8. Visual Data Dimensionality Reduction

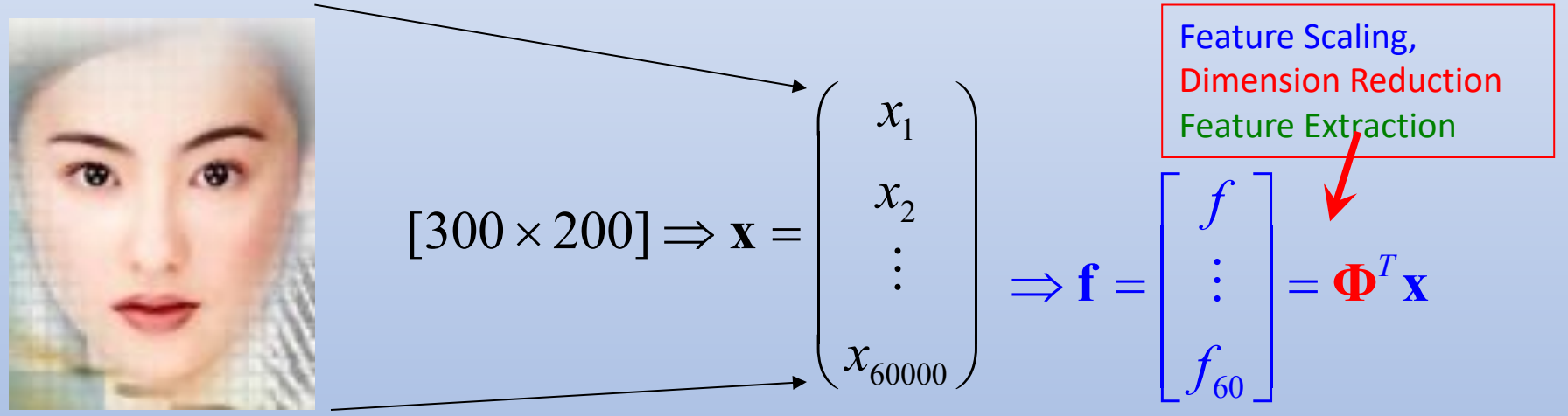
- High data dimensionality often imposes great burdens on the robust and accurate recognition due to insufficient knowledge about the data population and limited number of training samples.
- Dimensionality reduction is thus a critical module of such recognition systems.
- Extracting the discriminative and reliable features or dimensions is always the key step for image recognition and computer vision.

  - Feature extraction and dimensionality reduction can be based on
    - Human expert knowledge
    - Image local structures
    - Image global structure
    - Machine learning from training database

8. Visual Data Dimensionality Reduction

Problems:

- Poor or even no human knowledge about the effective features.
- Image patterns are so complex that it is impossible to incorporate all different variations into the deterministic computer program.

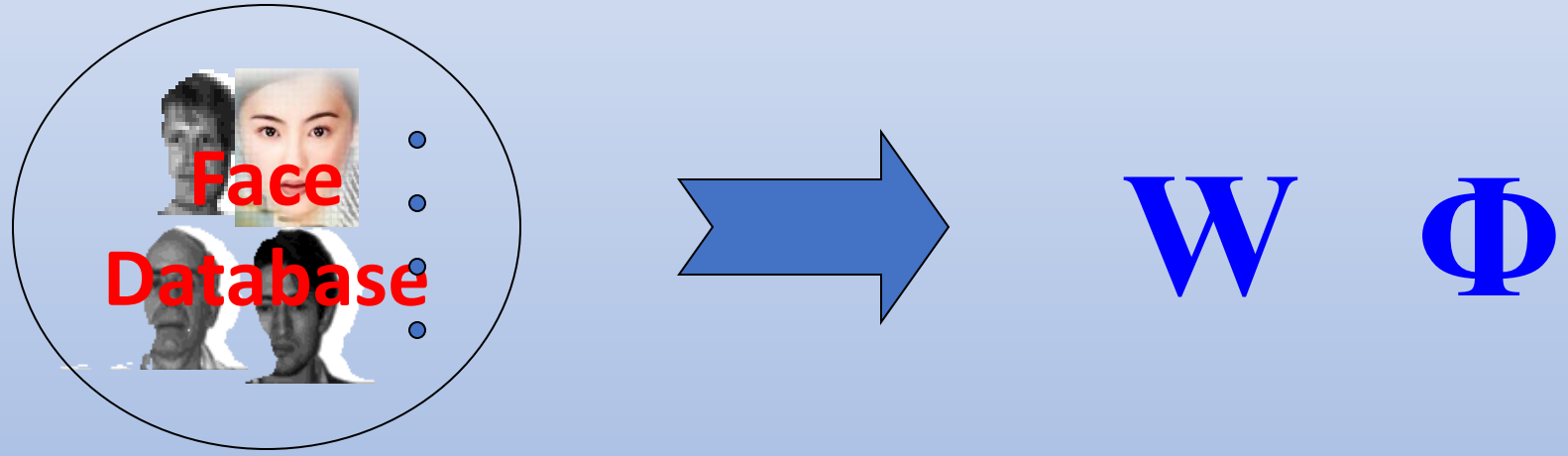


Solution: Take all pixels of the whole image as initial features and derive the effective features based on machine learning.

A real application example: J. Ren, X.D. Jiang and J. Yuan, "A Complete and Fully Automated Face Verification System on Mobile Devices," Pattern Recognition, vol. 46, no. 1, pp. 45-56 January 2013.

8. Visual Data Dimensionality Reduction

The transform parameter W or Φ that weight the features and reduce the feature dimensionality is determined by machine (computer) learning (training) from a image database.



Why is the transform necessary?
what principles are used to derive W ?
What are the potential problems?
How to alleviate these problems?

PCA: Principal Component Analysis

Given a data set of q n-dimensional training samples

$$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$$

Each sample represented by a column vector is a point in an n-dimensional space.

How to use one point, $\mathbf{x}_0$ to best represent all training data?

i.e.

$$\varepsilon^2 = \sum_{i=1}^{q} \| \mathbf{x}_0 - \mathbf{x}_i \|^2 = \sum_{i=1}^{q} (\mathbf{x}_0 - \mathbf{x}_i)^T (\mathbf{x}_0 - \mathbf{x}_i) \Rightarrow \text{minimum}$$

- It is very easy to prove that the solution is the sample mean

$$\mathbf{x}_0 = \boldsymbol{\mu} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_i$$

Just for symbolic simplicity, we centralize training samples and define a training data matrix by

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}, \qquad \mathbf{X} = \begin{bmatrix} \tilde{\mathbf{x}}_1 & \tilde{\mathbf{x}}_2 & \dots & \tilde{\mathbf{x}}_q \end{bmatrix}$$

PCA: Principal Component Analysis

- Now we want to just use one dimension $\phi, \quad \| \phi \|^2 = \phi^T \phi = 1$ to best represent all samples, $\tilde{\mathbf{x}}_i$.
- The *n*-dimensional data $\tilde{\mathbf{x}}_i$ are reduced to one-dimensional data.

$$a_i = \phi^T \tilde{\mathbf{x}}_i$$

- The best dimension ϕ makes reconstruction error minimum.

$$\varepsilon^2 = \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i - a_i \phi \|^2 \Rightarrow \text{minimum}$$

- It is easy to have

$$\varepsilon^2 = \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i - a_i \phi \|^2 = \sum_{i=1}^{q} (\tilde{\mathbf{x}}_i - a_i \phi)^T (\tilde{\mathbf{x}}_i - a_i \phi)$$

$$= \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i \|^2 - \sum_{i=1}^{q} a_i^2 \qquad \left(\text{note: } a_i \phi^T \tilde{\mathbf{x}}_i = a_i^2, \ a_i \phi^T a_i \phi = a_i^2 \phi^T \phi = a_i^2 \right)$$

$$= \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i \|^2 - \sum_{i=1}^{q} \phi^T \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \phi$$

$$= \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i \|^2 - \phi^T \left(\sum_{i=1}^{q} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \right) \phi$$

PCA: Principal Component Analysis

- The sample covariance matrix of all training data,

$$\mathbf{S}^t = \frac{1}{q}\sum_{i=1}^{q}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{q}\sum_{i=1}^{q}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$$

  is also called the total scatter matrix in the literature.

- To minimize $\varepsilon^2 = \sum_{i=1}^{q} \| \tilde{\mathbf{x}}_i \|^2 - q\phi^T\mathbf{S}^t\phi$

is to maximize $\phi^T\mathbf{S}^t\phi$ with the constraint of $\| \phi \|^2 = \phi^T\phi = 1$

- Use Lagrange optimization method

$$f(\phi, \lambda) = \phi^T\mathbf{S}^t\phi - \lambda(\phi^T\phi - 1) \Rightarrow \max imum$$

$$\frac{\partial f}{\partial \phi} = 2\mathbf{S}^t\phi - 2\lambda\phi = 0$$

$$\mathbf{S}^t\phi = \lambda\phi$$

- The solution is eigenvalue and eigenvector of matrix $\mathbf{S}^t$

PCA: Principal Component Analysis

- We see that if ϕ is the eigenvector of the covariance matrix $\mathbf{S}^t$,

  $\phi^T \mathbf{S}^t \phi$ is maximum and the reconstruction error is minimum.

