EE269
Signal Processing for Machine Learning

Lecture 9

Instructor : Mert Pilanci

Stanford University

October 12, 2020

Recap:Covariance Estimation

- Suppose $x_1, x_2, ...x_n$ i.i.d. $\sim N(\mu, \Sigma)$
- Estimating means

  $\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_n$
- Estimating covariances

  $\Sigma_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$

  Regularized covariance: $\hat{\Sigma} = (1 - \alpha)\,\mathrm{diag}(\Sigma_{ML}) + \alpha \Sigma_{ML}$

Recap:Covariance Estimation

- Suppose $x_1, x_2, ...x_n$ i.i.d. $\sim N(\mu, \Sigma)$
- Estimating means

  $\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_n$
- Estimating covariances

  $\Sigma_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$

  Regularized covariance: $\hat{\Sigma} = (1 - \alpha)\,\mathrm{diag}(\Sigma_{ML}) + \alpha \Sigma_{ML}$
- LDA

  Estimate μ_k, for $k = 1..., K$ and Σ

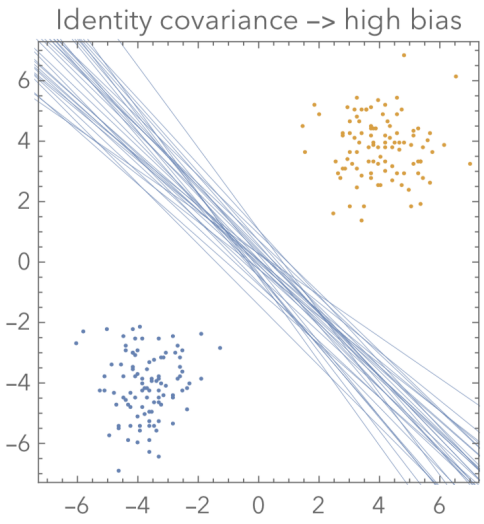
  $Kn + \binom{n}{2} + n$ parameters
- QDA

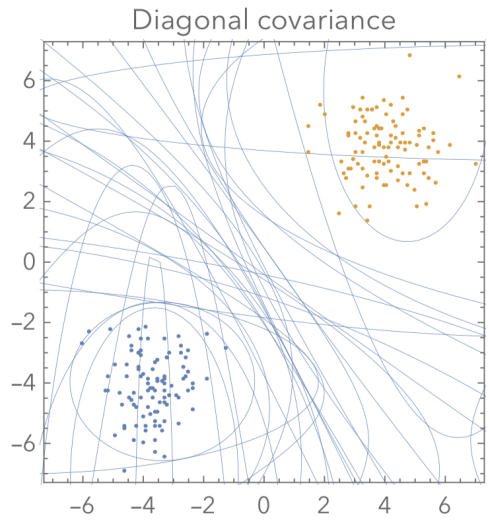
  Estimate μ_k, Σ_k for $k = 1..., K$

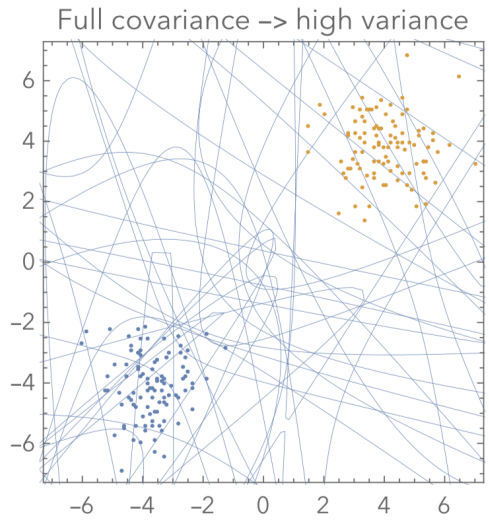
  $Kn + K\left(\binom{n}{2} + n\right)$ parameters

Recap: Stability of Covariance

- Estimating covariance on 3-sample subsets







slide credit: P. Smaragdis

Recap: Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

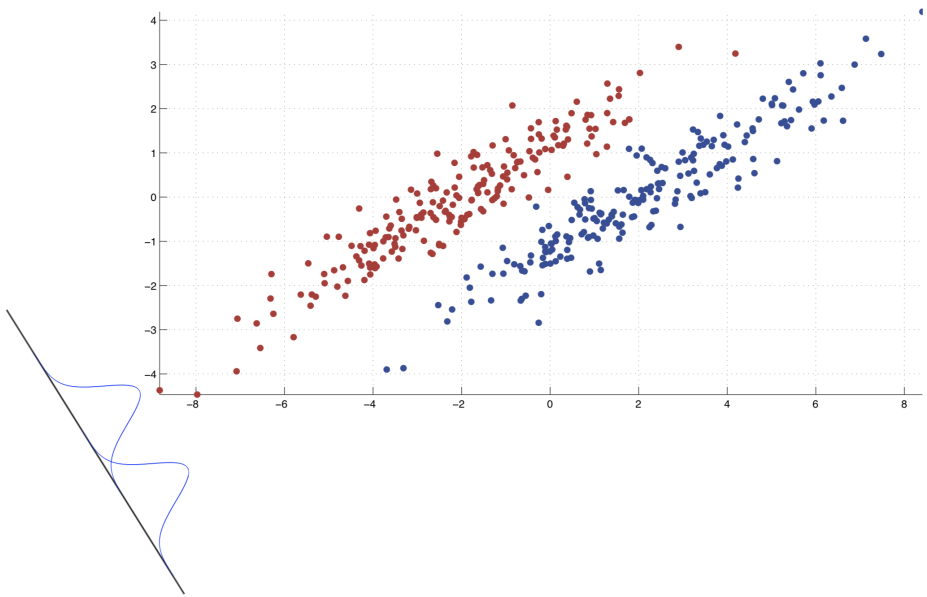
Recap: Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

  $\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}]$

  $\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}]$

$$\max_a \frac{(\beta_1 - \beta_2)^2}{\sigma_1^2 + \sigma_2^2}$$



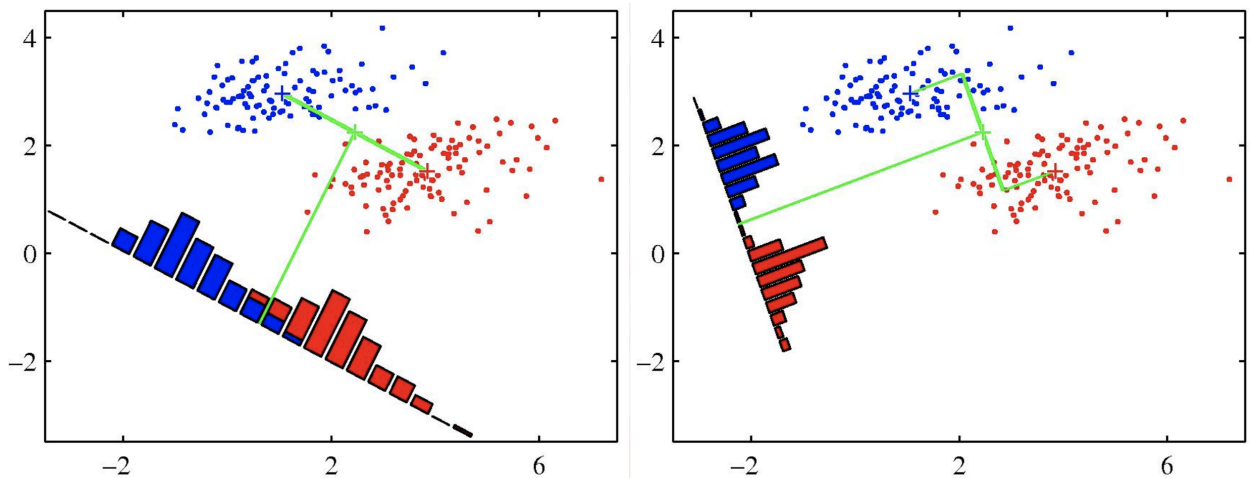
Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

