Coo-2310-7

CONF-7910136--1

IRIA

Institut de Recherche
d'Informatique
et d'Automatique

SECONDES JOURNÉES INTERNATIONALES ANALYSE DES DONNÉES ET INFORMATIQUE

SECOND INTERNATIONAL SYMPOSIUM ON DATA ANALYSIS AND INFORMATICS

ÉDITION PROVISOIRE

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

VERSAILLES

17 - 18 - 19 Octobre 1979

EB

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

APPROACHES TO ANALYSIS OF DATA THAT CONCENTRATE NEAR HIGHER-DIMENSIONAL MANIFOLDS

Jerome H. Friedman[1], *John W. Tukey*[2,3] and *Paul A. Tukey*[3]

Stanford Linear Accelerator Center, Stanford, California 94305[1]
Princeton University, Princeton, New Jersey 08540[2]
Bell Laboratories, Murray Hill, New Jersey 07974[3]

The need to explore structure in high-dimensional clouds of data-points which may concentrate near (possibly non-linear) manifolds of lower dimension has led to the current development of three new approaches. The first is a computer-graphic system (PRIM'79) which facilitates interactive viewing and manipulation of an ensemble of points. The other two are automatic procedures for separating a cloud into more manageable pieces. One of these (BIDEC) performs successive partitioning of the cloud using hyperplanes; the other (Cake Maker) explores expanding sequences of neighborhoods. Both procedures provide facilities for examining the resulting pieces and the relationships among them.

INTRODUCTION

This paper is about investigation in progress, rather than investigation completed. For this and other reasons, the present account is a sketch rather than the paper to be presented at the Symposium. It seemed an appropriate subject, however, since the ideas involved, and the class of problems that drew them forth, are relatively unfamiliar.

One of the main kinds of data provided by high-energy particle physics is the complete description of each of many instances of a single nuclear reaction, that reaction being specified by a list of the particles (usually 2) entering a collision and a list of the particles (often 3 to 7) produced by it.

Any such reaction is likely to proceed through several channels, each of which can be thought of as specifying a tree of transient resonances or other intermediate conditions. For our present purposes, it may be important to think of these channels in bundles which share a common tree structure (but where individual channels may involve different, but analogous, resonances, etc. at a given branch of the tree).

If the final list is of k particles, knowledge of their momenta, a total of $3k$ numbers, completely describes an individual event. Because of conservation of momentum, conservation of energy, and rotational symmetry around the direction of the incoming particle, $3+1+1 = 5$ relationships allow us to use only $3k-5$ coordinates. Thus 3-body final-state reactions require 4 coordinates, while 7-body final-state reactions would require 16.

Prepared in part in connection with research at Princeton University sponsored by the U.S. Department of Energy.

The coordinates most likely to simplify the appearance of the $(3k-5)$-dimensional point clouds corresponding to a collection of individual instances of a specific k-body final-state reaction involve the choice of a pair of particles, say A and B -- or, for $k>3$, two subsets of particles, similarly labelled -- and a sign. The resulting coordinates are of the form

$$(E_A-M_A \cdot M_A)+(E_B-M_B \cdot M_B) \quad \text{(an ``invariant mass'')}$$

or of the form

$$(E_A-M_A \cdot M_A)-(E_B-M_B \cdot M_B) \quad \text{(the ``square of a 4--momentum transfer'')}$$

where E_A and E_B are the energies of the A^{th} and B^{th} particle or particle subset, and M_A and M_B are the corresponding (three-dimensional) momenta.

Any particular bundle of channels can be made "flat", can have the point clouds corresponding to each of its channels concentrated near a flat lower-dimensional manifold, if we make a proper choice of such coordinates. For $k>3$, however, no single choice can be expected to flatten all bundles of channels at the same time. Thus the empirical study of data for k-body final-state reactions with $k>3$ should be facilitated by good techniques for recognizing and isolating parts of point clouds that are located *near* manifolds of intermediate dimension that may be *twisted* as well as flat.

