### 3.8.1 Principal Component Analysis (PCA)

We begin by considering the problem of representing all of the vectors in a set of $n$ $d$-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by a single vector $\mathbf{x}_0$. To be more specific, suppose that we want to find a vector $\mathbf{x}_0$ such that the sum of the squared distances between $\mathbf{x}_0$ and the various $\mathbf{x}_k$ is as small as possible. We define the squared-error criterion function $J_0(\mathbf{x}_0)$ by

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_k||^2, \tag{78}$$

and seek the value of $\mathbf{x}_0$ that minimizes $J_0$. It is simple to show that the solution to this problem is given by $\mathbf{x}_0 = \mathbf{m}$, where $\mathbf{m}$ is the sample mean,

$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k. \tag{79}$$

This can be easily verified by writing

$$\begin{aligned}
J_0(\mathbf{x}_0) &= \sum_{k=1}^{n} ||(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})||^2 \\
&= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2\sum_{k=1}^{n} (\mathbf{x}_0 - \mathbf{m})^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\
&= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2(\mathbf{x}_0 - \mathbf{m})^t \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\
&= \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 + \underbrace{\sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2}_{independent\ of\ \mathbf{x}_0}.
\end{aligned} \tag{80}$$

Since the second sum is independent of $\mathbf{x}_0$, this expression is obviously minimized by the choice $\mathbf{x}_0 = \mathbf{m}$.

The sample mean is a zero-dimensional representation of the data set. It is simple, but it does not reveal any of the variability in the data. We can obtain a more interesting, one-dimensional representation by projecting the data onto a line running through the sample mean. Let $\mathbf{e}$ be a unit vector in the direction of the line. Then the equation of the line can be written as

$$\mathbf{x} = \mathbf{m} + a\mathbf{e}, \tag{81}$$

where the scalar $a$ (which takes on any real value) corresponds to the distance of any point $\mathbf{x}$ from the mean $\mathbf{m}$. If we represent $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$, we can find an "optimal" set of coefficients $a_k$ by minimizing the squared-error criterion function

$$\begin{aligned}
J_1(a_1, \ldots, a_n, \mathbf{e}) &= \sum_{k=1}^{n} ||(\mathbf{m} + a_k\mathbf{e}) - \mathbf{x}_k||^2 = \sum_{k=1}^{n} ||a_k\mathbf{e} - (\mathbf{x}_k - \mathbf{m})||^2 \\
&= \sum_{k=1}^{n} a_k^2 ||\mathbf{e}||^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2.
\end{aligned} \tag{82}$$

Recognizing that $||\mathbf{e}|| = 1$, partially differentiating with respect to $a_k$, and setting the derivative to zero, we obtain

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m}). \tag{83}$$

Geometrically, this result merely says that we obtain a least-squares solution by projecting the vector $\mathbf{x}_k$ onto the line in the direction of $\mathbf{e}$ that passes through the sample mean.

SCATTER MATRIX

This brings us to the more interesting problem of finding the *best* direction $\mathbf{e}$ for the line. The solution to this problem involves the so-called *scatter matrix* $\mathbf{S}$ defined by

$$\mathbf{S} = \sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t. \tag{84}$$

The scatter matrix should look familiar—it is merely $n - 1$ times the sample covariance matrix. It arises here when we substitute $a_k$ found in Eq. 83 into Eq. 82 to obtain

$$\begin{aligned} J_1(\mathbf{e}) &= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\sum_{k=1}^{n}[\mathbf{e}^t(\mathbf{x}_k - \mathbf{m})]^2 + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\sum_{k=1}^{n} \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\mathbf{e} + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2 \\ &= -\mathbf{e}^t\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2. \end{aligned} \tag{85}$$

Clearly, the vector $\mathbf{e}$ that minimizes $J_1$ also maximizes $\mathbf{e}^t\mathbf{S}\mathbf{e}$. We use the method of Lagrange multipliers (described in Section A.3 of the Appendix) to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to the constraint that $||\mathbf{e}|| = 1$. Letting $\lambda$ be the undetermined multiplier, we differentiate

$$u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^t\mathbf{e} - 1) \tag{86}$$

with respect to $\mathbf{e}$ to obtain

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e}. \tag{87}$$

Setting this gradient vector equal to zero, we see that $\mathbf{e}$ must be an eigenvector of the scatter matrix:

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}. \tag{88}$$

In particular, because $\mathbf{e}^t\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^t\mathbf{e} = \lambda$, it follows that to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$, we want to select the eigenvector corresponding to the largest eigenvalue of the scatter matrix. In other words, to find the best one-dimensional projection of the data (best in the least-

