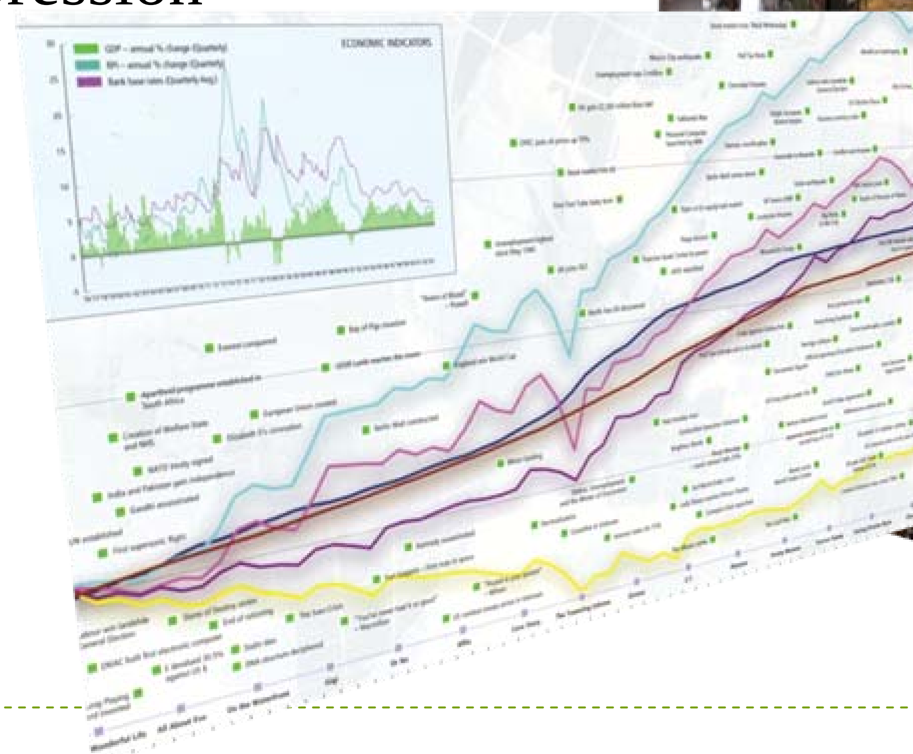
Unsupervised Feature Selection for Multi-Cluster Data

