Multivariate Empirical Dynamical Quantiles

Benjamin Hines*and Guoqi Qian†

March 10, 2020

Abstract

In time series analysis we use empirical dynamical quantiles to separate the time series data set into quantiles that we can use to help us summaries the data. In this paper we generalise empirical dynamical quantiles to a multivariate time series setting. To do this, we create a new method for ordering multivariate points in a data set by using a generalised depth measure. We call this new multivariate ordering *Directional Depth*.

AMS classifications: 62M10, 62G30, 62H10.

Keywords: Time-Series Analysis, Multivariate Statistics, Empirical Distribution Theory.

1 Introduction

Quantiles are points that divide the range of a probability distribution into intervals where each interval contains a specified amount of probability. Given a probability distribution function F for some random variable X, the p^{th} quantile of the

*B. Hines is with Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia; email: bhines@student.unimelb.edu.au

†G. Qian is with Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia; email: g.qian@ms.unimelb.edu.au.

distribution $\pi^{(p)}$ is defined as

$$\pi^{(p)} = \inf_{x \in \mathbb{R}} \{x | F(x) \geq p\} \tag{1}$$

that is, from $-\infty$ to the point $\pi^{(p)}$ there is p amount of probability present. If the distribution is inveritble over its domain (such as an exponential distribution) then equation (1) simply becomes

$$\pi^{(p)} = F^{-1}(p). \tag{2}$$

We call $\pi^{(p)}$ for $p \in [0, 1]$ a theoretical quantile.

In a situation where we have a sample of data from some unknown univariate distribution, we use quantiles to help us explain the nature of the data and its underlying distribution by comparing these sample quantiles to the theoretical quantiles of a known distribution. In a univariate sample of n observations, the p^{th} quantile $q^{(p)}$ is defined by

$$q^{(p)} = \inf_{x \in \mathbb{R}} \left\{ x | \hat{F}(x) \geq p \right\} \tag{3}$$

where $\hat{F}$ is the empirical distribution function (EDF) of the sample [1]. This empirical quantile can also be expressed by

$$q^{(p)} = \arg\min_{y \in \mathbb{R}} \left\{ \sum_{x_i \geq y} p|x_i - y| + \sum_{x_i < y} (1-p)|x_i - y| \right\} \tag{4}$$

[2].

In a situation where we have a n different time-series observations across T different time periods (e.g Spatio-Temporal data), $\mathbf{x}_i^T = (x_{i,1}, \ldots, x_{i,T})$ for $i = 1, \ldots, n$, we can put the different time-series into quantiles via a method know as empirical dynamical quantiles (EDQ). In EDQ, for the observed data C_n, the p^{th} empirical dynamic quantile is the time-series $\mathbf{x}_i$ in C_n that satisfies the equation

$$q_t^{(p)} = \arg\min_{y_t \in C_n} \left\{ \sum_{t=1}^{T} \left(\sum_{x_{i,t} \geq y_t} p|x_{i,t} - y_t| + \sum_{x_{i,t} < y_t} (1-p)|x_{i,t} - y_t| \right) \right\} \tag{5}$$

where $C_n = \{\mathbf{x}_i | i = 1, \ldots, n\}$ is the set of n observed time-series [3]. As an example of EDQ, consider figure 1 which shows 500 simulated stationary $AR(1)$ time-series over 200 time intervals, where each initial value is generated independently from a

