# Supplementary paper to: Practical Outlier Detection in Functional Data Analysis

Michael L. Creutzinger
Department of Statistics, Colorado State University
and
Julia L. Sharp
Mathematical Statistician, National Institue of Standards and Technology

February 10, 2024

**Abstract**

Existing functional data outlier detection methodology requires the use of a functional data depth measure, functional principal components, likelihood ratio, and/or an outlyingness measure like Stahel-Donoho. Although effective, these functional outlier detection methods may not be easily interpreted. In this paper, a new method is proposed for functional outlier detection: Practical Outlier Detection (POD). POD is developed with the use of summary statistics to be easily interpretable and is found to be as good or better (in terms of accuracy, precision, and Matthew's Correlation Coefficient) than competing methodology, especially for the identification of shape outliers. POD is fully assessed and compared to other methodology using simulation studies and a case study with World Population Growth data.

*Keywords:* POD, outlier detection, functional data analysis, world population growth, magnitude outliers, shape outliers

# 1 Additional Figures

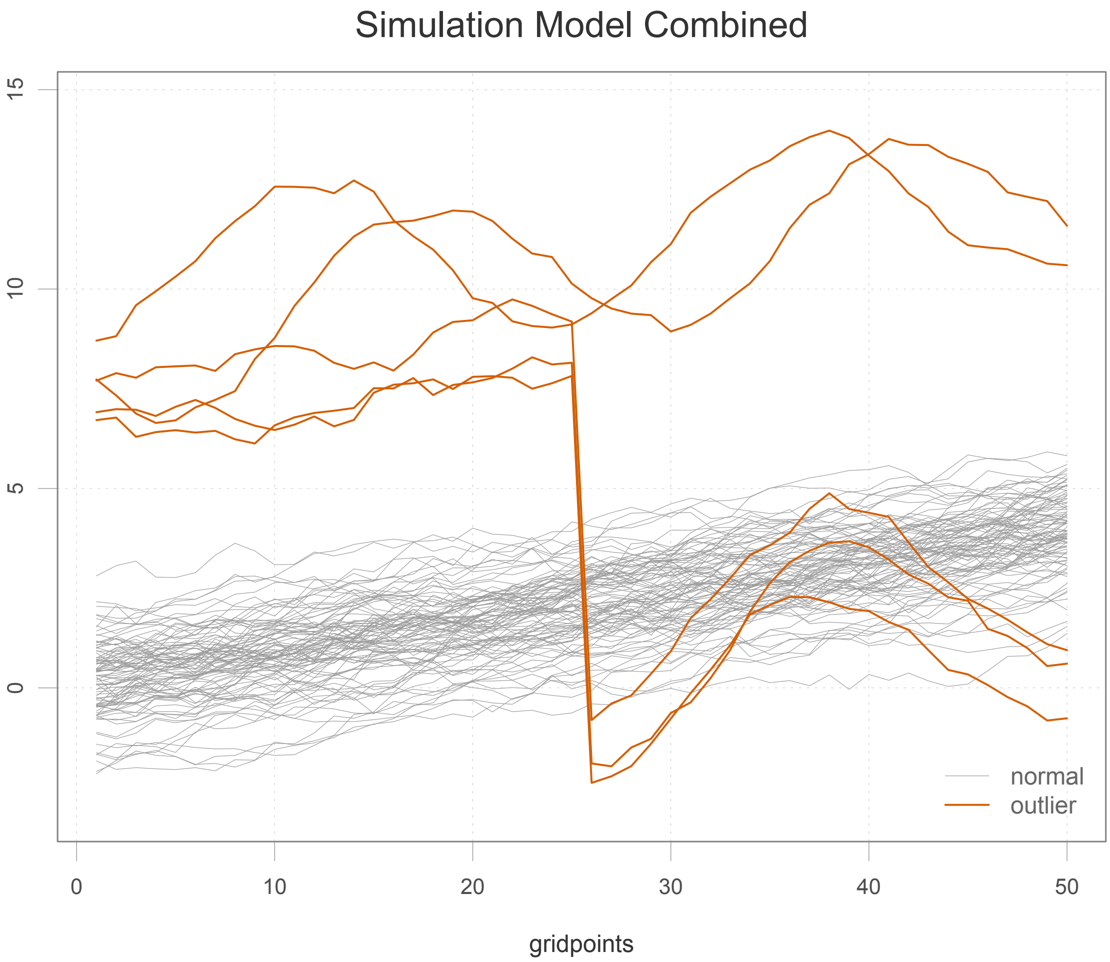

Figure 1: Random sample generated from Simulation Model "Combined", which generates combined magnitude and shape outliers. The random sample presented has $n = 100$ observations, $T = 100$ sampling points, and outlier rate, $\Delta = 0.05$.

