Lect 4-5 Dimension reduction techniques

Common techniques for dimension reduction.

1. Principal component analysis (PCA): [generative, global, linear]
2. Fisher linear discriminant analysis: [discriminative, global, linear]
3. Independent component analysis (ICA)
4. Multi-dimensional scaling (MDS) [generative, global, non-linear]
5. Local Linear embedding (LLE) [generative, local, linear]
6. Transformed component analysis (TCA)

Some features are generative and some are discriminative.

Lecture 4: Dimension Reduction I

One of the recurring problems encountered in applying statistical techniques to pattern recognition problem is the so called *curse of dimensionality*: [note that this curse is on modeling, not necessarily for discrimination].

For example, let the feature space be 50 dimensional, i.e. $\mathbf{x} = (x_1, x_2, \ldots, x_{50})$ and suppose that each dimension is divided to $L = 20$ discrete levels, then

$$\mathbf{x} \in \Omega = \{1, 2, \ldots, 20\}^{50}.$$

Ω has 20^{50} cells, on which the class models $p(\mathbf{x}|\omega_i)$ are defined. However the number of observed samples n is, in general, much smaller than 20^{50}, thus no observations are available for most of the cells in Ω.

One common method is to assume smooth density functions in empty spaces, e.g.

- Maximum entropy – the parametric methods

$$p(\mathbf{x}) = \frac{1}{Z} e^{-\sum_{i=1}^{k} \lambda_i \phi_i(\mathbf{x})}$$

- Window function – the non-parametric methods

$$p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta_n(\mathbf{x} - \mathbf{x}_i).$$

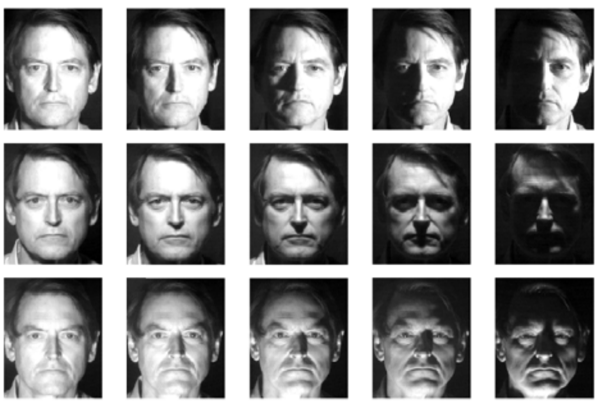
Properties of data in space

In the following, we will illustrate and characterize the properties of the data in space to draw some intuitions.

We will use image data as example as they are very meaningful and visible to our eyes.

Overview: what are the real dimensions of your data?

For a human face image of 128 x 128 pixels, what is the dimension of all images of a same person under varying illumination? It must be quite small. Appearance Variations



The non-linearity occurs when there is cast shadow on face.

Geometric variations: Expression

The the geometric deformations are also low-dimensional.

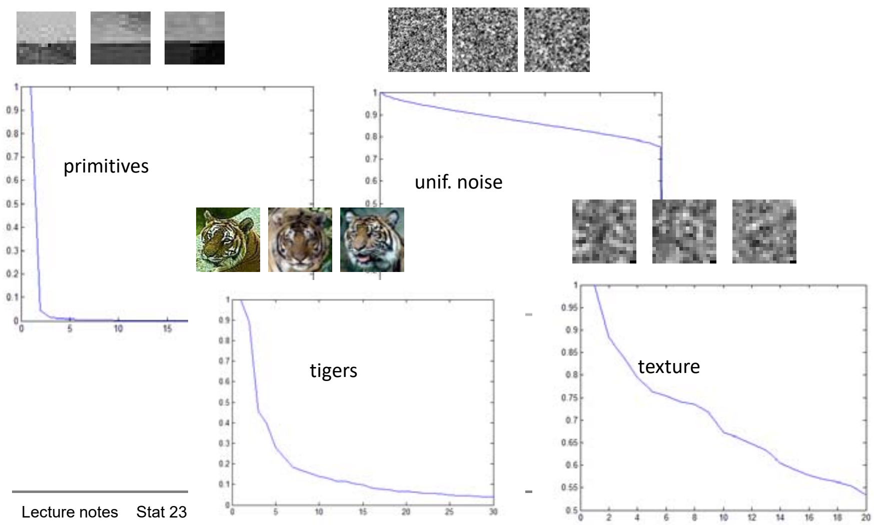
So, our project no.1 will explore the geometry and appearance dimensions.

