Empirical Dynamic Quantiles for Time Series

Daniel Peña (Joint work with R.S.Tsay and R. Zamar)
Universidad Carlos III de Madrid

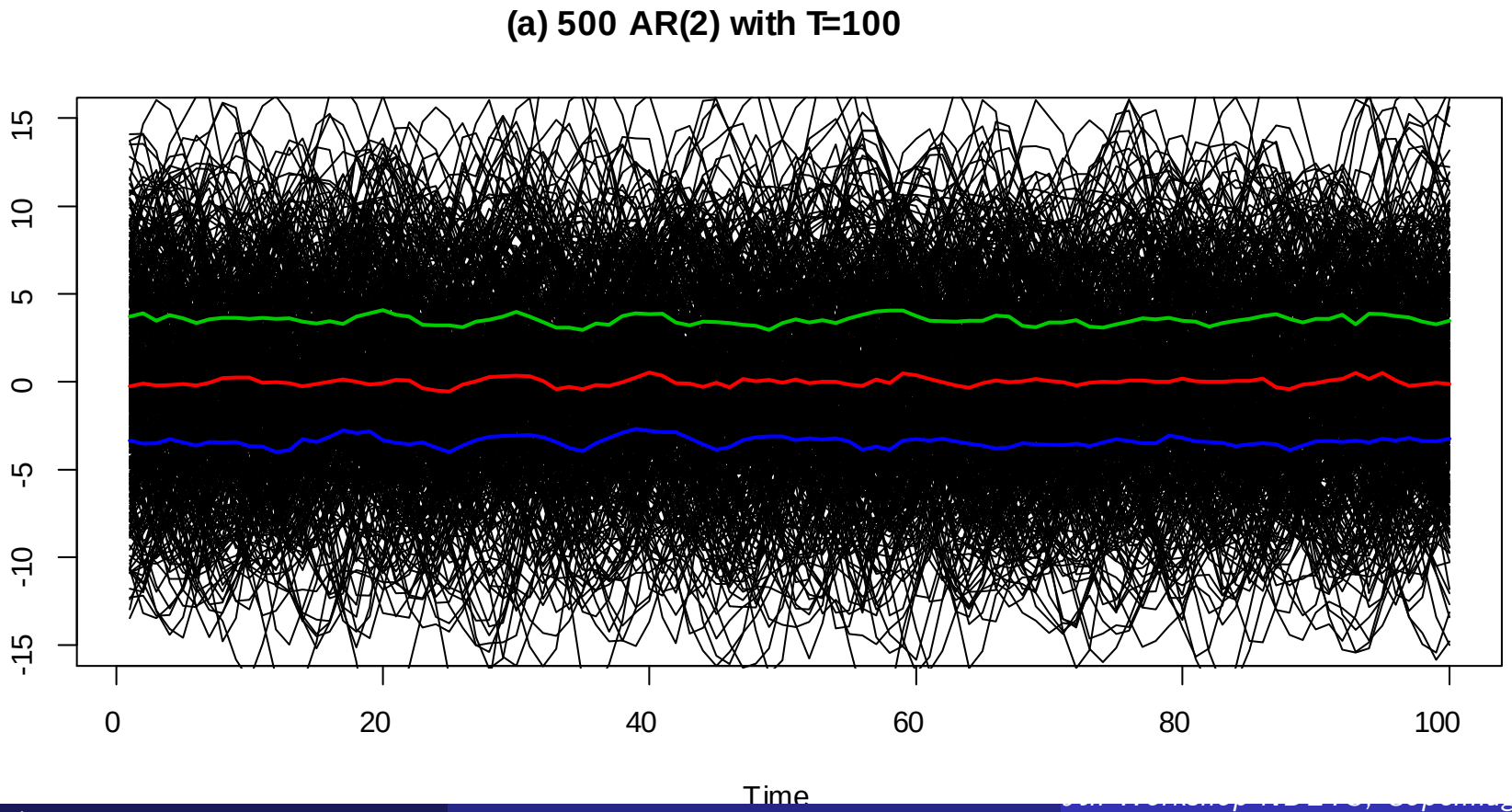
9th Workshop NDETS, Copenhagen 20-21 September

Outline

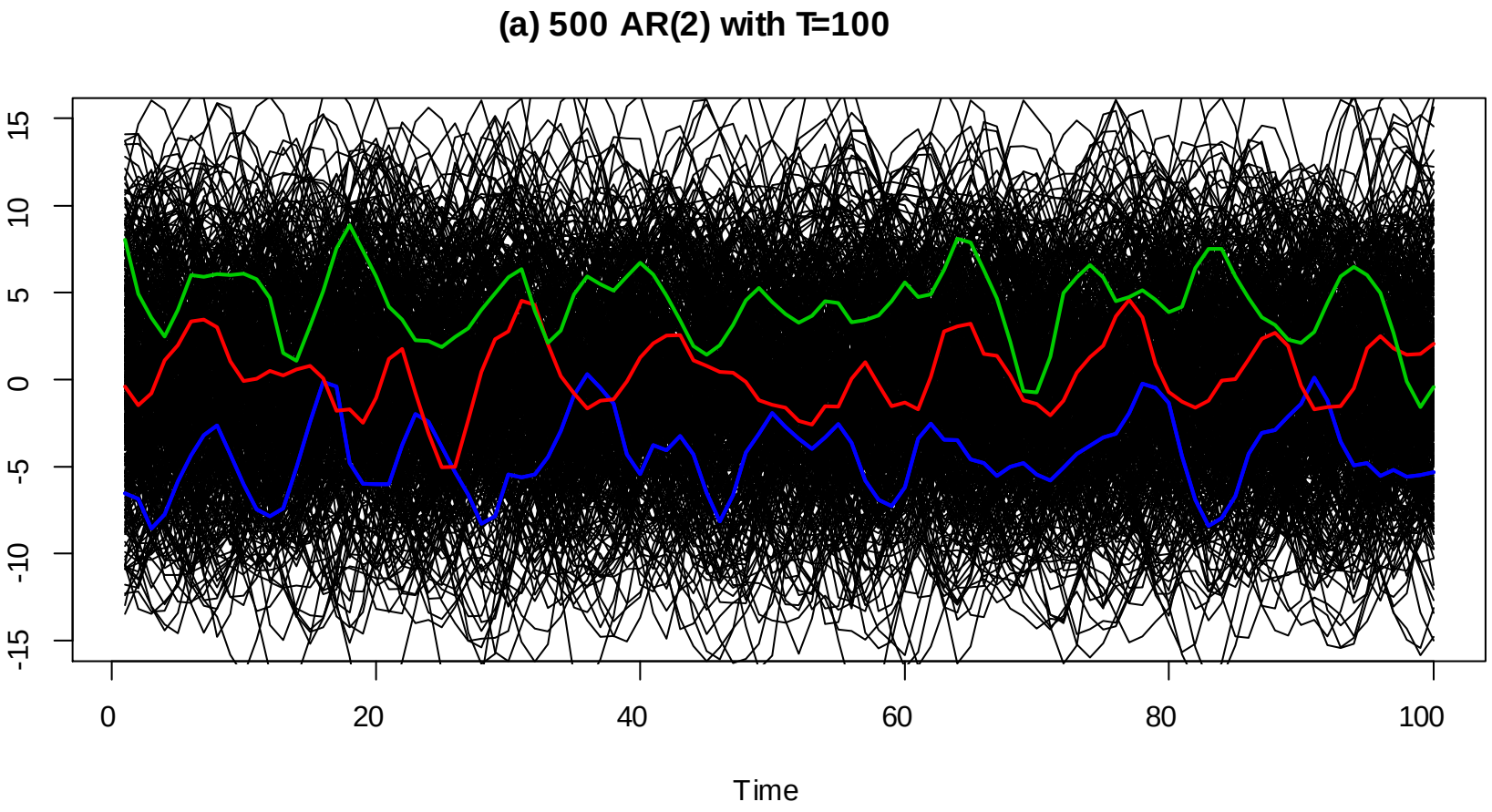
- Introduction: Quantiles in TS
- Empirical Dynamic Quantiles (EDQ)
- Computing EDQ
- Consistency of EDQ
- Two examples
- Extensions: Functional Dynamic quantiles
- Conclusions

Introduction

We want to use quantiles to visualize large sets of time series.
The standard definition of time series quantiles as timewise quantiles (TQ) is not usuful for this objective.
For instance, quantiles .25, .50 and .75 of a set of AR(2) processes



We present a definition of empirical time series quantiles, EDQ, that shows the dynamic of the series
EDQ for .25,.50 and .75 in the same data set.



We suppose a set of m time series, at times $1, .., T$.
These series can be represented as:
(a) T vectors $\mathbf{x}_t$,of dimension m, $\{\mathbf{x}_t | 1 \leq t \leq T, \mathbf{x}_t \in R^m\}$, a multivariate time series value.
(b) m vectors $\mathbf{x}_i$, with dimension T,$\{\mathbf{x}_i | 1 \leq i \leq m, \mathbf{x}_i \in R^T\}$, the univariate time series.
The p-th quantile at time t_0 is computed using the marginal distribution of the vector $\mathbf{x}_{t_0}$

Suppose first that the m series are realizations of the same univariate process. Calling $F_{t_0}(x)$ to the marginal distribution of the process at time t_0 the p-th quantile is defined as

$$Q_{t_0}^{(p)} = \inf_{x \in R} \{x | F_{t_0}(x) \geq p\}.$$

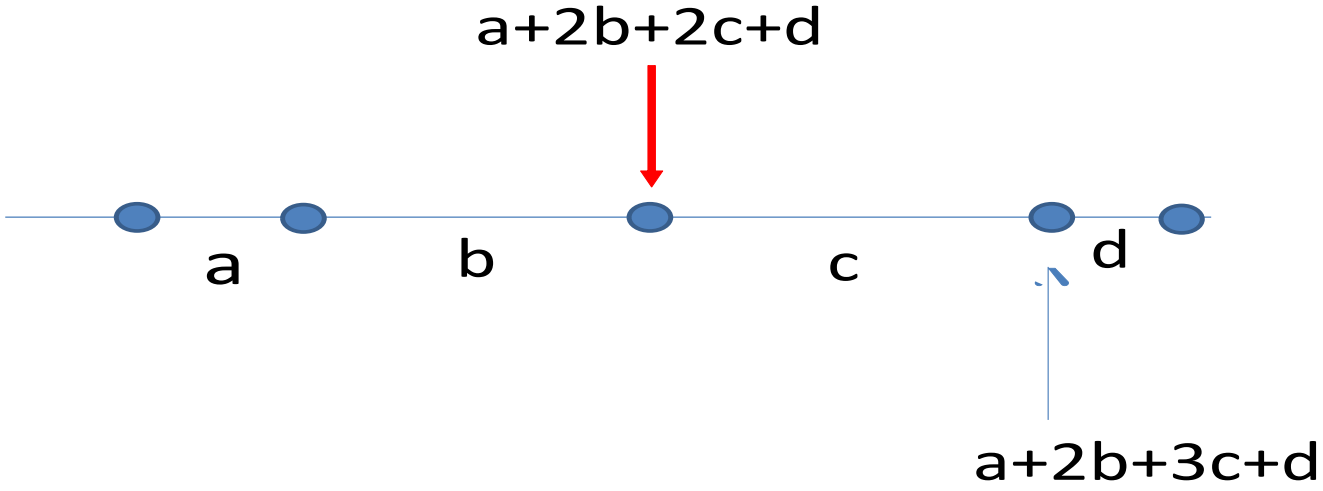
and the set of values $Q_t^{(p)}$ for given p and $1 \leq t \leq T$ are the timewise quantiles, TQ.
The estimation of these quantiles can be done by using the empirical marginal distribution $\widehat{F}_{t_0}(x)$,and the empirical p-th timewise quantile (ETQ) is defined by

$$q_t^{(p)} = \inf_{x \in R} \left\{ x | \widehat{F}_t(x) \geq p \right\}$$

ETQ can be computed by (1) order statistics; (2) solving a minimization problem.

