Projection Pursuit

A time-honored method for detecting unanticipated 'structure' – clusters, outliers, skewness, concentration near a line or a curve – in bivariate data is to look at a scatterplot, using the ability of the human perceptional system for instantaneous pattern discovery. The question is how to bring this human ability to bear if the data are high-dimensional.

Scanning all 45 pairwise scatterplots of a 10-dimensional data set already tests the limits of most observers' patience and attention span, and it is easy to construct examples where there is obvious structure in the data that will not be revealed in any of those plots. This fact is illustrated in Figures 1.

Figure 1 shows a two-dimensional data set consisting of two clearly separated clusters. We added eight independent standard Gaussian 'noise' variables and then rotated the resulting 10-dimensional data set into a random orientation. Visual inspection of all 45 pairwise scatterplots of the resulting 10-dimensional data fails to reveal the clusters; the scatterplot which, subjectively, appears to be most structured is shown in Figure 2.

However, we know that there do exist planes for which the projection is clustered; the question is how to find one.

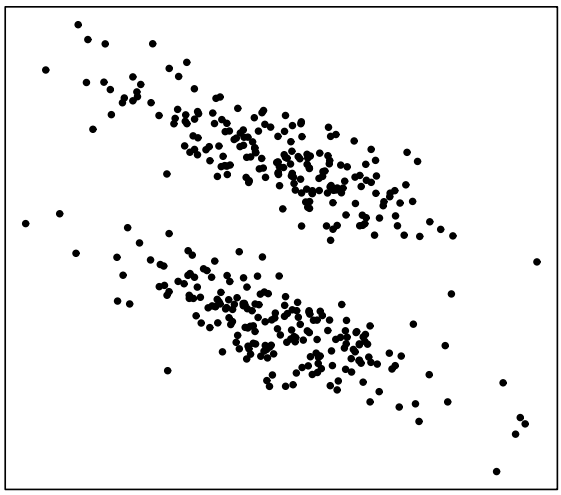

Figure 1 Sample from a bivariate mixture of two Gaussians

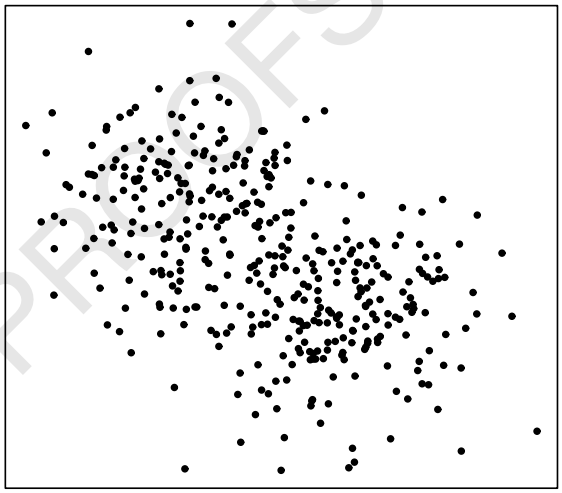

Figure 2 Most 'structured' pairwise scatterplot

Looking for Interesting Projections

The basic idea of projection Pursuit, suggested by Kruskal [15] and first implemented by Friedman and Tukey [10], is to define a *projection index* $I(\mathbf{u}, \mathbf{v})$ measuring the degree of 'interestingness' of the projection onto the plane spanned by the (orthogonal) vectors **u** and **v** and then use numerical optimization to find a plane maximizing the index.

A key issue is the choice of the projection index. Probably the most familiar projection index is the variance of the projected data. A plane maximizing this index can be found by linear algebra – it is spanned by the two largest principal components (*see* **Principal Component Analysis**). In our example, however, projection onto the largest principal components (Figure 3) does not show any clustering – variance is not necessarily a good measure of 'interestingness'.

Instead, a better approach is to first sphere the data (transform it to have zero mean and unit covariance) and then use an index measuring the deviation of the projected data from a standard Gaussian distribution. This choice is motivated by two observations. First, if the data are multivariate Gaussian (*see* **Bivariate Heritability**), then all projections will be Gaussian and projection pursuit will not find any interesting projections. This is good, because a multivariate Gaussian distribution is completely

