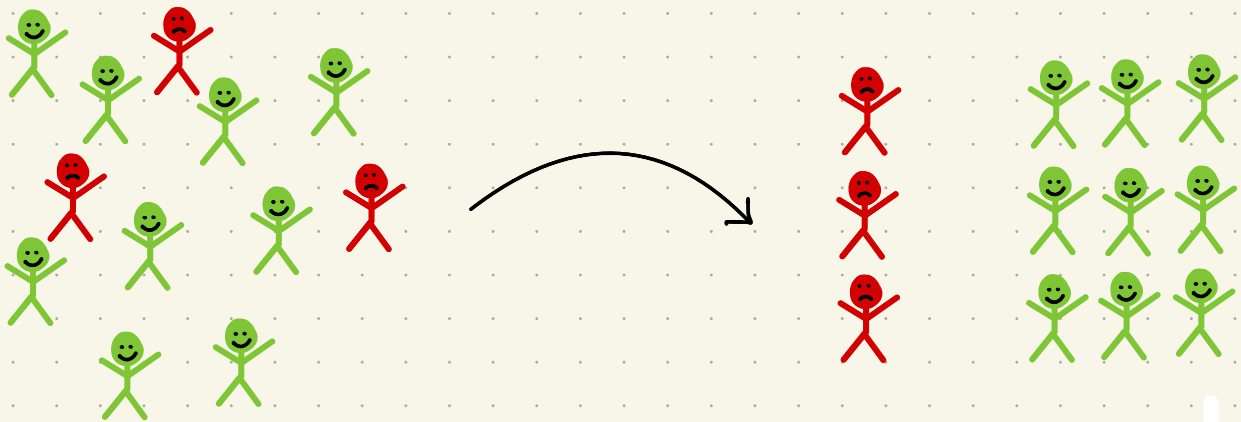
Classification



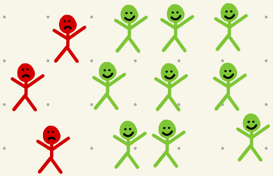
Classification task

Given feature matrix X and response y

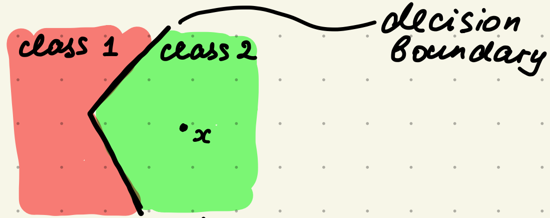
$$X = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix} \in \mathbb{R}^{n \times p} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

where $y_i \in \{1, \dots, K\}$ gives class of i-th observation.

Denote by $C_k \subseteq \{1 \dots K\}$ the subset of observations that belong to class k and $n_k = |C_k|$.



Data $\{x_i, y_i\}_{i=1}^n$



$h: \mathcal{X} \longrightarrow \{1 \dots K\}$
decision rule

Examples:

- predict if a patient is sick, $\quad y_i \in \{\text{sick}, \text{healthy}\}$
- predict image class $\quad y_i \in \{\text{cat}, \text{dog}, \text{bird}\}$
- predict movie genre $\quad y_i \in \{\text{comedy}, \text{thriller}, \text{drama}\}$

Linear discriminant analysis

Denote by z a random variable giving the class label.

$$z = \begin{cases} 1 & \text{with probability } \pi_1 \\ \vdots & \\ k & \text{with probability } \pi_k \end{cases}$$ priors

We want to build Bayes classifier

$$h(x) = \underset{k=1\dots k}{\operatorname{argmax}} P(z=k \mid X=x)$$

Equivalently, $h(x) = \underset{k=1..k}{\operatorname{argmax}} P(X=x \mid Z=k)\, \pi_j$

$$P(z=k \mid X=x) = \frac{P(X=x \mid Z=k)\pi_k}{\sum_{j=1}^{k} P(X=x \mid Z=j)\, \pi_j}$$ ← common

$$h(x) = \underset{k=1\dots k}{\operatorname{argmax}} P(X=x \mid Z=k) \cdot \pi_k$$