For instance, the empirical timewise median, or ETQ(.50), is obtained at every time point by taking the observation in the middle of the ordered sample or by minimizing

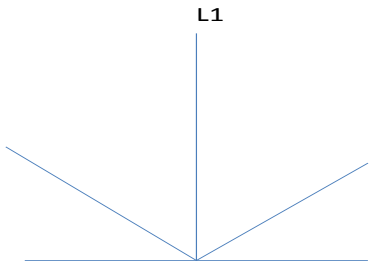
$$q_{t_0}^{(.5)} = \arg\min_{y \in R} \left[\sum |x_{it_0} - y| \right].$$

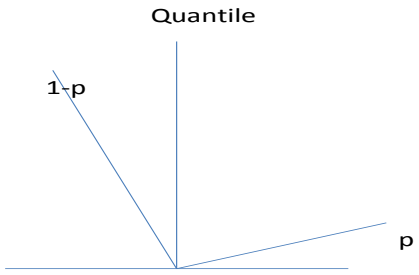


In general, the *pth* ETQ can be computed by ordering the sample and taking the corresponding order statistic or by minimizing

$$q_{t_0}^{(p)} = \arg\min_{y \in R} \left[p \sum_{x_{it_0} \geq y} |x_{it_0} - y| + (1-p) \sum_{x_{it_0} < y} |x_{it_0} - y| \right].$$

Note that for $p = .5$ we got the L_1 distance. For p small, most of the observations will be greater than the quantile and are given a small weight, and the few that are smaller than the quantile are given a large weight $(1-p)$





For time series:

- The ETQ (empirical timewise quantiles) only convey information about the marginal time behavior, but ignore the dynamics of the series.
- All the stationary processes with the same marginal distribution will have the same quantiles, it does not matter if they are AR(1) or AR(10) or MA processes
- For all Gaussian stationary TS the quantiles q_t^* will be roughly constant straight lines.
- For non stationary TS, the timewise quantiles describe the change over time of the marginal distributions.

Suppose now that the m series are independent realizations of a vector process of dimension $m_T \leq m$.Then, the m values x_{i,t_0} come from some univariate mixture distribution

$$F_{t_0}(x) = \sum_{j=1}^{m_T} \alpha_j F_{j,t_0}(x)$$

where $F_{j,t_0}(x)$ is the marginal distribution of the jth component of the mixture and $\sum_{j=1}^{m_T} \alpha_j = 1$. The p-th quantile at time t_0 is defined as

$$Q_{t_0}^{(p)} = \inf_{x \in R} \{x | F_{t_0}(x) \geq p\}.$$

Using the empirical marginal distribution $\widehat{F}_{t_0}(x)$,the empirical p-th quantile (ETQ) is defined by

$$q_{t_0}^{(p)} = \inf_{x \in R} \left\{x | \widehat{F}_{t_0}(x) \geq p\right\}$$

and the definition of ETQ can be extended to these cases.

Summary

We propose a new definition of empirical time series quantiles which conveys the dependence of the observations over time that characterizes the given data set.

We will call empirical dynamic quantile (EDQ) to the series in the set of observed data that minimizes the quantile distance function.

- Thus, we add in the minimization problem the restriction that the solution must be a member of the time series set.
- We prove that for independent realizations of a stationary process the EDQ, as ETQ, are consistent estimates for the population quantiles.

Definition of Empirical Timewise Quantiles (ETQ)

Notation: For a scalar a, and $0 \leq p \leq 1$, the *quantile function is*

$$\rho_p(a) = pa^+ + (1-p)\, a^-.$$

where $a^+ = aI(a \geq 0)$ and $a^- = -aI(a \leq 0)$. Thus, $\rho_p(a) = pa$ when $a \geq 0$ and $\rho_p(a) = (1-p)\,|a|$ for $a \leq 0$.
For a vector $\mathbf{a} = (a_1, .., a_n)'$,the *quantile function* is

$$\rho_p(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}(pa_i^+ + (1-p)\, a_i^-)$$

The ETQ for a given p is the sequence $\{q_t^*\}$ (We drop the p from now on) that minimizes

$$\{q_t^*\} = \arg\min_{\{y_t\}\in R}\left[\sum_{t=1}^{T}\rho_p(\mathbf{x}_t - \mathbf{1}y_t)\right].$$

where $\mathbf{1} = (1, ...,1)' \in R^m$.

Definition of EDQ

Given a set of time series $\mathbb{C}_m = \{x_{it}, 1 \leq i \leq m, 1 \leq t \leq T\}$, the empirical dynamic *pth* quantile (EDQ) is the series $\{q_t\}$ (again we drop the order of the quantile, p) in the set $\mathbb{C}_m$ that satisfies

$$\{q_t\} = \arg \min_{\{y_t\} \in \mathbb{C}_m} \left[\sum_{t=1}^{T} \left(\sum_{x_{it} \geq y_t} p \left| x_{it} - y_t \right| + \sum_{x_{it} \leq y_t} (1-p) \left| x_{it} - y_t \right| \right) \right]$$

or, using the quantile function

$$\{q_t\} = \arg \min_{\{y_t\} \in \mathbb{C}_m} \left[\sum_{t=1}^{T} \rho_p \left(\mathbf{x}_t - \mathbf{1} y_t \right) \right].$$

Remark: always y_t and therefore q_t is one of the observed series

Computing the EDQ

The computation of these quantiles requires the evaluation of the objective function

$$S_p(y_t) = \sum_{t=1}^{T} \rho_p \left(\mathbf{x}_t - \mathbf{1} y_t\right)$$

for each $y_t \in \mathbb{C}_m$. The series that leads to a smaller value of $S_p(q_t)$ is the empirical dynamic quantile (EDQ).
The direct computation of the EDQ requires $O(m^2)$ operations (distances between two series) plus $m \log(m)$ sorting operations. This is easy to do for moderate m but for large m, the direct estimation can be slow. We show next an algorithm that gives a good approximation to the dynamic quantiles with $O(m)$ computations and can be applied to very large sets of time series.

Some Intuition

- If a series is always the nearest to the timewise quantiles this is obviously the dynamic quantile.
- In general this series does not exist, and the series that is the nearest to the timewise quantile varies at every time.
- However, the dynamic quantile must be included in a small set of series that have L_1 distance smaller than C to the timewise quantile, but C is unknown.

Computing Algorithm for EDQ

The EDQ will be a member of a certain set of time series which are not far from the timewise quantile.
The timewise quantile q_t^* minimizes at each time t the distance $\rho_p\left(\mathbf{x}_t - \mathbf{1}y_t\right)$, so

$$\rho_p\left(\mathbf{x}_t - \mathbf{1}y_t\right) \geq \rho_p\left(\mathbf{x}_t - \mathbf{1}q_t^*\right)$$

and hence, we can write

$$\rho_p\left(\mathbf{x}_t - \mathbf{1}y_t\right) = \rho_p\left(\mathbf{x}_t - \mathbf{1}q_t^*\right) + c_t\left|y_t - q_t^*\right|$$

and the objective function we want to minimize verifies

$$S_p(y_t) = S_p(q_t^*) + \sum_{t=1}^{T} c_t\left|y_t - q_t^*\right|.$$

where the coefficients $c_t \geq 0$ are given by

$$c_t = \left(\rho_p\left(\mathbf{x}_t - \mathbf{1}y_t\right) - \rho_p\left(\mathbf{x}_t - \mathbf{1}q_t^*\right)\right) / \left|y_t - q_t^*\right|.$$

Algorithm

As the term $S_p(q_t^*)$ is constant for all the series in the set, the solution implies minimizing $\sum_{t=1}^{T} c_t \left| y_t - q_t^* \right|$

Step 1.(Finding h initial values) Set $c_t = 1$, for all $t = 1, ..., T$. The series in $\mathbb{C}_m$ can sorted in increasing value of

$$D(y_t) = \sum_{t=1}^{T} \left| y_t - q_t^* \right|$$

and denoted by $\{x_{k_1 t}\}$,..., $\{x_{k_h t}\}$ in $\mathbb{C}_m$, the h series nearest to q_t^* are those with minimum value of $D(y_t)$. Each of these h series is the starting value for the following iterative algorithm.

Algorithm

Step 2 (Iteration). For each series $x_{k_j t}$ from **Step 1,** we compute:
(a) the weights $c_{k_j t}$ for this series as follows: if $x_{k_j t} \neq q_t^*$

$$c_{k_j t} = \left(\rho_p\left(\mathbf{x}_t - \mathbf{1}x_{k_j t}\right) - \rho_p\left(\mathbf{x}_t - \mathbf{1}q_t^*\right)\right) / \left|x_{k_j t} - q_t^*\right|$$

otherwise $c_{k_j t} = 0$;
(b) the distances $D(y_t)$ of all the series to q_t^* with the new weights $c_t = c_{k_j t}$ and select the one that minimizes $S_p(y_t)$. If the new solution improves the objective function take this series as the new $x_{k_j t}$ and go to (a) and iterate. Otherwise, stop and output the index of the best series found in this **Step 2** and its value in the objective function.
Step 3 (Final selection). Select as EDQ the series with the smallest value of the objective function among the h candidates found in **Step 2.**

Monte Carlo Results

We show the percentage of times that our computing algorithm finds the overall constrained minimizer that defines the EDQ

- Quantile orders, p = (.01,.05,.25,.50,.75,.95,.99)
- TS Models(1) AR(1): $x_t = c + \phi x_{t-1} + a_t$,; (2) MA(1): $x_t = c + a_t - \theta a_{t-1}$.
- In all cases $c \sim N(0, 2)$ and $a_t \sim N(0, 1)$ and ϕ and θ are $U(-.9, .9)$.
- $m = 50, 200,\ T = 100, 200$
- The number of replicates is equal to 300.

Monte Carlo Results

Percentage of succes for an AR(1) with m=100 and T=100.

$h \backslash p$	.01	.05	.25	.50	.75	.95	.99
1	0.50	0.7	0.82	0.9	0.82	0.7	0.49
10	0.96	1.00	1.00	1.00	1.00	1.00	0.99
15	0.99	1.00	1.00	1.00	1.00	1.00	0.98
20	1.00	1.00	1.00	1.00	1.00	1.00	0.99

Percentage of succes for an MA(1), m=100 and T=100.

$h \backslash p$	.01	.05	.25	.50	.75	.95	.99
1	0.37	0.64	0.76	0.81	0.84	0.64	0.32
10	0.96	0.99	1.00	1.00	1.00	1.00	0.97
15	0.98	1.00	1.00	1.00	1.00	1.00	0.98
20	1.00	1.00	1.00	1.00	1.00	1.00	0.99

Percentage of succes for an AR(1), m=500 and T=200.

$h \backslash p$	.01	.05	.25	.50	.75	.95	.99
1	0.62	0.66	0.80	0.82	0.79	0.64	0.66
10	1.00	0.99	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Percentage of succes for an MA(1), m=500 and T=200.

$h \backslash p$	.01	.05	.25	.50	.75	.95	.99
1	0.56	0.63	0.74	0.82	0.74	0.68	0.58
10	1.00	0.98	1.00	1.00	1.00	0.98	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Consistency of the EDQ

It can be proved that:
(a) Given an ensemble of time series $\{\mathbf{x}_i\}_{i=1}^{m}$, with $\mathbf{x}_i \in R^T$, independent realizations of some stochastic process.
(b) Let Q_t be the theoretical *pth* quantile at time t and $\mathbf{Q} \in R^T$ the time series of such quantiles
(c) The empirical dynamic quantiles (EDQ) converge when m goes to infinity to $\mathbf{Q}$, the population quantiles.

