

# Supervised classification and experimental designs

Gilbert Saporta

CEDRIC- CNAM

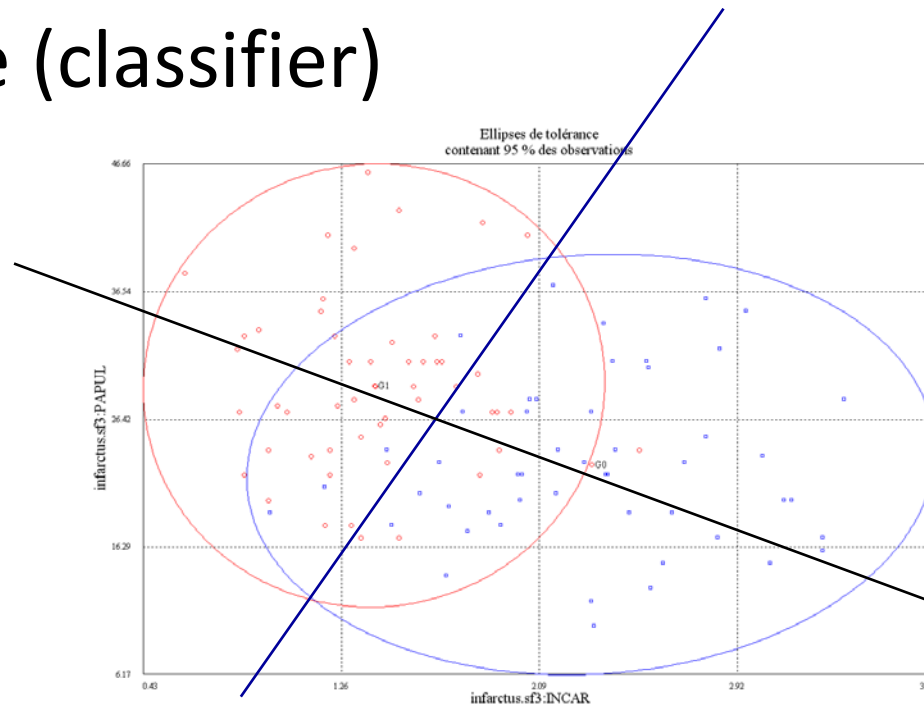
[gilbert.saporta@cnam.fr](mailto:gilbert.saporta@cnam.fr)

# Outline

- Reminder of main classification methods
- Experimental designs
- Behaviour of main methods
- Conclusions and perspectives

# Supervised classification

- Binary response  $Y$  ;  $p$  numerical predictors  $X_j$
- Two sets of points in  $\mathbb{R}^p$
- Looking for a frontier or a score (classifier)



# two linear classifiers

- Fisher's LDA

$$S(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \ln\left(\frac{p_1}{p_2}\right) - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

- Logistic regression

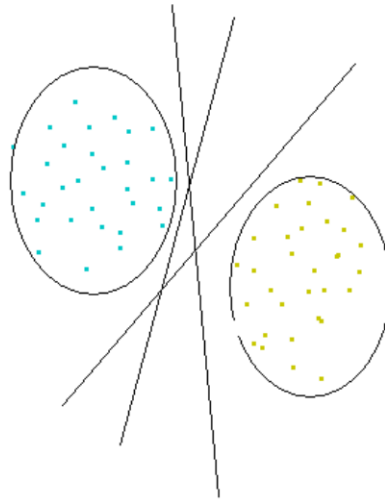
$$\pi(\mathbf{x}) = P(Y = 1 / \mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

- Geometrical background for Fisher's LDA: classify in the closest group according to Mahalanobis distance to centroid.

However: for two multinormal distributions with equal covariance matrices one has:

$$P(G_1 / \mathbf{x}) = \frac{\exp(S(\mathbf{x}))}{1 + \exp(S(\mathbf{x}))}$$

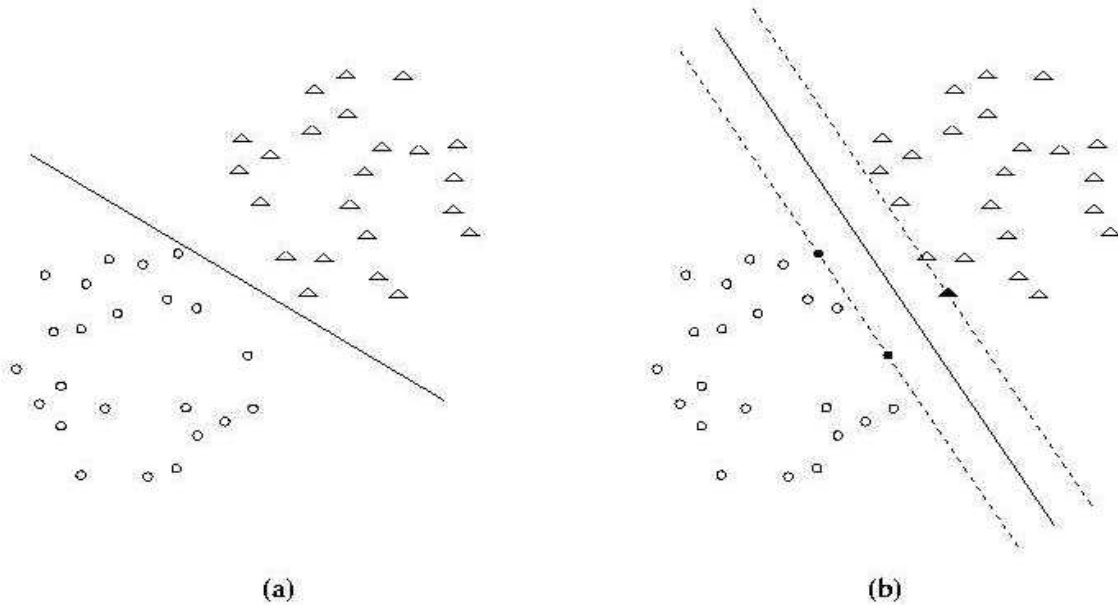
- The case of complete separation



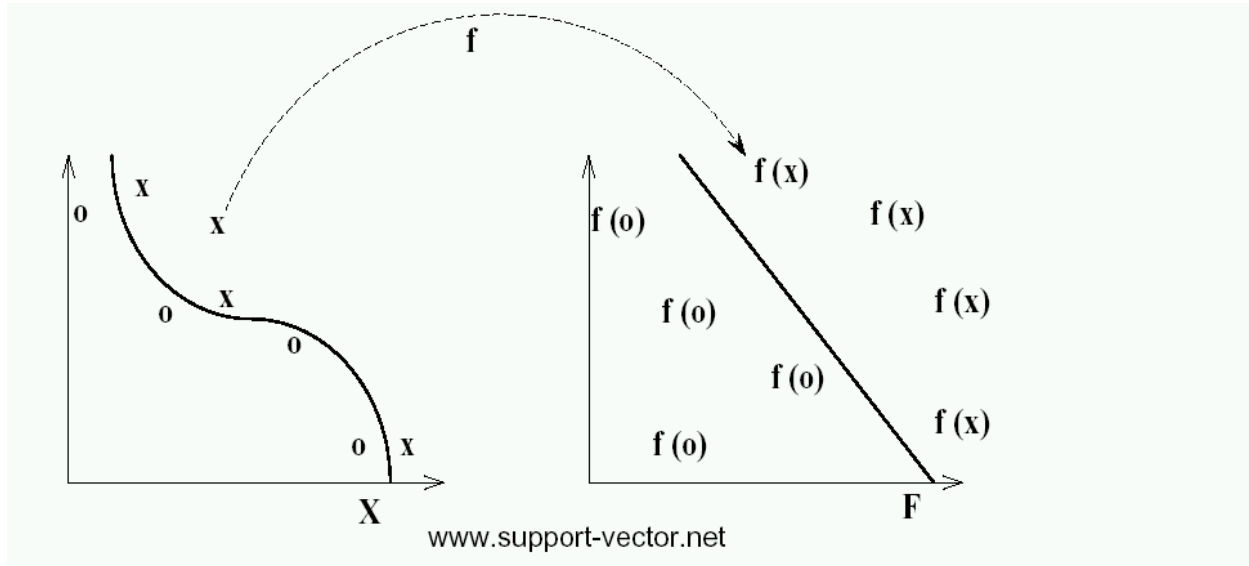
- No solution for logistic regression
- Fisher's LDA still exists as long as centroids are different
- Infinite number of frontiers

# SVMs

- Linear SVM: frontier with the maximal margin (thick hyperplane)