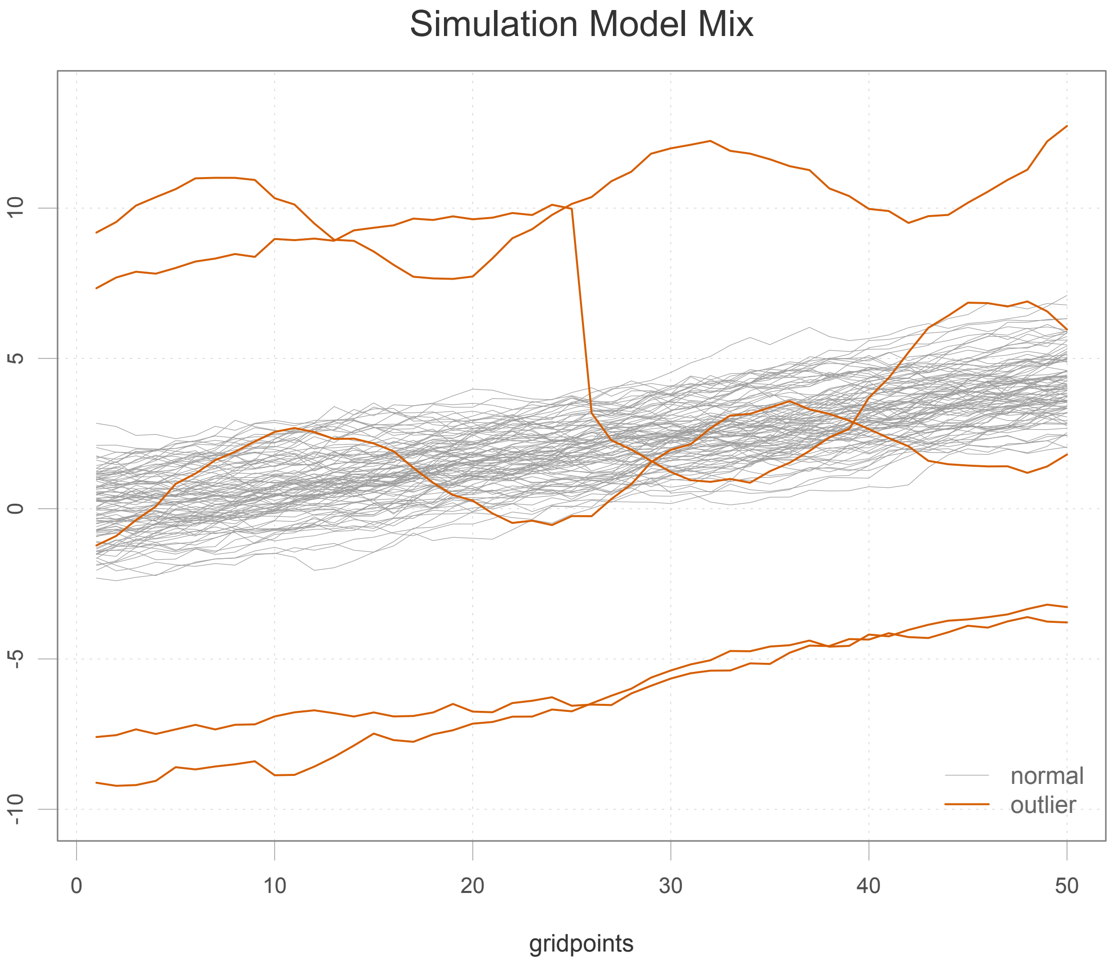

Figure 2: Random sample generated from Simulation Model "Mix", which generates magnitude outliers, shape outliers, and combined magnitude and shape outliers. The random sample presented has $n = 100$ observations, $T = 100$ sampling points, and outlier rate $\Delta = 0.05$.

