Matching Quantiles Estimation: selecting representative portfolios

Qiwei Yao
Department of Statistics, London School of Economics
q.yao@lse.ac.uk

Joint work with
Nikolaos Sgouropoulos, Barclays Bank
Claudia Yastremiz, Bank of England
Motivations: Basel III backtesting

MQE: matching quantiles estimation
- Lasso MQE – sparse representation
- matching a part of distribution

Convergence of the algorithm

Asymptotic properties of MQE

Goodness-of-match: a measure and a test

Numerical illustration
- simulation
  - a real data example

Tracking portfolios
**Basel III**: a set of comprehensive reform measures that puts in place a global regulatory standard on bank capital adequacy, stress testing and market liquidity, it was developed by the Basel Committee on Banking Supervision and agreed in September 2010.

**The CCR backtesting**: the counterparty credit risk backtesting is one of the mandatory requirements under Basel III. It compares the risk measures implied by the pricing models employed by the bank with the realised exposure calculated based on the traded prices.

**A counterparty representative portfolio**: Basel III allows banks to construct representative portfolios for each counterparty. A representative portfolio may consist of a subset of the trades between the two banks. Banks are left to decide the number and trades to be included in the portfolio, but they have to justify their choices to their supervisors (i.e. the FSA).
\( Y: \) the total portfolio of a counterparty
\( \mathbf{X} = (X_1, \cdots, X_p)': p \) mark-to-market values of trades.

**Goal:** identify a representative portfolio

\[
\mathbf{\beta}' \mathbf{X} = \mathbf{\beta}_1 X_1 + \cdots + \mathbf{\beta}_p X_p
\]

which provides an adequate approximation for \( Y \).

The approximation could be two-fold:

- \( E(\mathbf{\beta}' \mathbf{X}) \approx E(Y) \), or
- \( \mathcal{L}(\mathbf{\beta}' \mathbf{X}) \approx \mathcal{L}(Y) \)

Basel III requires that the representative portfolios match the whole distribution including the characteristics such as mean, variance, VaR, and the sensitivities of risk factors.
Matching the whole distribution can be achieved by searching for \( \beta \) to minimize

\[
\int_0^1 \left\{ Q_Y(\alpha) - Q_{\beta'X}(\alpha) \right\}^2 d\alpha,
\]

where \( Q_\xi(\alpha) \) denotes the \( \alpha \)-th quantile of r.v. \( \xi \), i.e.

\[
P\{\xi \leq Q_\xi(\alpha)\} = \alpha, \quad \text{for } \alpha \in [0, 1].
\]

This leads to the MQE:

\[
\hat{\beta} = \arg \min_{\beta} \sum_{j=1}^{n} \{ Y(j) - (\beta'X)(j) \}^2,
\]

where

\[
(\beta'X)(1) \leq \cdots \leq (\beta'X)(n)
\]

are the order statistics of \( \{\beta'X_j\} \).
Remarks

1. $Y(j)$ is the $\frac{j}{n}$-th sample quantile for $Y$. Hence the MQE tries to match two sample quantiles at all levels.

2. We match the two quantile functions, instead of CDF or PDF, as the extreme quantiles are more important.

3. To match the means, we should use the OLS:

$$\hat{\beta} = \arg\min_{\beta} \sum_{j=1}^{n} \{Y_j - \beta'X_j\}^2.$$ 

The calculation for MQE is carried out by an iterative OLS algorithm.

4. The MQE and OLS estimate different $\beta$.

5. Discard the pairing between $Y_j$ and $X_j$. 
An iterative algorithm for computing the MQE $\hat{\beta}$:

**Step 1** Set an initial value $\beta^{(0)}$. (For example, $\beta^{(0)} = \tilde{\beta}$.)

**Step 2** For $k \geq 1$, let $X_{(1)}^{(k-1)}, \ldots, X_{(n)}^{(k-1)}$ be a permutation of $\{X_j\}$ such that

$$(\beta^{(k-1)})'X_{(1)}^{(k-1)} \leq \cdots \leq (\beta^{(k-1)})'X_{(n)}^{(k-1)}.$$

Define $\beta^{(k)} = \arg \min_\beta R_k(\beta)$, where

$$R_k(\beta) = \frac{1}{n} \sum_{j=1}^{n} \left( Y_j - \beta'X_{(j)}^{(k-1)} \right)^2.$$

We stop the iteration when $|R_k(\beta^{(k)}) - R_{k-1}(\beta^{(k-1)})|$ is small, and set $\hat{\beta} = \beta_k$. 

Toy example 1

Let $Y = X + Z$, where $X, Z$ are independent $N(0, 1)$.

