Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Instructor: Dr. Brandon Corbett

Lecture 9 for ChE 765 | Sep 767, McMaster University

Agenda

1. Assignment submission
2. Handling Qualitative Variables
    1. X-Space: Binary variables
    2. Y-Space: Classification!

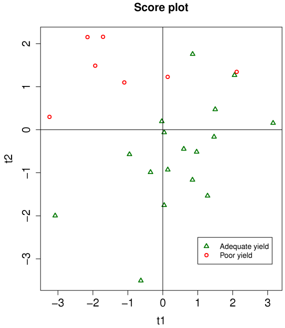
Qualitative in X-Space

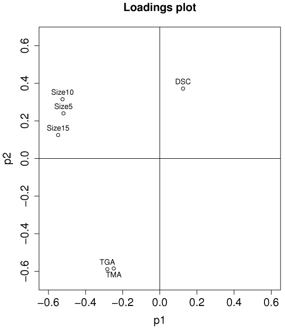
Handling qualitative variables

Binary variables

- (yes/no) **or** (on/off) **or** (present/absent)
- Also called "dichotomous variables"
- Included and used like any other variable
- Centering and scaling affected by the relative number of rows from each category
- *illustration on the board*

Or just use variable to colour-code scores by:





Handling qualitative variables

Unordered indicators must be expanded into extra columns

- aka "dummy variables", or "categorical variables"
- can be done with **X**-space and/or **Y**-space variables

e.g. reactor T, reactor D, reactor H

$$\begin{array}{c|c|c|l} 1 & 0 & 0 & \longleftarrow \text{T} \\ 0 & 1 & 0 & \longleftarrow \text{D} \\ 0 & 0 & 1 & \longleftarrow \text{H} \end{array}$$

Should then block scale the group of columns, especially if number of levels is high

We will use this in the class on "Classification"

Ordered categorical variable: ordinal variable

e.g. Primary school, High school, College, Bachelor's, Masters

- Can convert them to a single column of integers. e.g.
  - 1 = Primary school
  - 3 = College
  - 5 = Masters
- You may choose to leave them as a single column then
  - e.g. months of the year: Jan=01, Feb=02, *etc*
- Loadings interpretation: same as a continuous variable
- As a predictor in the **Y**-space: round prediction to closest integer

Caution: In many cases the gap from say 1 to 2 is not the same as the gap from say 3 to 4.

Rather expand into columns then.

Qualitative in Y-Space

Outline

1. Some definitions and examples as background
2. Classical classifiers
3. 3 latent variable classifiers, with a case study
4. Judging a classification model
5. Case studies

Examples of classification

A periodic table classifies the elements

- into groups (columns)
  - entries in a column have similar configuration of outer electron shells (halogens: F, $C\ell$, Br, I, ...), and similar behaviour
- into periods (rows)
  - entries within a row have same number of electron shells

Group → / ↓ Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba		72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra		104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Uuq	115 Uup	116 Uuh	117 Uus	118 Uuo
Lanthanides				57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
Actinides				89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Examples of classification

Counterfeit detection of drugs

- Require a quick method to answer: "counterfeit" or "real" ?
- Uses the Raman spectral signature measured on the compound
- Step 1: first identify the compound
- Step 2: have a library of "real" and "counterfeit" profiles to compare to
  - Method must allow for the possibility of a new, unidentified "counterfeit"

More details in the readable paper "Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics"

Examples of classification: Kaggle Competitions

Competitions current running:

- "Give Me Some Credit": predict if client will experience financial distress in the next 2 years
- "Don't Get Kicked!": predict if a car purchased at auction is a *kick* (bad buy)
- "Stay Alert!": detect whether driver is alert or not using vehicle, environment and driver data acquired while driving.

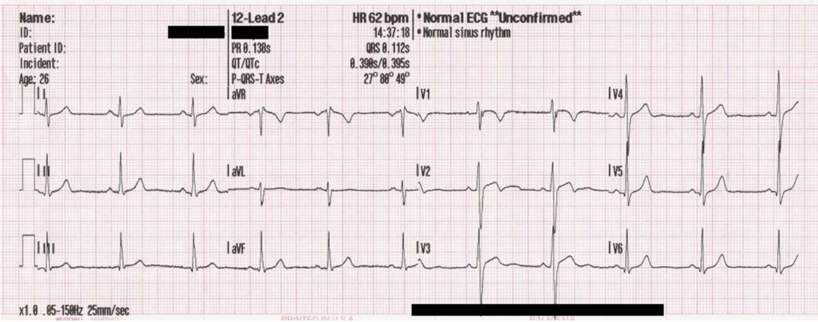
I highly recommend reading forums on closed competitions to see the techniques used.

