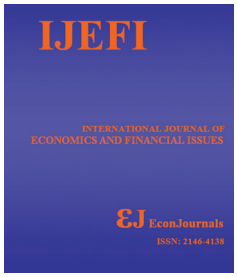


International Journal of Economics and Financial Issues

ISSN: 2146-4138

available at http: www.econjournals.com

International Journal of Economics and Financial Issues, 2024, 14(2), 203-206.

Visualisation of Mahalanobis Distances for Trivariate JOINT Distributions

Emily Groenewald[1], Gary Van Vuuren[2]*

[1]School of Economics, University of Cape Town, Cape Town, South Africa, [2]Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, 2351, South Africa. *Email: vvgary@hotmail.com

Received: 04 December 2023 **Accepted:** 08 March 2024 **DOI:** https://doi.org/10.32479/ijefi.15663

ABSTRACT

The Mahalanobis distance is a statistical measure used to quantify the distance between elliptic distributions with distinct locations and shared shapes, while accounting for the variables' covariance structure. It is applicable to both estimative and predictive estimation approaches, where variations are limited to location, and it assesses the similarity or dissimilarity between data and the mean (centroid) of a multivariate distribution, within the family of multivariate elliptic distributions. It is thus useful for outlier identification. The aim of the study is to provide, for the first time, a three-dimensional visualisation of the Mahalanobis distance when the underlying framework comprises three jointly connected variables (rather than the standard two variables presented in textbooks). Data with Mahalanobis distances exceeding a predefined threshold, determined using a χ^2 distribution, are considered outliers. This approach is analogous to identifying outliers for univariate distributions based on critical values derived from confidence levels. While the literature mainly discusses the Mahalanobis distance formulation for bivariate distributions, we extend the discussion to include one additional variable and provide a visualisation of the resulting Mahalanobis distance for a trivariate distribution. An empirical example is presented to illustrate a practical application of a trivariate Mahalanobis distance. Visualising outliers alongside other historical events within three-factor systems can offer valuable insights into the risk profile of the current environment and assess the probability of future extreme events.

Keywords: Mahalanobis Distance, Trivariate, Elliptic Distributions, Outliers
JEL Classifications: C38, G17, E37

1. INTRODUCTION AND BACKGROUND

Distance concepts play a fundamental role in various multivariate statistical techniques. Among these, the Mahalanobis distance (Mahalanobis, 1936) is a popular measure in the context of multivariate normality. Mahalanobis (1936) extended Rao's (1945) method for calculating distances between members of a well-behaved parametric distribution family.

In finance, problems such as asset allocation and risk management often involve multiple random variables. Market participants are interested in understanding the similarity between realisations of these variables and identifying objective indicators that identify the breakdown of previously stable relationships. During the 2008/2009 subprime crisis, for example, noticeable deviation in market behavior were observed compared to previous patterns. These aberrations were first manifest in certain instruments such as asset-backed securities, but as the crisis unfolded, this atypical behaviour rapidly permeated other instruments and markets (Calice, 2011). Return distribution changes can also render optimised portfolio weights suboptimal, and losses arise from trading strategies relying on outdated market price regularities (Han and Park, 2022) particularly during severely stressed conditions, such as those experienced during the recent COVID-19 pandemic.

The Mahalanobis distance has also found applications in finance, including asset classification (Zuo and Serfling, 2000), portfolio

surveillance (Kritzman and Li, 2010), and outlier frequency and severity detection (Penny, 1996; Ghorbani, 2019; Li et al., 2019; Clarke and Grose, 2023; Singh et al., 2023). Its usefulness in the identification of multivariate outliers arises from its selection of robust and independent centroid and covariance matrix measures. Because it can be combined with techniques which reduce the impact of outliers, it helps detect atypical patterns in multivariate observations (Mitchell and Krzanowski, 1985).

Because it is impossible to visualise the probability density for -variate (for n > 2)distributions, descriptions of the Mahalanobis distance in the literature focus mainly on bivariate distributions (Warren et al., 2011; Ghorbani, 2019) despite its wider applicability to multivariate distributions. Visualisation of the Mahalanobis distance, however, *is* possible for a trivariate setup and instructive for assessing outliers in such a case.

2. METHODOLOGY

In the description which follows, $\vec{p}$ may comprise observations sampled from many variables, but the development employs a bivariate case for ease of notation and explanation.