Lecture notes Stat 231-CS276A © S.C. Zhu

A wide spectrum of categories from low to high entropy

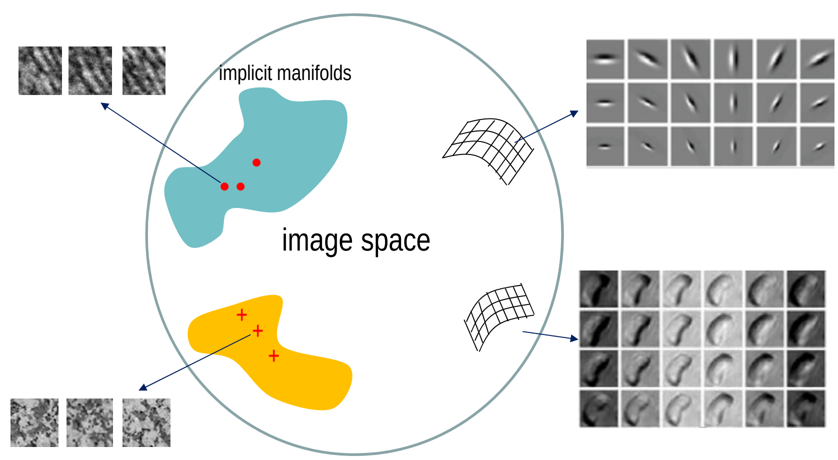
Entropy ~ Dimension ~ Log volume(manifold)

Take 16x16 image patches (256-space), run PCA for each category, and plot the eigen-values in decreasing order.



A look at the space of image patches

Each is an "equivalency class" of images, and different metrics are used in such spaces.



By Analogy: a cosmology picture

The real dataset is often

a mixture of many subspaces of different dimensions.

A clustering technique is to separate subspaces.

Don't use clustering technique blindly.
We must be aware of the structures of the data. Key issues:

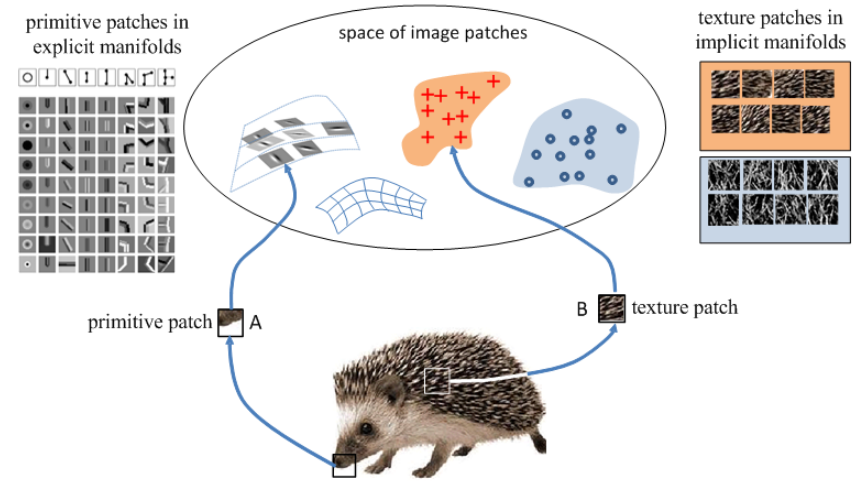
1, Varying dimensions
2, Scaling and transition
3, Compositional structures

For images, Information scaling leads to transitions !

Scaling (zoom-out) increases the image entropy (dimensions)

Ref: Wu, Zhu, Guo, "From Information Scaling of Natural Images to Regimes of Statistical Models," *Quarterly of Applied Mathematics*, 2007.

