THE UNIVERSITY OF MINNESOTA

Statistics 5401 November 28, 2005

Notes on Classification

The Classification Problem

Suppose you observe a p-dimensional random vector $\mathbf{x}$ whose components represent measurements on an individual known to belong to one of g distinct groups or populations π_1, π_2, ..., π_g, but you don't know for certain to which population the individual belongs. How do you use $\mathbf{x}$, together with what you know about the populations, to *guess* which population $\mathbf{x}$ comes from as accurately as possible? This is the essence of the *classification problem*.

You would probably consider a procedure for guessing - a *classification procedure* - to be good when the probability of making a mistake, that is classifying an individual in the wrong population, is small. When some mistakes are worse than others, you would want the probability of making expensive mistakes to be small.

Diagnosis

Each population π_i consists of individuals with a particular health condition from a list of specified conditions, perhaps including "no health problem," and the elements of $\mathbf{x}$ represent items in the patient's medical history and/or the patient's results from medical diagnostic procedures. The classification problem is to *diagnose* the health condition on the basis of these data. To misdiagnose someone with tuberculosis as having a cold is probably a more serious mistake than misdiagnosing someone with hay fever as having a cold.

Prediction

Population π_1 might consist of individuals who will declare bankruptcy in the next 12 months, and π_2 consist of individuals that will not declare bankruptcy. Here g = 2. The components of $\mathbf{x}$ might be items in the individuals credit report as well as demographic information. The classification problem is the problem of *predicting* whether or not a particular individual will declare bankruptcy in the next 12 months.

Identification

The populations are different varieties of a particular type of plant, and the components of $\mathbf{x}$ are measurements on various characteristics such as petal length and width. The classification problem is now a problem of *identification*.

Prior Probabilities and Misclassification Probabilities

In some cases you may reasonably know or be able to estimate *prior probabilities*, $p_i = P(\pi_i)$. p_i is the probability that an individual belongs to population π_i, *prior to observing* $\mathbf{x}$ or when you are ignorant of the value of $\mathbf{x}$. Since we are assuming that $\pi_1, \pi_2, \ldots \pi_g$ is an exhaustive set of possibilities, $\sum_{1 \le i \le g} p_i = 1$.

For medical *diagnosis*, p_i would specify how prevalent medical condition i is, that is, the probability that a randomly selected patient has medical condition i.

For *predicting* bankruptcy, $p_1 = 1 - p_2$ would be the probability that a randomly selected individual will declare bankruptcy in the next 12 months. Almost certainly p_1 would depend on overall economic conditions.

For *identifying* plant varieties, p_i = proportion of plants of variety i out of all the plants of that type. More realistically, p_i might also reflect the difficulty in finding a specimen of variety variety i.

By an application of Bayes' theorem, once $\mathbf{x}$ has been observed, the *posterior probability* that it was derived from population π_i is

$$P(\pi_i \mid \mathbf{x}) = p_i f_i(\mathbf{x}) / \{\sum_{1 \le j \le g} p_j f_j(\mathbf{x})\},$$

where $f_i(\mathbf{x})$ is the density of $\mathbf{x}$ for an individual in population π_i. The denominator is what is needed to make $\sum_{1 \le j \le g} P(\pi_i \mid \mathbf{x}) = 1$.

When you know $P(\pi_i \mid \mathbf{x}) \approx 1$, you can be practically certain that $\mathbf{x}$ came from π_i. When $P(\pi_i \mid \mathbf{x}) \approx 0$, then it is very unlikely that $\mathbf{x}$ came from π_i. When $P(\pi_i \mid \mathbf{x})$ has an intermediate value, you would be quite uncertain as to whether it did or did not come from population $\mathbf{x}$.

There are many procedures you might use to classify $\mathbf{x}$. As a generic symbol for a classification procedure I use the notation $\hat{\pi}$, with $\hat{\pi}(\mathbf{x})$ signifying the population the procedure assigns when $\mathbf{x}$ is observed. This notation is based on the idea that the unknown identity of the population is analogous to an unknown *parameter* since, once it is known, the distribution of the observation is determined. When you observe $\mathbf{x}$ and select the population to classify the case into you are "*estimating*" this parameter. If the procedure selects π_i based on data $\mathbf{x}$, we write $\hat{\pi}(\mathbf{x}) = \pi_i$. The possible "values" for $\hat{\pi}(\mathbf{x})$ are $\pi_1, \pi_2, \ldots, \pi_g$.

We are not yet concerned with the problem of estimating a procedure on the basis of a sample of data, a so called *training sample*; we are looking at how to compare classification methods when we have *complete* information about the different populations, not just information about a sample. We assume we know the prior probabilities p_j and the distribution $f_j(\mathbf{x})$ of $\mathbf{X}$ in population j, j = 1, ..., g.

Classification probabilities

I use the notation

$$P(i \mid j) = P(\hat{\pi}(\mathbf{x}) = \pi_i \mid \pi_j),\ 1 \le i,j \le g$$

to represent the probability of classifying $\mathbf{x}$ as coming from population π_i when it actually comes from population π_j. A more complete notation would be $P_{\hat{\pi}}(i \mid j)$ because $P(i \mid j)$ depends on the particular classification rule $\hat{\pi}$. I assume that the classification rule always makes a definite choice so $\sum_{1 \le i \le g} P(i \mid j) = 1$ for every j.

In this notation, $P(j \mid j)$ is the probability of *correctly* classifying an individual from π_j and $1 - P(j \mid j) = \sum_{i \ne j} P(i \mid j)$ is the probability of *misclassifying* an individual from π_j.

In diagnosis situations, $P(j \mid j)$ is the probability of making a <u>correct</u> diagnosis of someone with medical condition j, and $1 - P(j \mid j)$ is the probability of making an <u>incorrect</u> diagnosis.

We can display the classification probabilities $P(i \mid j)$ in a g by g table:

Pop.	Prior P	π_1	π_2	π_3	. . .	π_g
π_1	p_1	<u>$P(1 \mid 1)$</u>	$P(2 \mid 1)$	$P(3 \mid 1)$	. . .	$P(g \mid 1)$
π_2	p_2	$P(1 \mid 2)$	<u>$P(2 \mid 2)$</u>	$P(3 \mid 2)$	. . .	$P(g \mid 2)$
π_3	p_3	$P(1 \mid 3)$	$P(2 \mid 3)$	<u>$P(3 \mid 3)$</u>	. . .	$P(g \mid 3)$
. . .	. . .	. . .	. . .	. . .	. . .	. . .
π_8	p_g	$P(1 \mid g)$	$P(2 \mid g)$	$P(3 \mid g)$	. . .	<u>$P(g \mid g)$</u>

The diagonal elements (<u>underlined</u>) of the table are the probabilities of *correct* classification and the off-diagonal elements, $P(i \mid j)$, $i \ne j$, are probabilities of *incorrect* classification or of errors. There is obviously no reason to assume the table is symmetric.