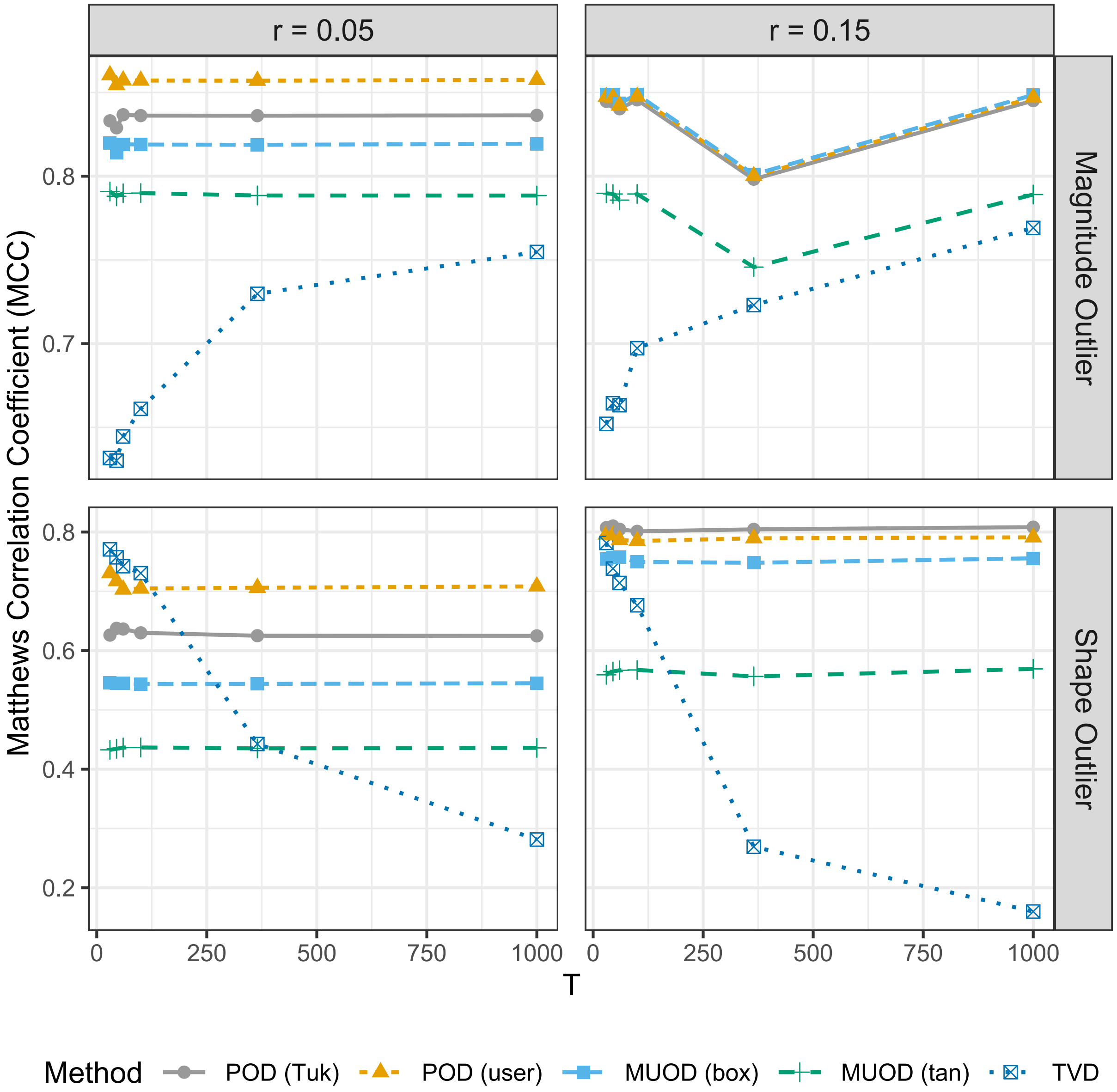

Figure 3: Average Matthew's Correlation Coefficient ($MCC$) for classifying magnitude and shape outliers on a varying number of sampling points ($T = 30, 45, 60, 100, 365, 1000$), faceted on the proportion of outliers ($\Delta = 0.05, 0.15$) and the type of outlier (rows), and linetype and color change by the method type (POD (Tuk), POD (user), MS-Plot, MUOD (box), and TVD.

# 2 Practical Outlier Detection (POD): Development of the Method

## 2.1 Deciding the Number of Intervals

The first priority in developing our practical outlier method is to minimize the number of tuning parameters. Specifically, the number of intervals used to bin the sampling points of each observation, had a direct effect on the performance of the method. Rather than allowing a user to specify the number of intervals, it is decided to optimize the algorithm in terms of accuracy, precision, recall (or sensitivity), etc. with respect to the number of intervals used to bin the data. When developing the simulation, the whole domain, one interval, two intervals, ..., up to a number of intervals that would allow at least three sampling points per interval, are all decided. For each number of intervals, a sequence of 30 threshold values, using $\delta$ from 0.01 to 0.99, is used for deciding the outlying observations. Using a sequence of thresholds allowed the calculation of Precision-Recall Area Under the Curve ($PR - AUC$) to compare the performance of each number of intervals used. Precision-Recall is preferred over the traditional Receiver-Operating-Characteristic curve because outliers tend to represent a small proportion of total observations.

