L10: Linear discriminants analysis

Linear discriminant analysis, two classes

Linear discriminant analysis, C classes

LDA vs. PCA

Limitations of LDA

Variants of LDA

Other dimensionality reduction methods

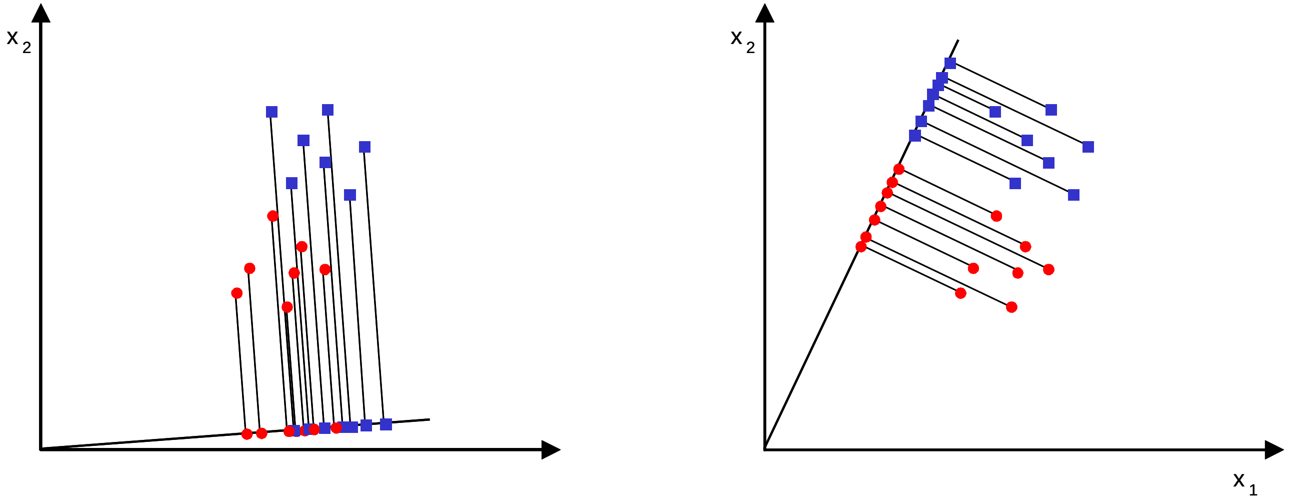
Linear discriminant analysis, two-classes

Objective

- LDA seeks to reduce dimensionality while preserving as much of the class discriminatory information as possible
- Assume we have a set of D-dimensional samples $\{x^{(1}, x^{(2}, \dots x^{(N)}\}$, N_1 of which belong to class ω_1, and N_2 to class ω_2
- We seek to obtain a scalar y by projecting the samples x onto a line

$$y = w^T x$$

- Of all the possible lines we would like to select the one that maximizes the separability of the scalars



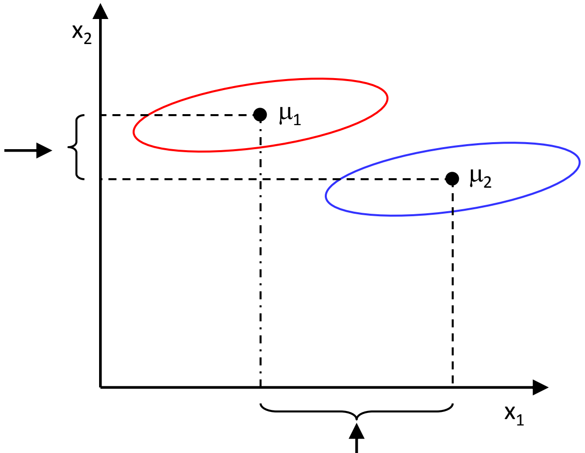
- In order to find a good projection vector, we need to define a measure of separation
- The mean vector of each class in x-space and y-space is

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \text{ and } \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T\mu_i$$

- We could then choose the distance between the projected means as our objective function

$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T(\mu_1 - \mu_2)|$$

  - However, the distance between projected means is not a good measure since it does not account for the standard deviation within classes



This axis yields better class separability

This axis has a larger distance between means

Fisher's solution

- Fisher suggested maximizing the difference between the means, normalized by a measure of the within-class scatter
- For each class we define the scatter, an equivalent of the variance, as