A classification rule might not actually make use of **x**. For example, when $p_1 >> p_j$, $j \neq 1$, that is, the prior probability of π_1 is much greater than the prior probability of any other population, a defensible rule might be to *ignore* **x** and *always* classify any case as belonging to π_1. Then $P(1 \mid 1) = P(1 \mid 2) \ldots = P(1 \mid g) = 1$, and $P(j \mid k) = 0$, $j \neq 1$. This would probably not be a good approach if there was a high cost associated with failing to recognize a rare population π_j, $j \neq 1$.

When all mistakes are equally bad, one reasonable way to evaluate a classification rule $\hat{\pi}$ is its

TPM = *Total Probability of Misclassification* = P(wrong selection).

Explicitly, TPM is defined as

$$\begin{aligned} TPM = TPM(\hat{\pi}) &\equiv \sum_{1 \le i \le g} p_i \left\{ \sum_{j \neq i} P(j \mid i) \right\} = \sum_{1 \le i \le g} p_i \left\{ 1 - P(i \mid i) \right\} \\ &= 1 - \sum_{1 \le i \le g} p_i P(i \mid i) \end{aligned}$$

This weights each $1 - P(i \mid i)$, the probability of misclassifying an individual from π_i, by the prior probability p_i of π_i. TPM is the probability that an individual that is randomly selected from a population chosen with probability p_1 would be misclassified. The notation $TPM(\hat{\pi})$ emphasizes that TPM is a characteristic of the rule $\hat{\pi}$. Different classification rules will generally have different TPM values. For the example, in the preceding paragraph, the rule that always classifies an individual as π_1 has $TPM = 1 - p_1$.

Costs of Misclassification

Earlier I recognized the possibility that some mistakes might be worse than others. You can formalize this idea by supposing that there are specific *costs* associated with misclassifying an individual. Such costs normally will depend on both the true population π_j that **x** comes from and the guessed population $\hat{\pi}(\mathbf{x})$. For instance, the cost of misclassifying a poisonous mushroom as being edible is probably greater than the cost of misclassifying an edible mushroom as poisonous because the former misclassification can result in someone's injury or death.

Let $C(j \mid i)$ represent the cost incurred when $\hat{\pi}(\mathbf{x}) = \pi_j$ when in fact the **x** comes from π_i. We will see below that we can assume, without any loss of generality, that $C(i \mid i)$, the cost of correct classification, is zero. However, for the moment, I don't make that assumption. Indeed, it is probably reasonable to assume $C(i \mid i) \le 0$ (a negative "cost" is a "benefit"). You can

display values of $C(j\,|\,i)$ in a g by g table similar to that for $P(j\,|\,i)$.

Given that **x** actually comes from π_i, the expected cost of applying rule $\hat{\pi}$ is $EC(i) \equiv \sum_{1\le j\le g} P(j\,|\,i)C(j\,|\,i)$. Averaging EC(i) over the populations weighting by prior probabilities p_i, the overall *expected cost* EC of $\hat{\pi}$ will be

$$EC = EC(\hat{\pi}) = \sum_{1\le i\le g} p_i EC(i) = \sum_{1\le i\le g} p_i \{\sum_{1\le j\le g} P(j\,|\,i)C(j\,|\,i)\}.$$

The expected cost EC(i) involved in classifying an individual from population π_i is weighted by the prior probability p_i of π_i. Because $\sum_{1\le j\le g} P(j\,|\,i) = 1$, $P(i\,|\,i) = 1 - \sum_{j\ne i} P(j\,|\,i)$ and a little algebraic manipulation yields

$$\begin{aligned} EC &= \sum_i p_i \left\{ C(i\,|\,i) + \sum_{j\ne i} P(j\,|\,i)(C(j\,|\,i) - C(i\,|\,i)) \right\} \\ &= \sum_i p_i C(i\,|\,i) + \sum_i p_i \left\{ \sum_{j\ne i} P(j\,|\,i)\tilde{C}(j\,|\,i) \right\}, \end{aligned}$$

where

$$\tilde{C}(j\,|\,i) \equiv C(j\,|\,i) - C(i\,|\,i)$$

is the *penalty* for misclassifying as π_j when π_i is correct.

The second term in the expression for EC is the *expected penalty* of misclassification. Because the first term, $\sum_i p_i C(i\,|\,i)$, does not depend on the classification rule $\hat{\pi}$ used, when you select $\hat{\pi}$ to minimize the expected *penalty* of misclassification you also minimize the expected *cost* (EC) and vice versa. But the expected penalty has the same form as EC when $C(i\,|\,i) = 0$, $i = 1,\ldots,g$, with $\tilde{C}(j\,|\,i)$ replacing $C(j\,|\,i)$. This is the basis of the claim that there is no harm in assuming that all $C(i\,|\,i) = 0$, that is, that there is no cost, positive or negative, incurred in making a correct classification. With this assumption $C(j\,|\,i) = \tilde{C}(j\,|\,i)$ and the expected cost is the ECM = *Expected Cost of Misclassification*

$$ECM(\hat{\pi}) = ECM = \sum_i p_i \sum_{j\ne i} P(j\,|\,i)C(j\,|\,i).$$

For the simple example of always classifying as π_1, $ECM = \sum_{i\ne 1} p_i C(1\,|\,i)$. Even when all p_i are small, $i \ne 1$, when $C(1\,|\,i)$, the cost of misclassifying a member of π_i as being a member of π_1, is very large, ECM may be unacceptably high.

You can formalize the situation when all mistakes are equally bad by fixing all $C(i\,|\,j)$ to be the same, say, $C(i\,|\,j) = c$, $i \ne j$. Then $ECM(\hat{\pi}) = c \times TPM(\hat{\pi})$.

Comparing Classification Rules

When there are identifiable costs of misclassification, it seems reasonable to prefer $\hat{\pi}_a$ to $\hat{\pi}_b$ when $ECM(\hat{\pi}_a) < ECM(\hat{\pi}_b)$, that is, when the expected cost of

using $\hat{\pi}_a$ is less than the expected cost of using $\hat{\pi}_b$. From this point of view, the "best" possible rule would be one with the *lowest possible* ECM.

When costs are equal or you cannot reasonably specify them, you would prefer $\hat{\pi}_a$ to $\hat{\pi}_b$ when $TPM(\hat{\pi}_a) < TPM(\hat{\pi}_b)$, that is, when the probability of a classification error when using $\hat{\pi}_a$ is less than the probability of a classification error when using $\hat{\pi}_b$. The "best" rule would be one that has the *smallest possible* TPM. Since TPM = ECM when $C(j \mid i) = 1$, $j \neq i$, you can use a general method for determining the minimum ECM rule to determine the minimum TPM rule

Two group case

The simplest situation is when $g = 2$.