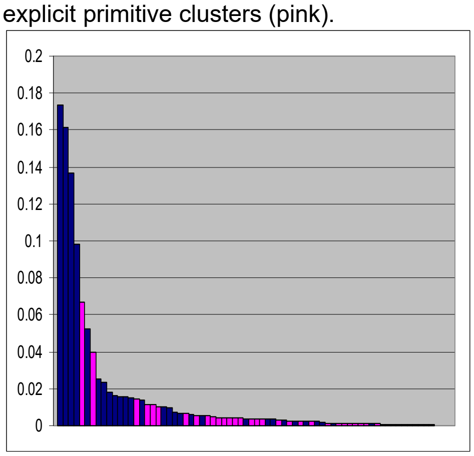
Another common observation: Compositionality

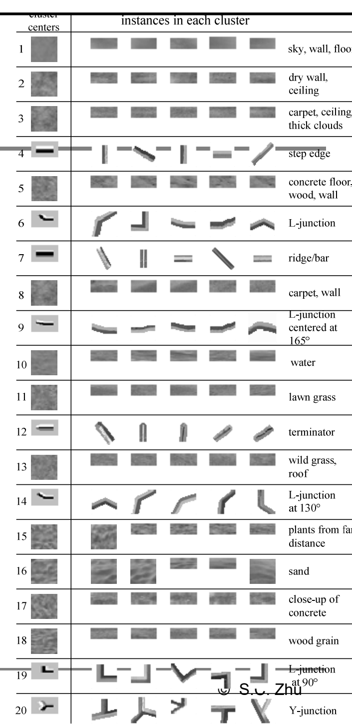


Z. Si and S.C. Zhu, "Learning Hybrid image Template (HiT) by Information Projection," IEEE Trans. on PAMI, 2012.

Clusters in natural images

implicit texture clusters (blue),
explicit primitive clusters (pink).



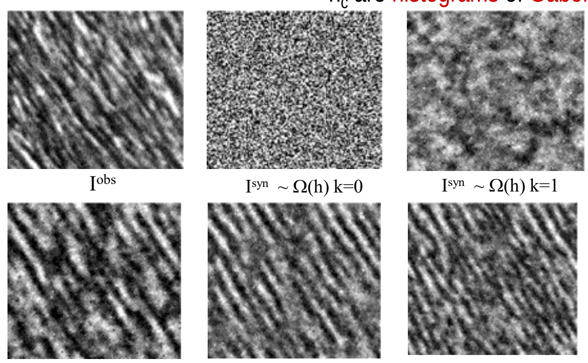


Dimension reduction in the texture spaces: PCA does not working here

$$\text{a texture} = \Omega(h_c) = \{\, I : \ h_i(I) = h_{c,i} \ , i = 1,2,\ldots,K \,\}$$

h_c are histograms of Gabor filters

This example also shows how we may define a *pattern* as a *set*, and how we can visualize a pattern by simulation through Markov chain Monte Carlo (MCMC)

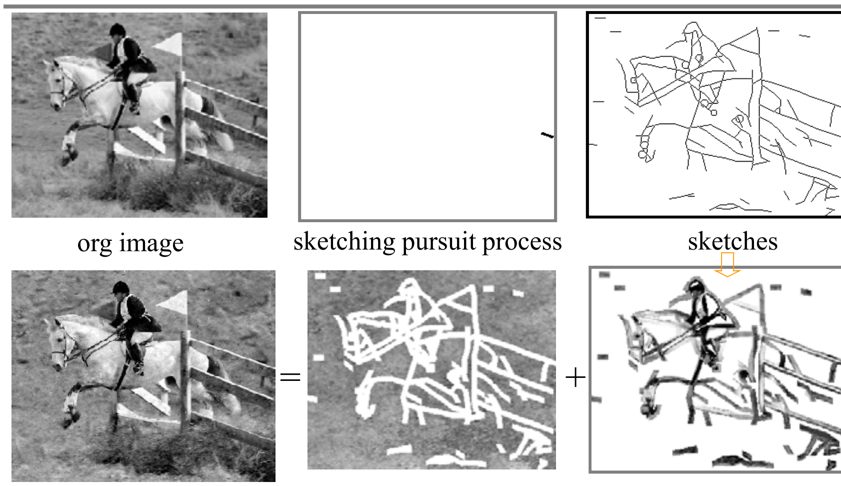


I^{obs} $\quad$ $I^{syn} \sim \Omega(h)$ k=0 $\quad$ $I^{syn} \sim \Omega(h)$ k=1

$I^{syn} \sim \Omega(h)$ k=3 $\quad$ $I^{syn} \sim \Omega(h)$ k=4 $\quad$ $I^{syn} \sim \Omega(h)$ k=7

Advanced topic: integrating the two regimes for coding

Here we use two metrics in the two regimes. Marr, 1982, Guo, Zhu, and Wu, 2003-05.

