Are you normal? A new projection pursuit index to assess a sample against a multivariate null distribution

Annalisa Calvi, Ursula Laa, Dianne Cook

July 4, 2024

Abstract

1 Introduction

Linear projections are useful in many aspects of statistical analysis of multivariate data, and especially useful for visualising the data. A linear projection provides a dimension reduction while maintaining interpretability. For example, a biplot (Gabriel 1971, Gower & Hand n.d.) shows the structure creating the maximum variance in the data, and also visualizing the projection matrix to learn which variables contribute to it. We might find clusters of outliers that were hiding in high dimensions.

More generally, projection pursuit (Friedman & Tukey 1974, Jones & Sibson 1987, Huber 1985) defines a quantitative criterion for the interestingness of a projection (a projection pursuit index), and searches the space of possible projections for the most interesting one to display. We can also define sequences of interpolated linear projections to better understand a multivariate distribution. Animating a randomly selected interpolated sequence of linear projections is called a grand tour (Asimov 1985, Buja & Asimov 1986, Buja et al. 2005, Cook et al. 2006, Lee et al. 2022). The combination of these two approaches would then use a projection pursuit index to select interesting projections, but display them via an interpolated path to provide context. This is called a guided tour (Cook et al. 1995).

The question is whether we can use these techniques to assess new data samples in the context of an established normal, such as a specific multivariate normal distribution. In physics, the normal distribution may describe experimental results, or a global fit for a selected model, and we might want to compare to a set of other models. In medical applications, the normal distribution might summarize historic data of a healthy population and we compare it to samples from new patients. In outlier detection we might use robust measures to define the normal distribution and look for anomalies.

This paper describes a new projection pursuit index which is optimized by projections where a new sample is most distant from the existing normal distribution. It is organised as follows. Section 2 provides more context for the methods and visualisation. Section 3 provides the details of the new index, and example use is illustrated in Section 6.

2 Background

To compare a new sample with an existing norm, like a multivariate normal distribution, in higher than two dimensions, we have typically used two samples of points. The norm is represented by points on the surface of a p-dimensional ellipsoid, corresponding to a confidence level. A sample of points uniformly distributed on a p-dimensional sphere is generated by

1. Simulating a sample of observations ($\boldsymbol{x}$, which are p-D vectors) from $N_p(\boldsymbol{\mu}, \Sigma)$.
2. Transforming each observation to have unit distance from the mean, $\frac{\boldsymbol{x}^\top}{||\boldsymbol{x}^\top||}$.

To convert this to points on the surface of a confidence ellipsoid,

3. transform the shape using a specific variance covariance, and shift to center on the mean vector.

Finally, new observations can be visually compared with this ellipsoid by

4. plotting them together.

Figure 1 illustrates this process for 2-D. This is easiest way to view this normal region relative to a new sample for any p-D problem.

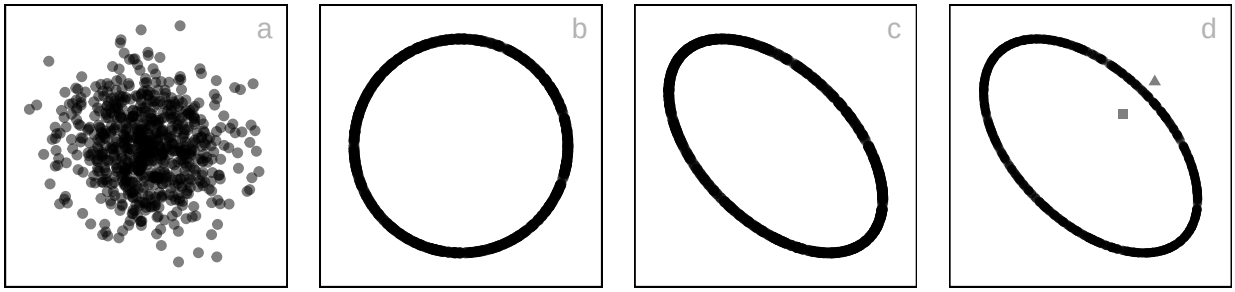


Figure 1: Simulating a uniform sample on a sphere involves sampling from a multivariate normal (a) and transforming each observation to have length equal to 1. A confidence ellipsoid is generated by transforming the sphere relative to a specified variance-covariance matrix (c), and new observations can be visually assessed to be inside or outside by plotting with the ellipsoid (d).

<PLOTS FROM USCHI's PHYSICS EXAMPLE HERE>

Although this is flexible, this does not make it easy to guide the tour towards the directions (projections) where the samples are most different from the normal. What would be desirable is to analytically define the confidence ellipsoid, compute flag observations that are outside, steer the tour to projections that reveal the extent of the difference. And also display the projected ellipsoid as a geometric shape rather than a sample of points. These are the procedures that are described in the next section.

