8.0 Cluster analysis procedures

The cluster analysis is mainly used to find "best" classes [Spaeth83]. Figure 8.01 represents two-dimensional patterns. Which patterns belong together? This is about metric data i.e. a metric is existing. Non-metric data e.g. enumerating (colour, sex and s.o.) can be provided with a metric by a mapping function as previously mentioned. The number of the required classes is the only given component. The classes should be as homogeneous as possible and heterogeneous in opposite to other classes.

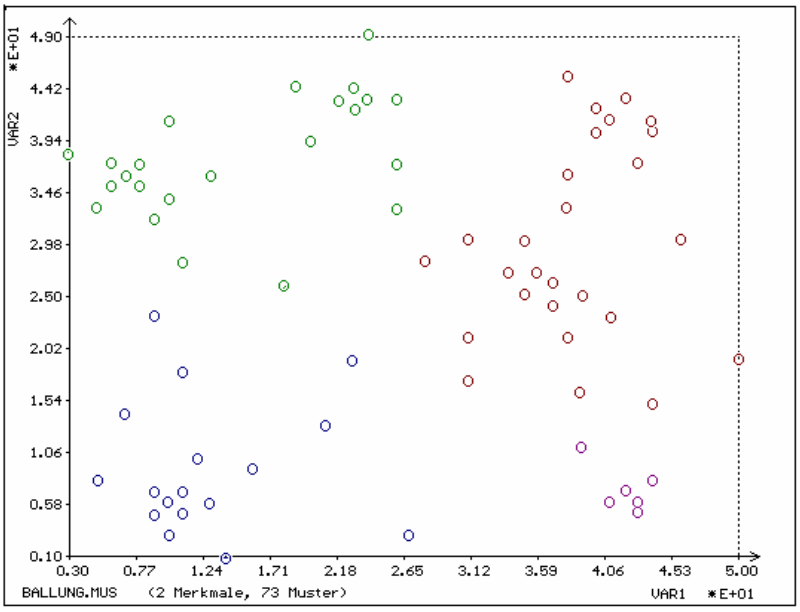

Figure 8.01:
A distribution of spherical clusters.

8.1 The variance criterion

For a cluster a centre is calculated, which has the property that the sum of the quadratic distances from the mean vector to the single pattern has a minimum.

$$d_k^2(x,\mu_k)=(x-\mu_k)^T(x-\mu_k)$$

In the following figure you see the result.

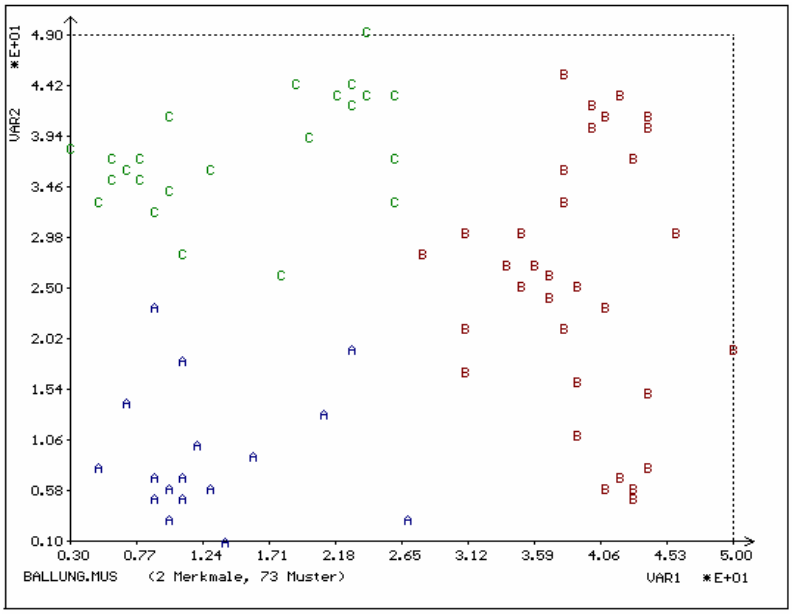

Figure 8.02:
The variance criterion with the option to find three clusters works out fine the class A, B and C

8.2 The determinant

criterion

With the Variance criterion you find invariance towards scale transformation. The determinant criterion does not possess this disadvantage.

$$d^2_{kG}(x,\mu_k) = (x-\mu_k)^T G (x-\mu_k)$$

The derminant criterion is suits the best for finding ellipsoidal clusters.

