Feature Generation: LDA and PCA

Theodoridis Chs. 5.8, 6.1-6.3 (see also DHS, Ch 3.8)

Feature Generation

Purpose:

Given a training set, transform existing features to a smaller set that maintains as much classification-related information as possible

- i.e. 'Pack' information into a smaller feature space, removing redundant feature information

Linear Discriminant Analysis (LDA)

Goal

Find a line in feature space on which to project all samples, such that the samples are well (maximally) separated

$$y = \frac{\mathbf{w}^T \mathbf{x}}{||\mathbf{w}||} \qquad \tilde{\mu}_i = \mathbf{w}^T \mu_i$$

Projection

w is a unit vector (with length one): points projected onto line in direction of w

- Magnitude of w is not important (scales y)

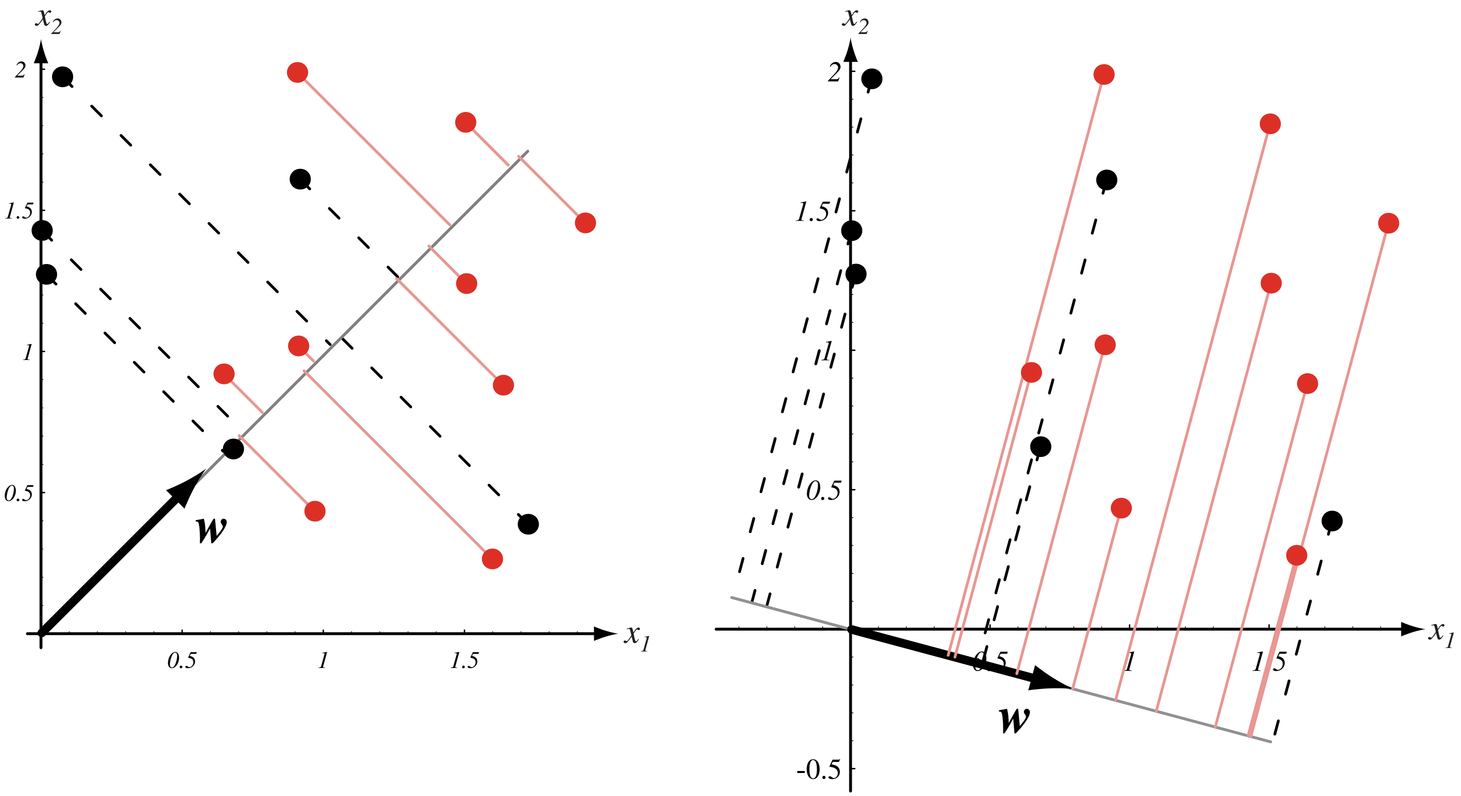


FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Criterion: FDR

Criterion

We use Fisher's Discriminant Ratio to evaluate how well a particular projection separates classes on the projection line

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

FDR for LDA

Recall: $\tilde{\mu}_i = \mathbf{w}^T \mu_i \qquad FDR = \dfrac{\boxed{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}}{\boxed{\tilde{\sigma}_1{}^2 + \tilde{\sigma}_2{}^2}}$

