

MULTIVARIATE PATTERN RECOGNITION FOR CHEMOMETRICS

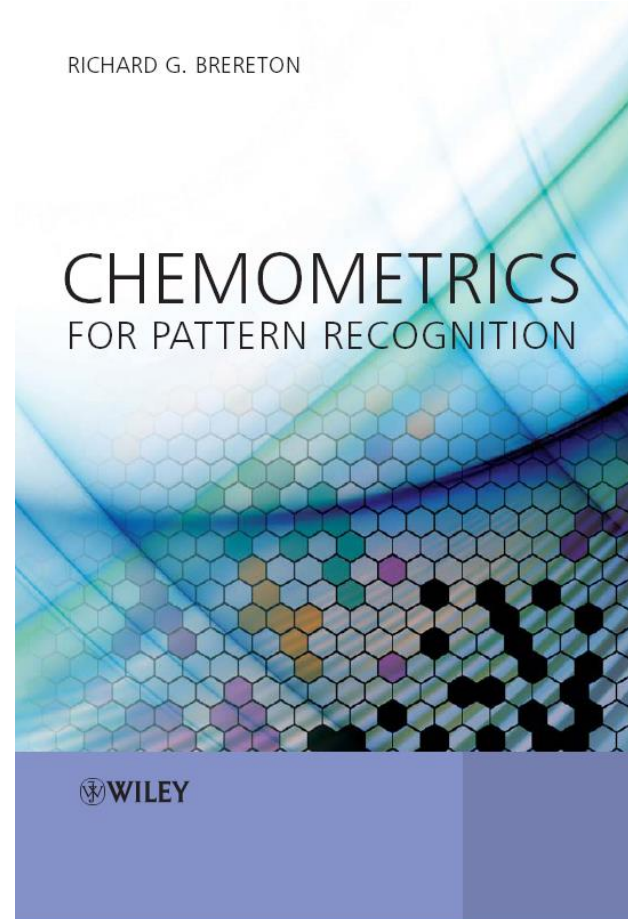
Richard Brereton

r.g.brereton@bris.ac.uk

Pattern Recognition

Book

**Chemometrics
for Pattern
Recognition,
Wiley, 2009**



Pattern Recognition

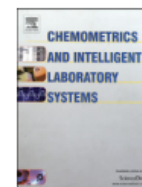
Chemometrics and Intelligent Laboratory Systems 149 (2015) 90–96



Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



Pattern recognition in chemometrics



Richard G. Brereton

School of Chemistry, Cantocks Close, Bristol BS8 1TS, United Kingdom

ARTICLE INFO

Article history:

Received 26 March 2015

Received in revised form 16 June 2015

Accepted 18 June 2015

Available online 26 June 2015

Keywords:

Pattern recognition

Partial least squares discriminant analysis

SIMCA

Linear discriminant analysis

Historic review

Support vector machines

ABSTRACT

The origins of chemometrics within chemical pattern recognition of the 1960s and 1970s are described. Trends subsequent to that era have reduced the input of pattern recognition within mainstream chemometrics, with a few approaches such as PLS-DA and SIMCA becoming dominant. Meanwhile vibrant and ever expanding literature has developed within machine learning and applied statistics which has hardly touched the chemometric community. Within the wider scientific community, chemometric originated pattern recognition techniques such as PLS-DA have been widely adopted largely due to the existence of widespread packages, but are widely misunderstood and sometimes misapplied.

© 2015 Elsevier B.V. All rights reserved.

Pattern Recognition

Special Issue - Tutorial

Journal of
CHEMOMETRICS

Received: 27 November 2013,

Revised: 28 January 2014,

Accepted: 04 February 2014,

Published online in Wiley Online Library: 18 March 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2609

Partial least squares discriminant analysis: taking the magic away

Richard G. Brereton^{a*} and Gavin R. Lloyd^b

Partial least squares discriminant analysis (PLS-DA) has been available for nearly 20 years yet is poorly understood by most users. By simple examples, it is shown graphically and algebraically that for two equal class sizes, PLS-DA using one partial least squares (PLS) component provides equivalent classification results to Euclidean distance to centroids, and by using all nonzero components to linear discriminant analysis. Extensions where there are unequal class sizes and more than two classes are discussed including common pitfalls and dilemmas. Finally, the problems of overfitting and PLS scores plots are discussed. It is concluded that for classification purposes, PLS-DA has no significant advantages over traditional procedures and is an algorithm full of dangers. It should not be viewed as a single integrated method but as step in a full classification procedure. However, despite these limitations, PLS-DA can provide good insight into the causes of discrimination via weights and loadings, which gives it a unique role in exploratory data analysis, for example in metabolomics via visualisation of significant variables such as metabolites or spectroscopic peaks. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: Partial Least Squares; Discrimination; Classification; Two Class Classifiers