Outline of the proof

Consider the average check distance of a series to the compac set $C_m = (\mathbf{x}_1, ..., \mathbf{x}_m)$,and call

$$R_m(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\rho_p(\mathbf{x}_i - \mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T}\rho_p\left(\mathbf{x}_{it} - \mathbf{1}y_t\right)$$

Then It can be proved that:
(a) $R_m(\mathbf{y})$ is Lipschitz,that is,

$$\left|R_m\left(\mathbf{y}_1\right) - R_m\left(\mathbf{y}_2\right)\right| \leq d_1\left(\mathbf{y}_1, \mathbf{y}_2\right) = \sum_{t=1}^{T}\left|y_{1t} - y_{2t}\right|$$

and (b) Given two compacts set $\mathcal{C}_m \subset R^T$ and $\mathcal{C}_m \subset \mathcal{C}_{m+1}$ call $\mathbf{q}_m$ the EDQ in $\mathcal{C}_m$ and $\mathbf{q}_{m+1}$in $\mathcal{C}_{m+1}$, then

$$R_{m+1}\left(\mathbf{q}_m\right) \geq R_{m+1}\left(\mathbf{q}_{m+1}\right).$$

Consider a stochastic process $\mathbf{x} = (x_1, ..., x_T)$ in $[1, T]$ and define $R(\mathbf{y}) = E\{\varrho_p(\mathbf{x}, \mathbf{y})\}$ For any compact set $\mathcal{D} \subset R^T$ we have

$$\lim_{m \to \infty} \sup_{\mathbf{y} \in \mathcal{D}} |R_m(\mathbf{y}) - R(\mathbf{y})| = 0.$$

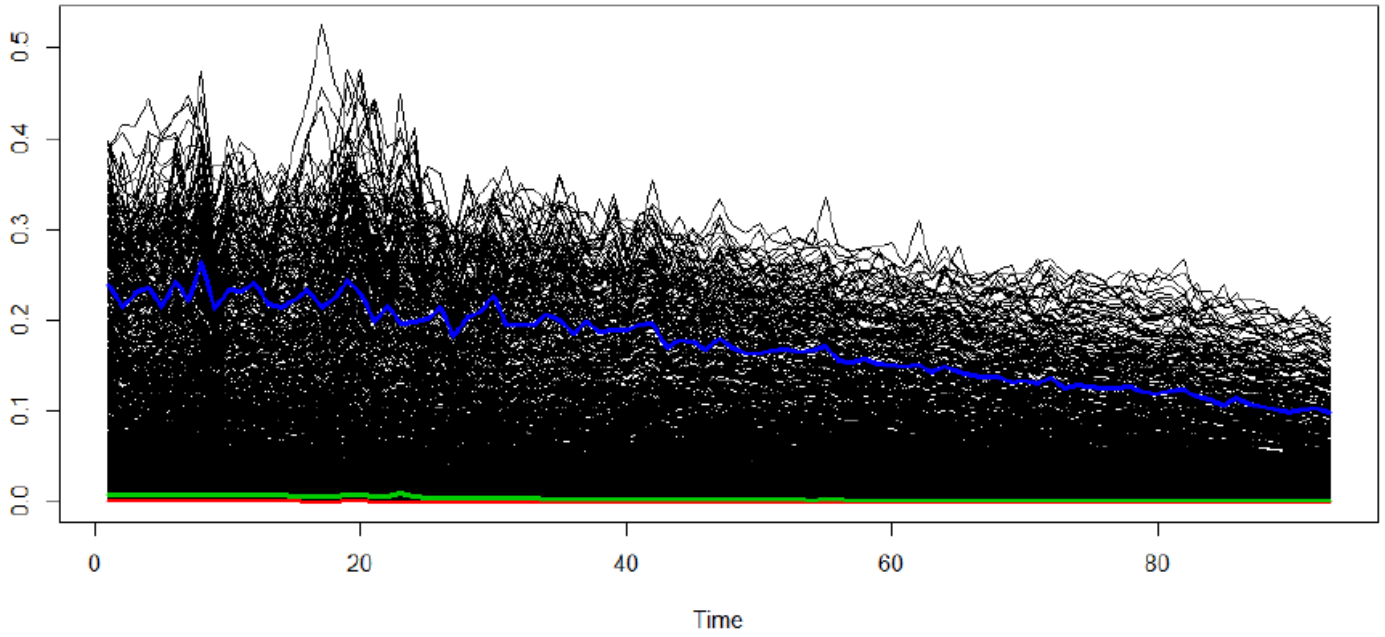
Let $\mathbf{q}_m$ be the constrained minimizer of $R_m(\mathbf{y})$ for $\mathcal{C}_m = \{\mathbf{x}_i\}_{i=1}^m$. Then, for all $\varepsilon > 0$,

$$\lim_{m \to \infty} P(d(\mathbf{q}_m, \mathbf{Q}) < \varepsilon) = 1.$$

Mortality rate

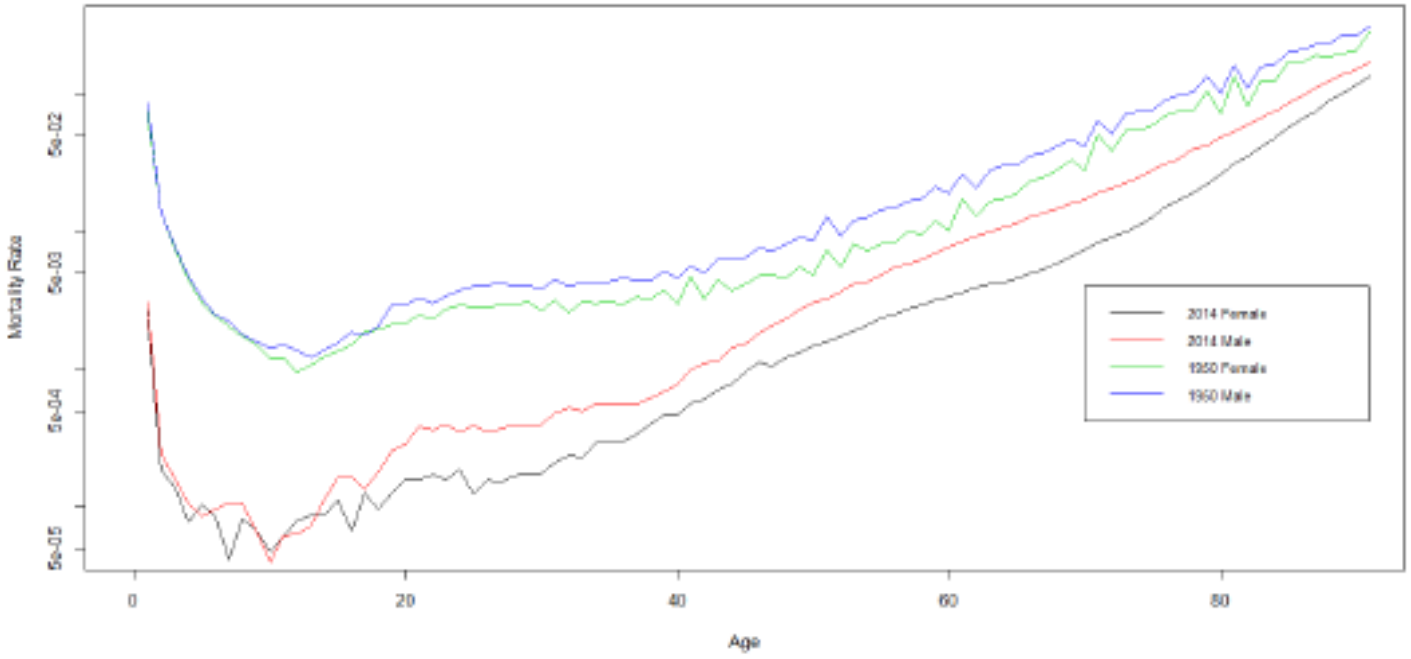
We consider mortality rates series. Each series shows the evolution of the mortality rate for a gender, age and country in the period: 1922 to 2014. The cases considered are female and male, ages from 0 to 90 years and ten countries. Hence, $m = 1820 = 2 \times 91 \times 10$ and
$T = 93(2014 - 1922 + 1)$.
The series are named by first giving the initial of the gender (F or M), followed by the year (from 00 to 90) and the country. For instance, the series F11DEN corresponds to the mortality rate for females of age 11 in Denmark. The countries are Belgium (BEL), Denmark (DEN), France, (FRA), Italy (ITA), Netherlands (NED), Norway (NOR), Spain (ESP), Sweden (SWE), Switzerland (SWI) and United Kingdom (UK).(Same geographic area in this period).

Figure: plot of the 1820 time series and the EDQ for $p = (0.05;\ 0.5; 0.95)$



The U shape of mortality rate

Cross section analysis of male and female mortality in a country in two years: 1960 and 2014



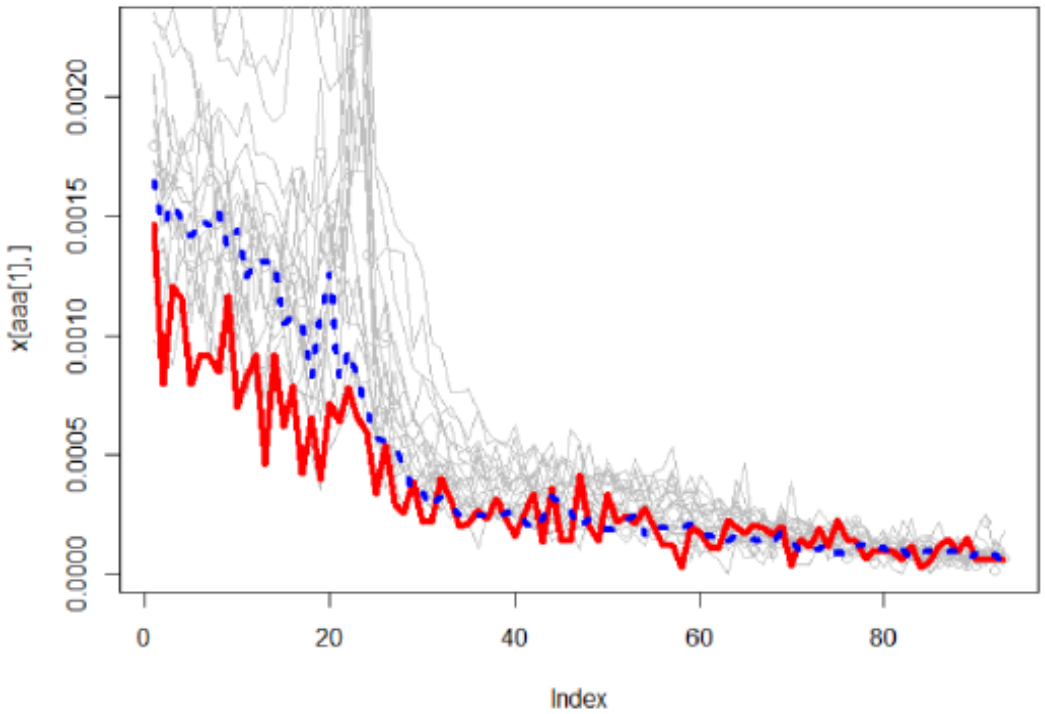
EDQ for the mortality rate in Europe

Table 5: EDQ por series of mortality rate in European countries

EDQ	Series	EDQ	Series
.001	F11DEN	.75	F70UK
.01	F11UK	.95	F87UK
.05	M10UK	.99	M88BEL
.25	F32UK	.999	M90FRA
.50	F47FRA		

