STATISTICS IN TRANSITION new series, March 2023
Vol. 24, No. 2, pp. 1–16, DOI 10.59170/stattrans-2023-017
Received – 01.05.2021; accepted – 27.10.2022

Outlier detection based on the functional coefficient of variation

Ipek Deveci Kocakoç[1], **Istem Köymen Keser**[2]

Abstract

The coefficient of the variation function is a useful descriptive statistic, especially when comparing the variability of more than two curve groups, even when they have significantly different mean curves. Since the coefficient of variation function is the ratio of the mean and standard deviation functions, its particular property is that it shows the acceleration more explicitly than the standard deviation function. The aim of the study is twofold: to show that the functional coefficient of variation is more sensitive to abrupt changes than the functional standard deviation and to propose the utilisation of the functional coefficient of variation as an outlier detection tool. Several simulation trials have shown that the coefficient of the variation function allows the effects of outliers to be seen explicitly.

Key words: coefficient of variation function, outlier detection, functional data analysis.

1. Introduction

The desire to interpret many curvilinear data together has increased the requirements for functional data analysis methods, and with the development of these methods, it is aimed to reveal the underlying structures of functional data called curves or surfaces. With the development of functional equivalents of multivariate statistical analysis techniques, functional data analysis (FDA) techniques have found a wide range of applications such as financial data (Wang et al., 2021), medical data (Ullah and Finch, 2013), climate change (Ghumman et al., 2020), air quality (Martinez Torres et al., 2020), management science (Dass and Shropshire, 2012), pandemics (Tang et al., 2020), etc. A detailed review of applications of functional data analysis can be found in Ullah and Finch (2013).

In functional data analysis, firstly, Ramsay (1982), Ramsay and Dalzell (1991), Rice and Silverman (1991), and Ramsay and Silverman (1997) introduced basic descriptive

[1] Corresponding Author. Econometrics Dept., Dokuz Eylul University, Izmir, Turkey. E-mail: ipek.deveci@deu.edu.tr. ORCID: https://orcid.org/0000-0001-9155-8269.

[2] Econometrics Dept., Dokuz Eylul University, Izmir, Turkey. E-mail: istem.koymen@deu.edu.tr. ORCID: https://orcid.org/0000-0003-2123-188X.

statistics such as the mean function, the standard deviation function, and variance-covariance surfaces, which form the basis of functional multivariate methods. Detailed information on descriptive statistics of functional data can be found in Shang (2015). Besides, Keser et al. (2016) have discussed the coefficient of variation (CV) function, which is the functional equivalent of the coefficient of variation. Krzysko and Smaga (2019) examined the multivariate coefficient of variation for functional data as a value, not as a function.

In this paper, our aim is to use the coefficient of variation function for detection of the effect of outliers by utilizing its sensitivity to abrupt changes in the data. The CV function is necessary to compare the variation between curve groups especially when the mean curves are different between curve groups or when we want to emphasize the main variation in time points. Suppose the height of boys and girls have different mean curves. If we want to compare the variation of height, we need the CV function. This study aims to use this feature of the CV function to detect abrupt changes caused by outliers in the data.

FDA has some methods for outlier detection which are extensions of classic statistical methods. Integrated squared error (Hyndman and Ullah, 2007), depth-based weighting and trimming (Febrero et al., 2007 and 2008), functional bagplot and functional highest density region (HDR) boxplot (Hyndman and Shang, 2010), trimmed estimators (Gervini, 2012), functional boxplot, and adjusted functional boxplot (Sun and Genton, 2011, 2012), robust functional principle component analysis procedure (Sawant et al., 2012), projection-based trimming (Fraiman and Svarc, 2013), outliergram (Arribas-Gil and Romo, 2014), and probabilistic modelling (de Pinedo et al., 2020) are prominent. Hubert et al. (2015) studied multivariate functional outlier detection. In this study, we propose the utilization of the coefficient of variation function for detection of the effect of outliers.

The paper is organized as follows: Section 2 presents the coefficient of variation function via the basis function approach. In Section 3, the results of simulation studies conducted to compare the standard deviation function and the coefficient of variation function in terms of sensitivity to abrupt changes are given. Outlier detection utilization of the CV function is explained in detail. Finally, Section 4 deals with some conclusions and suggestions.

2. Coefficient of variation function

Classical descriptive statistics for univariate data can similarly be applied to functional data with minor modifications. If the variability of multiple functional data groups needs to be compared, especially when mean functions of these data groups are

different, the coefficient of variation function can be used instead of the standard deviation function (Keser et al., 2016).

$$\text{CV function} = \frac{\text{Standard deviation function}}{mean\ function} \quad (1)$$

In comparison with other approaches, the basis function approach is mainly used in functional data analysis when estimating the curves. According to the basis function method, the estimates of mean, standard deviation, and the CV function are as follows.

Here, **B** is the (nxK) basis function matrix which consists of $B_i(t_j)$, i=1,2,…, K, j=1,2,…, n values, where t_j denotes j-th time point, K denotes the number of basis functions, and n denotes the number of curves. **C** is the variance-covariance matrix of coefficients which are obtained by the roughness penalty method or the least sum of squares method. According to the basis function approach, the variance-covariance matrix for n curves is $\mathbf{V} = \mathbf{B}*\mathbf{C}*\mathbf{B}^{\mathbf{T}}$.

i) Calculation of the coefficient vector of the standard deviation function (*std*):

As $s = sqrt(diag(\mathbf{V}))$, $s = \mathbf{B}*std$. In order to find *std,* the coefficient vector of the standard deviation function, the following calculations are carried out.