Pattern Recognition

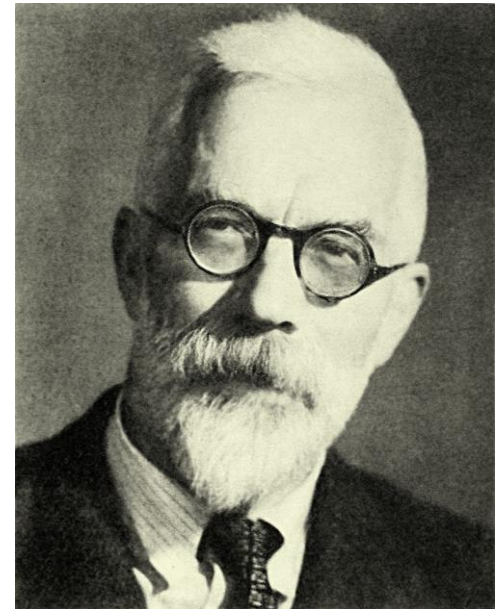
- Many definitions
 - Most modern definitions involve classification
 - Not just classification algorithms
 - Is there enough evidence to be able to group samples?
 - Are there outliers?
 - Are there unsuspected subgroups?
 - What are the most diagnostic variables / features / markers?
 - Is the method robust to future samples with different correlation structures?
- Etc.

Pattern Recognition

- **Supervised Pattern Recognition**
 - Known or hypothesised classes in advance
 - Majority of applications
- **Unsupervised Pattern Recognition**
 - Class structure not known or hypothesised

Historic Origins

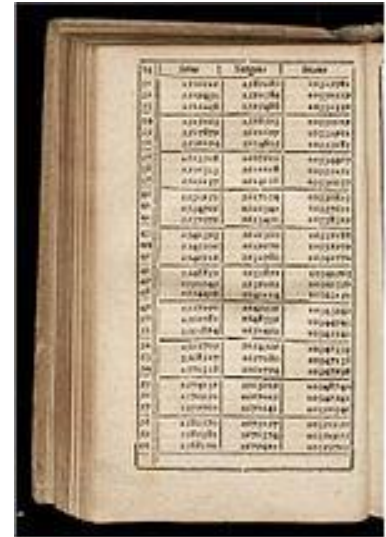
- 1920s – 1930s
 - UK agricultural industry
 - Old landowners had to improve methods after social change
 - Statisticians hired to make more efficient
 - R A Fisher and colleagues develop multivariate methods.
 - Early papers eg “Fisher iris data”



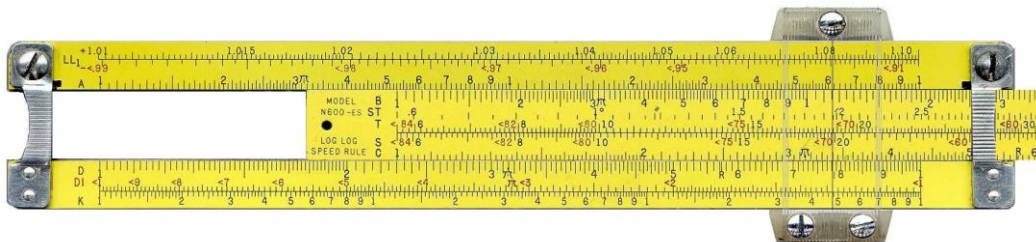
Historic Origins

- Postwar

- Gradual development and acceptance of multivariate pattern recognition by statisticians
- Limited because of computing power
- A 1921 paper by R.A.Fisher calculated to take
 - 8 months of 12 hour daysjust to calculate the numbers in the tables at 1 minute per number



The image shows a large, multi-page table with many columns and rows of numbers. The table is printed on aged, yellowish paper. The columns are labeled with letters and numbers, and the rows contain numerical data. The table is a complex statistical table, likely a chi-square distribution table or a similar statistical table from the early 20th century.



