Unsupervised and supervised dimension reduction: Algorithms and connections

Jieping Ye
Department of Computer Science and Engineering

Evolutionary Functional Genomics Center
The Biodesign Institute

Arizona State University

Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component Analysis (PCA)
  - Issues and extensions
- Linear Discriminant Analysis (LDA)
  - Issues and extensions
- Summary

Overview of my research

- Protein structure analysis
  - Pairwise and multiple structure alignment
- Gene expression pattern analysis (joint work with Sudhir Kumar's group)
  - Dimension reduction
  - Clustering
  - Biclustering
  - Classification
  - Semi-supervised learning

Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component
  - Issues and extensions
- Linear Discriminant Analysis
  - Issues and extensions
- Summary

Gene expression pattern analysis

- *In Situ* staining of a target mRNA at several time points during the development of a D. melanogaster embryo gives a detailed spatial-temporal view of the expression pattern of a given gene.
  - Capture spatial gene interactions based on computational analysis of images

- Microarray gene expression data reveals only the average expression levels.
  - Fail to capture any pivotal spatial patterns

Challenges in gene expression pattern analysis

- High dimensionality (312*120 pixels)
- Noise (images taken under different conditions)
- Large database (about 50000 images)
  - Growing rapidly
- Natural solution
  - Apply dimension reduction as a preprocessing step

Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component
  - Issues and extensions
- Linear Discriminant Analysis
  - Issues and extensions
- Summary

What is dimension reduction?

- Embed the original high-dimensional data in a lower-dimensional space.
  - Critical information should be preserved.

- Motivation
  - Curse of dimensionality
  - Intrinsic dimensionality
  - Visualization
  - Noise removal

What is dimension reduction?

- Algorithms
  - Unsupervised versus supervised
  - Linear versus nonlinear
  - Local versus global
- Many other applications
  - Microarray data analysis
  - Protein classification
  - Face recognition
  - Text mining
  - Image retrieval

Unsupervised versus supervised dimension reduction

- Unsupervised
  - Principal Component Analysis
  - Independent Component Analysis
  - Canonical Correlation Analysis
  - Partial Least Square
- Supervised
  - Linear Discriminant Analysis

PCA and LDA

- Well known for dimension reduction
- Face recognition
  - Eigenface versus Fisherface
- Most unsupervised dimension algorithms are closely related to PCA
- The criterion used in LDA is shared with many other clustering and classification algorithms

Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component Analysis
  - Issues and extensions
- Linear Discriminant Analysis
  - Issues and extensions
- Summary

Principal Component Analysis (PCA)

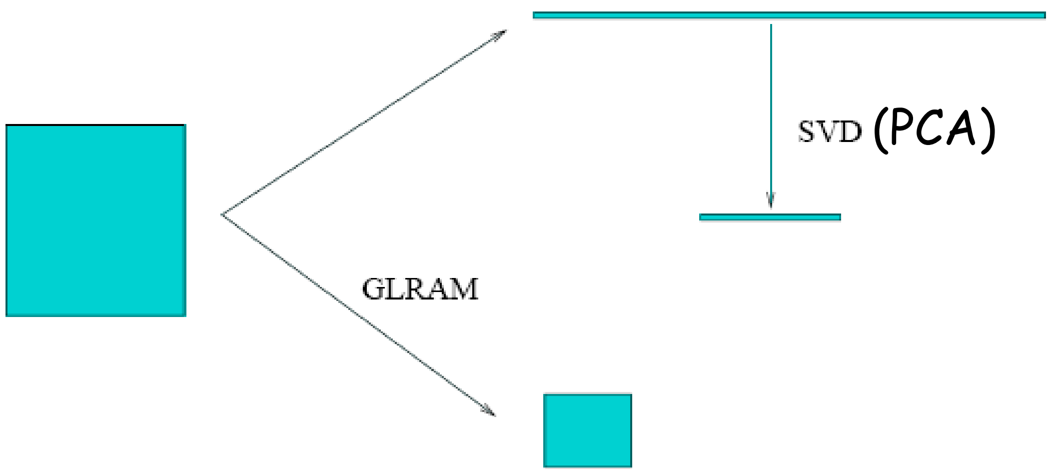
- Keep the largest variance
  - Capture global structure of the data
- Computed via Singular Value Decomposition (SVD)
- Achieve the minimum reconstruction error
- Applied as a preprocessing step for many other algorithms in machine learning and data mining.