PCA

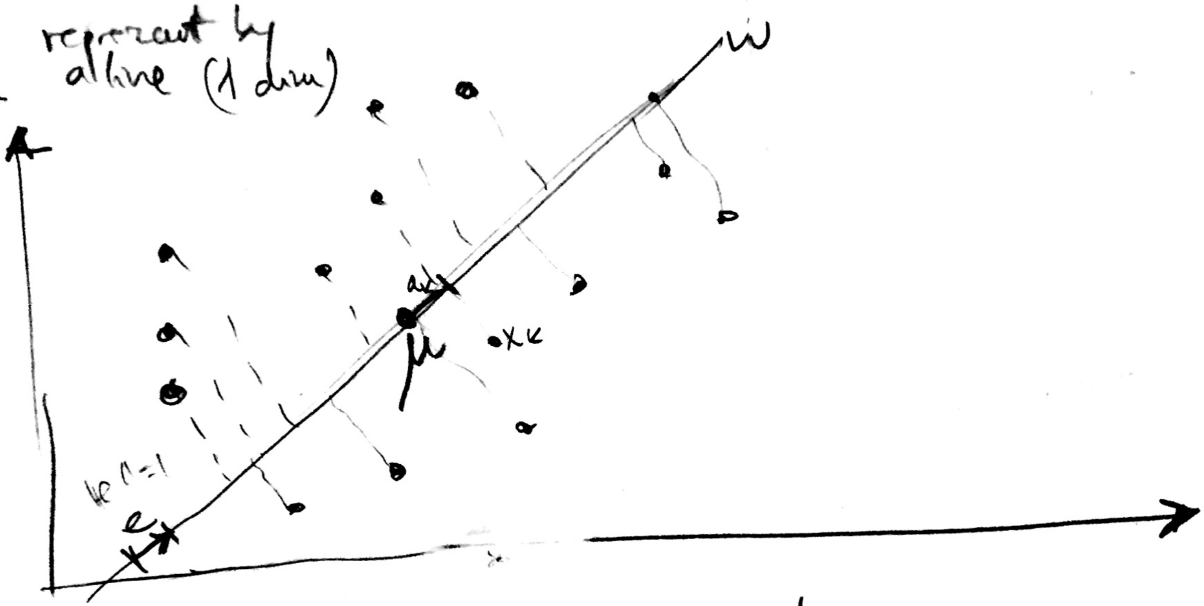
(1) REPREZENT BY A POINT (0 dimensione)

that would be the mean

$$\mu = \frac{1}{n}\sum_k x_k$$

$$\mu = \underset{x}{\mathrm{argmin}} \sum_k \|x - x_k\|^2$$

② PCA represent by a line (1 dim)




the line will pass throug the mean

we need to define and find a good w.

$x_k = \mu + a_k e$ $\quad$ $\|e\|=1$ is the direction of the line

$a_k$ = scalar

$$J(a_1, a_2, \ldots, e) = \sum_k \| \mu + a_k e - x_k \|^2 = \sum_k \| a_k e - (x_k - \mu) \|^2 =$$

$$= \sum_k a_k^2 \|e\|^2 - 2 \sum a_k e^T (x_k - \mu) + \sum \| x_k - \mu \|^2$$

$$\frac{\partial J}{\partial a_i} = 0 \Rightarrow a_k = e^T (x_k - \mu)$$

$$E[\mu + a_k e] = \mu + E[a_k e] = \mu + e^T E[x_k - \mu] e = \mu$$

MOST IMPORTANT: what $\boxed{e}$ is a good direction?

- minimize $J$
- maximize the variance of the projections on $e$.

---

variance of projections

$$E\left[\left(\cancel{\mu} + a_k e - E[\mu + a_k e]\right)^2\right] = E\left[\cancel{\mu} + a_k e - \cancel{\mu}\right] = E\left[(a_k e)^2\right]$$

$$= E\left[e^T(x_k-\mu) \cdot e^T(x_k-\mu)\right] = E\left[e^T(x_k-\mu)(x_k-\mu)^T e\right] =$$

$$= e^T \Sigma e$$

where $$\Sigma = \sum_k (x_k-\mu)(x_k-\mu)^T = \begin{array}{cc} & j \\ i & \begin{bmatrix} \sum_k (x_k^1-\mu^1)(x_k^1-\mu^1) & & \sum_k (x_k^1-\mu^1)(x_k^d-\mu^d) \\ & \sum_k (x_k^i-\mu^i)(x_k^j-\mu^j) & \\ \sum_k (x_k^d-\mu^d)(x_k^1-\mu^1) & & \sum_k (x_k^d-\mu^d)(x_k^d-\mu^d) \end{bmatrix} \end{array}$$

$$= \begin{bmatrix} \sigma_{11} & & \sigma_{1d} \\ & \sigma_{ij} & \\ \sigma_{dd} & & \sigma_{dd} \end{bmatrix}$$ – covariance matrix

