

# Lecture 4

## Discriminant Analysis

Discriminant analysis uses continuous variable measurements on different groups of items to highlight aspects that distinguish the groups and to use these measurements to classify new items. Common uses of the method have been in biological classification into species and sub-species, classifying applications for loans, credit cards and insurance into low risk and high risk categories, classifying customers of new products into early adopters, early majority, late majority and laggards, classification of bonds into bond rating categories, research studies involving disputed authorship, college admissions, medical studies involving alcoholics and non-alcoholics, anthropological studies such as classifying skulls of human fossils and methods to identify human fingerprints.

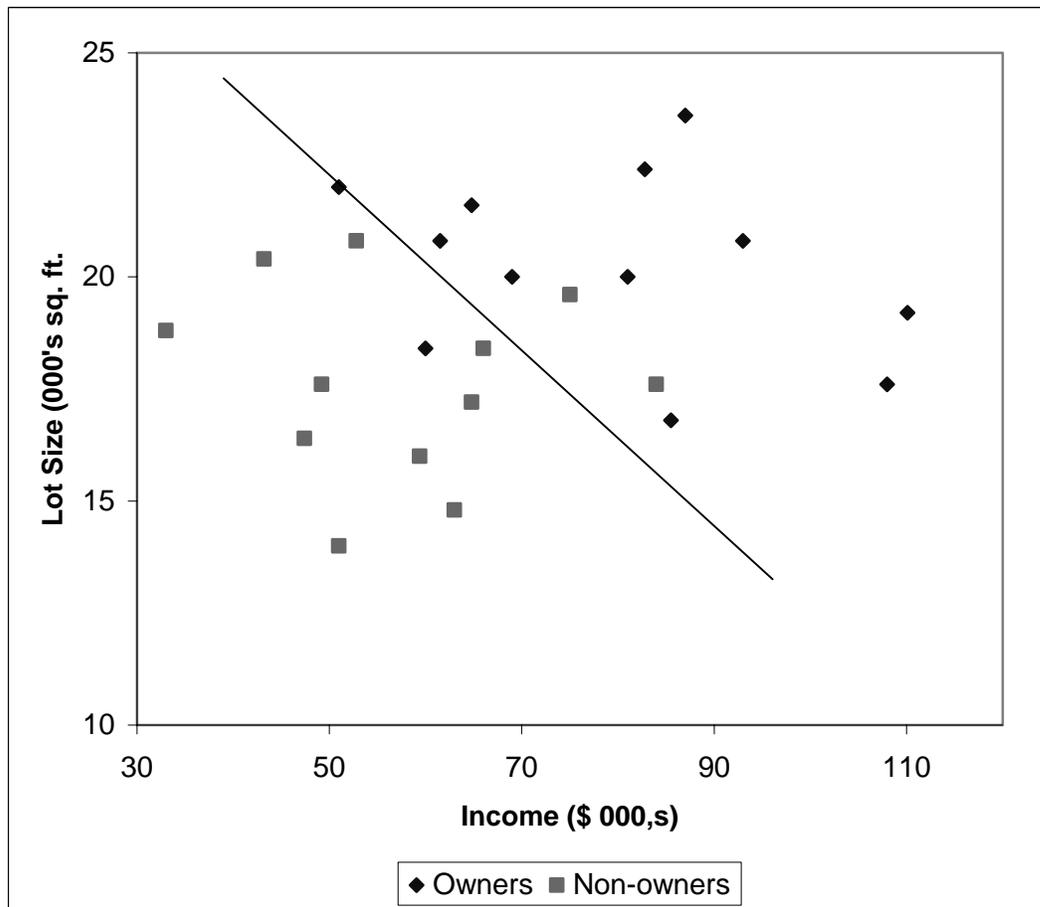
Example 1 (Johnson and Wichern)

