Adaptive Discriminant Analysis by Minimum Squared Errors

Haesun Park
Div. of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, U.S.A.

(Joint work with Barry L. Drake and Hyunsoo Kim)

Stanford, June 2006

Adaptive Dimension Reduction for Clustered Data

- Linear Discriminant Analysis (LDA) and its Generalizations for undersampled problems, LDA/GSVD
- Extension to kernel-based nonlinear method KDA/GSVD
- Relationship to Classifier design by MSE
- Adaptive feature subspace tracking method
- Test results: Facial recognition, efficient cross-validation by downdating, etc.

Clustered Data: Facial Recognition

The 1st sample

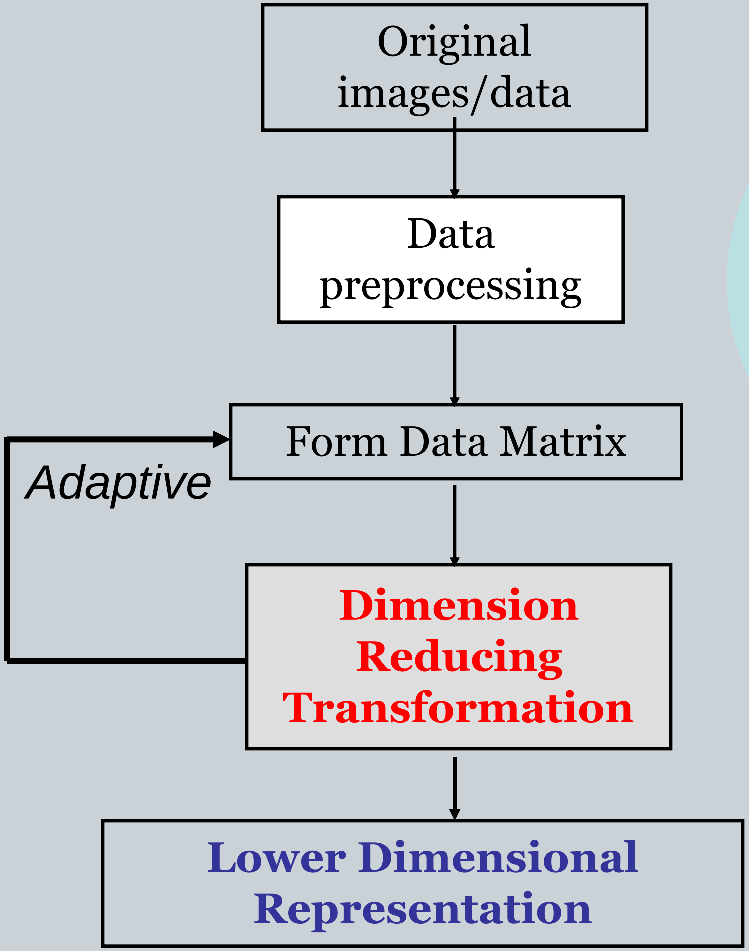
The 35th sample

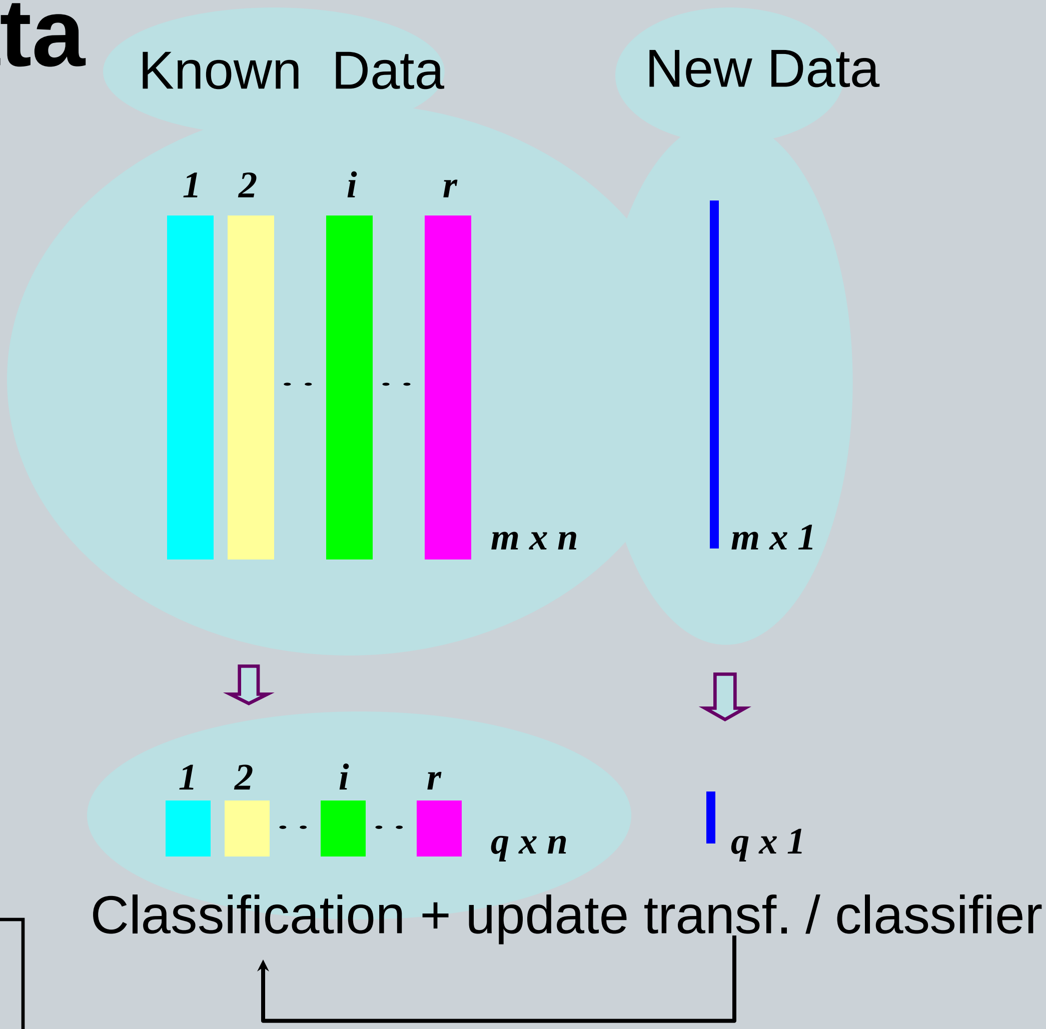
AT&T (ORL) Face Database

- 400 frontal images = 40 person x 10 images each, variations in pose, facial expression
- image size :92 x 112
