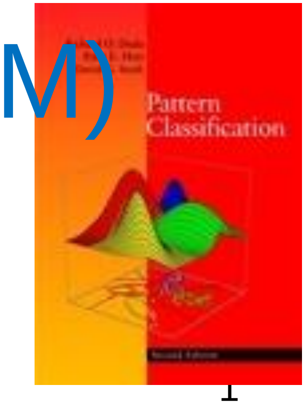
Chapter 3 (part 3)

- Problems of Dimensionality
- Maximum-Likelihood and Bayesian Parameter Estimation
- Fisher Linear Discriminant
- Expectation-Maximization (EM)



All materials used in this course were taken from the textbook *"Pattern Classification"* by Duda et al., John Wiley & Sons, 2001 with the permission of the authors and the publisher

- Problems of Dimensionality
  - Problems involving 50 or 100 features (binary valued)

1. Classification accuracy depends upon the dimensionality and the amount of training data.
2. The computational complexity of designing the classifier.

Case of two classes multivariate normal with the same covariance:

It can be shown:

$$P(error) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-\mu^2/2}\, du$$

$$where: \quad r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\lim_{r \to \infty} P(error) = 0$$

r is the squared Mahalanobis distance

- The probability of error decreases as *r* increases, approaching zero as *r* approaches infinity.
- If features are independent then:

$$\mathbf{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{i=d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- Most useful features are the ones for which the difference between the means is large relative to the standard deviation.

- It has frequently been observed in practice that, beyond a certain point, the inclusion of additional features leads to worse rather than better performance: we have the wrong model !

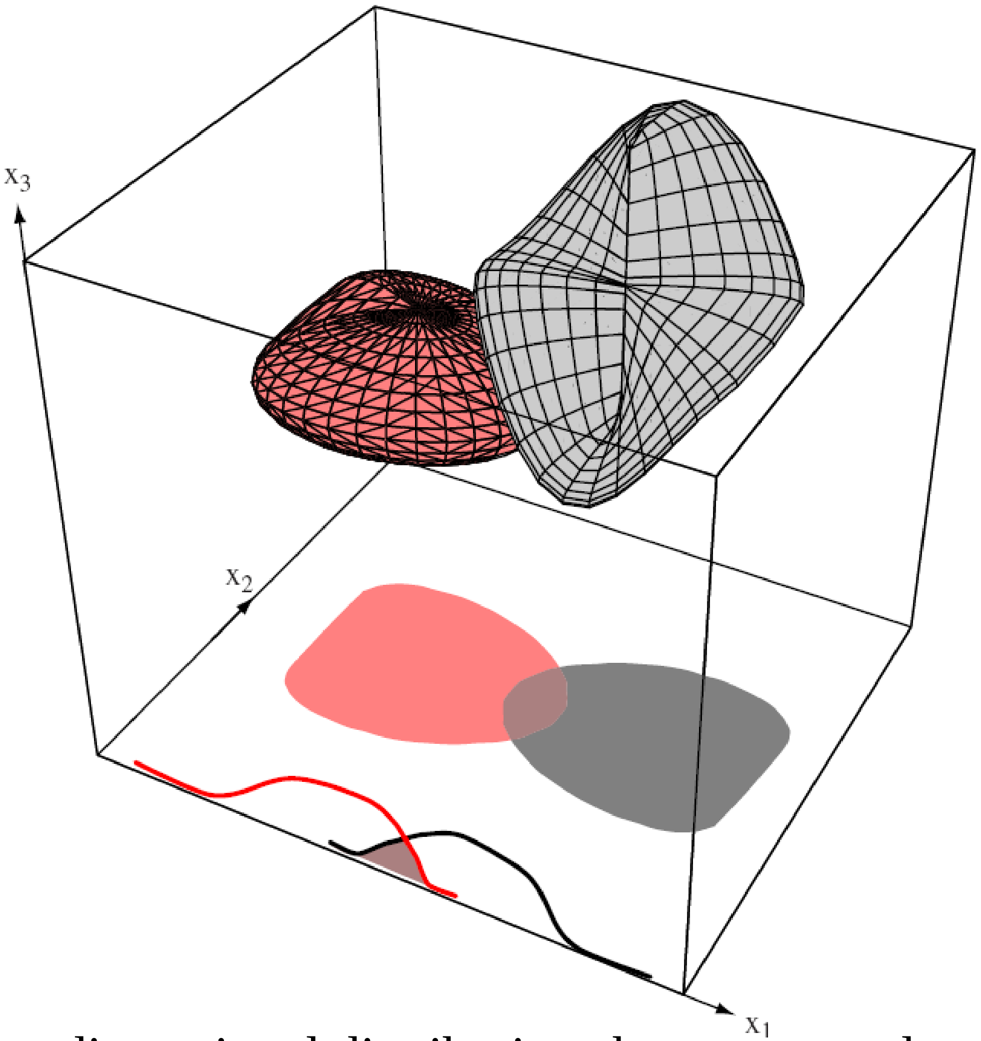


Figure 3.3: Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace — here, the two-dimensional x_1-x_2 subspace or a one-dimensional x_1 subspace — there can be greater overlap of the projected distributions, and hence greater Bayes errors.

- Fusing of different types of information, referred to as feature fusion, is a good application for Principal Components Analysis (PCA).
- Increasing the feature vector dimension can significantly increase the memory (e.g., the number of elements in the covariance matrix grows as the square of the dimension of the feature vector) and computational complexity.
- Good rule of thumb: 10 independent data samples for every parameter to be estimated.
- For practical systems, such as speech recognition, even this simple rule can result in a need for vast amounts of data.

- Computational Complexity

  – Our design methodology is affected by the computational difficulty

    - "big oh" notation
      $f(x) = O(h(x))$ "big oh of $h(x)$"

If: $\exists(c_0, x_0) \in \Re^2; \ \left|f(x)\right| \leq c_0\left|h(x)\right| \text{ for all } x \succ x_0$

(An upper bound on $f(x)$ grows no worse than $h(x)$ for sufficiently large x!)

$f(x) = 2+3x+4x^2$

$g(x) = x^2$

$f(x) = O(x^2)$

- "big oh" is not unique!

  $f(x) = O(x^2)$; $f(x) = O(x^3)$; $f(x) = O(x^4)$

- "big theta" notation

  $f(x) = \Theta(h(x))$

If:

$$\exists (x_0, c_1, c_2) \in \Re^3; \forall x > x_0$$

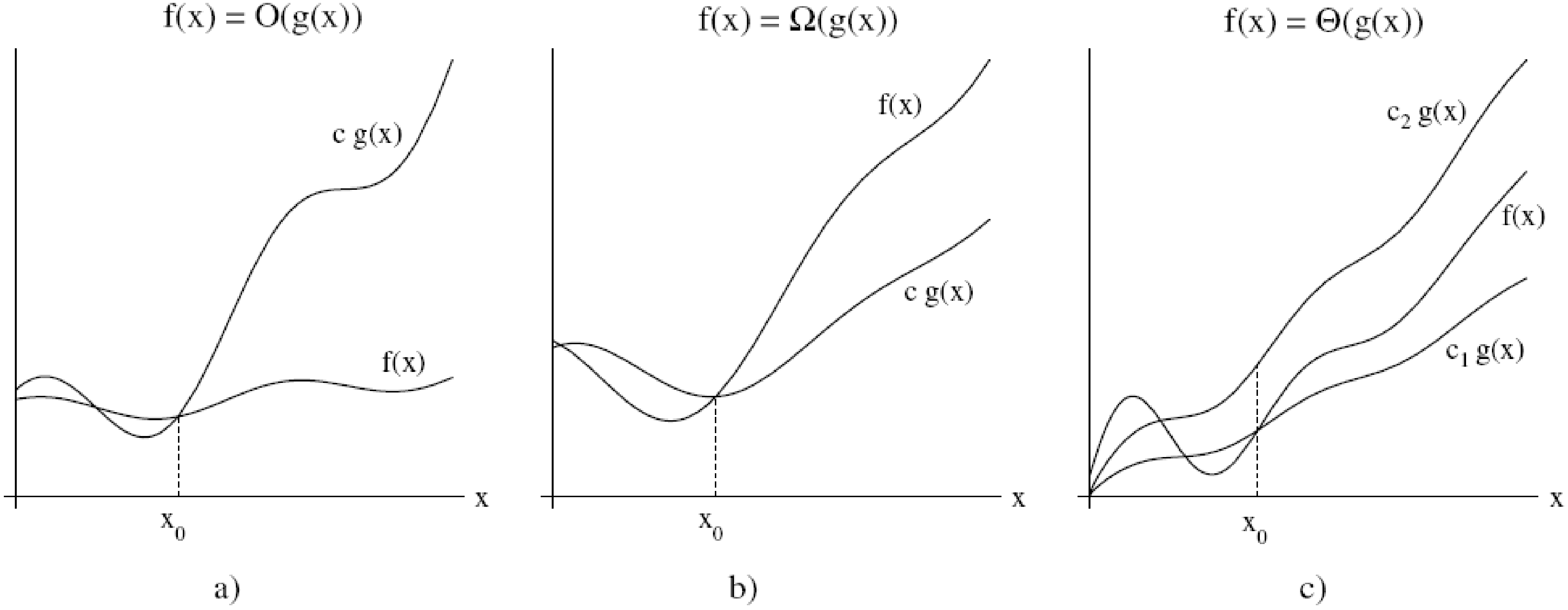
$$0 \le c_1 h(x) \le f(x) \le c_2 h(x)$$

$f(x) = \Theta(x^2)$ but $f(x) \neq \Theta(x^3)$

Asymptotic upper bound $O(g(x)) = \{f(x)$: there exist positive constants c and x_0 such that $0 \leq f(x) \leq cg(x)$ for all $x \geq x_0\}$

Asymptotic lower bound $\Omega(g(x)) = \{f(x)$: there exist positive constants c and x_0 such that $0 \leq cg(x) \leq f(x)$ for all $x \geq x_0\}$

Asymptotically tight bound $\Theta(g(x)) = \{f(x)$: there exist positive constants c_1, c_2, and x_0 such that $0 \leq c_1 g(x) \leq f(x) \leq c_2 g(x)$ for all $x \geq x_0\}$



Complexity of the ML Estimation

- Gaussian priors in *d* dimensions classifier with *n* training samples for each of *c* classes

- For each category, we have to compute the discriminant function

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \overbrace{\hat{\boldsymbol{\mu}}}^{O(d.n)})^t \overbrace{\hat{\boldsymbol{\Sigma}}^{-1}}^{O(n.d^3)} (\mathbf{x} - \hat{\boldsymbol{\mu}}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \underbrace{\frac{1}{2}\ln\left|\hat{\boldsymbol{\Sigma}}\right|}_{O(d^3)} + \underbrace{\ln P(\omega)}_{O(n)}$$

- The computational complexity of finding the sample mean $\hat{\boldsymbol{\mu}}$ is *O*(*nd*), since for each of the *d* dimensions we must add *n* component values.

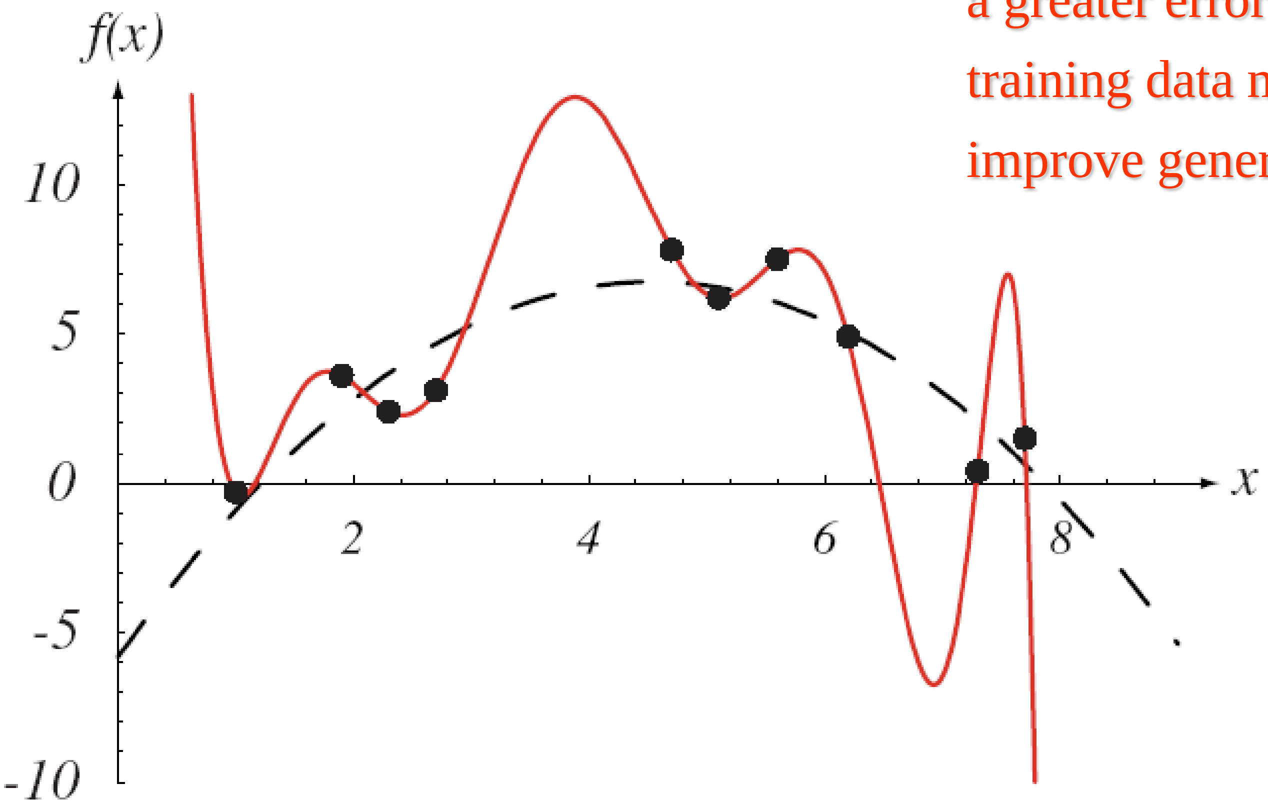
- For each of the *d*(*d* + 1)/2 independent components of the sample covariance matrix **Σ** there are *n* multiplications and additions, giving a complexity of O(d^2n).

- Determinant of **Σ** is an $O(d^3)$ calculation, as we can easily verify by counting the number of operations in matrix “sweep” methods.
- The inverse can be calculated in $O(d^3)$ calculations, for instance by Gaussian elimination.
- The complexity of estimating $P(\omega)$ is of course $O(n)$.
- Total = $O(d^3.n)$ Total for c classes = $O(cd^3.n) \approx \mathrm{O}(d^3.n)$
  Cost increase when d and n are large!
- Bayesian learning has higher complexity as a consequence of integrating over model parameters $\boldsymbol{\theta}$.
- … space complexity – time complexity ….

