ISSN: 2317-0840

Scaling the k_{max} criterion in the DDCAM methodology

Thayssa Marum Rezende[1] , Josino José Barbosa[1†] Helgem de Souza Ribeiro Martins[1] , Gabriel Lima de Souza[1] , Anderson Ribeiro Duarte[1] .

[1]*Universidade Federal de Ouro Preto, Instituto de Ciências Exatas e Biológicas, Departamento de Estatística; Ouro Preto-MG, Brasil.*

Abstract: *Outlier detection in multivariate data is a critical challenge with applications in fields such as finance, medicine, and industrial process monitoring. This study investigates the Data-Driven Cluster Analysis Method, designed to improve the identification of atypical observations through adaptive clustering strategies. Specifically, the research examines the role of the stopping criterion k_{max} — the maximum number of clusters considered — in determining the method's efficiency and accuracy. Using Monte Carlo simulations with contaminated normal, exponential, and point mass distributions, the study evaluates whether excessively large k_{max} values contribute meaningfully to model performance or merely increase computational cost. Results demonstrate that the optimal number of clusters, selected via the Bayesian Information Criterion (BIC), consistently falls well below the imposed k_{max} threshold, regardless of dimensionality, or contamination level. Furthermore, as sample size increases, the gap between the selected k and the k_{max} limit widens, while runtime grows proportionally. These findings suggest that overly conservative settings for k_{max} are unnecessary and can be replaced by more parsimonious values without compromising detection accuracy. The study reinforces DDCAM's robustness and stability while highlighting opportunities for computational optimization.*

Keywords: *Multivariate outliers; Monte Carlo simulation; Cluster analysis; Data-Driven Cluster Analysis Method.*

Introduction

Statistical analysis of data sets, regardless of their size, requires a careful descriptive approach. It is crucial to conduct research before applying more advanced statistical techniques. We must pay particular attention to identifying unusual observations, often referred to as outliers, which can significantly impair the statistical analysis (Montgomery and Runger, 2019).

Detecting outliers in univariate data is not a trivial task from a theoretical standpoint. The complexity increases even further when searching for multivariate outliers, that is, outliers in data sets involving two or more variables analyzed simultaneously (Barbosa et al., 2020).

According to Hawkins (1980), outliers are observations or sets of information that appear inconsistent when compared to the rest of the data set. Barnett and Lewis (1994) conceptualize outliers as observations that arouse suspicion because they are significantly deviated from the rest of the data set, suggesting the possibility that they were generated by a mechanism other than that which typically generates the data.

The concept of multivariate outliers refers to the distance in k-dimensional subspaces defined by the variables involved in the analysis. According to Jolliffe and Cadima (2016), an observation can be considered a multivariate outlier even if it is not a univariate outlier, that is, even if it is not an atypical observation concerning each of the variables individually considered.

[†]Autor correspondente: josino@ufop.edu.br
Received: 25/08/2025. Revised: 09/10/2025. Accepted:10/10/2025.

A large number of studies across various fields employ methodologies to detect outliers. Many of these applications are based on data generated by production processes. When these processes exhibit anomalous behavior, outside the usual pattern of predictability, outlier values are produced Aggarwal (2017). This observation highlights the significant importance of outlier analysis. Valuable information about rare but present characteristics in the data under investigation is identified and influences data generation processes.

Developing strategies to detect these unusual characteristics has numerous practical applications. Aggarwal (2017) provides examples of several such applications, including fraud detection in the financial system and credit transactions, medical diagnostic procedures, and sensors for event detection, among others.

The study proposed by Barbosa et al. (2018) presents a multivariate outlier detection technique performed through cluster analysis, called the cluster analysis method (CAM). Duarte et al. (2024) proposed a significant improvement on CAM, called the Data-Driven Cluster Analysis Method (DDCAM). DDCAM has shown considerable potential in multivariate outlier detection procedures compared to its predecessor technique, CAM, and other techniques based on the Mahalanobis distance.

Research related to multivariate outlier detection has broad applicability. The possibility of further investigating this approach is itself a significant motivation. The motivation for this study stems from the opportunity to evaluate potential enhancements to the outlier detection method proposed by Duarte et al. (2024). DDCAM is a sophisticated method that seeks to accurately and efficiently identify outliers in complex datasets through a combination of statistical and adaptive techniques. The adaptive selection of the number of clusters and the consideration of the distance from the centroid to the median of the data are unique aspects that contribute to the robustness and efficiency of the method in outlier detection.

The importance of outlier detection procedures motivates a wide range of studies. The various existing methods are generally validated through Monte Carlo simulations. In these studies, data are simulated with or without the presence of outliers, which are generated by simulation engines, and then the detection procedures are applied. In most cases, simulated test data are created by generating contaminated multivariate populations using the mixture of multivariate distributions technique through these widely used Monte Carlo simulations. The main objective of this study is to investigate the effects of the stopping criterion for choosing the number of cluster in DDCAM Duarte et al. (2024).

