Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

Functional boxplots for multivariate curves

Wenlin Dai and Marc G. Genton*

Received 21 March 2018; Accepted 23 May 2018

A two-stage functional boxplot is introduced for the visualization and exploratory data analysis of multivariate curves. Specifically, the original functional boxplot is combined with an outlier-detection procedure on the basis of the functional directional outlyingness, which accounts for both the magnitude and shape outlyingness of functional data. This combination is robust to various types of outliers and, hence, captures the data structures more accurately than does the functional boxplot alone. It also allows for both marginal and joint analysis of the multivariate curves. We apply the proposed tool to Spanish weather data in an illustrative example. © **2018 John Wiley & Sons, Ltd.**

Keywords: directional outlyingness; functional boxplot; multivariate functional data; outlier detection; visualization; visuanimation

1 Introduction

Data visualization is a necessary complement to statistical analysis that intuitively demonstrates the features of a data set. One popularly implemented graphical tool is the univariate boxplot proposed by Tukey (1975). Bivariate extensions of boxplots were investigated by Goldberg & Iglewicz (1992) and Rousseeuw et al. (1999). Catering to the demand of exploratory analysis for functional data, which are frequently recorded owing to the evolution of technology, Sun & Genton (2011) proposed the functional boxplot as an analogue.

The functional boxplot of Sun & Genton (2011) is constructed by ordering a group of univariate curves from the centre outward according to the modified band depth (MBD) (López-Pintado & Romo, 2009) or any other user-provided functional ranking. Specifically, the envelope of the 50% deepest curves forms the 50% central region; by inflating this region by 1.5 times its vertical range, one can obtain two fences to detect outliers. Eventually, the envelope of the central 50% region, the median curve and the maximum non-outlying envelope are demonstrated as descriptive statistics; detected outlying curves are also visualized. Besides, Sun & Genton (2012) provided an adaptive way to determine the inflating factor by accounting for the dependence structure of the functional data. Popularly used functional depths include, for instance, the band depth and the MBD (López-Pintado & Romo, 2009), the spatial depth (Chakraborty & Chaudhuri, 2014) and the extremal depth (Narisetty & Nair, 2016; Myllymäki et al., 2017). Functional boxplots constructed with other types of functional depths were investigated by Martin-Barragan et al. (2016), Narisetty & Nair (2016) and Serfling & Wijesuriya (2017).

The functional boxplot of Sun & Genton (2011) flags the curves that cross its fences as outliers. As a result, it cannot handle some types of outliers well, for example, isolated outliers, shape outliers and covariance outliers, according to the taxonomy of functional outliers by Hubert et al. (2015). Hence, the shape of the resultant central region and fences may be deformed and fail to accurately capture the general structure of the underlying data set.

Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
***Email: marc.genton@kaust.edu.sa**

© 2018 John Wiley & Sons, Ltd

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

The ISI's Journal for the Rapid Dissemination of Statistics Research

To address these drawbacks, we eliminate the negative effects of outliers by combining a functional boxplot with an outlier-detection procedure based on functional directional outlyingness (Dai & Genton, 2018a,2018b). The functional directional outlyingness effectively measures the centrality of multivariate curves, allowing us to generalize the functional boxplots to multivariate functional data.

The remainder of this paper is organized as follows. The outlier-detection procedure is described in Section 2. The two-stage functional boxplot is presented for both univariate and multivariate curves in Section 3. An application to Spanish weather records is illustrated in Section 4. The paper ends with a discussion in Section 5.

2 Outlier-detection procedure

Functional directional outlyingness (Dai & Genton, 2018a) is a measure that accounts for the direction of an underlying observation's point-wise deviation from the bulk of the data, thereby revealing both the magnitude and the shape of that observation's outlyingness. More specifically, Dai & Genton (2018a) defined directional outlyingness for point-wise data as

$$\mathbf{O}(\mathbf{Y}, F_{\mathbf{Y}}) = \{1/d(\mathbf{Y}, F_{\mathbf{Y}}) - 1\} \cdot \mathbf{v}, \quad d(\mathbf{Y}, F_{\mathbf{Y}}) > 0,$$

where $F_{\mathbf{Y}}$ denotes the distribution of a random variable $\mathbf{Y}$, d is a conventional depth notion and $\mathbf{v}$ is the unit vector pointing from the median of $F_{\mathbf{Y}}$ to $\mathbf{Y}$. Dai & Genton (2018a) also defined two quantities to measure the magnitude and shape outlyingness of a curve:

$$\mathbf{MO}(\mathbf{X}, F_{\mathbf{X}}) = \int_{\mathcal{I}} \mathbf{O}(t)\, dt \quad \text{and} \quad \mathrm{VO}(\mathbf{X}, F_{\mathbf{X}}) = \int_{\mathcal{I}} \{\mathbf{O}(t) - \mathbf{MO}\}^{\mathsf{T}} \{\mathbf{O}(t) - \mathbf{MO}\}\, dt,$$

where $\mathbf{X}$ is a p-dimensional functional random vector defined on the interval $\mathcal{I}$ and $F_{\mathbf{X}}$ denotes the distribution of $\mathbf{X}$. $\mathbf{O}(t) = \mathbf{O}(\mathbf{X}(t), F_{\mathbf{X}(t)})$, $\mathbf{X}(t)$ is a p-dimensional random vector, and $F_{\mathbf{X}(t)}$ denotes its corresponding distribution.