Historic Origins

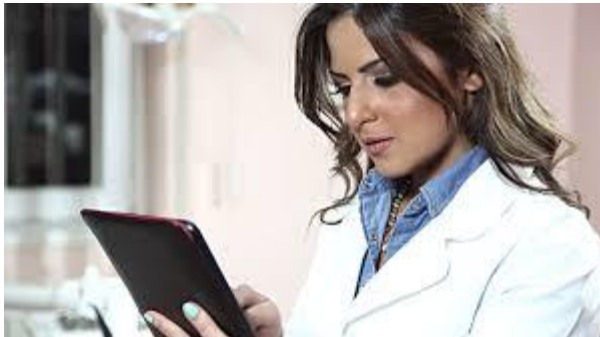
- 1960s – 1980s
 - Chemical Pattern Recognition
 - Facile computer power and good programming languages
 - No longer needed to be a statistician
 - Origins of chemometrics
 - Renamed in mid 1970s by Svante Wold
 - The name chemometrics took off in late 1970s / early 1980s
 - Early pioneers often regarded themselves as doing pattern recognition.

Historic Origins

- 1980s – 1990s
 - Growth of chemometrics
 - Pattern recognition small element others such as
 - Signal Analysis
 - Multivariate Curve Resolution / Factor Analysis
 - Experimental Design
 - Multivariate Calibration
 - Primarily instrumental analytical chemistry
 - Often small datasets, eg. 20 samples and 10 HPLC peaks

Historic Origins

- **Modern Day**
 - Large datasets possible
 - Applications to new areas outside mainstream analytical chemistry
 - Cheap and fast computer power



Univariate Classifiers

- Traditional approach to classification
- Select one or more marker compounds
- Measure
 - HPLC peak height
 - GCMS peak height
 - NMR
- Determine
 - Presence / absence
 - Concentration
 - Peak ratios

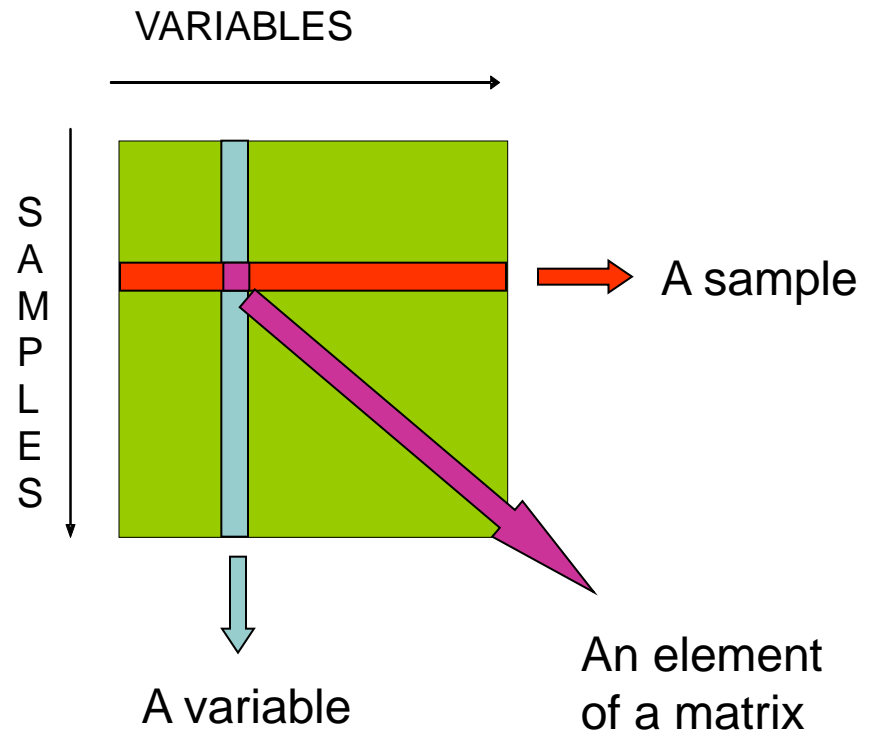
Univariate Classifiers

- Traditional approach
- Problems
 - Quantitative analysis is often difficult and very dependent on instrument and reference standards.
 - GCMS, HPLC, extraction may be expensive and time consuming whereas spectroscopic methods such as NIR may be faster and cheaper
 - Food contains many compounds and as such using traditional methods only a small number of markers are studied
 - Many differences are quite subtle especially when detecting adulteration, different phenotypes, different factories etc.
 - Some minor differences are important