Background

Most studies on multivariate outlier detection have been based on the classical Mahalanobis distance. Rousseeuw and Van Zomeren (1990) presented a method that uses the robust minimum volume ellipsoid estimator (MVE). This estimator is obtained from the smallest volume ellipsoid that encompasses at least k sample points, where $n/2 < k < n$.

Filzmoser (2005) and Filzmoser et al. (2005) introduced a method based on the robust minimum covariance determinant (MCD) estimator. The MCD estimator is derived from a subset that minimizes the determinant of the sample covariance matrix. The size of this subset is defined by a window h, with $n/2 < h < n$. The mean of these h points serves as the location estimate, while the scale estimate is proportional to the corresponding covariance matrix. The width of the window h influences the robustness of the estimator.

Cerioli (2010) presented a comprehensive set of multivariate tests for outlier detection, based on the MCD estimator. The study also proposes an approximation for the robust distance distribution and introduces FSRMCD, a procedure based on the robust Mahalanobis distance.

Furthermore, a new iteration step was incorporated into the FSRMCD procedure, resulting in the iterated and reweighted MCD method, called IRMCD. The IRMCD method demonstrates greater detection power for samples with high levels of contamination, while FSRMCD focuses more on controlling false positives.

Leys et al. (2018) discusses the importance of accurately identifying multivariate outliers and criticizes the method that uses the basic Mahalanobis distance. This criticism is based on the fact that this indicator relies on the multivariate sample mean and the covariance matrix, which are particularly sensitive to the presence of outliers.

Zhu et al. (2017) discusses the challenges faced by real-time outlier detection algorithms when applied to GPS trajectories. The work aims to identify anomalous routes using historical trajectory data and popular routes, simultaneously considering irregularities in space and time. As a solution, the authors propose the time-dependent popular routes-based (TPRO) algorithm, which combines an offline preprocessing step with an online detection step, adopting a real-time multivariate approach.

Kutsuna and Yamamoto (2017) proposes an outlier detection technique based on binary decision diagrams. They introduce leave-one-out density as a new metric, calculated by the ratio of the number of data points in a region to its volume after excluding the point of interest. The approach has been applied to both synthetic data and data sets used in machine learning, and the results demonstrate high efficiency on large data volumes due to its near-linear computational complexity.

Luo et al. (2018) presents a multivariate outlier detection approach for medical image vectors using variograms. This geostatistical tool is used to assess the spatial dependence of displacement vectors. Because atypical vectors exhibit distinct spatial correlation patterns compared to valid vectors, they can be effectively detected as outliers using variogram analysis.

Wang et al. (2019) introduces the virtual outlier score (VOS), a novel multivariate outlier detection model. The technique constructs a similarity graph based on the k nearest neighbors and incorporates a virtual node connected by fictitious edges. A personalized Markov random walk is then performed on this graph, assuming that outliers are more likely to be visited. The method has been tested on synthetic and real datasets, demonstrating superior performance compared to state-of-the-art algorithms according to the ROC metric.

Wang and Mao (2019) argue that, in industrial settings, ensuring product quality and process performance requires the accurate identification of outliers. They propose algorithms based on Gaussian process models, using specific mean, covariance, likelihood, and inference functions. Three regression models and one classification model were developed. The results demonstrate that these approaches require fewer assumptions than traditional methods and are more suitable for the industrial environment.

Wahid and Rao (2019) develops a distance-based outlier detection algorithm using the classical particle swarm optimization (PSO) algorithm. The approach assigns each data point a degree of outlier probability, calculated by the sum of the distances between the point and its k nearest neighbors. PSO is used to locate peripheral subspaces where outliers may be present, being effective for high-dimensional data.

Kamalov and Leung (2020) proposes an outlier detection technique in multivariate data based on principal component analysis. The method, called PCOut, exploits features of principal component decomposition. It stands out for its low computational cost and good applicability to large-dimensional datasets. PCOut performs well in both high- and low-dimensional contexts.

Lejeune et al. (2020) investigates outlier detection in multivariate functional data, which are composed of multivariate functions that depend on continuous variables, such as in time series. The complexity of the problem stems from the relevance of both individual behaviors and the dynamic correlation between parameters. The proposed solution uses differential geometric mappings to capture peripheral aspects and identify anomalies.

Several studies on outlier detection were discussed in this review. In particular, the study by Duarte et al. (2024) stands out for proposing the DDCAM approach, a data-driven strategy for identifying multivariate outliers. This proposal constitutes the main focus of this research and is presented in detail in the subsequent methodological section.

Methodological Tools

This research primarily focuses on investigating the potential for improving DDCAM Duarte et al. (2024). The objective is to present experimental justifications regarding the impact of the stopping criterion (k_{max}) in defining the number of clusters in DDCAM.

DDCAM methodology

DDCAM builds on the premises of its predecessor, the CAM technique, incorporating an adaptive component that allows the model to adjust based on specific characteristics of the analyzed data, thus justifying the term 'data-driven'. DDCAM differentiates itself by combining cluster analysis strategies with a dynamic mechanism for defining the number of clusters, which enhances outlier detection, particularly in complex datasets.

The DDCAM methodological structure consists of several stages, each contributing to increased accuracy in identifying outliers. The main steps of the process are:

- estimation of the parameter δ (maximum proportion of observations that can be considered outliers in the data set): an efficient estimator of this parameter (denoted by $\hat{\delta}$) is obtained through univariate analysis, which takes into account the sample mean and standard deviation;

- refinement I: through the k-means cluster analysis method, several possible values of k (number of clusters) are tested up to a maximum limit (k_{max}), determined by the ratio between the total number of observations (n) and the logarithm of this number. The validity of the clusters is assessed based on the rule that the smallest group cannot contain more than $\hat{\delta} \times n$ observations.

- refinement II: the set of k values is further refined, taking into account the distance between the cluster centroids and the median of the data, in addition to the limitation on the maximum size of each cluster, k values that do not simultaneously meet these criteria are eliminated;

- final selection of the k value: among the remaining k values, the Bayesian Information Criterion (BIC) is used to select the most appropriate number of clusters. The BIC penalizes overly complex models, helping to avoid overfitting and select a more parsimonious solution.

In summary, DDCAM represents an advanced methodology that combines statistical tools with adaptive mechanisms to promote more accurate outlier detection in complex contexts. Its ability to automatically adjust the number of clusters and consider the distribution

of centroids relative to the data median reinforces its robustness and efficiency in identifying outliers.

Evaluation of the effects of the stopping criterion (k_{max}) for choosing the value k

In the first DDCAM refinement process, the method begins by forming a cluster with $k = 2$. It verifies whether the group with the smallest number of elements contains at most $\hat{\delta} \times n$ elements. If this condition is met, the cluster with $k = 2$ is considered valid; otherwise, it is classified as invalid and will not be considered in the following steps. This procedure is repeated successively for $k = 3$, $k = 4$, and so on, up to the maximum value $k_{max} = {}^{n}/_{\log(n)}$. At the end of this process, only the clusters considered valid proceed to the second refinement.

It is clear that the definition of k_{max} directly influences DDCAM's runtime. The larger the sample size n, the larger the value of k_{max} and, consequently, the longer the method runtime. Given the growing need to handle large volumes of data, this study aims to investigate the optimal value of k, obtained through the BIC criterion (the last step of DDCAM), and analyze whether it is sufficiently far from the maximum threshold k_{max} to indicate an excessive and avoidable computational cost, without significant gains in performance or accuracy. To elucidate this issue, we used the simulation procedures described by Duarte et al. (2024) to perform several computational experiments.

In studies involving simulations with multivariate distributions, the multivariate normal distribution is commonly used in various contexts. This distribution plays a central role in multivariate analysis, particularly in simulations that aim to represent data with outliers. In these cases, a common approach is to use a mixture of multivariate normal distributions, known as a contaminated multivariate normal distribution.

The simulation starts from the definition of two mean vectors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and two covariance matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. The vector $\boldsymbol{\mu}_1$ is generated from a univariate standard normal distribution $\mathcal{N}(0, 1)$ for each of its coordinates. The vector $\boldsymbol{\mu}_2$ has its coordinates simulated from a $\mathcal{N}(\pm 2, 1)$, so that its values are shifted approximately two standard deviations about $\boldsymbol{\mu}_1$. For simplification, it is assumed that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, with this matrix being generated randomly based on the method proposed by Duarte et al. (2025), which allows adjusting the level of correlation between the variables, making the simulations closer to real scenarios. Considering a random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^p$, with a contaminated multivariate normal distribution, its distribution function can be expressed as: $\mathcal{F}_p = (1 - \delta)\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \delta\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Where $(1 - \delta)$ is the probability that the process is executed by $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, δ is the probability that the process is executed by $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\boldsymbol{\mu}_i$ is the vector of means with $i = 1, 2$ and $0 \leqslant \delta \leqslant 1$.

In addition to normally distributed data, the study also addressed the generation of data with a contaminated multivariate exponential distribution. To achieve this, a constant mean vector equal to 1 and a covariance matrix Σ were used. We achieved the contamination by shifting the exponential distribution by two units, equivalent to two standard deviations. Details on the density of the multivariate exponential distribution can be found at Marshall and Olkin (1967).

Under the contamination condition, the resulting distribution function is given by $\mathcal{F}_p = (1 - \delta)\mathcal{D}_p + \delta\mathcal{D}_p^*$, where $\mathcal{D}_p$ represents the original distribution and $\mathcal{D}_p^*$ the distribution of the contaminated data. Point mass contamination, as described by Ruwet and Haesbroeck (2011) Ruwet and Haesbroeck (2011), was also considered. In this case, $\mathcal{D}_p^*$ corresponds to a Dirac Delta distribution Δ_x, concentrating all the probability in a single point. In the simulation procedure, first n observations are generated from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, to simulate point contamination, for each coordinate $i \in \{1, \ldots, p\}$, a fixed value two standard deviations away from the mean $\mu_i \pm 2\sigma_i$ is defined, where μ_i is the i-th coordinate of the vector $\boldsymbol{\mu}$ and σ_i^2 is the value of the