Medical classification

There are serious consequences for incorrect classification in the medical and legal areas.

Medical classification:

- Many diagnostic tests use more than one input: blood, urine, X-rays, and other measured factors to make a decision
- Another example:



Electronic nose for TB detection

New Delhi-based International Centre for Genetic Engineering and Biotechnology:

- Aim to develop a handheld device to detect TB by 2013
- Detect tuberculosis from biomarkers in breath
- Reduces the cost, and waiting for diagnosis, faster treatment and lower transmission of the disease
- Current procedure: analyze spit sample coughed up from the lungs
- Other breath signatures: lung cancer, pneumonia, multiple sclerosis

Other interesting classification examples

- Automatic detection of spam *vs* non-spam ("ham")
  - appearance of certain words: "business", "free", *etc*
  - capitalized words
  - punctuation and symbols, such as: ! # ($;
- Drug tests for sporting events
- Canada Revenue Agency divides taxpayers into 6 classes according to certain features
  1. *Law abiders*: female, 65 years+, less educated, retired
  2. *Altruistic compliers*: 45-64 years, married, work full-time, university educated, incomes over $100,000
  3. *Rationalizers*: oldest, least-educated, lowest-income, male, retired, living in Quebec, born in Canada
  4. *Underground economy*: female, under 30, single, students, Ontarians, born outside Canada, household incomes over $100,000, more educated
  5. *Over-taxed opportunists*: female, work full-time, Ontarians,report paid employment
  6. *Outlaws*: second least-educated, second lowest-income group, under 30 years of age, male, self-employed.

Classifying painters

- 57 paintings for each: Van Gogh, Monet, Pollock, Kandinsky, Rothko, Dali, Ernst, and de Chirico
- Pixels cropped to 600×600 region
- Extracted 4027 numerical descriptors for each image:
  - edge and shape statistics
  - textures
  - histograms
  - Fourier transform, Wavelet transforms, and others
- Obviously many columns are correlated, many are useless
- Found surprising **similarities**, not normally considered by art critics
  - e.g. Van Gogh and Jackson Pollock

Read more about the study

Definitions

Class: a group of observations that are coherent in some way (belong together)

Class label: text, or numeric indicator that tell which class the observation belongs to. For example:

- "No" = 0, "Yes=1"
- "Good", "Adequate", "Bad"

Classification model: at a minimum, when given K measured values for a new observation will tell which class a new observation belongs to. Hopefully the model can do more than this.

Definitions

Unsupervised classifier: the model does not use class labels
Supervised classifier: the classification model has access to the class labels when building the model. The model is "taught" to classify between the different classes.

For example:

- *unsupervised*: babies recognize familiar faces soon after birth
- *supervised*: reading and writing must be learned – cannot be acquired in an unsupervised manner

Pattern recognition: often used as a synonym for classification

Machine learning: also a synonym for classification, but also includes continuous Y-variables.

- training (model building)
- generalizing (predictions from new observations)

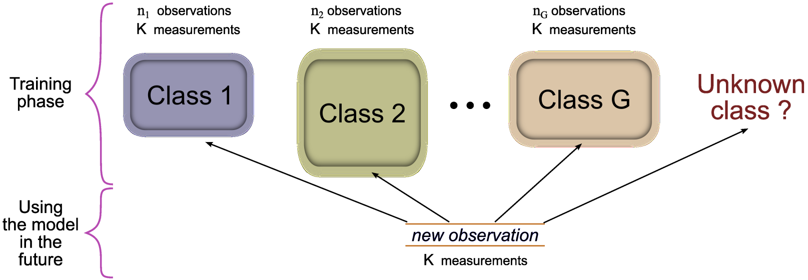
Objectives for classification

We usually have one or more of these objectives in mind when building classification models. Take the taxpayer example:

- How many groups do we have in our data?
  - Let the data naturally cluster
- What are the characteristics within each class (group)?
- What separates one class from another?
- For a new observation (person): which class do they fall in?
- Should there maybe a new class if we can't classify an observation?