- SVM with a non-linear kernel



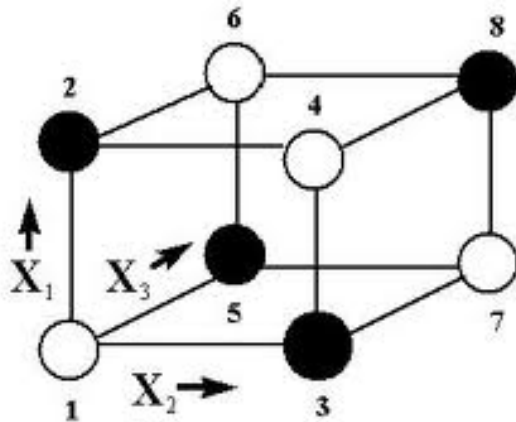
classifier: 
$$f(\mathbf{x}) = \sum_{i \in \text{supports}} \alpha_i y_i K(\mathbf{x}_i; \mathbf{x}) + b$$



# Experimental designs

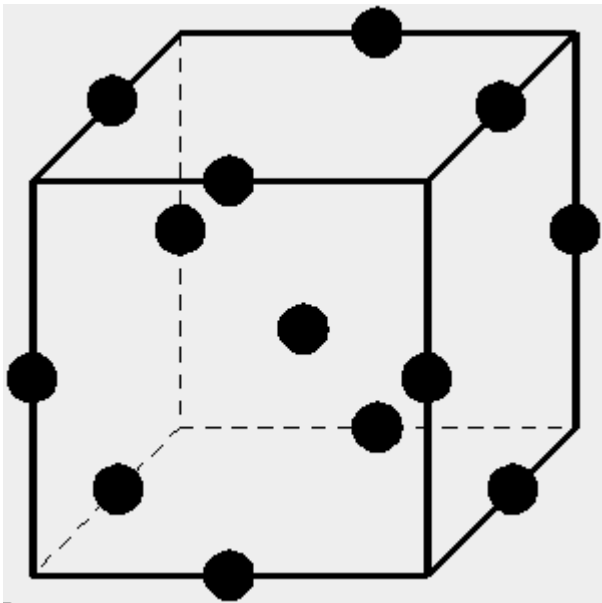
- Controlled experiments: observed  $x$  are fixed in advance, not random like in a sampling scheme
- Typically a few points with repeated observations
- Classical designs for an optimal estimation of a linear model (with possibly quadratic effects and interactions)

- Design points often at the extremes of the experimental domain
  - predictors with finite ranges, standardised to  $[-1, +1]$

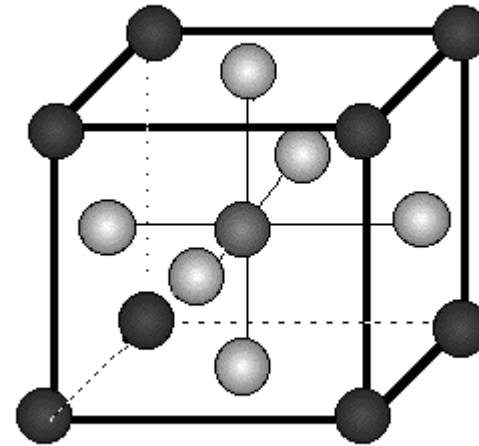


Full and half-fractional design for  
3 factors

- Some other examples:

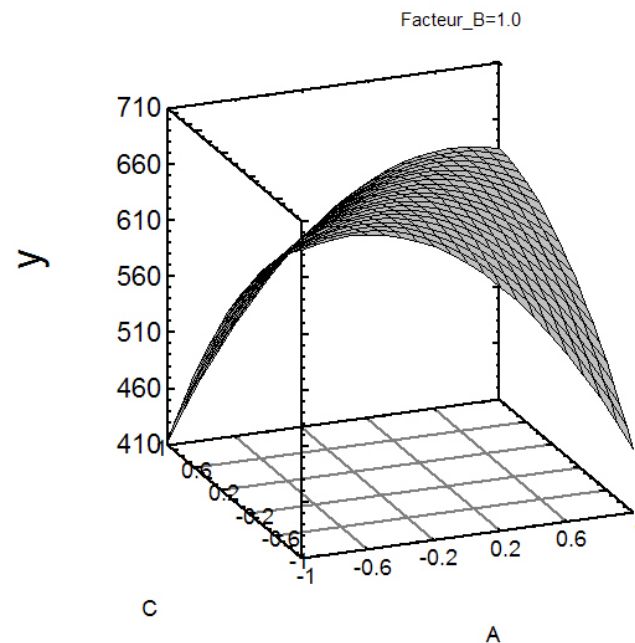


Box-Behnken



Central composite

- Designs used to build response surface when the response is numerical, which is not well fitted for a binary response



