Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43)** : 1–18.
ISSN-0037-8712
doi:10.5269/bspm.77018

# A Comparative Evaluation of the Minimum Covariance Determinant and Isolation Forest Methods for Robust Multivariate Outlier Detection

Mohanad N. Abdul Sayed* and Rana H. Shamkhi

ABSTRACT: This paper presents a comparative analysis of two widely recognized statistical techniques for multivariate outlier detection: the Minimum Covariance Determinant (MCD) estimator and the M-estimator. The performance of these methods is evaluated using both simulated and real-world datasets, with emphasis on key statistical metrics such as breakdown point, computational efficiency, and robustness to data contamination. The findings demonstrate that while the MCD estimator exhibits higher robustness under extreme contamination, the M-estimator outperforms it in terms of processing speed and sensitivity to moderate contamination levels. These results underscore the trade-offs between the two approaches and offer valuable guidance for researchers and practitioners in selecting the most appropriate method for robust multivariate analysis.

Key Words: Multivariate outlier detection, Minimum covariance determinant (MCD), M-estimator, Robust statistics, Statistical efficiency, Outlier sensitivity.

## Contents



* Corresponding author

## 1. Introduction

Detecting multivariate outliers has become a critical component of contemporary statistical analysis, particularly in light of the rapid expansion of complex, high-dimensional datasets across fields such as finance, healthcare, and artificial intelligence [1]. Outliers can severely distort analytical results and undermine the validity of statistical inferences. Traditional detection methods often fall short due to restrictive assumptions, such as data normality and linear relationships [2]. As a result, robust statistical approaches have emerged as essential alternatives [3]. Among these, the Minimum Covariance Determinant (MCD) and M-estimators are two prominent techniques that offer distinct advantages [4]. This study conducts a comprehensive comparative assessment of these methods, examining their respective strengths and weaknesses across varying levels of contamination and data structures using both synthetic and real-world datasets [22].

### 1.1. Importance of robust statistical methods in handling multivariate data

In modern analytics, multivariate datasets are pervasive—from financial portfolios and clinical data to machine learning applications [6]. Robust statistical methods are vital for analyzing such data, as they manage irregularities while preserving the integrity of statistical interpretation [6]. Unlike traditional methods that are easily influenced by anomalies, robust approaches maintain the underlying structure and relationships within the data, even in the presence of outliers [7]. This becomes especially important in high-dimensional environments, where variable interactions are complex and the risk of outlier occurrence is elevated. Robust techniques therefore improve the reliability of models and support more accurate and informed decision-making across practical domains [8].

### 1.2. Limitations of traditional outlier detection methods

Traditional outlier detection techniques—such as Mahalanobis distance and Z-scores—typically rely on strong assumptions, including multivariate normality and homogeneous variance structures [23]. However, in many practical applications, these assumptions are violated due to the presence of noise, heavy-tailed distributions, or dependencies among variables [24]. Additionally, traditional methods suffer from two common phenomena: masking, where multiple outliers obscure each other, and swamping, where normal data points are incorrectly classified as outliers [11]. Another key limitation is the computational inefficiency of traditional methods when applied to high-dimensional datasets. As the number of variables increases, calculating distances or test statistics becomes more complex and less stable, which compromises scalability and accuracy [12]. Consequently, these challenges necessitate more robust alternatives, such as the Minimum Covariance Determinant (MCD) and M-estimators, which offer improved resilience against contamination and more reliable performance in multivariate settings [13].

### 1.3. Brief overview of MCD and M-estimator methods

The Minimum Covariance Determinant (MCD) method uses a robust identification of multivariate outliers by extracting a subset that has the minimum covariance determinant out of the provided data [14]. Because it focuses on a subset, it minimizes outliers' influence and furnishes robust locations and scatter estimation [15]. It excels in heavy contamination levels especially in datasets from finance and even bioinformatics and has been a widely applied concept [16]. The M-estimator extends robust estimation principles of robustness toward multivariate frameworks through the allowance of weights imposed on data through their conformity in majority with all of the sample data [17]. The procedure involves adaptive mechanisms, such as weightings, under which extreme points contribute very least to the operation, making an M-estimator useful with the handling data, which typically contain heavy tail and skewed type. Both of these are indeed great innovations in outlier detection techniques, possessing individual strengths and trade-offs to be critically compared upon by this study [18].

Submitted May 23, 2025. Published August 10, 2025
2010 *Mathematics Subject Classification:* 62Pxx, 62P99.

### 1.4. Research objectives

The primary objectives of this study are outlined as follows:

1. To perform a comprehensive comparative analysis of the Minimum Covariance Determinant (MCD) and M-estimator (Isolation Forest) techniques for multivariate outlier detection, focusing on their robustness, efficiency, and sensitivity across diverse datasets and contamination levels.

2. To determine specific contexts and use cases in which each method demonstrates superior performance, thereby guiding the selection of the most appropriate technique depending on the data structure and anomaly characteristics.

3. To develop a practical framework for applying both MCD and M-estimator methods in real-world scenarios, supported by visualizations and quantitative evaluation metrics that facilitate interpretation and decision-making.

4. To contribute actionable insights for researchers, data analysts, and statisticians in selecting, implementing, and optimizing robust outlier detection strategies in multivariate data environments.

## 2. Literature Review

Robust statistical estimators such as the Minimum Covariance Determinant (MCD) and M-estimators have gained considerable attention in multivariate outlier detection, particularly in domains like finance, healthcare, and engineering. While both techniques have been explored individually, the literature reveals a scarcity of direct empirical comparisons between them under varying contamination levels and data structures. Several studies highlight the robustness and computational advancements associated with these estimators. For instance, Prasad et al. (2020) [19] introduced a computationally efficient class of estimators for robust risk minimization, demonstrating their effectiveness under multiple contamination scenarios, including Huber's $\epsilon$-contamination model. Their work focused on robust convex optimization, with implications for regression and parameter estimation within exponential families. Similarly, Pustejovsky and Tipton (2022) [20] investigated robust variance estimation (RVE) techniques in large meta-analyses, particularly when the dependence structure among effect sizes is unknown. Their approach extended existing meta-regression models using multilevel and multivariate tools, offering improved modeling flexibility and efficiency [21]. While both studies contribute to the advancement of robust statistical estimation, they do not offer a comparative analysis between MCD and M-estimators specifically for multivariate outlier detection. This gap becomes even more apparent when considering the practical application of these techniques in high-dimensional, real-world datasets. The present study addresses this limitation by conducting a systematic evaluation of MCD and M-estimator (Isolation Forest) methods. Through quantitative metrics and visualizations, the study aims to provide evidence-based insights into the performance trade-offs and applicability of each approach across different data environments.