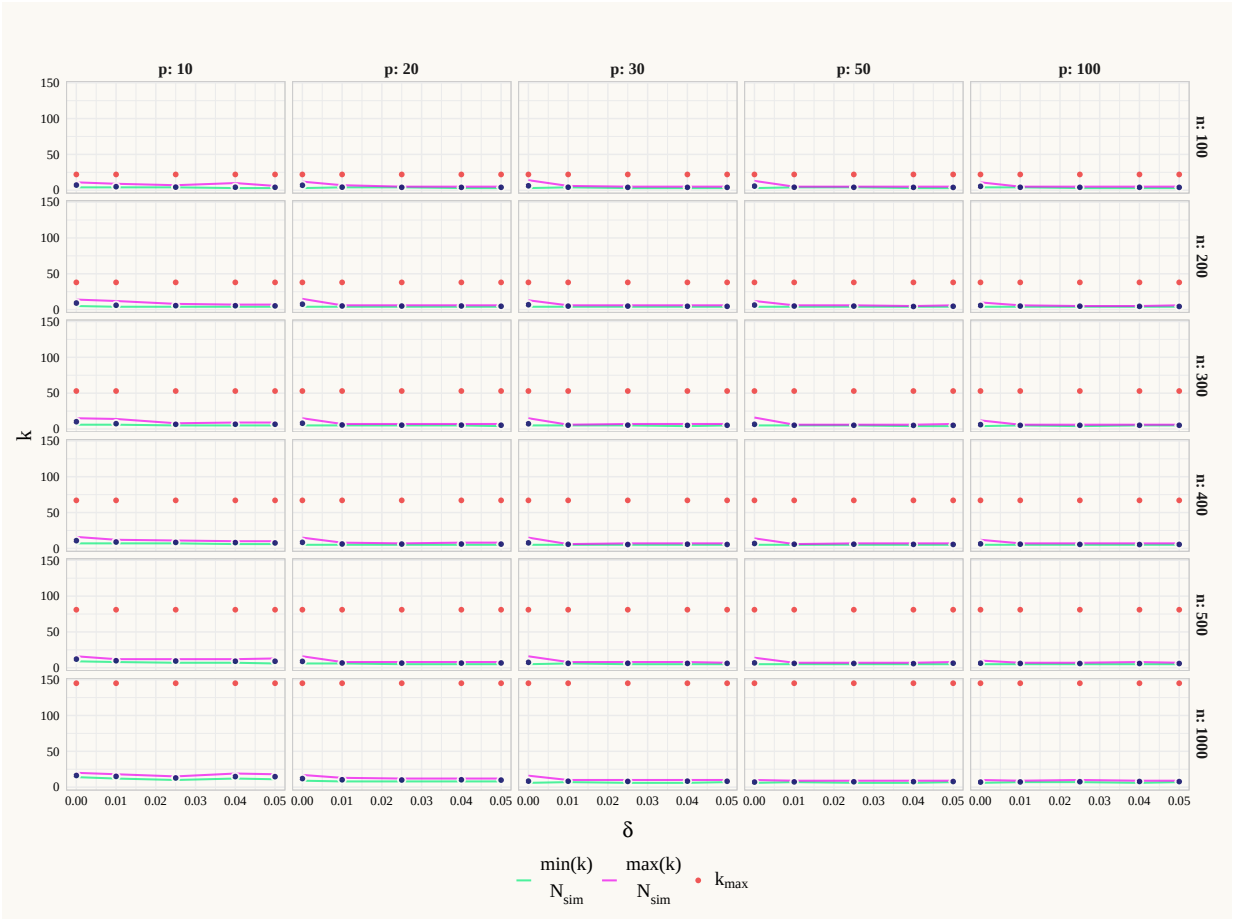
i-th entry on the diagonal of $\boldsymbol{\Sigma}$. A random draw determines whether the displacement will be positive or negative. Finally, the $\delta \times n$ contaminated observations in the sample are replaced by this fixed point $(x_1, \ldots, x_p)$.

Numerical Results and Insights

Our computational experiments, therefore, consider the three data models described above: contaminated normal, contaminated exponential, and point mass contamination. All codes used were implemented in the statistical software R R Core Team (2025). The computational experiments were conducted on a 2.80 GHz Intel® Core™ i7-1165G7 processor with Windows 11 Home 64-bit and 8 GB of RAM.

The experimental procedure with p-dimensional normal data was conducted with $p \in \{10, 20, 30, 50, 100\}$, the mixing ratios were $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}05\}$, the sample sizes were $n \in \{100, 200, 300, 400, 500, 1000\}$, for all scenarios, $N_{sim} = 200$ runs were performed. Figure 1 presents the results of the experiments. In the columns, we present the number of variables considered in the simulated database, representing the values of p for p-varied data. In the rows, we indicate the sample sizes used in the simulation, corresponding to the values of n.

Figure 1: Observed k values for normal data.