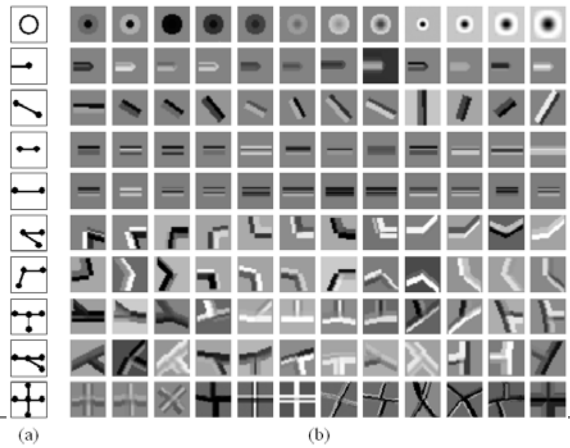


org image $\quad$ sketching pursuit process $\quad$ sketches

syn image = synthesized textures + sketch image

Image primitives as simple ASM model (just like the face)

Learned texton/primitives (dictionary, codebook) with some landmarks that transform and warp the patches.



Primal sketch example

original image   synthesized image   sketching pursuit process

One common dimension reduction technique: PCA

The principal component analysis (PCA), also called Karhunen-Loeve transform in functional space, is widely used for dimension reduction. In vision, it becomes popular by the eigen-face example. There are many ways to derive PCA, here we study it from the perspective of dimension reduction.

Given: a number of n samples { $x_1, x_2, \ldots, x_n$} in d-space.

Objective: project it in a d'< d space, that is, approximate each vector x_k by

$$m + \sum_{i=1}^{d'} a_{ki} e_i \quad \rightarrow \quad x_k$$

Criterion: minimize the sum of squared error.

$$J_{d'}(m, a, e) = \sum_{k=1}^{n} \| (m + \sum_{i=1}^{d'} a_{ki} e_i) - x_k \|^2$$

PCA

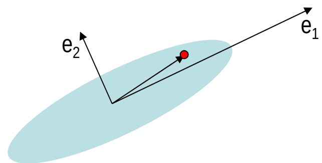
The result of minimizing the error is:

m is the sample mean,

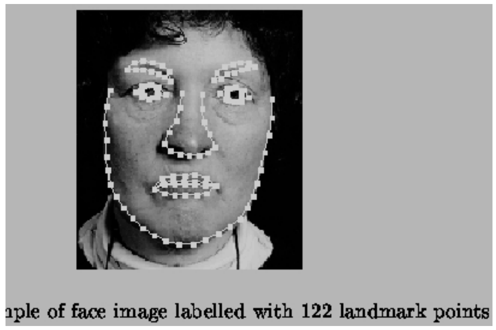
e_i is the i-th largest eigen-vector of the co-variance matrix

a_{ki} is the projection of x_k to e_i

The book derives this in three separate steps. As this is so well-known, we don't unfold the details.

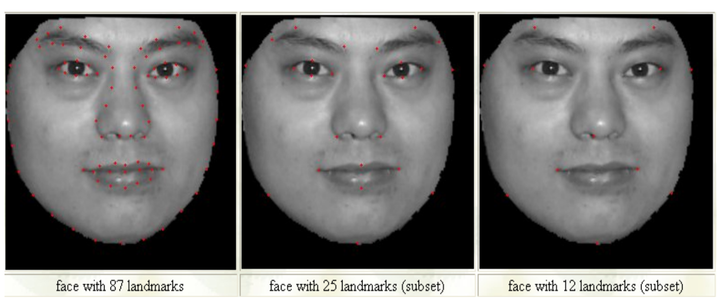


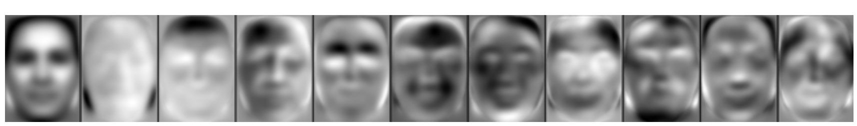
Example on face representation



400 images each labeled with 122 points.