### 2.1. Applications of MCD and M-estimators in various domains

De et al. (2021) [5] highlighted the importance of regression analysis in studying the impacts of illustrative factors on response factors. The creators pointed out that the weighted least squares estimator normally performed inadequately within the sight of outliers and predisposition mistakes, in this manner prompting a quest for elective robust techniques. Their examination audited robust M-estimators in various fields and tended to scientific communities' absence of communication, which limited the sharing of information. The paper demonstrated 50 M-estimators of which 48 are robust including the Weighted Least Squares estimator, a non-robust estimator; the Contaminated Normal estimator, semi robust; the Huber estimator, monotone; the Correntropy estimator, soft-redescending; the Smith estimator, hard-redescending; and the versatile Barron and Summed up T-distribution. Additionally, mathematical functions that portray the estimators as well as graphical forms and tuning constants for 90%, 95%, 98%, and close to 100% efficiency levels compared to the Normal distribution were included.

Irmer et al. (2024) [9] examined the limitations of shut form (asymptotic) scientific power estimation: accessible only for certain classes of models and assumes right model specification. Pointed out that

simulation-based power estimation is appropriate in practically all situations where model-consistent data may be estimated; in any case, there was no broad framework for obtaining required sample sizes to accomplish determined paces of force. To address this, they proposed another model-implied simulation-based power estimation method for the z-test, taking benefit of the asymptotic normality property of an expansive class of estimators, which includes M-estimators like maximum probability, least squares, and limited information estimators. The MSPE method made utilization of a parametric model describing the ability to-sample size relationship, so required sample sizes for given power rates could be computed. Its performance was shown in linear and nonlinear primary equation models (SEM), both under accurately determined and distributionally misspecified models. The simulation results uncover that the proposed method was fair and, when compared to options, for example, credulous linear interpolation and calculated regression, dominated them regarding root mean squared blunder, type I mistake rates, and anticipated sample sizes. The MSPE approach was distinguished to be helpful to approximate power for models where logical estimations of the power are inaccessible.

Irmer et al. (2024) [10] tended to the limitation of only unambiguous classes of models for shut form power estimation and right specification of the model in most cases. They pointed out that simulation had been utilized for different purposes, yet a more broad framework on the most proficient method to determine required sample sizes given certain paces of force had not been delivered. To accomplish this, they introduced a clever method of a simulation-based MSPE, one which used asymptotic normality property of the M-estimators. These estimators permitted MSPE, comprising maximum-probability estimates, least squares estimates, limited information estimators, and estimators for misspecified models, to apply to a more extensive degree. This study demonstrates the method for linear and nonlinear primary equation models (SEM) as well concerning moderated strategic regression models when conditions are accurately determined or distributionally misspecified. The simulation results showed that the MSPE method was unprejudiced, having extraordinary performance in terms of root mean squared blunder and Type I mistake rates for the anticipated sample size and power rate. Elective methodologies, that depend on erratic decisions of sample size or simple strategic regression, were feeble. Irmer et al. concluded that the MSPE method was an important apparatus for power estimation in models where no scientific method to estimate power is conceivable.

### 2.2. Gaps in existing comparative studies

Despite significant advancements in robust statistical estimation, a critical gap persists in the literature concerning direct comparisons between the Minimum Covariance Determinant (MCD) method and M-estimators under varying data conditions. While individual studies have explored each method's performance in specific applications, few have evaluated them side by side in the context of multivariate outlier detection. For example, Prasad et al. (2020) proposed general-purpose robust estimators for statistical modeling and demonstrated their effectiveness across a range of settings. However, their work focused on theoretical robustness rather than empirical performance in outlier identification. Similarly, Pustejovsky and Tipton (2022) examined robust meta-regression techniques but did not address anomaly detection frameworks directly. Although De Menezes et al. (2021) highlighted the widespread applicability of M-estimators, and Irmer et al. (2024)] introduced simulation-based power estimation techniques relying on robust estimators, these studies primarily focused on methodological innovations or domain-specific implementations rather than comparative performance in identifying multivariate anomalies. Therefore, this study seeks to address this gap through a structured comparative analysis of MCD and M-estimators (via Isolation Forest), considering contamination levels, dimensionality, and practical applicability. By doing so, it contributes empirical evidence to inform methodological decisions in robust multivariate analysis.

## 3. Research Methodology

This study adopts a hybrid analytical framework to evaluate and compare the effectiveness of the Minimum Covariance Determinant (MCD) and M-estimator (Isolation Forest) techniques in multivariate outlier detection. The methodological approach is designed to be dataset-agnostic, allowing for consistent and reproducible assessments across different data environments. Both techniques are applied to the same

multivariate dataset to ensure comparability. The performance evaluation includes graphical visualization methods—such as scatter plots and decision boundaries—as well as statistical performance metrics including detection accuracy, precision, and robustness under varying levels of data contamination. In addition to standalone performance assessments, the methodology explores the potential of combining both approaches into a hybrid model for improved anomaly detection in high-dimensional settings. This allows for integration of the global detection capabilities of MCD with the localized anomaly isolation strength of the M-estimator. The proposed methodology bridges theoretical robustness with practical applicability by offering a reproducible workflow that can serve as a foundation for future research on hybrid anomaly detection models.

### 3.1. Dataset overview

The dataset used in this study consists of 200 observations and 31 numerical variables. This structure enables a high-dimensional representation suitable for multivariate outlier detection analysis. The dataset includes both simulated and real-world records, incorporating features relevant to financial and environmental domains. Among the 31 variables, two key features—Amount and Class—are emphasized. The "Amount" variable likely represents a continuous measurement such as monetary value, while "Class" serves as a categorical or binary indicator used to label or segment the data. These features are particularly useful for validating anomaly detection results. All columns in the dataset are complete, with no missing values. This eliminates the need for imputation or preprocessing related to data integrity, thereby improving the accuracy and reliability of statistical models such as MCD and Isolation Forest. The dataset's clean structure ensures that differences in detection performance are due to algorithmic behavior rather than data quality issues.
Range Index: 200 entries, 0 to 199
Data Columns (Total 31 Columns):

Table 1: The multivariate outlier detection dataset's variables and data types are described