A riding-mower manufacturer would like to find a way of classifying families in a city into those that are likely to purchase a riding mower and those who are not likely to buy one. A pilot random sample of 12 owners and 12 non-owners in the city is undertaken. The data are shown in Table I and plotted in Figure 1 below:

Table 1

Observation	Income (\$ 000's)	Lot Size (000's sq. ft.)	Owners=1, Non-owners=2
1	60	18.4	1
2	85.5	16.8	1
3	64.8	21.6	1
4	61.5	20.8	1
5	87	23.6	1
6	110.1	19.2	1
7	108	17.6	1
8	82.8	22.4	1
9	69	20	1
10	93	20.8	1
11	51	22	1
12	81	20	1
13	75	19.6	2
14	52.8	20.8	2
15	64.8	17.2	2
16	43.2	20.4	2
17	84	17.6	2
18	49.2	17.6	2
19	59.4	16	2
20	66	18.4	2
21	47.4	16.4	2
22	33	18.8	2
23	51	14	2
24	63	14.8	2

Figure 1



We can think of a linear classification rule as a line that separates the  $x_1$ - $x_2$  region into two parts where most of the owners are in one half-plane and the non-owners are in the complementary half-space. A good classification rule would separate out the data so that the fewest points are misclassified: the line shown in Fig.1 seems to do a good job in discriminating between the two groups as it makes 4 misclassifications out of 24 points. Can we do better?

We can obtain linear classification functions that were suggested by Fisher using statistical software. You can use XLMiner to find Fisher's linear classification functions. Output 1 shows the results of invoking the discriminant routine.

## Output 1

### Prior Class Probabilities

Prior class probabilities	According to relative occurrences in training data
---------------------------	--

Class	Probability
1	0.5
2	0.5

### Classification Functions

Variables	Classification Function	
	1	2
Constant	-73.160202	-51.4214439
Income (\$ 000's)	0.42958561	0.32935533
Lot Size (000's sq. ft.)	5.46674967	4.68156528

### Canonical Variate Loadings

Variables	Variate1
Income (\$ 000's)	0.01032889
Lot Size (000's sq. ft.)	0.08091455

### Training Misclassification Summary

Classification Confusion Matrix		
Actual Class	Predicted Class	
	1	2
1	11	1
2	2	10

Error Report			
Class	# Cases	# Errors	% Error
1	12	1	8.33
2	12	2	16.67
Overall	24	3	12.50

We note that it is possible to have a misclassification rate that is lower (3 in 24) by using the classification functions specified in the output. These functions are specified in a way that can be easily generalized to more than two classes. A family is classified into Class 1 of owners if Function 1 is higher than Function 2, and into Class 2 if the reverse is the case. The values given for the functions are simply the weights to be associated with each

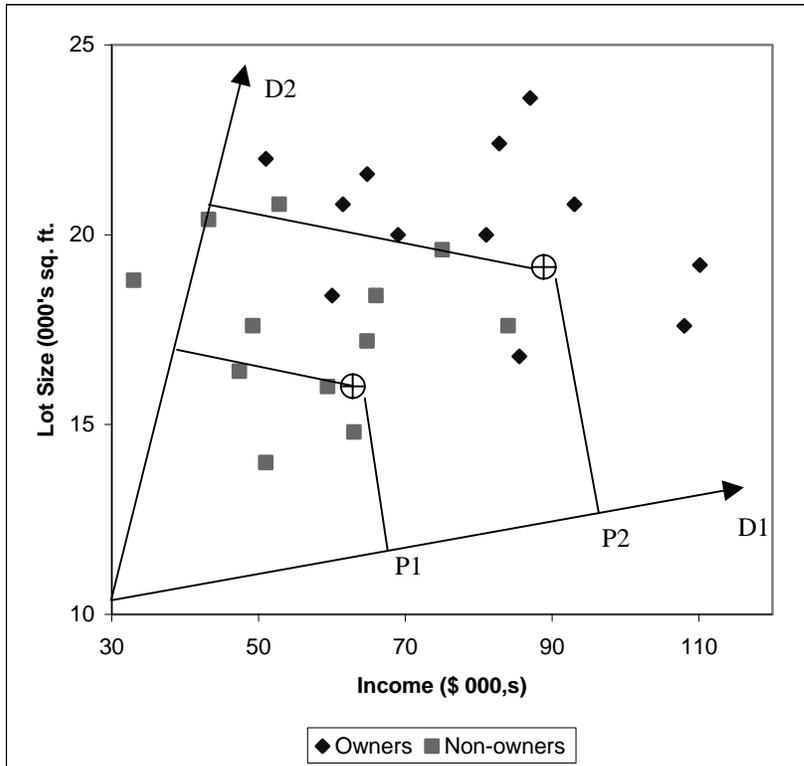
variable in the linear function in a manner analogous to multiple linear regression. For example, the value of the Classification function for class1 is 53.20. This is calculated using the coefficients of classification function1 shown in Output 1 above as  $-73.1602 + 0.4296 \times 60 + 5.4667 \times 18.4$ . XLMiner computes these functions for the observations in our dataset. The results are shown in Table 3 below.