Project I. Human face modeling





Eigen-faces without alignment

Eigen-vectors for Geometry and Photometry

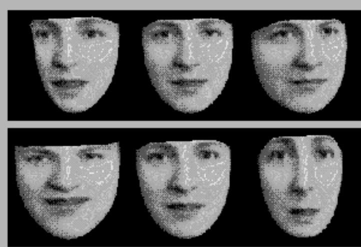


Fig. 2. First two modes of shape variation (± 3 sd)

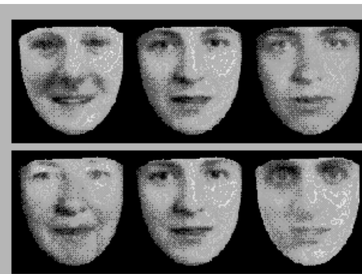
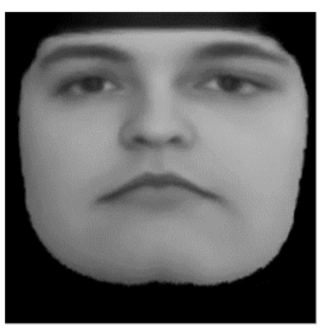


Fig. 3. First two modes of grey-level variation (± 3 sd)

http://vimeo.com/user1158726/videos

Demo made by Maria Pavlovskaia: showing the meanings of various axes.



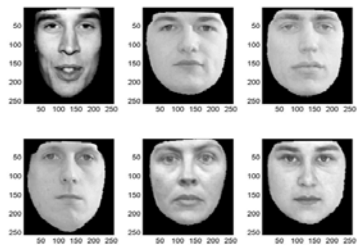
Sliding in an axis for geometric changes

Sliding an axis for appearance changes

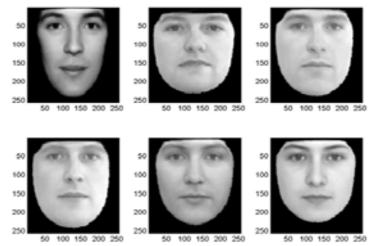
Project I. Human face modeling

Results from previous student:

Face examples in the dataset

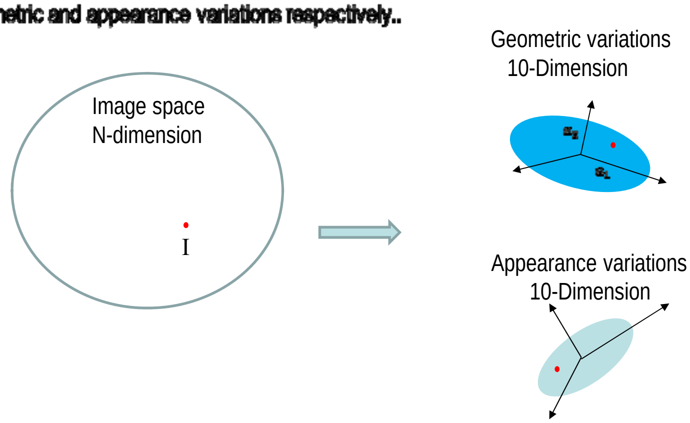


Face (lossy) reconstructed by 20 numbers



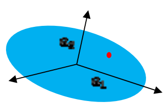
Project I. Human face modeling

The essence of this AAM (Active Appearance Model) is to transfer each face image I in N (=number of pixels) dimensional image space to $(a_1, \ldots a_{10})$ and $(b_1, \ldots, b_{10})$ in two independent subspaces spanned by the two groups of eigen-vectors for geometric and appearance variations respectively..

