Stochastic Search Variable Selection in Quantile Regression Based on Empirical Likelihood

Ruibin Xi
Peking University
第10届全国概率统计会议
Joint work with Yunxiao Li and Yiming Hu
Outline

- Introduction
  - Quantile regression
  - Empirical likelihood
- Bayesian model selection in quantile regression based on empirical likelihood
  - Asymptotic property
  - Gibbs sampler
  - Simulation
  - Real Data analysis
- Discussion
Quantile regression

- In linear model setup
  response = signal + i.i.d. error
  OLS for parameter estimating
Quantile regression

- In linear model setup
  response = signal + i.i.d. error
  OLS for parameter estimating
Quantile regression

- In linear model setup
  response = signal + i.i.d. error

OLS for parameter estimating
Quantile regression

- In linear model setup
  - response = signal + i.i.d. error
  - OLS for parameter estimating
The Check function

- We define a loss function
  \[ \rho_\tau(u) = \begin{cases} 
  \tau u & \text{if } u > 0 \\
  (\tau - 1)u & \text{if } u \leq 0
  \end{cases} \]

- Note that if \( \tau = 0.5 \), \( \rho_\tau(u) = \frac{|u|}{2} \)

- Quantiles solve a simple optimization problem
  \[ \hat{\alpha}(\tau) = \operatorname{argmin} \mathbb{E} \rho_\tau(Y - \alpha) \]
Quantile regression

The usual linear regression solves

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T b)^2$$

Quantile regression solves (Koenker and Bassett 1978; Koenker 2005)

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_T(y_i - x_i^T b)$$
Bayesian Analysis of quantile regression

- Bayesian methods of quantile regression
  - Skewed Laplace distribution (Yu and Moyeed 2001, Li et al. 2010)
  - Dirichlet process (Kottas and Krnjajic 2009)
  - Empirical likelihood (Yang and He 2012; Kim and Yang 2011)

- Advantage of Bayesian analysis
  - Easily incorporate prior information
  - Exact inference when sample size is small
Skewed Laplace distribution-based Bayesian analysis

- The skewed Laplace distribution
  \[ f(u|\sigma) = \tau(1-\tau)\sigma \exp\{-\sigma \rho_\tau(u)\} \]

- The joint likelihood
  \[ y_i = x_i^T \beta + u_i, \quad i = 1, \ldots, n \]
  \[ u_i \quad \text{i.i.d. skewed-Laplace} \]

- The joint likelihood
  \[ f(y|X) = \tau^n(1-\tau)^n\sigma^n \exp\{-\sigma \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta)\} \]

- Maximizing the likelihood is equivalent to minimizing
  \[ \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T b) \]
Bayesian model selection

- Bayesian Lasso
  \[ \pi(\beta_k | \sigma, \lambda) = \sigma \lambda \exp\{-\sigma \lambda |\beta_k|\} \]

- Bayesian elastic net
  \[ \pi(\beta_k | \eta_1, \eta_2) = C(\eta_1, \eta_2) \frac{\eta_1}{2} \exp\{-\eta_1 |\beta_k| - \eta_2 \beta_k^2\} \]

- Bayesian group lasso
  \[ \pi(\beta_g | \eta) = C_{dg} \sqrt{\det(K_g)} \eta^{dg} \exp\left(-\eta \|\beta_g\|_{K_g}\right) \]
Empirical likelihood (EL)

- First introduced by Owen (1988)
  - constructing confidence interval for the mean

- Linear model (Owen 1991), general estimating equation (Qin and Lawless 1994)

- Given an estimating equation \( \sum_{i=1}^{n} g(x_i, \theta) = 0 \), the EL is defined as

\[
L(\theta) = \sup \left\{ \prod_{i=1}^{n} p_i \mid \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(x_i, \theta) = 0, \text{ and } 0 \leq p_i \leq 1 \right\}.
\]
Empirical likelihood

- Asymptotic properties
  - Wilk’s theorem: the EL ratio $\xrightarrow{d} \chi^2_p$
  - The maximum EL estimator (MELE) is asymptotically normally distributed.

