Interface Foundation of America

Grand Tour and Projection Pursuit

Author(s): Dianne Cook, Andreas Buja, Javier Cabrera and Catherine Hurley

Source: *Journal of Computational and Graphical Statistics*, Vol. 4, No. 3 (Sep., 1995), pp. 155-172
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: https://www.jstor.org/stable/1390844
Accessed: 23-10-2018 16:31 UTC

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/1390844?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Institute of Mathematical Statistics, American Statistical Association, Interface Foundation of America, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Computational and Graphical Statistics*

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

Grand Tour and Projection Pursuit

Dianne COOK, Andreas BUJA, Javier CABRERA, and Catherine HURLEY

The grand tour and projection pursuit are two methods for exploring multivariate data. We show how to combine them into a dynamic graphical tool for exploratory data analysis, called a projection pursuit guided tour. This tool assists in clustering data when clusters are oddly shaped and in finding general low-dimensional structure in high-dimensional, and in particular, sparse data. An example shows that the method, which is projection-based, can be quite powerful in situations that may cause grief for methods based on kernel smoothing. The projection pursuit guided tour is also useful for comparing and developing projection pursuit indexes and illustrating some types of asymptotic results.

Key Words: Data visualization; Interactive dynamic graphics; Data projections; Exploratory data analysis.

1. INTRODUCTION

In this article we show that two graphical methods for exploring high (say p) dimensional data—the *grand tour* (Asimov 1985; Buja and Asimov 1986; Buja, Asimov, Hurley, and McDonald 1988), a dynamic tool, and *projection pursuit* (Kruskal 1969; Friedman and Tukey 1974; Huber 1985), a static tool—naturally complement each other and can be combined to enhance each's performance in detecting low-dimensional structure. A grand tour attempts to provide the viewer with an overview of a multivariate point scatter by presenting a continuous (dynamic) sequence of low (d, usually $= 1, 2, 3$) dimensional projections, which, within time constraints, are representative of all possible projections of the data. In contrast, projection pursuit seeks out only low-dimensional projections that expose interesting features of the high-dimensional point cloud. It does this by optimizing a criterion function, called the projection pursuit index, over all possible d-dimensional (d-D) projections of p-dimensional (p-D) data. Projection pursuit results in a number of static plots of projections that are deemed interesting, in contrast to the dynamic movie of arbitrary projections that is provided by a grand tour. Unfortu-

Dianne Cook is Assistant Professor, Department of Statistics, 323 Snedecor Hall, Iowa State University, Ames, IA 50011-1210, dicook@iastate.edu. Andreas Buja is Member of the Technical Staff, AT&T Bell Labs, Room 2C-261, 600 Mountain Avenue, PO Box 636, Murray Hill, NJ 07974-0636. Javier Cabrera is Associate Professor, Department of Statistics, Hill Center, Busch Campus, Rutgers University, New Brunswick, NJ 08903. Catherine Hurley is with the Department of Mathematics, St. Patrick's College, Maynooth, Ireland.

©1995 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 4, Number 3, Pages 155–172

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

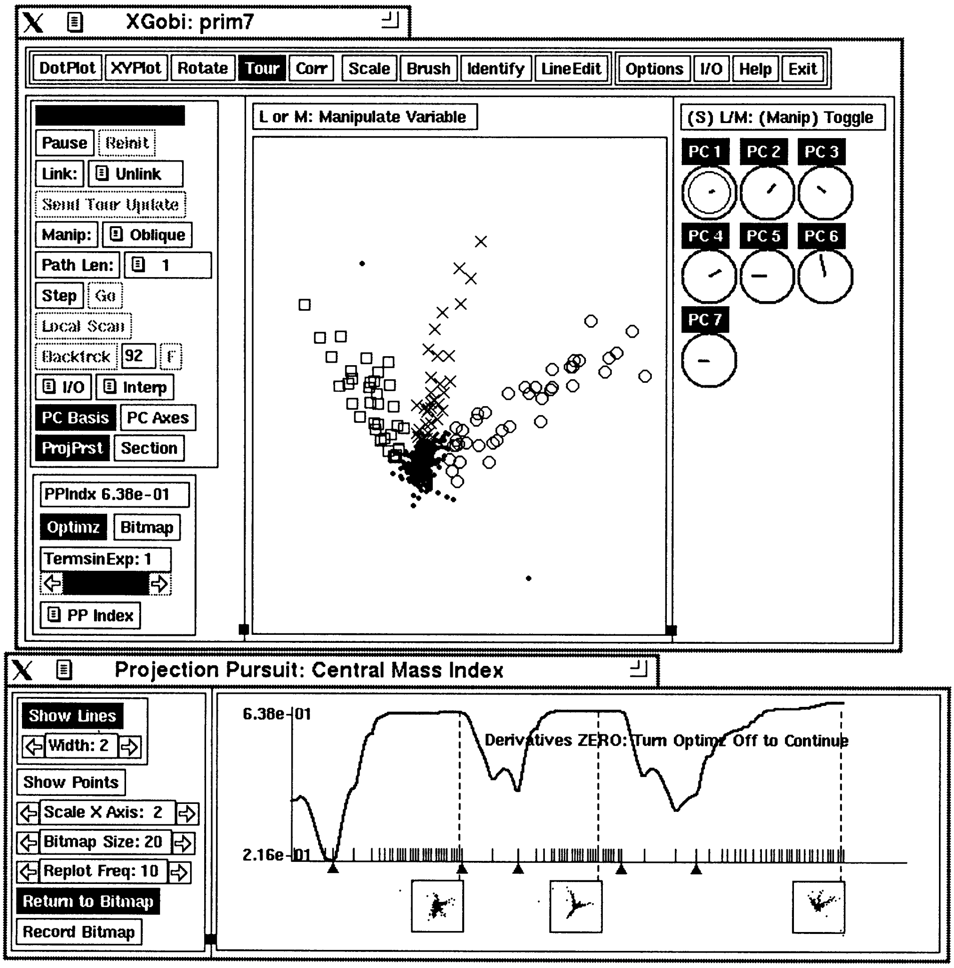

Figure 1. Implementation of Projection Pursuit Guided Tour in XGobi.

nately, static plots suffer from a lack of context because they have been removed from their neighborhood in the projection space, and although a grand tour provides the neighborhood context it has a tendency to spend too much time away from, or indeed never visit, the interesting projections. The two methods combined in an interactive, dynamic framework provide powerful tools for exploring high-dimensional data using projections. In particular, when the data is sparse in relation to its dimensionality, methods based on projections have advantages over those based on kernel smoothing. The work discussed in this article fills gaps in research on exploring high-dimensional data.