Deng Cai, Chiyuan Zhang, Xiaofei He
Zhejiang University

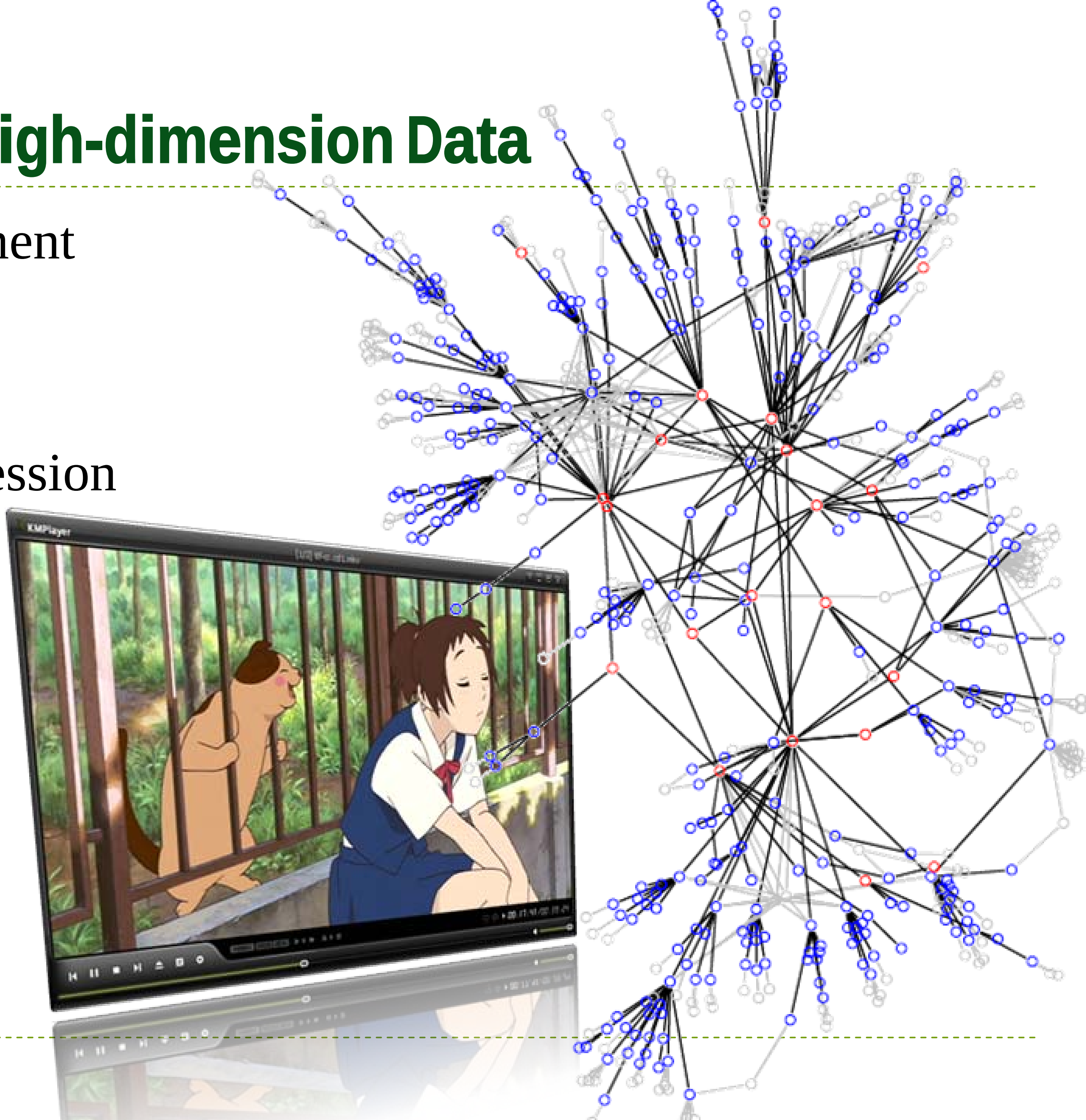
Problem: High-dimension Data

- Text document
- Image
- Video
- Gene Expression
- Financial
- Sensor
- …

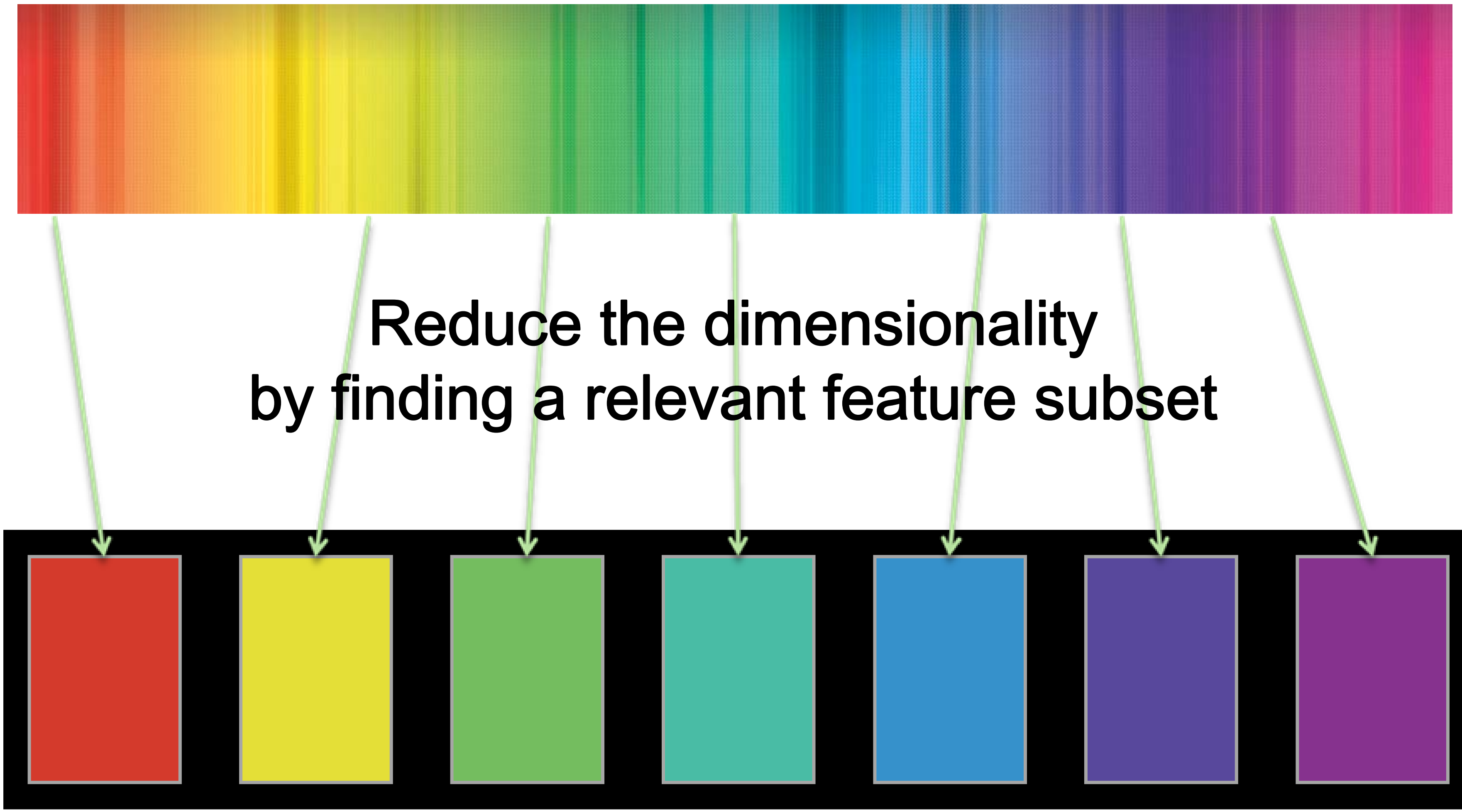


Problem: High-dimension Data

- Text document
- Image
- Video
- Gene Expression
- Financial
- Sensor
- …



Solution: Feature Selection