Table 3

Observation	Classes		Classification Function Values			Input Variables	
	Predicted Class	Actual Class	Max Value	Value for Class - 1	Value for Class - 2	Income (\$ 000's)	Lot Size (000's sq. ft.)
1	2	1	54.48067856	53.203125	54.48067856	60	18.4
2	1	1	55.41075897	55.41075897	55.38873291	85.5	16.8
3	1	1	72.75873566	72.75873566	71.04259491	64.8	21.6
4	1	1	66.96770477	66.96770477	66.21046448	61.5	20.8
5	1	1	93.22903442	93.22903442	87.71740723	87	23.6
6	1	1	79.09877014	79.09877014	74.72663116	110.1	19.2
7	1	1	69.44983673	69.44983673	66.54447937	108	17.6
8	1	1	84.86467743	84.86467743	80.71623993	82.8	22.4
9	1	1	65.81620026	65.81620026	64.93537903	69	20
10	1	1	80.49964905	80.49964905	76.5851593	93	20.8
11	1	1	69.01715851	69.01715851	68.37011719	51	22
12	1	1	70.97122192	70.97122192	68.88764191	81	20
13	1	2	66.20701599	66.20701599	65.03888702	75	19.6
14	2	2	63.3450737	63.23031235	63.3450737	52.8	20.8
15	2	2	50.44370651	48.70503998	50.44370651	64.8	17.2
16	2	2	58.31063843	56.91959	58.31063843	43.2	20.4
17	1	2	59.13978195	59.13978195	58.63995361	84	17.6
18	2	2	47.17838669	44.19020462	47.17838669	49.2	17.6
19	2	2	43.04730988	39.82517624	43.04730988	59.4	16
20	2	2	56.45681	55.78063965	56.45681	66	18.4
21	2	2	40.96767044	36.85684967	40.96767044	47.4	16.4
22	2	2	47.46070862	43.79101944	47.46070862	33	18.8
23	2	2	30.917593	25.28316116	30.917593	51	14
24	2	2	38.61510849	34.81158447	38.61510849	63	14.8

Notice that observations 1, 13 and 17 are misclassified as we would expect from the output shown in Table 2.

Let us describe the reasoning behind Fisher's linear classification rules. Figure 3 depicts the logic.



Consider various directions such as directions D1 and D2 shown in Figure 2. One way to identify a good linear discriminant function is to choose amongst all possible directions the one that has the property that when we project (drop a perpendicular line from) the means of the two groups onto a line in the chosen direction the projections of the group means (feet of the perpendiculars, e.g. P1 and P2 in direction D1) are separated by the maximum possible distance. The means of the two groups are:

	Income	Area
Mean1	79.5	20.3
Mean2	57.4	17.6

We still need to decide how to measure the distance. We could simply use Euclidean distance. This has two drawbacks. First, the distance would depend on the units we choose to measure the variables. We will get different answers if we decided to measure area in say, square yards instead of thousands of square feet. Second, we would not be taking any account of the correlation structure. This is often a very important consideration especially when we are using many variables to separate groups. In this case often there will be variables which by themselves are useful discriminators between groups but in the presence of other variables are practically redundant as they capture the same effects as the other variables.