Multivariate approach

- Multivariate data matrix

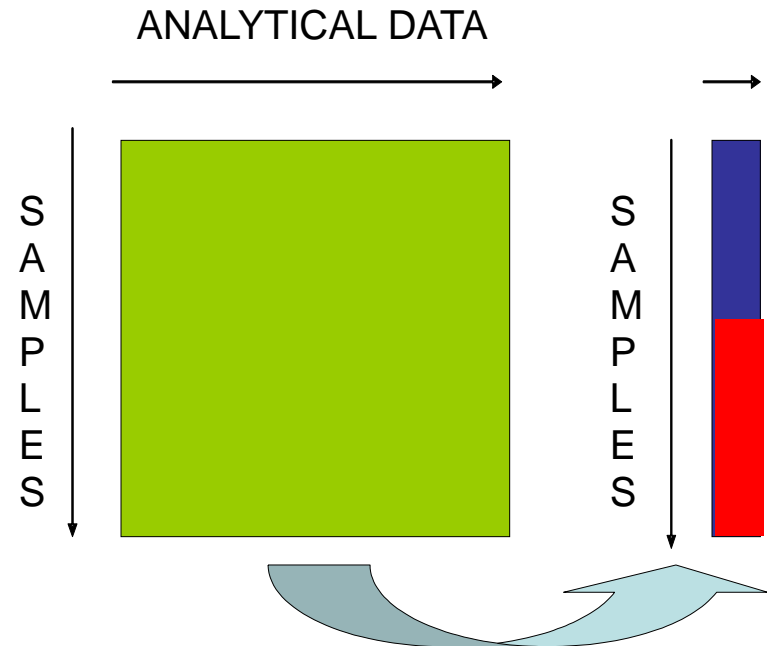
- We measure variables on samples e.g.
 - chromatographic intensities of chromatograms
 - concentrations of compounds in reaction mixtures
- The elements of a matrix consist of the size of the measured variable in a specific sample e.g.
 - the intensity of a specific peak in a specific chromatogram
 - The intensity of an absorbance by NIR



Multivariate approach

- Classification

- A way of grouping samples
- Predictive modelling
 - Predict the origins of samples
- Hypothesis tests
 - Is there a relationship between the analytical signal and their origins?



Modern Chemometrics and Analytical Chemistry

- In the modern world we can obtain many measurements per sample very easily.
- Many methods in textbooks are quite old as there is a long time lapse between writing texts and accepting new methods, often 20 years.
- Much traditional analytical chemistry involves optimisation, can we get better separations or better efficiencies.
- In chemometrics this is not always so: we often do not know the training set perfectly, there can be outliers, artefacts, misclassifications or even imperfect techniques.

Modern Chemometrics and Analytical Chemistry

- Traditional problems eg [Fisher's iris data](#), the answer is known for certainty in advance



