PCA and LDA

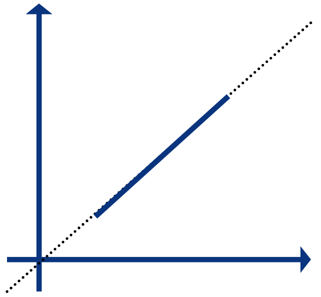
Nuno Vasconcelos

ECE Department, UCSD

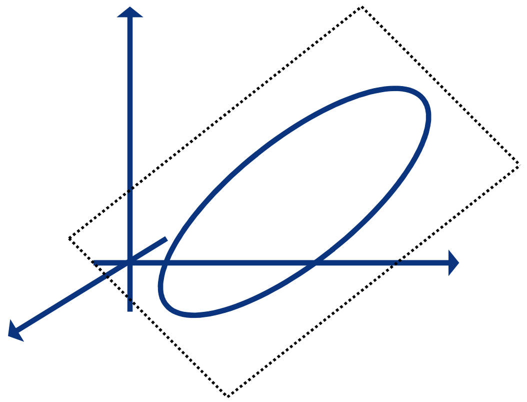
Principal component analysis

- basic idea:
  - if the data lives in a subspace, it is going to look very flat when viewed from the full space, e.g.

1D subspace in 2D



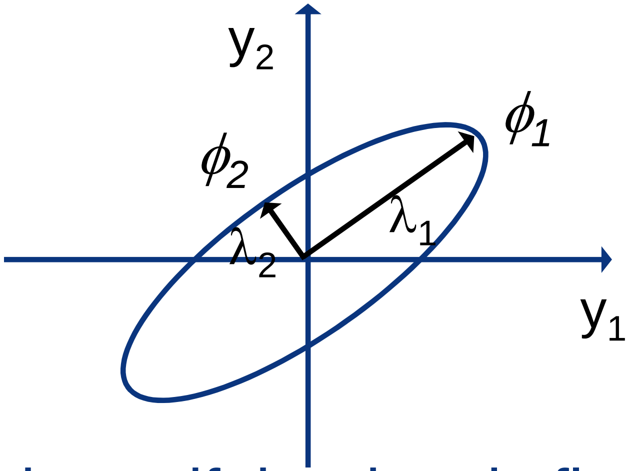
2D subspace in 3D



  - this means that if we fit a Gaussian to the data the equiprobability contours are going to be highly skewed ellipsoids

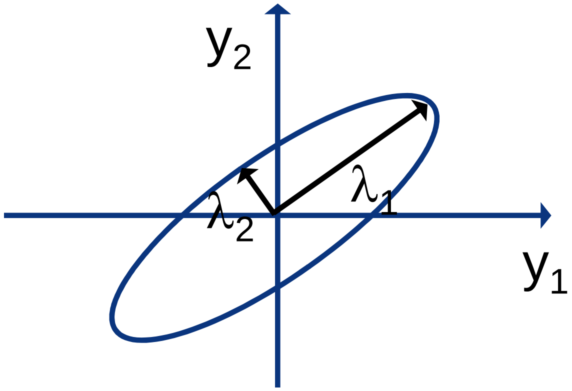
Principal component analysis

- If y is Gaussian with covariance Σ, the equiprobability contours are the ellipses whose
  - principal components ϕ_i are the eigenvectors of Σ
  - principal lengths λ_i are the eigenvalues of Σ

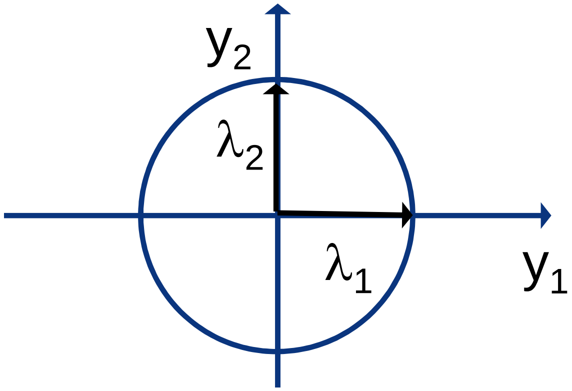


- by computing the eigenvalues we know if the data is flat

$\lambda_1 >> \lambda_2$: flat



$\lambda_1=\lambda_2$: not flat



Principal component analysis (learning)

