高维数据中低维结构的可视探索

中南大学
夏佳志

LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets

Jiazhi Xia[1], Fenjin Ye[1], Wei Chen[2*], Yusi Wang[1], Weifeng Chen[3], Yuxin Ma[2], Anthony K.H. Tung[4]
[1]**Central South University**
[2]ZheJiang University
[3]Zhejiang University of Finance & Economics
[4]National University of Singapore

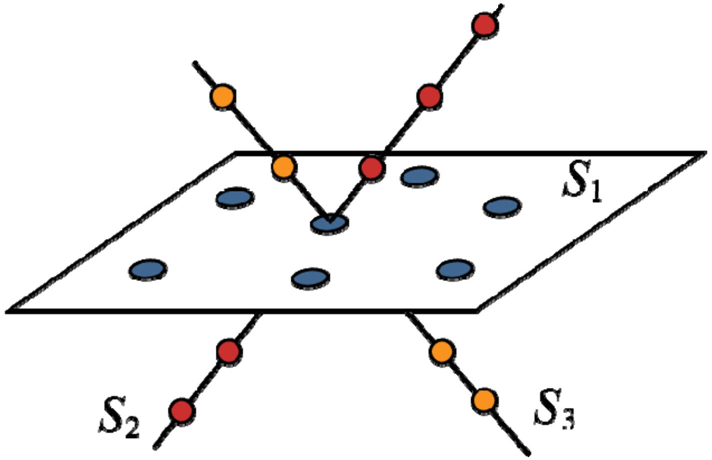
高维数据

- 通常指高维(Multidimensional)多元(Multivariate)数据
  - Multidimensional：数据具有多个独立属性
  - Multivariate：数据具有多个相关属性

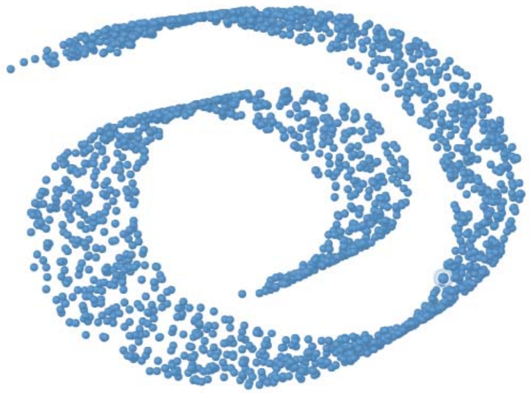
- 可视分析中的定义
  - 维度太高，以致难以从中提取可理解的维度关联信息
  - 一般来说，高于10维的数据可称为高维数据

  -E. Bertini, A. Tatu, D. Keim. Quality Metrics in High-Dimensional Data Visualization: An overview and Systematization. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2203-2212

高维数据中的低维结构



A dataset with three clusters, in a 2D **subspace**, and two 1-D **subspaces**, respectively.



Swiss roll in a 2D **manifold** embedded in 3D space.

Motivation：探索潜在的低维结构

- 给定一个未知的高维数据，如何探索其中潜在的低维结构?

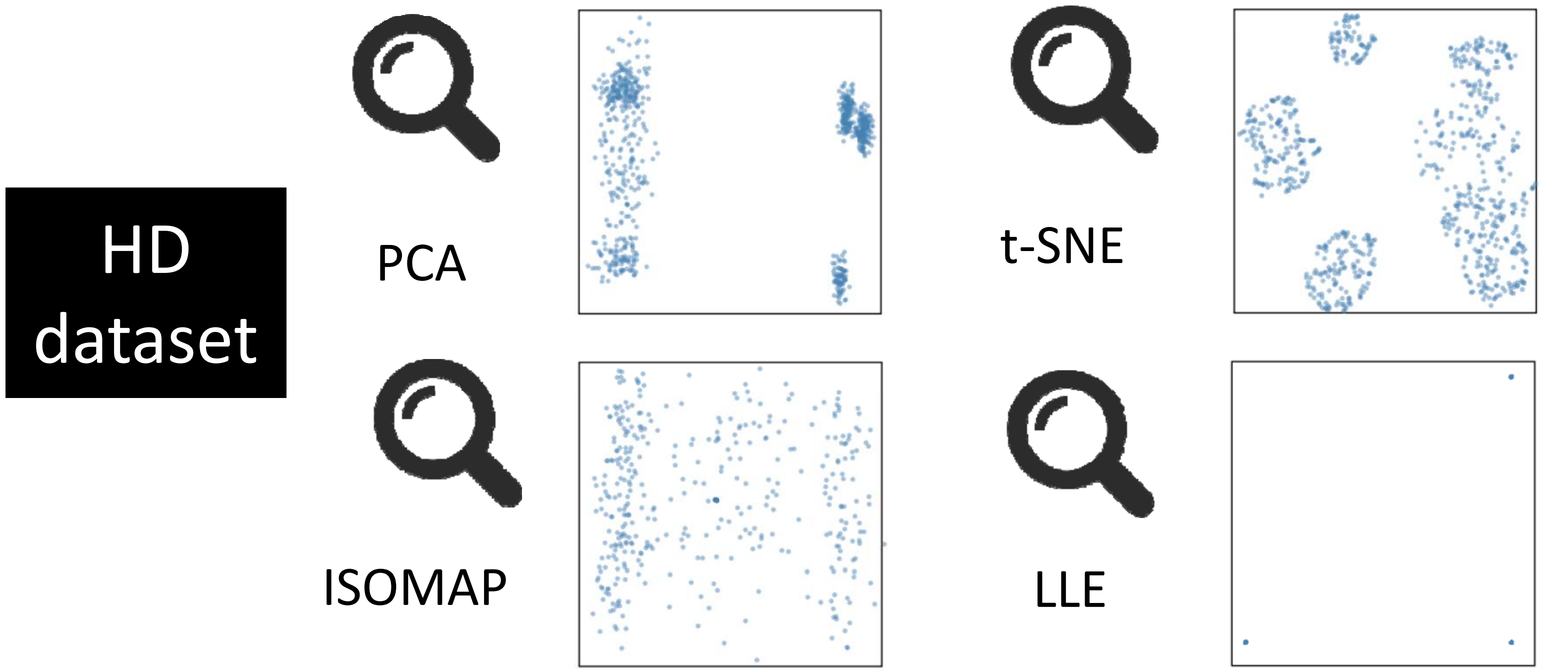
降维/聚类 三个阶段中的可视化

Stage	Methods
Post-processing	**Automatic models & Conventional Visualization**
Intra-processing	Interactive subspace analysis
Pre-processing	

自动化降维/聚类方法

Number of Subspaces / Manifolds	Linear	Non-linear
Single	**Linear DR**	**Manifold Learning**
Multiple	Subspace Clustering	Manifold Clustering

如何选择模型?



自动化降维/聚类方法

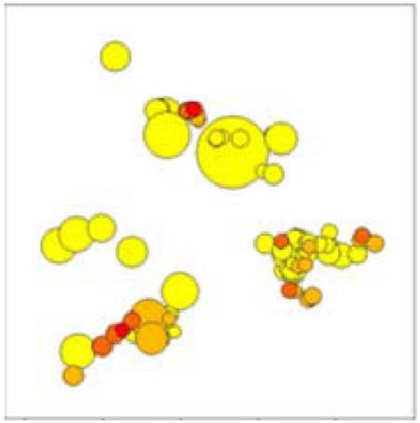
Number of Subspaces / Manifolds	Linear	Non-linear
Single	Linear DR	Manifold Learning
Multiple	**Subspace Clustering**	**Manifold Clustering**

子空间聚类&流形聚类

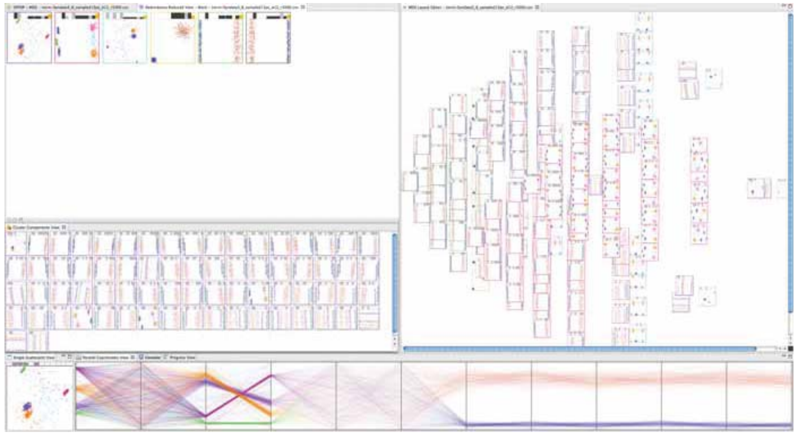
- 需要输入先验信息/满足先验假设
  - 聚类个数
  - 本真维度
  - 聚类分布