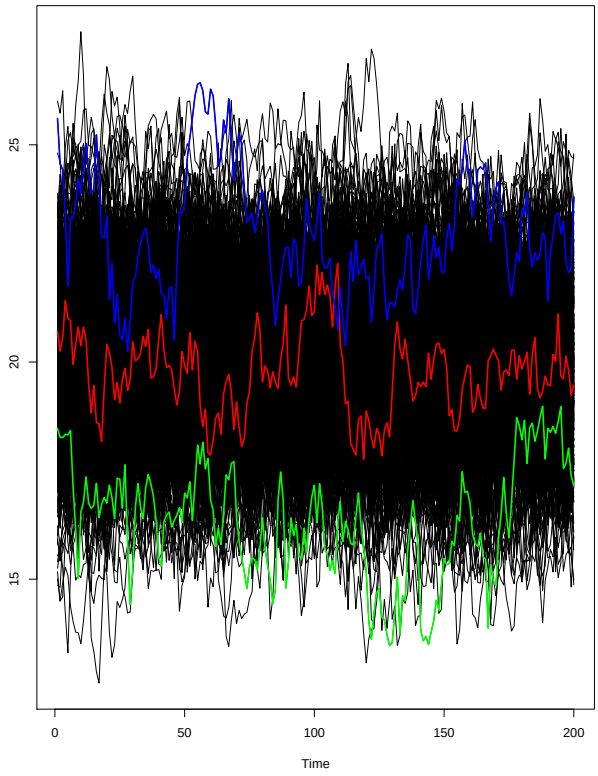

Figure 1: 500 simulated $AR(1)$ time-series over 200 time intervals with the empirical dynamical quantiles for $p = 0.05$ (green), $p = 0.5$ (red), $p = 0.95$ (blue).

$\mathcal{N}(20, 1)$. We can see the time-series that represent the 0.05^{th} (green), 0.5^{th} (red) and 0.95^{th} (blue) percentiles. We can see that there is a lot of variance in the different time-series given by the EDQ computation, even so much so the 0.95^{th} quantile time-series and the 0.5^{th} quantile time-series cross over.

EDQ is a useful tool in situations such as a spatio-temporal setting where we have a time-series at many different locations. Sometimes we have so many locations that any attempt at modelling the spatio-temporal process becomes a highly computationally intensive task so much so that it cannot be done, thus we can use EDQ to reduce the dimensionality of the data set such that we can use a subset of locations to explain the trend of the data.

2 Statistical Depth

With non-univariate data, we cannot calculate quantiles by the same means as the univariate data. Consider the case where we have n $d-$dimensional random samples $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots n$ from some $d-$dimensional multivariate random variable. For a sample of multivariate data, we use what is known as depth to give a quantile like description of the data set. There are several different methods for calculating the depth of a point in a data set. For example, the Mahalanobis depth of a point $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T \in S_n \subset \mathbb{R}^d$ is defined as

$$D(\mathbf{x}_i; S_n) = \left[1 + (\mathbf{x}_i - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right]^{-1} \tag{6}$$

where S_n is an $n \times d$ data matrix, $\bar{\mathbf{x}}$ is the vector of dimensional means of S_n and S is the sample covariance matrix of S_n [4]. Another method for finding the depth of a point is Location (Tukey) depth, where for a point $\mathbf{x}_i \in S_n$ the depth is defined as the smallest number of data points in a closed halfspace with the boundary through $\mathbf{x}_i$ [5]. The depth of a point relative to a given data set gives a measurement of how deep that point lies in the data cloud [6].

Figure 2 shows semi-correlated 2−dimensional data simulated from a bivariate normal distribution. The blue triangle represents the Mahalanobis deepest (largest depth) point in the data set which is as we can see, is roughly the centre of the data cloud.

For any given depth function, we have certain properties which are desirable for the function have [7].

1. *Affine invariance.* The depth of a point $\mathbf{x}_i \in S_n$ should not depend on the underlying coordinate system or on the scale of the underlying measurements.

2. *Maximality at center.* For a distribution having a uniquely defined center, the depth function should attain maximum value at this center.

3. *Monotonicity relative to deepest point.* As a point $\mathbf{x}_i \in S_n$ moves away from the deepest point along any fixed ray through the center, the depth of $\mathbf{x}_i$ should decrease monotonically.

4. *Vanishing at infinity.* The depth of a point $\mathbf{x}$ should approach zero as $||\mathbf{x}_i||_1$ approaches infinity.

Clearly shifting the entire data set by

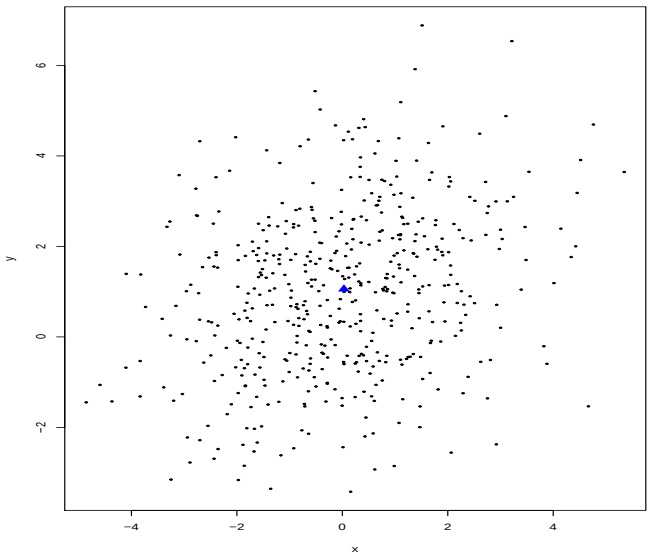

Figure 2: Mahalanobis deepest point for semi-positive correlated bivariate data simulated from a normal distribution.

