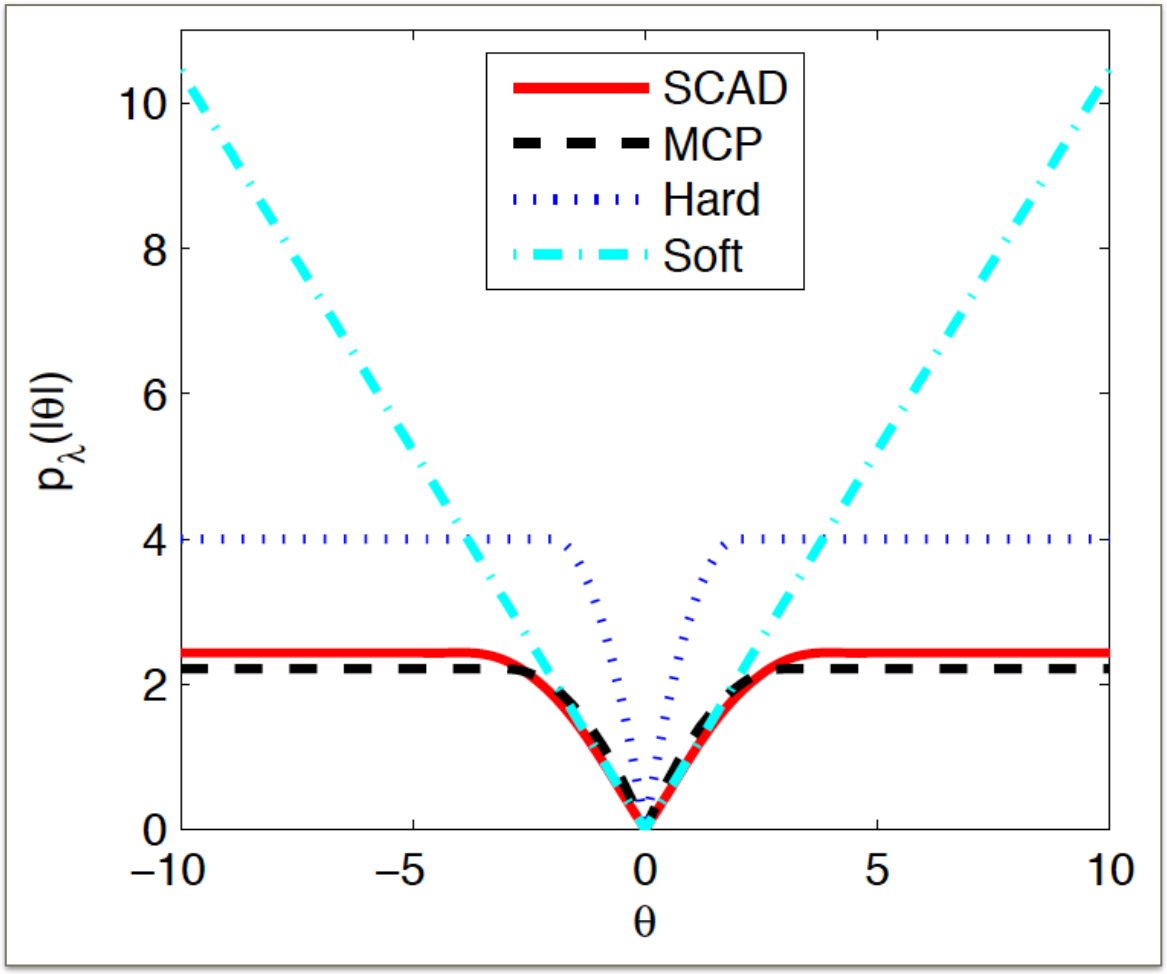
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\intercal \beta) + \lambda \|\beta\|_1,$$

where $\lambda \asymp \sqrt{\tau(1-\tau) \cdot \log(p)/n}$.