- 可能产生大量冗余的结果
  - 难以理解
  - 难以解释

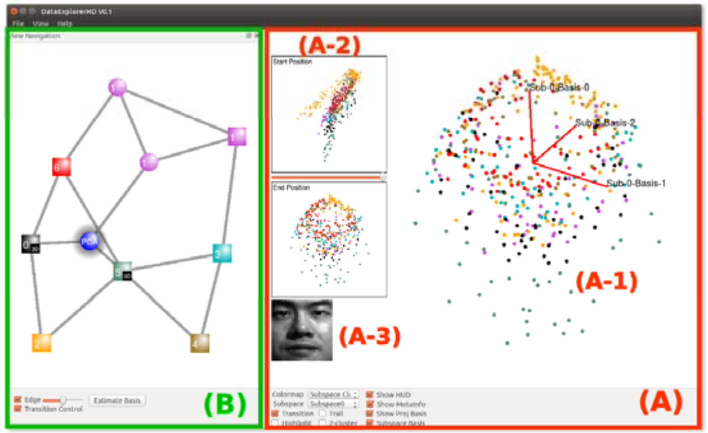
子空间聚类+ 可视化



I. Assent, et al. Visa: Visual subspace clustering analysis. SIGKDD, 2007.

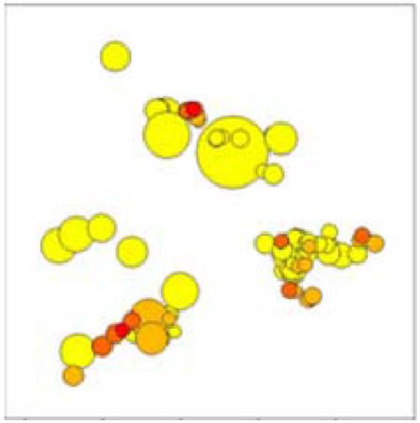


A. Tatu, et al. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. IEEE VAST, 2012.

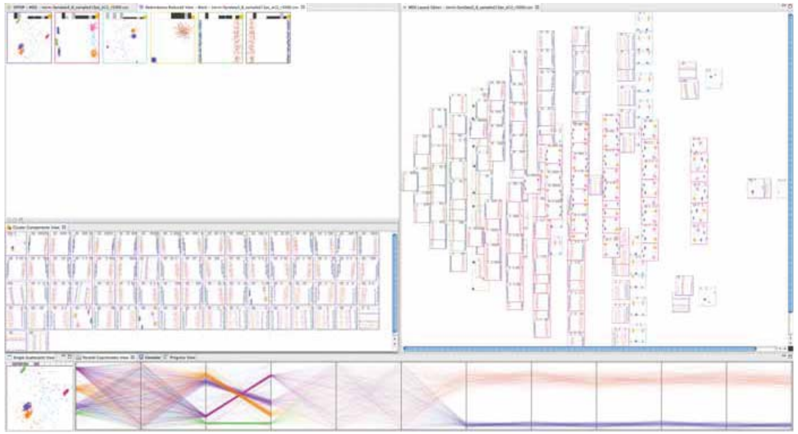


S. Liu, et al. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. EuroVis, 2015.

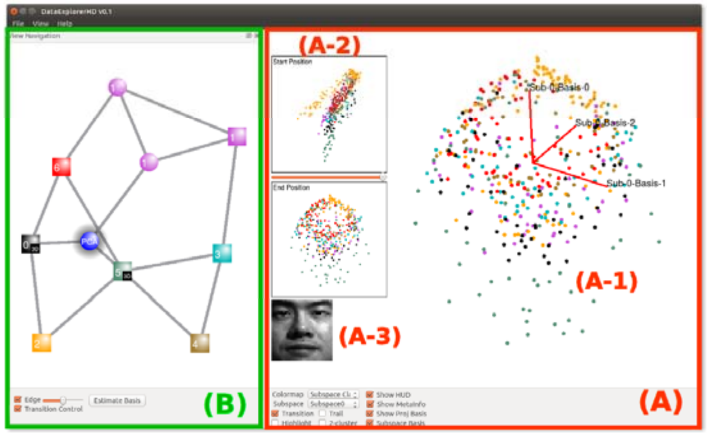
子空间聚类+可视化



I. Assent, et al. Visa: Visual subspace clustering analysis. SIGKDD, 2007.



A. Tatu, et al. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. IEEE VAST, 2012.



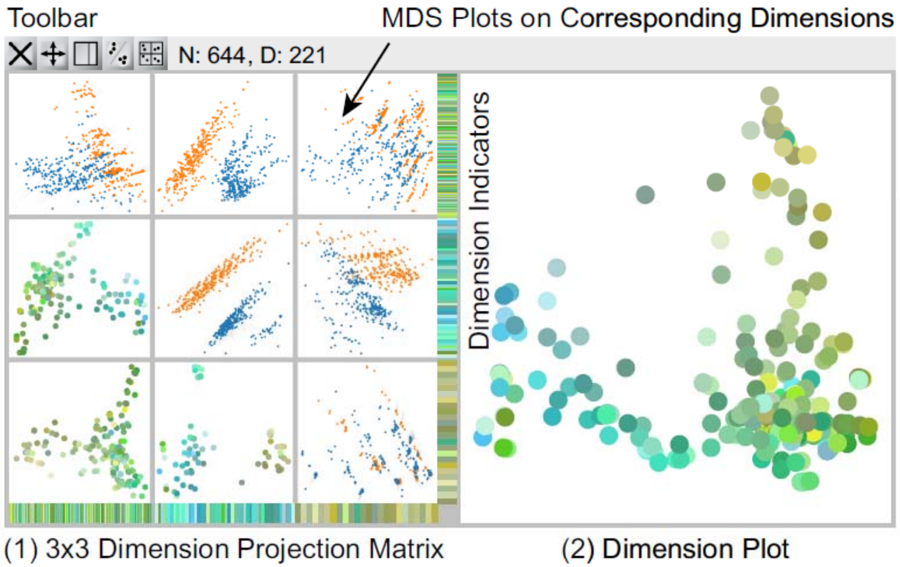
S. Liu, et al. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. EuroVis, 2015.

仅仅对降维/聚类结果的结构进行分析，难以验证结果的正确性

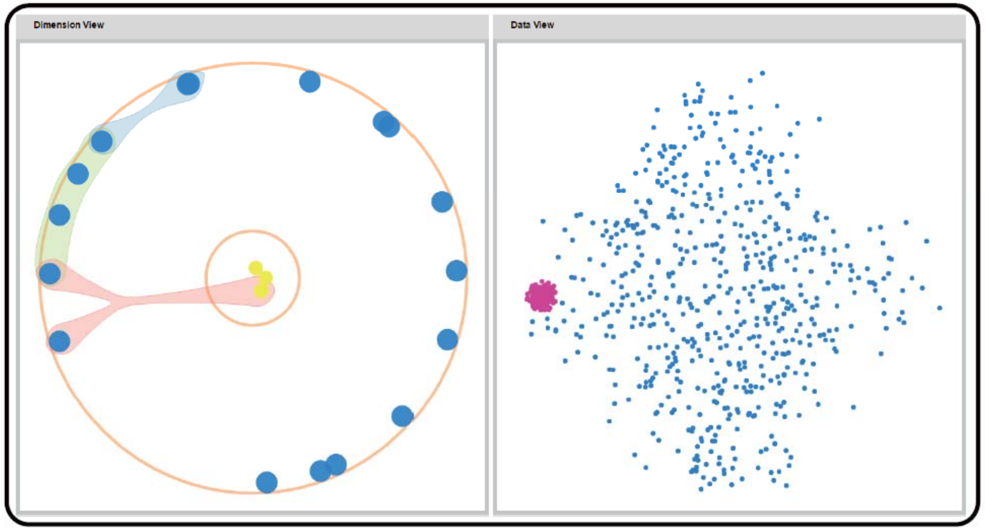
降维/聚类 三个阶段中的可视化

Stage	Methods	Limitation
Post-processing	Automatic models & Conventional Visualization	Blind model-choosing
Intra-processing	**Subspace exploration**	
Pre-processing		

子空间可视探索



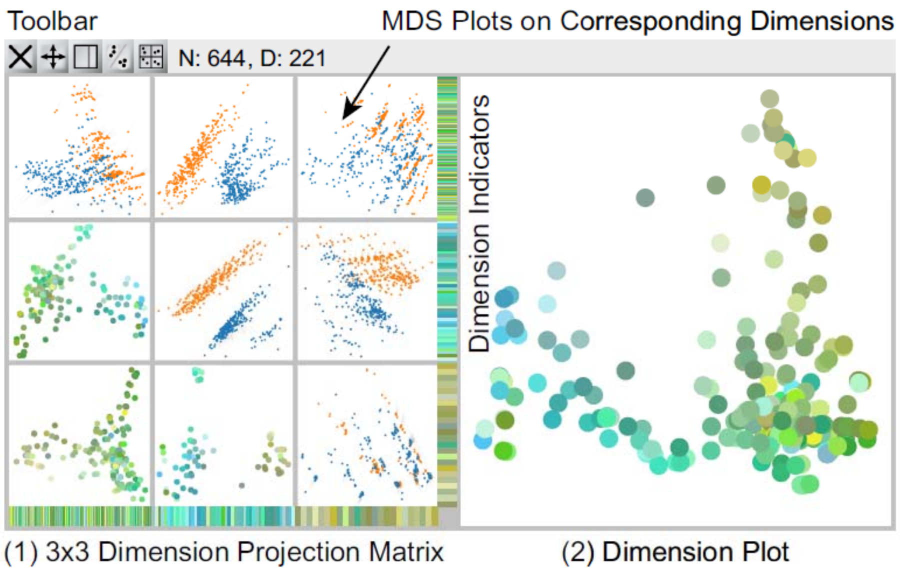
X. Yuan, et al. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. IEEE TVCG. 2013.

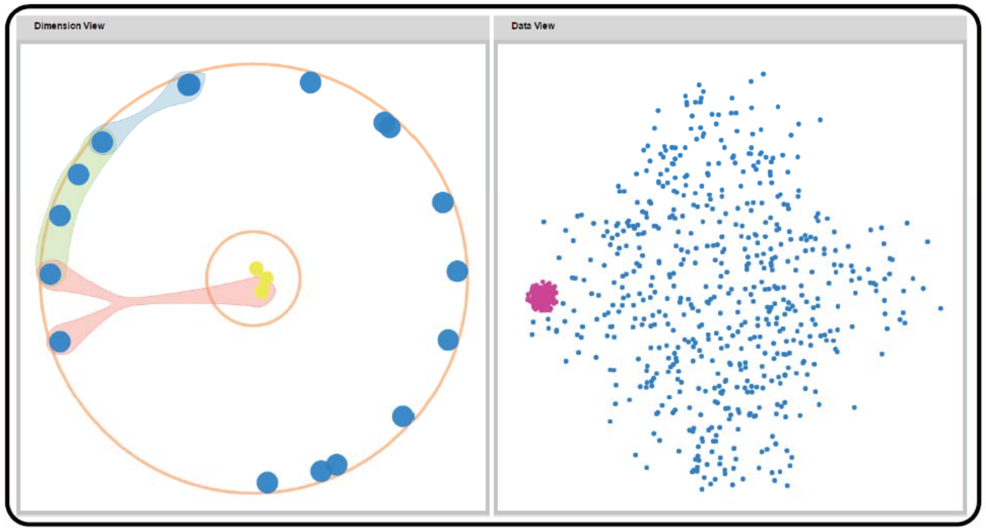


J. Xia, et al. Visual subspace clustering based on dimension relevance. JVLC. 2017

子空间可视探索

缺乏引导，冗长的试错循环





X. Yuan, et al. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. IEEE TVCG. 2013.

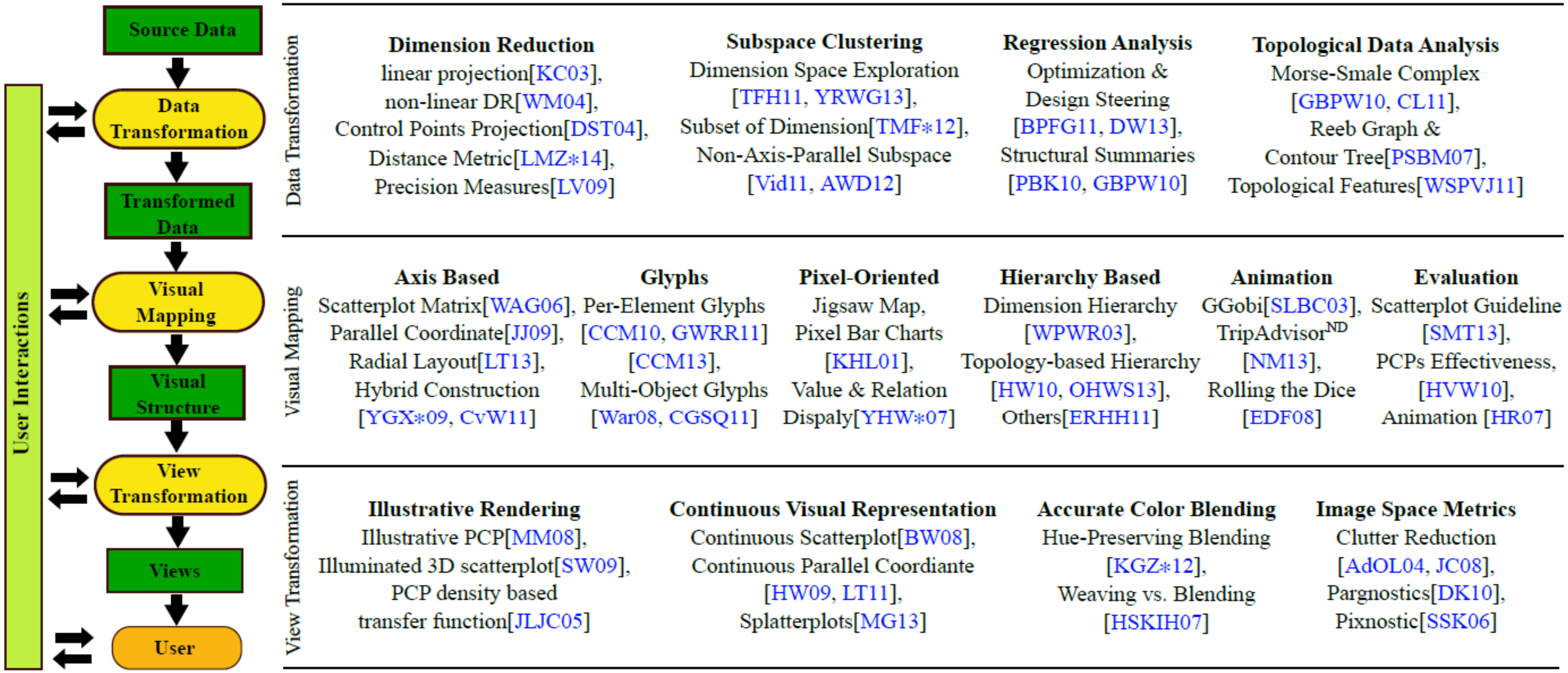
J. Xia, et al. Visual subspace clustering based on dimension relevance. JVLC. 2017

在建模之前，能否对数据进行预审察，
探索其潜在的低维结构?