A curve $\mathbf{X_0}$ is flagged as an outlier if its corresponding $(\mathbf{MO}_0^{\mathsf{T}}, \mathrm{VO}_0)^{\mathsf{T}}$ is detected as outlying with respect to the population distribution of $(\mathbf{MO}^{\mathsf{T}}, \mathrm{VO})^{\mathsf{T}}$. Specifically, Dai & Genton (2018a) showed that the empirical version of $(\mathbf{MO}^{\mathsf{T}}, \mathrm{VO})^{\mathsf{T}}$ can be well approximated by a $(p+1)$-dimensional normal distribution. A robust Mahalanobis distance (RMD) is calculated for each pair of $(\mathbf{MO}^{\mathsf{T}}, \mathrm{VO})^{\mathsf{T}}$, and the covariance matrix is estimated by the minimum covariance determinant estimator (Rousseeuw, 1985). The right tail distribution of RMD^2 is approximated by Fisher's F distribution, F_{RMD} (Hardin & Rocke, 2005). Then, $(\mathbf{MO}_0^{\mathsf{T}}, \mathrm{VO}_0)^{\mathsf{T}}$ is recognized as an outlier when

$$\mathrm{RMD}^2_{(\mathbf{MO}_0^{\mathsf{T}}, \mathrm{VO}_0)^{\mathsf{T}}} \geq C_{F_{\mathrm{RMD}}, \alpha}, \tag{1}$$

where $C_{F_{\mathrm{RMD}}, \alpha}$ is the $1-\alpha$ quantile of F_{RMD} and α is the significance level.

3 Two-stage functional boxplot

The aforementioned outlier-detection method is effective for various types of functional outliers. Thus, we improve the robustness of the functional boxplot to abnormal curves by combining it with the outlier-detection criterion (1) in a two-stage procedure. For a set of univariate curves X_k, $k \in \mathcal{S}$, the procedure is described as follows:

S1	Obtain indexes of abnormal curves, $\mathcal{S}_O$, with the outlier-detection criterion (1).
S2	Apply the functional boxplot to the remaining non-outlying curves, X_k, $k \in \mathcal{S} \backslash \mathcal{S}_O$, and add the detected outlying curves to the functional boxplot.

© 2018 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

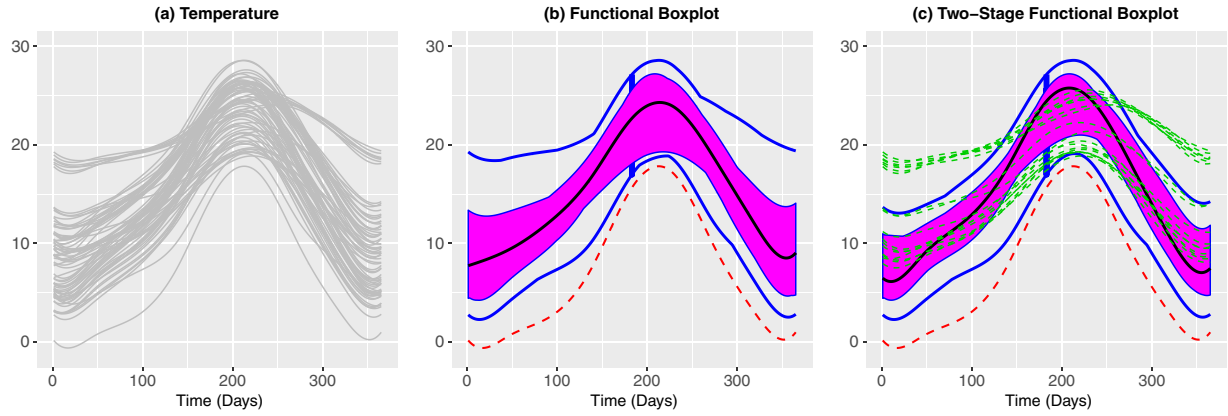


Figure 1. (a) Smoothed average annual temperatures curves. (b) Conventional functional boxplot: The red dashed curve denotes one detected outlier. (c) Two-stage functional boxplot generated by our proposed procedure: The green dashed curves denote outliers detected by criterion (1) in step S1, and the red dashed curve denotes one outlier detected by the functional boxplot in step S2.