- The variance of the data projected onto the dimension spanned by the eigenvector is maximum.

$$\phi^T \mathbf{S}^t \phi = \frac{1}{q} \sum_{i=1}^{q} \phi^T (\mathbf{x}_i - \boldsymbol{\mu}) \left[\phi^T (\mathbf{x}_i - \boldsymbol{\mu}) \right]^T = \frac{1}{q} \sum_{i=1}^{q} a_i^2$$

- Eigenvalue of a covariance matrix is the variance of the data projected onto the eigenvetor.

$$\because \mathbf{S}^t \phi = \lambda \phi \quad \therefore \phi^T \mathbf{S}^t \phi = \phi^T \lambda \phi = \lambda \phi^T \phi = \lambda$$

PCA: Principal Component Analysis

- Reducing the n-dimensional data $\mathbf{x}_i$ into lower m-dimensional data $\mathbf{y}_i$ by

$$\mathbf{y}_i = \mathbf{\Phi}^T (\mathbf{x}_i - \boldsymbol{\mu})$$

where $\mathbf{\Phi} = [\phi_1 \ \phi_2 \ \ldots \phi_m], \quad m < n$

consists of m eigenvectors corresponding to the m largest eigenvalues of the total scatter matrix $\mathbf{S}^t$.

- We can reconstruct the original n-dimensional data $\mathbf{x}_i$ using the lower m-dimensional data $\mathbf{y}_i$ using

$$\hat{\mathbf{x}}_i = \mathbf{\Phi}\mathbf{y}_i + \boldsymbol{\mu}$$

The reconstruction error is minumum.

$$\varepsilon^2 = \sum_{i=1}^{q} \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|^2 \Rightarrow \text{minimum}$$

PCA: Principal Component Analysis

- Therefore, the famous principal component analysis PCA transforms the *n*-dimensional *X* into *m*-dimensional *Y* by:

$$\mathbf{y} = \mathbf{\Phi}^T (\mathbf{x} - \boldsymbol{\mu})$$

where $\mathbf{\Phi} = [\phi_1 \ \phi_2 \ \ldots \phi_m], \quad m < n$

consists of m eigenvectors corresponding to the m largest eigenvalues of the total scatter matrix $\mathbf{S}^t$.

The reconstruction error is: $\Delta = \mathbf{x} - \hat{\mathbf{x}} = \mathbf{x} - \mathbf{\Phi}\mathbf{y} - \boldsymbol{\mu}$

and the mean squared reconstruction error

$$E[\| \Delta \|^2] = E[\Delta^T \Delta] = \sum_{k=m+1}^{n} \lambda_k$$

where λ_k are eigenvalues sorted in descending order.

PCA: Principal Component Analysis

- Note that:

$$\mathbf{S}^t = \frac{1}{q}\sum_{i=1}^{q}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{q}\sum_{i=1}^{q}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T = \frac{1}{q}\mathbf{X}\mathbf{X}^T$$

- The rank of the total scatter matrix $\mathbf{S}^t$ is min(q-1,n) at most. Therefore, it has q-1 nonzero eigenvalues at most.

- Thus, PCA with q-1 dimensional $\mathbf{y}$ will fully represent the original n-dimensional data $\mathbf{x}$, q-1$\leq n$, because the reconstruction error is zero.

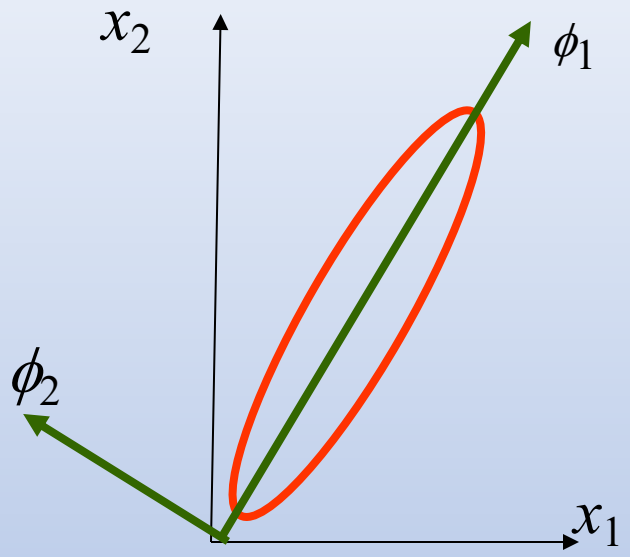
$$E[\| \Delta \|^2] = \sum_{k=q}^{n}\lambda_k = 0$$

PCA: Principal Component Analysis

$$PCA: \quad \max trace[\mathbf{\Phi}^T \mathbf{S}^t \mathbf{\Phi}] = \sum_{k=1}^{m} \lambda_k^t$$

$$\mathbf{\Phi}: \ [n \times m], \text{ e.g. } =[60000 \times 80]$$

$$\mathbf{y} = \mathbf{\Phi}^T (\mathbf{x} - \boldsymbol{\mu})$$

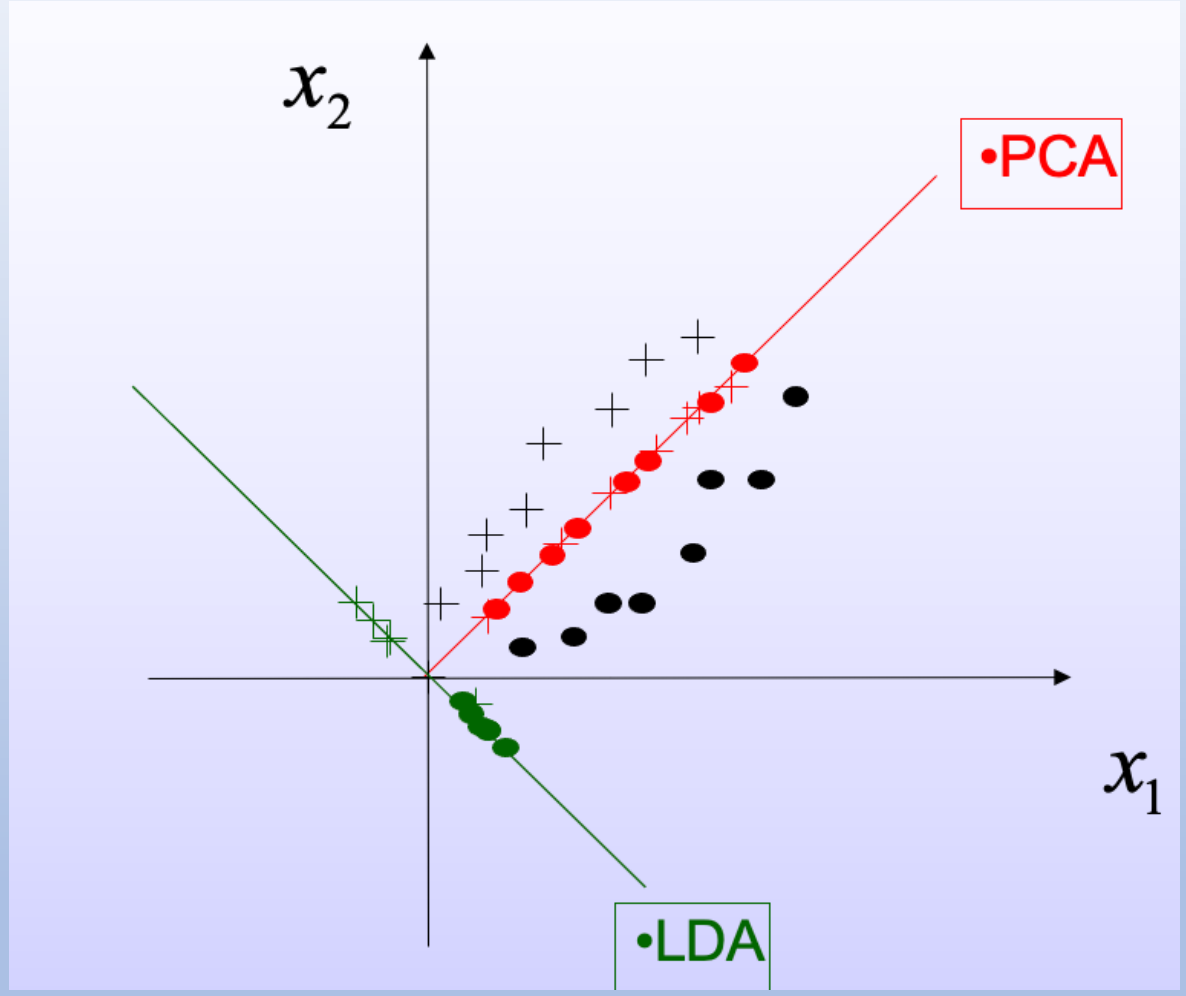


- PCA is an unsupervised method
  that minimizes the reconstruction error
  or maximize the variation (information carried by the data) in the reduced sub-space.
  This dimension reduction will reduce the computational complexity of the subsequent processing.
- However, any information lost may reduces the accuracy of the subsequent processing. Further more, the lost information, though less important for data representation, could be critical for discriminating the data class membership.

