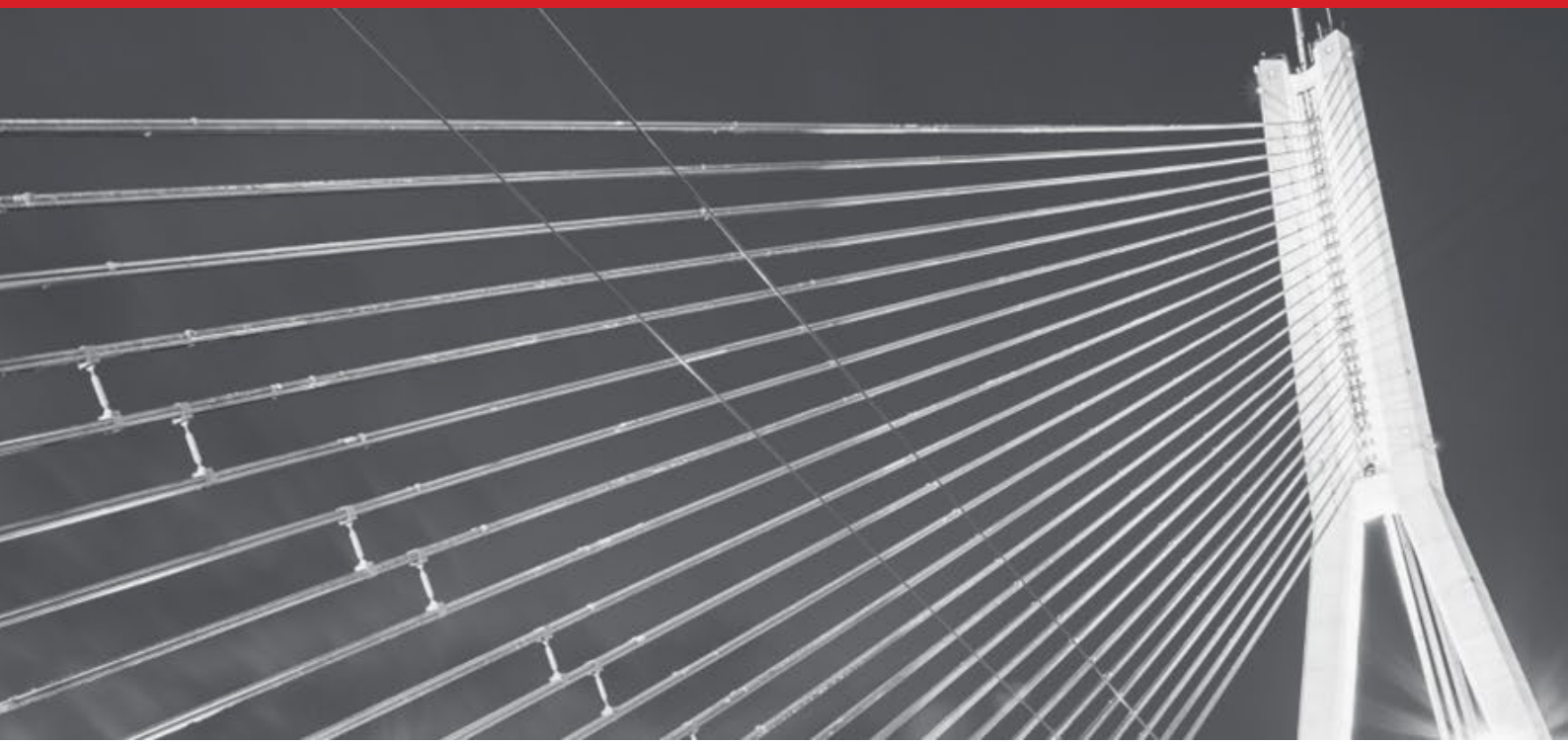
A Reacfin Research White Paper in Quantitative Finance