Let $\lambda(\mathbf{x}) \equiv f_1(\mathbf{x})/f_2(\mathbf{x})$ be the *likelihood ratio*. By a variant of the Neyman-Pearson lemma it can be demonstrated that the minimum ECM rule uses only the value of $\lambda(\mathbf{x})$ in selecting a population. Large values of $\lambda(\mathbf{x})$ classify a case as π_1 and small values as π_2. The dividing value is of the minimum ECM rule is $(p_2/p_1)\{C(1 \mid 2)/C(2 \mid 1)\}$. Specifically the rule is,

$$\text{When } \lambda(\mathbf{x}) \geq (p_2/p_1)\{C(1 \mid 2)/C(2 \mid 1)\} \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$

$$\text{When } \lambda(\mathbf{x}) < (p_2/p_1)\{C(1 \mid 2)/C(2 \mid 1)\} \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2.$$

The ratio $P(\pi_1 \mid \mathbf{x})/P(\pi_2 \mid \mathbf{x})$ of posterior probabilities is $p_1 f_1(\mathbf{x})/p_2 f_2(\mathbf{x}) = (p_1/p_2)\lambda(\mathbf{x})$. Therefore, you can also state the minimum ECM rule as:

$$\text{When } P(\pi_1 \mid \mathbf{x})/P(\pi_2 \mid \mathbf{x}) \geq C(1 \mid 2)/C(2 \mid 1) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$

$$\text{When } P(\pi_1 \mid \mathbf{x})/P(\pi_2 \mid \mathbf{x}) < C(1 \mid 2)/C(2 \mid 1) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2.$$

In words, when the *posterior odds* that π_1 is correct exceed the ratio of misclassification costs, classify in π_1; otherwise, classify in π_2.

Another equivalent statement of the rule is

$$\text{When } P(\pi_1 \mid \mathbf{x})C(2 \mid 1) \geq P(\pi_2 \mid \mathbf{x})C(1 \mid 2) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$

$$\text{When } P(\pi_2 \mid \mathbf{x})C(1 \mid 2) > P(\pi_1 \mid \mathbf{x})C(2 \mid 1) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2.$$

Since $C(1 \mid 1) = C(2 \mid 2) = 0$,

$$P(\pi_1 \mid \mathbf{x})C(2 \mid 1) = P(\pi_1 \mid \mathbf{x})C(2 \mid 1) + P(\pi_2 \mid \mathbf{x})C(2 \mid 2)$$
$$= \text{the } \textit{posterior expected cost} \text{ of selecting } \pi_2.$$

Similarly $P(\pi_2 \mid \mathbf{x})C(1 \mid 2)$ = posterior expected cost of selecting π_1. Thus the

minimum ECM rule can be summarized as "select the population with the minimum posterior expected misclassification cost."

By "posterior expected cost" of a population I mean the expected cost of choosing that population once you know $\mathbf{x}$, the expectation being computed conditional on the value of $\mathbf{x}$. When the value of $\mathbf{x}$ makes it almost certain that π_1 is the correct choice, that is, $P(\pi_2 \mid \mathbf{x}) \approx 0$, the expected cost is near zero. When it is a 50-50 proposition, that is, $P(\pi_2 \mid \mathbf{x}) \approx .5$, the expected cost $\approx .5 \times C(1 \mid 2)$.

When the misclassification costs are all equal, the minimum ECM rule is also the minimum TPM rule and is

$$\text{When } P(\pi_1 \mid \mathbf{x})/P(\pi_2 \mid \mathbf{x}) \geq 1 \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$
$$\text{When } P(\pi_1 \mid \mathbf{x})/P(\pi_2 \mid \mathbf{x}) < 1 \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2.$$

In words, this says, "select π_1 when the posterior odds favor π_1 and select π_2 when the posterior odds favor π_2." Another way to state this rule is

$$\text{When } P(\pi_1 \mid \mathbf{x}) \geq P(\pi_2 \mid \mathbf{x}) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$
$$\text{When } P(\pi_2 \mid \mathbf{x}) > P(\pi_1 \mid \mathbf{x}) \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2.$$

In words, this is, "select the population with the higher posterior probability conditional on $\mathbf{x}$."

More than two groups

When $g > 2$, these optimal classification rules generalize nicely. The minimum ECM rule is "select the population with the minimum posterior expected cost," that is, $\hat{\pi}(\mathbf{x}) = \pi_k$ where the minimum value of

$$\sum_{j \neq i} P(\pi_j \mid \mathbf{x}) C(i \mid j) = \left\{ \sum_{j \neq i} p_j f_j(\mathbf{x}) C(i \mid j) \right\} / \left\{ \sum_j p_j f_j(\mathbf{x}) \right\}, \; i = 1, \ldots, g.$$

occurs when $i = k$.

Since, for given $\mathbf{x}$, the quantity in the denominator, $\sum_j p_j f_j(\mathbf{x})$, is constant, the minimum ECM rule can also be stated as "select the population with the smallest value of $\sum_{j \neq i} p_j f_j(\mathbf{x}) C(i \mid j)$."

For the minimum TPM rule (that is, when you assume the costs of misclassification are equal), this rule becomes "select the population with the smallest value of $\sum_{j \neq i} P(\pi_j \mid \mathbf{x}) = 1 - P(\pi_i \mid \mathbf{x})$"; equivalently, "select π_k with the largest posterior probability $P(\pi_k \mid \mathbf{x})$ conditional on $\mathbf{x}$."

As when $g = 2$, you can ignore $\sum_j p_j f_j(\mathbf{x})$ in the denominator This leads to the rule "select the population with the largest value of $p_i f_i(\mathbf{x})$" or, equivalently,

"select the population with the largest value of $\log(p_i f_i(\mathbf{x})) = \log(p_i) + \log(f_i(\mathbf{x}))$."

Equal Variance Multivariate Normal Case - Linear Classification

We illustrate these rules for the situation where the distribution associated with population π_i is $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, that is, $\mathbf{x}$ has density

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}\{\det(\boldsymbol{\Sigma}_i)\}^{-1/2}\exp\{-(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu})/2\}$$

and log density

$$\log(f_i(\mathbf{x})) = -(p/2)\log(2\pi) - \log(\det(\boldsymbol{\Sigma}_i))/2 - (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu})/2$$

The notation exp{. . .} means $e^{(\ldots)}$.

First make the simplifying assumption that all populations have the same variance matrix, that is $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$.