降维/聚类 三个阶段中的可视化

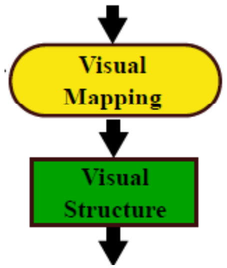
Stage	Methods	Limitation
Post-processing	Automatic models & Conventional Visualization	Blind model-choosing
Intra-processing	Subspace exploration	Trail-and-error
Pre-processing	**Exploratory Analysis ?**	

高维数据可视化技术分类



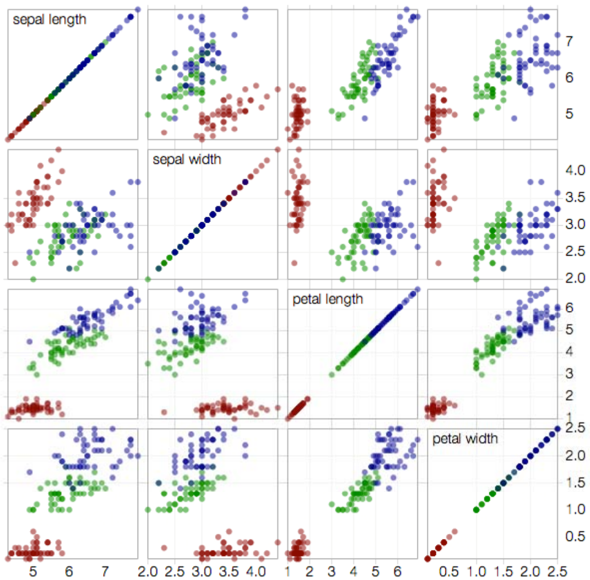
-Shusen Liu, Dan Maljovec, Bei Wang, Peer-Timo Bremer, and Valerio Pascucci. Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics* 23(3), 1249-1268, 2017.

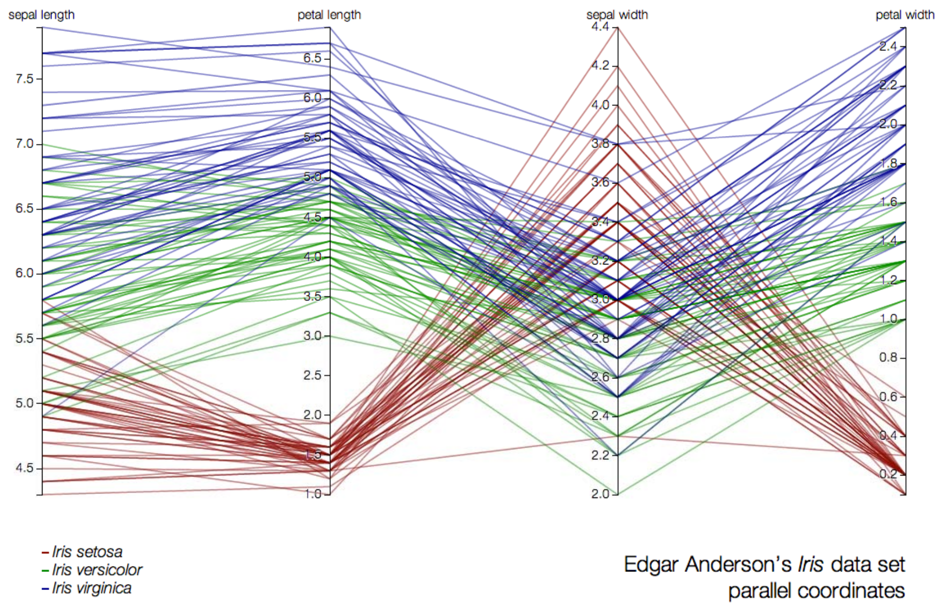
Visual Mapping



Visual Mapping

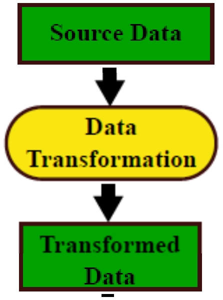
Axis Based	Glyphs	Pixel-Oriented	Hierarchy Based	Animation	Evaluation
Scatterplot Matrix[WAG06], Parallel Coordinate[JJ09], Radial Layout[LT13], Hybrid Construction [YGX*09, CvW11]	Per-Element Glyphs [CCM10, GWRR11] [CCM13], Multi-Object Glyphs [War08, CGSQ11]	Jigsaw Map, Pixel Bar Charts [KHL01], Value & Relation Dispaly[YHW*07]	Dimension Hierarchy [WPWR03], Topology-based Hierarchy [HW10, OHWS13], Others[ERHH11]	GGobi[SLBC03], TripAdvisorND [NM13], Rolling the Dice [EDF08]	Scatterplot Guideline [SMT13], PCPs Effectiveness, [HVW10], Animation [HR07]





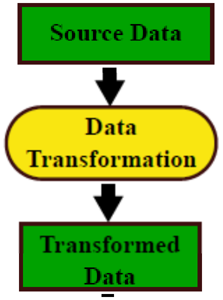
Edgar Anderson's *Iris* data set
parallel coordinates

Data Transformation



Data Transformation	**Dimension Reduction** linear projection[KC03], non-linear DR[WM04], Control Points Projection[DST04], Distance Metric[LMZ∗14], Precision Measures[LV09]	**Subspace Clustering** Dimension Space Exploration [TFH11, YRWG13], Subset of Dimension[TMF∗12], Non-Axis-Parallel Subspace [Vid11, AWD12]	**Regression Analysis** Optimization & Design Steering [BPFG11, DW13], Structural Summaries [PBK10, GBPW10]	**Topological Data Analysis** Morse-Smale Complex [GBPW10, CL11], Reeb Graph & Contour Tree[PSBM07], Topological Features[WSPVJ11]

Data Transformation



	Dimension Reduction	Subspace Clustering	Regression Analysis	Topological Data Analysis
Data Transformation	linear projection[KC03], non-linear DR[WM04], Control Points Projection[DST04], Distance Metric[LMZ∗14], Precision Measures[LV09]	Dimension Space Exploration [TFH11, YRWG13], Subset of Dimension[TMF∗12], Non-Axis-Parallel Subspace [Vid11, AWD12]	Optimization & Design Steering [BPFG11, DW13], Structural Summaries [PBK10, GBPW10]	Morse-Smale Complex [GBPW10, CL11], Reeb Graph & Contour Tree[PSBM07], Topological Features[WSPVJ11]

可视分析设计由任务驱动

分析目标

- 探索潜在的低维结构
- 为 选择/使用/调整 自动化模型提供信息

Overview

Analysis Tasks

- T1: NO. of clusters
- T2: Linear or Non-linear
- T3: Intrinsic dimensionality
- T4: Distribution
- T5: Locality Assumption

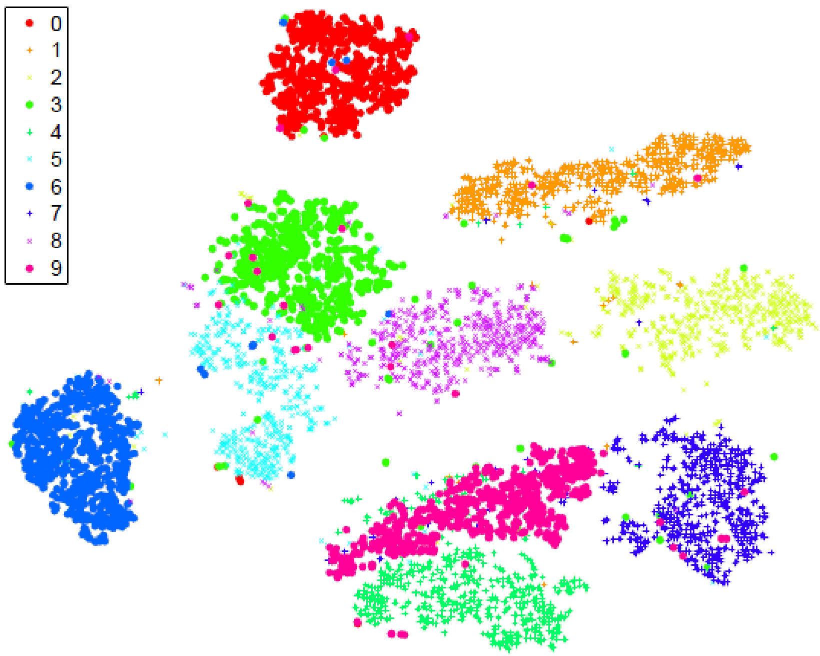
Data/Feature Abstraction

- Partition
- Geodesic distance
- Local tangent space
- LTS Diversity
- k-neighborhood locality

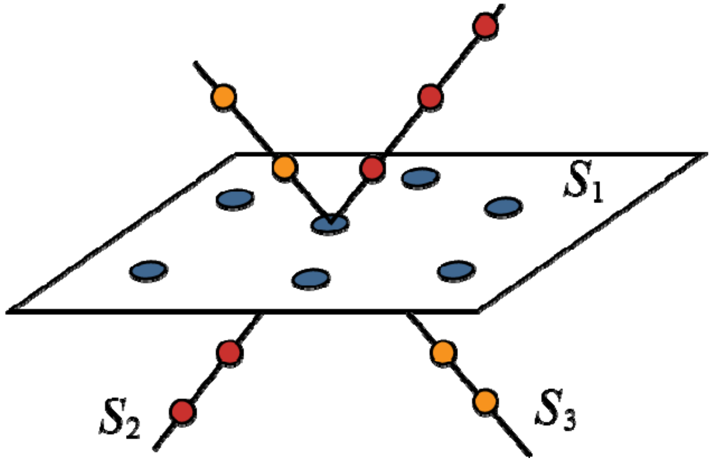
Visual Design

- LTSD-GD View
- T-SNE View
- Est. Local dimensionality
- Scree plot of point-wise LTSs
- Scree plot of structures
- Identified structures view

