UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Chimiche e Geologiche

PLS -DA

Implementation, Practical work

Marina Cocchi

Dipartimento di Scienze Chimiche e Geologiche

marina.cocchi@unimore.it

- Context - General concepts
- Discrimination
- Discrimination based on PLS
  - PLS-DA

The classification task

- Find a criterion to assign an object (sample) to one category (class) based on a set of measurements performed on the object itself
- Category or class is a group of objects sharing similar characteristics
- In classification categories are defined *a priori* (Supervised Methods*)

- * *Clustering can highlight the presence of categories but does not use class membership as a criterion (unsupervised)*

General steps in Classification Task

→ Collect a representative set of samples for each category [calibration set]

→ Measure for each sample some (many) features/descriptors that can be good to characterise the different categories [classes]

→ Define Classification Rules that on the basis of the values of the descriptors can decide class membership for each sample

→ Evaluate the performance of the classification model [calibration / internal validation]

→ Applicate the classification rules to “new” samples [test set]

→ Evaluate the goodness of predictions [validation]

Critical Issues

→ Collect a representative set of samples for each category [calibration set] ➡ Representativeness of sampling

→ Measure for each sample some (many) features/descriptors that can be good to characterise the different categories [classes] ➡ Efficacy of descriptors

→ Define Classification Rules that on the basis of the values of the descriptors can decide class membership for each sample ➡ Classes imbalance, Outliers

Classification Methods

Many methods are available and a distinction among them can be made according to different criteria:

1. A first one is between the methods which:

focus on discrimination among the classes ➛ *discriminant analysis*

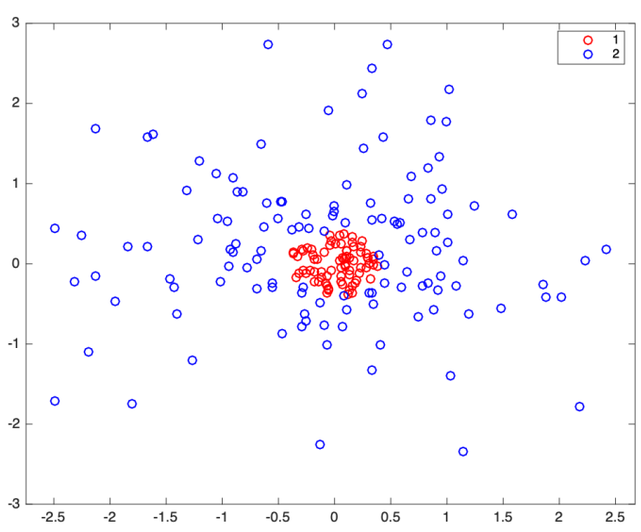
focus on modelling the single category ➛ *class modelling*

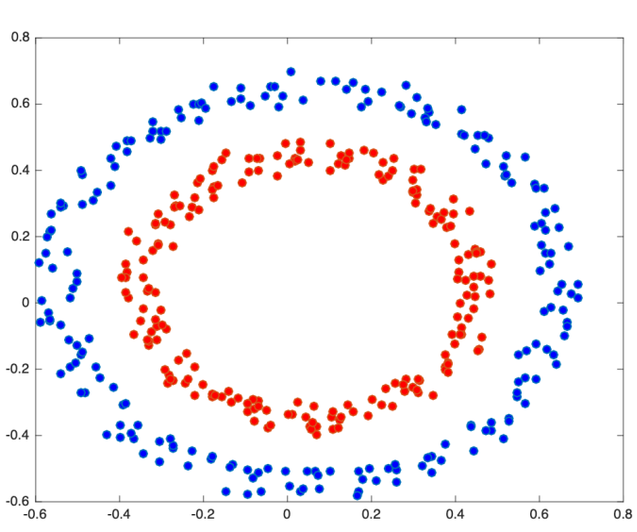
2. A second is considering the nature/type of methods:
   - based on distribution [*probabilistic]*
     - ➛ Parametric: reference statistical distributions (mean, variance,..) [LDA, QDA,..]
     - ➛ Non-Parametric: density of grouping [potential functions]
   - based on distances: ➛ between objects [e.g. KNN]
     - ➛ to class model [e.g. SIMCA]
   - based on "try and correct procedure" [automatic learning, e.g. ANN, Decision Tree]

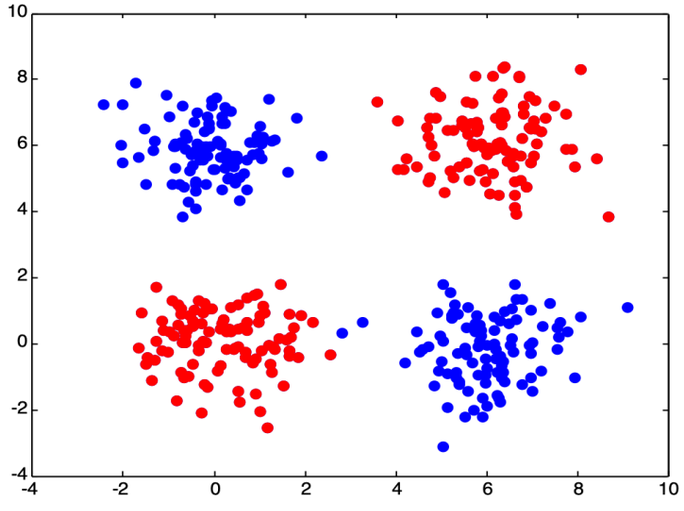
Classification Methods

Many methods are available and a distinction among them can be made according to different criteria:

3. A third one is between
   - ➛ *linear (e.g. LDA, PLS-DA, SIMCA)*
   - ➛ *non-linear (e.g. ANN, KNN)*





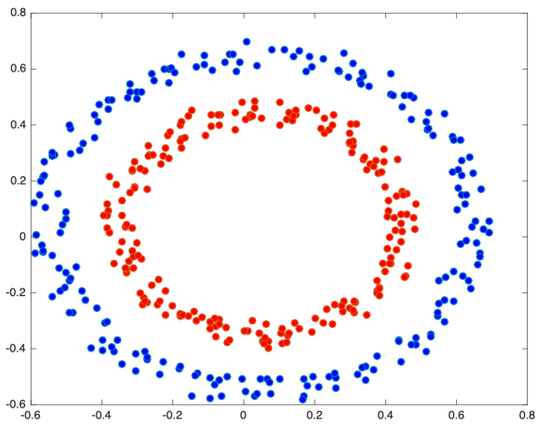


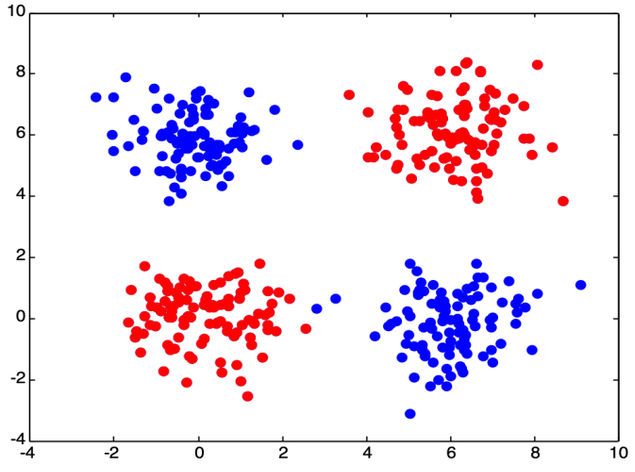
in these cases classes are not linearly separable

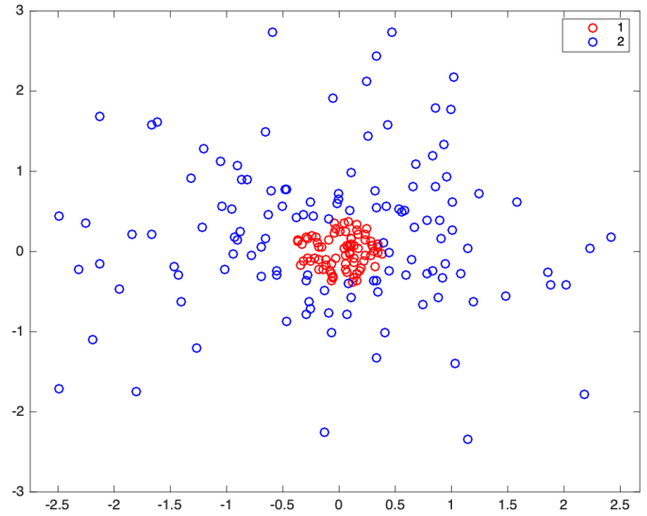
Classification Methods

linear methods may be applied to "non-linear" separable classes if

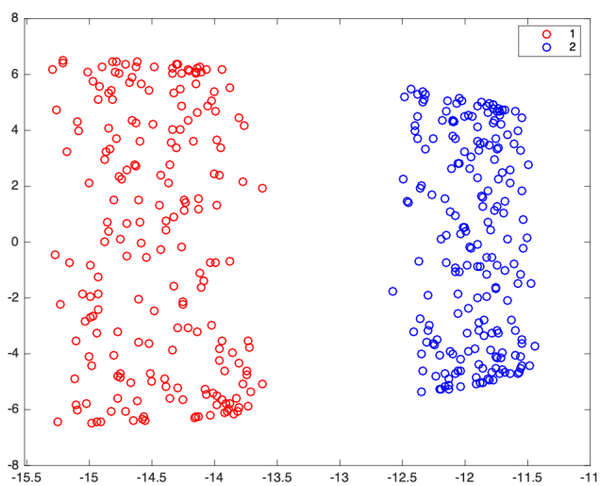
➔ *data are first mapped by non linear Kernel, e.g. Euclidean distances, Gaussian, etc…*

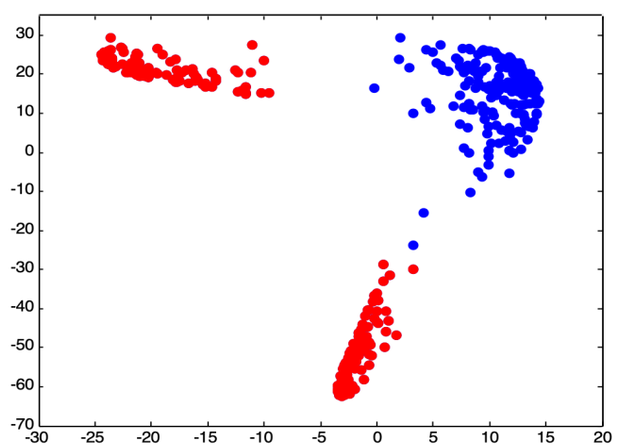




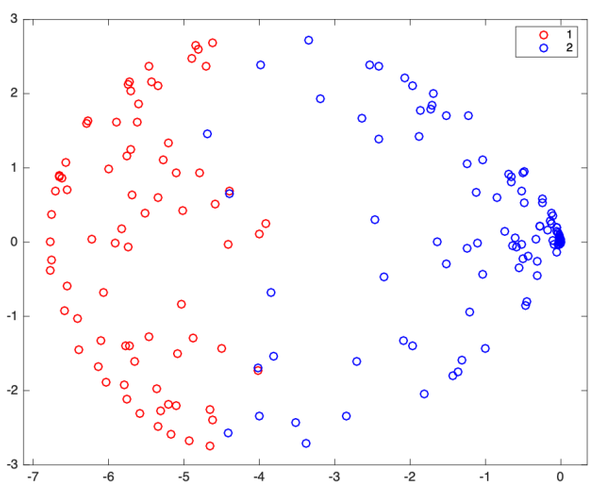


PCA on D_{Euc}





PCA on Gaussian Kernel

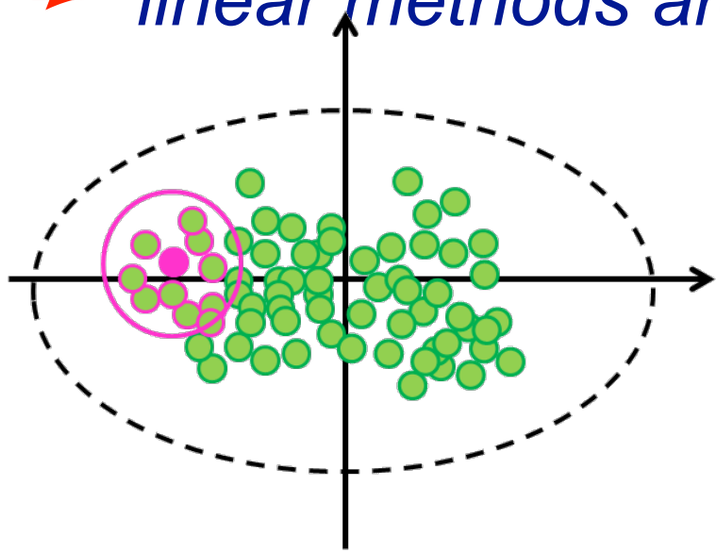


P. Zerzucha, B. Walczak, Trends in Analytical Chemistry, Vol. 38, 2012

Classification Methods

linear methods may be applied to "non-linear" separable classes if

➔ *linear methods are applied locally, e.g. Locally Weighted PLS-DA[1]*



- among the calibration samples find the N_l nearest neighbours to the sample to be assessed
- use only these N_l neighbours to build the model

(weights are applied on the basis of distance)

