Functional Skewness and Kurtosis Based on Functional Quantiles

Juan David Otálora

EAFIT University, Medellín, Colombia.

E-mail: jotalor1@eafit.edu.co

Francisco Zuluaga

EAFIT University, Medellín, Colombia.

Summary. Functional data analysis has translated several concepts from classical statistics. In this article we present functional Skewness and Kurtosis, based in the quantiles of the functional data. We develop an interpretation for these estimators and test them in real-world data.

1. Introduction

Functional data analysis is a new branch of statistics which, unlike classical statistics, take the data as continuous functions (curves) over time or other continuous variable. In fact, in several applied sciences the data obtained, with the advances in computational capacity, are sampled over a finer grid, so it can be assumed that the data comes from a family of continuous curves, hence allowing for the use of methodologies from curves. Some of the mentioned applied science areas include: environmetrics, chemometrics, biometrics, medicine, econometrics Ferraty and Vieu (2006). The main purpose of functional data analysis is to present a statistical description of the data. Therefore many of the classical statistics concepts and methods has been translated into functional data analysis. As in classical descriptive statistics, the main descriptive parameters to be estimated are: the mean, which describes the central tendency of the data, the standard deviation, which accounts the dispersion of the data, the Skewness, which indicates the symmetry of the data, and the Kurtosis, which gives a notion of the shape of the tails

of the data.

There are several estimators for the Skewness and Kurtosis. Estimators based on quantiles have been of special importance due to its capacity to replace the sensibility to extreme data that the mean based estimators posess. The definition of the Kurtosis has been attributed as the peakedness of the density distribution, however, there are counterexamples like the t-student distribution, that put on doubt this affirmation. In fact, an estimator based on quantiles can resolve this query. It is important to use more clear definitions of Kurtosis, like the ones which are based on the tails of the probability distribution of the data.

The main purpose of the present research proposal is to establish an estimator, for functional data, of the Skewness and Kurtosis curves. This translation is planned to be based on the quantiles of the data, in order to contribute to the theoretical development of this recent area of research.

2. State of the art

The Skewness and Kurtosis have been responsible for the shape characteristics of the probability distribution of data. As described in Fernandez and Fuentes (1995) there are many estimators for the Skewness of the data, and of particular importance stands out the Yule asymmetry coefficient which is based on the quantiles of the data. The definition of the Kurtosis that this book presents aims at the peakedness of the density function.

The parameters as Skewness and Kurtosis have been interpreted as the asymmetry and peakedness of the data distribution, respectively. However, Moors (1988) presents Kurtosis as the variations around the points $\mu \pm \sigma$. In the latter article an estimation of the Kurtosis based on the octiles is presented. The advantages of this formulation are: 1) it exists even when the moments of the distribution does not, 2) it does not depend on the extreme tails of the distribution, 3) the calculation is simpler and graphically in-

terpretable. The authors implement this estimator in various distribution and compare the results. These estimations are based on univariate quantiles, which for that case characterize the distribution of the data.

For the translation to FDA we need to account for the quantiles which have been explored in the literature of functional data analysis. In Mengmeng Guo and Hardle (2015) the authors use a functional data approach to estimate the quantiles of a generalized regression. For this, a PCA methodology is used in order to minimize a loss function. From this methodology they find the quantile curves, and apply it to real world data of weather stations in China. Fatima Benziadi and Tebboune (2016) use a kernel estimation of the conditional quantiles for functional data for ergodic process and determine consistency in this estimator.Walter (2011) explores a way of defining functional quantiles for the real functional data, defined over a continuous interval. The author uses the estimated quantiles to estimate the Skewness of a functional data-base for finance to test whether the data has a positive tendency.

3. Methodology

In the present section the general terminology and definitions necessary to describe the proposed methodology of functional skewness and kurtosis. First, a definition of skewness and kurtosis based on quantiles is presented. Then the terminology of functional data analysis is described and the definition of the functional quantiles. The theory and definition presented is taken from Ferraty and Vieu (2006) and Walter (2011).