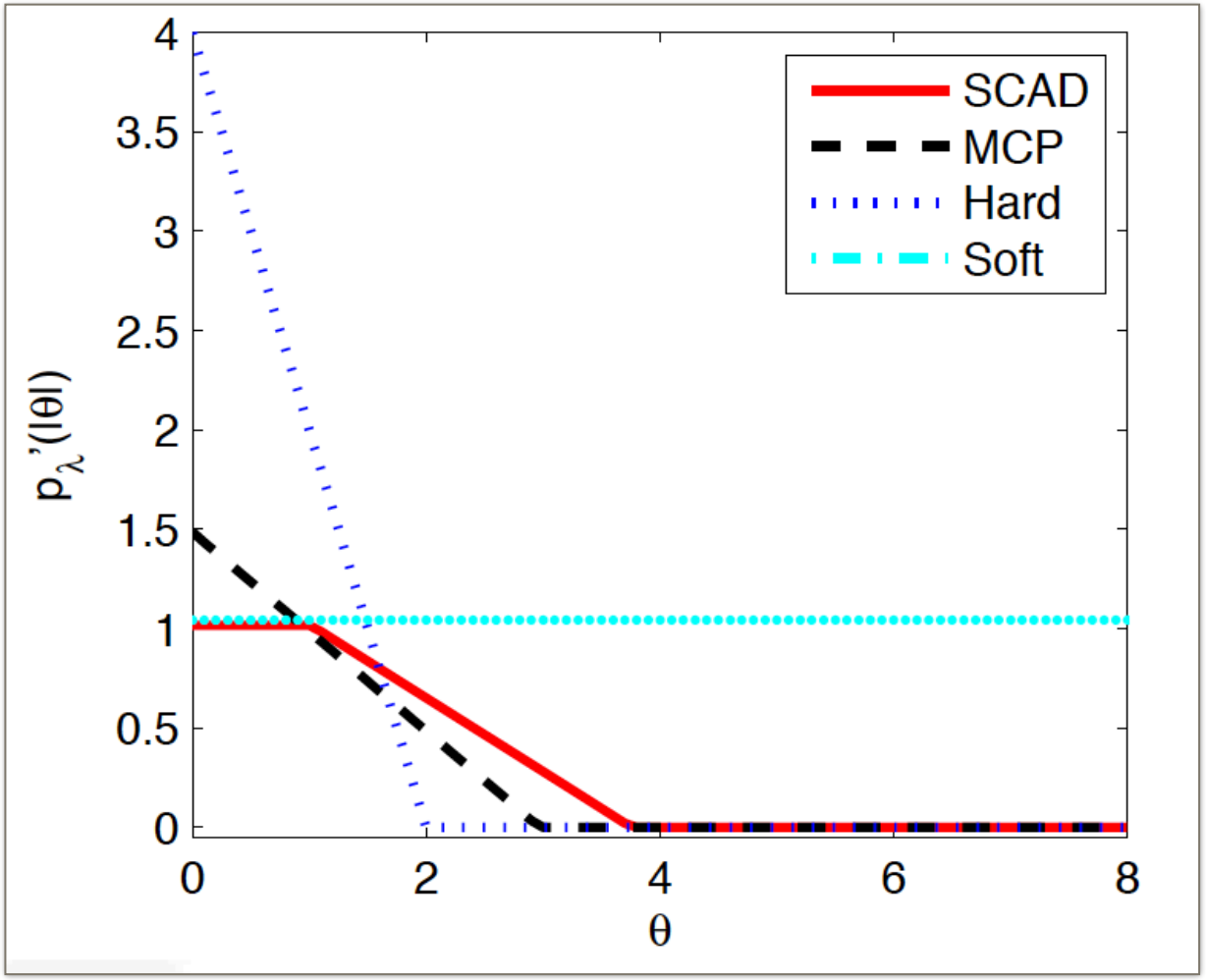
- R package: "quantreg", "FHDQR", "rqPen", etc.

- ℓ_1 penalty: introduce non-negligible estimation bias.

Concave Regularization



Smoothly Clipped Absolute Deviation: Fan & Li (2001)



Minimax Concave Penalty: C.-H. Zhang (2010)

Iteratively Reweighed ℓ_1-conquer

Starting with an initial estimate $\hat{\beta}^{(0)} = (\hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_p^{(0)})^\intercal$, at iteration $t = 1, \ldots, T$, solve the convex optimization problem

$$\min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_h(y_i - \mathbf{x}_i^\intercal \beta)}_{\hat{Q}_h(\beta)} + \underbrace{\sum_{j=2}^{p} q'_\lambda(|\hat{\beta}_j^{(t-1)}|)|\beta_j|}_{\|\lambda^{(t-1)} \circ \beta_-\|_1}$$

and obtain $\{\hat{\beta}^{(t)}\}_{t=1}^{T}$.

- $q_\lambda(t) = \lambda^2 q(t/\lambda)$, $q : [0,\infty) \to [0,\infty)$ is concave, increasing and $q(0) = 0$.
  - SCAD: $q'(t) = \min\left\{1, \left(1 - \frac{t-1}{a-1}\right)_+\right\}$;
  - MCP: $q'(t) = (1 - t/a)_+$;
  - Capped ℓ_1: $q'(t) = 1(t \le a/2)$.

Here $a > 2$ is a constant, say $a = 3.7$.

- We apply an Iterative Local Adaptive Majorize-Minimize (ILAMM) algorithm—a proximal gradient descent method—to compute weighted ℓ_1-penalized conquer.

- Given the previous iterate $\beta^{(\ell-1)}$, define an isotropic quadratic objective function

$$F(\beta; \phi, \beta^{(\ell-1)})$$
$$= \hat{Q}_h(\beta^{(\ell-1)}) + \langle \nabla \hat{Q}_h(\beta^{(\ell-1)}), \beta - \beta^{(\ell-1)} \rangle + \frac{\phi}{2} \|\beta - \beta^{(\ell-1)}\|_2^2.$$

- Minimizing $F(\beta; \phi, \beta^{(\ell-1)}) + \|\lambda \circ \beta\|_1$ yields

$$\beta^{(\ell)} = \mathrm{S}_{\mathrm{soft}}\left(\beta^{(\ell-1)} - \nabla \hat{Q}_h(\beta^{(\ell-1)})/\phi, \lambda/\phi\right),$$

where $\mathrm{S}_{\mathrm{soft}}(x, c) = \mathrm{sign}(x) \max(|x| - c, 0)$ is the soft-thresholding operator.

- The quadratic coefficient ϕ is chosen such that

$$F(\beta^{(\ell)}; \phi, \beta^{(\ell-1)}) \geq \hat{Q}_h(\beta^{(\ell-1)}).$$

Starting at a small value, say $\phi = 0.01$, iteratively increase ϕ by a factor of $\gamma = 1.25$ until the majorization property is met.

Thank you for your attention.

Softwares

- R: https://cran.r-project.org/web/packages/conquer/
- Python: https://github.com/WenxinZhou/Conquer

Manuscripts

- He, X., Pan, X., Tan, K. M. & Zhou, W.-X. (2020). Smoothed quantile regression with large-scale inference. *arXiv:2012.05187*.
- Tan, K. M., Wang, L. & Zhou, W.-X. (2020). High-dimensional quantile regression: convolution smoothing and concave regularization. *Preprint*.