LEA
Leonardo Electronic Almanac

Vol **16** Issue **6 – 7**

INTERACTIVE DATA EXPLORATION WITH TARGETED PROJECTION PURSUIT

Joe Faith, School of Computing, Engineering and Information Sciences, Northumbria University, Newcastle, NE2 1XE, UK
E-mail: <joe.faith@northumbria.ac.uk>

Abstract

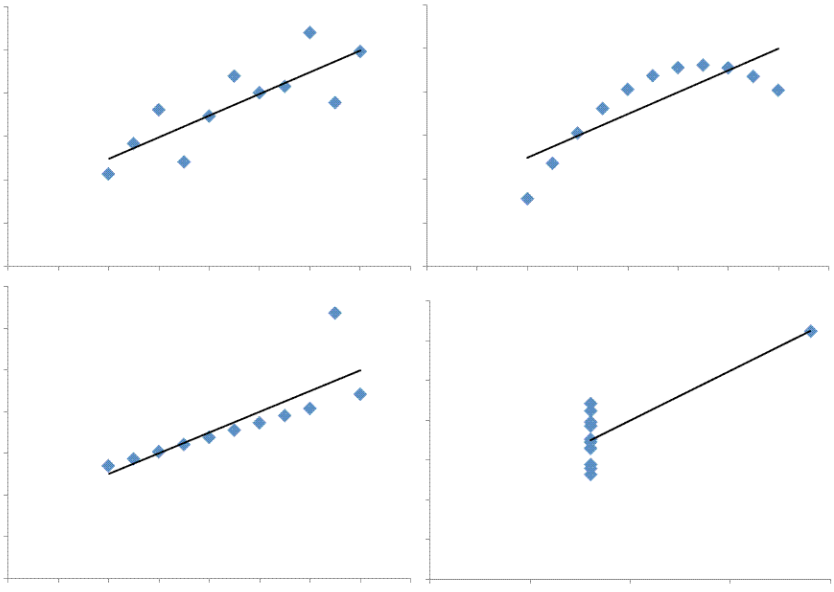
Data exploration is a vital, but little considered, part of the scientific process; but few visualisation tools can cope with truly complex data. Targeted Projection Pursuit (TPP) is an interactive data exploration technique that provides an intuitive and transparent interface for data exploration. A prototype has been evaluated quantitatively and found to outperform algorithmic techniques on standard visual analysis tasks.
Keywords: Data Exploration, Data Visualisation, Visual Analytics

Once a scientist has completed an experiment, the first question they ask is whether the data confirms the hypothesis; but the second question is what else that data is telling them. To do this the experimenter must explore their data, must *eyeball* it. This process is inherently graphical, iterative, and interactive as the scientist tries to visualize their data using techniques such as the familiar scatter plots and line graphs for two or three-dimensional data, supplemented by bubble, radar/spider and, more recently, parallel coordinate plots to visualise up to a dozen dimensions. Sometimes this exploration makes the experimenter fundamentally re-evaluate their data, illustrated by the case of Anscombe's Quartet [2, and Figure 1]: a set of hypothetical data sets with identical statistical properties but which tell very different stories when graphed. In such cases the experimenter is looking for the qualitative nature of the data behind the statistical quantities.

The term Exploratory Data Analysis was coined by Tukey in 1977 [1] to describe this process of analysing data in order to formulate hypotheses to test, and contrasted this with the conventional process of confirmatory data analysis, where statistical techniques are applied to the data to test hypotheses. He thus highlighted the essentially creative process of hypotheses formation that is often neglected in more positivist accounts of the scientific process, in which there is a linear process from hypothesis to experiment to data to confirmation, ignoring how the 'experimental loop' is closed as that data is subsequently explored to suggest further hypotheses and experiments.

However conventional visualisation techniques are reaching their limits as the complexity of data grows. This is especially the case in biology where the adoption of 'lab on a chip' technologies has increased by orders of magnitude the volume and complexity of the data available to scientists who are not naturally numerate [3]. This has led to the birth of the 'omics' sciences (genomics, proteomics, interactomics, etc) and the application of machine learning and data mining techniques to the resulting data (known as bioinformatics). DNA microarrays, for example, measure the expression levels of tens of thousands of genes in samples, such as from cancer tumours. The experimenter then tries to understand the relationship between those gene expression levels and the nature of the cancer observed by a clinician [4].

