Journal of Statistical Research *https://doi.org/10.3329/jsr.v59i1.83681*
2025, Vol. 59, No. 1, pp. 9-30 *ISSN 0256 - 422 X*

A NEW GRAPHICAL DEVICE BASED ON TRIMMED MEAN

SUBHRA SANKAR DHAR

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Kanpur 208016, India

Email: subhra@iitk.ac.in

SHALABH*

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Kanpur 208016, India

Email: shalab@iitk.ac.in

AAYUSH

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Kanpur 208016, India

Email: paayush@iitk.ac.in

HIMANSHU SHEKHAR DAS

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Kanpur 208016, India

Email: iamhimanshushekhar13@gmail.com

ABSTRACT

Trimmed mean has been considered a robust estimator of location parameters over the last five decades. The issue of outlier detection has been considered using analytical and graphical statistical tools. This article proposes a graphical device based on the trimmed mean to check whether data has any outliers or not, and in the presence of outliers, the proposed graphical device enables the estimation of the proportion of the outliers as well. The extension of the methodology to the high dimensional data is also outlined. Furthermore, the proposed visualization toolkit is implemented on economic data and gives us an idea of the presence/absence of influential observations/outliers.

Keywords and phrases: Breakdown point; Cauchy distribution; Laplace distribution; Location parameter; Normal distribution; Robust estimator; Trimmed mean.

* Corresponding author

1 Introduction

Detecting outliers in the data is an important issue in any statistical analysis. The presence of outliers in the data disturbs the optimal properties of the statistical tools which leads to impious, incorrect and invalid statistical inferences. Various concepts have been deployed in the literature to identify the outliers in the data. It becomes more challenging to identify the outliers in higher dimensional data sets. The outlier detection can be performed using analytical and graphical devices. The graphical devices have their advantages as they are easy to understand in many practical and complicated situations in datasets. The concept of the trimmed mean can be used to detect the outliers in the data; see Pratap et al. (2021).

The trimmed mean is a well-known robust estimator for the location parameter. Specifically, it can achieve good efficiency with a good breakdown point, which is an out-of-ordinary property of any estimator. In this context, it is important to mention that the sample mean is the asymptotically most efficient estimator of the location parameter when the data follow the normal distribution. In contrast, the sample median is the asymptotically most efficient estimator when the data follow the Laplace distribution. Note that the trimmed mean coincides with the sample mean when the trimming proportion equals zero, whereas it coincides with the sample median when the trimming proportion equals $(1/2)$. Overall, the trimmed mean bridges the sample mean and the sample median (see, e.g., Lehmann (1983) for a detailed discussion).

In the literature, to the best of our knowledge, Tukey and McLaughlin (1963) first time proposed the trimmed mean to use the trimmed version of t-statistic, and after a few years, Hogg (1967) also used the concept of trimmed mean for some practical purposes. During the same period of time, the asymptotic distribution of trimmed mean was derived by Bickel (1965), and Stigler (1973) worked on the same issue under weaker conditions. In the few years, many other works and applications on trimmed mean had been done by Dhar and Chaudhuri (2009), Dhar and Chaudhuri (2012), Dhar (2016), Dhar et al. (2022) and Dhar et al. (2022) and a few references therein.

The outliers detection by signal subspace matching is considered in Wax and Adler (2024), and Farnè and Vouldis (2024) presented a conditional outlier detection methodology for high-dimensional data. Further, Song et al. (2024) considered the outlier detection using penalized likelihood estimation for general spatial models, Murph et al. (2024) presented visualisation and outlier detection for probability density function ensembles, and Hu et al. (2024) described regularized Huber regression for outlier detection. A gradient test statistic for outlier detection in generalized estimating equations is discussed in Osorio et al. (2024), whereas Amin et al. (2022) and Rashid et al. (2022) considered outlier detection in gamma regression, and in high-dimensional data using support vector regression, respectively. A survey of outlier detection in high dimensional data streams is presented in Souiden et al. (2022) and Smiti (2020). The nonparametric tests for detection of high dimensional outliers are discussed in Modarres (2022), and Cabana et al. (2021) presented outlier detection in a multivariate setup using robust Mahalanobis distance with shrinkage estimators. Besides outlier detection in robust statistics point of view, Fan et al. (2021) studied a selective overview of recent advances in high dimensional factor models in the presence of the outliers and their applications to statistics, and a decade ago, Karoui et al. (2013) investigated regression model with high dimensional predictors with outliers. The readers are referred to the references in Karoui et al. (2013) as well in this context.

In this paper, we propose a graphical device to detect the outliers using the concept of trimmed mean. Let us now denote $\alpha \in (0, \frac{1}{2})$ as the trimming proportion of the trimmed mean, and it is a well-known result in the literature (see, e.g., Hampel et al. (1985) and Dhar and Chaudhuri (2009)) that the asymptotic breakdown point of α-trimmed mean is α. In other words, it indicates that the α-trimmed mean won't explode to infinity unless the proportion of outliers is larger than α, and this fact follows from the gross error sensitivity of the population version of the α-trimmed mean as well (see Hampel et al. (1985)). Using this idea, we propose a new graphical device and implement it on well-known business data. Moreover, the theoretical justification of the graphical device is also provided. The proposed idea is extended to high-dimensional data sets.

The rest of the article is organized as follows: The α-trimmed mean is defined, and its relevant statistical properties are discussed in Section 2. Section 3 proposes the graphical device based on the α-trimmed mean, and well-known economic data are analyzed using the proposed methodology in Section 4. The extension of the proposed method to high-dimensional data is provided in Section 5. Finally, some concluding remarks are presented in Section 6. The technical details and the Python code used for all numerical studies are provided in the Appendix.