Outliers' detection for non-Gaussian distributions

by Olivier Cheffert and Jean Dessain

® Reacfin White Paper Vol.1 2024 – November 2024



ABSTRACT

Detecting outliers in financial or insurance datasets is critical for accurate risk assessment and decision-making, particularly in the context of non-Gaussian distributions characterized by skewness and heavy tails (e.g. financial returns series). Traditional outlier detection methods often rely on assumptions of normality, which can lead to misleading results in the presence of such data anomalies. A robust outlier detection system does more than just identify anomalies; it provides significant business advantages. By enhancing data quality, this approach optimizes the performance of predictive models that depend on this data, leading to better-informed strategic decisions and more accurate risk management. Indeed, poor input data quality will impact the accuracy of the model and prevent it from considering the real underlying dynamics in the dataset. Thus, while impacting its accuracy, it will also impact its interpretability since the model will try to fit to outliers.

In this paper, we show a framework for outlier detection able to deal with skewed and heavy-tailed distributions. The goal is to effectively identify outliers while maintaining sensitivity to the underlying distributional characteristics. Through practical use cases, we will illustrate its applicability as part of the preprocessing task.

Reacfin

TVA: BE 0862.986.729
BNP Paribas 001-4174957-56
RPM Nivelles

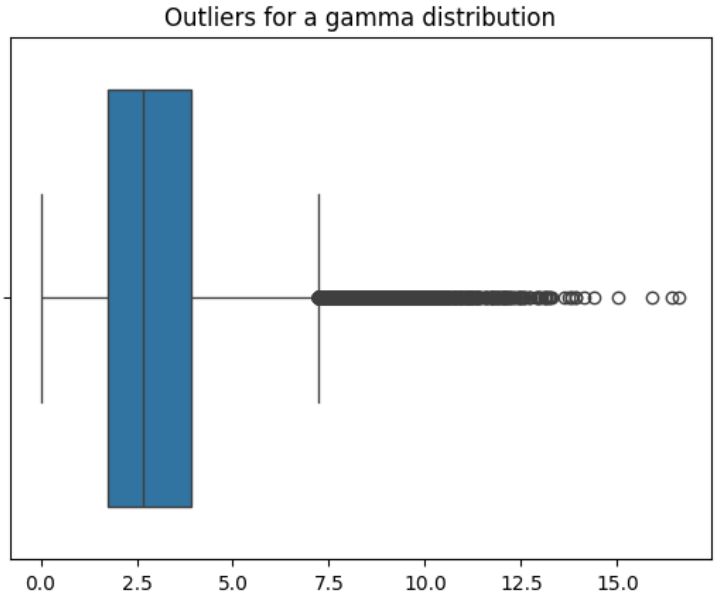
Tel: +32 (0)10 68 86 07
info@reacfin.com
www.reacfin.com

Reacfin s.a./n.v.
Place de l'Université 25
B-1348 Louvain-la-Neuve
Belgium

1. CLASSICAL TECHNIQUE: THE INTERQUARTILE RANGE

The interquartile range technique (IQR) is designed to exclude 0.7% of the normal distribution. If Q_i is the i^{th} quartile of given samples, then the interquartile is defined as IQR = Q_3 -Q_1 and each sample outside of [Q_1 – 1.5 IQR; Q3 + 1.5 IQR] is labelled as an outlier.

While this technique is easy to understand and to implement, the assumption of normality is strong, and this technique may not work in practice. For example, if a given variable is sampled 100 000 times from a gamma distribution, the skewness and tail heaviness of this variable will lead the IQR method to label true data as outliers:



Applied on this gamma (α=3, β=1) distribution, 2.6% of the samples are labelled as outliers. In a non-life insurance context, the variable can be skewed, heavy tailed (severity) and even discrete (frequency). In finance, these non-normal characteristics can also appear, for example, in the study of Loss Given default in credit risk, stock returns, ….