Recall normal density:

$$f(x;\mu,\Sigma)=\left(\frac{1}{2\pi}\right)^{p/2}\frac{1}{|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Linear discriminant analysis (LDA) assumes

$$X=x\,/\,Z=k \sim N_p(\mu_k,\Sigma)$$

same variance

- The decision rule is $h(x)=\underset{k=1\dots K}{\operatorname{argmax}}\,(a_k+b_k^T x)$

$$h(x)=\underset{k=1\dots K}{\operatorname{argmax}}\left(\log f(x;\mu_k,\Sigma)+\log \pi_k\right)=$$

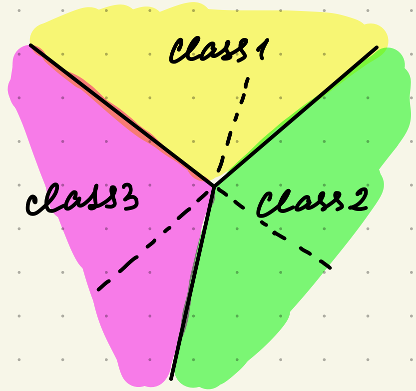
$$=\underset{k=1\dots K}{\operatorname{argmax}}\left(\cancel{\log\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}}-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)+\log\pi_k\right)$$

$$=\underset{k=1\dots K}{\operatorname{argmax}}\left(-\frac{1}{2}(\cancel{x^T\Sigma^{-1}x}-2\mu_k^T\Sigma^{-1}x+\mu_k^T\Sigma^{-1}\mu_k)+\log\pi_k\right)=$$

$$=\underset{k=1\dots K}{\operatorname{argmax}}\,(a_k+b_k^T x)$$ – linear functions of x

with $a_k=\log\pi_k-\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$ and $b_k=\mu_k^T\Sigma^{-1}$

- The decision boundaries are linear.



Lets take two classes k, k'

$i \in C_k$ if $(a_k + b_k^T x) > (a_{k'} + b_{k'}^T x)$

Thus, $a_k - a_{k'} + (b_k - b_{k'})^T x > 0$

(it's a half-space)

- In practice, we don't know π_k, μ_k, Σ, so we need to estimate them.

$$\hat{\pi}_k = \frac{n_k}{n} \qquad \hat{\mu}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T = \sum_{k=1}^{K} \frac{n_k}{n} S_k$$

where $S_k = \frac{1}{n_k} \sum_{i \in C_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$ - sample covariance for C_k

- If Σ is not common for groups, we appey quadratic discriminat analysis (QDA)

$$h(x) = \underset{k=1..K}{\operatorname{argmax}} \left(-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k \right)$$

$$-\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k =$$

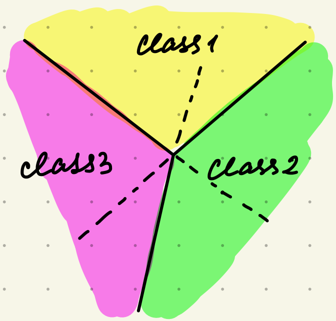
$$= -\frac{1}{2}\log|\Sigma_{k'}| - \frac{1}{2}(x-\mu_{k'})^T\Sigma_{k'}^{-1}(x-\mu_{k'}) + \log \pi_{k'}$$

$$-\frac{1}{2} x^T \underbrace{\left(\Sigma_k^{-1} - \Sigma_{k'}^{-1}\right)}_{Q} x + \underbrace{\left(\mu_k^T\Sigma_k^{-1} - \mu_{k'}^T\Sigma_{k'}^{-1}\right)}_{P^T} x \underbrace{- \frac{1}{2}\left(\mu_k^T\Sigma_k^{-1}\mu_k - \mu_{k'}^T\Sigma_{k'}^{-1}\mu_{k'}\right)}$$

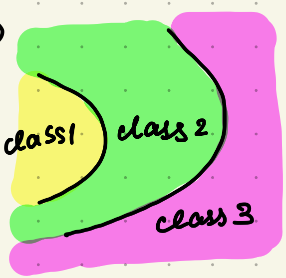
$$\underbrace{+ \log \pi_k - \log \pi_{k'}}_{R} = x^T Q x + P^T x + R \text{ - quadratic function}$$

Estimates for $\hat{\Sigma}_k = S_k = \frac{1}{n_k}\sum_{i\in C_k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