Issues in PCA

- PCA does not scale to large databases
  - Solution: GLRAM
- PCA does not capture local structure
  - Solution: Local PCA

GLRAM for large databases

- GLRAM extends PCA to 2D data.



- GLRAM scales to large databases.
  - Time: Linear on both the sample size and the data dimension
  - Space: Independent of the sample size

Local PCA

- Find clusters in the data
- Compute the transformation matrix considering the cluster structure.
  - Consider the local information
- Computation: Borrow techniques from GLRAM.
  - Low rank approximations on covariance matrices of all clusters.

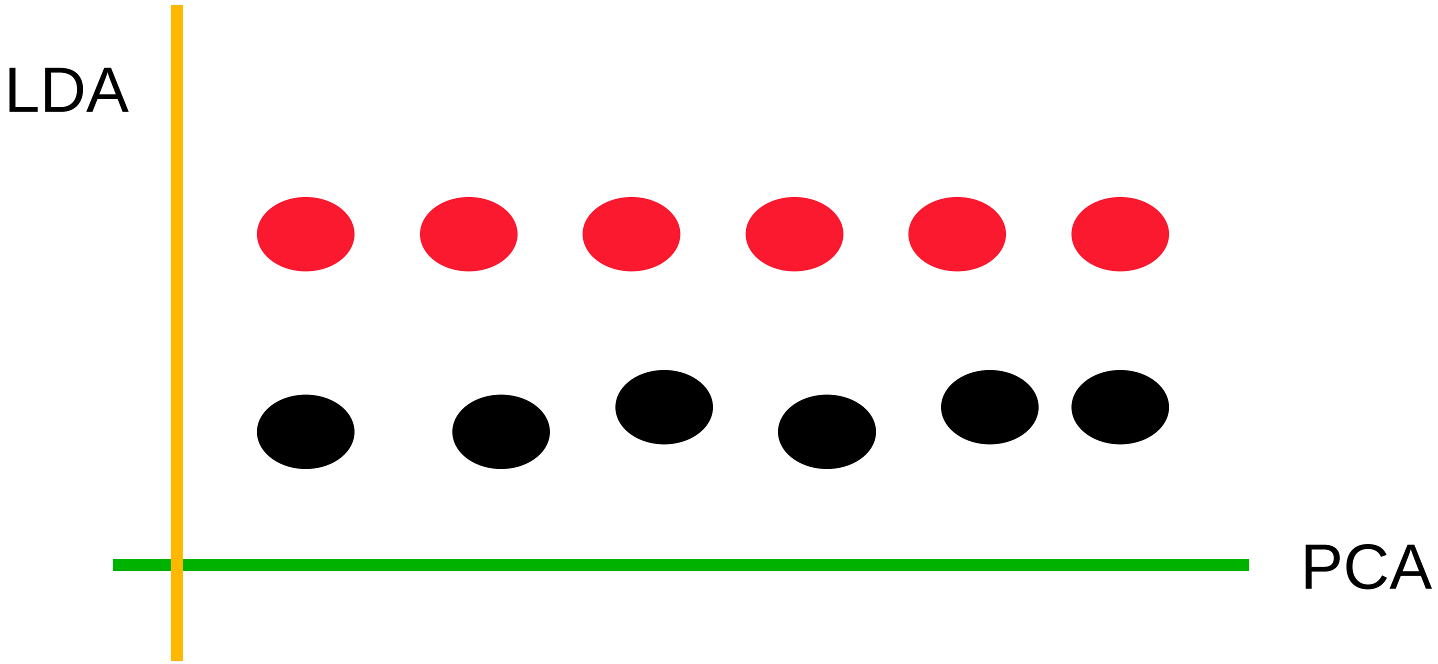
Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component Analysis
  - Issues and extensions
- Linear Discriminant Analysis
  - Issues and extensions
- Summary

Linear Discriminant Analysis (LDA)

- Find the projection with maximum discrimination
  - Effective for classification
- Computed by solving a generalized eigenvalue problem
- Optimal when each class is Gaussian and has the same covariance matrix

LDA versus PCA



PCA only considers the global structure of the data, while LDA utilizes the class information (maximum separation).