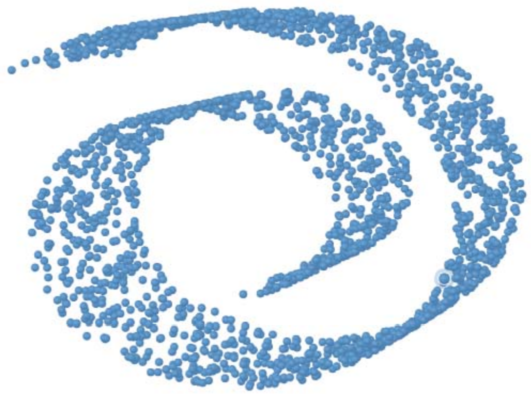
分析任务一：低维结构（聚类）的个数



分析任务二：低维结构是线性还是非线性？



线性子空间



流形

分析任务三：低维结构的本真维度

- 大量的子空间聚类/流形聚类方法需要本真维度作为算法的输入

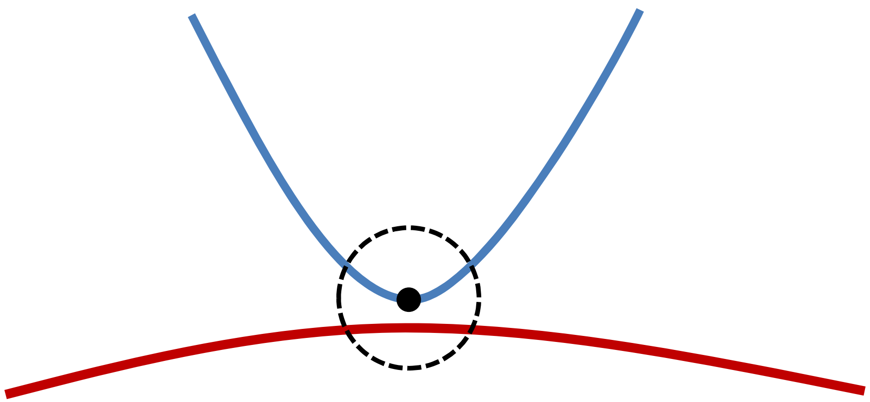
- 先验信息？ 试错？

分析任务四：低维结构的分布情况

- 低维结构之间的距离
- 低维结构所在空间之间的距离
- 低维结构之间是否交叉

分析任务五：局部性假设是否成立

- 局部性假设：流形上一点，与其邻域内所有的点都位于同一个流形上



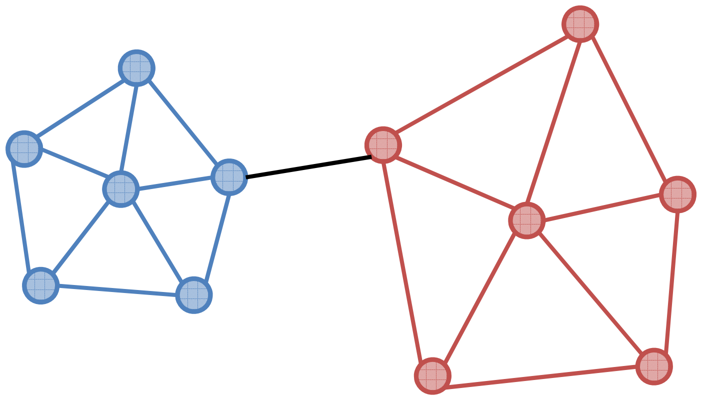
低维结构的表达

- 子空间
  - 基于线性系统的表达
- 流形
  - 邻域图（来自流形学习方法）

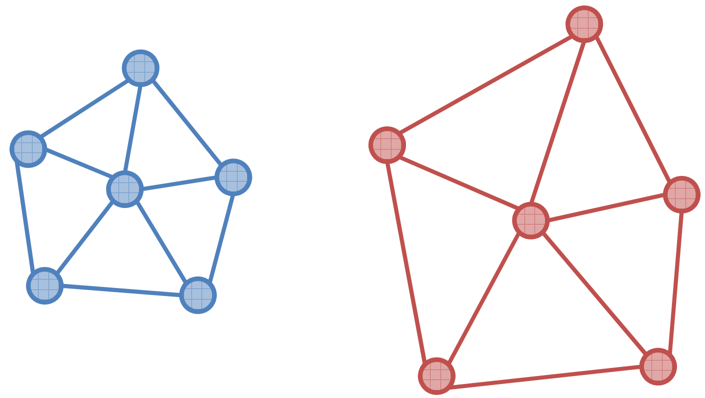
- 给定一个未知数据集
  - 需要一个统一的表达
  - 邻接图

特征提取: 分割

- Shared Nearest Neighbors (SNN) graph

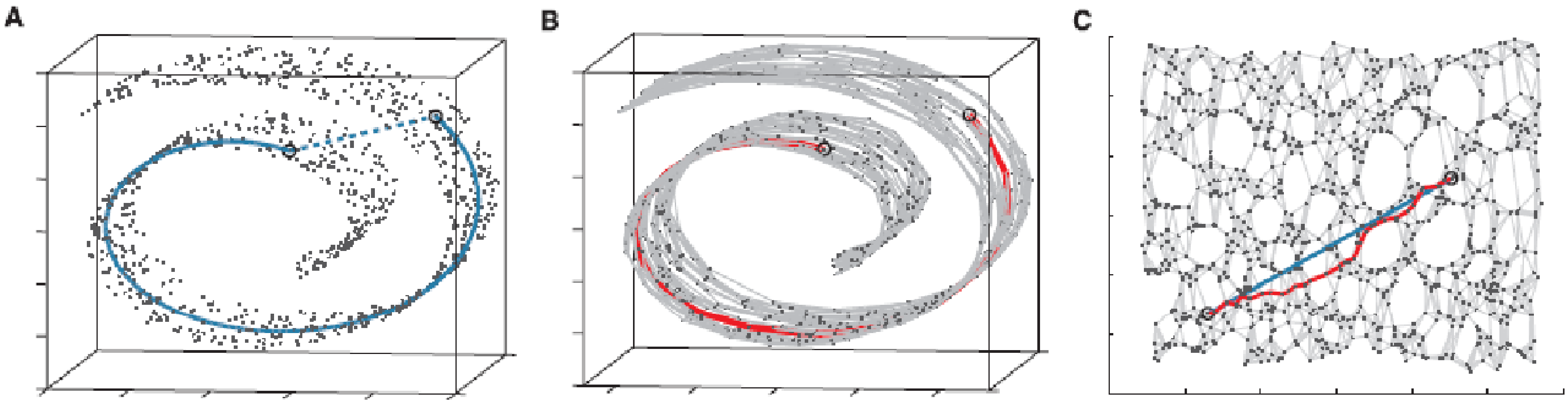


k-NN



SNN

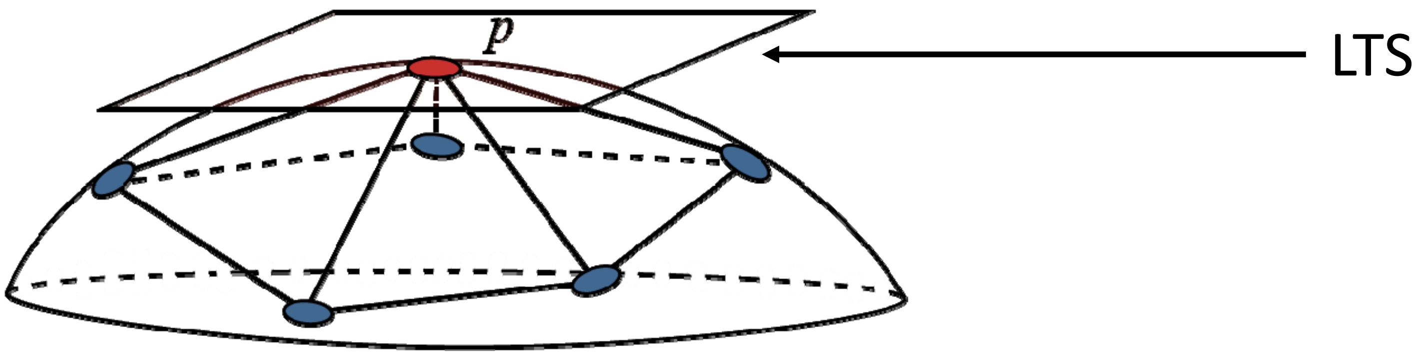
特征提取: 测地距离



-Joshua B. Tenenbaum, et al. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science. 2000.

特征提取: Local Tangent Space (LTS)

- Locality assumption: for each point, there is a small neighborhood, which contains only points of the same manifold



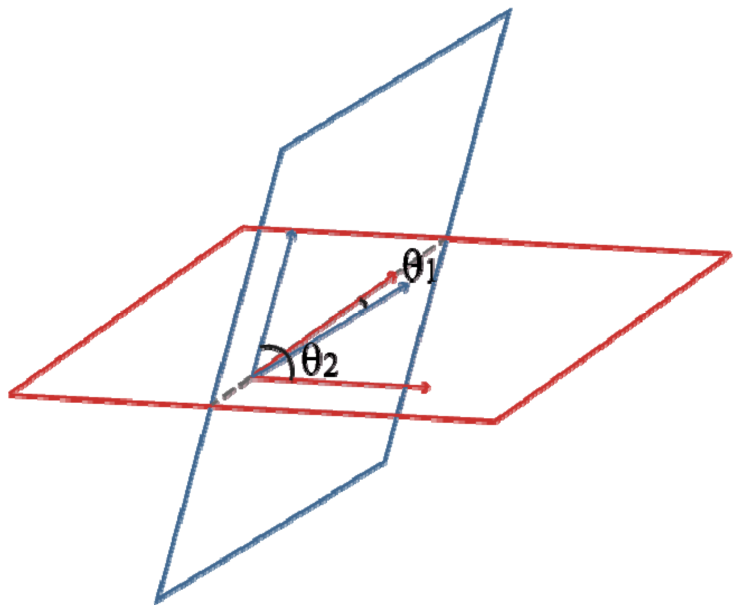
  – The local tangent space is fit by *k*-nearest neighbors (SVD)

特征提取: Local Tangent Space Divergence (LTSD)

- A divergence measurement between two subspaces

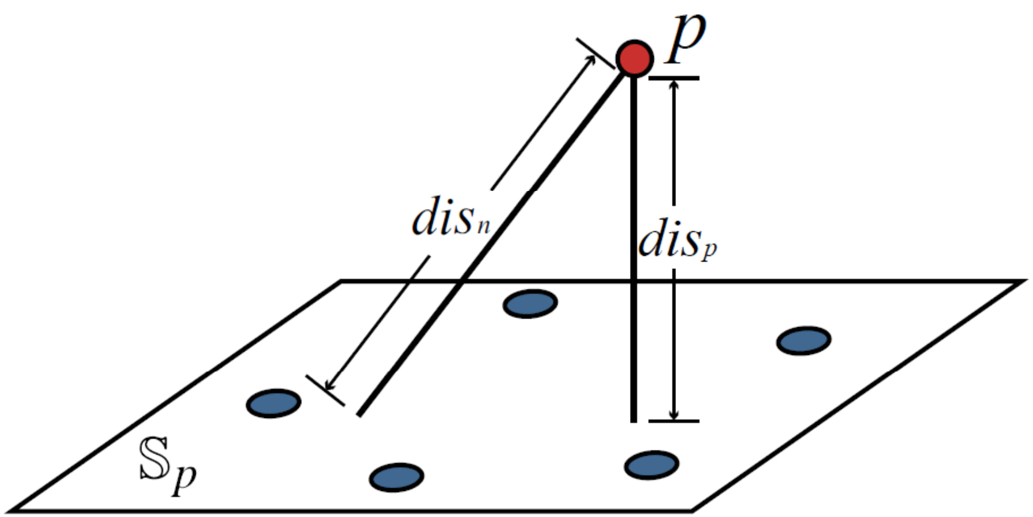
- It is defined by the principal angels between two subspaces

$$1 - \sqrt{\frac{cos^2\theta^{(1)} + \ldots + cos^2\theta^{(d_p \wedge d_q)}}{d_p \wedge d_q}}$$