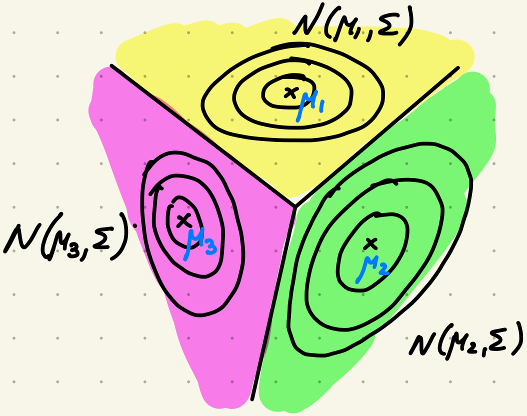
LDA



QDA



Geometry of LDA



Is the decision boundary the same as Voronoi tesselation?

$$h_V(x) = \underset{k=1\dots K}{\operatorname{argmin}} \|x - \mu_k\|^2$$

$$h(x) = \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2} (x-\mu_k)^T \Sigma^{-1} (x-\mu_k) - \log \pi_k \right) =$$

$$= \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2} (x-\mu_k)^T \Sigma^{-1/2} \Sigma^{-1/2} (x-\mu_k) - \log \pi_k \right) =$$

$$= \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2} \|\Sigma^{-1/2}(x-\mu_k)\|^2 - \log \pi_k \right) =$$

$$= \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2} \|x^* - \mu_k^*\|^2 - \log \pi_k \right)$$

Where x^* and μ_k^* are points in the transformed space

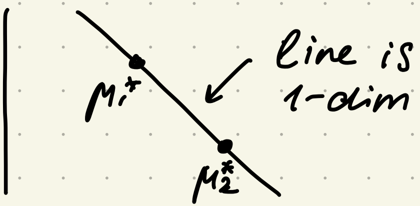
- The transformation $x^* = \Sigma^{-1/2} x$ is sphering, i.e. $\operatorname{cov}(x^*) = I$.

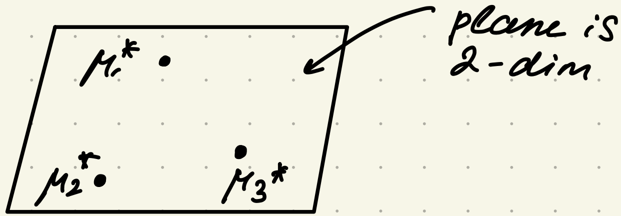
  $\operatorname{cov}(\Sigma^{-1/2} x) = \Sigma^{-1/2} \Sigma \, \Sigma^{-1/2} = I$

- Therefore, LDA classifies the points in the transformed space according to

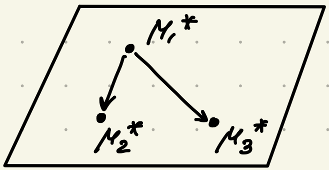
$$h(x^*) = \underset{k=1\ldots K}{\operatorname{argmin}} \Big(\underbrace{\tfrac{1}{2} \| x^* - \mu_k^* \|^2}_{\text{nearest centroid}} - \underbrace{\log \pi_k}_{\text{adjustment by the class size.}} \Big)$$

- $\mu_1^* \ldots \mu_K^*$ lie in a plane M of dimension $\leq K-1$.





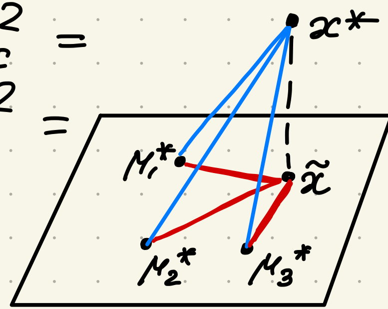
- Denote by P_M the projection operator onto this low-dimensional plane.



Find basis Q with QR for $(\mu_2^* - \mu_1^*, \dots, \mu_k^* - \mu_1^*)$. Then $P_M = QQ^T$.

- For the projected points $\tilde{x} = P_M x_*$ the value $h(\tilde{x})$ is the same as $h(x_*)$.

$$\|x^* - \mu_k^*\|^2 = \|P_M(x^* - \mu_k^*) + P_{M_\perp}(x^* - \mu_k^*)\|_F^2 =$$
$$= \|P_M(x^* - \mu_k^*)\|^2 + \|P_{M_\perp}(x^* - \mu_k)\|^2 =$$
$$= \|\tilde{x} - \mu_k^*\|^2 + \|P_{M_\perp}(x^*)\|^2$$



$$h(x^*) = \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2}\|x^* - \mu_k^*\|^2 - \log \pi_k \right) =$$
$$= \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2}\|\tilde{x} - \mu_k^*\|^2 + \frac{1}{2}\|P_{M_\perp}(x^*)\|^2 - \log \pi_k \right) =$$
$$= \underset{k=1\dots K}{\operatorname{argmin}} \left(\frac{1}{2}\|\tilde{x} - \mu_k^*\|^2 - \log \pi_k \right) = h(\tilde{x})$$

LDA procedure:

1. Compute $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}$
2. Sphere the data and transform centroids, project sphered data onto the plane M containing the sphered centroids.