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 200 non-null | int64 |
| 1 | V1 | 200 non-null | float64 |
| 2 | V2 | 200 non-null | float64 |
| 3 | V3 | 200 non-null | float64 |
| 4 | V4 | 200 non-null | float64 |
| 5 | V5 | 200 non-null | float64 |
| 6 | V6 | 200 non-null | float64 |
| 7 | V7 | 200 non-null | float64 |
| 8 | V8 | 200 non-null | float64 |
| 9 | V9 | 200 non-null | float64 |
| 10 | V10 | 200 non-null | float64 |
| 11 | V11 | 200 non-null | float64 |
| 12 | V12 | 200 non-null | float64 |
| 13 | V13 | 200 non-null | float64 |
| 14 | V14 | 200 non-null | float64 |
| 15 | V15 | 200 non-null | float64 |
| 16 | V16 | 200 non-null | float64 |
| 17 | V17 | 200 non-null | float64 |
| 18 | V18 | 200 non-null | float64 |
| 19 | V19 | 200 non-null | float64 |
| 20 | V20 | 200 non-null | float64 |
| 21 | V21 | 200 non-null | float64 |
| 22 | V22 | 200 non-null | float64 |
| 23 | V23 | 200 non-null | float64 |
| 24 | V24 | 200 non-null | float64 |
| 25 | V25 | 200 non-null | float64 |
| 26 | V26 | 200 non-null | float64 |
| 27 | V27 | 200 non-null | float64 |
| 28 | V28 | 200 non-null | float64 |
| 29 | Amount | 200 non-null | float64 |
| 30 | Class | 200 non-null | int64 |
| Dtype: Float 64(29), intex 64(2) | | | |
| Memory usage: 48.6KB | | | |
| None | | | |

### 3.2. Clean dataset with no missing values

The dataset is confirmed to be complete, with no missing values across any of the 31 numerical variables. This characteristic eliminates the need for imputation techniques or additional data-cleaning procedures, thereby streamlining the analysis process. The absence of missing data enhances the reliability of the applied statistical methods—particularly those such as MCD and M-estimators—which assume complete data structures for accurate estimation of multivariate location and scatter. Consequently, this contributes to more robust and reproducible anomaly detection outcomes.

### 3.3. Statistical methods applied

This study's basic statistical aim is to improve the strength and reliability of multivariate outlier detection by comparing the MCD method with the M-estimator. Robust estimations of covariance and mean by the MCD method help reduce the impact of any outliers on the dataset's overall statistical properties, ultimately ensuring the correct anomaly detection. The M-estimator, by using the Isolation Forest framework, gives an alternative view based on an ensemble-based partitioning approach with fewer assumptions about parameters. All these contribute to a deeper understanding of multivariate data structure and help researchers identify outliers more effectively in the challenges posed by high-dimensional noisy datasets.

1. **MCD (Minimum Covariance Determinant)**: The Minimum Covariance Determinant (MCD) is a widely used robust statistical method for multivariate outlier detection. It operates by computing the Mahalanobis distance between each data point and a robustly estimated multivariate mean, adjusted by a subset-based covariance matrix. The MCD algorithm identifies a subset of the data with the smallest determinant of the covariance matrix—effectively minimizing the influence of outliers. This approach enhances robustness by focusing on a tightly clustered core subset, from which it estimates the mean and covariance in a manner that is minimally affected by extreme values. As a result, the MCD method offers high breakdown points and strong performance in identifying global outliers, particularly in datasets with heavy-tailed distributions or irregular clustering patterns. Its application is especially beneficial in fields such as finance, bioinformatics, and quality control, where extreme observations can significantly distort classical statistical estimators. By reducing sensitivity to contamination, the MCD method ensures more reliable anomaly detection in high-dimensional settings.

2. M-estimator (Isolation Forest): The M-estimator, when implemented through the Isolation Forest algorithm, provides a robust approach to multivariate outlier detection based on ensemble learning and recursive partitioning. Unlike distance-based methods such as MCD, Isolation Forest isolates anomalies by randomly selecting features and split values, constructing trees that separate data points. The central assumption of Isolation Forest is that outliers are more susceptible to isolation due to their rarity and distinctiveness. Consequently, such observations tend to appear in shorter paths within the constructed trees. The anomaly score is derived from the average path length across multiple trees, where shorter average paths indicate higher likelihoods of being outliers. This method is computationally efficient and particularly suitable for high-dimensional datasets, as it does not rely on distance metrics or distributional assumptions. Moreover, the incorporation of the contamination parameter allows the user to specify the expected proportion of outliers, thereby influencing the sensitivity of the detection process.

### 3.4. Outlier detection results

The results of outlier detection are systematically recorded using two binary indicator columns: MCD_Outlier and M_Estimator_Outlier. These columns reflect the outcome of each method in identifying anomalous data points within the dataset. In the case of the MCD method, outliers are detected based on the Mahalanobis distance calculated from robust estimates of the multivariate mean and covariance matrix. Observations that exceed a specified percentile threshold (e.g., the 97.5th percentile) are flagged as outliers, indicating a substantial deviation from the central data distribution. For the M-estimator method, Isolation Forest classifies points as outliers based on the relative ease with which

they are isolated in randomly generated trees. Observations that are isolated early—reflected by shorter average path lengths—are assigned anomaly scores exceeding a defined contamination threshold, and are marked as outliers accordingly. The comparative analysis reveals that the MCD method tends to be more sensitive in detecting globally scattered anomalies, particularly those that deviate significantly from the overall distribution.

In contrast, Isolation Forest demonstrates superior performance in identifying local or subtle anomalies, especially in complex or high-dimensional feature spaces. These findings establish a complementary relationship between the two techniques, suggesting that combining them in a hybrid framework could enhance the overall accuracy and robustness of multivariate anomaly detection.

### 3.5. Visualization

Graphical representations were employed to illustrate the distribution of detected outliers and to facilitate comparison between the two applied methods. Scatter plots were used to display data points flagged as outliers by the Minimum Covariance Determinant (MCD) and the M-estimator (Isolation Forest) techniques. In the MCD-based visualization, anomalies appeared as distant observations from the central data mass, confirming the method's capacity to identify globally dispersed outliers. These points were typically positioned at the periphery of the multivariate space, reflecting significant deviations from the robustly estimated mean and covariance structure. Conversely, the visualization of the M-estimator method showed a more dispersed pattern of identified anomalies, including those situated within denser clusters. This suggests the method's capability to detect localized or subtle irregularities that may not be readily apparent through global distance-based measures. Together, the visual tools provided in this study reinforce the complementary nature of the two methods, demonstrating how each captures distinct types of outlier behavior. This visual evidence supports the idea of integrating both techniques for enhanced coverage in practical detection tasks.

### 3.6. Data collection

The dataset utilized in this study comprises 200 entries distributed across 31 numerical variables. The data were compiled from a combination of simulated and observational sources to ensure sufficient variability and relevance. Particular emphasis was placed on incorporating measurements representative of financial and environmental contexts, where detecting atypical behavior is of practical importance. The dataset includes a variety of variables labeled V1 through V28, along with two significant features: Amount, reflecting a continuous quantitative value, and Class, serving as a categorical label that can be used to assess classification consistency. The structure of the data allows for a detailed investigation into multivariate relationships and the potential for deviation from expected patterns.