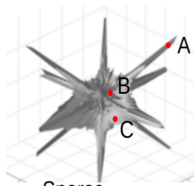


Discussion on advanced topics

The assumption of Gaussian distributions in the projected space is roughly right, but not precise. The point cloud of real data usually is not really an ellipsoid, but has “horns” or “spikes” – which indicates sparse representation



Compact Gaussian



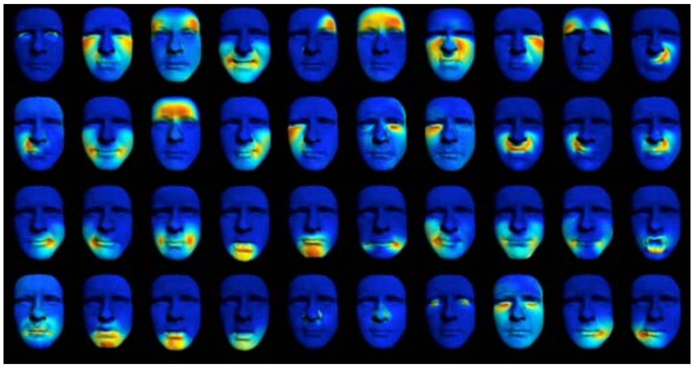
Sparse Super-Gaussian

In this simple example, Data points live in respectively
1D (point A)
2D (point B)
3D (point C)
subspaces.

Discussion on advanced topics

What does the “horn” or “spike” mean?

Suppose we have K=40 muscles that controls our expressions, each time, we may only use 2-4 muscles (i.e. sparsity) for each of our expression, thus the variations of each expression lies in a 2-4 dimensional subspace --- the horn. We have a Conbinatoral number of such sub-spaces (selecting 2-4 from 40).



Generative Basis from Yu et al., *Computer & Graphics*, 2012 (from P. Schyns)

Discussion on advanced topics

Are these animals smiling?

How did we (humans) perceive this, even though we never saw a smiling sheep before?

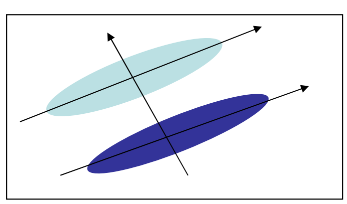
Reasoning and learning by mirroring actions

What is the pig doing? How do you figure out?

Another problems with PCA: in classification

The PCA reduces dimension for each class individually. The resulting components are good representation of the data in each class, but they may not be good for discrimination purpose.

For example, suppose the two classes have 2D Gaussian-like densities, represented by the two ellipsis. They are well separable. But if we project the data to the first principal component (i.e. from 2D to 1D), then they become inseparable (with a very low Chernoff information). The best projection is the short axis.



In fact, this is typical problem in studying generative models vs discriminative models. The generative models aim at representing the data faithfully, while discriminative models target telling objects apart.

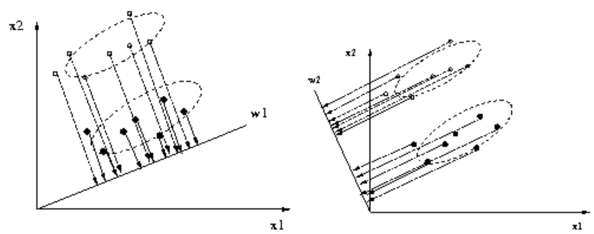
Fisher linear discriminant

In a two-class classification problem, given:

n samples $x_1, x_2, ..., x_n$ in a d-dimensional feature space, n_1 in subset χ_1 labeled ω_1 and n_2 in subset χ_2 labeled ω_2.

goal:

to find a vector w, and project the n samples on this axis $y = w^t x = < w, x >$, so that the projected samples are well separated.



Fisher linear discriminant

Definition:

The **sample mean** for class ω_i:

$$m_i = \frac{1}{n_i} \sum_{x \in \chi_i} x, \qquad i = 1, 2.$$

The **scatter matrix** for class ω_i:

$$S_i = \sum_{x \in \chi_i} (x - m_i)(x - m_i)^t, \qquad i = 1, 2.$$

The **between-class scatter matrix**

$$S_B = (m_1 - m_2)(m_1 - m_2)^t$$

The **within-class scatter matrix**

$$S_W = S_1 + S_2$$

Fisher linear discriminant

The **sample mean** of the projected points in class ω_i:

$$\tilde{m}_i = \frac{1}{n_i} \sum_{\mathrm{x} \in \chi_i} w^t \mathrm{x} = w^t \mathrm{m}_i, \qquad i = 1, 2.$$

The **scatter** of the projected points in class ω_i:

$$\tilde{s}_i = \sum_{\mathrm{x} \in \chi_i} (w^t \mathrm{x} - w^t \mathrm{m}_i)^2 = w^t \mathrm{S}_i w, \qquad i = 1, 2.$$

These are 1D variables

Fisher linear discriminant

Fisher's linear discriminant $y = w^t \mathrm{x}$: choose w to maximize

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{w^t \mathrm{S}_B w}{w^t S_W w}.$$

i.e. the between-class distance should be as large as possible, and the within class scatter should be as small as possible.