   This can be combined in a single linear transformation

   $$\tilde{x}_i = A x_i \qquad \tilde{\mu}_i = A \hat{\mu}_i$$

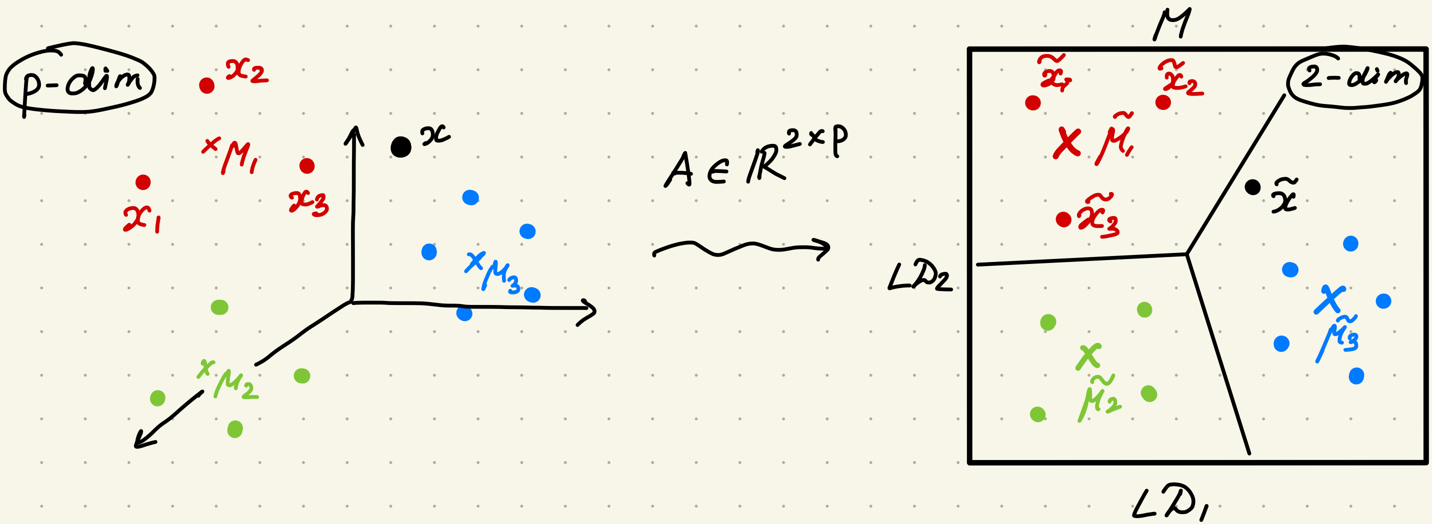
3. Given a new point $x \in \mathbb{R}^p$ transform it to $\tilde{x} = Ax$, then classify to the nearest centroid $\tilde{\mu}_i$, adjusting for class proportions

- Decision boundaries in the M space are also linear.

$$h(\tilde{x}) = \underset{k=1..K}{\operatorname{argmin}} \left(\frac{1}{2} \|\tilde{x} - \tilde{\mu}_k\|^2 - \log \pi_k \right)$$

$$= \underset{k=1...n}{\operatorname{argmax}} \Big(\underbrace{\log \pi_k - \frac{1}{2} \|\tilde{\mu}_k\|^2}_{\tilde{a}_k} + \underbrace{\tilde{\mu}_k^T}_{\tilde{b}_k} \tilde{x} \Big) = \underset{k=1..K}{\operatorname{argmax}} (\tilde{a}_k + \tilde{b}_k^T \tilde{x})$$

<u>Example</u>: If $K=3$, M is two-dimensional.



Reduced-rank DA (Fisher)

For $K > 3$ we can find an L-dimensional subspace of M to project onto $(L < K-1)$

Choose the subspace that spread out the projected centroids (like in PCA!)

