Classification I
SIMCA

Jens C. Frisvad

What is classification?

- Finding discrete clouds of points in multidimensional feature space with a certain empty space in between to other such clouds of points
- Putting into boxes
- ”The goal of ~~classification~~ identification is to assign new objects to the class to which they show the largest similarity” (p. 335)

SIMCA is a supervised method

- The classification has already been made!
- SIMCA is then a method to find the boundaries of the classes known and to help assign new objects to those classes (identification)
- Outliers are objects outside the class boundaries
- Aliens are object inside the boundaries that do not belong to the class

SIMCA:
Soft Independent Modelling of Class Analogy

- A supervised soft modelling method
- The classes are known beforehand and separate PCA analyses are made for each class
- The number of components in each class is determined by cross-validation
- SIMCA is used on PCA classes, but can in principle also be used for PLS

SIMCA

- The classes in the training set should consist of many objects representing the diversity of the class (suggestion: use at least 25 objects for each class)
- Variable selection should be made, preferably the same number of variables should be present in each set
- The SIMCA model can be tested with a test set
- Tabulate the number of outliers and aliens for each class

SIMCA identification

- Any new object can be allocated to one of the SIMCA classes, or may be an outlier to all the classes (identification)
- If the classes are overlapping, an object may be member of more than one class*
- The distance of any object to the different classes can be calculated
- The distance between classes can be calculated
- *Here fuzzy clustering may be used

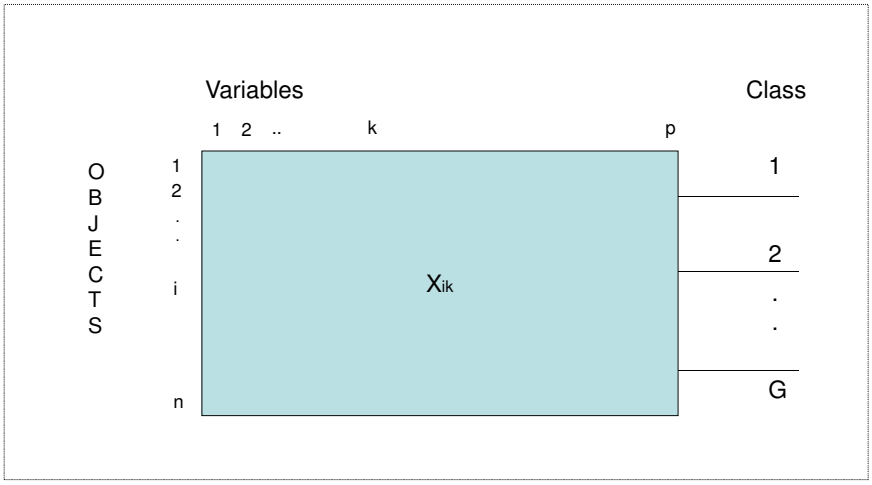
Pretreatment of data

- Use log (x+1) if the highest values in a variable are 8 times greater than the lowest values
- Chromatographic data should often be logarithmated
- Autoscaling (standardization) of often a very good idea
- Do not autoscale variables with a very low variance

SIMCA, the asymmetric case

- Occasionally one class is well defined, but the rest is just outside the class
- Example: All healthy people could be modelled into a SIMCA class, whereas people with different diseases can rarely be modelled by another class

Training set



PCA & SVD

- PCA: X-1μ = T P‘ + E (μ is the average)
- SVD: X-1μ = Uλ P‘ + E (λ is the square root of the eigenvalues in a diagonal matrix)
- Biplot: scores T = U λ and loadings multiplied with λ: P' λ
- For A components:

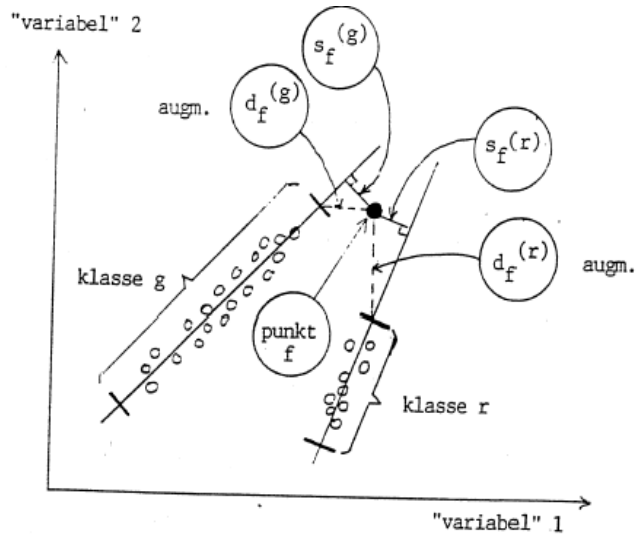
  $X_{np} - 1\ \mu_{np} = U_{na}\ \lambda_{aa}\ P'_{ap} + E_{np}$