Consider a bivariate observation comprising two linked values, x and y, drawn from two normally distributed populations, X and Y with means μ_x and μ_y and standard deviations σ_x and σ_y respectively. The pair of values may be represented by a column vector:

$$\vec{p} = \begin{bmatrix} x \\ y \end{bmatrix} = [x \quad y]'$$

and the means of the X and Y distributions define the centroid:

$$\vec{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = [\mu_x \quad \mu_y]'$$

If the correlation between the variables is ρ_{xy}, the $p \times p$ covariance matrix Σ is

$$£ = \begin{bmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho_{xy} \\ \sigma_x\sigma_y\rho_{xy} & \sigma_y^2 \end{bmatrix}$$

The Mahalanobis distance[1], $d_m^2(\vec{p})$, of a particular bivariate pair $\vec{p}$ from the centroid is $[\vec{p}-\vec{\mu}]'£^{-1}[\vec{p}-\vec{\mu}]$ or

$$d_m^2(\vec{p}) = \begin{bmatrix} x-\mu_x & y-\mu_y \end{bmatrix} £^{-1} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}$$

Where comprises two variables, the locus of $d_m^2(\vec{p})$ (not $d_m(\vec{p})$) defines an ellipse (in two dimensions) with centroid at $\vec{\propto}$. The boundary of such an ellipse is an equiprobability density contour and the probability associated with each d_m^2 contour follows a χ^2 distribution with p degrees of freedom. The locus of $\vec{p}$ coordinates satisfies $[\vec{p}-\vec{\mu}]'£^{-1}[\vec{p}-\vec{\mu}] \leq \chi_p^2(\alpha)$ is 1-α where α is a specified confidence level, so d_m^2 may be used to detect outliers in multivariate normal data (Warren et al., 2011; Ghorbani, 2019), that is:

$$d_m^2(\vec{p}) \sim \chi_p^2$$

Where $\vec{p}$ is trivariate, any given $d_m^2(\vec{p})$ defines an ellipsoid (in three dimensions) with centroid $\vec{\propto}$ (μ_x,μ_y,μ_z) and boundary reflecting an equiprobability density ellipsoidal surface. The probability associated with observations in the ellipsoidal volume bounded by this surface again follows a χ^2 distribution with p degrees of freedom.

3. DATA AND RESULTS

Consider a variable described by a univariate standard normal ($N(\mu,\sigma^2) \sim N(0,1)$) distribution as shown in Figure 1a. In this example, the probability of observing a datum $x < -1.645$ is $N^{-1}(-1.645) = 5\%$. Visualisation of this situation involves projecting the two-dimensional probability density onto the x-axis (one dimension): observations for which $x > -1.645$ occur with frequency 95%.

Next consider a bivariate joint distribution (Figure 1a), comprising two normally distributed variables with $N(\mu_x,\sigma_x^2)$ and $N(\mu_y,\sigma_y^2)$ and correlation $\rho_{xy} = -0.5$. The joint distribution is best visualised as a three-dimensional volume with orthogonal axes (Figure 1b).

A plan view of such a joint probability density is a projection onto a surface beneath the probability density volume (Figure 2a and 2b with ellipses representing equal $d_m^2(\vec{p})$ contours at varying distances from the centroid). Coordinates lying on a given ellipse are equally far (statistically) from the centroid, so the ellipse boundary represents a probability threshold. Points within the ellipse occur with probability α while those outside occur with probability 1-α. Coordinates on concentric ellipse boundaries are equiprobable. Visualisation of this situation involves projecting the three-dimensional probability density volume onto a two-dimension area.

Visualising probability densities for $n > 2$ are impossible, but the projection – the probability space – for a joint trivariate probability density is a three-dimensional volume of data, as shown in Figure 3a. A given ellipsoid, nesting concentrically around the centroid, encapsulates observations from a trivariate joint distribution that occur with probability α. Observations outside of the ellipsoid again occur with probability 1-α. Visualisation of this situation involves projecting the four-dimensional joint probability density distribution (impossible to visualise) into a three-dimension volume. Figure 3a shows empirical semi-annual GDP/inflation/unemployment coordinates spanning 1990–2023 with a 95% equiprobable boundary. Figure 3b presents the same empirical example with some outlier coordinates labelled according to the date on which they occurred. The ellipsoid boundary is set such

1 Euclidian distances are inadequate measures of distance in such cases since they ignore information provided by the variables' standard deviations and correlation between them.