- Given sample $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}, \ x_i \in \mathcal{R}^d$

  - compute sample mean: $\hat{\mu} = \frac{1}{n}\sum_i(\mathbf{x}_i)$

  - compute sample covariance: $\hat{\Sigma} = \frac{1}{n}\sum_i(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T$

  - compute eigenvalues and eigenvectors of $\hat{\Sigma}$

$$\hat{\Sigma} = \Phi\Lambda\Phi^T, \ \Lambda = diag(\sigma_1^2, \dots, \sigma_n^2) \ \Phi^T\Phi = I$$

  - order eigenvalues $\sigma_1^2 > ... > \sigma_n^2$

  - if, for a certain k, $\sigma_k << \sigma_1$ eliminate the eigenvalues and eigenvectors above k.

Principal component analysis

- Given principal compoenents $\phi_i, i \in 1, \ldots, k$ and a test sample $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\},\ \ t_i \in \mathcal{R}^d$

  - subtract mean to each point $\mathbf{t}'_i = \mathbf{t}_i - \widehat{\mu}$

  - project onto eigenvector space $\mathbf{y}_i = \mathbf{A}\mathbf{t}'_i$ where

$$\mathbf{A} = \begin{bmatrix} \phi_1^T \\ \vdots \\ \phi_k^T \end{bmatrix}$$

  - use $\mathcal{T}' = \{\mathbf{y}_1, \ldots \mathbf{y}_n\}$ to estimate class conditional densities and do all further processing on **y**.

Principal component analysis

- there is an alternative manner to compute the principal components, based on singular value decomposition
- SVD:
  - any real n x m matrix (n>m) can be decomposed as

$$\boxed{A = \mathrm{M}\Pi\mathrm{N}^T}$$

  - where M is a n x m column orthonormal matrix of left singular vectors (columns of M)
  - Π a m x m diagonal matrix of singular values
  - N^T a m x m row orthonormal matrix of right singular vectors (columns of N)

$$\boxed{\mathrm{M}^T\mathrm{M} = I \qquad \mathrm{N}^T\mathrm{N} = I}$$

PCA by SVD

- to relate this to PCA, we consider the data matrix

$$\boxed{X = \begin{bmatrix} | & & | \\ x_1 & \ldots & x_n \\ | & & | \end{bmatrix}}$$

- the sample mean is

$$\boxed{\mu = \frac{1}{n}\sum_i x_i = \frac{1}{n}\begin{bmatrix} | & & | \\ x_1 & \ldots & x_n \\ | & & | \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n}X1}$$

PCA by SVD

- and we can center the data by subtracting the mean to each column of X
- this is the centered data matrix

$$\boxed{\begin{aligned} X_c &= \begin{bmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{bmatrix} - \begin{bmatrix} | & & | \\ \mu & \dots & \mu \\ | & & | \end{bmatrix} \\ &= X - \mu 1^T = X - \frac{1}{n} X 1 1^T = X\left(I - \frac{1}{n} 1 1^T \right) \end{aligned}}$$

PCA by SVD

- the sample covariance is

$$\boxed{\Sigma = \frac{1}{n}\sum_i (x_i - \mu)(x_i - \mu)^T = \frac{1}{n}\sum_i x_i^c \left(x_i^c\right)^T}$$

where x_i^c is the i^{th} column of X_c

- this can be written as

$$\boxed{\Sigma = \frac{1}{n}\begin{bmatrix} | & & | \\ x_1^c & \dots & x_n^c \\ | & & | \end{bmatrix}\begin{bmatrix} - & x_1^c & - \\ & \vdots & \\ - & x_n^c & - \end{bmatrix} = \frac{1}{n} X_c X_c^T}$$