When g = 2, the log likelihood ratio is

$$\log(f_1(\mathbf{x})/f_2(\mathbf{x})) = \log(f_1(\mathbf{x})) - \log(f_2(\mathbf{x}))$$

$$= -\left\{(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)/2 - (\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)/2\right\}$$

$$= \boldsymbol{\ell}'(\mathbf{x} - (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2), \text{ where } \boldsymbol{\ell} \equiv \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

The minimum ECM rule selects π_1 when

$$\boldsymbol{\ell}'(\mathbf{x} - (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2) \geq \log(p_2/p_1) + \log\{C(1\,|\,2)/C(2\,|\,1)\}$$

and selects π_2 otherwise. Note that the right hand side (the "cutpoint") is a combination of log ratios of prior probabilities and misclassification costs.

Equivalently, the rule selects π_1 when

$$\boldsymbol{\ell}'\mathbf{x} \geq \boldsymbol{\ell}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 + \log(p_2/p_1) + \log\{C(1\,|\,2)/C(2\,|\,1)\}$$
$$= m + \log(p_2/p_1) + \log\{C(1\,|\,2)/C(2\,|\,1)\},\ m \equiv \boldsymbol{\ell}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2,$$

and selects π_2 otherwise. The left side, $\boldsymbol{\ell}'\mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}$, is Fisher's *linear discriminant function*. The right hand side is a constant *threshold* or cut-off value that separates values of $\boldsymbol{\ell}'\mathbf{x}$ favoring π_1 from those favoring π_2.

The more the prior *odds ratio* p_2/p_1 *favors* π_2 or the more the error *cost ratio* $C(1\,|\,2)/C(2\,|\,1)$ *disadvantages* π_1, the stronger is the evidence provided by $\boldsymbol{\ell}'\mathbf{x}$, required to select π_1. When $p_1 = p_2$ and $C(1\,|\,2) = C(2\,|\,1)$, the threshold

is m, the value of $\boldsymbol{\ell}'\mathbf{x}$ when $\mathbf{x}$ lies halfway between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

The posterior probability of π_i based on $\mathbf{x}$ is (because the factor $(2\pi)^{-p/2}\{\det \boldsymbol{\Sigma}\}^{-1/2}$ cancels out)

$$P(\pi_i \mid \mathbf{x}) = \frac{p_i \exp\left\{-(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)/2\right\}}{p_1\exp\left\{-(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)/2\right\} + p_2\exp\left\{-(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)/2\right\}}$$

You can simplify each exponential:

$$\exp\{-(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)/2\} = \exp\{-\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}/2\}\times\exp\{\boldsymbol{\ell}_i'\mathbf{x}\}\times\exp\{-c_i\},$$

where $\boldsymbol{\ell}_i \equiv \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ and $c_i \equiv \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i/2 = \boldsymbol{\ell}_i'\boldsymbol{\mu}_i/2$. When you substitute this in the expression for $P(\pi_i \mid \mathbf{x})$, you can cancel $\exp(-\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}/2)$ to get

$$P(\pi_i \mid \mathbf{x}) = p_i\exp(\boldsymbol{\ell}_i'\mathbf{x}-c_i)/\{p_1\exp(\boldsymbol{\ell}_1'\mathbf{x} - c_1) + p_2\exp(\boldsymbol{\ell}_2'\mathbf{x} - c_2)\},\ i = 1, 2.$$

Moreover, if you replace $\boldsymbol{\ell}_i'\mathbf{x}-c_i$ by $\boldsymbol{\ell}_i'\mathbf{x}-c_i - K(\mathbf{x})$, i = 1,2, where $K(\mathbf{x})$ may depend on $\mathbf{x}$ but not on i, you multiply both numerator and denominator by $\exp(-K(\mathbf{x}))$ leaving the ratio unchanged.

The minimum TPM classification rule which says "select the population with the higher $P(\pi_i \mid \mathbf{x})$," becomes, "select the population with the larger $p_i\exp(\boldsymbol{\ell}_i'\mathbf{x}-c_i)$," that is the rule

$$\text{When } \boldsymbol{\ell}_1'\mathbf{x} - c_1 + \log(p_1) \geq \boldsymbol{\ell}_2'\mathbf{x} - c_2 + \log(p_2) \text{ then } \pi_1$$
$$\text{When } \boldsymbol{\ell}_1'\mathbf{x} - c_1 + \log(p_1) < \boldsymbol{\ell}_2'\mathbf{x} - c_2 + \log(p_2) \text{ then } \pi_2$$

When $p_1 = p_2 = 1/2$, this specifies choosing the population with the larger $\boldsymbol{\ell}_i'\mathbf{x} - c_i$. Because the quantities being compared are linear combinations of the elements of $\mathbf{x}$, the rule is a *linear classification* rule.

You can assess the strength of the evidence in favor of π_1 or π_2 by computing the posterior probabilities $P(\pi_i \mid \mathbf{x})$, i = 1, 2. When $P(\pi_1 \mid \mathbf{x})$ is close to 1, you can confidently classify $\mathbf{x}$ as coming from π_1. On the other hand, when $P(\pi_1 \mid \mathbf{x})$ is not near 0 or 1, you should be in considerable doubt as to the correctness of classification.

When g > 2 and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, you can still do minimum TPM classification using linear functions of $\mathbf{x}$. Exactly as in the preceding paragraph, the posterior probabilities are

$$P(\pi_i \mid \mathbf{x}) = \frac{p_i \exp\{\boldsymbol{\ell}_i'\mathbf{x} - c_i - K(\mathbf{x})\}}{\sum_{1 \le j \le g} p_j \exp\{\boldsymbol{\ell}_j'\mathbf{x} - c_j - K(\mathbf{x})\}}$$

where $\boldsymbol{\ell}_i \equiv \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ and $c_i \equiv \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i/2 = \boldsymbol{\ell}_i'\boldsymbol{\mu}_i/2$, $i = 1,\ldots,g$, and $K(\mathbf{x})$ is an arbitrary function of $\mathbf{x}$ that is the same for all i.