Feature Selection Techniques

- Supervised
  - Fisher score
  - Information gain
- **Unsupervised (discussed here)**
  - Max variance
  - Laplacian Score, NIPS 2005
  - Q-alpha, JMLR 2005
  - MCFS, KDD 2010 (Our Algorithm)
  - …

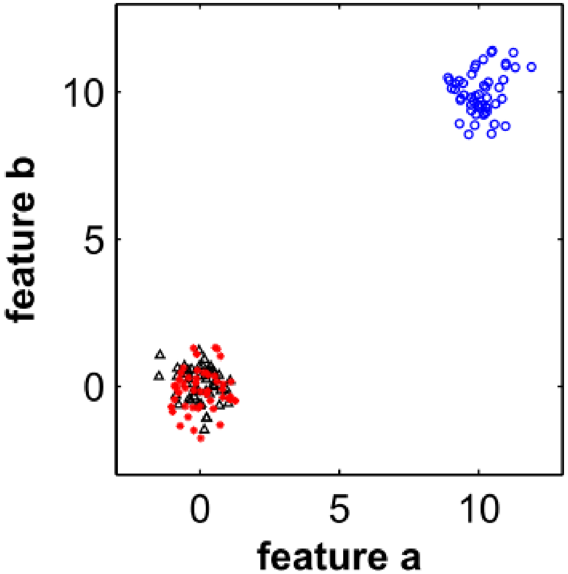
Outline

- Problem setting
- *Multi-Cluster Feature Selection* (MCFS) Algorithm
- Experimental Validation
- Conclusion