LDA: Linear Discriminant Analysis

- Problem of PCA in classification?

$$\mathbf{y} = \mathbf{\Phi}^T (\mathbf{x} - \boldsymbol{\mu})$$



➢ Desired? properties to determine the projection vector for classification:

Maximize separation between projected class means AND Minimize projected within-class scatter (variance)

LDA: Linear Discriminant Analysis

- Given q n-dimensional training samples of c classes

$$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_q$$

- The number of training samples of class ω_j is q_j, j = 1, 2,... c.
- The covariance matrix of class ω_j is computed as

$$\boldsymbol{\Sigma}_j = \frac{1}{q_j} \sum_{X_i \in \omega_j} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T, \quad \text{where } \boldsymbol{\mu}_j = \frac{1}{q_j} \sum_{X_i \in \omega_j} \mathbf{x}_i$$

- The within class scatter matrix of the c classes is defined as

$$\mathbf{S}^w = \sum_{j=1}^{c} \frac{q_j}{q} \boldsymbol{\Sigma}_j$$

- The between class scatter matrix of c classes is defined as

$$\mathbf{S}^b = \sum_{j=1}^{c} \frac{q_j}{q} (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T, \text{ obviously: } \boldsymbol{\mu} = \frac{1}{q} \sum_{i=1}^{q} \mathbf{x}_i = \sum_{j=1}^{c} \frac{q_j}{q} \boldsymbol{\mu}_j$$

LDA: Linear Discriminant Analysis

- Instead of $PCA: \ \max trace[\mathbf{\Phi}^T \mathbf{S}^t \mathbf{\Phi}] = \sum_{k=1}^{m} \lambda_k^t$

$$LDA: \ \min trace\left[\mathbf{\Phi}^T \mathbf{S}^w \mathbf{\Phi}\right] \ \& \ \max trace\left[\mathbf{\Phi}^T \mathbf{S}^b \mathbf{\Phi}\right]$$

$$\Rightarrow LDA: \ \max trace\left[\mathbf{\Phi}^T \mathbf{S}^{w^{-1}} \mathbf{S}^b \mathbf{\Phi}\right] = \sum_{k=1}^{m} \lambda_k^{b/w}$$

- Note that:

$trace(\mathbf{S}^t) \Rightarrow$ total variation amount of all samples

$trace(\mathbf{S}^w) \Rightarrow$ total variation amount of samples within classes

$trace(\mathbf{S}^b) \Rightarrow$ total variation amount of samples between classes

- It is easy to prove: $\mathbf{S}^t = \mathbf{S}^w + \mathbf{S}^b$

LDA: Linear Discriminant Analysis

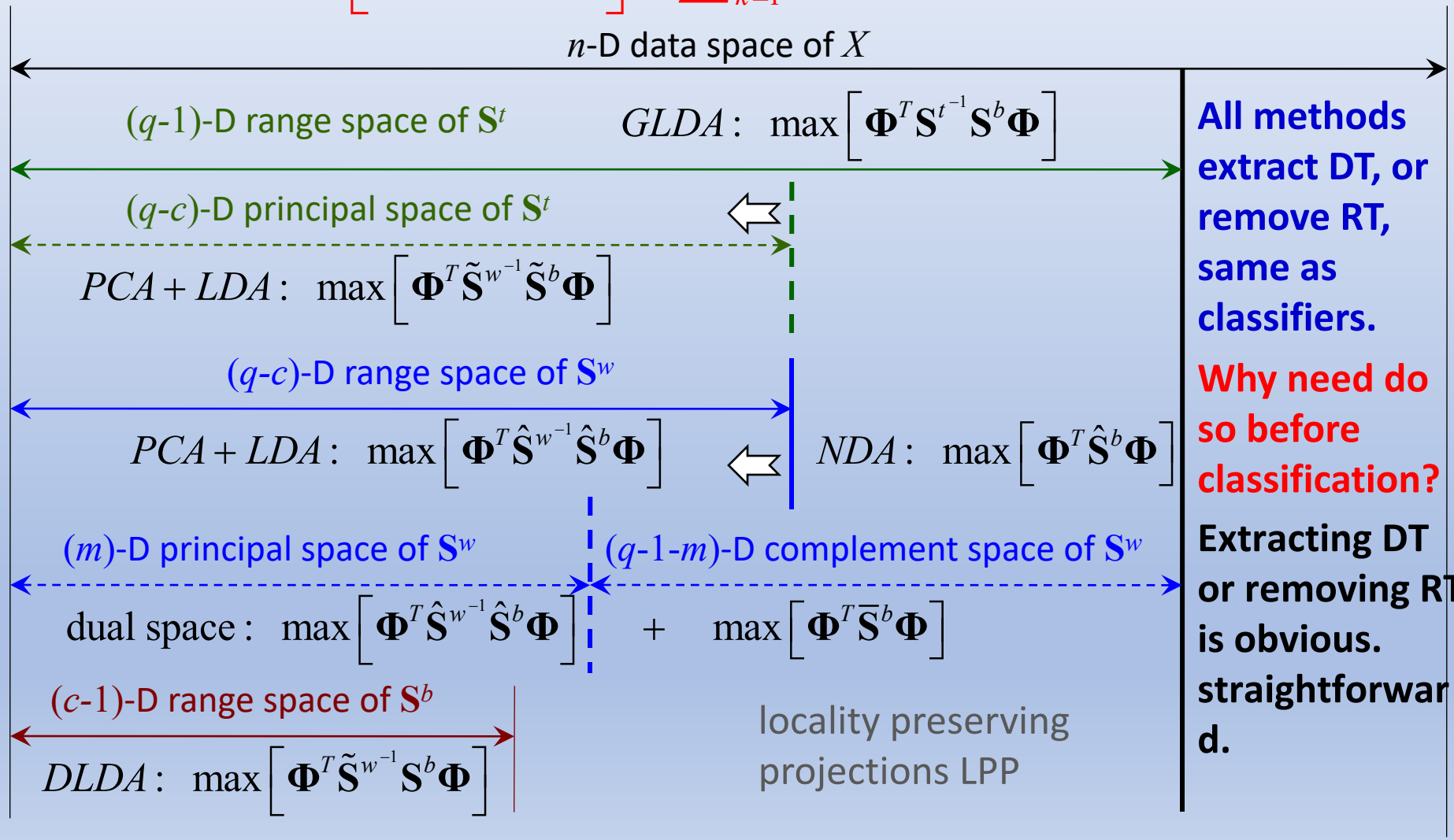
- As pattern recognition is not to represent the original data, but to discriminate the class membership of the data, obviously, LDA extracts much more discriminative features than PCA. Therefore, the vast majority of researchers prefer LDA to PCA for feature extraction and dimensionality reduction for pattern recognition.

$$LDA: \quad \max trace\left[\mathbf{\Phi}^T \mathbf{S}^{w^{-1}} \mathbf{S}^b \mathbf{\Phi}\right] = \sum_{k=1}^{m} \lambda_k^{b/w}$$

- The rank of $\mathbf{S}^w$ is $min(q\text{-}c,n)$ at most. For many applications $q\text{-}c<<n$. So, we have problem because $\mathbf{S}^w$ is singular and its inverse does not exist.
- Thousands of research papers proposed various LDA variants trying to solve this problem!

Major Efforts in LDA Variants

$$LDA: \ \max trace\left[\mathbf{\Phi}^T \mathbf{S}^{w^{-1}} \mathbf{S}^b \mathbf{\Phi}\right] = \sum_{k=1}^{m} \lambda_k^{b/w} \qquad \mathbf{S}^t = \mathbf{S}^w + \mathbf{S}^b$$



Major efforts in LDA variants

- PCA extracts the most representative information in the sense of the least square error. LDA extracts the most discriminative information in the linear constrain.

- As discriminative information is used for classification, the vast majority of researchers prefer LDA to PCA. Numerous efforts were made in the past two decades to make LDA workable or propose even more discriminative approaches than LDA.

- Although they are suboptimal (GLDA, dual space), or have lost some discriminative information before LDA (PCA+LDA, DLDA, NDA), numerous LDA variants makes LDA workable.

Questions on DR, PCA and LDA

- Any DR loses information. Extracting the most discriminative information just means losing least information.
- If higher accuracy can be achieved with less information, why the criterion of DR is set to extract the most discriminative information? If not, why do we need DR?
- Superficial study will cause misunderstanding and mistakes and hence insignificant (trivial) or fake research.

Objectives of FE/DR for PR

- We need rethink hard many related issues in-depth.
- This study (from this slide to the end of topic 12) has been published in:

  X.D. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, March 2011.

  X.D. Jiang, B. Mandal and A. Kot, "Eigenfeature Regularization and Extraction in Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383-394, March 2008.

  X.D. Jiang, "Asymmetric Principal Component and Discriminant Analyses for Pattern Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931-937, May 2009.

- Can Dimensionality reduction by PCA and LDA help enhance the classification accuracy? Why or why not? If yes, to what extent and how?

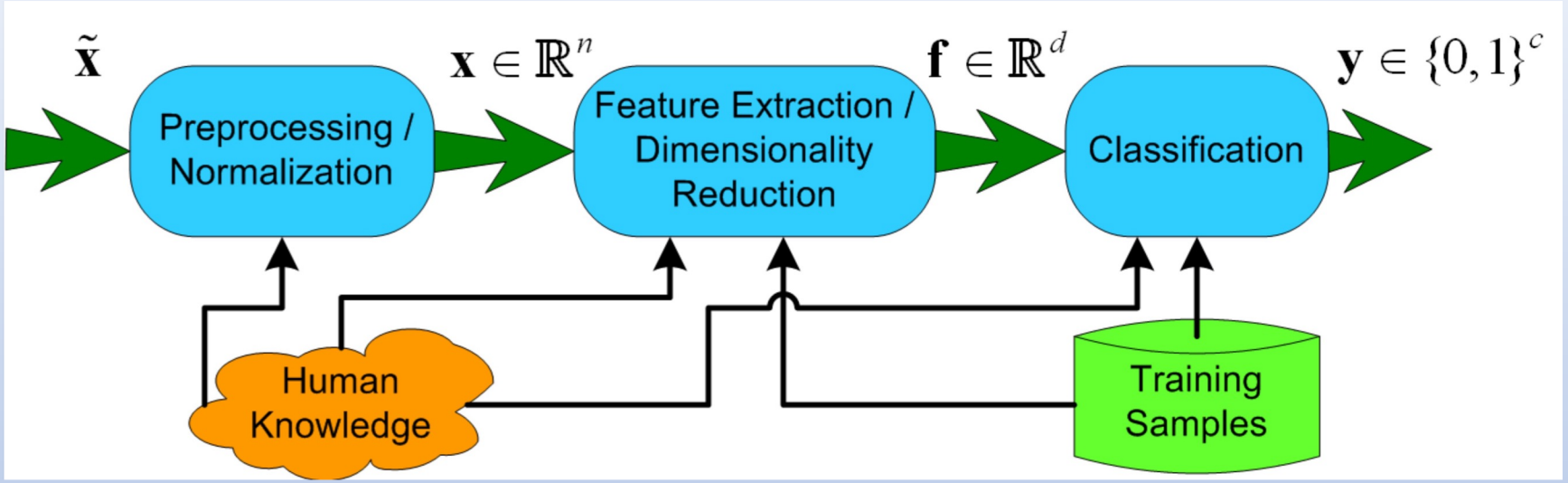
- This is far from straightforward.

Objectives of FE/DR for PR

- It is theoretically proven: classification accuracy increases or at least does not decrease as the data dimensionality increases, as long as the decision is based on the knowledge about the whole data population.
- However, it is also well known that high dimensionality often degrades the classification performance in practice for a fixed number of training data (curse of dimensionality).

- This paradox can be resolved by distinguishing the discriminative information about data population from that on training data.
- Although the ultimate objective of all modules of a recognition system is to extract the most discriminative information, it is the most discriminative information about the whole data population, not on a specific training set.

Objectives of FE/DR for PR



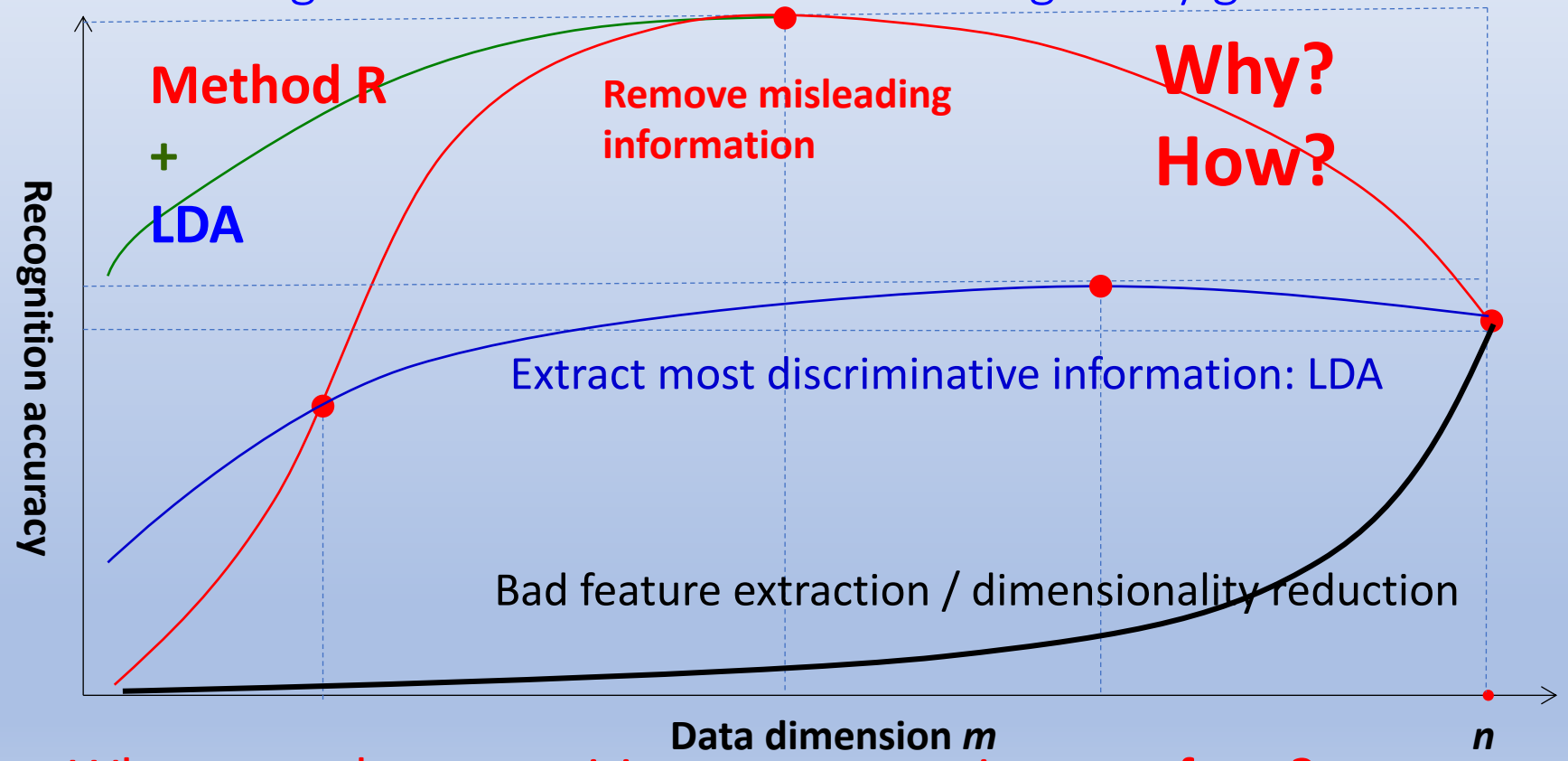
- All modules of a recognition system extract the discriminative information on population (**DP**), or remove the redundant information on population (**RP**).
- A classifier is trained to extract discriminative information on training data (**DT**), or remove the redundant information on training data (**RT**).

Objectives of FE/DR for PR

Recognition accuracy in general decreases with decreasing dimensionality

Information = discriminative information + redundant information

- Minimizing the loss doesn't mean we can get any gain!



Where can the recognition accuracy gain come from?

Information = reliable information + misleading information

A Simple Example shows DP, RP, DT, RT and MLI

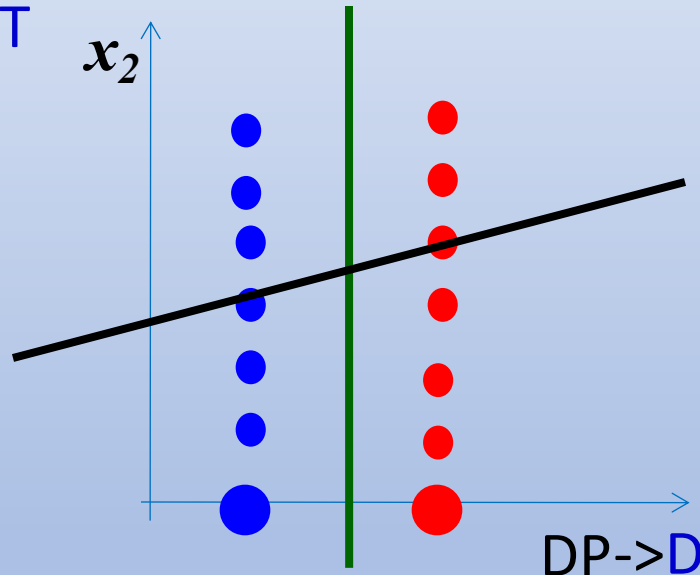
How can dimension reduction enhance the classification accuracy?

A proper Classifier extracts DT and removes / ignores RT.

RP ->DT ->MLI

=RT

MLI = RP∩DT



DP->DT

=DT

DP∩DT

Removing RT or extracting DT before classification cannot enhance the recognition accuracy.

It can only simplify or speed up the classification.

