EPJ Web of Conferences **245**, 06018 (2020) https://doi.org/10.1051/epjconf/202024506018
CHEP 2019

High-dimensional data visualisation with the grand tour

Ursula Laa[1,2,*]

[1]School of Physics and Astronomy, Monash University, Melbourne VIC-3800
[2]Department of Econometrics and Business Statistics, Monash University, Melbourne VIC-3800

Abstract. In physics we often encounter high-dimensional data, in the form of multivariate measurements or of models with multiple free parameters. The information encoded is increasingly explored using machine learning, but is not typically explored visually. The barrier tends to be visualising beyond 3D, but systematic approaches for this exist in the statistics literature. I use examples from particle and astrophysics to show how we can use the "grand tour" for such multidimensional visualisations, for example to explore grouping in high dimension and for visual identification of multivariate outliers. I then discuss the idea of projection pursuit, i.e. searching the high-dimensional space for "interesting" low dimensional projections, and illustrate how we can detect complex associations between multiple parameters.

1 Introduction

Data visualisation is an important part of the statistical analysis of data, and can often provide insights beyond what is found with standard summary statistics. This is illustrated by Anscombe's quartet [1], four datasets with the same statistical properties, including the mean and variance along each variable, as well as the correlation and regression line. The same datasets show clearly distinct dependencies that are revealed in a scatter plot, as can be seen in Fig. 1.

Visualisation is equally important in physics, where it is used to guide our intuition on phenomena beyond direct experience. We use visualisation for the analysis of our results, in order to understand and interpret them. It is also a powerful method for diagnosing problems, and often preferred for communicating results.

Typical physics data is however high-dimensional, both in the case of experimental measurements (where observations are commonly multivariate) and in theoretical studies (where models typically have multiple free parameters). Most visualisations thus rely on some form of dimension reduction. The most common methods are projecting, marginalising or profiling, which is done either for all parameter combinations (as in a scatter plot matrix) or for a selection based on prior knowledge. Here we show how the grand tour can be used to look at the data in more than three dimensions instead.

2 The grand tour

The grand tour [3] shows multivariate distributions by displaying a smoothly interpolated sequence of randomly selected low-dimensional linear projections. We can imagine rotating

*e-mail: ursula.laa@monash.edu

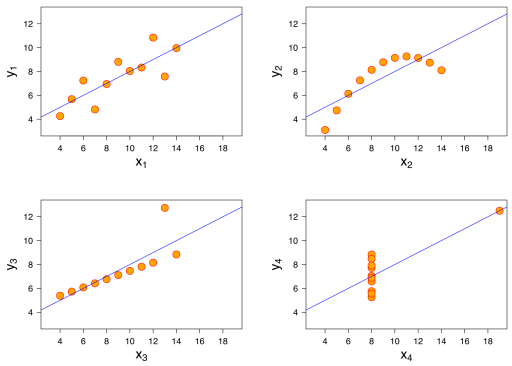

Figure 1. Anscombe's quartet (taken from [2]), four datasets that have nearly identical summary statistics but we can clearly identify differences in a graph.

the high-dimensional shape and looking at low-dimensional views from different angles, allowing us to extrapolate to the multidimensional shape. This is also how we visualise 3D objects on a 2D screen, and the grand tour enables us to look at distribution in any number of dimensions.

The grand tour can be used to visualise higher dimensional geometric shapes, for example hypercubes or tori in multiple dimensions. These visualisations help us understand structures in high-dimensional Euclidean space. A library of examples has been presented in [4], with a large number of example animations available here [5].

2.1 Exploring structure in high dimensions

We can use the grand tour to identify and explore grouping, and to find outlying points and other types of structure in a high-dimensional distribution. This was done in [6] for a study of the sensitivity of hadronic experiments to nucleon structure. The sensitivity is encoded in "fit residuals" [7], defining a 56 dimensional parameter space. The visualisation is used to study:

- Grouping: how do different types of measurements constrain different directions in parameter space.
- Relevance to the fit: the residuals encode the sensitivity of the global fit to single measurements, so finding observables that are different from the main distribution (i.e. the other measurements) corresponds to finding those which are expected to be important in future fits.