We can of course use depth for univariate data. In the univariate setting, the deepest point corresponds to the median of the data. As an example, consider the univariate data $1, 2, 3$, the depth of these points are $\frac{1}{3}, \frac{2}{3}$ and $\frac{1}{3}$ respectively. The issue with using depth as a substitute for quantiles is that we lose a sense of direction, that is the depth function is not a monotonic function. Thus, we can make an improvement to the notion of depth, to make a directional depth such that it is more comparable to a quantile definition.

Let $\mathbf{x}^{(m)}$ be the deepest point with respect to the data set S_n for some depth function $D(\mathbf{x}_i; S_n)$. If we consider the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of the sample covariance matrix S then the principal eigenvector $\mathbf{e}_1$ can give the depth function a sense of direction. If we add the principal eigenvector to the deepest point then we obtain another vector to this from the deepest point $\vec{\mathbf{v}} := \mathbf{x}^{(m)} \to \mathbf{x}^{(m)} + \mathbf{e}_1$, which of course is perpendicular to all other eigenvectors. Consider the vector from the deepest point to the $\mathbf{x}_i$, $\vec{\mathbf{u}}_i := \mathbf{x}^{(m)} \to \mathbf{x}_i$, if $\vec{\mathbf{u}}_i$ has a positive projection onto $\vec{\mathbf{v}}$, then we can consider the point $\mathbf{x}_i$ to have a larger directional depth than $\mathbf{x}^{(m)}$. The sign of the projection of $\vec{\mathbf{u}}_i$ onto $\vec{\mathbf{v}}$ can be expressed in terms of the angle between these vectors. If we let θ_i be the angle between $\vec{\mathbf{u}}_i$ and $\vec{\mathbf{v}}$, then the projection of $\vec{\mathbf{u}}_i$ onto $\vec{\mathbf{v}}$ is positive if $0 \leq \theta_i \leq \frac{\pi}{2}$ and negative if $\frac{\pi}{2} < \theta_i \leq \pi$, where $\theta_i = \arccos \frac{\vec{\mathbf{u}}_i \cdot \vec{\mathbf{v}}}{||\vec{\mathbf{u}}_i||_2 ||\vec{\mathbf{v}}||_2}$ and

$||\mathbf{a}||_2 = (\sum_{i=1}^{n} |a_i|^2)^{\frac{1}{2}}$. We can express this new depth of the point $\mathbf{x}_i$ as

$$D'(\mathbf{x}_i; S_n) = \begin{cases} D\big(\mathbf{x}^{(m)}; S_n\big) + \big(D\big(\mathbf{x}^{(m)}; S_n\big) - D(\mathbf{x}_i; S_n)\big) & , \text{ if } 0 \leq \theta_i \leq \frac{\pi}{2} \\ D(\mathbf{x}_i; S_n) & , \text{ otherwise} \end{cases} \tag{7}$$

Figure 3 shows an example of this for the bivariate case in the data above. The arrow from the deepest point indicates the direction of the principal eigenvector. The green points have a positive projection from the deepest point onto the vector $\vec{\mathbf{v}}$ and therefore are considered to have a larger directional depth than $\mathbf{x}^{(m)}$ whereas the red points have a negative projection from the deepest point onto $\vec{\mathbf{v}}$ and are therefore considered to have a smaller directional depth than $\mathbf{x}^{(m)}$. There are of course times