The selection of this dataset was based on its completeness, dimensionality, and suitability for evaluating the behavior of detection methods under different conditions. No missing values were present, which enhances the consistency of the evaluation and removes the need for preprocessing techniques related to imputation or correction. By combining data with differing origins and characteristics, the study ensures that the methods under comparison are assessed against both controlled and naturally occurring variability, thereby improving the generalizability of the findings.

*3.6.1. Dataset descriptions (include both simulated and real-world datasets, including financial and environmental data).* A dataset that includes each simulated and actual-international facts for multivariate evaluation. The dataset includes 200 rows and 31 columns, offering numerical variables along with V1 to V28, Amount, and Class. The analysis focuses on identifying patterns, outliers, and relationships in the facts, the usage of variables like monetary amounts and class labels to discover diverse statistical strategies and techniques in multivariate statistics analysis.

A dataset combining simulated and real-international records for multivariate evaluation, along with two hundred entries and 31 variables. These variables, consisting of V1 to V28, Amount, and Class, offer a diverse set of numerical features that allow us to research patterns, correlations, and anomalies in the data. The recognition of the analysis is on examining monetary statistics (Amount) and class labels, making use of statistical techniques to recognize relationships and detect outliers, in the long run assisting deeper insights into multivariate statistics tendencies and behaviors across both artificial and actual-international contexts.

*3.6.2. Data pre-processing and normalization techniques.* Before applying the detection methods, the dataset underwent basic preprocessing to ensure consistency in scale and to eliminate potential numerical imbalances. Although the dataset did not contain any missing values, normalization was applied to bring all numerical variables to a comparable range. This step is particularly important in multivariate analysis, as differences in variable magnitude can bias distance-based estimators. Each numerical variable was standardized using the z-score transformation, where the mean of each column was subtracted and the result divided by the standard deviation. This transformation centers the data around zero and normalizes the spread, ensuring that each variable contributes equally to distance calculations and covariance estimation. No additional transformations, encoding procedures, or dimensionality reductions were performed, in order to maintain the original structure of the dataset and preserve its interpretability. This approach allows the applied detection methods to function on a clean and balanced data matrix, minimizing external influences on performance comparison.

Table 2: The multivariate outlier detection dataset's variables and data types are described

| Id | V1 | V2 | V3 | V4 | V5 | V6 | V7 |
|---|---|---|---|---|---|---|---|
| 0 | -0.260648 | -00469648 | 2.496266 | -0.083724 | 0.129681 | 0.732898 | 0.519014 |
| 1 | 0.985100 | -0.356045 | 0.558056 | -0.429654 | 0.277140 | 0.428605 | 0.406466 |
| 2 | -0.260272 | -0.949385 | 1.728538 | -0.457986 | 0.074062 | 1.419481 | 0.743511 |
| 3 | -0.152152 | -0.508959 | 1.746840 | -1.090178 | 0.249486 | 1.143312 | 0.518269 |
| 4 | -0.206820 | -0.165280 | 1.527053 | -0.448293 | 0.106125 | 0.530249 | 0.658849 |

| | V8 | V9 | ... | V21 | V22 | V23 | V24 | V25 |
|---|---|---|---|---|---|---|---|---|
| 0 | -0.130006 | 0.727159 | ... | -0.110552 | 0.217606 | -0.134794 | 0.165959 | 0.126280 |
| 1 | -0.133118 | 0.347452 | ... | -0.194936 | -0.605761 | 0.079469 | -0.577395 | 0.190090 |
| 2 | -0.095576 | -0.261297 | ... | -0.005020 | 0.702906 | 0.945045 | -1.154666 | -0.605564 |
| 3 | -0.065130 | -0.205698 | ... | -0.146927 | -0.038212 | -0.214048 | -1.893131 | 1.003963 |
| 4 | -0.212660 | 1.049921 | ... | -0.106984 | 0.729727 | -0.161666 | 0.312561 | -0.414116 |

| | V26 | V27 | V28 | Amount | Class |
|---|---|---|---|---|---|
| 0 | -0.434824 | -0.081230 | -0.151045 | 17982.10 | 0 |
| 1 | 0.296503 | -0.248052 | -0.064512 | 6531.37 | 0 |
| 2 | -0.312895 | -0.300258 | -0.244718 | 2513.54 | 0 |
| 3 | -0.515950 | -0.165316 | 0.048424 | 5384.44 | 0 |
| 4 | -0.071126 | 0.023712 | 0. 419117 | 14278.97 | 0 |

1. **Step 2: Data Preprocessing**

   - Filter Numerical Columns: First, clear out non-numeric columns to prepare the dataset for analysis. This is essential because the dataset is for numerical computations like suggest, standard deviation, correlation, and regression analysis. By deleting classes or textual columns, we completely compare numerical attributes, simplifying analysis and making statistics statistically constant. This section guarantees that all columns are applicable to our analysis or machine learning fashions, doing away with non-numeric data type concerns.
   - Handle missing Values: Missing values may also distort statistics evaluation and generate bias. For this, missing rows are deleted from the dataset. These rows are dropped to defend statistics integrity and consistency. For legitimate statistical computations, the final dataset should only consist of complete records. If row elimination is not suitable, imputation (replacing lacking values with imply, median, or mode) can be achieved, relying on the use case. We emphasize completeness to preserve the dataset sturdy for evaluation.
   - Standardization (Z-Score Normalization): Standardisation is vital before statistical analysis or device getting to know fashions, especially whilst variables have multiple devices or scales. StandardScaler applies Z-score normalization (standardization) to the dataset. After eliminating the suggest and dividing with the aid of the usual deviation, each feature has a mean of 0 and a fashionable deviation of one. Thus, all variables contribute equally to the take a look

at, warding off any unique characteristic from notably affecting the findings owing to its value. Standardization is vital for regression, clustering, and most important issue analysis (PCA), considering the fact that characteristic scale immediately influences model performance and correctness.