**OLS:** $\beta = \arg\min_\beta \sum_{j=1}^n (Y_j - \beta X_j)^2$

**True value:** $1 = \arg\min_\beta E(Y_j - \beta X_j)^2$

**MQE:** $\hat{\beta} = \arg\min_\beta \sum_{j=1}^n \{Y(j) - (\beta X)(j)\}^2$

**True value:** $\sqrt{2} = \arg\min_\beta \int_0^1 \{Q_Y(\alpha) - Q_{\beta X}(\alpha)\}^2 d\alpha. \quad (L(Y) = L(\sqrt{2} X))$

1000 samples of size $n = 100$

$rMSE(\tilde{\beta}) = .0107$

$rMSE(\hat{\beta}) = .0109$

The algorithm converge after 2 iterations with initial value $\tilde{\beta}$. 

- p.8
Note $-\sqrt{2} = \arg\min_{\beta} \int_{0}^{1} \left\{ Q_Y(\alpha) - Q_{\beta X}(\alpha) \right\}^2 d\alpha$.

The algorithm converge after 2 iterations with initial value $-\tilde{\beta}$.

Algorithm
**Toy example 2**

Let \( Y = X_1 + X_2 + 1.414Z \), where \( X_1, X_2, Z \) are indep \( N(0, 1) \).

**OLS** \((\tilde{\beta}_1, \tilde{\beta}_2)\) estimate the true value \( (1, 1) \).

As for MQE, many true values as

\[
\mathcal{L}(Y) = N(0, 4) = \mathcal{L}(\beta_1 X_1 + \beta_2 X_2), \quad \forall (\beta_1^2 + \beta_2^2)^{1/2} = 2
\]

A likely true value: \((\sqrt{2}, \sqrt{2})\)
With random initial values $\beta_1^{(0)}, \beta_2^{(0)} \sim U(-2, 2)$
Remarks

1. There may exist more than one minimizers for

\[ \int_0^1 \{ Q_Y(\alpha) - Q_{\beta'X}(\alpha) \}^2 d\alpha \]

when, e.g., two of the components of \(X\) are iid.

There exists a unique LS approximation \(\mathcal{L}(\beta'X)\) for \(\mathcal{L}(Y)\).

2. To only match a part of \(\mathcal{L}(Y)\), say, between the \(\alpha_1\)-th quantile and the \(\alpha_2\)th quantile, let \(n_i = \lceil n\alpha_i \rceil\) and define

\[ R_k(\beta; \alpha_1, \alpha_2) = \frac{1}{n} \sum_{j=n_1+1}^{n_2} (Y(j) - \beta'X^{(k-1)}(j))^2. \]

3. Lasso MQE:

\[ R_k(\beta) = \frac{1}{n} \sum_{j=1}^{n} (Y(j) - \beta'X^{(k-1)}(j))^2 + \lambda \sum_{i=1}^{p} |\beta_i| \]
Convergence of the algorithm

Recall $\beta^{(k)} = \text{arg min}_\beta R_k(\beta)$, where

$$R_k(\beta) = \frac{1}{n} \sum_{j=1}^{n} \left( Y(j) - \beta'X(j)^{(k-1)} \right)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

and $\{(\beta^{(k-1)})'X(j)^{(k-1)}\}$ are the order statistics of $\{(\beta^{(k-1)})'X_j\}$.

**Theorem 1.** As $k \to \infty$, $R_k(\beta^{(k)}) \to c$, where $c \geq 0$ is a constant.

**Proof.**

$$R_k(\beta^{(k)}) = \frac{1}{n} \sum_{j=1}^{n} \left( Y(j) - \beta^{(k)}'X(j)^{(k-1)} \right)^2$$

$$\leq \frac{1}{n} \sum_{j=1}^{n} \left( Y(j) - \beta^{(k-1)}'X(j)^{(k-1)} \right)^2 \leq \frac{1}{n} \sum_{j=1}^{n} \left( Y(j) - \beta^{(k-1)}'X(j)^{(k-2)} \right)^2$$

$$= R_{k-1}(\beta^{(k-1)}).$$

**Lemma 1.** For any real numbers $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$,

$$\sum_{i=1}^{n} (a(i) - b(i))^2 \leq \sum_{i=1}^{n} (a_i - b_i)^2.$$
Asymptotic properties of the MQE

For any given $0 \leq \alpha_1 < \alpha_2 \leq 1$, let $\beta_0$ be a minimizer of

$$
S'(\beta) \equiv S(\beta; \alpha_1, \alpha_2) = \int_{\alpha_1}^{\alpha_2} \left\{ Q_Y(\alpha) - Q_{\beta'}X(\alpha) \right\}^2 d\alpha + \lambda \sum_{j=1}^{p} |\beta_j|.
$$