$$\boxed{(\tilde{\mu}_1 - \tilde{\mu}_2)^2} = \mathbf{w}^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \mathbf{w} \propto \boxed{\mathbf{w}^T S_b \mathbf{w}}$$

Between class scatter

$$\tilde{\sigma}_i^2 = E[(y - \tilde{\mu}_i)^2] = E[\mathbf{w}^T(x - \mu)(x - \mu_i)^T \mathbf{w})] = \mathbf{w^T \Sigma_i w}$$

Covariance matrix

$$\boxed{\tilde{\sigma}_1{}^2 + \tilde{\sigma}_2{}^2} \propto \boxed{\mathbf{w}^T S_w \mathbf{w}}$$

Within class scatter

Modified Criterion for LDA:
(Raleigh Quotient)

$$FDR(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

Finding the Optimal Projection Direction *w*

Our Goal: Find w maximizing FDR(w)

- achieved if w chosen such that:

$$S_b \mathbf{w} = \lambda S_w \mathbf{w}$$

- where lambda is the largest eigenvalue of $S_w^{-1} S_b$
- For two classes, to get the direction of w, use:

$$\mathbf{w} = S_w^{-1}(\mu_1 - \mu_2)$$

- This is the optimal reduction of m features to one for class separation

$$\boxed{FDR(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}}$$

A Classifier for ‘Free’

Linear classifier also defined by LDA:

$$g(x) = (\mu_1 - \mu_2)^T S_w^{-1} x + w_0$$

(class 1 if >= 0, class two if < 0)

w0 not defined directly by LDA; for Gaussians with identical covariances optimal classifier is:

$$g(x) = (\mu_1 - \mu_2)^T S_w^{-1} \left(x - \frac{1}{2}(\mu_1 + \mu_2) \right) - \ln \frac{P(\omega_2)}{P(\omega_1}$$