PCA by SVD

- the matrix

$$\boxed{X_c^T = \begin{bmatrix} - & x_1^c & - \\ & \vdots & \\ - & x_n^c & - \end{bmatrix}}$$

is real n x d. Assuming n > d it has SVD decomposition

$$\boxed{X_c^T = \mathrm{M}\Pi\mathrm{N}^T} \qquad \boxed{\mathrm{M}^T\mathrm{M} = I \qquad \mathrm{N}^T\mathrm{N} = I}$$

and

$$\boxed{\Sigma = \frac{1}{n} X_c X_c^T = \frac{1}{n}\mathrm{N}\Pi\mathrm{M}^T\mathrm{M}\Pi\mathrm{N}^T = \frac{1}{n}\mathrm{N}\Pi^2\mathrm{N}^T}$$

PCA by SVD

$$\boxed{\Sigma = \mathrm{N}\left(\frac{1}{n}\Pi^2\right)\mathrm{N}^T}$$

- noting that N is d x d and orthonormal, and Π^2 diagonal, shows that this is just the eigenvalue decomposition of Σ
- it follows that
  - the eigenvectors of Σ are the columns of N
  - the eigenvalues of Σ are

$$\boxed{\lambda_i = \frac{1}{n}\pi_i^2}$$

- this gives an alternative algorithm for PCA

PCA by SVD

- computation of PCA by SVD
- given X with one example per column
  - 1) create the centered data-matrix

$$\boxed{X_c^T = \left(I - \frac{1}{n} 11^T \right) X^T}$$

  - 2) compute its SVD

$$\boxed{X_c^T = \mathrm{M\Pi N}^T}$$

  - 3) principal components are columns of N, eigenvalues are

$$\boxed{\lambda_i = \frac{1}{n} \pi_i^2}$$

Limitations of PCA

- PCA is not optimal for classification
  - note that there is no mention of the class label in the definition of PCA
  - keeping the dimensions of largest energy (variance) is a good idea, but not always enough
  - certainly improves the density estimation, since space has smaller dimension
  - but could be unwise from a classification point of view
  - the discriminant dimensions could be thrown out
- it is not hard to construct examples where PCA is the worst possible thing we could do

Example

- consider a problem with
  - two n-D Gaussian classes with covariance $\Sigma=\sigma^2 I$, $\sigma^2 = 10$

$$X \sim N(\mu_i, 10I)$$

  - we add an extra variable which is the class label itself

$$X^{'} = [X,\ i]$$

  - assuming that $P_Y(0)=P_Y(1)=0.5$

$$E[Y] = 0.5\times 0 + 0.5\times 1 = 0.5$$

$$\mathrm{var}[Y] = 0.5\times(0-0.5)^2 + 0.5\times(1-0.5)^2$$

$$= 0.125 < 10$$

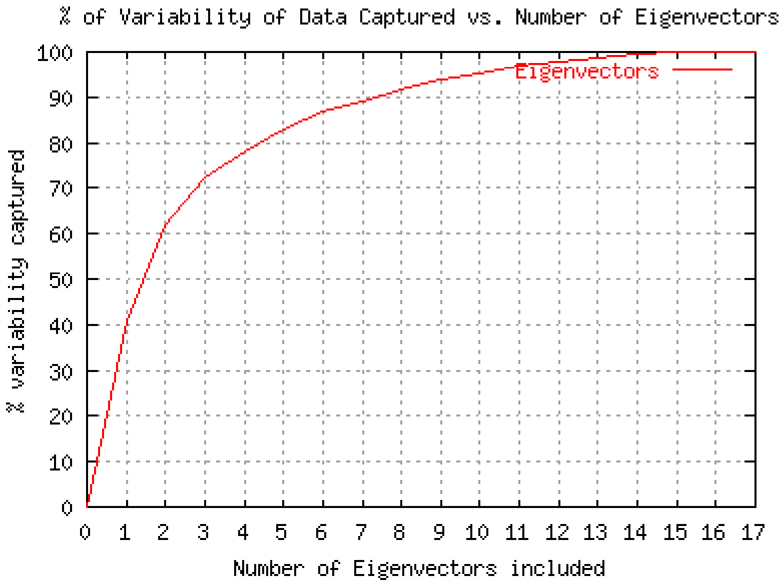
  - dimension *n+1* has the smallest variance and is the first to be discarded!

Example

- this is
  - a very contrived example
  - but shows that PCA can throw away all the discriminant info
- does this mean you should never use PCA?
  - no, typically it is a good method to find a suitable subset of variables, as long as you are not too greedy
  - e.g. if you start with n = 100, and know that there are only 5 variables of interest
  - picking the top 20 PCA components is likely to keep the desired 5
  - your classifier will be much better than for n = 100, probably not much worse than the one with the best 5 features
- is there a rule of thumb for finding the number of PCA components?

Principal component analysis

- a natural measure is to pick the eigenvectors that explain p % of the data variability
  - can be done by plotting the ratio r_k as a function of k

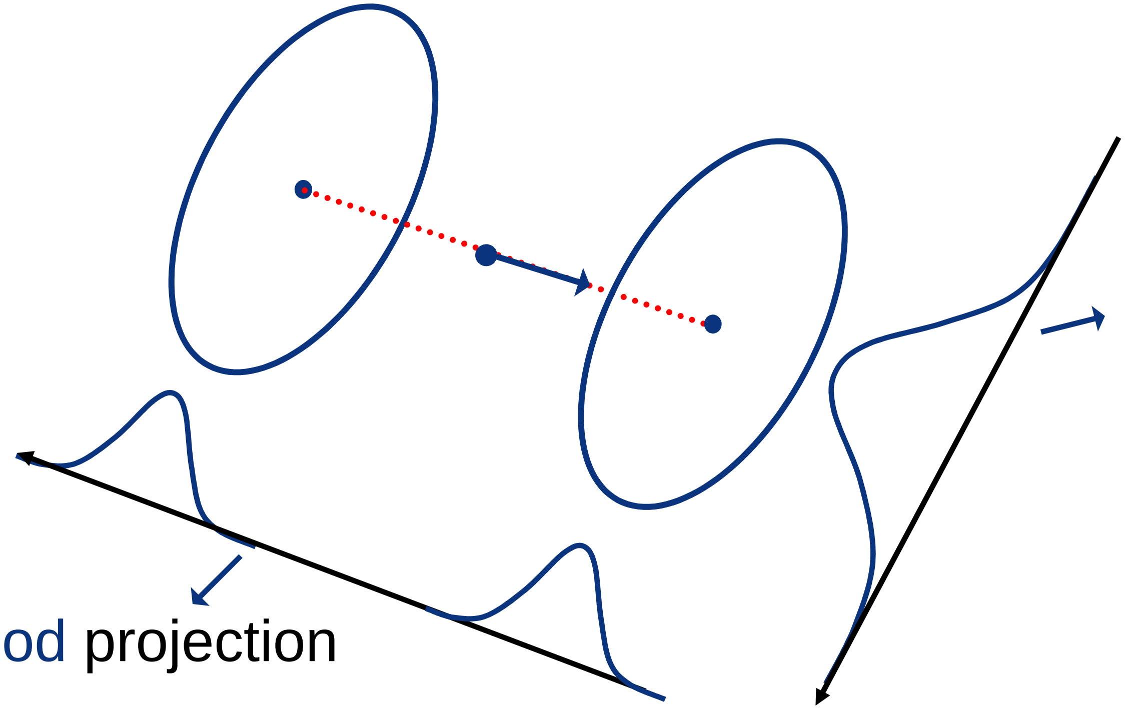


$$r_k = \sum_{i=1}^{k} \lambda_i^2 \Big/ \sum_{i=1}^{n} \lambda_i^2$$

  - e.g. we need 3 eigenvectors to cover 70% of the variability of this dataset

Fischer’s linear discriminant

- what if we really need to find the best features?
  - harder question, usually impossible with simple methods
  - there are better methods at finding discriminant directions
- one good example is linear discriminant analysis (LDA)
  - the idea is to find the line that best separates the two classes