The first consideration is to look for patterns in summary statistics collected on the whole domain, with respect to how many intervals would be the optimal choice for maximizing $PR - AUC$. Then, the data would drive the methodology, without direct user influence. Clear patterns when investigating trends of summary statistics against $PR - AUC$ and the number of intervals used are not observed. However, patterns in the number of intervals used and measurements of classification diagnostics (precision, recall, selectivity, etc.) are observed. In general, as the number of intervals used to bin the data increases, the number of outliers identified also increases. In all nine simulation models, the Practical Outlier method improved when using two intervals versus the whole domain, and when using three intervals versus two intervals. Yet, for several simulation settings, after reaching enough intervals used, the performance of Practical Outlier would drop off: precision, recall, and accuracy all decreased. The unexpected drop in performance led to further investigation of interval sizes based off the number of sampling points, $T$. Further investigation ultimately led to the interval rule described in Section 2.2 of the main paper.

A second simulation is conducted to decide between the use of 15 intervals, 20 intervals, and two techniques of combining their results (when T $\geq$ 60). One way to combine the results is to identify a list of outliers using 15 intervals, identify a list using 20 intervals, then combine the two lists as the final list of outliers. The second technique is to implement the method with 15 intervals and save the data frame containing the total count of each observation's "out" features. Repeat the same steps with 20 intervals used. Then, sum up the total count of "out" features found with 15 and with 20 intervals. Lastly, use this data frame of counts to identify the outliers.

In the simulation comparing combination of results, it is found that combining the counts of "out" features, before identifying the outliers with a threshold cutoff, outperformed the first technique of identifying two separate lists of outliers first, and then combining.

## 2.2 Determining Extreme Summary Statistics

The collection of summary statistics considered for POD do not all have symmetric sampling distributions. For example, if the population of a random sample is normally distributed, the sampling distribution of the sample variance will be $\chi^2$, which is right-skewed. Therefore, in order to find the extreme values of a summary statistic, it seems necessary to use a rule that is robust to skewed distributions. Seo (2002) reviews and compares common methods in outlier detection for non-functional, univariate data sets. Of particular interest, Vanderviere and Huber (2004) introduced the adjusted boxplot, which alters Tukey's original boxplot rule by leveraging a robust measure of skewness, called the medcouple. Medcouple is first introduced and compared to other robust measures of skewness (namely, quartile skewness (QS) and octile skewness (OS)) by Brys et al. (2004). When compared, the medcouple has the sensitivity of OS to detect skewness and the robustness of QS towards outliers that are present (Seo, 2002). It is also important to note that the medcouple is a bounded measure of skewness with a breakdown value of 25 % (Brys et al., 2004). That is, it would require that 25 %, or more, of the observations are replaced with outlying values, before medcouple becomes useless (Rousseeuw and Leroy, 1987).

To calculate the adjusted boxplot "fences," the first and third quartiles, the inner quartile range, and the medcouple must all be estimated first. Denote the first and third quartiles, for each summary statistic and each interval, as $Q_1\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}$ and $Q_3\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}$. The medcouple is estimated with respect to each summary statistic and each interval (resulting in 90 measures of medcouple). Denote the medcouple as $MC\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}$, which is defined as $MC\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}:=$

$$\mathrm{med}\left\{\frac{\left(\left[S_{type}\left(\mathrm{Int}_a\right)\right]_i-\mathrm{med}\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}\right)-\left(\mathrm{med}\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}-\left[S_{type}\left(\mathrm{Int}_a\right)\right]_{i'}\right)}{\left[S_{type}\left(\mathrm{Int}_a\right)\right]_{i'}-\left[S_{type}\left(\mathrm{Int}_a\right)\right]_i}\right\},$$

for $\left[S_{type}\left(\mathrm{Int}_a\right)\right]_i\leq\mathrm{med}\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}\leq\left[S_{type}\left(\mathrm{Int}_a\right)\right]_{i'}$.

The "fences" for the adjusted boxplot, with respect to each summary statistic and interval, are denoted as $(L,U)\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}$. Each fence is defined as

$$\begin{cases}\left(Q_1-1.5\times\exp\left(-3.5\times MC\right)\times IQR, Q_3+1.5\times\exp\left(4\times MC\right)\times IQR,\right) & \text{if } MC\geq 0\\ \left(Q_1-1.5\times\exp\left(-4\cdot MC\right)\cdot IQR, Q_3+1.5\cdot\exp\left(3.5\cdot MC\right)\cdot IQR,\right) & o.w.\end{cases},$$

where $Q_1$, $Q_3$, and $MC$ are calculated with respect to same summary statistic and interval, $\left\{S_{type}\left(\mathrm{Int}_a\right)\right\}$. Note that the adjusted boxplot is equivalent to Tukey's classical boxplot, when the medcouple is zero (symmetric distribution).