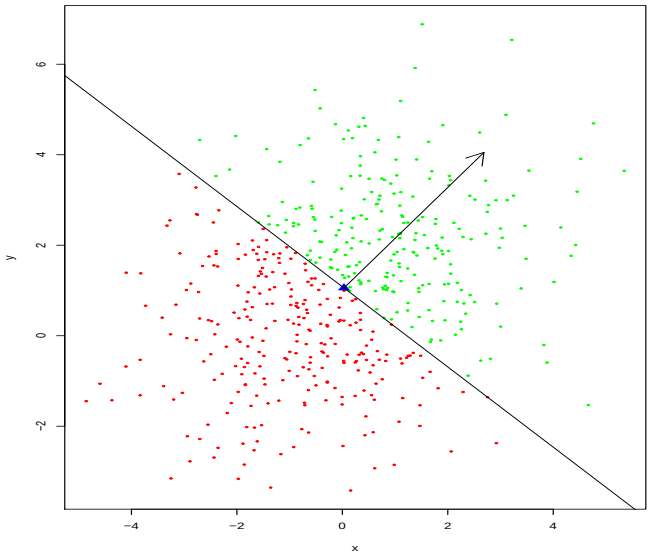

Figure 3: Directional depth based on the direction of the principal eigenvector from the deepest point.

where the principal eigenvector direction may be negative in one or more directions. In a bivariate case, this can occur when the two variables are negatively correlated and thus the principal eigenvector will be negative in one direction and positive in the other. As we are trying give direction to multi-dimensional data, this will not be a problem as it will still help describe the trend of the data.

1. $D'\big(A\mathbf{x}_i + \mathbf{b}; AS_n + \mathbf{b}\mathbf{1}_n^T\big) = D'(\mathbf{x}_i; S_n)$ for an $d \times d$ orthogonal matrix A and $d \times 1$ vector $\mathbf{b}$

2. $D'(\mathbf{x}_i; S_n) \to 0$ as $-\mathbb{I}\big(\frac{\pi}{2} < \theta_i < \pi\big)||\mathbf{x}_i||_2 \to -\infty$ and
   $D'(\mathbf{x}_i; S_n) \to 2D\big(\mathbf{x}^{(m)}; S_n\big)$ as $\mathbb{I}\big(0 \leq \theta_i \leq \frac{\pi}{2}\big)||\mathbf{x}_i||_2 \to \infty$

3 Multivariate EDQ

There are situations where we are dealing with a multivariate time-series, that is, we have two different observed variables which both come from the same source. For example, consider a rain gauge measurement and satellite precipitation estimate at some location for a given time period. If we want to calculate the empirical dynamical quantiles for a set of multivariate time-series, we can make altercations to equation (4) to achieve this. The univariate EDQ equation uses the quantiles of the data at each time interval. We can instead use the directional depth function as in equation (7) and we also replace the absolute difference with an ℓ_2-norm function (note that other distance metrics could be used). These multivariate empirical dynamical quantiles are defined as following

$$\mathbf{q}_t^{(p)} = \underset{\mathbf{y}_t \in C_n}{\arg\min} \left\{ \sum_{t=1}^{T} \left(\sum_{D'(\mathbf{x}_{i,t}; S_{n,t}) \geq D'(\mathbf{y}_t; S_{n,t})} p||\mathbf{x}_{i,t} - \mathbf{y}_t||_2 + \sum_{D'(\mathbf{x}_{i,t}; S_{n,t}) < D'(\mathbf{y}_t; S_{n,t})} (1-p)||\mathbf{x}_{i,t} - \mathbf{y}_t||_2 \right) \right\} \quad (8)$$

where $\mathbf{x}_{i,t}$ is the i^{th} multi-dimensional vector at time t and $S_{n,t}$ is the multivariate data at time t.

We can see in figure 4, a simulated example of the multivariate EDQ where we have two sets of time-series where the i^{th} time-series in each set is semi-positively correlated. That is, each time-series is a simulated AR(1) model where the i^{th} time-series in each set share a common initial value relative to the mean of the set with some error. In other words,

$$\begin{aligned} x_{i,t}^{(1)} &= \phi_1 x_{i,t-1}^{(1)} + \varepsilon_{i,t}^{(1)} \\ x_{i,t}^{(2)} &= \phi_1 x_{i,t-1}^{(2)} + \varepsilon_{i,t}^{(2)} \end{aligned} \quad (9)$$

where $\phi_1 = 0.8897$, $\varepsilon_{i,t}^{(1)} \overset{d}{=} \varepsilon_{i,t}^{(2)} \sim \mathcal{N}(0, 0.4)$, $x_{i,1}^{(1)} \sim Z_s + \varepsilon_{i,1}^{(1)}$, $x_{i,1}^{(2)} \sim Z_s + 1 + \varepsilon_{i,1}^{(2)}$ and $Z_s \sim \mathcal{N}(0,1)$. These two sets of realisations are semi-correlated, in that they have initial values that are up to some error, the same relative to the mean of the entire set and as $|\phi_1| < 1$ these time-series are stationary [8]. Figure 4 shows the multivariate EDQ for the 0.05^{th} (green), 0.5^{th} (red) and 0.95^{th} (blue) quantile time-series.