Interpreting EDQ

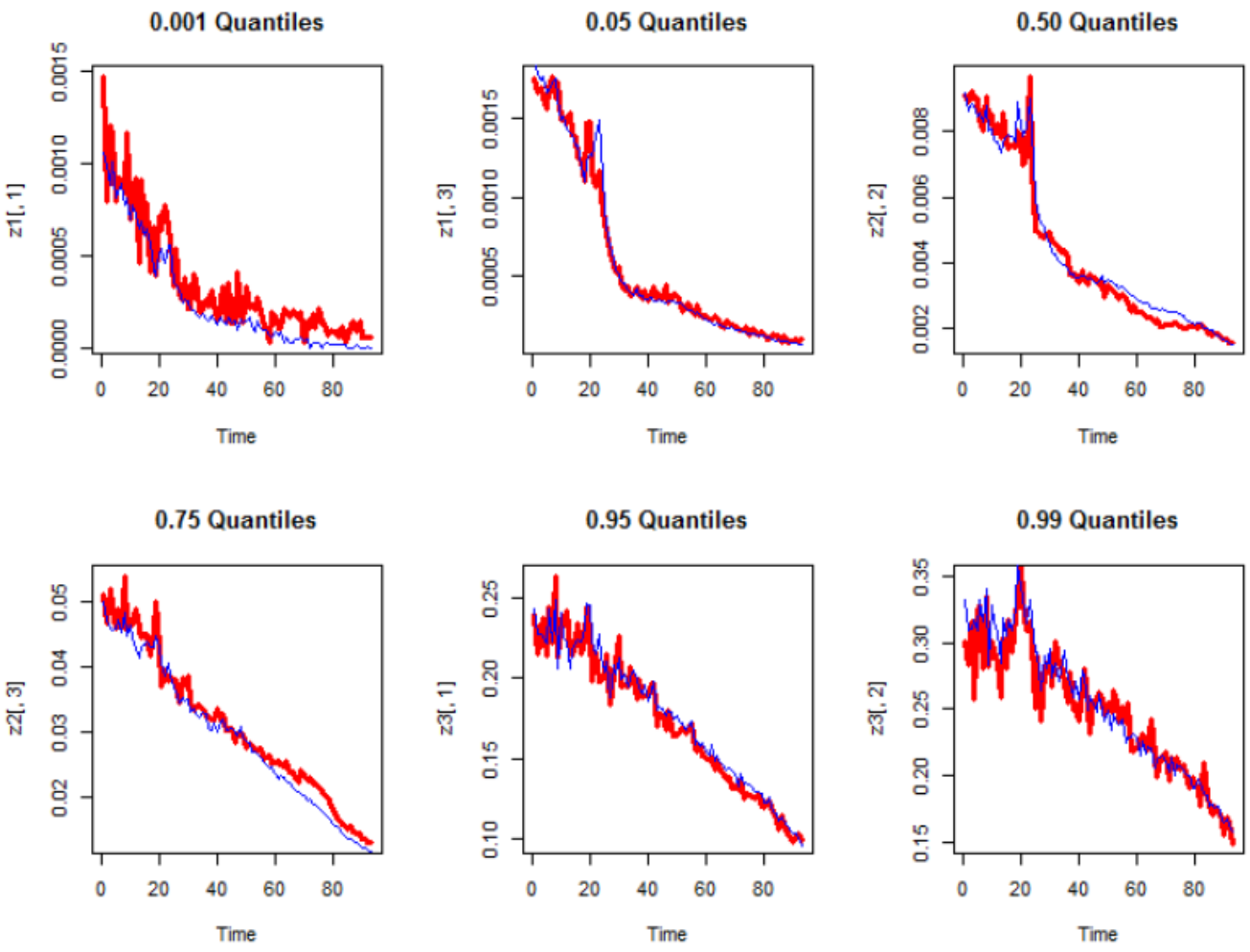
The EDQ have a clear interpretation. For instance, the series of 11 years old children are



The smallest mortality rate globally happens for females in Denmark, that is the EDQ(.001), with thick solid line red and the EDQ(0.01) corresponds to females children in the UK with blue dashed line.

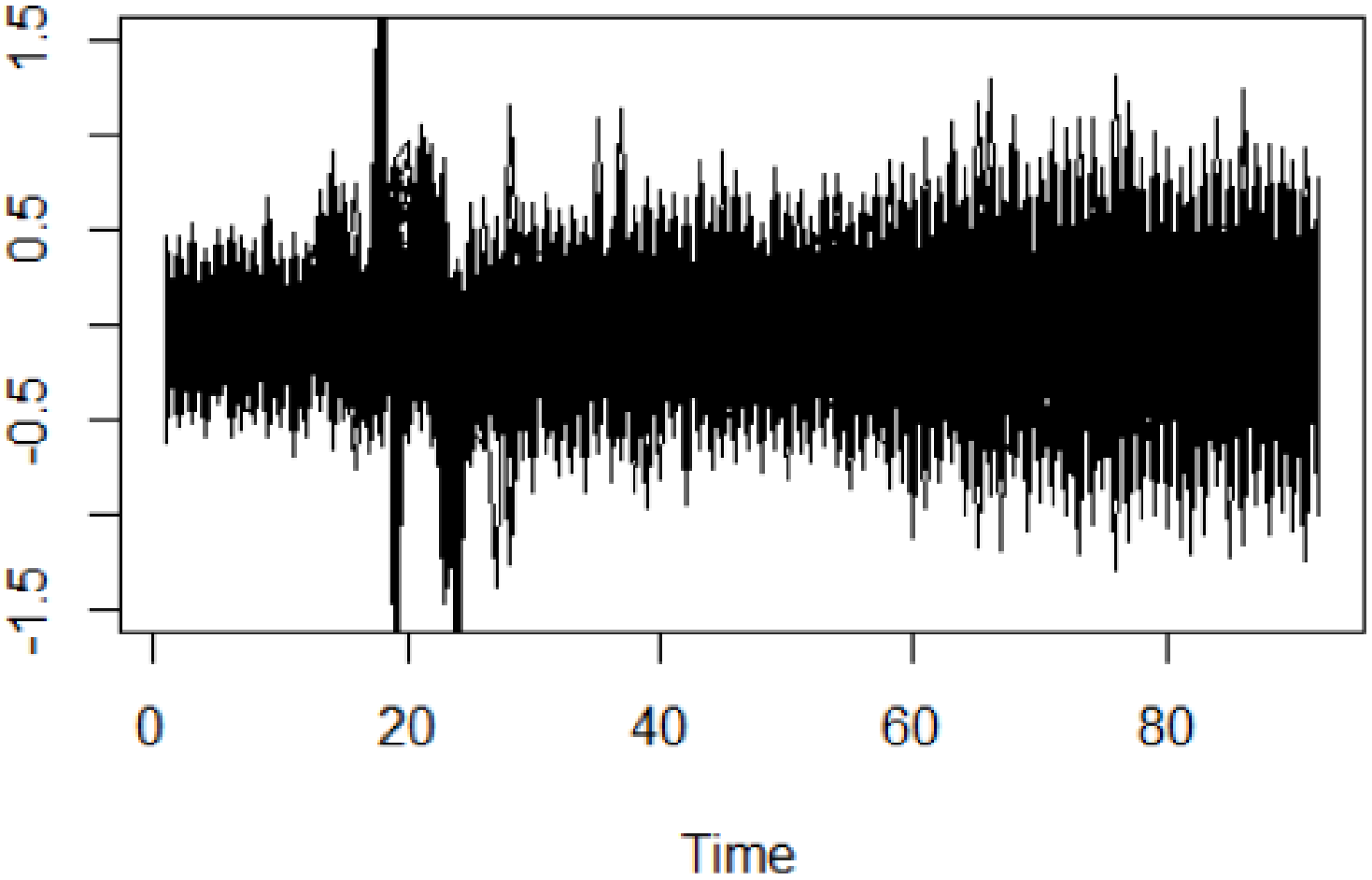
Comparing EDQ and ETQ

Comparing EDQ and ETQ: smoother the TQ and interpretation in DQ



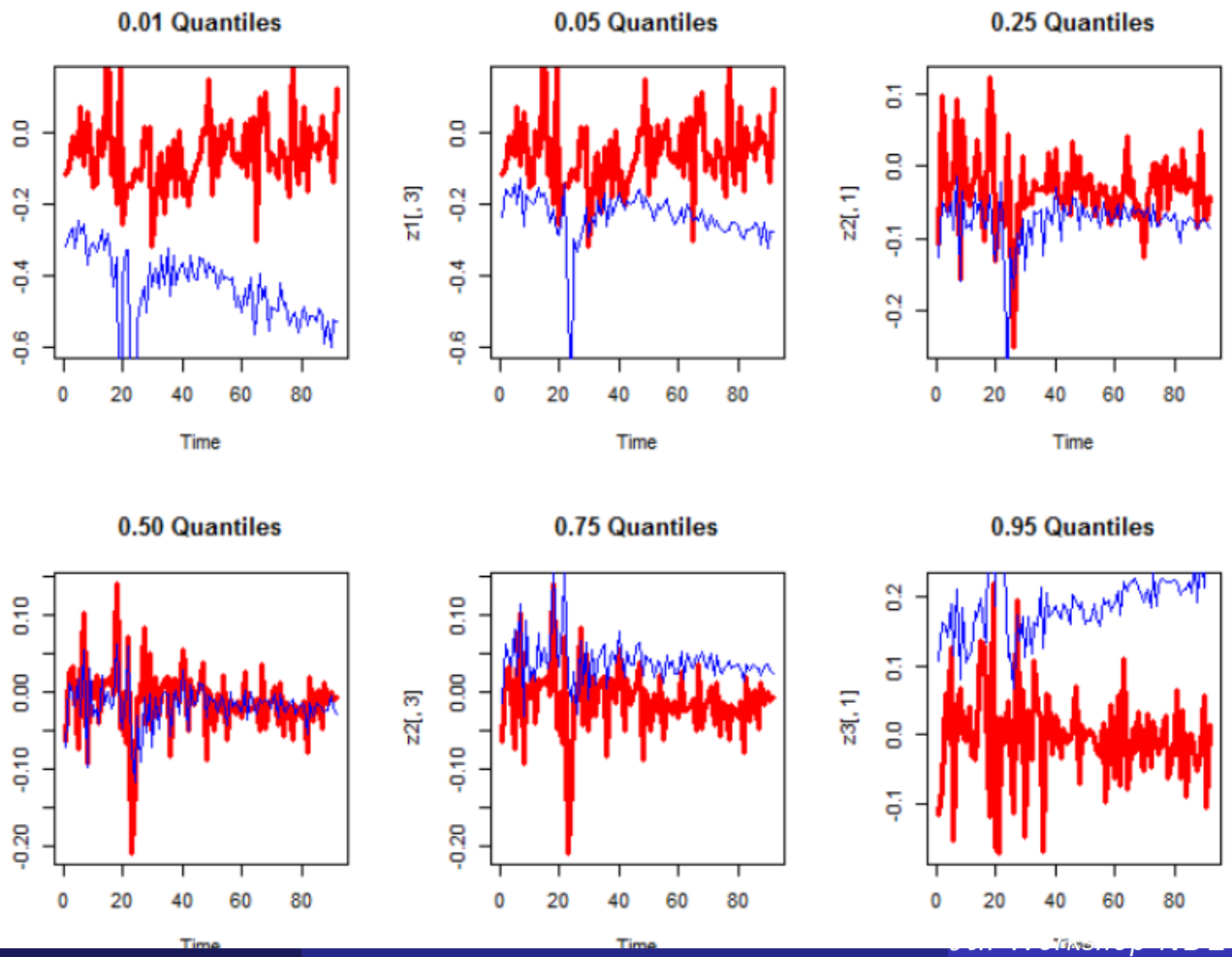
The rate of change in mortality rate

Series of relative change of mortality rate ($\nabla \log x_t$)



The ETQ and EDQ for the rate of change in mortality rate

Quantiles for the series $\nabla \log x$ (rate of change). The trends of ETQ (in blue lines) are not in any series and therefore they do not appear in the EDQ (red lines).