Since the **B** matrix may not be a square matrix, it is converted into a square matrix by multiplying both sides by $\mathbf{B}^{\mathbf{T}}$ because of the inverse problem.

$$\mathbf{B}^{\mathbf{T}}*s = \mathbf{B}^{\mathbf{T}}*\mathbf{B}*std$$

$\mathbf{B}^{\mathbf{T}}*\mathbf{B}$ matrix may not be invertible by the singular value decomposition. In this case, the Cholesky decomposition or Ridge regression may be used.

$$\mathbf{B}^{\mathbf{T}}*s = \mathbf{D}$$

$$\mathbf{B}^{\mathbf{T}}*\mathbf{B} = \mathbf{E}$$

While **R** is an upper triangular matrix, according to the Cholesky decomposition:

$$\mathbf{E} = \mathbf{R}^{\mathbf{T}}*\mathbf{R}.$$

$$\mathbf{D} = \mathbf{R}^{\mathbf{T}}*\mathbf{R}*std$$

$$(\mathbf{R}^{-1})^{\mathbf{T}}\mathbf{D} = (\mathbf{R}^{-1})^{\mathbf{T}}*\mathbf{R}^{\mathbf{T}}*\mathbf{R}*std$$

$$(\mathbf{R}^{-1})*(\mathbf{R}^{-1})^{\mathbf{T}}*\mathbf{D} = (\mathbf{R}^{-1})*(\mathbf{R}^{-1})^{\mathbf{T}}*\mathbf{R}^{\mathbf{T}}*\mathbf{R}*std$$

So, the coefficient vector of the standard deviation function is

$$std = (\mathbf{R}^{-1})*(\mathbf{R}^{-1})^{\mathbf{T}}*\mathbf{D}.$$

ii) Calculation of the coefficient vector of the mean function $\bar{c}$:

While $\bar{y}$ is the mean coordinate vector, $\bar{\boldsymbol{c}}$ may be obtained in a similar way as the *std*:

$$\bar{\boldsymbol{c}} = (\mathbf{B}^{\mathbf{T}}*\mathbf{B})*\mathbf{B}^{\mathbf{T}}*\bar{y}.$$

iii) Calculation of the coefficient vector of the CV function:

We propose that the coefficients of the CV function be calculated as:

$$CV = \frac{\mathbf{B}*std}{\mathbf{B}*\bar{\boldsymbol{c}}}$$

As denoted by Keser et al. (2016), especially when the curves have a mean function very close to zero, the CV function gets affected. This should be considered as a drawback of the statistic itself and exists in the original CV concept.

3. Simulation study

Since this CV function is a ratio of two functions, it is affected more by consecutive abrupt changes and can easily detect time points where essential changes of data occur. It is also proven by simulation studies in the next section that the CV function and its derivatives are as efficient as the standard deviation for detecting significant changes between time points for data with and without outliers.

In this study, we propose that CV functions may also be promoted as an outlier detection tool. Inspecting CV functions by using one-curve-out method may especially be used for this purpose. In order to show this usage of CV functions, outliers are investigated by CV functions and also by adjusted functional boxplots and adjusted outliergrams simultaneously.

3.1. The behaviour of CV function in abrupt changes

In this section simulation studies are conducted in order to compare the CV function with the standard deviation function for the detection of abrupt changes in functional data sets. For this purpose, n curves with 100 discrete points are generated by the main model $X(t) = sin4\pi t + \xi(t)$, t ∈ [0, 1], where $\xi(t)$ is normally distributed with 0 mean and some covariance structure between [0.2,0.8]). All analyses are done for n=50, 100 and 150 data sets. Since the size of the data set did not change the results, for the sake of easy interpretation and graphical readability, only n=50 case is reported here.

Computed descriptive statistics for the data set are given in Figure 1a and Figure 1b. The green lines show 50 curves, the red line shows the mean curve, the dotted line shows the standard deviation curve and the blue line shows the CV curve. Since the scales are not very compatible, presenting two different figures is preferred. The peaks on the CV function show the points where abrupt changes occur, which cannot be easily determined by the standard deviation function.

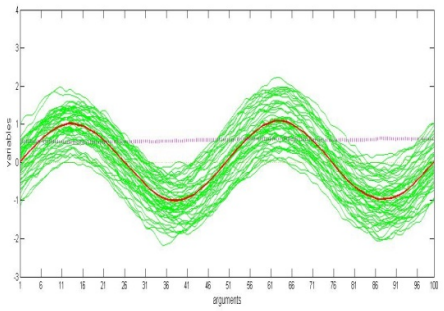


Figure 1a: Sample Data, mean and standard deviation function

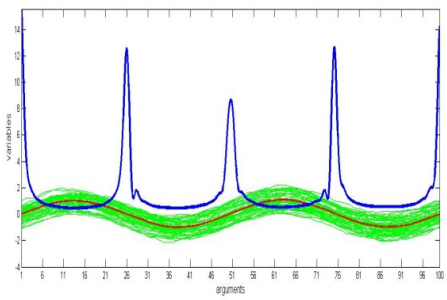


Figure 1b: Sample Data, mean and CV function

The time points in which the variability increases or decreases are not easily detectable in Figure 1a. Since CV is the proportion of two values, abrupt changes can be more easily detected especially when the range of standard deviation and mean functions are small. Here, the focus is not the value of the function at that point, but the abruptness of the situation. A partial view of CV values for one of the data sets is given in Table 1. It can be seen that there are abrupt changes for knots 1, 25, 26, 75,76, 99, and 100, which all have small changes in either their means or standard deviations, but distinctive CV values, as also can be seen from Figure 1b.