This is the Rayleigh quotient.

Fisher linear discriminant

Proposition The vector w that maximizes the criterion function

$$J(w) = \frac{w^t S_B w}{w^t S_W w}$$

is

$$w = S_W^{-1}(m_1 - m_2)$$

This is the so-called Rayleigh quotient, whose maxima are the eigen-values.

[Proof]: Suppose w^* is the optimal solution and

$$\lambda = J(w^*) = \frac{{w^*}^t S_B w^*}{{w^*}^t S_W w^*}$$

is the maximum. As $J(aw) = J(w)$, we only need to check $\|w\| = 1$. Note that $J(w)$ is differentiable.

As $\frac{d}{dw}(w^t S w) = 2Sw$ for a symmetric matrix S, we have

$$\frac{dJ(w)}{dw} = \frac{2S_B w}{w^t S_W w} - J(w)\frac{2S_W w}{w^t S_W w}.$$

As $\frac{dJ(w^*)}{dw} = 0$, we have,

$$S_B w^* = \lambda S_W w^*$$

and thus

$$S_W^{-1} S_B w^* = \lambda w^*$$

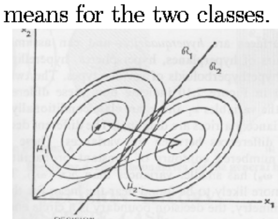
The conclusion follows $S_B w = c(m_1 - m_2)$. which is by definition of S_B

Fisher linear discriminant

Observation: In the Bayes decision study before, for two classes of Gaussian distribution with variance matrix $\Sigma_1 = \Sigma_2 = \Sigma$, the Bayes decision boundary is a straight line whose normal is the Fisher linear discriminant

$$w^t x + w_0 = 0, \quad w = \Sigma^{-1}(\mu_1 - \mu_2)$$

where μ_1, μ_2 are the means for the two classes.



In this special case the Fisher's linear discriminant coincided with the Bayes decision boundary.

$$w = S_W^{-1}(m_1 - m_2)$$

with $S_W = 2\Sigma$. Note w is perpendicular to the decision boundary.

Multiple discriminant analysis

For c classes, we compute c-1 discriminants, that is, to project the d-dimensional features into (c-1)-space. The linear discriminant is a special case with c=2.

For example, C=3.
The 2-discriminants span a 2D plane, the left projection is better than the right

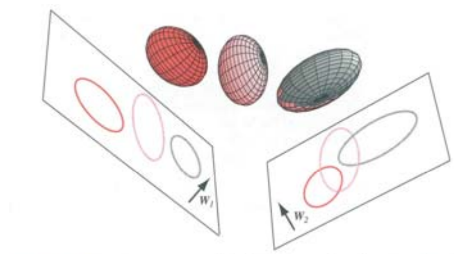


FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$.

Multiple discriminant analysis

For c-class problem, we first generalize the scatter measures.

The **within-class scatter matrix**:

$$S_W = S_1 + S_2 + \ldots + S_{c-1}$$

where S_i is the scatter matrix computed from samples inside class ω_i

The **between-class scatter matrix**:

$$S_B = S_{total} - S_W = \sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^t$$

where S_{total} is the total scatter matrix computed from all samples treat in one big class .

The scalar measure of the scatter matrix S is the determinant of the matrix $|S|$.

Multiple discriminant analysis

We seek vectors $w_i, i = 1, 2, ..., c-1$, and project the samples from d-dimension feature space $\mathrm{x} = (x_1, x_2, ..., x_d)$ to the $c-1$ dimensional space $\mathrm{y} = (y_1, y_2, ..., y_{c-1})$:

$$\mathrm{y} = (w_1^t \mathrm{x}, w_2^t \mathrm{x}, ..., w_{c-1}^t \mathrm{x}) = W^t \mathrm{x}$$

where W is a $(c-1) \times d$ matrix with w_i being the i-th column.