- Severely Undersampled:

10304 x 400

Adaptive Dimension Reduction of Clustered Data





Want: *Adaptive Dimension Reducing Transformation* *that can be effectively applied* ***across many application areas***

Measure for Cluster Quality

$A = [a_1 \ldots a_n]$: mxn, clustered data
N_i = items in class i, $|N_i| = n_i$, total r classes
c_i = average of data items in class i, *centroid*
c = global average, *global centroid*

(1) Within-class scatter matrix

$$S_w = \sum_{1 \le i \le r} \sum_{j \in N_i} (a_j - c_i)(a_j - c_i)^T$$

(2) Between-class scatter matrix

$$S_b = \sum_{1 \le i \le r} \sum_{j \in N_i} (c_i - c)(c_i - c)^T$$

(3) Total scatter matrix

$$S_t = \sum_{1 \le i \le n} (a_i - c)(a_i - c)^T$$

NOTE: $S_w + S_b = S_t$

Trace of Scatter Matrix

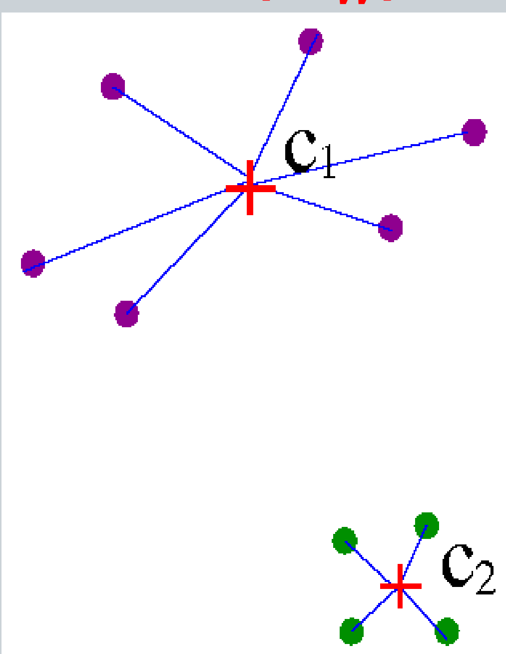
$$trace\,(S_w) = \sum_{1\le i\le r} \sum_{j\in N_i} \|a_j - c_i\|_2^2$$

