Grand Tours, Projection Pursuit Guided Tours, and Manual Controls

Created by Weber, YC Huang (note)

P1

Introduce myself~

P2 Summary

In the beginning, I would like to …

How do we find structure in multidimensional data when computer screens are only two-dimension?

It is a tricky problem that many of the Mathematicians, Statisticians and Data Scientists have been researched for a long time.

In this chapter, we will gonna to take a birds view of Multivariate Statistical Analysis with Data Visualization. And the methods introduced in this chapter are based on the application named "GGobis" which is an open source visualization program for exploring high-dimensional data. (Noted that it can be use in R through package “rggobis”, see ggobis).

In addition, three methods of GGobis are discussed in this chapter.

P3 Contents

Table of content

P4

Intro

P5 What is projection ? (1/2)

What is a projection? We can think of a projection as the shadow of an object. Especially if it is a 2-D projection, then the projection is the shadow the object casts under a bright light. If the object rotates in the light, we see many different 2-D shadows and we can infer the shape of the object itself.

P6 What is projection ? (2/2)

Mathematically, a projection of data is computed by multiplying an n x p data matrix, X, having n sample points in p dimensions, by an orthonormal p x d projection matrix, A, yielding a d-dimensional projection. For example, to project a 3-D object (3 columns, or variables, of data) onto a 2-D plane (the shadow of the object), we would use an orthonormal 3 x 2 matrix.

P7 The blind men and the elephant

What is hidden from the user who views only a few static projections? There could be a lot. You may be familiar with an ancient fable from India about the blind men and the elephant. One grabbed his tail and swore the creature was a rope. Another felt the elephant’s ear and yelled it was a hand fan. Yet another grabbed his trunk and exclaimed he’d found a snake. They argued and argued about what the elephant was, until a wise man settled the fight. They were all correct, but each described different parts of the elephant. Looking at a few static projections of multivariate data is like the blind men feeling parts of the elephant and inferring the nature of the whole beast.

P8 Interpolation methods

Static projections can be strung together into a movie using interpolation methods, providing the viewer with an overview of multivariate data.

These interpolation methods are commonly called tours.

Allowing the viewer to mentally connect disparate views, and thus supporting the exploration of a high-dimensional space.

We use tours to explore multivariate data like we might explore a new neighborhood: walk randomly to discover unexpected sights, employ a guide, or guide ourselves using a map.

These modes of exploration are matched by three commonly available types of tours. They are the tours available in the software, GGobi

§Grand tour

§PP guide tour

§Manual control

§When we use tours, what are we looking for in the data? We search for data projections that are not bell-shaped and, hence, not normally distributed, for example, clusters of points, outliers, nonlinear relationships, and low-dimensional substructures.

P9

Tours

P10 Movement of a projection

Most of us are familiar with 3-D rotation, which is something we can do in the real world. We can take an object and rotate it with our hands or walk around an object to view it from all sides.

Movement of a projection plane is achieved by selecting a starting plane and a target plane and computing intermediate planes using a geodesic interpolation. A geodesic is a circular path, which is generated by constraining the planar interpolation to produce orthonormal descriptive frames. This is the method used in GGobi.

Differences in the method of selecting the target plane yield different types of tours. The grand tour uses a random selection of target plane. The guided tour quantifies the structure present in each projection and uses this to guide the choice of target plane. Manual controls let the user choose the target direction by manipulating the values of projection coefficients.

P11 Grand Tour

The grand tour method for choosing the target plane is to use random selection. A frame is randomly selected from the space of all possible projections.

A target frame is chosen randomly by standardizing a random vector from a standard multivariate normal distribution: sample p values from a standard univariate normal distribution, resulting in a sample from a standard multivariate normal. Standardizing this vector to have length equal to one gives a random value from a (p-1)-dimensional sphere, that is, a randomly generated projection vector. Do this twice to get a 2-D projection, where the second vector is orthonormalized on the first.

Fig. Some views of 1-D grand tour paths in 3 dimensions (top let) and 6 dimensions (bottom). The path consists of a sequence of points on a 2-D and 6-D sphere respectively. Each point corresponds to a projection from 3 dimensions (or 6 dimensions) to 1 dimension. he solid circle indicates the first point on the tour path corresponding to the starting frame, yielding the 1-D data projection (top right) shown for the 3-D path. he solid square indicates the last point in the tour path, or the last projection computed

P12 Projection Pursuit Guided Tour

In a guided tour the next target basis is selected by optimizing a PP index function. The index function numerically describes what is interesting in a projection: higher values correspond to more interesting structure in the projections.

Using a PP index function to navigate the high-dimensional data space has the advantage over the grand tour of increasing the proportion of interesting projections visited.

The optimization procedure is an important part of a PP guided tour. The purpose of PP optimization is to find all of the interesting projections, so an optimization procedure needs to be flexible enough to find global and local maxima. It should not doggedly search for a global maximum, but it should spend some time visiting local maxima.

Fig. (top two plots) shows a PP guided tour path (1-D in three dimensions). It looks very similar to a grand tour path, but there is a big difference: the path repeatedly returns to the same projection and its negative counterpart (both highlighted by large solid black circles). The middle plot traces the PP index value over time. The path iterates between optimizing the PP function and random target basis selection. The peaks (highlighted by large solid black circles) are the maxima of the PP index, and for the most part, these are at the same projection. The corresponding data projections (approximately positive and negative of the same vector) are shown in the bottom row. The index is responding to a bimodal pattern in the data.

There are numerous PP indices. Like PCA, Holes, Central Mess, etc.

P13 Manual Controls

Manual controls enable the user to manually rotate a single variable into or out of a projection. This gives fine-tuning control to the analyst.

Fig. Manual controls used to examine the sensitivity of the clustering revealed by the LDA index to PC6 and PC2 is explored. Top left to top right plots: coefficient on PC6 reduced from 0.122 to 0.004 gives a smaller gap between clusters, but the clusters are still separable. Bottom left to bottom right: coefficient for PC2 reduced from 0.439 to 0.026 which removes the cluster structure

P14

Using …

P15

Tours are useful when used along with numerical methods for certain data analyses, such as dimension reduction and supervised and unsupervised classification.

Tours can help numerical approaches in many ways

§Choose which of the tools at hand works best for a particular problem.

§Understand how the tools work in a particular problem.

§Overcome the limitations of a particular tool to improve the solution.

P16

End notes

P17

§Tours support exploring real-valued data.

§They deliver many projections of real-valued data in an organized manner, allowing the viewer to see the data from many sides.