The criterion for the optimal W is

$$J(W) = \frac{|W^t \mathrm{S}_B W|}{|W^t S_W W|}$$

The solution is the eigenvectors whose eigenvalues are the $c-1$ largest in

$$\mathrm{S}_B w = \lambda \mathrm{S}_W w$$

There are standard c/Pascal/Fortran codes for computing the eigenvalues and eigenvectors in the book of *Numeric Analysis*

Examples in applications

Generally speaking, in pattern recognition, vision and speech, the feature extraction using filters can be viewed as a problem of dimension reduction.

For example, raw images and speech signals are represented as very long vectors, each filter response in the feature space is actually a function (or linear combination) of the original data representation.

How to design filters and wavelets?

This problem has been intensively studied in Neurosciences (including neural networks). As neurons at the primary visual cortex are considered as feature extractors, the principles for dimension reduction should hold the key to understanding the functions of nerve cells in the visual cortex. As people believe that the nerve system should adopt the optimal coding strategy in the evolution process.

Example

Task: glass vs no-glass in a face image

Compare the principal conponent analysis and Fisher discriminant analysis

a face image

a Fisher-face image

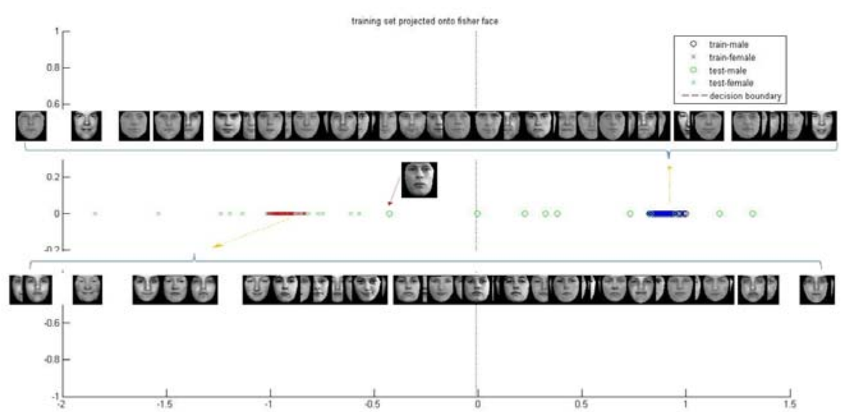
COMPARATIVE RECOGNITION ERROR RATES FOR GLASSES/ NO GLASSES RECOGNITION USING THE YALE DATABASE

Glasses Recognition		
Method	Reduced Space	Error Rate (%)
PCA	10	52.6
Fisherface	1	5.3

Ref. P.N. Belhumeur et al, "Eigenfaces vs FisherFaces…", IEEE Trans. PAMI Vol19, no7, 1997.

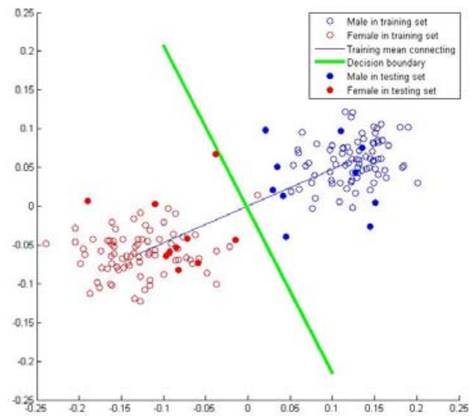
Project I, part 2: gender discrimination

Using only 1 Fisher face



Project I, part 2: gender discrimination

Using 2 Fisher faces: one for geometry and one for appearance



Discussion

In this FLD example, we showed that we may find 1-2 dimensions, i.e. ω so that the projected point on this axis can separate the two classes:

$$< \omega, I > \quad \le \quad < \omega, J > \qquad \forall I \in \Omega_{male}, \ \forall J \in \Omega_{female}.$$

In other methods, i.e. the Support Vector Machines, one simply map the image into a higher Dimensional space, such that,

$$< \omega, \phi(I) > \quad \le \quad < \omega, \phi(J) > \qquad \forall I \in \Omega_{male}, \quad \forall J \in \Omega_{female}.$$

Advanced examples: gender recognition

Results by Seyoung Park, UCLA CS Ph.D student.