  $\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}]$

  $\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}]$

$$\max_a \frac{(\beta_1 - \beta_2)^2}{\sigma_1^2 + \sigma_2^2}$$



Fisher's LDA

$$\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}] = a^T \mu_k$$
$$\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}] =$$
$$\mathbb{E}[(a^T(x - \mu_k))^2] = \mathbb{E}[(a^T(x - \mu_k)(x - \mu_k)^T a] = a^T \Sigma_k a$$

$$\max_a \frac{(\beta_1 - \beta_2)^2}{\sigma_1^2 + \sigma_2^2} = \max_a \frac{(a^T(\mu_1 - \mu_2))^2}{a^T(\Sigma_1 + \Sigma_2)a}$$
$$= \max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

Maximizing quadratic forms

$$\max_{a} \frac{a^T Q a}{a^T a}$$

Maximizing quadratic forms

$$\max_a \frac{a^T Q a}{a^T a}$$

- Eigenvalue Decomposition $Q = U \Lambda U^T$
- Change of basis $b = U^T a$, i.e., $Ub = a$

$$\max_a \frac{a^T U \Lambda U^T a}{a^T a} = \max_b \frac{b^T \Lambda b}{b^T U^T U b}$$
$$= \max_b \frac{b^T \Lambda b}{b^T b}$$

Maximizing quadratic forms

$$\max_a \frac{a^T Q a}{a^T a}$$

- Eigenvalue Decomposition $Q = U \Lambda U^T$
- Change of basis $b = U^T a$, i.e., $Ub = a$

$$\max_a \frac{a^T U \Lambda U^T a}{a^T a} = \max_b \frac{b^T \Lambda b}{b^T U^T U b}$$
$$= \max_b \frac{b^T \Lambda b}{b^T b}$$

- Optimum is given by $b = \delta[n - k^*]$ where

$$k^* = \arg\max_k \Lambda_{kk} = 1$$

Solution: $a = u_1$ maximal eigenvector, i.e., $Qu_1 = \lambda_1 u_1$

Optimal value : λ_1

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Theorem (Simultaneous Diagonalization)

  Let $P, Q \in \mathbb{R}^{n \times n}$ real symmetric matrices, and P is positive definite, then there exists a matrix V such that

  $$V^T P V = I$$
  $$V^T Q V = \Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$$

  where V, Λ satisfies the generalized eigenvalue equation:

  $$Q v_i = \lambda_i P v_i$$

Maximizing quadratic forms: two quadratics

- Theorem (Simultaneous Diagonalization)

  Let $P, Q \in \mathbb{R}^{n \times n}$ real symmetric matrices, and P is positive definite, then there exists a matrix V such that

$$V^T P V = I$$
$$V^T Q V = \Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$$

  where V, Λ satisfies thegeneralized eigenvalue equation:

$$Q v_i = \lambda_i P v_i$$

  Proof: Let $P = U_P \Lambda_P U_P^T$ be its Eigenvalue Decomposition

  $V' = U_P \Lambda_P^{-\frac{1}{2}}$ will only diagonalize P

  Let $V'^T Q V' = U' \Lambda' U'^T$ be its EVD

  Set $V = V' U'$ □

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Let V and Λ satisfy the generalized eigenvalue equation

$$Q v_i = \lambda_i P v_i$$

Basis change $a = Vb$, i.e., $b = V^T a$

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Let V and Λ satisfy the generalized eigenvalue equation

  $$Q v_i = \lambda_i P v_i$$

  Basis change $a = Vb$, i.e., $b = V^T a$

  $$\max_b \frac{b^T V^T Q V b}{b^T V^T P V b} = \max_b \frac{b^T \Lambda b}{b^T b}$$

- Solution: $a = v_1$, maximal generalized eigenvector

  Optimal value: λ_1 maximum generalized eigenvalue

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

$$P^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T a = \lambda a$$

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

$$P^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T a = \lambda a$$

$a = \text{constant} \times P^{-1}(\mu_1 - \mu_2)$

can be normalized as $a := \frac{P^{-1}(\mu_1 - \mu_2)}{||P^{-1}(\mu_1 - \mu_2)||_2}$

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

$$P^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T a = \lambda a$$

$a = \text{constant} \times P^{-1}(\mu_1 - \mu_2)$

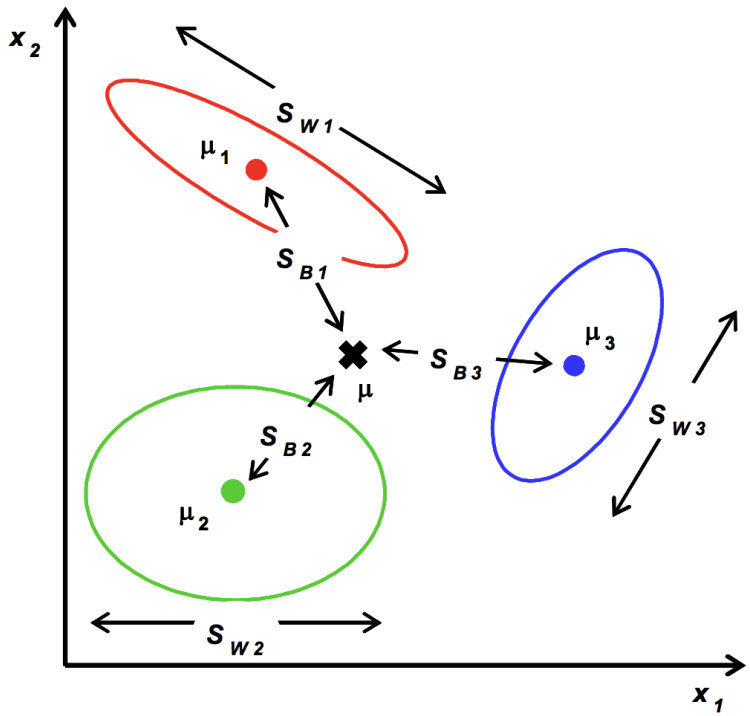
can be normalized as $a := \frac{P^{-1}(\mu_1 - \mu_2)}{||P^{-1}(\mu_1 - \mu_2)||_2}$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)

- Consider $y = A^T x$, where A^T is $m \times n$

  $\mu_k = \frac{1}{N_k} \sum_{j \in \text{class k}} x_j$ and $\mu = \frac{1}{N} \sum_{j=1}^{N} x_j$

  $S_k = \sum_{j \in \text{class k}} (x_j - \mu_k)(x_j - \mu_k)^T$

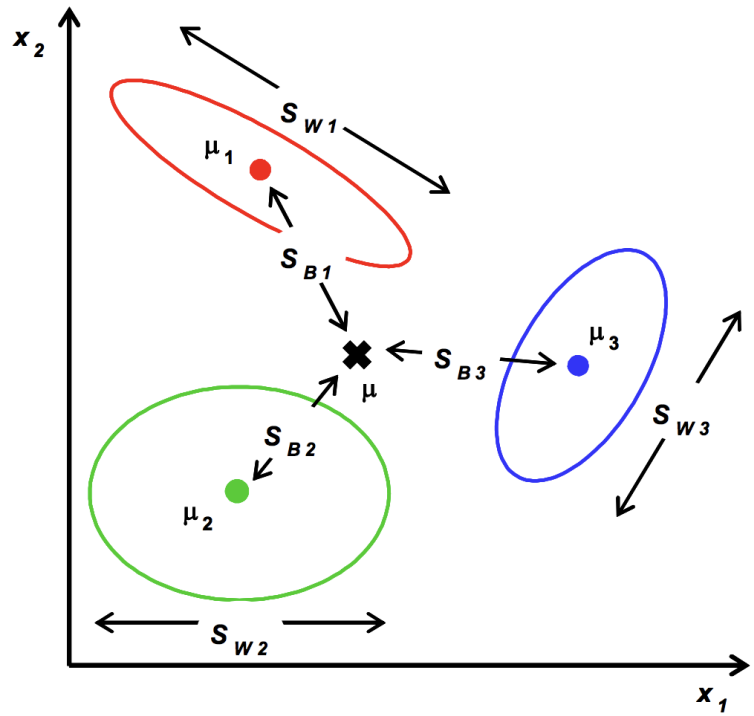


Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)