We can ask the same questions for classifiers developed on chemical processes, e.g. raw materials.

Data requirements for classification



- We require K measurements from 2 or more classes to build the models
- We test class membership for a new observation based on its K values

Illustration inspired by Wold's paper: "Pattern recognition by means of disjoint principal components models"

Data set we will use to illustrate ideas: Italian olive oils

$K = 8$ fatty acids found in the lipid fraction of $N = 572$ Italian olive oils. The oils are from $G = 3$ regions:

1. Southern Italy (North and South Apulia, Calabria, Sicily)
2. Sardinia (Inland and Coastal)
3. Northern Italy (Umbria, East and West Liguria)

Data source: Forina *et al.* (1983), "Classification of olive oils from their fatty acid composition".

Strategy for this section

For each method we will look at how we:

1. building the model
2. learning from the models
3. using the model on a new data point to classify it as:
   - belonging to one class
   - belonging to another class
   - belonging to a new, unknown class
4. relative advantages and disadvantages

Methods we will consider

Unsupervised (no teacher)

1. PCA

- We rely on the data to separate itself
- Human decides where the class boundaries go

Supervised (teacher)

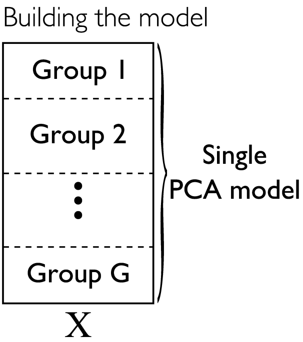
1. SIMCA: Soft independent models for class analogy
2. PLSDA: PLS discriminant analysis

- `student` is the model
- `teacher` is the objective function
- `training` is building the model to make the correct predictions

Unsupervised classification: PCA

Building and Learning

- Build a single PCA from all data
- Observe clustering in the scores
  - use different colours/shapes for each class
  - look at all score combinations
  - (same idea as masking in score images)
- Use group-to-group contributions
- Use loadings plots to understand separation
- Screen for non-class members in SPE



Unsupervised classification: PCA

Using the model on a new observation

1. Use the score space to decide which class a new observation lies in
2. Use SPE to identify non-class members

Why classification works with PCA:

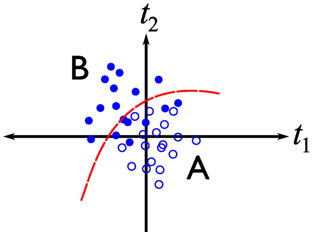
PCA just explains variance. We get a classifier if those directions of variance also happen to separate observations into clusters. Reasonable to assume observations *within a class* are similar (they will cluster together).

More likely to happen if clusters are relatively "tight" and separate from each other.

Unsupervised classification: PCA

Advantages and disadvantages

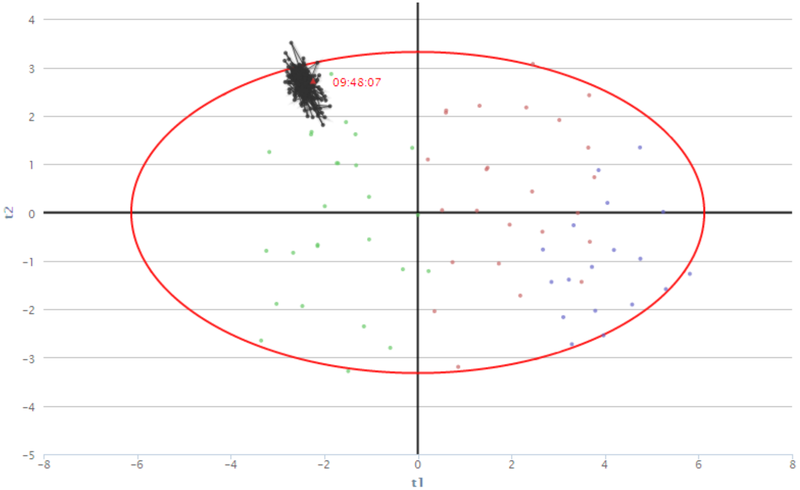
- Number of clusters gives an indication of number of classes
- Loadings help explain why classes differ from each other
- One model to explain everything: not very accurate always
- Modeller must decide on the class boundaries
- If too many classes: hard to find the boundaries with low error
- So we add extra components to help discriminate, but ...
- **Dis**: Can be tough to see if boundary is across more than 2 score directions
- **Adv**: Manual boundaries allow us to take soft-constraints regarding misclassification into account