Fig 1. Anscombe's data sets have identical statistical properties, but have different qualitative properties to a human observer.



From Dimension Reduction to the Grand Tour

Traditional visualisation techniques cannot represent this complexity, so one response is to reduce the dimensionality of the data. Two basic dimension reduction techniques are commonly used: linear projections, and non-linear techniques such as multi-dimensional scaling (MDS) [5]. A linear projection can be thought of as a view of a high dimensional data space through a two-dimensional 'window': each position of that window yields a different view of the data. Principal Components Analysis (PCA), for example, is a widely used projection-based dimension-reduction technique that positions the projection window such that the data points are as widely spread as possible. MDS, on the other hand, 'squeezes' the original high-dimensional data onto a two-dimensional plane such that the relationships between the points are captured as accurately as possible.

Both approaches have their limitations. MDS inevitably misrepresents relationships in the data to some extent – known as the degree of stress – and this can mislead the experimenter. If MDS is used to visualise the gene expression levels of cancer tumours, for example, stress may show one sample as being more closely related to those of another cancer type than is actually the case. Another limitation on MDS is the 'curse of dimensionality': in high-dimensional space there is little variance in distances between points, so any lower-dimensional mapping that reproduces these distances will show data points evenly dispersed, losing any underlying structure and blurring patterns.

Linear projections, on the other hand, do not suffer the problem of stress and

can mitigate the curse of dimensionality, but they can only show a single view, can only represent one aspect, of the data at a time. The resulting views are strictly accurate, but partial.

There are two solutions to the partial nature of linear projections. *Projection pursuit* is one solution, which finds the best or most informative or interesting projection -- usually defined operationally as statistical non-normality. But this is still yields just one view, so Asimov proposed a Grand Tour [7] – described as an attempt to look at the data 'from all possible angles'. A Grand Tour is a video sequence in which each frame shows the result of a single projection of the data, with the sequence as a whole including all possible projection planes. However, the Grand Tour replaces the quality of projection pursuit with quantity: a grand tour in high dimensional space is long and mostly uninformative.

Ideally we need some way of guiding the tour, to use our perception of the data to find regions of interest. One interface that tackles this problem is GGobi, which allows the user to pause and rewind a given Grand Tour, and altering the resulting views by moving the projection window [8]. But this interface is opaque, in the sense that the user can rarely anticipate the effect of their actions. The user has n controls to manipulate, the effect of each will be unknown and which will have unpredictable effects in combination. Only where the user has strong intuitions about the nature of the data can use this interface to reveal the structure more clearly. In other words, once the user knows what they are looking for then such an interface will help them find it. But it is unsuited to true exploration of the data. The user can do little more than random search – which has its place, but is of little use when faced with a complex set where n may be measured in the hundreds or thousands.

Targeted Projection Pursuit

The human visual system is superb at spotting patterns. We are especially good at spotting partial or obscured patterns, ignoring noise, and disregarding outliers: tasks that pattern recognition algorithms struggle with. We are also extremely efficient at recognising structure in correlated movements, as illustrated in Johansson's classic 'point light' experiments [9]; and at detecting the effect of our actions on stimuli, which is the basis of our ability to act reliably in a dynamic environment. Targeted Projection Pursuit (TPP, available online [10]) is a tool for exploring complex data sets based on these strengths: it is an interactive scatter plot in which, instead of the user moving the coordinate system of the data, they manipulate the data itself.

