Linear methods for classification: Reduced Rank LDA

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

February 5, 2020

LDA and PCA

- no natural tuning parameter for LDA
- can use PCA dimension reduction on inputs
- number of PCs used is tuning parameter
- no information about the outcome used in PCA
- reduced rank LDA similar, but uses outcome information
- prerequisite: SVD and Eigen decomposition

Linear algebra review

- see LA_Examples link on wiki
- "diagonal" matrix only diagonal elements are non-zero
- easy to invert

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{bmatrix}$$

$$D^{-1} = \begin{bmatrix} \frac{1}{d_1} & 0 & 0 & 0 \\ 0 & \frac{1}{d_2} & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{d_p} \end{bmatrix}$$

Linear algebra review

- "orthogonal" matrix columns have correlation zero
- also called "linearly independent"
- easy to invert; transpose is inverse
- if V is an orthogonal matrix

$$V^{-1} = V^T$$

$$V^T V = I$$

- I is the "identity" matrix

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Singular Value Decomposition

- say X is $n \times p$ matrix
- SVD is $X = UDV^T$
- U - $n \times p$ - orthogonal - "left singular vectors"
- D - $p \times p$ - diagonal - $d_1 \geq \cdots \geq d_p$ "singular values"
- V - $p \times p$ - orthogonal - "right singular vectors"
- SVD exists for all matrices
- if any $d_j = 0$, X is "singular"; cols of X are linearly dependent
- `svd()` function in R will compute U, D and V

Eigen decomposition of $X^T X$

$$
\begin{aligned}
X^T X &= (UDV^T)^T UDV^T \\
&= VDU^T UDV^T \\
&= VD^2V^T
\end{aligned}
$$

- columns of V are eigenvectors (also right singular vectors)
- diagonal elements of D^2 are eigenvalues of $X^T X$

$X^T X$ is proportional to $\mathrm{cov}(X)$

- if columns of X are centered (mean zero), then

$$\mathrm{cov(X)} = \Sigma = \frac{1}{n} X^T X$$

$$\begin{aligned} \Sigma &= \frac{1}{n} X^T X \\ &= \frac{1}{n} V D^2 V^T \end{aligned}$$

- can do PCA with X or Σ, get the same V and PCs

Principal components from SVD or Eigen

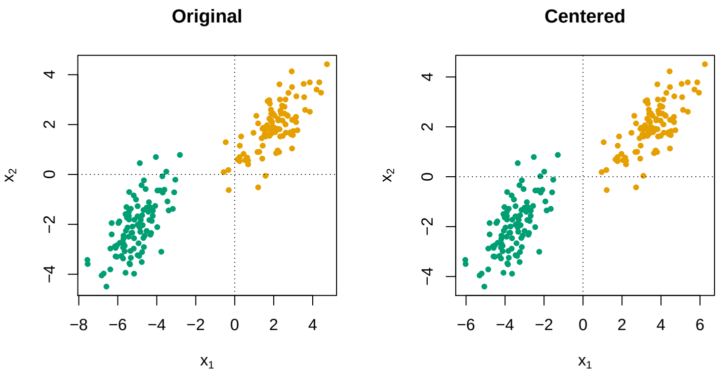
- the principal components of a matrix X are simply

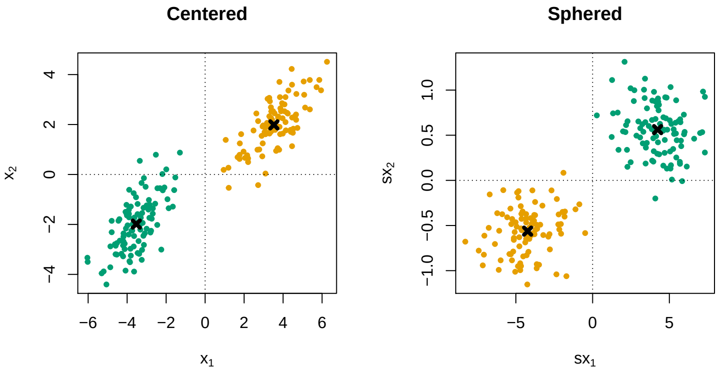
$$Z = XV$$

- eigenvectors (cols of V) are "principal component directions"
- diagonal elements of D^2 are eigenvalues of $X^T X$
- eigen values are related to variance of PCs

Sphereing

- consider $\mathrm{cov}(x) = \Sigma = VDV^T$
- sphered inputs are $x^* = x\Sigma^{-1/2} = xVD^{-1/2}$
- $\mathrm{cov}(x^*) = I_p$


Original
Centered
x_1
x_2
−8
−6
−4
−2
0
2
4
6

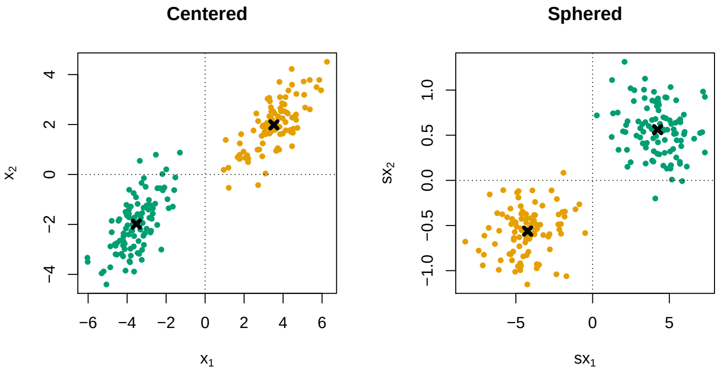

Centered
x1
x2
−6 −4 −2 0 2 4 6
−4 −2 0 2 4
Sphered
sx1
sx2
−5 0 5
−1.0 −0.5 0.0 0.5 1.0

LDA and Eigen-decomposition of Σ

- $\Sigma^{-1/2} = VD^{-1/2}$
- $\Sigma^{-1/2}(\Sigma^{-1/2})^T = \Sigma^{-1}$

$$\begin{aligned}
\delta_k(x) &= \log \pi_k - \frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)^T \\
&= \log \pi_k - \frac{1}{2}(x-\mu_k)\Sigma^{-1/2}(\Sigma^{-1/2})^T(x-\mu_k)^T \\
&= \log \pi_k - \frac{1}{2}[(x-\mu_k)\Sigma^{-1/2}][(x-\mu_k)\Sigma^{-1/2}]^T \\
&= \log \pi_k - \frac{1}{2}[x^* - \mu_k^*][x^* - \mu_k^*]^T
\end{aligned}$$

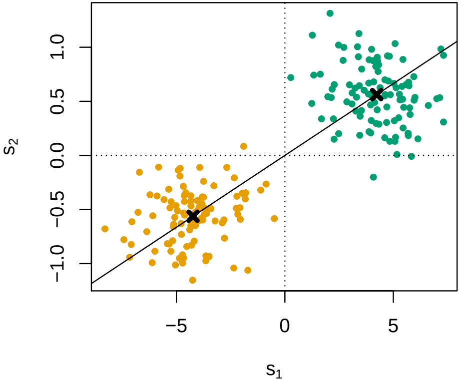
- x^* are "sphered" inputs
- μ^* are "sphered" centers
- why sphere inputs and centers?
- only distances from sphered centers are important
- for new x_0^*, classify to class with nearest μ_k^*


Centered
Sphered
x1
x2
sx1
sx2

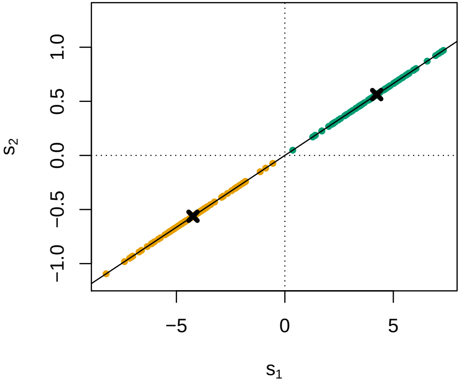
LDA as a reduced dimension classifier

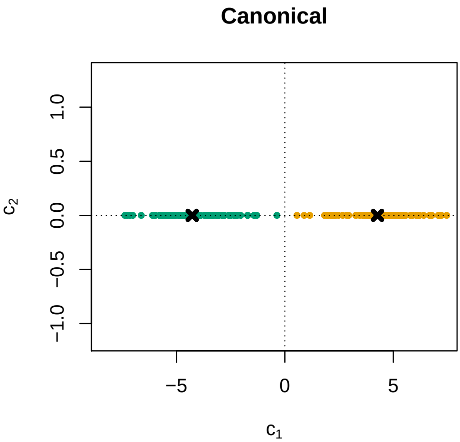
- consider a 2-classes ($K = 2$) and 2-dim input ($p = 2$)
- 2 sphered centers spanned by a 1-dim plane (i.e., a line)
- distances orthogonal to this line do not affect classification
- might as well project input onto line without loss
- projected variables are called "canonical" or "discriminant"
- original $\rightarrow$ sphered $\rightarrow$ canonical/discriminant
- when $K << p$, substantial dimension reduction of input

Sphered



Sphered




Canonical
1.0
0.5
0.0
−0.5
−1.0
c_2
−5
0
5
c_1

Code example

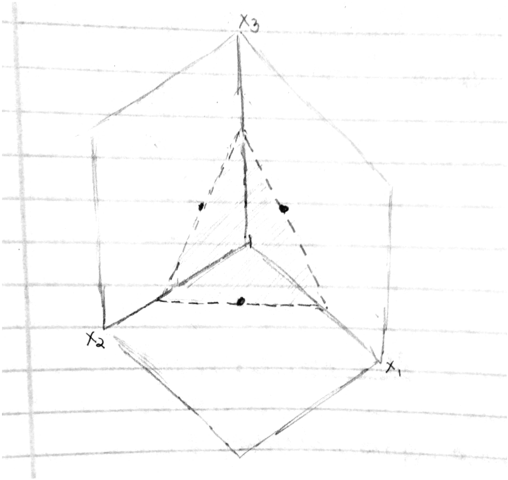
`sphered-and-canonical-inputs.R`

LDA as a reduced dimension classifier

- consider a 3-classes ($K = 3$) and 3-dim input ($p = 3$)
- 3 sphered centers spanned by a 2-dim plane
- can project onto plane without loss
- can we project onto lower dimension (i.e., reduce the rank) without much loss of discrimination?
- degree of dimension reduction is tuning parameter in reduced rank LDA

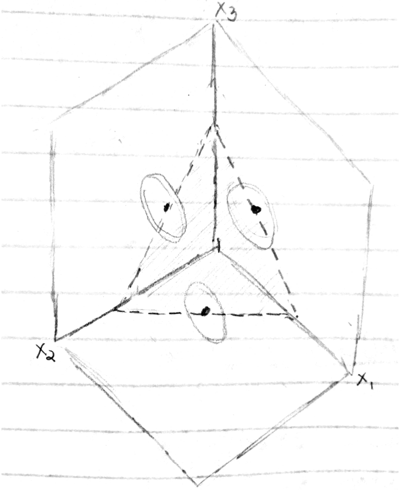
LDA as a reduced dimension classifier

3-class problem ($K = 3$) and 3-dimensional sphered input ($p = 3$)



LDA as a reduced dimension classifier

3-class problem ($K = 3$) and 3-dimensional sphered input ($p = 3$)



error: sphered data contours should not indicate correlation

How to compute reduced rank LDA

- do PCA on sphered centers $\mu^* = [\mu_1^*, \dots, \mu_K^*]^T$
- let $B = \mathrm{cov}(\mu^*)$
- compute $B = U_B D_B V_B^T$
- let $1 \leq l \leq K - 1$ and V_B^l the first l columns of V_B
- compute canonical variables and centers
- $x^l = x^* V_B^l$
- $\mu^l = \mu^* V_B^l$

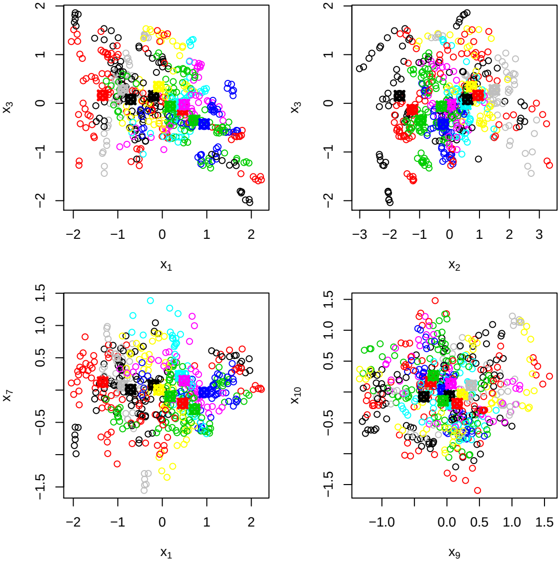
How to compute reduced rank LDA

- $x^l = x^* V_B^l$
- $\mu^l = \mu^* V_B^l$
- use canonical variable and centers in discriminant
- $\delta_k(x) = \log \pi_k - \frac{1}{2}[x^l - \mu_k^l]^T [x^l - \mu_k^l]$
- to classify x, compute $x^l = x\Sigma^{-1/2} V_B^l$ then find closest μ_k^l
- number of canonical variables l is tuning parameter
- $l = K - 1$ is same as regular LDA
- $l < K - 1$ makes model less flexible
- select l by minimizing estimate of EPE

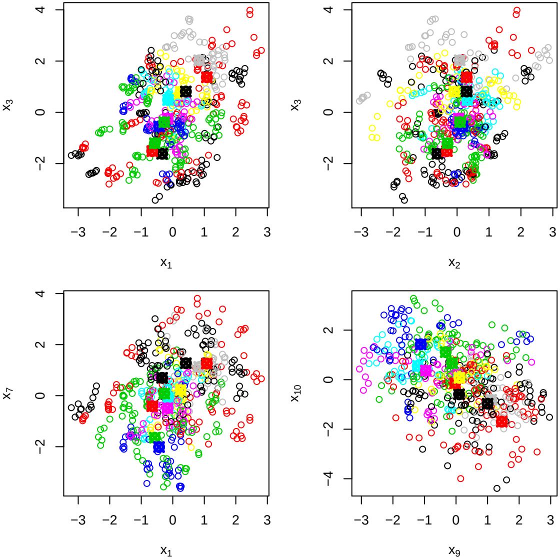
Vowel data

- well-known data for testing classifiers
- $K = 11$ classes (vowels)
- $p = 10$ inputs
- 0.40 is best attained EPE (using zero-one loss)

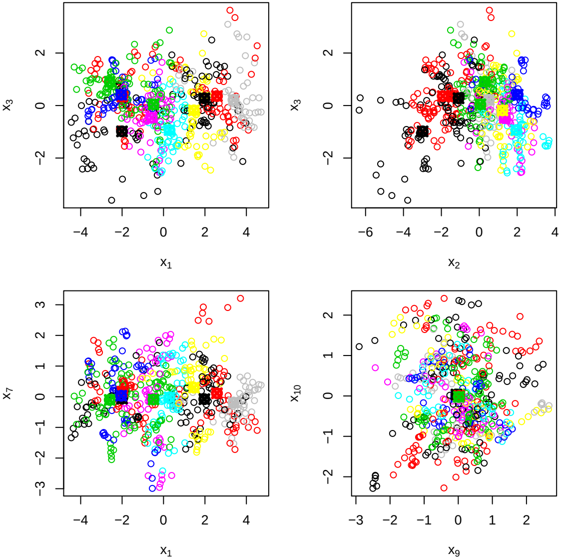
Original vowel data



Sphered vowel data



Canonical vowel data



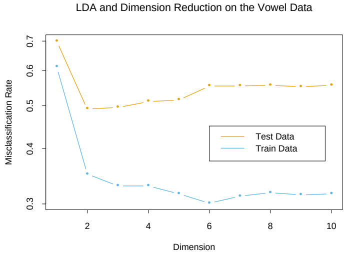


FIGURE 4.10. *Training and test error rates for the vowel data, as a function of the dimension of the discriminant subspace. In this case the best error rate is for dimension 2. Figure 4.11 shows the decision boundaries in this space.*

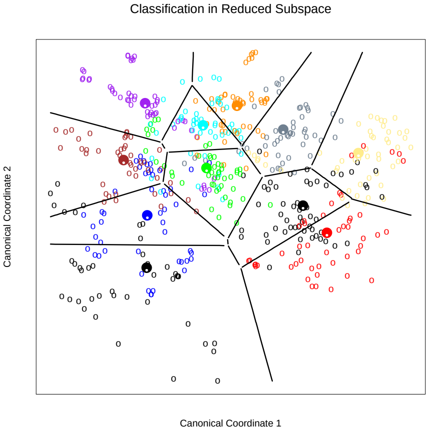


FIGURE 4.11. *Decision boundaries for the vowel training data, in the two-dimensional subspace spanned by the first two canonical variates. Note that in any higher-dimensional subspace, the decision boundaries are higher-dimensional affine planes, and could not be represented as lines.*