Table 1: A partial view of CV values for one of the data sets

Knots (Time points)	Std. Deviation coefficient	Mean coefficient	CV coefficient
1	**0.4712**	**-0.0303**	**15.5499**
2	**0.4744**	**0.0951**	**4.9908**
3	0.4735	0.2173	2.1794
.			
.			
.			
24	0.4606	0.1987	2.3177
25	**0.4652**	**0.0748**	**6.2150**
26	**0.4602**	**-0.0369**	**12.4740**
27	**0.4568**	**-0.1611**	**2.8350**
28	0.4675	-0.2741	1.7055
29	0.4826	-0.3855	1.2520
.			
.			
.			
73	0.5155	0.3025	1.7042
74	0.5128	0.1784	2.8746
75	**0.5036**	**0.0397**	**12.6714**
76	**0.5064**	**-0.1005**	**5.0384**
77	0.5115	-0.2339	2.1867
78	0.5221	-0.3560	1.4668
79	0.5145	-0.4744	1.0846
.			
.			
.			
97	0.5537	-0.3924	1.4109
98	**0.5501**	**-0.2702**	**2.0363**
99	**0.5542**	**-0.1529**	**3.6237**
100	**0.5463**	**-0.0383**	**14.2749**

Here, the CV function is apparently helpful for detecting abrupt changes in data. The same results are valid for data sets with outliers as can be seen in the next section.

3.2. The behaviour of CV function for data with outliers

There are two types of outliers in functional data analysis: magnitude outliers and shape outliers. In general, magnitude outliers appear far apart from other curves, and shape outliers have a distinct pattern from other curves. In this study, contamination models for outliers are chosen from Arribas-Gil and Romo's (2014) R file from their supplementary materials. In order to examine the behaviour of CV functions for data sets with and without outliers, outlier curves are generated with contamination models given below. Outlier curves are entered as the 51st curve after 50 non-outlier curves (generated from the main model) in the data set.

The contamination model for shape outlier (assuming that $\xi(t)$ is standard normally distributed) is:

$$X(t) = cos(4\pi t - 0.25) + 0.05\,\xi(t)\,.$$

The contamination model for magnitude outlier (assuming that $\xi(t)$ is standard normally distributed) is:

$$X(t) = sin(4\pi t) + 2 + 0.05\,\xi(t)$$

A representation of one sample data set from one of the trials is given in Figure 2 with 50 curves of non-outlier data, one shape, and one magnitude outlier.

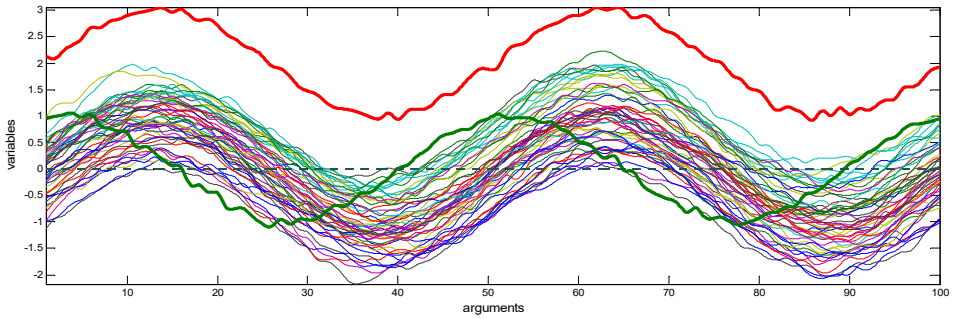


Figure 2: Data curves with shape and magnitude outliers
(Green bold line: shape outlier, Red bold line: magnitude outlier)

In our study, along with CV, Sun and Genton's (2011) adjusted functional boxplot and Arribas_Gill and Romo's (2014) adjusted outliergram are utilized simultaneously to detect outliers for practical purposes and reconfirmation. Graphs of those detection methods for the sample data set in Figure 2 are given in Figure 3.

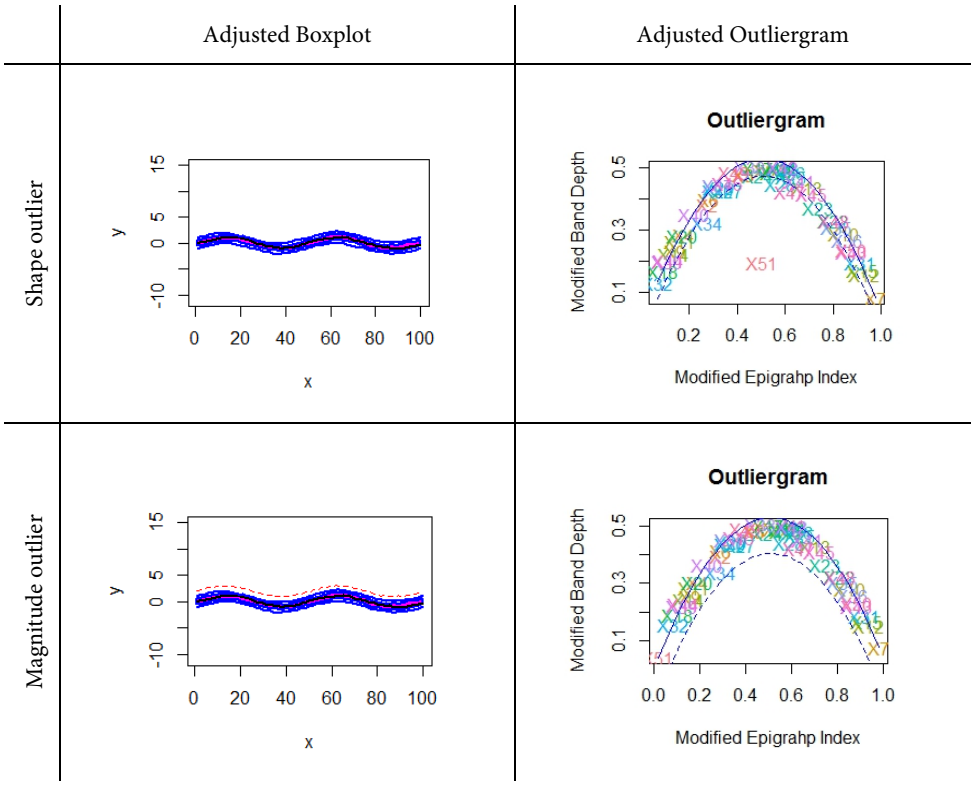


Figure 3: Outlier Detection by Adjusted Boxplot and Adjusted Outliergram

The functional adjusted boxplot is designed to detect outliers of magnitude, while the adjusted outliergram is designed to detect outliers of shape that are more difficult to find.