- Note: $g(x_i, \theta)$ usually need to be sufficiently smooth in $\theta$ for technical reasons
EL for quantile regression

- Taking directional derivative about $\beta$, the quantile regression estimates solves

$$\sum_{i=1}^{n} \phi_{\tau}(y_i - x_i^T \beta) x_i \approx 0,$$

$$\phi_{\tau}(t) = \tau - I_{[t<0]}$$

The EL for quantile regression is

$$L(\beta) = \sup \left\{ \prod_{i=1}^{n} p_i \left| \sum_{i=1}^{n} p_i \phi_{\tau}(y_i - x_i^T \beta) x_i = 0, \sum_{i=1}^{n} p_i = 1, 0 \leq p_i \leq 1 \right. \right\}$$
Model selection in EL settings

- Maximizing the penalized EL (Tang and Leng 2010)
  \[ \log(L(\theta)) - n \sum_{j=1}^{p} p\lambda(||\theta_j||), \]

- The penalty can be Lasso (Tibshirani), elastic net (Zou and Hastie 2005), SCAD (Fan and Li 2001) ....

- Difficulty:
  - Computationally expensive, especially for quantile regression
  - Choice of the tuning parameter
Bayesian Model selection in EL

- Put a “spike and slab” prior on $\beta_i$

$$\theta_i I_{\{\beta_i=0\}} + (1-\theta_i) I_{\{\beta_i\neq 0\}} N(0, \sigma^2)$$

- The hierarchical model is

$$Y|X, \beta \sim L(\beta|X, Y) = \sup \{ \prod_{i=1}^{n} p_i \mid \Sigma_{i=1}^{n} p_i \phi_r(y_i - x_i^T \beta)x_i = 0, \Sigma_{i=1}^{n} p_i = 1, 0 \leq p_i \leq 1 \}$$

$$\beta_i|\theta_i, \sigma \sim \theta_i I_{\{\beta_i=0\}} + (1-\theta_i) I_{\{\beta_i\neq 0\}} N(0, \sigma^2), \ i = 1, \ldots, p$$

$$\theta_i \sim U(0, 1), \ i = 1, \ldots, p$$

$$1/\sigma^2 \sim \Gamma(a, b) \ a > 0, b > 0.$$
Asymptotic property

Theorem 1 Under some regularity conditions, we have

- if $\beta_j = 0$, then $P(\beta_j = 0| \, X, Y) \to 1$ in probability.
- if $\beta_j \neq 0$, the posterior distribution of $\beta_j$ is approximately normally distributed.

Proof:

$$L(\beta|X, Y) = \exp\left\{ -\frac{n}{2}(\beta - \bar{\beta})^T V_{12} T V_{11}^{-1} V_{12} (\beta - \bar{\beta}) + O_p(n^{-1/2}) \right\}$$
Parameter estimation (1)

- Maximize the posterior likelihood?

\[ f(\beta, \theta, \sigma^2 | X, Y; \alpha, \eta) \propto L(\beta | X, Y) \prod_{i=1}^{p} \pi(\beta_i | \theta_i, \sigma^2) I_{(0,1)}(\theta_i) \Gamma\left(\frac{1}{\sigma^2}; a, b\right) \]

- Use Gibbs sampler

\[ f(\sigma^{-2} | \beta, \theta, X, Y) \propto \Gamma(a + h/2, b + \frac{1}{2} \sum_{j \in H} \beta_j^2) \]
\[ f(\theta_j | \beta, \theta_{-j}, \sigma^{-2}, X, Y) \propto \text{Beta}(1 + I(\beta_j = 1), 1 + I(\beta_j \neq 1)) \]
\[ f(\beta_j | X, Y, \theta, \sigma^2, \beta_{-j}) \propto L(\beta | X, Y) \pi(\beta_j | \theta_j, \sigma^2). \]

where \( H = \{j : \beta_j \neq 0\} \) \( h = \#H \)
Parameter estimation (1)

Maximize the posterior likelihood?

\[ f(\beta, \theta, \sigma^2 | X, Y; \alpha, \eta) \propto L(\beta | X, Y) \prod_{i=1}^{p} \pi(\beta_i | \theta_i, \sigma^2) I_{(0,1)}(\theta_i) \Gamma\left(\frac{1}{\sigma^2}; a, b\right) \]

Use Gibbs sampler

\[ f(\sigma^{-2} | \beta, \theta, X, Y) \propto \Gamma(a + h/2, b + \frac{1}{2} \sum_{j \in H} \beta_j^2) \]

\[ f(\theta_j | \beta, \theta_{-j}, \sigma^{-2}, X, Y) \propto \text{Beta}(1 + I(\beta_j = 1), 1 + I(\beta_j \neq 1)) \]

\[ f(\beta_j | X, Y, \theta, \sigma^2, \beta_{-j}) \propto L(\beta | X, Y) \pi(\beta_j | \theta_j, \sigma^2). \]

where \( H = \{ j : \beta_j \neq 0 \} \quad h = \# H \)