- Consider $y = A^T x$, where A^T is $m \times n$

  $\mu_k = \frac{1}{N_k} \sum_{j \in \text{class k}} x_j$ and $\mu = \frac{1}{N} \sum_{j=1}^{N} x_j$

  $S_k = \sum_{j \in \text{class k}} (x_j - \mu_k)(x_j - \mu_k)^T$



- within-class scatter $S_W = \sum_{k=1}^{K} S_k$
- between-class scatter $S_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)

- Consider the transformation $y = A^T x$, where A^T is $m \times n$
- Transform all samples $y_j = A^T x_j$, $\forall j$.
- Mean and scatter matrices in the transformed y domain:

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)

- Consider the transformation $y = A^T x$, where A^T is $m \times n$
- Transform all samples $y_j = A^T x_j$, $\forall j$.
- Mean and scatter matrices in the transformed y domain:
  $\tilde{\mu}_k = \frac{1}{N_k} \sum_{j \in \text{class k}} y_j = \frac{1}{N_k} \sum_{j \in \text{class k}} A^T y_j = A^T \mu_k$
  $\tilde{\mu} = \frac{1}{N} \sum_{j=1}^{N} y_j = \frac{1}{N} \sum_{j=1}^{N} A^T x_j = A^T \mu$

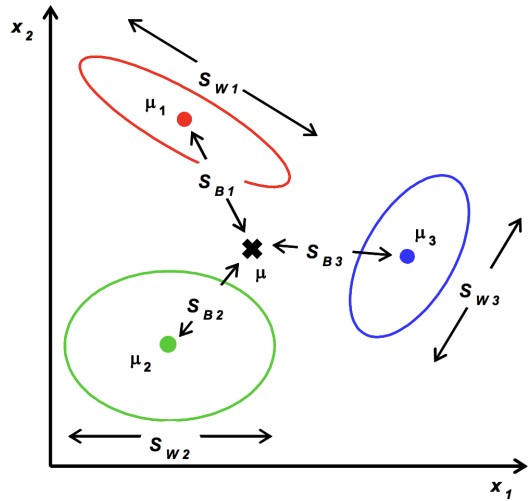
$$\tilde{S}_k = \sum_{j \in \text{class k}} (y_j - \tilde{\mu}_k)(y_j - \tilde{\mu}_k)^T$$

$$= \sum_{j \in \text{class k}} (A^T x_j - A^T \mu_k)(A^T y_j - A^T \tilde{\mu}_k)^T = A^T S_k A$$

within-class scatter $\tilde{S}_W = \sum_{k=1}^{K} \tilde{S}_k = A^T S_W A$

between-class scatter $\tilde{S}_B = A^T S_B A$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)



- Consider $y = A^T x$, where A^T is $m \times n$

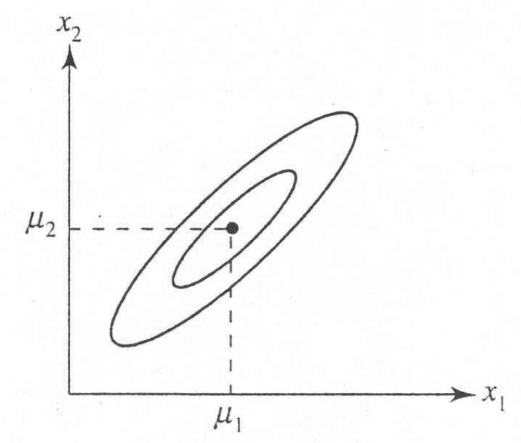
  within-class scatter $\tilde{S}_W = A^T S_W A$

  between-class scatter $\tilde{S}_B = A^T S_B A$
- What is the right objective function ?

Bivariate Gaussian

$$x = [x_1, ...x_n] \sim N(\mu, \Sigma),\ n = 2,\ \Sigma = \begin{bmatrix} \sigma_1^2 & \\ & \sigma_2^2 \end{bmatrix}$$

90% probability contour



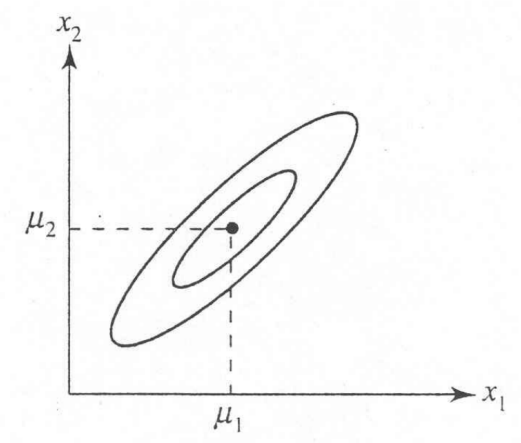
$$P(x_1, ...x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

whitening $x \sim N(\mu, \Sigma) \implies z = \Sigma^{-\frac{1}{2}}(x - \mu) \sim N(0, I)$

Bivariate Gaussian

$x = [x_1, ...x_n] \sim N(\mu, \Sigma)$, $n = 2$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \\ & \sigma_2^2 \end{bmatrix}$

90% probability contour



$$P(x_1, ...x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

whitening $x \sim N(\mu, \Sigma) \implies z = \Sigma^{-\frac{1}{2}}(x - \mu) \sim N(0, I)$

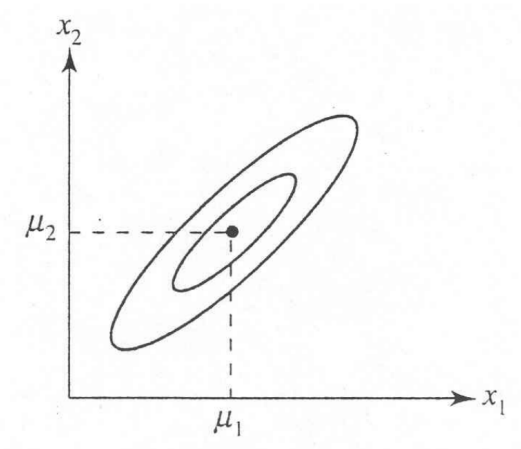
- area of the ellipse containing $1 - \alpha$ probability

  area$= \pi \chi_2^2(\alpha) \sigma_1 \sigma_2$

  chi-square upper percentile $\chi_2^2(0.1) \approx 4.61$, $\chi_2^2(0.01) \approx 9.21$