A typical example is illustrated in Figure 1, which utilizes the averaged annual temperature data recorded at 73 weather stations in Spain during the period 1980–2009, from the R package *fda.usc*. The raw data were discretely recorded and preprocessed with the smoothing spline method; see Figure 1(a). The conventional functional boxplot based on the MBD for this data set is presented in Figure 1(b), and our proposed two-stage functional boxplot is presented in Figure 1(c). We compare the two types of functional boxplots in terms of both their outlier-detection and description of the data structure. The conventional functional boxplot diagnoses only one shifted outlier (the red dashed curve), which is parallel to the median curve but shifted downward. The two-stage functional boxplot detects not only the shifted outlier but also a group of shape outliers (the dashed green curves). These shape outliers possess moderate magnitudes on average but shaped differently from the bulk of the data. In visualizing the data structure, the two plots produce the same lower fence, but the conventional one provides a higher upper fence for all seasons except summer owing to the undetected shape outliers. The central regions also reveal similar differences between the two plots.

Inspired by the idea of visuanimation proposed by Genton et al. (2015), we provide in Movie 1 a comparison of the two-stage functional boxplot with the conventional functional boxplots constructed with rankings from different functional depth notions, for example, the MBD, the integral depth (FM, Fraiman & Muniz (2001)), the mode depth (Cuevas et al., 2006), the L^∞ depth (Long & Huang, 2015), the random projection depth (Cuevas et al., 2007) and the extremal depth (Narisetty & Nair, 2016; Myllymäki et al., 2017). Six outliers (in colour) of different types are added to 100 non-outlying curves (grey), with their level of outlyingness increasing from frame 1 to frame 40. None of the six conventional functional boxplots successfully recognize the pure shape outlier (blue) or the isolated outlier (green), which are quite well handled by the two-stage functional boxplot. This suggests that the proposed two-stage functional boxplot is more robust against outliers and, thus, more reliable for describing data structures.

4 Simulations

4.1 Univariate curves

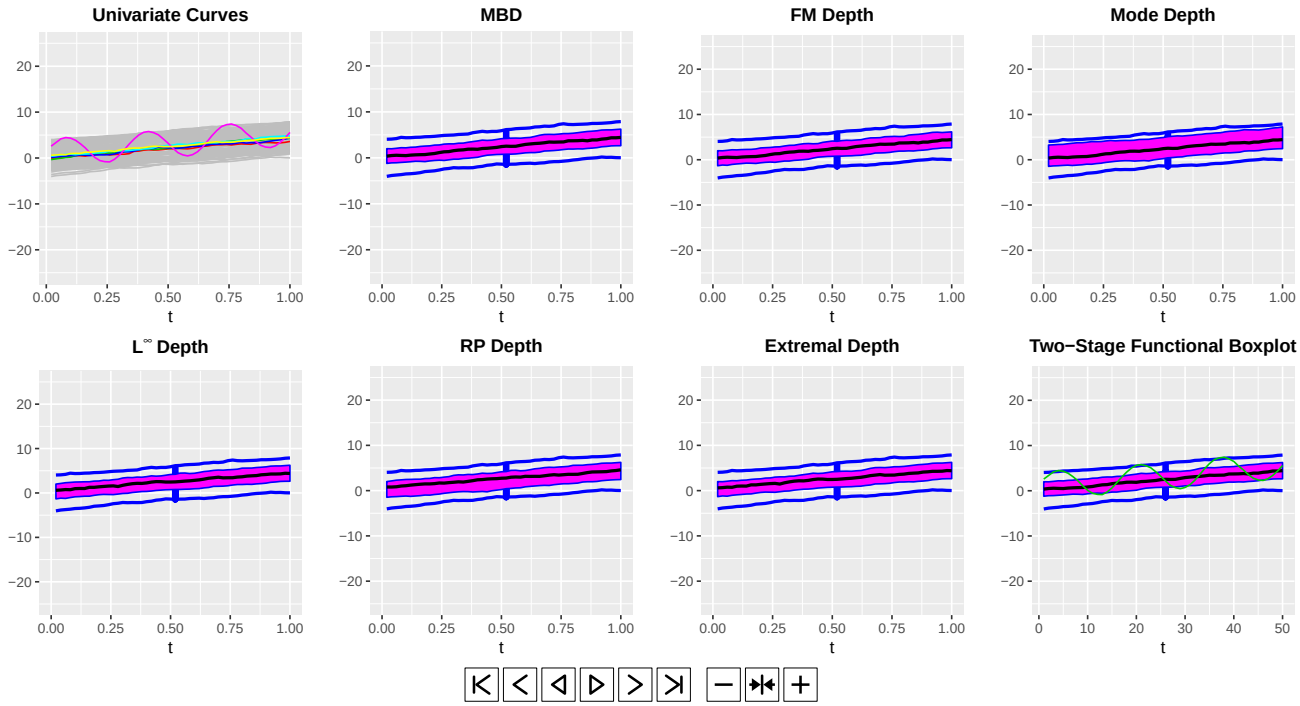
We comprehensively demonstrate the advantages of the two-stage functional boxplot in Movie 2, which visualizes the behaviour of conventional and two-stage functional boxplots contaminated by different types of outliers. Specifically, we consider the following settings:

 © 2018 John Wiley & Sons, Ltd

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

The ISI's Journal for the Rapid Dissemination of Statistics Research



Movie 1. First plot: 100 non-outlying curves (grey) and six different types of outliers (in colour). Second to seventh plots: functional boxplots based on MBD, FM depth, mode depth, L^∞ depth, RP depth and extremal depth, respectively. Last plot: two-stage functional boxplot [To ensure that the visuanimation will play properly, use Adobe Acrobat Reader to view the pdf file.]

- Model 0 (main model). $X(t) = 4t + e(t)$, where $e(t)$ is a Gaussian process with zero mean and the covariance function $\gamma(s,t) = \exp\{-|t-s|\}$. We generated 100 samples from the main model as the bulk of the data; the contaminated models are described subsequently in Models 1–3.
- Model 1 (rotation outlier; first row of Movie 2). Three rotation outliers are introduced for each frame, and there are 40 frames in total. The outliers are generated by rotating the non-outlying curves around a fixed point.
- Model 2 (isolated outlier; second row of Movie 2). One isolated outlier is introduced for each frame, and there are 80 frames in total. The outlier is generated by partially shifting one non-outlying curve with different outlying intervals and outlying levels.
- Model 3 (covariance outlier; third row of Movie 2). Two covariance outliers are introduced for each frame, and there are 40 frames in total. The outliers are generated by increasing the oscillation level of the covariance function: $\tilde{\gamma}(s,t) = C_\sigma \exp\{-|t-s|\}$, with C_σ ranging from 0 to 5.

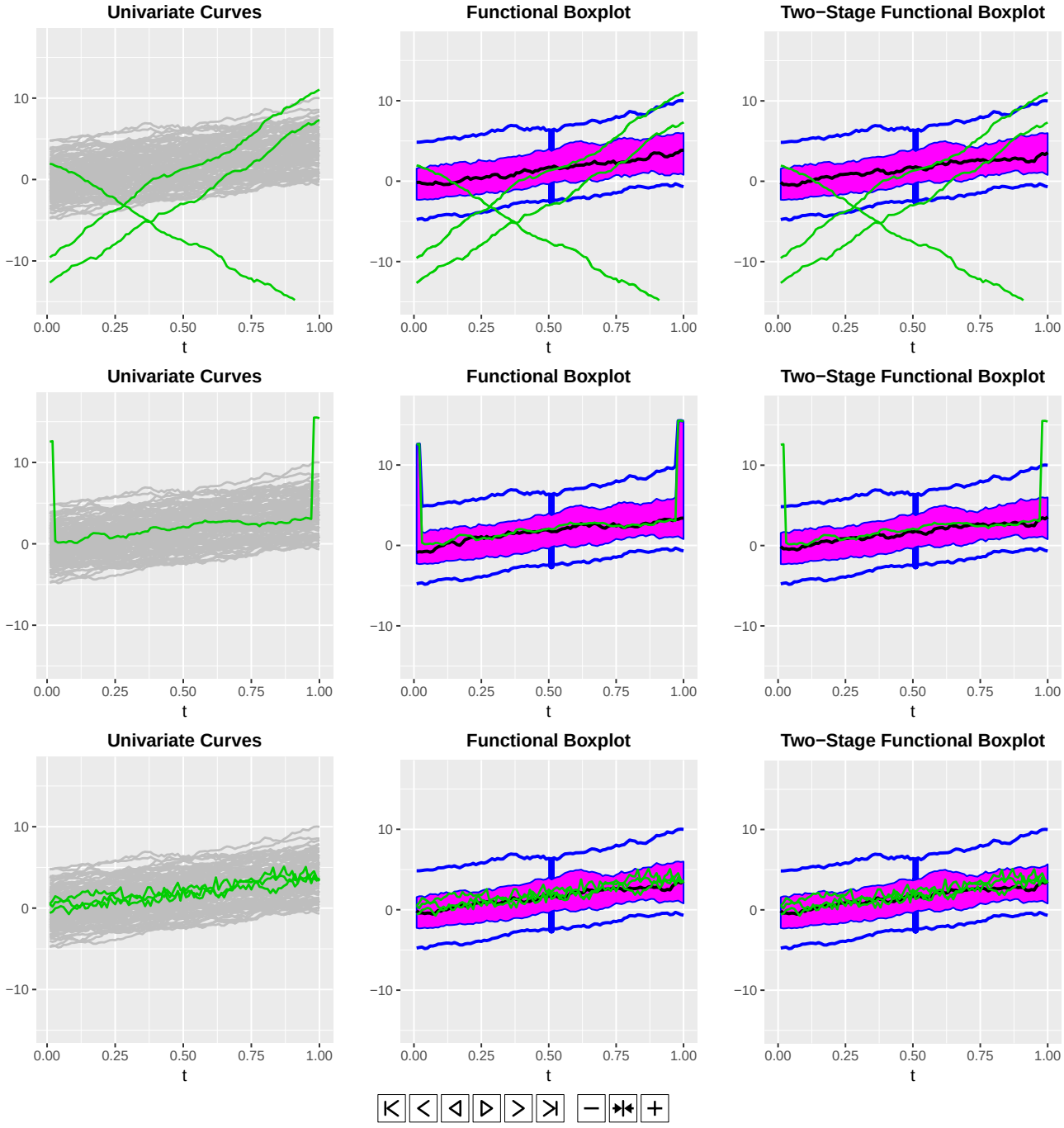
Movie 2 visualizes only the central region and fences of functional boxplots without detected outliers. Additionally, the manually introduced outliers are plotted as well. The purpose is to illustrate the robustness of the two methods to various types of outliers. Apparently, the two-stage functional boxplots eliminate the influence of outliers more effectively; hence, the resulting data structure is more accurate.