The descriptive statistics are the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope for the adjusted functional boxplot (Sun and Genton, 2011), based on the centre outward ordering induced by band depth for functional data. By inflating the internal region (the envelope), the outer region (the fence) is obtained. Any curve that crosses the fences is depicted as possible outliers. For our sample data set, as expected, the adjusted functional boxplot easily detects magnitude outliers (dotted curve in the bottom left corner in Figure 3), however, shape outlier is not detected (top left corner in Figure 3).

Adjusted outliergram gives the boundaries for dashed parabola for the detection of shape outliers. Any value outside these boundaries is determined as an outlier. The further the curves in the sample are identical and straight, the nearer the points in the outliergram to the dashed parabola. On the other hand, the noisier the curves and the large number of crossing points between them, the more dispersed the points in the

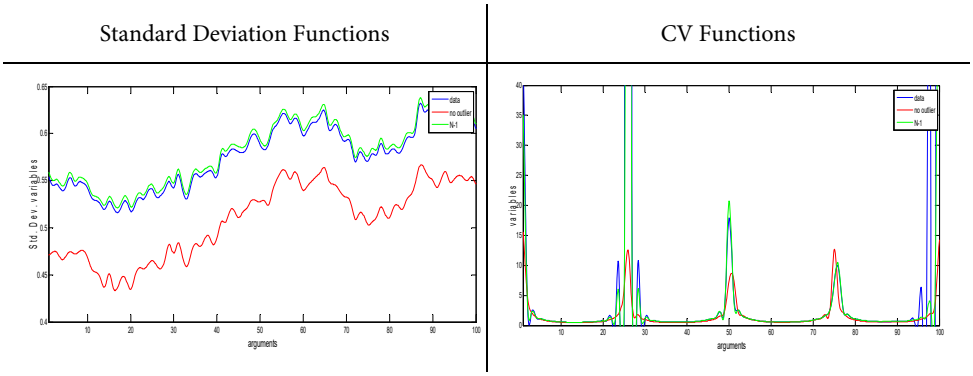
outliergram. The points with the largest distances to the parabola represent the most outlying curves, in terms of shape, of the sample (Arrabas Gil, Romo, 2014). For our sample data set, the 51st curve entered as the shape outlier lies out and far from the boundaries (top right corner in Figure 3) whereas the 51st curve entered as the magnitude outlier lies inside the parabola but far away from the other curves (bottom right corner in Figure 3).

Since the FDA is a visual method, many random data sets are generated from the models but, for clarity and conciseness, only three are randomly selected from the models in the study. Yet, it should be noted that all other data sets have similar results and may be provided by the authors. Changes in both standard deviation functions and CV functions are examined in every generated data set for the following three cases for both shape and magnitude outliers and the results are presented in Figure 4 and Figure 5 for those three selected data sets. Outliers are included separately in the data sets in order to avoid masking effect over each other.

- Case 1: Data with one outlier – (total: N = 51 curves, 51st curve is the outlier)
- Case 2: Data with one outlier when one of the non-outlier curves is excluded – (total: N-1 = 50 curves)
- Case 3: Data with no outliers – (total: N-1 = 50 curves)

In Case 2, every non-outlier curve is excluded one at a time. Since there are 50 curves, the computations are made 50 times for each data set and similar conclusions are made, so only one of the results is reported here.

Figure 4 shows results for data sets with one magnitude outlier and Figure 5 shows results for data sets with one shape outlier. In both figures, the left panel of each row shows the standard deviation function for three randomly selected cases while the right panel shows the proposed the coefficient of variation function. The same analysis is conducted also with normalized data but no difference is found and therefore not reported here. Absolute values of CV are used for reflecting the variability better and easier comparison with standard deviations.