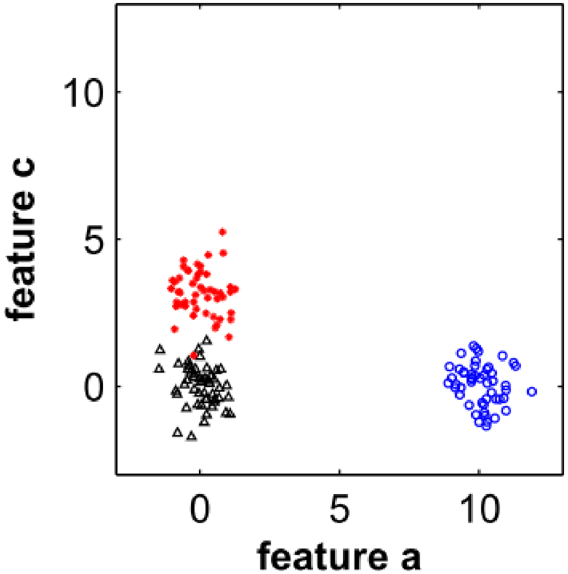
Problem setting

- Unsupervised Multi clusters/classes Feature Selection

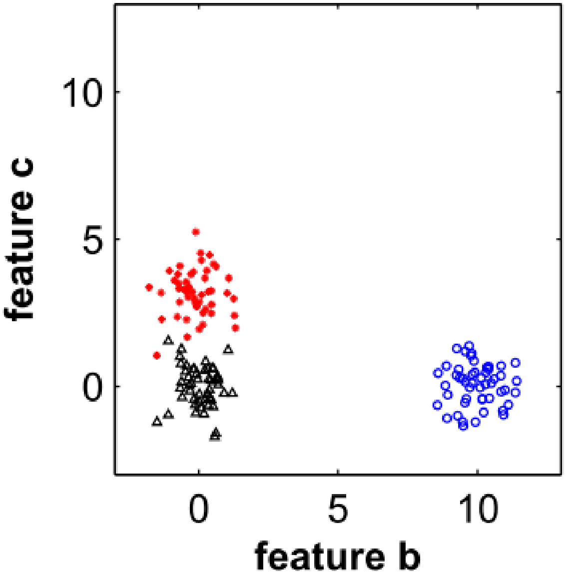
- How traditional score-ranking methods fail:



(a) plane $a \otimes b$



(b) plane $a \otimes c$



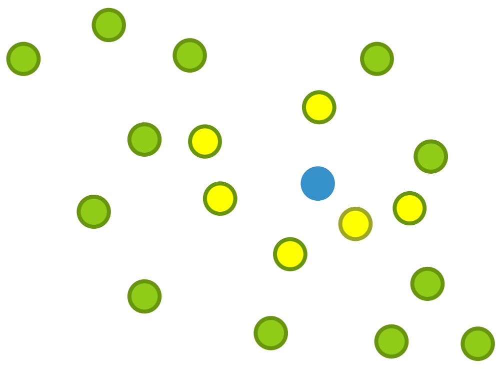
(c) plane $b \otimes c$

Multi-Cluster Feature Selection (MCFS) Algorithm

- Objective
  - Select those features such that the multi-cluster structure of the data can be well preserved
- Implementation
  - Spectral analysis to explorer the intrinsic structure
  - L1-regularized least-square to select best features

Spectral Embedding for Cluster Analysis

- Laplacian Eigenmaps
  1. Weight matrix W
     - 0-1 weighting
     - Heat kernel weighting
     - Cosine weighting
  2. Graph Laplacian
     - $L = D - W$ where $D_{ii} = \sum_j W_{ij}$
  3. Generalized Eigen-problem
     - $Ly = \lambda Dy$



Spectral Embedding for Cluster Analysis

- Laplacian Eigenmaps
  - Can *unfold* the data manifold and provide the *flat* embedding for data points
  - Can reflect the data distribution on each of the data clusters
  - Thoroughly studied and well understood

Learning Sparse Coefficient Vectors

- LASSO Regression
  - $\min_{a_k} \|y_k - X^T a_k\|^2 + \beta |a_k|$
  - a_k contains the combination coefficients for different features in approximating y_k
  - With L1-norm regularization, some coefficients will be shrunk to zero if β is large enough
  - LARs can be used to solve the problem efficiently
    - and conveniently (explicitly control the sparsity)
  - Solved the problem of feature correlation & combination

Feature Selection on Sparse Coefficient Vectors

- Select d features from M feature candidates
- Obtain K sparse coefficient vector $\{a_k\}_{k=1}^{K}$, each of cardinality d
- Assign a MCFS score for each feature as
  - $MCFS(j) = \max_k |a_{k,j}|$
- Select the d features with top MCFS scores

Algorithm Summary

1. Construct p-nearest neighbor graph W
2. Solve generalized eigen-problem to get K eigenvectors corresponding to the smallest eigenvalues
3. Solve K L1-regulairzed regression to get K sparse coefficient vectors
4. Compute the MCFS score for each feature
5. Select d features according to MCFS score