$$trace\,(S_b) = \sum_{1\le i\le r} \sum_{j\in N_i} \|c_i - c\|_2^2$$

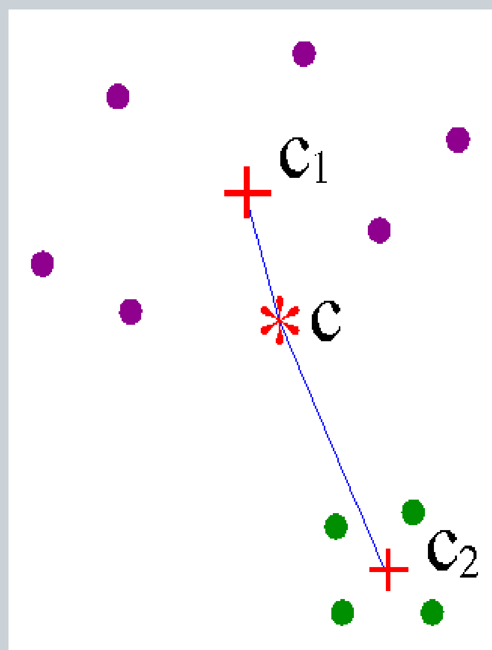
$$trace\,(S_t) = \sum_{1\le i\le r} \sum_{j\in N_i} \|a_j - c\|_2^2$$

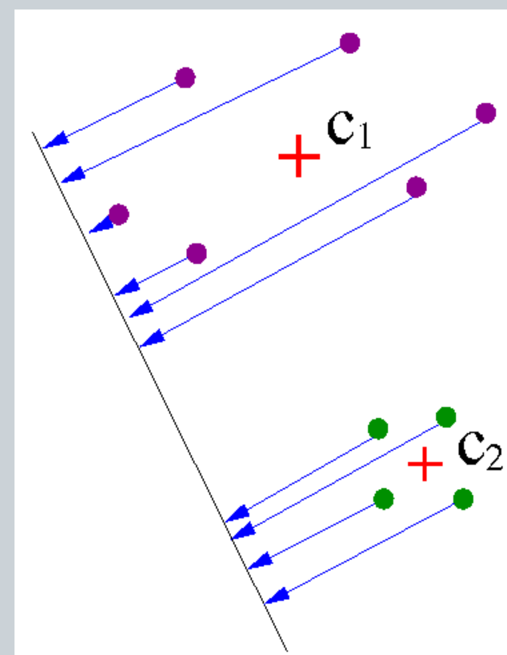
Dimension Reducing Transformation

trace (S_w)



trace (S_b)





Optimal Dimension Reducing Transformation

$$y : mx1 \xrightarrow{G^T : qxm} G^T y : qx1, \quad q << m$$

High quality clusters have

small trace(S_w) **&** ***large trace(S_b)***

Want: G

s.t. ***min trace($G^T S_w G$)*** **&** ***max trace($G^T S_b G$)***

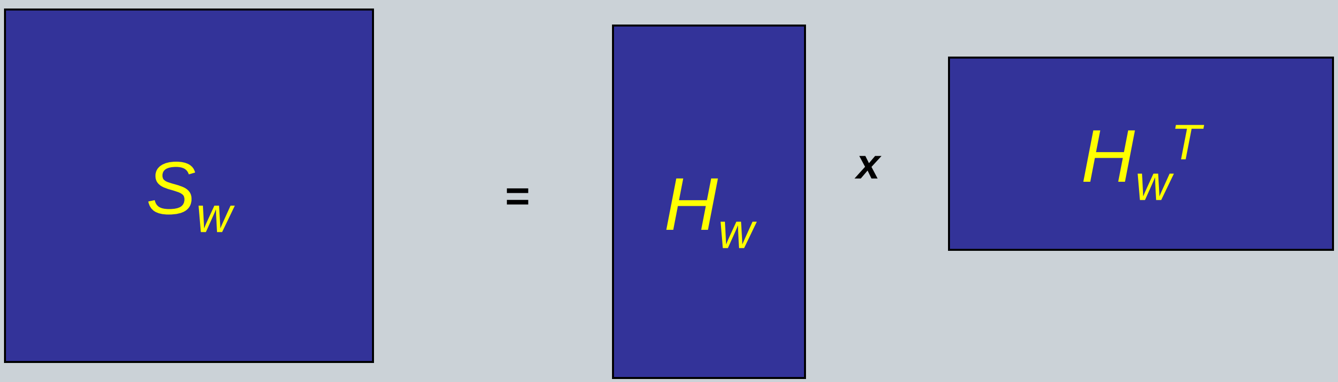
- **max trace (($G^T S_w G$)$^{-1}$ ($G^T S_b G$))** → **LDA** ***(Fisher 36, Rao 48)***
- **max trace ($G^T S_b G$)** → **Orthogonal Centroid** ***(Park et al. 03)*** $G^TG=I$
- **max trace ($G^T (S_w+S_b) G$)** → **PCA** ***(Hotelling 33)*** $G^TG=I$
- **max trace ($G^T A A^T G$)** → **LSI** ***(Deerwester et al. 90)*** $G^TG=I$