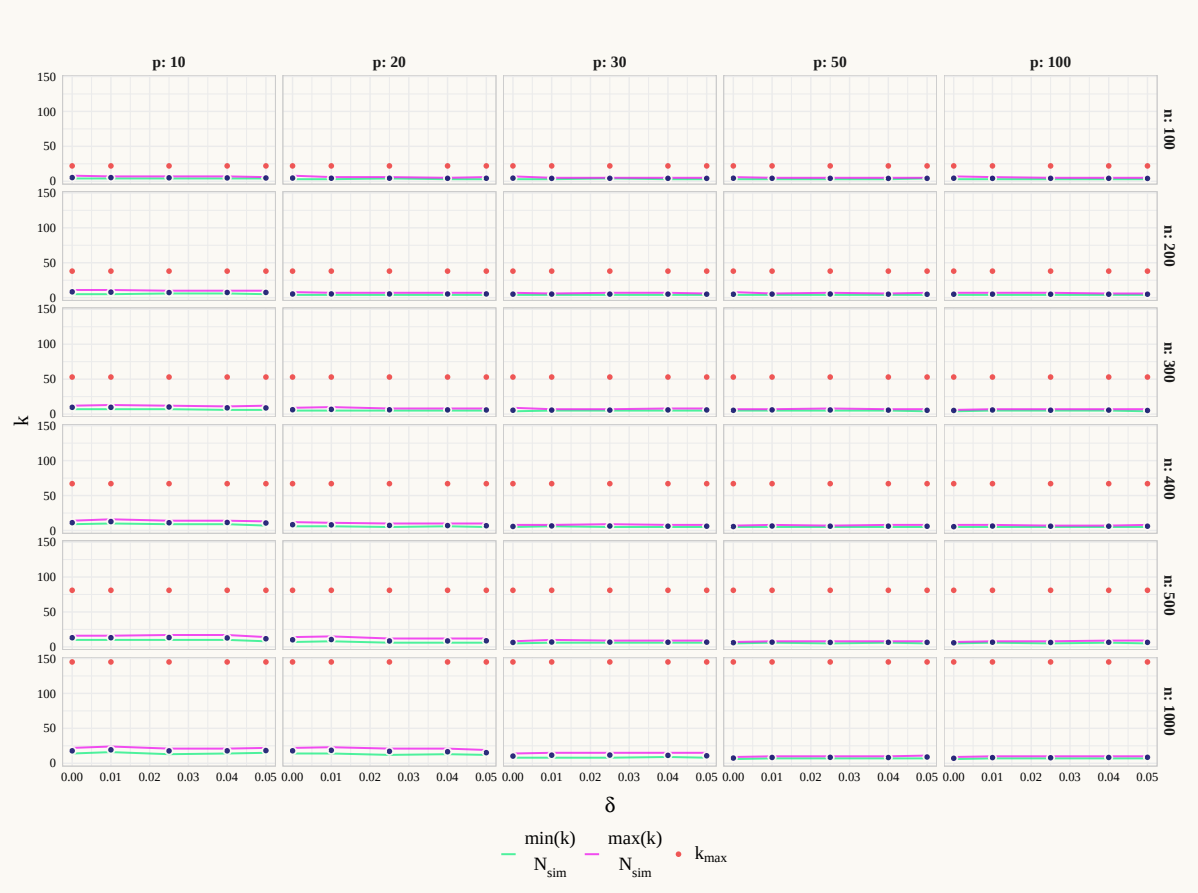


Source: from the authors (2025).

The results obtained with normally distributed data indicate that, in all scenarios analyzed, the minimum, maximum, and mean values of k are very close to each other. This behavior demonstrates the method's stability in selecting the most appropriate value of k using the BIC criterion. It is also observed that the number of variables has a minimal influence on this process, underscoring the robustness of our method. As the sample size increases, the optimal value of k, as defined by the BIC, remains significantly below the maximum threshold k_{max}. The distance between the chosen value k and the threshold k_{max} increases noticeably with the sample size.

A second experimental procedure was conducted with p-dimensional exponential data and simulated exponential data, with $p \in \{10, 20, 30, 50, 100\}$, the mixing ratios were $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}05\}$, the sample sizes were $n \in \{100, 200, 300, 400, 500, 1000\}$. Figure 2 presents the results with the same row and column configuration verified in the normal data.

Figure 2: Observed k values for exponential data.