2 Trimmed Mean and Its Properties

Let $X_1, X_2, \ldots, X_n$ be a random sample from a location-scale family with the form of distribution $F(X, \theta)$, where F is an absolutely continuous Lebesgue measurable distribution function, and θ is the unknown location parameter. The α-trimmed mean, which is introduced by Tukey (1948), based on the random sample $X_1, X_2, \ldots, X_n$ is defined as

$$\bar{X}_{n,\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha+1]}^{n-[n\alpha]} X_{(i)},$$

where $X_{(i)}$ is the i-th order statistic of the random sample $X_1, X_2, \ldots, X_n$, and $\alpha \in \left(0, \frac{1}{2}\right)$ is the trimming proportion. Here $[\cdot]$ denotes the largest integer contained in $[\cdot]$. Further, observe that $\bar{X}_{n,\alpha}$ coincides with the sample mean when $\alpha \to 0$ and with the sample median when $\alpha \to \frac{1}{2}$. For details about the α-trimmed mean, the readers may refer to Dhar and Chaudhuri (2009), Dhar and Chaudhuri (2012), and a few references therein. The population version of the trimmed mean (denoted as $\theta(\alpha)$) can be defined as follows.

$$\theta(\alpha) = \frac{1}{1 - 2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x f(x) dx,$$

where $f(\cdot)$ is the probability density function of $F(.)$. Note that $\theta(\alpha) = \theta$ for all $\alpha \in (0, \frac{1}{2})$ if and only if f is a symmetric probability density function with center of symmetry $= 0$.

We now assume some technical conditions that will be required for some theoretical properties of the trimmed mean.

(A1) Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution function F, where F is continuously differentiable, and the derivative of F (i.e., f) is positive on the entire real line.

Theorem 1. *Under (A1), for any* $0 < \alpha_1 < \alpha_2 < \frac{1}{2}$, *we have* $\sup_{\alpha_1<\alpha<\alpha_2} |\bar{X}_{n,\alpha} - \theta(\alpha)| = O_p(n^{-\frac{1}{2}})$.

Theorem 1 asserts that $\bar{X}_{n,\alpha}$ can approximate $\theta(\alpha)$ arbitrarily well for a sufficiently large sample size over any open interval of $(0, \frac{1}{2})$. In other words, the feature of $\theta(\alpha)$ can be captured by $\bar{x}_{n,\alpha}$ over any open interval of $(0, \frac{1}{2})$ when the sample size is large enough. This fact enables us to develop the graphical device based on $\bar{x}_{n,\alpha}$.

We now would like to recall the concept of quantitative breakdown point (see, e.g., Hampel et al. (1985)) of $\theta(\alpha)$. The definition of the breakdown point of any functional is as follows: The breakdown point of a functional T at a distribution P is defined as

$$\epsilon^*(T, P, d) = \inf\{\epsilon > 0 : |T(P) - T(Q)| = \infty \text{ for some } Q \text{ such that } d(P, Q) < \epsilon\},$$

where d is a suitable metric measuring the discrepancy between P and Q. The following Theorem states the breakdown point of $\theta(\alpha)$.

Theorem 2. *Under (A1),* $\epsilon^*(\theta(\alpha), F, d) = \alpha$, *where* d *is a pseudo metric.*

Theorem 2 asserts that the breakdown point of $\theta(\alpha)$ is α. In other words, the trimmed mean functional $\theta(\alpha)$ will break down if the original distribution P and the contaminated distribution Q are apart from each other by at least α. This idea, along with the assertion in Theorem 1, motivates a graphical device to check whether the data has outliers or not, which will be studied in the next section.

3 Graphical Device for Detecting Outliers

Note that the assertion in Theorem 1 indicates that $\bar{x}_{n,\alpha}$ can approximate $\theta(\alpha)$ arbitrary well for a sufficiently large sample, and it follows from the assertion in Theorem 2 that $\theta(\alpha)$ can break down when the original distribution P and the contaminated distribution Q are apart from each other by at least α, i.e., in view of sample analogue, when the data generated from Q has at least α proportion of data, which are located far apart from the data cloud generated from P. *Hence, as* $\bar{x}_{n,\alpha}$ *can approximate* $\theta(\alpha)$ *arbitrary well for a sufficiently large sample, one can plot* $\bar{x}_{n,\alpha}$ *with respect to* α *(i.e., the trimming proportion) for a given data to estimate the proportion of outliers in data.*

We conduct a simulation experiment using the data from different distributions to understand the implementation of the proposed graphical device. Let us now consider the following three examples.

Example 3.1. Data are obtained from the following mixture of normal distributions (i) $0.7N(0,1)+0.3N(10,1)$ and (ii) $0.7N(0,1)+0.3N(100,1)$. Here $N(\mu, \sigma)$ denotes the normal distribution with location parameter μ and the scatter parameter σ.

Example 3.2. Data are obtained from the following mixture of Laplace distributions (i) $0.7L(0,1)+0.3L(10,1)$ and (ii) $0.7L(0,1) + 0.3L(100,1)$ distributions. Here $L(\mu, \sigma)$ denotes the Laplace distribution with location parameter μ and the scatter parameter σ.

Example 3.3. Data are obtained from the following mixture of Cauchy distributions (i) $0.7C(0,1)+0.3C(10,1)$ and (ii) $0.7C(0,1)+0.3C(100,1)$ distributions. Here $C(\mu,\sigma)$ denotes the Cauchy distribution with location parameter μ and the scatter parameter σ.

Note that for all three examples, i.e., Examples 1, 2, and 3, the models are well-known location contamination model (see, e.g., Dhar and Chaudhuri (2012)). The choice contamination model (location contamination model is a special case) has been used in the context of robust statistics since 1970s (see Huber (1981), p. 9 and 11), and the same concept was used by Dhar et al. (2016) (see Section 2 in this article) in measuring the robustness of various association index. For a lucid understanding of this concept, let us consider the following example. Suppose that the data are generated from $0.9N(0,1)+0.1N(100,1)$ distribution. That means with probability 0.9, the data are generated from $N(0,1)$, which is the main data cloud, and with probability 0.1, the data are generated from $N(100,1)$ distribution, which are the outlier observations. Observe that for a sufficiently large sample, 99.73% observations from the main data cloud are lying between -3 and 3 whereas 0.27% observations (i.e., outliers) are lying between 97 and 103. For real data, concerning outliers, such a location contamination model has been used in analyzing many real data sets including well-known Crabs Study data and Wine data set. In the location contamination model (broader sense, mixture of Gaussian distributions), these two data sets are analyzed in Clark and McNicholas (2024) (see Sections 4.3 and 4.4 in this article). Besides, such type of mixture distribution has been used in statistical modelling and forecasting as well (see, e.g., Shalabh et al. (2024)).