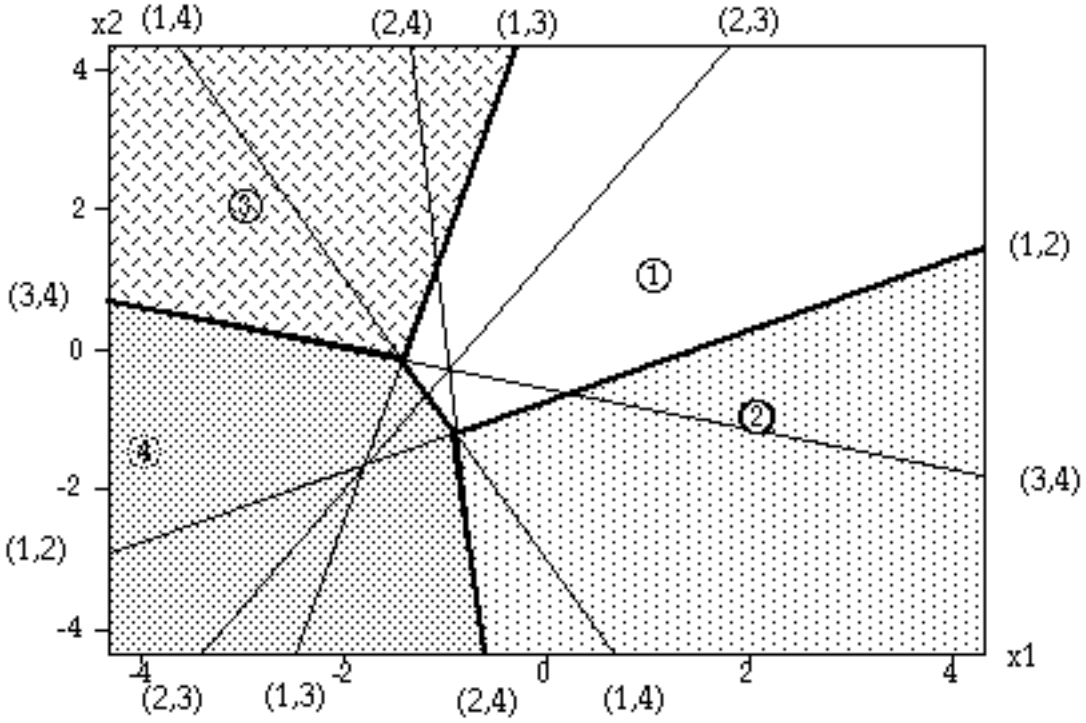
When any $\boldsymbol{\ell}_i'\mathbf{x} - c_i$ is large you may need some $K(\mathbf{x})$ so as to make $\exp\{\boldsymbol{\ell}_j'\mathbf{x} - c_j - K(\mathbf{x})\}$ representable in a computer. $K(\mathbf{x}) = \max_i\{\boldsymbol{\ell}_i'\mathbf{x} - c_i\}$ is often a good choice, because all

$$\exp\{\boldsymbol{\ell}_j'\mathbf{x} - c_j - K(\mathbf{x})\} = \exp\{\boldsymbol{\ell}_i'\mathbf{x} - c_i\}/\max_j(\exp\{\boldsymbol{\ell}_j'\mathbf{x} - c_j\}) \le 1.$$

For minimum TPM classification, you compute the g posterior probabilities $P(\pi_i \mid \mathbf{x})$ and select the population with largest $P(\pi_i \mid \mathbf{x})$. Alternatively, you can select the population with the largest $\boldsymbol{\ell}_i'\mathbf{x} - c_i + \log(p_i)$. That is, classification amounts to comparing the values of g linear functions. As always, $P(\pi_i \mid \mathbf{x})$ measures the strength of the evidence in favor of π_i, $i = 1,\ldots,g$.

In geometrical terms, the minimum TPM rule divides up p-dimensional space into g regions separated by the (p-1)-dimensional planes. Each region consists of all the possible values of $\mathbf{x}$ that would be classified into a single population.

Here is an example when $p = 2$ and $g = 4$, assuming $p_1 = p_2 = p_3 = p_4$ and $\Sigma = I_2$.



Classification regions when $g = 4$, $p = 2$

The circled numbers indicate the four population means ($\mu_1 = [1,1]'$, $\mu_2 = [2,-1]'$, $\mu_3 = [-3, 2]'$, $\mu_4 = [-4,-1.5]'$). This gives $\ell_i = \Sigma^{-1}\mu_i = \mu_i$ and $(c_1, c_2, c_3, c_4) = (1, 2.5, 6.5, 9.125)$. A line labeled (i, j) separates the plane into a part where π_i is preferred and a part where π_j is preferred. Its equation is $(\ell_i - \ell_j)'x = c_i - c_j$. The area where a population is preferred has a boundary (heavy lines) made up of straight lines. The is a feature of linear classification.

When $p_1 = p_2 = \ldots = p_g$, the minimum TPM rule selects the population with the largest value of $\exp\{-(x-\mu_i)'\Sigma^{-1}(x-\mu_i)/2\}$, or the smallest value of $(x-\mu_i)'\Sigma^{-1}(x-\mu_i)$. But the latter is the Mahalanobis or generalized distance between x and μ_i. Thus the minimum TPM rule selects the "nearest" population when distance is the Mahalanobis distance to the mean.

Normal Unequal Variance Case - Quadratic Classification

When the variance matrices differ, the situation is more complicated.

Start with the case when $g = 2$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Then

$$\log(f_1(\mathbf{x})) - \log(f_2(\mathbf{x})) = -\log\left\{\det(\boldsymbol{\Sigma}_1)/\det(\boldsymbol{\Sigma}_2)\right\}/2 - \left\{(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}/2$$

$$= \log(\det(\boldsymbol{\Sigma}_2))/2 - \log(\det(\boldsymbol{\Sigma}_1))/2 + (q_1(\mathbf{x}) + \boldsymbol{\ell}_1'\mathbf{x}) - (q_2(\mathbf{x}) + \boldsymbol{\ell}_2'\mathbf{x}) + c_2 - c_1$$

where

$$q_i(\mathbf{x}) = -\mathbf{x}'\boldsymbol{\Sigma}_i^{-1}\mathbf{x}/2,\ \boldsymbol{\ell}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i,\ c_i = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i/2 = \boldsymbol{\ell}_i'\boldsymbol{\mu}_i/2,\ i = 1, 2$$

Since $q_i(\mathbf{x})$ involves squares and products of the elements of $\mathbf{x}$, $q_i(\mathbf{x}) + \boldsymbol{\ell}_i'\mathbf{x}$ is a *quadratic* function of $\mathbf{x}$ rather than a linear function like $\boldsymbol{\ell}_i'\mathbf{x}$.

