Computational Intelligence: Methods and Applications

Lecture 7
Discriminant Component Analysis

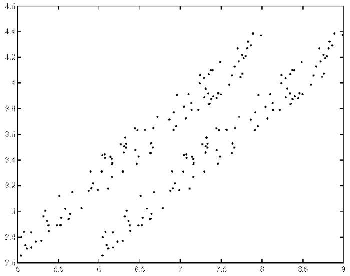
Włodzisław Duch

SCE, NTU, Singapore

Google: Duch

PCA problem

PCA transformation for 2D data:



PCA give worst possible solution here from the point of view of seeing the structure of the data.

PCA is completely unsupervised, knows only about variance, but nothing about different classes of data.

Goal: find direction of projection that shows more structure in the data, using class structure; this will then be supervised approach.

“Discriminant coordinates” should reveal class structure much better.

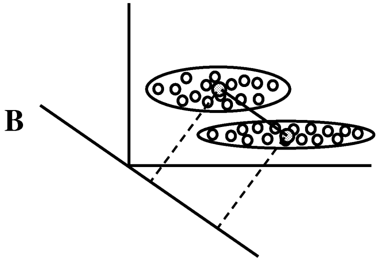
Maximize separation

PCA does not find the best combination to reveal the structure of data. No information about data structure, class labels, is used in PCA.

If class info is available (or clusterization has been performed):

- find mean vectors,
- find projection direction that maximizes separation of means
- for 2 classes:

$$\mathbf{B} = \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) \big/ \left\|\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right\|$$



B direction ||B||=1 defines line passing through 0, the value of projection $Y=B^T \cdot X$

If data clusters are spherical B is a good choice; find more projection directions B_i in the space orthogonal to B_1=B.
Here direction X_2 is a better choice!

Reminder: orthogonal projection

If one interesting direction B_1 has been found in the n-vector $X=[X^{(1)},X^{(2)\ldots}X^{(n)}]$ data space, each with d dimensions $X^{(1)}=(X_1 \ldots X_d)$ then orthogonal space is created using projection operator:

$$\mathbf{P}_1 = \mathbf{I} - \mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}$$

Here P_1 is a dxd matrix that applied to X rotates the data space leaving only components that are orthogonal to B_1 (normalized vector, $||B_1||=1$).

$$\mathbf{P}_1\mathbf{B}_1 = \left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right)\mathbf{B}_1 = \mathbf{B}_1 - \mathbf{B}_1 = 0$$

$$\mathbf{P}_1\mathbf{B}_2 = \left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right)\mathbf{B}_2 = \mathbf{B}_2$$

if B_2 is orthogonal to B_1

Create X’ = P_1 X and repeat the same procedure to find an interesting orthogonal vector B_2.
Projection operator orthogonalizing to the $[B_1,B_2]$ space is simply $P_2 = I-B_1B_1^T-B_2B_2^T$, adding more terms for higher P_n.

Greater separation

This is not an optimal solution, although it is better than PCA ...
Covariance of the two distributions should be taken into account:

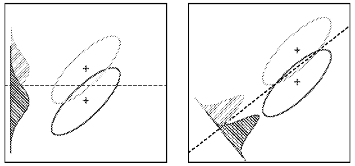


Figure from: Chapter 4, Elements of statistical learning.
By Hasti, Tibshirani and Friedman 2001

Within-class scatter

Find transformation Y = W$^{\mathrm{T}}$·X, that maximizes the distance of projected mean values:

$$\left|\bar{Y}_1-\bar{Y}_2\right|=\left|\mathbf{W}^{\mathrm{T}}\cdot\left(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\right)\right|$$

But scaling W may create arbitrarily large differences!

This distance should be large relatively to the variance (scatter):

$$s_{Y,i}^2=\sum_{j=1}^{n(C_i)}\left(Y_i^{(j)}-\bar{Y}_i\right)^2$$

Within-class scatter, or variance without normalization constant 1/(n-1).
i=1,2 class of vectors, j=1..$n(C_i)$ sample

$$s_Y^2=\sum_{i=1}^{N_C}s_{Y,i}^2$$

Total within-class scatter;
variances could be also used.

Fisher criterion

Maximize the Fisher criterion function:

$$J(\mathbf{W})=\frac{\left|\bar{\mathbf{Y}}_1-\bar{\mathbf{Y}}_2\right|^2}{s_{Y,1}^2+s_{Y,2}^2}$$

This is maximum for large separation and small within-class variance.

