Simfit

Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.
https://simfit.org.uk
https://simfit.silverfrost.com

Discriminant analysis provides methods for allocating new observations to existing training sets, i.e. groups that have been defined on the basis of previous investigations.

From the main SIMFIT menus choose [Statistics], [Multivariate, [Discriminant analysis] then read in the default test file `g03daf.tf1` which has the following data.

```
1   1.1314    2.4596
1   1.0986    0.2624
1   0.6419   -2.3026
1   1.3350   -3.2189
1   1.4110    0.0953
1   0.6419   -0.9163
2   2.1163    0.0000
2   1.3350   -1.6094
2   1.3610   -0.5108
2   2.0541    0.1823
2   2.2083   -0.5108
2   2.7344    1.2809
2   2.0412    0.4700
2   1.8718   -0.9163
2   1.7405   -0.9163
2   2.6101    0.4700
3   2.3224    1.8563
3   2.2192    2.0669
3   2.2618    1.1314
3   3.9853    0.9163
3   2.7600    2.0281
begin{values}
    1.6292   -0.9163
    2.5572    1.6094
    2.5649   -0.2231
    0.9555   -2.3026
    3.4012   -2.3026
    3.0204   -0.2231
end{values}
```

This data set has three groups, indicated by the nondecreasing integers in columns 1, for three types of Cushing's syndrome, the variables in columns 2 and 3 are logarithms of urinary excretion rates (mg/hr) for two steroid metabolites, and the values below the data are additional observations for allocating to one of the three groups. Such extra observations can also be added interactively and expanded training sets containing the newly assigned data can be saved as SIMFIT MANOVA type files.

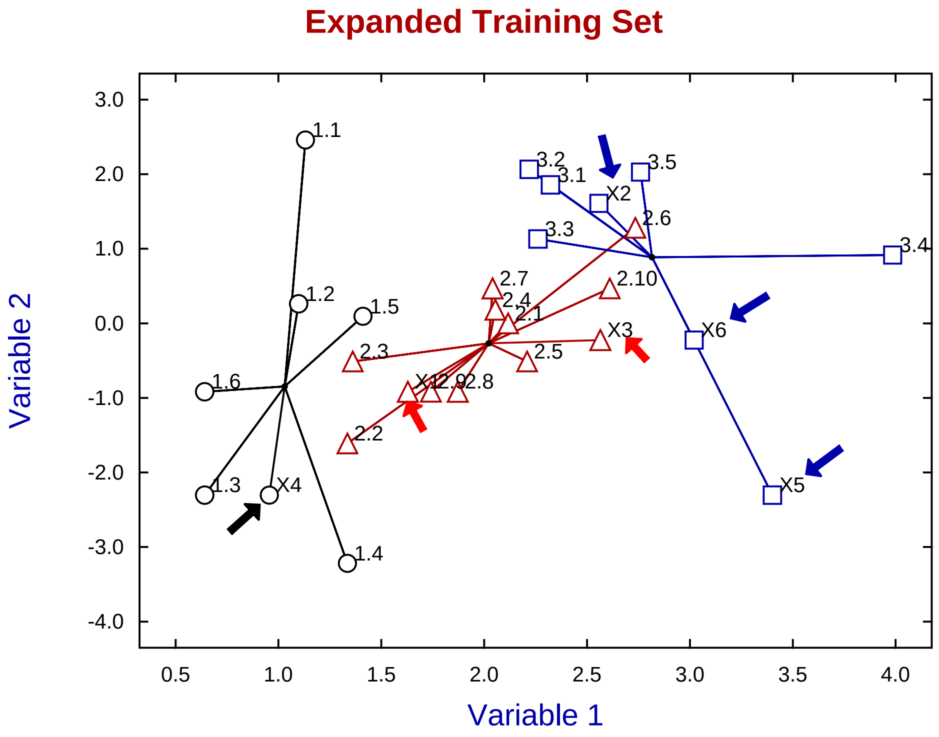
Assigning new observations to groups defined by training sets can be made more objective by employing Bayesian techniques than by simply using distance measures, but only if a multivariate normal distribution can be assumed. For instance, the next table displays the results from assigning the six observations appended to `g03daf.tf1` to groups defined by using the data as a training set, under the assumption of unequal variance-covariance matrices and equal priors.

