

 ***Visualization & Data Mining for High Dimensional Datasets***

★ ★ ★ *KDIR 2012 Tutorial* ★ ★ ★

Alfred Inselberg¹
School of Mathematical Sciences
Tel Aviv University
Tel Aviv, Israel
aiisreal@math.tau.ac.il * www.math.tau.ac.il/~aiisreal



¹Senior Fellow San Diego SuperComputing Center 1996,
Distinguished Professor National University of Singapore 2011

Distinguished Professor Korea University 2008,

Abstract

A dataset with M items has 2^M subsets any one of which may be the one fulfilling our objectives. With a good data display and interactivity our fantastic pattern-recognition can not only cut great swaths searching through this combinatorial explosion, but also extract insights from the visual patterns. These are the core reasons for data visualization. With parallel coordinates (abbr. ||-cs) the search for relations in multivariate datasets is transformed into a 2-D pattern recognition problem. The foundations are developed interlaced with applications. Guidelines and strategies for knowledge discovery are illustrated on several real datasets (financial, process control, credit-score, intrusion-detection etc) one with hundreds of variables. A geometric classification algorithm is presented and applied to complex datasets. It has low computational complexity providing the classification rule explicitly and *visually*. The minimal set of variables required to state the rule (features) is found and ordered by their predictive value. Multivariate relations can be modeled as hypersurfaces and used for decision support. A model of a (real) country's economy reveals sensitivities, impact of constraints, trade-offs and economic sectors unknowingly competing for the same resources. An overview of the methodology provides foundational understanding; learning the patterns corresponding to various multivariate relations. These patterns are robust in the presence of errors and that is good news for the applications. We stand at the threshold of breaching the gridlock of multidimensional visualization.

The parallel coordinates methodology has been applied to collision avoidance and conflict resolution algorithms for air traffic control (3 USA patents), computer vision (1 USA patent), data mining (1 USA patent), optimization, decision support and elsewhere.

KEYWORDS: Exploratory Data Analysis, Classification for Data Mining, Multidimensional Visualization, Parallel Coordinates, Multidimensional/Multivariate Applications

Alfred Inselberg received a Ph.D. in Mathematics and Physics from the University of Illinois (Champaign-Urbana) and was Research Professor there until 1966. He held research positions at IBM, where he developed a Mathematical Model of Ear (TIME Nov. 74), concurrently having joint appointments at UCLA, USC and later at the Technion and Ben Gurion University. Since 1995 he is Professor at the School of Mathematical Sciences at Tel Aviv University. He was elected Senior Fellow at the San Diego Supercomputing Center in 1996, Distinguished Visiting Professor at Korea University in 2008 and Distinguished Visiting Professor at National University of Singapore in 2011. Alfred invented and developed the multi-dimensional system of Parallel Coordinates for which he received numerous awards and patents (on Air Traffic Control, Collision-Avoidance, Computer Vision, Data Mining). The textbook *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*”, Springer (October) 2009, has a full chapter on Data Mining and was acclaimed, among others, by *Stephen Hawking*.

Audience

The accurate visualization of multidimensional problems and multivariate data unlocks insights into the role of dimensionality. The tutorial is designed to provide such insights for people working on complex problems.

Half-Day Tutorial

Introductory Level

Lecture notes with annotated visuals and references will be provided

Outline

Introduction

- *Visualization – Insight from Images*: incorporating our tremendous **pattern recognition** into the problem solving process. Overview of information visualization methodologies and examples of visualization successes in data exploration.
- *Parallel Coordinates* : Motivation and definition, point \leftrightarrow line duality in 2-D. Mapping multivariate relations into distinct planar patterns. Comparison with other visualization methodologies. Interactive demos: visual patterns and duality reveal relations in multivariate GIS satellite data Fig. 1, financial data Fig. 2, intrusion detection data Fig. 3.

Linear and near-linear multivariate relations

- *Multidimensional Lines*: Their visualization, transformations, minimum distance between pairs of lines in N-dimensions, **collision avoidance in air traffic control (3 USA patents)** – demo of real air traffic conflict scenarios and their resolution.
- *Planes, Hyperplanes & Flats*: Visualization of hyperplanes and detection of coplanarity, **Industrial and commodity markets – buy/sell indicators, risk analysis**. Transformations, approximate coplanarity.