The main data cloud is generated from standard normal, Laplace, and Cauchy distributions, and the outliers in the data cloud are generated from the same distribution but non-zero location parameter. The sample sizes considered are $n = 50, 100, 200$ and 500. Here we have taken the values of the location parameter as 10 (see the diagrams of the first and the second rows in Figures 1, 2 and 3) and the location parameter as 100 (see the diagrams in the third and the fourth rows in Figures 1, 2 and 3). In general, the diagrams in Figures 1, 2 and 3, the values of $\bar{x}_{n,\alpha}$ is reasonably large when α is smaller than 0.3 (approximately), and the values of $\bar{x}_{n,\alpha}$ is close to zero when α is greater than 0.3 (approximately). This feature is more prominent in Figures 1 and 2 as normal and Laplace distributions are light-tailed, whereas the Cauchy distribution is heavy-tailed. Moreover, in each of Figures 1, 2 and 3, the change of curvature is sharper in the diagrams in the first and the second rows compared to that in diagrams in the third and the fourth rows as the distributions with the location parameter as 100 are more skewed than the distributions with the location parameter as 10. The sample sizes are denoted as num_sample inside the Figures 1, 2 and 3.

Therefore, our overall suggestion is the following to estimate the proportion of outliers in the data:

Step 1 : For a given data, plot $\bar{x}_{n,\alpha}$ for various choices of $\alpha \in (0, \frac{1}{2})$.

Step 2 : Check range of α, where $\bar{x}_{n,\alpha}$ is much larger than zero. Let $[0, \beta]$ be that range.

Step 3: The estimated proportion of outliers in the data is β.

In this context, we would like to mention that the behavior of the trimmed mean for such cases can be characterized by the second derivative of the trimmed mean with respect to the trimming proportion, as considered in Dhar and Chaudhuri (2012) (see Section 3 in this article). Technically speaking, suppose that $\theta^{''}(\alpha)$ denotes the second derivative of $\theta(\alpha)$, where $\theta(\alpha)$ is the same as

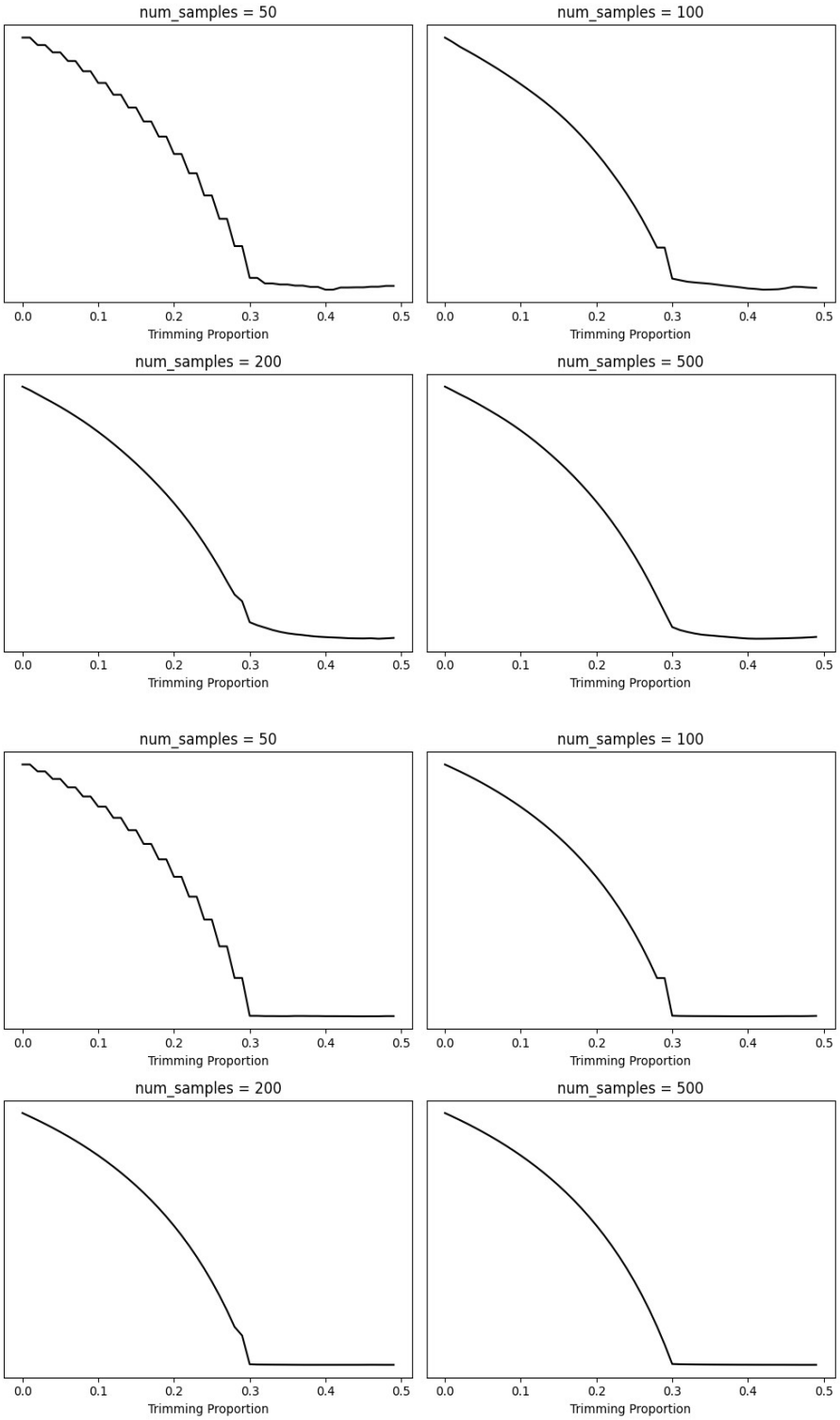

Figure 1: *The diagrams of the first and the second rows plot* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7N(0,1) + 0.3N(10,1)$ *for different choices of* α*, and the diagrams in the third and the fourth rows plot* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7N(0,1) + 0.3N(100,1)$ *for different choices of* α*. For details, see Example 1.*