3 Anomaly index

4 Projecting an ellipsoid

A *p-D* ellipsoid corresponding to a given variance-covariance (Σ) is defined by

$$(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^T = c^2$$

where c a constant that depends on a specific confidence level.

Theorem. *The projection of this p-D ellipsoid in 2-D has the equation*

$$(\boldsymbol{y} - \boldsymbol{\mu}_p)(P^T \Sigma P)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_p)^T = c^2.$$

Proof. The projection of an ellipsoid onto 2-D is an ellipse, where the curve of the ellipse is defined through the set of points $\boldsymbol{x}$ for which the gradient is parallel to the projection plane. That is, the curve consists of $\boldsymbol{x}$ satisfying

$$\nabla(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^T = 2(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1} = 2\boldsymbol{s}P^T$$

for some $\boldsymbol{s} \in \mathbb{R}^2$, where P is a $(p \times 2)$ orthonormal basis defining the projection. We can write $\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{s}P^T\Sigma$. Making this substitution in the p-D ellipsoid equation yields

$$c^2 = (\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^T = \boldsymbol{s}P^T\Sigma P\boldsymbol{s}^T$$

We call points in the projection that are on the curve $\boldsymbol{y}$, so that $\boldsymbol{y} = \boldsymbol{x}P$ for $\boldsymbol{x}$ on the p-D curve. Then $\boldsymbol{y} - \boldsymbol{\mu}_p = (\boldsymbol{x} - \boldsymbol{\mu})P = \boldsymbol{s}P^T\Sigma P$, where $\boldsymbol{\mu}_p = \boldsymbol{\mu}P$ is the projected mean. We then substitute $(\boldsymbol{y} - \boldsymbol{\mu}_p)(P^T\Sigma P)^{-1}$ for $\boldsymbol{s}$ in the equation $c^2 = \boldsymbol{s}P^T\Sigma P\boldsymbol{s}^T$. From this we can compute the analog equation for the projection as

$$(\boldsymbol{y} - \boldsymbol{\mu}_p)(P^T\Sigma P)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_p)^T = c^2$$

as claimed. □

This means the matrix $(P^T\Sigma P)^{-1}$ is defining the ellipse in the 2-D projection. In general c could be any constant, but typically we would select it as a quantile of the χ^2 distribution,

so that the size of the ellipse corresponds to a selected probability.

4.1 Index specification

To define a measure of an interesting projection is to maximize the average Mahahlanobis distance (Mahalanobis 1936) in the projection for a subset of points, W. The set of points could be chosen in different ways, but the default is those that are outside the specified ellipsoid in p-D. Alternatives could be to select a set of observations with the largest Mahalanobis distance, manually select observations or possibly a group of points identified via clustering of the extremes.

The index is written as

$$\sum_{\boldsymbol{w} \in W} (\boldsymbol{w} - \boldsymbol{\mu}) P (P^T \Sigma P)^{-1} P^T (\boldsymbol{w} - \boldsymbol{\mu})^T$$

where by default $W = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}) \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^T > c^2\}$, is the set of observations outside the p-D ellipsoid.