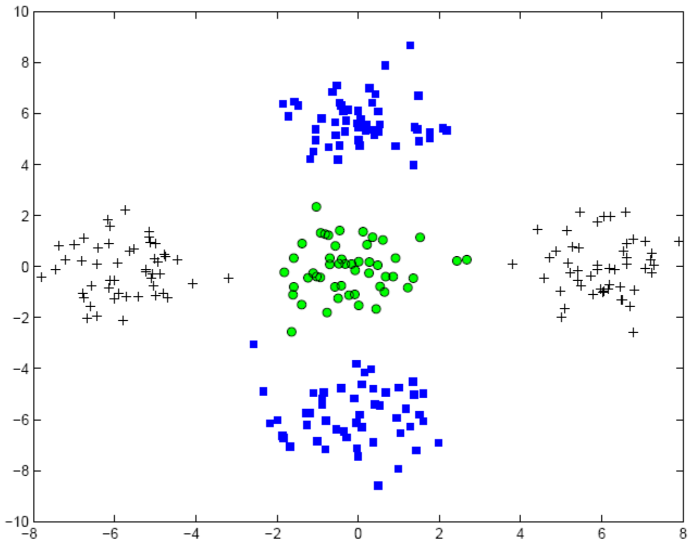
Issues in LDA

- Not effective when the assumptions are violated
  - The class centroids coincide
  - Class covariances vary
- Singularity or undersampled problem
  - Data dimension is larger than sample size

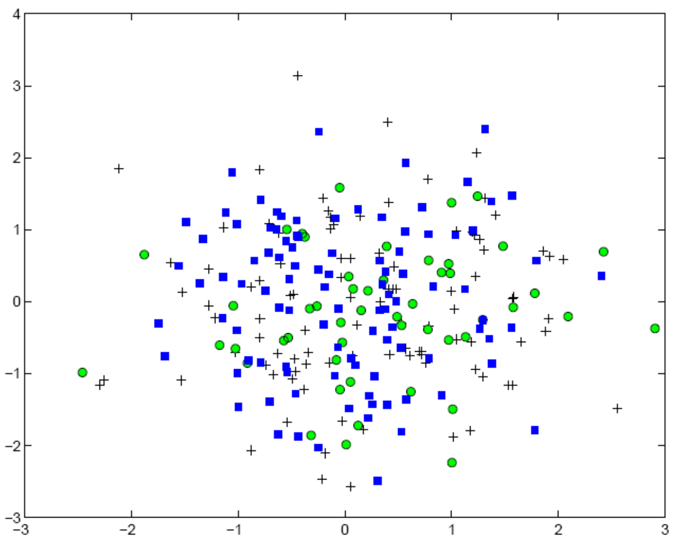
Covariance-preserving Projection

- CPM considers the class covariance information
  - CPM preserves the class covariance information, while maximizing class discrimination

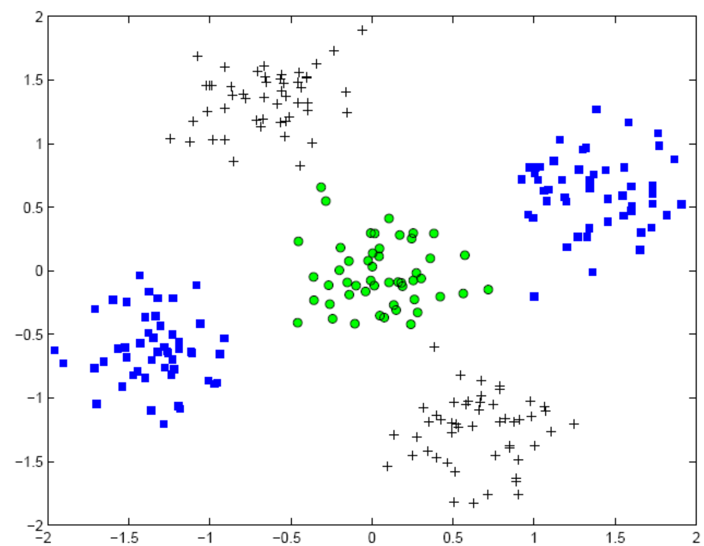
- Applicable when class centroids coincide or class covariances vary