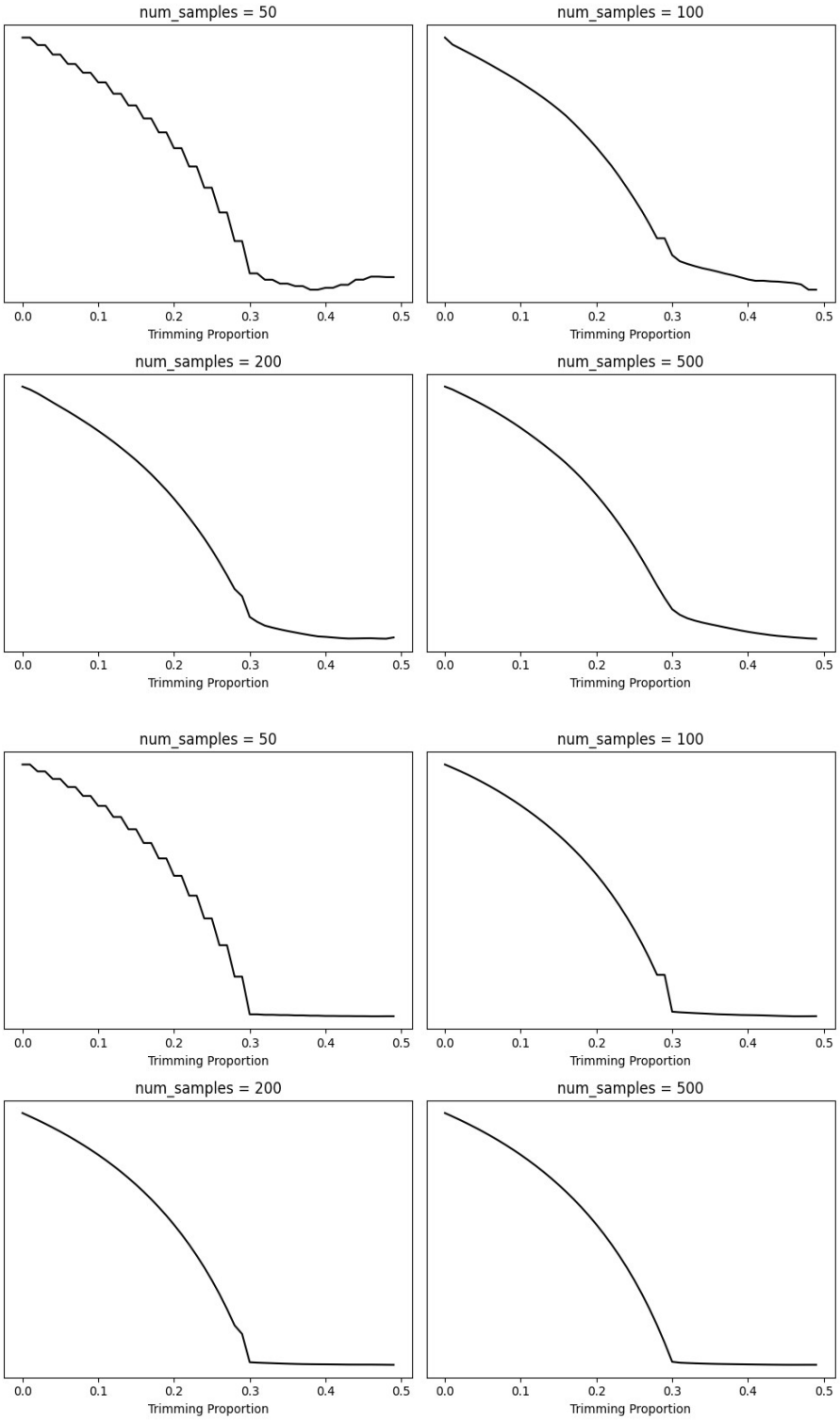

Figure 2: *The diagrams in the first and the second rows plot* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7L(0,1) + 0.3L(10,1)$ *for different choices of* α*, and the diagrams in the third and the fourth rows plot* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7L(0,1) + 0.3L(100,1)$ *for different choices of* α*. For details, see Example 2.*