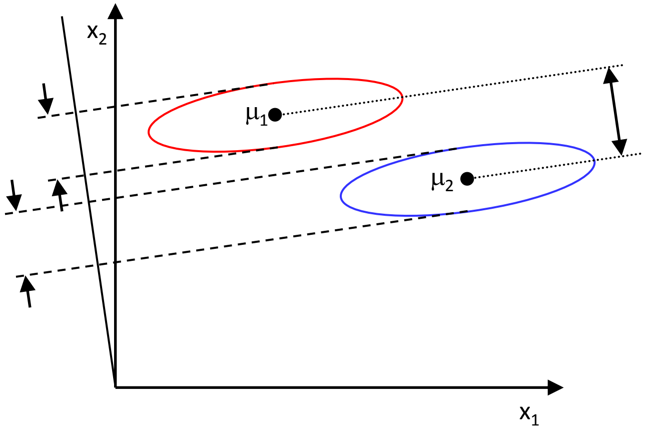
$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

  - where the quantity $(\tilde{s}_1^2 + \tilde{s}_2^2)$ is called the within-class scatter of the projected examples

- The Fisher linear discriminant is defined as the linear function $w^T x$ that maximizes the criterion function

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- Therefore, we are looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible



To find the optimum w^*, we must express $J(w)$ as a function of w

- First, we define a measure of the scatter in feature space x

$$S_i = \sum_{x \in \omega_i} (x-\mu_i)(x-\mu_i)^T$$
$$S_1 + S_2 = S_W$$

  - where S_W is called the within-class scatter matrix

- The scatter of the projection y can then be expressed as a function of the scatter matrix in feature space x

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y-\tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 =$$
$$= \sum_{x \in \omega_i} w^T (x-\mu_i)(x-\mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w$$

- Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \underbrace{(\mu_1-\mu_2)(\mu_1-\mu_2)^T}_{S_B} w = w^T S_B w$$

  - The matrix S_B is called the between-class scatter. Note that, since S_B is the outer product of two vectors, its rank is at most one

- We can finally express the Fisher criterion in terms of S_W and S_B as

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

- To find the maximum of $J(w)$ we derive and equate to zero

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0 \Rightarrow$$

$$[w^T S_W w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_W w]}{dw} = 0 \Rightarrow$$

$$[w^T S_W w] 2 S_B w - [w^T S_B w] 2 S_W w = 0$$

- Dividing by $w^T S_W w$

$$\left[\frac{w^T S_W w}{w^T S_W w}\right] S_B w - \left[\frac{w^T S_B w}{w^T S_W w}\right] S_W w = 0 \Rightarrow$$

$$S_B w - J S_W w = 0 \Rightarrow$$

$$S_W^{-1} S_B w - J w = 0$$

- Solving the generalized eigenvalue problem ($S_W^{-1} S_B w = J w$) yields

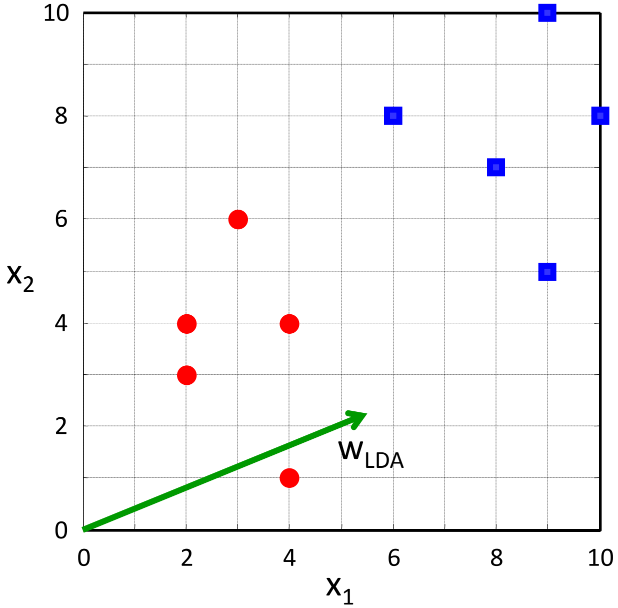
$$w^* = \arg\max\left[\frac{w^T S_B w}{w^T S_W w}\right] = S_W^{-1}(\mu_1 - \mu_2)$$

- This is know as Fisher's linear discriminant (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension

Example

Compute the LDA projection for the following 2D dataset

$$X1 = \{(4,1),(2,4),(2,3),(3,6),(4,4)\}$$
$$X2 = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$$