3.1. Skewness and Kurtosis based in quantiles

In Fernandez and Fuentes (1995) a definition of skewness based in quantiles is presented. This definition of skewness allows a more meaningful insight into the definition of skewness.

In figure 1 two probability distribution are presented: a normal distribution and a gamma distribution. The following example is presented to identify the skewness estimated with the quantiles of the distribution, the first with a symmetrical distribution

and the second with an asymmetrical distribution. The formula of the estimator is presented as follows:

$$B = \frac{(Q_3 - Q_2) + (Q_1 - Q_2)}{Q_3 - Q_1}$$

Where Q_i is the *ith* quartile. The logic behind this equation, also known as the Yule asymmetry coefficient, is that we compare the distance between the first and third quantile to the median, which corresponds to the second quantile. For a positive skewness, the median will be closer to the first quantile, and for a negative skewness, the median will be closer to the third quantile. So the sign of the numerator of B will be translated into the sign of the skewness. The denominator eliminates the units so B becomes dimensionless.

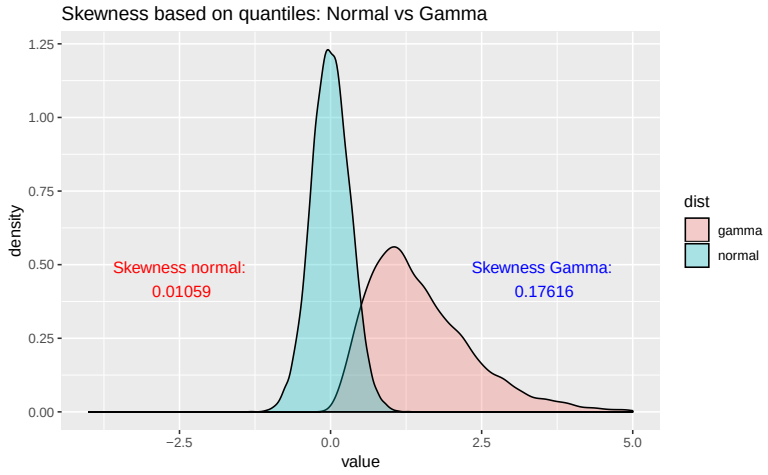

Fig. 1: Skewness

As we can see in figure 1, the normal distribution is symmetric, while the gamma distribution is not. This is why the normal distribution has an estimated skewness of approximated zero, and the gamma distribution has positive estimated skewness. The latter is biased to the left.

In Moors (1988) a definition for kurtosis based in quantiles is presented. In classical statistics kurtosis have been wrongly defined as the peakedness of the distribution.

In figure 2, a normal distribution and a t-student distribution are illustrated. The t-student is a counterexample for the definition of Kurtosis as the peakedness of the distribution. Here, the Kurtosis is estimated by the octiles of the distribution with the following formula:

$$T = \frac{(E_7 - E_5) + (E_3 - E_1)}{E_6 - E_2}$$

This formula estimates the dispersion around the values $\mu \pm \sigma$. The argument behind this formula is that if the distance between each of the terms in the numerator is large, then there is small dispersion around the points $\mu \pm \sigma$, and vice-versa.