In the last decade most projection pursuit indexes (for example, Jones and Sibson 1987; Friedman 1987; Hall 1989; Morton 1989; Cook, Buja, and Cabrera 1993a; Posse 1994) have been anchored on the premise that to find the structured projections one should search for the most nonnormal projections. Good arguments for this can be found in Huber (1985) and Diaconis and Freedman (1984). (We should point out that searching for the most nonnormal directions is also discussed by Andrews, Gnanadesikan, and Warner

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

[1971] in the context of transformations to enhance normality of multivariate data.) This clarity of purpose makes it relatively simple to construct indexes that "measure" how distant a density estimate of the projected data is from a standard normal density. (Note that the data is usually sphered before beginning projection pursuit to remove mean and variance effects from the search, and in this sense the comparison with a *standard* normal density is justified.) The projection pursuit index, a function of all possible projections of the data, invariably has many "hills and valleys" and "knife-edge ridges" because of the varying shape of underlying density estimates from one projection to the next. To accommodate the optimization of such a function Friedman (1987) proposed a projection pursuit algorithm that entails an initial rough global search for relatively high values of the function from which to, secondly, start derivative-based searches to find the global maximum.

In the last few years, with the assistance of powerful desktop computing hardware, research on the grand tour has concentrated on user interaction. The tools for user interaction, suggested to date, take the form of motion alteration and restriction, such as a facility to retrace the tour path and restriction of movement to subspaces, such as, principal component, canonical correlation, or discriminant coordinate space (Hurley and Buja 1990). We now add to this bag of tricks projection pursuit guidance. The grand tour is used to move the viewing plane arbitrarily through the projection space, which acts to provide random starting points for derivative-based optimization of the projection pursuit index. The actual time point at which the optimization is initiated may be determined by the viewer, or in an automated implementation by some predetermined initiation mechanism. In our implementation we have concentrated on the former to provide a highly interactive user controlled interface.

Figure 1 shows a window dump of the implementation of a projection pursuit guided tour in XGobi (Swayne, Cook, and Buja 1991), which is a software system that is publicly available from StatLib. [To get started using StatLib, send the one-line e-mail message `send index` to `statlib@lib.stat.cmu.edu`. A program will read your request and send further instructions. StatLib can also be accessed by FTP, Gopher, and WWW. The e-mail reply from StatLib will contain instructions for the other methods of access.] XGobi is designed for analysis of high-dimensional data through manipulation of scatterplots. It offers such plotting techniques as textured dotplots (Tukey and Tukey 1990), pairwise plots, and 3-D rotation as well as the tour, and includes interactive operations on the data such as scaling, linked brushing, and identification of points. It is written in C and uses the X Window System (trademark of MIT). Although it is possible to construct a projection pursuit guided tour for any projection dimension, the implementation in XGobi only uses 2-D projections, which is natural for 2-D display devices.

To give some familiarity with the graphical appearance of the XGobi guided tour see Figure 1. Two windows are shown. The main window displays a paused grand tour (in principal component space) surrounded by many controls and the bottom window displays the projection pursuit index that has been plotted over time as the tour progressed. At the top of the main window is a line of mode buttons where it can be seen that tour mode is highlighted. Associated with the tour mode is the panel of controls to the left of the plot window that includes tools for interacting with a grand tour and controls for

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

the projection pursuit guidance. To the right of the plot window is a collection of circles and labels representing the variables of the data set.

The next section discusses implementing a projection pursuit guided tour, using the example of XGobi, and the tools that we have found naturally assist user interaction. The third section gives examples of both exploring data and viewing functions with the projection pursuit guided tour.

2. IMPLEMENTATION

2.1 Basic Ideas: Optimization Adaptation of Tour Movement

The *grand tour* is defined as a continuous one-parameter (time, usually) family of d-D projection planes that is dense in the set of all d-D planes in p-space ($d < p$). The space of all unoriented d-D planes through the origin in Euclidean p-space is called a *Grassman manifold*, which we denote as $G_{d,p}$. In contemplating an implementation of a grand tour this definition lends itself to a variety of interpretations. One approach depends on the construction of a filling curve that systematically traverses $G_{d,p}$. (See Asimov [1985] for a discussion of some attempts at constructing good deterministic paths, which is, as yet, an unresolved problem.) Alternatively, a random sampling of $G_{d,p}$ combined with the construction of a continuous path between pairs of sampled planes can be used.

The second approach is the simplest and most easily adaptable grand tour construction. It is the method that we concentrate on and we call it an interpolation tour. The construction procedure is described in detail in Buja, Asimov, and Hurley (1989), but in simple terms there are two basic steps that are iterated. Initialization is from a predetermined *starting plane*, $\mathcal{V}_{(0)}$:

1. Sample, randomly, for a d-D plane in p-space, which we call the *target plane*, $\mathcal{V}_{(1)}$. (To do this generate d vectors in $\mathbb{R}^p$ by orthonormalizing d p-D standard normal vectors, for example. This results in a random orthonormal basis, denoted $\boldsymbol{u}_{(1)}$, for a random plane.)
2. Interpolate from the starting plane, $\mathcal{V}_{(0)}$, to the target plane, $\mathcal{V}_{(1)}$, set this to be the new starting plane, and return to Step 1. (The interpolation is implemented in discrete steps that appear continuous to the eye, and the size of the steps can be adjusted to simulate apparent speed changes. We call the starting planes and target planes *basis planes*. Knowing the basis plane sequence allows the tour path to be reconstructed. The orthonormal basis for $\mathcal{V}_{(0)}$ is denoted as $\boldsymbol{u}_{(0)}$.)

As indicated earlier (p. 156), however, this type of grand tour may not provide the user with a view of any interesting projections—a problem that becomes worse as p increases. The objective is to use the derivatives of the projection pursuit index to select the new target plane in a more judicious manner; this adaptation of Step 1 generates the projection pursuit guided tour which we now explain in more detail. Let $\boldsymbol{z}$ be a p-D random vector, with $\mathbf{0}$ mean, and identity covariance matrix, $\boldsymbol{x}= (x_1, \ldots, x_d) = \boldsymbol{u}'\boldsymbol{z}$, where $\boldsymbol{u}$ is an orthonormal basis for an arbitrary d-plane in p-space, and let $I(\boldsymbol{x})$ be a d-dimensional projection pursuit index. (I is a function of the projected data matrix and the domain is all possible projections. For our purposes we have restricted ourselves to continuously differentiable functions, but it is possible to relax this condition if appropriate optimiza-

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

tion methods are used.) Using this notation, the target plane $\mathcal{V}_{(1)}$, characterized by the orthonormal basis $\boldsymbol{u}_{(1)}$, is chosen as the result of orthonormalization of

$$\boldsymbol{u}_{(0)} + k \cdot \frac{\partial I(\boldsymbol{x})}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{u}}\bigg|_{\boldsymbol{u}_{(0)}},$$

where k is the step size parameter of the optimization. In terms of dynamic graphics, k is a path length parameter because it determines the distance to the next target plane. We consider the maximum of the index I to be reached when its value no longer increases by further movement in the derivative direction—that is, in practical terms, the difference between the index values of the previous interpolation step and the current is below a tolerance value.

This is exactly steepest ascent optimization with respect to each component vector of $\boldsymbol{u}$. (It is also possible to use conjugate gradient methods by a simple adaptation of the definition of the target plane, and, of course, other methods by more radical adaptations.) At some time point the local maximum will be reached, which means that the tour must stop because the target plane is nearly identical to the starting plane. To continue motion when this happens we propose to revert the target plane selection procedure back to random sampling for some period of time before engaging in optimization again. The effect is analogous to performing steepest ascent optimization from multiple random starting points. The difference, of course, is that here the entire optimization procedure is visualized, and the viewer may determine the starting points for the optimization by using visual cues. We call the real-time process of periodically switching the target plane selection between random sampling and derivative-based selection, a *projection pursuit guided tour*.