Classical LDA

(Fisher '36, Rao '48)

max trace $((G^T S_w G)^{-1} (G^T S_b G))$

- G : leading (r-1) e.vectors of $S_w^{-1} S_b$
  Fails when m>n (undersampled), S_w *singular*



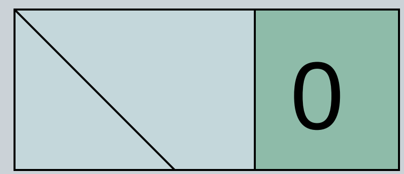
- $S_w = H_w H_w^T$, $H_w = [a_1 - c_1, a_2 - c_1, \ldots, a_n - c_r]$: $m \times n$
- $S_b = H_b H_b^T$, $H_b = [1/\sqrt{n_1}(c_1 - c), \ldots, 1/\sqrt{n_r}(c_r - c)]$: $m \times r$

LDA based on GSVD (LDA/GSVD)

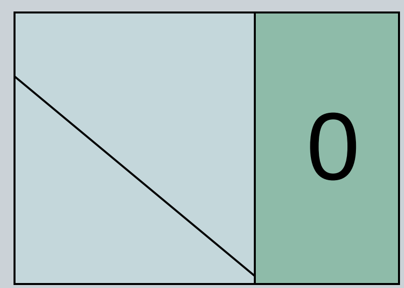
(Howland, Jeon, Park, SIMAX03, Howland and Park, IEEE TPAMI 04)

- Works regardless of singularity of *scatter matrices*
- $S_w^{-1} S_b x = / \ x \rightarrow S_b x = / \ S_w x \rightarrow \ {}^2 H_b H_b^T x = g^2 H_w H_w^T x$
- G comes from leading (r-1) generalized singular vectors of H_b^T and H_w^T

$$U^T H_b^T X = (S_b \ \ 0) =$$



$$V^T H_w^T X = (S_w \ \ 0) =$$



$X^T H_b H_b^T X = X^T S_b X$ and $X^T H_w H_w^T X = X^T S_w X$

Classical LDA is a special case of LDA/GSVD

Generalized SVD

(Paige and Saunders '81)

$$S_b x = / \, S_w x \rightarrow \quad {}^2 H_b H_b^T x = g^2 H_w H_w^T x$$

$$X^T S_b X = \begin{bmatrix} I & & & \\ & D_b & & \\ & & 0 & \\ & & & 0 \end{bmatrix} \qquad X^T S_w X = \begin{bmatrix} 0 & & & \\ & D_w & & \\ & & I & \\ & & & 0 \end{bmatrix}$$

Want ***G*** s.t. ***max trace*** $(G^T S_b G)$ and ***min trace*** $(G^T S_w G)$

X = [X1 X2 X3 X4]	g	δ	x_i belongs to
X_1	1	0	$null(S_b)^c \cap null(S_w)$
X_2	$1 > \beta > 0$	$0 < \delta < 1$	$null(S_b)^c \cap null(S_w)^c$
X_3	0	1	$null(S_b) \cap null(S_w)^c$
X_4	any	any	$null(S_b) \cap null(S_w)$

Generalization of LDA for Undersampled Problems

- **Regularized LDA** *(Friedman '89, Zhao et al. '99 …)*
- **LDA/GSVD** : Solution $G = [\ X_1 \ \ X_2\ ]$ *(Howland, Jeon, Park '03)*
- Solutions based on $Null(S_w)$ and $Range(S_b)$…

(Chen et al. '00, Yu & Yang '01, Park & Park '03 …)

- **Two-stage methods**:

  - Face Recognition: PCA + LDA *(Swets & Weng '96, Zhao et al. 99)*
  - Information Retrieval: LSI + LDA *(Torkkola '01)*
  - Mathematical Equivalence: *(Howland and Park '03)*