Fisher's method gets over these objections by using a measure of distance that is a generalization of Euclidean distance known as Mahalanobis distance. This distance is defined with respect to a positive definite matrix  $\Sigma$ . The squared Mahalanobis distance

between two p-dimensional (column) vectors  $y_1$  and  $y_2$  is  $(y_1 - y_2)' \Sigma^{-1} (y_1 - y_2)$  where  $\Sigma$  is a symmetric positive definite square matrix with dimension p. Notice that if  $\Sigma$  is the identity matrix the Mahalanobis distance is the same as Euclidean distance. In linear discriminant analysis we use the pooled sample variance matrix of the different groups. If  $X_1$  and  $X_2$  are the  $n_1 \times p$  and  $n_2 \times p$  matrices of observations for groups 1 and 2, and the respective sample variance matrices are  $S_1$  and  $S_2$ , the pooled matrix  $S$  is equal to  $\{(n_1-1) S_1 + (n_2-1) S_2\} / (n_1 + n_2 - 2)$ . The matrix  $S$  defines the optimum direction (actually the eigenvector associated with its largest eigenvalue) that we referred to when we discussed the logic behind Figure 2. This choice of Mahalanobis distance can also be shown to be optimal\* in the sense of minimizing the expected misclassification error when the variable values of the populations in the two groups (from which we have drawn our samples) follow a multivariate normal distribution with a common covariance matrix. In fact it is optimal for the larger family of elliptical distributions with equal variance-covariance matrices. In practice the robustness of the method is quite remarkable in that even for situations that are only roughly normal it performs quite well.

If we had a prospective customer list with data on income and area, we could use the classification functions in Output 1 to identify the sub-list of families that are classified as group 1. This sub-list would consist of owners (within the classification accuracy of our functions) and therefore prospective purchasers of the product.

### **Classification Error**

What is the accuracy we should expect from our classification functions? We have an training data error rate (often called the re-substitution error rate) of 12.5% in our example. However this is a biased estimate as it is overly optimistic. This is because we have used the same data for fitting the classification parameters as well for estimating the error. In data mining applications we would randomly partition our data into training and validation subsets. We would use the training part to estimate the classification functions and hold out the validation part to get a more reliable, unbiased estimate of classification error.

So far we have assumed that our objective is to minimize the classification error and that the chances of encountering an item from either group requiring classification is the same. . If the probability of encountering an item for classification in the future is not equal for both groups we should modify our functions to reduce our expected (long run average) error rate. Also we may not want to minimize misclassification rate in certain situations. If the cost of mistakenly classifying a group 1 item as group 2 is very different from the cost of classifying a group 2 item as a group 1 item, we may want to minimize the expect cost of misclassification rather than the error rate that does not take cognizance of unequal misclassification costs. It is simple to incorporate these situations into our framework for two classes. All we need to provide are estimates of the ratio of the

---

\* This is true asymptotically, i.e. for large training samples. Large training samples are required for  $S$ , the pooled sample variance matrix, to be a good estimate of the population variance matrix.

chances of encountering an item in class 1 as compared to class 2 in future classifications and the ratio of the costs of making the two kinds of classification error. These ratios will alter the constant terms in the linear classification functions to minimize the expected cost of misclassification. The intercept term for function 1 is increased by  $\ln(C(2|1)) + \ln(P(C_1))$  and that for function 2 is increased by  $\ln(C(1|2)) + \ln(P(C_2))$ , where  $C(i|j)$  is the cost of misclassifying a Group  $j$  item as Group  $i$  and  $P(C_j)$  is the apriori probability of an item belonging to Group  $j$ .