Removing MLI enhances the classification/decision accuracy and robustness

However, Can this be achieved by the criteria of PCA or LDA?

How to find the unreliable or harmful or misleading dimensions?

Objectives of FE/DR for PR

- Curse of dimensionality, small sample size problem. Dimensionality reduction may solve this problem. These are common sense. However, what dimensions should be extracted or what else should be removed?

- Extracting maximal information just means minimizing loss. Minimizing loss does not mean any gain.

- Therefore, it is unlikely that the dimensionality reduction by minimizing the loss of discriminative information can boost the classification accuracy.

Objectives of FE/DR for PR

A classifier is trained to capture the most discriminative information on the training samples. If some statistics learnt on the training data deviate from those of the data population, misclassification rate on the novel data increases.

Therefore, to boost the classification accuracy, the dimensionality reduction before classification should be targeted at removing the dimensions misleading or harmful for the classification.

Problems of classification in high dimension

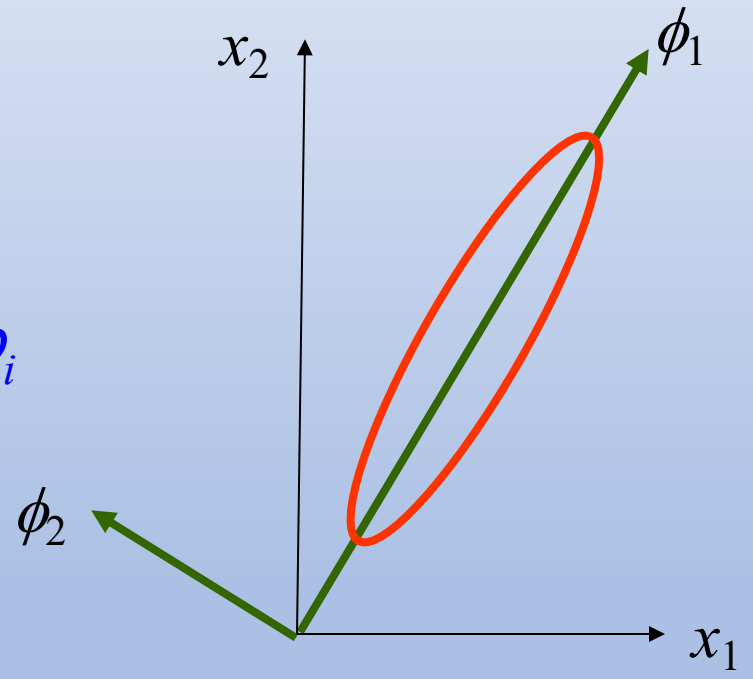
- In the eigen-space spanned by ϕ_k, we can handle the complex inverse of the *n* by *n* covariance matrix in a convenient scalar form!

$$\mathbf{z} = \{z_k\}_1^n = \mathbf{\Phi}_i{}^T \mathbf{x} = \{\phi_k\}_1^n{}^T \mathbf{x}$$

$$\bar{z}_k = \phi_k{}^T \boldsymbol{\mu}_i$$

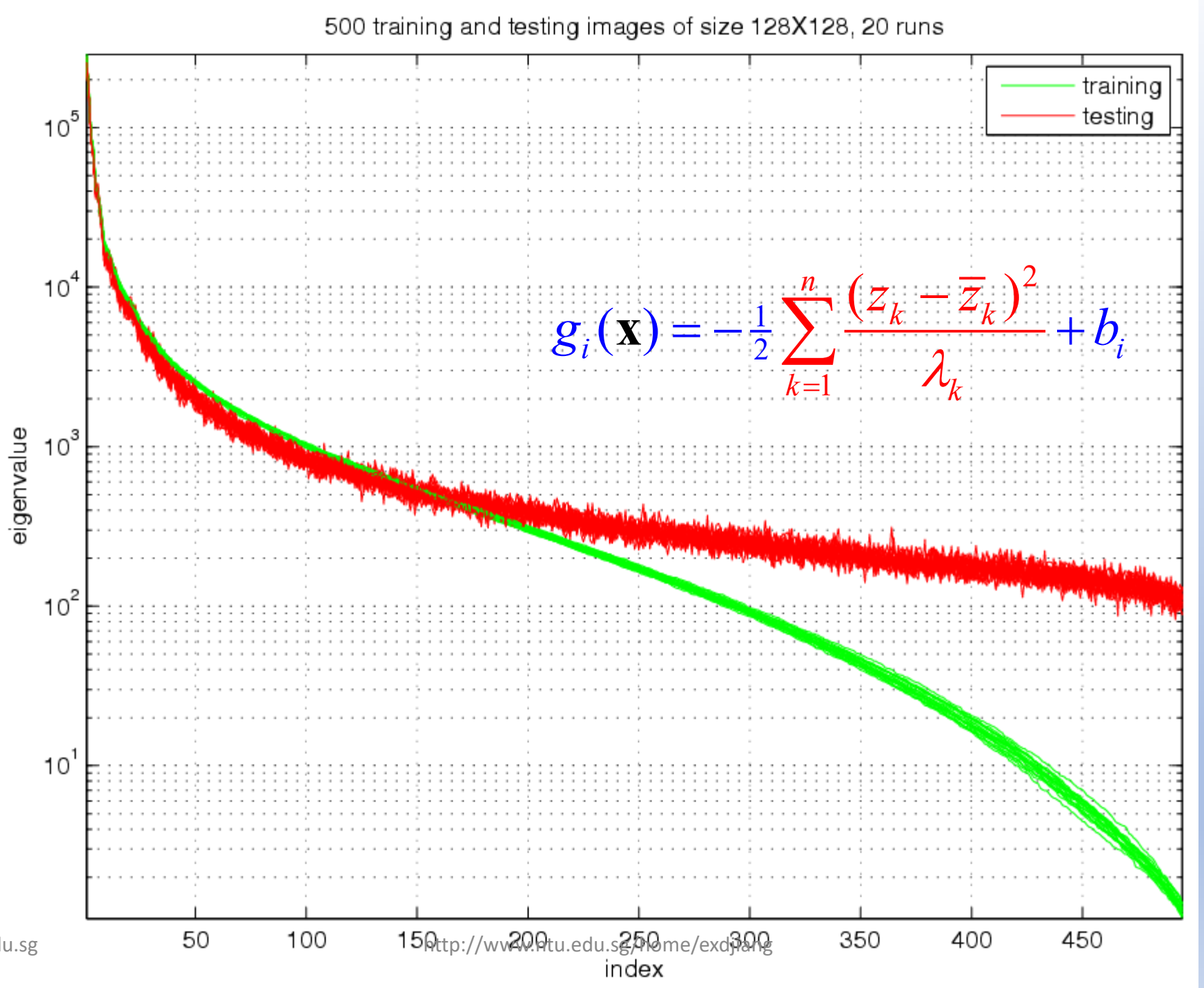
$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

$$= -\tfrac{1}{2}\sum_{k=1}^{n}\frac{(z_k-\bar{z}_k)^2}{\lambda_k} + b_i$$



- Problem is now evident if some eigenvalues are unreliable, especially if some eigenvalues are small and approach to zero.