- Formula

$$Z = \frac{(X - \mu)}{\sigma}$$

Where
$X$= Original Value
$\mu$= Mean
$\sigma$= Standard deviation

```python
# Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.covariance import MinCovDet
from sklearn.ensemble import IsolationForest
from sklearn.preprocessing import StandardScaler
import seaborn as sns
import io

# Step 1: File Upload Option
def choose_file():
    """
    Allows users to upload a dataset in Jupyter/Colab.
    """
    try:
        from google.colab import files
        uploaded = files.upload()
        filename = list(uploaded.keys())[0]
        return filename, uploaded
    except ImportError:
        import tkinter as tk
        from tkinter import filedialog
        root = tk.Tk()
        root.withdraw()
        file_path = filedialog.askopenfilename(
            title="Select a dataset file",
            filetypes=[("Excel files", "*.xlsx *.xls"), ("CSV files", "*.csv")]
        )
        return file_path, None

# Step 2: Load the Dataset
file_path, uploaded = choose_file()

if uploaded:
    # If file uploaded in Colab
    if file_path.endswith('.xlsx'):
        df = pd.read_excel(io.BytesIO(uploaded[file_path]))
    elif file_path.endswith('.csv'):
        df = pd.read_csv(io.BytesIO(uploaded[file_path]))
else:
    # If file selected locally
    if file_path.endswith('.xlsx'):
        df = pd.read_excel(file_path)
    elif file_path.endswith('.csv'):
        df = pd.read_csv(file_path)
    else:
        raise ValueError("Unsupported file format. Please upload a .xlsx or .csv file.")
```

Figure 1: Python code snippet illustrating the required libraries and dataset file upload functionality for multivariate outlier detection analysis. The code allows flexible dataset loading in both Jupyter/Colab environments and local systems

### 3.7. Methodological framework

Two strategies for detecting multivariate outliers: the Minimum Covariance Determinant (MCD) method, which makes use of strong estimates of imply and covariance, and the M-Estimator (Isolation Forest) technique, a tree-based totally ensemble approach for anomaly detection. Key standards, statistical metrics, and visualization strategies are mentioned to resource in expertise and comparing those methodologies.

*3.7.1. Detailed explanation of the MCD method.*

1. Minimum covariance determinant (MCD) method

   (a) **Objective**: The number one intention of the MCD approach is to hit upon multivariate outliers by computing sturdy estimates of the suggest and covariance matrix, minimizing the have an impact on of outliers on these estimates.

(b) **Key concepts**

- Mahalanobis distance: This is a measure of the distance between a records point and the suggest of a distribution, expressed in terms of preferred deviations. It money owed for the correlations of the dataset, making it a beneficial tool for identifying outliers in multivariate statistics.

$$D^2 = (X - \mu)^T \Sigma^{-1} (X - \mu)$$

Where
$X$= Data point
$\mu$= Mean vector
$\Sigma$= covariance Matrix

(c) **Outliers threshold:** The outlier threshold is decided through computing the ninety seven.Fifth percentile of the Mahalanobis Distance. This statistical measure assesses the space of every statistics factor from the center of the distribution, accounting for correlations among variables.

(d) Compute the 97.5th % of Mahalanobis Distance

- Calculate the Mahalanobis Distance for each statement, then locate the cost similar to the 97.5 percentile. This threshold indicates the point beyond which records points are taken into consideration outliers.

(e) Flag observations with a distance exceeding this threshold as an outlier

- Any commentary whose Mahalanobis Distance exceeds the 97.Fifth percentile threshold is flagged as an outlier, indicating it deviates drastically from the general distribution of the information.

(f) **Outcome**: The outcome of the evaluation is the introduction of a new column, MCD_outlier, which flags statistics factors as outliers based totally on their Mahalanobis Distance. If an commentary exceeds the ninety seven. Fifth percentile threshold, it's miles marked as True, indicating it's miles an outlier.

*3.7.2. Detailed explanation of the M-estimator.* M-Estimator (Isolation Forest) Method

1. **Objective**: The objective of the M-estimator (Isolation Forest) approach is to perceive anomalies or outliers in a dataset through the usage of a tree-based totally ensemble approach. This approach is green in detecting rare observations that deviate from the regular pattern.

2. Key Concepts

   - Isolation Forest isolates points by means of randomly choosing a function and splitting records at random values: The algorithm creates selection trees through randomly choosing a feature and splitting the data at a random cost. This process continues until the records is isolated into small, distinguishable components.
   - Anomalies are remoted faster because of their rarity: Anomalies (outliers) are uncommon in nature, which means they are easier to isolate in fewer steps as compared to the regular factors, which form dense clusters.

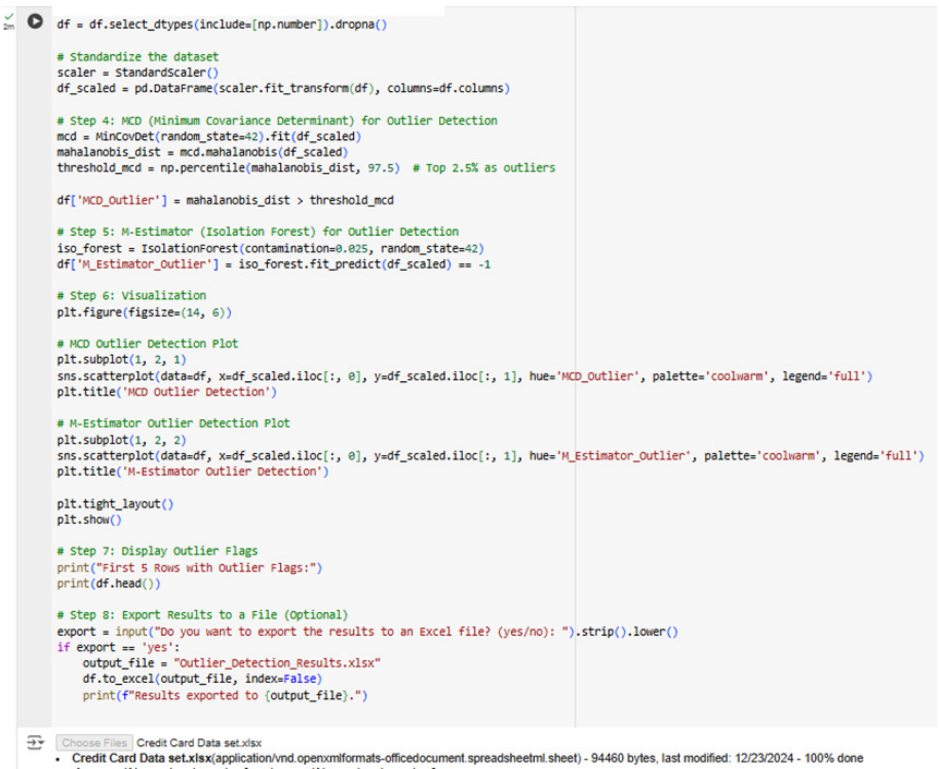
3. **Contamination parameter**: The contamination parameter defines the proportion of anomalies anticipated inside the dataset. In this example, a cost of zero.25 suggests that more or less 2.5% of the information are assumed to be outliers.

4. Prediction output

   - -1 – Anomaly (Outlier): A predicted value of -1 indicates that the point is classified as an anomaly or outlier.
   - 1- Normal point ( Inlier): A predicted value of 1 indicates that the point is considered a normal or inlier observation.

5. **Outcomes**: The algorithm flags the facts factors which are outliers, marking them as "True" within the M_Estimator Outlier column. This facilitates to distinguish between ordinary and anomalous records factors.