- Misclassifying observation "**A**" as "**B**" is more costly/life-threatening
- Make the boundary conservative for class "**B**" and wider for class "**A**"

PCA classification example

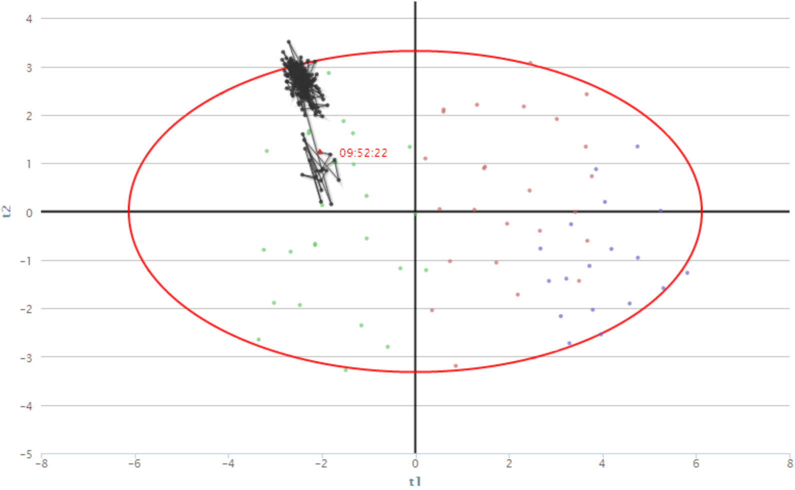
This company combines monitoring with an unsupervised classifier



Current operation is too close to the boundary for comfort ... move

PCA classification example

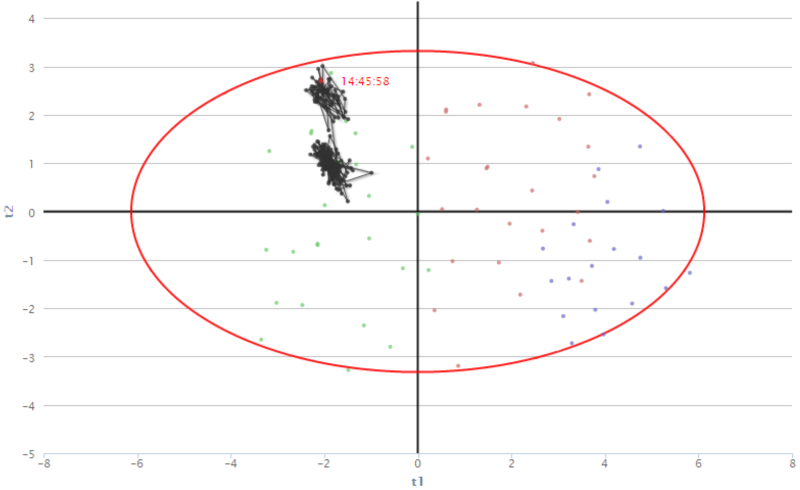
Operator can quickly confirm the process change was successful



Background dots (G, R, B) are historical good data for 3 grades

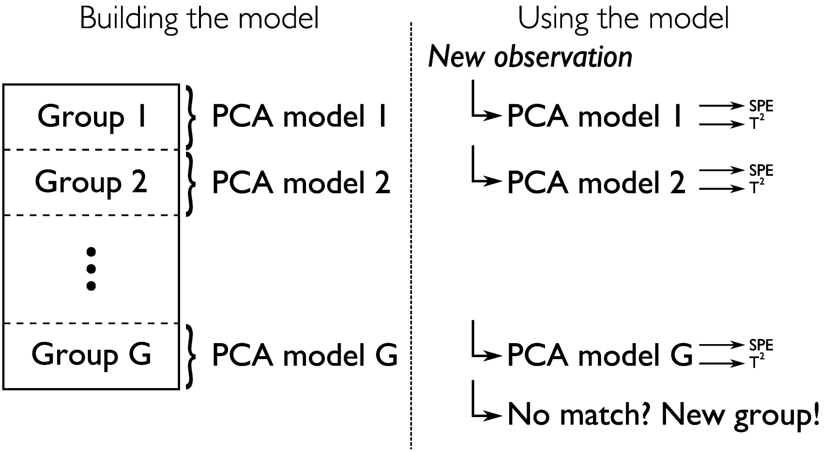
PCA classification example

Operators use it to check the settings for the desired grade



Slight overlap between red and blue grades is expected and known.