The minimum ECM rule is now

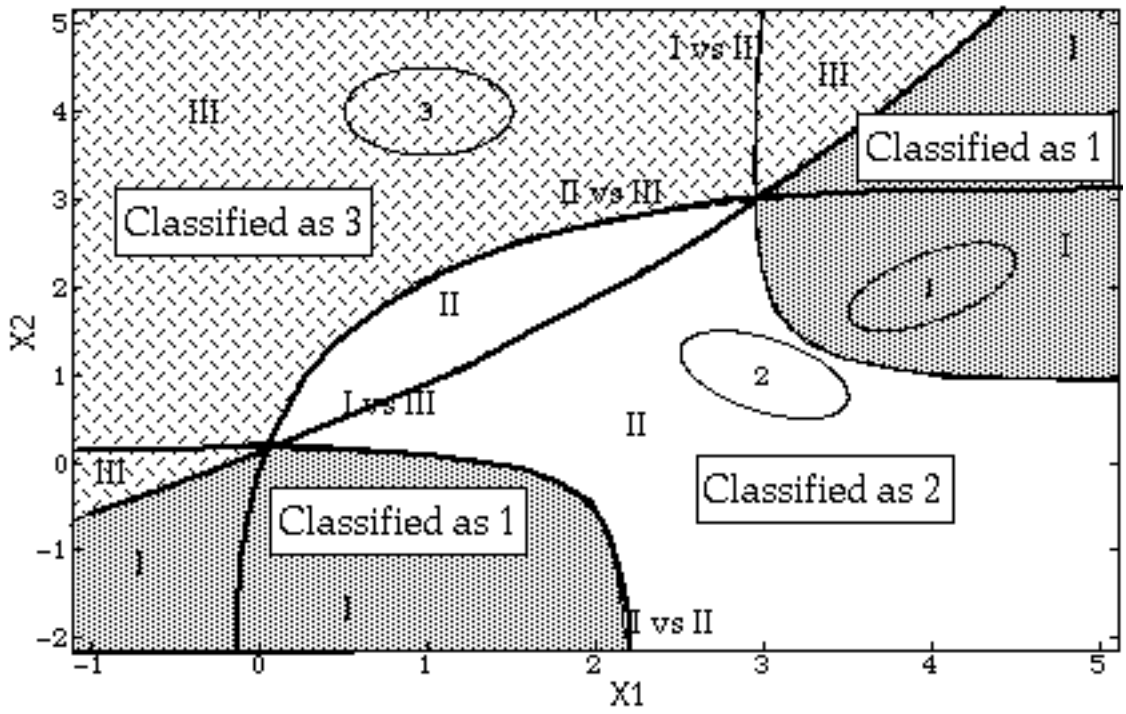
$$\text{When } d_1^Q(\mathbf{x}) - d_2^Q(\mathbf{x}) \geq \log\left\{C(1\,|\,2)/C(2\,|\,1)\right\} \text{ then } \hat{\pi}(\mathbf{x}) = \pi_1$$

$$\text{When } d_1^Q(\mathbf{x}) - d_2^Q(\mathbf{x}) < \log\left\{C(1\,|\,2)/C(2\,|\,1)\right\} \text{ then } \hat{\pi}(\mathbf{x}) = \pi_2,$$

where $d_i^Q(\mathbf{x}) \equiv -\log(\det(\boldsymbol{\Sigma}_i))/2 + q_i(\mathbf{x}) + \boldsymbol{\ell}_i'\mathbf{x} - c_i + \log(p_i)$, a *quadratic* function of $\mathbf{x}$ rather than a linear function. This implies that the surface in p-dimensional space (line when $p = 2$) that separates values of $\mathbf{x}$ that are favorable to π_1 from values that are favorable to π_2 is *curved* rather than flat or straight. In fact, the surface can even be a closed surface like an ellipsoid, with the inside, say, favoring π_1 and the outside favoring π_2.

When $g > 2$, the minimum TPM rule again depends on $d_i^Q(\mathbf{x})$, $i = 1,\ldots,g$. The rule is "select π_i such that $d_i^Q(\mathbf{x})$ is smallest."

Here is an example of the shape of classifying regions with quadratic discrimination. There are three populations with means $\mu_1 = [4,2]$, $\mu_2 = [3,1]$, and $\mu_3 = [1,4]$, correlations $\rho_1 = .6$, $\rho_2 = 0$ and $\rho_3 = 0$, with all variances the same.



The ellipses indicate the shapes and orientations of the countours of bivariate normal populaions. The dark lines are boundaries of equal preference between two groups. Note that the area in which population 1 is preferred consists of two pieces. The same is true for the other populations, although only one piece shows in the plot.

Classification when the parameters are not known

The discussion so far has been in terms of the *true* distributions or the *population* means and variance matrices. In practice, you never have such complete information. You have to estimate the distributions from "training samples" selected from the g populations.

The difference between a training sample and a "target sample", that is a sample whose elements you might want to classify, is that in a training sample you know which population or group each observed **x** comes from, while you don't know that for the target sample. For the types of problems discussed earlier, training samples might derived from past records of patients with a definitive diagnosis, from the actual bankruptcy history of a sample of individuals in the recent past, or from samples of plant specimens

that have been classified by experts.

In the most general case, when you can't make any assumptions about the distributions, you need methods of multivariate density estimation that directly estimate $f_j(\mathbf{x})$. Since that is beyond the scope of this course, I will look only at the multivariate normal case when you can assume $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$.

We have seen that in this case, optimal discrimination rules are linear with the rule depending on $\boldsymbol{\mu}_i$, i = 1,...,g, and $\mathbf{\Sigma}$ (and prior probabilities and costs). It seems natural use the "plug in" classification rule. That is, the rule found by substituting for $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$, the estimates $\hat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}_i}$, i = 1,...,g and $\hat{\mathbf{\Sigma}} = \mathbf{S}_{pooled} = \mathbf{S} = (N-g)^{-1}\sum_i (n_i-1)\mathbf{S}_i = (N-g)^{-1}\mathbf{E}$, where $\mathbf{E}$ is the within-group MANOVA error matrix. The resulting classification rules are almost certainly not optimal since they will differ from the rules based on the unknown true values of parameters.

An important problem is to estimate the TPM or the ECM for the rule. You might think that all you need to do is to apply the estimated classification rule to the training data and see how well it does, that is find the *apparent error rate*,

$$\text{APER} = (\text{number of } \mathbf{x}\text{'s misclassified by } \hat{\pi})/N$$

or the apparent cost. Unfortunately, this is almost always optimistic, in the sense that you can expect the actual error rate or cost incurred when you apply the rule to a *different* independent data set to be *larger* than when it is applied to the training set. This is because, to some degree, the estimation procedure customizes the classification rule to peculiarities of the training samples which will not be present in a different data set. It is, in fact, a hard problem to use the training sample to estimate the actual error rate of an empirically derived rule.

Fisher Discriminant Functions

When the populations are multivariate normal with equal variance matrices, the linear discriminant method given above requires the computation of g linear functions. Actually you can get by with the $f_h = g - 1$ linear functions $(\boldsymbol{\ell}_1 - \boldsymbol{\ell}_g)'\mathbf{x},\ldots, (\boldsymbol{\ell}_{g-1} - \boldsymbol{\ell}_g)'\mathbf{x}$. However, it may be possible to find a classification rule based on a smaller number of functions that allows almost as good classification as the minimum TPM rule. When $p < g-1$, you can in fact *exactly* express the minimum TPM rule in terms of only p linear functions. The approach is very similar to the determination of MANOVA

canonical variables.