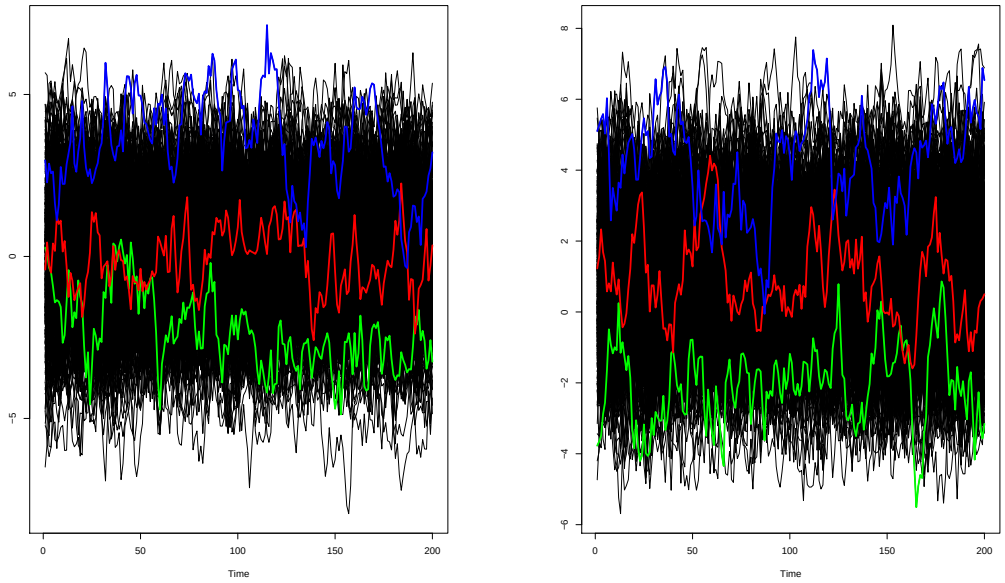

Figure 4: Two sets of semi-positively correlated stationary $AR(1)$ time-series with the empirical dynamical quantiles for $p = 0.05$ (green), $p = 0.5$ (red), $p = 0.95$ (blue).

As we can see, the EDQ time-series given exhibit strong positive correlation in their respective position in their data sets. Thus we call this type of result strong positive multivariate EDQ (MEDQ).

Another example of multivariate EDQ can come from the situation where the variables are no longer positively correlated but now negatively correlated, and we can also increase the dimensionality of the problem so that now we have four variables for each time series:

$$\begin{aligned} x_{i,t}^{(1)} &= \phi_1 x_{i,t-1}^{(1)} + \varepsilon_{i,t}^{(1)} \\ x_{i,t}^{(2)} &= \phi_1 x_{i,t-1}^{(2)} + \varepsilon_{i,t}^{(2)} \\ x_{i,t}^{(3)} &= \phi_1 x_{i,t-1}^{(3)} + \varepsilon_{i,t}^{(3)} \\ x_{i,t}^{(4)} &= \phi_1 x_{i,t-1}^{(4)} + \varepsilon_{i,t}^{(4)} \end{aligned} \tag{10}$$

where $x_{i,1}^{(1)} \sim Z_s + \varepsilon_{i,1}^{(1)}$, $x_{i,1}^{(2)} \sim -2Z_s + 10 + \varepsilon_{i,1}^{(2)}$, $x_{i,1}^{(3)} \sim Z_s - 5 + \varepsilon_{i,1}^{(3)}$ and $x_{i,1}^{(4)} \sim -Z_s - 10 + \varepsilon_{i,1}^{(4)}$ (ϕ_1, Z_s and $\varepsilon_{i,t}^{(1,2,3,4)}$ are defined the same as above).