It does not depend on the norm of W, only on the direction.

$$s_{Y,k}^2=\sum_{j=1}^{n(C_k)}\left(\mathbf{W}^{\mathrm{T}}\cdot\mathbf{X}^{(j)}-\mathbf{W}^{\mathrm{T}}\cdot\bar{\mathbf{X}}_k\right)^2$$

$$=\sum_{j=1}^{n(C_k)}\mathbf{W}^{\mathrm{T}}\left(\mathbf{X}^{(j)}-\bar{\mathbf{X}}_k\right)\left(\mathbf{X}^{(j)}-\bar{\mathbf{X}}_k\right)^{\mathrm{T}}\mathbf{W}=\mathbf{W}^{\mathrm{T}}\mathbf{S}_k\mathbf{W}$$

$$\mathbf{S}_k=\sum_{j=1}^{n(C_k)}\left(\mathbf{X}^{(j)}-\bar{\mathbf{X}}_k\right)\left(\mathbf{X}^{(j)}-\bar{\mathbf{X}}_k\right)^{\mathrm{T}}$$

$\mathbf{S}_k$ is the within-class scatter matrix for class C_k.

Fisher linear discriminant

Total within-class scatter matrix $S_I = S_1+S_2$

The difference of the means may be written as:

$$\left|\bar{Y}_1-\bar{Y}_2\right|^2=\mathbf{W}^{\mathrm{T}}\left(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\right)\left(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\right)^{\mathrm{T}}\mathbf{W}=\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}$$

S_B is the between-class scatter matrix.

Fisher criterion is:

$$\max_{\mathbf{W}}J(\mathbf{W})=\frac{\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}}{\mathbf{W}^{\mathrm{T}}\mathbf{S}_I\mathbf{W}}$$

This is a Rayleigh quotient, and the solution to this maximization problem is quite simple using variational techniques.

This procedure defines FDA = Fisher Discriminant Analysis.

FDA solution

Small perturbation of W around maximum should not influence *J*(W), therefore the two scalars below should be proportional:

$$\mathbf{S}_B\mathbf{W} = \lambda\mathbf{S}_I\mathbf{W}$$

This is generalized eigenvalue equation, but since S_I may be inverted:

$$\mathbf{S}_I^{-1}\mathbf{S}_B\mathbf{W} = \lambda\mathbf{W}$$

it becomes standard eigenvalue problem. First direction is found from:

$$\mathbf{S}_B\mathbf{W} = \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^{\mathrm{T}}\mathbf{W} = \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)\alpha\left(\mathbf{W}\right)$$

$$\mathbf{W} = \mathbf{S}_I^{-1}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)\alpha\left(\mathbf{W}\right)/\lambda; \quad \mathbf{W} \leftarrow \mathbf{W}/\|\mathbf{W}\|$$

because only the direction of W, not the length, is important, and α(W) is a scalar that does not change direction.
More directions are found from the eigenequation or repeating the procedure in subspace orthogonal to W.

FDA second vector

FDA is frequently used for classification, projecting data on a line.

For visualization generating the second FDA vector in a two-class problem is not so trivial.

This is due to the fact that the rank of the S_B matrix for the 2-class problems is 1 (why?), or for the K class problems it is K-1.

Also S_I may be ill-conditioned.

Solutions:

pseudoinverse for S_I makes it more stable;

perturbations of the S_I values may help.

Please consult A. Webb, Statistical pattern analysis, Chapter 4.3.3 for detailed discussion of numerical problems.

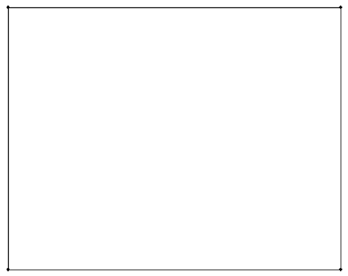
Example: FDA projections of hypercubes (using Matlab).

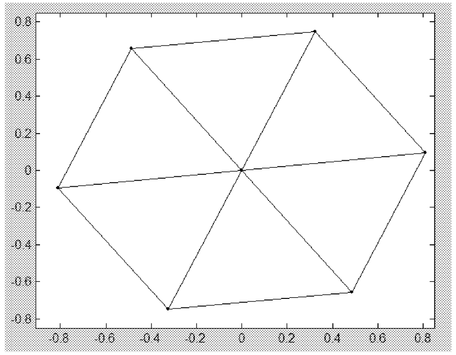
PCA for 3D cube

Example: Matlab PCA solution for projections of hypercubes.

For 3D cube taking all 8 vertices, from (0,0,0) to (1,1,1):

PCA selects one of the walls.