- predict the sample

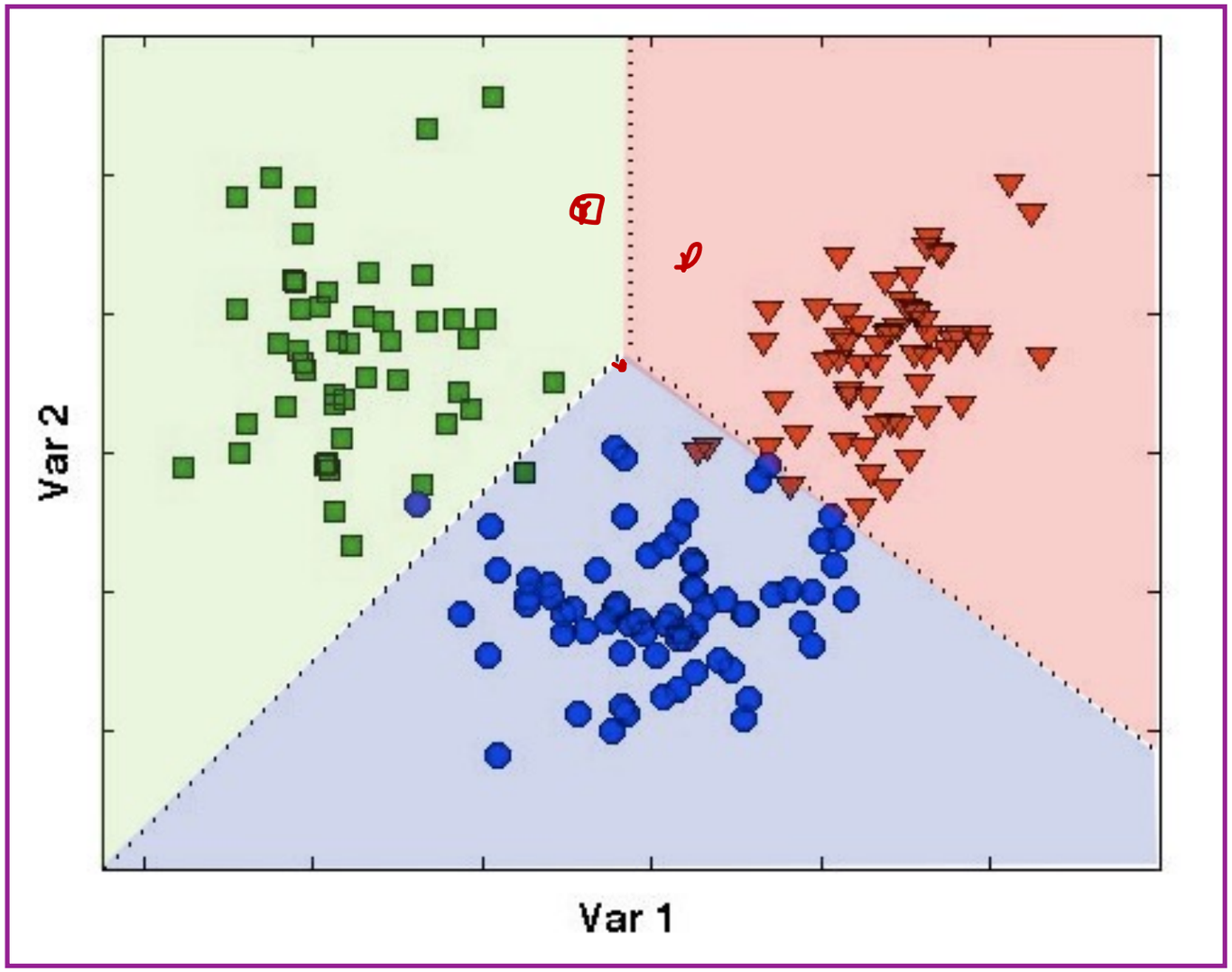
[1] M. Bevilacqua, F. Marini / Analytica Chimica Acta 838 (2014) 20–30

Discriminant classification

- The **discriminant methods** implicitly or explicitly try to identify the boundaries which separate the different classes in the multidimensional space.

- The corresponding outcome is always the classification to one of the C available categories [classes].

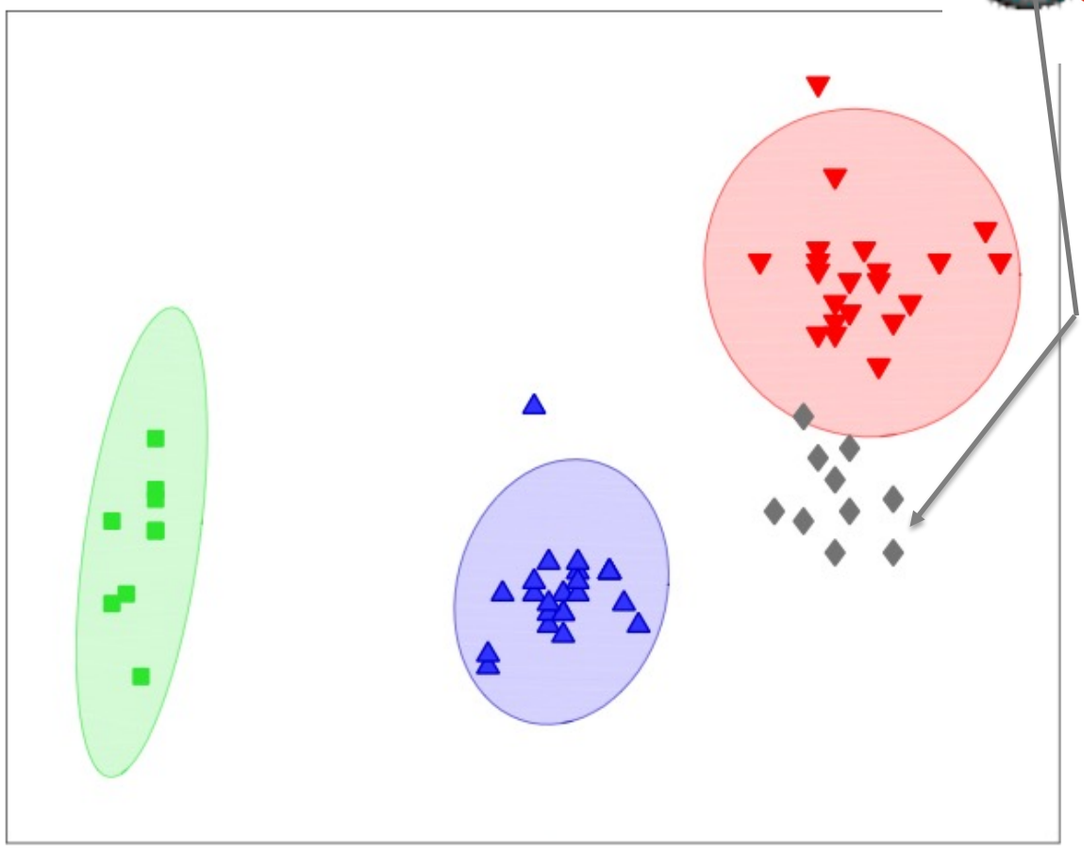
- All classes information used



✦ MODELLING AUTHENTICITY

Class Modelling eg: SIMCA, UNEQ

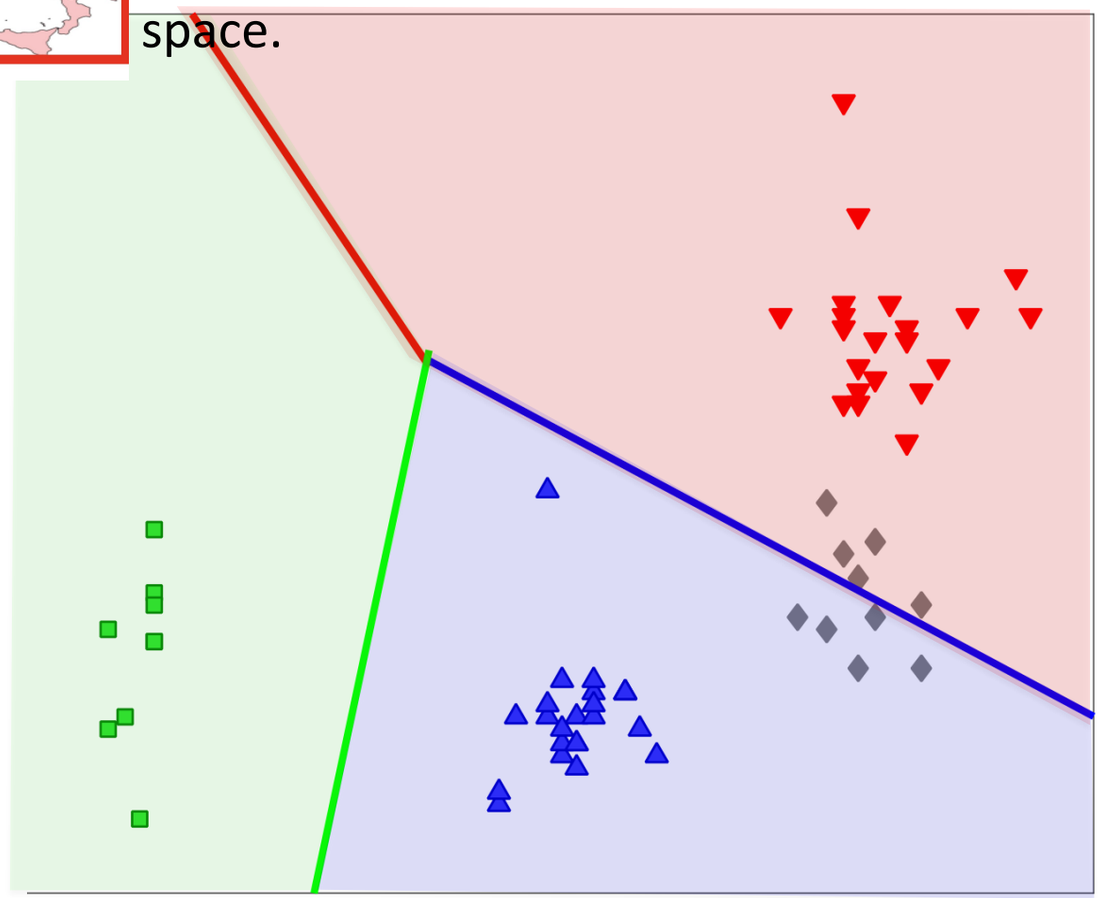
- Focus on looking for similarities among samples belonging to the same class.
- Each category is modelled individually.



- single class information is used
- e.g. use in authentication task, assessment of compliance

Discriminant classification eg: LDA, PLS-DA

- Discriminant methods aim at identifying the boundaries which separate the different classes in the multidimensional space.



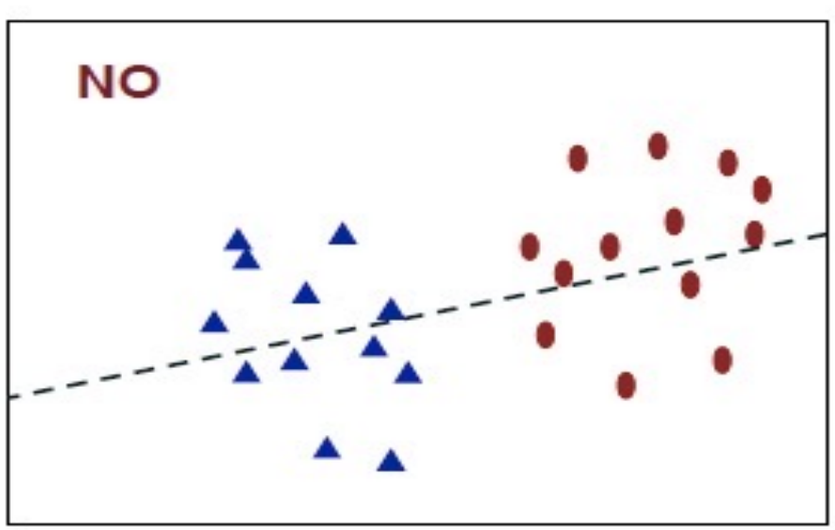
- other/others class/es information always used
- NOT appropriate to contrast a “category” from the rest

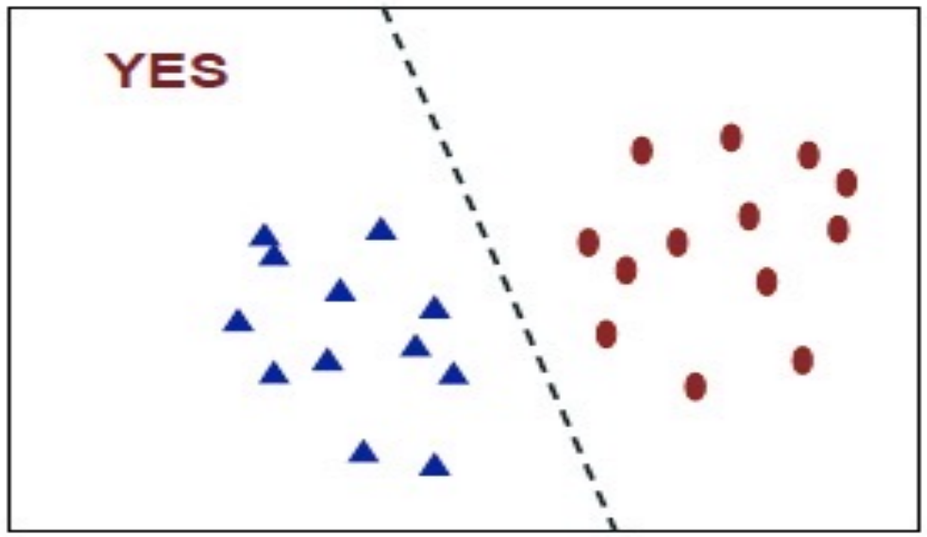
The Discriminant approach to classification

- Class boundaries are built to minimise classification error:

$$E_T = \sum_{g=1}^{G} e_g = \sum_{g=1}^{G} \frac{n_{g,wrong}}{n_g}$$

total classification error:
% of wrongly assigned samples
over all classes





The classification rule which minimises E is the so-called Bayes' rule:

"a sample is assigned to the class to which it has the maximum probability of belonging"

The Discriminant approach to classification

- Probability is defined as:

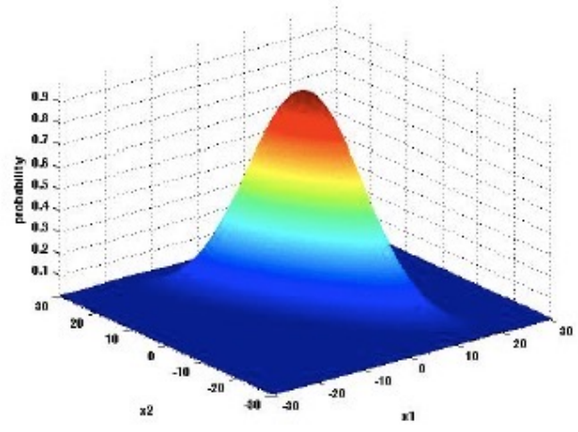
prior probability = 1/number of classes

posterior probability ← $p(g|\mathbf{x}_i) = \dfrac{P_g f(\mathbf{x}_i|g)}{\sum_k P_k f(\mathbf{x}_i|k)}$ → probability distribution of **x** for samples of class c

normalisation factor

Probabilistic method make assumption on f(x/g), which is derived from the reference distribution function, e.g. Linear and Quadratic Discriminant analysis assume a Multivariate Gaussian

$$f(\mathbf{x}_i|g) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_g|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}_i - \bar{\mathbf{x}}_g)^T \mathbf{S}_g^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_g)\right]$$

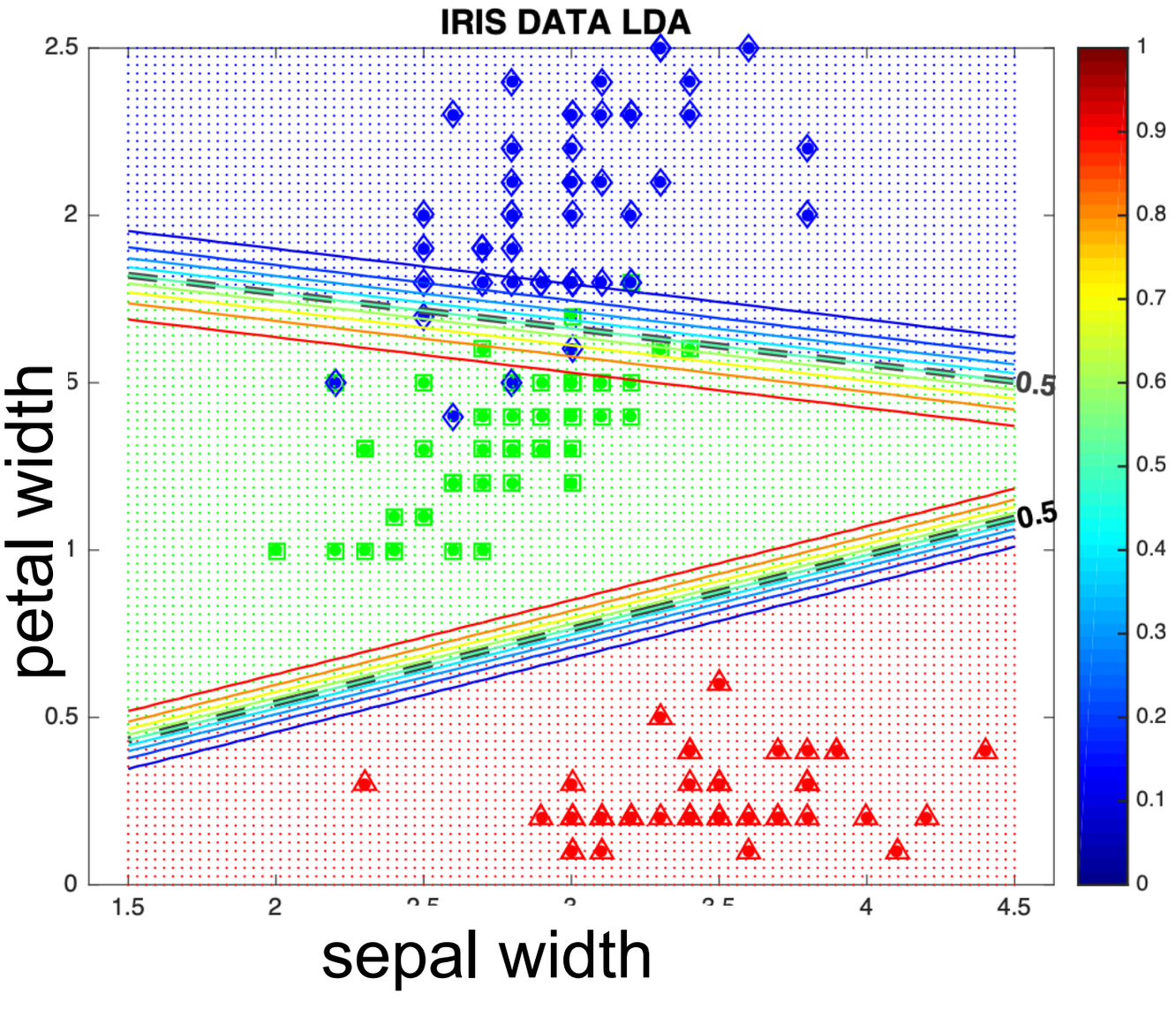


discriminant methods may differ in the way of defining and estimating probabilities.

iris data set: 3 classes

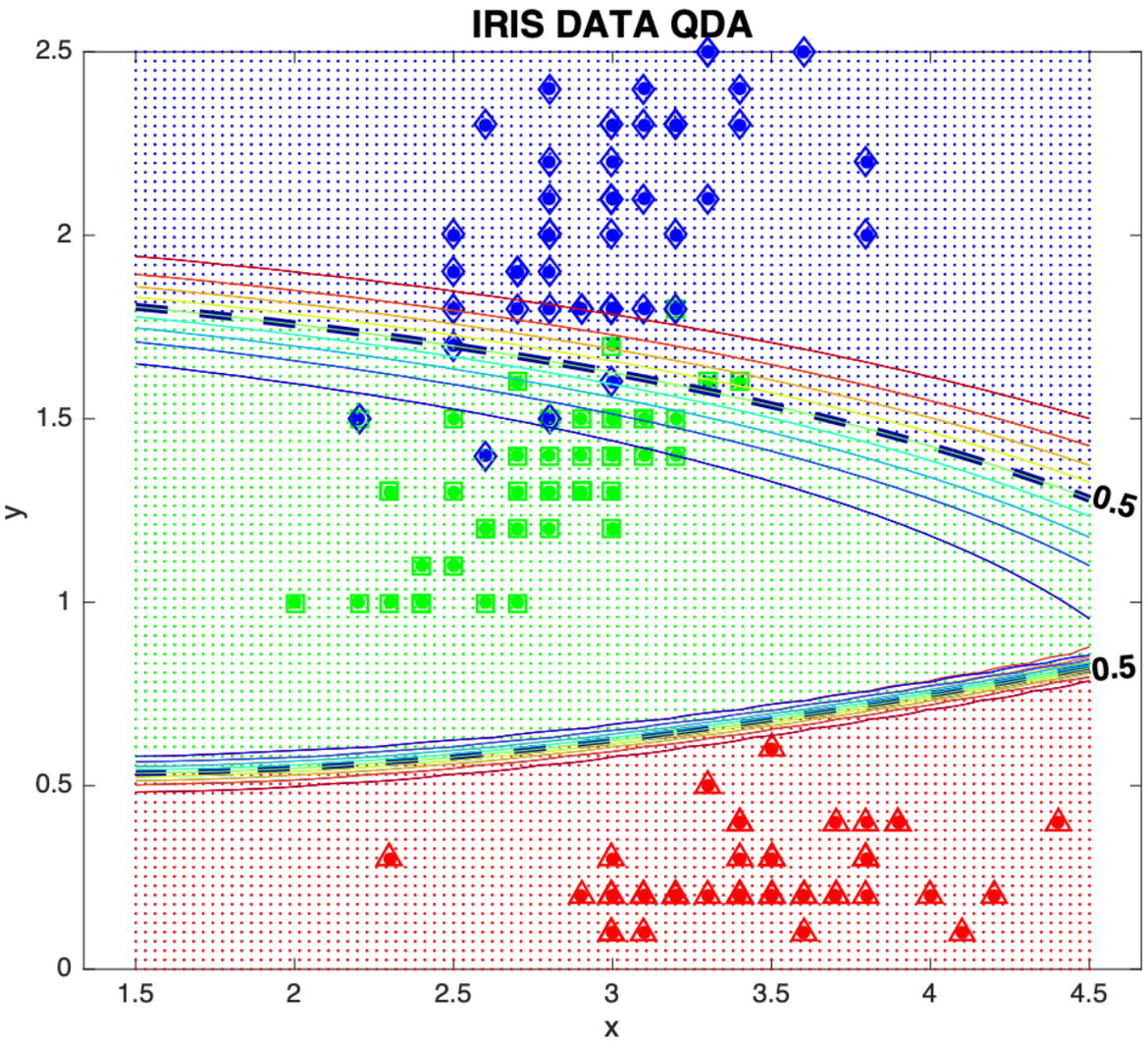
LDA

linear delimiter



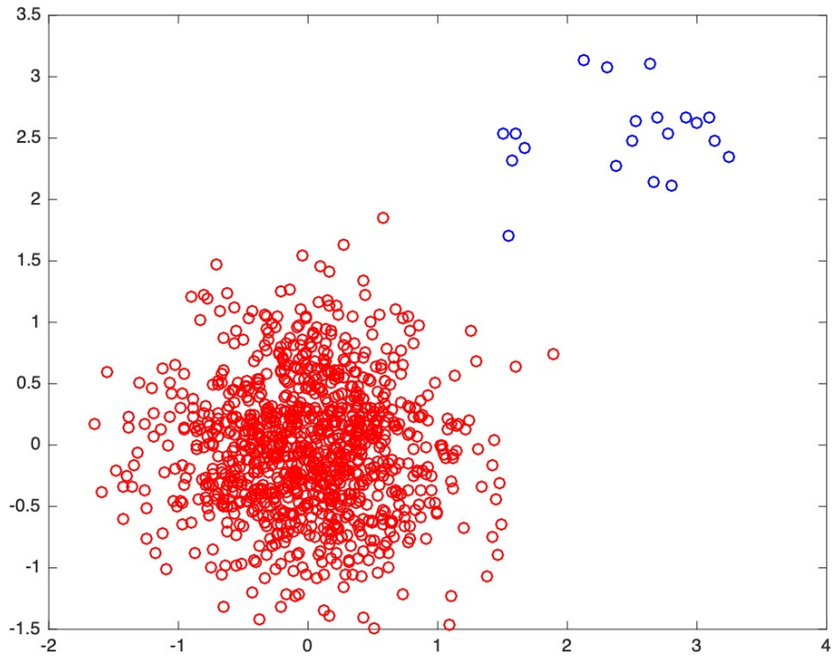
QDA

quadratic delimiter

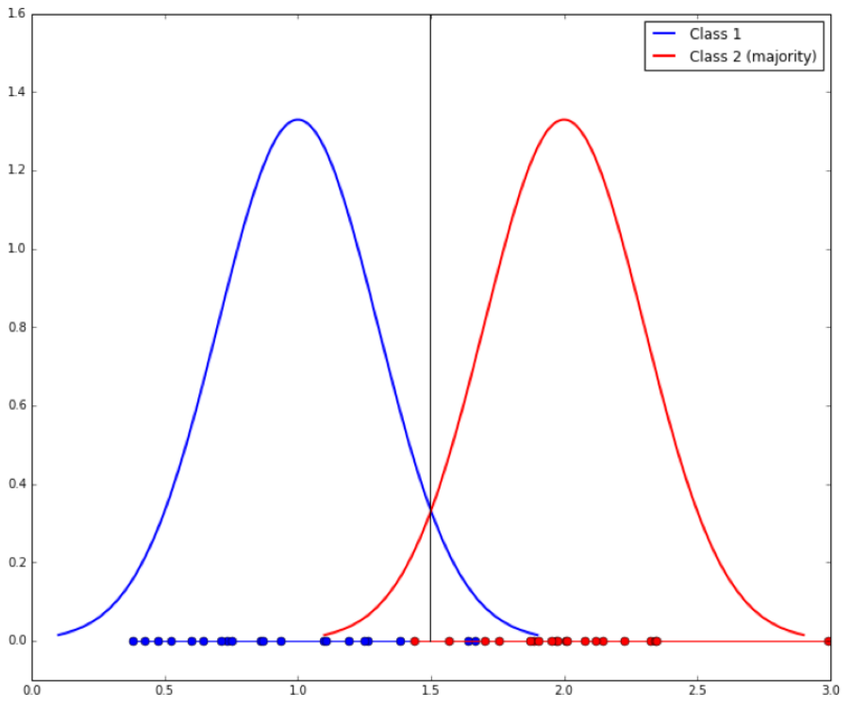


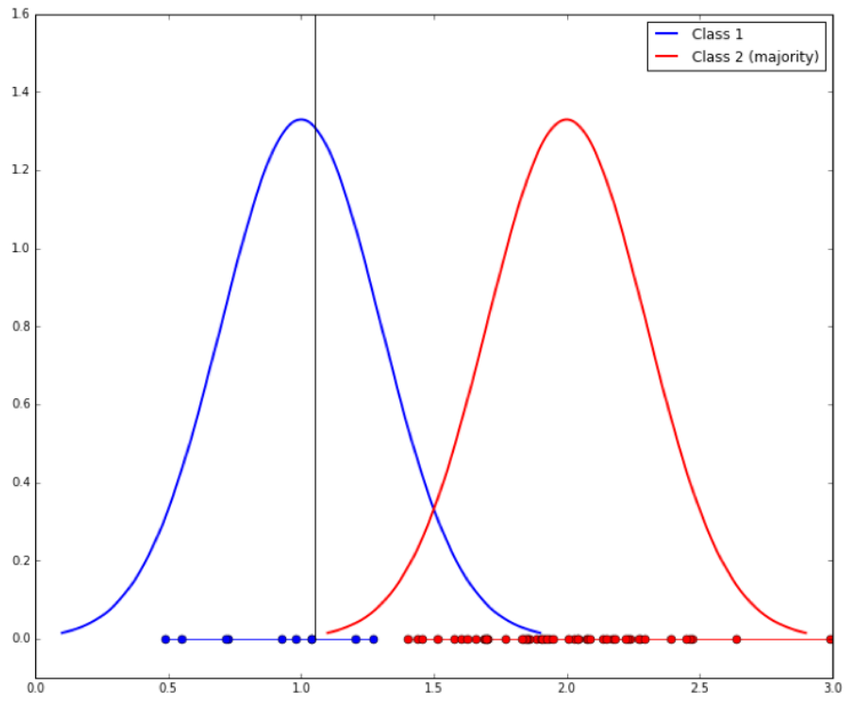
The Discriminant approach to classification