Supervised classification: SIMCA



- Called SIMCA
- Soft Independent Modelling of Class Analogy

Illustration of the SIMCA principle

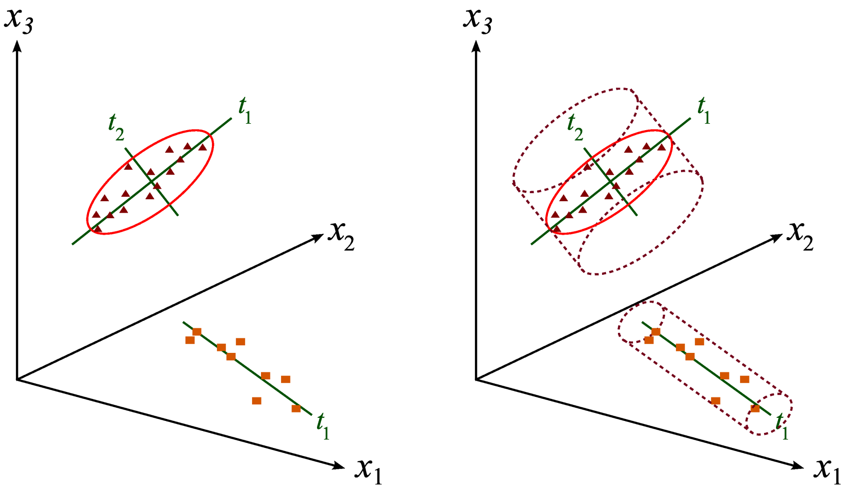


Illustration based on Wold *et al.*, 1984

SIMCA

Soft Independent Modelling of Class Analogy: **why it works**

- Objects within a class are similar to each other (*analogous*)
- We have an *independent model* for each class
- The *soft* descriptor means an observation may be classified *in the future* as belonging to more than one class

It is a supervised classifier, because we use the class information to build "the model".

Actually "the model" is a sequence of PCA models

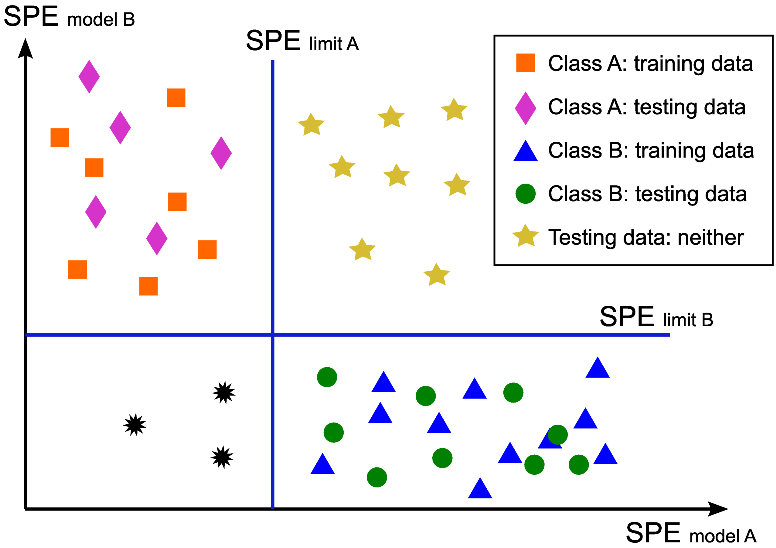
How to use SIMCA on a new observation

For each PCA class model (i.e. the g^{th} model):

- Preprocess: $\mathbf{x}_{\text{new,raw}} \xrightarrow{g} \mathbf{x}_{\text{new}}$
- Project to get scores: $\mathbf{t}'_{\text{new}} = \mathbf{x}'_{\text{new}}\mathbf{P}_g$
- Calculate the T^2 value. Below limit?
- Calculated predicted $\widehat{\mathbf{x}}'_{\text{new}} = \mathbf{t}'_{\text{new}}\mathbf{P}'_g$
- Calculate SPE from $\mathbf{e}'_{\text{new}} = \mathbf{x}'_{\text{new}} - \widehat{\mathbf{x}}'_{\text{new}}$. Below limit?