Multivariate Gaussian

$$x = [x_1, ...x_n] \sim N(\mu, \Sigma)$$



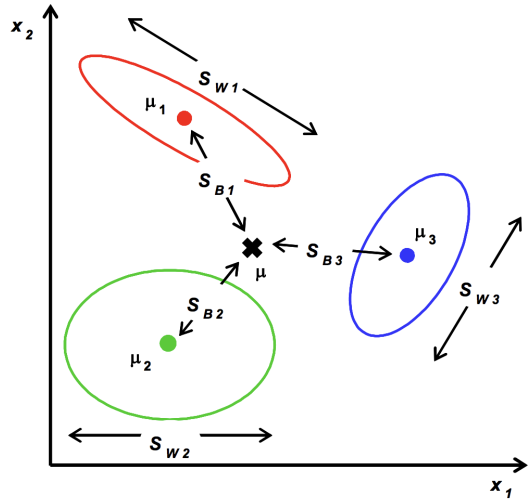
$$P(x_1, ...x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

The smallest region such that there is probability $1 - \alpha$ that a randomly selected observation will fall into is

an n-dimensional ellipsoid with volume

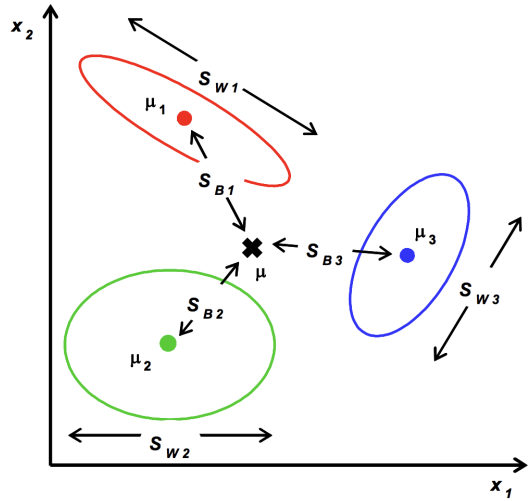
$$\frac{(2\pi)^{n/2}}{n\Gamma(n/2)} (\chi_p^2(\alpha)^{n/2}) |\Sigma|^{1/2}$$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)



- Consider $y = A^T x$, where A^T is $m \times n$
- Mean and scatter matrices in the transformed y domain:
  within-class scatter $\tilde{S}_W = A^T S_W A$
  between-class scatter $\tilde{S}_B = A^T S_B A$
- Objective function $J(A) = \frac{|A^T S_B A|}{|A^T S_W A|}$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)

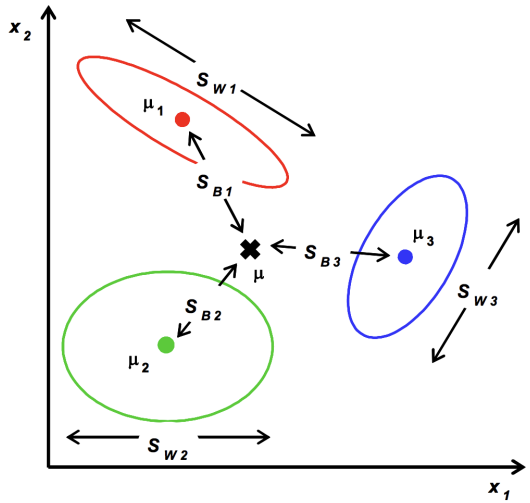


- Consider $y = A^T x$, where A^T is $m \times n$
- Mean and scatter matrices in the transformed y domain:
  within-class scatter $\tilde{S}_W = A^T S_W A$
  between-class scatter $\tilde{S}_B = A^T S_B A$
- Objective function $J(A) = \frac{|A^T S_B A|}{|A^T S_W A|}$
  Columns of the optimal A satisfy

$$S_B a_i = \lambda_i S_W a_i$$

  i.e., Eigenvectors of ${S_W}^{-1} S_B$

Multi-class Fisher LDA (K classes with $N_1, ..., N_K$ examples)



$\mu_k = \frac{1}{N_k} \sum_{j \in \text{class k}} x_j$ and $\mu = \frac{1}{N} \sum_{j=1}^{N} x_j$

$S_k = \sum_{j \in \text{class k}} (x_j - \mu_k)(x_j - \mu_k)^T$

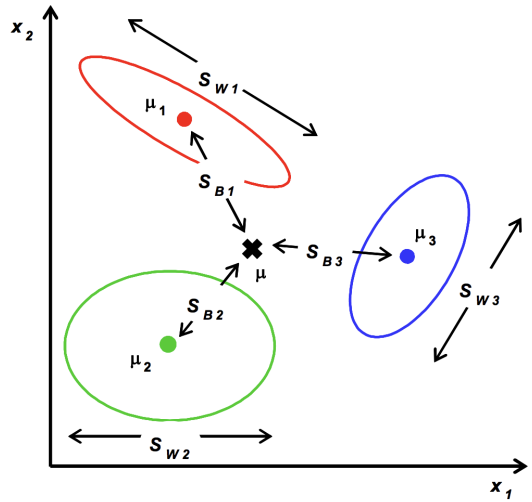
within-class scatter $S_W = \sum_{k=1}^{K} S_k$

between-class scatter $S_B = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T$

- Objective function $J(A) = \frac{|A^T S_B A|}{|A^T S_W A|}$

  Columns of the optimal A are the Eigenvectors of ${S_W}^{-1} S_B$

Another objective function

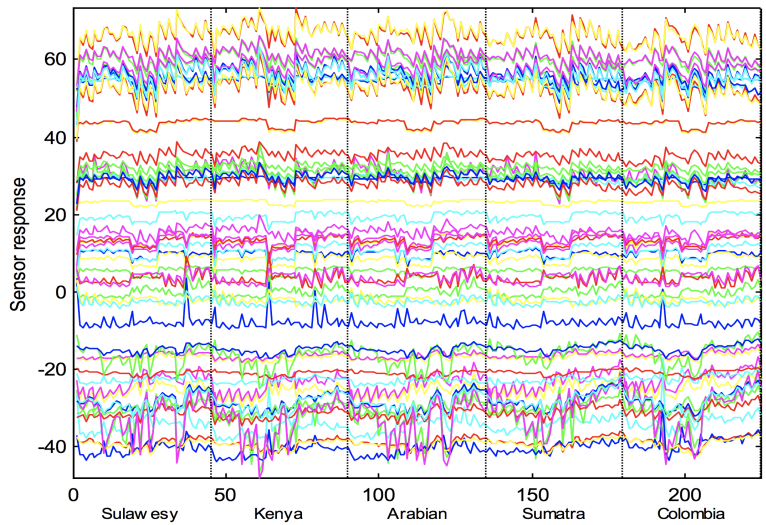


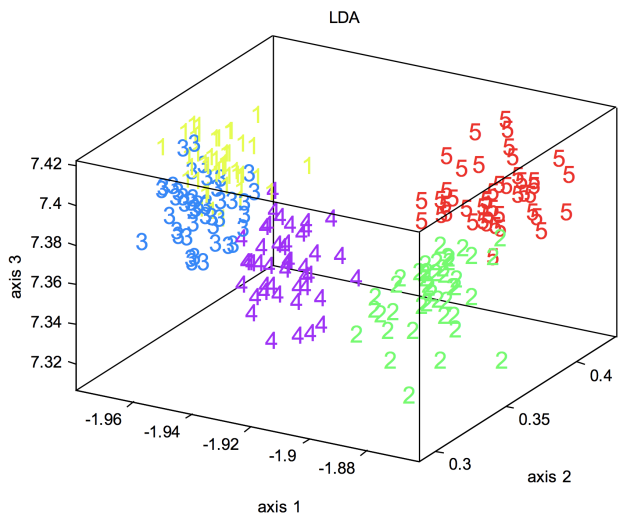
- Objective function $J(A) = \frac{|A^T S_B A|}{|A^T S_W A|}$
- Alternative objective $J(A) = \text{trace}[(A^T S_W A)^{-1}(A^T S_B A)]$
- Same solution