Iris setosa



Iris versicolor

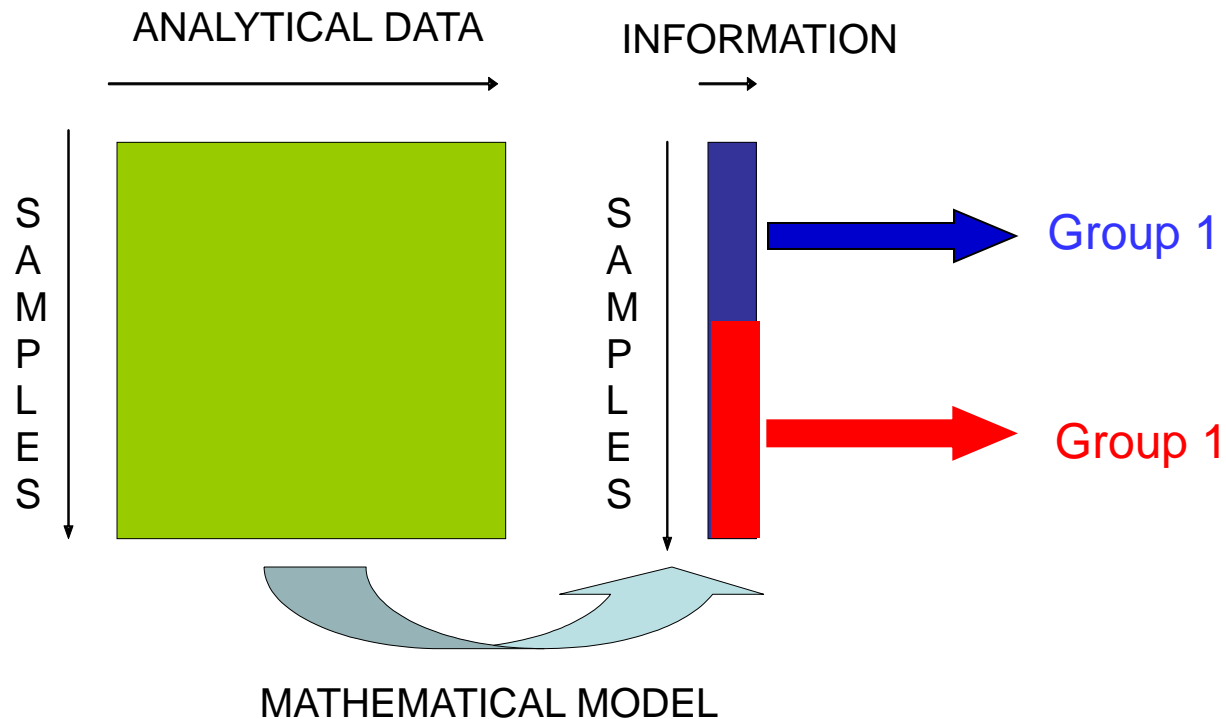


Iris virginica

- The aim is to reach this well established answer as well as we can
- We might then ask which variables (in the iris data, the physical measurements) are most useful (in modern terminology marker compounds) for example or to predict the origins of an unknown
- In many modern situations we do not know the answer in advance

Predictive models

- Form a mathematical model between the analytical data and the factor of interest. Can be more than two groups.



Predictive models

- Training set

- Traditional approach. Divide samples into training and test set.
- Develop a method that works very well on training set.

- Weakness

- The training set in itself may not be perfect
- Numerous reasons
- So 95% correctly classified may not necessarily be “better” than 85%

Predictive models

- Is it possible to predict membership of a group?
- Can we take an unknown sample and say which group it belongs to using analytical data?
- Can we classify an unknown sample to a group?
- Is the data good enough (of sufficient quality)?
- Are there subgroups in training set?
- Are there outliers in training set?

Predictive models

- Can we predict the origins of a food stuff according to its country?
- By overfitting, yes, but in practice this means nothing.
- Many examples of “perfect” separation!!!!
- With sophisticated modern methods possible but of no meaning



Predictive models

- What is the best method?
 - No real answer
 - Can do on simulations, but these will not incorporate real life issues
 - Simulations good for developing algorithms to check they work
 - In real life we often need controls, eg “null” datasets, permutations

Multivariate Classification Techniques

- Too much emphasis on named techniques
- What matters is formulating the question well
 - Choosing an appropriate training set
 - Choosing an appropriate test set
 - Deciding what problems you will look at
 - Eg are you interested in outliers
 - Are you interested in distinguishing two or more groups
 - How confident are you about the training set
 - Is the analytical technique appropriate and reliable