Data Mining – Multidimensional Detective

- *Visual data mining (exploratory data analysis - EDA) (USA Patent)*
 - The case for visualization.
 - Information-preserving data display, user interface, interactivity.
 - Query Design – atomic and complex queries formed with Boolean operations.
 - Navigating the discovery process from *visual cues* without prior hypothesis or biases.
 - Interactive exploration on several multivariate datasets (one with hundreds of variables)
- *Automatic Classification – a geometric classification algorithm*.
 - Finds explicit classification rule, displays it as a hypersurface see Fig. 4, and measures the rule's precision.
 - Finds the minimal subset of parameters (features) needed to describe the rule – this is not an approximation.

- Orders the features by their predictive value.
- Live illustration on real complex datasets e.g. identifying hostile vehicles from afar by their noise signatures.
- Comparison with 22 well-known classifiers using bench-mark datasets.
- A divide-and-conquer strategy partitioning the dataset into more manageable subsets e.g. identifying and classifying *different types of water mines*.

Non-linear multivariate relations – recognition from the visual patterns

- *Curves*
- *Surfaces & Hypersurfaces* represented by $(N - 1)$ planar regions visually revealing the hypersurfaces' properties. This representation is preferable for some applications even in 3-D Fig. 5-9.
- Classes of hypersurfaces – corresponding to classes of multivariate relations
 1. developables (see Fig. 7) ruled and more general surfaces.
 2. **Viewing convexity in any dimension.**
 3. **Non-convex, recognizing folds, crevices, bumps, dimples (see Fig. 8), swirls and non-orientable (see Fig. 9).**
 4. Transformations and their dualities – see Fig. 6
- Interior point construction and display algorithms – feasible states. Applications to **process control and decision support** Fig. 10. Demo constructing a *visual* model of a (real) country's economy and interactively finding sensitivities, criticalities, trade-offs and discovering that two economic sectors unknowingly compete for the same resources.

Concentrating Relational Information into Patterns

- We are at the threshold of breaking the gridlock of multidimensional visualization: inputting a massive dataset and outputting planar patterns – the “graphs” of multivariate relations in the data. An outline of a plan for achieving this together with other research problems will be presented at the end.

WEBSITE — www.math.tau.ac.il/aaisreal

EYE-CANDY

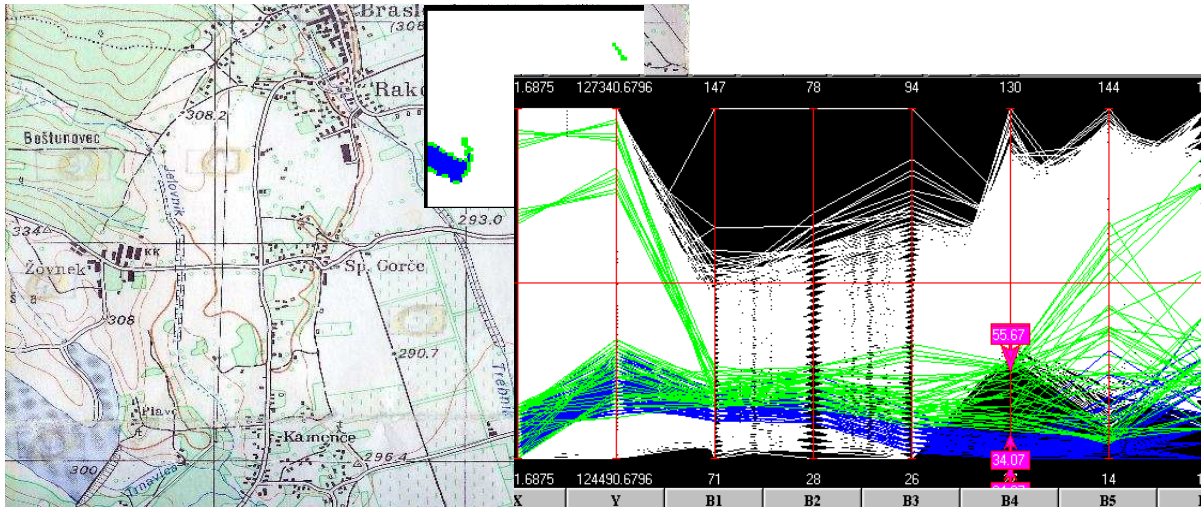


Figure 1: Ground emissions measured by satellite on a region of Slovenia (left) are displayed on the right. The water and lake's edge (middle) are discovered with two queries.