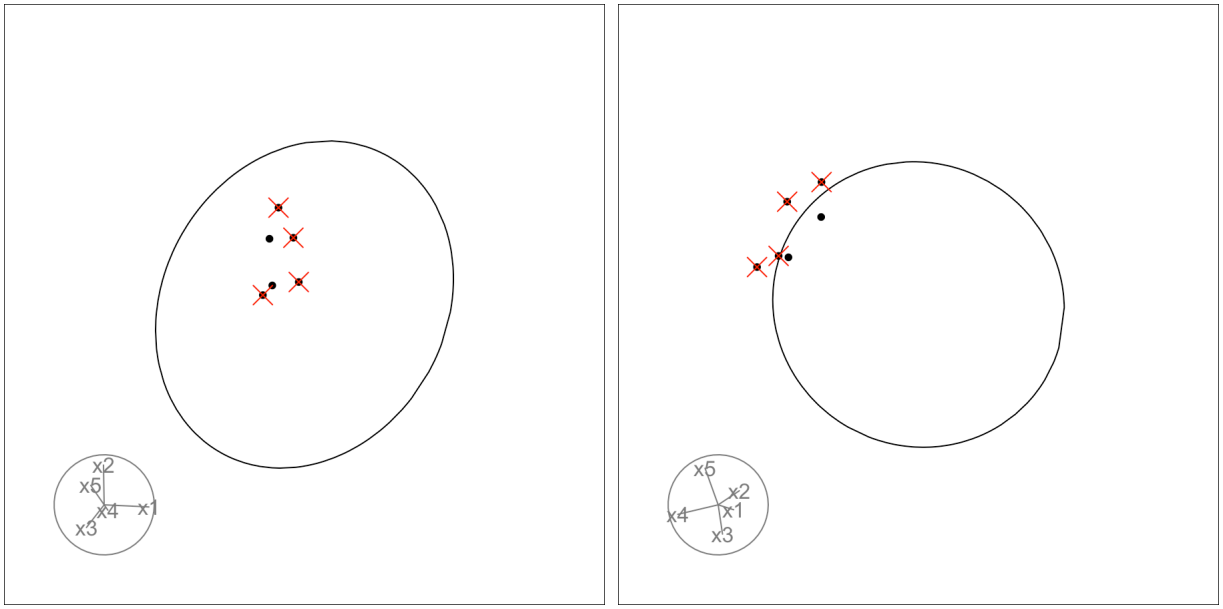
4.2 Additional considerations

If the observations in W are primarily departing from the normal range in the same direction, the index will be expected to perform well in finding this average direction. However, if the observations have very different departures from the norm, it may be useful to break them into groups, and separately optimize on these subsets. One could consider clustering these observations using angular distance to find groups of observations that have similar directions of departure.

5 Implementation

This is implemented in the `tourr` (Wickham et al. 2011, 2024) package, where the projected ellipsoid can be drawn for each projection. The guided tour will take arguments specifying the data, and the null variance-covariance matrix.

```
library(tourr)
library(mulgar)
set.seed(929)
vc_null <- matrix(rep(0.5, 5*5), ncol=5)
diag(vc_null) <- 1
m_null <- rep(0, 5)
vc_samp <- matrix(rep(0, 5*5), ncol=5)
diag(vc_samp) <- 1
vc_samp[4,5] <- -0.47
vc_samp[5,4] <- -0.56
vc_samp <- vc_samp*0.1
m_samp <- c(0, 0, 0, 1.9, 2.3)
samp <- as.data.frame(rmvn(6,
                           mn = m_samp,
                           vc = vc_samp))
animate_xy(samp, guided_anomaly_tour(anomaly_index(),
  ellipse=vc_null), ellipse=vc_null,
  axes = "bottomleft", half_range=5, center=FALSE)
```



(a) Random projection. (b) Optimal projection.

Figure 2: Two projections of simulated example data corresponding to the sample code: (a) random projection where sample is inside the 2-D ellipse, (b) optimal projection from index, showing most of the sample outside. A red cross indicates that the point is outside the p-D ellipse. The optimal projection uses mostly variables x_4, x_5, which is expected because these are the two directions where the sample most differs from the norm.

6 Examples

6.1 Health: liver function tests

This example is motivated by a problem posed during consulting with a pharmaceutical company, but the data shown here is simulated, simply to illustrate the application. Liver function tests commonly provide measurements on albumin, protein, bilirubin, gamma-glutamyl- transferase (GGT), aspartate aminotransferase (AST), alkaline phosphatase (ALP) and alanine aminotransferase (ALT). There are normal ranges on these measurements reported by Lala et al. (2023) and listed in Table 1.

Table 1: Normal ranges provided for liver function tests.

	albumin (g/L)	protein (g/L)	bilirubin (µmol/L)	GGT (IU/L)	AST (IU/L)	ALP (IU/L)	ALT (IU/L)
min	35	60	0	2	5	30	5
max	50	80	20	44	30	120	40