### **Extension to more than two classes**

The above analysis for two classes is readily extended to more than two classes. Example 2 illustrates this setting.

Example 2: Fisher's Iris Data This is a classic example used by Fisher to illustrate his method for computing classification functions. The data consists of four length measurements on different varieties of iris flowers. Fifty different flowers were measured for each species of iris. A sample of the data are given in Table 4 below:

Table 4

OBS#	SPECIES	CLASSCODE	SEPLEN	SEPW	PETLEN	PETW
1	Iris-setosa	1	5.1	3.5	1.4	0.2
2	Iris-setosa	1	4.9	3	1.4	0.2
3	Iris-setosa	1	4.7	3.2	1.3	0.2
4	Iris-setosa	1	4.6	3.1	1.5	0.2
5	Iris-setosa	1	5	3.6	1.4	0.2
6	Iris-setosa	1	5.4	3.9	1.7	0.4
7	Iris-setosa	1	4.6	3.4	1.4	0.3
8	Iris-setosa	1	5	3.4	1.5	0.2
9	Iris-setosa	1	4.4	2.9	1.4	0.2
10	Iris-setosa	1	4.9	3.1	1.5	0.1
...	...	...	...	...	...	...
51	Iris-versicolor	2	7	3.2	4.7	1.4
52	Iris-versicolor	2	6.4	3.2	4.5	1.5
53	Iris-versicolor	2	6.9	3.1	4.9	1.5
54	Iris-versicolor	2	5.5	2.3	4	1.3
55	Iris-versicolor	2	6.5	2.8	4.6	1.5
56	Iris-versicolor	2	5.7	2.8	4.5	1.3
57	Iris-versicolor	2	6.3	3.3	4.7	1.6
58	Iris-versicolor	2	4.9	2.4	3.3	1
59	Iris-versicolor	2	6.6	2.9	4.6	1.3
60	Iris-versicolor	2	5.2	2.7	3.9	1.4
...	...	...	...	...	...	...
101	Iris-virginica	3	6.3	3.3	6	2.5
102	Iris-virginica	3	5.8	2.7	5.1	1.9
103	Iris-virginica	3	7.1	3	5.9	2.1
104	Iris-virginica	3	6.3	2.9	5.6	1.8

105 Iris-virginica	3	6.5	3	5.8	2.2
106 Iris-virginica	3	7.6	3	6.6	2.1
107 Iris-virginica	3	4.9	2.5	4.5	1.7
108 Iris-virginica	3	7.3	2.9	6.3	1.8
109 Iris-virginica	3	6.7	2.5	5.8	1.8
110 Iris-virginica	3	7.2	3.6	6.1	2.5

The results from applying the discriminant analysis procedure of Xlminer are shown in Output 2:

## Output 2

### Classification Functions

Variables	Classification Function		
	1	2	3
Constant	-86.3084793	-72.8526154	-104.368332
SEPLEN	23.5441742	15.6982136	12.4458504
SEPW	23.5878677	7.07251072	3.68528175
PETLEN	-16.4306431	5.21144867	12.7665491
PETW	-17.398407	6.43422985	21.0791111

### Canonical Variate Loadings

Variables	Variate1	Variate2
SEPLEN	0.06840593	0.00198865
SEPW	0.12656119	0.17852645
PETLEN	-0.18155289	-0.0768638
PETW	-0.23180288	0.23417209

### Training Misclassification Summary

Classification Confusion Matrix			
Actual Class	Predicted Class		
	1	2	3
1	50	0	0
2	0	48	2
3	0	1	49

Error Report			
Class	# Cases	# Errors	% Error
1	50	0	0.00
2	50	2	4.00
3	50	1	2.00
Overall	150	3	2.00

For illustration the computations of the classification function values for observations 40 to 55 and 125 to 135 are shown in Table 5.