SOLUTION (by hand)

- The class statistics are

$$S_1 = \begin{bmatrix} .8 & -.4 \\ & 2.64 \end{bmatrix} \quad S_2 = \begin{bmatrix} 1.84 & -.04 \\ & 2.64 \end{bmatrix}$$

$$\mu_1 = [3.0\ 3.6]^T;\ \ \mu_2 = [8.4\ 7.6]^T$$

- The within- and between-class scatter are

$$S_B = \begin{bmatrix} 29.16 & 21.6 \\ & 16.0 \end{bmatrix} \quad S_W = \begin{bmatrix} 2.64 & -.44 \\ & 5.28 \end{bmatrix}$$

- The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1} S_B v = \lambda v \Rightarrow \left| S_W^{-1} S_B - \lambda I \right| = 0 \Rightarrow \begin{vmatrix} 11.89 - \lambda & 8.81 \\ 5.08 & 3.76 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda = 15.65$$

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \Rightarrow \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} .91 \\ .39 \end{bmatrix}$$

- Or directly by

$$w^* = S_W^{-1}(\mu_1 - \mu_2) = [-.91\ \ -.39]^T$$

LDA, C classes

Fisher’s LDA generalizes gracefully for C-class problems

- Instead of one projection y, we will now seek $(C-1)$ projections $[y_1, y_2, \dots y_{C-1}]$ by means of $(C-1)$ projection vectors $w_i$arranged by columns into a projection matrix $W = [w_1|w_2|\dots|w_{C-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

Derivation

- The within-class scatter generalizes as

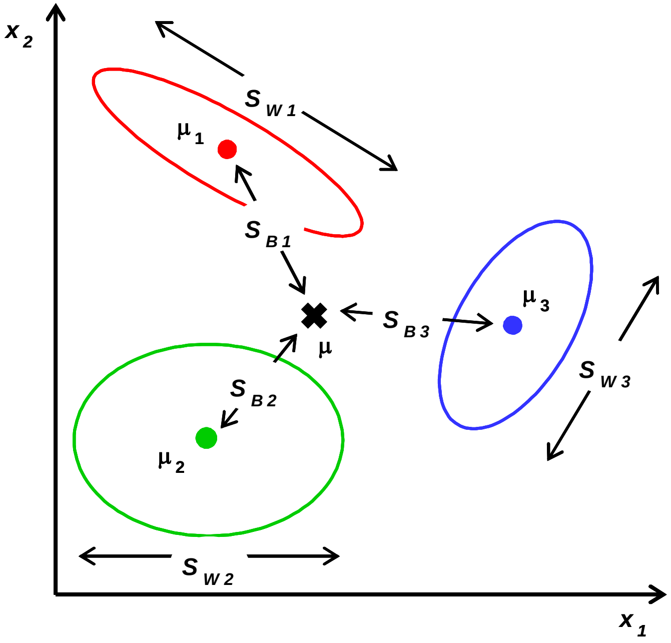
$$S_W = \quad {}_{i=1}^{C} S_i$$

  - where $S_i = \sum_{x\in\omega_i}(x-\mu_i)(x-\mu_i)^T$ and $\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x$

- And the between-class scatter becomes

$$S_B = \quad {}_{i=1}^{C} N_i(\mu_i-\mu)(\mu_i-\mu)^T$$

  - where $\mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N} \quad {}_{i=1}^{C} N_i\mu_i$



- Matrix $S_T = S_B + S_W$ is called the total scatter

- Similarly, we define the mean vector and scatter matrices for the projected samples as

$$\tilde{\mu}_i = \frac{1}{\mathrm{N}_i} \sum\nolimits_{y \in \omega_i} y \qquad \tilde{S}_W = \ {}_{i=1}^{C} \sum\nolimits_{y \in \omega_i} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T$$

$$\tilde{\mu} = \frac{1}{N} \sum\nolimits_{\forall y} y \qquad \tilde{S}_B = \ {}_{i=1}^{C} N_i (\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^T$$

- From our derivation for the two-class problem, we can write

$$\tilde{S}_W = W^T S_W W$$
$$\tilde{S}_B = W^T S_B W$$

- Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has $C - 1$ dimensions), we use the determinant of the scatter matrices to obtain a scalar objective function

$$J(W) = \frac{\left|\tilde{S}_B\right|}{\left|\tilde{S}_W\right|} = \frac{|W^T S_B W|}{|W^T S_W W|}$$