Complexity Analysis

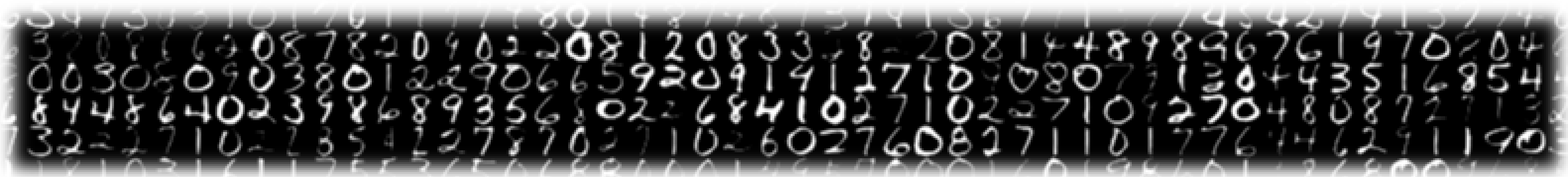
1. Graph construction
   - $O(N^2M)$ to compute pairwise distance
   - $O(N^2p)$ to find p neighbors for each data point
2. Lanczos algorithm for eigen-problem
   - $O(KNp)$
3. LARs for LASSO solving
   - $O(Kd^3 + NKd^2)$
4. MCFS score computation
   - $O(KM)$
5. Feature selection
   - $O(M \log M)$

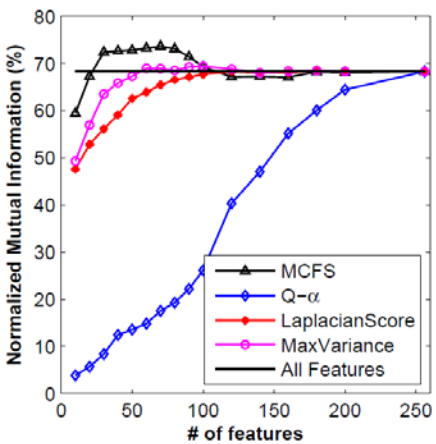
Experiments

- Unsupervised feature selection for
  - Clustering
  - Nearest neighbor classification
- Compared algorithms
  - MCFS
  - Q-alpha
  - Laplacian score
  - Maximum variance

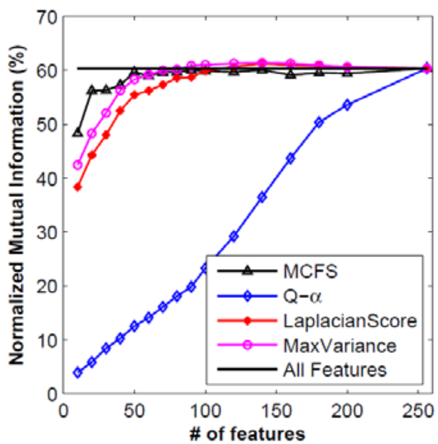
Experiments (USPS Clustering)

- USPS Hand Written Digits
  - 9298 samples, 10 classes, 16x16 gray-scale image each

