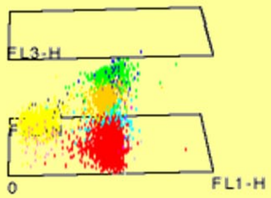
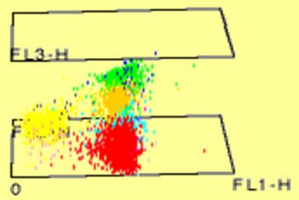
RobustMap: A Fast and Robust Algorithm for Dimension Reduction and Clustering

Lionel F. Lovett, II
Jackson State University



Research Alliance in Math and Science
Computer Science and Mathematics Division
Mentors: George Ostrouchov and Houssain Kettani
http://www.csm.ornl.gov/Internships/Websites05/l_lovett/abstract.html



Abstract

Databases can be very large due to the number of items and due to the number of attributes (high-dimensionality) associated with each item. Clustering reduces the number of items to their representative clusters and dimension reduction reduces the number of attributes. In addition, visualization of high-dimensional data requires reduction to lower-dimensional views that are often displayed as two or three dimensional plots. Traditional dimension reduction algorithms such as the singular value decomposition based principal components are computationally demanding and can be very slow. As the size of databases continues to grow, so does the demand for faster methods to visualize the data. RobustMap is a new, fast and robust dimension reduction method for high-dimensional datasets, which can separate outlying clusters from the main body of the data while computing a low-dimensional representation. It relies on stochastic concepts and on statistical distance distributions. The algorithm considers distance distributions from random and from extreme points to determine projection axes and clusters for dimension reduction. In determining the clusters, RobustMap focuses on the largest cluster, excluding outlying clusters. The visualization applications of this algorithm may be implemented in a range of disciplines, which include: medical databases, images, time series, music, and data mining.

Project Goals/Tasks

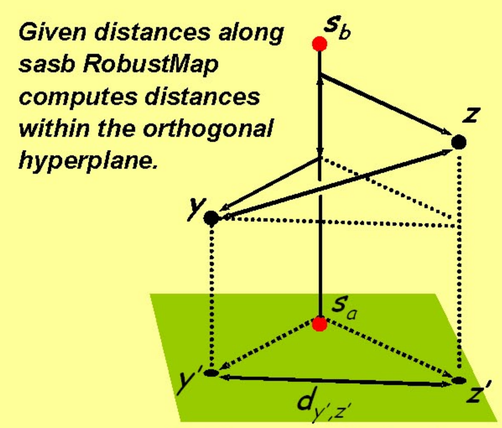
The focus of this project was to develop a fast and robust dimension reduction method for large, high-dimensional data sets. The project improves on an algorithm that is set up for indexing, data-mining and creating visualizations of traditional and multimedia datasets.

Dimension Reduction

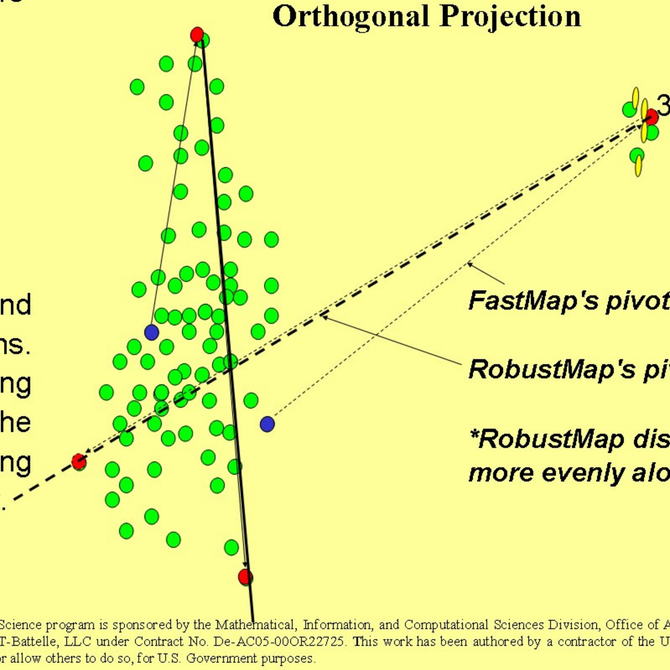
- Visualizations require low dimensional views 2-D or 3-D
- Fast similarity searching requires fast algorithms to compute for traditional and multimedia databases

RobustMap Algorithm

RobustMap is based on FastMap and statistical properties of distance distributions. FastMap is too sensitive to outliers. Using robust statistics, RobustMap extracts the largest cluster from a dataset, while identifying outlying clusters and reducing dimensionality.



Orthogonal Projection



RobustMap's Processes

1. Compute *n* distances from first object
   - Take point of largest distance
   - Repeat
2. Plots and Clusters
   - Create diagnostic histograms for distances
   - Estimate probability density of distances
   - Find ratio of actual to expected distances
   - Exclude high ratio objects as outlying clusters
3. Finish projection using only remaining objects

Results

- RobustMap correctly extracts largest cluster
- RobustMap performs dimension reduction
- RobustMap maps data to k dimensions in O(nk) time
- RobustMap exploits robust statistics

Applications

- Similarity searching in string databases
- Medical databases, where 1-D objects (e.g., ECGs), 2-D images (e.g., X-rays) and 3-d images (e.g., MRI brain scans) are stored
- Time series, Data Mining, and Visualizations

Future Research

- Compute ratio threshold based on probability and data
- Create loop to reduce dimensionality for remaining clusters
- Develop additional theory for RobustMap

Office of Science U.S. DEPARTMENT OF ENERGY

The Research Alliance in Math and Science program is sponsored by the Mathematical, Information, and Computational Sciences Division, Office of Advanced Scientific Computing Research, U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725. This work has been authored by a contractor of the U.S. Government, accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

UT-BATTELLE
OAK RIDGE NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY