The Rayleigh Quotient

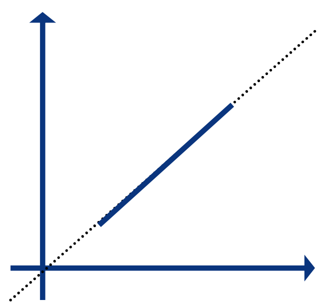
Nuno Vasconcelos

ECE Department, UCSD

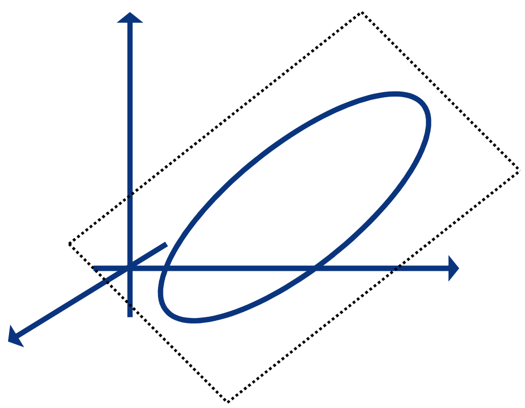
Principal component analysis

- 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

The role of the mean

- note that the mean of the entire data is a function of the coordinate system
  - if X has mean μ then X – μ has mean 0
- we can always make the data have zero mean by centering
  - if and

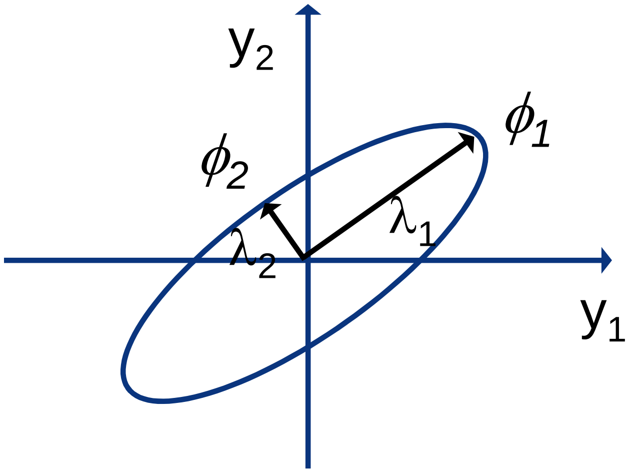
$$\boxed{X = \begin{bmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{bmatrix}} \qquad \boxed{X_c^T = \left(I - \frac{1}{n} 1 1^T \right) X^T}$$

  - then X_c has zero mean
- can assume that X is zero mean without loss of generality

Principal component analysis

- If y is Gaussian with covariance Σ
- the equiprobability contours

$$\boxed{y^T \Sigma^{-1} y = K}$$



- are the ellipses whose
  - principal components ϕ_i are the eigenvectors of Σ
  - principal lengths λ_i are the eigenvalues of Σ
- by detecting small eigenvalues we can eliminate dimensions that have little variance
- this is PCA

PCA by SVD

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

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

  - 2) compute its SVD

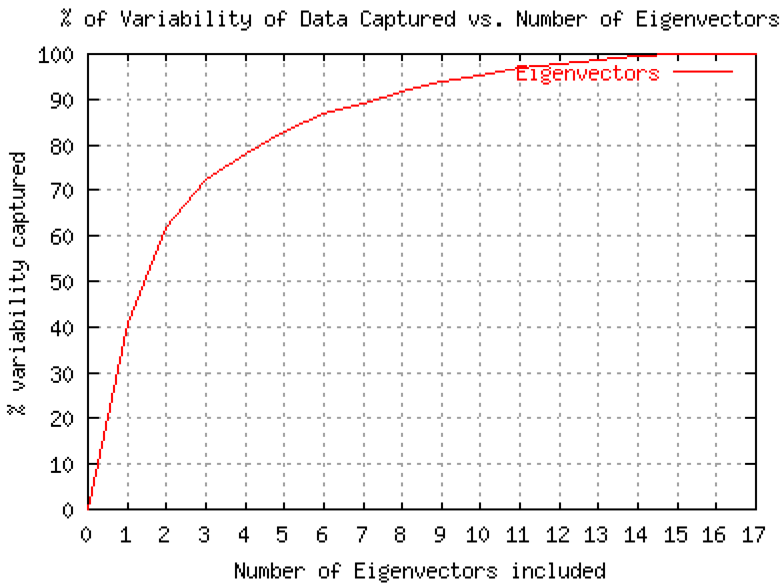
$$\boxed{X_c^T = \mathrm{M}\Pi\mathrm{N}^T}$$

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

$$\boxed{\lambda_i = \pi_i^2 / n}$$

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



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

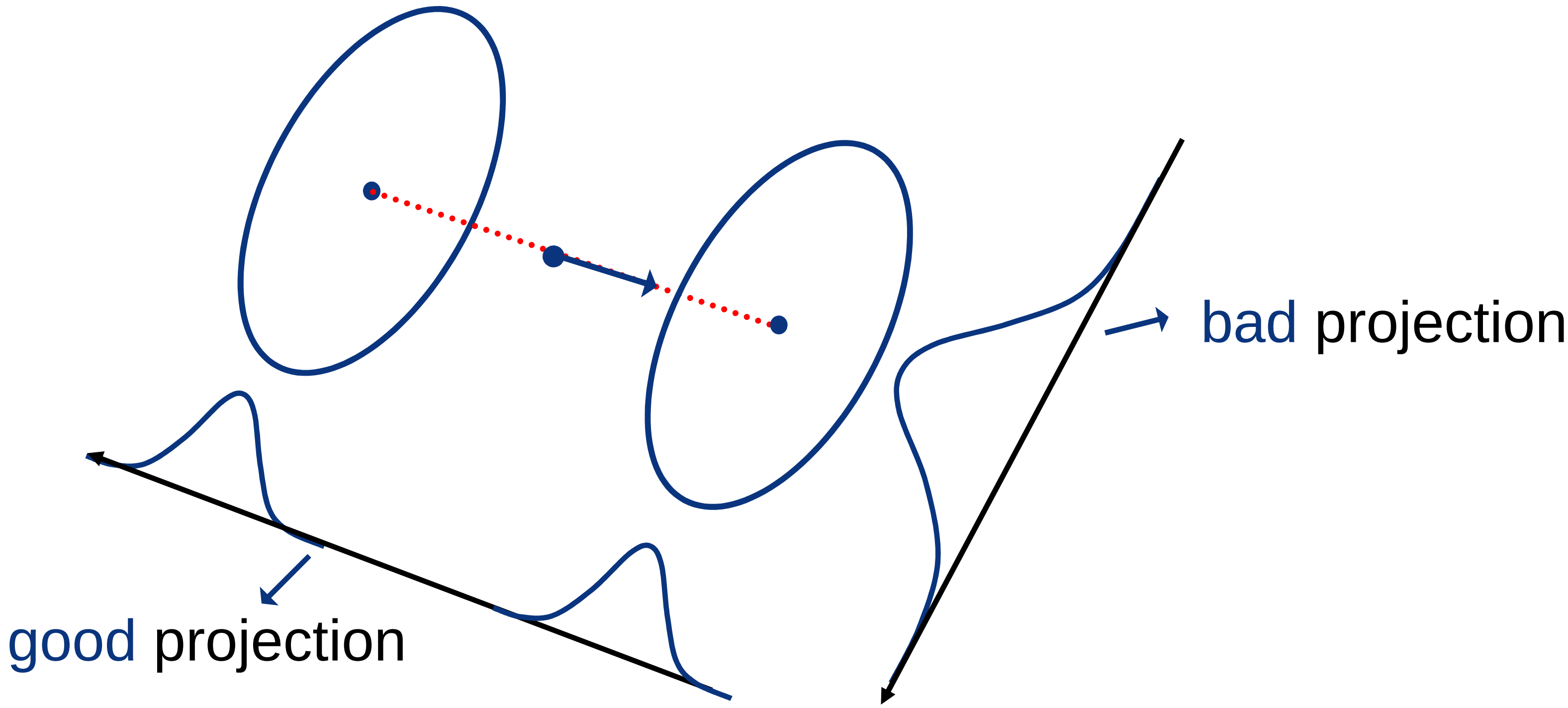
  - e.g. we need 3 eigenvectors to cover 70% of the variability of this dataset

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

Fischer's linear discriminant

- find the line $z = w^T x$ that best separates the two classes



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

Linear discriminant analysis

- this 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

- optimal solution is

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

- BDR after projection on z is equivalent to BDR on x if
  - the two classes are Gaussian and have equal covariance
- otherwise, LDA leads to a sub-optimal classifier

The Rayleigh quotient

- it turns out that the maximization of the Rayleigh quotient

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

$$\boxed{\begin{array}{l} S_B, S_W, \textit{symmetric} \\ \qquad\qquad \textit{positive semidefinite} \end{array}}$$

  appears in many problems in engineering and pattern recognition

- we have already seen that this is equivalent to

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

  and can be solved using Lagrange multipliers

The Rayleigh quotient

- define the Lagrangian

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

- maximize with respect to *w*

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

- to obtain the solution

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

- this is a generalized eigenvalue problem that you can solve using any eigenvalue routine
- which eigenvalue?

The Rayleigh quotient

- recall that we want

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

  and the optimal w satisfies

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

- hence

$$\boxed{(w^*)^T S_B w^* = \lambda (w^*)^T S_W w^* = \lambda K}$$

  which is maximum for the largest eigenvalue

- in summary, we need the generalized eigenvector $S_B w = \lambda S_W w$ of largest eigenvalue

The Rayleigh quotient

- case 1: S_w invertible
  - simplifies to a standard eigenvalue problem

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

  - w is the largest eigenvalue of $S_w^{-1} S_B$
- case 2: S_w not invertible
  - this is case is more problematic
  - in fact the cost can be unbounded
  - consider $w = w_r + w_n$, w_r in the row space of S_w and w_n in the null space

$$\boxed{\begin{aligned} w^T S_W w &= (w_r + w_n)^T S_W (w_r + w_n) = (w_r + w_n)^T S_W w_r \\ &= w_r^T S_W (w_r + w_n) = w_r^T S_W w_r \end{aligned}}$$

The Rayleigh quotient

- and

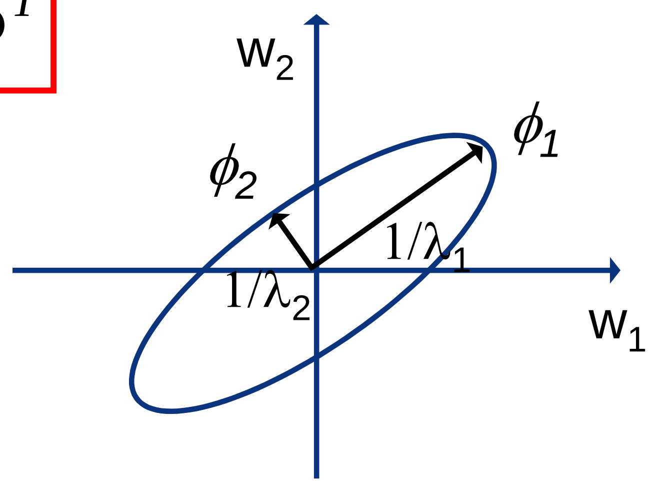
$$\boxed{\begin{aligned} w^T S_B w &= (w_r + w_n)^T S_B (w_r + w_n) \\ &= w_r^T S_B w_r + 2 w_r^T S_B w_n + \underbrace{w_n^T S_B w_n}_{\geq 0} \end{aligned}}$$

- hence, if there is a (w_r, w_n) such that $w_r^T S_B w_n \geq 0$,
  - we can make the cost arbitrarily large
  - by simply scaling up the null space component w_n
- this can also be seen geometrically

The Rayleigh quotient

- recall that

$$\boxed{w^T S_W w = K \quad \text{with} \quad S_W = \Phi \Lambda \Phi^T}$$

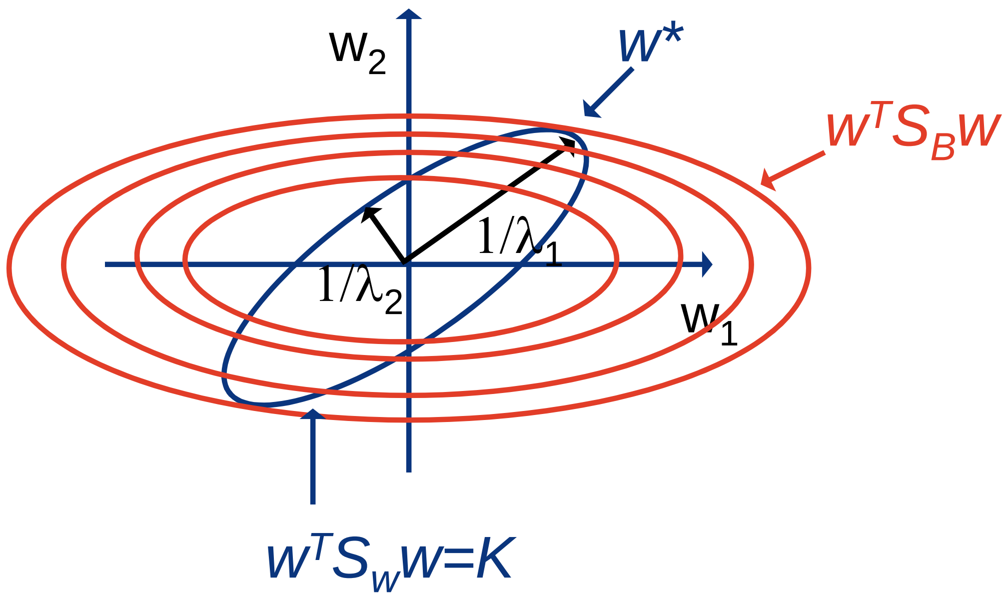


- are the ellipses whose
  - principal components ϕ_i are the eigenvectors of S_w
  - principal lengths are $1/\lambda_i$
- when the eigenvalues go to zero, the ellipses blow up
- consider the picture of the optimization problem

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

The Rayleigh quotient

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

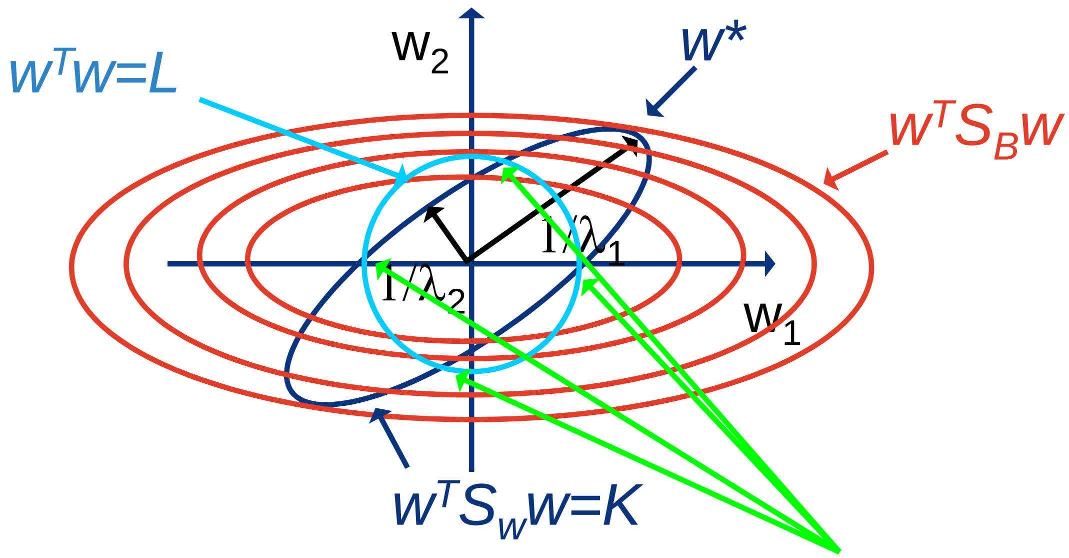


- the optimal solution is where the outer red ellipse (cost) touches the blue ellipse (constraint)
  - in this example, as λ_1 goes to *0*, $||w^*||$ and the cost go to infinity

The Rayleigh quotient

- how do we avoid this problem?
  - we introduce another constraint

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



  - restricts the set of possible solutions to these points (surfaces in high dimensional case)

The Rayleigh quotient

- the Lagrangian is now

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

- and the solution satisfies

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

or

$$\left(S_B - \lambda\left[S_W + \gamma I\right]\right)w = 0, \quad \gamma = \beta / \lambda$$

- but this is exactly the solution of the original problem with $S_w + \gamma I$ instead of S_w

$$\boxed{\max_w w^T S_B w \quad \text{subject to} \quad w^T\left[S_W + \gamma I\right]w = K}$$

The Rayleigh quotient

- adding the constraint is equivalent to maximizing the regularized Rayleigh quotient

$$J(w) = \frac{w^T S_B w}{w^T \left[S_W + \gamma I\right] w}$$

$$S_B, S_W, \text{symmetric positive semidefinite}$$

- what does this accomplish?
  - note that

$$\begin{aligned} S_W = \Phi\Lambda\Phi^T \Rightarrow S_W + \gamma I &= \Phi\Lambda\Phi^T + \gamma\Phi I \Phi^T \\ &= \Phi\left[\Lambda + \gamma I\right]\Phi^T \end{aligned}$$

  - this makes all eigenvalues positive
  - the matrix is no longer non-invertible

The Rayleigh quotient

- in summary

$$\boxed{\max_w \frac{w^T S_B w}{w^T S_W w}}$$

$$\boxed{\begin{array}{r} S_B, S_W, symmetric \\ positive\ semidefinite \end{array}}$$

- 1) S_w invertible
  - w* is the eigenvector of largest eigenvalue of $S_w^{-1}S_B$
  - the max value is λK, where λ is the largest eigenvalue
- 2) S_w not invertible
  - regularize: $S_w \rightarrow S_w + \gamma I$
  - w* is the eigenvector of largest eigenvalue of $[S_w + \gamma I]^{-1}S_B$
  - the max value is λK, where λ is the largest eigenvalue

Regularized discriminant analysis

- back to LDA:
  - when the within scatter matrix is non-invertible, instead of

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

  - we use

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

regularized within class scatter

Regularized discriminant analysis

- this is called regularized discriminant analysis (RDA)
- noting that

$$\boxed{\begin{aligned} S_W &= \Sigma_1 + \Sigma_0 + \gamma I \\ &= \Sigma_1 + \gamma_1 I + \Sigma_0 + \gamma_0 I \end{aligned}} \qquad \boxed{\gamma_1 + \gamma_0 = \gamma}$$

- this can also be seen as regularizing each covariance matrix individually
- the regularization parameters γ_i are determined by cross-validation
  - more on this later
  - basically means that we try several possibilities and keep the best

Principal component analysis

- back to PCA: given X with one example per column
  - 1) create the centered data-matrix

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

this has one point per row

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

  - note that the projection of all points on principal component ϕ is

$$\boxed{z = X_c^T \phi}$$

$$\boxed{\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} (x_1 - \mu)^T \phi \\ \vdots \\ (x_n - \mu)^T \phi \end{bmatrix}}$$

Principal component analysis

- and, since

$$\frac{1}{n}\sum_i z_i = \frac{1}{n}\sum_i (x_i - \mu)^T \phi = \left(\frac{1}{n}\sum_i x_i - \mu\right)^T \phi = 0$$

- the sample variance of Z is given by its norm

$$\boxed{\mathrm{var}(z) = \frac{1}{n}\sum_i z_i^2 = \|z\|^2}$$

- recall that PCA looks for the largest variance component

$$\boxed{\max_\phi \|z\|^2 = \max_\phi \left\|X_c^T \phi\right\|^2 = \max_\phi \left(X_c^T \phi\right)^T X_c^T \phi = \max_\phi \phi^T X_c X_c^T \phi}$$

Principal component analysis

- recall that 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

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

Principal component analysis

- hence the PCA problem is

$$\boxed{\max_{\phi} \phi^T X_c X_c^T \phi = \max_{\phi} \phi^T \Sigma \phi}$$

- as in LDA, this can be made arbitrarily large by simply scaling ϕ
- to normalize we constrain ϕ to have unit norm

$$\boxed{\max_{\phi} \phi^T \Sigma \phi \quad \text{subject to} \quad \|\phi\| = 1}$$

- which is equivalent to

$$\boxed{\max_{\phi} \frac{\phi^T \Sigma \phi}{\phi^T \phi}}$$

- shows that PCA = maximization of a Rayleigh quotient

Principal component analysis

- in this case

$$\boxed{\max_{w} \frac{w^T S_B w}{w^T S_W w}} \qquad \boxed{\begin{array}{r} S_B, S_W, symmetric \\ positive\ semidefinite \end{array}}$$

- with $S_B = \Sigma$ and $S_w = I$
- S_w is clearly invertible
  - no regularization problems
  - w* is the eigenvector of largest eigenvalue of $S_w^{-1}S_B$
  - this is just the largest eigenvector of the covariance Σ
  - the max value is λ, where λ is the largest eigenvalue

The Rayleigh quotient dual

- let's assume, for a moment, that the solution is of the form

$$\boxed{w = X_c \alpha}$$

  - i.e. a linear combination of the centered datapoints

- hence the problem is equivalent to

$$\boxed{\max_{\alpha} \frac{\alpha^T X_c^T S_B X_c \alpha}{\alpha^T X_c^T S_W X_c \alpha}}$$

- this does not change its form, the solution is
  - α^* is the eigenvector of largest eigenvalue of $(X_c^T S_W X_c)^{-1} X_c^T S_B X_c$
  - the max value is λK, where λ is the largest eigenvalue

The Rayleigh quotient dual

- for PCA
  - $S_w = I$ and $S_B = \Sigma = 1/n\ X_c X_c^T$
  - the solution satisfies

$$\boxed{S_B w = \lambda S_W^{-1} w \Leftrightarrow \frac{1}{n} X_c X_c^T w = \lambda w \Leftrightarrow w = X_c \underbrace{\frac{1}{n\lambda} X_c^T w}_{\alpha}}$$

  - and, therefore, we have
    - w* eigenvalue of $S_B = X_c X_c^T$
    - α* eigenvalue of $(X_c^T S_W X_c)^{-1} X_c^T S_B X_c = (X_c^T X_c)^{-1} X_c^T X_c X_c^T X_c = X_c^T X_c$
  - i.e. we have two alternative manners in which to compute PCA

Principal component analysis

- primal
  - assemble matrix
  - $\Sigma = X_c X_c^T$
  - compute eigenvectors ϕ_i
  - these are the principal components

- dual
  - assemble matrix
  - $K = X_c^T X_c$
  - compute eigenvectors α_i
  - the principal components are
  - $\phi_i = X_c \alpha_i$

- in both cases we have an eigenvalue problem
  - primal on the sum of outer products $\boxed{\Sigma = \sum_i x_i^c \left(x_i^c\right)^T}$
  - dual on the matrix of inner products $\boxed{K_{ij} = \left(x_i^c\right)^T x_j^c}$

The Rayleigh quotient

- this is a property that holds for many Rayleigh quotient problems
  - the primal solution is a linear combination of datapoints
  - the dual solution only depends on dot-products of the datapoints
- whenever both of these hold
  - the problem can be kernelized
  - this has various interesting properties
  - we will talk about them
- many examples
  - kernel PCA, kernel LDA, manifold learning, etc.

Any questions?