General model in SIMCA

$$x_{ik} - \mu_k = \sum_{a=1}^{A} \beta_{ka} \Theta_{ai} + e_{ik}$$

e_{ik} has the standard deviation s_0, which is the typical distance for any object in the class to the class itself

$$s_0 = \sqrt{\sum_{k=1}^{p} \sum_{i=1}^{n} e_{ik}^{\ 2} / (p - A)(n - A - 1)}$$



Distance from an object to a class

$$d_{f,g} = s_{f,g}(augm.) = \sqrt{s^2{}_{f,g} + (t_{af} - \Theta_{a,\lim})^2 \cdot \Phi^2{}_a}$$

$$where: \Phi_a = s_{f,g} / s_{\Theta,g}(a)$$

Confidence cylinder

(box in higher dimensions)

- Confidence radius:

$$\sqrt{F_{95\%}} \cdot s_{g0} = \sqrt{F_{95\%}} \cdot \sum_i s_i{}^2 / n_g$$

$$s_i = \sqrt{\sum_k v_k{}^2 \cdot e_{ik}{}^2 \frac{n_g}{n_g - A_g - 1}} / (p - A_g)$$

The weight v_k can be set to 1, but else it is $v_k = mpow_k$, when mpow is near 1 and 0.1 if mpow is less than 0.1
The F value is with $(p\text{-}A_g)$ and $(p\text{-}A_g)(n_g - A_g - 1)/2$ degrees of freedom

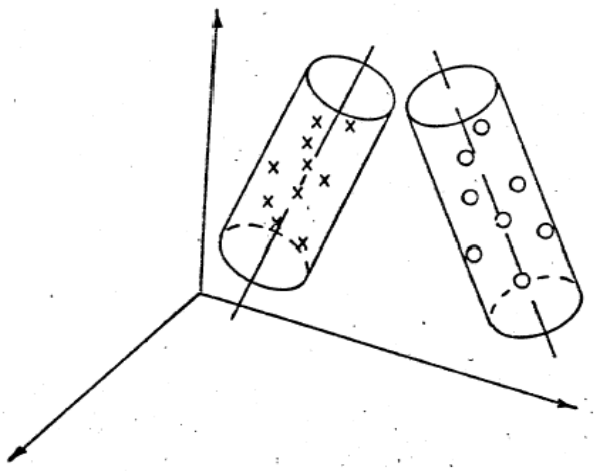
Top and bottom of cylinder

- Top: $\theta_{max} + t^*/2\ s_t$, bottom $\theta_{min} - t^*/2\ s_t$

t^* is the t-distribution with n_g degrees of freedom

$$s_{t,g}^{\ 2}(a) = \sum_{k=1} \Theta_{ak,g}^{\ 2} / n_g$$

SIMCA cylinders



Distance from new object to class

$$\left| x_{kf} w_k - \mu_{k,g} \right| = \sum_{a=1}^{A_g} t_{ar} \beta_{ka} + e_{kf}$$

Autoscaling gives the weights w_k

Standard deviation for new object

Degrees of freedom ($p-A_g$)

$$s_{f,g} = \sqrt{\sum_{k=1}^{p} v_k {e_{ik}}^2 \Psi / (p - A_g)}$$

V_k can be set to one or as mpow (see later)
If the object is in the class proper, the expression has to corrected with the factor $n_g/(n_g-A_g-1) = \Psi$, else it is 1.

Test whether object is in the class

- Degrees of freedom of F-test: $(p\text{-}A_g)$ and $(p\text{-}A_g)(n_g\text{-}A_g)/2$

$$F = s_{i,g}{}^2 / s_0{}^2$$

Test for outliers in training set

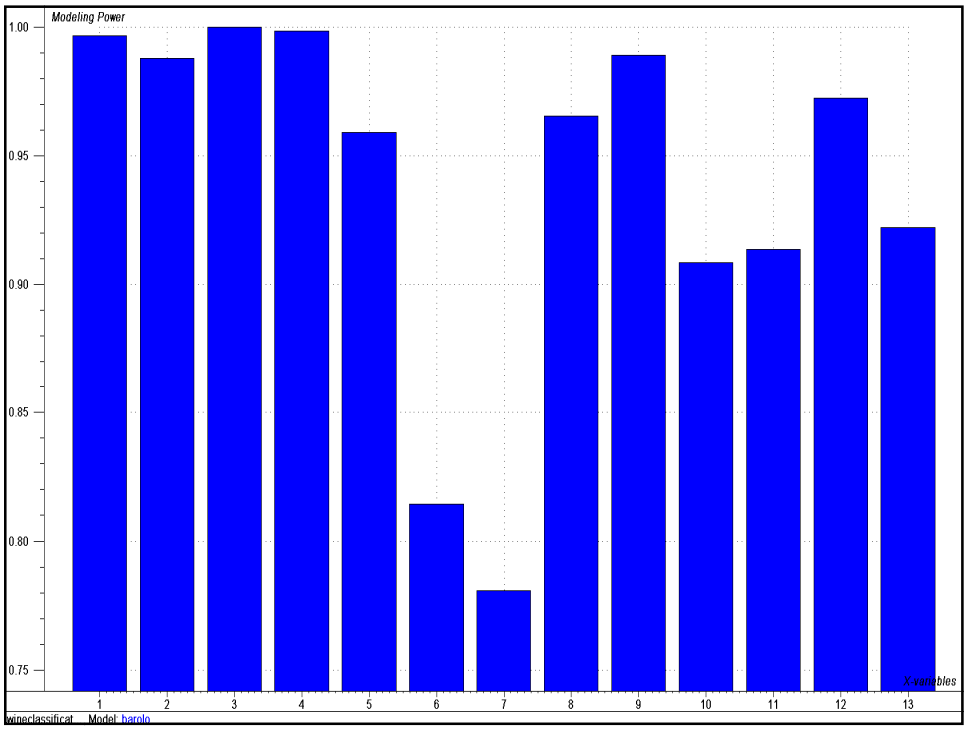
Degrees of freedom in F-test:
$(p\text{-}A_g)/\Psi_{(squared)}$ and $(p\text{-}A_g)(n_g - A_g - 1)$

$$F = \Psi^2 \cdot s_{i,g}{}^2 / s_0{}^2$$

Modelling power

$$mpow_k = 1 - \frac{\sqrt{\sum_{i=1}^{n_g} e_{ik}^{\ 2} / (n_g - A_g - 1)}}{\sqrt{\sum_{i=1}^{n_g} (x - \mu_k)^2 / n_g - 1}}$$

Relevance of variable k in class g, less than 0.1 is low, and the variable may be irrelevant



Discrimination power

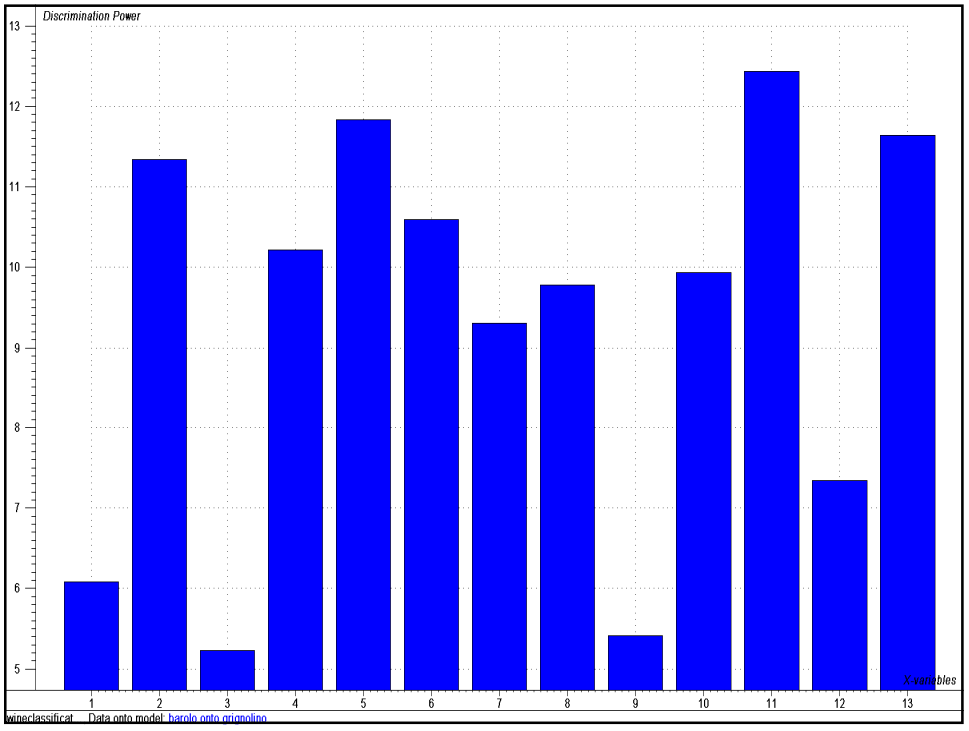
(of variable)

$$d_k(r,g) = \sqrt{\frac{s_{k,r}^{\ 2}(g) + s_{k,g}^{\ 2}(r)}{s_{k,r}^{\ 2} + s_{k,g}^{\ 2}}}$$