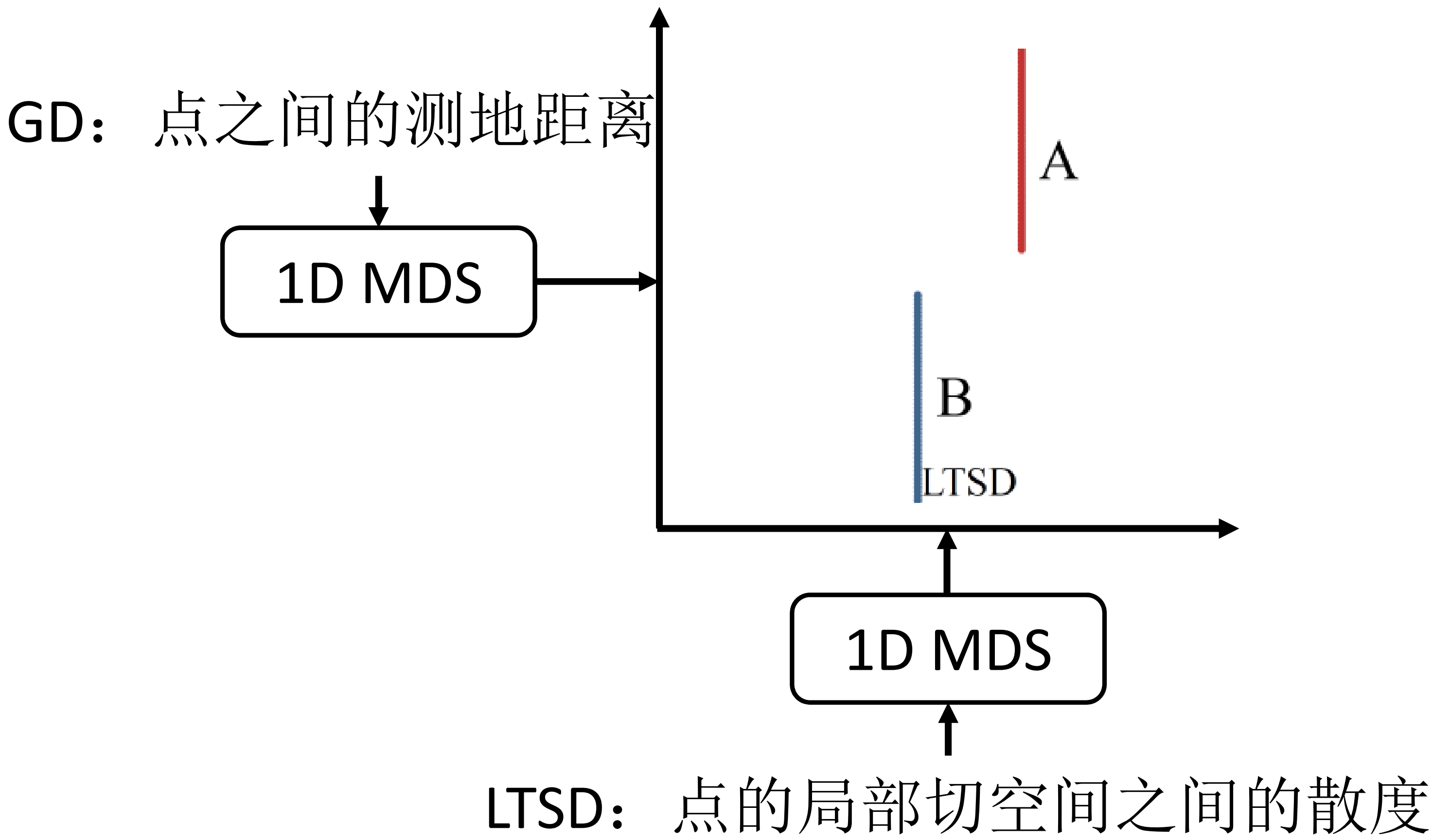


特征提取：局部性估计

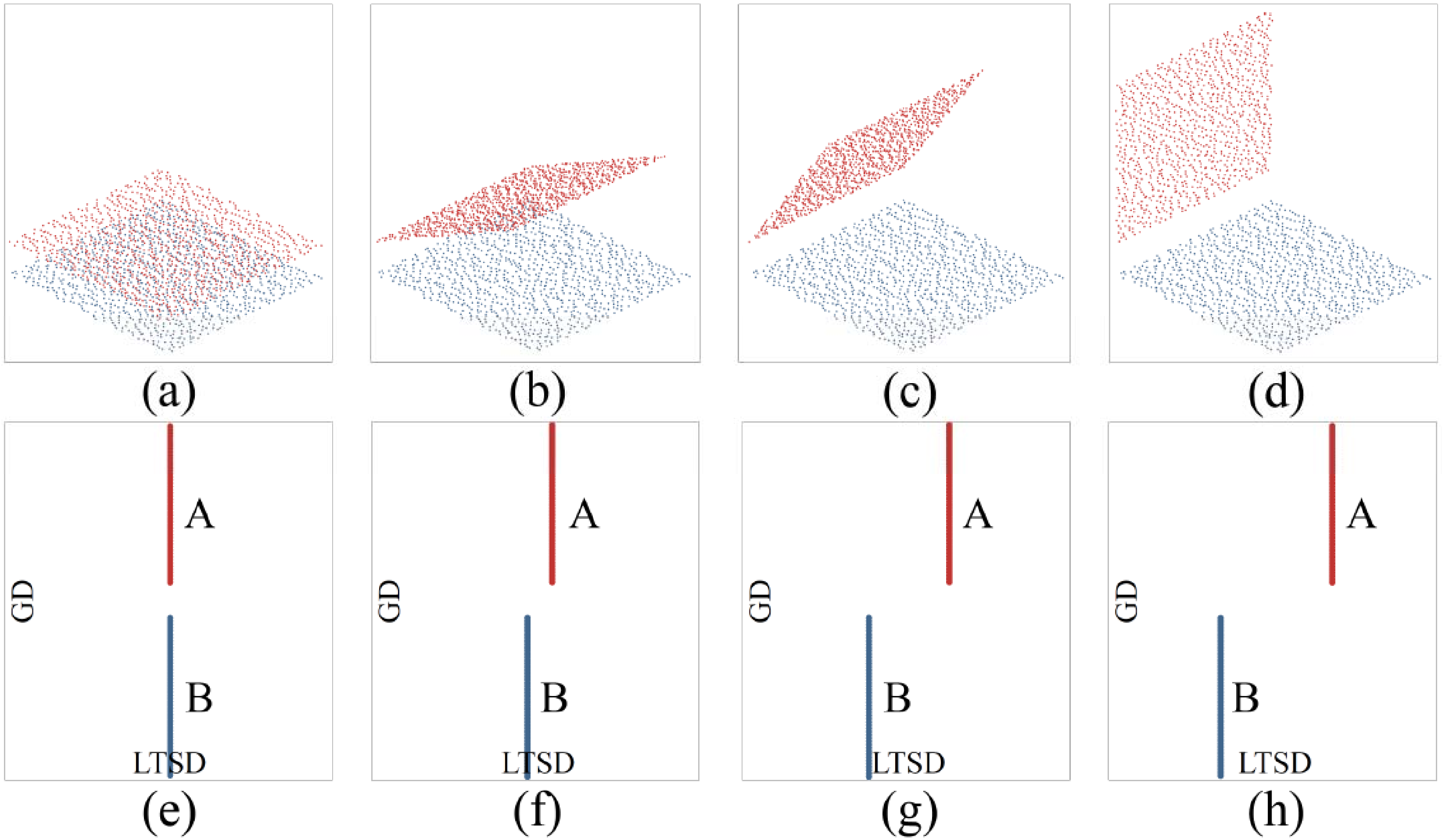
- k-neighborhood locality
  - The locality assumption may not hold due to noise and complicated structure, e.g. intersection.