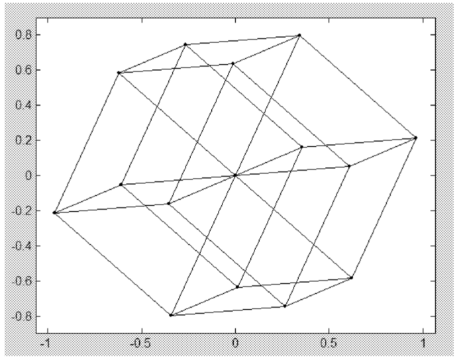


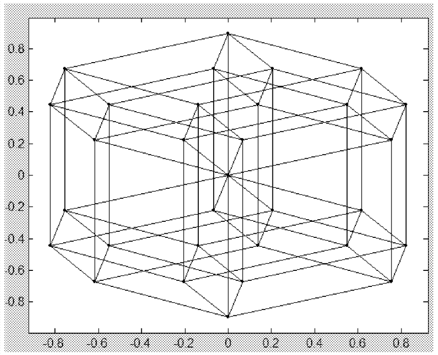
Removing (0,0,0) and using only 7 vertices better structure is shown.

PCA for 4/5D hypercube

For 4D cube taking 15 vertices, from (0,0,0,1) to (1,1,1,1):

For 5D cube taking 31 vertices, from (0,0,0,0,1) to (1,1,1,1,1):

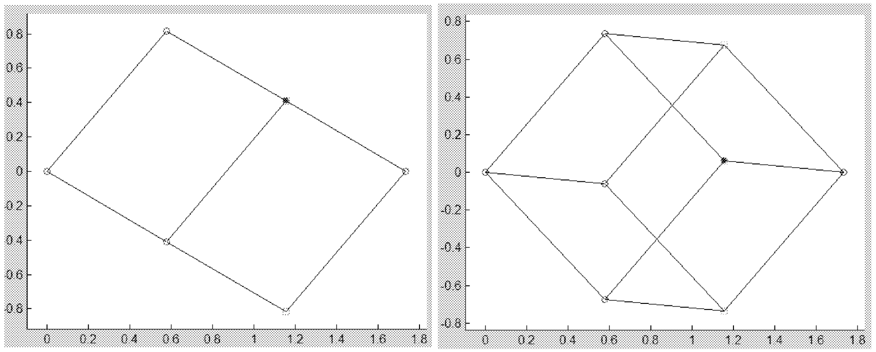




FDA for 3D cube

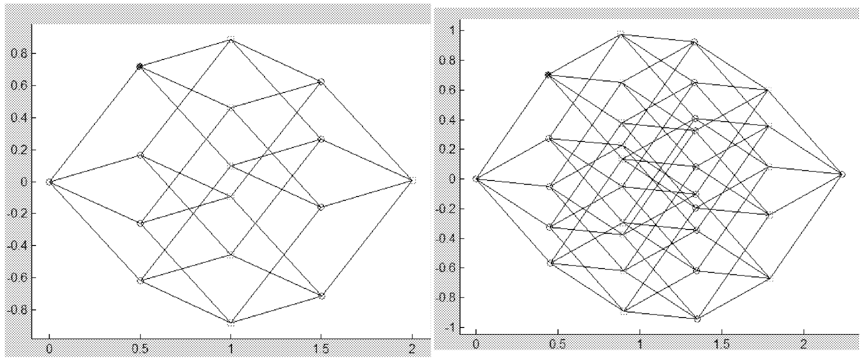
FDA requires classes: here odd/even parity of vertex bit strings.

Left: pseudoinverse S_I , right – perturbed S_B matrix



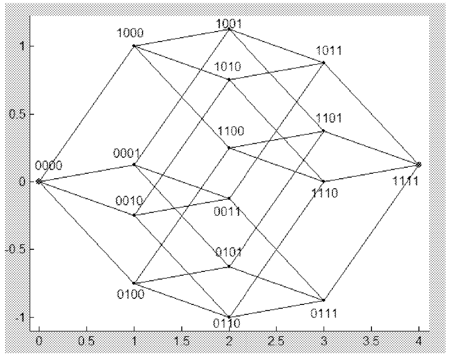
FDA for 4/5D cube

With perturbed S_B matrix: note that vertices from the same class (same parity) are projected on a vertical line; perturbation does not make them uniformly spaced along this line.



Lattice projection

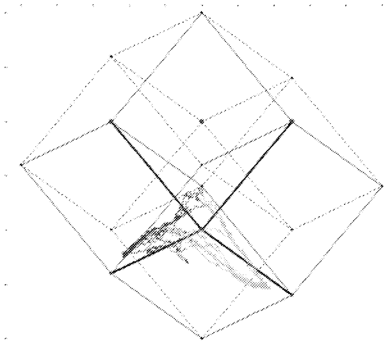
For normalized data $X_i \in [0,1]$ FDA projection is close to the lattice projection, defined as W_1=[1,1,..1] direction and W_2 maximizing separation of the points with fixed number of 1 bits.