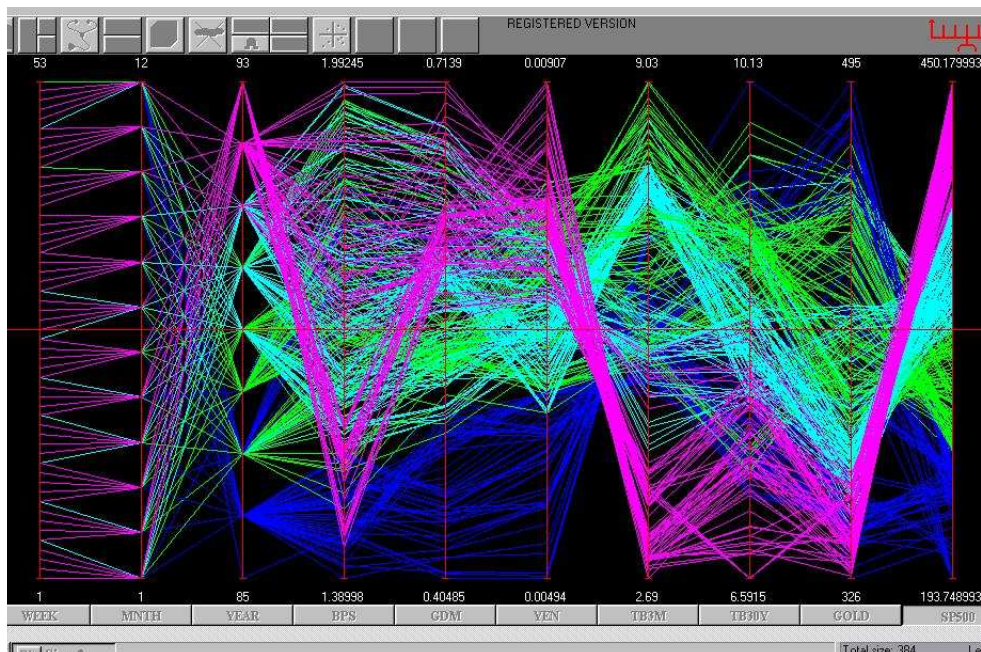


Figure 2: Multidimensional contouring query applied to a financial dataset. Quickly reveals interrelationships between the variable intervals. Note those for the highest SP500 (last axis) range and the other variables.