- $$S = \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x})(x_i - \bar{x})^T}_{\text{total covariance}} =$$

$$= \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T}_{W\text{-within-class}} + \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T}_{B\text{-between-class}}$$

$$W = \sum_{k=1}^{K} \frac{n_k}{n} \cdot S_k \simeq \sum_{k=1}^{K} \pi_k \Sigma_k \;(= \Sigma \text{ if common covariance})$$

$$B = \sum_{k=1}^{K} \frac{n_k}{n} \left(\bar{x}_k - \sum_{k=1}^{K} \frac{n_k}{n} \bar{x}_k\right)\left(\bar{x}_k - \sum_{k=1}^{K} \frac{n_k}{n} \bar{x}_k\right)^T \simeq \sum_{k=1}^{K} \pi_k (\mu_k - \mu)(\mu_k - \mu)^T$$

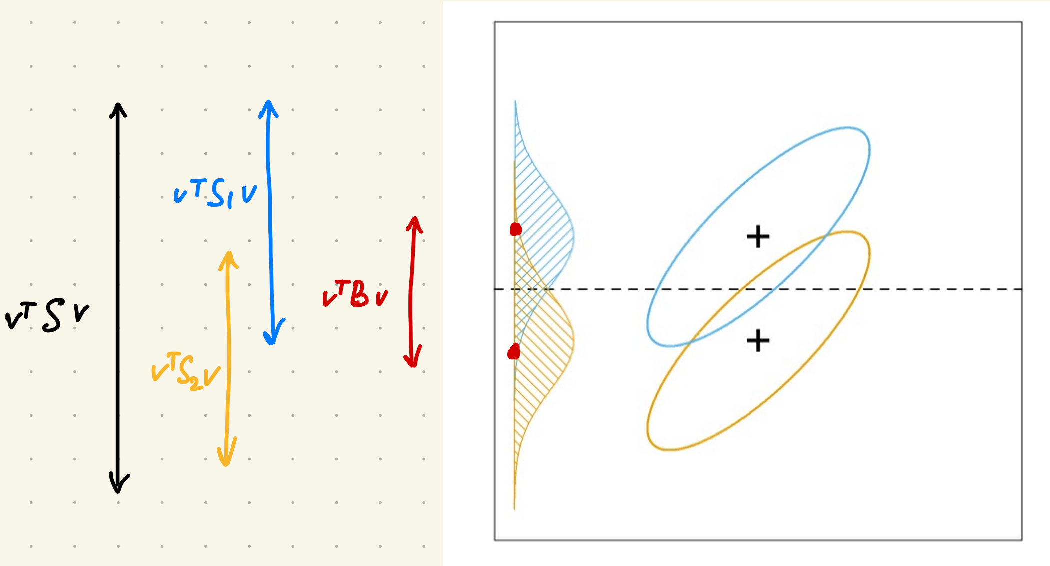
with $\mu = \sum_{k=1}^{K} \pi_k \mu_k$ – Covariance & mean of centroids weighted by π_k

- Suppose we want to project data onto direction v, then the variance of projections is