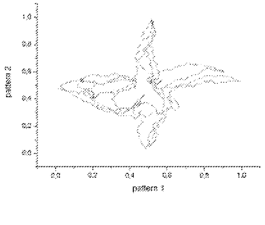


Lattice projections may be quite useful up to 6 dimensions.

Dynamical lattice projection

The state of a dynamical $X(t)$ system in 4 dimensions changes in time. Where are the attractors and what is this system doing?





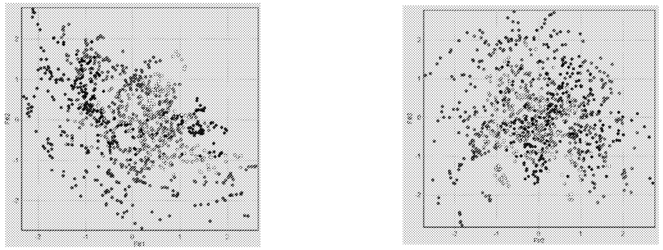
Seems to be going between 1000 and 0100 and 1100.

Second system presented in 2 PCA coordinates has initially 4 attractors but becomes finally chaotic.

Vowel example

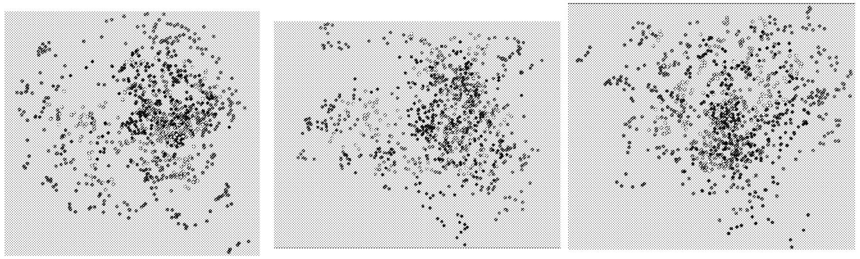
11 vowels were segmented from speech of a number of people, and 10 features were derived for each vowel.

Examples of 2D scatterograms for standardized data are below.



Vowel PCA

PCA components 1-2, 1-3 i 2-3



Vowel FDA

FDA components 1-2

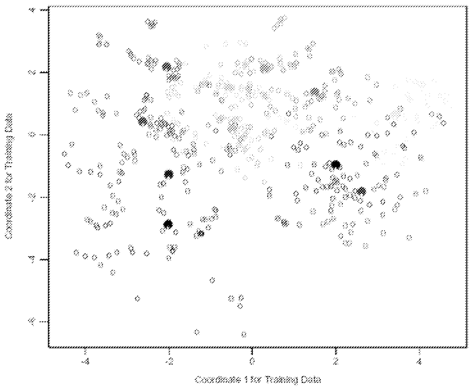


Figure from: Chapter 4, Elements of statistical learning.
By Hasti, Tibshirani and Friedman 2001

FDA higher components

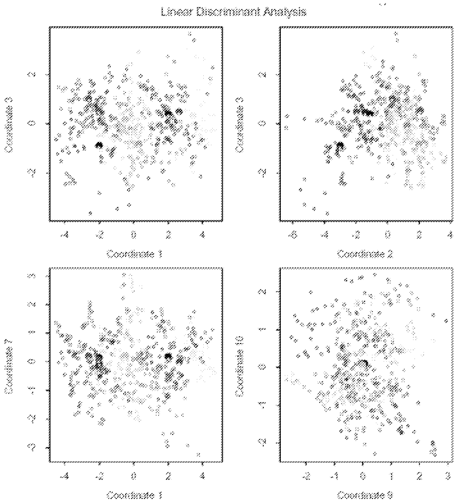


Figure from: Chapter 4, Elements of statistical learning.
By Hasti, Tibshirani and Friedman 2001

FDA prototypes

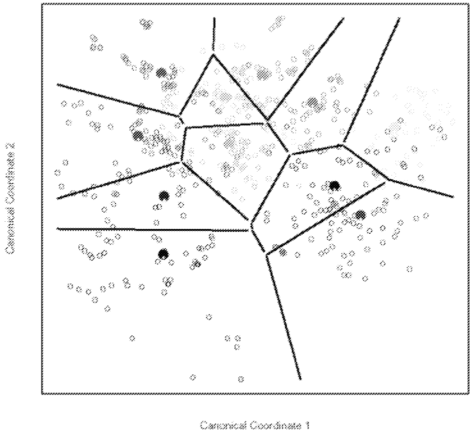


Figure from: Chapter 4, Elements of statistical learning.
By Hasti, Tibshirani and Friedman 2001