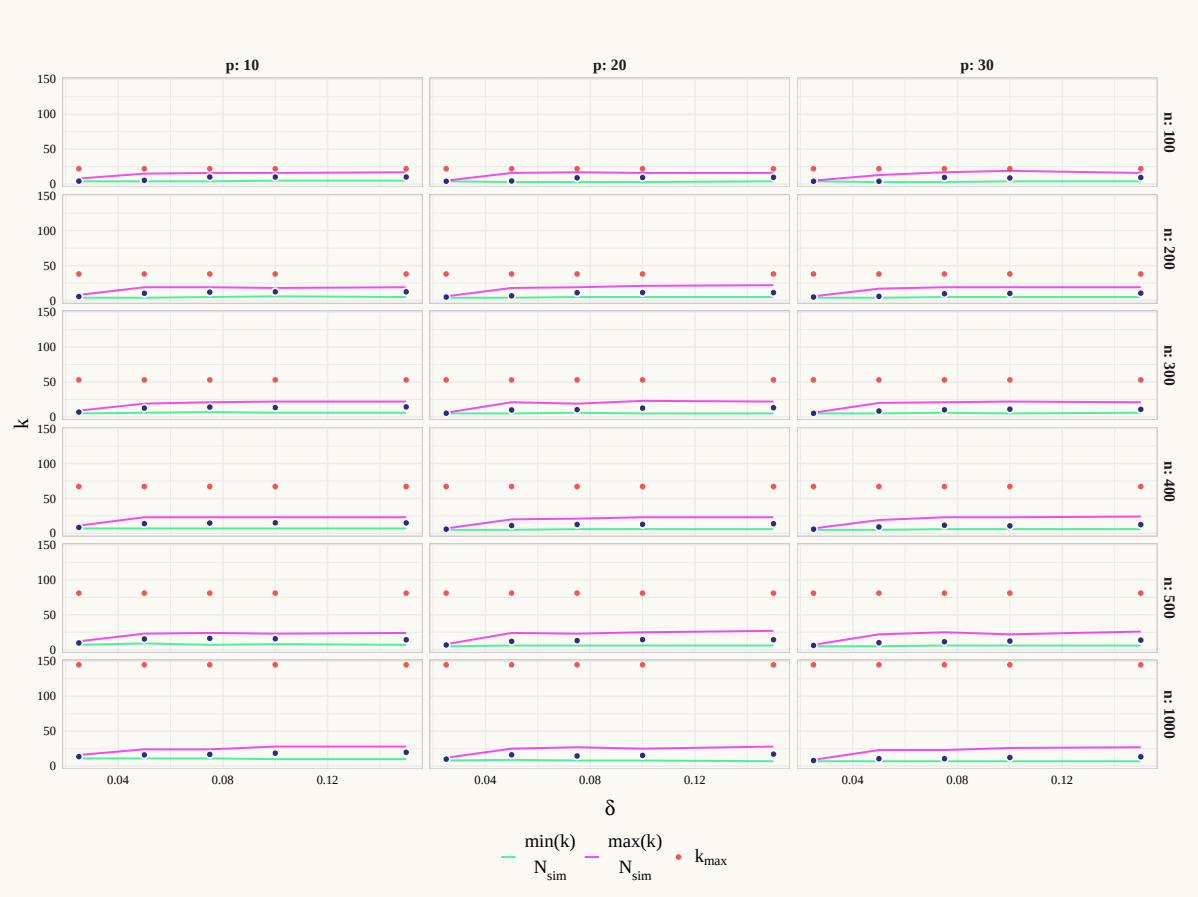


Source: from the authors (2025).

The results for exponentially distributed data also show that, for each scenario evaluated, both the smallest and largest values of k are very close to the mean k, once again confirming the stability of the DDCAM method in choosing the most appropriate k using the BIC criterion. Furthermore, it was also possible to observe that the number of variables (p) again has little influence on this process and that, as the sample size (n) increases, the optimal value of k, determined using the BIC criterion, is consistently well below the maximum limit k_{max}. Also, for exponential data, as the sample size increases, the distance between the chosen value k and the limit k_{max} increases significantly.

Finally, a third experiment was also conducted by us. We used point mass contamination model. The experiments were performed with simulated p-dimensional data, with $p \in \{10, 20, 30\}$, mixing ratios were $\delta \in \{0{,}025; 0{,}05; 0{,}075; 0{,}10; 0{,}15\}$, and sample sizes were set at $n \in \{100, 200, 300, 400, 500, 1000\}$. For the probabilistic point mass contamination model proposed in this study, investigations with $\delta = 0$ are not applicable, as they become quite similar to the study of normal data with $\delta = 0$. For all scenarios, 200 runs were performed. Figure 3 presents the results of the experiments performed in the same previous visualization settings.

Figure 3: Observed k values for point mass contamination.