MQE: $\hat{\beta} = \arg\min_{\beta} S_n(\beta), \quad n_i = \lfloor n\alpha_i \rfloor$

$$
S_n(\beta) \equiv S_n(\beta; \alpha_1, \alpha_2) = \frac{1}{n} \sum_{j=n_1+1}^{n_2} \left\{ Y(j) - (\beta'X)(j) \right\}^2 + \lambda \sum_{j=1}^{p} |\beta_j|.
$$

As $S(\beta)$ may have many minimizers, let

$$
B_0 = \{ \beta : S(\beta) = S(\beta_0) \}, \quad d(\hat{\beta}, B_0) = \min_{\beta \in B_0} \| \hat{\beta} - \beta \|.
$$
Asymptotic properties of the MQE

Theorem 2. Let conditions (i)–(iii) hold. Then as $n \to \infty$,

(i) $S_n(\hat{\beta}) \xrightarrow{P} S(\beta_0)$

(ii) $d(\hat{\beta}, B_0) \to 0$.

Two keys in the proof:
the Bahadur-Kiefer bounds for quantiles, and

Lemma 2. Let $b_i = a_i + \delta_i$ for $i = 1, \cdots, n$. Then

$$\max_{1 \leq i \leq n} |a(i) - b(i)| \leq \max_{1 \leq i \leq n} |\delta_i|.$$
Conditions (i)–(iii)

(i) Let \( \{Y_j\} \) be iid, and \( \{X_j\} \) be iid. Both \( f_Y(\cdot) \) and \( f_X(\cdot) \) exist.

(ii) (The Kiefer condition.) It holds for any fixed \( \beta \) that

\[
\sup_{\alpha_1 \leq \alpha \leq \alpha_2} |f'_{\beta'X}(Q_{\beta'X}(\alpha))| < \infty, \quad \inf_{\alpha_1 \leq \alpha \leq \alpha_2} f_{\beta'X}(Q_{\beta'X}(\alpha)) > 0.
\]

Furthermore

\[
\sup_{\alpha_1 \leq \alpha \leq \alpha_2} |f'_Y(Q_Y(\alpha))| < \infty, \quad \inf_{\alpha_1 \leq \alpha \leq \alpha_2} f_Y(Q_Y(\alpha)) > 0.
\]

(iii) \( X \) has a bounded support.

Note. The Kiefer condition implies, e.g.

\[
\sup_{\alpha_1 \leq \alpha \leq \alpha_2} \left| \sqrt{n}f_Y(Q_Y(\alpha))\{Q_{n,Y}(\alpha) - Q_Y(\alpha)\} + \sqrt{n}\{F_{n,Y}(Q_Y(\alpha)) - \alpha\} \right| = O_P\left(n^{-1/4}(\log n)^{1/2}(\log \log n)^{1/4}\right).
\]
Goodness-of-match

Let \( F(y) = P(Y \leq y) \), \( g(\cdot) \) be the PDF of \( F(\beta'X) \).

When \( \mathcal{L}(Y) = \mathcal{L}(\beta'X) \), \( F(\beta'X) \sim U(0, 1) \) and \( g(x) = I(0 < x < 1) \).

**Measure for GoM:** \( \rho = 1 - \frac{1}{2} \int_0^1 |g(x) - 1| \, dx = 1 - 0.5 \int_0^1 |dG - dx| \).

Then \( \rho = 1 \) iff \( \mathcal{L}(Y) = \mathcal{L}(\beta'X) \), and \( 0 \) iff the supports of \( \mathcal{L}(Y) \) and \( \mathcal{L}(\beta'X) \) do not overlap.

Let

\[
U_i = F_n(\beta'X_i), \quad \text{where} \quad F_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(Y_j \leq x).
\]

\[
\hat{\rho} = 1 - \frac{1}{2} \sum_{j=1}^{[n/k]} |C_j - k/n|, \quad \text{where} \quad C_j = \frac{1}{n} \sum_{i=1}^{n} I\left(\frac{(j-1)k}{n} < U_i \leq \frac{jk}{n}\right).
\]
**Goodness-of-match test**

To test $H_0: \mathcal{L}(Y) = \mathcal{L}(\beta'X)$, we define a test statistic

$$T_n = \sqrt{n} \sum_{j=1}^{[n/k]} |C_j - k/n|.$$ 

$T_n$ is distribution-free under $H_0$, we list the critical values obtained from simulation with $n = 1000$