Intrinsic to an interactive and dynamic implementation of a projection pursuit guided tour are a number of tools that are discussed in the next few sections. Recall that a global picture of the controls of the projection pursuit guided tour in XGobi is shown in Figure 1 (p. 156).

2.2 MONITORING WINDOW

A vital accompaniment of the projection pursuit guided tour is a monitoring window (Fig. 2). This window keeps a running plot of the projection pursuit index values for the sequence of projections displayed in the main tour window of Figure 1 over time. This involves storing a vector of index values and, in our implementation, the vector has a fixed length that depends on the size of the monitoring window. During on-screen motion, as the vector becomes filled, old values are replaced by new ones, and thus a shifting window of the most recent index values is maintained. The plot also rescales itself vertically when a new index value is above or below previous maximum and minimum values because it is assumed that global extreme values are not known a priori.

Along the horizontal axis (time) are a number of "landmarks," short vertical lines above the axis and triangles below the axis. The short vertical lines indicate when a new target plane is chosen. The triangles indicate the time point when optimization is either turned on or off. During optimization the index values increase with time. In the

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

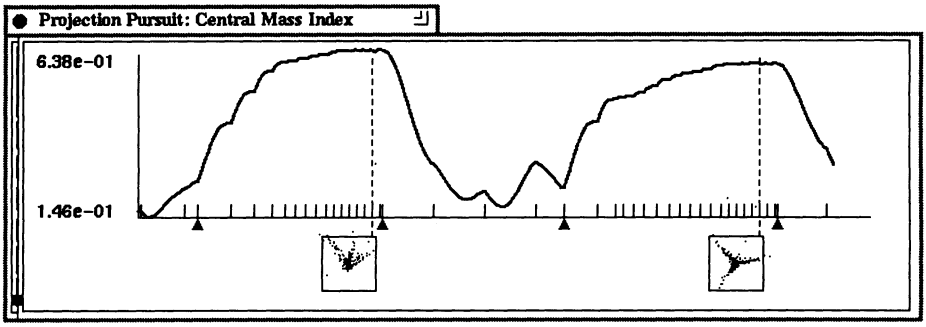

Figure 2. Monitoring Window for Projection Pursuit Guided Tour in XGobi.

figure optimization was turned on at the leftmost triangle, so the index value increases until the second triangle when optimization was turned off. It was turned on again at the third triangle and off at the fourth. (From the plot, it may appear that using a projection pursuit guided tour to search for interesting projections of high-dimensional data is a "heat up/cool down" process, such as simulated annealing, for finding maxima of an index. However, Figure 2 is a record of a real-time user-controlled procedure, and simulated annealing is an example of an automated procedure which is a possible alternative when real-time computations are not feasible.)

Marking the time of the two local maximum index values are two *bitmaps*. These are copies of the projection displayed in the tour window at the time the local maximum index values were reached, as indicated by the stabilizing of the index value. Their presence assists in mental reconstruction of the tour path by recording important features. In XGobi a bitmap can be generated at any time during a projection pursuit guided tour by a simple button click, but we have found it to be most useful to record local maxima.

2.3 BITMAP INTERFACE

There are two important additional uses of the bitmaps. The first is to direct the tour to return to the particular view provided by a bitmap accessed through a mouse click on the bitmap of interest. (In fact, this facility was incorporated after observing that people using the projection pursuit guided tour exhibited a natural tendency to want to return the tour to the previous bitmap views.) This behavior, though, depends on the bitmap remaining visible in the monitor window, which it will only do for the length of time represented by the width of the window. There is no scroll facility to retrieve invisible bitmaps. The second use is to "stack up" views that have been found in order to "replay" them later. This approach depends on the existence of a history mechanism in the tour. In XGobi this is provided by a backtrack feature in which a running linked-list of basis planes provides a mechanism for retracing the path of a tour. In addition, a prerecorded set of basis planes may be read in to describe a particular path to be traveled. This facility can be combined with a recorded list of basis planes that represent the bitmaps, or local maxima of the projection pursuit index.

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

2.4 NAVIGATIONAL TOOLS

When a structured projection is found it is important to understand the relationship between the constituent variables. With 3-D data the contribution of variables to a projection is often represented by a tripodal axis. This readily extends to higher dimensions in which a p-podal axis tree illustrates the linear combination of variables contributing to a projection. The disadvantage, however, is that it suffers from clutter as more variables are added. The solution provided by Buja et al. (1988) and Hurley and Buja (1990) is to take each axis stem out of the p-podal representation and embed it in its own icon, specifically a reference unit circle. We call these the variable circles and the radial bar represents the relative contribution of each variable to the displayed projection. These are the primary navigational tools. In Figure 1 (p. 156) they can be seen to the right of the main plot window. (They also serve a utility function in XGobi in that clicking on a variable circle adds or removes the variable from the tour.)

2.5 INDEX CHOICES: MENU, PARAMETER ADJUSTMENT

One of the most powerful features of dynamic graphics is the ability to quickly "twiddle" parameters and make option selections. The menu of indexes in XGobi includes the 2-D natural Hermite (Cook et al. 1993a), Hermite (Hall 1989), Legendre (Friedman 1987), Friedman–Tukey style (Friedman and Tukey 1974), and entropy (Jones and Sibson 1987) indexes, as well as three simple template-like indexes (Cook et al. 1993a) designed to detect projections with "holes" in the center (holes index) or concentration of mass in the center (central mass index), or skewness (skewness index). For complete information on the different indexes the reader is encouraged to consult the appropriate references.

2.6 IMPACT OF SPHERING

It is usual that the data is sphered before beginning projection pursuit to remove the influence of location and scale on the search for structured projections. This is especially necessary for indexes that "measure" the departure of the projected data density from a standard normal density because location and scale differences may dominate the other structural differences. However, sphering has an unfortunate side effect. It visibly changes the data. For example, consider points uniformly distributed on a cylinder that has a small length to radius ratio, as in points painted on a short piece of tube (Figure 3a). Sphering is analogous to increasing the length of the tube (Figure 3b), resulting in the hole being less visible. Hence, sphering is graphically distracting because it changes the shape of the data and may in some cases hide features that were previously visible.

Nevertheless, sphering is essential to the effectiveness of the current selection of projection pursuit indexes in XGobi, so the data is sphered before beginning a projection pursuit guided tour. In displaying the procedure, however, one can choose to use the sphered or unsphered data space. Our preference is to show the projection pursuit guided tour on the sphered data, although in XGobi it is possible to also display the corresponding unsphered data projections using the linked tour facility (see Sec. 2.8). (The projection

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

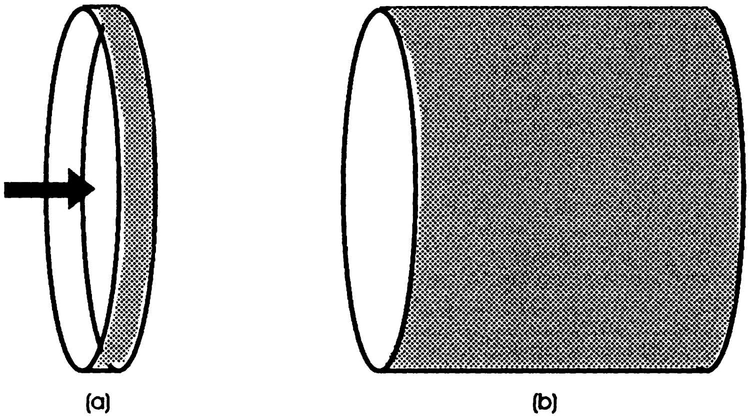

Figure 3. Visual Effect of Sphering Data. (a) before sphering the hole is easy to see; (b) after sphering the hole not as easy to see.

coordinates, u, from the sphered space are "back-transformed" to the corresponding coordinates in the unsphered data space.)