bad projection

good projection

Linear discriminant analysis

- we have two classes such that

$$E_{X|Y}\left[X \mid Y = i\right] = \mu_i$$

$$E_{X|Y}\left[(X - \mu_i)(X - \mu_i)^T \mid Y = i\right] = \Sigma_i$$

- and want to find the line

$$z = w^T x$$

that best separates them

- one possibility would be to maximize

$$\left(E_{Z|Y}\left[Z \mid Y = 1\right] - E_{Z|Y}\left[Z \mid Y = 0\right]\right)^2 =$$

$$\left(E_{X|Y}\left[w^T x \mid Y = 1\right] - E_{X|Y}\left[w^T x \mid Y = 0\right]\right)^2 = \left(w^T\left[\mu_1 - \mu_0\right]\right)^2$$

Linear discriminant analysis

- however, this $\left(w^T[\mu_1 - \mu_0]\right)^2$

  can be made arbitrarily large by simply scaling w
- we are only interested in the direction, not the magnitude
- need some type of normalization
- Fischer suggested

$$\boxed{\begin{gathered}\max \frac{between\ \ class\ \ scatter}{within\ \ class\ \ scatter} = \\ \max_w \frac{\left(E_{Z|Y}[Z \mid Y=1] - E_{Z|Y}[Z \mid Y=0]\right)^2}{\operatorname{var}[Z \mid Y=1] + \operatorname{var}[Z \mid Y=0]}\end{gathered}}$$

Linear discriminant analysis

- we have already seen that

$$\left(E_{Z|Y}[Z \mid Y = 1] - E_{Z|Y}[Z \mid Y = 0]\right)^2 = \left(w^T[\mu_1 - \mu_0]\right)^2$$

$$= w^T[\mu_1 - \mu_0][\mu_1 - \mu_0]^T w$$

- also

$$\mathrm{var}[Z \mid Y = i] = E_{Z|Y}\left\{\left(z - E_{Z|Y}[Z \mid Y = i]\right)^2 \mid Y = i\right\}$$

$$= E_{Z|Y}\left\{\left(w^T[x - \mu_i]\right)^2 \mid Y = i\right\}$$

$$= E_{Z|Y}\left\{w^T[x - \mu_i][x - \mu_i]^T w \mid Y = i\right\}$$

$$= w^T \Sigma_i w$$

Linear discriminant analysis

- and

$$J(w) = \frac{\left(E_{Z|Y}\left[Z \mid Y = 1\right] - E_{Z|Y}\left[Z \mid Y = 0\right]\right)^2}{\operatorname{var}\left[Z \mid Y = 1\right] + \operatorname{var}\left[Z \mid Y = 0\right]}$$

$$= \frac{w^T(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w}{w^T(\Sigma_1 + \Sigma_0)w}$$

- which can be written as

between class scatter

$$\boxed{J(w) = \frac{w^T S_B w}{w^T S_W w}} \qquad \boxed{\begin{aligned} S_B &= (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \\ S_W &= (\Sigma_1 + \Sigma_0) \end{aligned}}$$

within class scatter

Linear discriminant analysis

- maximizing the ratio

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

  - is equivalent to maximizing the numerator while keeping the denominator constant, i.e.

$$\boxed{\max_w w^T S_B w \quad \text{subject to} \quad w^T S_W w = K}$$

  - and can be accomplished using Lagrange multipliers
  - define the Lagrangian

$$\boxed{L = w^T S_B w - \lambda\left(w^T S_W w - K\right)}$$

  - and maximize with respect to both *w* and λ

Linear discriminant analysis

- setting the gradient of

$$L = w^T\left(S_B - \lambda S_W\right)w \; + \lambda K$$

  with respect to *w* to zero we get

$$\nabla_w L = 2\left(S_B - \lambda S_W\right)w \; = 0$$

  or

$$\boxed{S_B w = \lambda S_W w}$$

- this is a generalized eigenvalue problem
- the solution is easy when $\boxed{S_w^{-1} = \left(\Sigma_1 + \Sigma_0\right)^{-1}}$ exists