There is no trend in the individual time series of relative change in mortality

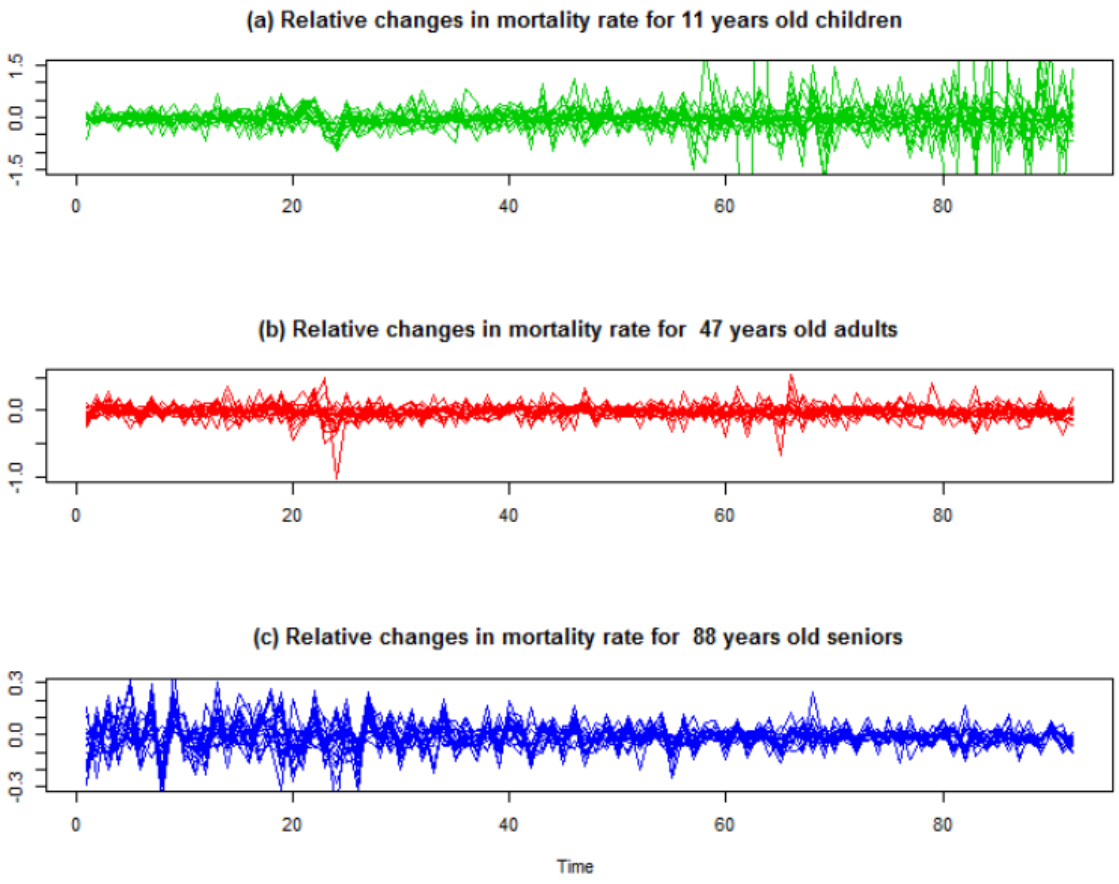
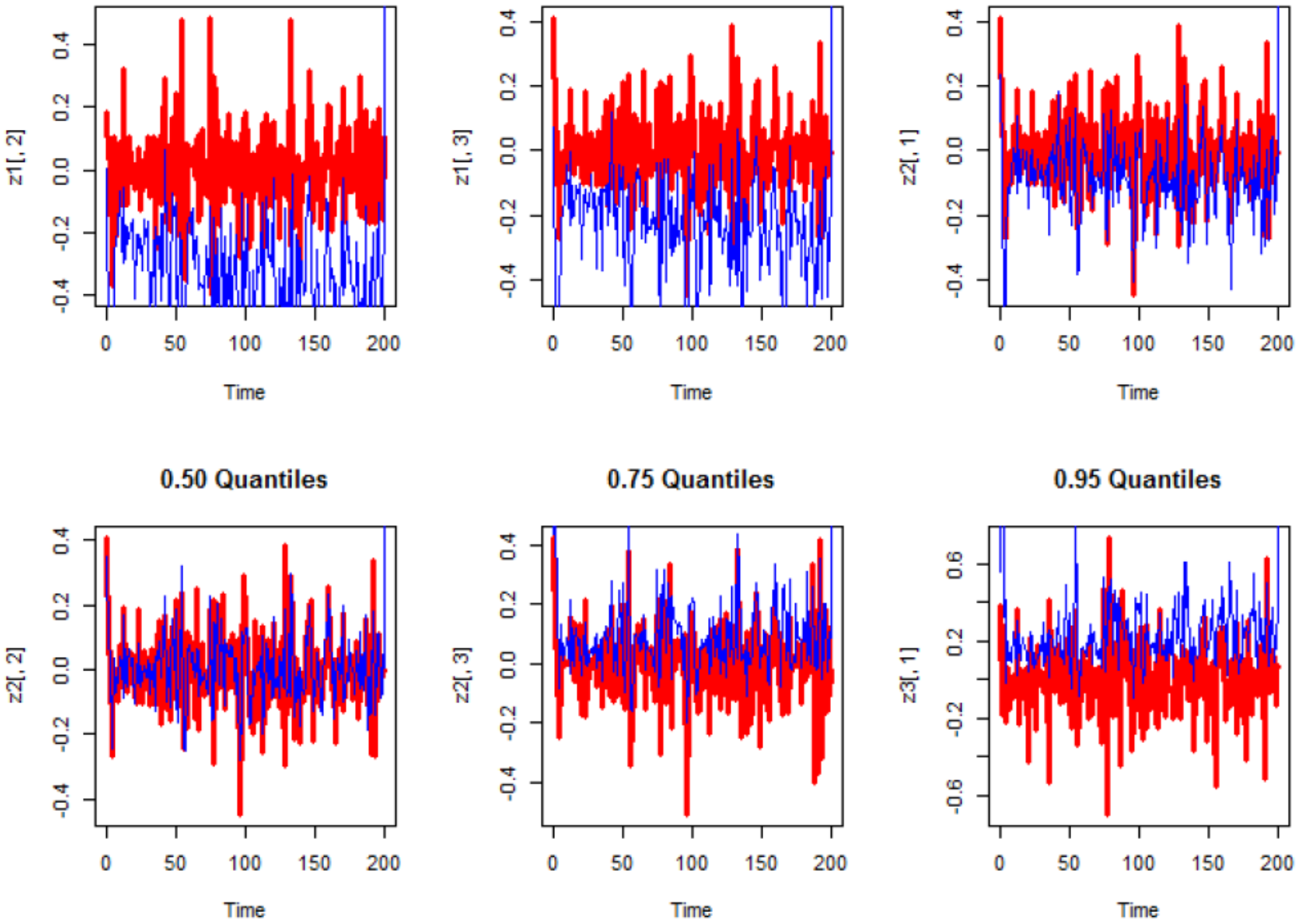


Figure 6: Mortality rate relative changes in Europe for three ages

This example shows that the TQ can be misleading in the interpretation of the behaviour of the series and that combining both type of quantiles you have a better representation of the data

Electricity Prices

Our second example corresponds to the hourly day-ahead electricity prices for eight zones in the New England electric market.
The data set is available at `www.iso-ne.com`.
We have 1334 $(24 \times 8 \times 7)$ time series of hourly price of electricity for 8 regions in each of the 7 days of the week.
In total we have $T = 200$ weeks, starting January 2004.
We analyze the series $\nabla \log P_t = \log P_t - \log P_{t-1}$, where P_t is the price of electricity at week t.


z1[, 2]
z1[, 3]
z2[, 1]
0.50 Quantiles
0.75 Quantiles
0.95 Quantiles
z2[, 2]
z2[, 3]
z3[, 1]
Time

As in the previous example the timewise quantiles show the marginal properties of the distribution of the price distribution in every hour and day of the week with different behavior of the dynamics of the series. Note that the differences between the quantiles appear only in the extreme order quantiles, especially for the lower order. However, both median series are fairly similar, with smaller variability for the ETQ.

Extensions

For m small we can relax the condition that the EDQ is one of the given series

We may define

$$\mathcal{C}_m = \{\mathbf{y} : \min |\mathbf{y} - \mathbf{x}_i| \leq \delta\}$$

or

$$\mathcal{C}_m = \left\{\mathbf{y} : \prod_{i=1}^{m} D\left(|\mathbf{y} - \mathbf{x}_i|\right) \leq \delta\right\}$$

for some small $\delta > 0$, where $D\left(a\right)$ is a distance function.

More general is to solve the following optimization problem

$$\{u_t\} = \arg \min_{\{y_t\} \in \mathbb{C}} \left[\sum_{t=1}^{T} \rho_p \left(\mathbf{x}_t - \mathbf{1} y_t \right) + \lambda \sum_{t=1}^{T} \left| y_t - q_t^* \right| \right]$$

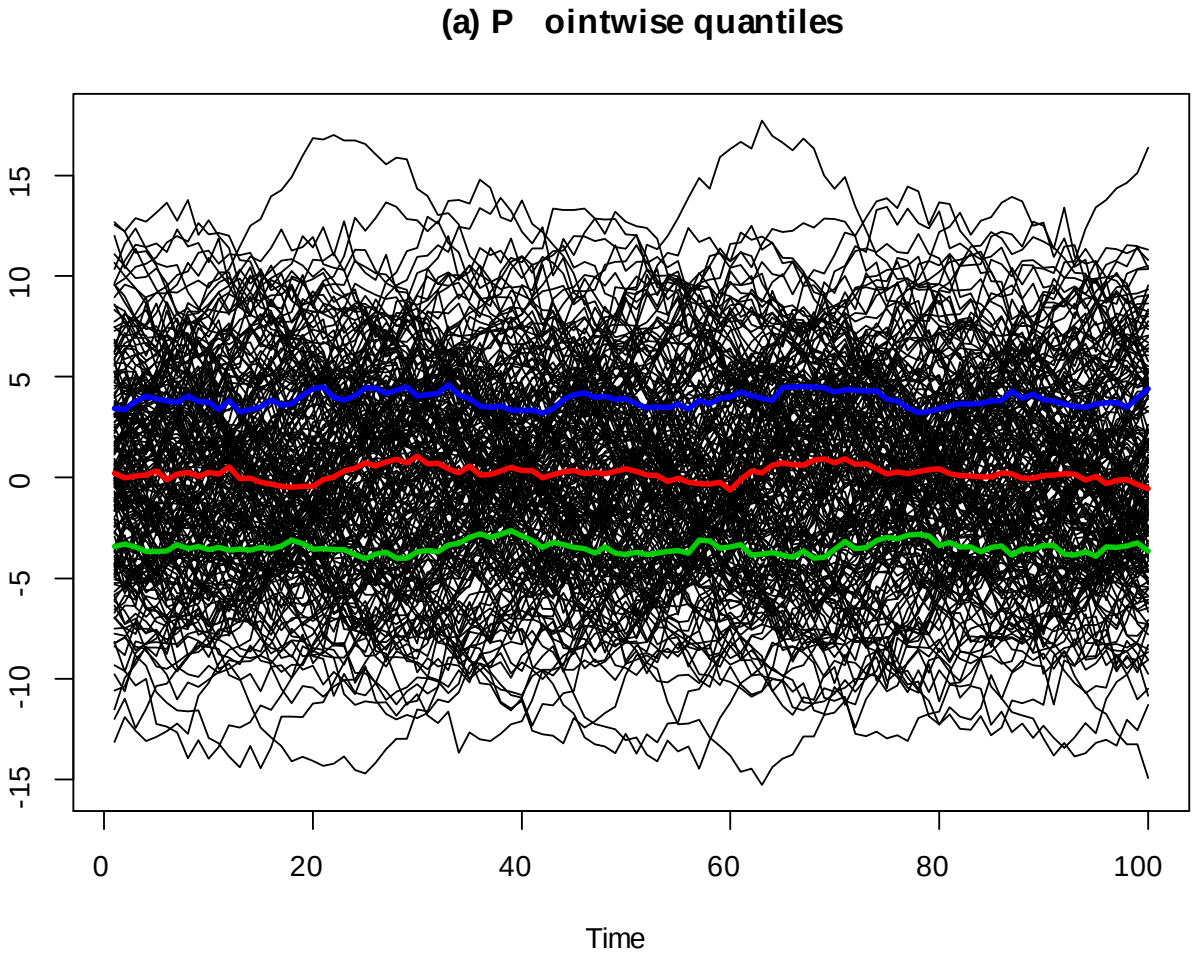
the solution will be a compromise between the EDQ, q_t, and the ETQ, q_t^*. An approximate solution of this problem is to define the quantiles as