2.7 Inclusion of User-Defined Index Functions

The implementation in XGobi is set up to make it feasible for users to include their own index functions with a minimal knowledge of C and X. Essentially two functions need to be provided—one for calculating the index value at a particular projection and another for calculating derivatives. The interested reader should read the distributional notes of XGobi for further information (see `statlib` note on p. 157).

2.8 Linked Tours

One solution to the problem of sphering is to show both projections—the projection of the sphered data and the equivalent projection in unsphered data space. This is facilitated by linking two XGobis. The XGobi running a projection pursuit guided tour sends its projection to one showing the data in unsphered data space. (The linking function inverts the projection coordinates appropriately.)

The linked tour facility can be used to compare different projection pursuit indexes and for cross-validation of data, for example, checking if one interesting projection proves interesting for both halves of a data set.

3. EXAMPLES

3.1 Finding Low-Dimensional Structure in Data

The particle physics data set that we use to illustrate the projection pursuit guided tour was initially used to introduce projection pursuit by Friedman and Tukey (1974). The data

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

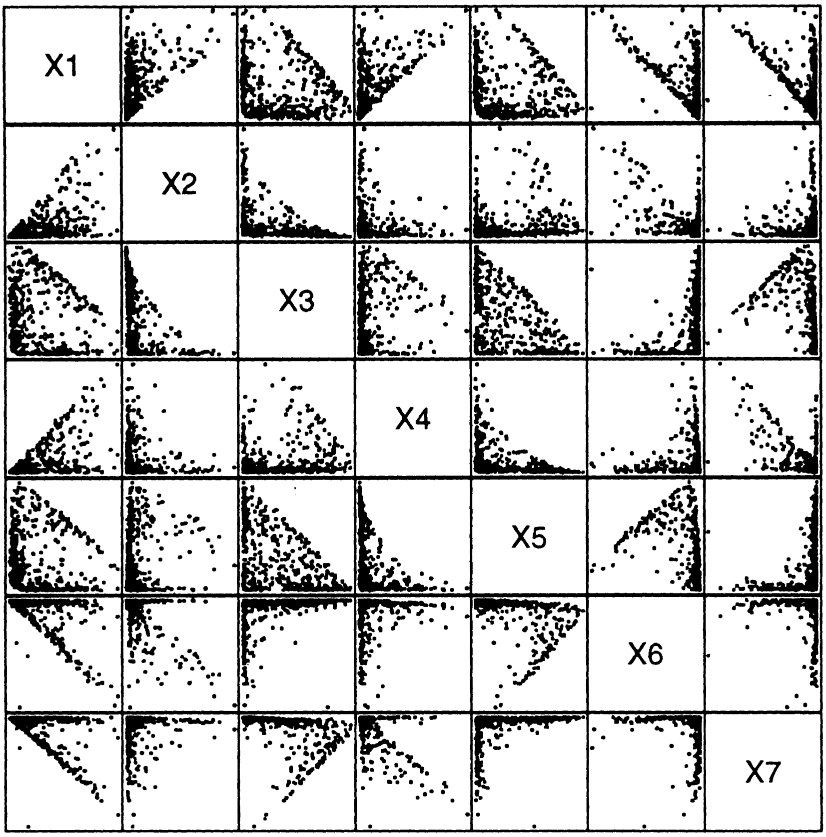

Figure 4. Pairwise Plots of 7-D Particle Physics Data.

is old, and the reaction recorded by the data is not interesting to contemporary physicists, but it is important to statisticians for the reason that the inherent structure has never been completely described. The combination of the grand tour and projection pursuit contributes significantly to revealing the nature of the variable relationships in seven dimensions. Recently, Koschat and Swayne (1992, in press) have used the projection pursuit guided tour in XGobi to explore telecommunications data, and indeed found previously undetected structure.

3.1.1 7-D Particle Physics Data

The 7-D particle physics data (often called "prim7 data") contains 500 observations taken from a high-energy particle physics scattering experiment that yields four particles. The reaction can be described completely by seven independent measurements. For this reaction, $\pi_b^+ p_t \longrightarrow p\pi_1^+\pi_2^+\pi^-$, the following measurables (squared invariant mass) were used: $X_1 = \mu^2(\pi^-, \pi_1^+, \pi_2^+), X_2 = \mu^2(\pi^-, \pi_1^+), X_3 = \mu^2(p, \pi^-), X_4 = \mu^2(\pi^-, \pi_2^+), X_5 = \mu^2(p, \pi_1^+), X_6 = \mu^2(p, \pi_1^+, -p_t), X_7 = \mu^2(p, \pi_2^+, -p_t)$. Here, $\mu^2(A, B, \pm C) = (E_A + E_B \pm E_C)^2 - (P_A + P_B \pm P_C)^2$, and $\mu^2(A, \pm B) = (E_A \pm E_B)^2 - (P_A \pm P_B)^2$, where E and P represent the particle's energy and momentum, respectively,

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

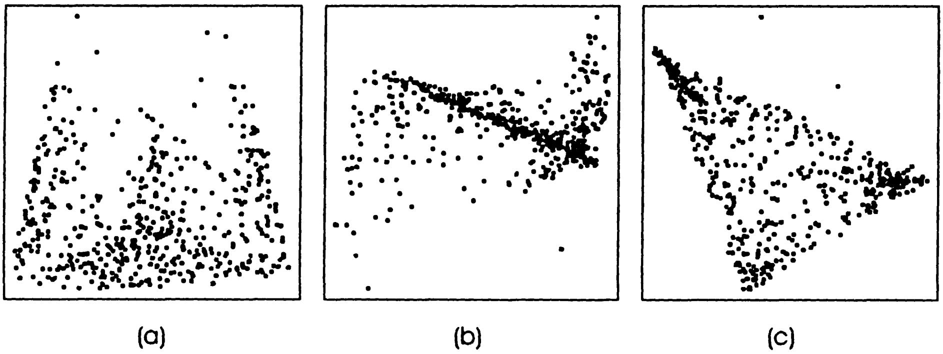

Figure 5. 7-D Particle Physics Data. (a) First two principal components; (b) projection similar to that found by Friedman and Tukey (1974); and (c) projection found by Jee (1985).

as measured in billions of electron volts. The notation $(p)^2$ represents the inner product P/P. The ordinal assignment of the two π^+'s was done randomly. The data is originally from Ballam et al. (1971), which contains a more complete description of the reaction. Important features of the data are short-lived, intermediate reaction stages which appear as clusters or clumpiness along low-dimensional linear subspaces ("arms").