We are wholly ignorant of the extent to which such techniques will be helpful in studying, empirically, other kinds of data in 4 or more dimensions. We have only had techniques of a few kinds, particularly: (1) those that assumed the point cloud consists of ellipsoidal blurs, (2) those, like multiple regression, that assumed the point cloud is close to a manifold easily described in terms of carriers (given functions of the coordinates), (3) those, like simple-structure factor analysis, that assume the point-cloud would be concentrated near a simple arrangement of hyperplanes. Without a way to look for other kinds of structure than such conventional instances, we are unlikely to know whether or not other kinds of structure are (a) never present, (b) occasional or (c) frequent. Thus, flexible methods of analyzing multiresponse data, which are guided by the particle physics picture, may -- or may not -- prove to be very useful in other areas of inquiry.

In summary, we are interested in exploring clouds of points in high dimensional space without any prior knowledge about their configuration, and we seek tools to help discover whether the points are concentrated near manifolds of lower dimension. We want to be able to deal with manifolds that are flat or curved, disjoint or intersecting, possibly of low absolute dimension, possibly of low "negative dimension" (that is, of dimension $n-k$ embedded in n dimensional space, where k is small).

In what follows, we sketch three approaches to the analysis of such data:

1) A computer-graphic approach to a more flexible display system for looking at and manipulating such data, called PRIM'79 (planning by Mary Ann Fisherkeller, Jerome H. Friedman, Werner Stuetzle, John W. Tukey; implementation by Mary Ann Fisherkeller, Werner Stuetzle, Mathis Thoma, and Jerome H. Friedman). This system is not expected to be fully operational before the Symposium, but some of the untried ideas seem worthy of description.

2) A binary-decomposition approach (developed by Jerome H. Friedman, with comments from others in the group). Here we hope to report on some initial experience.

3) An approach of locating, pinching, and assembling pieces of manifolds, conveniently referred to as "cake making" (developed by John W. Tukey with major inputs from Paul A. Tukey; implemented by the latter). Here we may or may not have relevant experience to report at the Symposium.

PRIM'79

This display system is a descendant of PRIM-9, developed and implemented at SLAC (Stanford Linear Accelerator Center) in 1972 (Tukey, Friedman and Fisherkeller 1976, Bin 88 Productions 1973) and, to a certain extent, of the PRIMS-ETH system developed and implemented at the ETH in Zürich by Werner Stuetzle and Mathis Thoma in 1978.

These systems emphasized rigid rotations of a data hypervolume, which we may as well think of as spherical, supplemented by an ability to use any two of the current coordinates as the visible axes. PRIM-9 also involved (a) an ability to mask with boundaries parallel to any current-coordinate coordinate hyperplane, and (b) an ability to convert (coordinate-fixed) masking into (data-fixed) isolation, possibly step by step.

Applied to 3-body final-state reactions, where the manifolds of concentration are one-dimensional and can be flattened simultaneously, PRIM-9 worked smoothly identifying "resonances", doing this less strongly when the resonances appeared as lines of concentration in the display, but more strongly when they appeared as spots of concentration. (The usefulness of this led to the development of an iterative batch program, "projection pursuit" (Friedman and Tukey 1974), designed to enhance the spottiness of a two-dimensional display.) Applied to 4-body final-state reactions, where we expect 2-dimensional manifolds of concentration, not all simultaneously flattenable, projection pursuit gave helpful but not definitive results, producing pictures obviously in need of further flattening (untwisting).

PRIM'79 differs from both ancestors in allowing certain non-rigid deformations of the unit ball into itself as well as rotations. As a consequence, undoing an arbitrary string of manipulations no longer reduces to applying the inverse of an appropriate non-singular matrix.

One principal instance of such a nonlinear deformation is a computationally simplified version of a twist around an axis seen as horizontal, in a three-space, two of whose three coordinates are visible, by an amount described by a Chebyshëv polynomial in horizontal location. This capability should be helpful in untwisting concentration manifolds of intermediate dimension.

An important non-rotation of a different sort is a facility for sharpening scatter-diagrams. This can be done in two ways, both dependent upon near-neighbor points. One can assess local densities, perhaps in terms of the (reciprocal of the) volume of the smallest sphere around a point that contains *k* points, and then delete those points whose local densities lie in a lowest fraction of such densities. Alternatively, one can move each point toward its local kCG, that is, toward the center of gravity of its *k* nearest neighbors. Both processes tend to clear away "field points" and sharpen up structures. Neither has a natural inverse, so that, if back-stepping is to be feasible, memory has to be provided for individual point coordinates and not just for parametrized deformations.

PRIM'79's basic isolation mechanism follows its Zürich ancestor in using a cursor to identify polygon corners for masking or isolation, rather than being restricted, as PRIM-9 was, to masking parallel to a current coordinate axis.

Wherever reasonable, PRIM'79's control will involve the "1,1,2,4,8,...and back" procedure used in PRIM-9 to drive rotations and masks. In particular, coordinate selection is to be done in this way, as is the *x*– and *y*–control of a cursor (which involves a four-step cycle of right, up, left and down).

Instead of a light-pen controlled menu, like that added to PRIM-9 at a later stage, the qualitative control of PRIM'79 will involve a recursive use of menus and submenus, with control exercised by driving a cursor and pressing a "yes" button.

It is hoped to link, through menus and submenus, the display manipulation with a collection of analytical programs (such as projection pursuit, for instance), running in a largish host computer.

BINARY DECOMPOSITION (BIDEC)

This approach for detecting and describing concentrations of points near lower-dimensional manifolds proceeds "top down". It aims to partition the multidimensional data space into disjoint regions that are as large as possible, subject to the constraint that, within each, the lower-dimensional manifold near which the points lie be as linear as possible. The principal components of the points that lie within the region are then used as a local description of the manifold. The ensemble of all such principal-components solutions represents a piecewise linear approximation to the manifold. The interrelationships among these regions must then be studied systematically to lead to an understanding of the overall manifold.

With this procedure, the construction of the regions is very important. If a region is too large, then the manifold curvature within it may be too great to allow a reasonable linear approximation. If a region is too small it may lie entirely within the manifold. In this case, the principal-components solution will reflect only the shape of the region and give little information about the manifold itself.

Even within these constraints, the specific placement and shape of each region must be chosen carefully. The size, shape and orientation of each one must be such that the manifold within it is as linear as possible. Because of the finite (usually small) sample size, each region should be as large as possible, subject to the above goals. Where the curvature of the manifold is high, there should be many regions with narrowest dimension normal to the direction(s) of maximum manifold curvature. Places where the manifold is nearly linear should be represented by single large regions.

The local regions are chosen using binary decomposition. Consider a set S_o of n multivariate observations. Based on these observations, choose a direction $\underset{\sim}{d}$ and a (scalar) split point s. The sample S_o is then divided into two subsamples, S_1 and S_2, such that:

$$\left.\begin{array}{ll} \underset{\sim}{x}_i \in S_1 & \text{if } \underset{\sim}{x}_i \cdot \underset{\sim}{d} < s \\ \underset{\sim}{x}_i \in S_2 & \text{otherwise} \end{array}\right. \qquad (i=1,\ldots,n)$$

this partitioning procedure is then reapplied to the subsamples S_1 and S_2 obtaining four subsamples; each of these is then similarly partitioned, and so on. The recursive application of this decomposition procedure is continued until all subsets finally meet a terminal condition. These terminal subsets represent mutually exclusive subsamples of the complete data sample and lie in mutually exclusive convex regions defined by the various split directions and split points. These terminal regions collectively cover the entire data space.

The size, shape and orientation of these terminal regions are determined by the specific prescription for choosing the split direction and split point at each stage of the binary decomposition, and the terminating condition that ends the splitting. These prescriptions are usually cast in the form of optimizing some figure of merit. The goal of the partitioning is to choose regions for applying local principal-components analysis. The common property to be shared by the data in each terminal region is that the manifold near which the points lie be as linear as possible within it.