---

$$J(e) = \sum_k a_k^2 - 2\sum a_k^2 + \sum_k \|x_k - \mu\|^2 =$$

$$= -\sum_k \left(e^T(x_k-\mu)\right)^2 + \sum_k \|x_k - \mu\|^2 =$$

$$= \underbrace{-\sum_k e^T(x_k-\mu)(x_k-\mu)^T e}_{-e^T\Sigma e} + \sum_k \|x_k-\mu\|^2$$

So minimizing the error $\Longleftrightarrow$ maximize the variance of projections.

maximize $e^T \Sigma e$

subject to $\|e\| = 1 \Leftrightarrow e^T e = 1$

Lagrangian $L = \max \left[ e^T \Sigma e - \alpha (e^T e - 1) \right]$

$$\frac{\partial L}{\partial e} = 0 \Leftrightarrow 2\Sigma e - 2\alpha e = 0 \Rightarrow \Sigma e = \alpha e$$

$$\Downarrow$$

$$\boxed{\begin{array}{l} e = \text{eigen vector of } \Sigma \\ \alpha = \text{eigen value of } \Sigma \end{array}}$$

$$e^T \Sigma e = e^T \alpha e = \alpha$$

$$\Downarrow$$

We need to choose the (eigen vector, eigen value) pair with the biggest eigen value.

---

Say we want a second "biggest" dimension.

- constrained that is orthogonal on the first dim $e_1 = e$ so that measures a diff variance component

Lagrangian: $\max L = \left[ e_2^T \Sigma e_2 - \alpha(e_2^T e_2 - 1) - \beta (e_2^T e_1 - 0) \right]$

$$\partial L = 0 \Rightarrow 2\Sigma e_2 - 2\alpha e_2 - \beta e_1 = 0$$

$$\Updownarrow$$

$$\underbrace{2e_1^T \Sigma e_2}_{\substack{0 \\ e_1^T e_2 = 0}} - \underbrace{2e_1^T \alpha e_2}_{\substack{0 \\ e_1^T e_2 = 0}} - e_1^T \beta e_1 = 0 \Rightarrow \beta = 0$$

$$\Downarrow$$

$$\Sigma e_2 = \alpha e_2 \Rightarrow e_2 \text{ eigenvector}$$

$$\alpha = \lambda_2 = \text{second eigenvalue}$$

Spectral decomposition $\Sigma$ symmetric, pos def? $\Rightarrow e_i$ orthogonal.

$$C = \begin{pmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_d \\ | & | & & | \end{pmatrix}$$ $e_i$ = eigenvectors of $\Sigma$, normalized

$e_i^T e_j = 0$ $i \neq j$; $\|e_i\| = 1$. Then.

$C \cdot C^T = I_d$

$$\Sigma = \Sigma C C^T = \Sigma \begin{pmatrix} e_1 & e_2 & \cdots & e_d \end{pmatrix} C^T = \left(\Sigma e_1, \Sigma e_2 \ldots \Sigma e_d\right) C^T$$

$$= \left(\lambda_1 e_1, \lambda_2 e_2, \ldots \lambda_d e_d\right) C^T = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T \ldots + \lambda_d e_d e_d^T$$

$$= C D C^T \text{ where } D = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_d \end{pmatrix}$$

$$\Sigma = \begin{bmatrix} | & | & \\ e_1 & e_2 & \cdots \\ | & | & \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_d \end{bmatrix} \begin{bmatrix} — e_1 — \\ — e_2 — \\ \vdots \\ — e_d — \end{bmatrix}$$