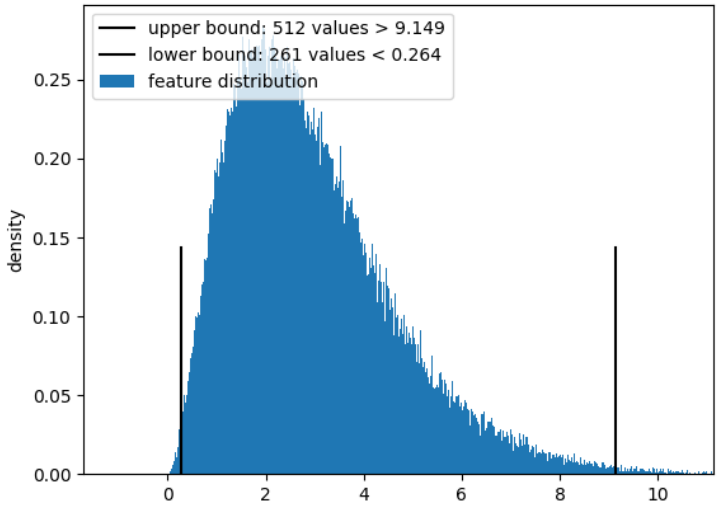
Fortunately, some new techniques allow to deal with specificities of a given sample but most of them focus only on a particular characteristic like tail heaviness. Therefore, a collection of different tools is required to detect outliers in practice which may increase the operational risk.

2. STAR OUTLIERS

STAR outliers, standing for Skew and Tail-heaviness Adjusted Removal of outliers, is a univariate outliers detection method developed by Gregg and Moore (Gregg & Moore, 2023). Through transformation functions, this technique deals with tail heaviness, skewness, multimodality, monotonicity and discrete distributions.

STAR outliers use a generalized method based on IQR, as proposed by Verardi and Vermandele (Verardi & Vermandele, 2018) which defines the concept of asymmetrical outlying (ASO). Initially, a ratio is calculated by subtracting each data point from the median and then dividing by the interquartile range. Subsequently, this ratio is scaled to fit within the unit interval and subjected to a probit transformation. If these ASO values are uniformly distributed on the unit interval before the probit transform, the final values are normally distributed. However, if it's not the case, this gives a transformed normal distribution with skew and tail heaviness. Such a distribution is called a Tukey g and h distribution which adds a skewness and a kurtosis parameter. Furthermore, multimodality is detected using the Hartigan's dip test (Hartigan & Hartigan, 1985) and a Tukey mixture will be fit to the data. Highly skewed variables are passed into mirror transforms before using the ASO technique. Finally, the outlier threshold is computed from the fitted Tukey distribution.

Using the same sampled gamma distribution as before, 0.77% of the sample is considered outliers:



This case study shows that STAR can deal with highly skewed and heavy tailed distributions like this gamma distribution.

On discrete (frequency) laws such as Poisson, this method can also be used and will label 1.1% of data as outliers for a Poisson(λ=3). When using STAR on a standard normal distribution, it detects 0.8% or outliers with a slightly narrower interval than IQR which detects 0.73% of outliers. Thus, for normally distributed data, the Interquartile Range (IQR) method, due to its simplicity, may be sufficient.

3. TAKEAWAYS

Preprocessing is essential when working with financial and insurance datasets, as inaccuracies in input data can compromise the performance of machine learning models. A critical step in preprocessing is outlier detection, which helps prevent bias in the model. Indeed, poor input data quality can severely impact the accuracy of predictive models, preventing them from capturing the true underlying dynamics of the dataset. This degradation in accuracy not only undermines the model's performance but also affects its interpretability, as the model may erroneously try to fit to outliers rather than the relevant trends.

This paper demonstrates that the classical interquartile range (IQR) method can significantly misestimate the target of 0.7% outliers, incorrectly labeling valid data as outliers on non-Gaussian data. In the case of real Loss Given Default data, even higher proportions of outliers were observed, highlighting the limitations of this method (Cheffert, 2024). For instance, some financial ratios of defaulted companies exhibited outlier rates as high as 25% according to the IQR approach. Thus, critical applications such as risk management for a bond portfolio can fail to accurately reflect the true risk due to poor data quality. Therefore, practitioners must recognize the presence of non-Gaussian data in business applications; an incorrect preprocessing step can lead to misguided predictions, regardless of the model used thereafter.