  Columns of the optimal A are the Eigenvectors of ${S_W}^{-1} S_B$

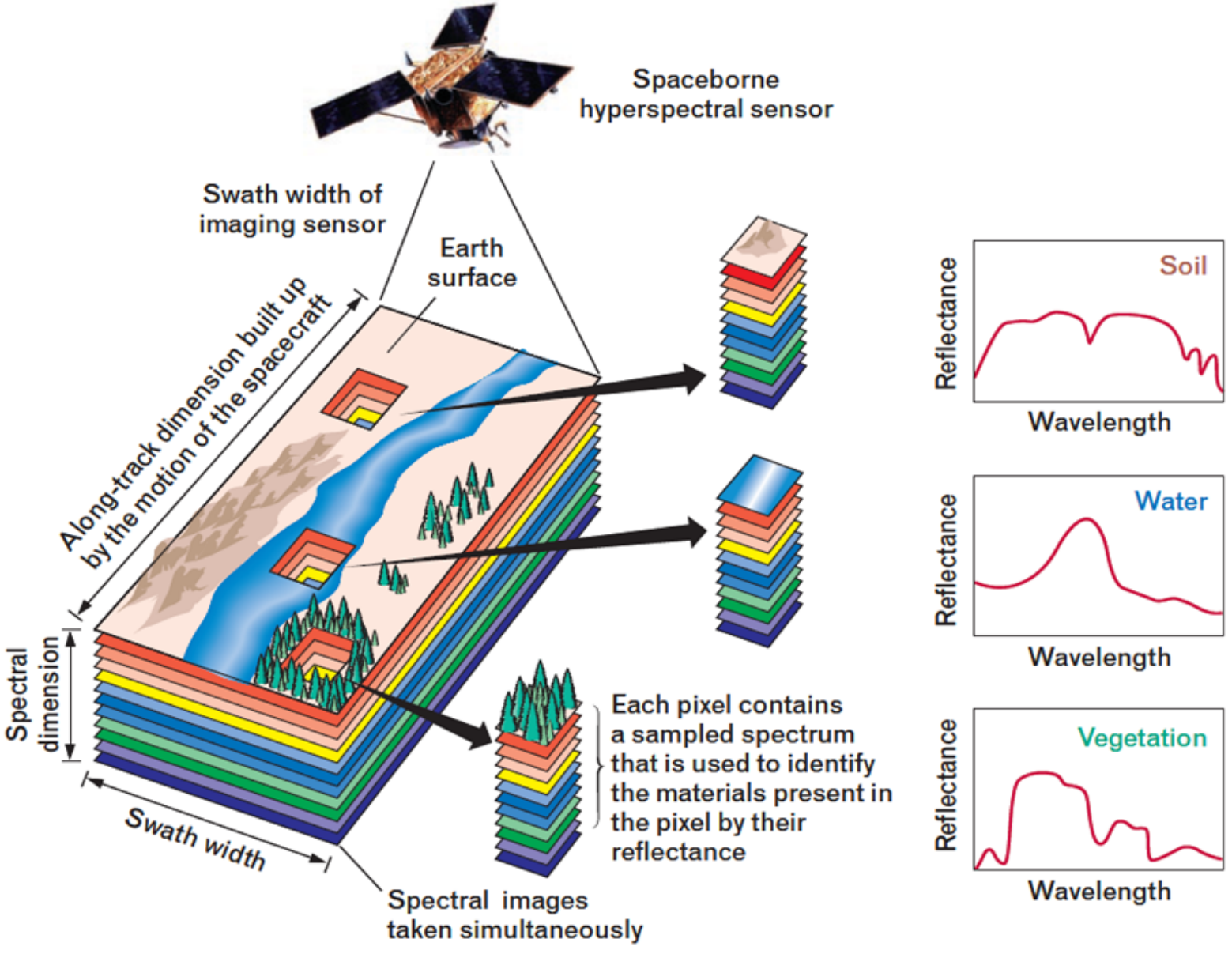
Coffee bean dataset

- 5 types of coffee beans were presented to an array of chemical gas sensors.