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$

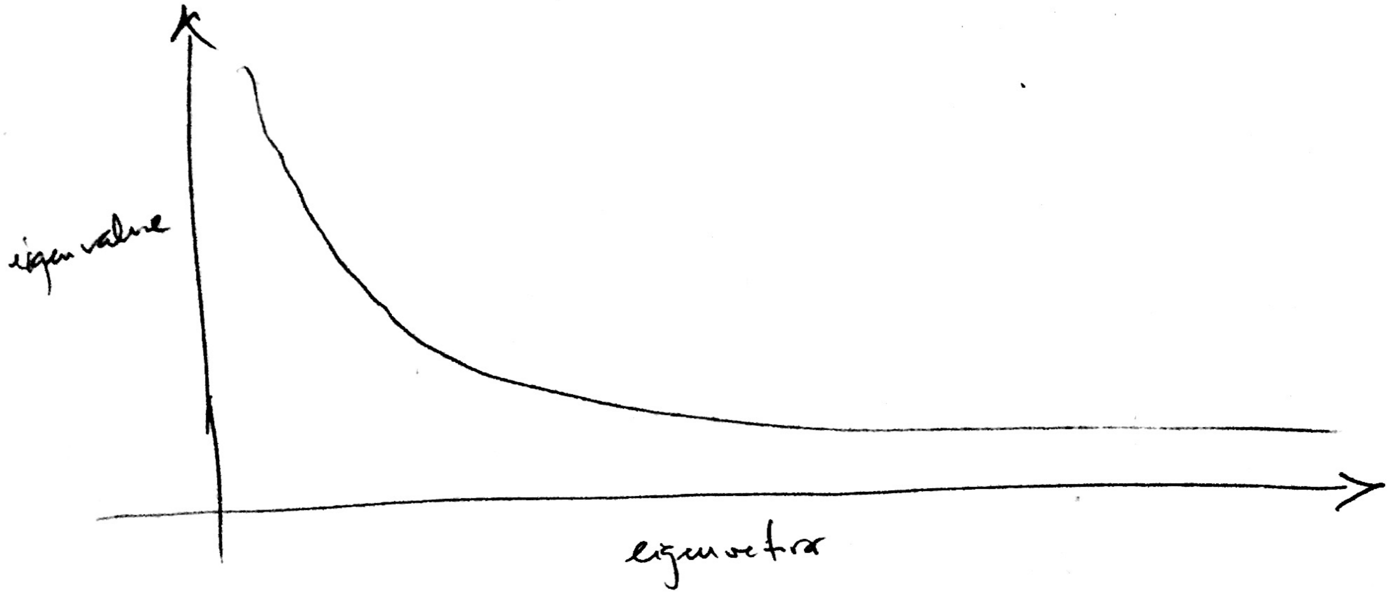
- if some $\lambda_{t+1} = \lambda_{t+2} = \lambda_d = 0$ then we can only keep $e_1, e_2 \ldots e_t$; $\lambda_1, \lambda_2 \ldots \lambda_t$ so we reduced to $t$ dimens.
- if <u>approximate</u> to $t$ dimensions eliminate $t+1 \to d$ eigenv. even if they are <u>not</u> 0