<table>
<thead>
<tr>
<th>Significance level</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k/n = 0.1$</td>
<td>4.49</td>
<td>4.85</td>
<td>5.16</td>
<td>5.52</td>
<td>5.79</td>
</tr>
<tr>
<td>$k/n = 0.05$</td>
<td>5.98</td>
<td>6.36</td>
<td>6.67</td>
<td>6.99</td>
<td>7.24</td>
</tr>
<tr>
<td>$k/n = 0.025$</td>
<td>8.13</td>
<td>8.44</td>
<td>8.76</td>
<td>9.08</td>
<td>9.33</td>
</tr>
</tbody>
</table>
Proposition 1. Let \( \{\xi_1, \cdots, \xi_n\} \) and \( \{\eta_1, \cdots, \eta_n\} \) be two independent random samples from two distributions \( F \) and \( G \), and \( F \) be a continuous distribution. Let \( F_n(x) = \frac{1}{n} \sum_i I(\xi_i \leq x) \) and \( U_i = F_n(\eta_i) \). Let \( C_j \) and \( T_n \) are defined as above. Then the distribution \( T_n \) is independent of \( F \) and \( G \) provided \( F(\cdot) \equiv G(\cdot) \).

Proof. (i) \( U_i = \frac{1}{n} \sum_{j=1}^{n} I\{F(\xi_j) \leq F(\eta_i)\} \) almost surely, and (ii) \( \{F(\xi_i)\} \) and \( \{F(\eta_i)\} \) are two independent samples from \( U[0, 1] \) when \( F(\cdot) \equiv G(\cdot) \).
Simulation

Setting: \( Y_j = \beta'X_j + Z_j = \beta_1 X_{j1} + \cdots + \beta_p X_{jp} + Z_j, \quad j = 1, \ldots, n \)

Use OLS \( \tilde{\beta} \) as initial value

Maximal 500 iterations, and terminate and let MQE \( \hat{\beta} = \beta^{(k)} \) if

\[
\left| \left\{ R_k(\beta^{(k)}) \right\}^{1/2} - \left\{ R_{k-1}(\beta^{(k-1)}) \right\}^{1/2} \right| < 0.001
\]

Estimation errors: \( \text{rMSE}(\hat{\beta}) = \{ R_k(\beta^{(k)}) \}^{1/2} \)

Sample size \( n = 300 \) or 800, Dimension \( p = 50, 100 \) or 200

"Noise-to-signal" ratio: \( r \equiv \frac{\text{STD}(Z_j)}{\text{STD}(\beta_1 X_{j1} + \cdots + \beta_p X_{jp})} = 0.5, 1 \) or 2.

No. of replications: 1000

Post-sample (of size 300) predictive errors:

\[
\text{rMPE}(\beta) = \left( \frac{1}{300} \sum_{j=1}^{300} \left\{ y_{(j)} - (\beta'x)_{(j)} \right\}^2 \right)^{1/2}
\]
\[ X_j = (X_{j1}, \cdots, X_{jp})' = AU_j + \varepsilon_j, \quad \text{where} \]

all components of \( \varepsilon_j \) are independent \( t_4 \),

\( A \) is a \( p \times 3 \) constant factor loading matrix,

\( U_j \) consists of 3 independent AR(1) with \( \pm \) centered log-\( N(0, 1) \) innovations.

\( Z_j \) are independent \( N(0, \sigma^2) \) with \( \sigma = r \text{STD}(\beta'X_j) \).

For each sample,

\( \beta_j \) are drawn independently from \( U[-0.5, 0.5] \),

the elements of \( A \) are drawn independently from \( U[-1, 1] \),

AR coefficients for 3 factors are drawn independently from \( U[-0.95, 0.95] \).
\[ \mathcal{L}(\beta'X) \neq \mathcal{L}(Y) \]

As \( n \uparrow \), rMSE ↓

As \( r \uparrow \), rMSE ↑

As \( p \uparrow \), rMSE ↓

– p.22
Means and standard deviations of the No. of iterations required for computing $\hat{\beta}$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.5</td>
<td>22.2</td>
<td>18.1</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>27.1</td>
<td>20.7</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>31.3</td>
<td>22.2</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>6.0</td>
<td>4.3</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.1</td>
<td>4.8</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.0</td>
<td>5.0</td>
<td>2.3</td>
</tr>
<tr>
<td>800</td>
<td>Mean</td>
<td>30.4</td>
<td>31.5</td>
<td>25.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.3</td>
<td>38.0</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>53.0</td>
<td>44.6</td>
<td>31.7</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>8.5</td>
<td>6.9</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.5</td>
<td>8.0</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.9</td>
<td>9.9</td>
<td>5.5</td>
</tr>
</tbody>
</table>