Overfitting

- It frequently happens that the number of available samples is inadequate, and the question of how to proceed arises.
- One possibility is to reduce the dimensionality, either by redesigning the feature extractor, by selecting an appropriate subset of the existing features, or by combining the existing features in some way (Ch10).
- Another possibility is to assume that all c classes share the same covariance matrix (pooled covariance).
- Yet another alternative is to look for a better estimate for $\mathbf{\Sigma}$ (e.g., use Bayesian parameter estimate).

- If a priori estimate $\boldsymbol{\Sigma}_0$ is available, a Bayesian or pseudo-Bayesian estimate of the form $\lambda\boldsymbol{\Sigma}_0 + (1-\lambda)\hat{\boldsymbol{\Sigma}}$ might be employed.
- For example, one might assume that all covariances for which the magnitude of the correlation coefficient is not near unity are actually zero. An extreme of this approach is to assume statistical independence, thereby making all the off-diagonal elements be zero, regardless of empirical evidence to the contrary.
- Even though such assumptions are almost surely incorrect, the resulting heuristic estimates sometimes provide better performance than the maximum likelihood estimate of the full parameter space. → paradox. → problem of insufficient data (an analogous problem in curve fitting).



a greater error on training data might improve generalization!

FIGURE 3.4. The “training data” (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2+bx+c+\varepsilon$ where $p(\varepsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because it would lead to better predictions for new samples.

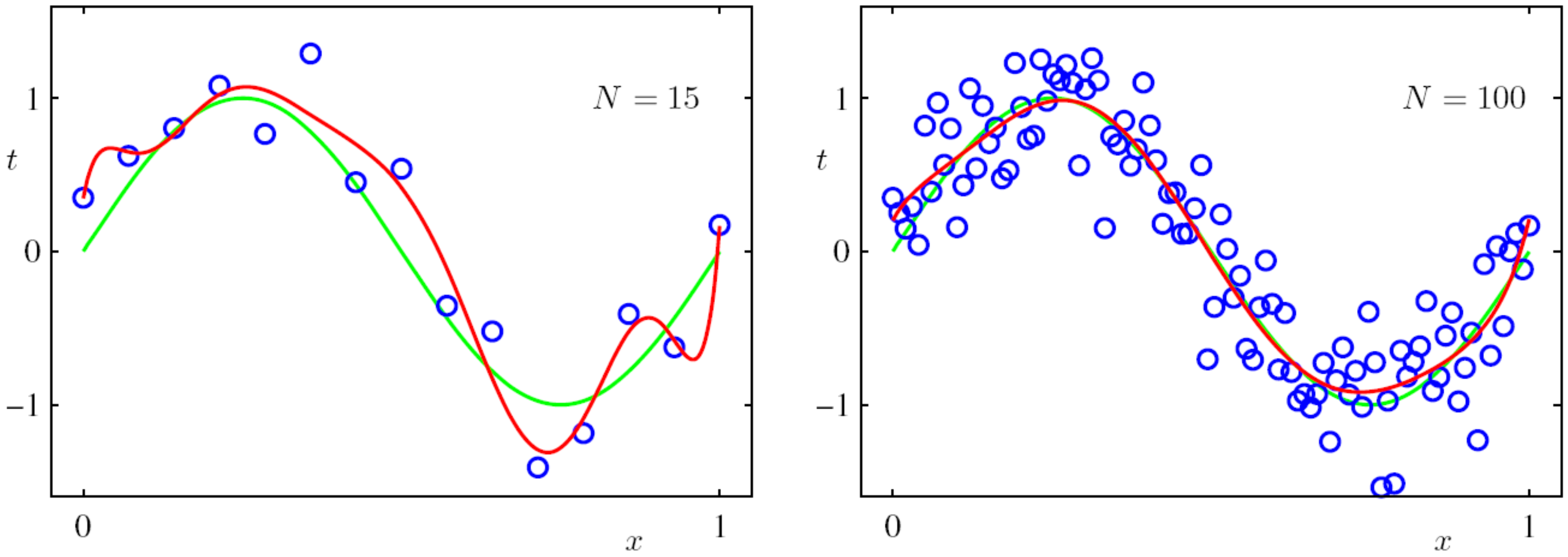


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the *M* = 9 polynomial for *N* = 15 data points (left plot) and *N* = 100 data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

- In fitting the points in Fig. 3.4, then, we might consider beginning with a high order polynomial (e.g., 10th order), and successively smoothing or simplifying our model by eliminating the highest-order terms. While this would in virtually all cases lead to greater error on the "training data," we might expect the generalization to improve.
- There are a number of heuristic methods that can be applied in the Gaussian classifier case.
- For instance, suppose we wish to design a classifier for distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and we have reason to believe that we have insufficient data for accurately estimating the parameters.

- We might make the simplification that they have the same covariance, i.e., $N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$, and estimate $\boldsymbol{\Sigma}$ accordingly. Such estimation requires proper normalization of the data.
- An intermediate approach is to assume a weighted combination of the equal and individual covariances, a technique known as *shrinkage*, (also called regularized discriminant analysis) since the individual covariances "shrink" toward a common one. If i is an index on the c categories in question, we have

$$\boldsymbol{\Sigma}_i(\alpha)=\frac{(1-\alpha)n_i\boldsymbol{\Sigma}_i+\alpha n\boldsymbol{\Sigma}}{(1-\alpha)n_i+\alpha n} \quad \text{for } 0<\alpha<1.$$

- We could “shrink” the estimate of the (assumed) common covariance matrix toward the identity matrix, as

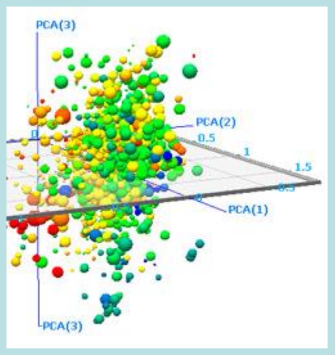
$$\boldsymbol{\Sigma}(\beta) = \left(1-\beta\right)\boldsymbol{\Sigma} + \beta\mathbf{I} \qquad \text{for } 0 < \beta < 1$$

- Such methods for simplifying classifiers have counterparts in regression, generally known as *ridge regression* (رگرسیون ستیغی). In ANN this is called *weight decay.*

Component Analysis and Discriminants

- Goal: Combine features in order to reduce the dimension of the feature space
  - Linear combinations are simple to compute and tractable
  - Project high dimensional data onto a lower dimensional space
  - Three classical approaches for finding “optimal” linear transformation
    - PCA (Principal Component Analysis) “Projection that best represents the data in a least- square sense”
    - MDA (Multiple Discriminant Analysis) “Projection that best separates the data in a least-squares sense”
    - Independent Component Analysis (ICA): projection that minimizes the mutual information of the components.

Principal Component Analysis



- Given *n* d-dimensional samples $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$. Representing the set by $\mathbf{x}_0$ (finding 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

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

and seek the value of $\mathbf{x}_0$ that minimizes J_0.