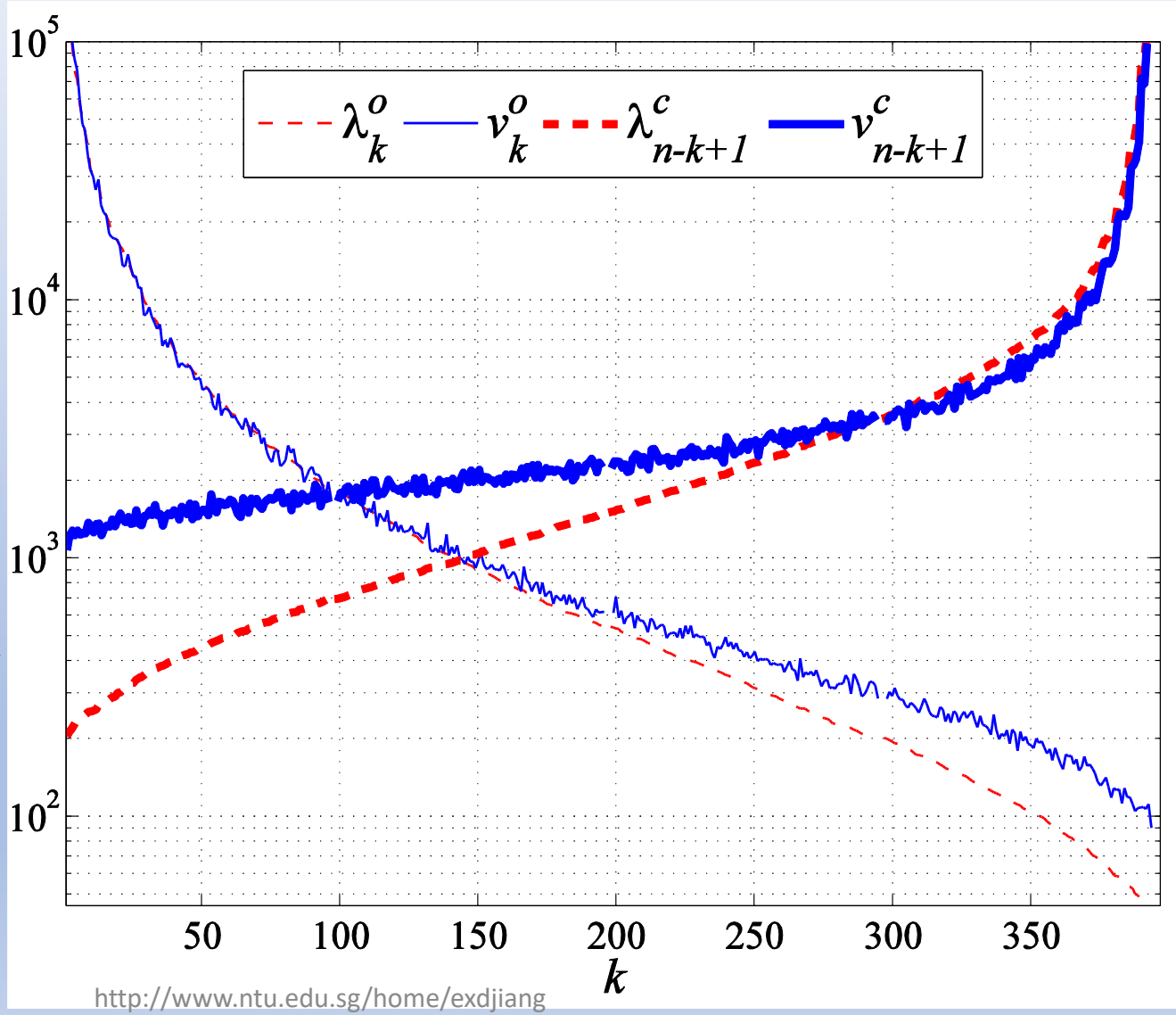

500 training and testing images of size 128X128, single run
training
testing
eigenvalue
index
10^5
10^4
10^3
10^2
10^1
50
100
150
200
250
300
350
400
450


500 training and testing images of size 128X128, 20 runs
training
testing
$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{n}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} + b_i$
eigenvalue
index
10^5
10^4
10^3
10^2
10^1
50
100
150
200
250
300
350
400
450

Problems of Small Eigenvalues

Problems of the dimensions corresponding to the unreliable small eigenvalues.

X.D. Jiang, "Asymmetric Principal Component and Discriminant Analyses for Pattern Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931-937, May 2009.



Solution1: adding a small constant

- One solution is to regularize the covariance matrix. A common practice in classification and data regression is to add a constant to its diagonal elements.

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{\Sigma}_i + a\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i$$

- Although this method was originally proposed to circumvent the singularity of $\boldsymbol{\Sigma}_i$ and the numerical instability of its inverse, numerous algorithms for classification, data regression, dimensionality reduction and manifold learning adopt this classical technique.
- Why this technique can improve the classification accuracy?
- The underlying principle can been seen by its equivalence to adding the constant to all eigenvalues

$$(\mathbf{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{\Sigma}_i + a\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \sum_{k=1}^{n}\frac{1}{\lambda_k + a}(z_k - \bar{z}_k)^2$$

Solution1: adding a small constant

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{\Sigma}_i + a\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + b_i = \sum_{k=1}^{n}\frac{-1/2}{\lambda_k + a}(z_k - \bar{z}_k)^2 + b_i$$

- The regularized eigen-spectrum $\lambda^a_k = \lambda_k + a$ can be very close to the population variances as shown in the Figures latter.

- This method has no dimensionality reduction effect.

Solution2: probabilistic subspace learning

Another solution, called probabilistic subspace learning decomposes the discriminant function into two parts and replaces the small eigenvalues by a constant

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{n}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} + b_i \quad \Downarrow$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{m}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} - \frac{1}{2}\sum_{k=m+1}^{n}\frac{(z_k - \bar{z}_k)^2}{\rho_{av}} + b_i$$

The constant is computed by

$$\rho_{av} = \frac{1}{n-m}\sum_{k=m+1}^{n}\lambda_k$$

as it is the optimal approximation to λ_k *for* $m < k < n$.

This method leads to one of the best performers called Bayesian algorithm in the face recognition community

Solution3: enhanced probabilistic subspace learning

However, the purpose of regularization is not best approximating to the eigenspectrum but to the population variances.

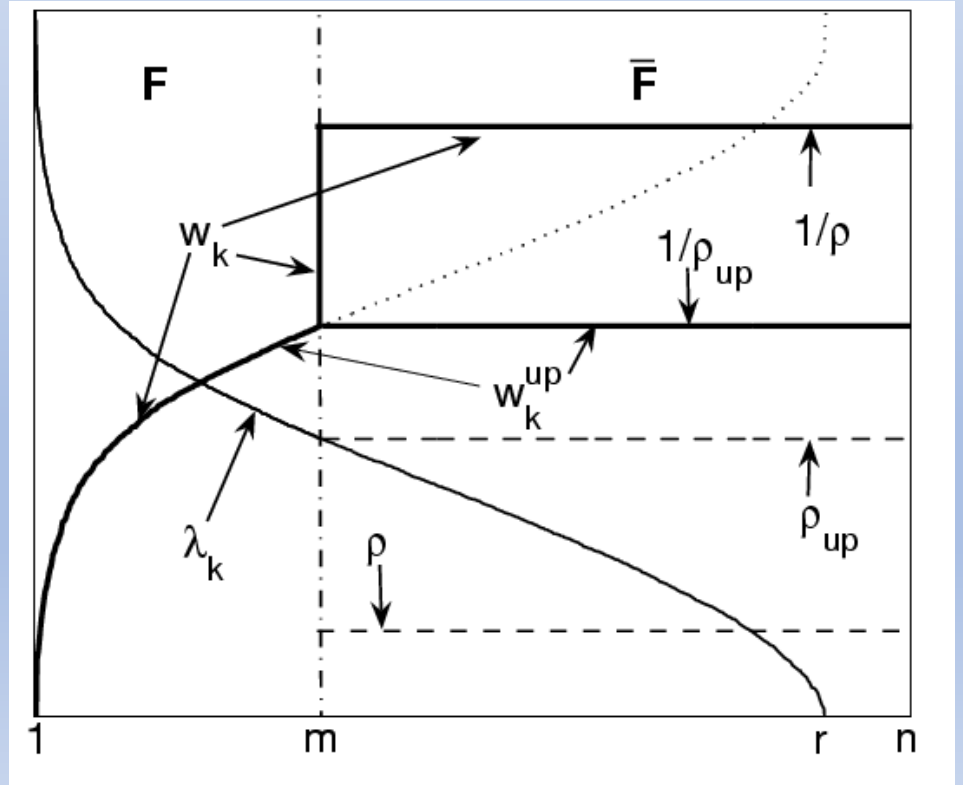
$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{m}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} - \frac{1}{2}\sum_{k=m+1}^{n}\frac{(z_k - \bar{z}_k)^2}{\rho_{up}} + b_i$$

$$\rho_{up} = \lambda_{m+1}$$

$$= \max(\lambda_k,\ k = m+1,\ldots,n)$$

X.D. Jiang, B. Mandal and A. Kot, Enhanced Maximum Likelihood Face Recognition, *Electronics Letters*, vol. 42, no. 19, pp. 1089-1090, September 2006.

recognition rate	number of training images		
method	500	1000	1400
ρ_{av}	89.76	90.44	89.70
ρ_{up}	91.03	93.03	93.64
$\rho_{up}\mathrm{R}^r$	92.90	95.94	96.46