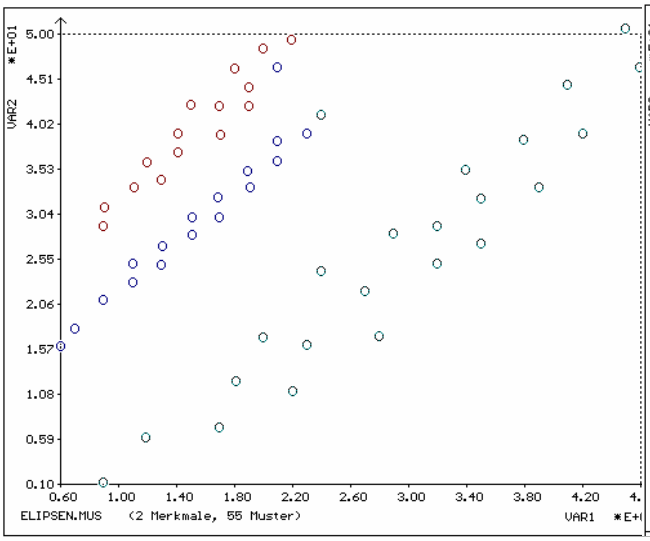

Figure 8.03: A distribution of three ellipsoidal clusters

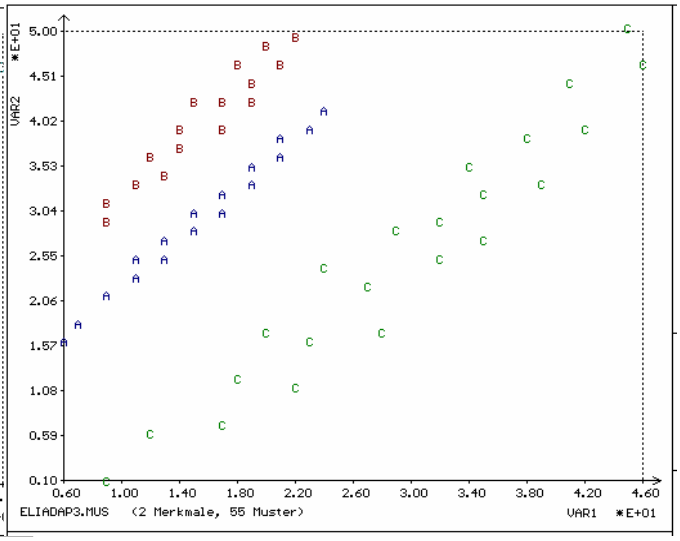

Figure 8.04: The determinant criterion with the option of three clusters, finds three ellipsoidal classes A, B, and C

8.3 The criterion of the adaptive distances

At the determinant criterion a metric was introduced, which considers the distribution of all patterns of a partition. (G without index) A further step is a class specific metric, which defines a different distance measure for each cluster (G with index).

$$d^2_{kG_j}(x,\mu_k) = (x-\mu_k)^T G_j (x-\mu_k)$$

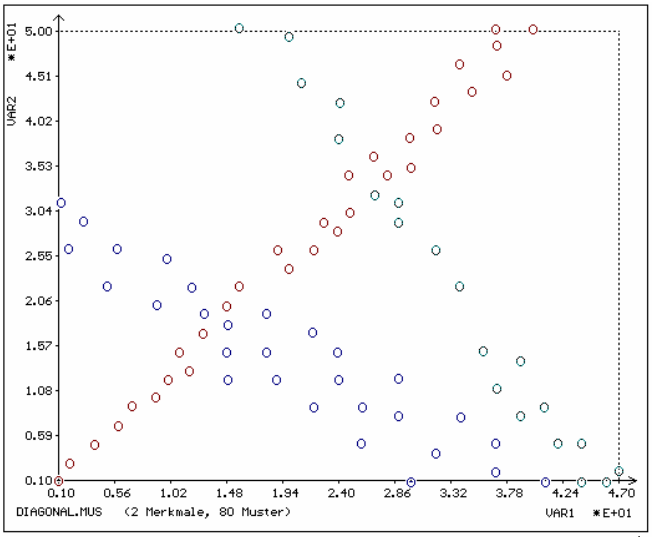

Figure 8.05: A distribution of three penetrating clusters

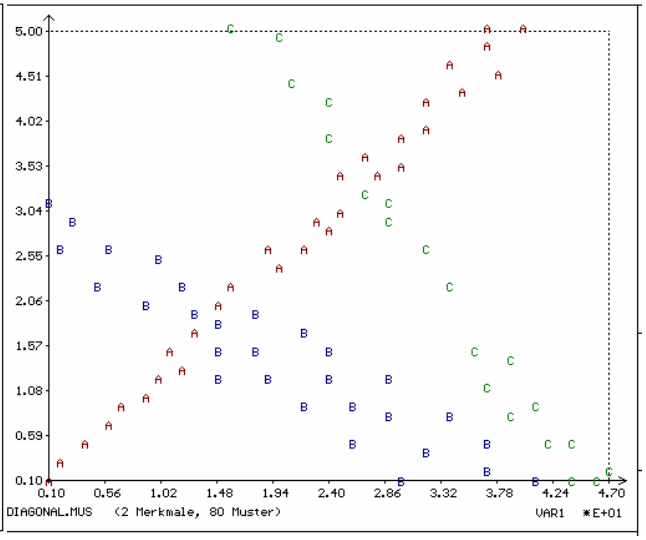

Figure 8.06: The criterion of adaptive distances with the option of three clusters, finds three penetrating diagonal classes A, B, and C

8.4 Comparision of the three criterions

The formulas shown above represent three possibilities of forming clusters. In order to understand the effectiveness, the formulas are applied to the same data in the two-dimensional area. A distinguisher of the formulas is given by the respective metriks.
Looking at the following illustrations you will see, that the variance criterion is suitable best for finding spherical clusters. The determinant criterion recognizes ellipsoidal clusters, and the criterion of the adaptive distances is able to recognize overlapping and/or penetrating clusters.

What happens to the feature vectors presented in figure 8.01, if you use the determinant criterion and the criterion of adaptive distances?

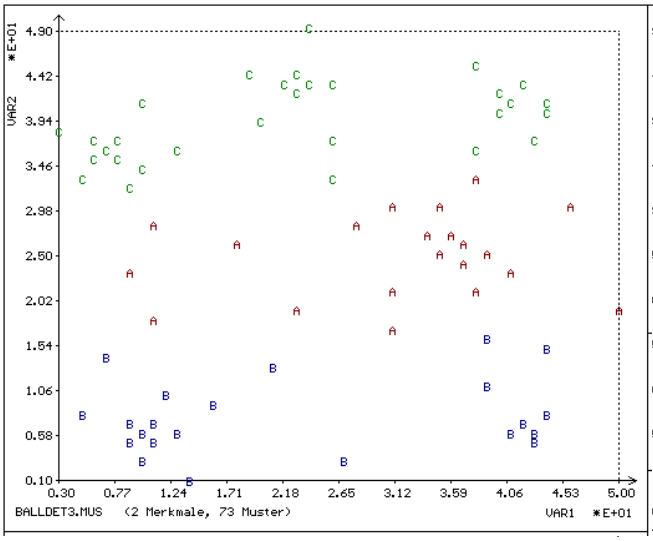

Figure 8.07: The determinant criterion with the option of three clusters recognizes ellipsoidal clusters

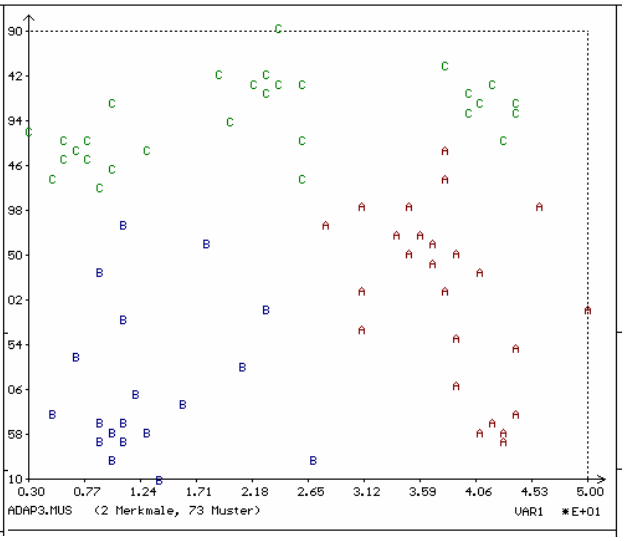

Figure 8.08: The criterion of adptive distances with the option of three clusters, finds also ellipsoidal clusters

What happens to the feature vectors presented in figure 8.03, if you use the variance criterion and the criterion of adaptive distances?

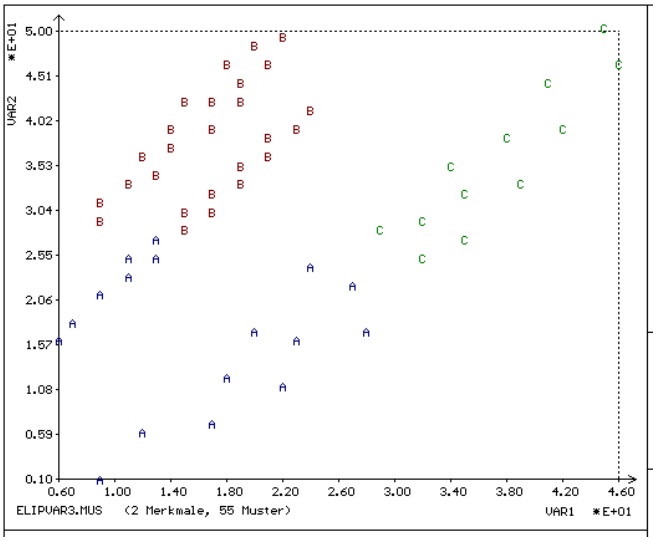

Figure 8.09: The variance criterion with the option of three clusters tries to find out spherical clusters