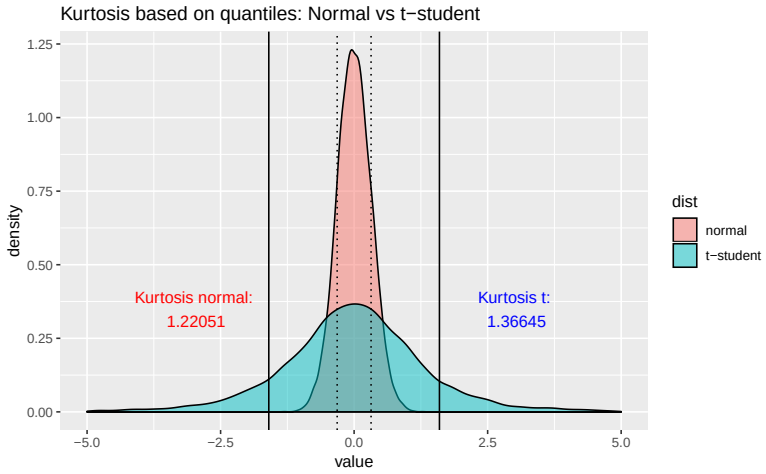

Fig. 2: Kurtosis

In figure 3, the cumulative distribution functions of the distribution from figure 2 are presented. We can observe the octiles E_1, E_3, E_5, E_7 of both distribution as the vertical dotted lines. It is clear that the octiles E_1, E_2 and E_5, E_7 are further apart for the t-student distribution, which can be translated that the dispersion around the points $\mu \pm \sigma$ is lower for the t-student distribution. Hence, the t-student distribution has a higher kurtosis. Thus, solving the higher kurtosis of the t-student distribution without referring to the peakedness of the distribution.

Now, several important definitions of the functional data analysis are presented. It

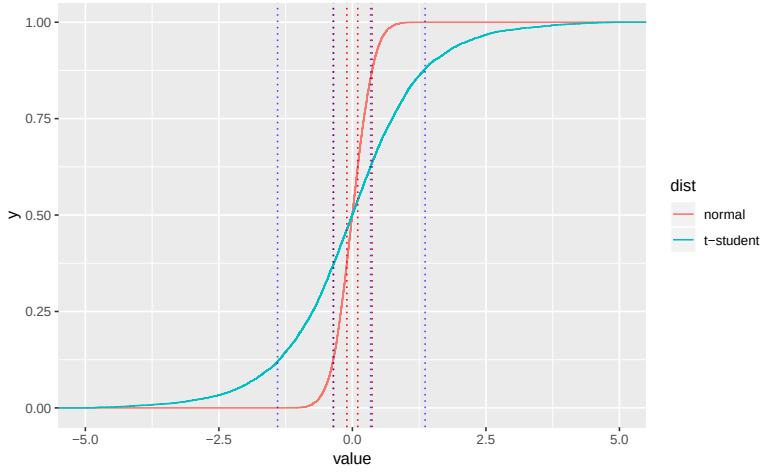

Fig. 3: Kurtosis CDF

is needed to understand how to define functional quantiles and how to estimate them empirically. The first important definition is the functional random variable, as it is defined as follows.

3.2. Functional Random Variable

A random variable $\mathcal{X} : \Omega \to E$ is a functional random variable iff every $X \in E$ is a function $X : M \to F$ for some non-degenerate continuum M.

$\mathcal{X}$ is real functional random variable iff every $X \in E$ is a function mapping $\mathcal{I} \to \mathbb{R}$, where $\mathcal{I} = [a, b], a, b \in \mathbb{R}$ and $a < b$

The cumulative distribution can also be defined for functional random variables. As we know univariate quanitles, cumulative distribution function is important to define quantiles. Distributional functional are defined as follows:

3.3. Distribution Functional

If $\mathcal{X} : \Omega \to E$ is a real functional random variable then the distributional functional is:

$$F_{\mathcal{X}}(X) = P[\mathcal{X}(t) \leq X(t) \; for \; all \; t \in \mathcal{I}]$$

Now that we have the definition of functional random variable and distribution functional, we can define the functional quantiles. This definition is based in the marginal distribution function. The definitions of functional quantiles is defined as follows:

3.4. Functional quantiles

If $\mathcal{X} : \Omega \to E$ is a real functional random variable with marginal distribution function $F_{\mathcal{X}}(X)$ at $t \in \mathcal{I}$ then the α quantile of $\mathcal{X}$ is:

$$\mathcal{Q}_\alpha(t) = inf\{x \in \mathbb{R} : F_{\mathcal{X}}(X) \geq \alpha\},\ for\, t \in \mathcal{I}$$