Solution4: Dimensionality Reduction

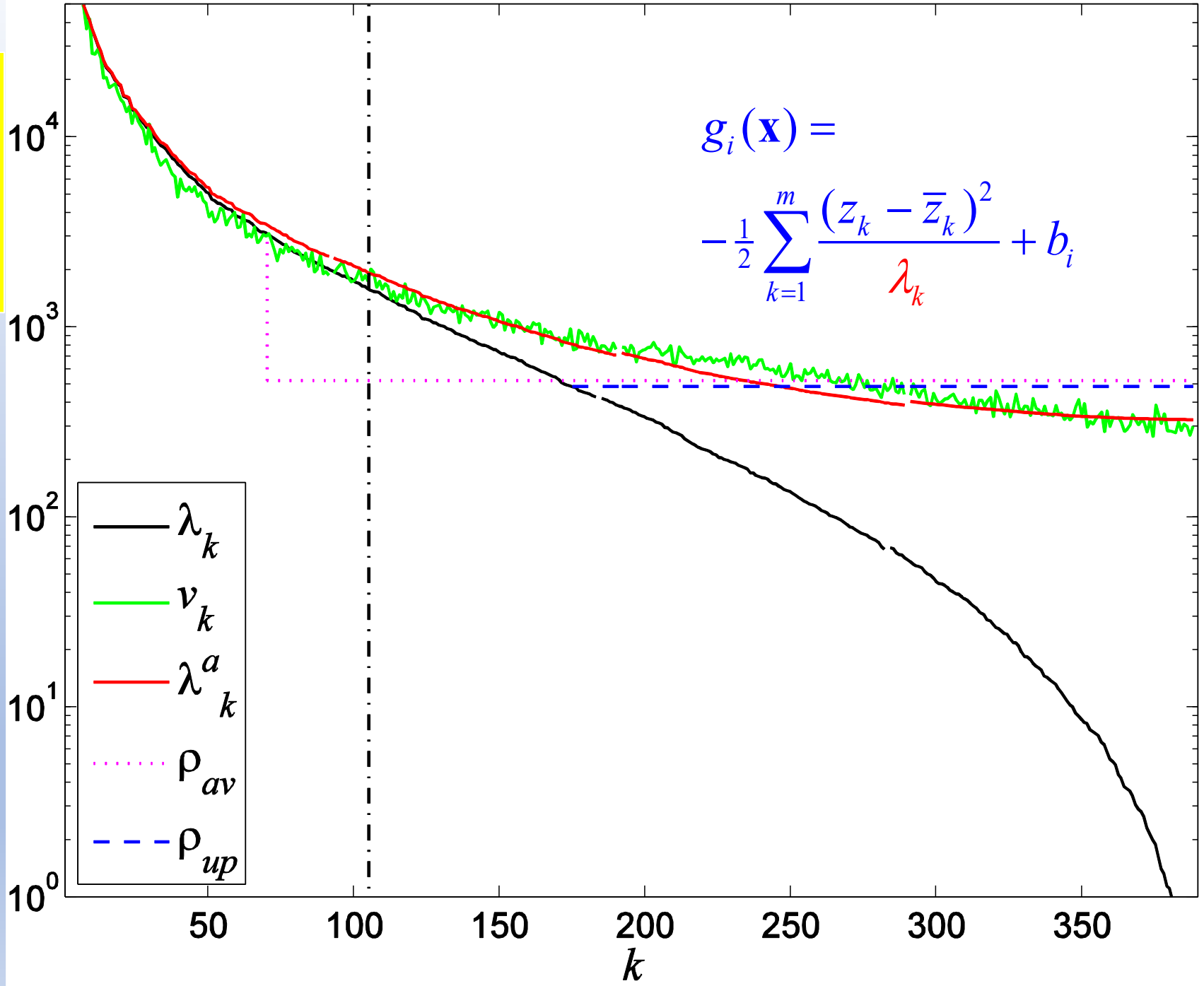
$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{m}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} - \frac{1}{2}\sum_{k=m+1}^{n}\frac{(z_k - \bar{z}_k)^2}{\rho} + b_i$$

- Why PCA, though an unsupervised method that minimizes the reconstruction error rather than maximizes the discrimination of classes, can improve the classification accuracy?

The underline principle behind PCA for classification is to remove dimensions corresponding to the unreliable small eigenvalues of the class-conditional covariance matrices.

$$\mathbf{S}^t = \mathbf{S}^w + \mathbf{S}^b = \sum_{j=1}^{c}\frac{L_j}{L}\boldsymbol{\Sigma}_j + \mathbf{S}^b$$

The black curve shows an eigen-spectrum obtained from 400 face images and the green curve shows the variances v_k of other 8,500 face images projected on the eigenvectors ϕ_k, in logarithm scale.



X.D. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, March 2011.

exdjiang@ntu.edu.sg

Solution6: Supervised/Asymmetric PCA

- Is the unsupervised PCA that minimizes the data reconstruction error optimal in the dimensionality reduction for classification?

$$\mathbf{S}^t = \frac{1}{q}\sum_{i=1}^{L}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$$

$$= \mathbf{S}^w + \mathbf{S}^b = \sum_{j=1}^{c}\frac{q_j}{q}\boldsymbol{\Sigma}_j + \mathbf{S}^b$$

- What is the problem of PCA for classification?

  Classes with more training samples are heavier weighted so more dimensions unreliable for theses classes are removed.

- How to modify PCA from an unsupervised method to a supervised one for an even better classification?
- How to modify PCA to handle unbalanced data and asymmetric classes?

Solution6: Supervised/Asymmetric PCA

PCA

$$\mathbf{S}^t = \frac{1}{q}\sum_{i=1}^{L}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$$

$$= \mathbf{S}^w + \mathbf{S}^b = \sum_{j=1}^{c}\frac{q_j}{q}\boldsymbol{\Sigma}_j + \mathbf{S}^b$$

It is not the more but the less reliable covariance matrix that should be heavier weighted in the covariance mixture so that more dimensions characterized by the small variances of this class can be removed.

New methods were proposed in

X.D. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, March 2011.

X.D. Jiang, "Asymmetric Principal Component and Discriminant Analyses for Pattern Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931-937, May 2009.

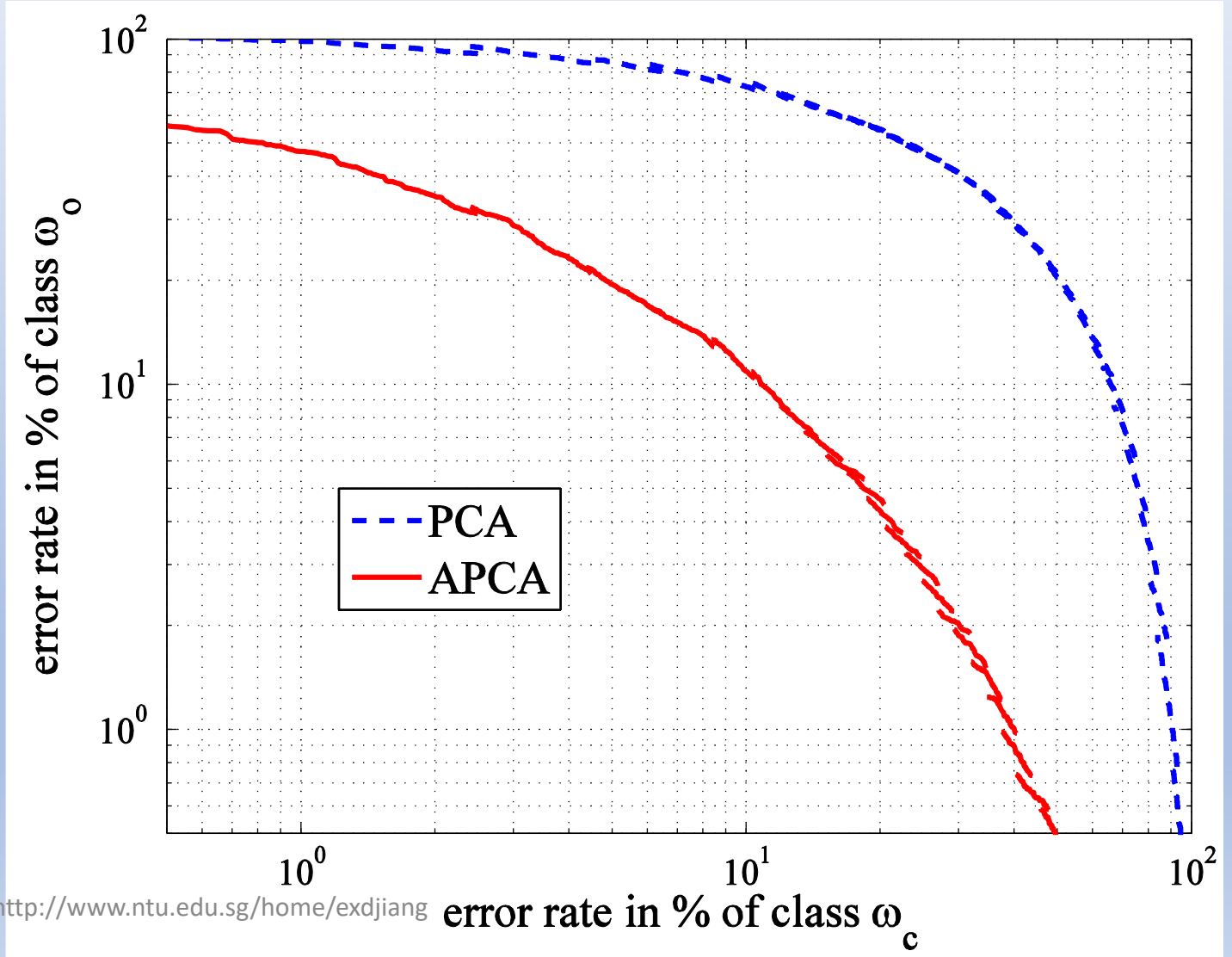
Asymmetric PCA: APCA $\quad \boldsymbol{\Sigma}_a = a\boldsymbol{\Sigma}_o + \beta\boldsymbol{\Sigma}_c + \eta\mathbf{S}^b$

Supervised PCA: SPCA $\quad \boldsymbol{\Sigma}_a = \sum_{j=1}^{c} a_j\boldsymbol{\Sigma}_j + \eta\mathbf{S}^b$