- Could estimate a classifier

- Orthogonal designs give usually better estimates than sampling
  - Ex: simple regression

$$V(b) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$X \in [-3, 3] \quad n=6$$

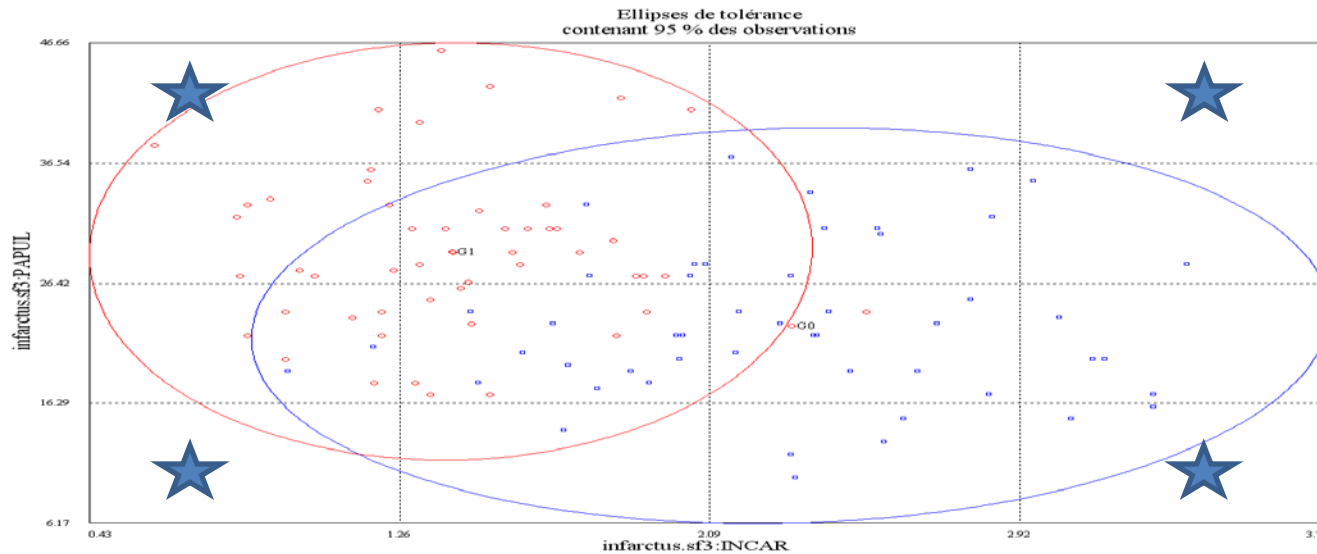
- First case  $x_i = -3 \ -2 \ -1 \ 1 \ 2 \ 3$       $\sum_{i=1}^n (x_i - \bar{x})^2 = 28$
- Second case : 3 obs at -3 and 3 at +3      $\sum_{i=1}^n (x_i - \bar{x})^2 = 54$
- Still better with 3 obs at -2.16 and + 2.16

# Classifying with designs

- Advantages of orthogonal or quasi orthogonal designs:
  - $X$  variables become « independent »; but they are no longer variables (not random)
  - covariance matrices (  $W, V$  ) are regular
  - Variable selection becomes easy and non ambiguous