10
8
6
4
2
0
−2
−4
−6
−8
−10
−8 −6 −4 −2 0 2 4 6 8

Data


4
3
2
1
0
−1
−2
−3
−3 −2 −1 0 1 2 3

LDA


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2 −1.5 −1 −0.5 0 0.5 1 1.5 2

CPM

Other related work

- Sliced Average Variance Estimator (SAVE). Cook and et al.
- Heteroscedastic Discriminant Analysis (HDA). Kumar and et al.
- Another method by Zhu and Hastie
- Theoretical result
  - CPM is an approximation of HDA
    - Efficiency and robustness
  - SAVE is a special case of CPM

Singularity problem

- PCA+LDA: Apply PCA before the LDA stage [Swet and Weng, TPAMI 1996; Belhumeour and et al., TPAMI 1997]
- RLDA: Modify the scatter matrix to make it nonsingular [Friedman, JASA, 1989; Ye and et al., 2005]
- ULDA: The features in the transformed space are uncorrelated. [Ye and et al., TCBB 2004]
- OLDA: Orthogonal transformation. [Ye. JMLR 2005]
  - More robust to noise compared with ULDA
- 2DLDA: Extend LDA to 2D data [Ye. NIPS 2005]

Efficient model selection for RLDA

- Key idea in RLDA
  - Perturb the scatter matrix to make it nonsingular
- Limitation
  - How to choose the best regularization parameter (amount of perturbation) efficiently?
- Proposed approach
  - Compute RLDA with a large number of parameters with approximately the same time as running RLDA a small number of times.

m	Doc1	Doc2	GCM	ALL	ORL	PIX
1	6.68	19.14	16.04	20.85	39.95	22.66
2	6.68	19.19	16.10	22.23	39.98	22.57
4	6.77	19.23	16.15	22.11	40.42	22.67
8	6.86	19.32	16.33	22.34	40.75	23.15
16	6.96	19.53	16.43	22.48	41.84	23.72
32	7.13	20.35	16.46	22.92	44.15	24.38
64	7.32	21.47	17.18	23.31	48.25	26.30
128	7.80	22.90	17.92	23.63	56.85	30.24
256	8.87	26.84	19.91	23.99	74.24	37.59
512	11.01	34.36	23.36	24.66	107.9	52.92
1024	15.36	49.59	30.15	28.14	176.7	81.74
T(1024)/T(1)	2.30	2.59	1.88	1.35	4.42	3.61
k/d(× 1e3)	1.39	1.33	0.87	0.48	3.88	3.00

Table 1: Computational time (in seconds) of RLDA for different $m = |\Lambda|$.

Outline of talk

- Overview of my research
- Challenges in gene expression pattern analysis
- What is dimension reduction?
  - Unsupervised versus Supervised
- Principal Component Analysis
  - Issues and extensions
- Linear Discriminant Analysis
  - Issues and extensions
- Summary

Summary

- Challenges in gene expression pattern analysis
- PCA and LDA are important techniques for dimension reduction
- Limitations of PCA and LDA
- Recent extensions of PCA and LDA

Related publication

- J. Ye. Generalized low rank approximations of matrices. ***Machine Learning,*** 2005.
- *J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems.* ***Journal of Machine Learning Research,*** *2005.*
- *J. Ye and et al. Two-dimensional Linear Discriminant Analysis.* ***NIPS,*** *2005.*
- J. Ye and et al. Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics,*** *2004.*
- J. Ye and et al. Efficient Model Selection for Regularized Linear Discriminant Analysis. TR-05-006, Dept. of CSE, ASU.
-
- J. Ye and et al. CPM: Covariance-preserving Projection Method. TR-05-007, Dept. of CSE, ASU.