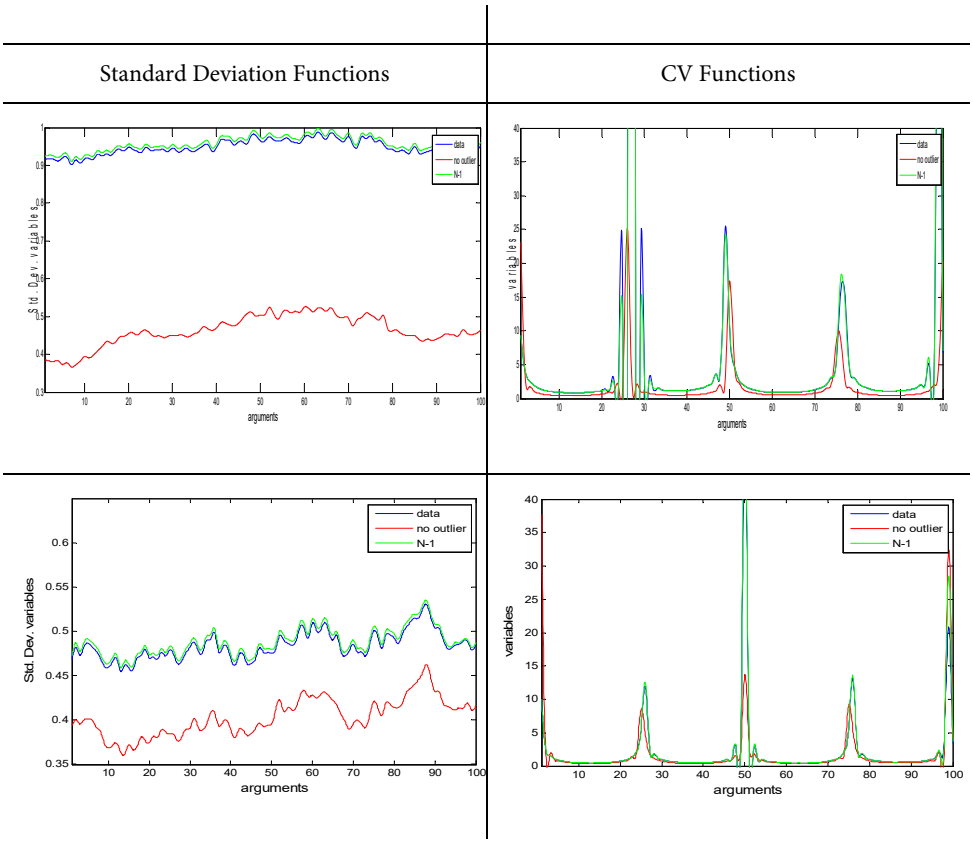
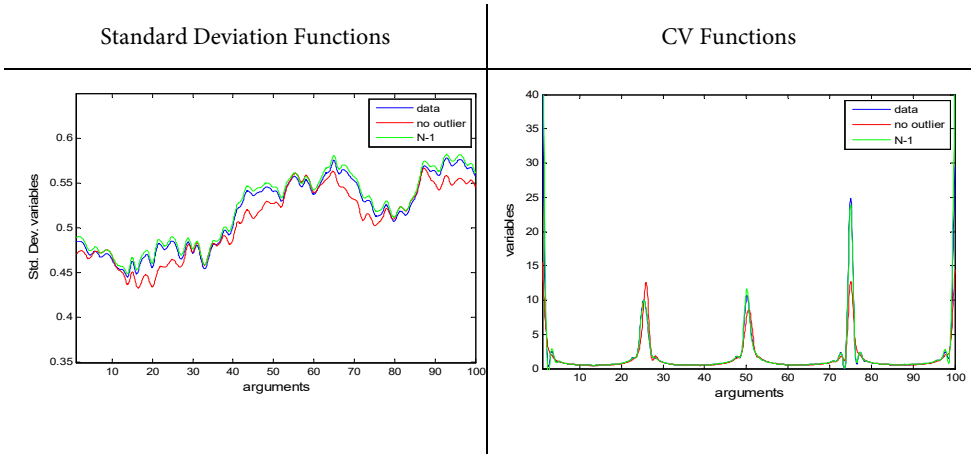


Figure 4: Standard Deviation and CV Functions for Magnitude Outlier for Randomly Selected Cases



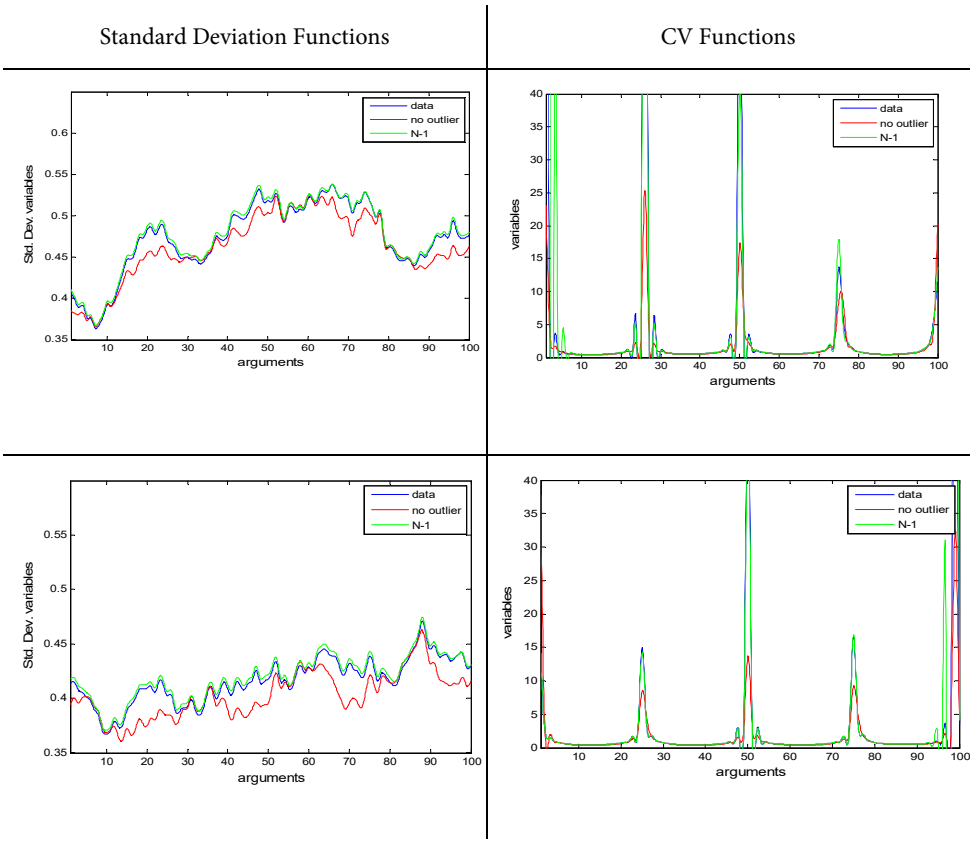


Figure 5: Standard Deviation and CV Functions for Shape Outlier

When examining standard deviation functions for all data sets, it can be seen that the standard deviation function of data with outlier lies distant from the standard deviation function of non-outlier data, as expected.