- And we will seek the projection matrix W^* that maximizes this ratio

- It can be shown that the optimal projection matrix W^* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem

$$W^* = [w_1^* | w_2^* | \dots w_{C-1}^*] = \arg\max \frac{|W^T S_B W|}{|W^T S_W W|} \Rightarrow (S_B - \lambda_i S_W) w_i^* = 0$$

NOTES

- S_B is the sum of C matrices of rank ≤ 1 and the mean vectors are constrained by $\frac{1}{C} \quad {}_{i=1}^{C} \mu_i = \mu$
  - Therefore, S_B will be of rank $(C-1)$ or less
  - This means that only $(C-1)$ of the eigenvalues λ_i will be non-zero
- The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1} S_B$
- LDA can be derived as the Maximum Likelihood method for the case of normal class-conditional densities with equal covariance matrices

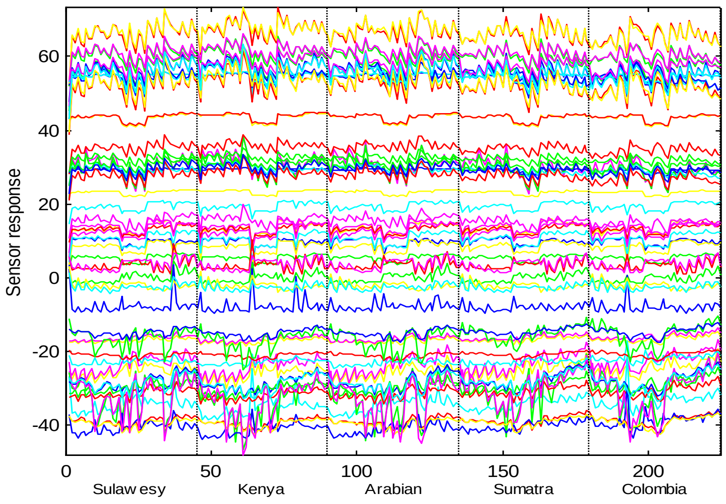
LDA vs. PCA

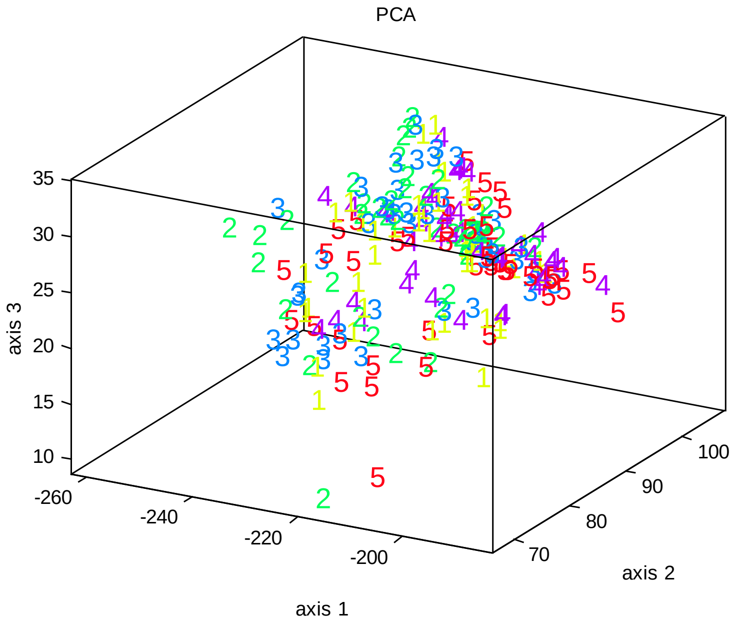
This example illustrates the performance of PCA and LDA on an odor recognition problem