Figure 1: (a) Univariate, standard normal distribution. The dashed line represents a position on a one-dimensional line beyond which there is a 5% probability of observing an observation more severe ($N^{-1}(d_m) = N^{-1}(-1.645) = 5\%$), and (b) joint probability density for two marginal normal distributions correlated at $\rho = +0.5$

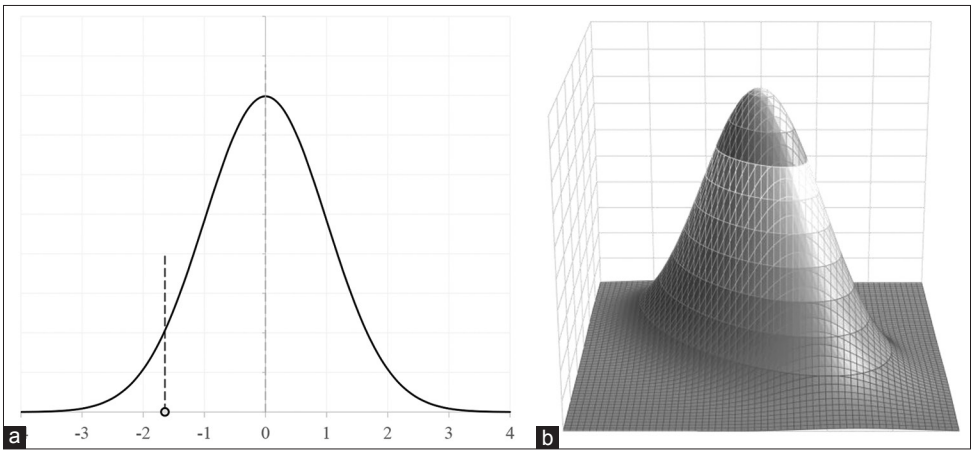


Figure 2: (a) Projection of joint probabilities onto a two-dimensional surface. Ellipses reflect points statistically equidistant from the centroid and, for any given ellipse, the probability of a joint observation as or less extreme occurring is □, while a joint event more extreme occurring is the area outside the ellipse, with a probability of 1-□ ($0 \leq □ \leq 1$) and (b) schematic representation of a joint bivariate distribution density (□ = −0.5) onto a surface and an equiprobable ellipse boundary defined such that 5% observations (outliers) occur outside the ellipse and 95% within.

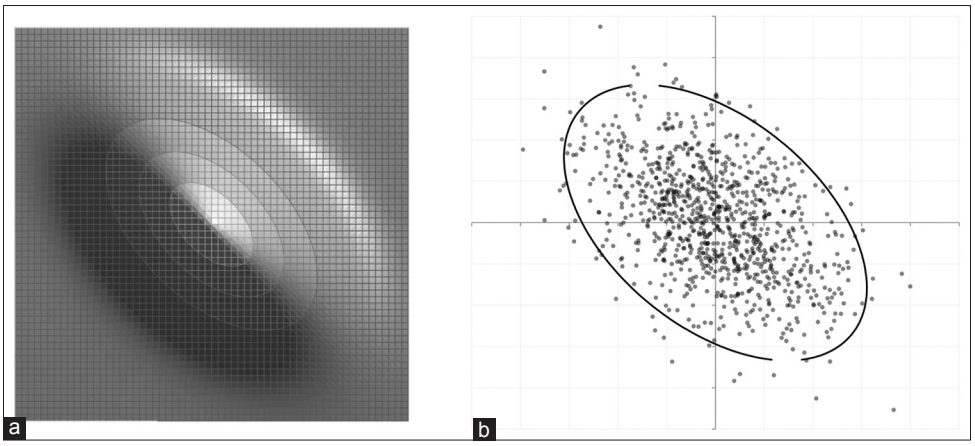
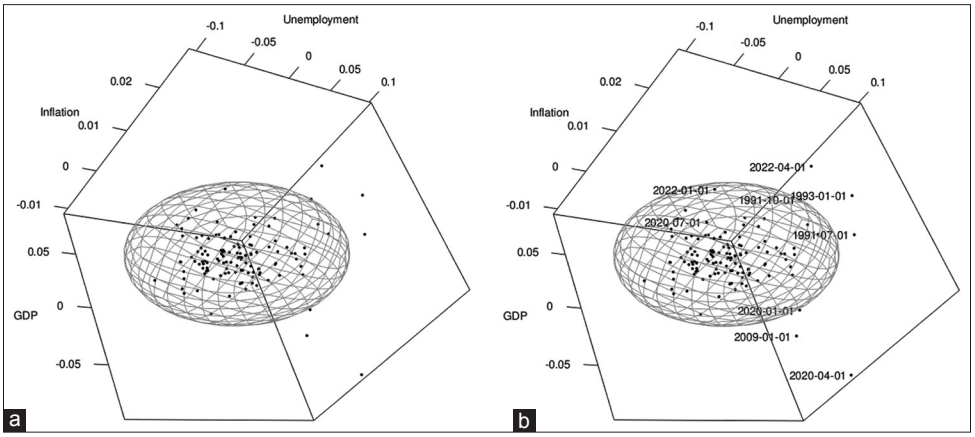