Usually we focus only on the SPE's

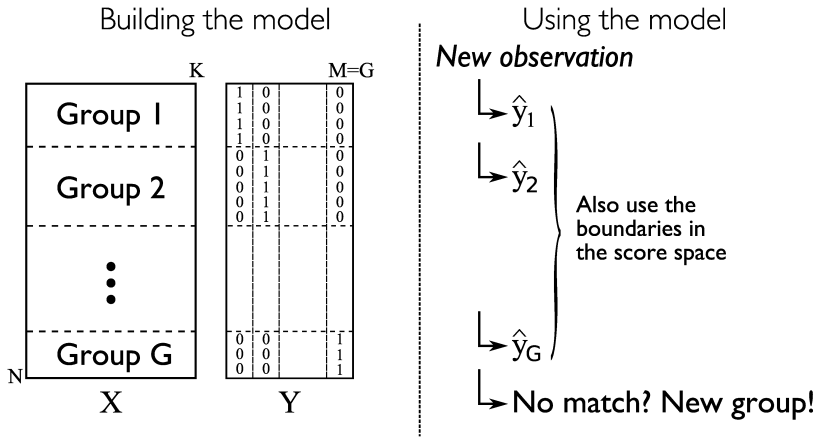
Cooman's plot



SIMCA: advantages and disadvantages

- Learning: understand relationships between variables of a class: loadings, VIP
- Each class can have different number of components
- We can add a new class later on without rebuilding previous models
- Detect outliers within each class by using SPE and T^2
- **Dis**: Hard to interpret why the classes separate
- **Dis**: What if two or more classes claim new observation i?
  - Implement a voting scheme
  - Could weight the votes in proportion to $\dfrac{\text{SPE}_i}{\text{SPE}_g^{\text{lim}}}$
  - $\text{SPE}_g^{\text{lim}} = $ SPE limit from model for class g

Supervised classification: PLS-DA



- Just an ordinary PLS model with a special **Y**-space
- Main characteristic of the **Y**-space?

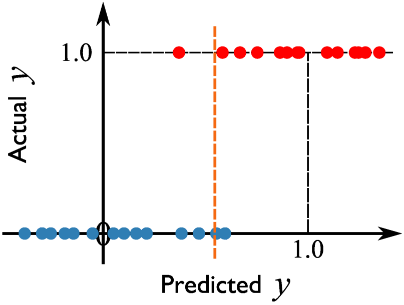
Supervised classification: PLS-DA

Why it works:

- PCA is one model – its objective is not to separate groups
- What if we re-orient the latent variables to separate groups!
- Create the orthogonal **Y**-space to "encourage" model to discriminate
- Recall the 3 objectives of PLS:
  1. Best explanation of the **X**-space
  2. Best explanation of the **Y**-space
  3. Maximize relationship between **X**- and **Y**-space

Using the PLSDA model on a new observation

- Preprocess: $\mathbf{x}_{\text{new,raw}} \longrightarrow \mathbf{x}_{\text{new}}$
- Project to get scores: $\mathbf{t}'_{\text{new}} = \mathbf{x}'_{\text{new}}\mathbf{W}*$
- Calculate the T^2 value. Below limit?
- Calculated predicted $\widehat{\mathbf{x}}'_{\text{new}} = \mathbf{t}'_{\text{new}}\mathbf{P}'$
- Calculate SPE from $\mathbf{e}'_{\text{new}} = \mathbf{x}'_{\text{new}} - \widehat{\mathbf{x}}'_{\text{new}}$. Is SPE below limit?
- Calculated predicted $\widehat{\mathbf{y}}'_{\text{new}} = \mathbf{t}'_{\text{new}}\mathbf{C}'$
- Note that $\widehat{\mathbf{y}}'_{\text{new}}$ is a vector: a prediction for each class



Example: let $G = 4$, and predict observation belonging to class 3:

- Ideal prediction:
  $\widehat{\mathbf{y}} = [0,\ 0,\ \mathbf{1},\ 0]$
- In practice:
  $\widehat{\mathbf{y}} = [-0.2,\ 0.4,\ \mathbf{0.92},\ 0.1]$

Where do the boundaries go?