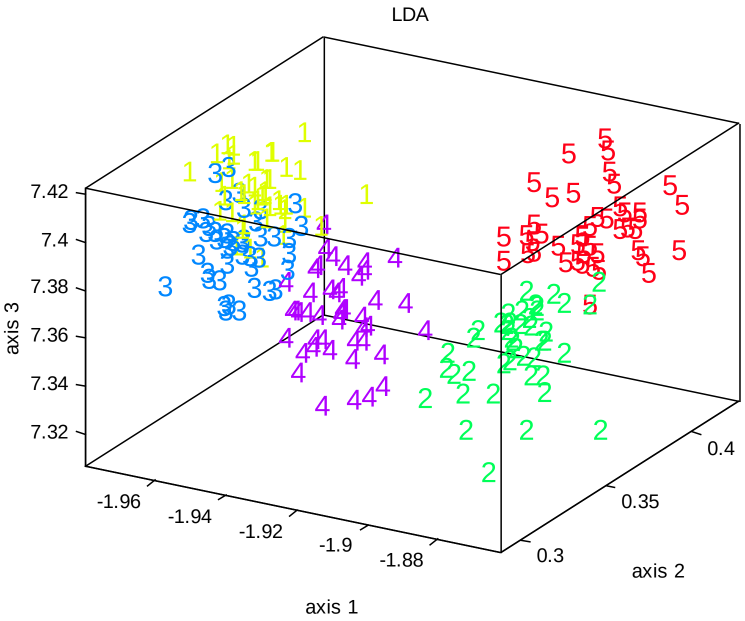
- Five types of coffee beans were presented to an array of gas sensors
- For each coffee type, 45 “sniffs” were performed and the response of the gas sensor array was processed ir order to obtain a 60-dimensional feature vector

Results

- From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
- This is one example where the discriminatory information is not aligned with the direction of maximum variance







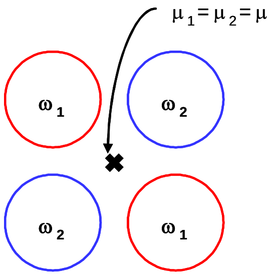
Limitations of LDA

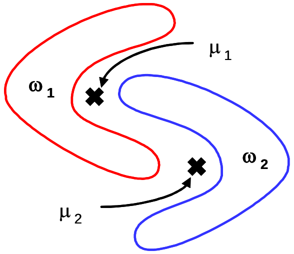
LDA produces at most $C-1$ feature projections

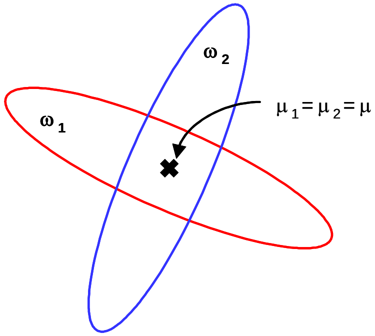
- If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features

LDA is a parametric method (it assumes unimodal Gaussian likelihoods)

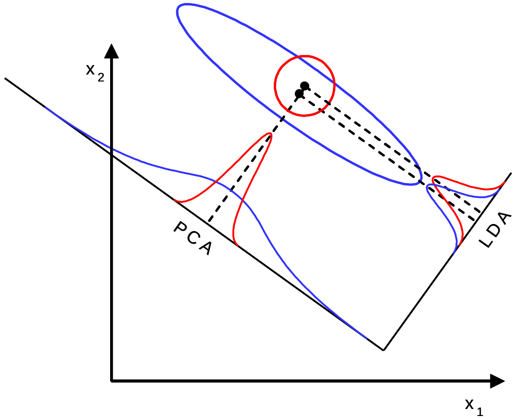
- If the distributions are significantly non-Gaussian, the LDA projections may not preserve complex structure in the data needed for classification







LDA will also fail if discriminatory information is not in the mean but in the variance of the data



Variants of LDA

Non-parametric LDA (Fukunaga)

- NPLDA relaxes the unimodal Gaussian assumption by computing S_B using local information and the kNN rule. As a result of this
  - The matrix S_B is full-rank, allowing us to extract more than $(C-1)$ features
  - The projections are able to preserve the structure of the data more closely

Orthonormal LDA (Okada and Tomita)

- OLDA computes projections that maximize the Fisher criterion and, at the same time, are pair-wise orthonormal
  - The method used in OLDA combines the eigenvalue solution of $S_W^{-1}S_B$ and the Gram-Schmidt orthonormalization procedure
  - OLDA sequentially finds axes that maximize the Fisher criterion in the subspace orthogonal to all features already extracted
  - OLDA is also capable of finding more than $(C-1)$ features

Generalized LDA (Lowe)

- GLDA generalizes the Fisher criterion by incorporating a cost function similar to the one we used to compute the Bayes Risk
  - As a result, LDA can produce projects that are biased by the cost function, i.e., classes with a higher cost C_{ij} will be placed further apart in the low-dimensional projection

Multilayer perceptrons (Webb and Lowe)