More iterations are required when

- $r$ increases, or
- $n$ increases, or
- $p$ decreases.
Scatter plots of $\text{rMPE}(\hat{\beta})$ against $\text{rMPE}(\tilde{\beta})$: $n = 800$
Scatter plots of $r\text{MPE}(\hat{\beta})$ against $r\text{MPE}(\tilde{\beta})$: $n = 300$

$p=50, r=0.5$

$p=50, r=1$

$p=50, r=2$

$p=100, r=0.5$

$p=100, r=1$

$p=100, r=2$

$p=200, r=0.5$

$p=200, r=1$

$p=200, r=2$
Mean and (STD) of the estimated goodness-of-match measures

<table>
<thead>
<tr>
<th>$p$</th>
<th>$r$</th>
<th>$n = 300$</th>
<th></th>
<th></th>
<th>$n = 800$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>OLS</strong></td>
<td><strong>MQE</strong></td>
<td></td>
<td><strong>OLS</strong></td>
<td><strong>MQE</strong></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>.89 (.02)</td>
<td>.95 (.01)</td>
<td>.89 (.02)</td>
<td>.88 (.01)</td>
<td>.92 (.01)</td>
<td>.89 (.02)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.85 (.03)</td>
<td>.95 (.01)</td>
<td>.89 (.02)</td>
<td>.83 (.02)</td>
<td>.84 (.03)</td>
<td>.92 (.01)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.76 (.04)</td>
<td>.95 (.01)</td>
<td>.88 (.02)</td>
<td>.71 (.03)</td>
<td>.72 (.04)</td>
<td>.93 (.01)</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>.89 (.02)</td>
<td>.96 (.01)</td>
<td>.89 (.02)</td>
<td>.86 (.01) ; .87 (.02)</td>
<td>.96 (.01) ; .89 (.02)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.84 (.02)</td>
<td>.96 (.01)</td>
<td>.88 (.02)</td>
<td>.83 (.02)</td>
<td>.84 (.03)</td>
<td>.96 (.01)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.79 (.03)</td>
<td>.96 (.01)</td>
<td>.87 (.03)</td>
<td>.74 (.03)</td>
<td>.75 (.04)</td>
<td>.94 (.01)</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>.89 (.02)</td>
<td>.97 (.01)</td>
<td>.88 (.02)</td>
<td>.86 (.01)</td>
<td>.87 (.02)</td>
<td>.96 (.01)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.87 (.02)</td>
<td>.97 (.01)</td>
<td>.84 (.04)</td>
<td>.83 (.01)</td>
<td>.84 (.02)</td>
<td>.96 (.01)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.85 (.02)</td>
<td>.97 (.01)</td>
<td>.78 (.04)</td>
<td>.78 (.02)</td>
<td>.79 (.04)</td>
<td>.96 (.01)</td>
</tr>
</tbody>
</table>

Post-sample correlations are in **BLUE**.

With the MQE, $\hat{\rho} \geq 0.92$ for the in-sample matching, and $\hat{\rho} \geq 0.87$ for the post-sample matching (except the overfitting case when $n = 300$ and $p = 200$).

With the OLS, the minimum value of $\hat{\rho}$ is 0.71 for the in-sample matching, and is 0.72 for the post-sample matching.
Mean and (STD) of sample correlations between $Y$ and $\tilde{\beta}'X$, and between $Y$ and $\hat{\beta}'X$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$r$</th>
<th>$n = 300$ OLS</th>
<th>$n = 300$ MQE</th>
<th>$n = 800$ OLS</th>
<th>$n = 800$ MQE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.5</td>
<td>.95 (.02) .93 (.02) .95 (.02) .92 (.03)</td>
<td>.95 (.02) .94 (.02) .94 (.02) .93 (.02)</td>
<td>.84 (.04) .81 (.06) .81 (.04) .78 (.06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.86 (.04) .79 (.06) .84 (.04) .76 (.06)</td>
<td>.68 (.06) .56 (.09) .58 (.05) .50 (.08)</td>
<td>.63 (.06) .56 (.09) .58 (.05) .50 (.08)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.68 (.06) .51 (.10) .65 (.06) .47 (.10)</td>
<td>.63 (.06) .56 (.09) .58 (.05) .50 (.08)</td>
<td>.63 (.06) .56 (.09) .58 (.05) .50 (.08)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>.96 (.01) .92 (.03) .96 (.01) .91 (.03)</td>
<td>.95 (.02) .94 (.02) .95 (.02) .93 (.02)</td>
<td>.85 (.04) .80 (.06) .83 (.04) .77 (.06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.89 (.03) .74 (.07) .88 (.03) .72 (.08)</td>
<td>.66 (.06) .53 (.09) .63 (.05) .48 (.09)</td>
<td>.66 (.06) .53 (.09) .63 (.05) .48 (.09)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.75 (.05) .43 (.10) .74 (.05) .40 (.10)</td>
<td>.66 (.06) .53 (.09) .63 (.05) .48 (.09)</td>
<td>.66 (.06) .53 (.09) .63 (.05) .48 (.09)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>.98 (.01) .85 (.05) .98 (.01) .84 (.05)</td>
<td>.96 (.01) .92 (.03) .95 (.01) .92 (.03)</td>
<td>.87 (.03) .76 (.07) .86 (.03) .74 (.07)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.95 (.02) .60 (.10) .94 (.02) .59 (.10)</td>
<td>.72 (.04) .46 (.10) .71 (.04) .44 (.10)</td>
<td>.72 (.04) .46 (.10) .71 (.04) .44 (.10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.89 (.02) .28 (.10) .88 (.02) .28 (.10)</td>
<td>.72 (.04) .46 (.10) .71 (.04) .44 (.10)</td>
<td>.72 (.04) .46 (.10) .71 (.04) .44 (.10)</td>
<td></td>
</tr>
</tbody>
</table>

