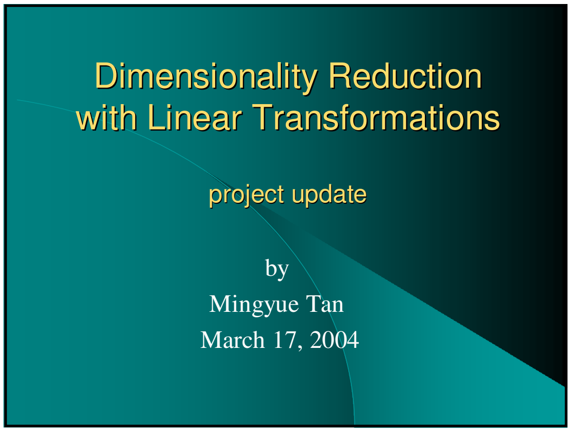


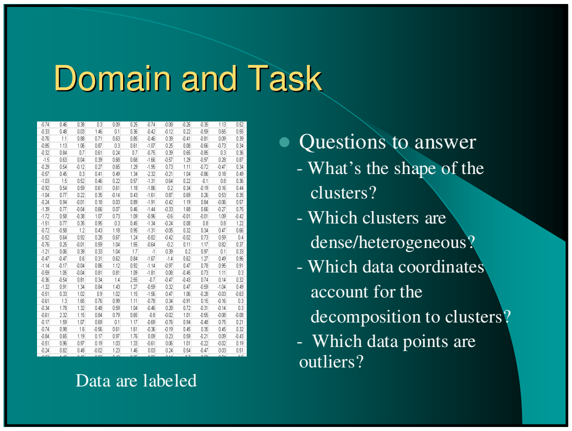
Dimensionality Reduction with Linear Transformations

project update

by

Mingyue Tan

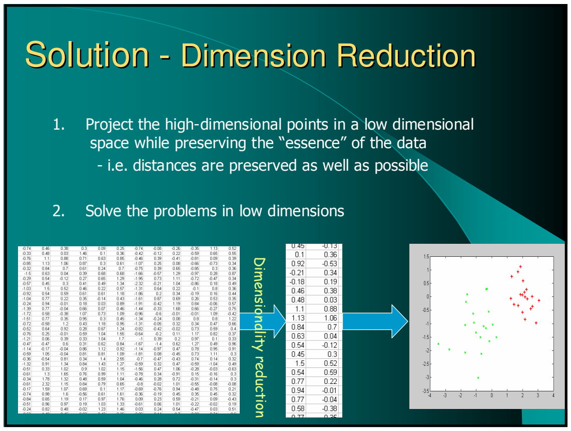
March 17, 2004



Domain and Task

Data are labeled

- Questions to answer
  - What's the shape of the clusters?
  - Which clusters are dense/heterogeneous?
  - Which data coordinates account for the decomposition to clusters?
  - Which data points are outliers?

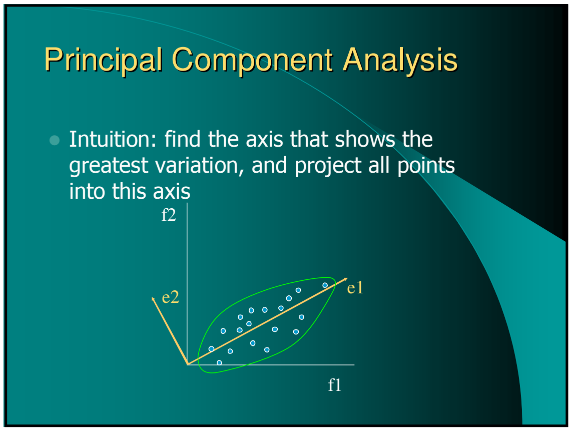


Solution - Dimension Reduction

1. Project the high-dimensional points in a low dimensional space while preserving the "essence" of the data
   - i.e. distances are preserved as well as possible
2. Solve the problems in low dimensions

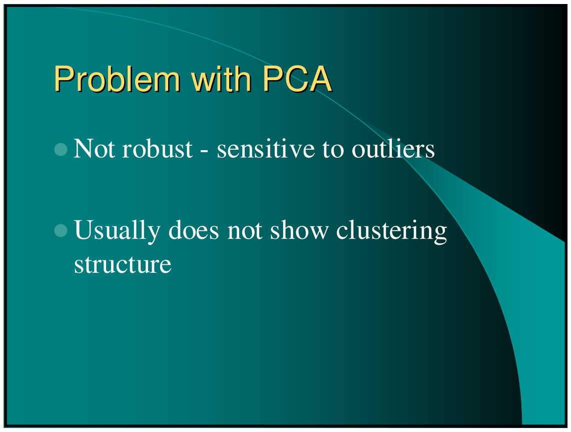
Dimensionality reduction

0.45	-0.13
0.1	0.36
0.92	-0.53
-0.21	0.34
-0.18	0.19
0.46	0.38
0.48	0.03
1.1	0.88
1.13	1.06
0.84	0.7
0.63	0.04
0.54	-0.12
0.45	0.3
1.5	0.52
0.54	0.59
0.77	0.22
0.94	-0.01
0.77	-0.04
0.58	-0.38
0.77	-0.35