The figure of merit for determining the relative goodness of a particular binary decomposition of a point sample is the degree to which the manifold in each of the two resulting subregions can be represented by a linear approximation. The partitioning direction $\underset{\sim}{d}$ and split point s that yield the maximum improvement in this property are the ones chosen.

The degree to which a particular set of principal-components represents a multivariate point set is difficult to determine in an absolute sense, but relative improvement can be estimated. This estimate is based on the fact that the determinant (product of the eigenvalues) of the covariance matrix of a point sample is an estimate of the square of the relative volume occupied by the sample in the p-dimensional space. This estimate is biased toward large values with the degree of bias being proportional to the curvature of the lower-dimensional manifold near which the points lie. If the manifold is nearly linear, then the bias is small.

The volume of a point sample can be estimated conveniently by a principal-components analysis over the entire sample, or by partitioning the sample and taking the sum of separate estimates in each of the two subsamples. If the sample tends to lie near a lower-dimensional manifold that is curved, the latter estimate will tend to be smaller. The smallest estimate will be obtained when the partitioning direction is parallel to the direction of maximum curvature of the manifold.

The present criterion for partitioning each subsample in the binary decomposition described above is to choose the direction and split point that minimize the sum of volumes of the two subsamples as estimated by the square roots of the determinants of their respective sample covariance matrices. It is not computationally feasible to perform the search over all possible directions and split-point values for this minimum. A reasonable subset of search directions is the set of principal axes of the sample, partitioning each at the projected median. After determining the best of these for a partitioning direction, a search can be made for the best split point by trial partitions at a fixed number of percentiles along the split direction.

A reasonable terminating condition is to stop the partitioning of a region when the best possible split of a subsample yields little or no improvement toward reducing the estimated volume. There are situations, however, when this criterion can terminate partitioning too soon. There can be data point configurations for which a single split yields little improvement, but subsequent partitionings result in great improvement. Lack of improvement is not a sufficient criterion for termination.

Partitioning must be terminated when the subsample size becomes too small for reliable estimation of the covariance matrix and its roots. Thus, a sufficient condition for termination is that the terminal region sample size be below some cutoff value b_o. Using this criterion as the sole one for termination can, however, cause oversplitting, which can cause terminal regions to lie totally within the manifold.

A good termination rule is to stop splitting a subsample at that point at which no further decomposition to any subsequent level yields improvement in the volume estimate, or when the sample size is below b_o. This can be accomplished by continued splitting until all subsets have sample sizes less than b_o, and then subsequent recombining of the subsets in inverse order of the splitting, checking at each stage for improvement in the volume estimate. Subregions are recombined in this bottom up manner until further recombination results in degradation (increase) of the volume estimate.

The result of applying the binary decomposition procedure is a collection of mutually exclusive convex subregions of the data space. The subsamples assocated with these regions can be used to estimate the local properties of the lower-dimensional manifold. Although knowledge of these properties is important, it is useful to combine them to form a global interpretation of the manifold as well.

An important ingredient to a global interpretation is the connectivity structure of the terminal regions. For example, it is important to know which regions border each other.

which are close and far from each other, and which regions lie on the same manifold. Some of this information can be obtained by studying the partitioning sequence (binary tree) that gave rise to the terminal region set. For example, "brother" terminal regions (regions produced by a single split of a common parent region) border each other, while "cousin" regions (those for which the common ancestor is once or only a few times removed) generally tend to be close. However, the information obtainable from the partitioning sequence is generally quite limited.

Another approach with which we have experimented defines a distance between all pairs of terminal regions and then applies standard clustering technology to the resulting interregion distance matrix. The distance between a pair of regions (i,j) is taken to be $d_{ij} = \log v_{ij} - \log v_i - \log v_j$, where v_i and v_j are the volumes of the individual regions, as estimated from the covariance matrices of their individual point samples, and v_{ij} is a similar volume estimate based on the covariance matrix of the pooled sample. This distance will tend to be large if the regions are physically separated or if they intersect at a large angle. It will be small if the regions border each other and intersect at a small angle.

We have so far applied single-linkage analysis to the interregion distance matrix with mixed results, and plan to apply more powerful multidimensional scaling techniques in the hope of being able to describe the global properties of detected structures.