Post-sample correlations are in BLUE.

- Corr($Y, \hat{\beta}'X$) $\leq$ Corr($Y, \tilde{\beta}'X$), but the loss due to the disrespect of the pairing is not substantial.

- When $n \uparrow$, in-sample correlation decreases, post-sample correlations increases

- When $p \uparrow$, in-sample correlation increases, post-sample correlations decreases

- When $r \downarrow$, both correlations increase.

- Overfitting when $n = 300$ and $p = 200$, reflected by the large differences between in-sample and post-sample correlations.
A real data example

Select a representative portfolio for a total counterparty portfolio $Y$ from $p = 146$ trades.

The total data contains 1000 records of the total portfolio and the trades at tenor 1 month.

Data have been rescaled to protect the innocent.

We use the first $n = 700$ data points for estimation, the last 300 points to check the performance.

In-sample Corr($Y, \beta'X$): 0.566 (OLS), 0.558 (MQE)

Post-sample Corr($Y, \beta'X$): 0.248 (OLS), 0.230 (MQE)

No. of iterations for calculating the MQE: 7
Set $k/n = 0.05$.

In-sample $\hat{\rho}$: 0.741 (OLS), 0.905 (MQE)

Post-sample $\hat{\rho}$: 0.785 (OLS), 0.855 (MQE)

The goodness-of-match test statistic for the post sample:

$$T_n = 5.023$$

Reject the OLS matching at the 0.5% significance level.

Not reject the MQE matching even at the 10% level.
QQ-plots of the post-sample

The blue straight lines mark the diagonal $y = x$ on which the two quantiles are equal.
The post-sample quantiles of the total counterparty portfolio $Y$, and both the OLS representative portfolio $\tilde{\beta}'X$ and the MQE representative portfolio $\hat{\beta}'X$. The numbers in parentheses are the absolute estimation errors.

<table>
<thead>
<tr>
<th>Quantile level</th>
<th>Total portfolio</th>
<th>OLS representative portfolio</th>
<th>MQE representative portfolio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>-.548</td>
<td>-.308 (.240)</td>
<td>-.543 (.005)</td>
</tr>
<tr>
<td>5%</td>
<td>-.391</td>
<td>-.237 (.154)</td>
<td>-.409 (.018)</td>
</tr>
<tr>
<td>10%</td>
<td>-.332</td>
<td>-.186 (.146)</td>
<td>-.349 (.017)</td>
</tr>
<tr>
<td>30%</td>
<td>-.159</td>
<td>-.077 (.082)</td>
<td>-.140 (.019)</td>
</tr>
<tr>
<td>50%</td>
<td>.018</td>
<td>.016 (.002)</td>
<td>.017 (.001)</td>
</tr>
<tr>
<td>70%</td>
<td>.107</td>
<td>.094 (.013)</td>
<td>.152 (.045)</td>
</tr>
<tr>
<td>90%</td>
<td>.288</td>
<td>.200 (.088)</td>
<td>.354 (.066)</td>
</tr>
<tr>
<td>95%</td>
<td>.383</td>
<td>.254 (.129)</td>
<td>.427 (.044)</td>
</tr>
<tr>
<td>97.5%</td>
<td>.485</td>
<td>.312 (.173)</td>
<td>.575 (.090)</td>
</tr>
</tbody>
</table>
Tracking portfolios

\(Y\): return of a target to be tracked (e.g. S&P500, FTSE100)

\(X_1, \ldots, X_p\): returns of \(p\) securities used to track \(Y\).