FIRST PAGE PROOFS

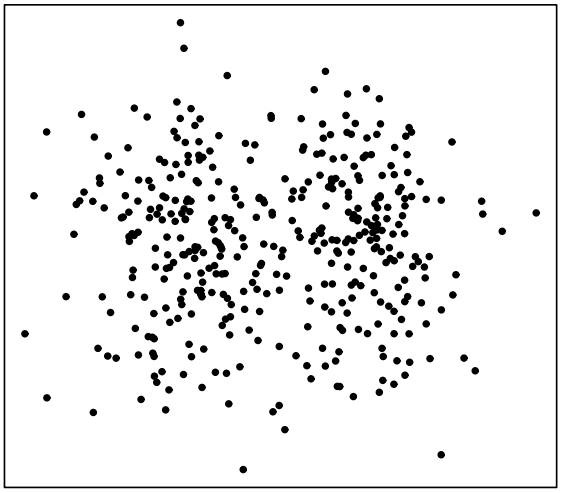

Figure 3 Projection onto largest principal components

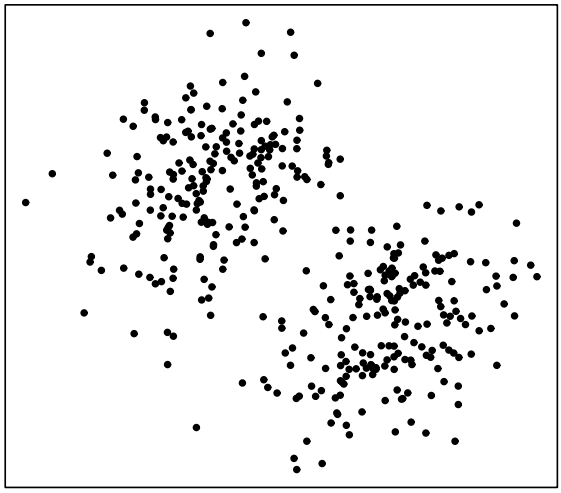

Figure 4 Projection onto plane maximizing the 'holes' index

specified by its mean and covariance matrix, and there is nothing more to be found. Second, Diaconis and Freedman [3] have shown that under appropriate conditions most projections of multivariate data are (approximately) Gaussian, which suggests regarding non-Gaussian projections as interesting.

Many projection indices measuring deviation from Gaussianity have been devised; see, for example [2,11–14]. Figure 4 shows projection of our simulated data onto a plane maximizing the 'holes' index [1]; the clusters are readily apparent.

Example: The Swiss Banknote Data

The Swiss Banknote data set [4] consists of measurements of six variables (width of bank note; height on left side; height on right side; lower margin; upper margin; diagonal of inner box) on 100 genuine and 100 forged Swiss bank notes. Figure 5 shows a projection of the data onto the first two principal components. The genuine bank notes, labeled '+', are clearly separated from the false ones.

Applying projection pursuit (with a Hermite index of order 7) results in the projection shown in Q1 •Figure 6 (adapted from [14]).

This picture (computed without use of the class labels) suggests that there are two distinct groups of forged notes, a fact that was not apparent from Figure 5.

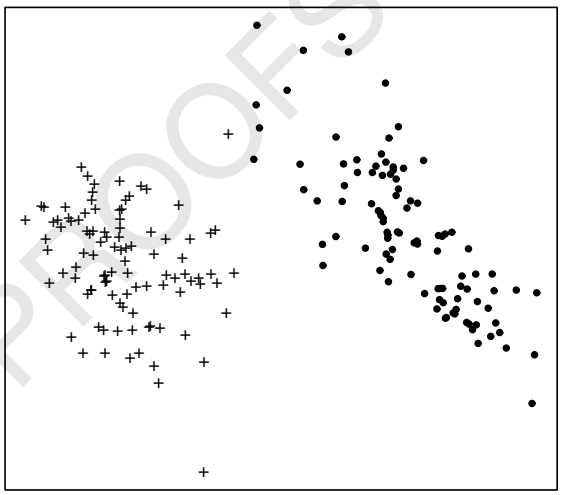

Figure 5 Projection of Swiss Banknote data onto largest principal components

Projection Pursuit Modeling

In general there may be may be multiple interesting views of the data, possibly corresponding to multiple local maxima of the projection index. This suggests using multiple starting values for the nonlinear optimization, such as planes in random orientation (*see* **Optimization Methods**). A more principled bsa460 approach is to remove the structure revealed in consecutive solution projections, thereby deflating the

FIRST PAGE PROOFS

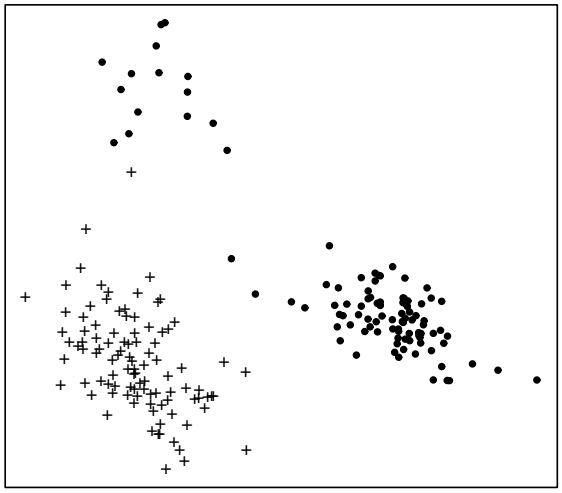

Figure 6 Projection of Swiss Banknote data onto plane maximizing the 'Hermite 7' index

corresponding local maxima of the index. In the case where a solution projection shows multiple clusters, structure can be removed by partitioning the data set and recursively applying projection pursuit to the individual clusters. The idea of alternating between projection pursuit and structure removal was developed into a general *projection pursuit paradigm for multivariate analysis* by Friedman and Stuetzle [9]. The projection pursuit paradigm has been applied to density estimation [6, 8, 12, 13], regression [7], and classification [5].

Software

Projection Pursuit is one of the many tools for visualizing and analyzing multivariate data that together make up the *Ggobi Data Visualization System*. Ggobi is distributed under an AT&T open source license. A self-installing Windows binary or Linux/Unix versions as well as accompanying documentation can be downloaded from `www.ggobi.org`.

References

[1] Cook, D., Buja, A. & Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions, *Journal of Computational and Graphical Statistics* **2**(3), 225–250.

[2] Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics* **4**(3), 155–172.

[3] Diaconis, P. & Freedman, D. (1984). Asymptotics of graphical projection pursuit, *Annals of Statistics* **12**, 793–815.

[4] Flury, B. & Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces, *Journal of the American Statistical Association* **76**, 757–765.

[5] Friedman, J.H. (1985). Classification and multiple regression through projection pursuit, Technical Report LCS-12, Department of Statistics, Stanford University.

[6] Friedman, J.H. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**, 249–266.

[7] Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.

[8] Friedman, J.H., Stuetzle, W. & Schroeder, A. (1984). Projection pursuit density estimation, *Journal of the American Statistical Association* **79**, 599–608.

[9] Friedman, J.H. & Stuetzle, W. (1982). Projection pursuit methods for data analysis, in *Modern Data Analysis*, R.L., Launer & A.F., Siegel, eds, Academic Press, New York, pp. 123–147.

[10] Friedman, J.H. & Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* **C-23**, 881–890.

[11] Hall, P. (1989). Polynomial projection pursuit, *Annals of Statistics* **17**, 589–605.

[12] Huber, P.J. (1985). Projection pursuit, *Annals of Statistics* **13**, 435–525.

[13] Jones, M.C. & Sibson, R. (1987). What is projection pursuit, (with discussion), *Journal of the Royal Statistical Society, Series A* **150**, 1–36.

[14] Klinke, S. (1995). Exploratory projection pursuit-the multivariate and discrete case, in W. Kloesgen, P. Nanopoulos & A. Unwin, eds, *Proceedings of NTTS '95*, Bonn, 247–262.

[15] Kruskal, J.B. (1969). Towards a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new "index of condensation", in *Statistical Computation*, R.C., Milton & J.A., Nelder, eds, Academic Press, New York, pp. 427–440.

(*See also* **Hierarchical Clustering**; ***k*-means Algorithm**; **Minimum Spanning Tree**; **Multidimensional Scaling**) bsa103 bsa104 bsa391 bsa415

WERNER STUETZLE

FIRST PAGE PROOFS

bsa506

4 Projection Pursuit

Abstract: Projection pursuit was conceived as a method for finding 'interesting' one- or two-dimensional projections of multivariate data revealing unanticipated structure such as clusters, skewness, or presence of outliers. The initial idea of exploring high-dimensional data using low-dimensional marginals was later extended into a general paradigm for modeling multivariate data that has been applied to density estimation, regression, and classification.

Keywords: Visualization; cluster analysis; density estimation; principal component analysis

Q2 • **Author Contact Address:** University of Washington, Seattle, WA, USA

FIRST PAGE PROOFS

bsa506

QUERIES TO BE ANSWERED BY AUTHOR (SEE MARGINAL MARKS Q..)

IMPORTANT NOTE: You may answer these queries by email. If you prefer, you may print out the PDF, and mark your corrections and answers directly on the proof at the relevant place. Do NOT mark your corrections on this query sheet. Please see the proofing instructions for information about how to return your corrections and query answers.

Q1. Please confirm if you have obtained permission to use this figure. If yes, please provide us with a copy of the grant letter.
Q2. Please check your Affiliation Details and update accordingly. Please ignore this query and accept our apologies if you have already sent the updated affiliation.

FIRST PAGE PROOFS