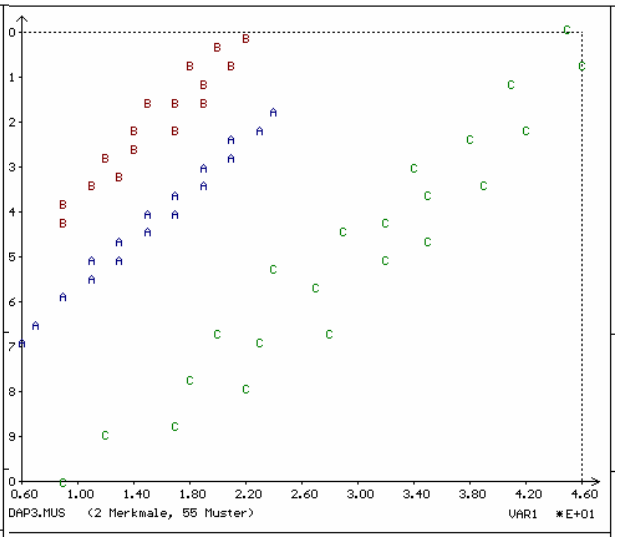

Figure 8.10: The criterion of adptive distances with the option of three clusters, finds also ellipsoidal clusters

What happens to the feature vectors presented in figure 8.05, if you use the variance criterion and the determinant criterion?

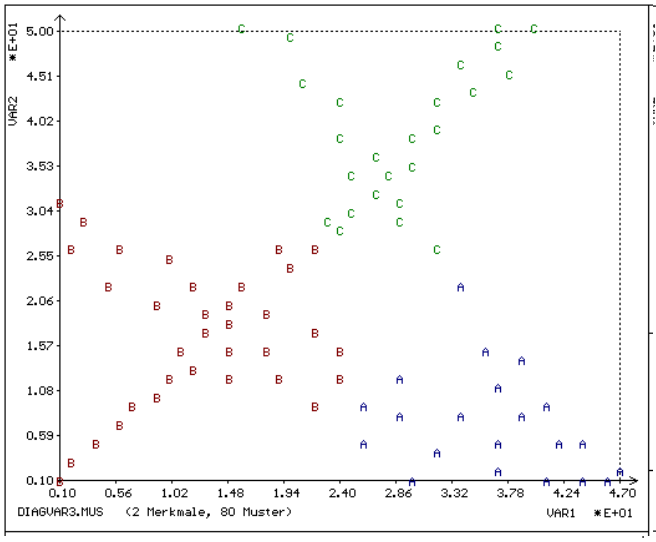

Figure 8.11: The variance criterion with the option of three clusters tries to find out spherical clusters

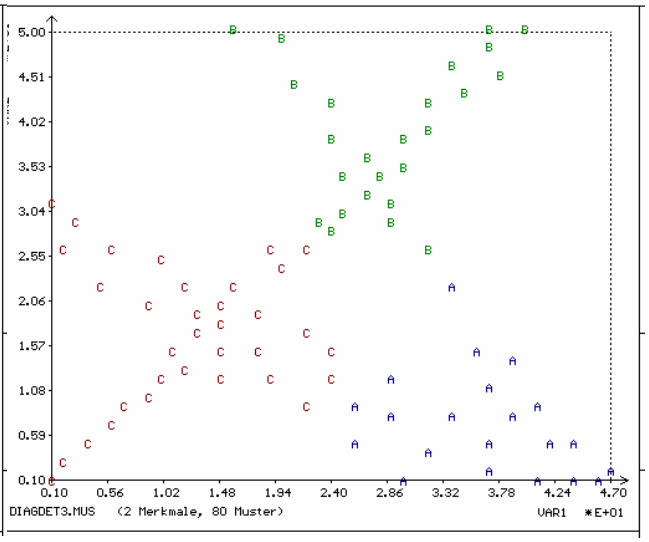

Figure 8.12: The determinant criterion with the option of three clusters, finds also spherical clusters

Investigations with two-dimensional patterns would be interesting. The characters one to zero could be taken, like in chapter 3.0 figure 3.13, and therby found an automatic classification of the patterns.