Three different methods are investigated for deciding the extreme summary statistics: use of Tukey's classical boxplot for all statistics, use of adjusted boxplot for all statistics, and the use of Tukey's classical boxplot for roughly symmetric sampling distributions, and otherwise an adjusted boxplot. Through statistical theory, backed by an empirical study, the sampling distributions of the mean, minimum, maximum, median, $AUC$, and coefficient of variation are found to be symmetric, while the sampling distributions of the variance, range, and roughness are found to be positively skewed.

Of all three methods, it is found that using Tukey's classical boxplot for all summary statistics outperformed using an adjusted boxplot for all summary statistics and using a combination of Tukey's classical and adjusted boxplot.

## 2.3 Classifying the Type of Outlier

In Section 2, it is mentioned that the measures of location would be ideal for identifying magnitude outliers, while the measures of dispersion and roughness would be ideal for identifying shape (and amplitude) outliers. Simulation studies are used to observe the frequency of extreme summary statistics for each observation identified as a functional outlier by POD, and then to identify patterns in the count of extreme summary statistics found.

Four different data generation models are considered: a model with magnitude outliers only (see Model 1, Figure 1 in the main paper), a model with shape outliers only (see Model 7, Figure 1 in the main paper), a model with combined (magnitude and shape) outliers only, and a model with all types present (see Figure 2). The model with combined outliers only is similar to Figure 2, without the strictly magnitude and/or shape outliers present. Each model is simulated using all combinations of sample size $n$ = 30, 45, 60, number of sampling points $T = 30$, 100, 250, rate of outliers $\Delta = 0.05$, 0.15, and covariance roughness $\beta = 0.1$, 0.5, 0.9. Each unique combination of model, sample size ($n$), number of sampling points ($T$), rate of outliers ($\Delta$), and covariance roughness ($\beta$) is simulated 250 iterations. On each iteration, a random sample of functional data is generated by the chosen model with the chosen simulation parameters. POD is implemented to identify the extreme statistics of every observation on every interval. Only the data pertaining to the true functional outliers in the sample, as identified by the simulation model, is retained.

Investigation of the results found a pattern related to the count of extreme summary statistics per interval and the class of functional outlier. Specifically, the total number of extreme summary statistics in each group, location vs variation, found per interval is computed. For each observation, an interval that has more than two extreme location stats is identified as Magnitude outlying, and an interval that has more than one extreme variation stat is identified as Shape outlying. If at least one third of the intervals are identified as Magnitude outlying, then the functional outlier is classified as a Magnitude outlier. If at least one fifth of the intervals are identified as Shape outlying, then the functional outlier is classified as a Shape outlier. In rare cases, if a functional outlier has neither one third or more Magnitude outlying intervals nor one fifth or more Shape outlying intervals, then the functional outlier is identified as a Shape outlier. This typically occurs when a functional outlier is strictly Magnitude outlying over one part of the domain and strictly Shape outlying over another part of the domain.

This pattern gave the ability to accurately classify 96.5 % of all 373,400 simulated functional outliers. Of the 16,803 simulation functional outliers that are improperly classified, 3116 of them are truly both Shape and Magnitude outlying, but falsely identified as Magnitude outlier only; 2956 of them are truly Magnitude outlying only, but falsely identified as both Shape and Magnitude outlying; 10,722 of them are truly Shape outlying only, but falsely identified as both Shape and Magnitude outlying; and nine of them are truly Shape outlying only, but identified as a Magnitude outlier only. Since this pragmatic rule is able to reach more than 95 % accuracy, it is chosen to be implemented in the methodology of POD for classification of type.

### Disclaimer

This work is U.S. Government work and not protected by U.S. copyright.

### Acknowledgements

This work utilized the Alpine high performance computing resource at the University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, and Colorado State University.

## References

Brys, G., M. Hubert, and A. Struyf (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics 13*(4), 996–1017.

Rousseeuw, P. J. and A. Leroy (1987). Robust regression and outlier detection. *John Wiley & Sons.*

Seo, S. (2002). A review and comparison of methods for detecting outliers in univariate data sets. pp. 59.

Vanderviere, E. and M. Huber (2004). An adjusted boxplot for skewed distributions. *Computational Statistics*, 8.