PCA+ LDA/GSVD = LDA/GSVD

LSI +LDA/GSVD = LDA/GSVD

More efficient = QRD + LDA/GSVD

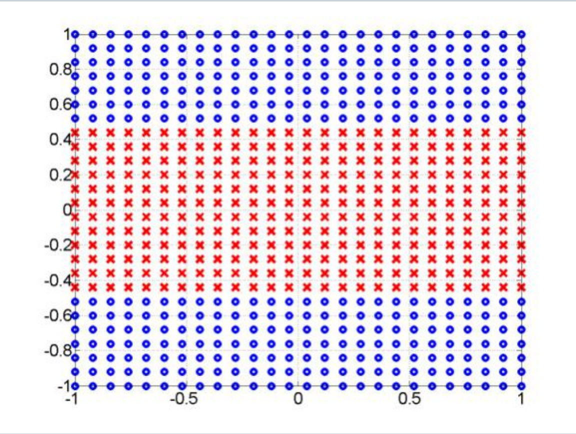
Nonlinear Dimension Reduction by Kernel Functions

Ex. Feature mapping F

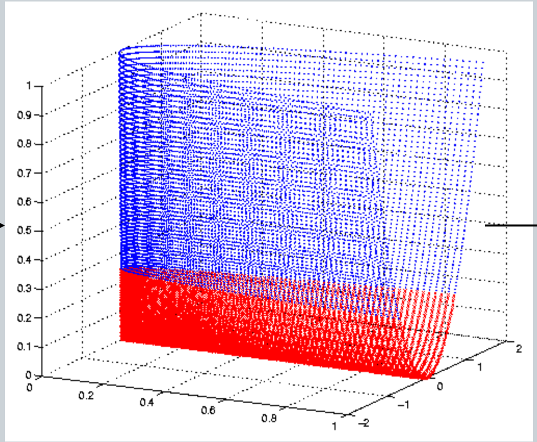
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longrightarrow F(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{bmatrix},$$

$$k(x, y) = \langle F(x), F(y) \rangle = \langle x, y \rangle^2$$

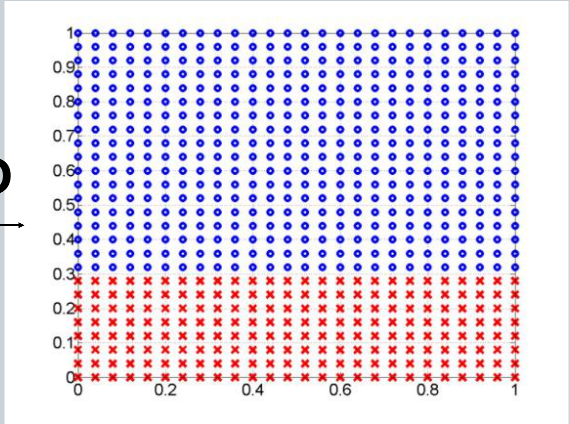
(a polynomial kernel function)



F



2D



Nonlinear Dimension Reduction by Kernel Functions

If $k(x,y)$ satisfies Mercer's condition, then there is a mapping F to an inner product space,

$$k(x,y) = < F(x), F(y) >$$

$$A \xrightarrow{F} F(A)$$

$$< x, y > \qquad k(x,y) = < F(x), F(y) >$$

Mercer's Condition for $A=[a_1, \cdots ,a_n]$:

kernel matrix $K = [\, k(a_i, a_j)\,]_{1 \le i,\, j \le n}$

is **positive semi-definite**.

Ex) RBF Kernel Function: $k(a_i, a_j) = \exp(-s \|a_i - a_j\|^2)$

Nonlinear Discriminant Analysis based on Kernel Functions (KDA/GSVD)

(C. Park and H. Park, SIMAX 04)

Assume a feature mapping: $f : a: mx1 \rightarrow f(a): px1$, m<<p

and apply LDA/GSVD to S_w and S_b in feature space

G : leading (r-1) generalized singular vectors of (H_w^f , H_b^f)

$$S_w^f = H_w^f \times H_w^{f\,T}$$

- $S_w^f = H_w^f H_w^{fT}$, $H_w^f = [a_1^f - c_1^f, a_2^f - c_1^f, \ldots, a_n^f - c_r^f]$: pxn
- $S_b^f = H_b^f H_b^{f\,T}$, $H_b^f = [\, a_1(c_1^f - c^f), \ldots, a_r (c_r^f - c^f)]$: pxr

- *f unknown but problem can be formulated to utilize kernel fcn.*

Classifier by MSE and Dimension Reduction by LDA/GSVD

(Binary Case) *(Duda et al. 01, C. Park and H. Park 04)*

- **MSE**

$$f(a) = a^T w + b$$

$$= \begin{cases} n/n_1 & \text{if } a \in \text{class } 1 \\ -n/n_2 & \text{if } a \in \text{class } 2 \end{cases}$$

$$\min \left\| \begin{bmatrix} 1 & a_1^T \\ \vdots & \\ 1 & a_n^T \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} - \begin{bmatrix} n/n_1 \\ \vdots \\ -n/n_2 \\ \vdots \end{bmatrix} \right\|_2 \longrightarrow$$