A mixture of the point mass at zero and a continuous distribution. Hard to sample
Parameter estimation (2)

- Use a Metropolis-Hastings (M-H) step to sample from

\[ f(\beta_j | X, Y, \theta, \sigma^2, \beta_{-j}) \]

- The M-H algorithm

  - Target: sample from \( \pi(x) \)
  - choose a proposing distribution \( q(x, y) \)
    - Generate \( y \) from \( q(x^{(i)}, \cdot) \) and \( u \) from \( U(0, 1) \)
    - If \( u \leq \alpha(x^{(i)}, y) \)
      - set \( x^{(i+1)} = y \).
    - Else
      - set \( x^{(i+1)} = x^{(i)} \).

where

\[
\alpha(x, y) = \min \left[ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right], \quad \text{if } \pi(x)q(x, y) > 0
\]

\[ = 1, \quad \text{otherwise.} \]
Parameter estimation (3)

- How to choose the proposing distribution?
    Given $\beta_i^{(t)}$ at the t-th step, the proposing distribution is
    $$q(\beta - \beta_i^{(t)})$$
    what q?
  - Kim and Yong (2011) proposed using a pre-specified distribution
Parameter estimation (3)

- How to choose the proposing distribution?

Random walk (Tierney 1994, Roberts et al. 1997):
Given at the $t$-th step, the proposing distribution is $q$.

Kim and Yong (2011) proposed using a pre-specified distribution.
Parameter estimation (4)

- If the EL $L(\beta)$ were smooth, the likelihood function can be approximated by

$$l(\beta_j) = \log(L(\beta_j, \beta_{-j}))$$

$$l(\beta_j) \approx l(\bar{\beta}_j) + \frac{1}{2} l''(\bar{\beta}_j)(\beta_j - \bar{\beta}_j)^2$$

where $l(\cdot)$ is maximized at $\bar{\beta}_j$

- Since the likelihood function convex, $v_j^{-2} = -l''(\bar{\beta}_j) > 0$

approximately

$$f(\beta_j|X, Y, \theta, \sigma^2, \beta_{-j}) \propto \exp\left\{-\frac{1}{2} v_j^{-2}(\beta_j - \bar{\beta}_j)^2\right\} \pi(\beta_j|\theta_j, \sigma^2)$$
Parameter estimation (4)

- If the EL $L(\beta)$ were smooth, the likelihood function can be approximated by

$$l(\beta_j) = \log(L(\beta_j, \beta_{-j}))$$

$$l(\beta_j) \approx l(\bar{\beta}_j) + \frac{1}{2} l''(\bar{\beta}_j)(\beta_j - \bar{\beta}_j)^2$$

where $l(\cdot)$ is maximized at $\bar{\beta}_j$

- Since the likelihood function convex, $v_j^-2 = -l''(\bar{\beta}_j) > 0$ approximately

$$f(\beta_j | X, Y, \theta, \sigma^2, \beta_{-j}) \propto \exp\left\{-\frac{1}{2} v_j^-2(\beta_j - \bar{\beta}_j)^2\right\}\pi(\beta_j | \theta_j, \sigma^2)$$

A mixture of the point mass at zero and a normal distribution. Easy to sample
Parameter estimation (5)

- $L(\beta)$ is not differentiable, we take $\bar{\beta}_j$ as the value that minimizes
  \[ \sum_{i=1}^{n} \rho_r(\tilde{y}_i - x_{ij}\beta_j). \]

  where \[ \tilde{y}_i = y_i - \sum_{l \neq j} x_{il}\beta_l \]

- Take $v_j^{-2}$ as the bootstrap variance of $\bar{\beta}_j$

- The proposing distribution is chosen as
  \[ \exp\{-\frac{1}{2}v_j^{-2}(\beta_j - \bar{\beta}_j)^2\} \pi(\beta_j|\theta_j, \sigma^2) \]
Parameter estimation (5)

- If parameter estimation is not differentiable, we take $\hat{\theta}$ as the value that minimizes:

$$
\text{where} \quad \hat{\theta} = \arg\min_{\theta} \left( \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2 \right)
$$
Bayesian quantile regression weighted at multiple quantiles (1)

Consider the model with homogeneous errors

\[ Y_i = \mu_1 + x_i \beta + u_i \]

with the \( \tau_1 \)th quantile of \( u_i \) being zero

The asymptotic variance of \( \beta \) is inversely proportional to \( f(\xi_{\tau_1}) \) (\( f \) is the density of \( u \))

If \( f(\xi_{\tau_2}) > f(\xi_{\tau_1}) \) for the \( \tau_2 \)th quantile \( \xi_{\tau_2} \)

\[ Y_i = \mu_2 + x_i^T \beta + w_i \]

\[ \mu_2 = \mu_1 + \xi_{\tau_2} - \xi_{\tau_1} \]

\[ w_i = u_i - (\xi_{\tau_2} - \xi_{\tau_1}) \]
Bayesian quantile regression weighted at multiple quantiles (2)

We may consider minimizing

$$\sum_{i=1}^{n} \left[ \rho_{\tau_1}(y_i - \mu_1 - x_i \beta) + \rho_{\tau_2}(y_i - \mu_2 - x_i \beta) \right].$$

More generally, given a set of quantile points $\tau_k \in (0, 1)$ we may minimize

$$\sum_{k=1}^{m} a_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mu_k - x_i^T \beta)$$

The corresponding EL is

$$L(\beta, \mu) = \sup \left\{ \prod_{i=1}^{n} p_i \mid \sum_{k=1}^{m} a_k \sum_{i=1}^{n} p_i \phi_{\tau_k}(y_i - \mu_k - x_i^T \beta)x_i = 0, \right.$$ 

$$\sum_{i=1}^{n} p_i \phi_{\tau_k}(y_i - \mu_k - x_i^T \beta) = 0, \forall 1 \leq k \leq m, \sum_{i=1}^{n} p_i = 1, 0 \leq p_i \leq 1 \right\}$$
Bayesian quantile regression weighted at multiple quantiles (3)

Similarly define the corresponding Bayesian model and get the following asymptotic property

**Theorem 2**  Under some regularity conditions, we have

- if $\beta_j = 0$, then $P(\beta_j = 0| X, Y) \to 1$ in probability.
- if $\beta_j \neq 0$, the posterior distribution of $\beta_j$ is approximately normally distributed.
Simulation Study

In the simulations, we compared

- **LASSO**
- **QR**: quantile regression
- **qrLasso**: QR with Lasso (Li and Zhu 2008)
- **bqrLasso**: Bayesian regularized QR with Lasso (Li et al. 2010)
- **BEQR**: Bayesian EL-based QR
- **BEQR.W**: Bayesian EL-based QR weighted at multiple quantiles
Simulation Study-homogeneous errors

Simulation setup

\[ y_i = x_i^T \beta_0 + u_i, \quad i = 1, \ldots, n \]
\[ \beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0) \]

where \( u_i \)'s have the \( \tau \)th quantile equal to 0.

The error distributions

\[ N(\mu, \sigma^2), \text{ with } \mu = 0, \sigma^2 = 9. \]
\[ \text{Laplace}(\mu, b), \text{ with } \mu = 0, b = 3. \]
\[ 0.6N(\mu_1 - a, \sigma^2) + 0.4N(\mu_2 - a, \sigma^2) \]
\[ \mu_1 = 2, \mu_2 = -2, \sigma^2 = 9, \text{ and } a = 0.4 \]
\[ 0.6\text{Laplace}(\mu_1 - a, b) + 0.4\text{Laplace}(\mu_2 - a, b) \]
Simulation Study-homogeneous errors

<table>
<thead>
<tr>
<th>quantile</th>
<th>Method</th>
<th>Error Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>normal</td>
</tr>
<tr>
<td>$\theta = 0.9$</td>
<td>Lasso</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>qrLasso</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>BEQR,W</td>
<td>0.41</td>
</tr>
<tr>
<td>$\theta = 0.5$</td>
<td>Lasso</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>qrLasso</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td><strong>0.37</strong></td>
</tr>
<tr>
<td></td>
<td>BEQR,W</td>
<td>0.39</td>
</tr>
<tr>
<td>$\theta = 0.1$</td>
<td>Lasso</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>qrLasso</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>BEQR,W</td>
<td><strong>0.41</strong></td>
</tr>
</tbody>
</table>
## Simulation Study-homogeneous errors