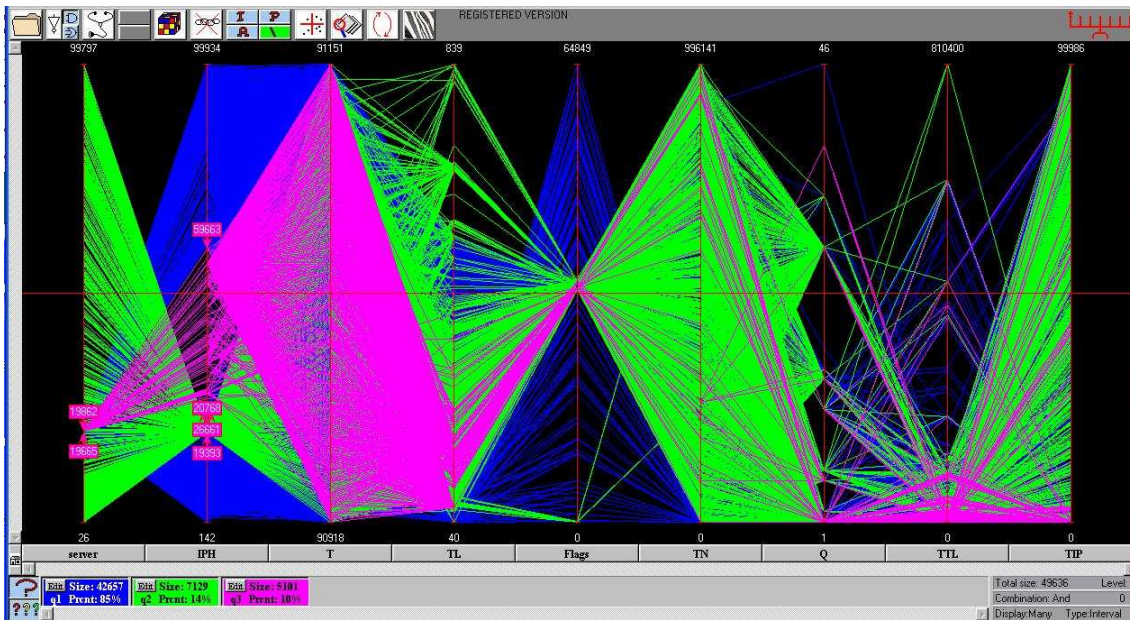


Figure 3: Detecting Network Intrusion from Internet Traffic Flow Data. Note the many-to-one relations. A server (marked on the leftmost axis) is “bombarding” many servers (shown in the 2nd axis) and there others examples – how many can you spot?

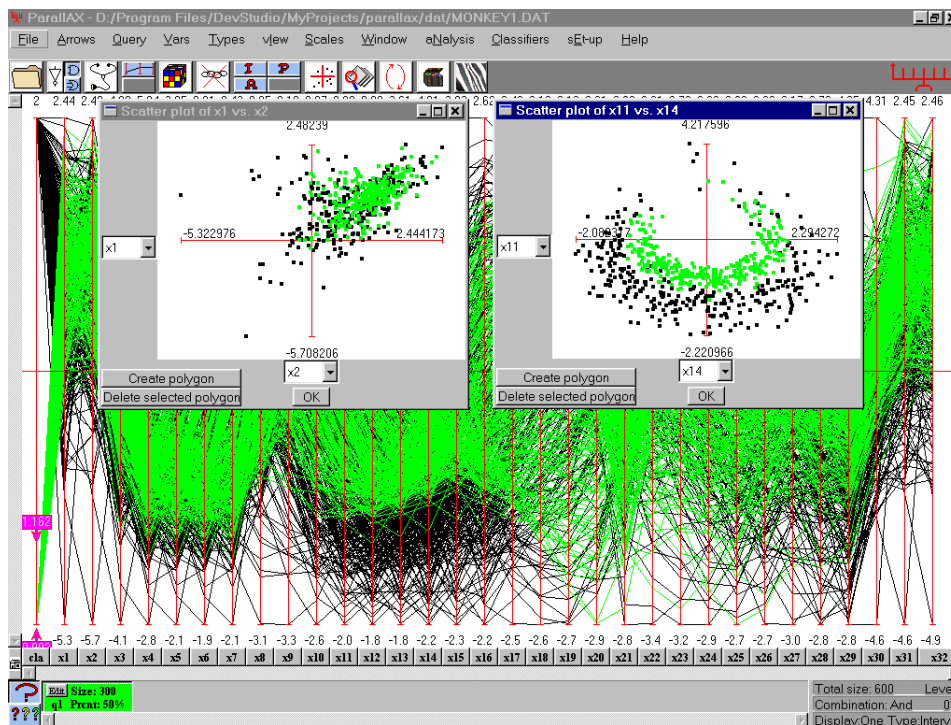


Figure 4: In the background is a dataset with 32 variables and 2 categories. Classifier finds the *nine* variables *features* needed to describe the classification rule with 4% error, and orders these variables according to their predictive value. On left is plot of first two variables and on the right the best two variables after classification.