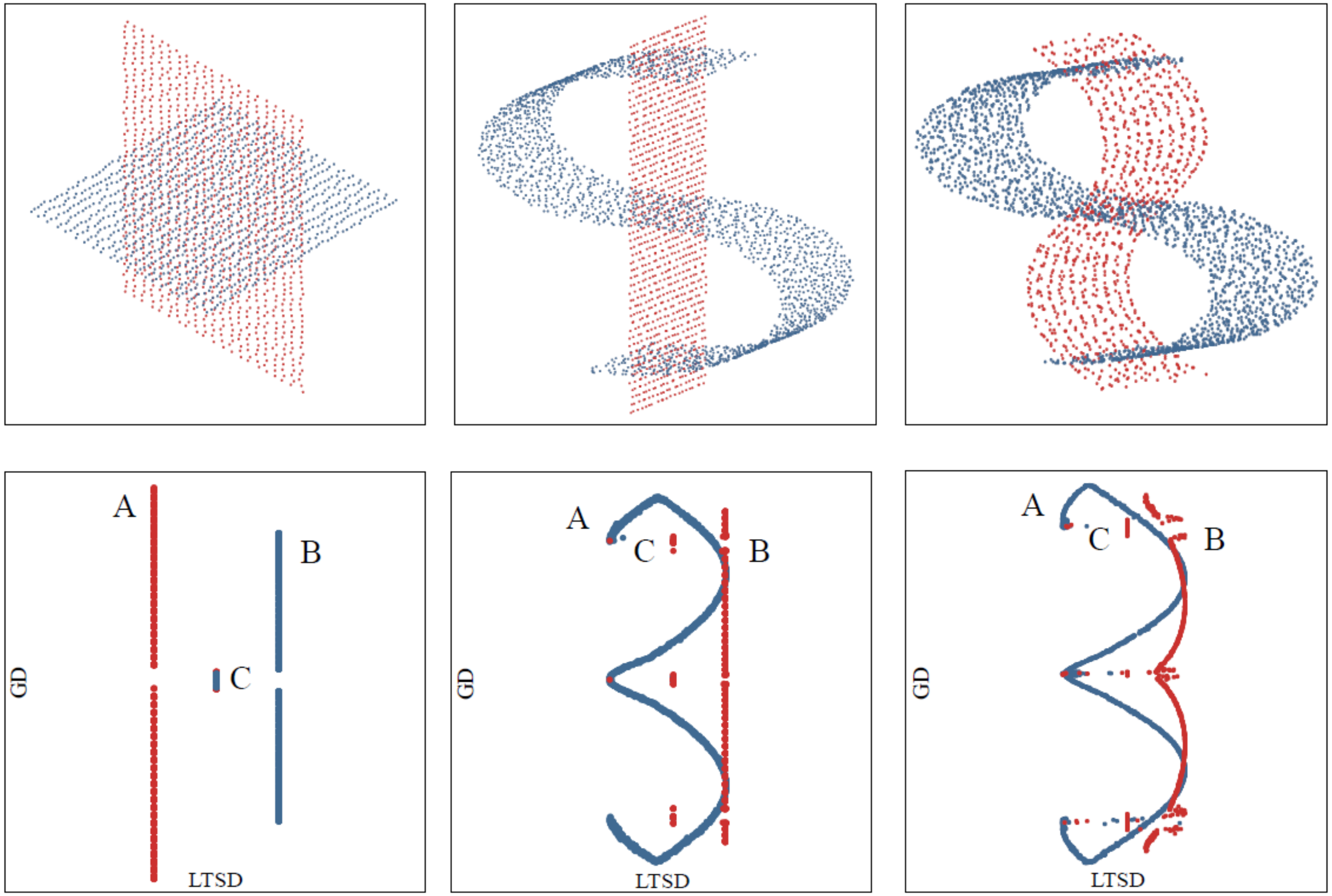
The LTSD-GD view



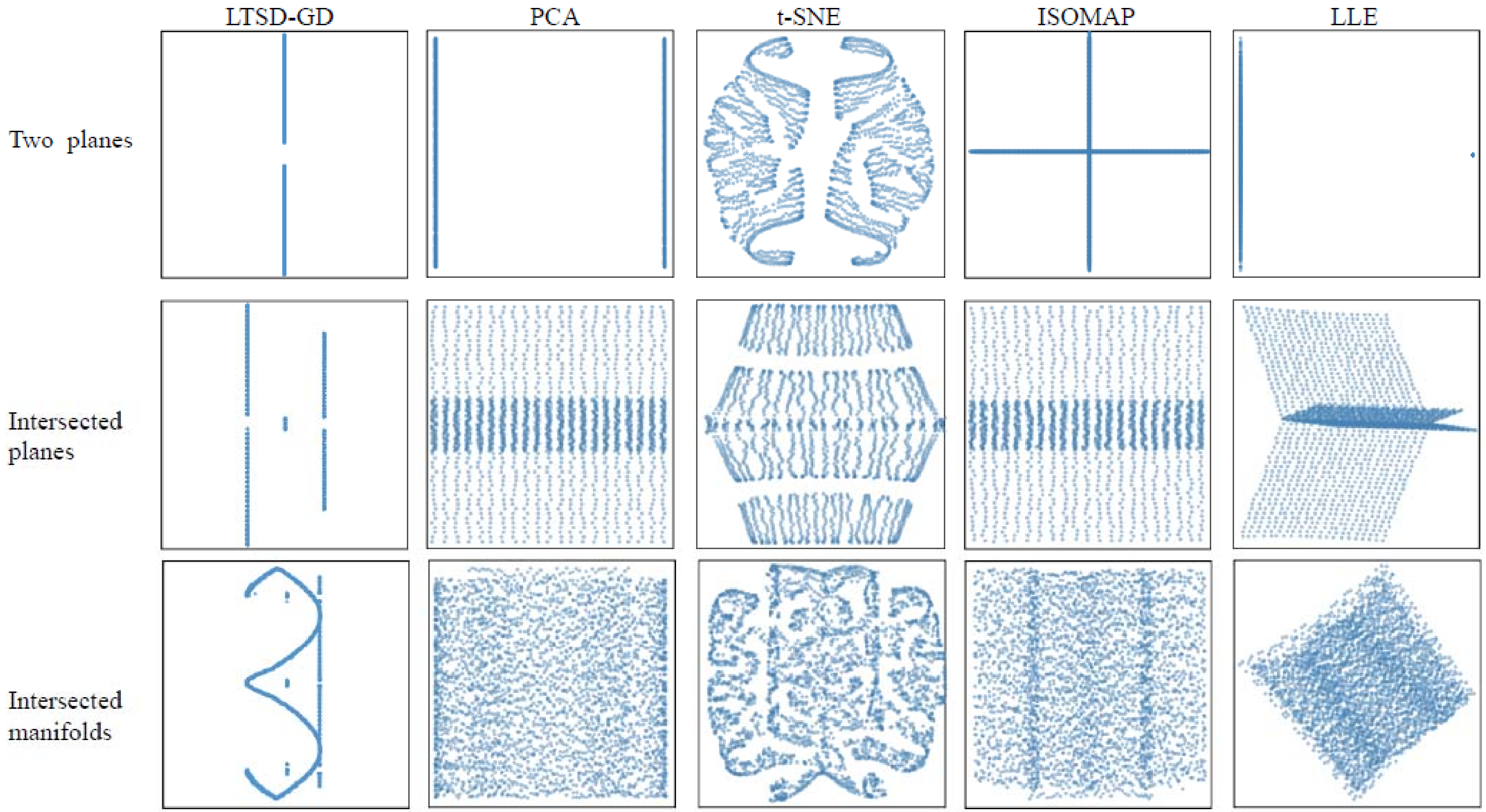
LTSD-GD 视图：两个子空间



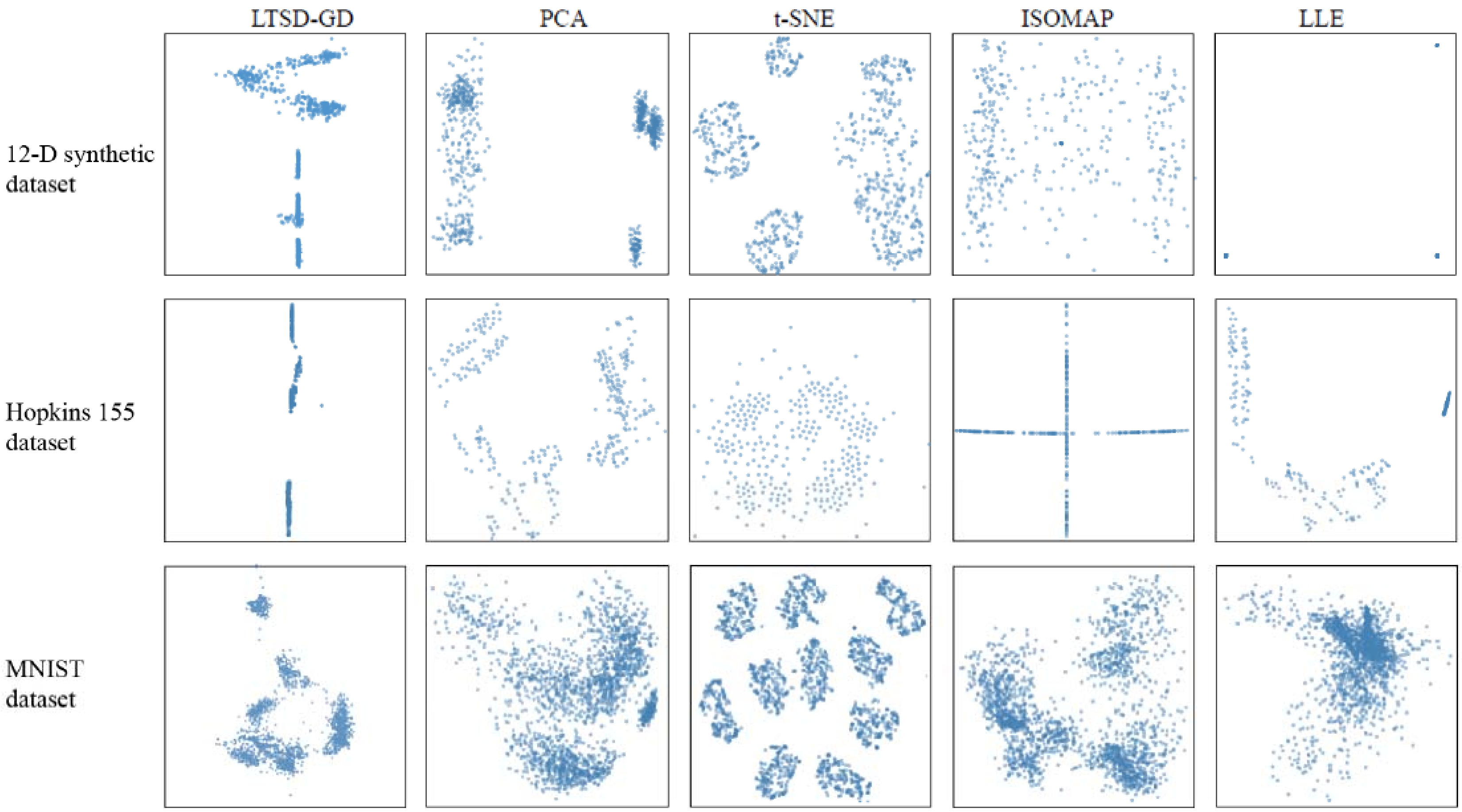
LTSD-GD 视图：两个相交的流形



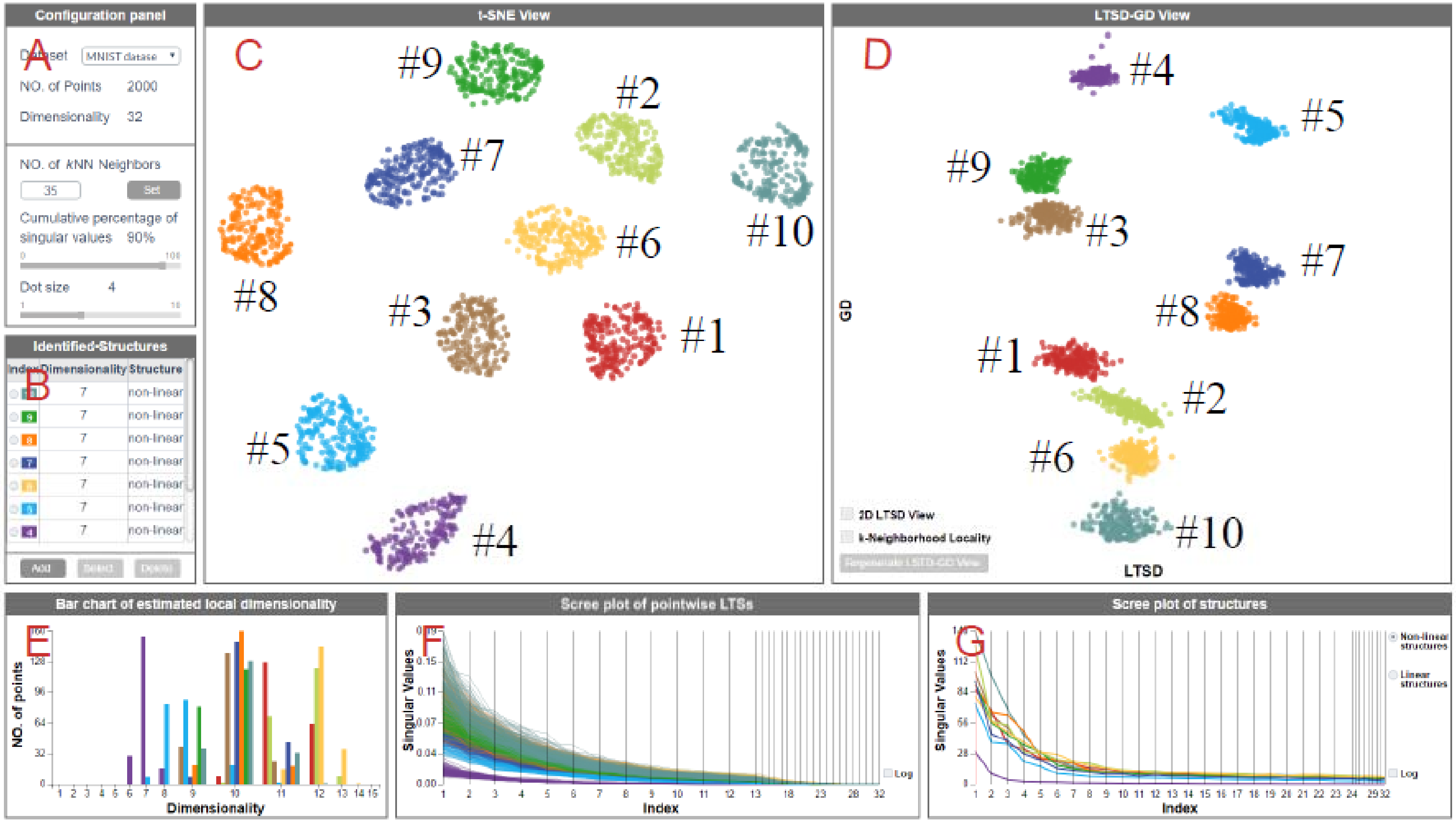
Comparisons



Comparisons



Visual analysis interface

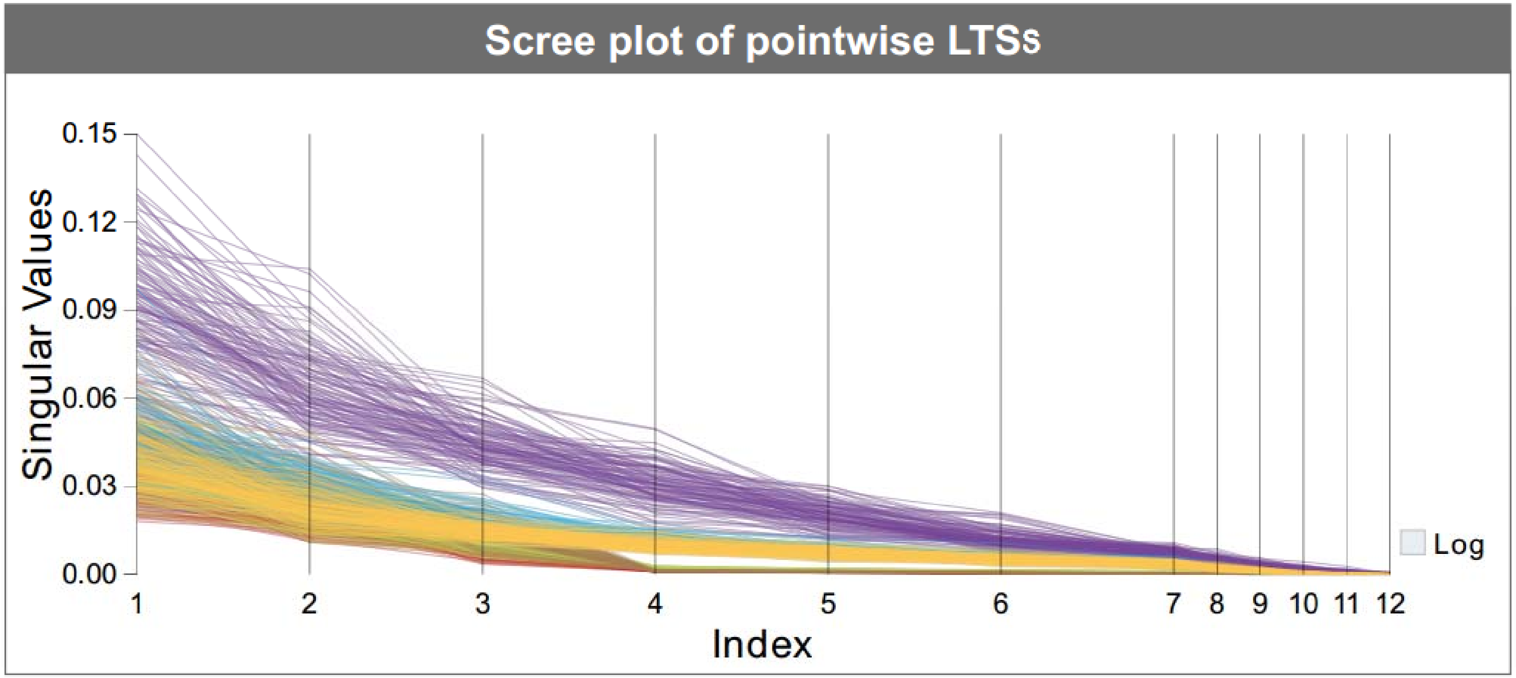


t-SNE view

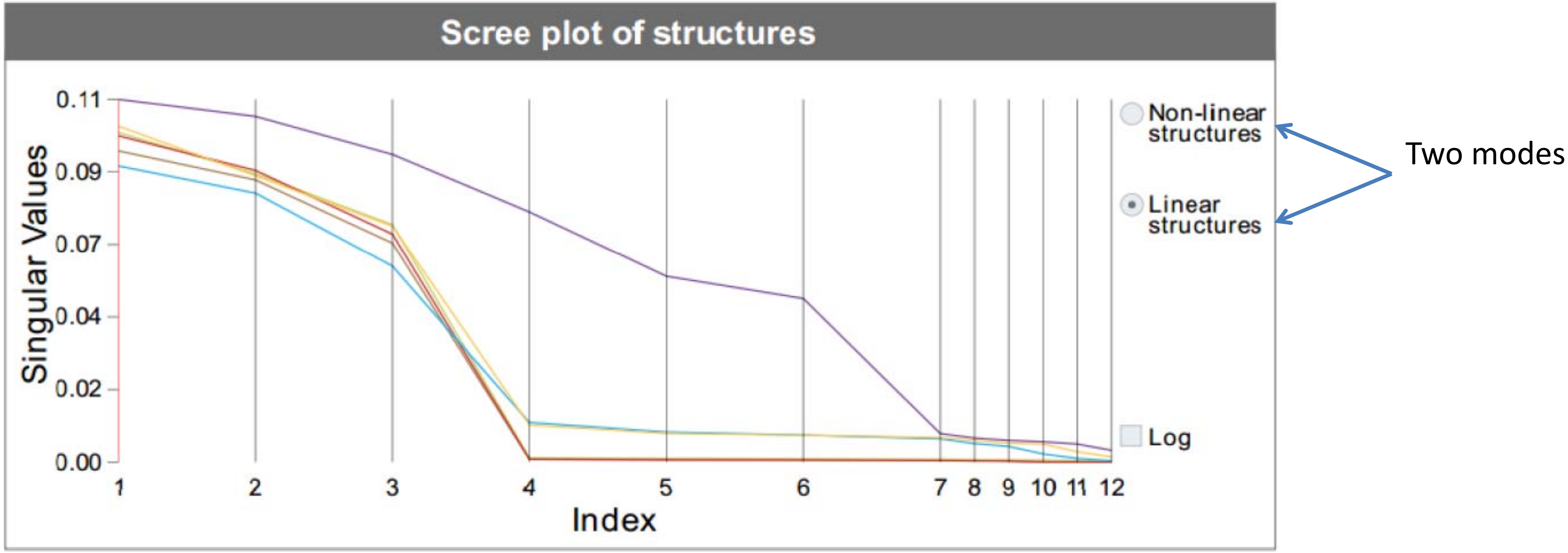