CAKE MAKING

Another approch to the problem of detecting concentrations of points near lower-dimensional manifolds is to look from the inside out, in other words, to examine local pieces of the point cloud, discover their structure, then patch the pieces together, gradually gaining an understanding of the overall structure.

The process for examining pieces of the point cloud, which we shall call cake making, starts by focusing attention in the neighborhood of some point in a region of fairly high local density. Since we expect that noise (including both measurement variation and actual variation) in the individual coordinates of our data will spread out the manifold and give it a certain thickness in the full n dimensions, we will expect very small spherical neighborhoods to be fully occupied with points, and therefore uninformative. So we look at successive disjoint spherical shells containing M points each, all centered at the original point of interest. As the shells begin to protrude from the manifold, the configurations of points they contain will begin to look like hyper-rings.

By doing principal-components analysis on the points in each successive shell and looking for a large gap in each set of eigenvalues, we can assess the inherent dimensionality of each ring. As we move outward from shell to shell, we expect to see a decrease in dimensionality, then several successive shells with low dimensionality, and finally an increase as the shells extend beyond the neighborhood in which the manifold has simple and fairly smooth structure. The principal-components analysis of each shell also yields a decomposition of n-space into two sets of directions, one set parallel to the hyper-ring, and a complementary set perpendicular to it. As we move outward among the shells with minimum dimensionality, we also check the parallel spaces and do not allow them to tilt too far out of alignment with each other. In other words, we stop if we reach a radius at which the manifold starts to bend too sharply.

Having accepted some set of shells expanding outward from our point of interest, having determined the local dimensionality of the manifold, and having determined, as well, the directions that are locally parallel and perpendicular to it (by suitable averaging of the directions obtained from the principal-components analyses of the shells), we next seek to tidy up all the points within the accepted shells into a cake. This is done in two stages. First, we fit a quadratic hypersurface to the accepted points by regressing each of the perpendicular components on a set of quadratic functions of the parallel components. We then consider the sum of each vector residual (from these regressions) and the mean vector of the accepted points. We can think of the resulting point cloud as a flattened (that is, no longer curved) version of the cloud of accepted points. We process these adjusted points with a "pinching" operation which allows each point to migrate toward an adjacent region of higher local density, the motion being modulated by the local density. In this way, points close to the (now linear) manifold move closer still, and points farther away move less or not at all. The net effect is to pinch the cake into a more sharply defined form. An important consideration in doing this is to prevent local density concentrations from pulling points aside, thus leaving holes in the manifold, so it is important that the motion of each point during pinching be confined to the directions that are locally orthogonal to the manifold. After pinching is complete, the curved components that were modelled in the regression are added back in, so that the pinched points lie close to the actual curved manifold.

This completes the making of a cake centered at the original point of interest. We then shift attention to some adjoining region and repeat the whole process. When this has been done a sufficient number of times to exhaust most of the points, or at least most of the points in some larger region, then the individual cakes must be pieced together in an informative fashion. This problem is similar to that of piecing together the regions determined by the binary decomposition procedure described above. One way being explored is to use multidimensional scaling with the cakes as objects, but tailored to consider the orientations of the cakes as well as the locations of their centers of gravity. Thus, cakes with nearby centers of gravity but with very different orientations will not be treated as near one another in the scaling process. One way to do this is to use both distance between centers of gravity and distance between projection matrices.

REFERENCES

Bin 88 Productions (1973). PRIM-9, a 16mm color motion picture.

Friedman, J. H., Fisherkeller, M. A. and Tukey, J. W. (1974). PRIM-9: An interactive multidimensional data display and analysis system. A.E.C. Scientific Computer Information Exchange Meeting, May 2-3 (unpublished).

Tukey, J. W., Friedman, J. H. and Fisherkeller, M. A. (1976). PRIM-9, An Interactive Multidimensional Data Display and Analysis System. *Proceedings of the 4th International Congress for Stereology,* Sept. 4-9, 1975, Gaithersburg, Maryland.