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

Annalisa Calvi, Ursula Laa, Dianne Cook

June 25, 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 (REF).

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 4.

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 ellipse, 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 from $N_p(\mu, \Sigma)$.
2. Transforming each observation to have unit distance from the mean.

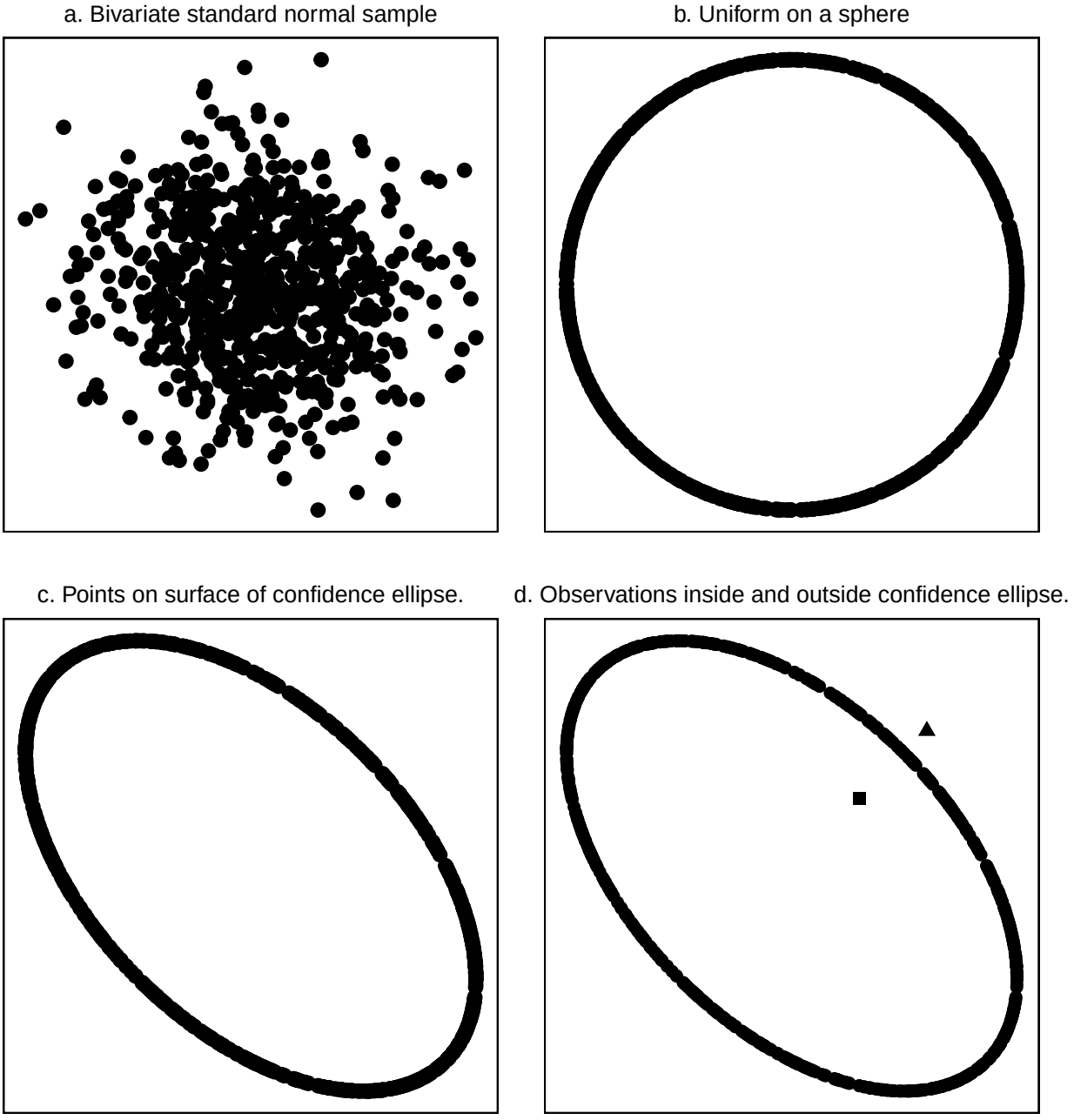

Figure 1: Simulating a uniform sample on a sphere involves sampling from a multivariate normal (a) and transforming each observation to have length equal to 1. A confidence ellipse is generated by transforming the sphere relative to a specified variance-covariance matrix (c), and new observations can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

3 Anomaly index

4 Examples