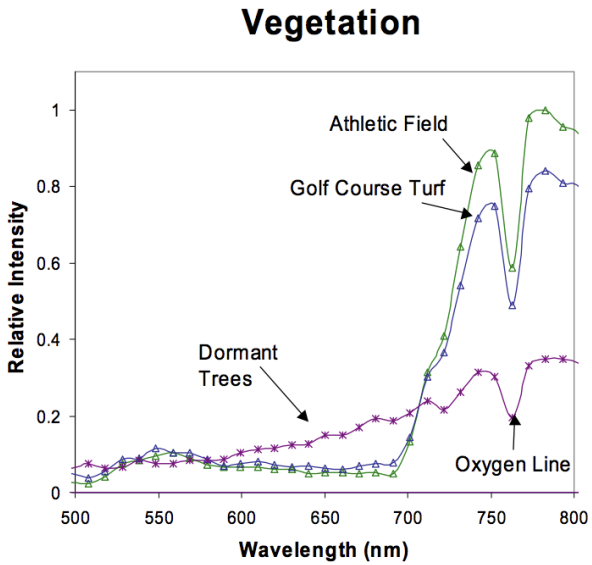
Application: Hyperspectral Imaging



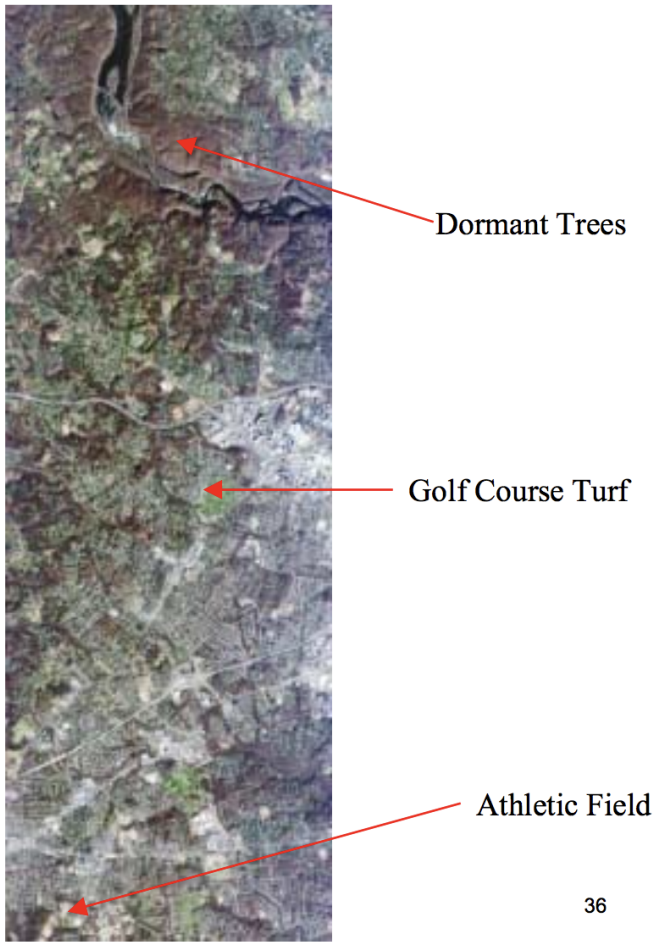
Application: Hyperspectral Imaging

- E0-1 Satellite (Hyperion imaging spectrometer) 200 bands

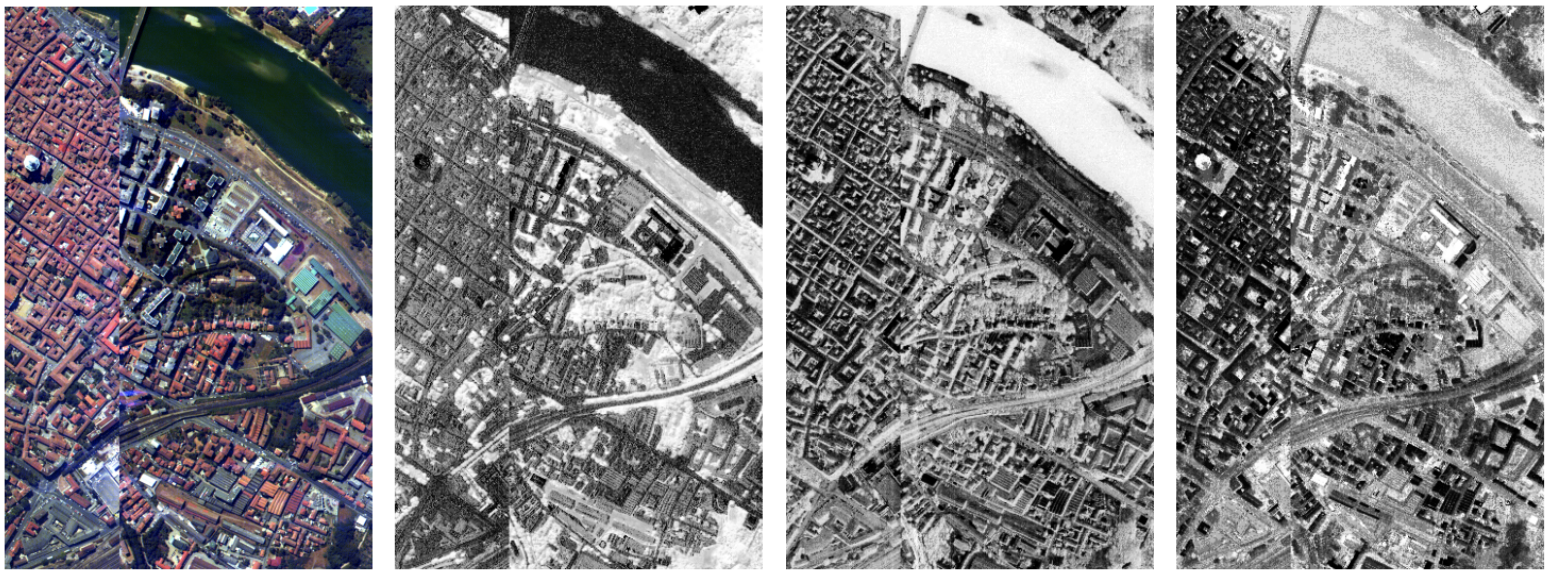
Image taken by Hyperion shows the relative chlorophyll content of vegetation in Fairfax County. The spectral profiles indicate healthy grass in the athletic field and golf course. The spectral profile of the trees indicates dormant vegetation.



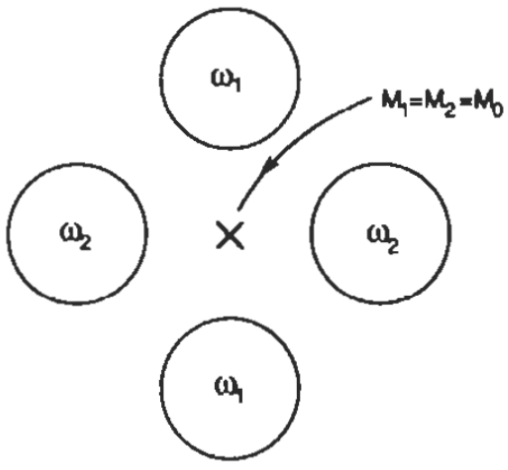
Oxygen in the atmosphere is detected by the spectral profiles in the near infrared wavelength.