Let the g populations be $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, ..., $N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$, with prior probabilities p_1, ..., p_g. Then $\overline{\boldsymbol{\mu}} = \sum_i p_i \boldsymbol{\mu}_i$ be the prior expectation of $\mathbf{x}$. For any linear function $y = \boldsymbol{\ell}'\mathbf{x}$ let $\mu_i(y) = E[y \mid \pi_i] = \boldsymbol{\ell}'\boldsymbol{\mu}_i$ be its mean and $\sigma^2(y) = V[y \mid \pi_i] = \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}$ be its variance when $\mathbf{x}$ is known to come from π_i. Then the prior expectation of y is $\overline{\mu}(y) = \sum_i p_i \mu_i(y) = \boldsymbol{\ell}'\overline{\boldsymbol{\mu}}$. We start by seeking a single linear function $y_1 = \boldsymbol{\ell}_1'\mathbf{x}$ which maximally separates the populations, in the sense that it has the largest possible *non-centrality* parameter

$$\delta^2(y) = \sum_i p_i \left\{ \mu_i(y) - \overline{\mu}(y) \right\}^2 / \sigma^2(y) = \{ \sum_i p_i \boldsymbol{\ell}'(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'\boldsymbol{\ell} \} / \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}$$
$$= \boldsymbol{\ell}'\mathbf{B}_0\boldsymbol{\ell} / \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}, \text{ where } \mathbf{B}_0 \equiv \sum_i p_i (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})'.$$

Note $\mathbf{B}_0$ differs from $\mathbf{B}_\mu$ in Johnson and Wichern (eq. 11-58 p. 629) which tacitly assumes equal prior probabilities. Even when $p_1 = p_2 = \ldots = p_g = g^{-1}$, $\mathbf{B}_0$ as defined here differs from Johnson and Wichern's $\mathbf{B}_\mu$ by a factor of 1/g.

Now $\boldsymbol{\ell}'\mathbf{B}_0\boldsymbol{\ell} / \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}$ is maximized by choosing $\boldsymbol{\ell} = \mathbf{e}_1$, where $\mathbf{e}_1$ is the eigenvector of $\mathbf{B}_0$ *relative* to $\boldsymbol{\Sigma}$ (eigenvector of $\boldsymbol{\Sigma}^{-1}\mathbf{B}_0$) corresponding to the largest relative eigenvalue λ_1. With the usual normalization, $\mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{e}_1 = 1$, the distribution of $y_1 = \mathbf{e}_1'\mathbf{x}$ when $\mathbf{x}$ comes from π_i is $N_1(\mathbf{e}_1'\boldsymbol{\mu}_i, 1^2)$. Hence the posterior probability of π_i based only on the value of y_1 is

$$P(\pi_1 \mid y_1) = p_i \exp\{-(y_1 - \mathbf{e}_1'\boldsymbol{\mu}_i)^2/2\} / (\sum_j p_j \exp\{-(y_1 - \mathbf{e}_1'\boldsymbol{\mu}_j)^2/2\}).$$

When $p_1 = \ldots = p_g$, this implies that maximum TPM classification based on only y_1 amounts to selecting the population π_i with the smallest $(y_1 - \mathbf{e}_1'\boldsymbol{\mu}_i)^2 = (\mathbf{e}_1'(\mathbf{x} - \boldsymbol{\mu}_i))^2$, the square of the distance from y_1 to $\mathbf{e}_1'\boldsymbol{\mu}_i$.

If you now seek a second linear function, y_2 of $\mathbf{x}$, uncorrelated with y_1, that maximally separates the populations, you get $y_2 \equiv \mathbf{e}_2'\mathbf{x}$, where $\mathbf{e}_2$ is the eigenvector of $\mathbf{B}_0$ relative to $\boldsymbol{\Sigma}$ corresponding to the second largest relative eigenvalue λ_2. When $\mathbf{x}$ comes from π_i, y_2 is $N(\mathbf{e}_2'\boldsymbol{\mu}_i, 1^2)$. Continuing, you can find $s \equiv \min(p, g-1)$ linear functions $y_i = \mathbf{e}_i'\mathbf{x}$ where $\mathbf{e}_i$ is the eigenvector of $\mathbf{B}_0$ relative to $\boldsymbol{\Sigma}$ which corresponds to the i^{th} largest relative eigenvalue λ_i.

The non-centrality parameters of these linear functions are $\delta^2(y_i) = \lambda_i$. If $p > g-1$, any additional linear functions $y_i = \mathbf{e}_i'\mathbf{x}$ uncorrelated with $y_1, \ldots, y_s$ have $\delta^2(y_i) = 0$ since $E(y_i \mid \pi_j) = \mathbf{e}_i'\boldsymbol{\mu}_j = \mathbf{e}_i'\overline{\boldsymbol{\mu}}$ does not depend on j. Hence, for $i > s$

= min(g-1,p), y_i is of no use in classifying $\mathbf{x}$.

Variables $y_1 = \mathbf{e}_1'\mathbf{x}$, ..., $y_s = \mathbf{e}_s'\mathbf{x}$ are the *Fisher* discriminant functions and are linear in $\mathbf{x}$ and contain all the linearly extracting information for classification. They can be thought of as "true" rather than estimated MANOVA canonical variables.

When g = 2 and s = 1, y_1 is proportional to $\boldsymbol{\ell}'\mathbf{x}$, where $\boldsymbol{\ell} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

The overall Mahalanobis distance of $\mathbf{x}$ from $\boldsymbol{\mu}_j$ is

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) &= \textstyle\sum_{1\le i\le s}(y_i - \mathbf{e}_i'\boldsymbol{\mu}_j)^2 + \sum_{s+1\le i\le p}(y_i - \mathbf{e}_i'\boldsymbol{\mu}_j)^2 \\ &= \textstyle\sum_{1\le i\le s}(y_i - \mathbf{e}_i'\boldsymbol{\mu}_j)^2 + \sum_{s+1\le i\le p}(y_i - \mathbf{e}_i'\overline{\boldsymbol{\mu}})^2 \\ &\equiv Q_{js}(\mathbf{x}) + R_s.\end{aligned}$$

The second term (R_s) is either missing ($g-1 \ge p$) or does not depend on j ($g-1 < p$) because $\mathbf{e}_i'\overline{\boldsymbol{\mu}} = \mathbf{0}$, $i > s$. Therefore the posterior probabilities given $\mathbf{x}$ are

$$P(\pi_j \mid x) = p_j\exp\{-Q_{js}(\mathbf{x})/2\}/\textstyle\sum_{1<i<g} p_i\exp\{-Q_{is}(\mathbf{x})\}.$$

Since $Q_{js}(\mathbf{x}) = \sum_{1\le i\le s}(y_i^2 - 2(\mathbf{e}_i'\boldsymbol{\mu}_j)\,\mathbf{e}_i'\mathbf{x} + 2c_{ij})$, where $c_{ij} = (\mathbf{e}_i'\boldsymbol{\mu}_j)^2/2$, you can express the posterior probabilities in terms of the s linear functions $\mathbf{e}_i'\mathbf{x}$, i = 1, ..., s. Minimum TPM classification amounts to selecting π_j with the smallest

$$\textstyle\sum_{1\le i\le s}\{(\boldsymbol{\mu}_j'\,\mathbf{e}_i)y_i - c_{ij}\} + \log(p_j).$$

Note that $\sum_{1\le i\le s}(y_i - \mathbf{e}_i'\boldsymbol{\mu}_j)^2$ is the squared Euclidean distance between the vector $\mathbf{y} = [y_1, \ldots, y_s]'$ and $E[\mathbf{y} \mid \pi_j] = [\mathbf{e}_1,\ldots,\mathbf{e}_s]'\boldsymbol{\mu}_j$.