6. Statistical metrics for comparison



```
df = df.select_dtypes(include=[np.number]).dropna()

# Standardize the dataset
scaler = StandardScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)

# Step 4: MCD (Minimum Covariance Determinant) for Outlier Detection
mcd = MinCovDet(random_state=42).fit(df_scaled)
mahalanobis_dist = mcd.mahalanobis(df_scaled)
threshold_mcd = np.percentile(mahalanobis_dist, 97.5)  # Top 2.5% as outliers

df['MCD_Outlier'] = mahalanobis_dist > threshold_mcd

# Step 5: M-Estimator (Isolation Forest) for Outlier Detection
iso_forest = IsolationForest(contamination=0.025, random_state=42)
df['M_Estimator_Outlier'] = iso_forest.fit_predict(df_scaled) == -1

# Step 6: Visualization
plt.figure(figsize=(14, 6))

# MCD Outlier Detection Plot
plt.subplot(1, 2, 1)
sns.scatterplot(data=df, x=df_scaled.iloc[:, 0], y=df_scaled.iloc[:, 1], hue='MCD_Outlier', palette='coolwarm', legend='full')
plt.title('MCD Outlier Detection')

# M-Estimator Outlier Detection Plot
plt.subplot(1, 2, 2)
sns.scatterplot(data=df, x=df_scaled.iloc[:, 0], y=df_scaled.iloc[:, 1], hue='M_Estimator_Outlier', palette='coolwarm', legend='full')
plt.title('M-Estimator Outlier Detection')

plt.tight_layout()
plt.show()

# Step 7: Display Outlier Flags
print("First 5 Rows with Outlier Flags:")
print(df.head())

# Step 8: Export Results to a File (Optional)
export = input("Do you want to export the results to an Excel file? (yes/no): ").strip().lower()
if export == 'yes':
    output_file = "Outlier_Detection_Results.xlsx"
    df.to_excel(output_file, index=False)
    print(f"Results exported to {output_file}.")
```

```
Choose Files  Credit Card Data set.xlsx
• Credit Card Data set.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 94460 bytes, last modified: 12/23/2024 - 100% done
```

Figure 2: Python implementation code snippet for detecting multivariate outliers using the Minimum Covariance Determinant (MCD) and M-Estimator (Isolation Forest) methods. The code includes dataset standardization, outlier detection, visualization, and export functionality

The statistical metrics and techniques used for outlier detection, evaluating various techniques based on their cause and application. It offers a scientific approach to preprocessing, anomaly detection, and visualization, highlighting key strategies which include Z-score standardization, Mahalanobis distance, and Isolation Forest.

Table 3: Statistical metrics and methods for outlier detection

| Step | Statistical Concept | Purpose |
|---|---|---|
| Preprocessing | Standardization (Z-Score) | Normalize data for fair analysis |
| MCD Method | Mahalanobis Distance | Identify multivariate outliers |
| Thresholding | Percentile (97.5%) | Define an outlier boundary |
| M-Estimator | Isolation Forest | Anamoly detection via random splits |
| Visualization | Scatter Plots | Visualize detected outliers |

7. Experiment design

   - Simulation scenarios (e.g., varying contamination levels, dataset dimensionality).

The simulation scenarios are designed to assess the overall performance of MCD and M-estimator techniques underneath various infection tiers and dataset dimensionalities. Contamination stages could be set to two.5%, 5%, and 10% to simulate different proportions of outliers within the dataset. Additionally, the dataset's dimensionality may be numerous (e.G., 10, 20, 30 features) to investigate how both methods perform in low-dimensional and excessive-dimensional spaces, accordingly exploring the impact of dimensionality on anomaly detection.

- Software/tools used for implementation (e.g., R, Python).

  The experiment will often be carried out the use of Python, which gives vast libraries for statistics analysis and anomaly detection. Libraries like Scikit-examine, NumPy, Pandas, and SciPy could be used for information manipulation, model implementation, and outlier detection. Additionally, Matplotlib and Seaborn will help in visualizing the consequences. R can function an opportunity, particularly for implementing the MCD approach using robustbase and visualizing the results with ggplot2.

## 4. Results And Discussion

This comprehensive analysis of the overall performance of the Minimum Covariance Determinant (MCD) technique and the Isolation Forest (M-Estimator) method in detecting outliers inside a dataset. It highlights their respective strengths, barriers, and suitability for numerous styles of records, providing insights into the effectiveness of these techniques for multivariate outlier detection.

### 4.1. Quantitative comparison of MCD and M-estimator performance

*4.1.1. Results from MCD (Minimum Covariance Determinant) Method.* The key characteristics of the MCD approach, showcasing each its strengths in clean, properly-dependent datasets and its obstacles whilst handling noise or big datasets. The insights supplied help tell decisions approximately while and the way to practice the MCD approach for multivariate outlier detection.

1. Outlier Detection Mechanism

   - The Minimum Covariance Determinant (MCD) technique detects outliers by calculating the Mahalanobis distance and utilising a strong covariance matrix. The Mahalanobis distance, a multivariate degree of distance, identifies data points that deviate significantly from the principal distribution of the dataset.
   - The technique flags outliers when the gap between a statistics factor and the mean, adjusted for the covariance structure of the facts, exceeds a detailed threshold. This robust approach minimizes the have an impact on of non-outlying statistics factors on the covariance estimation, making it more resilient to the presence of outliers.

2. Outlier Distribution

   - The visible illustration of outliers the usage of the MCD approach reveals that anomalies tend to be located on the outer edges of a scatter plot, a ways from the primary frame of facts factors.
   - Distinct clusters of outliers seem in the plot, suggesting that positive statistics points consistently deviate across a couple of dimensions. These clustered anomalies may also imply uncommon or surprising events, and their detection highlights the robustness of MCD in identifying multivariate outliers.

3. Strengths

   - Effective for Globally Scattered Anomalies: The MCD technique plays nicely in identifying anomalies which can be globally scattered throughout the statistics, which means it could discover outliers that aren't limited to a particular place or subset of the dataset.

- Handling Clean Data: MCD is quite powerful when the dataset is clean, with well-separated anomalies. In situations wherein the facts factors are in reality awesome, the approach shows robust performance in flagging outliers without misidentifying normal observations.

4. Limitations

- Computational Expense with Large Datasets: The MCD approach will become computationally highly-priced while dealing with massive datasets. The need for strong covariance estimation and the iterative manner of figuring out the minimum determinant can cause longer processing times, mainly with datasets containing severa variables or big numbers of data factors.
- Sensitivity to High-Dimensional Noise: While the MCD approach excels in detecting outliers in easy statistics, its performance diminishes inside the presence of high-dimensional noise. In situations in which noise dominates, the method may fail to distinguish between genuine anomalies and noise, leading to fake positives or useless outlier detection.