When examining CV functions for all data sets, it can mostly be seen that the CV function of data with outliers lies distant from the CV function of data without outliers similar to the results for the standard deviation function. When any one of the non-outlier curves are excluded (Case 2), both standard deviation and CV functions have very close results to data with outliers. This may be a supporting result that outlier curves affect standard deviation and CV functions.

Since cubic B-splines are used to obtain the standard deviation and CV functions, the first and second derivative functions of these functions are also continuous. Therefore, the movements of the curves can easily be examined. Interpretations of derivative functions rather than the functions themselves may provide a stronger inference. Besides, utilizing derivative functions when comparing the curves makes them more comparable with respect to the origin. The first and second derivative

functions of standard deviation and CV functions for magnitude outliers are given in Figures 6 and 7 respectively. The first derivative function shows the velocity while the second one shows the acceleration.

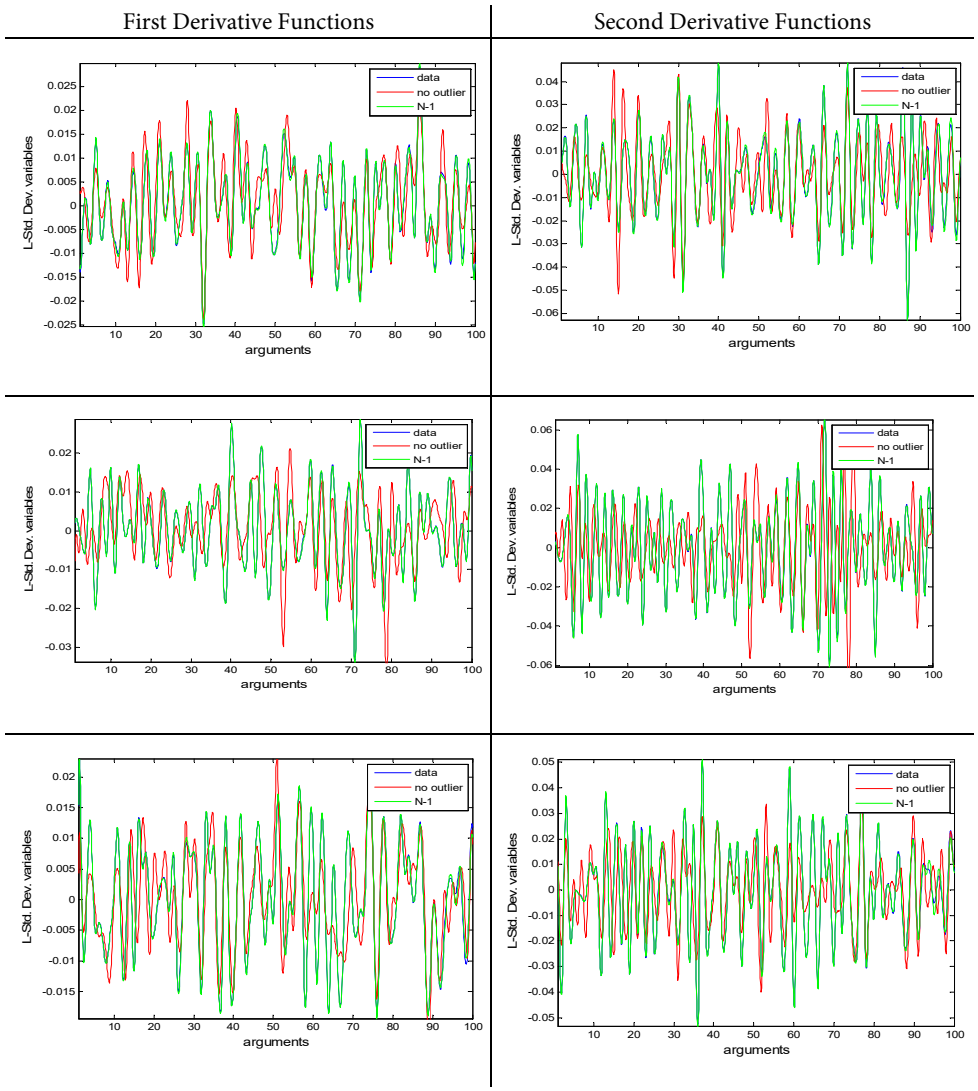


Figure 6: Derivative functions of standard deviation function for magnitude outlier

When Figure 6 is examined, data with outliers and data with (N-1) curves show a very similar, even overlapping, behaviour for both the first and second derivative functions while non-outlier data lies distantly and with shifts. By utilizing derivative functions, the dimensions of ups and downs have become more comparable.