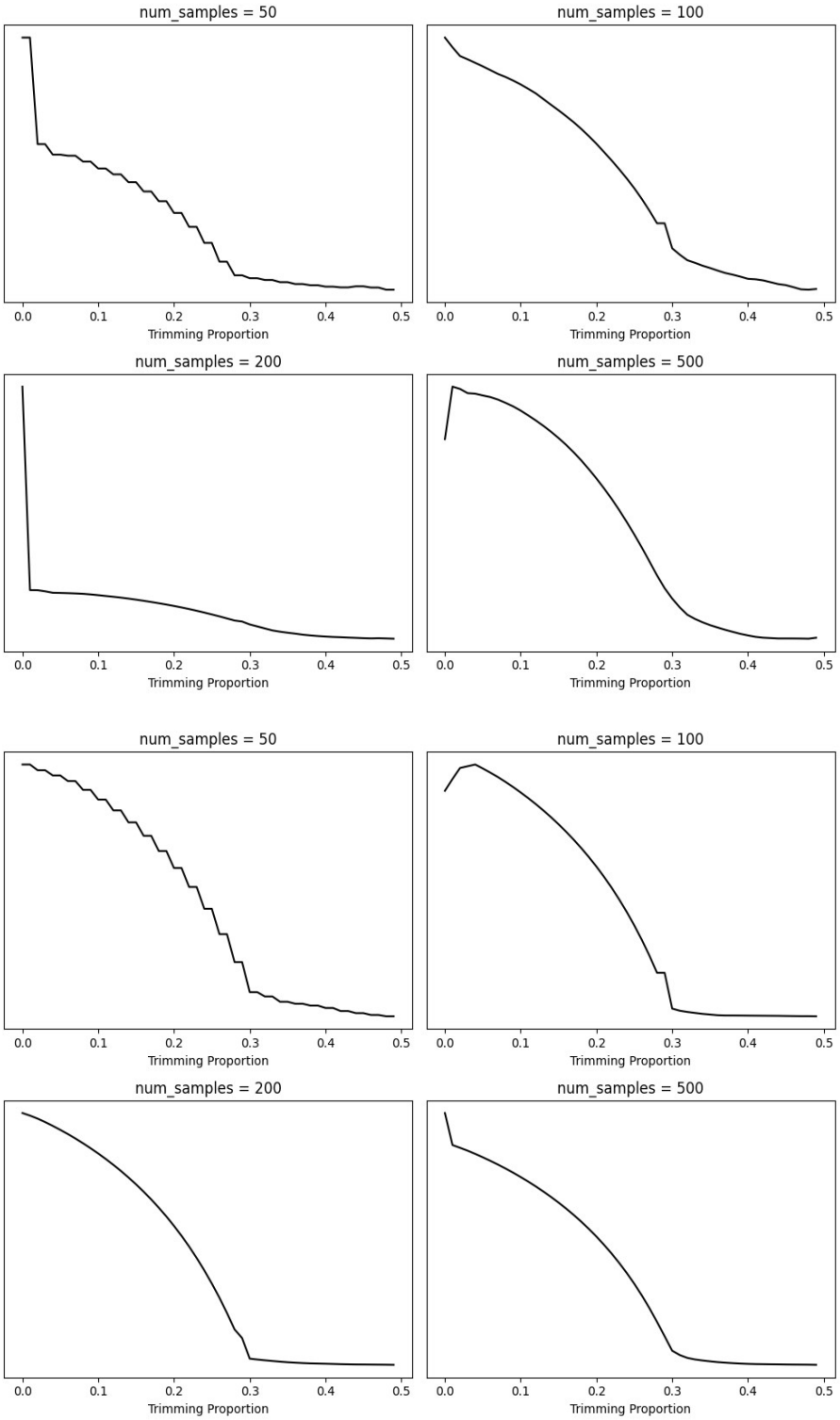

Figure 3: *The diagrams in the first and the second row plots* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7C(0,1)+0.3C(10,1)$ *for different choices of* α*, and the diagrams in the third and the fourth row plot* $\bar{x}_{n,\alpha}$ *when the data obtained from* $0.7C(0,1) + 0.3C(100,1)$ *for different choices of* α*. For details, see Example 3.*

defined at the beginning of Section 2, and let $\hat{\theta}^{''}(\alpha)$ be a consistent estimator of $\theta^{''}(\alpha)$. Suppose that the estimator of β, which is denoted by $\hat{\beta}_n$, is defined as

$$\hat{\beta}_n = arg \max_{\alpha \in [a_1, a_2]} \hat{\theta}^{''}(\alpha),$$

where $0 < a_1 < a_2 < \frac{1}{2}$. It follows from the assertion in Theorem 5 of Dhar and Chaudhuri (2012) that $\hat{\beta}_n$ will be a "good" estimator of β as long as $2\{1 - H(\frac{\phi}{2})\}$ is small enough, where the form of the location contamination model is $(1-\beta)H(.) + \beta H(. - \phi)$. At present, to the best of our knowledge, this is the best possible optimal property of $\hat{\beta}_n$ in estimating β known to us. In Dhar and Chaudhuri (2012), we have also seen that $\hat{\beta}_n$ performs better than EM algorithm based estimator of β for many examples.

The next section will implement this graphical device on a real data set.

4 Real Data Analysis

Here, we implement the proposed graphical device on a well-known real data set described in the following.

Bike sales data: The bike sales data set encompasses the data about the sales of bikes with different variables affecting the decision of sales data. This dataset contains variables like age, sex, marital status, homeowner, children, and their bike sale decisions. The bike sales data set has 89 rows with diverse attributes, which is available at `https://www.kaggle.com/datasets/ratnarohith/uncleaned-bike-sales-data`. The data set is also available at `https://home.iitk.ac.in/~shalab/dhar_shalabh_trim_graphical/uncleanedbikesalesdata.xlsx`. The data set provides comprehensive information, including the date of sale, customer demographics (age, gender, age group), geographic details (country, state), and transaction specifics (product category, sub-category, order quantity). In this study, the "Revenue" column is used as it is expected to have a handful number of outliers. Figure 4 illustrates the scatter plot of the data, which indicates the presence of many outliers. Therefore, for this data set, to have an idea about the average "Revenue", using the sample mean will lead to a misleading result. However, as our proposed graphical device reveals 42% (approximately) outliers in the data, one can use a highly robust estimator of the center to get an idea about the average "Revenue" in the data set.

We now plot $\bar{x}_{n,\alpha}$ (here $n = 89$) with respect to α (i.e., trimming proportion) for this data data in Figure 5. The diagram in Figure 5 indicates that the curvature of $\bar{x}_{n,\alpha}$ becomes parallel to horizontal axis when the trimming proportion is larger than 0.42 (approximately), i.e., in other words, this data set has approximately 42% outliers.

5 Extension to High Dimensional Data

In the recent past, due to the advancement of technology, we have been seeing a lot of high dimensional data in various fields like Medicine, Finance, Engineering, Robotics and so on. Here,

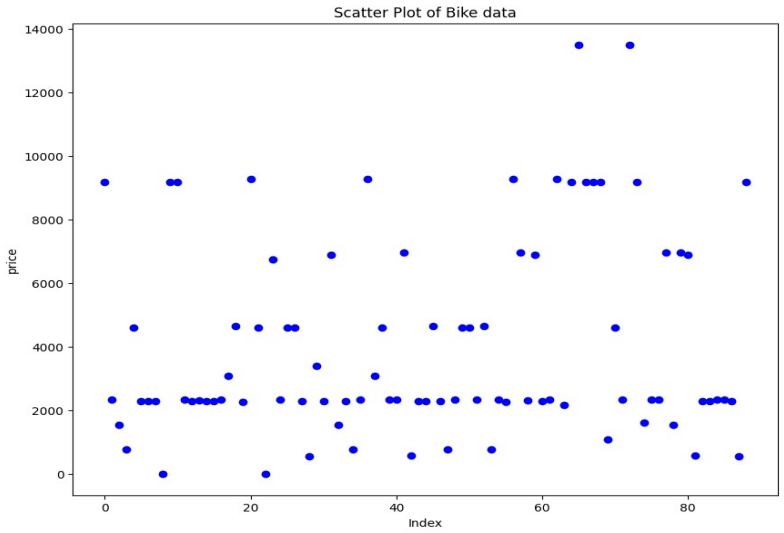

Figure 4: *Scatter plot of Bike sales data*

we make an effort to address how the proposed methodology can be extended for high-dimensional data.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an n observed values of some d-dimensional random vector $\mathbf{X}$, where the dimension d can be larger than the sample size n, and it can be treated as one of the notions of high dimensional data. Now, let $\mathbf{X}_{(i;n)}$ be the i-th order statistic in $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ (here $i = 1, \ldots, n$ and $n \in \mathbb{N}$), although the concept of ordering for high dimensional data (strictly speaking, any data with dimension more than two) is not straightforward like linear ordering. Hence, one needs to appropriately define $\mathbf{X}_{(i;n)}$. In the following, we discuss the possibilities.

The most straightforward one is defining order statistics componentwise. However, there are two fundamental issues involved in this procedure. Firstly, a particular componentwise order statistic may not be a member of the sample, although defining the α-trimmed mean

$$\bar{\mathbf{X}}_{n,\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} \mathbf{X}_{(i;n)}$$

is possible. Secondly, the more important issue is that the componenetwise order statistic cannot capture the dependence structure among the components.