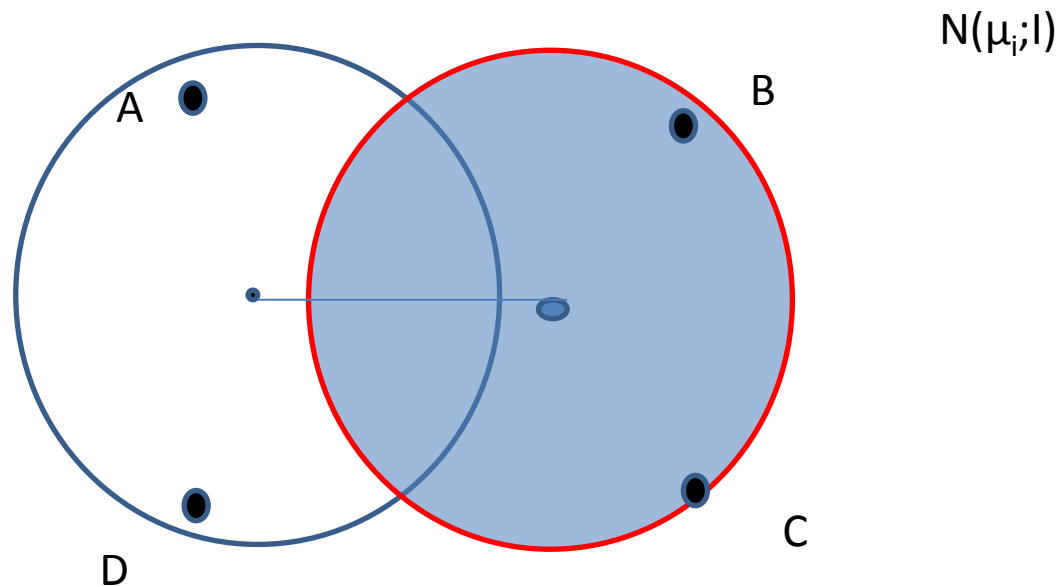
- Drawbacks

- High probability of complete separation: not good for logistic regression



- Not well fitted for SVM: points too far from the support vectors

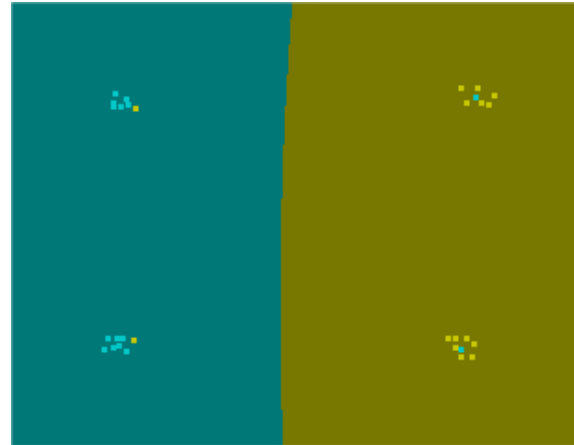
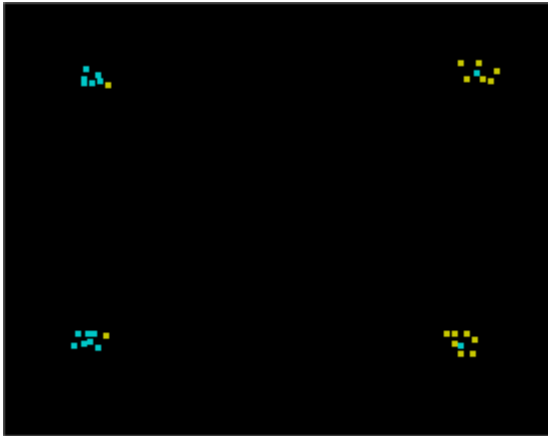
- Probability of complete separation
  - Depends on the Mahalanobis distance between both groups





- Simple example:
  - points are away from their centroids of one standard deviation
  - 2 repetitions for each (n=8)
  - Complete separation occurs if  $Y=-1$  for A and D, and  $Y=1$  for B and C
  - When Mahalanobis distance  $\Delta=3$ , the probability of complete separation is 0.99; if  $\Delta=2$  prob=0.86, if  $\Delta=1$  prob=0.2

# SVM



SVM-toy

# Conclusions and perspectives

- Estimating a classification rule when data comes from an experimental design works well for linear models (Fisher's LDA), but not for non linear ones.
- This is easy to understand since most designs are made for linear models

- Designs should be optimized according to the model
  - Difficult problem: define new criteria (different from D or A optimality) like X-optimality; new algorithms
  - Few publications

**Min Yang, Bin Zhang and Shuguang Huang** (2011)

Optimal designs for generalized linear models with multiple design variables  
*Statistica Sinica* **21**, 1415-1430

**J.-P. Vila and J.-P. Gauchi.** (2007) Optimal designs based on exact confidence regions for parameter estimation of a nonlinear regression model.

*Journal of Statistical Planning and Inference*, **137**(9): 2935-2953

Thanks for your attention!