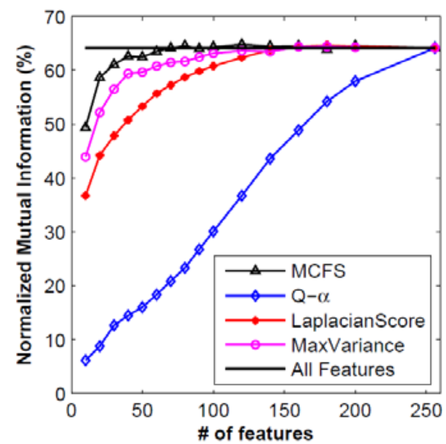




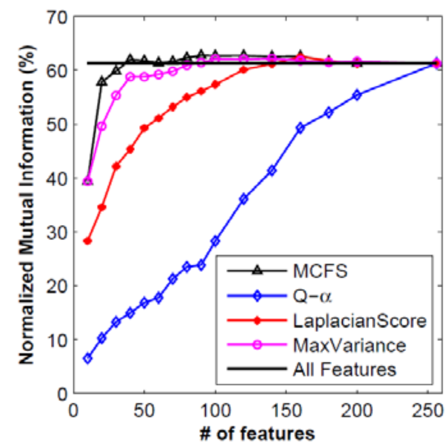
(a) 3 Clusters



(b) 5 Clusters



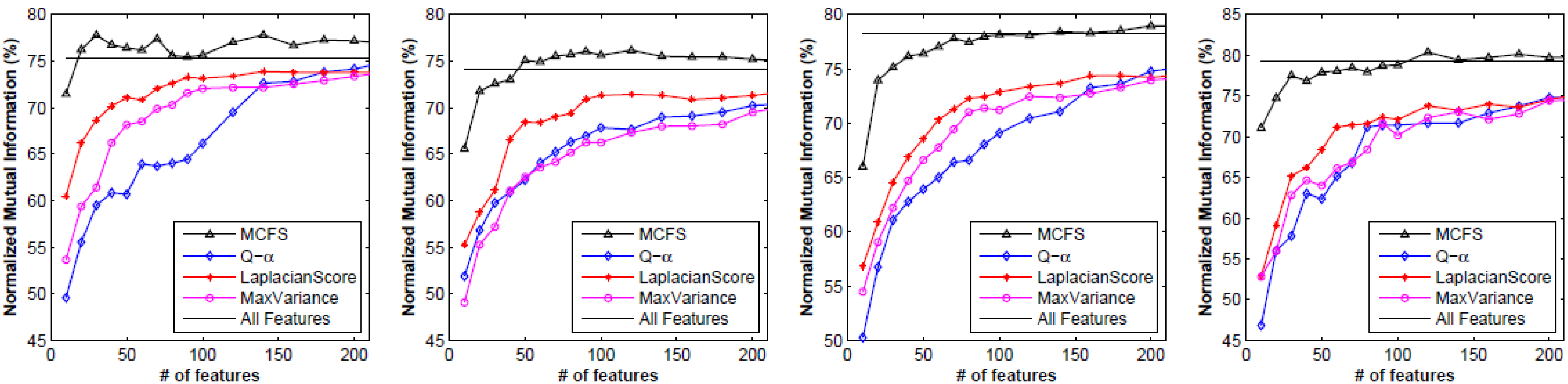
(c) 7 Clusters



(d) 10 Clusters

Experiments (COIL20 Clustering)

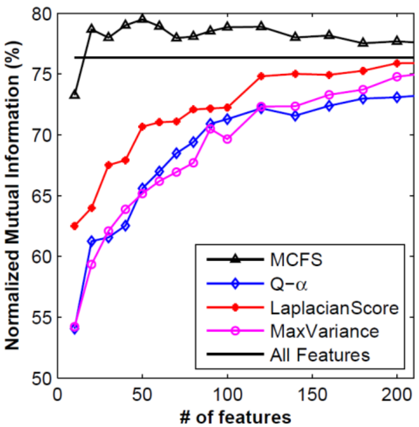
- COIL20 image dataset
  - 1440 samples, 20 classes, 32x32 gray-scale image each



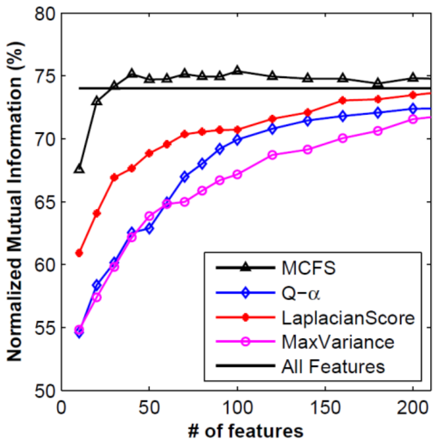
(a) 5 Clusters (b) 10 Clusters (c) 15 Clusters (d) 20 Clusters

Experiments (ORL Clustering)

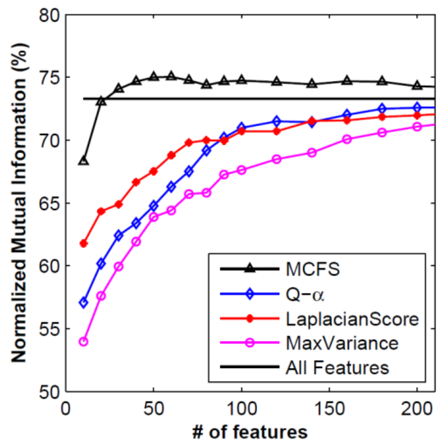
- ORL face dataset
  - 400 images of 40 subjects
  - 32x32 gray-scale images