Figure 4 shows the pairwise plots of the seven measurements. It is clear there are some linear relationships between the variables because of the clumpiness along the coordinate axes and diagonals. There are also three aberrant points visible in the plot of X1 versus X6, X1 versus X7, and X3 versus X6.

Figure 5a shows a plot of the first two principal components. This view indicates the presence of structure, perhaps three clusters, but it is not lucid enough to distinguish between them. In their original projection pursuit-based analysis, Friedman and Tukey (1974) found a projection in which the points lie on a "Z" shape (similar to the projection in Fig. 5b). With a projection pursuit index based on Fisher information, Jee (1985) found a projection in which the points lie on a triangle, with heavier concentrations at the vertices (Fig. 5c). Although they are interesting, these three views do little to divulge the basic shape of the point cloud. Using the projection pursuit guided tour the data points appear to form a very simple pattern: a basic triangle with two linear, or wedge-shape, structures extending from each vertex. We relate the interactive procedure that led to this description in the next few paragraphs. Although the session is summarized by the plots in Figure 6a–d, we must emphasize that these plots only represent instantaneous snap-shots of projections obtained during the projection pursuit guided tour. In reality, of course, the user experiences a movie-like representation of the evolving projections along the tour path. (Video footage of the projection pursuit guided tour on the particle physics data is available in Cook et al. 1993b.)

Figure 6a is the projection corresponding to a local maximum of the holes index, showing the triangle with two wrapped arms. We painted the two arms as *crosses* and *rectangles*, and identify them as arms CS and OR, respectively. (Note that color may also be used to further enhance the identification of the arms.) The job of classifying points in the intersection is made easier by the on-screen motion, the sense of which cannot be adequately portrayed by these flat sheets of paper, as indicated in the previous paragraph.

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

In the on-screen environment, a 3-D sense accompanies the movement of these arms: the tips of the arms rock against each other as the maximum is approached. This view, as mentioned previously, is a local maximum and, interestingly, the projection given by the global maximum is not very informative! This is not altogether unexpected. Although the holes index is successful in detecting the arms it is theoretically maximized by points distributed on a unit circle. In the process of projection pursuit the optimal index value corresponds to the projection that best approximates this extremal distribution. The view given in Figure 6a does not approximate the extremal distribution very well so it is not surprising that there is another projection of this data that has a higher index value. The holes index is sensitive to a very specific type of structure, whereas the more omnibus-

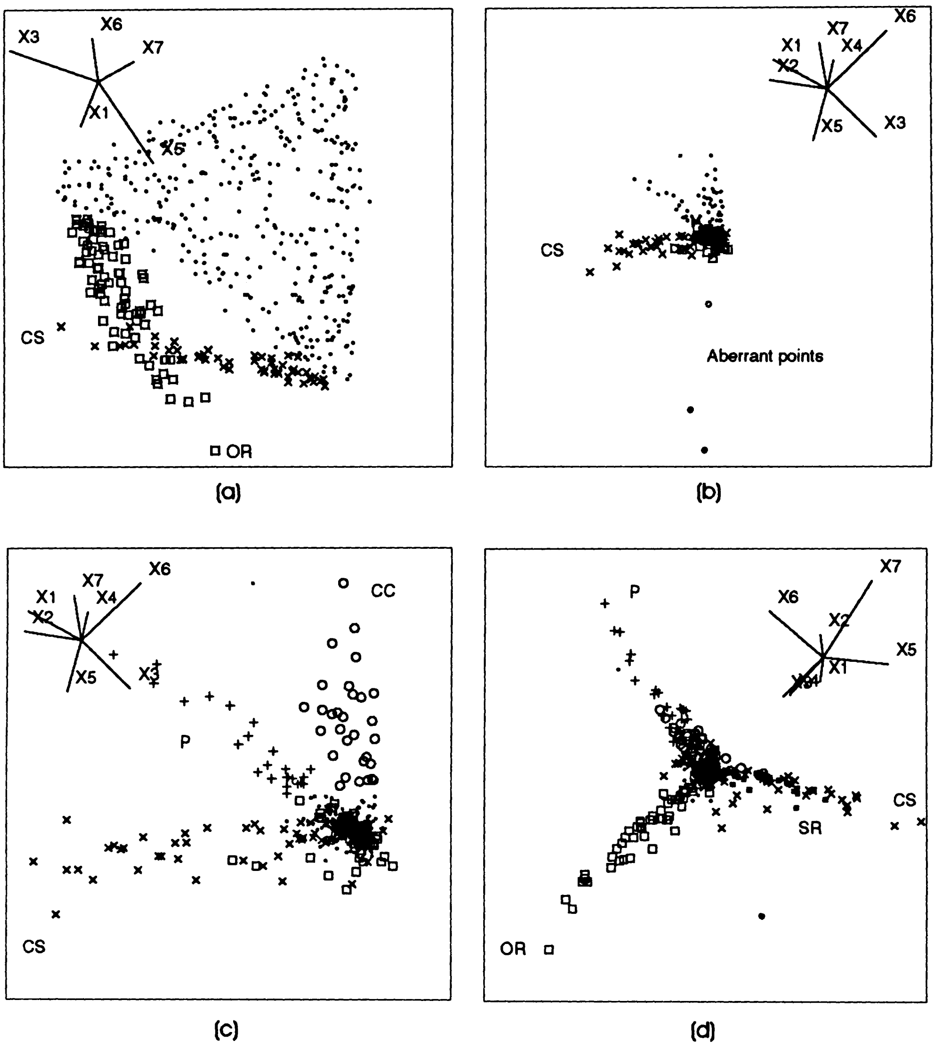

Figure 6. Analysis of 7-D Particle Physics Data. (a) local maximum of holes index; (b) global maximum of central mass index; (c) magnified view of (b); (d) local maximum of central mass index.

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

type indexes, such as those based on nonnormality measures, are sensitive to a much broader range of structure, and when using these indexes this situation will be more common.

Figure 6b is the projection given by the global maximum of the central mass index, and one can now see several new structures in the data—two more arms and three aberrant points. Figure 6c is the same projection magnified to focus more on the previously unseen arms, painted as circles (arm CC) and plusses (arm P). Figure 6d shows a projection corresponding to a local maximum of the central mass index. One more arm (small solid rectangles, arm SR) is visible, although difficult to see clearly in the view because the points also lie along arm OR. (In XGobi it is very easy to mask out the arm OR to brush points on underlying arm.) With further exploration, another arm (call it U for unbrushed at this point!) can be seen.