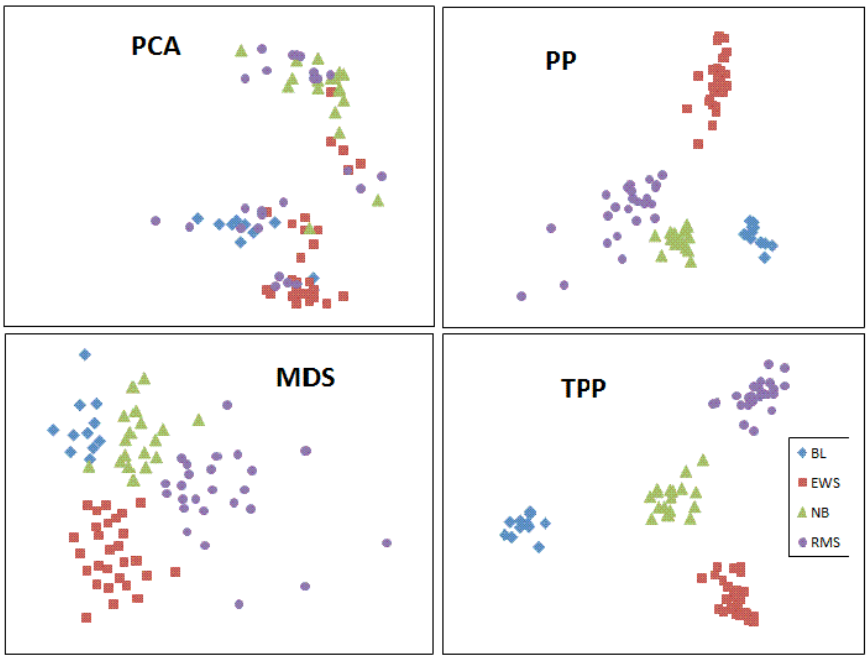


Fig. 2. Comparing views of the same data

Suppose an experimenter graphs some data using a linear projection such as PCA and they notice what seems to be a pattern, such as some clustering or a trend. The immediate question is whether another view of the data would show the pattern more clearly, or whether it is just the result of noise. In this situation, the natural impulse is to try to 'grab' those points that fail to fit the perceived pattern and try to move them into place. TPP allows the user to do just this, by playing with the data to explore possible views. Using TPP the user can try to move the data points to better fit that pattern by using a simple rubber-band drag-and-drop interface, such as selecting one of the clusters to separate it from the other data points. If there is a projection in which these changes can be found, then the points move as the user drags them. If not, then the points stick.

The overall process is thus one of hypothesis- formation and testing: by attempting to move some points, the experimenter is suggesting a hypothesis about the structure of the data; if the data fits the hypothesis, then the result is shown.

Thus the data points respond to the user's actions in real time as the tool tries to find a view of the data under which the user's hypothesis about the nature of the data may be satisfied, and applies that projection to the data to create a new view. The effect is akin to playing with a semi-pliable hyperdimensional solid.

The principal advantage of such an interface is that it is transparent, in the sense that the response of the system is intuitive and predictable. If the user spots a partial pattern then they manipulate the elements of that pattern directly, rather than controls whose effects on the pattern are unpredictable. Thus this tool uses the full power of the human visual system to do what it is best at – spotting patterns – while the computer based tool is left to do dumb linear algebra.

Evaluation

TPP is designed to be used by experimenters, particularly those within bioinformatics. Therefore it has been evaluated on genuine gene expression data sets available in the public domain, and on tasks where a quantitative comparison with standard data analysis techniques is possible: finding views of classified samples, feature selection, and

detecting misclassified samples.

In the first task the user is presented with views of gene expression data sets in which each sample is of a known diagnostic class and used the tool to find views that best show the separation between classes. The resulting two-dimensional view of the data is then tested using standard statistical measures of class separation, and compared with standard dimension reduction techniques. The result is that the user was able to find views of gene expression data that showed a much clearer separation between classes than standard methods (for experimental procedures and detailed results see [11,12]).

For example, Figure 2 shows three views of the same data set produced by conventional projection pursuit, PCA, MDS and TPP respectively. This dataset comprises cDNA microarray analysis of small, round blue cell childhood tumors (SRBCT), including neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt Lymphoma (BL) and members of Ewings family of tumors (EWS) [17]. The view produced by the TPP method shows a clear separation between all classes, compared to the others. For example, the view produced using MDS shows one aspect of the 'curse of dimensionality': the small variance between points in high dimensional space results in a view with very little difference between intra-class and inter-class point distances. In other words there is little 'bunching' or clear separation between classes.

Also note that the view of the data produced by human-driven project pursuit was better (in the sense of separating sample classes) than that produced by a conventional algorithmic projection pursuit. In other words, a human was more effective at searching the extremely large space of all possible projections than an algorithm; a result which reinforces the value of the 'division of labour' between human user and machine mentioned above.