- Solution: $\mathbf{x}_0=\mathbf{m}$, where $\mathbf{m}$ is the sample mean,

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

- Proof:

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

$$= \sum_{k=1}^{n}\|(\mathbf{x}_0 - \mathbf{m})\|^2 - 2\sum_{k=1}^{n}(\mathbf{x}_0 - \mathbf{m})^t(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n}\|(\mathbf{x}_k - \mathbf{m})\|^2$$

$$= \sum_{k=1}^{n}\|(\mathbf{x}_0 - \mathbf{m})\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^t\cancelto{0}{\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})} + \sum_{k=1}^{n}\|(\mathbf{x}_k - \mathbf{m})\|^2$$

$$= \sum_{k=1}^{n}\|(\mathbf{x}_0 - \mathbf{m})\|^2 + \underbrace{\sum_{k=1}^{n}\|(\mathbf{x}_k - \mathbf{m})\|^2}_{\textit{independent of } \mathbf{x}_0} \rightarrow \text{Minimized by } \mathbf{x}_0=\mathbf{m}$$

- The sample mean is a zero-dimensional representation of data set. 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, $\mathbf{x}=\mathbf{m}+a\mathbf{e}$, where **e** is a unit vector in the direction of the line.
- If we represent $\mathbf{x}_k$ by $\mathbf{m}+a_k\mathbf{e}$**,** we can find an "optimal" set of coefficient a_k by minimizing the squared-error criterion function

$$J_1(a_1,\ldots,a_n,\mathbf{e}) = \sum_{k=1}^{n} \|(\mathbf{m}+a_k\mathbf{e})-\mathbf{x}_k\|^2 = \sum_{k=1}^{n} \|a_k\mathbf{e}-(\mathbf{x}_k-\mathbf{m})\|^2$$

$$= \sum_{k=1}^{n} a_k{}^2 \|\mathbf{e}\|^2 - 2\sum_{k=1}^{n} a_k\mathbf{e}^t(\mathbf{x}_k-\mathbf{m}) + \sum_{k=1}^{n} \|(\mathbf{x}_k-\mathbf{m})\|^2 \qquad (82)$$

Recognizing that ||**e**||=1, partially differentiating with respect to a_k, and setting the derivative to zero, we obtain

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

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

Finding the best direction **e** for the line. → Scatter Matrix **S.**

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

- Scatter matrix **S** is *n*-1 times the sample covariance matrix. Using Eqs 82 , 83 →

$$J_1(\mathbf{e}) = \sum_{k=1}^{n} a_k{}^2 - 2\sum_{k=1}^{n} a_k{}^2 + \sum_{k=1}^{n} \|(\mathbf{x}_k - \mathbf{m})\|^2$$

$$= -\sum_{k=1}^{n} [\mathbf{e}^t(\mathbf{x}_k - \mathbf{m})]^2 + \sum_{k=1}^{n} \|(\mathbf{x}_k - \mathbf{m})\|^2$$

$$= -\sum_{k=1}^{n} \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t\mathbf{e} + \sum_{k=1}^{n} \|(\mathbf{x}_k - \mathbf{m})\|^2$$

$$= -\mathbf{e}^t\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} \|(\mathbf{x}_k - \mathbf{m})\|^2$$

- The vector **e** that minimizes J_1 also maximizes $\mathbf{e}^t\mathbf{S}\mathbf{e}$. We use the method of Lagrange multiplier to maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to the constraint that $\|\mathbf{e}\|=1$.

Letting λ be the undetermined multiplier we differentiate

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

with respect to **e** and equating to zero to obtain

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0 \Rightarrow \mathbf{S}\mathbf{e} = \lambda\mathbf{e} \Rightarrow \mathbf{e}^{t}\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^{t}\mathbf{e} = \lambda$$

To maximize $\mathbf{e}^{t}\mathbf{S}\mathbf{e}$ we want to select the eigenvector corresponding to the largest eigenvalue of the scatter matrix.

This interesting result can be extended from one-dimensional projection to a *d*-dimensional projection.

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$

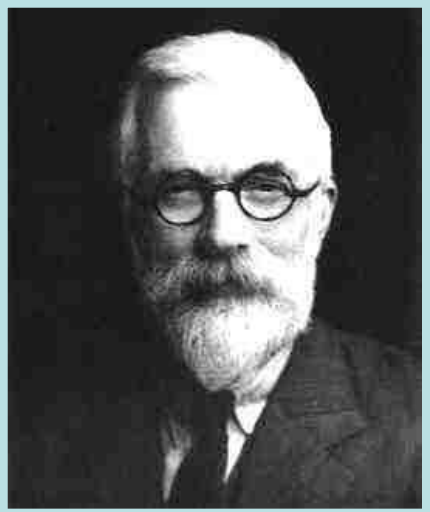
Where $d' < d$. It can be shown that

$$J_{d'} = \sum_{k=1}^{n} \| (\mathbf{m} + \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i) - \mathbf{x}_k \|^2$$

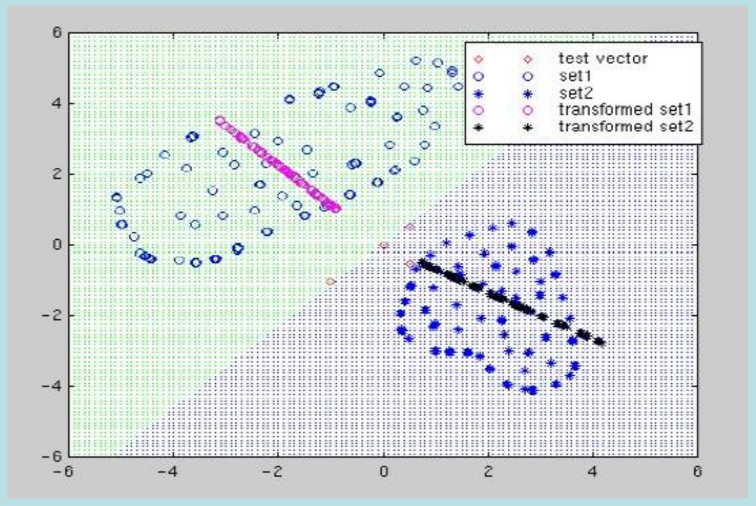
is minimized when the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{d'}$ are the d' eigenvectors of the scatter matrix having the largest eigenvalues. Because the scatter matrix is real and symmetric, these eigenvectors are orthogonal.

They form a natural set of basis vectors for representing any feature vector **x**. The *coefficients* a_i are the components of **x** in that basis and are called *principal components*.

Fisher Linear Discriminant



- Although PCA finds components that are useful for representing data, there is no reason to assume that these components must be useful for discriminating between data in different classes.
- Discriminant analysis seeks directions that are efficient for discrimination.



- We can reduce the dimensionality from *d* dimensions to one dimension if we merely project the *d*-dimensional data onto a line.
- By moving the line around, we might be able to find an orientation for which the projected samples are well separated. This is exactly the goal of classical discriminant analysis.
- Suppose that we have a set of *n* *d*-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, n_1 in the subset D_1 labeled ω_1 and n_2 in the subset D_2 labeled ω_2. If we form a linear combination of the components of **x**, we obtain the scalar dot product

$$y = \mathbf{w}^t\mathbf{x}$$

and a corresponding set of n samples $y_1, \ldots, y_n$ divided into the subsets Y_1 and Y_2.

Geometrically, if $\|\mathbf{w}\| = 1$, each y_i is the projection of the corresponding $\mathbf{x}_\mathrm{i}$ onto a line in the direction of **w**.

The magnitude of **w** is not important but its direction is.

If we imagine that the samples labeled ω_1 fall more or less into one cluster while those labeled ω_2 fall in another, we want the projections falling onto the line to be well separated, not thoroughly intermingled.