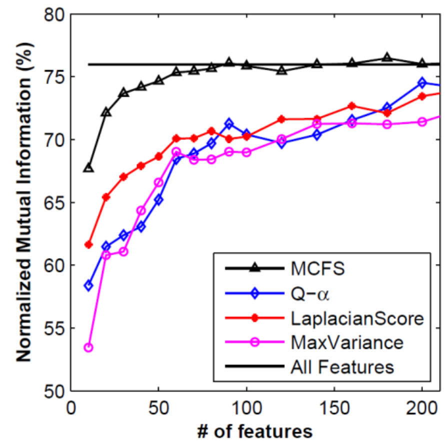
10 Classes



20 Classes



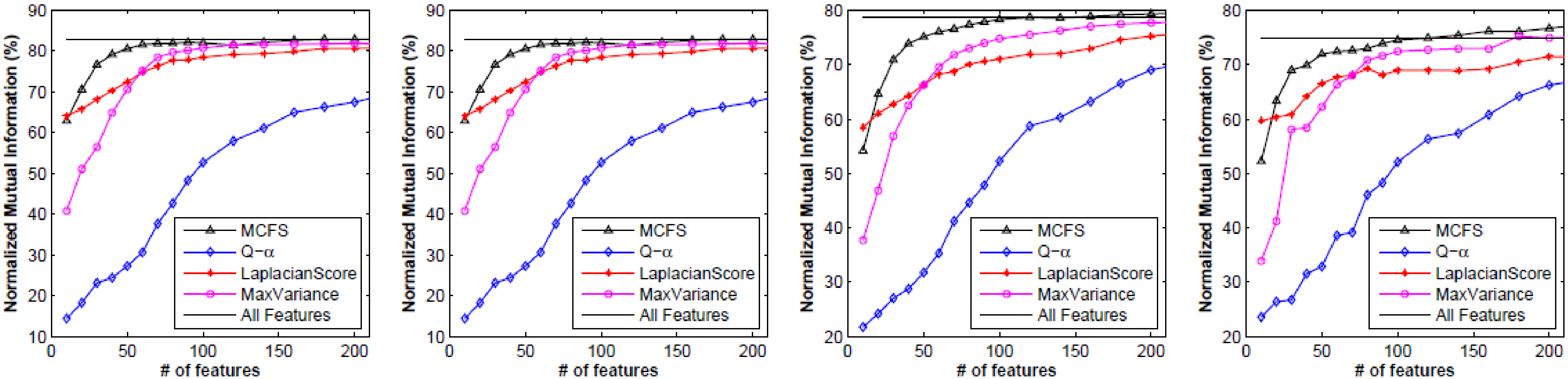
30 Classes



40 Classes

Experiments (Isolet Clustering)

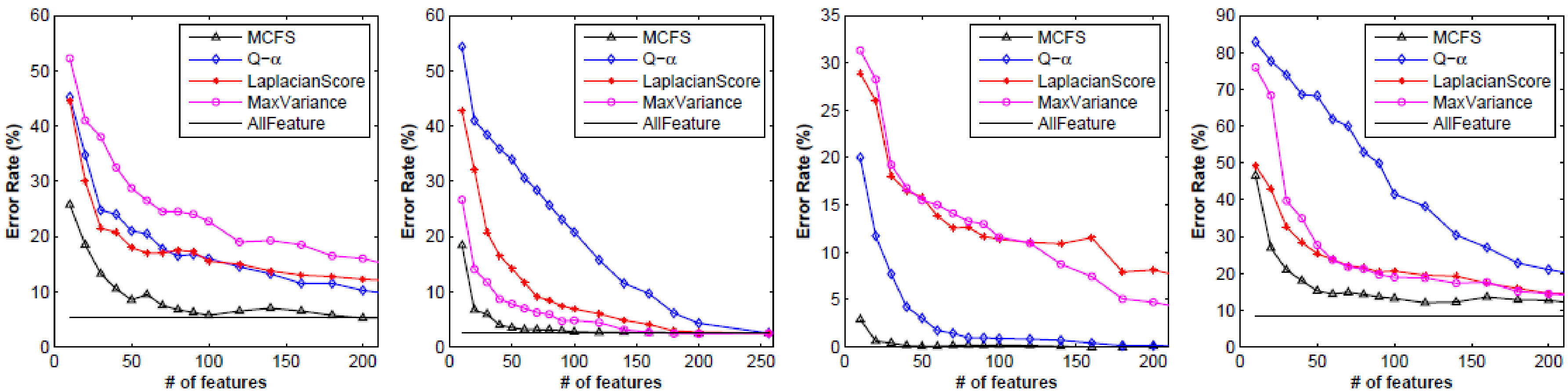
- Isolet spoken letter recognition data
  - 1560 samples, 26 classes
  - 617 features each sample



(a) 10 Clusters (b) 15 Clusters (c) 20 Clusters (d) 26 Clusters

Experiments (Nearest Neighbor Classification)