<table>
<thead>
<tr>
<th>$\theta$/mean</th>
<th>Method</th>
<th>Error Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TP/FP</td>
</tr>
<tr>
<td>mean</td>
<td>Lasso</td>
<td>3.00/2.30</td>
</tr>
<tr>
<td>$\theta = (0.9, 0.5, 0.1)$</td>
<td>BEQR.W</td>
<td>3.00/0.18</td>
</tr>
<tr>
<td>$\theta = 0.9$</td>
<td>qrLasso</td>
<td>3.00/2.78</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>3.00/0.86</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td>2.96/0.19</td>
</tr>
<tr>
<td>$\theta = 0.5$</td>
<td>qrLasso</td>
<td>3.00/2.01</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>3.00/0.20</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td>2.98/0.12</td>
</tr>
<tr>
<td>$\theta = 0.1$</td>
<td>qrLasso</td>
<td>3.00/2.78</td>
</tr>
<tr>
<td></td>
<td>bqrLasso</td>
<td>2.99/0.86</td>
</tr>
<tr>
<td></td>
<td>BEQR</td>
<td>2.97/0.19</td>
</tr>
</tbody>
</table>
Simulation Study-heterogeneous errors

Consider the model

\[ y_i = \beta_{10}x_{i1} + \sum_{j=2}^{8} \beta_{j0}x_{ij} + x_{i1}\epsilon_i \]

\( \epsilon_i \) are generated as in the i.i.d. case

\[ \beta_0 = (\beta_{10}, \cdots, \beta_{80}) = (3, 1.5, 0, 0, 2, 0, 0, 0) \]
## Simulation Study-heterogeneous errors

<table>
<thead>
<tr>
<th>quantile</th>
<th>Method</th>
<th>Error Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>normal</td>
</tr>
<tr>
<td>θ = 0.90</td>
<td>Lasso 2.05</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>qrLasso 0.95</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>bqrLass 0.37</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>BEQR 0.28</td>
<td>0.46</td>
</tr>
<tr>
<td>θ = 0.50</td>
<td>Lasso 0.41</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>qrLasso 0.31</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>bqrLass 0.26</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>BEQR 0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>θ = 0.10</td>
<td>Lasso 1.95</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>qrLasso 0.55</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>bqrLass 0.35</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>BEQR 0.35</td>
<td>0.64</td>
</tr>
</tbody>
</table>
An application

- microRNA: small non-coding RNA binds to 3-UTR region of mRNAs

- Contradicting opinion about microRNA
  - Canalization effect (reduce gene expression variance)
    Hornstein and Shomron, Nature Genetics, 2006
    Wu et al. Genome Research 2009
  - Increase gene expression variation
    Lu and Clark, Genome Research 2012
An application

- microRNA: small non-coding RNA binds to 3'UTR region of mRNAs

- Contradicting opinion about microRNA

- Canalization effect (reduce gene expression variance)
  Hornstein and Shomron, Nature Genetics, 2006
  Wu et al. Genome Research 2009

- Increase gene expression variation
  Lu and Clark, Genome Research 2012
An application

- Data: RNAseq data from 70 individuals
- \(~ 20000\) genes
- \(Y\): expression variation for each gene
- Covariates:
  - Mean Expression
  - \# of microRNA targets, target Score of 3`-UTR
  - \# of SNP on 3`-UTR, Gene Length, length of 3`-UTR
<table>
<thead>
<tr>
<th>Method</th>
<th>Error Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau = 0.1$ $\tau = 0.3$ $\tau = 0.5$ $\tau = 0.7$ $\tau = 0.9$</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.08109 0.12607 0.13156 0.16285 0.11509</td>
</tr>
<tr>
<td>qrLasso</td>
<td>0.03792 0.12233 0.20715 0.22757 0.13412</td>
</tr>
<tr>
<td>qrLasso.S</td>
<td>0.03750 0.09058 0.12150 0.13073 0.10171</td>
</tr>
<tr>
<td>QR</td>
<td>0.03749 0.09056 0.12144 0.12743 0.09053</td>
</tr>
<tr>
<td>bqrLasso</td>
<td>0.03750 0.09057 0.12143 0.12742 0.09053</td>
</tr>
<tr>
<td>BEQR</td>
<td>0.03750 0.09059 0.12146 0.12743 0.09062</td>
</tr>
</tbody>
</table>
Conclusion

- Developed an EL based Bayesian model selection method in quantile regression

- Asymptotic property

- Simulation study shows that BEQR and BEQR.W performs better in general

- Disadvantage: cannot handle $p >= n$. 
Postdoc position on bioinformatics available!

ruibinx@math.pku.edu.cn