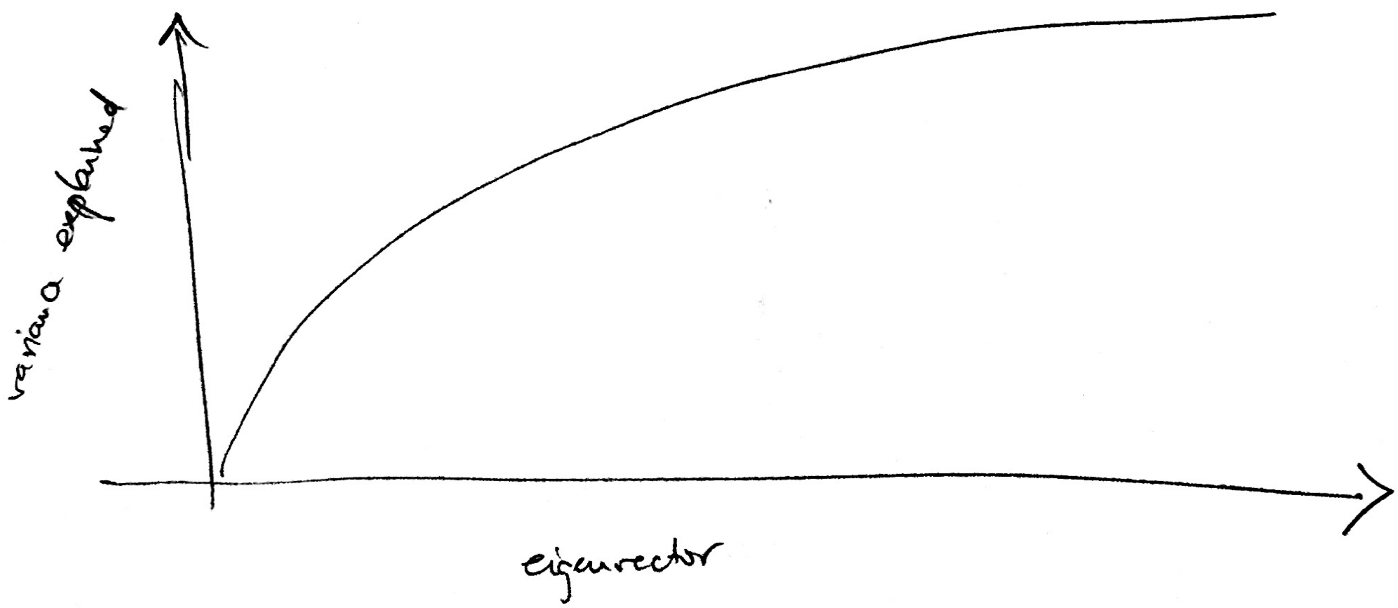
---

$\Sigma$ symmetric $\Rightarrow e_i$ orthogonal:

$$(\lambda_1 e_1)^T e_2 = (A e_1)^T e_2 = e_1^T A^T e_2 = e_1^T (A e_2) = e_1^T \lambda_2 e_2.$$

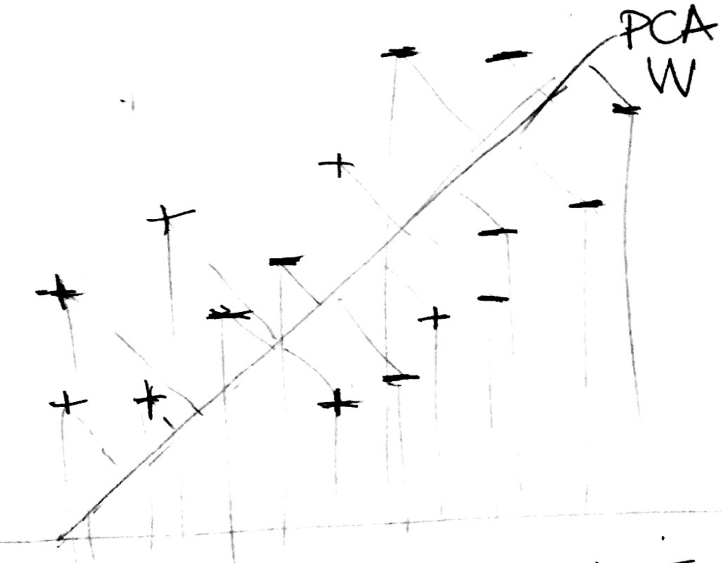
But $\lambda_1 \neq \lambda_2 \Rightarrow e_1^T e_2 = 0$.


eigenvalue
eigenvector


variance explained
eigenvector

# FISHER LINEAR DISCRIMINANT

Find dimensions that help classify dat. (PCA might loose them)




class better dim.

$m_1$ points $\in C_1$
$m_2$ points $\in C_2$

$$\mu_1 = \frac{1}{m_1} \sum_{x \in C_1} x$$

$$\mu_2 = \frac{1}{m_2} \sum_{x \in C_2} x$$

$\|w\|=1$ w = the line ; projected points $z = wx$

projected means: $\bar{\mu}_1 = \frac{1}{m_1} \sum_{C_1} w^T x = w^T \mu_1$

distance between projected means is $|\bar{\mu}_1 - \bar{\mu}_2| = |w^T(\mu_1 - \mu_2)|$

$$\Sigma_i = \sum_{C_i} (x - \mu_i)(x - \mu_i)^T \qquad \Sigma_W = \Sigma_1 + \Sigma_2$$

$$\sigma_i^2 = \mathrm{Var}_{C_i}[w^T x] = w^T \Sigma_i w \qquad \sigma_1^2 + \sigma_2^2 = w^T \Sigma_1 w + w^T \Sigma_2 w = w^T \Sigma_W w$$

$$\Sigma_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$(\bar{\mu}_1 - \bar{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \Sigma_B w$$

Fihers linear disc criteria

$$(\text{maximize}) \quad J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2} =$$

$$= \frac{w^T \Sigma_B w}{w^T \Sigma_W w}$$ (sometimes called "Rayleigh" quotient)

Solution must (max J) satisfy $\Sigma_B w = \lambda \Sigma_W w$

(A) if $\Sigma_W$ nonsingular $\Rightarrow \Sigma_W^{-1} \Sigma_B w = \lambda w \Rightarrow$
$(\lambda, w)$ eigens of $\Sigma_W^{-1} \Sigma_B$

(B) for us, we do not need $\Sigma_W^{-1} \Sigma_B$ because $\Sigma_B w$ is in the direction of $(\mu_1 - \mu_2)$, so $w = \Sigma_W^{-1} (m_1 - m_2)$

(C) All it is left is to find a threshold on the projections.