To study this data we first group the experimental observations according to the type of measurement into: deep-inelastic scattering (DIS), vector boson production (VBP) and jet production (jets) measurements. In addition we use principal component analysis to reduce the size of the parameter space to be visualised from 56 down to 6 dimensions.

A first look at the data with the grand tour already provides interesting insights. A static example is shown in Fig. 2, and the full animation is available here [8]. The animation shows that the three groups are orthogonal in the high-dimensional space. We also find that all points in the jets data fall in a 2D plane in this 6D space.

By focusing on one type of measurement we can get a more detailed understanding of the results. As an example we consider the jets cluster and we look at the data in the first four principal components obtained for this group. The data is separated into individual experimental studies for a detailed comparison in a grand tour display of this 4D space. A static view is shown in Fig. 3, the full animation can be found here [9].

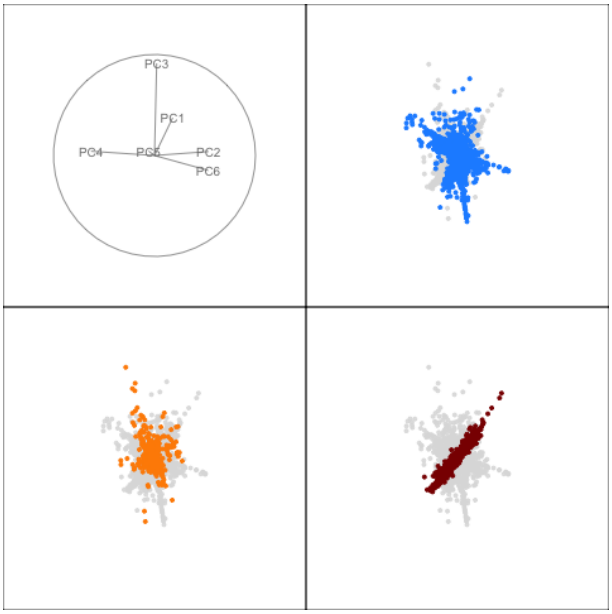

Figure 2. Example projection of the full dataset viewed with the tour, see here for full animation. The points are grouped into DIS (blue), VBP (orange) and jets (red) type measurements and the example illustrates the orthogonality between these groups.

The visualisation clearly highlights outlying points in the distribution, and we have used different marker symbols to match the most interesting ones to the metadata associated with each observation. Several of the outlying points are found to genuinely extend the reach in parameter space, in a consistent way for groups of similar measurements. These are for example measurements at large rapidity and large momentum transfer, e.g. points with $|y| > 2$ and $\mu > 1000$ GeV are marked with downward pointing triangles. In the tour animation we can study their distribution in the multidimensional parameter space, and find that they indeed extend in a similar direction of parameter space, away from the remaining observations.

One point is found to be inconsistent with expectations and should be investigated more closely. This is the ATLAS7new [10] measurement in the last rapidity bin ($|y| > 2.5$, $\mu >$ 950 GeV), and is marked wit a star symbol. This point is clearly visible as a multivariate outlier in Fig. 3.

For the complete mapping between marker symbols and kinematic regions see [6].

3 Projection pursuit

The idea of projection pursuit [11, 12] is to find interesting low-dimensional projections of high-dimensional data by optimizing an index function over all possible projections. This idea can be combined with the grand tour, defining the guided tour [13]. The guided tour uses the index when selecting the next plane in the sequence of projections. By only selecting planes with larger index values than the current one, the guided tour path is moving towards more interesting views of the data as the animation progresses.