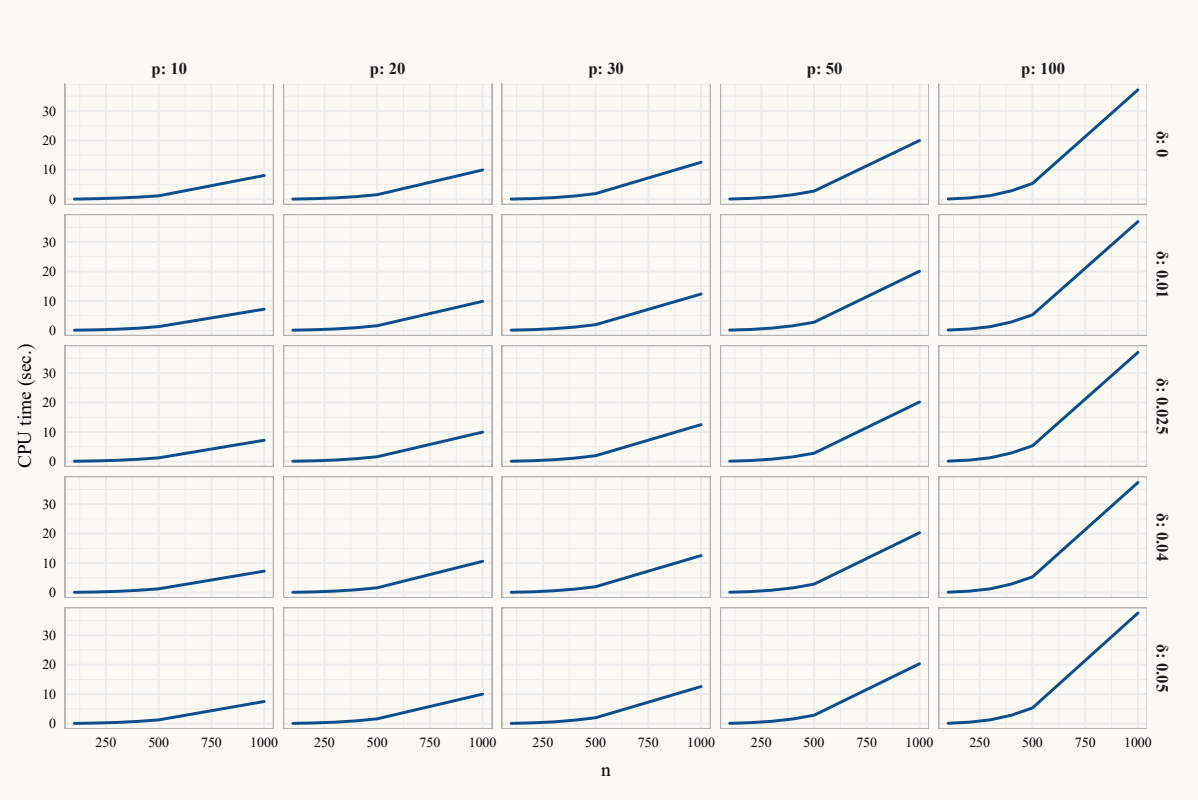
Source: from the authors (2025).

The results obtained for the data generated by point mass contamination also show that, in each scenario evaluated, across the 200 iterations, the minimum and maximum values of k remain close to the mean value of k. This behavior again confirms the robustness of the DDCAM method in defining the optimal number of clusters using the BIC criterion, regardless of this hyperparameter. Furthermore, it was observed that the number of variables (p) again had little influence on the process and that, as the sample size (n) increased, the optimal value of k, estimated by BIC, was consistently well below the upper threshold k_{max}, and this difference widened with increasing sample size.

After evaluating the behavior of the k_{max} criterion about the optimal value of k using the BIC criterion for the three proposed contaminations, we present an analysis of the CPU time required to execute the DDCAM method. Figure 4 presents the results of the CPU time consumed for the executions with normally distributed data. The other distributions do not present significant changes; therefore, their results will not be shown in this study. In the columns of

the Figure 4, we present the quantities (p) of variables considered. In the rows, we indicate the mixing ratios (δ) used.

Figure 4: CPU time consumed by DDCAM.



Source: from the authors (2025).

The results are presented in terms of average run time, considering the 200 simulations performed. As observed by Duarte et al. (2024), the mixing rate does not significantly impact the computational time of the DDCAM method. On the other hand, the processing time increases proportionally with the increase in both the number of variables and the sample size.

Increasing the sample size results in higher values for the parameter k_{max}, which directly influences the computational time required to run DDCAM. It is also observed that the optimal value of k, determined by the Bayesian Information Criterion (BIC), remains consistently much lower than the upper threshold k_{max}, especially as the sample size increases. These findings suggest that, in practice, setting excessively high values for k_{max} may represent an unnecessary and avoidable computational cost, without commensurate benefits in terms of method performance or accuracy.

Final Remarks

The main objective of this study was to comparatively investigate the stopping criterion (k_{max}) for choosing the number of clusters in cluster analysis using the multivariate outlier detection method Data-Driven Cluster Analysis Method (DDCAM), using the value k determined according to the Bayesian Information Criterion (BIC). We observed that defining k_{max} as $^{n}/_{\log(n)}$ directly impacts the method's runtime — primarily because it includes the sample size n in its definition — and that further discussion of this stopping criterion is crucial due to the possibility of analyzing large databases.

The results obtained through the simulation procedures demonstrate the impact on computational time of choosing k_{max}. Across the 200 runs, across all contaminated p-dimensional datasets analyzed (normal, exponential, and generated by point mass contamination) and in all scenarios considered (with different values of p, δ, and n), as the sample size increased, the optimal value of k consistently fell well below the maximum limit k_{max}, and the number of variables did not influence the results. Furthermore, we found that the minimum and maximum values of k always remained close to the mean value of k, demonstrating the stability of the DDCAM method in selecting the most appropriate value of k through BIC.