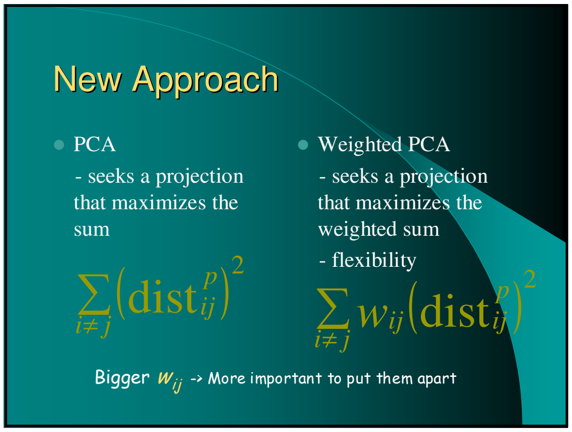
Principal Component Analysis

- Intuition: find the axis that shows the greatest variation, and project all points into this axis



Problem with PCA

- Not robust - sensitive to outliers
- Usually does not show clustering structure



New Approach

- PCA
  - seeks a projection that maximizes the sum

$$\sum_{i\neq j}\left(\mathrm{dist}_{ij}^{p}\right)^2$$

- Weighted PCA
  - seeks a projection that maximizes the weighted sum
  - flexibility

$$\sum_{i\neq j} w_{ij}\left(\mathrm{dist}_{ij}^{p}\right)^2$$

Bigger w_{ij} -> More important to put them apart

Weighted PCA

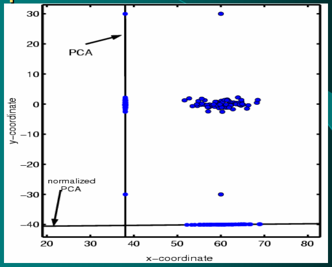
Varying w_{ij} gives:

- Weights specified by user
- Normalized PCA – robust towards outliers

$$\sum_{i\neq j} w_{ij}\left(\mathrm{dist}_{ij}^{p}\right)^{2} \qquad w_{ij}=\frac{1}{\mathrm{dist}_{ij}}$$

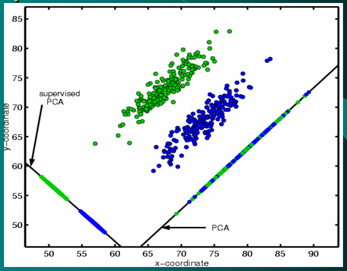
- Supervised PCA – shows cluster structures
  - If i and j belong to the same cluster ➔ set $w_{ij}=0$
  - *Maximize inter-cluster scatter*

Comparison – with outliers



- PCA: Outliers typically govern the projection direction

Comparison – cluster structure

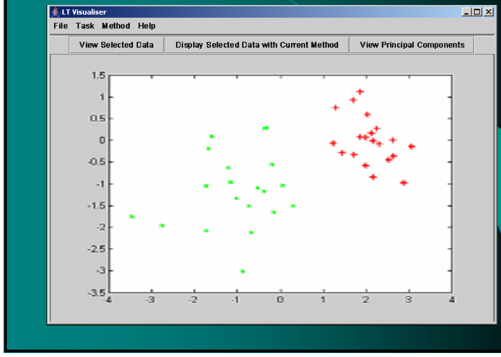


- Projections that maximize scatter ≠ Projections that separate clusters

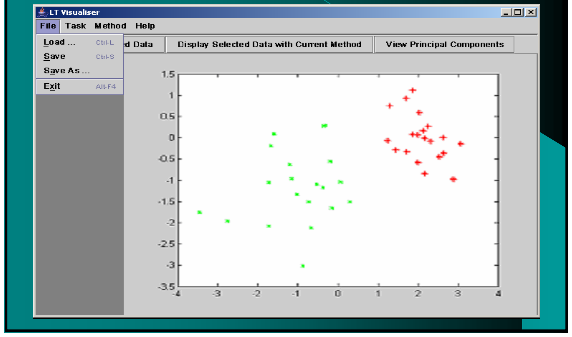
Summary

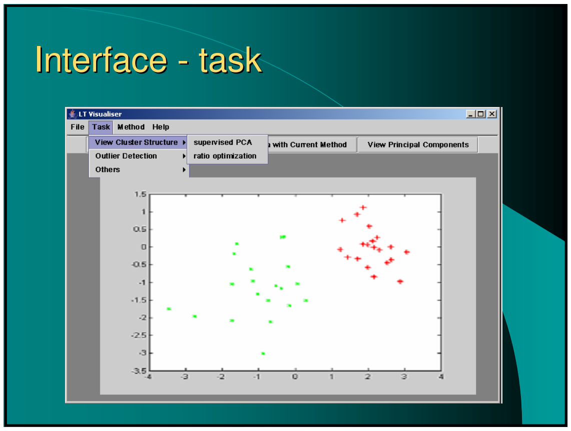
Method	Tasks
Naïve PCA	Outlier Detection
Weights-specified PCA	General view
Normalized PCA	Robustness towards Outliers
Supervised PCA	Cluster structure
Ratio optimization	Cluster structure (flexibility)