This raises the question of how to define the "interestingness" of low-dimensional projections. Typical index functions aim to detect departures of the projected distribution from

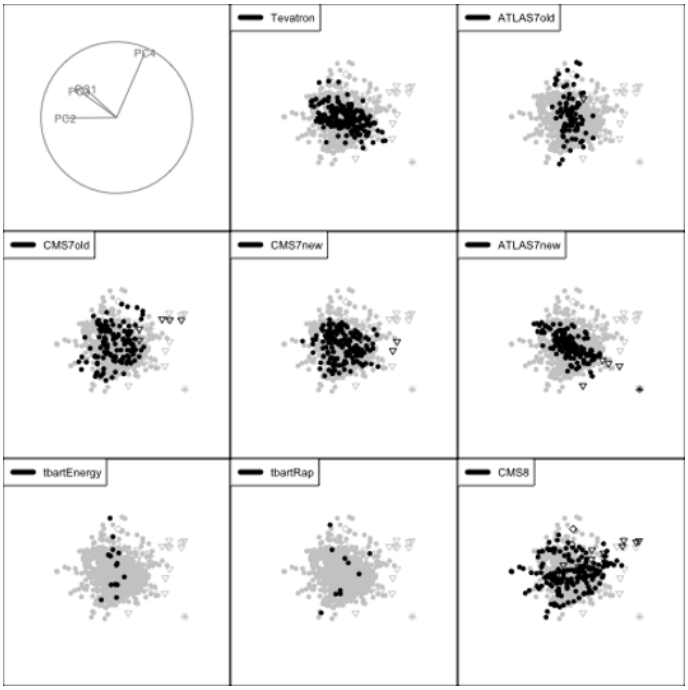

Figure 3. Example projection of the jets dataset viewed with the tour, see here for full animation. The points are grouped by experimental analysis, and selected kinematic regions are highlighted with different marker symbols. We can identify multivariate outliers from the visualisation.

a known distribution (most commonly the normal distribution). In physics research finding complex bivariate patterns in projections is of particular interest. This was discussed in [14] which considered index functions developed for variable selection in large datasets, and evaluated their potential as projection pursuit indexes. The considered indexes include the scagnostics measures [15] for characterising 2D scatter plots, and indexes based on mutual information [16].

These index functions were applied to posterior samples from fitting gravitational wave signals. The example in Fig. 4 shows the views identified by optimising different indexes for the 11D posterior draws obtained when fitting a simulated gravitational wave signal from a binary black hole merger [17]. In this case projection pursuit identified the strong association between the two parameters describing the position in the sky, right ascention (ra) and declination (dec), and the time of the event (time). Similar views were identified by the mutual information based TIC index and the spline based index from [18]. On the other hand, the scagnostics index convex failed to optimise the view for this dataset and only shows a nuisance distribution.

4 Graphical interface

A graphical interface for using both the grand and guided tour is available in the R [19] package galahr [20], and is based on the implementation of tour methods in R in the tourr package [21]. The interface allows the user to upload a datafile, and select different options for setting up the tour in an input tab. The results tab then displays the resulting animation,

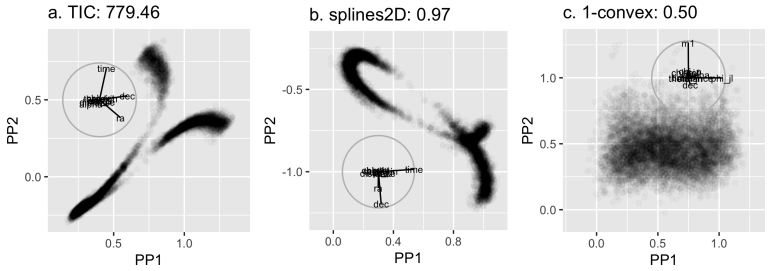

Figure 4. Optimal views of an 11D posterior sample fitting a simulated gravitational wave signal, obtained through projection pursuit optimising the TIC (left), splines2d (middle) and 1-convex (right) index function. The first two views show structure, while the 1-convex index failed to find anything interesting.

and has several interactive features, including a timeline that lets the user move along the tour path, linked brushing between 1D projections along each input parameter and the tour display, and tooltip information for the identification of the projected points.