If $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_r >> \lambda_{r+1} \ge \ldots \ge \lambda_s \ge 0$, that is, the first r relative eigenvalues are large when compared to the last s - r relative eigenvalues, then you can expect that classification based on the $r < s$ linear combinations $y_1, y_2, \ldots, y_r$ alone will do almost as well as classification based on all the y_i's which is equivalent to classification using all the x_i's.

The posterior probability of π_j using only $y_1, \ldots, y_r$ is

$$P(\pi_j \mid y_1,\ldots,y_r) = p_j \exp\{-Q_{jr}(\mathbf{x})/2\}/(\sum_i p_i \exp\{-Q_{ir}(\mathbf{x})\}),$$

where $Q_{jr}(\mathbf{x}) = \sum_{1\le i\le r}(y_i - \mathbf{e}_i'\boldsymbol{\mu}_j)^2 = \sum_{1\le i\le r}\left\{ y_i^2 - 2(\boldsymbol{\mu}_i'\mathbf{e}_j)\mathbf{e}_i'\mathbf{x} + 2c_{ij} \right\}$. $P(\pi_j \mid y_1,\ldots,y_r)$ on the data only through the r linear functions $y_i = \mathbf{e}_i'\mathbf{x}$, i = 1,...,r. Selecting the population π_j with the smallest value of $\sum_{1\le i\le r}\{(\boldsymbol{\mu}_j'\mathbf{e}_i)y_i - c_{ij}\} + \log(p_j)$ should be an "almost TPM" classification rule.

In practice, of course, you will need to estimate the unknown parameters. The sample version of $\mathbf{B}_0$ is

$$\hat{\mathbf{B}}_0 \equiv \sum_j p_j(\overline{\mathbf{x}}_j - \overline{\overline{\mathbf{x}}})(\overline{\mathbf{x}}_j - \overline{\overline{\mathbf{x}}})', \text{ where } \overline{\overline{\mathbf{x}}} \equiv \sum_j p_j \overline{\mathbf{x}}_j.$$

Note: Johnson and Wichern don't even try to estimate $\mathbf{B}_\mu$. Instead they use

$$\mathbf{B} = \mathbf{H} = \sum_j n_j(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})', \text{ where } \overline{\mathbf{x}} \equiv \sum_j n_j \overline{\mathbf{x}}_j/N,\ N = \sum_j n_j,$$

the among-groups hypothesis matrix from a one-way MANOVA. When $p_j \tilde{=} n_j/N$, $\mathbf{B}$ can be considered an estimate of $N\mathbf{B}_0$. When $n_1 = n_2 = \ldots = n_g = n$, $\mathbf{B}$ can be considered and estimate of $n\mathbf{B}_\mu$.

Let $\hat{\mathbf{e}}_1,\ldots,\hat{\mathbf{e}}_s$ be the eigenvectors of $\hat{\mathbf{B}}_0$ relative to $\mathbf{S}$, then the estimated Fisher discriminant functions are $\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x}$.

If you use the observed proportions $\hat{p}_j = n_j/N$ in place of prior probabilities p_j, $\hat{\mathbf{B}}_0$ becomes $\mathbf{H}/N = \sum_j (n_j/N)(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})'$. In that case, the estimated Fisher discriminant functions are proportional to exactly the MANOVA canonical variables. This is the only case Johnson and Wichern consider.

When you have no idea about the prior probabilities p_i, estimating them by $\hat{p}_i = n_j/N$ is *sometimes* a sensible thing to do. In that case, the estimated Fisher discriminant functions are $\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \sqrt{f_e}\,\hat{z}_i$, where $\hat{z}_i = \hat{\boldsymbol{\ell}}_i'\mathbf{x}$ is a MANOVA canonical variable ($\hat{\boldsymbol{\ell}}_i$ is an eigenvector of $\mathbf{H}$ relative to $\mathbf{E}$). This is because the relative eigenvectors $\hat{\mathbf{e}}_i$ of $\mathbf{H}/N$ relative to $\mathbf{S} = \mathbf{E}/f_e$ are proportional to the relative eigenvectors $\hat{\boldsymbol{\ell}}_i$ of $\mathbf{H}$ relative to $\mathbf{E}$, specifically $\hat{\mathbf{e}}_i = \sqrt{f_e}\,\hat{\boldsymbol{\ell}}_i$.

Using the first r of the estimated Fisher discriminant functions, you would classify an observation $\mathbf{x}$ to the population with the largest value of

$$\sum_{1\le i\le r}\{(\overline{\mathbf{x}}_j'\hat{\mathbf{e}}_i)\hat{y}_i - \hat{c}_{ij}\} + \log(p_j),\ \hat{c}_{ij} = (\overline{\mathbf{x}}_j'\hat{\mathbf{e}}_i)^2/2.$$

Since $\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x}$ and $\hat{c}_{ij} = (\overline{\mathbf{x}}_j'\hat{\mathbf{e}}_i)^2/2$,

$$\sum_{1\le i\le r}\left\{(\overline{\mathbf{x}}_j'\hat{\mathbf{e}}_i)\hat{y}_i - \hat{c}_{ij}\right\} = \overline{\mathbf{x}}_j'\left\{\sum_{1\le i\le r}\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'\right\}\mathbf{x} - \overline{\mathbf{x}}_j'\left\{\sum_{1\le i\le r}\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'\right\}\overline{\mathbf{x}}_j/2$$

$$= \overline{\mathbf{x}}_j'\hat{\mathbf{M}}_r\mathbf{x} - \overline{\mathbf{x}}_j'\hat{\mathbf{M}}_r\overline{\mathbf{x}}_j/2, \text{ where } \hat{\mathbf{M}}_j = \sum_{1\le i\le r}\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'.$$

Thus, letting, $c_j^* \equiv \overline{\mathbf{x}}_j'\hat{\mathbf{M}}_r\overline{\mathbf{x}}_j/2 = \sum_{1\le i\le r}\hat{c}_{ij}$ and $\boldsymbol{\ell}_j^* \equiv \hat{\mathbf{M}}_r\overline{\mathbf{x}}_j$, you would classify according to the largest value of $\boldsymbol{\ell}_j^{*\prime}\mathbf{x} - c_j^*$, $j = 1,\ldots,g$.