Furthermore, we performed an analysis of the effects on CPU time for the DDCAM execution. The results demonstrated that the mixing ratio δ did not significantly influence the CPU time consumed. However, the larger the sample size and the greater the number of variables, the more CPU time is required. The analysis once again demonstrated how increasing n impacts processing time, confirming the relevance of the research question addressed in this study.

In summary, we conclude that the current definition of the stopping criterion for selecting the number of clusters in the DDCAM method yields high values of k_{max}, without significant compensation in terms of method efficiency or accuracy. Our findings confirm that this configuration incurs an additional computational cost that is hardly justifiable. Therefore, further studies are needed to propose a new definition for k_{max} that reduces the impact of sample size on method runtime.

Conflicts of interest and the use of artificial intelligence

The authors declare that there is no conflict of interest in conducting this research, and the authors also confirm that no AI resources were used.

Acknowledgements

The authors would like to thank the Universidade Federal de Ouro Preto for partially supporting the development of this study.

References

C. C. Aggarwal. *An Introduction to Outlier Analysis*, pages 1–34. **Springer International Publishing**, 2017.

J. J. Barbosa, T. M. Pereira, and F. L. P. Oliveira. Uma proposta para identificação de outliers multivariados. ***Ciência e Natura***, 40(40):2–9, 2018.

J. J. Barbosa, A. R. Duarte, and H. S. R. Martins. A performance evaluation in multivariate outliers identification methods. ***Ciência & Natura***, 42:e16 1–14, 2020.

V. Barnett and T. Lewis. *Outliers in statistical data*, volume 3. **Wiley New York**, 1994.

A. Cerioli. Multivariate outlier detection with high-breakdown estimators. ***Journal of the American Statistical Association***, 105(489):147–156, 2010.

A. R. Duarte, J. J. Barbosa, H. S. R. Martins, and F. L. P. Oliveira. Data-driven cluster analysis method: a novel outliers detection method in multivariate data. ***Communications in Statistics-Simulation and Computation***, pages 1–21, 2024.

A. R. Duarte, H. S. R. Martins, and F. L. P. Oliveira. CM-generator: an approach for generating customized correlation matrices. ***Communications in Statistics-Simulation and Computation***, 54(2):510–529, 2025.

P. J. Filzmoser. Identification of multivariate outliers: a performance study. ***Austrian Journal of Statistics***, 34(2):127–138, 2005.

P. J. Filzmoser, R. G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. ***Computers & Geosciences***, 31(5):579–587, 2005.

D. M. Hawkins. *Identification of outliers*, volume 11. **Springer**, 1980.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. ***Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences***, 374(2065):20150202, 2016.

F. Kamalov and H. H. Leung. Outlier detection in high dimensional data. ***Journal of Information & Knowledge Management***, 19(01):2040013, 2020.

T. Kutsuna and A. Yamamoto. Outlier detection using binary decision diagrams. ***Data Mining and Knowledge Discovery***, 31(2):548–572, 2017.

C. Lejeune, J. Mothe, A. Soubki, and O. Teste. Shape-based outlier detection in multivariate functional data. ***Knowledge-Based Systems***, page 105960, 2020.

C. Leys, O. Klein, Y. Dominicy, and C. Ley. Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. ***Journal of experimental social psychology***, 74:150–156, 2018.

J. Luo, S. Frisken, I. Machado, M. Zhang, S. Pieper, P. Golland, M. Toews, P. Unadkat, A. Sedghi, and H. Zhou. Using the variogram for vector outlier screening: application to feature-based image registration. ***International Journal of Computer Assisted Radiology and Surgery***, 13(12):1871–1880, 2018.

A. W. Marshall and I. Olkin. A multivariate exponential distribution. ***Journal of the American Statistical Association***, 62(317):30–44, 1967.

D. C. Montgomery and G. C. Runger. *Applied statistics and probability for engineers*. **John wiley & sons**, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. **R Foundation for Statistical Computing**, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

P. J. Rousseeuw and B. C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

C. Ruwet and G. Haesbroeck. Impact of contamination on training and test error rates in statistical clustering. ***Communications in Statistics—Simulation and Computation***, 40(3):394–411, 2011.

A. Wahid and A. C. S. Rao. A distance-based outlier detection using particle swarm optimization technique. In *Information and Communication Technology for Competitive Strategies*, pages 633–643. **Springer**, 2019.

B. Wang and Z. Mao. Outlier detection based on gaussian process with application to industrial processes. ***Applied Soft Computing***, 76:505–516, 2019.

C. Wang, Z. Liu, H. Gao, and Y. Fu. Vos: A new outlier detection model using virtual graph. ***Knowledge-Based Systems***, 185:104907, 2019.

J. Zhu, W. Jiang, A. Liu, G. Liu, and L. Zhao. Effective and efficient trajectory outlier detection based on time-dependent popular route. ***World Wide Web***, 20(1):111–134, 2017.