➡ Classes imbalance

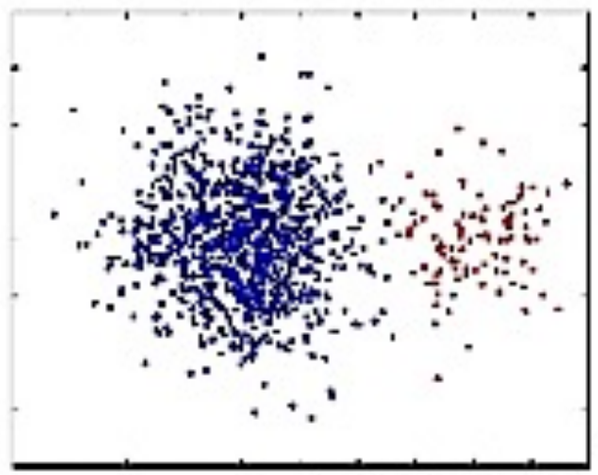


Discriminant Classifiers perform poorly in this cases



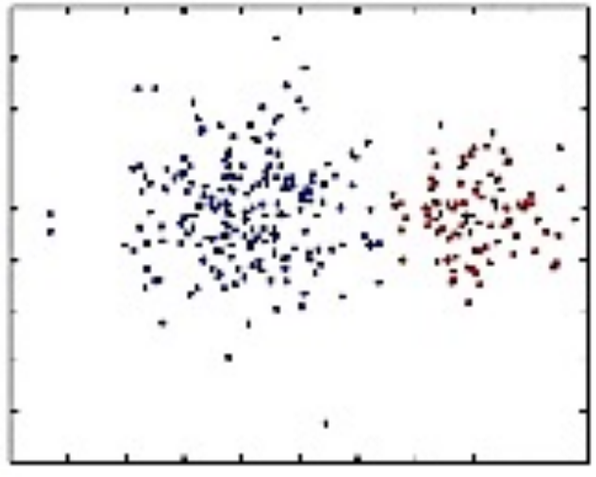


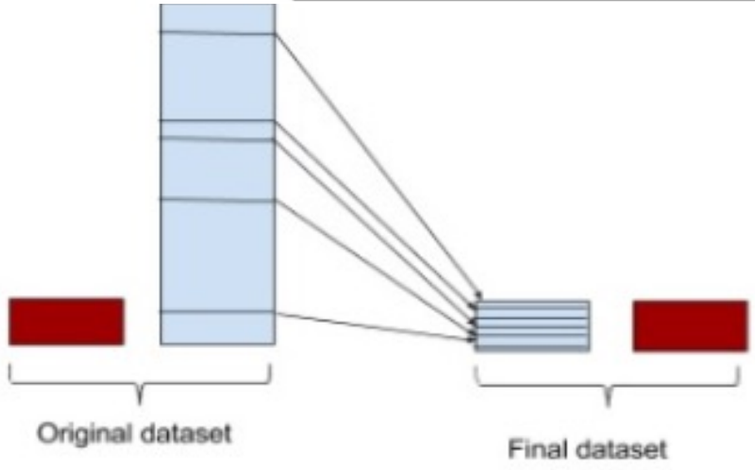
Possible solution:
Rebalance the data



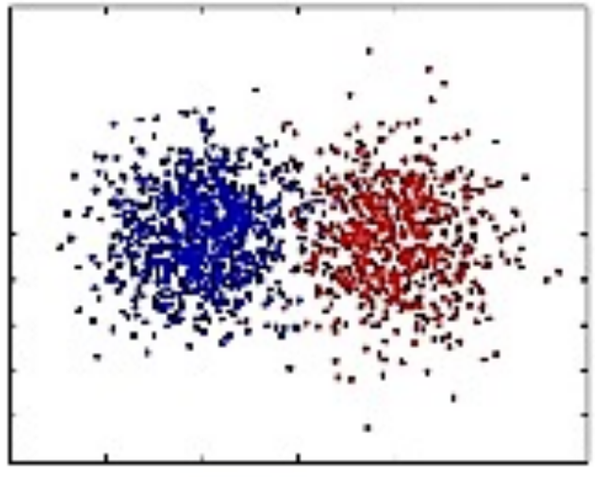
Imbalanced Data

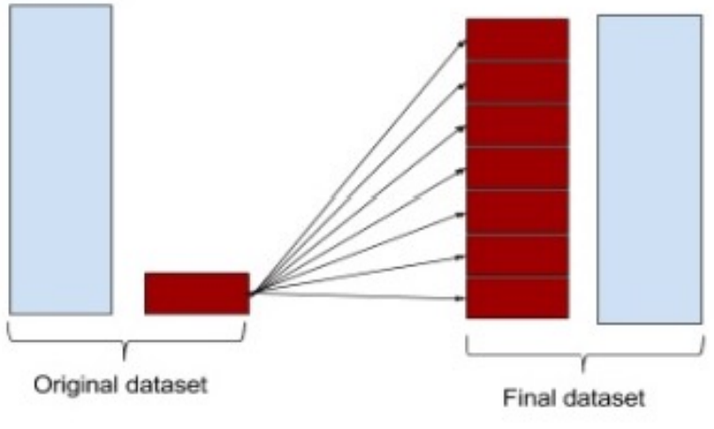
Under-sampling





Over-sampling





Performance Measures

- One way of summarizing the discriminant classification results is the so-called "confusion matrix".

		predicted		
		class1	class2	class3
true	class1	40 (95.24%)	2 (4.76%)	0 (0.00%)
	class2	1 (2.63%)	35 (92.11%)	2 (5.26%)
	class3	0 (0.00%)	3 (10.71%)	25 (89.29%)

diagonal elements:
TP (true positive)
[**TN** (true negative)]

extra-diagonal elements:
FN (false negative for "true")
FP (false positive for "predicted

- %Non-error rate (%NER) or %Correct classification rate (%CCR):

$$\%NER(g) = \frac{n_{g,correct}}{n_{g,tot}} \times 100 \qquad \%NER(tot) = \frac{\sum_{g=1}^{G} n_{g,correct}}{n_{tot}} \times 100 \quad = 100/108=92.6\ \%$$

⇩

- **Sensitivity**:
- % samples from class i, correctly classified as i

Class 1: 40/42 (95.24%); Class 2: 35/38 (92.11%); Class 3: 25/28 (89.29%)

- **Specificity:**
- % samples from class ≠i, correctly classified as not i.

Class 1: 65/66 (98.48%); Class 2: 65/70 (92.86%); Class 3: 78/80 (97.50%)

Performance Measures

to asses overall performance

CLASS MODELLING

Efficiency = $\sqrt{(\text{Sensitivity}*\text{Specificity})}$

Discriminant Classification

A sample is always assigned to one of the modelled classes

--> Sensitivity and Specificity are not independent

% Correct Classification Rate (%CCR)

LDA shares the same limitations/assumptions of MLR

LDA

- independent variables
- X variables are exact
- residual are normal

Needs more objects than variables!

Possible Solution:

- To overcome this limitation data reduction can be operated before applying discriminant analysis, e.g. by using principal component analysis, or reformulating the discriminant classification problem in regression terms as in extended canonical variate* (ECVA) or in Discriminant PLS

* ECVA utilizes Partial Least Squares regression as an engine for solving an eigenvector problem involving singular covariance matrices.

Discriminant PLS (DPLS, PLS-DA)

1. Regression on dummy variables

Common Frame

Y coding: as many dummy y-variables (1/0) as classes

Fit a PLS2 model

Predict Y values for future/unknown sample $\hat{\mathbf{Y}} = \mathbf{XB}$

2. Classification rules

True Discriminant (DPLS)

- Predict the dummy values ($\mathbf{y}_{i\_pred}$) and assign the object to the group with highest predicted value, k = argmax($\mathbf{y}_{i\_pred}$)*
- Use LDA, QDA, etc.. on PLS scores or $\mathbf{y}_{i\_pred}$ (not redundant, i.e. n° classes -1)

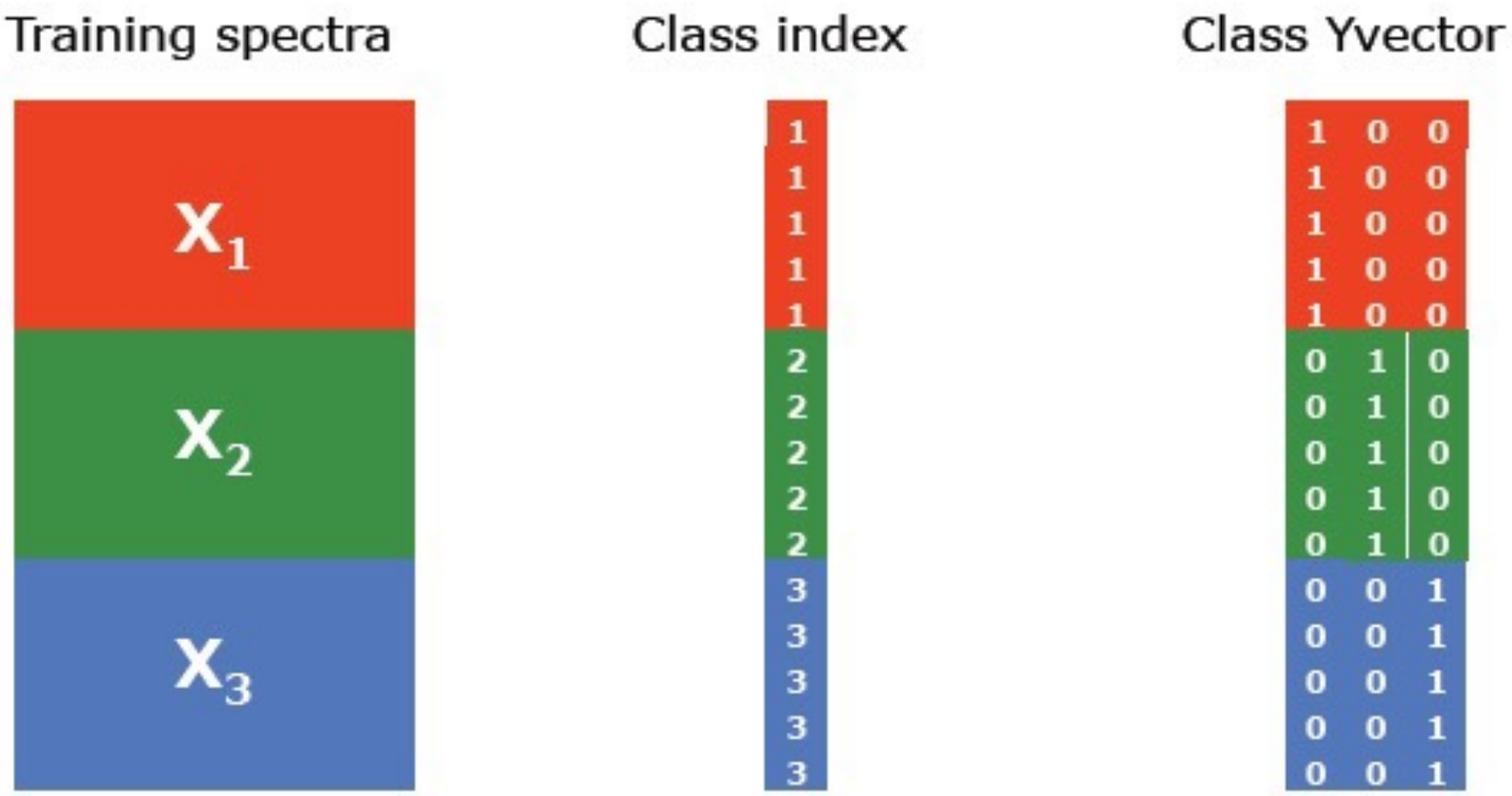
*with more than 2 categories can be sub-optimal (masking effect Hastie[1])

Hybrid (if more than 2 categories)

- Define an acceptance threshold on the basis of predicted the values ($\mathbf{y}_{i\_pred}$)