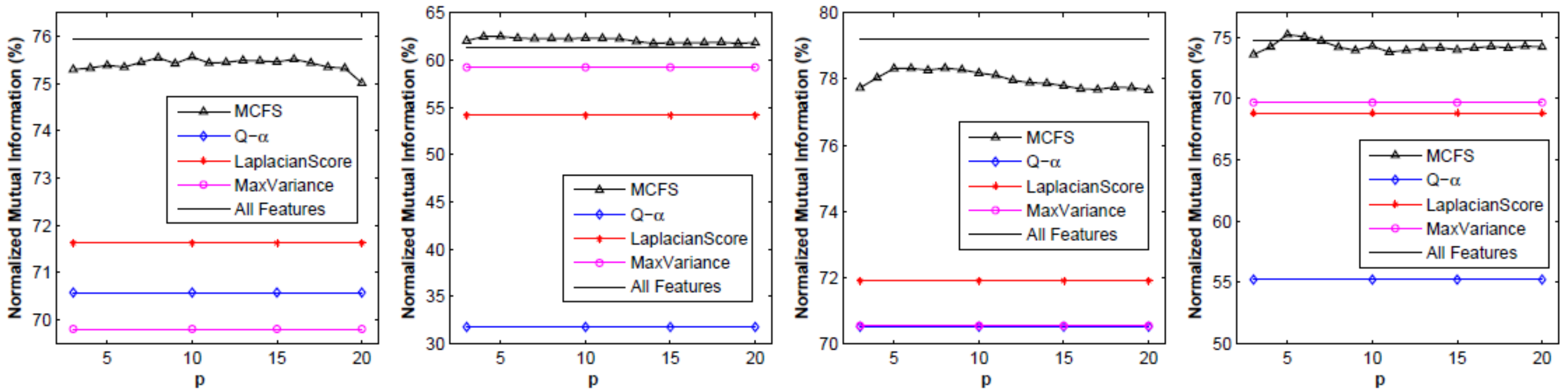
- Leave-one-out cross validation
- Measured by error rate
  - $ER = 1 - \frac{1}{N}\sum_{i=1}^{N}\delta(c(x_i), c(x_i'))$



(a) ORL (b) USPS (c) COIL20 (d) Isolet

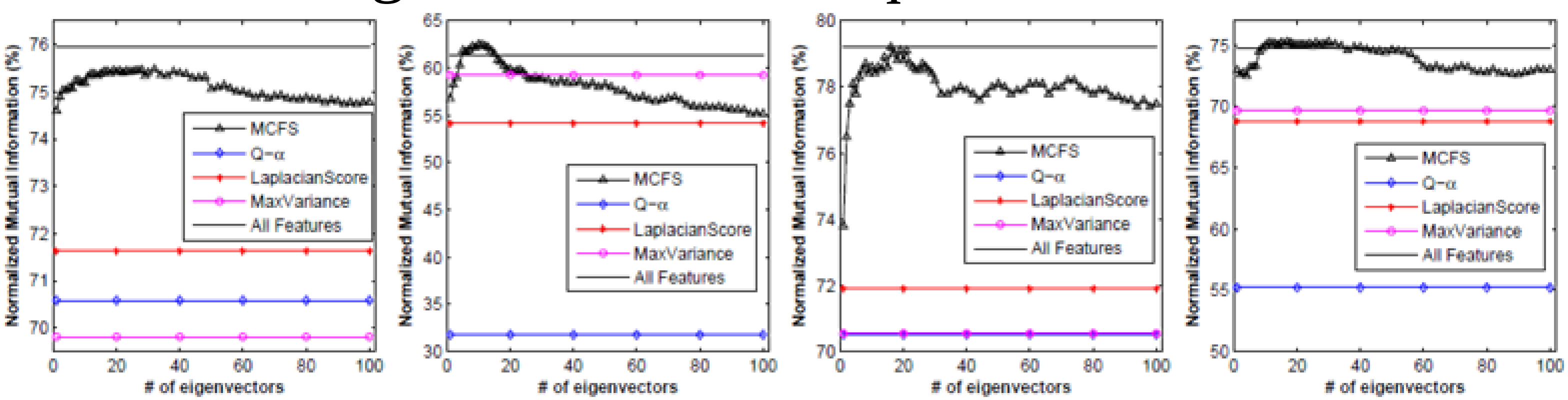
Experiments (Parameter Selection)

- Number of nearest neighbors p: stable



(a) ORL (b) USPS (c) COIL20 (d) Isolet

- Number of eigenvectors: best equal to number of classes



(a) ORL (b) USPS (c) COIL20 (d) Isolet

Conclusion

- MCFS
  - Well handle multi-class data
  - Outperform state-of-art algorithms
  - Performs especially well when number of selected features is small (< 50)

Questions

?