At this stage we can say there are six arms extending from the triangular region and arms CS and OR arise from separate vertices of the triangle. The relative location of

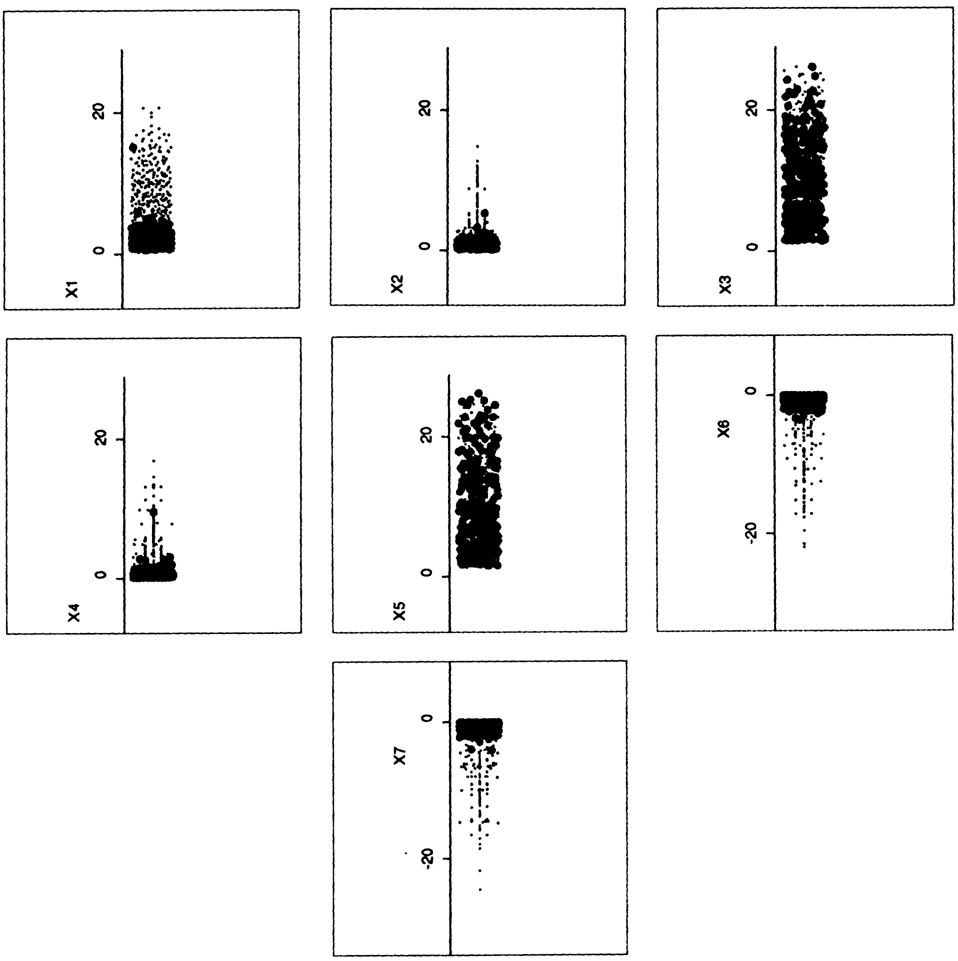

Figure 7. Textured Dotplot of Variables With Points in the Base Triangle Region Highlighted. The plots suggest that the triangular relationship is formed from variables 3 and 5

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

the others can be found by switching off projection pursuit guidance and watching the data touring, with the features identified, over an extended period of time. The motion provides a "Gestalt" sense of the proximity of points (and hence features). It is easy to see that arms CC and P extend together from the remaining vertex, and the short arm SR extends from the same vertex as the arm OR, and that U and CS extend from the third vertex.

Return to examining the plots in Figure 6. These indicate that each arm is approximately 1-D. Before making conclusions solely on these plots, remember that these are each 2-D projections of 7-D data, meaning there are 5 hidden "back" dimensions. Consider some facts about 2-D projections of solid 7-D geometric shapes: (1) a point (0-D object in $I\!R^7$) always projects to a point; (2) a line (1-D) projects as line or a point (0-D); (3) a plane (2-D) projects as a plane, line, or a point; and (4) a 3-D subspace projects as a plane, a line, or a point. These are solid shapes but serve the purpose of showing that the arms, as finite samples (including error) from the geometric shapes, may be higher than 1-D. (For more discussion of projections of geometric shapes see Furnas and Buja [1994].) Conclusions may be drawn if all possible projections are seen. Watching the data in a grand tour for an extended period of time is an approximation to all possible projections, and provides empirical information about each arm in the data. Each of the arms appears close to 0- or 1-D in most views shown by the grand tour suggesting to us that the relationship between the points in each arm is 1-D. The points in the triangle on the other hand always appear as approximately a triangle, a line, or a point. There are never more than three obvious vertices visible, which excludes higher dimensional shapes from consideration. So we conclude that these points do indeed lie close to a 2-D triangle in $I\!R^7$.

From a physicist's perspective, the next step is to relate the structure back to the original variables. As an example of the interpretation we concentrate just on the points in the base triangle, but note that points in the other regions can be examined in a similar manner. The points in the triangle are highlighted and examined in comparison to all the points in the univariate projections along the coordinate axes (Fig. 7). The triangle only has breadth in variables X3 and X5—that is, the squared invariant mass for a proton and a negative π-meson ($\mu^2(p, \pi^-)$), and the proton and a positive π-meson ($\mu^2(p, \pi_1^+)$), respectively. The interpretation is that these observations represent interactions between the particles p, π^-, π_1^+.

3.2 Viewing Functions

In this section we convey our experience using the projection pursuit guided tour for gaining intuition about functions defined on projections of p-space. An immediate use is in the comparison of different projection pursuit indexes. The second example that we show is an illustration of asymptotic results for 2-D projections, given in Diaconis and Freedman (1984).

