Bayesian Quantile Regression using a Mixture of Polya Trees

Minzhao Liu and Mike Daniels

Department of Statistics
University of Florida

2012.4.4
Outline

Quantile Regression

Polya Tree

QR with Polya Tree
What is Quantile Regression

Suppose random variable $Y$ has cdf $F$, then the $\tau$-th quantile of $Y$ is

$$Q_Y(\tau) = \inf \{ y : F(y) \geq \tau \},$$

Furthermore, if we have covariates $X$, then the quantile regression parameter $\beta$ satisfies this condition

$$Q_Y(\tau|x) = x'\beta(\tau),$$

if $F$ is continues , then

$$\Pr(y \leq x'\beta(\tau)) = \tau.$$
Why We Need Quantile Regression

Figure: Engel Curves for Food: This figure plots data taken from Engel’s (1857) study of the dependence of households’ food expenditure on household income.
Advantage over regular mean regression:
- more information provided,
- slope may vary with different $\tau$. 
Recall least squared estimator for mean regression model is:

\[ \beta = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x'_i b)^2 \]

In quantile regression, we have: The true \( \tau \)-th quantile regression parameter \( \beta(\tau) \) minimizes the expectation of the check function:

\[ \beta(\tau) = \arg \min_{b \in \mathbb{R}^p} \mathbb{E} (\rho_{\tau}(y - x'_i b)) , \]

\[ \rho_{\tau}(z) = z (\tau - I(z < 0)) . \]

Empirically, we take

\[ \beta(\tau) = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (\rho_{\tau}(y - x'_i b)) , \]
Asymmetric Laplace Distribution pdf:

\[ f_\tau(z; \mu, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp \left\{ -\rho_\tau \left( \frac{z - \mu}{\sigma} \right) \right\} \]

Property:

\[ F_z(\mu) = \tau \]

If we assume

\[ y = x^T \beta + \epsilon \]
\[ \epsilon \sim \text{ASL}_\tau(0, \sigma) \]

then,

\[ L(\beta, \sigma | Y) \propto \sigma^{-n} \exp \left\{ -\sum_{i=1}^{n} \rho_\tau \left( \frac{y_i - x_i^T \beta}{\sigma} \right) \right\} \]

\[ \Leftrightarrow \beta(\tau) = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (\rho_\tau(y - x_i^T b)) \]
Remarks

- no distribution assigned in check function
- one ALD distribution for all models
- for different $\tau$, separate model needed
Our Proposal

Consider location-shift model

\[ Y = X\beta + \epsilon, \quad \epsilon \sim F_\epsilon \]

\[ \Rightarrow \beta(\tau) = \beta + F_\epsilon^{-1}(\tau)e_1 \]

conditional quantiles

beta_0

beta_1

\[ y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i. \]
Heterogeneity

\[ Y = x\beta + (x\gamma)\epsilon, \quad \epsilon \sim F_\epsilon \]

\[ \Rightarrow \beta(\tau) = \beta + \gamma F^{-1}_\epsilon(\tau) \]

conditional quantiles

\[ y_i = \beta_0 + \beta_1 x_{i1} + (1 + x_{i1})\epsilon_i \]

Question: How to get \( F^{-1}_\epsilon(\tau) \),
A Flexible Nonparametric Prior: Polya Tree

PT parameters:

1. Partitions:
   \[ \Sigma = \{ B_0, B_1, B_{00}, \ldots \} \]
2. Weights
   \[ \mathcal{A} = \{ \alpha_0, \alpha_1, \ldots, \alpha_\epsilon, \ldots \} \]
3. \( Y_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}) \)
4. \( \alpha_{\epsilon 0} = \alpha_{\epsilon 1} = cj^2 \),
   absolutely continuous
5. Partial / finite PT, M level

\[ G | \Pi, \mathcal{A} \sim PT(\Pi, \mathcal{A}) \]
Why and advantages

1. flexible, nonparametric
2. posterior tractable

\[ x \mid G \sim G \]
\[ G \sim PT(B, A) \]
\[ \Rightarrow G \mid x \sim PT(B, A^*) \text{ with } \alpha^*_\epsilon = \begin{cases} 
\alpha_\epsilon + 1 & \text{if } x \in B_\epsilon \\
\alpha_\epsilon & \text{otherwise}
\end{cases} \]

3. easy to get posterior quantile
4. easy to fix median or other quantiles
5. can be absolutely continuous, while DP cannot
Since the behavior of polya tree prior highly depends on the partition parameter $\Pi$, thus a mixture of Polya tree prior was defined this way:

\[
\begin{align*}
    x | G & \sim G \\
    G & \sim PT(B, A) \\
    \Rightarrow G | \Pi^\theta, A^\theta & \sim PT(\Sigma^\mu, \sigma, A^c)
\end{align*}
\]

where $\Sigma$ partition was constructed based on baseline measure $N(\mu, \sigma^2)$. and then assign $\theta = (\mu, \sigma^2, c)$ a prior.