More generally, when datasets exhibit non-normal characteristics—such as excess kurtosis, skewness, multimodality, or integer distributions—the IQR method fails to address these issues, which can lead to accuracy problems in subsequent modeling. The STAR outlier detection method offers an advantage by accommodating non-Gaussian distributions and identifying a more accurate number of outliers. Additionally, it can handle missing data, making it a crucial preliminary step before imputation. Imputing data before detecting outliers or relying on ineffective outlier detection methods can result in erroneous and biased imputed values. However, the analysis made on normally distributed data has shown that IQR can still be used when working with a Gaussian distribution. Thus, a first step may be to use a normality test, such as the Jarque-Bera test, to decide whether IQR or STAR should be used.

4. REFERENCES

Cheffert, O. (2024). Loss Given Default modelling for corporate bonds.

Gregg, J., & Moore, J. (2023). STAR_outliers: a python package that separates univariate outliers from non-normal distributions. *BioData Mining*.

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, 70 - 84.

Verardi, V., & Vermandele, C. (2018). Univariate and Multivariate Outlier Identification for Skewed or Heavy-Tailed Distributions. *The Stata Journal: Promoting communications on statistics and Stata*, 517-532.

5. ABOUT REACFIN

We develop **sustainable actuarial, quantitative financial and AI for Finance solutions** in partnership with our clients (from **design and modeling** to **operationalization** in their systems), building on **strong data analytics** while securing **full transparency** and **integral knowledge transfer**.

Reacfin is the reliable bridge between academic excellence and market best practices.

The company started its activities in 2004 as a spin-off of department of UCLouvain School of Statistics, Biostatistics and Actuarial Science.

In its early days, we focused on actuarial consultancy services for Belgian Pension Funds, Insurance Companies and Mutual organizations. Rapidly, in the following years, we expanded our business internationally and broadened our scope of services to Risk Management, Quantitative Finance, Portfolio Management and Data Analytics for Financial Institutions in the broader sense (i.e. Insurers, Banks, Asset Managers, Pension Funds, Financial Market Infrastructures and Regulators). Today, Reacfin is extending its range of services to include process automation, the introduction of AI and, more generally, the optimal use of corporate and external data.

Based in Louvain-la-Neuve (Belgium), Reacfin employs today more than 35+ consultants most of which hold PhD's or highly specialized university degrees.

Over the years, we have now served in excess of 150 different financial institutions, the vast majority of which are recurrent clients, which we see as the most convincing indicator of our clients' satisfaction.

Missions we regularly perform consist of models design, developments & deployment, model validations, definition of risk- & portfolio management policies, organization & governance advisory, strategic asset allocations or specialized management consulting with regard to Risk & Portfolio management problems.

We organize our consulting services along 4 Centers of Excellence:

RISK MANAGEMENT & FINANCE

- ESG / climate risk management
- Implementation/calibration of stochastic models
- Valuation/pricing of financial instruments
- Development of AM & ALM models
- Credit Portfolio Management Models (incl. IRB, IFRS9, etc.)
- Asset allocation, (Automated) trading & hedging strategies
- Quantitative Risk Management models
- Strategic opportunities assessment and business valuations
- Industrialization of processes & organizational optimization
- Business intelligence, benchmarking & surveys
- Internal & regulatory reporting (KRI's & KPI's dashboards)
- Validations, model review frameworks and model documentation

LIFE, HEALTH AND PENSION

- Valuation frameworks: Solvency 2, IFRS17, Local GAAP
- Pricing, product development & reserving
- Dynamic Financial Analysis (DFA)
- Capital Requirement optimization
- Business valuation support & Actuarial function outsourcing
- Pension liabilities valuation