Another option is to define $\mathbf{X}_{(i;n)}$ through data depth, which a statistical toolkit to order a multivariate data For details on data depth, the readers are refereed to Liu et al. (1999). There are a few well known versions of data depth, and among them, the spatial depth (see Vardi and Zhang

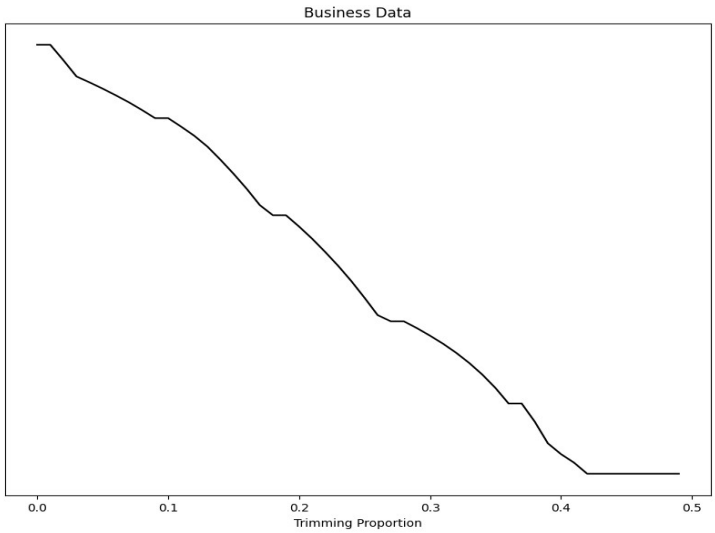

Figure 5: *Plot of $\bar{x}_{n,\alpha}$ for Bike sales data.*

(2000)) and the half space depth (see Zuo and Serfling (2000)) are applicable for high dimensional data as well. Strictly speaking, even when $d \geq n$, these two depth functions are computable, and in particular, exact computation of the spatial depth is always possible regardless of the value of d. In all, we can define spatial depth (SD) based α-trimmed mean as

$$\bar{\mathbf{X}}_{SD,n,\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} \mathbf{X}_{(SD;i;n)},$$

where $\mathbf{X}_{(SD;i;n)}$ is the i-th order spatial depth based order statistic in $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. The definition of the spatial depth is provided in the Appendix.

Conjectures: *The assertions in Theorems 1 and 2 holds for $\bar{\mathbf{X}}_{SD,n,\alpha}$.*

To summarize, one may use $\bar{\mathbf{X}}_{SD,n,\alpha}$ to estimate the proportion of outliers in high dimensional data analogously as $\bar{X}_{n,\alpha}$ is used in the case of univariate data for the same reason. Moreover, we would like to mention that the trimmed mean can be defined in an infinite-dimensional setting as long as the ordering can be defined in that setting, which follows from the definition of the trimmed mean, and in the statistics literature, there are a few methodologies to order the data lying in infinite dimensional space (see, e.g., Lopez-Pintado and Romo (2009)). Consequently, the concept of this graphical device can be extended for infinite dimensional (e.g., functional data) as long as the ordering can be defined.

6 Concluding Remarks

In this article, we have proposed a new graphical device based on a well-known trimmed mean to estimate the proportion of outliers in the data. As we often encounter, many data have outliers, and using traditional estimators often leads to a wrong result as the presence of outliers was unknown beforehand. To overcome this problem, one may use the proposed graphical device to check whether the data has outliers or not, and if yes, then the proportion of the outliers can be estimated using this visualization toolkit. Once this information is in hand, the practitioners will be able to make decisions on how they will analyse the data. Next, we propose a future extension to the problem considered in the article for further research. Suppose that $\hat{\beta}_n$ is the estimated proportion of the outliers in the data using the proposed graphical device, where $\beta \in (0, \frac{1}{2})$ is the actual proportion of outliers in the data. One may now check whether $\hat{\beta}_n$ converges in probability or almost surely to β or not, and it is needless to mention that such theoretical issues will be of interest for future research.

Acknowledgments

The authors are thankful to the Editor-in-Chief Professor B. M. Golam Kibria and two anonymous reviewers for their excellent feedback on the previous draft of the article, which improved the article significantly. In addition, Subhra Sankar Dhar gratefully acknowledges his core research grant (CRG/2022/001489), Government of India.

References

Amin, M., Afzal, S., Akram, M. N., Muse, A. H., Tolba, A. H., and Abushal, T. A. (2022), "Outlier detection in gamma regression using Pearson residuals: simulation and an application," *AIMS Mathematics*, 7(8), 15331–15347.

Bhattacharya, S., Kamper, F., and Beirlant, J. (2023), "Outlier detection based on extreme value theory and applications," *Scandinavian Journal of Statistics*, 50(3), 1466–1502.

Bickel, P. (1965), "On Some Robust Estimates of Location," *The Annals of Statistics*, 36, 847–858.

Billingsley, P. (1999), "Convergence of Probability Measures," Wiley and Sons, New York.

Cabana, E., Lillo, R. E., and Laniado, H. (2021), "Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators," *Statistical Papers*, 62(4), 1583–1609.

Clark, K. M., and McNicholas, P. D. (2024), "Finding Outliers in Gaussian Model-based Clustering," *Journal of Classification*, 41, 313–337.

DasGupta, A. (2008), "Asymptotic Theory of Statistics and Probability," Springer Verlag, New York.

Dhar, S. S. (2016), "Trimmed Mean Isotonic Regression," *Scandinavian Journal of Statistics*, 43, 202–212.

Dhar, S. S., Chatterjee, U., and Shalabh. (2022), "A Note on Asymptotic Distribution of Trimmed mean," *Journal of the Indian Society for Probability and Statistics*, 23, 327–335.

Dhar, S. S., and Chaudhuri, P. (2009), "A Comparison of Robust Estimators Based on Two Types of Trimming," *AstA Advances in Statistical Analysis*, 93, 151–158.

Dhar, S. S., and Chaudhuri, P. (2012), "On the Derivatives of the Trimmed mean," *Statistica Sinica*, 22, 655–679.

Dhar, S. S., Dassios, A., and Bergsma, W. (2016), "A study of the power and robustness of a new test for independence against contiguous alternatives," *Electronic Journal of Statistics*, 10, 330–351.