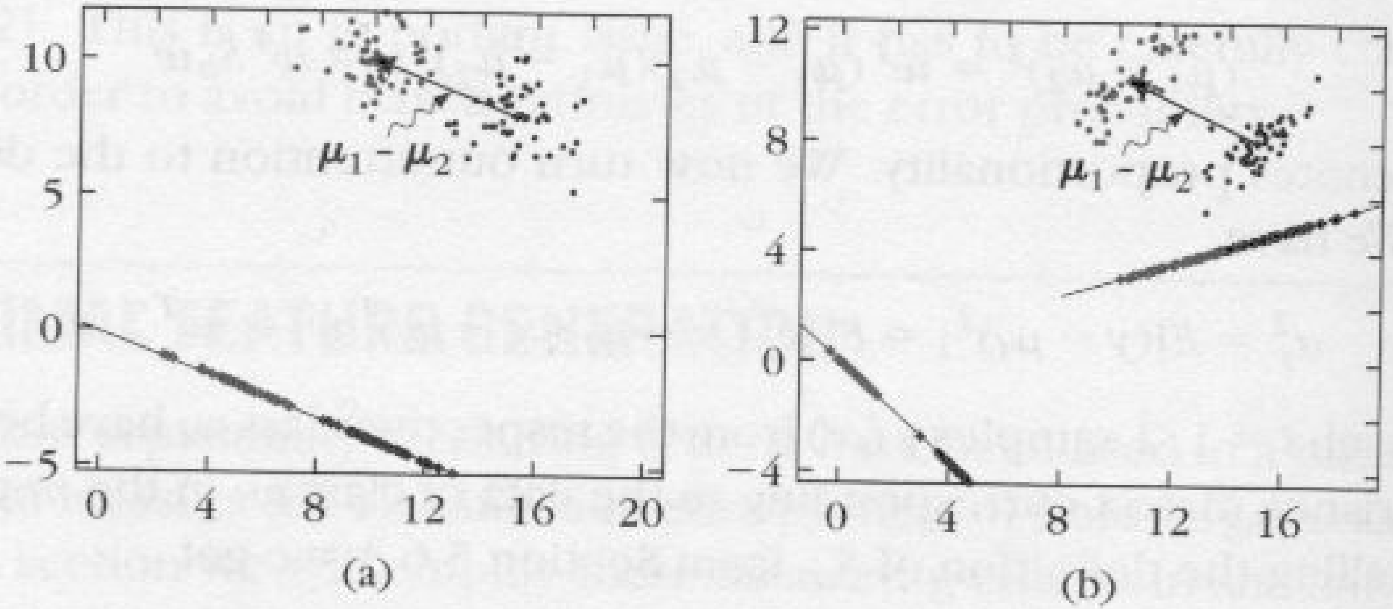


FIGURE 5.6

(a) The optimal line resulting from Fisher's criterion, for two Gaussian classes. Both classes share the same diagonal covariance matrix, with equal elements on the diagonal. The line is parallel to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. (b) The covariance matrix for both classes is nondiagonal. The optimal line is on the left. Observe that it is no more parallel to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The line on the right is not optimal and the classes, after the projection, overlap.

LDA, Cont'd

If original distributions multimodal and overlapping:

Classes for samples will overlap in the projection (little use)

Generalization for multiple classes is discussed further in the Theodoridis text.

Karhunen-Loève Transform (Principal Components Analysis - PCA)

Key Idea:

Model points in feature space by their deviation from the global mean in the *primary directions of variation in feature space*

- Defines a new, smaller feature space, often with more discriminating information

Directions of variation are computed from the global covariance matrix (*unsupervised*)

PCA Transform (Abridged)

1. Compute mean, covariance matrix for training set
   - e.g. MATLAB: m = mean(Train);   C = cov(Train.data);

2. Find the (unit-length) eigenvectors of the covariance matrix (see DHS Appendix A2.7) - *complexity O(D^3) for DxD matrix*
   - e.g. MATLAB: [V, L] = eig(C)

3. Sort eigenvectors by decreasing eigenvalue

4. Choose k eigenvectors with largest eigenvalues (principal components)

5. Return components as columns of a matrix, and associated eigenvalues (in a diagonal matrix)

Selection of Components

k Largest Eigenvalues

Correspond to eigenvectors in primary directions of variation within the data

- Large eigenvalues may be interpreted as the "inherent dimensionality" of 'signal' in the data
- Often only a small number of large eigenvalues

m - k Remaining Eigenvalues

Generally contain noise (random variation)

Why PCA?

Features are mutually uncorrelated (artifact of covariance matrix being real and symmetric)

The feature space reduction produced by a PCA with k components minimizes the mean-squared error between samples in the original space, and the newly transformed space, for any k-element transform matrix:

$$J_k = \sum_{i=1}^{n} ||(\mu_0 + \sum_{j=1}^{k} a_{ij} e_i) - x_i||^2$$

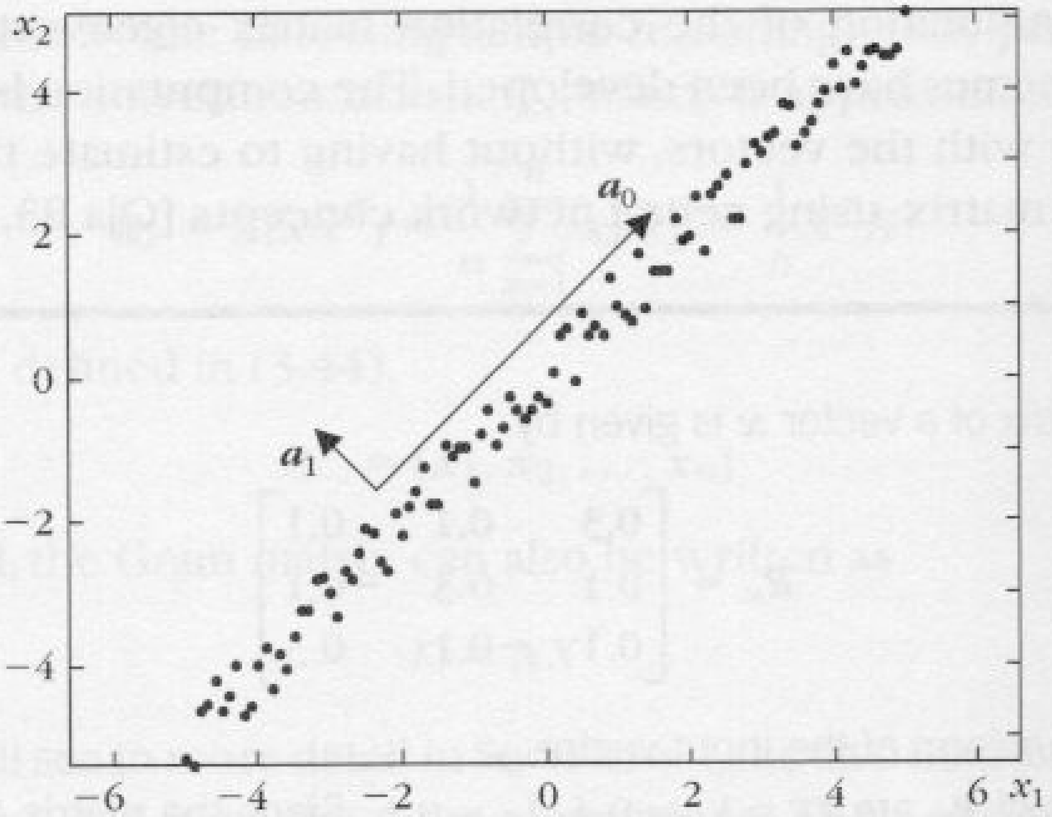


FIGURE 6.2

Points around the $\boldsymbol{x}_2 = \boldsymbol{x}_1$ line. The eigenvectors of the associated covariance matrix are $\boldsymbol{a}_0$ and $\boldsymbol{a}_1$. The principal eigenvector $\boldsymbol{a}_0$ points in the direction of maximum variance.

The New Order (Feature Space)

Feature Space after PCA:

Becomes *coefficients* of the principal components (first, including all *d* eigenvectors):

$$x = \mu_0 + \sum_{i=1}^{d} a_i e_i$$

To reduce feature space size, we limit the number of principle components to *k*:

$$\hat{x} = \mu_0 + \sum_{i=1}^{k} a_i e_i$$

Coefficients (Bishop, Ch. 12)

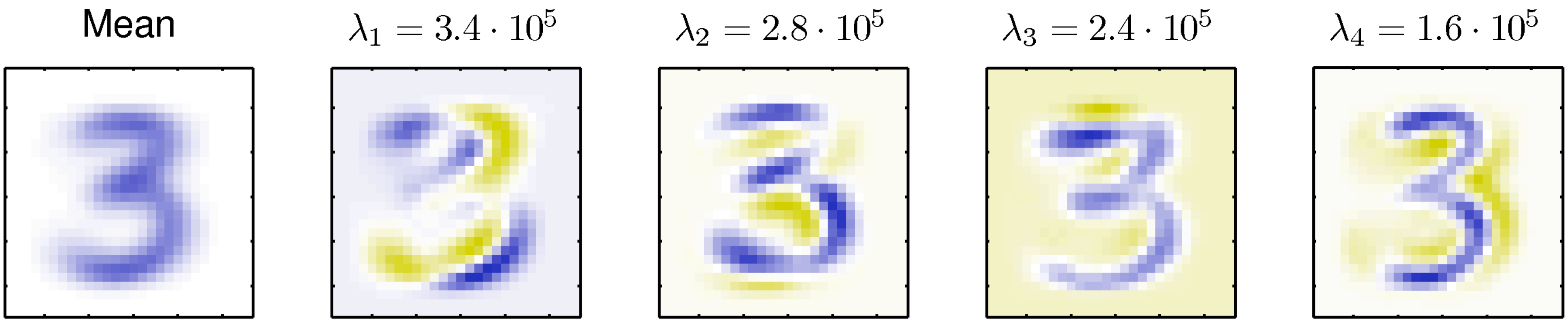
$$\tilde{x}_n = \sum_{i=1}^{k} (x_n^T e_i) e_i + \sum_{i=k+1}^{D} (\mu_0^T e_i) e_i$$

$$= \mu_0 + \sum_{i=1}^{k} (x_n^T e_i - \mu_0^T e_i) e_i$$

Coefficients (a_i)

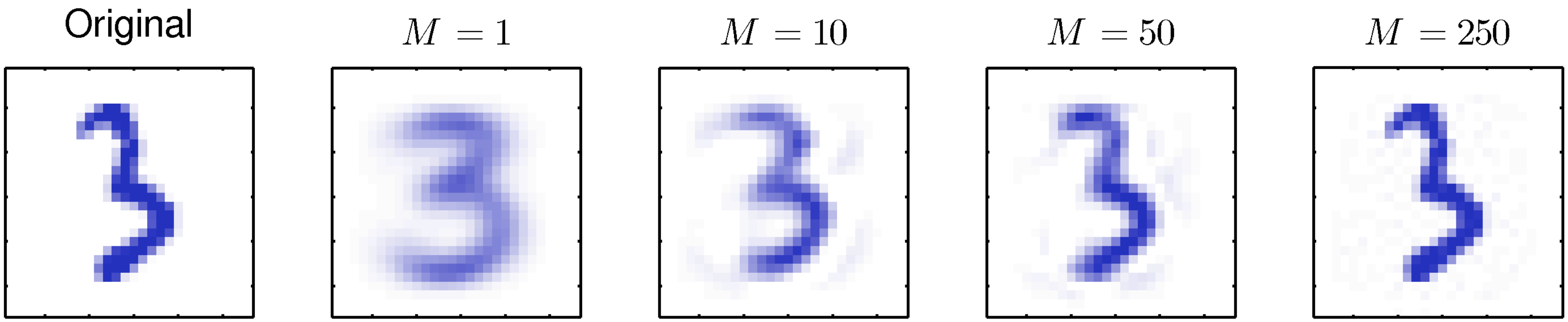
For each eigenvector used (component), difference between inner product with original sample and global mean

Example: MNIST (Bishop, Ch. 12)



Mean $\lambda_1 = 3.4 \cdot 10^5$ $\lambda_2 = 2.8 \cdot 10^5$ $\lambda_3 = 2.4 \cdot 10^5$ $\lambda_4 = 1.6 \cdot 10^5$

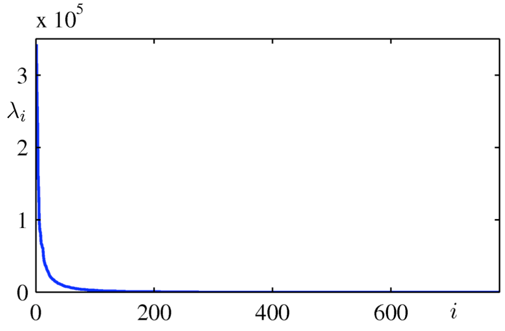
Eigenvectors shown in yellowish-green: eigenvalues above images



Original $M = 1$ $M = 10$ $M = 50$ $M = 250$

M: # Principal Components Utilized (max. components = 784 (28x28))

Eigenvalue spectrum for digit data:



(a)

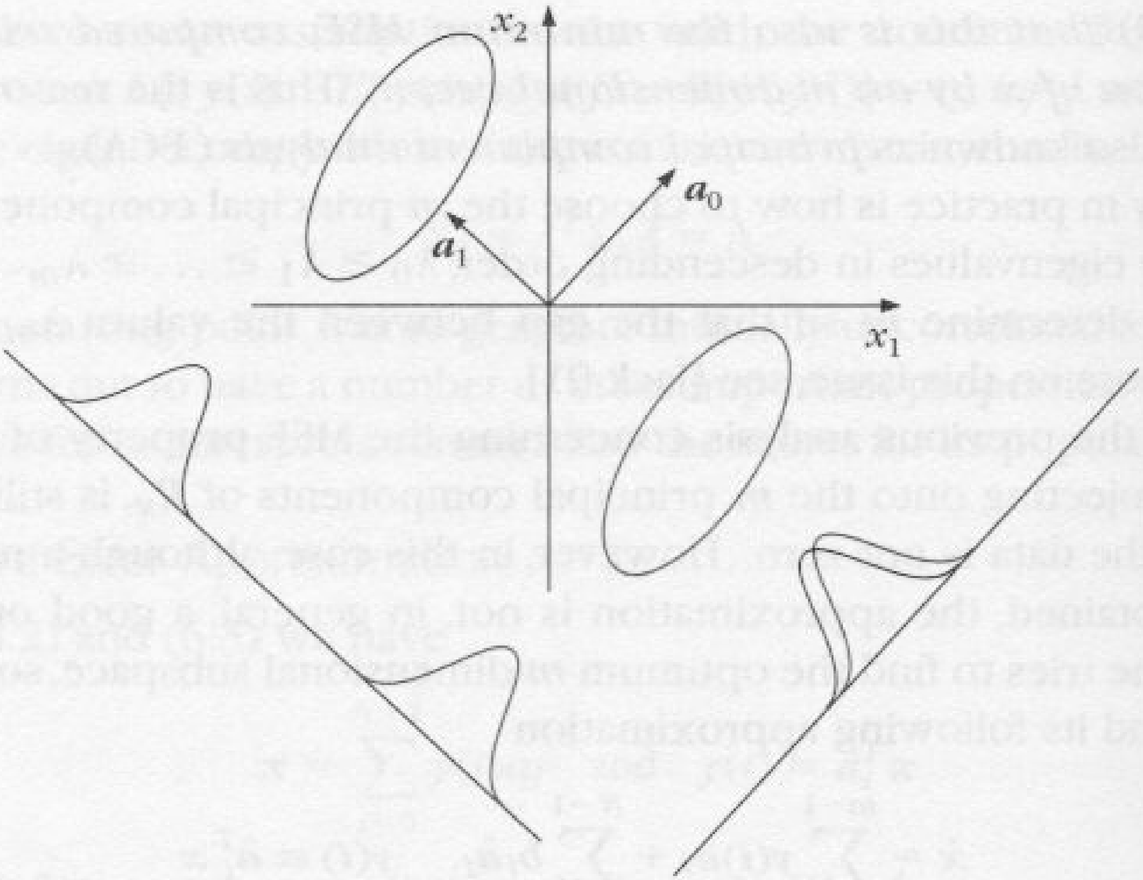


FIGURE 6.1

The KL transform is not always best for pattern recognition. In this example, projection on the eigenvector with the larger eigenvalue makes the two classes coincide. On the other hand projection on the other eigenvector keeps the classes separated.

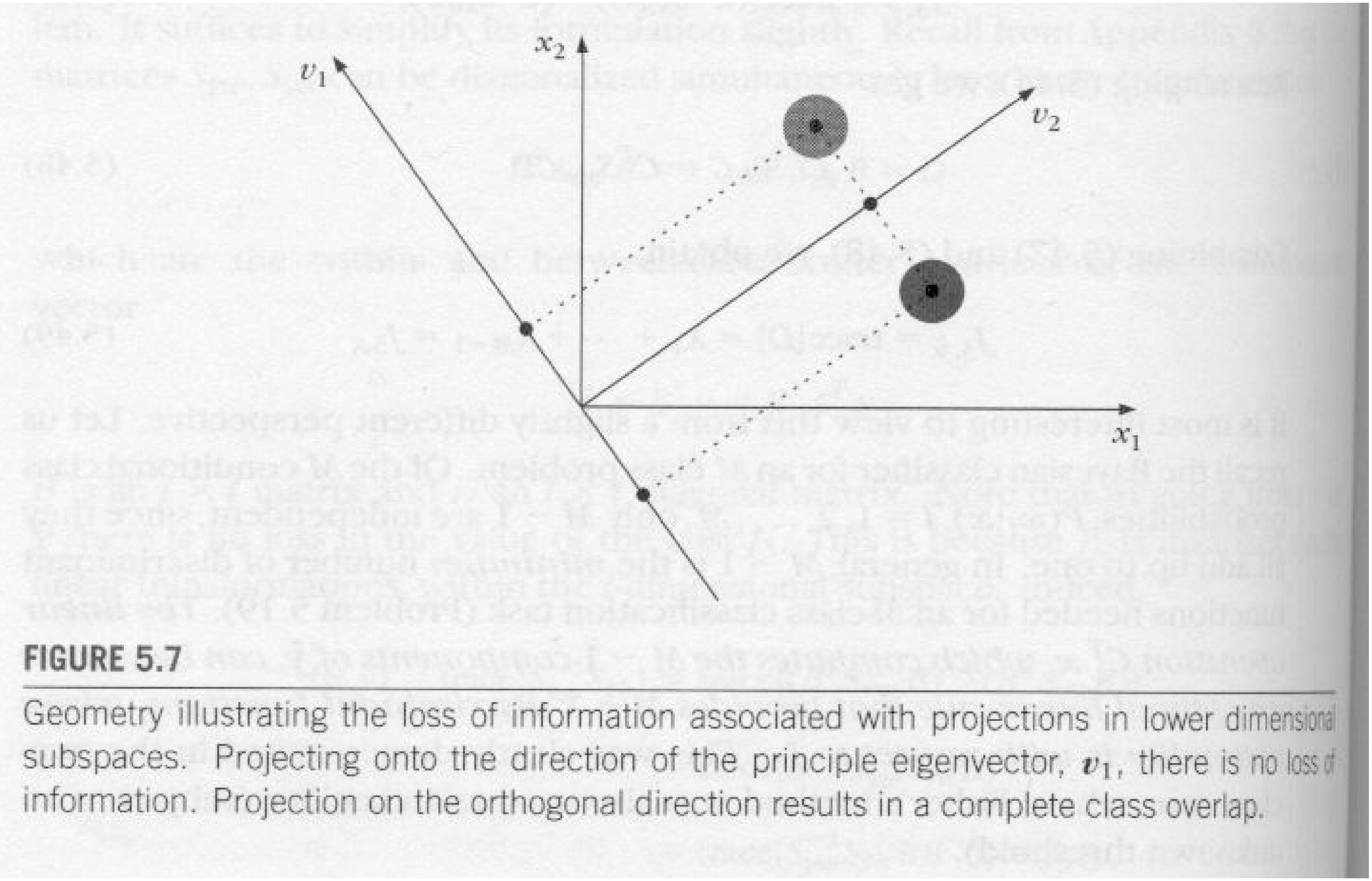


FIGURE 5.7

Geometry illustrating the loss of information associated with projections in lower dimensional subspaces. Projecting onto the direction of the principle eigenvector, $\boldsymbol{v}_1$, there is no loss of information. Projection on the orthogonal direction results in a complete class overlap.