Figure 3: (a) Projection of a trivariate joint distribution density into a three-dimensional ellipsoid. The boundary (shell) represents statistically equidistant/equiprobable points from the centroid: data within the shell (in this case) have P ≤ 95% probability of occurring and (b) empirical application of a trivariate Mahalanobis distance to German inflation, GDP, and unemployment (variables normal and standardised) showing the severity of the 2008/9 credit crisis, the impact of the COVID-19 pandemic, and other extreme market events. Coordinate labels indicate relevant dates associated with outliers



that events within it occur with probability 95% and those outside it, 5%. The severity of various market crises is evident.

4. CONCLUSION

4.1. Concluding Remarks

The Mahalanobis distance is a valuable tool in finance, where it has been used for asset classification, portfolio surveillance, outlier detection, and hypothesis testing. By measuring the statistical distance between observations and a centroid, the Mahalanobis distance can identify atypical patterns, detect outliers, and assess the severity of events. It is particularly useful in multivariate scenarios where multiple variables are involved, as it includes the covariance structure and provides a robust measure of dissimilarity.

4.2. Policy Implications

The Mahalanobis distance is a powerful tool for quantifying the similarity or dissimilarity of multivariate data. Its ability to capture distribution changes and identify breakdowns in previously stable relationships makes it an asset for market participants and researchers alike. Potential policy implications could involve requiring financial institutions to disclose – as a relative measure – the Mahalanobis distance for various extreme events which may have occurred in the recent past. The ability to visualise these outliers in comparison with other historical events in three-factor systems can help provide insight into the riskiness of the current milieu and establish the likelihood of future extreme events.

4.3. Suggestions for Future Studies

Future research could explore the knock-on effect of the COVID-19 pandemic. In addition, volatile energy prices in Europe, could be explored to identify why Germany and Italy had significant quarters during 2022, with the addition of inflation in the model, while South Africa and the USA did not. Stress tests and the clear identification of extreme events can be better visualised under trivariate factor systems.

REFERENCES

Calice, G. (2011), The subprime asset-backed securities market and the equity prices of large complex financial institutions. Journal of International Financial Markets, Institutions and Money, 21(4), 585-604.

Clarke, B.R., Grose, A. (2023), A further study comparing forward search multivariate outlier methods including ATLA with an application to clustering. Statistical Papers, 64(1), 395-420.

Ghorbani, H. (2019), Mahalanobis distance and its application for detecting multivariate outliers. FME Transactions - Mathematics and Informatics, 34(3), 583-595.

Han, C., Park, F.C. (2022), A geometric framework for covariance dynamics. Journal of Banking and Finance, 134(2022), 1-18.

Kritzman, M., Li, Y. (2010), Skulls, financial turbulence, and risk management. Financial Analysts Journal, 66, 30-41.

Li, X., Deng, S., Li, L., Jiang, Y. (2019), Outlier detection based on robust Mahalanobis distance and its application. Open Journal of Statistics, 9(1), 15-26.

Mahalanobis, P.C. (1936), On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta), 2(1), 49-55.

Mitchell, A.F.S., Krzanowski, W.J. (1985), The Mahalanobis distance and elliptic distributions. Biometrika, 72(2), 464-467.

Penny, K.I. (1996), Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Journal of the Royal Statistical Society, Series C (Applied Statistics), 45, 73-81.

Rao, C.R. (1945), Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 37(1), 81-91.

Singh, G.N., Bhattacharyya, D., Bandyopadhyay, A. (2023), Robust estimation strategy for handling outliers. Communications in Statistics - Theory and Methods. Available from: https://www.tandfonline.com/doi/abs/10.1080/03610926.2023.2218567.

Warren, R., Smith, R. E., Cybenko, A. K. (2011), Use of Mahalanobis Distance for Detecting Outliers and Outlier Clusters in Markedly Non-normal Data: A Vehicular Traffic Example. SRA International Air Force Research Laboratory, 711th Human Performance Wing, Human Effectiveness Directorate. Available from: https://apps.dtic.mil/sti/pdfs/ADA545834.pdf

Zuo, Y., Serfling, R. (2000), General notions of statistical depth function. Annals of Statistics, 28(5), 461-482.