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

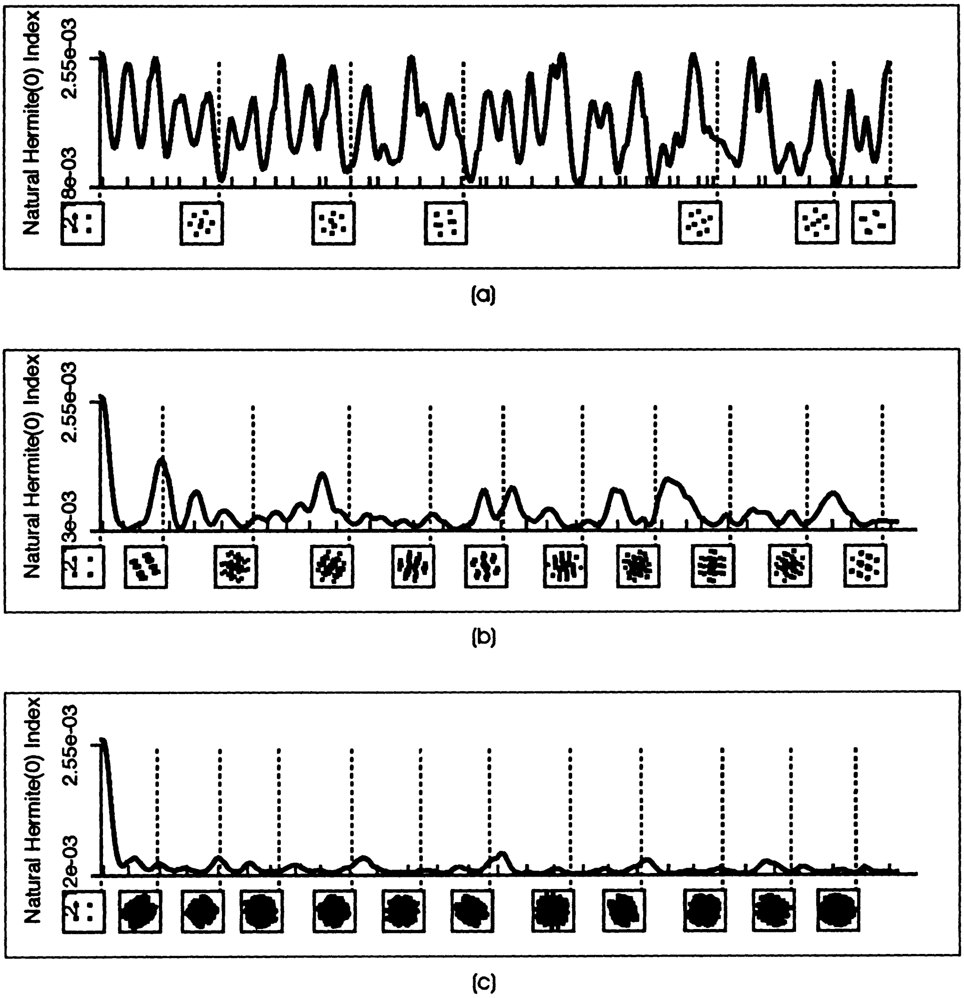

Figure 8. Illustration of Diaconis and Freedman's (1984) Result. Data generated by placing a point on each vertex of a 3-D (a), 5-D (b), and 9-D (c) cube.

3.2.1 Comparing Projection Pursuit Indexes

With the first implementation of projection pursuit into the dynamic framework of the grand tour we included simply the Legendre (Friedman 1987) and Hermite (Hall 1989) indexes. Hall's original motivation for proposing the Hermite index was based on an asymptotic argument that the Legendre index was shown to be overly susceptible to outliers. We did not observe this in practice, but rather we noticed that the Hermite index has a tendency to uncover projections of the data that have a "hole" in the center, which is quite a useful feature. The Legendre index also does this but to a lesser extent and seems more attracted to skewness. Differences such as these can be detected quickly by eye and used to direct further analytical work (Cook et al. 1993a).

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

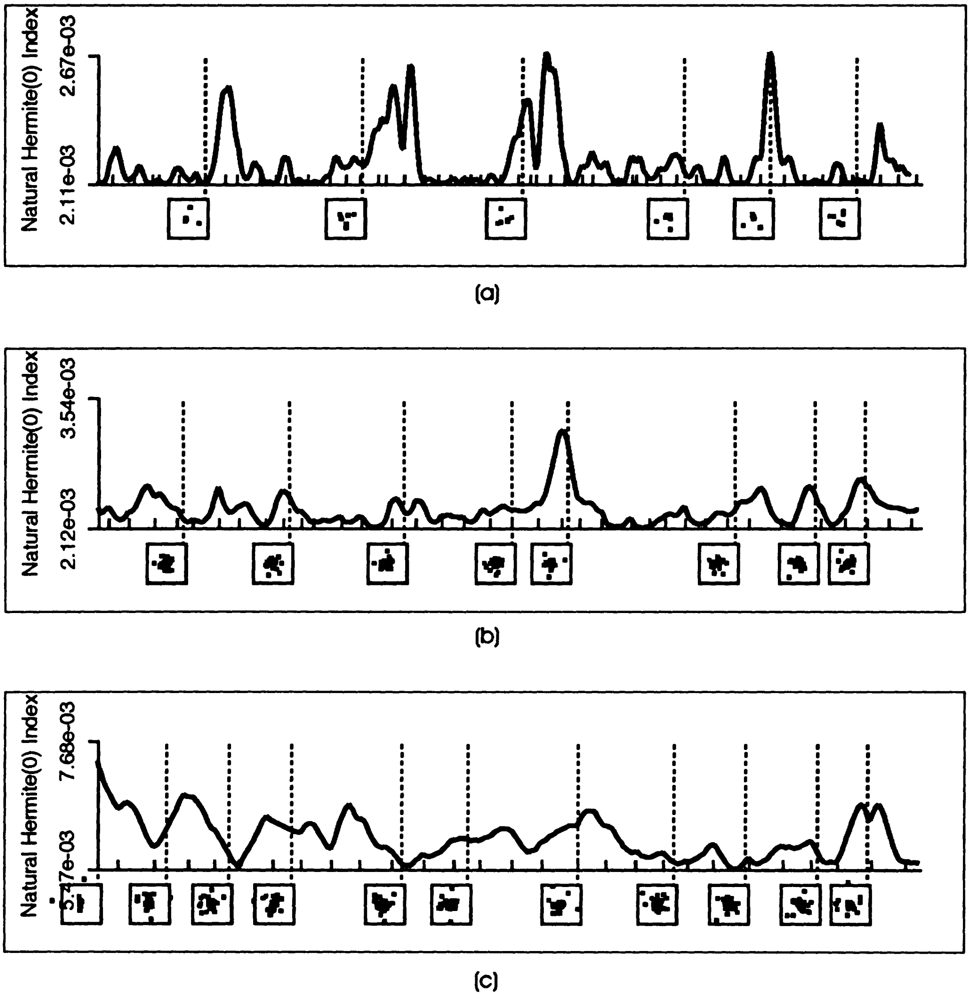

Figure 9. Projection Pursuit Guided Tour With Natural Hermite Index, Order 0, on a Sample From the Multivariate Cauchy Distribution. 8 points from 3-D (a), 32 points from 5-D (b), and 512 points from 9-D (c).

3.2.2 Illustrative Intuition of Fundamental Concepts

In analyzing multivariate data fundamental to the use of projections are theories as to the nature of projections from high dimensions down to low dimensions. For projection pursuit one fundamental underpinning is that *for many high-dimensional data sets most low-dimensional projections look approximately Gaussian (*)*. So to find the revealing, unusual projections one should search for the least Gaussian-looking projections. This is the premise on which many projection pursuit indexes have been based (see Sec. 1). We argue that this should not be the only premise on which indexes should be based and follow with an example (Fig. 9) to illustrate this. Nevertheless the premise is a good starting point and worth illustrating graphically as well as numerically.

Diaconis and Freedman (1984) formalized the basis on which the premise (*) is reasonable. We show an example that illustrates (*) on a sequence of data that conforms

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

to Diaconis and Freedman's constraints. A multivariate data set is constructed by placing a point on each vertex of a cube. Three such data sets are created: one 3-D, one 5-D, and one 9-D (n grows at the rate 2^p). Each data set is viewed in a tour: a segment displaying the sequence of index values is shown in Figure 8 (3-D cube (a); 5-D cube (b); 9-D cube (c)). The plotted index is the natural Hermite (0) index which is theoretically minimized by a Gaussian density. When the dimension is 3 almost every projection (a sample of these is shown in the bitmaps below the index plot) is revealing, but when the dimension is 9 almost every projection is not revealing in the sense of being close to Gaussian; the index plot is much flatter and close to the minimum value that would be obtained for a similar sample from a Gaussian distribution. As an aside it is interesting to note that visually the data set is clearly not Gaussian because it is far too regular, the points always lie in gridded, angular patterns. Nevertheless the most revealing projections are the ones that expose the method of construction which, in this case, are the projections along the marginal axes showing points on the vertices of a square (= 2-D cube). And projection pursuit using an index minimized by a Gaussian density serves the purpose of finding these revealing projections, among an increasing proportion of near-Gaussian views as p increases.