NON-LIFE

- Implementation or review of reserving methodologies
- Development of innovative pricing methodologies and tools
- Valuation & profitability analysis models
- Risk mitigation optimization
- Business valuation, capital management and actuarial function

ACTUARIAL ENGINEERING

- Processes streamlining and automation leveraging best-in-class software development techniques and deep learning (DL) algorithms
- Refactoring of legacy tools, supporting from design and development to deployment within client's infrastructure
- Innovative application of state-of-the-art machine learning (ML) and generative artificial intelligence (GenAI), like large language models (LLM)
- Data management & data quality pipe-line automation
- Data visualization (dynamic dashboards, automated reports, etc.)
- Training and coaching to help clients' teams to upskill their technological knowledge

We deploy material efforts at ensuring that Reacfin deliverables systematically have the following characteristics:

State of the art technical skills	▪ Expertise in most advanced quantitative modelling & academic excellence of a spin-off ▪ All our consultants hold multiple masters or Phd ▪ Best-in-class quantitative and qualitative risk management leveraging on highly experienced senior consultants
Balanced and pragmatic approach	▪ Client-centric solutions focused on deliverables ▪ Respecting the principle of proportionality ▪ Cost efficient within tight pre-agreed budgets
No black box solutions	▪ Hands-on implementation tested for real-life conditions ▪ Open source solutions on request ▪ Close cooperation with our clients
Clearly structured processes	▪ Lean & efficient tailored project management ▪ Regular progress reviews ▪ Agile approach to adapt to the evolving needs of our clients
Documentation, coaching & training	▪ Clear & comprehensive documentation compliant with existing or upcoming regulation ▪ Adapted trainings at all levels of the organisation ▪ Coaching support for implementation and operationnalisation of processes

We articulate our offer along 3 brands:

Reacfin Consulting — We offer consulting services in actuarial science & quantitative finance, including a.o. capital, portfolio, product, risk and liquidity management. We build our expertise on broad data analytics capacities.

Reaxii — By blending strong actuarial and financial business expertise with an in-depth understanding of cutting-edge IT and AI technologies, we enable our clients to become more competitive and focus on their core business such as complex analysis, strategic decision-making and innovation.

Reacfin Academy — We share our knowledge with our clients. We offer a comprehensive learning platform, including on-site trainings, e-learning modules, e-classrooms and webinars.

Reacfin's management puts great emphasis at sharing and embedding our driving values within the company:

Excellence: Our main feature	Innovation: Our founding ambition	Integrity: Our commitment	Solution-Driven: Our focus	Reliability: Our characteristic
We attract the best people We develop their skills and career through diversified missions and rigorous knowledge management We go the extra-mile to deliver the best quality in our work & services	By acting as a bridge linking academic excellence with best market practices, we select the latest research that best serves our client Through out of the box thinking, we apply state-of-the-art techniques that offer our clients pragmatic added-value solutions	We put work ethics, client's best interest and confidentiality as the foundation of our work We commit to promoting the greatest transparency and knowledge sharing in all our client's solutions	We are dedicated to clearly understanding the needs of our clients We deliver solutions that produce measurable value Our deliverables are tailored and actionable solutions to our client's challenges	We develop sustainable partnerships with our clients We never compromise on our commitments including level of quality, budgets & deadlines All our deliverables are designed, developed and tested to last over time with constant efficiency

6. CONTACT DETAILS

Xavier Maréchal
Managing Partner
xavier.marechal@reacfin.com

Feel free to check our online resources for more information and free material

Check our online resources on www.reacfin.com

Online apps & Tools demo's

Latest research & training programs

Reacfin

*We develop **sustainable actuarial, quantitative financial and AI for Finance solutions** in partnership with our clients (from **design and modeling** to **operationalization** in their systems), building on **strong data analytics** while securing **full transparency** and **integral knowledge transfer**.*

www.reacfin.com
info@reacfin.com
+ 32 (0)10 68 86 07

Linking Academic Excellence with Market Best Practice