\[
\theta \sim \pi(\theta)
\]
Predictive Density, Cumulative, and Quantiles

\[ X_1, \ldots, X_n | G \sim G \]
\[ G | \Pi \sigma^2, A^c \sim PT(\Pi \sigma^2, A^c) \]

where \( G_0 = \mathcal{N}(0, \sigma^2) \) is the baseline measure, and \( g_0(x) \) is its density function. Then, from the partial PT, the predictive density function is

\[
f(x | X_1, \ldots, X_n) = \left[ \prod_{j=2}^M \frac{c_j^2 + n(\epsilon(j, x)|X)}{2c_j^2 + n(\epsilon(j - 1, x)|X)} \right] 2^{M-1} g_0(x)
\]
Integrating the predictive density function,

\[ F(x|X_1, \ldots, X_n) = \sum_{i=1}^{N-1} P_i + P_N \left( G_0(x)2^M - (N - 1) \right) \]

\[ P_i = \frac{1}{2} \prod_{j=2}^{M} \frac{c_j^2 + n(\epsilon(j, x)|X)}{2c_j^2 + n(\epsilon(j - 1, x)|X)} \]

\[ N = \left[ 2^M G_0(x) + 1 \right] \]

By inverting the predictive cumulative density function,

\[ F_{X|X}^{-1}(\tau) = G_0^{-1} \left[ \frac{\tau - \sum_{i=1}^{N} P_i + N \cdot P_N}{2^M P_N} \right] \]

\[ N \text{ satisfy } \sum_{i=1}^{N-1} P_i < \tau \leq \sum_{i=1}^{N} P_i \]
Quantile Regression with PT

Suppose we have a location-shift model, with heterogeneity form:

\[ y_i = x_i \beta + (x_i \gamma) \epsilon_i \]
\[ \epsilon_i | G \sim G \]
\[ G | \Pi, \mu, \sigma, \mathcal{A} \sim PT(\Pi^{\mu,\sigma}, \mathcal{A}) \]

in order not to confound with location parameter \( \beta \), we fix \( \epsilon \)'s median as 0 (\( \mu = 0 \)). Also, to deal with the identifiability problem of \( \gamma \), the first element of \( \gamma \) is set to be 1.

Estimation:

\[ \beta(\tau) = \beta + \gamma F_{\epsilon}^{-1}(\tau) \]
\[ p(\beta(\tau) | Y) = p(h(\beta, \gamma, F_{\epsilon}^{-1}) | Y) \]
Compare our quantile regression model with Polya trees priors (HeterPTlm) with Koenker’s ’rq’ function in R package ’quantreg’.

- $Y_i = 1 + x_{i1} + x_{i2} + N(0, 1)$
- $Y_i = 1 + x_{i1} + x_{i2} + \text{Gamma}(3, 1)$
- $Y_i = 1 + x_{i1} + x_{i2} + \text{MixtureNormal}$
- $Y_i = 1 + x_{i1} + x_{i2} + (1 - 0.5x_{i1} + 0.5x_{i2})\text{MixtureNormal}$
- $Y_i = 1 + x_{i1} + x_{i2} + (1 - 0.5x_{i1} + 0.5x_{i2})\text{Gamma}(3, 1)$

where Mixture Normal $\sim 0.5N(-2, 1) + 0.5N(2, 1)$.

100 datasets and each dataset contains 100 observations. Evaluate $\tau = 0.5$ and $\tau = 0.9$. 
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>Coverage</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rq</td>
<td>HeterPTlm</td>
<td>rq</td>
</tr>
<tr>
<td>M1.5</td>
<td>1.39(0.13)</td>
<td>1.06(0.12)</td>
<td>0.82</td>
</tr>
<tr>
<td>M2.5</td>
<td>3.31(0.36)</td>
<td>3.18(0.35)</td>
<td>0.89</td>
</tr>
<tr>
<td>M3.5</td>
<td>17.2(1.48)</td>
<td>2.21(0.30)</td>
<td>0.85</td>
</tr>
<tr>
<td>M4.5</td>
<td>95.4(6.69)</td>
<td>9.87(1.09)</td>
<td>0.86</td>
</tr>
<tr>
<td>M5.5</td>
<td>14.3(1.36)</td>
<td>10.0(1.04)</td>
<td>0.89</td>
</tr>
<tr>
<td>M1.9</td>
<td>2.35(0.26)</td>
<td>2.07(0.23)</td>
<td>0.95</td>
</tr>
<tr>
<td>M2.9</td>
<td>15.2(1.51)</td>
<td>13.5(1.61)</td>
<td>0.9</td>
</tr>
<tr>
<td>M3.9</td>
<td>3.68(0.46)</td>
<td>3.94(0.56)</td>
<td>0.93</td>
</tr>
<tr>
<td>M4.9</td>
<td>25.1(2.66)</td>
<td>14.8(1.41)</td>
<td>0.88</td>
</tr>
<tr>
<td>M5.9</td>
<td>73.7(8.48)</td>
<td>60.2(8.14)</td>
<td>0.93</td>
</tr>
</tbody>
</table>

The length of the credible interval from HeterPTlm is also shorter than that from rq function.