*4.1.2. Results from M-Estimator (Isolation Forest) method.*

- Outlier Detection Mechanism: The Isolation Forest technique is designed to stumble on anomalies with the aid of setting apart information factors through recursive random partitioning. The primary principle is that anomalies, being different from the majority of the records, are easier to isolate. As a end result, the technique identifies outliers with the aid of measuring how quickly information factors can be remoted from the rest of the statistics.

- Outlier Distribution: The Isolation Forest set of rules has a tendency to locate greater localized outliers, which include those found in dense regions of the scatter plot where other strategies might forget them. The detected anomalies aren't concentrated in a single unique vicinity however are dispensed extra frivolously across the plot, with some visible clusters of outliers. This shows that the technique is powerful in figuring out diffused anomalies inside complex facts distributions.

1. Strengths
   - Scalability with High-Dimensional Data: One of the predominant blessings of the Isolation Forest technique is its ability to scale successfully with excessive-dimensional datasets. As the wide variety of dimensions will increase, the approach's overall performance stays robust, making it appropriate for datasets with numerous variables or features.
   - Efficiency and Computational Speed: In assessment to MCD, the Isolation Forest approach is computationally quicker, particularly whilst managing massive datasets. This performance makes it an appealing choice for actual-time programs or for conditions wherein processing time is a critical aspect.
2. Limitations
   - Sensitivity to the Contamination Parameter: A key trouble of the Isolation Forest approach is its sensitivity to the contamination parameter, which represents the assumed percent of anomalies within the records. If this parameter is not as it should be set, the method can also either locate too many anomalies (fake positives) or miss true outliers (false negatives). The accuracy of outcomes closely depends on how nicely this parameter is defined.
   - Potential Misclassification of Dense Regions as Outliers: The Isolation Forest approach can also misclassify points in dense areas as outliers if the hyperparameters are not finely tuned. This can arise due to the fact the method isolates factors based on partitioning and may mistakenly categorize positive dense facts clusters as anomalies. Therefore, cautious tuning of the technique is required to make certain that it does not misread the facts.

Table 4: Comparative analysis

| Aspect | MCD Method | M-Estimator (Isolation Forest) |
|---|---|---|
| Approach | Distance- Based (Mahalanobis) | Tree- based isolation mechanism |
| Outlier Focus | Global Anomalies | Local anomalies and isolated points |
| Speed | Computationally expensive | Faster and Scalable |
| Robustness | Robust Against small datasets | Scalable to high- dimensional data |
| Sensitivity | Sensitivity to covariance matrix | Sensitivity to contamination parameter |

### 4.2. Visualization of results using scatterplots

The scatterplot visualizations provide a clear view of the detected outliers, highlighting their distribution and relationships in the dataset. These plots allow for an instantaneous comparison of the outlier detection overall performance of the MCD and M-Estimator (Isolation Forest) methods, showcasing the diagnosed anomalies in a visually interpretable layout.

Figure 3 illustrates the comparative outlier detection results of the MCD method and the Isolation Forest (M-Estimator) method, highlighting their distinct anomaly identification patterns across the dataset.

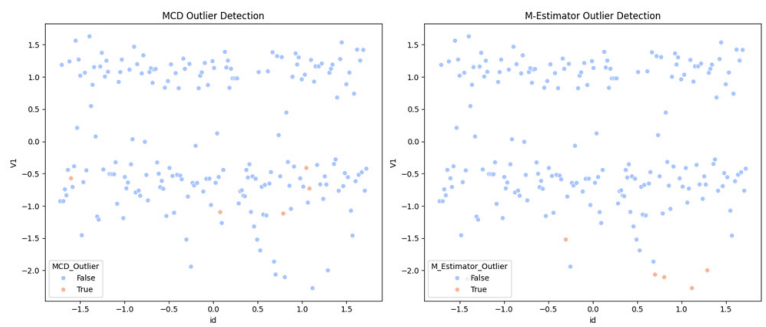



Figure 3: Outlier detection performance comparison (MCD) vs isolation forest

### 4.3. Analysis of robustness and sensitivity to different types of outliers

The desk provides the primary 5 rows of the dataset, along side outlier flags from both the MCD and M-Estimator techniques, indicating whether or not every information point is flagged as an outlier.

Table 5: MCD with M-Estimator sample data with outlier flags

| | Id | V1 | V2 | V3 | V4 | V5 | V6 | V7 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | -0.260648 | -0.469648 | 0-2.496266 | -0.083724 | 0.129681 | 0.732898 | 0.519014 |
| 1 | 1 | 0.985100 | -0.356045 | 0.558056 | -0.429654 | 0.277140 | 0.428605 | 0.406466 |
| 2 | 2 | -0.260272 | -0.949385 | 1.728538 | -0.457986 | 0.074062 | 1.419481 | 0.743511 |
| 3 | 3 | -0.152152 | -0.508959 | 1.746840 | -1.090178 | 0.249486 | 1.143312 | 0.518269 |
| 4 | 4 | -0.206820 | -0.165280 | 1.527053 | -0.448293 | 0.106125 | 0.530549 | 0.658849 |
| | | V8 | V9 ..... | V23 | V24 | V25 | V26 | V27 |
| 0 | | -0.130006 | 0.727159 | -0.134794 | 0.165959 | 0.126280 | -0.434824 | -0.081230 |
| 1 | | -0.133118 | 0.347452 | 0.079469 | -0.577395 | 0.190090 | 0.296503 | -0.248052 |
| 2 | | -0.095576 | -0.261297 | 0.945045 | -1.154666 | -0.605564 | -0.312895 | -0.300258 |
| 3 | | -0.065130 | -0.205698 | -0.214048 | -1.893131 | 1.003963 | -0.515950 | -0.165316 |
| 4 | | -0.212660 | 1.049921 | -0.161666 | 0.312561 | -0.414116 | 1.071126 | 0.023712 |
| | | V28 | Amount | Class | MCD Outlier | M-Estimator Outliers | | |
| 0 | | -0.151045 | 17982.10 | 0 | False | False | | |
| 1 | | -0.064512 | 6531.37 | 0 | False | False | | |
| 2 | | -0.244718 | 2513.54 | 0 | False | False | | |
| 3 | | 0.048424 | 5384.44 | 0 | False | False | | |
| 4 | | 0.419117 | 14278.97 | 0 | False | False | | |

The Table 4 offers the primary five rows of a dataset, showcasing the capabilities (V1 to V28), amount, magnificence, and outlier flags identified the usage of two anomaly detection strategies: the Minimum Covariance Determinant (MCD) and M-Estimator (Isolation Forest). Each row includes values for the functions, with the "Amount" column reflecting a specific quantitative metric. The "Class" column represents the category or label related to the records factors. The "MCD Outlier" and "M-Estimator Outliers" columns suggest whether the records point is flagged as an outlier by every technique, with all the records factors in this case being labeled as non-outliers (False) by way of each strategies, suggesting that the detected factors do not deviate appreciably from the expected styles in keeping with these strategies.

