Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

Using Scagnostics to find Interesting Projections of Multivariate Data

Ursula Laa
Monash University

Second Author
Affiliation

Abstract

The abstract of the article.

Keywords: scagnostics, tour, projection pursuit, multivariate data, statistical graphics, data visualisation.

1. Introduction

A grand tour provides dynamic projections from the high dimensions into some low dimensional representation, in particular also as a two dimensional scatter plot ADDREF. The idea of projection persuit can be coupled to the tour algorithm, to guide the tour towards more interesting views of the projected data (the so-called guided tour), by maximising a given projection persuit index ADDREF (Cook, Buja, Cabrera, Hurley, 1995). The guided tour thus allows the user to view interesting projections (local maxima or minima of given index) as well as their neighborhoods, allowing the user to see the projection in the context of the full parameter space. Typically projection persuit indices are built to detect deviations of the projected data from a normal distribution, and the nonnormalness is considered equivalent to interestingness of a given projection ADDREF (Cook, Buja, Cabrera, 1993). Moreover, indices for detecting holes, concentration of mass in the center or skewness have also been designed.

Here we aim to extend the definition of interesting projections and available projection persuit indices by studying potential new index functions. As a starting point we consdier the scagnostics measures that were proposed for selecting bivariate scatterplots in high-dimensional datasets ADDREF (??). Rather than considering the scagnostics measures for all possible bivariate combinations, we want to use one of the measures as an index function for a guided tour of 2 dimensional projections. We first study the feasability of this approach and discuss potential pitfalls. We then aim to define more robust measures … ???

2. Scagnostics measures

Eight scagnostics measures have been defined, the definitions are based on the convex and alpha hull as well as the minimum spanning tree. Give definitions here…

Note that in practice the measures are calculated after binning the data to avoid large computing times.

3. Analysis setup

To study the feasablility of using scagnostics measures as an index fuction for guided tours, we first consider the variation of each of the measures, as calculated from the projected scatter plot, when smoothly changing the projection via the tour algorithm. Moreover, we test the dependence on the sample size and on the outlier removal typically performed to make the scagnostics measures more robust. In addition the computing time is also important, as it might limit the dynamic viewing and require to first record a guided tour path which can later be viewed in a smooth display.

As we will see the scagnostics measures often exhibit a significant jump despite changing smoothly between projections, a consequence of the discreete nature of the definition, but possibly also due to the binning performed as a first step of calculation. We therefore explore options of ameliorating this situation using different smoothing methods, and by redefining the measures in a more robust fashion.

4. Application to particle physics data

4.1. Data description

For the following illustrations we work with the particle physics dataset (ATLAS Collaboration, 2015). This data includes a number of predictions for observables (in agreement with current limits) and the so-called finetuning measure for a large set of parameter points in a supersymmetric model. The observables describe predictions for decay branching ratios and other precision observables that are generally used to constrain new models of particle physics. We use the log transformed values for the variables XYZ but not the finetuning, and we rescale all variables to a [0,1] interval to obtain comparable scales. Contributions to the variable gminus2 can be negative or positive, we therefore add a small shift before the log transformation of this variable. Maybe include GGally pairs plot here?

4.2. Software setup

Use tourr and scagnostics package (ADDREF)

4.3. Scagnostics of grand tour projections

(ref:scagnostics-summary) Summary of evolution of scagnostics measures over 100 interpolated grand tour projections We see that three of the considered measures, Clumpy, Sparse and Striated, are not relevant for the considered data and projections, as the valuse are consistently low (clearly below 0.1). The remaining measures show large variation depending on the projection, sample size and also often strong dependence on the outlier removal before calculation of scagnostics measures. Moreover, we note that despite considering interpolated projections, i.e. a smooth transition between the randomly selected projection planes in the grand tour, the variation of the scagnostics measures does not follow a smooth evolution but exhibits several spikes. This behaviour is expected to result in difficulites when using scagnostics measures as a projection pursuit index in the context of a dynamic guided tour. To further understand this behaviour let us study the projected data points and also the binned version for some of the projections exhibiting a spike. We first show in Figure ADDREF projections t=18/19/20 of the small data set (500 data points). Not that t=19 has been marked by a vertical line in Figure ADDREF. The left column shows the projected data points, and the right column the binned data used to calculate the scagnositcs measures. The scagnostics measures in the three projections are summarised in Table ADDREF. We see that despite similar distribution of the points in all three projections, the outlying measure is similar for t=18 and t=20 but jumpig down for t=19. Differences can however be observed in the binned data (right column of Figure ADDREF), likely causing the behaviour of the measures.