- **LDA/GSVD**

$$d^2 S_b x = g^2 S_w x$$

$$\downarrow$$

$$\boxed{\begin{aligned} & w^T a + b \\ & = w^T (a-c) \\ & = a\, x^T (a-c) \end{aligned}}$$

* *Extended to* ***(non)linear multi-class relationship***

(C. Park and H. Park , SIMAX, 05)

Relationship between Kernelized MSE and KDA/GSVD

(Binary Case) *(Billings and Lee 02, C. Park and H. Park 05)*

- **MSE**

$$f(a) = f(a)^T w + b$$

$$= \begin{cases} n/n_1 & \text{if } a \in \text{class } 1 \\ -n/n_2 & \text{if } a \in \text{class } 2 \end{cases}$$

$$\min \left\| \begin{bmatrix} 1 & f(a_1)^T \\ \vdots & \\ 1 & f(a_n)^T \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} - \begin{bmatrix} n/n_1 \\ \vdots \\ -n/n_2 \\ \vdots \end{bmatrix} \right\|_2 \longrightarrow$$

$$= \min \left\| \begin{bmatrix} e & f(A)^T \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} - y \right\|_2$$

- **LDA/GSVD**

$$d^2 S_b^f x = g^2 S_w^f x$$

$$\downarrow$$

$$\boxed{\begin{aligned} & w^T f(a) + b \\ & = w^T (f(a) - c^f) \\ & = a\, x^T (f(a) - c^f) \end{aligned}}$$

However, f is not known.

Formulation of Kernelized MSE

f is unknown but nonlinearization is possible using kernel functions and the fact that $w = f(A)\, z$ for some z

$$\min \left\| \begin{bmatrix} e & f(A)^T \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} - y \right\|_2 = \min \left\| \begin{bmatrix} e, & f(A)^T f(A) \end{bmatrix} \begin{bmatrix} b \\ z \end{bmatrix} - y \right\|_2$$

$$= \min \left\| \begin{bmatrix} e & K \end{bmatrix} \begin{bmatrix} b \\ z \end{bmatrix} - y \right\|_2$$

- Let $G = [\, e \quad K]$: nx(n+1)
- K : symmetric positive semidefinite
- Solution related to KDA/GSVD is $\begin{bmatrix} b \\ z \end{bmatrix} = G^+ y$
- If rank(G) = n, then G^+ can be obtained from QRD of G^T

  Let $G^T = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$, then $G^+ = G^T (GG^T)^{-1} = Q \begin{bmatrix} R^{-T} \\ 0 \end{bmatrix}$

Adaptive KDA by Regularized MSE (KDA/RMSE)

Replace
$$\min \left\| \begin{bmatrix} e & K \end{bmatrix} \begin{bmatrix} b \\ z \end{bmatrix} - y \right\|_2$$

by
$$\min \left\| \begin{bmatrix} e & K + / \; I \end{bmatrix} \begin{bmatrix} b \\ z \end{bmatrix} - y \right\|_2$$

- $G_/ = [\, e \quad K + / \; I \,]$: nx(n+1), rank($G_/$) = n for l > 0
- Solution can be obtained by QRD of $G_/{}^T$
- Updated and downdated sol. can be obtained by QRD updating/downdating.
- At least an order of magnitude faster than GSVD updates.

Adaptive Kernel MSE

(Kim, Drake, and Park '05)

- Kernel MSE

$$G_{/} = [e \quad K + / \; I\,] = \begin{bmatrix} 1 & k_{1,1} + / & \dots & k_{1,n} \\ \vdots & \vdots & & \vdots \\ 1 & k_{n,1} & \dots & k_{n,n} + / \end{bmatrix}$$