The class boundaries may be drawn:

- in the score space, *and*
- in the model predictions for $\widehat{\mathbf{y}}$

You will likely require both boundary types, especially in multi-class classifiers.

Optimal boundary locations along $\widehat{\mathbf{y}}$ axis can be found via a cross-validation, or "by eye"

Learning from a PLSDA model

One of the most powerful advantages we get from a PLSDA model

- interpreting the loadings
- interpreting coefficients for each class in y
- summarizing many loadings with a VIP plot

Take a look at olive oil example

These 3 plots show us which **X**-matrix features are

- important to separate classes
- can move you from one class to the other
  - e.g. your product *vs* you competitor's product

Advantages and disadvantages of PLSDA

- Good model interpretations
- Anecdotally: PLSDA struggles with 5 or more classes
  - Use tree-based hierarchical models
  - Use multiple PLSDA models (next)
- A single observation can be classified into more than one category (soft classifier)

Multiple classes

If we have 3 or more classes we have some flexibility.

Let G = number of classes. For $G = 3$: A, B, C

1. One PLSDA model on all G classes (rarely works if $G > 4$)
   - Use predictions and scores
2. Build G PLSDA models: one-vs-rest
   - A *vs* B+C
   - B *vs* A+C
   - C *vs* A+B
3. Build $\frac{G(G-1)}{2}$ binary PLSDA models
   - A *vs* B
   - B *vs* C
   - C *vs* A

All 3 options allow an observation to belong to more than one class. Use a "voting scheme".

Interesting ways to use classification models

- Build on model on your product and the competitor's product. What information do you get from:
  - Loadings from an unsupervised PCA model
  - Coefficients from a PLS-DA
  - SIMCA models on the two classes of observations
- Customer complaints:
  - let 0 = number of complaints below a particular threshold
  - let 1 = number of complaints above a particular threshold
  - let $\mathbf{X}$ contain variables such as raw material properties, measurements from your process, *etc* and interpret VIP and coefficient plots from PLSDA
  - Could also use y=number of complaints in an ordinary PLS. How would this be different from PLSDA?
- Customer tracking
  - Websites track plenty of information about your visit
  - At the end: 0=did not buy and 1=bought product
  - What can we learn from the weights in PLSDA?

Proper validation of a classifier

A classification system often consists of more than one LV model.
For each model we must

- decide on number of components for each model
- set decision boundaries in the scores
- set decision boundaries on the $\hat{y}$
- suitable limits (e.g. 95% or 99% limits) to discriminate

With all these free parameters, it is easy to overfit.

Consider using 3 data sets:

- model building data with cross-validation to choose A
- a testing set to find suitable boundaries (scores, $\hat{y}$, SPE limits)
- a validation set to test performance

Proper validation of a classifier

As Bro points out in the paper on "Some common misunderstandings in chemometrics":

- build model on $K = 50$ totally random variables
- randomly assign class labels (e.g. A and B) to different rows
- Build a PLSDA model
- You will see near perfect separation
- But testing with a validation set will show the poor model performance

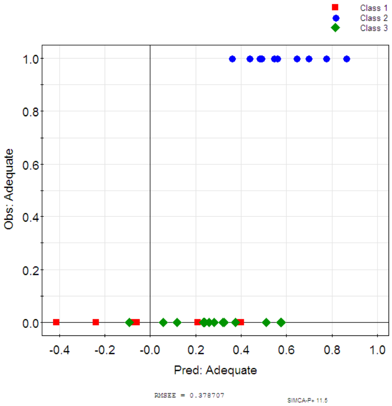
Judging performance of classification model

The usual measures of judging prediction performance are **not suitable** for classifiers:

- R^2 and Q^2
- RMSEE and RMSEP

They give approximate information, but are generally useless for comparing your various iterations as you build a classifier. Yet most software packages only provide this output.

For example, does RMSEE mean anything in the figure here?



Better performance metrics

- Receiver operating characteristic (not a single number)
- AUC (has shortcomings)
- Matthews correlation coefficient is supposedly one of the best[1]

[1] Not used it myself ... students with classification projects should use all 3