$$s_{k,r}^{\ 2} = \sum_{i=1}^{n_r} e_{ik}^{\ 2} / (n_r - A_r - 1)$$

$$s_{k,r}^{\ 2}(g) = \sum_{i=1}^{n_r} e_{kf}^{\ 2}(g) / n_r$$

Low discrimination power: around 1
High discrimination power: 3-4

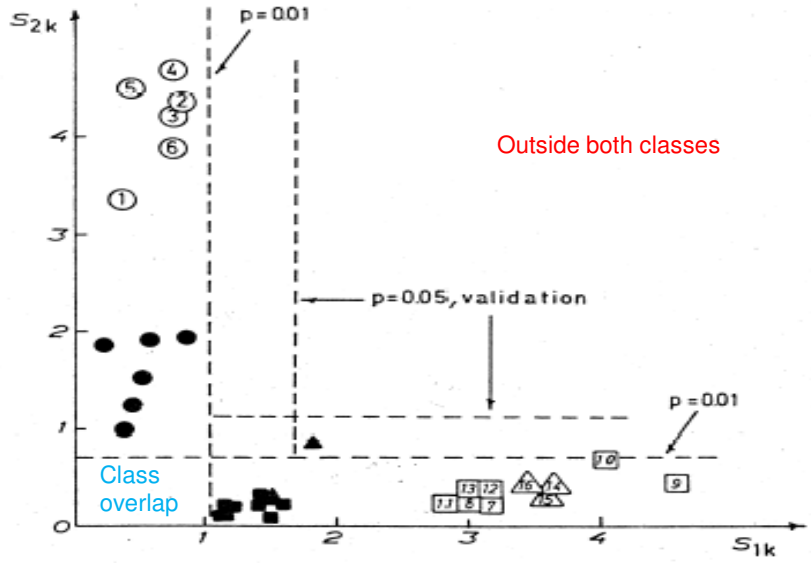


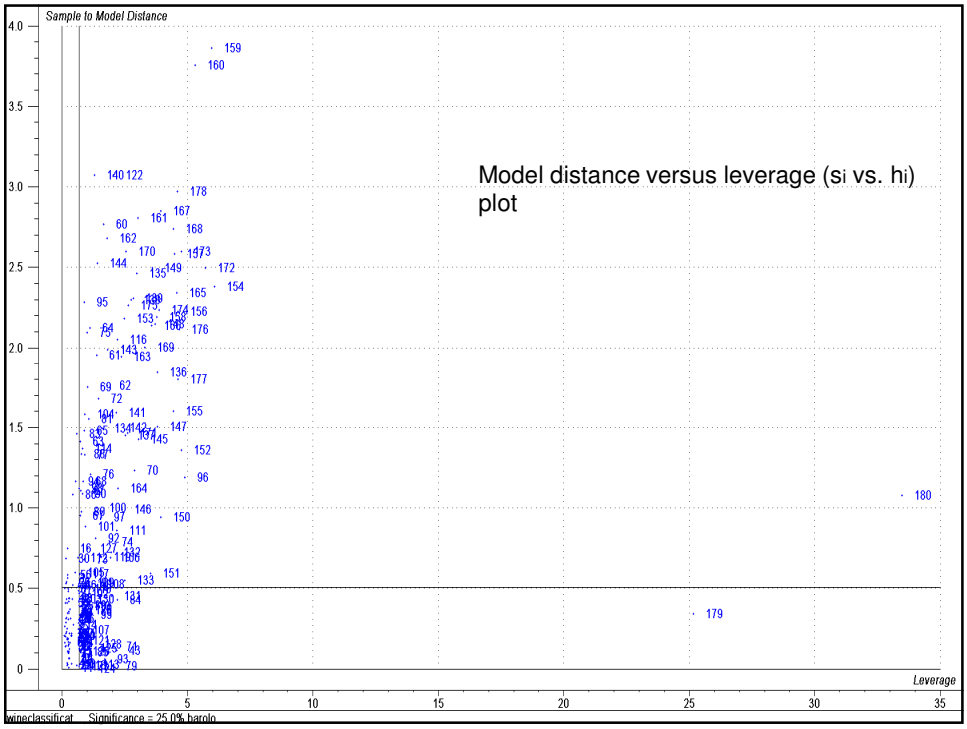
Distance between classes

$$d_{r,g} = \sqrt{\frac{\sum_{k=1}^{p} (s_{k,r}{}^2(g) + s_{k,g}{}^2(r))}{\sum_{k=1}^{p} (s_{k,r}{}^2 + s_{k,g}{}^2)}}$$

If d is less than 1, the classes are overlapping
If d is larger than 3-4 the classes are well separated

Coomans Plot





Model distance versus leverage (s_i vs. h_i) plot