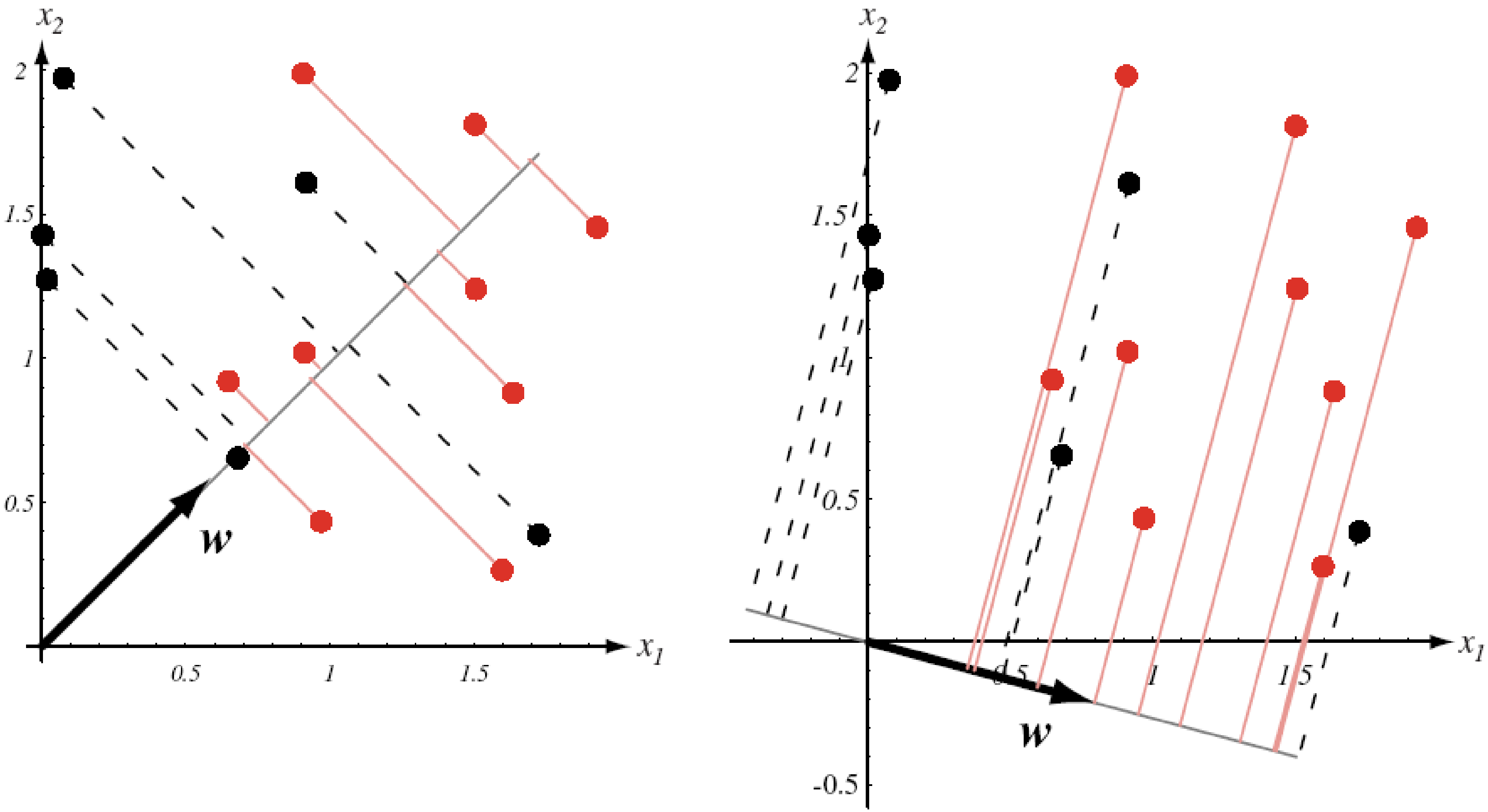


FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points.

- Finding the best such direction **w**, one we hope will enable accurate classification.
- A measure of the separation between the projected points is the difference of the sample means.
- If m$_i$ is the *d*-dimensional sample mean given by

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

the sample mean for the projected points is given by

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i$$

and is simply the projection of m_i.

The distance between the projected means is

$$\left|\tilde{m}_1 - \tilde{m}_2\right| = \left|\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2\right|$$

and that we can make this difference as large as we wish merely by scaling **w**.

To obtain good separation of the projected data we really want the difference between the means to be large relative to some measure of the standard deviations for each class.

Rather than forming sample variances, we define the *scatter* for projected samples labeled ω_i by

$$\tilde{s}_i^{\,2} = \sum_{y \in Y_i} \left(y - \tilde{m}_i\right)^2$$

Thus, $(\frac{1}{n})(\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2})$ is an estimate of the variance of the pooled data, and $(\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2})$ is called the total *within-class scatter* of the projected samples.

The *Fisher linear discriminant* employs that linear function $\mathbf{w}^t\mathbf{x}$ for which the criterion function

$$\boxed{J(\mathbf{w}) = \frac{\left|\tilde{m}_1 - \tilde{m}_2\right|^2}{\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2}}}$$

is maximum (and independent of $\|\mathbf{w}\|$).

While the **w** maximizing $J(\cdot)$ leads to the best separation between the two projected sets, we will also need a *threshold* criterion before we have a true classifier. We first consider how to find the optimal **w**, and later turn to the issue of thresholds.

To obtain $J(\cdot)$ as an explicit function of **w**, we define the *scatter matrices* $\mathbf{S}_i$ and scatter $\mathbf{S}_W$ by

$$\mathbf{S}_i = \sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

and $$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

Then we can write

$$\tilde{\mathbf{s}}_i^{\,2} = \sum_{y \in Y_i} \left(y - \tilde{m}_i \right)^2 = \sum_{\mathbf{x} \in D_\mathbf{i}} \left(\mathbf{w}^t \mathbf{x} - \mathbf{w}^t \mathbf{m}_i \right)^2$$

$$= \sum_{\mathbf{x} \in D_\mathbf{i}} \mathbf{w}^t \left(\mathbf{x} - \mathbf{m}_i \right) \left(\mathbf{x} - \mathbf{m}_i \right)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$$

therefore the sum of these scatters can be written

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2} = \mathbf{w}^t \mathbf{S}_W \mathbf{w}$$

Similarly, the separations of the projected means obeys

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left(\mathbf{w}^t \mathbf{m}_1 - \mathbf{w}^t \mathbf{m}_2 \right)^2$$

$$= \mathbf{w}^t \left(\mathbf{m}_1 - \mathbf{m}_2 \right) \left(\mathbf{m}_1 - \mathbf{m}_2 \right)^t \mathbf{w}$$

$$= \mathbf{w}^t \mathbf{S}_B \mathbf{w}$$

where $$\mathbf{S}_B = \left(\mathbf{m}_1 - \mathbf{m}_2 \right) \left(\mathbf{m}_1 - \mathbf{m}_2 \right)^t$$

We call $\mathbf{S}_W$ the ***W****ithin-class scatter matrix*. It is proportional to the sample covariance matrix for the pooled *d*-dimensional data. It is symmetric and positive semidefinite, and is usually nonsingular if $n > d$.

Likewise, $\mathbf{S}_B$ is called the ***B****etween class scatter matrix*. It is also symmetric and positive semidefinite, but because it is the outer product of two vectors, its rank is at most one. In particular, for any **w**, $\mathbf{S}_B\mathbf{w}$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$, and $\mathbf{S}_B$ is quite singular.

In terms of $\mathbf{S}_B$ and $\mathbf{S}_W$, the criterion function $J(\cdot)$ can be written as

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}$$

This expression is well known in mathematical physics as the generalized *Rayleigh quotient*.