4.2 Multivariate curves

Besides univariate curves, the two-stage functional boxplot is also applicable to multivariate curves. In the first step (S1) of a multivariate case, we provide two options for detecting outliers: Treat each dimension either separately

© 2018 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190



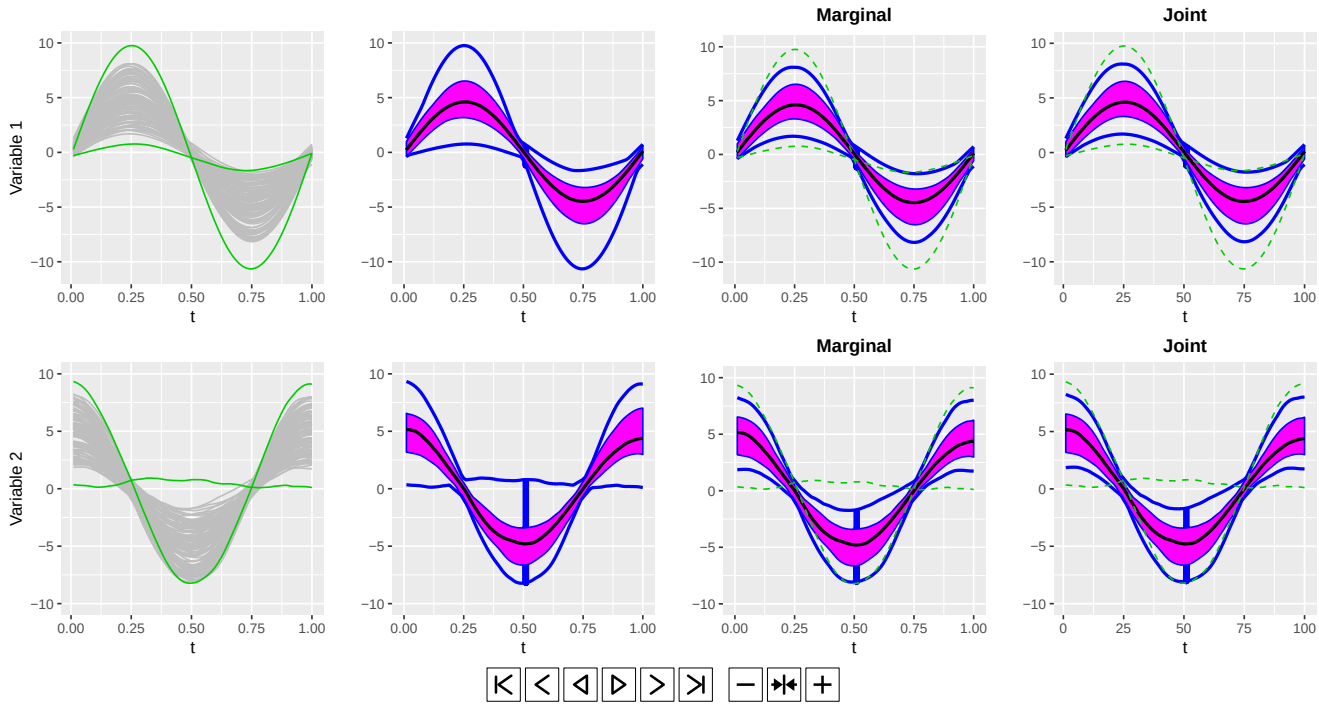
Movie 2. First column: 100 non-outlying curves (grey) and outliers (green). Second column: functional boxplots based on MBD. Third column: two-stage functional boxplots. First row: rotated outliers. Second row: isolated outliers. Third row: outliers generated by different covariance functions [To ensure that the visuanimation will play properly, use Adobe Acrobat Reader to view the pdf file.]