Table 5

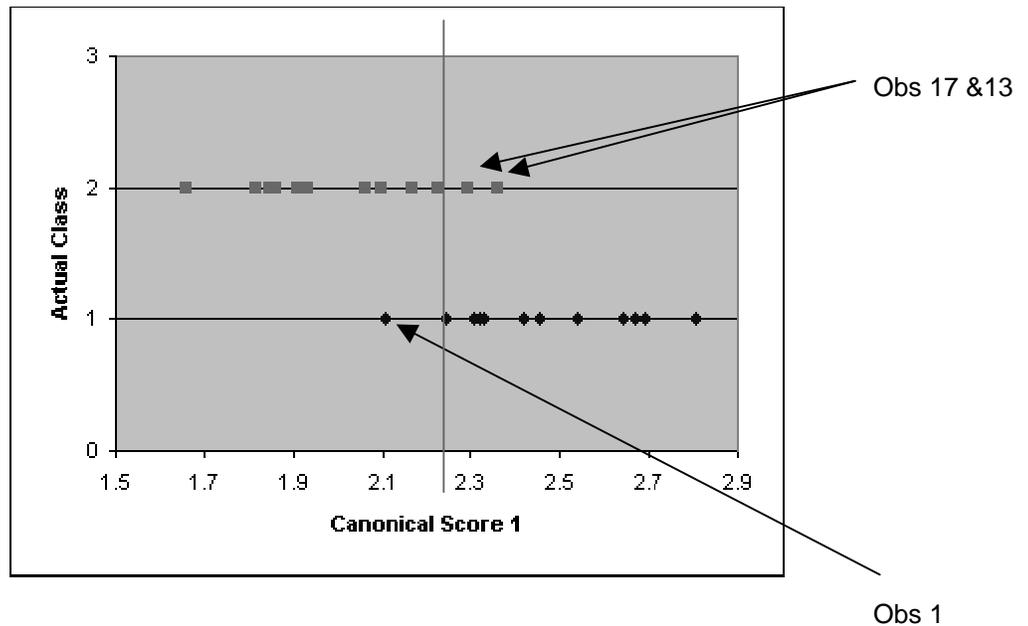
40	1	1	85.83991241	85.83991241	40.3588295	-4.99889183	5.1	3.4	1.5	0.2
41	1	1	87.39057159	87.39057159	39.09738922	-6.32034588	5	3.5	1.3	0.3
42	1	1	47.3130455	47.3130455	22.76127243	-16.9656143	4.5	2.3	1.3	0.3
43	1	1	67.92755127	67.92755127	26.91328812	-17.0013542	4.4	3.2	1.3	0.2
44	1	1	77.24185181	77.24185181	42.59109879	3.833352089	5	3.5	1.6	0.6
45	1	1	85.22311401	85.22311401	46.55926132	5.797660828	5.1	3.8	1.9	0.4
46	1	1	69.24474335	69.24474335	32.9426384	-9.37550068	4.8	3	1.4	0.3
47	1	1	93.63198853	93.63198853	43.70898056	-2.24812603	5.1	3.8	1.6	0.2
48	1	1	70.99331665	70.99331665	30.5740757	-13.2355289	4.6	3.2	1.4	0.2
49	1	1	97.62510681	97.62510681	45.620224	-1.40413904	5.3	3.7	1.5	0.2
50	1	1	82.76977539	82.76977539	37.56061172	-7.88865852	5	3.3	1.4	0.2
51	2	2	93.16864014	52.40012741	93.16864014	84.05905914	7	3.2	4.7	1.4
52	2	2	83.35085297	39.81990051	83.35085297	76.14614868	6.4	3.2	4.5	1.5
53	2	2	92.57727814	42.66094208	92.57727814	87.10716248	6.9	3.1	4.9	1.5
54	2	2	58.96462631	9.096075058	58.96462631	51.02903748	5.5	2.3	4	1.3
55	2	2	82.61280823	31.09611893	82.61280823	77.19327545	6.5	2.8	4.6	1.5
125	3	3	108.2157593	19.08612823	98.88184357	108.2157593	6.7	3.3	5.7	2.1
126	3	3	111.5763855	28.78975868	105.6568604	111.5763855	7.2	3.2	6	1.8
127	3	3	82.33656311	15.52721214	80.8759079	82.33656311	6.2	2.8	4.8	1.8
128	3	3	83.10569	16.2473011	81.24172974	83.10569	6.1	3	4.9	1.8
129	3	3	101.362709	1.872017622	90.11497498	101.362709	6.4	2.8	5.6	2.1
130	3	3	104.0701981	30.83799362	101.9132233	104.0701981	7.2	3	5.8	1.6
131	3	3	115.9760056	20.68055344	107.1320648	115.9760056	7.4	2.8	6.1	1.9
132	3	3	131.8220978	49.37147522	124.2605438	131.8220978	7.9	3.8	6.4	2
133	3	3	103.4706192	0.132176965	90.75839996	103.4706192	6.4	2.8	5.6	2.2
134	2	3	82.07889557	18.17195129	82.07889557	81.08737946	6.3	2.8	5.1	1.5
135	3	3	82.13652039	2.270064592	79.48704529	82.13652039	6.1	2.6	5.6	1.4