It is easy to show that a vector **w** that maximizes $J(\cdot)$ must satisfy

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$$

for some constant λ, which is a generalized eigenvalue problem.

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}.$$

Since $\mathbf{S}_B\mathbf{w}$ is in the direction of $\mathbf{m}_1$-$\mathbf{m}_2$ and the scale factor for **w** is immaterial the solution for the **w** that optimizes $J(\cdot)$ is

$$\boxed{\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)}$$

Thus, we have obtained **w** for Fisher's linear discriminant — the linear function yielding the maximum ratio of between-class scatter to within-class scatter (called *canonical variate*).

This mapping is many-to-one, and in theory can not possibly reduce the minimum achievable error rate if we have a very large training set.

- All that remains is to find the threshold, i.e., the point along the one-dimensional subspace separating the projected points.
- When the conditional densities $p(\mathbf{x}/\omega_i)$ are multivariate normal with equal covariance matrices $\boldsymbol{\Sigma}$, we can calculate the threshold directly.

The optimal decision boundary is $\mathbf{w}^t\mathbf{x} + w_0 = 0$, $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

and where w_0 is a constant involving **w** and the prior probabilities.

- Thus, for the normal, equal-covariance case, the optimal decision rule is merely to decide ω_1 if Fisher's linear discriminant exceed some threshold, and to decide ω_2 otherwise. (Choose w_0 where the posteriors in the one-dimensional distributions are equal).

$$g(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T S_w^{-1}\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) - \ln\frac{P(\omega_2)}{P(\omega_1)}$$

- Let's work some examples (class-independent PCA and LDA).

http://www.ece.msstate.edu/research/isip/projects/speech/software/demonstrations/applets/util/pattern_recognition/current/index.html

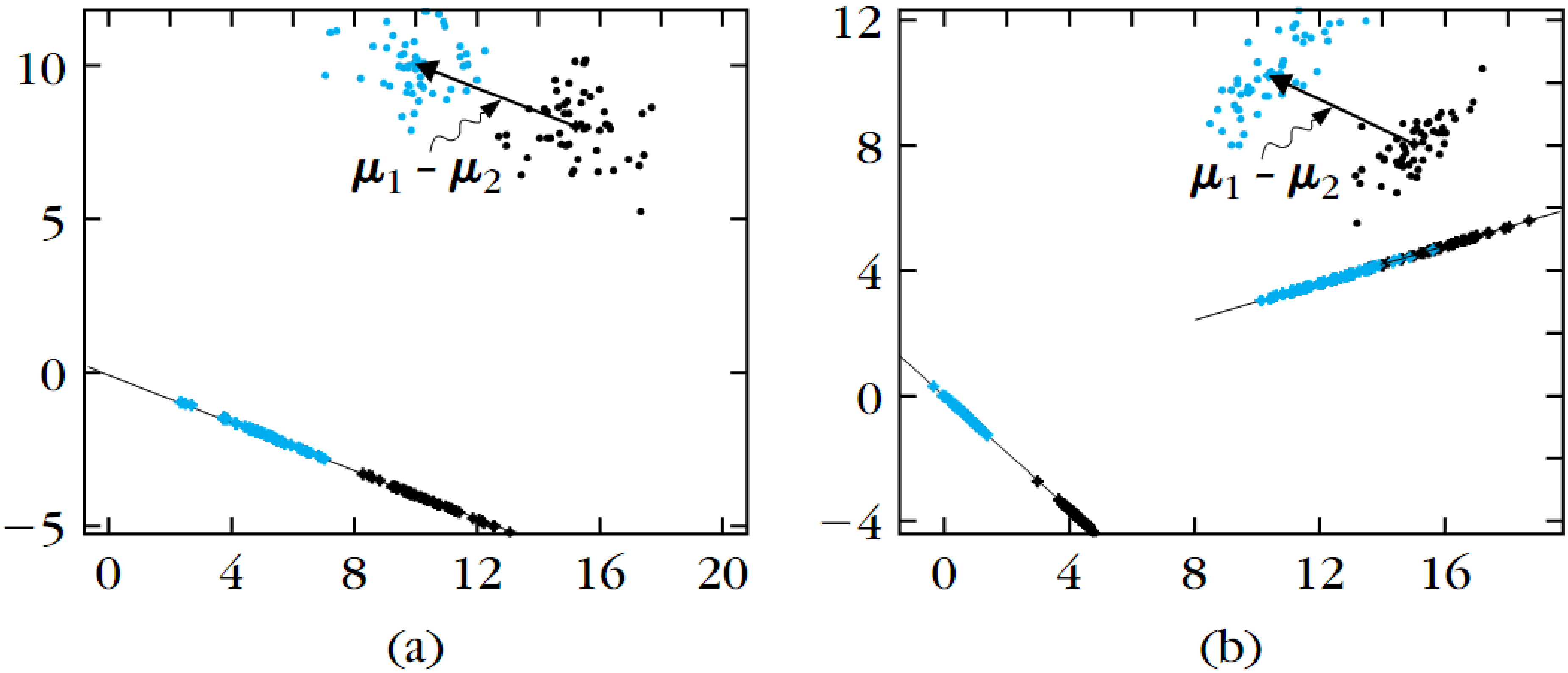


FIGURE 5.6 (a) The optimal line resulting from Fisher's criterion, for two Gaussian classes. Both classes share the same diagonal covariance matrix, with equal elements on the diagonal. The line is parallel to $\boldsymbol{\mu}_1$-$\boldsymbol{\mu}_2$. (b) The covariance matrix for both classes is nondiagonal. 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.

Multiple Discriminant Analysis

- For the *c*-class problem, the natural generalization of Fisher's linear discriminant involves *c* - 1 discriminant functions.
- Thus, the projection is from a *d*-dimensional space to a (*c* -1)-dimensional space, and it is tacitly assumed that $d \geq c$. Within-class scatter matrix

$$\mathbf{S}_W = \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

The proper generalization for $\mathbf{S}_B$ is not quite so obvious. Suppose that we define a *total mean vector* **m** and a *total scatter matrix* $\mathbf{S}_T$ by

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i$$

$$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$$

Then it follows that

$$\mathbf{S}_T = \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^t =$$

$$\sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t + \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t$$

$$= \mathbf{S}_W + \sum_{i=1}^{c} n_i (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t$$

The 2^{nd} term is defined as a *general Between-class scatter matrix*, so that the total scatter is the sum of the Within-class scatter and the between-class scatter:

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$$

For the 2-class case, the resulting between-class scatter matrix is $n_1 n_2/n$ times our previous definition.

The projection from d-dim to (c-1)-dim is done by c-1 **discriminant** functions:

$$y_i = \mathbf{w}_i^t \mathbf{x} \quad i = 1,\ 2,\ ...,\ c\text{-}1.$$

If the y_i are viewed as components of a vector **y** and the weight vectors $\mathbf{w}_i$ are viewed as the columns of a d-by-(c - 1) matrix **W**, then the projection can be written as a single matrix equation

$$\boxed{\mathbf{y} = \mathbf{W}^t\mathbf{x} = \left[\mathbf{w}_i^t\mathbf{x}\right] \quad i = 1,\ 2,\ ...,\ c - 1}$$

The samples $\mathbf{x}_1, \dots, \mathbf{x}_n$ project to a corresponding set of samples $\mathbf{y}_1, \dots, \mathbf{y}_n$, which can be described by their own mean vectors and scatter matrices. Thus, if we define

$$\tilde{\mathbf{m}}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in Y_i} \mathbf{y} \qquad \tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\mathbf{m}}_i$$

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{c} \tilde{S}_i = \sum_{i=1}^{c} \sum_{\mathbf{y} \in Y_i} \left(\mathbf{y} - \tilde{\mathbf{m}}_i\right)\left(\mathbf{y} - \tilde{\mathbf{m}}_i\right)^t$$

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i \left(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}\right)\left(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}\right)^t$$

It can be shown

$$\tilde{\mathbf{S}}_W = \mathbf{W}^t \mathbf{S}_W \mathbf{W}$$

$$\tilde{\mathbf{S}}_B = \mathbf{W}^t \mathbf{S}_B \mathbf{W}$$

These equations show how the within-class and between-class scatter matrices are transformed by the projection to the lower dimensional space.