(This is blueprint of how to show scatter and binning plots but maybe other projections will be more interesting.)

FIXME need to arrange plots better! Fix common axes for all corresponding plots FIXME binned plot needs better formatting so we can see relevant features easily

Another consideration relevant for the purpose of the guided tour is the computing time, as mentioned above it may limit the tour display to previously recorded tours. We study here the computing time depending on the sample size with and without the outlier removal when calculating the measures. Figure ADDREF shows this variation for 3 different sample selections. The time is measured over the 100 projections shown before and is the sum of computing time for projecting the data and calculating the scagnostics measures on the projected data. We see that overall the outlier removal increases the runtime by about a factor two. However we have seen above that skipping the step of outlier removal often leads to very different values of some of the scagnostics measures. (What to do with this information? Should not skip this step?) Interestingly the computing times are longest for intermediate sample sizes of a few hundred data points, and converge to lower run times for larger samples. This is because the binning of the input data is much faster than the computation of the measures which is increased for smaller samples.

4.4. Smooth transitions

We next turn to the question of how we can obtain smooth transitions of the scagnostics measures. (Below just an example of using ggplot smoothing to display what kind of behaviour we want to find. It follows the general trend without capturing the spikes.)

Want to explore jittering either the projection or the points, but this does not give the desired smooth behaviour (why?) Maybe show how it compares to the sketch of smooth polynomial function? Select parameters from previous study without including the detailed plots here..

Binning in high dimensions?

Outlier removal before binning?

4.5. Test with guided tour

try out scagnostics measures as guided tour index

Example below uses Convex as index function. We show the evolution of the index on the interpolated guided tour (top) and on the selected planes without interpolation (bottom). Note that the interpolated projections do not smoothly increase the index value from one selected plane to the next, but shows the unstable nature of the index computation.

Figure (ref:guided-tour-scatter) compares the scatter plots (left unbinned, right binned) of the start projection, the projection with the lowest index value (t=19 of the interpolated projections) and the final projection (t=28).

show computing time? for each index as function of sample size? (long to prepare such a plot..)

4.6. Additional indices to consider

Additions to the standard scagnostics measures have been implemented in the package mbgraphics, based on splines (spline2d) and distance correlation (dcor2d). Give definitions here.. We repeat the study from before, we first test in Figure ADDREF the behaviour of the index values over 100 interpolated projections that are randomly selected via the grand tour algorithm (the same projections as considered above). Note that here the measures are calculated on non-binned data. The general behaviour seems similar to the scagnostics monotonic index. The observed behaviour is smooth (much in contrast to the behaviour of the scagnostics measures) and we next test the behaviour in a guided tour, comparing again also the first and final projection. (which index to select, will try both below…) It seems they both maximise correlation, this might not be too useful? Seems like [,1] basically -[,2] in the final projection (except for [7,])? probably we see only structure of gminus2 variable, which has very narrow peak, only few outlying points is there a better way of transforming it to spread out the structure?

Physics: what is final projection, what is special about outlying points?

maybe move timing to the end, add also these indices in the same plot (instead of testing several samples, behaviour anyway similar)

Affiliation:

Ursula Laa
Monash University
First line Second line
E-mail: `ursula.laa@monash.edu`
URL: `http://rstudio.com`

Journal of Statistical Software `http://www.jstatsoft.org/`
published by the Foundation for Open Access Statistics `http://www.foastat.org/`
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
`doi:10.18637/jss.v000.i00` *Accepted:* yyyy-mm-dd