Figure 5 shows the four different time-series, where the top left plot shows the time-series corresponding to $\mathbf{x}^{(1)}$, which has positively correlated initial values to the

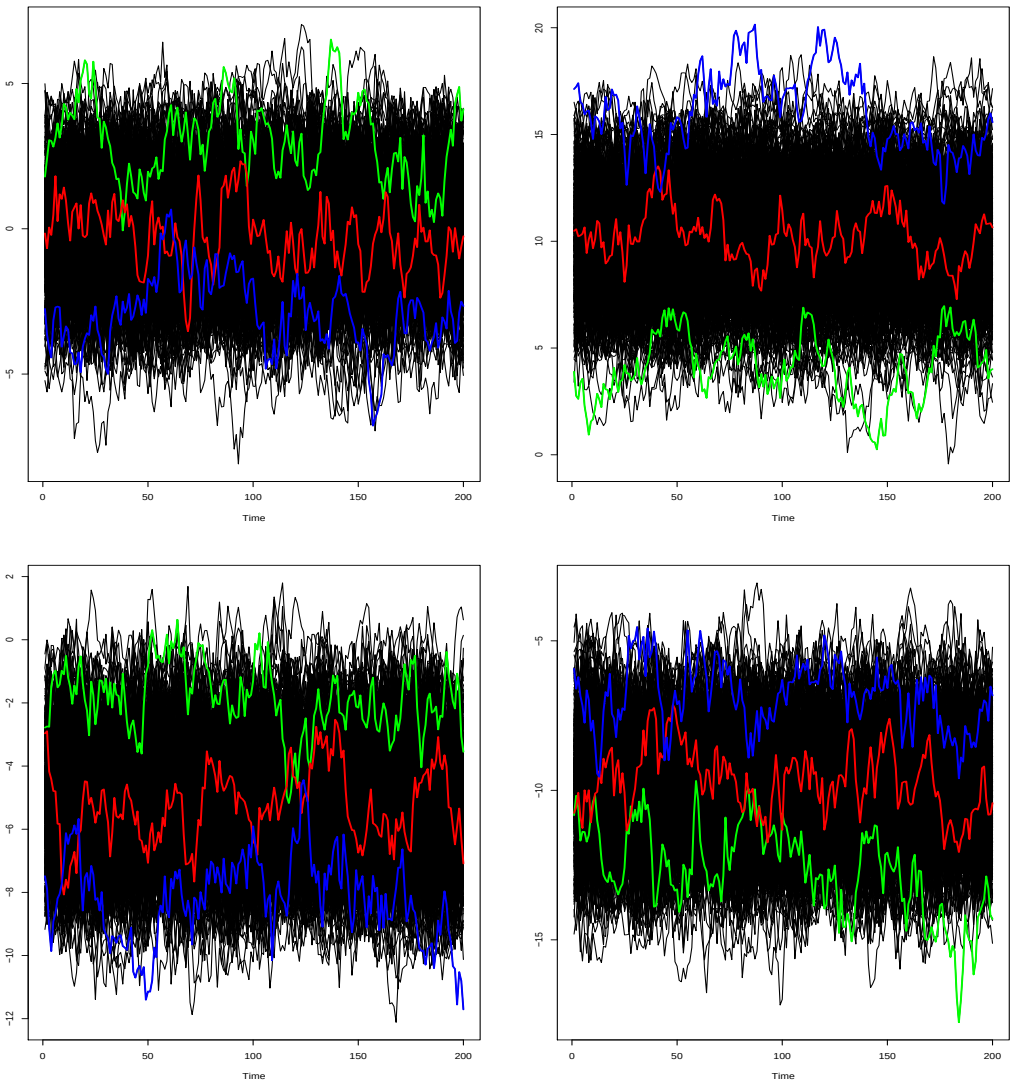

Figure 5: Four sets of positively and negatively semi-correlated stationary $AR(1)$ time-series with the empirical dynamical quantiles for $p = 0.05$ (green), $p = 0.5$ (red), $p = 0.95$ (blue).

time-series in the bottom left plot ($\mathbf{x}^{(3)}$) but negatively correlated with the time-series in the top right ($\mathbf{x}^{(2)}$) and bottom right ($\mathbf{x}^{(4)}$). Interestingly, having negatively correlated initial values does not seem to affect the overall results of the multivariate EDQ. The 0.05^{th} quantile (green) is in the upper parts of the top left and bottom left plots, but in the lower parts of the top right and top left plots. Similarly the 0.95^{th} quantile (blue) is in the lower parts of the top left and bottom left plots, but in the upper parts of the top right and bottom right plots. $\mathbf{x}^{(1)}$ exhibits strong positive MEDQ with $\mathbf{x}^{(3)}$ but exhibits strong negative MEDQ with $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(4)}$ in that a given upper calculated quantile in one set appears as a lower calculated quantile in

the other set.