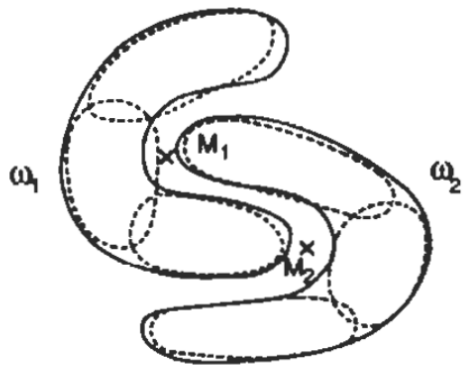


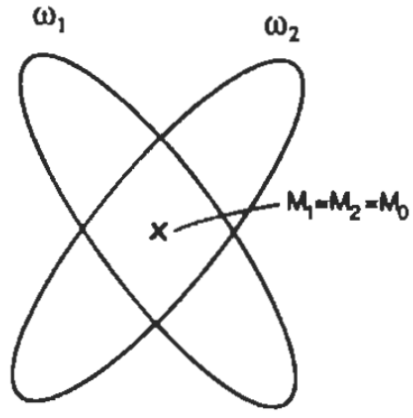
Fisher LDA for Hyperspectral Imaging



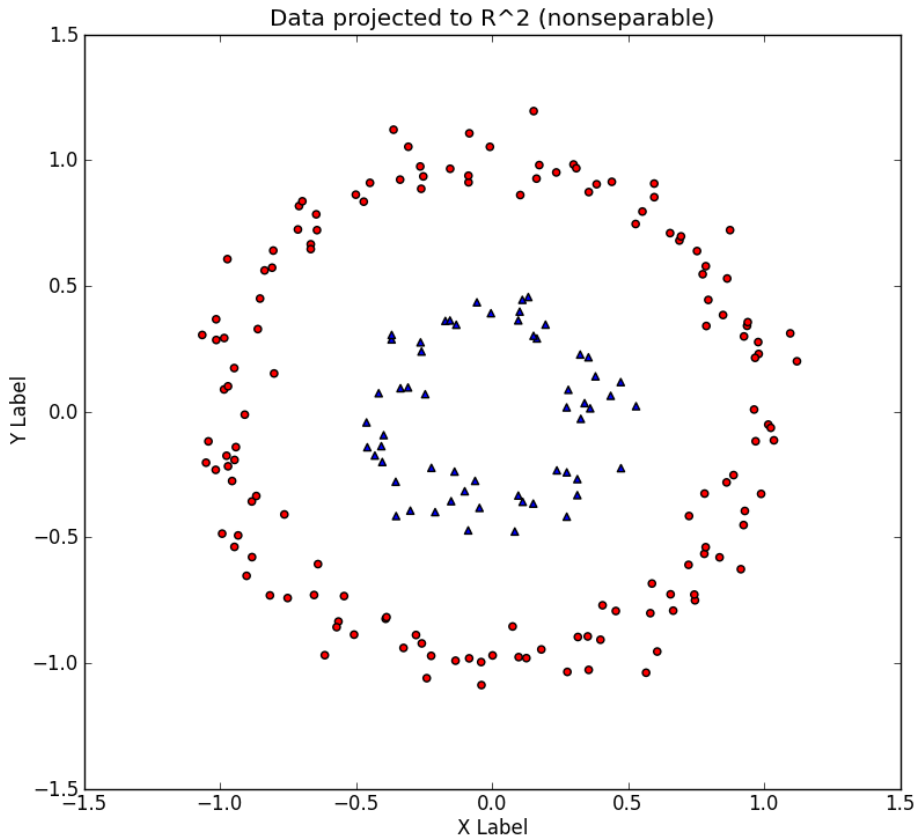
Problems with linear discrimination



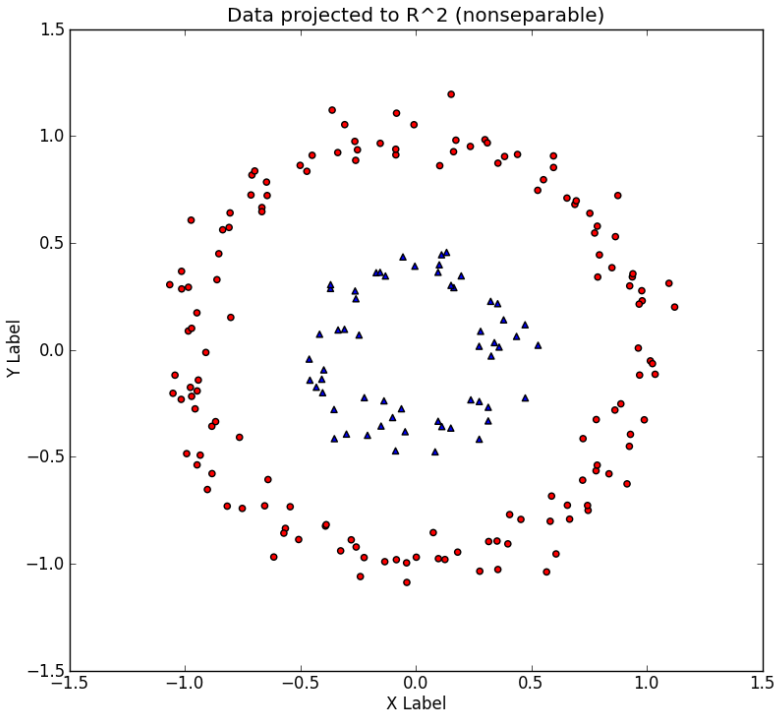


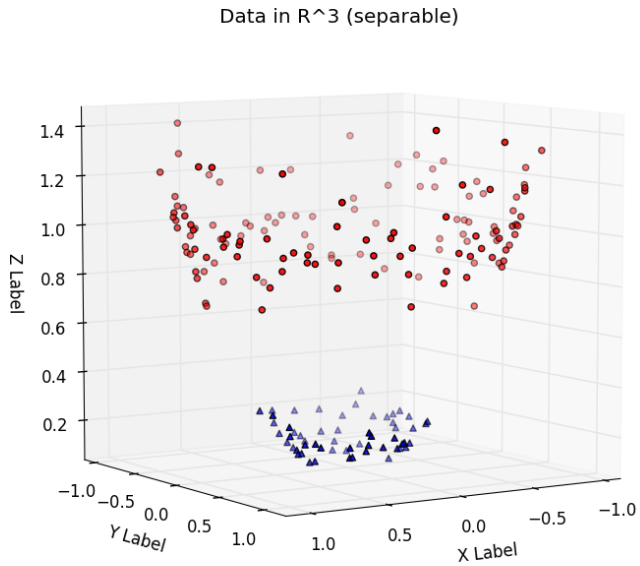


Problems with linear discrimination



Problems with linear discrimination





Stationary Processes and Quadratic Discriminants

- $h_k(x) = x^T W_k x + w_k^T x + w_{k0}$

  Classify as class k if $h_k(x) > h_{k'}(x) \quad \forall k' \neq k$

  $W_k = -\frac{1}{2}\Sigma_k^{-1}$

  $w_k = \Sigma_k^{-1}\mu_k$

  $w_{k0} = -\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$

  For stationary processes Σ_{kl} only depends on $(k-l)$
- Estimate $r[k] = \sum_n (x[n]-\mu)(x[n+k]-\mu)$

Stationary Processes and Quadratic Discriminants

- $h_k(x) = x^T W_k x + w_k^T x + w_{k0}$

  Classify as class k if $h_k(x) > h_{k'}(x) \quad \forall k' \neq k$

  $W_k = -\frac{1}{2}\Sigma_k^{-1}$

  $w_k = \Sigma_k^{-1}\mu_k$

  $w_{k0} = -\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$

  For stationary processes Σ_{kl} only depends on $(k-l)$
- Estimate $r[k] = \sum_n (x[n]-\mu)(x[n+k]-\mu)$

  Fourier transformation $y = Fx$ turns every Σ_k into diagonal
- i.e., $F\Sigma F^H =$ Diagonal
- Estimate only the diagonals

How to check if linear/quadratic discrimination works

- Plot $z_1 = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$ and $z_2 = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$ in the (z_1, z_2) space

How to check if linear/quadratic discrimination works

- Plot $z_1 = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$ and $z_2 = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2)$ in the (z_1, z_2) space
- 40 dimensional radar signal, two classes

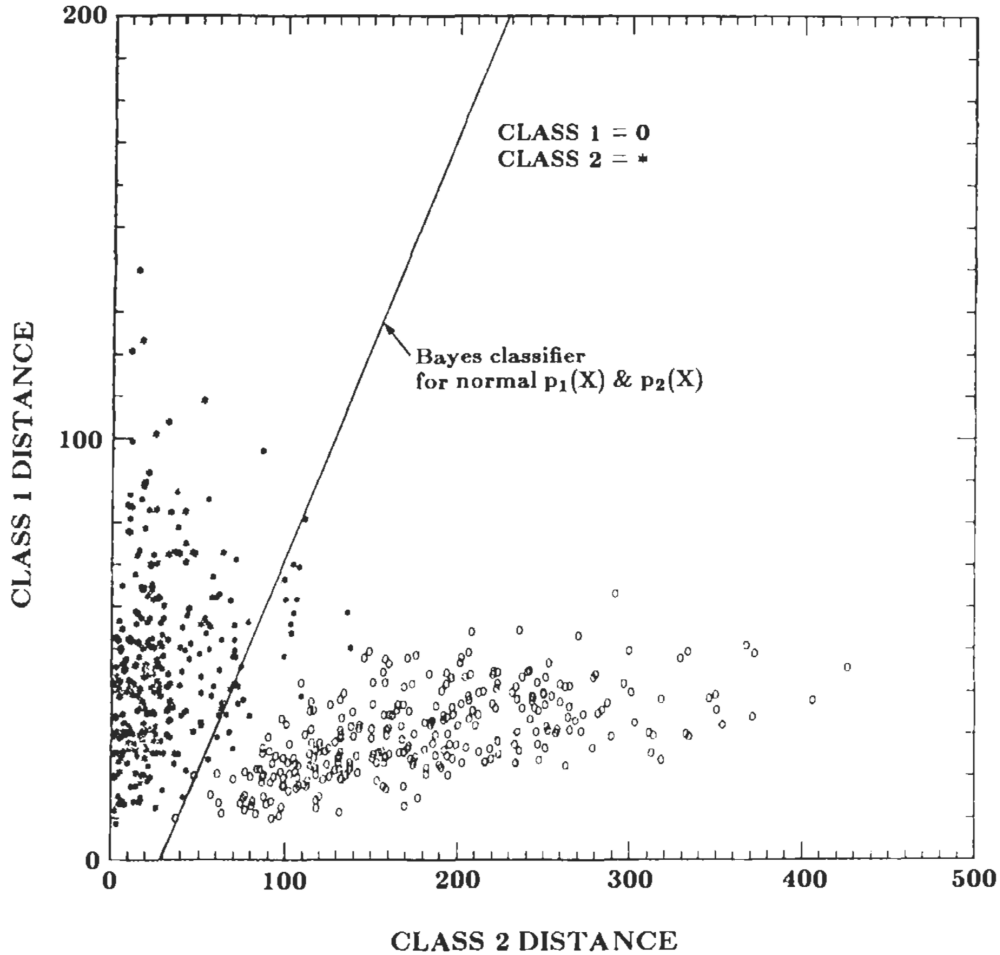


Fig. 4-9 d^2-display of a radar data.