What we seek is a transformation matrix **W** that in some sense maximizes the ratio of the between-class scatter to the within-class scatter.

A simple scalar measure of scatter is the determinant of the scatter matrix. The determinant is the product of the eigenvalues, and hence is the product of the "variances" in the principal directions, thereby measuring the square of the hyperellipsoidal scattering volume (ref. ch2).

$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|}$$

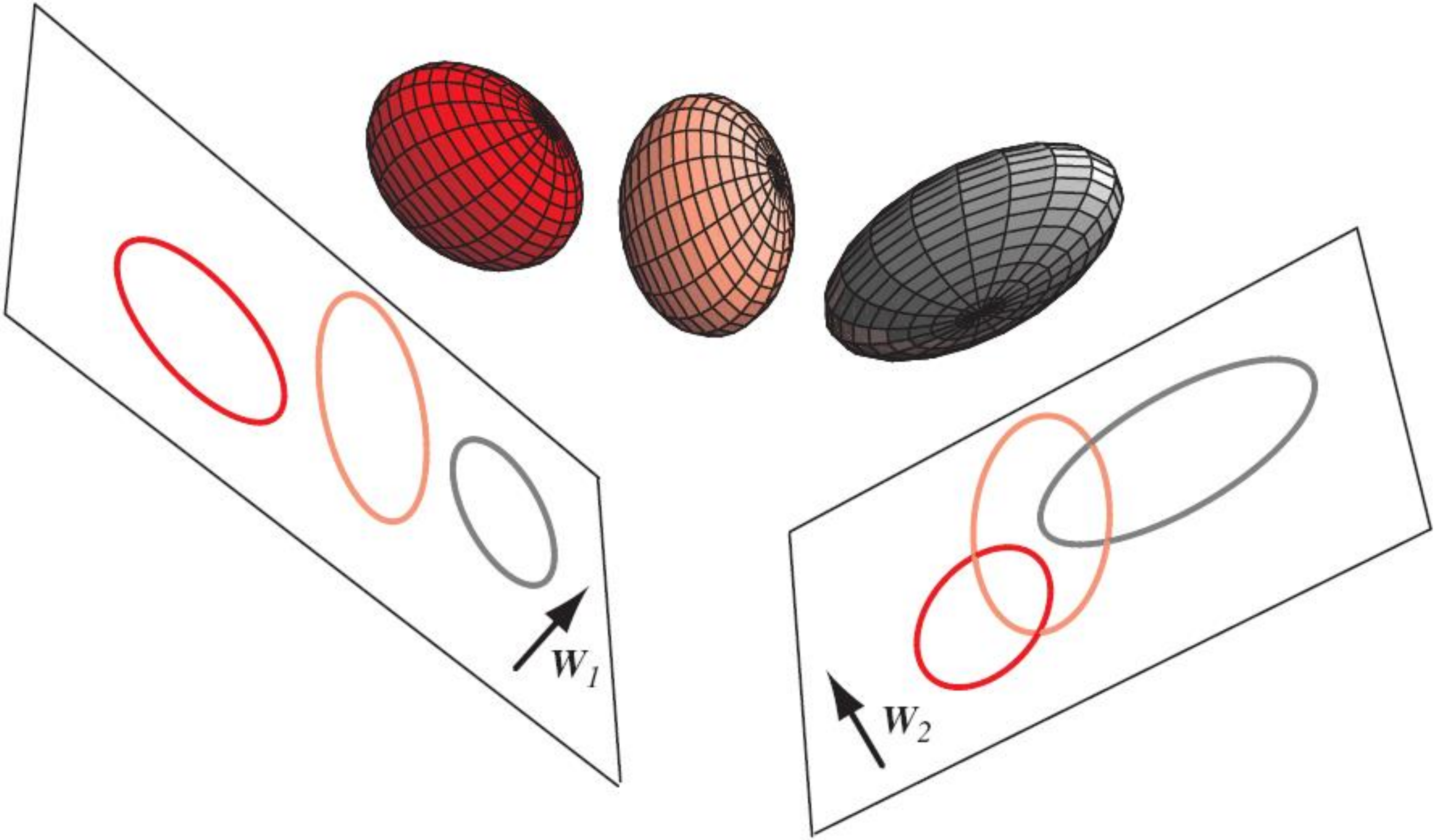


FIGURE 3.6. Three 3-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$.

Maximizing $J(.)$? → The columns of an optimal **W** are the generalized eigenvectors that correspond to the largest eigenvalues in

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i .$$

If $\mathbf{S}_W$ is non-singular, this can be converted to a conventional eigenvalue problem as before. However, this is actually undesirable, since it requires an unnecessary computation of the inverse of $\mathbf{S}_W$. Instead, one can find the eigenvalues as the roots of the characteristic polynomial

$$\left|\mathbf{S}_B - \lambda_i \mathbf{S}_W\right| = 0$$

and then solve

$$\left(\mathbf{S}_B - \lambda_i \mathbf{S}_W\right)\mathbf{w}_i = 0$$

directly for the eigenvectors $\mathbf{w}_i$.

Because $\mathbf{S}_B$ is the sum of c matrices of rank* one or less, and because only c-1 of these are independent, $\mathbf{S}_B$ is of rank c-1 or less. Thus, no more than c - 1 of the eigenvalues are nonzero, and the desired weight vectors correspond to these nonzero eigenvalues.

If the within-class scatter is isotropic, the eigenvectors are merely the eigenvectors of $\mathbf{S}_B$, and the eigenvectors with nonzero eigenvalues span the space spanned by the vectors $\mathbf{m}_i - \mathbf{m}$. In this special case the columns of $\mathbf{W}$ can be found simply by applying the Gram-Schmidt orthonormalization procedure to the c -1 vectors $\mathbf{m}_i - \mathbf{m}$, $i = 1, \ldots, c - 1$.

* *number of linearly independent rows or columns of a full matrix*

As in the two-class case, multiple discriminant analysis primarily provides a reasonable way of reducing the dimensionality of the problem. Parametric or nonparametric techniques that might not have been feasible in the original space may work well in the lower-dimensional space.