## 5. Conclusion

A hybrid anomaly detection framework combining the strengths of the MCD and M-Estimator (Isolation Forest) techniques gives an superior approach for detecting each global and localized outliers. The MCD technique excels in smaller, mild-sized datasets via efficaciously figuring out international anomalies based on distance measures, making it appropriate for instances where the primary challenge is the distribution of anomalies across the entire dataset. In contrast, the M-Estimator (Isolation Forest) technique thrives in excessive-dimensional, complex datasets, correctly detecting localized anomalies and last robust against noise. This technique is specifically tremendous for big, multifaceted statistics environments, uncovering anomalies now not handiest on the extremes however also within internal clusters. The one of a kind patterns determined inside the visible evaluation highlight the complementary strengths of both methods, with MCD detecting boundary outliers and Isolation Forest revealing a broader range of anomalies. Together, these strategies offer a comprehensive, scalable solution for effective anomaly detection in diverse statistics eventualities. Here are a number of the recommendations is covered:

- For smaller datasets with clean and nicely-described anomaly distributions: The MCD method is suggested because of its simplicity and efficiency in identifying outliers based totally on distance metrics. It is right when the dataset is not overly complicated and the paradox pattern is globally dispensed.

- For large, excessive-dimensional datasets with complex, noisy, or localized anomaly patterns: The M-Estimator, particularly through the Isolation Forest method, should be used. It is strong to noise and excels in detecting anomalies in massive, intricate datasets with numerous distributions, presenting a more adaptable technique for detecting a wide variety of outliers.

## References


1. Charu C. Aggarwal, *Supervised outlier detection*, Outlier Analysis, Springer, Cham, Switzerland, 2017, pp. 197–228.

2. M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, *Lof: identifying density-based local outliers*, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (Dallas, Texas, USA), ACM, May 16–18 2000, pp. 93–104.

3. S. Chakraborty, A. Basu, and A. Ghosh, *A componentwise estimation procedure for multivariate location and scatter: Robustness, efficiency and scalability*, arXiv:2410.21166 (2024), arXiv preprint.

4. V. Chandola, A. Banerjee, and V. Kumar, *Anomaly detection: A survey*, ACM Computing Surveys (CSUR) **41** (2009), no. 3, 1–58.

5. D.Q.F. De Menezes, D.M. Prata, A.R. Secchi, and J.C. Pinto, *A review on robust m-estimators for regression analysis*, Computers & Chemical Engineering **147** (2021), 107254.

6. P. D. Domański, *Statistical outlier labelling–a comparative study*, 2020 7th International Conference on Control, Decision and Information Technologies (CoDIT) (Prague, Czech Republic), vol. 1, IEEE, June 2020, pp. 439–444.

7. P. Filzmoser, R. G. Garrett, and C. Reimann, *Multivariate outlier detection in exploration geochemistry*, Computers & Geosciences **31** (2005), no. 5, 579–587.

8. V. Hodge and J. Austin, *A survey of outlier detection methodologies*, Artificial Intelligence Review **22** (2004), 85–126.

9. J.P. Irmer, A.G. Klein, and K. Schermelleh-Engel, *Model-implied simulation-based power estimation for correctly specified and distributionally misspecified models: Applications to nonlinear and linear structural equation models*, Behavior Research Methods **56** (2024), no. 8, 8955–8991.

10. Julien Patrick Irmer, Andreas G. Klein, and Karin Schermelleh-Engel, *A general model-implied simulation-based power estimation method for correctly and misspecified models: Applications to nonlinear and linear structural equation models*, OSF Preprint pe5bj, Center for Open Science, 2024.

11. J. Laurikkala, *Improving identification of difficult small classes by balancing class distribution*, Artificial Intelligence in Medicine. AIME 2001 (S. Quaglini, P. Barahona, and S. Andreassen, eds.), Lecture Notes in Computer Science, vol. 2101, Springer, Berlin, Heidelberg, 2001, pp. 494–503.

12. 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** (2018), 150–156.

13. F. T. Liu, K. M. Ting, and Z. H. Zhou, *Isolation forest*, 2008 Eighth IEEE International Conference on Data Mining (Pisa, Italy), IEEE, 2008, pp. 413–422.

14. R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with r)*, John Wiley & Sons, 2019.

15. M. Mayrhofer, *Robustness and explainable outlier detection for multivariate, matrix-variate, and functional settings*, Doctoral dissertation, Technische Universität Wien, Vienna, Austria, 2024.

16. M. A. A. M. Mokhtar, N. S. Yusoff, and C. Z. Liang, *Robust hotelling's $t^2$ statistic based on m-estimator*, Journal of Physics: Conference Series **1988** (2021), no. 1, 012116.

17. E. Nkum, *Robust multivariate estimation and inference with the minimum density power divergence estimator*, Doctoral dissertation, The University of Texas, Austin, Texas, 2024.

18. M.A. Pimentel, D.A. Clifton, L. Clifton, and L. Tarassenko, *A review of novelty detection*, Signal Processing **99** (2014), 215–249.

19. A. Prasad, A.S. Suggala, S. Balakrishnan, and P. Ravikumar, *Robust estimation via robust gradient estimation*, Journal of the Royal Statistical Society Series B: Statistical Methodology **82** (2020), no. 3, 601–627.

20. J.E. Pustejovsky and E. Tipton, *Meta-analysis with robust variance estimation: Expanding the range of working models*, Prevention Science **23** (2022), no. 3, 425–438.

21. C. Reimann and Wiley InterScience (Online service), *Statistical data analysis explained: Applied environmental statistics with r*, John Wiley & Sons, Chichester, UK, 2008.

22. P.J. Rousseeuw and M. Hubert, *Robust statistics for outlier detection*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1** (2011), no. 1, 73–79.

23. O. K. Sajana, *A study on robust multivariate techniques*, Doctoral dissertation, St. Thomas' College (Autonomous), Thrissur, University of Calicut, India, 2020.

24. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, *Estimating the support of a high-dimensional distribution*, Neural Computation **13** (2001), no. 7, 1443–1471.

*Mohanad N. Abdul Sayed,*
*Department of Computer Networking and Software Techniquses,*
*Technical Institute Qurna, Southern Technical University,*
*Basrah, Iraq.*
*E-mail address:* Mohanad87@stu.edu.iq

*and*

*Rana H. Shamkhi,*
*Department of Pharmacognosy and Medicinal Plants,*
*College of Pharmacy, University of Basrah,*
*Basrah, Iraq.*
*E-mail address:* rana413427@gmail.com