(slide credit: K. Fukunaga)

Separating Hyperplanes

- Linear/Quadratic Discriminant Analysis, Bayes Optimal Classifiers

  require modeling signals: $p(x)$, and class distributions $p(y|x)$
- **Vapnik's Principle**

  'When solving a problem if interest, do not solve a more general problem as an intermediate step'

Separating Hyperplanes

- Directly estimate hyperplanes $w^T x + b \geq 0$
  parameters $\theta = (w, b)$

Separating Hyperplanes

- Directly estimate hyperplanes $w^T x + b \geq 0$ parameters $\theta = (w, b)$
- Hyperplane: $H = \{x : \ w^T x + b = 0\}$

Separating Hyperplanes

- Directly estimate hyperplanes $w^T x + b \geq 0$

  parameters $\theta = (w, b)$
- Hyperplane: $H = \{x : \ w^T x + b = 0\}$

  distance between a point z and H

  $d(z, H) = \min_{h \in H} ||z - h||_2$

  Decompose $z = z_0 + \frac{w}{||w||_2} r$

  $w^T z + b = w^T z_0 + b + ||w||_2 r$

Separating Hyperplanes

- Directly estimate hyperplanes $w^T x + b \geq 0$

  parameters $\theta = (w, b)$
- Hyperplane: $H = \{x : \ w^T x + b = 0\}$

  distance between a point z and H

  $d(z, H) = \min_{h \in H} ||z - h||_2$

  Decompose $z = z_0 + \frac{w}{||w||_2} r$

  $w^T z + b = w^T z_0 + b + ||w||_2 r$

$$d(z, H) = |r| = \frac{|w^T z + b|}{||w||_2}$$

Margin

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

- Directly estimate hyperplanes $w^T x + b = 0$
- Margin ρ of a hyperplane is

$$\begin{aligned} \rho(w, b) &= \min_{i=1,...,n} d(x_i, H) \\ &= \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2} \end{aligned}$$

Margin

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

- Directly estimate hyperplanes $w^T x + b = 0$
- Margin ρ of a hyperplane is

$$\begin{aligned}\rho(w, b) &= \min_{i=1,...,n} d(x_i, H) \\ &= \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}\end{aligned}$$

- Maximum margin separating hyperplane is the solution of

$$\begin{aligned}&\max_{w,b} \rho(w, b) \\ &s.t. \;\; y_i(w^T x_i + b) \geq 0 \; \forall i\end{aligned}$$

Maximum margin hyperplane

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

- Margin $\rho(w, b) = \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$
- Maximum margin separating hyperplane is the solution of

$$\max_{w,b} \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$$
$$s.t. \;\; y_i(w^T x_i + b) \geq 0 \; \forall i$$

Maximum margin hyperplane

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

- Margin $\rho(w, b) = \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$
- Maximum margin separating hyperplane is the solution of

$$\max_{w,b} \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$$
$$s.t. \;\; y_i(w^T x_i + b) \geq 0 \; \forall i$$

- **not unique**

  $(\alpha w, \alpha b)$ gives the same hyperplane

Maximum margin hyperplane

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

- Margin $\rho(w, b) = \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$
- Maximum margin separating hyperplane is the solution of

$$\max_{w,b} \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||_2}$$
$$s.t. \;\; y_i(w^T x_i + b) \geq 0 \; \forall i$$

- **not unique**

  $(\alpha w, \alpha b)$ gives the same hyperplane
- Scale w and b by $\frac{1}{\min_{i=1,...,n} |w^T x_i + b|}$

  Now $\rho = \frac{1}{||w||_2}$

Maximum margin hyperplane

Data $x_1, ..., x_n$ and corresponding labels $y_1, ...y_n \in \{-1, +1\}$

$$\max_{w,b} \frac{1}{||w||_2}$$
$$s.t. \;\; y_i(w^T x_i + b) \geq 1 \; \forall i$$

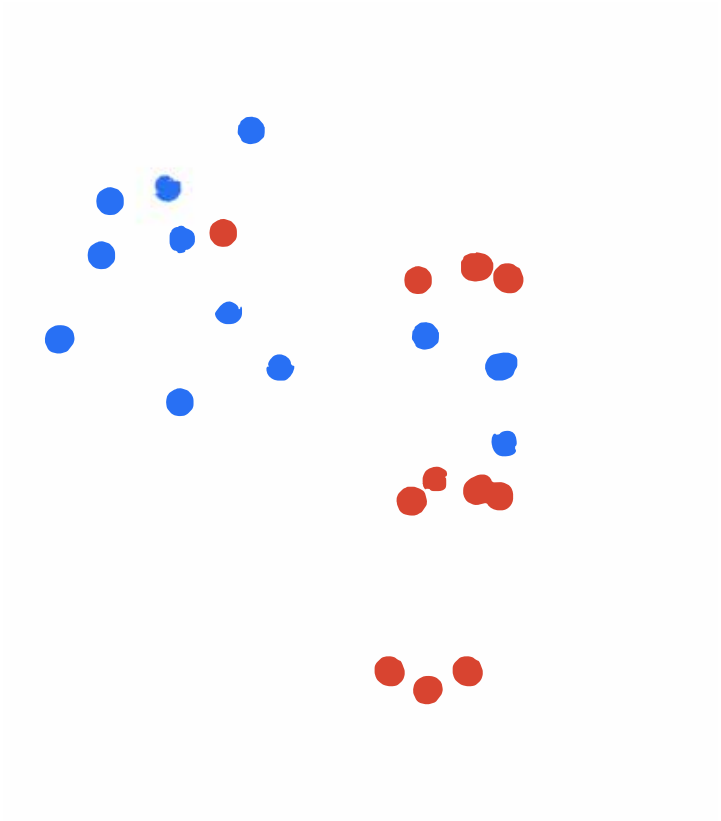
equivalently

$$\min_{w,b} ||w||_2^2$$
$$s.t. \;\; y_i(w^T x_i + b) \geq 1 \; \forall i$$

- Hard-margin support vector machine (SVM)

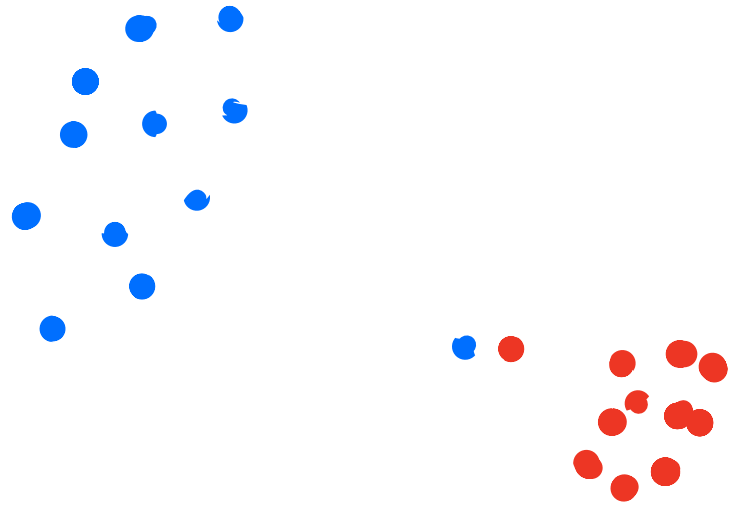
Problems with hard margin

- Separability



Problems with hard margin

- Sensitivity



Soft Margin Support Vector Machine

$$\min_{w,b,s_1,...,s_n} \frac{1}{2}||w||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - s_i \ \forall i$$
$$s_i \geq 0 \ \forall i$$

- $s_1, ..., s_n$ are slack variables
- C is a tuning parameter