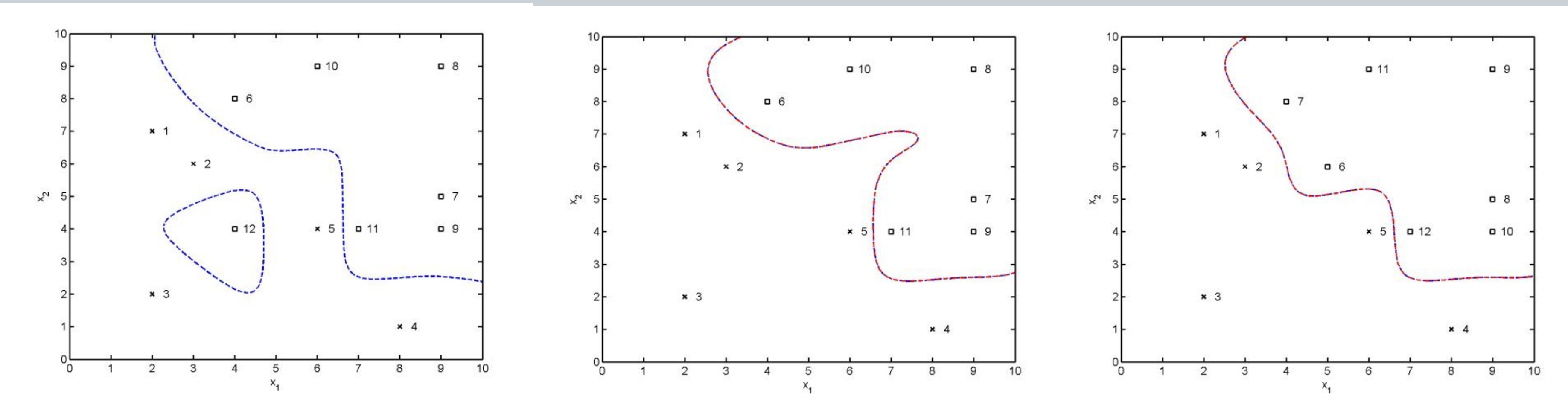
Appending a data point ***a'*** *: apply QRD updating twice*

$$G_{/}{}^{‘} = \left[\begin{array}{c|c|ccc} 1 & k_{a',a'} + / & k_{a',1} & \dots & k_{a',n} \\ \hline 1 & k_{1,a'} & & & \\ \vdots & \vdots & & K + / \; I & \\ 1 & k_{n,a'} & & & \end{array}\right]$$

Removing a data point $\boldsymbol{a_k}$ *: apply QRD downdating twice*

$$G_{/}{}^{‘} = \left[\begin{array}{ccc|cc|cc} 1 & k_{1,1} + / & \dots & k_{1,k-1} & k_{1,k+1} & \dots & k_{1,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \hline 1 & k_{k-1,1} & \dots & k_{k-1,k-1} + / & k_{k-1,k+1} & \dots & k_{k-1,n} \\ 1 & k_{k+1,1} & \dots & k_{k+1,k-1} & k_{k+1,k+1} + / & \dots & k_{k+1,n} \\ \hline \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & k_{n,1} & \dots & k_{n,k-1} & k_{n,k+1} & \dots & k_{n,n} + / \end{array}\right]$$

New Decision Boundaries after Deletion and Addition of Data Points



- New decision boundaries after deleting the 12th point and inserting (5,6)
- Dash-dotted contour : a decision boundary of the adaptive KDA/RMSE
- Dashed contour : a decision boundary obtained by computing sol. from scratch by the RMSE.

$$K(\mathbf{a}_1, \mathbf{a}_2) = \exp(-\gamma \|\mathbf{a}_1 - \mathbf{a}_2\|^2)$$

Comparison Between KDA and KDA/RMSE

Method	Thyroid	Diabetes	Heart	Titanic
KDA	3.9 +- 2.0	26.3 +- 2.2	16.1 +- 3.5	24.1 +- 2.7
KDA(75%) + adaptive KDA(25%)	3.9 +- 2.0	26.3 +- 2.2	16.1 +- 3.5	24.1 +- 2.7

Average and standard deviation of test set classification errors in % for 100 partitions

Face Recognition Training Dataset (AT&T)

- 10 persons x 5 images/person = 50 images total
- Each image: 46x56
- Five-fold CV using KDA
  SVD of $G = (e \;\; K)$: 40x41 is computed for each fold

  $K(a_1, a_2) = \exp(-\gamma \|a_1 - a_2\|^2)$

  $Err_{min} = 4.0\%$
- Five-fold CV using KDA/RMSE

  QRD of $G_\lambda = (e \;\; K+\lambda I)$: 50x51, and block downdate.