**Tracking:**

\[
\min_{\{w_i\}} E \left( Y - \sum_{i=1}^{p} w_i X_i \right)^2
\]

subject to \(\sum_i w_i = 1\) and \(\sum_i |w_i| \leq c\).

The constant \(c > 1\) controls short sales, as

\[
\sum_{w_i > 0} w_i = \frac{(1 + c)}{2}, \quad \sum_{w_i < 0} |w_i| = \frac{(c - 1)}{2}.
\]
MQE for tracking: MQE-LASSO

Solving:
\[
\min_{\{\beta_j\}} \int_{\alpha_1}^{\alpha_2} \left\{ Q_Y(\alpha) - Q_{\beta'X}(\alpha) \right\}^2 d\alpha
\]
subject to
\[
\sum_{j=1}^{p} |\beta_j| \leq \delta \sum_{i=1}^{p} |\hat{\beta}_i^{(0)}|,
\]
where \( \hat{\beta}^{(0)} = (\hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_p^{(0)})' \) is the unconstrained MQE, and \( \delta \in (0, 1) \) controls, indirectly, the total exposure to short-sales.

Let
\[
\hat{w}_i = \frac{\hat{\beta}_i}{\sum_{1\leq j\leq n}\hat{\beta}_j}, \quad i = 1, \ldots, p.
\]

Then \( \sum_i |\hat{w}_i| \leq c \) for any \( c \geq \delta \sum_i |\hat{\beta}_i^{(0)}|/| \sum_j \hat{\beta}_j| \).

Remark. For any \( c > 0 \), find the largest \( \delta > 0 \), satisfying this condition, from the whole Lasso solution path.
Tracking FTSE100 using 30 stocks

Using the data in 2004-2006 ($n = 758$) to estimate portfolios
Comparing the performances using the data in 2007 (253 trading days)

Methods included: OLS, MQE, OLS-LASSO, MQE-LASSO,
MQE-LASSO matching partial distribution

Data were downloaded from Yahoo!Finance

$NM = \text{the mean value of all the negative returns}$

Note. Financial markets in 2004-2007 were overall bullish!
# 30 actively traded stocks (included in FTSE100)

<table>
<thead>
<tr>
<th>Code</th>
<th>Company Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANTO</td>
<td>Antofagasta</td>
</tr>
<tr>
<td>ARM</td>
<td>ARM Holdings</td>
</tr>
<tr>
<td>BARC</td>
<td>Barclays</td>
</tr>
<tr>
<td>BATS</td>
<td>British American Tobacco</td>
</tr>
<tr>
<td>BG</td>
<td>BG Group</td>
</tr>
<tr>
<td>BLT</td>
<td>BHP Billiton</td>
</tr>
<tr>
<td>BP</td>
<td>BP</td>
</tr>
<tr>
<td>BSY</td>
<td>British Sky Broadcasting</td>
</tr>
<tr>
<td>BT-A</td>
<td>BT Group</td>
</tr>
<tr>
<td>CNA</td>
<td>Centrica</td>
</tr>
<tr>
<td>OML</td>
<td>Old Mutual</td>
</tr>
<tr>
<td>RBS</td>
<td>Royal Bank of Scotland Group</td>
</tr>
<tr>
<td>RIO</td>
<td>Rio Tinto</td>
</tr>
<tr>
<td>RSA</td>
<td>RSA Insurance Group</td>
</tr>
<tr>
<td>ULVR</td>
<td>Unilever</td>
</tr>
<tr>
<td>CRDA</td>
<td>Croda International</td>
</tr>
<tr>
<td>DGE</td>
<td>Diageo</td>
</tr>
<tr>
<td>GSK</td>
<td>GlaxoSmithKline</td>
</tr>
<tr>
<td>HSBA</td>
<td>HSBC Holdings</td>
</tr>
<tr>
<td>ITV</td>
<td>ITV</td>
</tr>
<tr>
<td>LGEN</td>
<td>Legal &amp; General Group</td>
</tr>
<tr>
<td>LLOY</td>
<td>Lloyds Banking Group</td>
</tr>
<tr>
<td>MKS</td>
<td>Marks &amp; Spencer Group</td>
</tr>
<tr>
<td>MRW</td>
<td>Morrison Supermarkets</td>
</tr>
<tr>
<td>NG</td>
<td>National Grid</td>
</tr>
<tr>
<td>PRU</td>
<td>Prudential</td>
</tr>
<tr>
<td>RDSB</td>
<td>Royal Dutch Shell</td>
</tr>
<tr>
<td>RR</td>
<td>Rolls-Royce Group</td>
</tr>
<tr>
<td>TSCO</td>
<td>Tesco</td>
</tr>
<tr>
<td>VOD</td>
<td>Vodafone Group</td>
</tr>
</tbody>
</table>
## Daily returns (in percentages) in 2007