$$\widetilde{q}_t = \alpha q_t + (1 - \alpha) q_t^*$$

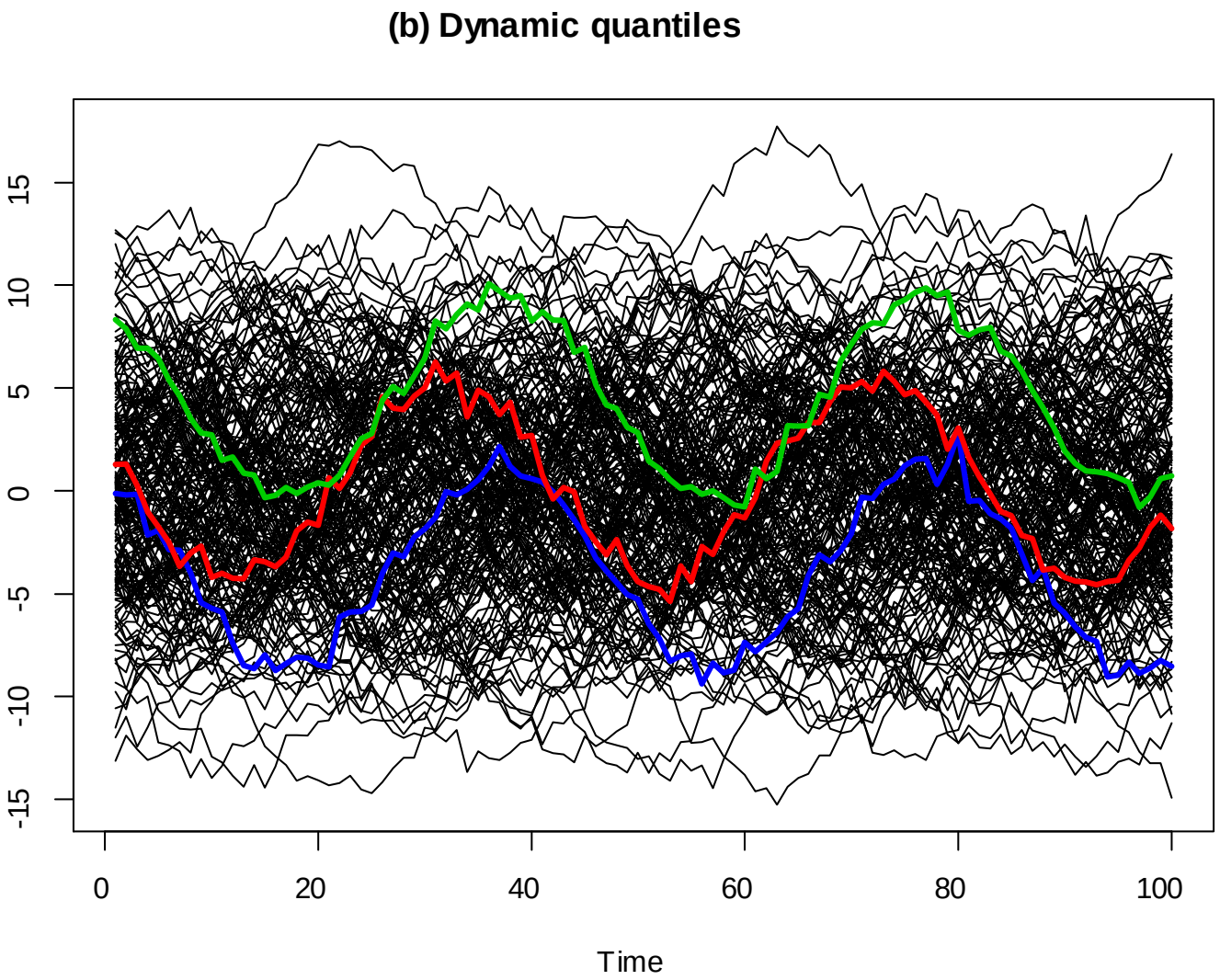
where the trade-off is controlled by $0 \leq \alpha \leq 1$.

Extension to functional data (with A. Arribas and J. Romo)

The same idea can be applied to functional data: pointwise quantiles do not show the structure of the functions.



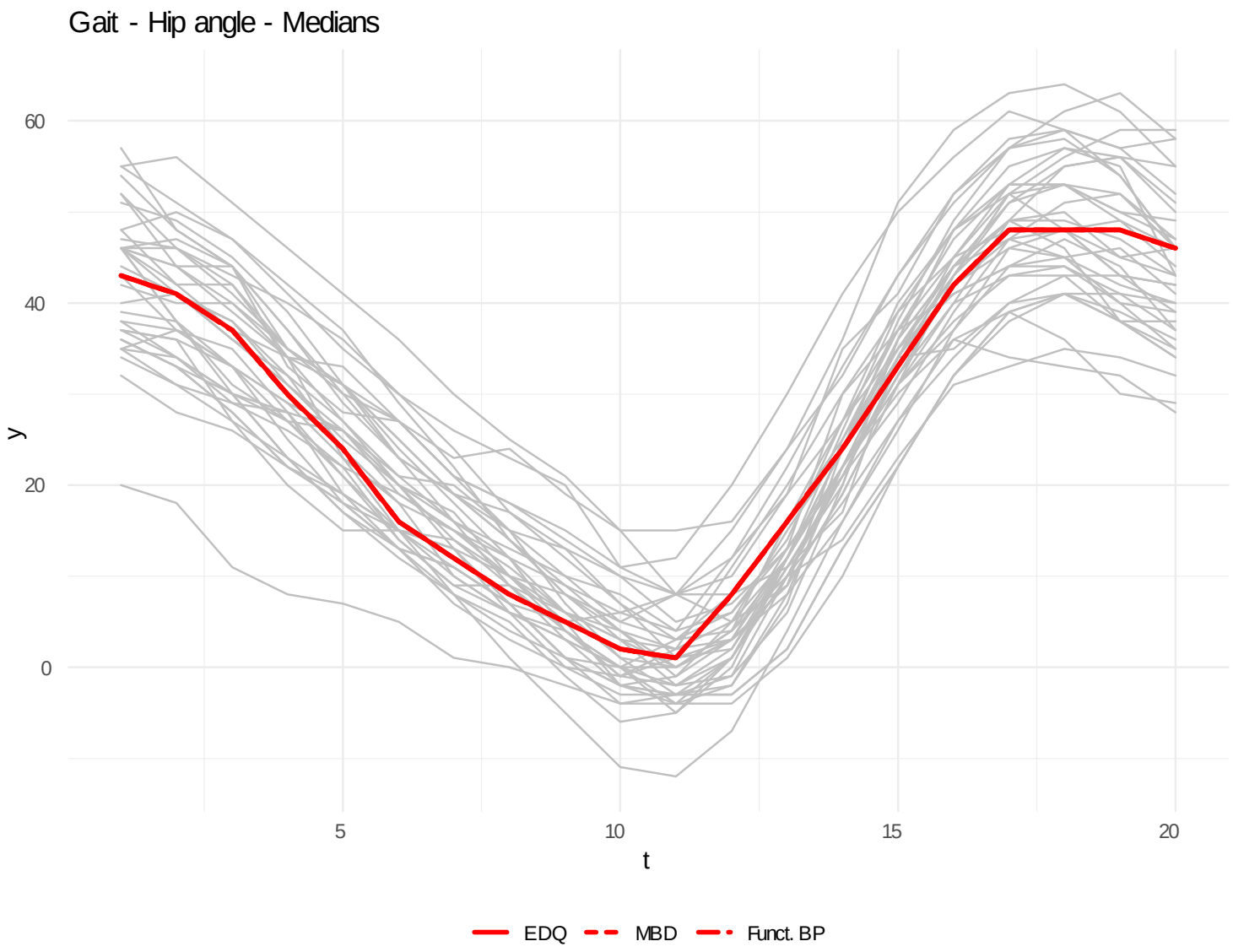
For the same data Dynamic functional quantiles



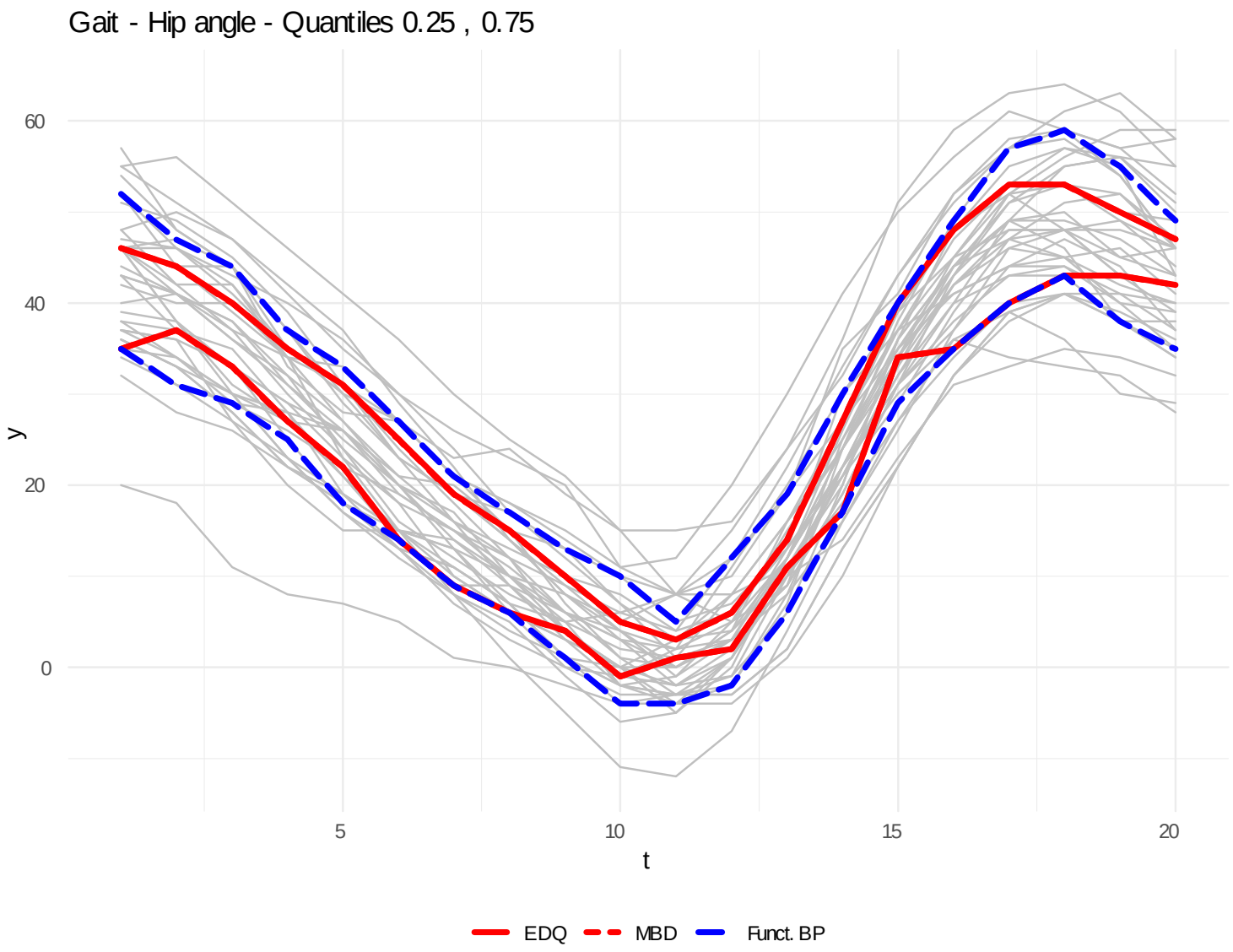
Data: $x_i(t) = \mu_i + \sin(wt + \alpha_i) + a_t,\ m = 200$

Depth measures can be used to build quantiles.
As an illustration we compare
(1) The functional quantiles computed as in Sun and Genton (2011) using the bands of the modified band depth measure proposed by Lopez Pintado and Romo (2009).
(2) The series that is the best L1 approximation to the previous functional quantiles.
(3) The dynamic functional quantiles, EDQ.