## Canonical Variate Loadings

The canonical variate loadings are useful for graphical representation of the discriminant analysis results. These loadings are used to map the observations to lower dimensions while minimizing loss of “separability information” between the groups.

Fig. 3 shows the canonical values for Example 1. The number of canonical variates is the minimum of one less than the number of classes and the number of variables in the data. In this example this is  $\text{Min}(2-1, 2) = 1$ . So the 24 observations are mapped into 24 points in one dimension ( a line). We have condensed the separability information into 1 dimension from the 2 dimensions in the original data. Notice the separation line between the x values and the mapped values of the misclassified points.

Obs.	Actual Class	Predicted Class	Canonical Score 1
1	1	2	2.10856112
2	1	1	2.242484535
3	1	1	2.417066352
4	1	1	2.318249375
5	1	1	2.80819681
6	1	1	2.690770149
7	1	1	2.5396162
8	1	1	2.667718012
9	1	1	2.33098441
10	1	1	2.64360941
11	1	1	2.30689349
12	1	1	2.45493109
13	2	1	2.36059193
14	2	2	2.228388032
15	2	2	2.061042332
16	2	2	2.096864868
17	2	1	2.29172284
18	2	2	1.932277468
19	2	2	1.908168866
20	2	2	2.17053446
21	2	2	1.816588006
22	2	2	1.86204691
23	2	2	1.65957709
24	2	2	1.84825541



In the case of the iris we would condense the separability information into 2 dimensions. If we had  $c$  classes and  $p$  variables, and  $\text{Min}(c-1,p) > 2$ , we can only plot the first two canonical values for each observation. In such datasets sometimes we still get insight into the separation of the observations in the data by plotting the observations in these two coordinates.

### **Extension to unequal covariance structures**

When the classification variables follow a multivariate normal distribution with variance matrices that differ substantially between different groups, the linear classification rule is no longer optimal. In that case the optimal classification function is quadratic in the classification variables. However, in practice this has not been found to be useful except when the difference in the variance matrices is large and the number of observations available for training and testing is large. The reason is that the quadratic model requires many more parameters that are all subject to error to be estimated. If there are  $c$  classes and  $p$  variables, the number of parameters to be estimated for the different variance matrices is  $cp(p + 1)/2$ . This is an example of the importance of regularization in practice.

### **Logistic discrimination for categorical and non-normal situations**

We often encounter situations in which the classification variables are discrete, even binary. In these situations, and where we have reason to believe that the classification variables are not approximately multivariate normal, we can use a more generally applicable classification technique based on logistic regression.