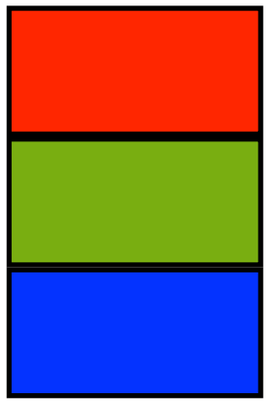
[1] Hastie T. et al. The element of statistical learning, Springer: New York,2001

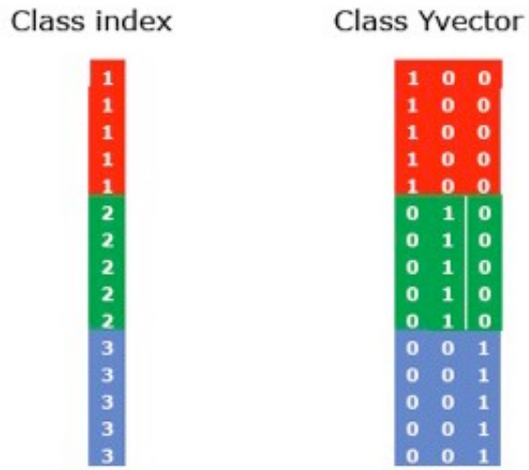
- Classification is accomplished through regression of X against a binary matrix containing class-membership information



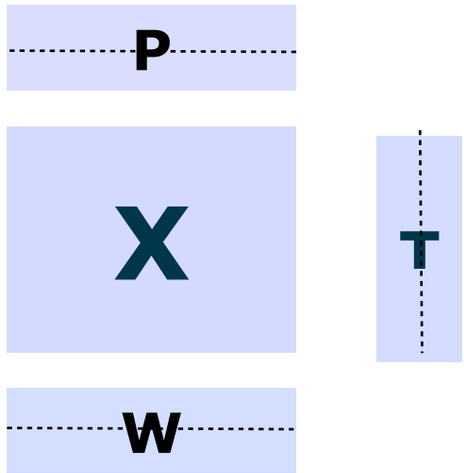
Regression on dummy variables (DPLS, PLS-DA)

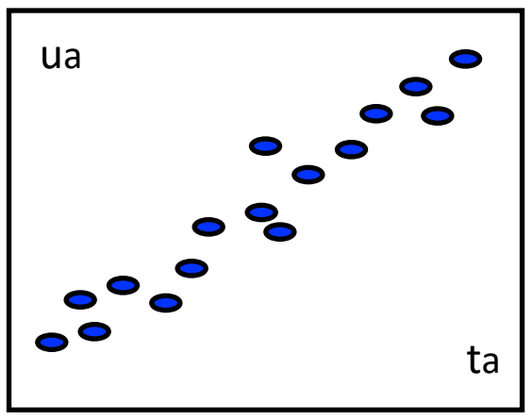
Y coding: as many dummy y-variables as classes 1/0

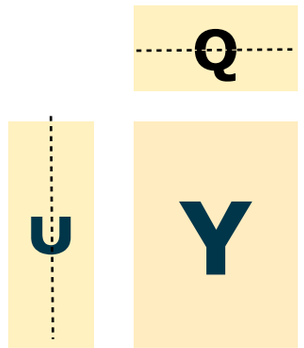




Fit a PLS2 model







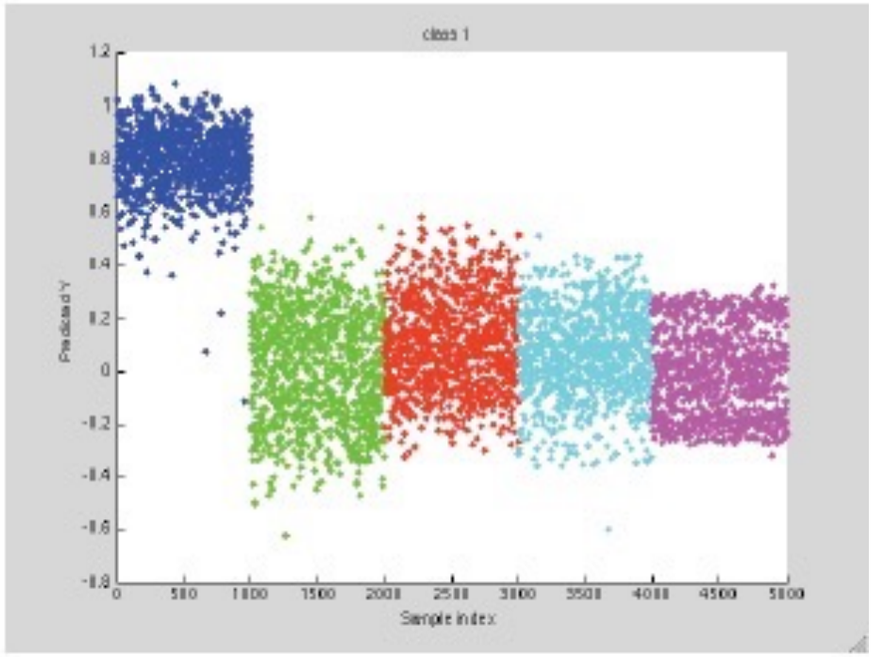
$$X = TP^T + E \qquad U = bT \qquad Y = UQ^T + F$$

$$B_{PLS} = W(P^TW)^{-1}\mathrm{diag}(b)Q \qquad \widehat{Y} = XB_{PLS}$$

Classification based on LV: Discriminant PLS

- Predicted y is real-valued:

"true" y			predicted y		
1	0	0	1.03	0.09	-0.10
1	0	0	0.68	0.21	0.08
1	0	0	0.99	-0.10	0.01
1	0	0	0.96	0.18	-0.14
1	0	0	0.79	0.02	0.25
0	1	0	0.14	0.94	0.07
0	1	0	-0.01	1.12	0.12
0	1	0	0.08	0.89	-0.02
0	1	0	0.33	0.45	0.25
0	1	0	0.15	0.72	0.06
0	0	1	0.13	-0.18	0.85
0	0	1	0.21	0.17	0.56
0	0	1	-0.09	0.32	0.69
0	0	1	0.12	0.06	1.01
0	0	1	0.02	-0.03	0.98



Possible Classification rules

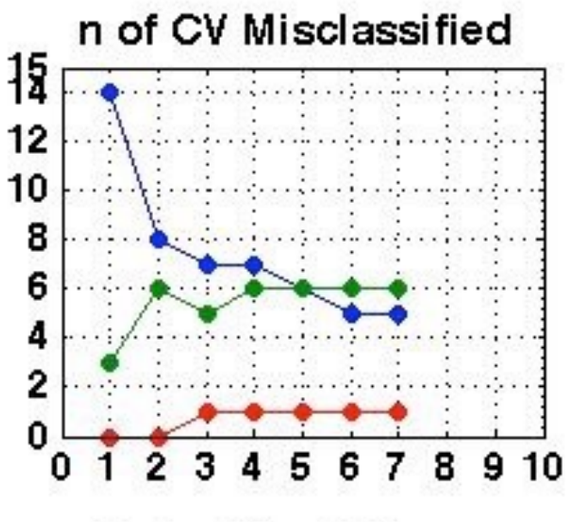
True Discriminant

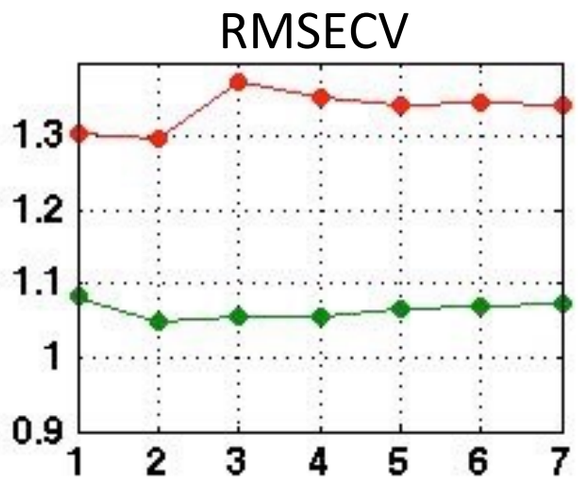
- Sample is assigned to the class corresponding to the highest y component

Hybrid

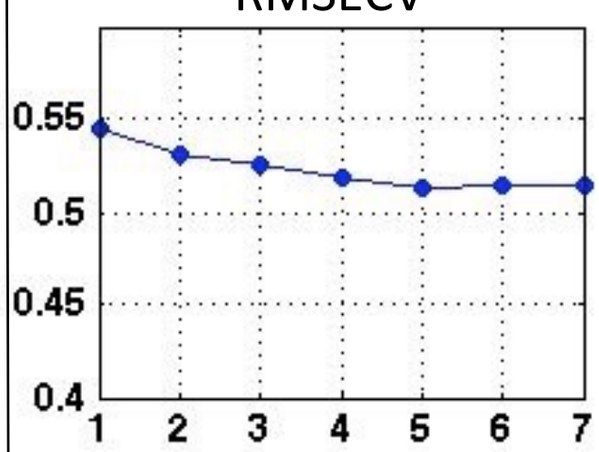
- When there are only two classes, threshold is set at 0.5
  - A Y-predicted threshold value is defined for each class which minimise the CV classification error

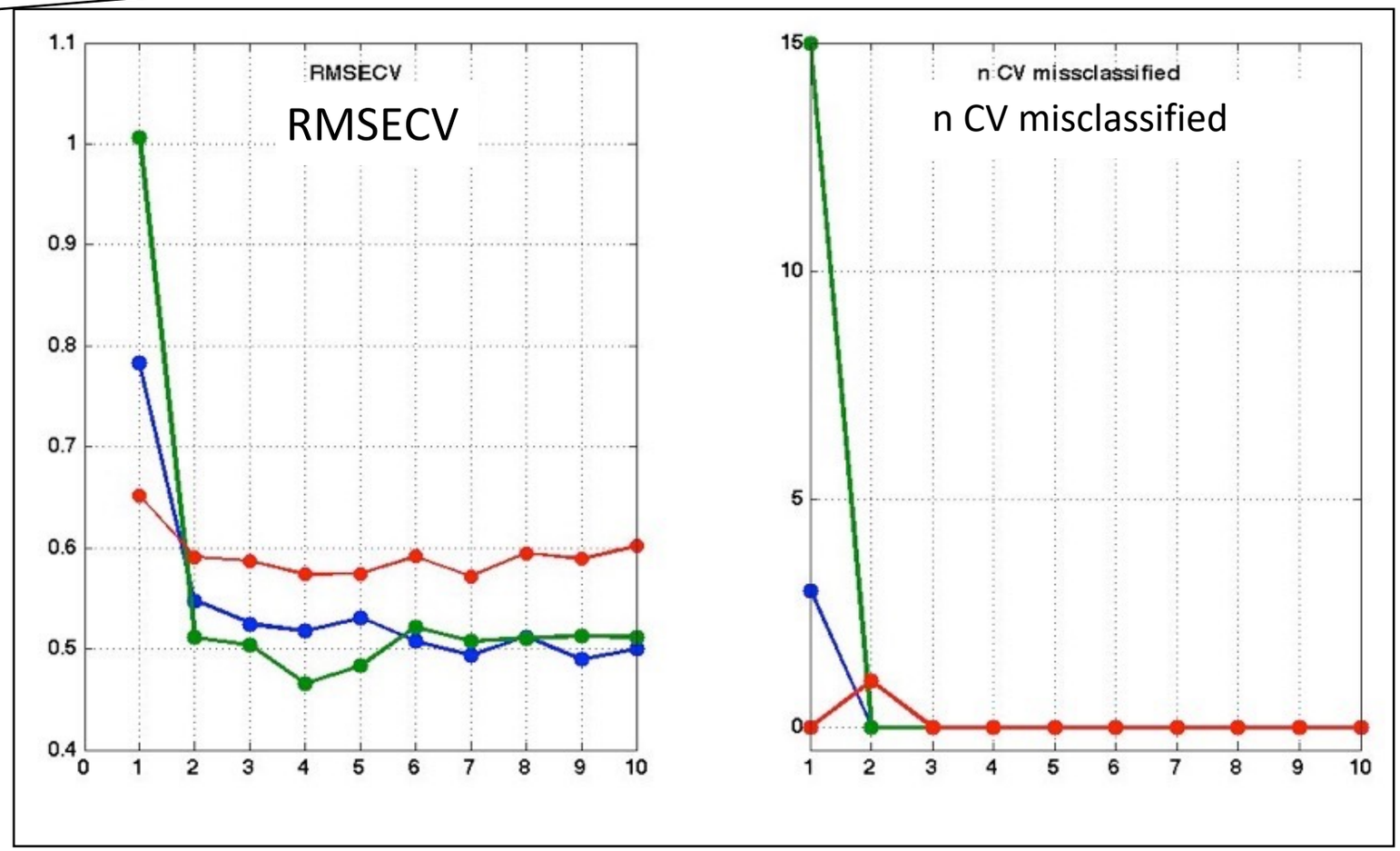
I.dimensionality (How many PLS components)?





