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

Annalisa Calvi, Ursula Laa, Dianne Cook

June 26, 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 (REF) 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 (REF) 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 (REF). 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.

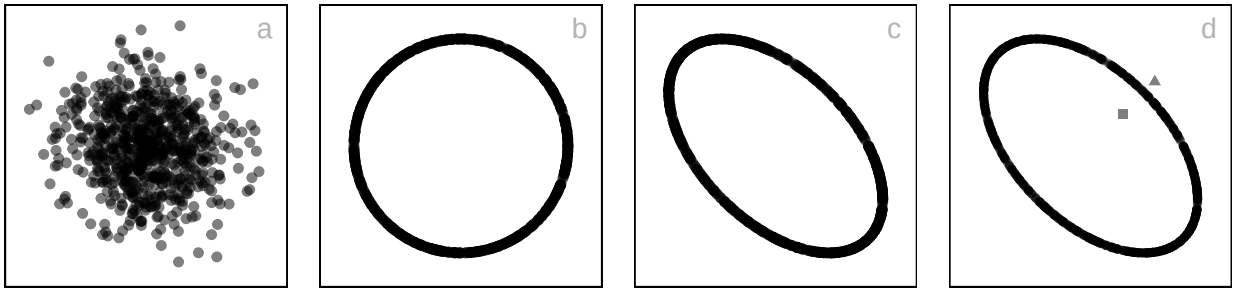


Figure 1: Simulating a uniform sample on a sphere involves sampling from a multivariate normal (a) and transforming each observation to have length equal to 1. A confidence ellipsoid is generated by transforming the sphere relative to a specified variance-covariance matrix (c), and new observations can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

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. We call points in the projection that are on the curve $\boldsymbol{y}$. From this we can compute the analog equation for the projection as

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

where P is a $(p \times 2)$ orthonormal basis defining the projection and $\boldsymbol{\mu}_p = \boldsymbol{\mu}P$ the projected mean. 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 Malahanobis distance 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

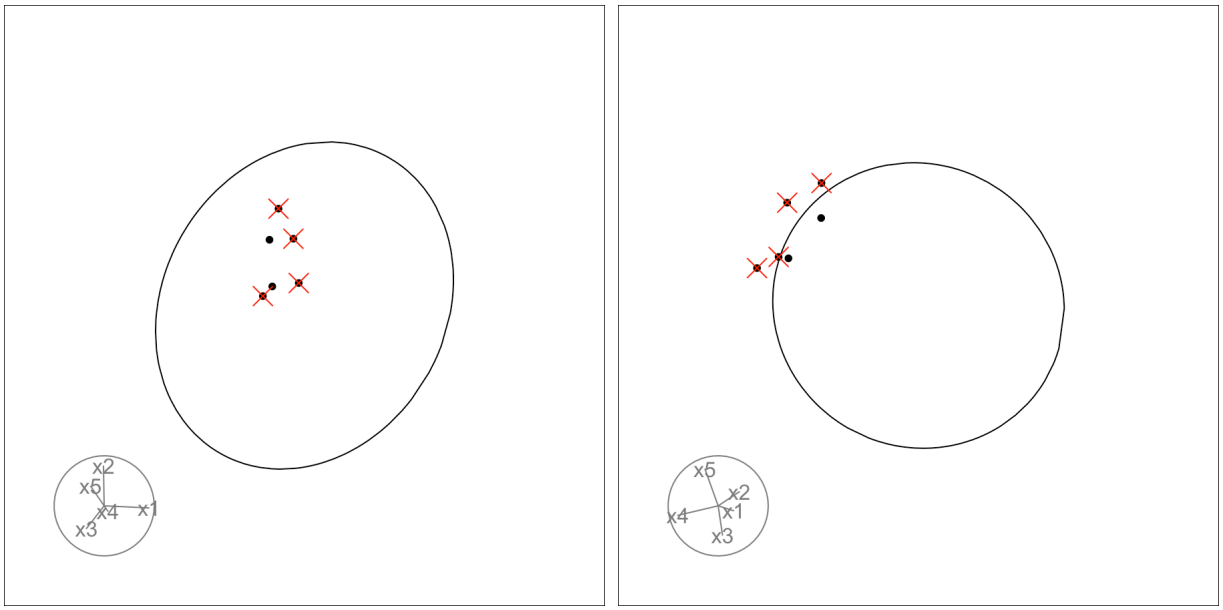
If the observations in W are primarily departing from the normal range in the same direction, the index will be expected to perform well in finding this average direction. However, if the observations have very different departures from the norm, it may be useful to break them into groups, and separately 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` (REF) 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 Medical patients

6.2 Robust statistics

6.3 Physics?

Conclusion

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*