An example where one of Diaconis and Freedman's (1984) restrictions (the vectors' length being proportional to p) is violated can be found by taking samples from a multivariate Cauchy distribution. Figure 9 shows segments of a tour displaying the natural Hermite (0) index on a sample of size 8 from a 3-D Cauchy, 32 from a 5-D Cauchy, and 512 from a 9-D Cauchy. In this case there is no flattening out of the index function as p increases. Projection pursuit with an index sensitive to nonnormality does not assist in determining the nature of this multivariate data set.

4. DISCUSSION

In this article we have introduced exploring high-dimensional data using the projection pursuit guided tour. The work is motivated by the desire to understand high-dimensional relationships in data and builds on graphical methods that have been developed in recent years. We have used XGobi as a development platform for the new tools. Although developing code in C is more cumbersome than using S (Becker, Chambers, and Wilks 1988) or LispStat (Tierney 1991), for example, the computational efficiency allows more flexibility for implementing computationally intensive methods such as those that we have examined. In Section 3, we have liberally used many of the other tools available in XGobi, thus illustrating the symbiotic nature of these tools for exploring data.

The implementation in XGobi uses exclusively 2-D projection pursuit indexes. These are desirable for finding fully 2-D relationships, for example a 2-D spiral amid noise directions. Extensions to 1-D and 3-D indexes and grand tours would prove useful for finding structures of these dimensions. We have restricted ourselves to smooth, differentiable projection pursuit indexes, but many others exist that are not smooth although they seem useful. For example, the fractal index (Cabrera and Cook 1992) shows particular promise in detecting structure lying on low-dimensional, nonlinear manifolds. The

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

simple-minded use of derivative-based optimization precludes the inclusion of such an interesting index, because derivatives of the fractal index are not available. Some excellent work to improve this situation has been done by Posse (1993) who proposed an efficient optimization algorithm for 2-D projection pursuit indexes, based on the algorithm for 1-D indexes given by Huber (1990), which does not require derivatives. In his article is also a very promising index based on the chi-squared distance of the observed bivariate data density and the expected bivariate normal density. This index requires derivative-free optimization also. Each of these considerations would greatly enhance the current implementation.

ACKNOWLEDGMENTS

Thanks are extended to colleagues at Bellcore, Rutgers University, and the Physics Department, Iowa State University for many helpful suggestions and discussions. This work was done while Dianne Cook was with Bellcore and Rutgers University and Andreas Buja was with Bellcore.

[Received January 1994. Revised April 1995.]

REFERENCES

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971), "Transformations of Multivariate Data," *Biometrics*, 27, 825–840.

Asimov, D. (1985), The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal on Scientific and Statistical Computing*, 6, 128–143.

Ballam, J., Chadwick, G. B., Guiragossiân, G. T., Johnson, W. B., Leith, D. W. G. S., and Moriyasu, K. (1971), "Van Hove Analysis of the Reaction $\pi^- p \longrightarrow \pi^- \pi^- \pi^+ p$ and $\pi^+ p \longrightarrow \pi^+ \pi^+ \pi^- p$ at 16 gev/c*," *Physical Review, D, Particles and Fields*, 4, 1946–1966.

Becker, R., Chambers, J., and Wilks, A. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Buja, A., and Asimov, D. (1986), "Grand Tour Methods: An Outline," *Computing Science and Statistics*, 17, 63–67.

Buja, A., Asimov, D., and Hurley, C. (1989), "Methods for Subspace Interpolation in Dynamic Graphics," Technical Memorandum, Morristown, NJ: Bellcore.

Buja, A., Asimov, D., Hurley, C., and McDonald, J. A. (1988), "Elements of a Viewing Pipeline for Data Analysis," in *Dynamic Graphics for Statistics*, W. S. Cleveland and M. E. McGill eds., Monterey, CA: Wadsworth, pp. 277–308.

Cabrera, J., and Cook, D. (1992), "Projection Pursuit Indices based on Fractal Dimension," *Computing Science and Statistics*, 24, 474–477.

Cook, D., Buja, A., and Cabrera, J. (1993a), "Projection Pursuit Indexes Based on Orthonormal Function Expansions," *Journal of Computational and Graphical Statistics*, 2, 225–250.

Cook, D., Buja, A., Cabrera, J., and Swayne, D. (1993b), "Grand Tour and Projection Pursuit," video available by request to dicook@iastate.edu.

Diaconis, P., and Freedman, D. (1984), "Asymptotics of Graphical Projection Pursuit," *The Annals of Statistics*, 12, 793–815.

Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J. H., and Tukey, J. W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computing C*, 23, 881–889.

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms

Furnas, G., and Buja, A. (1994), "Prosection Views: Dimensional Inference Through Sections and Projections," *Journal of Computational and Graphical Statistics*, 3, 323–385.

Hall, P. (1989), "Polynomial Projection Pursuit," *The Annals of Statistics*, 17, 589–605.

Huber, P. J. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435–525.

——— (1990), "Algorithms for Projection Pursuit," Technical Report PJH-90-3, Massachusetts Institute of Technology, Dept. of Mathematics.

Hurley, C., and Buja, A. (1990), "Analyzing High-Dimensional Data with Motion Graphics," *SIAM Journal on Scientific and Statistical Computing*, 11, 1193–1211.

Jee, J. R. (1985), "A Study of Projection Pursuit Methods," Technical Report TR 776-311-4-85, Rice University, dept. of Statistics.

Jones, M. C., and Sibson, R. (1987), "What is Projection Pursuit?" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 150, 1–36.

Koschat, M. A., and Swayne, D. F. (1992), "Visualizing Panel Data," video available by request to dfs@bellcore.com.

——— (in press), "Visualizing Panel Data," Journal of Business and Economic Statistics.

Kruskal, J. B. (1969), "Toward 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 and J. A. Nelder eds., New York: Academic Press, pp. 427–440.

Morton, S. C. (1989), "Interpretable Projection Pursuit," Technical Report 106, Laboratory for Computational Statistics, Stanford University.

Posse, C. (1993), "Projection Pursuit Exploratory Data Analysis," Technical Report 1993.2, Lausanne: Swiss Federal Institute of Technology.

——— (1995), "Tools for Two-dimensional Projection Pursuit," *Journal of Computational and Graphical Statistics*, 4, 83–100.

Swayne, D. F., Cook, D., and Buja, A. (1991), "XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S," in *ASA Proceedings of the Section on Statistical Graphics*, Alexandria, VA: American Statistical Association, pp. 1–8.

Tierney, L. (1991), *LispStat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, New York: John Wiley.

Tukey, J., and Tukey, P. (1990), "Strips Displaying Empirical Distributions: I. Textured Dot Strips," Technical Memorandum, Morristown, NJ: Bellcore.

This content downloaded from 149.157.61.199 on Tue, 23 Oct 2018 16:31:33 UTC
All use subject to https://about.jstor.org/terms