© 2018 John Wiley & Sons, Ltd

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

The ISI's Journal for the Rapid Dissemination of Statistics Research



Movie 3. First column: 100 non-outlying curves (grey) and outliers (green). Second column: functional boxplots based on MBD. Third column: marginal two-stage functional boxplots based on one variable. Fourth column: joint two-stage functional boxplots based on two variables [To ensure that the visuanimation will play properly, use Adobe Acrobat Reader to view the pdf file.]

or jointly. The corresponding tools are the marginal two-stage functional boxplot and the joint two-stage functional boxplot.

We consider an example of bivariate curves contaminated by outliers. Following the setting in López-Pintado et al. (2014), we have a bivariate random Gaussian process $\mathbf{X}(t) = \mathbf{e}(t)$, where $\mathbf{e}(t) = \{\mathrm{e}_1(t), \mathrm{e}_2(t)\}^{\mathrm{T}}$, with zero mean and the following cross-covariance function (Gneiting et al., 2010; Apanasovich et al., 2012):

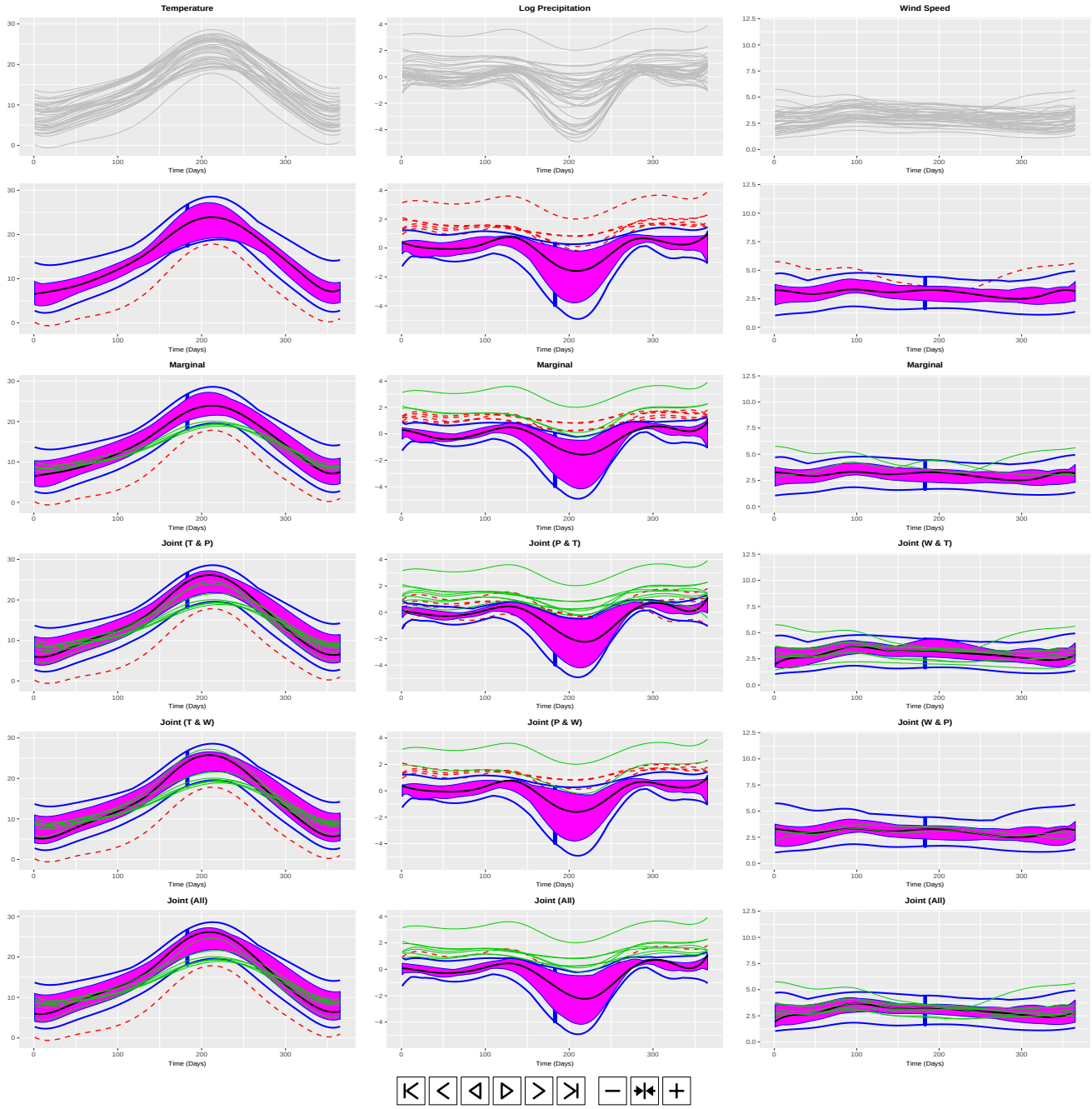
$$C_{ij}(s,t) = \rho_{ij}\sigma_i\sigma_j\mathcal{M}(|s-t|;\nu_{ij},\alpha_{ij}), \quad i,j = 1,2,$$