  $Err_{min} = 2.0\%$

Face Recognition Testing

1. Cross validation on training dataset

$$Err_{cv} = 2.0\%$$

2. Classification of test dataset using optimal parameters obtained from CV.

Updated Face Recognition Training Dataset

Target:
1st image

Efficient Computing

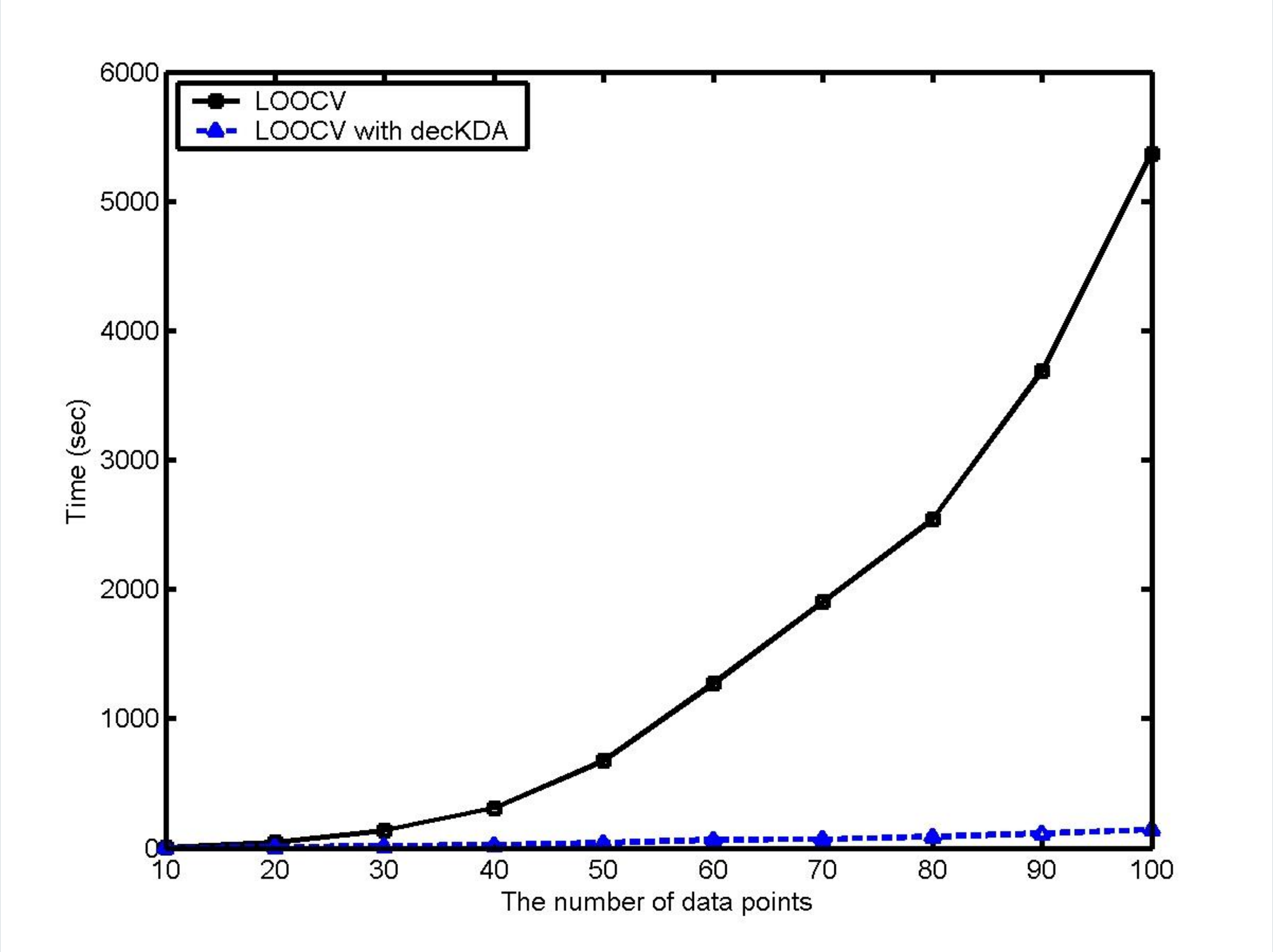
1. Removed an old image by decKDA
2. Appended a new image by incKDA

Result:

$$Err_{cv} = 2.0\%$$

$$Err_{tst} = 0.0\%$$

Computation time for leave-one-out cross validation (LOOCV): 2001 KDD cup drug design data, 8000 features



Solid black line : computation time of ordinary LOOCV

Dashed blue line : computation time of LOOCV using decKDA/RMSE

Summary / Future Research

Effective Algorithm for Adaptive Disc. Analysis

- Utilized the relationship between LDA/GSVD and MSE
- Replaced SVD up/down-dating by QRD up/down-dating
- Applicable to a wide range of problems (Facial recognition, text classification, faster cross validation algorithms …)

Current and Future Research

* Development of recursive feature tracking system
  based on recursive KDA/RMSE / parallel implementation
* Utilization of other efficient methods such as
  complete orthogonal decomposition
* Establish **Mathematical Relationships** among
  Dimension Reduction, Classifier Design, Data Reduction, …

Thank you !