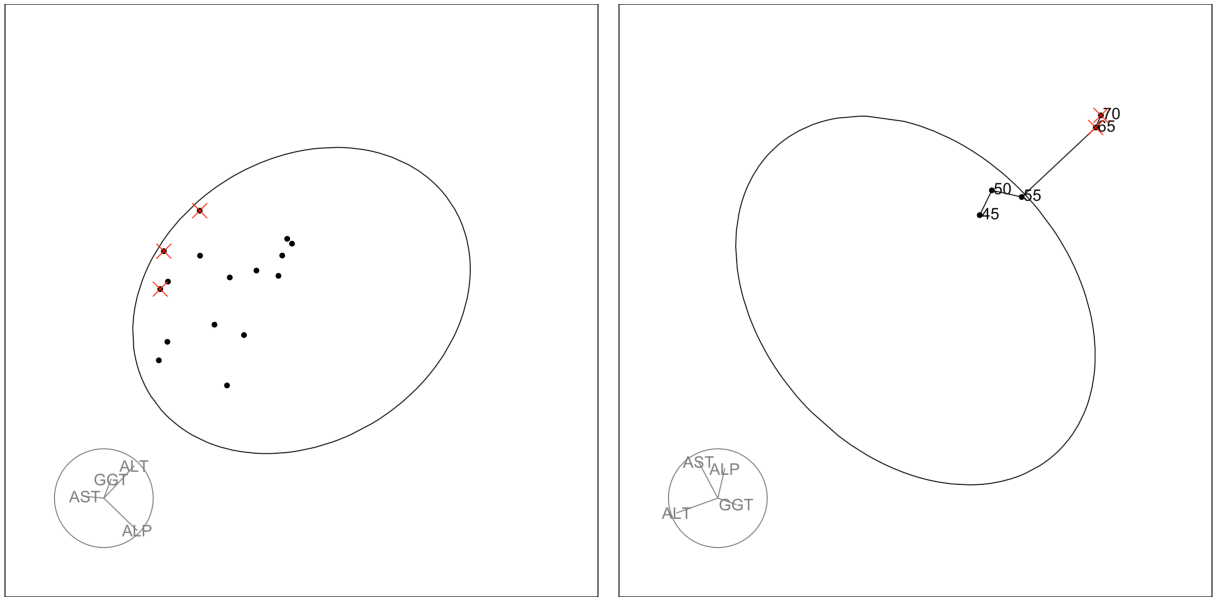
These measurements are also likely correlated, based on guidance like the *ratio of AST to ALT of 2:1 indicates possible alcohol abuse*. Although this was provided by the pharmaceutical company, correlation between these measurements for normal patients is not readily available. When a correlation matrix for normal patients is provided this would allow construction of the null ellipse upon which to examine new samples.

Figure 3 illustrates two examples. The first is similar to the consulting project. A sample of liver test scores for new patients was provided in order to examine their values relative to the normal range. Here only four tests are used, GGT, AST, ALP, ALT. Plot (a) shows a projection of this sample relative to the normal range. Three of the patients are outside the confidence ellipse but all of the patients are located away from the mean. What has been typical in the past is to compute normal values based on tests of healthy young males. The samples provided for the project were all recorded on women. We see that this sample has slightly lower ALT and ALP, which is consistent with what is reported in Lala et al. (2023).

The second example shows a longitudinal record of a single patient, measured at ages 45, 50, 55, 65 and 70. The lines connect the records in tome order. The projection corresponds to the maxima from a projection-pursuit guided tour using the anomaly index:

$$P = \begin{bmatrix} 0.371 & -0.128 \\ -0.388 & 0.732 \\ 0.140 & 0.599 \\ -0.832 & -0.297 \end{bmatrix}$$

From the axes representation in Figure 3 (b) of this projection, we can see that the direction that profile extends is primarily contrasting ALT (fourth row) and ALP (third row). This is consistent with aging, where ALP increases and ALT decreases.



(a) Sample of female patients. (b) Longitudinal profile of an aging patient.

Figure 3: Two projections of simulated example data corresponding to common liver tests: (a) sample of female patients, (b) longitudinal test results for a single patient. Red cross indicates observation is outide the 4-D confidence ellipse. The female patients tend to have lower ALP and ALT than the normal range. As the patient aged, the level of ALP increases and ALT decreases.

6.2 Robust statistics: weather extremes

This example is motivated by the weather data example from Filzmoser et al. (2018). We illustrate using the anomaly index to compare potential outliers with a reference normal

distribution that is derived using robust methods on the original data. The data contains 16 numerical measurements that are averages across the three summer months June, July and August, these are provided for 68 years (1955 - 2022), see Filzmoser et al. (2018) for more details.

To estimate the underlying normal distribution the data is first centered and scaled using the median and the median absolute deviation (MAD), before applying the minimum covariance determinant (MCD) estimator (Rousseeuw 1985) using the implementation in Maechler et al. (2024). The MCD estimates for the mean vector and the variance-covariance matrix are then used to define the reference normal distribution.

Here we will consider points to be outlying if they are more than 5σ away from that mean value, given $p = 16$ this corresponds to a Mahalanobis distance of 60 or larger. This will identify 20 out of the 68 observations as outlying. Since these are outlying in different combinations of variables, as found in Filzmoser et al. (2018), we will further separate the outlying points into clusters. To focus on the direction when grouping the points we first normalize observations to have length 1 and then apply k-means clustering. Considering different cluster validation statistics computed from Hennig (2024) suggests that $k = 4$ or $k = 5$ is preferred, for simplicity we will work with $k = 4$.