Finally, this quantiles need to be estimated from an empirical database. The say to do it is to replace the distributional functional by its empirical estimation. The definition of empirical estimation of functional quantiles is presented as follows:

3.5. Empirical estimation of functional quantiles

The estimation for the functional quantiles utilizes the empirical estimation for the functional empirical distribution function. The formula for estimating the latter is presented as follows:

$$\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$$

Let $\mathcal{X}$ be a real functional random variable and let $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$ be a functional dataset generated by $\mathcal{X}$. Let $\hat{F}_{\mathcal{X}(t)}$ be the empirical marginal distribution function at $t \in \mathcal{I}$ based on the dataset. The estimation of the $\alpha - quantile$ for the sample $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$ is:

$$\mathcal{Q}_\alpha(t) = inf\{x \in \mathbb{R} : \hat{F}_{\mathcal{X}(t)}(x) \geq \alpha\},\ for\, t \in \mathcal{I}$$

3.6. Functional Skewness and Kurtosis based on quantiles

Finally, the estimators are presented as follows:

3.6.1. Functional Skewness

Let $\mathcal{X}$ be a functional random variable with quantiles $\mathcal{Q}_\alpha(t)$. The functional Skewness is:

$$B(t) = \frac{(Q_3(t) - Q_2(t)) + (Q_1(t) - Q_2(t))}{Q_3(t) - Q_1(t)}$$

3.6.2. Functional Kurtosis

Let $\mathcal{X}$ be a functional random variable with quantiles $\mathcal{Q}_\alpha(t)$. The functional Kurtosis is:

$$T(t) = \frac{(E_7(t) - E_5(t)) + (E_3(t) - E_1(t))}{E_6(t) - E_2(t)}$$

3.7. Interpretation

Skewness in functional data analysis can be interpret as the direction of the tendency of the functions. If the skewness is positive, then the $quantile_3$ is further away from the median function than the $quantile_1$, then there is more probability in every t that the functions generated by the functional random variable will have an upward tendency over the continuos interval. The same reasoning applies in the opposite direction.

On the other hand, Kurtosis reflects several characteristics of the functional data-set. Low Kurtosis function can be seen as an indication of a bimodal distribution. This can be interpreted as the data-set coming from two different functional random variables. A high Kurtosis function can be an indication of high density around the mean function and the tails of the distribution. This definition of Kurtosis for functional data matches the definition of Kurtosis based on quantiles. In the latter, the Kurtosis is interpret as the inverse measure of density around the points $\mu \pm \sigma$. Therefore, when the Kurtosis is low, the points $\mu \pm \sigma$ show high density, the two modes of the bimodal functional distribution.

4. Results

4.1. Data Description

For the results, we are using two data-bases. The first one is taken from Febrero-Bande and Gonzalez-Manteiga (2008) and it contains trajectories measured hourly for contamination (NOx) levels in Pobleanu measured in $\mu g/m^3$ from the day February 23 to the day June in the year 2005. The second data-base, growth data-set, contains the heights of 39 boys and 53 girls between the age 1 to 18.