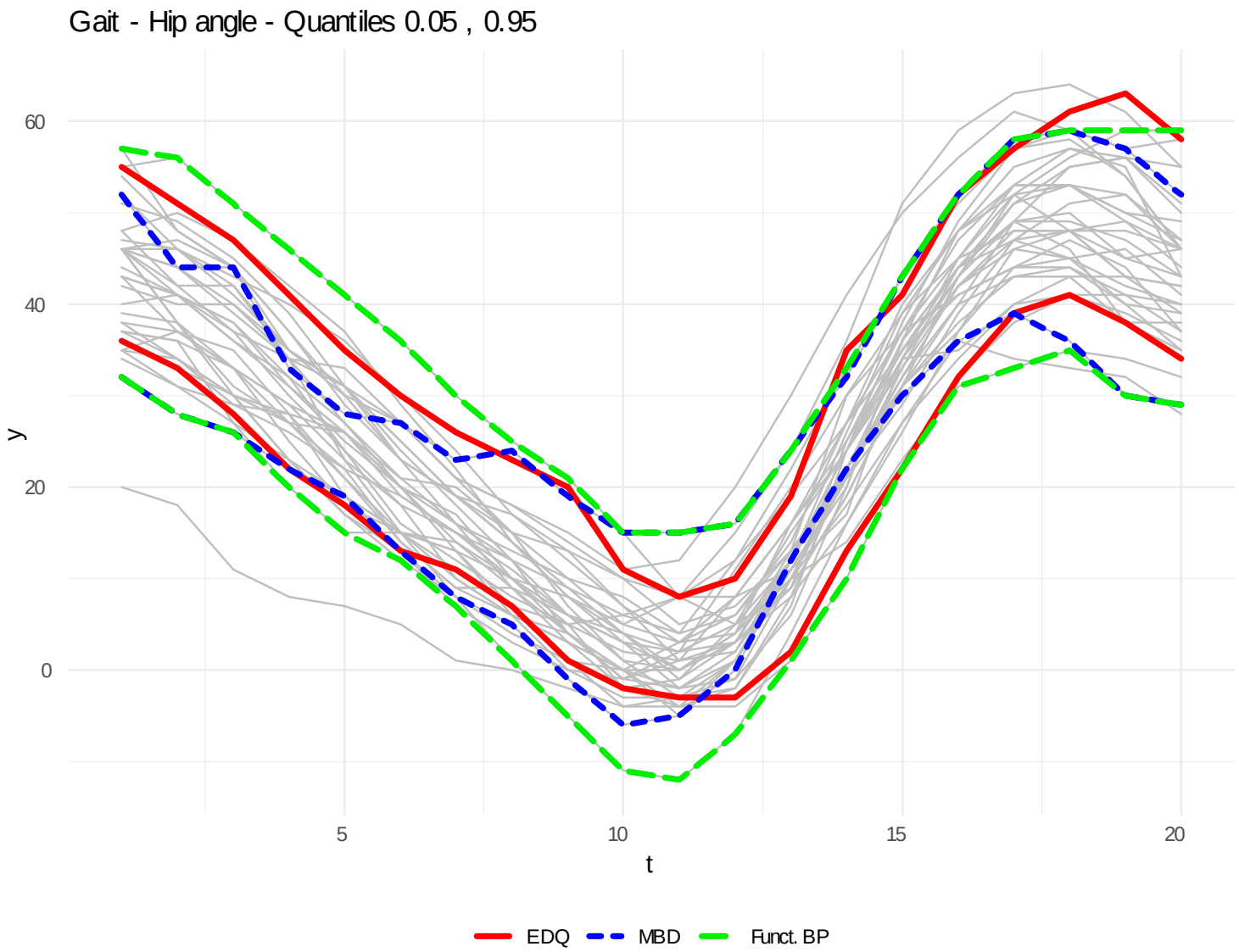
Gait data: hip angle. Medians are equal

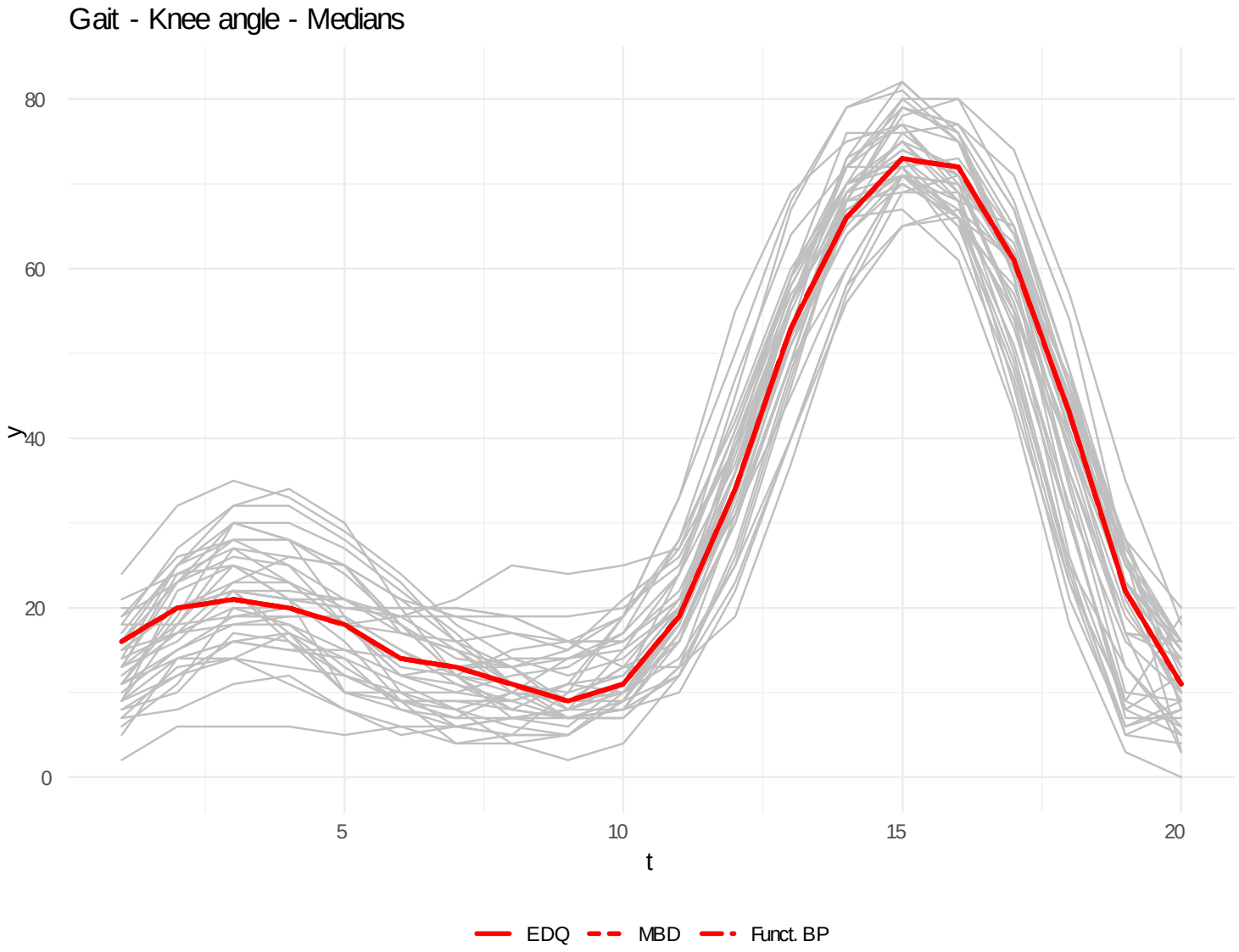


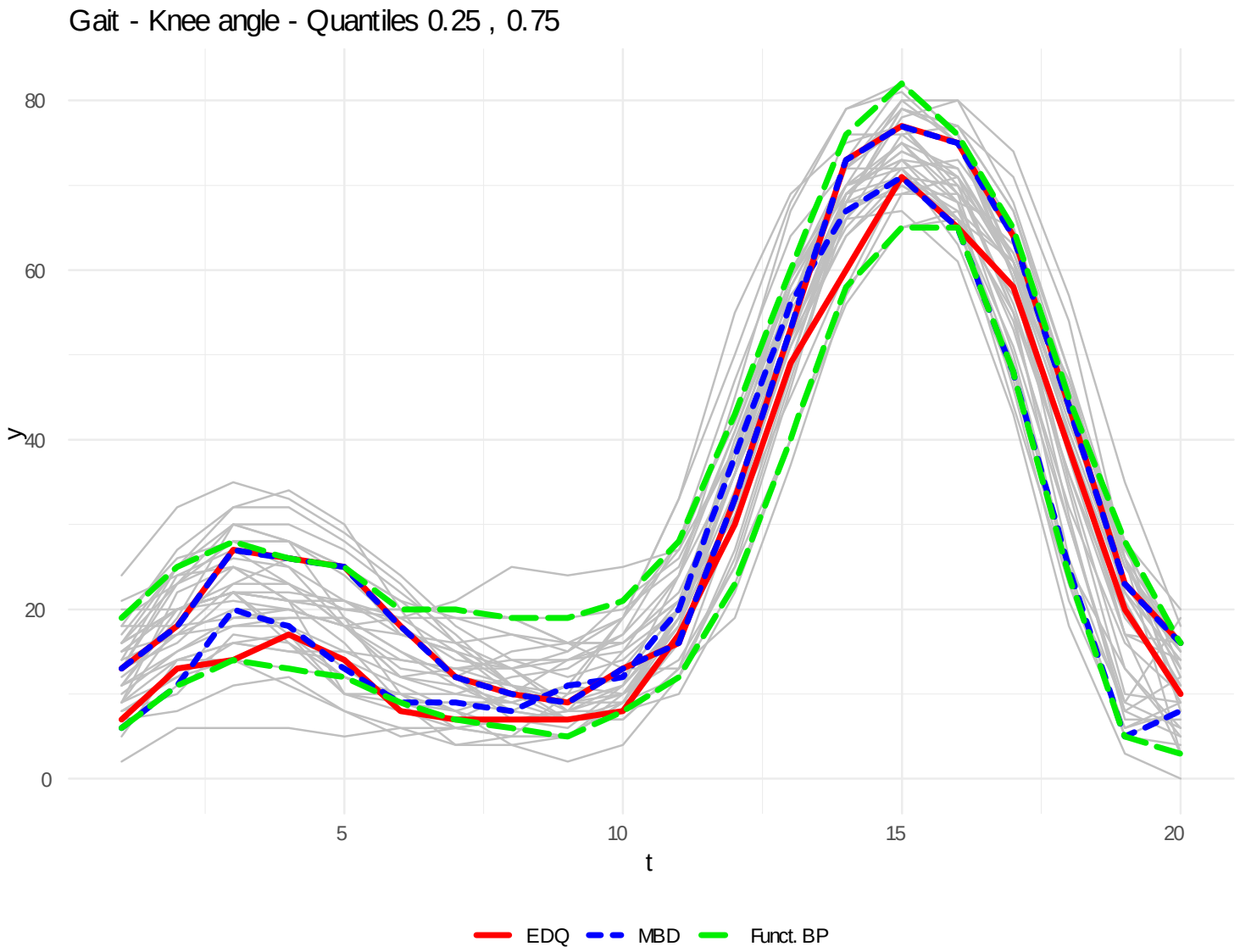
Gait data: hip angle. .25 and .75 quantiles