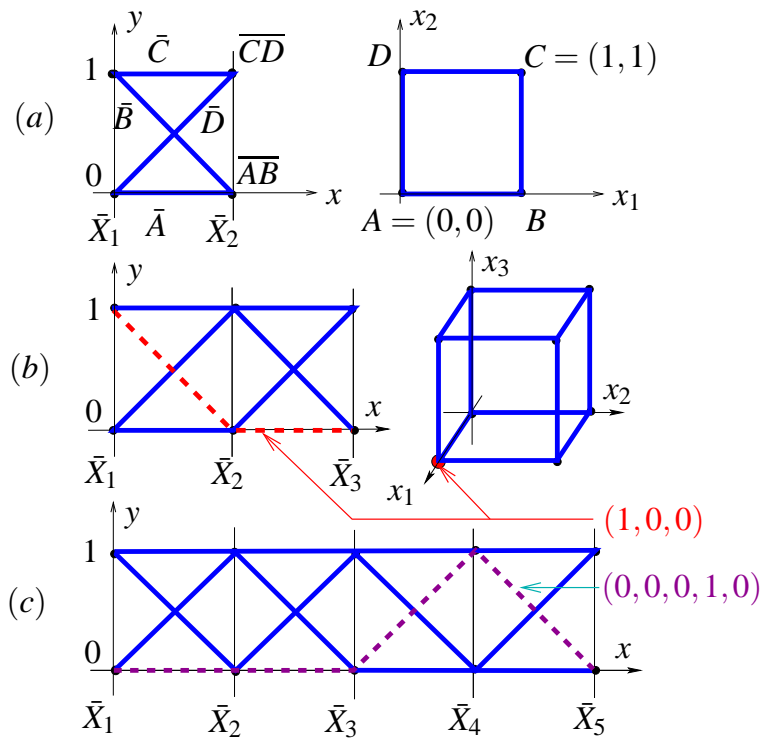


Figure 5: Square, cube and hypercube in 5-D

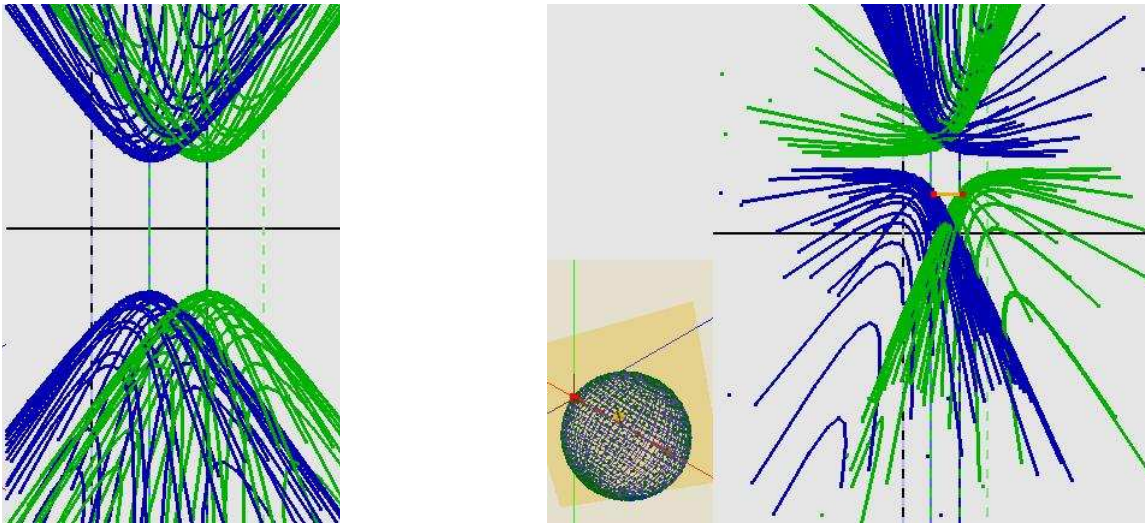


Figure 6: Representation of a sphere centered at the origin (left) and after a translation along the x_1 axis (right) causing the two hyperbolas to rotate in opposite directions. Note the *rotation* \leftrightarrow *translation* duality. In N-D a sphere is represented by $N - 1$ such hyperbolic regions — pattern repeats as for hypercube above.