(Pattern Recognition.4th-Theodoridis-Koutroumbas) روش دوم تحلیل متمایز کننده چند کلاسه

$$\mathbf{y} = \mathbf{W}^t\mathbf{x} = \left[\mathbf{w}_i^t\mathbf{x}\right] \quad i = 1,\ 2,\ \ldots,\ l, \quad \mathbf{W}(d\times l), \mathbf{x}(d\times 1), \mathbf{y}(l\times 1)$$

$$\tilde{\mathbf{S}}_W = \mathbf{W}^t\mathbf{S}_W\mathbf{W} \qquad \tilde{\mathbf{S}}_B = \mathbf{W}^t\mathbf{S}_B\mathbf{W}$$

$$J_3(\mathbf{W}) = tr\left\{\frac{\tilde{\mathbf{S}}_B}{\tilde{\mathbf{S}}_W}\right\} = tr\left\{\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B\right\} = tr\left\{\left(\mathbf{W}^t\mathbf{S}_W\mathbf{W}\right)^{-1}\left(\mathbf{W}^t\mathbf{S}_B\mathbf{W}\right)\right\}$$

$$\to \text{maximization} \quad \frac{\partial J_3(\mathbf{W})}{\partial \mathbf{W}} = 0$$

$$-2\mathbf{S}_W\mathbf{W}\left(\mathbf{W}^t\mathbf{S}_W\mathbf{W}\right)^{-1}\left(\mathbf{W}^t\mathbf{S}_B\mathbf{W}\right)\left(\mathbf{W}^t\mathbf{S}_W\mathbf{W}\right)^{-1} + 2\mathbf{S}_B\mathbf{W}\left(\mathbf{W}^t\mathbf{S}_W\mathbf{W}\right)^{-1} = 0$$

⇨ $$\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)\mathbf{W} = \mathbf{W}\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B$$

دو ماتریس $\hat{\mathbf{S}}_\mathrm{W}$ و $\hat{\mathbf{S}}_\mathrm{B}$ را میتوان همزمان توسط یک تبدیل خطی، قطری نمود.

$$\mathbf{B}^t\tilde{\mathbf{S}}_W^{-1}\mathbf{B}=\mathbf{I}, \qquad \mathbf{B}^t\tilde{\mathbf{S}}_B\mathbf{B}=\mathbf{D}$$

که ماتریسهای پراکندگی درون کلاسی و بین کلاسی بردار تبدیل یافته زیر هستند

$$\hat{\mathbf{y}}=\mathbf{B}^t\mathbf{y}=\mathbf{B}^t\mathbf{W}^t\mathbf{x}$$

ماتریس $\mathbf{B}(l\times l)$ و ماتریس $\mathbf{D}(l\times l)$ قطری است. لازم بذکر است در تبدیل $\mathbf{y}$ به $\hat{\mathbf{y}}$ هیچ تلفاتی در مقدار J_3 وجود نخواهد داشت زیرا:

$$J_{3,\hat{\mathbf{y}}}=tr\left\{\tilde{\mathbf{S}}_{W,\hat{y}}^{-1}\tilde{\mathbf{S}}_{B,\hat{y}}\right\}=tr\left\{\left(\mathbf{W}^t\tilde{\mathbf{S}}_W\mathbf{W}\right)^{-1}\mathbf{W}^t\tilde{\mathbf{S}}_B\mathbf{W}\right\}=tr\left\{\mathbf{B}^{-1}\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B\mathbf{B}\right\}$$

$$=tr\left\{\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B\mathbf{B}\mathbf{B}^{-1}\right\}=tr\left\{\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B\right\}=J_{3,\mathbf{y}} \quad \blacksquare$$

$$\left.\begin{aligned}&\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)\mathbf{W}=\mathbf{W}\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B\\ &\mathbf{B}^t\tilde{\mathbf{S}}_W^{-1}\mathbf{B}=\mathbf{I}\\ &\mathbf{B}^t\tilde{\mathbf{S}}_B\mathbf{B}=\mathbf{D}\end{aligned}\right\} \quad \underbrace{\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)}_{d\times d}\underset{d\times l}{\mathbf{C}}=\underset{(d\times l)\times(l\times l)}{\mathbf{CD}} \quad , \quad \mathbf{C}_{d\times l}=\mathbf{WB}$$

eigenvalue–eigenvector problem

ماتریس $\mathbf{D}$ قطری و شامل مقادیر ویژه ماتریس $\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ است.
ماتریس $\mathbf{C}$ شامل بردارهای ویژه ماتریس $\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ در ستونهای خود است.

سوال: کدام l بردار ویژه از بین d بردار ویژه بایستی انتخاب شوند؟
چون ماتریس $\mathrm{\mathbf{S}_B}$ رتبه c-1 دارد (c تعداد کلاسهاست) $\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ نیز دارای رتبه c-1 است بنابراین $\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ دارای c-1 مقدار ویژه غیر صفر است.

حالت اول: اگر $l=c$-1 انتخاب شود ماتریس $\mathbf{C}$ را که ستونهای آن شامل c-1 بردار ویژه با نرم واحد ماتریس $\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ است تشکیل میدهیم. سپس بردار تبدیل یافته $\hat{\mathbf{y}}=\mathbf{C}^t\mathbf{x}$ را تشکیل میدهیم. این حداکثر مقدار J_3 را تضمین میکند. یعنی در تبدیل d بعدی به c-1 بعدی در قدرت تفکیک پذیری کاهشی رخ نمیدهد. چون اثر یک ماتریس برابر مجموع مقادیر ویژه آن ماتریس است داریم:

$$J_{3,\mathbf{x}} = tr\left\{\mathbf{S}_W^{-1}\mathbf{S}_B\right\} = \lambda_1 + \ldots + \lambda_{c-1} + 0$$

$$J_{3,\hat{\mathbf{y}}} = tr\left\{\left(\mathbf{C}^t\mathbf{S}_W\mathbf{C}\right)^{-1}\left(\mathbf{C}^t\mathbf{S}_B\mathbf{C}\right)\right\}, \quad \mathbf{C}^t\mathbf{S}_B\mathbf{C} = \mathbf{C}^t\mathbf{S}_W\mathbf{C}\mathbf{D}$$

$$\Rightarrow J_{3,\hat{\mathbf{y}}} = tr\left\{\mathbf{D}\right\} = \lambda_1 + \ldots + \lambda_{c-1} = J_{3,\mathbf{x}}$$

تبدیل خطی $\mathbf{C}^t\mathbf{x}$ که *c-1 مولفه* $\hat{\mathbf{y}}$ را محاسبه میکند یک قاعده خطی اپتیمال است که *c-1 تابع متمایز کننده ارائه میدهد (بهینه گی نسبت به معیار* J_3 میباشد). در حالت خاص دوکلاسه *c=2* است و تنها یک مقدار ویژه غیر صفر دارد. پس

$$\hat{\mathbf{y}} = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^t \mathbf{S}_W^{-1}\mathbf{x}$$

که همان متمایز کننده خطی فیشر است.

حالت دوم: اگر *l<c-1* انتخاب شود ماتریس $\mathbf{C}$ شامل l بردار ویژه ماتریس متناظر با l *بزرگترین مقدار* ویژه تشکیل میدهیم. البته این حداکثر مقدار J_3 را حاصل میکند. ولی $J_{3,\hat{\mathbf{y}}}<J_{3,\mathbf{x}}$ است.

Other Scatter Matrices Criteria

- The sum of squared error is defined as

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

- The trace (sum of diagonal elements) is the simplest scalar measure of the scatter matrix, as it is proportional to the sum of the variances in the coordinate directions

$$tr\left[S_W\right] = \sum_{i=1}^{c} tr\left[S_i\right] = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 = J_e$$

$$J_d = |\mathbf{S}_W| = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|.$$

$$tr\left[S_B\right] = \sum_{i=1}^{c} n_i \left\|\mathbf{m}_i - \mathbf{m}\right\|^2$$

$$\frac{|\mathbf{S}_W|}{|\mathbf{S}_T|} = \prod_{i=1}^{d} \frac{1}{1+\lambda_i}.$$

$$tr\mathbf{S}_W^{-1}\mathbf{S}_B = \sum_{i=1}^{d} \lambda_i.$$

$$J_f = tr\mathbf{S}_T^{-1}\mathbf{S}_W = \sum_{i=1}^{d} \frac{1}{1+\lambda_i}$$