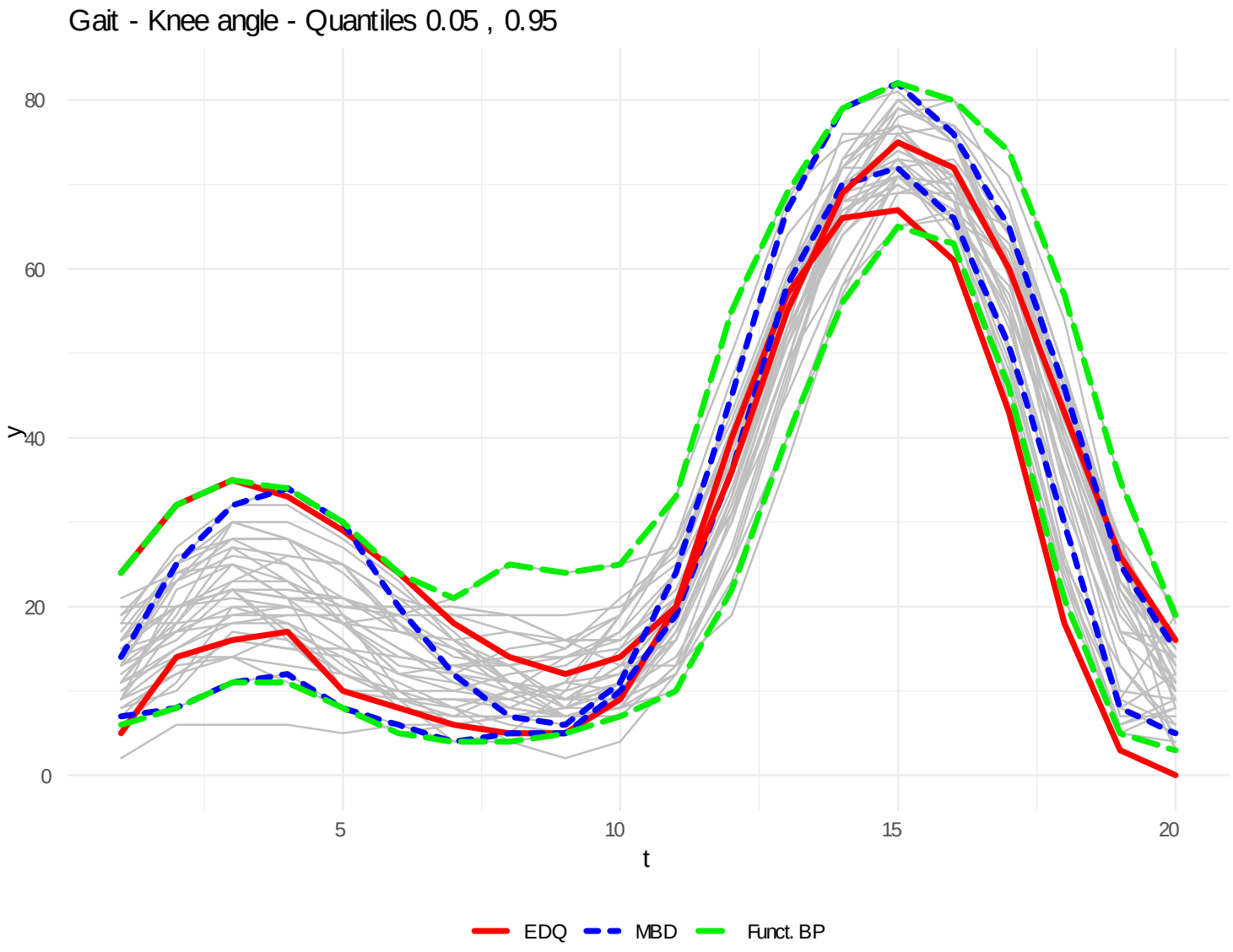


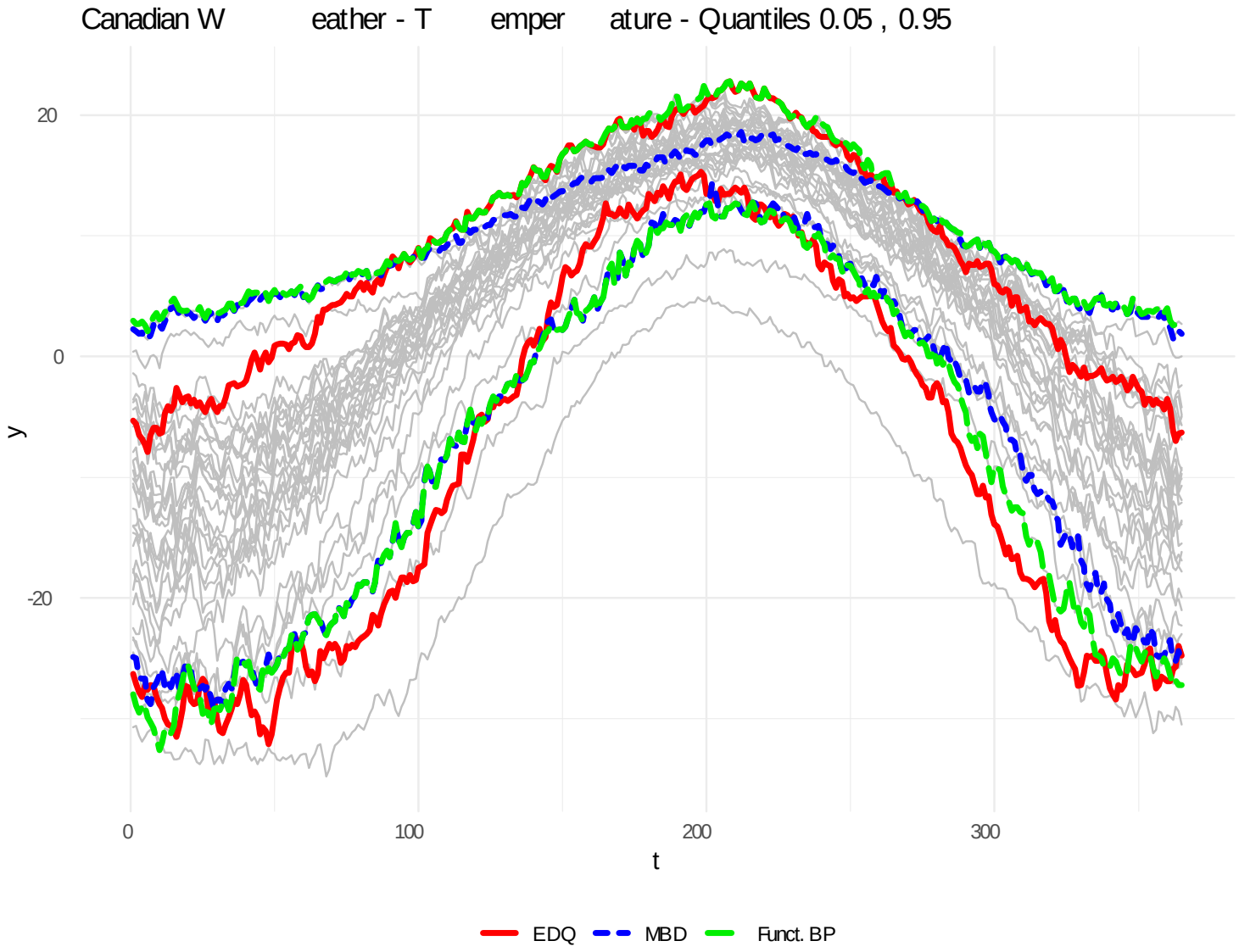
Gait data: hip angle. .25 and .75 quantiles

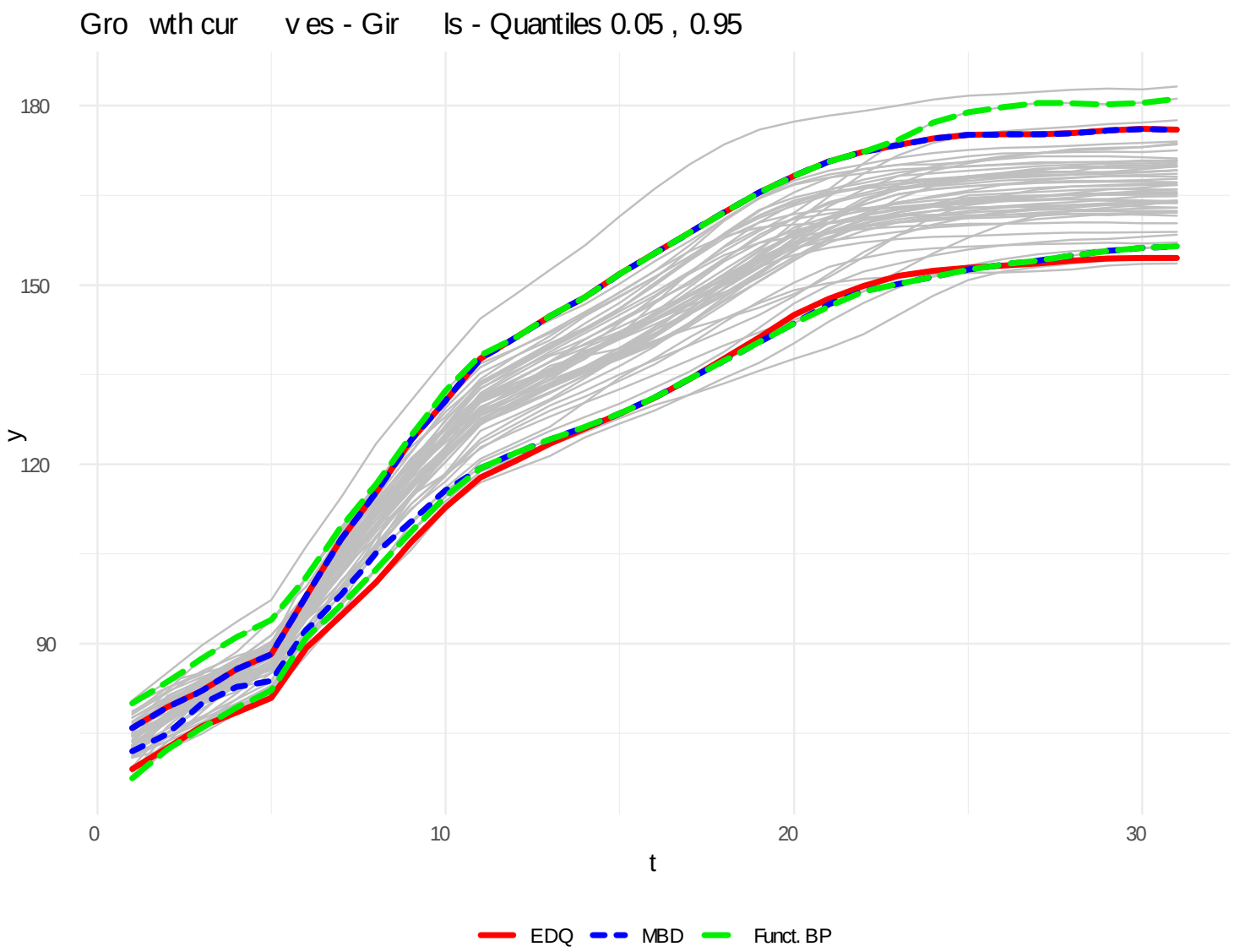



Gait - Knee angle - Medians
80
60
40
20
0
y
5
10
15
20
t
EDQ
MBD
Funct. BP


Gait - Knee angle - Quantiles 0.25 , 0.75
80
60
40
20
0
y
5
10
15
20
t
EDQ
MBD
Funct. BP


Gait - Knee angle - Quantiles 0.05 , 0.95
80
60
40
20
0
y
5
10
15
20
t
EDQ
MBD
Funct. BP


Canadian W eather - T emper ature - Quantiles 0.05 , 0.95
20
0
-20
y
0
100
200
300
t
EDQ
MBD
Funct. BP


Gro wth cur v es - Gir ls - Quantiles 0.05 , 0.95
180
150
120
90
y
0
10
20
30
t
EDQ
MBD
Funct. BP

Conclusions

The EDQ are shown to be useful for visualization of a large number of time series.
Compared to the timewise quantiles they have the following advantages:
(1) Convey the time dynamics of the series
(2) Can be identified with individual time series in the set enhancing its interpretation
(3) Are efficiently computed by a fast algorithm.
On the other hand, the ETQ give information of the evolution of the marginal distribution of the set of series. We have shown in the examples that the joint use of both quantiles provide a better description of the data set that the one given by using only the ETQ.
The idea can be extended for functional data.

Thank you for your attention