- 提供聚类信息

- 需要比较少的先验信息
  - 仅需要满足局部性假设

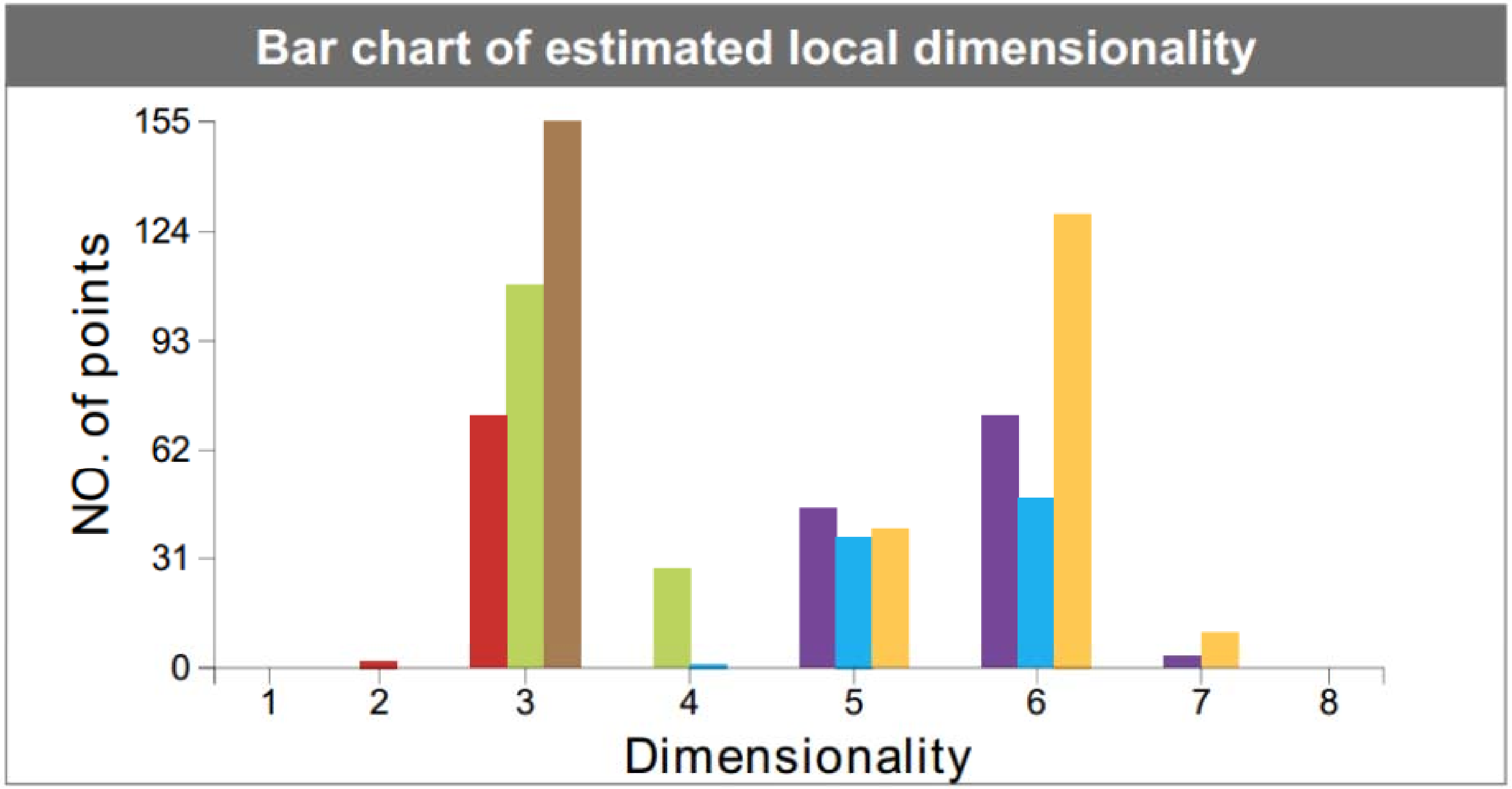
Scree plot of pointwise LTSs



Scree plot of structures

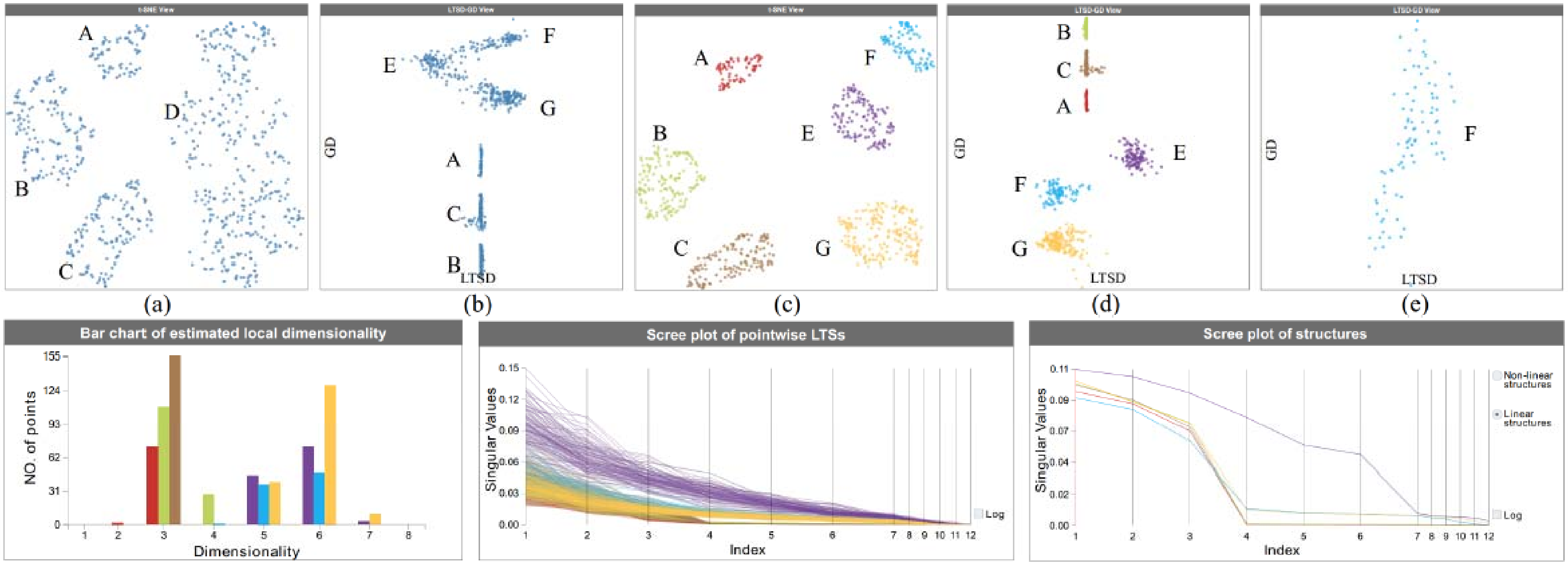


Bar chart of estimated local dimensionality



Case study: the synthetic 12-D dataset

12D, 750 points, in six clusters

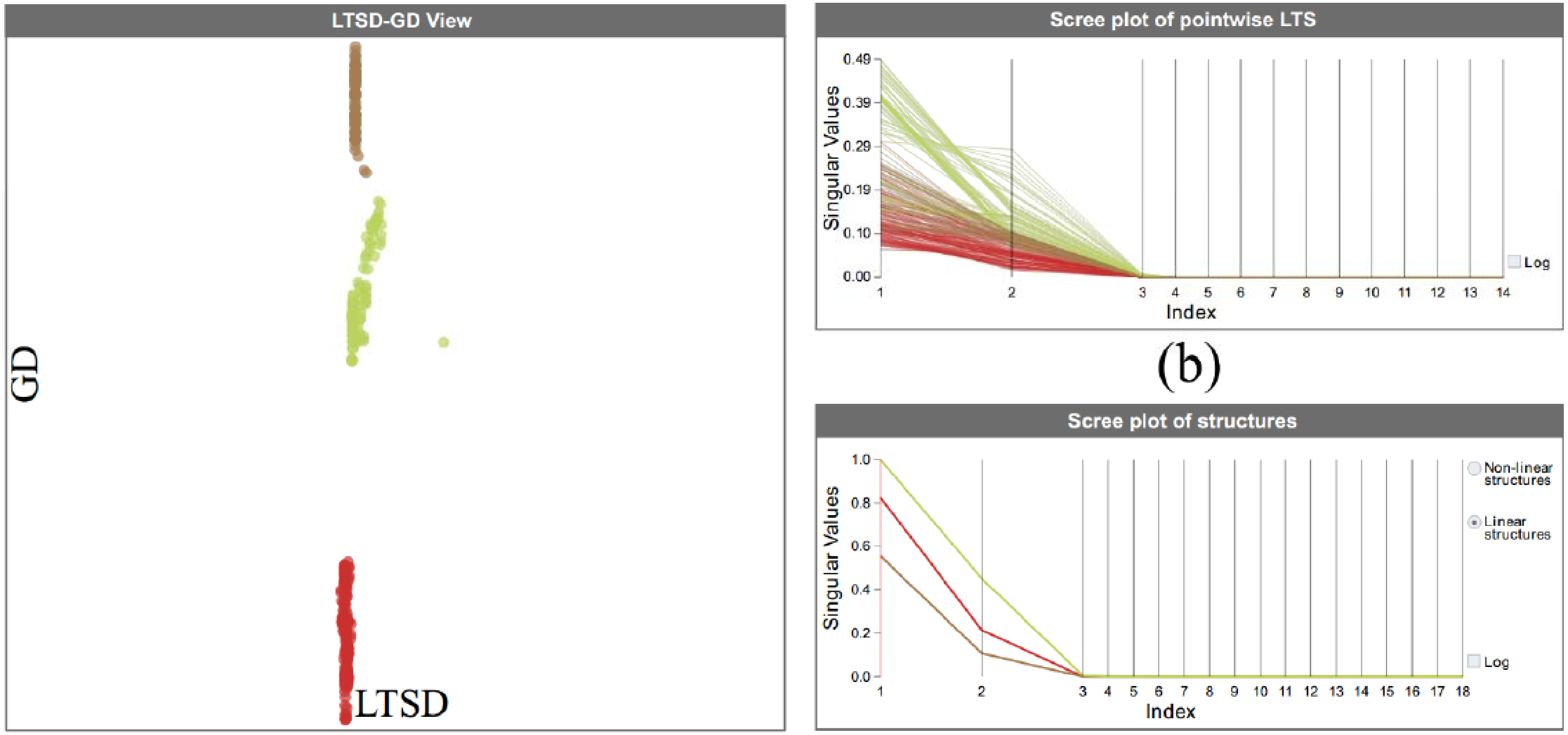


Exploring The Synthetic 12D Data