Not always RMSECV has the same trend as CV error
Remember Ypred are continuos
Yclass are dummies (1/0)





- **Interpreting Variable importance**

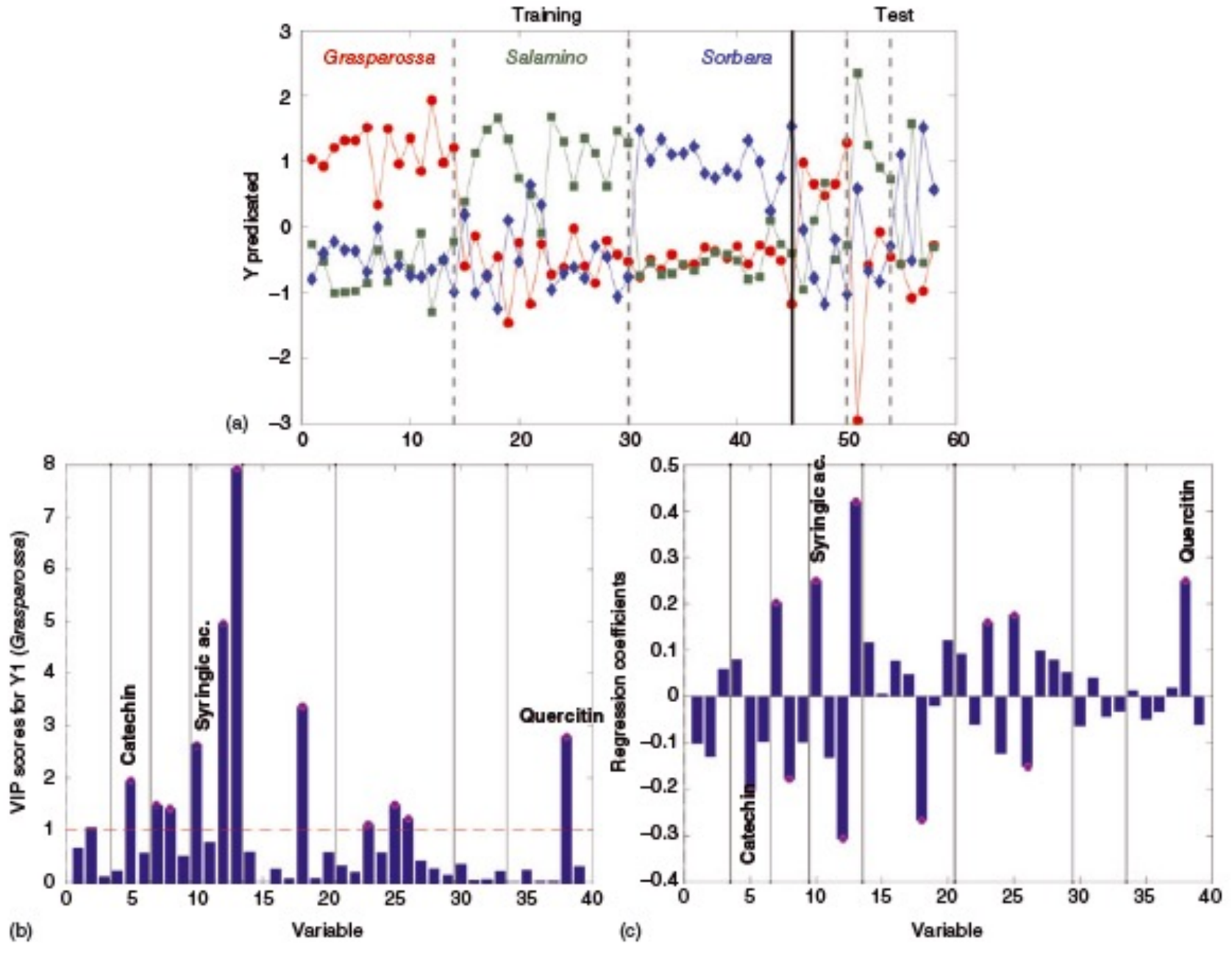


Figure 17 Results of PLS-DA model built by using the peak areas of the 39 resolved components by MCR (considering all elution windows). (a) Estimated Y's values vs. samples number, each class in different colors and symbols as reported in the figure. Dashed vertical gray lines separate the samples belonging to different classes, solid black line separate training from test set samples. VIP scores and regression coefficients line plots are shown in (b) and (c), respectively.

an Example

Data set

Animal Feed (flour):
bovine (45)
fish (43)
chicken (45)

MIR spectroscopy

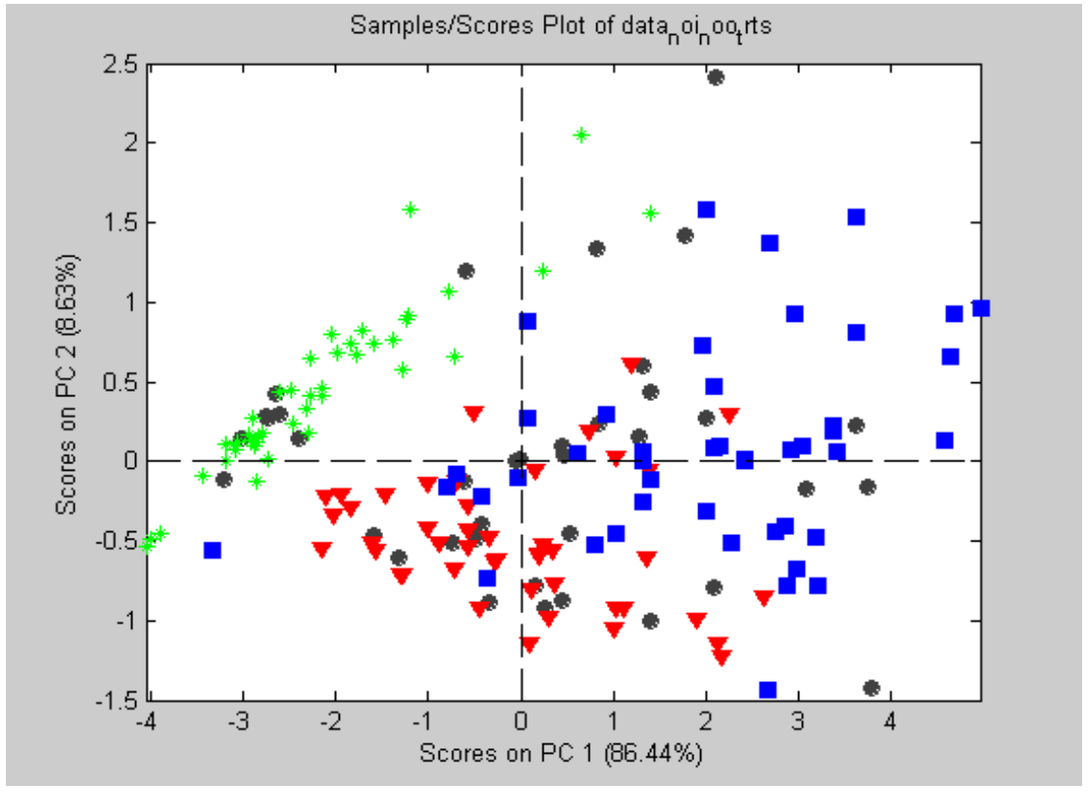
after drying
FT-MIR (850 – 1250 cm^{-1}):

Preprocessing: mean center;

Aim

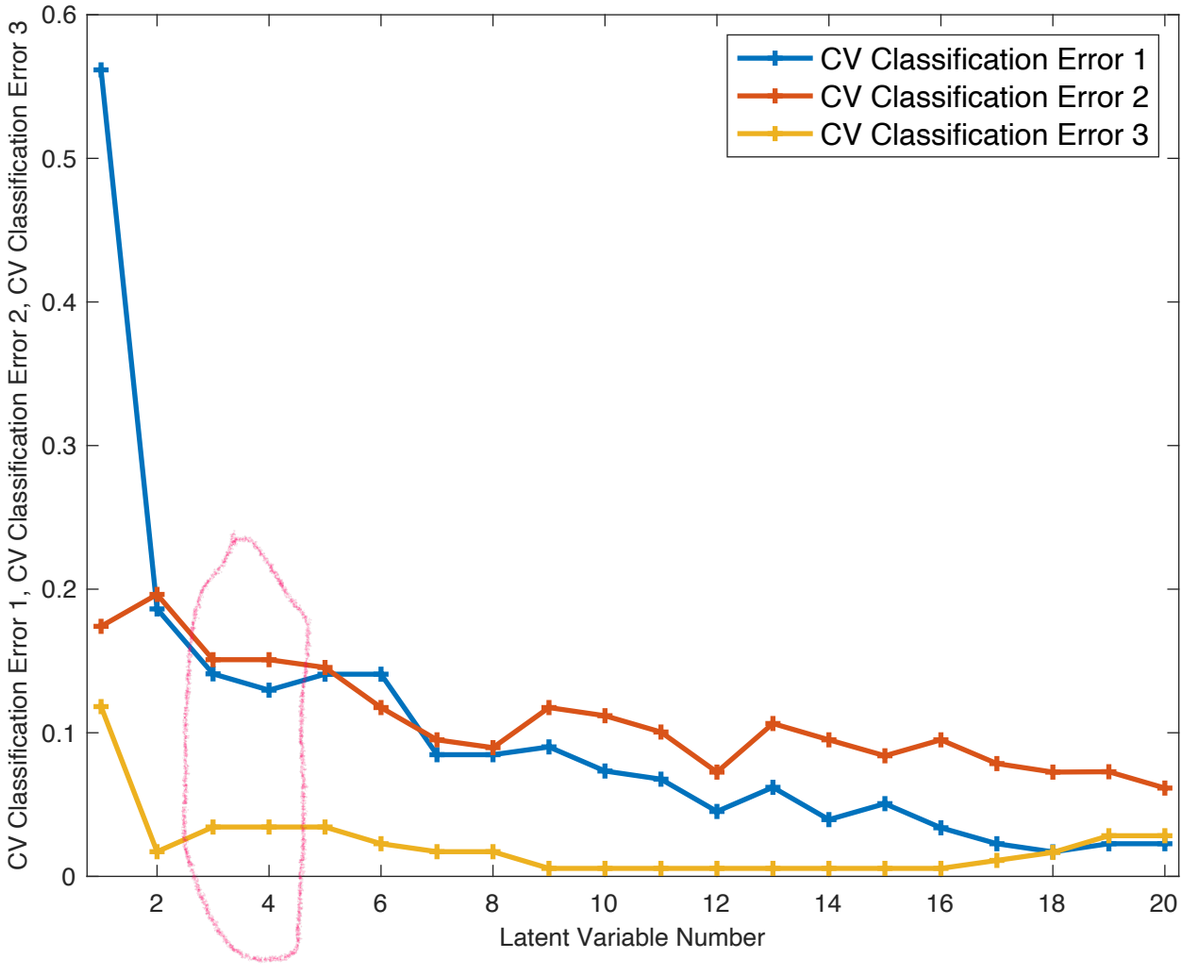
Assessing the animal species of the feed flour

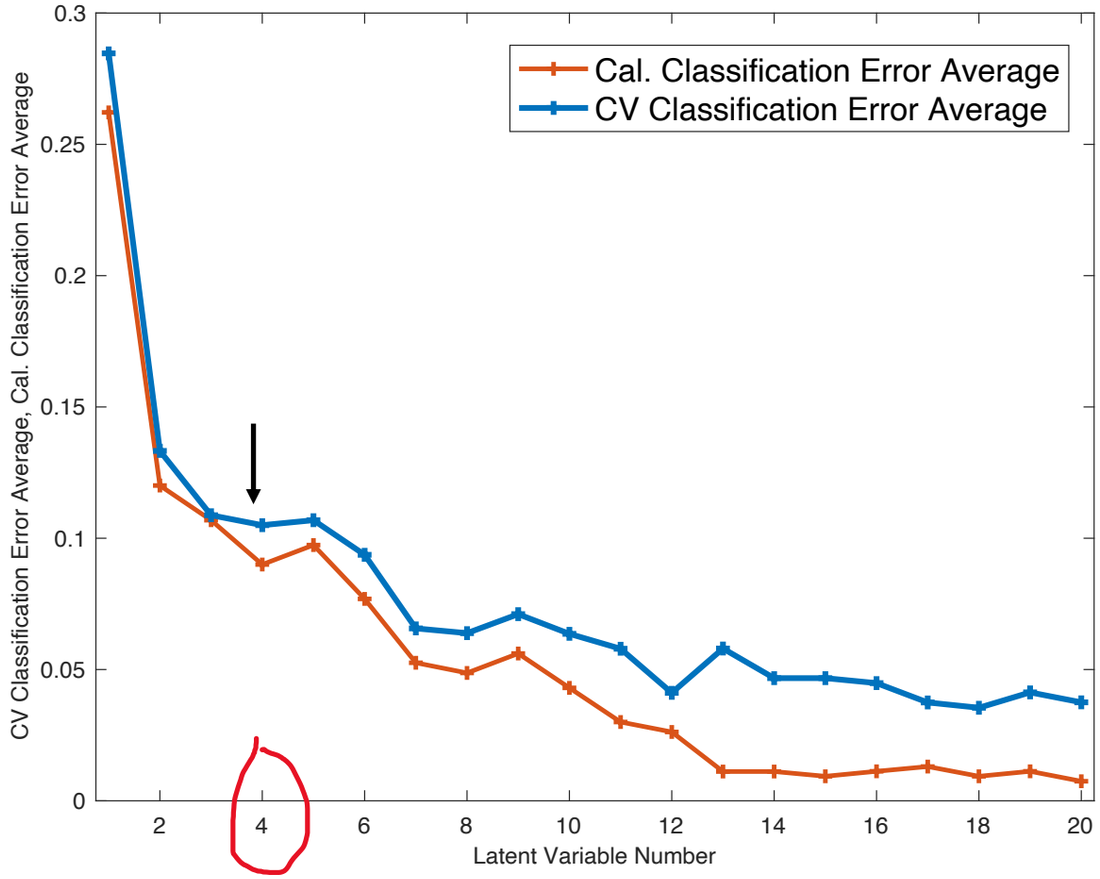
Explorative PCA results



Classification based on LV: Discriminant PLS

Animal Feed (flour):
bovine (45)
fish (43)
chicken (45)