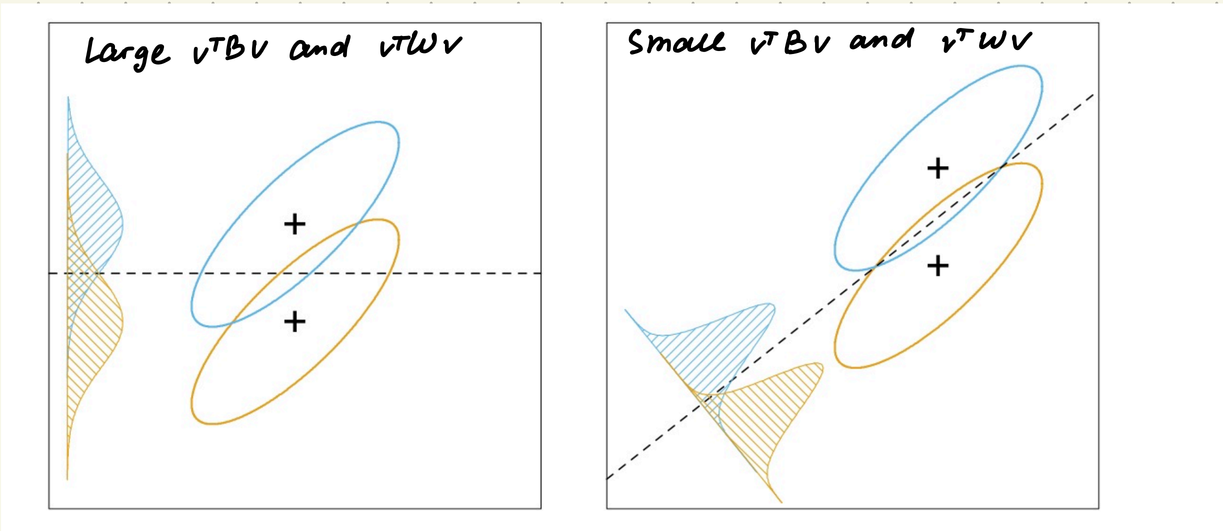
$$v^T S v = v^T W v + v^T B v$$

$$v^T W v = \sum_{k=1}^{n} \frac{n_k}{n} v^T S_k v = \sum_{k=1}^{K} \pi_k \cdot (\text{variance of projection in } C_k)$$

$$v^T B v = \text{variance of projectins of centroids (weighted by } \pi_k)$$



We want large $v^T B v$ and small $v^T W v$



Fisher's proposition:

$$\underset{v \in \mathbb{R}^p}{\text{maximize}} \quad \frac{v^T B v}{v^T W v}$$

Rayleigh quotient

- The solution v is the largest eigenvector of $W^{-1}B$.

Let $\tilde{v} = W^{1/2}v$ then $v = W^{-1/2}\tilde{v}$ and Fisher's problem

$$\underset{\tilde{v}\in\mathbb{R}^p}{\text{maximize}} \quad \frac{\tilde{v}^T(W^{-1/2}BW^{-1/2})\tilde{v}}{\|\tilde{v}\|^2}$$

Equivalently, if $\tilde{B} = W^{-1/2}BW^{-1/2}$ we solve

$$\underset{\tilde{v}\in\mathbb{R}^p}{\text{maximize}} \quad \tilde{v}^T\tilde{B}\tilde{v} \text{ subject to } \|\tilde{v}\| = 1$$

Now, $\tilde{v}$ is the largest e.vector of $\tilde{B} \Rightarrow \tilde{B}\tilde{v} = \lambda_1\tilde{v}$

$$W^{-1/2}BW^{-1/2}\tilde{v} = W^{-1/2}B\cancel{W^{-1/2}}\,\cancel{W^{1/2}}v = W^{-1/2}Bv = \lambda_1 W^{-1/2}v$$

Then, $W^{-1}Bv = \lambda_1 v$.

- If we assumed common covariance for classes ($W \simeq \Sigma$), then $\tilde{B} = \sum_{k=1}^{K} \pi_k (\tilde{\mu}_k - \tilde{\mu})(\tilde{\mu}_k - \tilde{\mu})^T$ with $\tilde{\mu}_k = A\mu_k$ where A is sphering + projection operator from LDA.

$$\Sigma^{-1/2}\mu_k = \mu_k^* = \tilde{\mu}_k = A\mu_k$$

$$\tilde{B} = \Sigma^{-1/2} B \Sigma^{-1/2} = \sum_{k=1}^{K} \pi_k \underbrace{\Sigma^{-1/2}(\mu_k - \mu)}\underbrace{(\mu_k - \mu)^T \Sigma^{-1/2}} =$$

$$= \sum_{k=1}^{K} \pi_k (\tilde{\mu}_k - \tilde{\mu})(\tilde{\mu}_k - \tilde{\mu})^T$$

- Thus $\tilde{v}$ is the top eigenvector of $\tilde{B}$ and is the first PC direction of $\tilde{\mu}_1 \ldots \tilde{\mu}_K$ (where observations in PCA are weighted by $\pi_1 \ldots \pi_K$)

Summary (LDA + dimension reduction)

- Compute centroids $\mu_1, \dots, \mu_K$
- Compute within class covariance W
- Transform centroids $\tilde{\mu}_k = W^{-1/2} \mu_k$
- Compute between class covariance $\tilde{B}$
- Denote by $\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_L$ the eigenvectors of $\tilde{B}$
- Project data onto $v_1 \dots v_\ell$ with $v_\ell = W^{-1/2} \tilde{v}_\ell$

LD directions