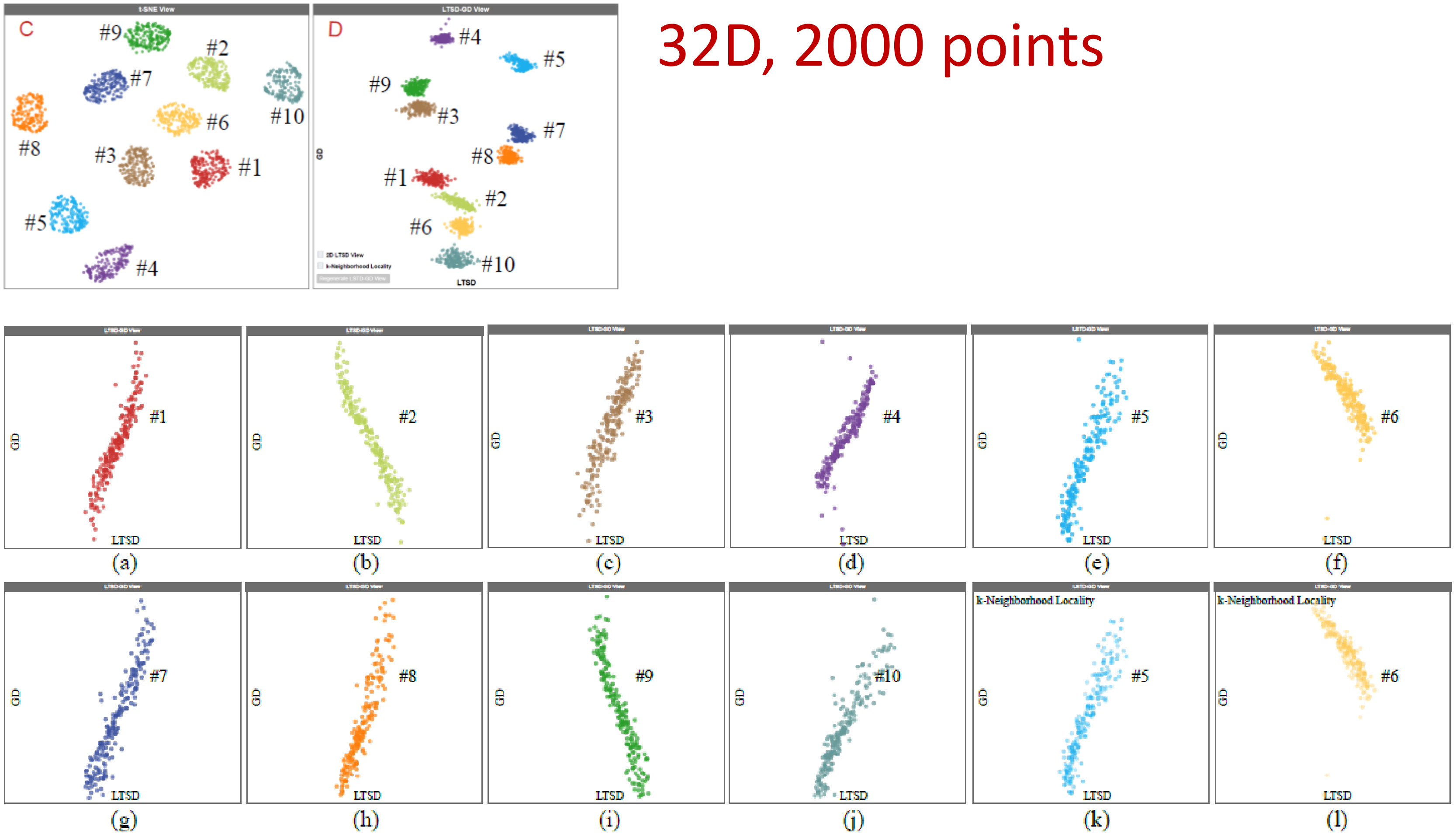
750 points, 12 dimensions (in 6 Gaussian clusters)

Case study: the Hopkins 155 dataset

62D, 297 points



Case study: the MINST dataset



32D, 2000 points

总结

- 提出了一种新的高维数据二维映射方法，以揭示潜在的低维结构

- 提出了一个对高维数据中潜在低维结构进行预审查的方法

正在进行的工作

- 大规模高维数据可视分析
- 面向领域问题的高维数据可视分析

谢谢！

夏 佳 志 中南大学

邮 箱： xiajiazhi@csu.edu.cn

个人 主页： http://faculty.csu.edu.cn/xiajiazhi