5 Summary and Outlook

The talk presented an overview of how tour methods can be used in physics, to learn about multivariate structures through visualisation. The examples show how we can understand grouping, find outlying points and uncover views that reveal structure hidden in combinations of multiple model parameters.

So far there are only few applications to physics problems, see references [6, 14, 22]. The graphical interface available in the galahr package was designed to make tour visualisations more accessible, to enable a broader application of the methods.

An important caveat of projection based visualisations is that they may hide concave or internal features. These structures can be revealed by looking at slices through the high-dimensional space instead. A simple approach to display interpolated slices based on projection planes obtained when running a tour was recently presented in [23].

Acknowledgments

This work was supported in part by the Australian Government through the Australian Research Council.

References

[1] F.J. Anscombe, The American Statistician **27**, 17 (1973), https://www.tandfonline.com/doi/pdf/10.1080/00031305.1973.10478966
[2] https://commons.wikimedia.org/w/index.php?curid=9838454
[3] D. Asimov, SIAM J. Sci. Stat. Comput. **6**, 128–143 (1985)
[4] B. Schloerke, H. Wickham, D. Cook, H. Hofmann, The R Journal **8**, 243 (2016)
[5] http://schloerke.github.io/geozoo/
[6] D. Cook, U. Laa, G. Valencia, Eur. Phys. J. **C78**, 742 (2018), 1806.09742
[7] B.T. Wang, T.J. Hobbs, S. Doyle, J. Gao, T.J. Hou, P.M. Nadolsky, F.I. Olness, Phys. Rev. **D98**, 094030 (2018), 1803.02777

[8] https://uschilaa.github.io/animations/pdfsense1.html
[9] https://uschilaa.github.io/animations/pdfsense2.html
[10] G. Aad et al. (ATLAS), JHEP **02**, 153 (2015), [Erratum: JHEP09,141(2015)], 1410.8857
[11] J.B. Kruskal, in *Statistical Computation*, edited by R.C. Milton, J.A. Nelder (Academic Press, New York, 1969), pp. 427–440
[12] J.H. Friedman, J.W. Tukey, IEEE Transactionson Computers **23**, 881 (1974)
[13] D. Cook, A. Buja, J. Cabrera, C. Hurley, Journal of Computational and Graphical Statistics **4**, 155 (1995)
[14] U. Laa, D. Cook, Computational Statistics (2020), https://doi.org/10.1007/s00180-020-00954-8, 1902.00181
[15] L. Wilkinson, A. Anand, R. Grossman, *Graph-theoretic scagnostics*, in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (2005), pp. 157–164, ISSN 1522-404X
[16] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Science **334**, 1518 (2011), http://science.sciencemag.org/content/334/6062/1518.full.pdf
[17] R. Smith, S.E. Field, K. Blackburn, C.J. Haster, M. Pürrer, V. Raymond, P. Schmidt, Phys. Rev. **D94**, 044031 (2016), 1604.08253
[18] K. Grimm, *mbgraphic: Measure Based Graphic Selection* (2017), r package version 1.0.0, https://CRAN.R-project.org/package=mbgraphic
[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2018), https://www.R-project.org/
[20] U. Laa, D. Cook, https://github.com/uschiLaa/galahr (2019)
[21] H. Wickham, D. Cook, H. Hofmann, A. Buja, Journal of Statistical Software **40**, 1 (2011)
[22] B. Capdevila, U. Laa, G. Valencia, Eur. Phys. J. **C79**, 462 (2019), 1811.10793
[23] U. Laa, D. Cook, G. Valencia (2019), 1910.10854