where ρ_{12} is the correlation between $X_1(t)$ and $X_2(t)$, $\rho_{11} = \rho_{22} = 1$; σ_i^2 is the marginal variance; and $\mathcal{M}(h;\nu,\alpha) = 2^{1-\nu}\Gamma(\nu)^{-1}(\alpha|h|)^{\nu}\mathcal{K}_{\nu}(\alpha|h|)$, $|h| = |s-t|$, is the Matérn (1960) class where $\mathcal{K}_{\nu}$ is a modified Bessel function of the second kind, $\nu > 0$ is a smoothness parameter and $\alpha > 0$ is a range parameter. Throughout the simulation, we choose the following parameters for the bivariate Matérn cross-covariance function: $\sigma_1 = \sigma_2 = 0.1$, $\alpha_{11} = 0.2$, $\alpha_{22} = 0.1$, $\alpha_{12} = 0.16$, $\nu_{11} = 2$, $\nu_{22} = 1.6$, $\nu_{12} = 1.8$ and $\rho_{12} = 0.6$. Then, the bivariate model is designed as follows:

- Model 4 (joint outlier; Movie 3). The non-outlying curves are defined by $X_1(t) = U_1\sin(2\pi t) + \mathrm{e}_1(t)$ and $X_2(t) = U_1\cos(2\pi t) + \mathrm{e}_2(t)$, where U_1 is generated from a uniform distribution on the interval $[2,8]$. Two outliers are introduced for each frame, and there are 40 frames in total. The outliers are generated from $X_{\mathrm{Out},1}(t) = U_2\sin(2\pi t) + \mathrm{e}_1(t)$ and $X_{\mathrm{Out},2}(t) = (10-U_2)\cos(2\pi t) + \mathrm{e}_2(t)$, with U_2 ranging from 1 to 10.

© 2018 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190



Movie 4. First row: smoothed average annual temperature, log precipitation and wind speed curves (from left to right). Second row: functional boxplots based on MBD. Third to fifth rows: two-stage functional boxplots based on pairwise combinations of variables. Sixth row: two-stage functional boxplots based on the combination of all three variables [To ensure that the visuanimation will play properly, use Adobe Acrobat Reader to view the pdf file.]

© 2018 John Wiley & Sons, Ltd

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

The ISI's Journal for the Rapid Dissemination of Statistics Research

In Model 4, most of the joint outliers are not marginally outlying, so it is impossible to detect them marginally. However, by handling the curves jointly, we may identify such outliers from the differences between the correlation structures of their two components. In Movie 3, we see that both the marginal and joint two-stage functional boxplots perform better than do the conventional ones. Furthermore, the joint two-stage functional boxplots are more robust than the marginal ones because the moving outliers do not deform the fences or the central regions.

5 Application to Spanish weather data

We apply our two-stage functional boxplot to the Spanish weather data mentioned in Section 3. In addition to the daily temperature records, we include the daily log precipitation and wind speed data in our illustration to create a three-dimensional functional data set. We present the performances of the original functional boxplot and the two-stage functional boxplot based on different combinations of the Spanish weather data in Movie 4. We randomly selected 40 of the 73 stations to generate the first frame of the movie, and then we added one more station per frame to visualize the performance of different types of boxplots in constructing the central regions and fences.

Compared with the conventional functional boxplot (the second row of Movie 4), the two-stage functional boxplots (both marginal and joint) can capture the structure of the data set more accurately, indicating more robustness when constructing the central regions and fences. For different research interests, various combinations of variables can be used to construct the two-stage functional boxplots. For example, the marginal temperature plot (the third plot in the first column of Movie 4) can be used to explore the pattern of annual temperature. To study the interaction between temperature and precipitation, the joint plots of these two variables (the fourth plots in the first and second columns of Movie 4, respectively) should be considered. In the final frame of both the marginal and joint temperature plots (third and fourth plots in the first column of Movie 4), different sets of curves are flagged as outliers. In the marginal plot, the outlying curves illustrate a smaller temperature variation than do the non-outlying curves across different seasons, whereas in the joint plot, several other curves are identified as outliers owing to their overall outlyingness in terms of the combination of temperature and log precipitation curves, although they appear similar to the median temperature curve. Thus, even the reason why a specific station is flagged as an outlier can be assessed.

6 Discussion

Our proposed two-stage functional boxplot, a tool for the visualization and exploratory data analysis of multivariate curves that combines outlier detection based on the functional directional outlyingness (Dai & Genton, 2018a) with the original functional boxplot (Sun & Genton, 2011), is more robust to outliers, especially to marginal shape outliers and joint outliers than is the functional boxplot used alone. Thus, the constructed central regions and fences are more accurate.