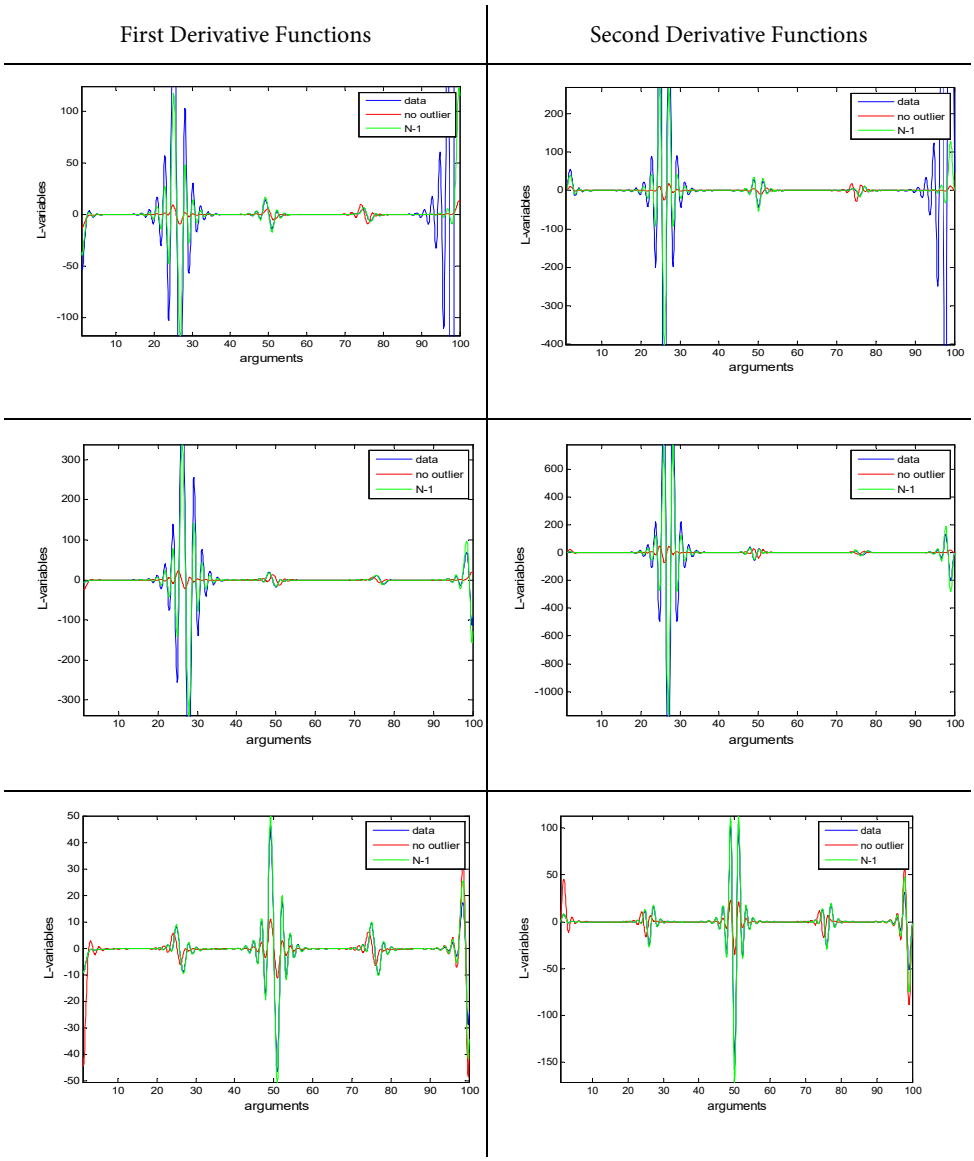


Figure 7: Derivative functions of CV function for magnitude outlier

When Figure 7 is examined, it can be seen that CV especially emphasizes time points in which abrupt changes appear. As in Figure 6, both derivative functions for non-outlier data lie distant from the other two data sets with outliers. Thus, it can be said that derivative functions have similar behaviour as the original curve functions. However, the derivative functions of the CV function do not get lost in small changes and can focus on abrupt changes better than that of the standard deviation function. The standard deviation function and its derivatives are strongly affected by the smallest

changes due to the effect of the mean. Examination of derivative functions enables better comparisons for ups and downs in all curves and even small changes can be determined by their help.

The behaviour of the CV function for data sets with and without outliers leads way to the one-out method as an outlier detection tool by itself. Since now we know that outliers affect the size of peaks in CV functions, any outlier curve can be detected by its different size of the CV function peaks. In order to show that, we examined 51 CV functions obtained by excluding one curve at a time and saw that the 51st curve has a very different peak structure than the others (especially for magnitude outlier), concluding that this curve may be an outlier.