The new anomaly index is first applied to the full dataset, such that the final projection will identify a compromise and provide a global picture of where the outlying points differ from the normal distribution.

7 Conclusion

Say something about the relationship with Peter J. Rousseeuw & Hubert (2018) and Stahel-Donoho outlyingness. This method could be implemented in the way we used to show ellipses, generate points on the surface of the irregular shape, overlay the data on this and make projections.

Difference from general outlier detection, see use of Mahalanobis distance in Filzmoser et al. (2018).

Potential new directions.

References

Asimov, D. (1985), 'The Grand Tour: A Tool for Viewing Multidimensional Data', *SIAM Journal of Scientific and Statistical Computing* **6**(1), 128–143.

Buja, A. & Asimov, D. (1986), 'Grand Tour Methods: An Outline', *Computing Science and Statistics* **17**, 63–67.

Buja, A., Cook, D., Asimov, D. & Hurley, C. (2005), Computational Methods for High-Dimensional Rotations in Data Visualization, *in* C. R. Rao, E. J. Wegman & J. L. Solka, eds, 'Handbook of Statistics: Data Mining and Visualization', Elsevier/North-Holland, Amsterdam, The Netherlands, pp. 391–414.

Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995), 'Grand tour and projection pursuit', *Journal of Computational and Graphical Statistics* **4**(3), 155–172.

URL: *https://www.tandfonline.com/doi/abs/10.1080/10618600.1995.10474674*

Cook, D., Lee, E.-K., Buja, A. & Wickham, H. (2006), Grand Tours, Projection Pursuit Guided Tours and Manual Controls, *in* C.-H. Chen, W. Härdle & A. Unwin, eds, 'Handbook of Data Visualization', Springer, Berlin.

Filzmoser, P., Hron, K. & Templ, M. (2018), *Applied Compositional Data Analysis*, Springer, Cham, Switzerland.
URL: *https://doi.org/10.1007/978-3-319-96422-5*

Friedman, J. H. & Tukey, J. W. (1974), 'A Projection Pursuit Algorithm for Exploratory Data Analysis', *IEEE Transactions on Computing C* **23**, 881–889.

Gabriel, K. R. (1971), 'The Biplot Graphical Display of Matrices with Applications to Principal Component Analysis', *Biometrika* **58**, 453–467.

Gower, J. C. & Hand (n.d.).

Hennig, C. (2024), *fpc: Flexible Procedures for Clustering*. R package version 2.2-12.
URL: *https://CRAN.R-project.org/package=fpc*

Huber, P. J. (1985), 'Projection Pursuit (with discussion)', *Annals of Statistics* **13**, 435–525.

Jones, M. C. & Sibson, R. (1987), 'What is Projection Pursuit? (with discussion)', *Journal of the Royal Statistical Society, Series A* **150**, 1–36.

Lala, V., Zubair, M. & Minter, D. A. (2023), 'Liver function tests'. Updated 2023 Jul 30.
URL: *https://www.ncbi.nlm.nih.gov/books/NBK482489/*

Lee, S., Cook, D., da Silva, N., Laa, U., Spyrison, N., Wang, E. & Zhang, H. S. (2022), 'The state-of-the-art on tours for dynamic visualization of high-dimensional data', *WIREs Computational Statistics* **14**(4), e1573.
URL: *https://github.com/dicook/wiley-isghdd*

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T. & Anna di Palma, M. (2024), *robustbase: Basic Robust Statistics*. R package version 0.99-2.

URL: *http://robustbase.r-forge.r-project.org/*

Mahalanobis, P. C. (1936), 'On the generalised distance in statistics', *Sankhya A* **80 (Suppl 1)**, 1–7.

URL: *https://doi.org/10.1007/s13171-019-00164-5*

Peter J. Rousseeuw, J. R. & Hubert, M. (2018), 'A measure of directional outlyingness with applications to image data and video', *Journal of Computational and Graphical Statistics* **27**(2), 345–359.

URL: *https://doi.org/10.1080/10618600.2017.1366912*

Rousseeuw, P. J. (1985), 'Multivariate estimation with high breakdown point', *Mathematical statistics and applications* **8**(283-297), 37.

Wickham, H., Cook, D., Hofmann, H. & Buja, A. (2011), 'tourr: An R Package for Exploring Multivariate Data with Projections', *Journal of Statistical Software* **40**(2).

URL: *http://www.jstatsoft.org/v40/i02/*

Wickham, H., Cook, D., Spyrison, N., Laa, N., Zhang, H. S. & Lee, S. (2024), *Tour Methods for Multivariate Data Visualisation*. R package version 1.2.0.

URL: *https://CRAN.R-project.org/package=tourr*