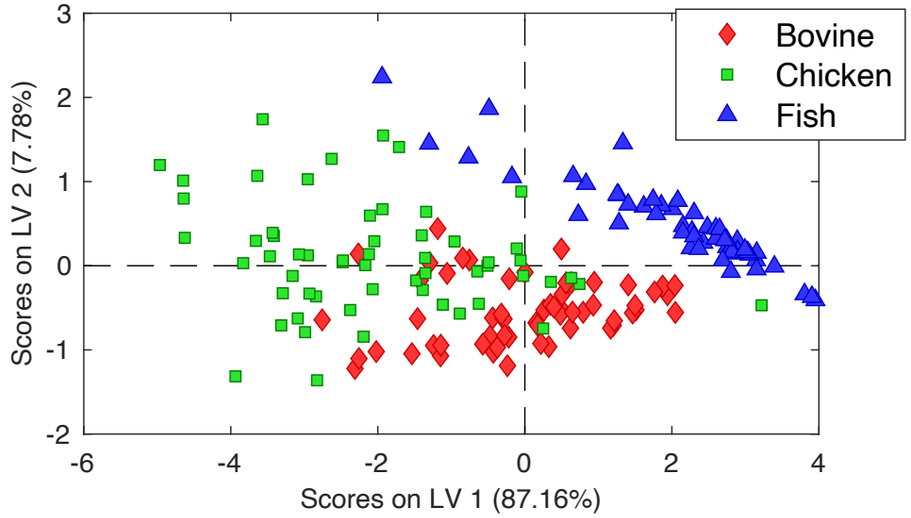
3 (4) LVs

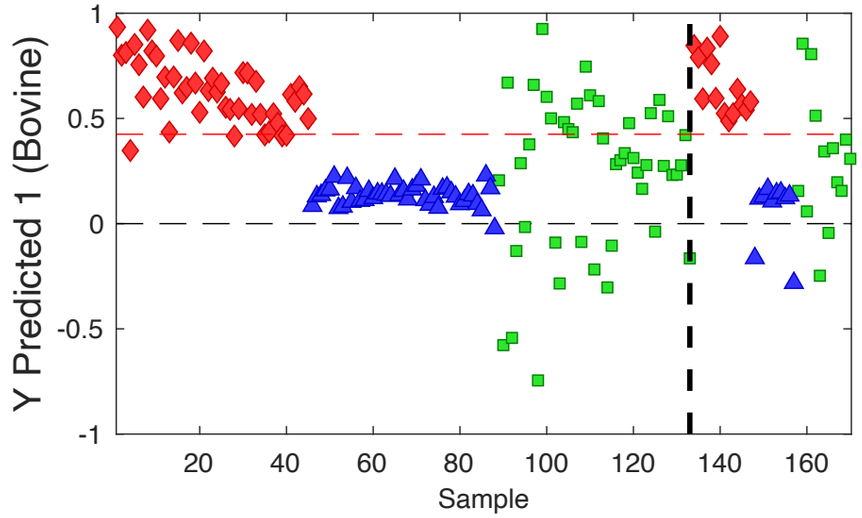
Classification based on LV: Discriminant PLS

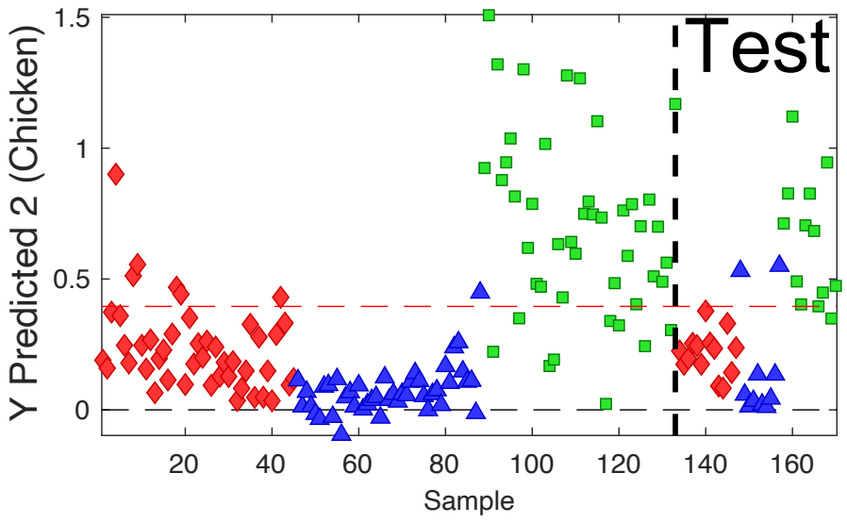
Animal Feed (flour):
bovine (45)
fish (43)
chicken (45)

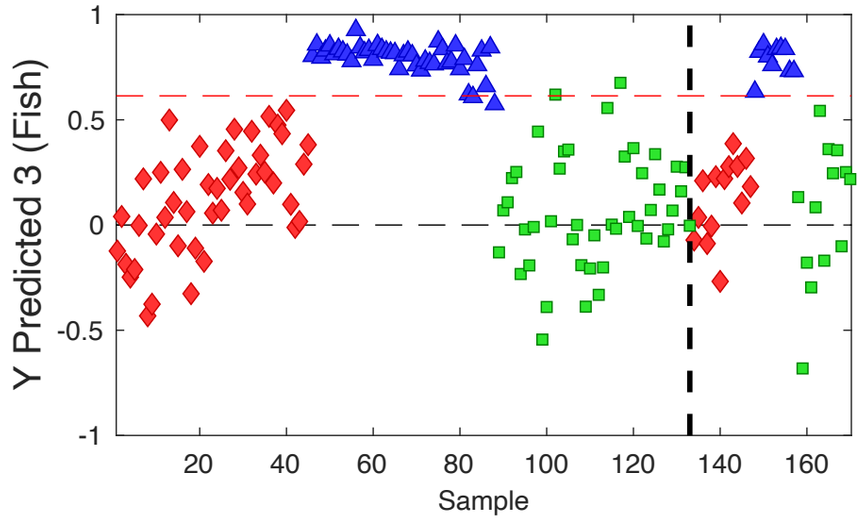
3 LVs

Bovine is not specific for chicken



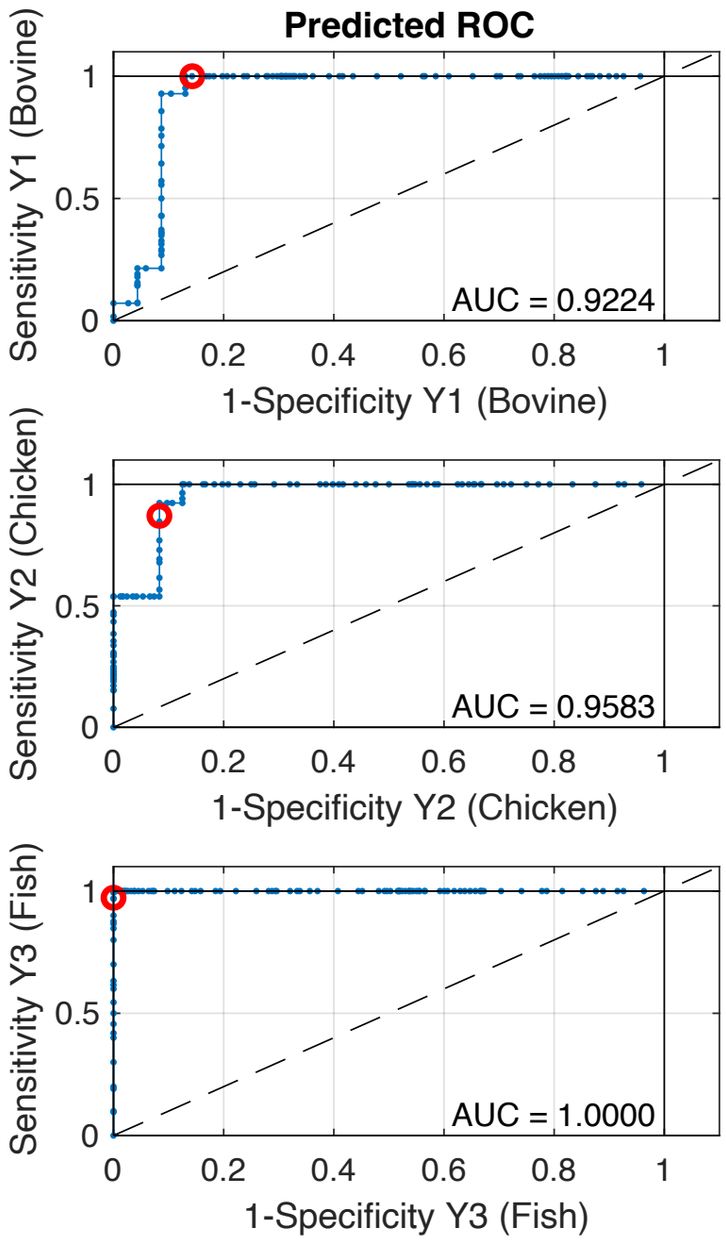


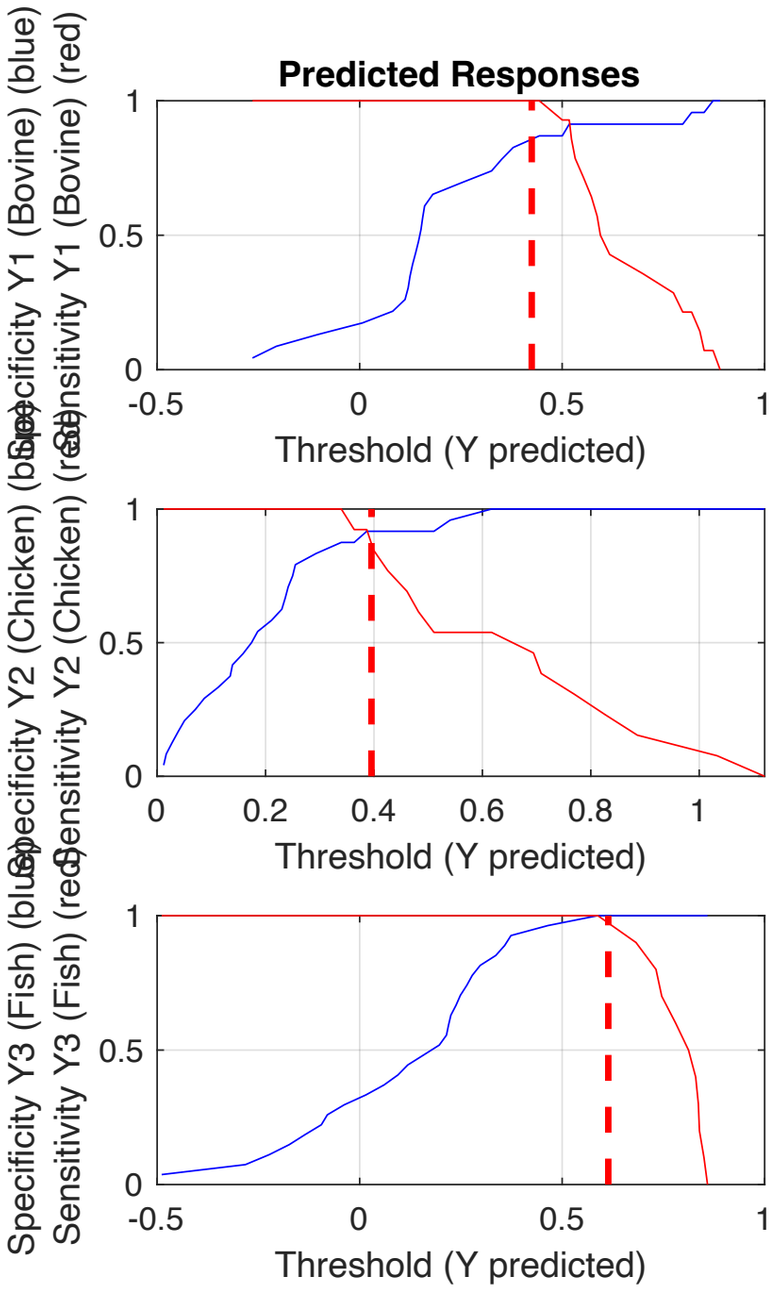




Animal Feed (flour):
bovine (45)
fish (43)
chicken (45)

