A Comparison and Integration of Quantile Regression and Finite Mixture Modeling

Richard J. Willke, Ph.D., Ching-Ray Yu, Ph.D., Birol Emir, Ph.D., Kelly H. Zou, Ph.D., PStat®, and Javier Cabrera, Ph.D.

1. Introduction
- Quantile regression (QR) and finite mixture models (FMM) are useful methods for estimating treatment effects at different quantiles in the outcome distribution while controlling for covariates.
- FMM is a parametric method with assumptions for the error distribution.
- However, QR and FMM may not yield similar estimates at different quantiles, which may lead to different interpretations of the results.
- There are 2 limitations that justify a non-parametric approach to the finite mixture: the error distribution may be unknown, and there may be outliers.

2. Objectives
- To propose new methodology, EQR MIX, for estimating FMM components by selecting a group of quantiles from QR to satisfy an optimality criterion.
- To integrate EQR MIX with FMM in a way that incorporates both nonparametric and parametric components to interpret estimated effects.

3. Algorithms for QRMIX and EQR MIX
- Choose the number of clusters k and a list of quantiles \(\{q_1, \ldots, q_k\}\) for which those k quantile regressions are the best alternatives to the sets of FMM estimates.
- Perform a quantile regression for the quantiles in Q, and extract the matrix of coefficients \(B\) for the models and the matrix of all residuals, where \(R\) has dimensions (sample size of the dataset)(number of quantiles in Q).
- For every set of k quantiles, \(\{q_1, \ldots, q_k\}\) of Q, assign each observation to the closest quantile line (i.e., the closest cluster).
- Define C(S), the residual sum of squares for all observations across all clusters. Minimize C(S) over all of the subsets of k quantiles on Q to obtain the best mixture.
- To make QRMIX more efficient, EQR MIX, calculate Huber’s M-estimators for regression models for each cluster and the corresponding residuals.
- Reassign each observation to the closest lines from the M-estimators.
- The criterion is the within-group residual sum of squares. A generalized weighted version may also be considered to improve the efficiency.
- Fig. 1 illustrates QRMIX showing 99 quantiles and the optimal clusters.

3.2. Constraints to Speed Up the Optimization Process
- If \(k=2\) and \(S=q_1,q_2\) is a subset of Q, then use the follow constraint \(q_1<0.5q_q\) and \(q_2>0.5q_q\) For convenience, \(q=0.03\).
- If \(k>2\) and \(S=q_1,q_2,\ldots,q_d\), then use the constraint \(q_1<0.5q_q\) and \(q_d>0.5q_q\) where \(d=1,\ldots, k-1\).

4. A Clinical Trial Example
- 4.1. Systolic Hypertension in the Elderly Program (SHEP) Clinical Trial
- In the SHEP trial, all of the patients (N=4064) had hypertension.
- This post-hoc illustration is aimed to validate the standard clinical ranges of BMI as low (BMI<25), medium (25<BM≤30), and high (BMI>30).
- Besides BMI, there are 27 additional variables, including demographics (e.g., race, gender, age), behavioral (e.g., smoking), comorbid conditions (e.g., stroke, diabetes), laboratory tests (e.g., sodium, glucose), and clinical variables (e.g., treatments, i.e., medication vs. placebo), etc.

4.2. Comparison of FMM and EQRMIX
- The 3 identified BMI clusters are low, medium, and high, along with their densities (Fig. 2), which are predicted by FMM and EQR MIX, respectively.
- The ranges of the 3 BMI groups estimated by EQR MIX correspond approximately to standard BMI classifications in clinical practice.

5. Monte-Carlo Simulations
- Each dataset has 1 response variable and 4 predictors in 3 clusters:
  - Cluster 1: BMI=20+0.1xSodium+0.001xTriglycerides+0.3xTreatment + e_1
  - Cluster 2: BMI=20+0.0015xTriglycerides+0.0025xCholes+0.3xTreatment+ e_2
  - Cluster 3: BMI=20+0.05xSodium+0.002xCholes+0.3xTreatment+ e_3
- Table 1 specifies several different error distributions in simulations.
- Fig. 3 represents an example of BMI density functions.

6. Conclusions
- EQR MIX is robust and yields fewer misclassifications than FMM.
- If FMM and EQR MIX yield different results, examine outliers or skewness.
- If data show strong departures from parametric assumptions, proceed with EQR MIX, or consider alternative assumptions for FMM.
- A caution is that FMM and EQR MIX may be biased if the overlap among the clusters is high (e.g., high variance or high skewness in large samples).

7. References

Table 1. Underlying Distributions.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>standard normal (N(0,1))</td>
</tr>
<tr>
<td>Long Tailed</td>
<td>t-distribution ((p=19))</td>
</tr>
<tr>
<td>Outliers</td>
<td>80% ((p=10)), 20% ((p=90))</td>
</tr>
</tbody>
</table>

Table 2. Number of Misclassifications by Method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Sample Size=500</th>
<th>Total Sample Size=4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQRMIX</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>FMM</td>
<td>0.43</td>
<td>0.41</td>
</tr>
<tr>
<td>QRMIX</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>EQRMIX</td>
<td>0.93</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Appendix

Fig. 1. QRMIX with 99 Quantile Lines and 3 Clusters.

Fig. 2. Densities of BMI by Cluster.
Blue=FMM;
Red=EQRMI

Fig. 3. Mixture Densities.

Fig. 4. Treatment Effect Estimates.
Left: Dark Blue=FMM;
Middle: Light Blue=QRMIX;
Right: Red=EQRMI.