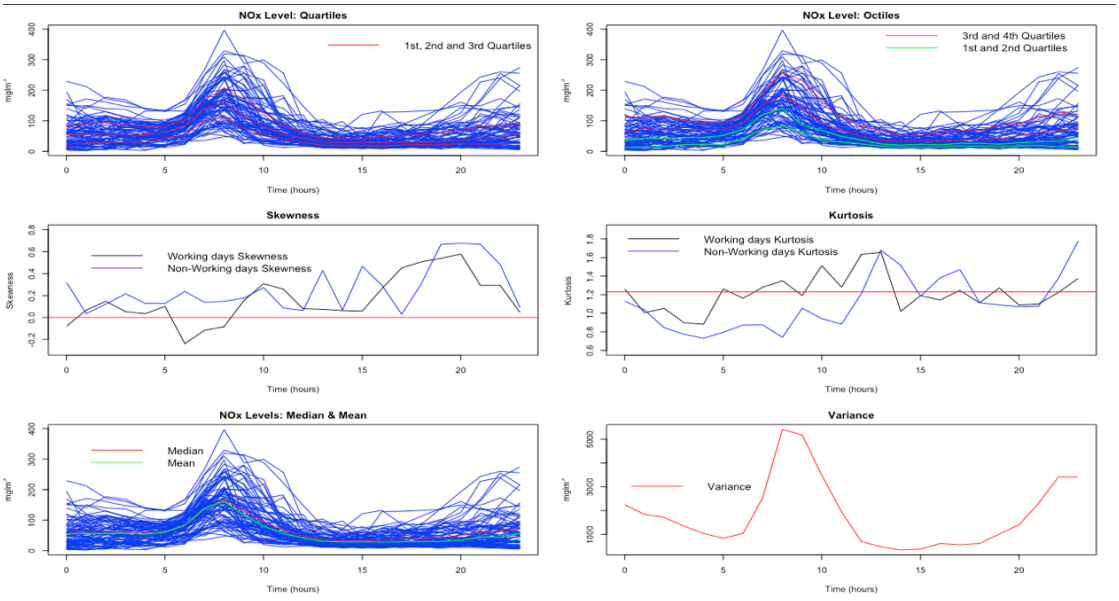

Fig. 4

As we can observe from figure 4 a complete description for the Poblenau data. In the first plot the functional data-set is presented and the first, second and third functional quartiles are the red lines. In the interval (0,5) hours the third quartile is slightly further from the the functional median than the first quantile. This result affects directly the Skewness, as we can observe in the figure of Skewness, the black line is slightly higher than the normal Skewness. For the next interval (5,13) hour the first and third functional quartiles are more separated than before due to the increase in the dispersion of the curves. The distance between the third and first quartile from the median mantains.

For the last interval (13,24) hour it is clear that the third quartile starts getting further away from the median than the first quartile, so the Skewness increases as well, as we observe in the Skewness figure. The curves, in the latter interval have a positive tendency.

The next figure illustrates the functional octiles for the data-set. The octiles help calculate the functional Kurtosis representend in the figure below. For the interval (0,5) hours the distance between the first and third octile, and the fifth and seventh octile remained generally constant. For the interval (5,11) hours the distance for the latter octiles starts increases, so the Kurtosis starts increasing as well. For the interval (12,24) the distance for the first and third octile reduces, but the distance from the fifth to the seventh octile starts increases, so the Kurtosis remains constant.

In the last two figures the classical functional estimators are presented: functional mean, median and variance. For the entire interval, the functional mean is slightly higher than the functional median. This reflects that the functional mean is biased towards the upper outliers. This also reflects that the functional random variable is for the most part, asymmetric, as well as presented in the Skewness figure, but with an upper tendency at the last 5 hours. The final figure represents the functional variance for the data-set.

For a better illustration for the definition of Kurtosis we use the growth data-set. As we can observe in figure 5 the blue curves represent the heights for the boys and the red curves represent the heights for the girls. For the interval (1,15) years the two types of curves seem to come from the same functional random variable. This matches with the high Kurtosis in this interval from the figure below, where the Kurtosis is higher than the Kurtosis for a Gaussian process. This deduction reflects a high density around the median of the curves and the tails. For the interval (15,18) years the two types of curves starts having different median. This matches with the definition of Kurtosis, where the density of the curves lies aruond the points $\mu \pm \sigma$, and from this point the two types of curves probably come from two different types of functional random variables.