Linear discriminant analysis

- in this case

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

- and, using the definition of S_B

$$S_W^{-1}(\mu_1 - \mu_0)(\mu_1 - \mu_0)^T w = \lambda w$$

- noting that $(\mu_1 - \mu_0)^T w = \alpha$ is a scalar this can be written as

$$S_W^{-1}(\mu_1 - \mu_0) = \frac{\lambda}{\alpha} w$$

- and since we don't care about the magnitude of w

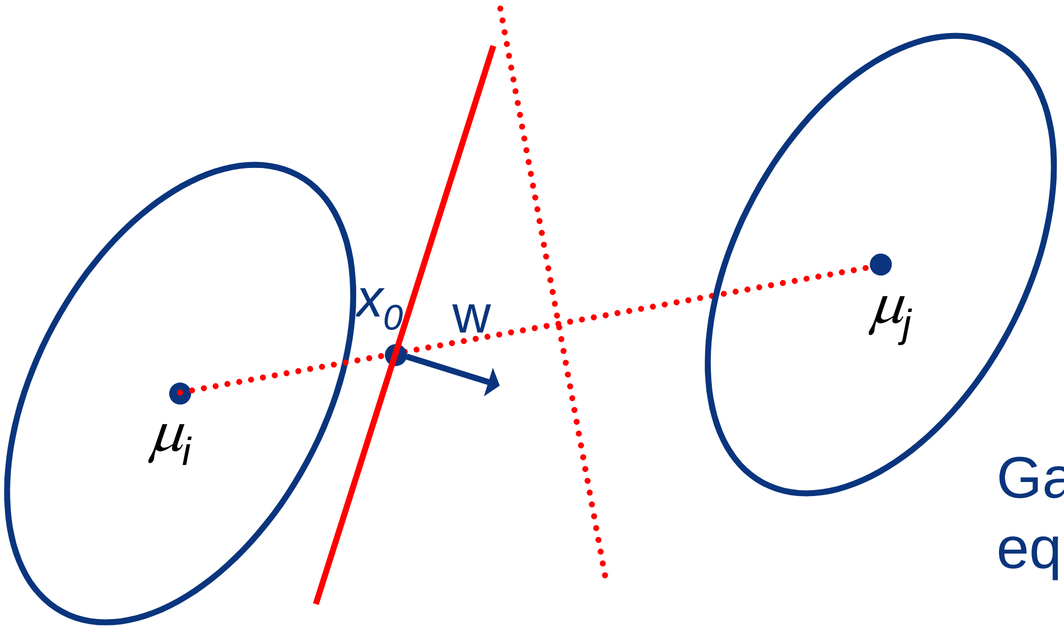
$$\boxed{w* = S_W^{-1}(\mu_1 - \mu_0) = (\Sigma_1 + \Sigma_0)^{-1}(\mu_1 - \mu_0)}$$

Linear discriminant analysis

- note that we have seen this before
  - for a classification problem with Gaussian classes of equal covariance $\Sigma_i = \Sigma$, the BDR boundary is the plane of normal

$$\boxed{w = \Sigma^{-1}\left(\mu_i - \mu_j\right)}$$

  - if $\Sigma_1 = \Sigma_0$, this is also the LDA solution



Gaussian classes,
equal covariance Σ

Linear discriminant analysis

- this gives two different interpretations of LDA
  - it is optimal if and only if the classes are Gaussian and have equal covariance
  - better than PCA, but not necessarily good enough
  - a classifier on the LDA feature, is equivalent to
    - the BDR after the approximation of the data by two Gaussians with equal covariance

Any questions?