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>No. of stocks</th>
<th>Return Mean</th>
<th>Return Max</th>
<th>Return Min</th>
<th>STD</th>
<th>NM</th>
<th>Short sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTSE100</td>
<td>100</td>
<td>0.014</td>
<td>3.444</td>
<td>-4.185</td>
<td>1.100</td>
<td>-0.889</td>
<td>0</td>
</tr>
<tr>
<td>OLS</td>
<td>30</td>
<td>0.014</td>
<td>3.532</td>
<td>-3.716</td>
<td>1.094</td>
<td>-0.851</td>
<td>0</td>
</tr>
<tr>
<td>MQE</td>
<td>30</td>
<td>0.013</td>
<td>3.552</td>
<td>-3.739</td>
<td>1.098</td>
<td>-0.869</td>
<td>0</td>
</tr>
<tr>
<td>OLS-L, δ = .7</td>
<td>23</td>
<td>0.021</td>
<td>3.943</td>
<td>-4.300</td>
<td>1.250</td>
<td>-0.965</td>
<td>0</td>
</tr>
<tr>
<td>MQE-L, δ = .7</td>
<td>21</td>
<td>0.049</td>
<td>4.062</td>
<td>-5.247</td>
<td>1.488</td>
<td>-1.150</td>
<td>0</td>
</tr>
<tr>
<td>LL, δ = .5</td>
<td>14</td>
<td>0.045</td>
<td>4.011</td>
<td>-4.963</td>
<td>1.415</td>
<td>-1.119</td>
<td>0</td>
</tr>
<tr>
<td>ML, δ = .5</td>
<td>10</td>
<td>0.119</td>
<td>5.05</td>
<td>-6.196</td>
<td>1.825</td>
<td>-1.336</td>
<td>0</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.7, 0, .5)</td>
<td>13</td>
<td>0.316</td>
<td>18.51</td>
<td>-8.015</td>
<td>2.805</td>
<td>-1.804</td>
<td>38.4</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.7, .25, .75)</td>
<td>11</td>
<td>-0.040</td>
<td>3.864</td>
<td>-4.715</td>
<td>1.567</td>
<td>-1.233</td>
<td>3.9</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.7, .5, 1)</td>
<td>15</td>
<td>1.608</td>
<td>52.77</td>
<td>-48.63</td>
<td>15.56</td>
<td>-11.93</td>
<td>885</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.5, 0, .5)</td>
<td>12</td>
<td>0.223</td>
<td>14.55</td>
<td>-6.936</td>
<td>2.330</td>
<td>-1.563</td>
<td>1.4</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.5, .25, .75)</td>
<td>15</td>
<td>0.077</td>
<td>7.858</td>
<td>-8.866</td>
<td>2.295</td>
<td>-1.743</td>
<td>0</td>
</tr>
<tr>
<td>ML, (δ, α₁, α₂) = (.5, .5, 1)</td>
<td>5</td>
<td>-0.036</td>
<td>4.375</td>
<td>-5.776</td>
<td>1.791</td>
<td>-1.119</td>
<td>22.0</td>
</tr>
</tbody>
</table>

**Note.** LL=OLS-LASSO, ML=MQE-LASSO
Daily returns in 2007

black cycles: FTSE100
mean=.014, NM=-.889

red cycles: MQE-lasso
$\delta = .7$
mean=.049, NM=-1.15

blue cycles: MQE-lasso
$(\delta, \alpha_1, \alpha_2) = (.5, 0, .5)$
mean=.223, NM=-1.563
A rolling window illustration

For each year in 2007 – 2013:

• use the data in previous three years for estimation to form the different portfolios

• calculate the mean daily returns and STD in the year

For LASSO estimation, use $\delta = 0.5$

The data for 2013 were only upto 10 September.
Mean annual returns

<table>
<thead>
<tr>
<th>Year</th>
<th>FTSE100</th>
<th>Lower-half</th>
<th>Middle-half</th>
<th>Upper-half</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard deviations of returns

<table>
<thead>
<tr>
<th>Year</th>
<th>FTSE100</th>
<th>Lower-half</th>
<th>Middle-half</th>
<th>Upper-half</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>