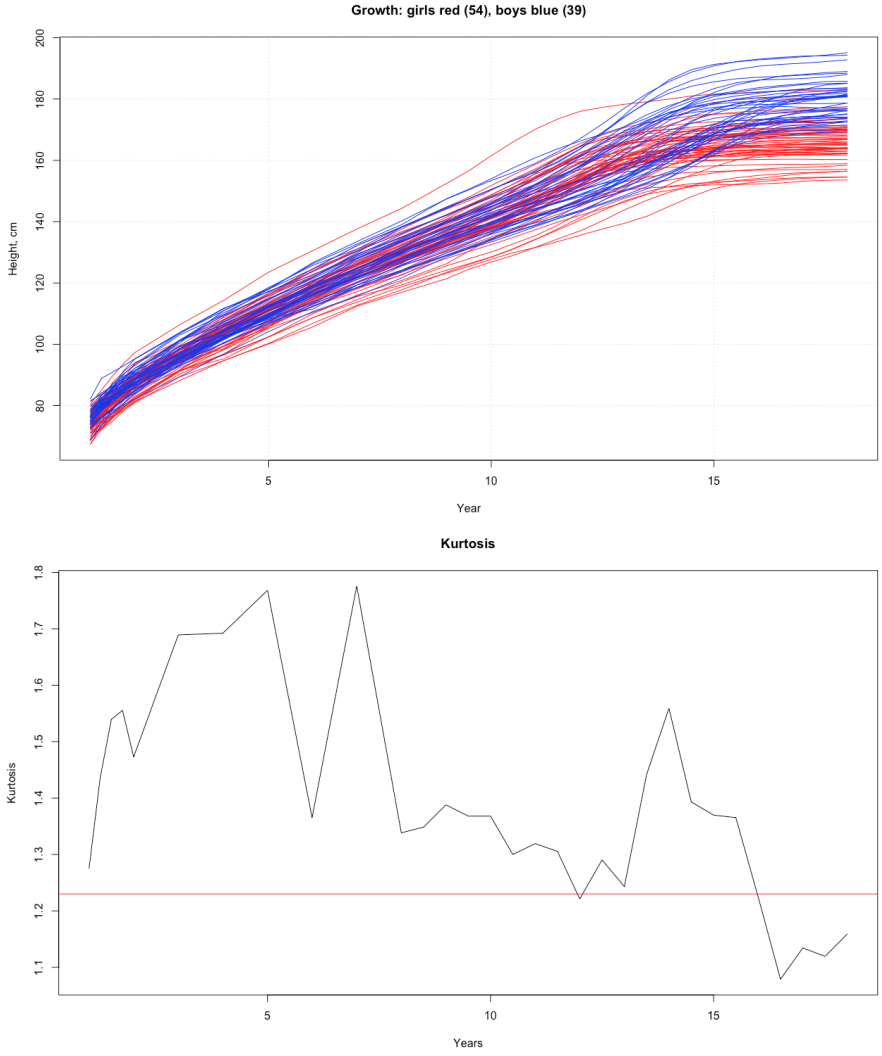

Fig. 5: Kurtosis for Growth data-set

5. Conclusions

In the present article we have translated the concepts of Skewness and Kurtosis, based on quantiles, to functional data analysis. In order to estimate these functional parameters, we used the estimators based on quantiles from classical statistics. Finally, we used the definition of functional quantiles to obtain the final estimators. We also developed an intuitive interpretation for these estimators for real-life use. Finally, we tested these estimators in real-world data and obtained a complete description for the curves.

References

Fatima Benziadi, A. L. and Tebboune, F. (2016) Note on conditional quantiles for functional ergodic data. *Comptes rendus - Mathématique*, **354**, 628–633.

Febrero-Bande, M, G. P. and Gonzalez-Manteiga, W. (2008) Outlier detection in functional data by depth measures with application to identify abnormal nox levels. *Environmetrics*, **19**, 331–345.

Fernandez, C. and Fuentes, F. (1995) *Curso de estadística descriptiva: Teoría y práctica.* Ariel, 1 edn.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis.* Springer, 1 edn.

Mengmeng Guo, Lan Zhou, J. Z. H. and Hardle, W. K. (2015) Functional data analysis of generalized regression quantiles. *Statistics and Computing*, **25**, 182–202.

Moors, J. (1988) A quantile alternative for kurtosis. *Royal Statistical Society*, **37**, 25–32.

Walter, S. (2011) *Defining Quantiles for Functional Data.* Ph.D. thesis, The University of Melbourne.