How threshold is established

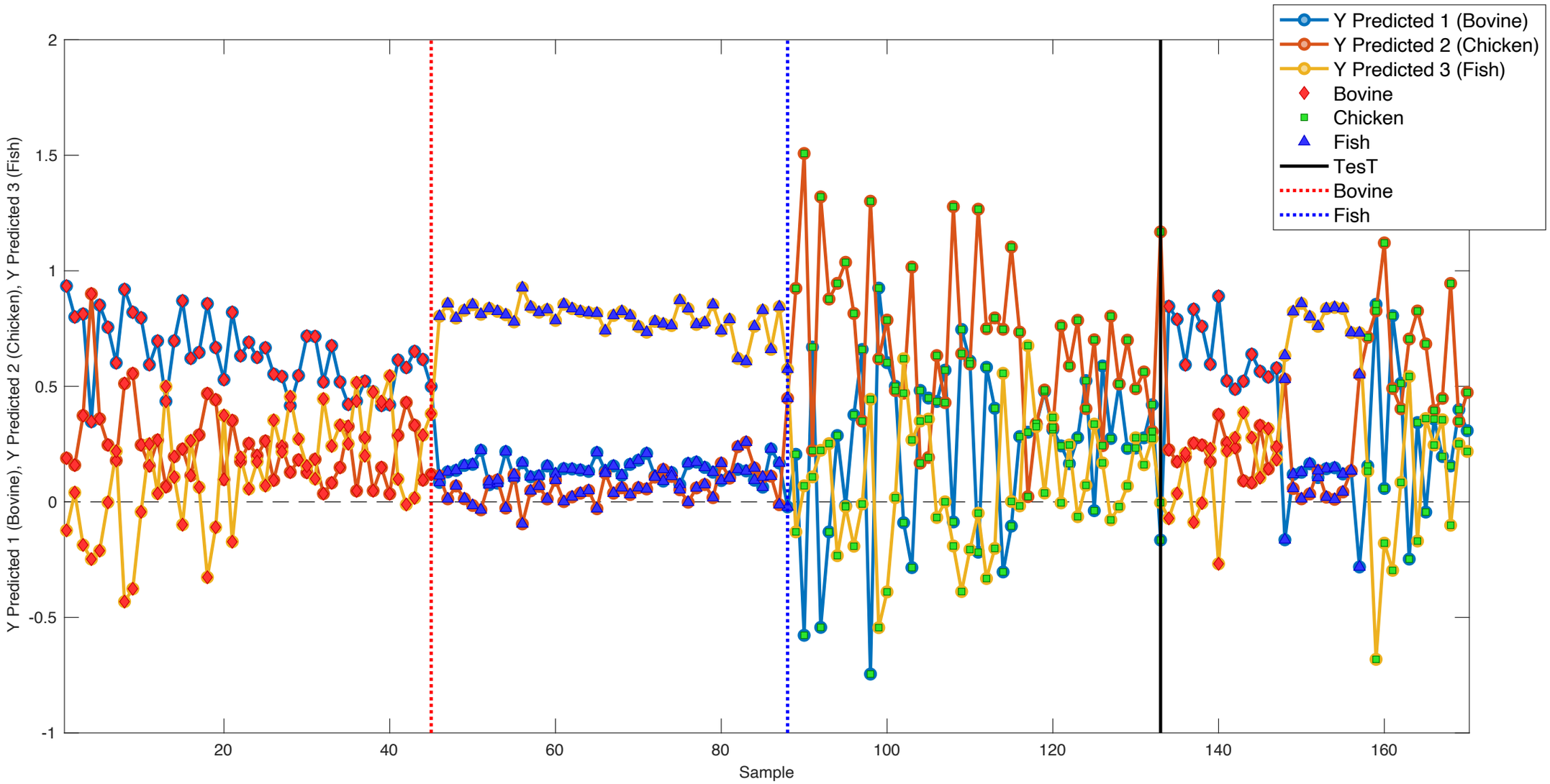




Classification based on LV: Discriminant PLS

Animal Feed (flour):
bovine (45)
fish (43)
chicken (45)

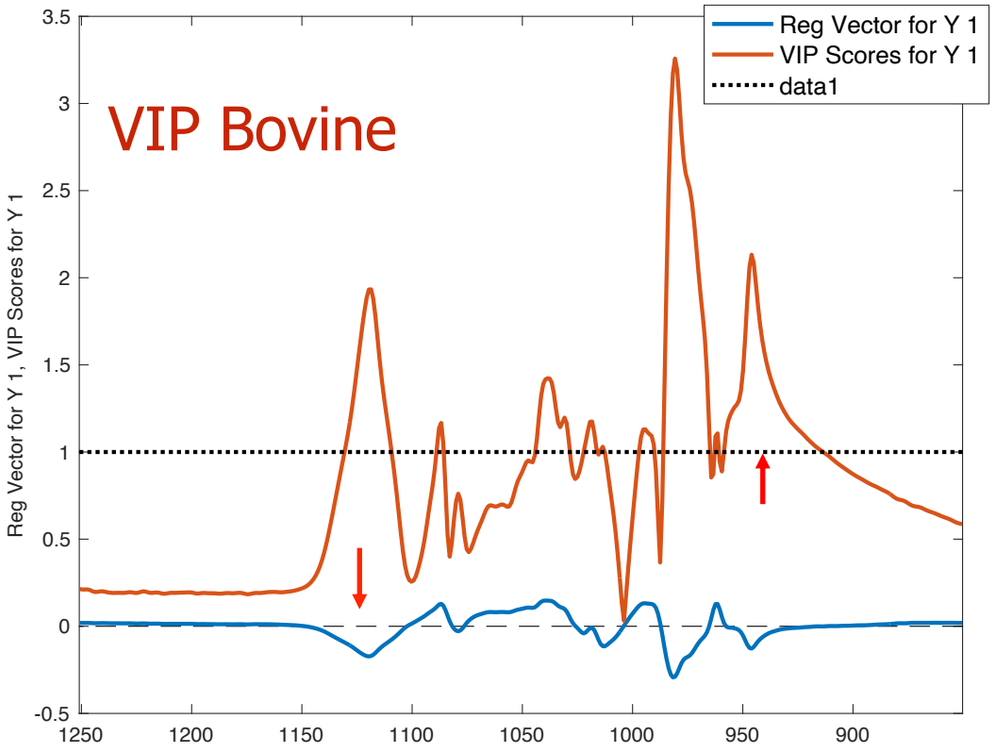
max(Ypred) True discriminant

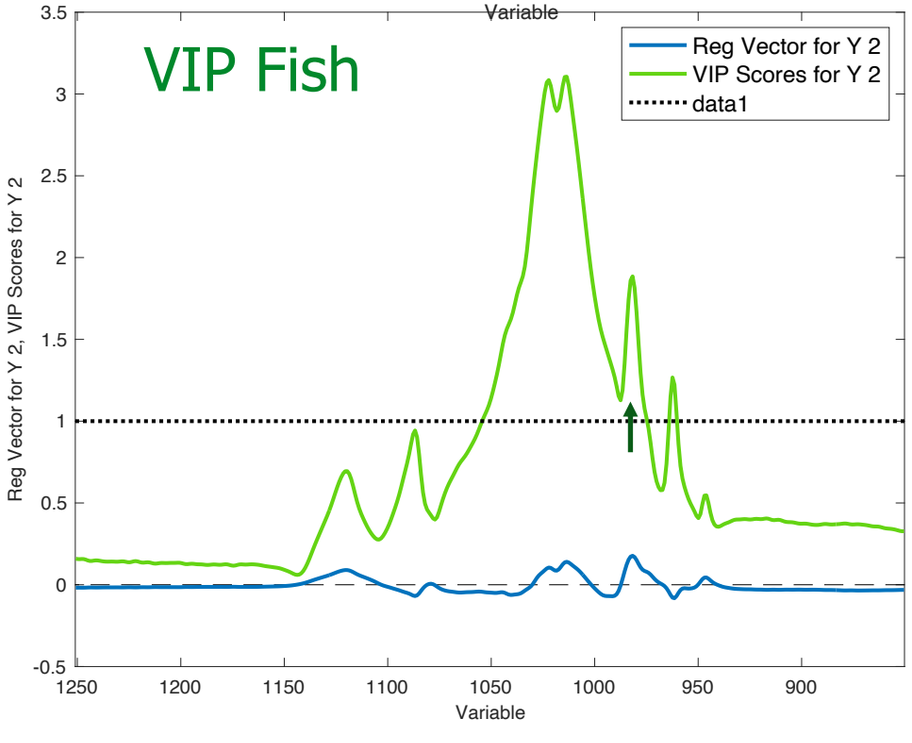


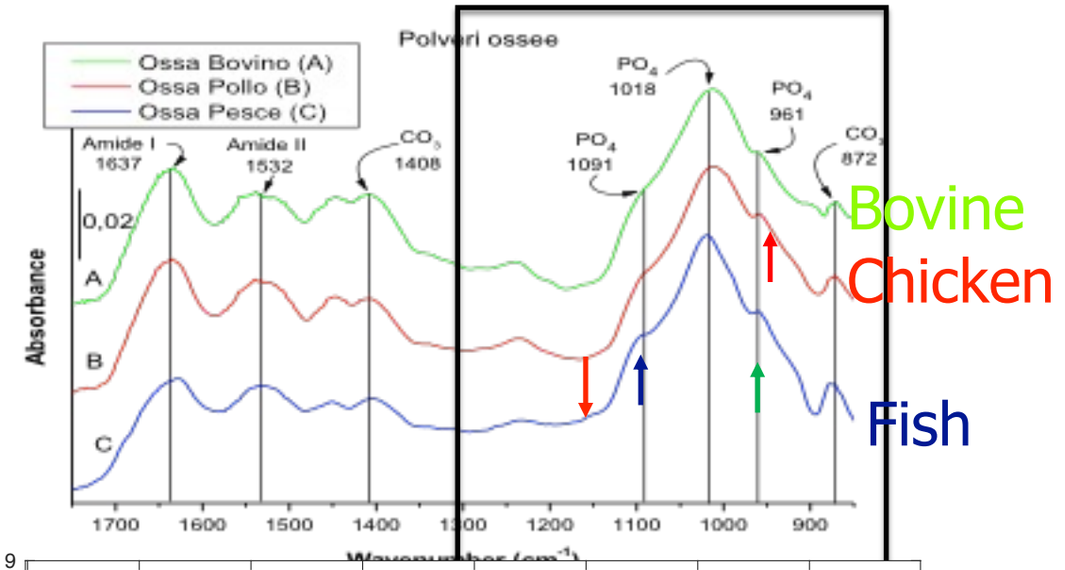
Less missclassified

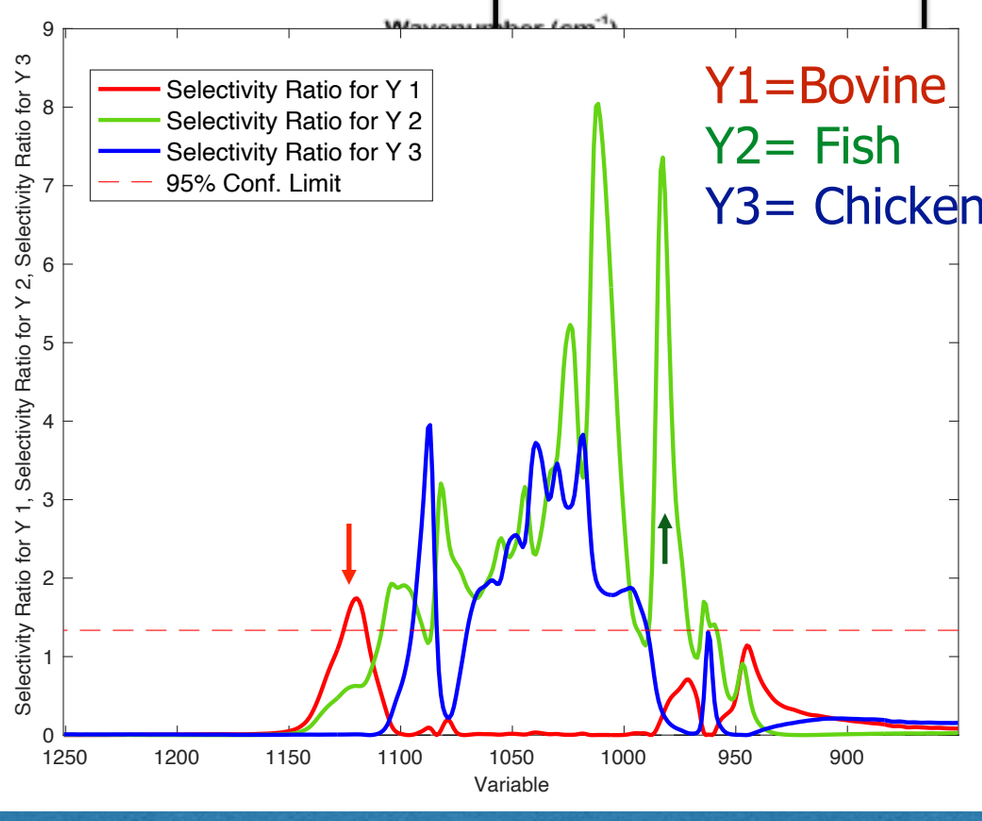
Classification based on LV: Discriminant PLS

- **Interpreting Variable importance**









Implementation in Software

- Y coding: as many dummy y-variables as classes 1/0
- Fit a PLS2 model

Common Frame

Classification rule

- **UNSCRAMBLER**:

Hard Threshold on Y-predicted values

y-predicted >= 0.5 (confidence limits for y-prediction are shown in plot)

It is also suggested to the user to use LDA on PLS scores

- **SIMCA-Umetrics**:

Hard Threshold on Y-predicted values

y-predicted <0.35 reject;
y-predicted 0.35-0.65 border-line;
y-predicted 0.65-1.35 accept;
y-predicted >1.35 do not belong, possible extreme/outlier

Options for class assignation

Given the rule above objects can fit one class, more than one class or none, then:

Unique assignation: sample assigned to the nearest class in term of lower probability

Multiple assignation: sample assigned to all the classes for which the threshold criterion is passed

Implementation in Software

- **PLS-toolbox**:

Classification rule

Threshold on Y-predicted values (can be in fit or in CV) considering each predicted y's independently.

Bayesan threshold [Computed using the distribution of calibration-samples Y-predictions (can also be in CV)]:

1. For each class separately fit a Gaussian to the y-predicted data of the class and a Gaussian to the y-predicted data of the rest of samples (all other categories).
2. The Threshold corresponds to the minimum overlap of the two gaussians.
3. Store as well the corresponding probabilities of belonging to the clasess (calculated form the fitted Gaussians) for a given y-predicted value.

Options for class assignation

F**alse negative (FN)**: objects of the class whose y-predicted is > of the threshold of the class

F**alse positive (FP):** objects NOT in the class but whose y-predicted is < of the threshold of the class.

Then assign according to one of the following:

mostprobable: sample assigned to the class for which the predicted probability value is higher

strict: sample is assigned to the class for which the y-predicted is higher than the threshold (Bayesan threshold or fixed to 0.5). If the threshold rule is passed for more than one class the sample is not assigned