The two-stage functional boxplot can be analogously generalized to multivariate images by combining the outlier detection with surface boxplots (Genton et al., 2014). Moreover, similar to the generalization of boxplots to bagplots (Rousseeuw et al., 1999), the pairwise interactions can be intuitively demonstrated by developing a three-dimensional tool to visualize the structure of a group of bivariate curves.

Acknowledgements

This research was supported by the King Abdullah University of Science and Technology (KAUST).

© 2018 John Wiley & Sons, Ltd

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

References

Apanasovich, TV, Genton, MG & Sun, Y (2012), 'A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components', *Journal of the American Statistical Association*, **107**, 180–193.

Chakraborty, A & Chaudhuri, P (2014), 'The spatial distribution in infinite dimensional spaces and related quantiles and depths', *The Annals of Statistics*, **42**, 1203–1231.

Cuevas, A, Febrero, M & Fraiman, R (2006), 'On the use of the bootstrap for estimating functions with functional data', *Computational Statistics and Data Analysis*, **51**, 1063–1074.

Cuevas, A, Febrero, M & Fraiman, R (2007), 'Robust estimation and classification for functional data via projection-based depth notions', *Computational Statistics*, **22**(3), 481–496.

Dai, W & Genton, MG (2018a), 'Directional outlyingness for multivariate functional data', *Computational Statistics and Data Analysis*. to appear. https://doi.org/10.1016/j.csda.2018.03.017.

Dai, W & Genton, MG (2018b), 'Multivariate functional data visualization and outlier detection', *Journal of Computational and Graphical Statistics*. to appear. https://doi.org/10.1080/10618600.2018.1473781.

Fraiman, R & Muniz, G (2001), 'Trimmed means for functional data', *TEST*, **10**, 419–440.

Genton, MG, Castruccio, S, Crippa, P, Dutta, S, Huser, R, Sun, Y & Vettori, S (2015), 'Visuanimation in statistics', *Stat*, **4**, 81–96.

Genton, MG, Johnson, C, Potter, K, Stenchikov, G & Sun, Y (2014), 'Surface boxplots', *Stat*, **3**, 1–11.

Gneiting, T, Kleiber, W & Schlather, M (2010), 'Matérn cross-covariance functions for multivariate random fields', *Journal of the American Statistical Association*, **105**, 1167–1177.

Goldberg, KM & Iglewicz, B (1992), 'Bivariate extensions of the boxplot', *Technometrics*, **34**, 307–320.

Hardin, J & Rocke, DM (2005), 'The distribution of robust distances', *Journal of Computational and Graphical Statistics*, **14**, 928–946.

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

Long, JP & Huang, JZ (2015), 'A study of functional depths', *arXiv preprint arXiv:1506.01332*.

López-Pintado, S & Romo, J (2009), 'On the concept of depth for functional data', *Journal of the American Statistical Association*, **104**, 718–734.

López-Pintado, S, Sun, Y, Lin, JK & Genton, MG (2014), 'Simplicial band depth for multivariate functional data', *Advances in Data Analysis and Classification*, **8**, 321–338.

Martin-Barragan, B, Lillo, R & Romo, J (2016), 'Functional boxplots based on epigraphs and hypographs', *Journal of Applied Statistics*, **43**, 1088–1103.

Matérn, B (1960), *Spatial Variation*, *Springer*, Heidelberg.

Myllymäki, M, Mrkvička, T, Grabarnik, P, Seijo, H & Hahn, U (2017), 'Global envelope tests for spatial processes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 381–404.

Narisetty, NN & Nair, VN (2016), 'Extremal depth for functional data and applications', *Journal of the American Statistical Association*, **111**, 1705–1714.

 © 2018 John Wiley & Sons, Ltd

Stat

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.190

The ISI's Journal for the Rapid Dissemination of Statistics Research

Rousseeuw, PJ (1985), Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications* in Grossmann, W, Pflug, G, Vincze, I & Wert, W (eds), Vol. B, *Reidel*, Dordrecht, 283–297.

Rousseeuw, PJ, Ruts, I & Tukey, JW (1999), 'The bagplot: A bivariate boxplot', *The American Statistician*, **53**, 382–387.

Serfling, R & Wijesuriya, U (2017), 'Depth-based nonparametric description of functional data, with emphasis on use of spatial depth', *Computational Statistics and Data Analysis*, **105**, 24–45.

Sun, Y & Genton, MG (2011), 'Functional boxplots', *Journal of Computational and Graphical Statistics*, **20**, 316–334.

Sun, Y & Genton, MG (2012), 'Adjusted functional boxplots for spatio-temporal data visualization and outlier detection', *Environmetrics*, **23**, 54–64.

Tukey, JW (1975), Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians*, Vol. 2, *Canad. Math. Congress*, Montreal, 523–531.

© 2018 John Wiley & Sons, Ltd