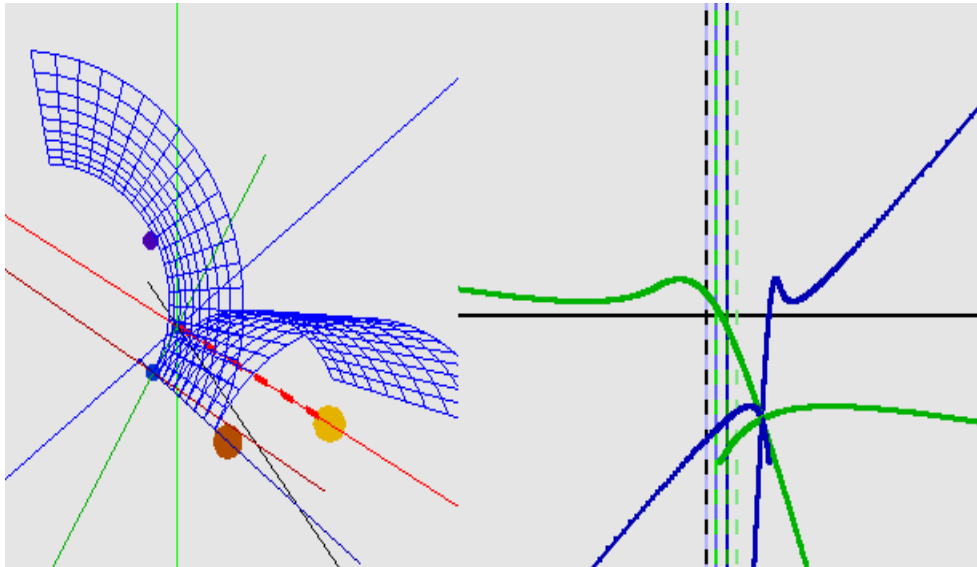


Figure 7: Note the two dualities $cusp \leftrightarrow inflection\ point$ and $bitangent\ plane \leftrightarrow crossing\ point$. Three such curves represent the corresponding hypersurface in 4-D and so on.

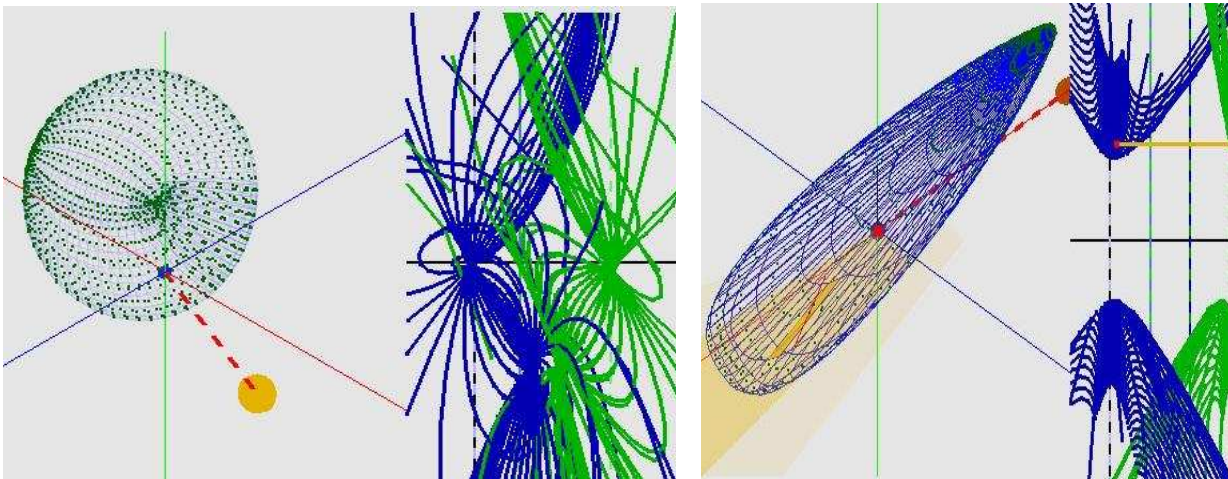


Figure 8: Representation of a surface with 2 “dimples” (depressions with cusp) which are mapped into pairs of “swirls” and are **all** visible. By contrast, in the perspective (left) one dimple is hidden. On the right is a convex surface represented by hyperbola-like (curves with two asymptotes) representation.