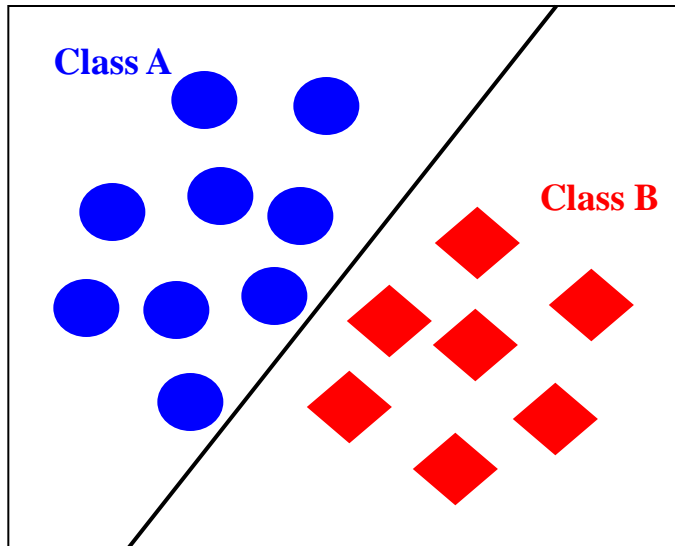
Class Boundaries

- Classification can be regarded as finding boundaries between groups of samples. The difference between techniques corresponds to the difference in establishing boundaries
- A classifier can be regarded as a method that finds a boundary between or around groups of samples, all common classifiers can be defined this way
- All classification methods can be formulated this way, including approaches based on PLS
- Sometimes techniques are presented in other ways e.g. projection onto lines, but these projections can be expressed as distance from boundaries, so the key to all techniques is to find a suitable boundary. Extensions e.g. class distance plots based on boundaries.

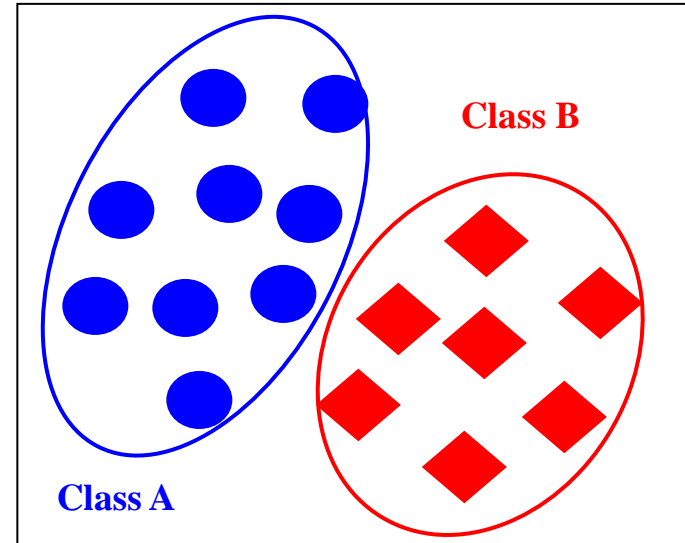
Class Boundaries

- **Two class classifiers** .
 - Model two classes simultaneously and try to form a boundary between them.
- **One class classifiers**
 - Model each class separately. Not all the classes need to be included.
 - Forms boundary around each class that is modelled often at a given confidence limit.
- **Multi class classifiers**
 - Model several classes simultaneously.

Class Boundaries



Two class classifier



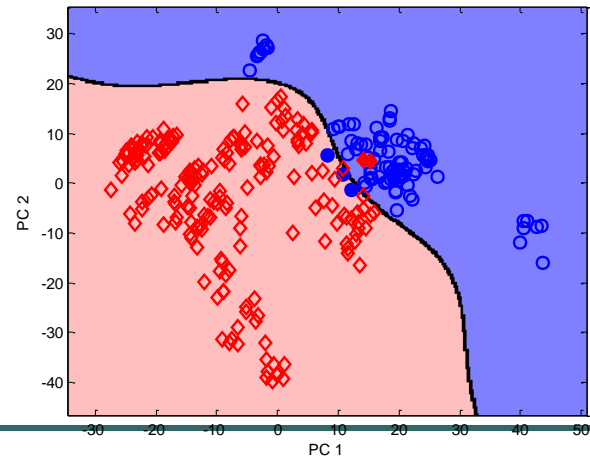
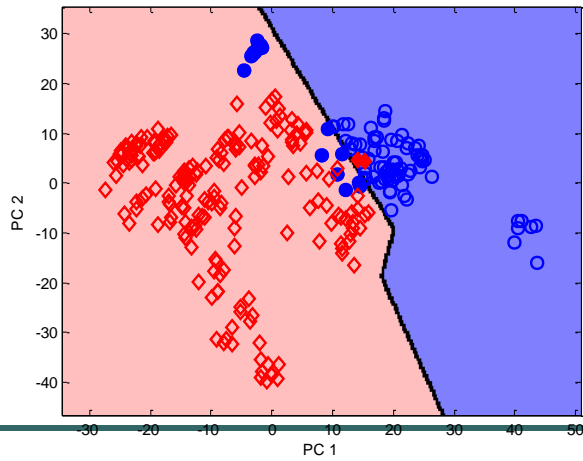
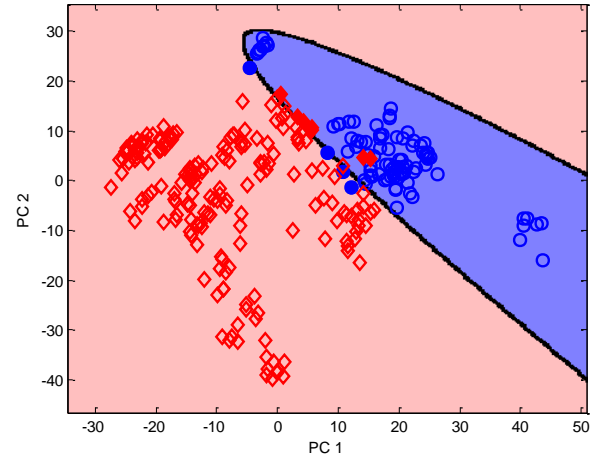
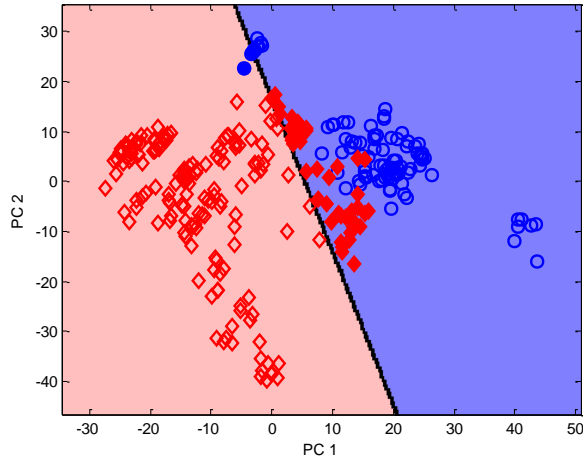
Two one class classifiers