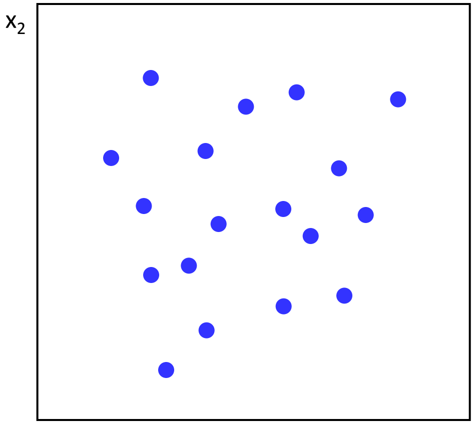
- It has been shown that the hidden layers of multi-layer perceptrons perform non-linear discriminant analysis by maximizing $Tr[S_B S_T^{\dagger}]$, where the scatter matrices are measured at the output of the last hidden layer

Other dimensionality reduction methods

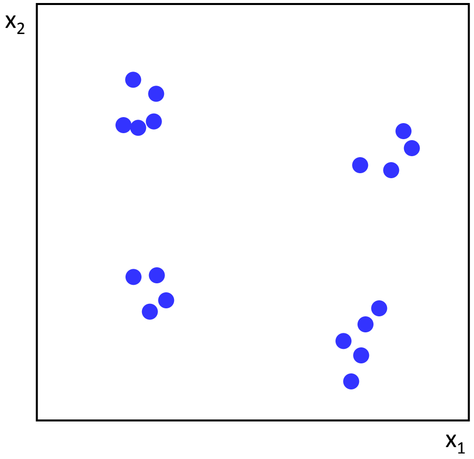
Exploratory Projection Pursuit (Friedman and Tukey)

- EPP seeks an M-dimensional (M=2,3 typically) linear projection of the data that maximizes a measure of “interestingness”
- Interestingness is measured as departure from multivariate normality
  - This measure is not the variance and is commonly scale-free. In most implementations it is also affine invariant, so it does not depend on correlations between features. [Ripley, 1996]
- In other words, EPP seeks projections that separate clusters as much as possible and keeps these clusters compact, a similar criterion as Fisher’s, but EPP does NOT use class labels
- Once an interesting projection is found, it is important to remove the structure it reveals to allow other interesting views to be found more easily

Uninteresting



Interesting

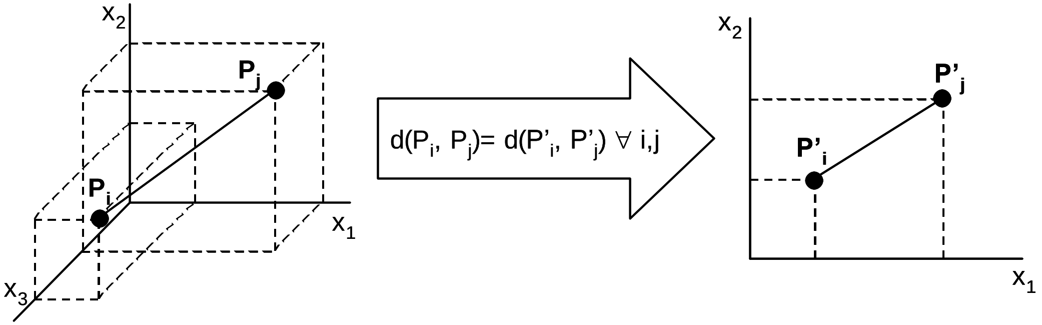


Sammon's non-linear mapping (Sammon)

- This method seeks a mapping onto an M-dimensional space that preserves the inter-point distances in the original N-dimensional space
- This is accomplished by minimizing the following objective function

$$E(d,d') = \quad {}_{i\neq j} \frac{\left[d\left(P_i,P_j\right)-d\left(P_i',P_j'\right)\right]^2}{d\left(P_i,P_j\right)}$$

  - The original method did not obtain an explicit mapping but only a lookup table for the elements in the training set
    - Newer implementations based on neural networks do provide an explicit mapping for test data and also consider cost functions (e.g., Neuroscale)
  - Sammon's mapping is closely related to Multi Dimensional Scaling (MDS), a family of multivariate statistical methods commonly used in the social sciences
    - We will review MDS techniques when we cover manifold learning


$d(P_i, P_j) = d(P'_i, P'_j)\ \forall\ i,j$