Solution6: Supervised/Asymmetric PCA

Classification performance comparison of different approaches

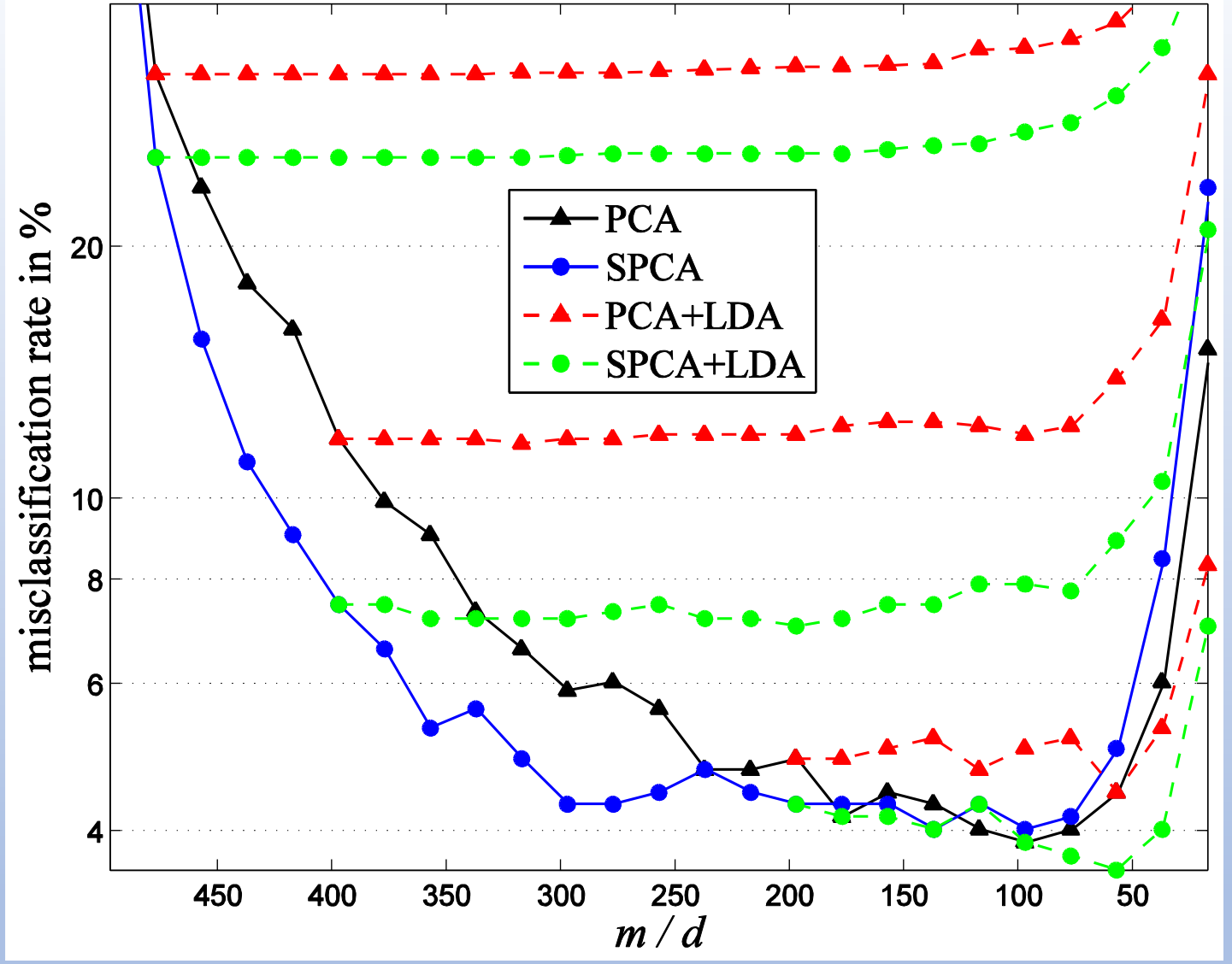
2,000/500 training samples and 20,000/5,000 testing samples of the 400 dimensional data of classes ω_o and ω_c. m=200, a=0.2



X.D. Jiang, "Asymmetric Principal Component and Discriminant Analyses for Pattern Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931-937, May 2009.

Misclassification rate against the reduced dimensionality by PCA/SPCA and PCA/SPCA+LDA of face identification problems. 497 people are randomly selected for training and other 697 people are used for testing Image size: 33X38.

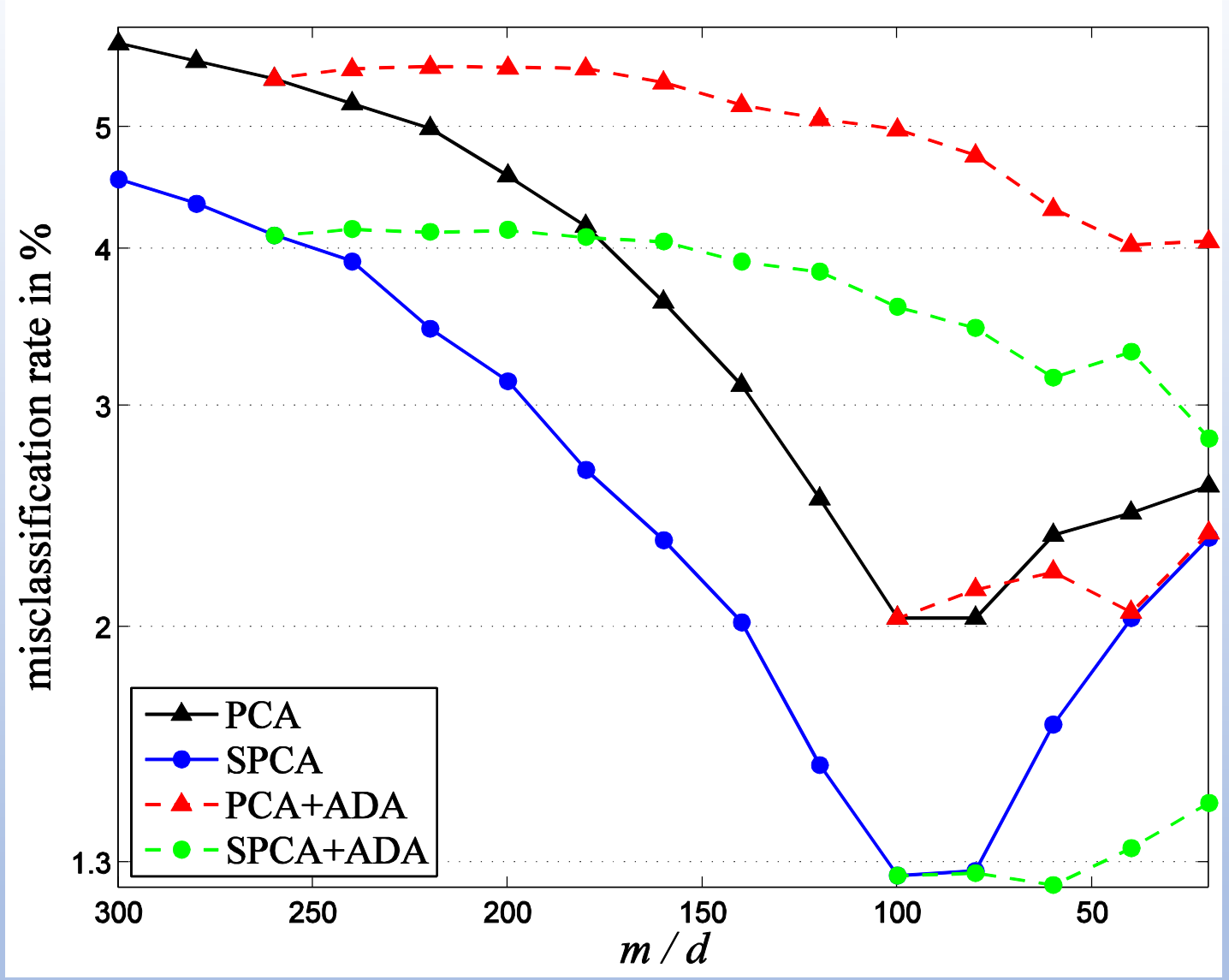
X.D. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, March 2011.



The most left point of each dashed curve indicates the dimensionality m of the PCA/SPCA subspace in which LDA/ADA further reduces it to d indicated by the other points on the same dashed curve.

Misclassification rate against the reduced dimensionality by PCA/SPCA and PCA/SPCA+ADA of face detection problems.
9000 face images and 9000 non-face images are used in training (75%) and testing (25%).
Image size: 20X20

X.D. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, March 2011.



The most left point of each dashed curve indicates the dimensionality m of the PCA/SPCA subspace in which LDA/ADA further reduces it to d indicated by the other points on the same dashed curve.

Feature Extraction/Dimension Reduction v. Machine Learning

- Although the ultimate objective of a recognition system is to extract the most discriminative information, it is on the whole data population, not on a specific training set. A classifier is trained to capture the most discriminative information on the training samples. If some statistics estimated on the training data deviate from those of the data population, misclassification rate on the novel data increases.

- The most discriminative criterion cannot well solve this problem because it in general just repeats the classification process. Maximizing the extracted just means minimize the loss, doe not mean gain. Redundant information does not mean harmful for classification.

- Therefore, to boost the classification accuracy, the dimensionality reduction by machine learning should be targeted at removing the dimensions unreliable (harmful) for the classification or regularizing (correcting) the unreliable statistics in these dimensions.