4 Australian Precipitation Data

Consider the monthly precipitation amounts in Australia, as an example of a multivariate time-series where the variables are the monthly rain gauge measurements and the monthly satellite precipitation estimate for how much precipitation has occurred in a given region. We have a set of multivariate time-series at 253 locations across 187 different time periods (over 15 years). In theory, a satellite can give a good estimate of how much precipitation that a rain gauge has recorded for a given time period up to some amount of error. Therefore, we can assume that the two variables are somewhat positively correlated and therefore using a multivariate EDQ equation in equation (8) to obtain the time-series quantiles of this data set, should give strong positive MEDQ results.

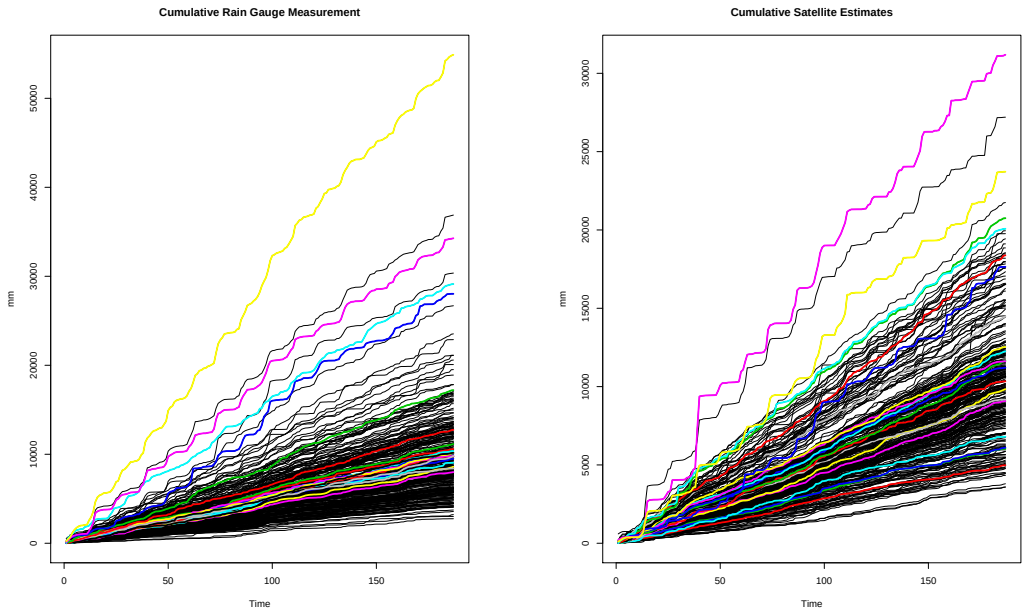

Figure 6: Cumulative Rain Gauge Measurements (left) and Cumulative Satellite Estimates (right) with multivariate EDQ.

In figure 6, we have the accumulated rain gauge measurements (left) and accumulated satellite precipitation estimates (right). We have performed multivariate EDQ on this data set and we can see an interesting result. We have calculated the quantiles for $p = 0, 0.01, \ldots, 1$ and obtain 22 unique multivariate time-series.

The yellow time series corresponds to the $p = 1$ EDQ and we can see that this is the largest observation in the cumulative rain gauge plot but the third largest observation in the cumulative satellite plot, where as the pink time-series corresponds to the $p = 0.99$ EDQ which is the largest observed time-series in the cumulative satellite plot and the third largest in the cumulative gauge plot. As the largest accumulated gauge measurement is significantly larger than the second largest gauge measurement compared to the satellite estimated where the difference is a lot smaller, thus the gauge value has more weight and thus it is given as the largest quantile.

References

[1] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[2] Thomas S Ferguson. *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press, 2014.

[3] Daniel Peña, Ruey S Tsay, and Ruben Zamar. Empirical dynamic quantiles for visualization of high-dimensional time series. *Technometrics*, pages 1–25, 2019.

[4] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. In . National Institute of Science of India, 1936.

[5] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[6] Maria Raquel Neto. The concept of depth in statistics. , .

[7] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

[8] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, NJ, 1994.