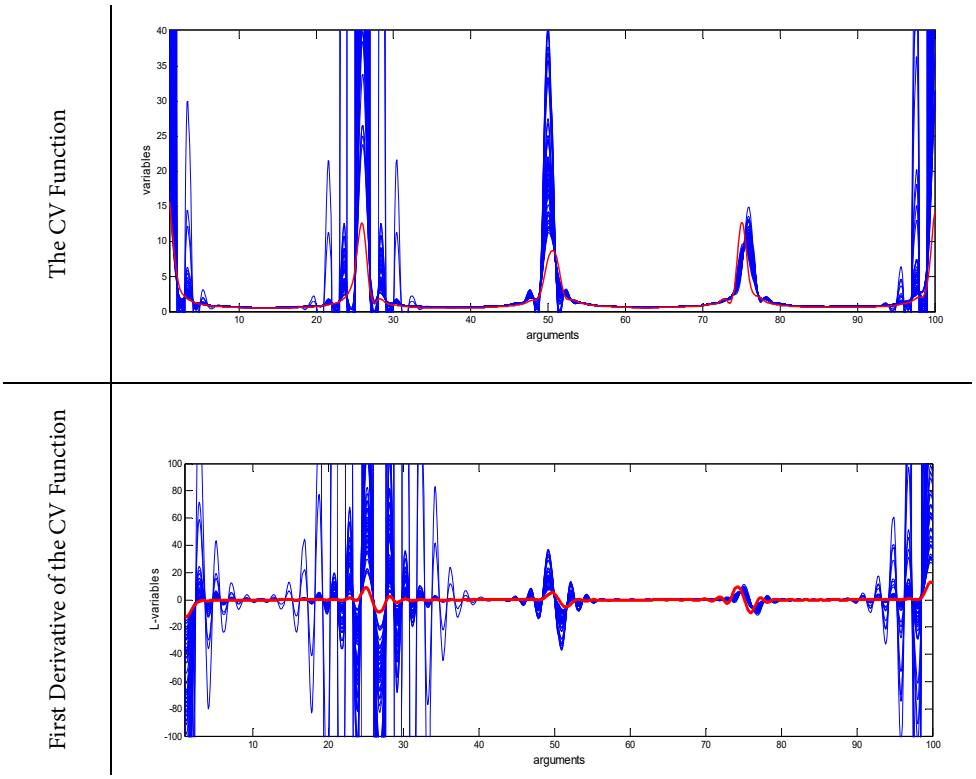


Figure 8: CV function and its first derivative for the one-out method

In Figure 8, the blue lines show the 50 curves which include the outlier, and the red curve shows the function when the outlier curve is excluded. Here, we also validate and confirm the outliers that we found by adjusted outliergram and adjusted functional boxplot. Therefore, the CV function can be utilized as a visual outlier curve detection tool.

4. Conclusions

The coefficient of variation function is proposed as an outlier identification method in this study. The CV function is a better descriptive statistics for determining abrupt changes than standard deviation. The availability of the first and second derivatives of the CV function also strengthens is utilization. In the case of outliers in the data set, it is also proven to be a useful statistic. By using the one-out method, outlier curves can easily be detected among others. Therefore, the CV function may be utilized in outlier detection as a confirmatory and complementary method to different outlier detection methods such as outliergram and functional boxplot.

More automated and easy detection of points with abrupt changes and curves which are outliers may be developed as a further study. Confidence intervals or probabilities for this detection may also be investigated.

References

Arribas-Gil, A., Romo, J., (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15, pp. 603–619.

de Pinedo, Á. R., Couplet, M., Marie, N., Marrel, A., Merle-Lucotte, E., and Sueur, R., (2020). Functional outlier detection through probabilistic modelling. In: Aneiros G., Horová I., Hušková M., Vieu P. (eds) *Functional and high-dimensional statistics and related fields*. IWFOS 2020. Contributions to Statistics. Springer, Cham. https://doi.org/10.1007/978-3-030-47756-1_30.

Dass, M., Shropshire, C., (2012). Introducing functional data analysis to managerial science. *Organizational Research Methods*, 15(4), pp. 693-721.

Febrero, M., Galeano, P., and Gonzalez-Manteiga, W., (2007). A functional analysis of NOx levels: Location and scale estimation and outlier detection. *Computational Statistics,* 22(3), pp. 411–427.

Febrero, M., Galeano, P., and Gonzalez-Manteiga, W., (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NO_X levels. *Environmetrics*, 19(4), pp. 331–345.

Fraiman, R., Svarc, M., (2013). Resistant estimates for high dimensional and functional databased on random projections. *Computational Statistics & Data Analysis*, 58, pp. 326–338.

Gervini, D., (2012). Outlier detection and trimmed estimation for general functional data. *Statistica Sinica*, 22(4), pp. 1639–1660.

Ghumman, A. R., Alodah, A., Haider, H., and Shafiquzzaman, M., (2020). Evaluating the impact of climate change on stream flow: integrating GCM, hydraulic modelling and functional data analysis. *Arabian Journal of Geosciences*, 13(17), pp. 1–15.

Hubert, M., Rousseeuw, P. J., and Segaert, P., (2015). Multivariate functional outlier detection. *Statistical Methods* & *Applications*, 24(2), pp. 177–202.

Hyndman, R. J., Ullah, M. S., (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics* & *Data Analysis*, 51(10), pp. 4942–4956.

Hyndman,R. J., Shang, H. L., (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19, pp. 29–49.

Krzysko, M., Smaga L., (2019). A multivariate coefficient of variation for functional data, *Statistics and Its Interface*, 12(4), pp. 647–658.

Keser, İ. K., Kocakoç, İ. D., and Şehirlioğlu, A. K., (2016). A new descriptive statistic for functional data: functional coefficient of variation. *Alphanumeric Journal*, 4(2), pp. 1–10.

Martínez Torres, J., Pastor Pérez, J., Sancho Val, J., McNabola, A., Martínez Comesaña, M., and Gallagher, J., (2020). A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. *Mathematics*, 8(2), p. 225.

Ramsay J. O., (1982). When the data are functions. *Psychometrika*, 47, pp. 379–396.

Ramsay, J. O., Dalzell, C. J., (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3), pp. 539–561.

Ramsay J. O, Silverman B. W., (1997). *Functional data analysis*. Springer-Verlag, New-York.

Rice John A., Silverman B. W., (1991). Estimating the mean and covariance structure when the data are curves. *Journal of the Royal Statistical Society,* Series B., Vol. 53, No.1, pp. 233–243.

Ullah S., Finch C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(43), pp. 539–572.

Sawant, P., Billor, N., and Shin H., (2012). Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27, pp. 83–102.

Shang, H. L., (2015). Resampling techniques for estimating the distribution of descriptive statistics of functional data. *Communications in Statistics – Simulation and Computation*, 44(3), pp. 614–635, doi: 10.1080/03610918.2013.788703.

Sun Y., Genton M. G., (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20, pp. 316–334.

Sun, Y., Genton M. G., (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, 23(1), pp. 54–64.

Tang, C., Wang, T., and Zhang, P., (2020). Functional data analysis: An application to COVID-19 data in the United States, *arXiv preprint arXiv:2009.08363*.

Wang, D., Li, X., Tian, S., He, L., Xu, Y., and Wang, X., (2021). Quantifying the dynamics between environmental information disclosure and firms' financial performance using functional data analysis. *Sustainable Production and Consumption*, 28, pp. 192–205.