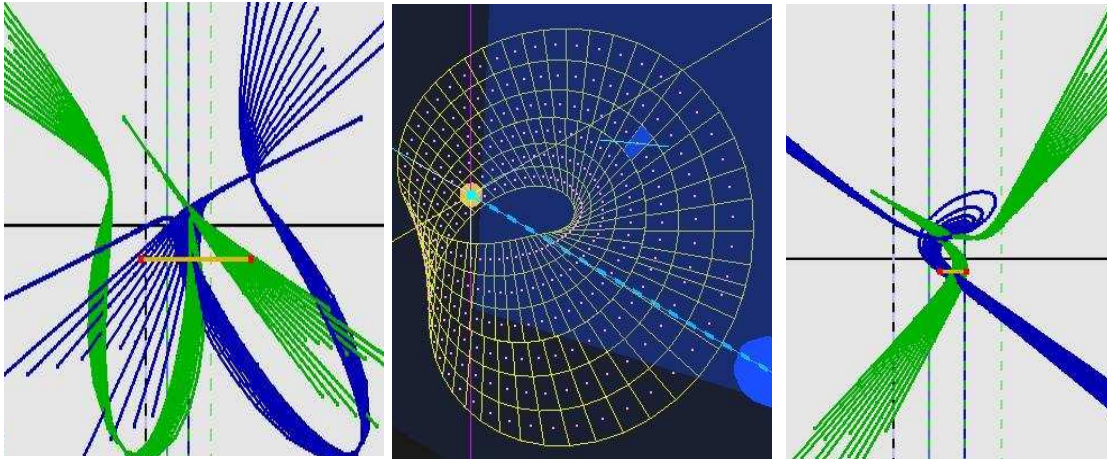


Figure 9: Möbius strip representation (left) has two cusps showing that the surface has an inflection point in 3-D (see Fig. 7 for *cusp* \leftrightarrow *inflection points* duality). This together with the upward/downward curves going to infinity in the same direction shows that the surface is closed and non-orientable. The curves and cusps may merge (right).

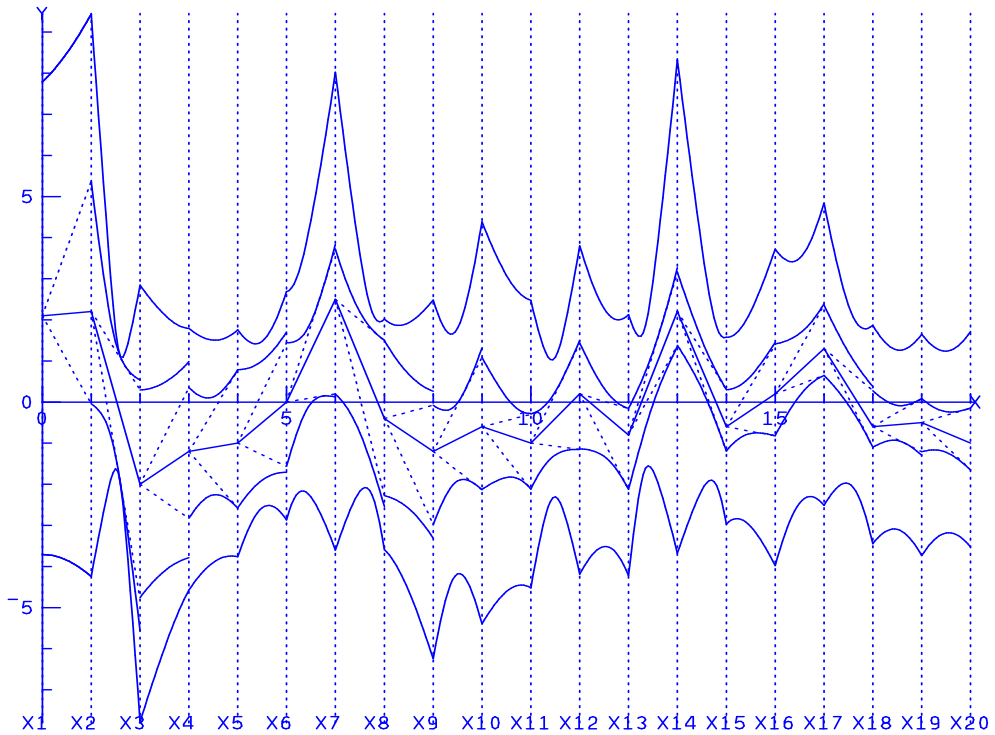


Figure 10: Interior point (polygonal line) construction algorithm shown for a convex hypersurface in 20 - D. A polygonal line touching any of the intermediate curves represents a point on the surface, and if it intersects one of the curves it represents an exterior point.