Interface

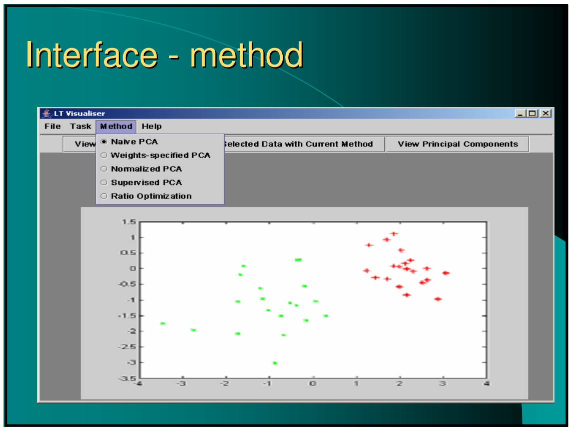


Interface - File

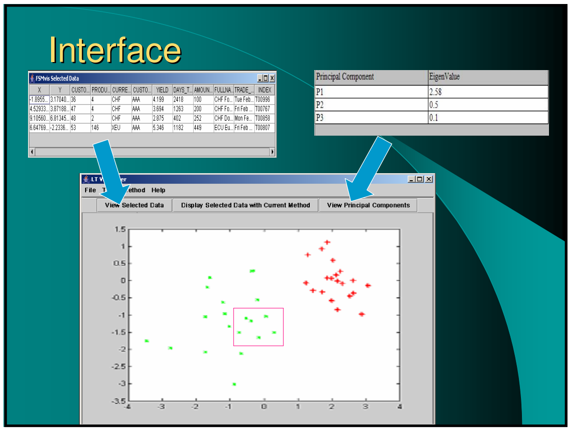




Interface - task



Interface - method



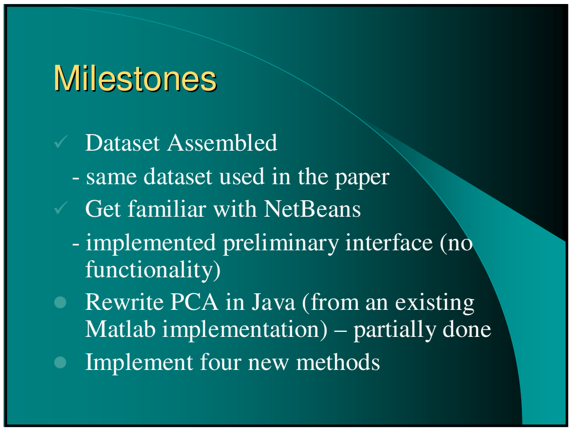
Interface

FSMvis Selected Data

X	Y	CUSTO...	PRODU...	CURRE...	CUSTO...	YIELD	DAYS_T...	AMOUN...	FULLNA...	TRADE_...	INDEX
-1.8955...	3.17040...	36	4	CHF	AAA	4.199	2418	100	CHF Fo...	Tue Feb...	T00996
4.52933...	3.87188...	47	4	CHF	AAA	3.694	1263	200	CHF Fo...	Fri Feb ...	T00767
9.10560...	6.81345...	48	2	CHF	AAA	2.875	402	252	CHF Do...	Mon Fe...	T00858
6.64769...	-2.2336...	53	146	XEU	AAA	5.346	1182	449	ECU Eu...	Fri Feb ...	T00807

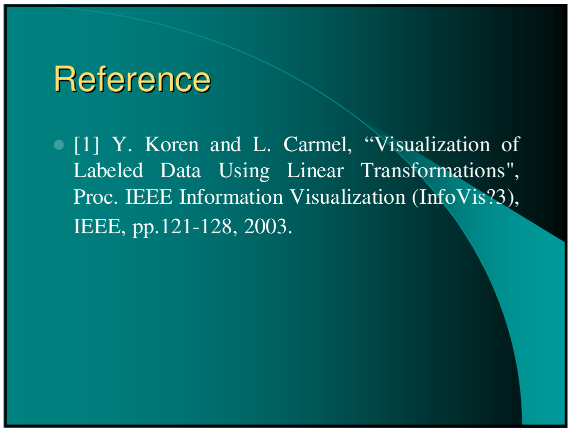
Principal Component	EigenValue
P1	2.58
P2	0.5
P3	0.1

LT V...er
File T... ethod Help
View Selected Data
Display Selected Data with Current Method
View Principal Components
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-3
-2
-1
0
1
2
3
4



Milestones

- ✓ Dataset Assembled
  - same dataset used in the paper
- ✓ Get familiar with NetBeans
  - implemented preliminary interface (no functionality)
- Rewrite PCA in Java (from an existing Matlab implementation) – partially done
- Implement four new methods



Reference

- [1] Y. Koren and L. Carmel, “Visualization of Labeled Data Using Linear Transformations", Proc. IEEE Information Visualization (InfoVis?3), IEEE, pp.121-128, 2003.