Dhar, S. S., Jha, P., and Rakshit, P. (2022), "The trimmed mean in non-parametric regression function estimation," *Theory of Probability and Mathematical Statistics*, 107, 133–158.

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021), "Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning," *Statistical Science*, 36, 303–327.

Farnè, M., and Vouldis, A. (2024), "ROBOUT: a conditional outlier detection methodology for high-dimensional data," *Statistical Papers*, 65(4), 2489–2525.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., and Stahel, W. A. (1985), "Robust Statistics: The Approach Based on Influence Function," Wiley and Sons, New York.

Hogg, R. V. (1967), "A partial review and some suggestions for future applications and theory," *Journal of the American Statistical Association*, 69, 909–923.

Hu, M., Li, M., and Kong, L. (2024), "Trustworthy regularized Huber regression for outlier detection," *Journal of Statistical Computation and Simulation*, 94(5), 1121–1137.

Huber, P. (1981), "Robust Statistics," Wiley and Sons, New York.

Karoui, N. E., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013), "On robust regression with high-dimensional predictors," *PNAS*, 110, 14557–14562.

Leger, C., and Romano, P. (1990), "Bootstrap adaptive estimation: The trimmed-mean example," *The Canadian Journal of Statistics*, 18, 297–314.

Lehmann, E. L. (1983), "Theory of Point Estimation," Wiley and Sons, New York.

Liu, R., Parelius, J. M., and Singh, K. (1999), "Multivariate analysis by data depth: descriptive statistics, graphics and inference," *The Annals of Statistics*, 27, 783–858.

Lopez-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734.

Modarres, R. (2022), "Nonparametric tests for detection of high dimensional outliers," *Journal of Nonparametric Statistics*, 34(1), 206–227.

Rashid, A. M., Midi, H., Dhhan, W., and Arasan, J. (2022), "Detection of outliers in high-dimensional data using ν-support vector regression," *Journal of Applied Statistics*, 49(10), 2550–2569.

Murph, A. C., Strait, J. D., Moran, K. R., Hyman, J. D., and Stauffer, P. H. (2024), "Visualisation and outlier detection for probability density function ensembles," *Stat*, 13(2), e662.

Osorio, F., Gárate, Á., and Russo, C. M. (2024), "The gradient test statistic for outlier detection in generalized estimating equations," *Statistics and Probability Letters*, 209, 110087.

Pratap, U., Canudas-de-Wit, C., and Garin, F. (2021), "Outlier detection and trimmed-average estimation in network systems," *European Journal of Control*, 60, 36–47.

Shalabh, Dhar, S. S., and Rajeshbhai, S. P. (2024), "Statistical Data-Driven Modelling and Forecasting: An Application to COVID-19 Pandemic," *Annals of Data Science*, https://doi.org/10.1007/s40745-024-00583-8.

Smiti, A. (2020), "A critical overview of outlier detection methods," *Computer Science Review*, 38, 100306.

Song, Y., Fang, M., Wang, Y., and Hou, Y. (2024), "Rapid outlier detection, model selection and variable selection using penalized likelihood estimation for general spatial models," *Spatial Statistics*, 61, 100834.

Souiden, I., Omri, M. N., and Brahmi, Z. (2022), "A survey of outlier detection in high dimensional data streams," *Computer Science Review*, 44, 100463.

Stigler, S. M. (1973), "The asymptotic distribution of the trimmed mean," *The Annals of Statistics*, 1, 472–477.

Tukey, J. (1948), "Some elementary problems of importance to small sample practice," *Human Biology*, 20, 205–214.

Tukey, J., and McLaughlin, D. H. (1963), "Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization," *Sankhya*, 25, 331–352.

Vardi, Y., and Zhang, C. H. (2000), "The multivariate L_1 median and associated data depth," *PNAS*, 97, 1423–1426.

Wax, M., and Adler, A. (2024), "Outliers detection by signal subspace matching," *IEEE Transactions on Signal Processing*, 72, 2498–2511.

Zuo, Y., and Serfling, R. (2000), "General notions of statistical depth function," *The Annals of Statistics*, 28, 461–482.

Received: August 22, 2024

Accepted: January 5, 2025

Appendix A : Technical Details

Proof of Theorem 1: The arguments of the proof are similar to the proof of Theorem 1 in Dhar and Chaudhuri (2012). For the sake of completeness, the outline of the arguments is provided here.
It follows from DasGupta (2008) that

$$\bar{x}_{n,\alpha} - \theta(\alpha)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{F^{-1}(\alpha)1_{(x_i \le F^{-1}(\alpha))} + x_i 1_{(F^{-1}(\alpha) \le x_i \le F^{-1}(1-\alpha))} + F^{-1}(1-\alpha)1_{(x_i \ge F^{-1}(1-\alpha))}}{1-2\alpha} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Using (**??**) and Slutsky's theorem, for any arbitrary choices of $\alpha_1, \ldots, \alpha_k$ and arbitrary choices of $l_1, \ldots, l_k$, we have $\sum_{i=1}^{k} l_i \sqrt{n}(\bar{x}_{n,\alpha_i} - \theta_{\alpha_i})$ converges weakly to a certain Gaussian distribution. Hence, for any arbitrary choices of $\alpha_1, \ldots, \alpha_k$, $\sqrt{n}(\bar{x}_{n,\alpha_1} - \theta(\alpha_1), \ldots, \bar{x}_{n,\alpha_k} - \theta(\alpha_k))$ converges weakly to a k-dimensional multivariate normal distribution.
Now, in order to prove the tightness of $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$, one needs to verify the two conditions in Theorem 13.2 in Billingsley (1999). Condition 2 related to stochastic process $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$ $(\alpha \in (\alpha_1, \alpha_2))$ follows from Theorem A.1 in Leger and Romano (1990) considering $\sum_{i=1}^{n} 1_{(x_i \le .)}$ and F instead of $\hat{G}_n$ and F_n, respectively. Next, Condition 1 holds for the stochastic process $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$ $(\alpha \in (\alpha_1, \alpha_2))$ because the $\bar{x}_{n,\alpha}$ is the average of certain quantiles. Hence, the stochastic process $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$ $(\alpha \in (\alpha_1, \alpha_2))$ is tight.
Therefore, in view of the fact that for any arbitrary choices of $\alpha_1, \ldots, \alpha_k$, $\sqrt{n}(\bar{x}_{n,\alpha_1} - \theta(\alpha_1), \ldots, \bar{x}_{n,\alpha_k} - \theta(\alpha_k))$ converges weakly to a k-dimensional multivariate normal distribution and the stochastic process $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$ $(\alpha \in (\alpha_1, \alpha_2))$ is tight, we can conclude that the stochastic process $\sqrt{n}(\bar{x}_{n,\alpha} - \theta(\alpha))$ $(\alpha \in (\alpha_1, \alpha_2))$ converges weakly to a Gaussian process under sup norm topology. Hence, $\sup_{\alpha_1 < \alpha < \alpha_2} \sqrt{n}|\bar{X}_{n,\alpha} - \theta(\alpha)| = O_p(1)$, which completes the proof. □
Proof of Theorem 2: See Example 3 in pp. 99-100 in Hampel et al. (1985). □

Appendix B : Python Code

The Python code used to analyze business data in Section 4 is presented below.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import scipy

mean_sample = 0
mean_outliers = 100
std_dev_sample = 1
std_dev_outliers = 1
num_samples = 200
```

```
# Generate random samples from the normal distribution
samples = np.random.normal(mean_sample, std_dev_sample, int(num_samples*0.7))
outliers = np.random.normal(mean_outliers, std_dev_outliers, int(num_samples*0.3))

all_samples = np.concatenate((samples, outliers))
np.random.shuffle(all_samples)

plt.scatter(range(len(all_samples)), all_samples, color='blue', marker='o')

trim_value = np.arange(0, 0.5, 0.01)

# Calculate and store trimmed means for each trim percentage
trimmed_means = []
for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

plt.plot(trim_value, trimmed_means, color = 'black')
plt.xlabel('Trimming Proportion')
plt.yticks([])
plt.show()

# 70% samples from  N(0,1)  and 30% samples from  N(10,1)

mean_sample = 0
mean_outliers = 10
std_dev_sample = 1
std_dev_outliers = 1
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:

  samples = np.random.normal(mean_sample, std_dev_sample, int(n*0.7))
  outliers = np.random.normal(mean_outliers, std_dev_outliers, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)
```

```
  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()

# 70% samples from  N(0,1)  and 30% samples from  N(100,1)
mean_sample = 0
mean_outliers = 100
std_dev_sample = 1
std_dev_outliers = 1
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:

  samples = np.random.normal(mean_sample, std_dev_sample, int(n*0.7))
  outliers = np.random.normal(mean_outliers, std_dev_outliers, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
```

```
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()

# 70% samples from  Cauchy(0,1)  and 30% samples from  Cauchy(10,1)
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:

  samples = np.random.standard_cauchy(int(n*0.7))
  outliers = stats.cauchy.rvs(10, 1, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
```

```
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()

70% samples from Cauchy(0,1) and 30% samples from  Cauchy(100,1)
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:

  samples = np.random.standard_cauchy(int(n*0.7))
  outliers = stats.cauchy.rvs(100, 1, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()
```

```
# 70% samples from  Laplace(0,1)  and 30% samples from  Laplace(10,1)
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:

  samples = np.random.laplace(0, 1, int(n*0.7))
  outliers = np.random.laplace(10, 1, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()

# 70% samples from  Laplace(0,1)  and 30% samples from  Laplace(100,1)
num_samples = [50, 100, 200, 500]

trim_value = np.arange(0, 0.5, 0.01)
trimmed_means_array = []
all_samples_list = []
for n in num_samples:
```

```
  samples = np.random.laplace(0, 1, int(n*0.7))
  outliers = np.random.laplace(100, 1, int(n*0.3))
  all_samples = np.concatenate((samples, outliers))
  all_samples_list.append(all_samples)

  trimmed_means = []
  for val in trim_value:
    trimmed_mean = stats.trim_mean(all_samples, val)
    trimmed_means.append(trimmed_mean)

  trimmed_means_array.append(trimmed_means)

plt.figure(figsize = (10, 8))
# Create a 2x2 grid of subplots
for i in range(len(num_samples)):

  plt.subplot(2, 2, i+1)
  plt.plot(trim_value, trimmed_means_array[i], label='Line 1', color='black')
  plt.xlabel('Trimming Proportion')
  # plt.ylabel('Y Axis')
  plt.yticks([])
  plt.title('num_samples = '+str(num_samples[i]))

# Adjust layout to prevent overlapping
plt.tight_layout()

# Display the plot
plt.show()

# Bike bike sales dataset
import pandas as pd

def convert_dollar_to_numeric(value):
    if isinstance(value, str) and '$' in value:
        value = value.replace('$', '').replace(',', '')
    try:
        if isinstance(value, pd.Timestamp):
            return value  # Return Timestamp as is
        return float(value)
    except ValueError:
        return value

def read_excel_with_dollars(filename):
    try:
        df = pd.read_excel(filename)
```

```
        for column in df.columns:
            df[column] = df[column].apply(convert_dollar_to_numeric)
        return df
    except Exception as e:
        print(f"Error: {e}")
        return None

# Provide the path to your Excel file
file_path = '/content/uncleaned bike sales data.xlsx'

data_frame = read_excel_with_dollars(file_path)

if data_frame is not None:
    print(data_frame)

data_frame["Revenue"]

trimmed_means = []
trim_value = np.arange(0, 0.5, 0.01)
for val in trim_value:
    trimmed_mean = stats.trim_mean(data_frame["Revenue"], val)
    trimmed_means.append(trimmed_mean)

plt.figure(figsize = (10, 8))
plt.plot(trim_value, trimmed_means, label='Line 1', color='black')
plt.xlabel('Trimming Proportion')
plt.yticks([])
plt.title("Business Data")

plt.show()
```

Definition of Spatial Depth : For a given data $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, where for $i = 1, \ldots, n$, $\mathbf{X}_i \in \mathbb{R}^d$ ($d \geq 1$), the spatial depth at any fixed point $\mathbf{x} \notin \mathcal{X}$ is defined by

$$SD(x) = 1 - \left|\left| \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{x} - \mathbf{X}_i}{||\mathbf{x} - \mathbf{X}_i||} \right|\right|.$$

For any fixed point $\mathbf{x} \in \mathcal{X}$ (say $\mathbf{X}_i$), the spatial depth is defined as

$$SD(\mathbf{X}_i) = 1 - \left|\left| \frac{1}{n} \sum_{j=1, j\neq i}^{n} \frac{\mathbf{X}_i - \mathbf{X}_j}{||\mathbf{X}_i - \mathbf{X}_j||} \right|\right|.$$