Method	Predictive
CV-mat	Unequal
Priors	Equal
Observation	Group-allocated
1	2
2	3
3	2
4	1
5	3
6	3

Posterior probabilities			Atypicality indices		
0.0939	0.9046	0.0015	0.5956	0.2539	0.9747
0.0047	0.1682	0.8270	0.9519	0.8360	0.0184
0.0186	0.9196	0.0618	0.9540	0.7966	0.9122
0.6969	0.3026	0.0005	0.2073	0.8599	0.9929
0.3174	0.0130	0.6696	0.9908	1.0000	0.9843
0.0323	0.3664	0.6013	0.9807	0.9779	0.8871

Plotting training sets and assigned observations

The next figure displays the training set from `g03daf.tf1`, together with the assignment of the extra observations appended to this test file as described previously. The additional observations allocated to the existing training set to create this expanded training set are emphasized by solid arrows, which confirm what the atypicality indices suggest, i.e. additional observation 5 is not particularly close to group 3.



Theory

The results from discriminant analysis will differ depending on whether it is assumed that the variables all have the same population covariance matrix so that this can be estimated from the pooled samples. Alternatively estimates from the separate groups can used. However, estimating variance-covariance matrices from multivariate samples requires sample sizes very much greater than the number of variables and, if this condition is not met, poor estimates can lead to incorrect allocations. So, unless sample sizes in all training sets are very much larger than the number of variables, it is probably best to use pooled estimates and ignore the tests suggesting unequal variance-covariance matrices.

The calculation is for g groups, each with n_j observations on m variables, and it is necessary to make assumptions about the identity or otherwise of the variance-covariance matrices, as well as assigning prior probabilities. Then Bayesian arguments lead to expressions for posterior probabilities q_j, under a variety of assumptions, given prior probabilities π_j as follows.

- Estimative with equal variance-covariance matrices (Linear discrimination)

$$\log q_j \propto -\tfrac{1}{2}D_{kj}^2 + \log \pi_j$$

- Estimative with unequal variance-covariance matrices (Quadratic discrimination)

$$\log q_j \propto -\tfrac{1}{2}D_{kj}^2 + \log \pi_j - \tfrac{1}{2}\log|S_j|$$

- Predictive with equal variance-covariance matrices

$$q_j \propto \frac{\pi_j}{((n_j+1)/n_j)^{m/2}\{1+[n_j/((n-g)(n_j+1))]D_{kj}^2\}^{(n-g+1)/2}}$$

- Predictive with unequal variance-covariance matrices

$$q_j \propto \frac{\pi_j\Gamma(n_j/2)}{\Gamma((n_j-m)/2)((n_j^2-1)/n_j)^{m/2}|S_j|^{1/2}\{1+(n_j/(n_j^2-1))D_{kj}^2\}^{n_j/2}}$$

Subsequently the posterior probabilities are normalized so that $\sum_{j=1}^{g} q_j = 1$ and the new observations are assigned to the groups with the greatest posterior probabilities. In this analysis the priors can be assumed to be all equal, proportional to sample size, or user defined. Also, atypicality indices I_j are computed to estimate how well an observation fits into an assigned group. These are

- Estimative with equal or unequal variance-covariance matrices

$$I_j = P(D_{kj}^2/2, m/2)$$

- Predictive with equal variance-covariance matrices

$$I_j = R(D_{kj}^2/(D_{kj}^2 + (n-g)(n_j-1)/n_j), m/2, (n-g-m+1)/2)$$

- Predictive with unequal variance-covariance matrices

$$I_j = R(D_{kj}^2/(D_{kj}^2 + (n_j^2-1)/n_j), m/2, (n_j-m)/2),$$

where $P(x,\alpha)$ is the incomplete gamma function, and $R(x,\alpha,\beta)$ is the incomplete beta function. Values of atypicality indices close to one for all groups suggest that the corresponding new observation does not fit well into any of the training sets, since one minus the atypicality index can be interpreted as the probability of encountering an observation as or more extreme than the one in question given the training set.

The assignment of extra observations to the training sets depends on the data transformation selected and variables suppressed or included in the analysis, and this must be considered when supplying extra observations interactively.