One advantage of using linear projections, such as TPP, for data visualisation is that they not only show an informative view of the data, but the weights of the projection itself include useful information For example, the projection that TPP found to generate the view in Figure 2 can also be used to determine which combinations of genes are the best predictors of the cancer classes shown; a process known as *feature selection* [12]. The performance of TPP as a feature selector has been tested empirically on a range of gene expression data sets, and the performance is found to be comparable with standard methods. Indeed, TPP is particularly effective in those cases where there are relatively few samples compared to the number of genes whose expression level is measured – a common situation in experiments involving human clinical cases.

The third task is to use TPP to spot outliers or cases of misdiagnosis in the data. This was tested by artificially changing the cancer class of some of the samples in the gene expression data sets used above. The resulting visualizations were then presented to users and they were asked to spot which samples had been altered. This process was compared with standard misclassification detection algorithms, and again the performance of TPP was comparable – another illustration of the power of human pattern recognition.

Conclusion

The human visual system has evolved to be superb at spotting patterns, and is still more reliable and powerful than artificial computer vision systems in applications ranging from face recognition to surveillance. Therefore it makes sense to exploit it when exploring data. The problem is how to present complex data in such a way as to let the experimenter eyeball it effectively, and to let their perception of the data guide their interaction with it. TPP achieves this by regarding the data visualisation not as a fixed object to be presented to the viewer, but as a construct to be actively and interactively explored. The resulting interface is both engaging and also appears to be statistically and scientifically useful.

Author Bio

oe Faith has a BSc in Mathematics, MSc in Computational Neuroscience, and PhD in Philosophy. Although now researching in Computer Science at Northumbria University (Newcastle, UK), he has always worked closely with artists and was a jurist on the Vida Artificial Life Arts prize, sponsored by Telefonica. His current research is in the visual analysis of complex data; trying to find ways in which science practitioners can not just visualize, but comprehend and explore high dimensional data spaces. His intuition is that by concentrating on the analytic benefits of visualisation, the aesthetic aspects will follow.

References and Notes

1. John Wilder Tukey, *Exploratory Data Analysis*. (Addison-Wesley, 1977). Tukey was a pioneering computer scientist who worked with Von Neumann and was also responsible for coining the terms 'bit' and 'software'.

2. F.J. Anscombe, "Graphs in Statistical Analysis", *American Statistician*, 27 (February 1973), *pp*17-21.

3. The situation in biology may be contrasted to experimental physics, where the data may be larger in volume but less complex in nature, and where the scientists have greater mathematical background.

4. T.R.Golub *et al* "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286(5439) (1999) *pp*531-7.

5. *A Survey of Dimension Reduction Techniques* (US DOE Office of Scientific and Technical Information, 2002), <www.llnl.gov/CASC/sapphire/pubs/148494.pdf>

6. J.B.Kruskal "Toward a Practical Method which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation which Optimizes a New 'Index of condensation'." *Statistical Computing*, 427-440 (1969), though the term was originated by J.H.Friedman and J.W.Tukey "A Projection Pursuit Algorithm for Exploratory Data Analysis." *IEEE Transactions on Computers*, C-23, *pp*881-890 (1974)

7. D.Asimov,"The Grand Tour: A Tool for Viewing Multidimensional Data". SIAM Journal of Scientific and Statistical Computing 6(1), *pp*128-31 (1985).

8. This interface is available in the GGobi data visualisation system. See Dianne Cook and Deborah F. Swayne, *Interactive and Dynamic Graphics for Data Analysis*, (Springer 2007) and available at < www.ggobi.org>

9. G.Johansson,"Visual perception of biological motion and a model for its analysis", *Perception & Psychophysics*, 14, *pp*201-211 (1973). Also see <www.psy.vanderbilt.edu/faculty/blake/BM/BioMot.html>

10. The prototype tool, along with sample data sets, is available from <computing.unn.ac.uk/staff/cgjf1/TPP.zip>

11. J.Faith and R.Mintram and M.Angelova, "Targeted Projection Pursuit for Gene Expression Data Classification and Visualisation", *Bioinformatics*, 22(21):2667 (2006).

12. J.Faith and M.Brockway, "Targeted Projection Pursuit Tool for Gene Expression Visualisation", *Journal of Integrative Bioinformatics*, 3(2):43 (2007).

13. I.Guyon and A.Elisseeff,"An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research* 3 (2003) *pp*1157-1182.

14. J.Faith and A.Enshaie," Data Exploration using Targeted Projection Pursuit", *In Preparation*.