Illustrated for bivariate classifiers but can be extended easily to multivariate classifiers

Two Class Classifiers

- Differ according to the complexity of the boundary
 - Model two classes simultaneously and try to form a boundary between them. Most classifiers can be expressed this way.
- Common Approaches
 - Euclidean Distance to Centroids
 - Linear Discriminant Analysis
 - Quadratic Discriminant Analysis
 - Partial Least Squares Discriminant Analysis
 - Support Vector Machines
 - K Nearest Neighbours

Two Class Classifiers



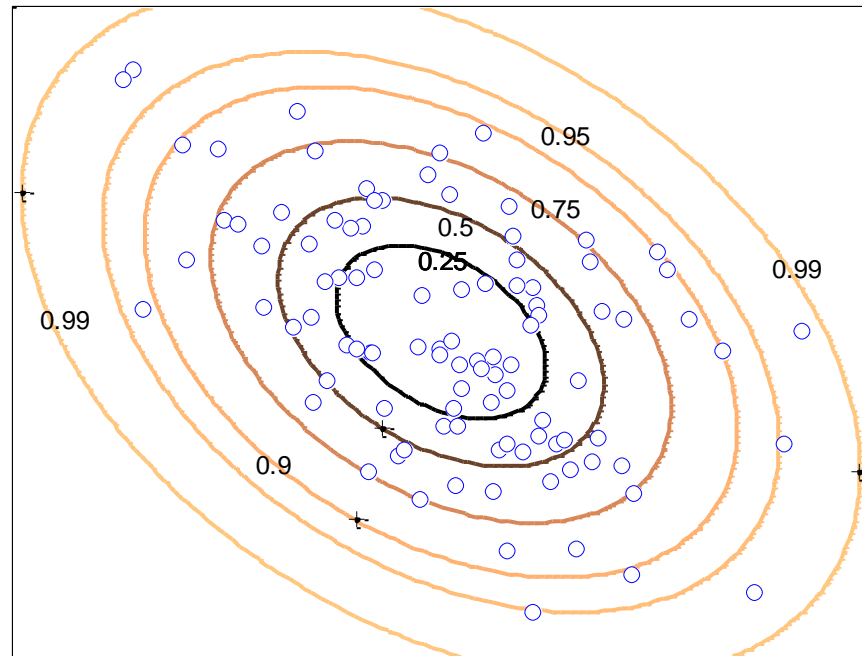
Two Class Classifiers

- No best method
 - The more complex boundaries, the better the training set model
 - The more complex boundaries, the bigger the risk of over-fitting, this means mistakes when classifying unknowns
 - Often over-optimistic models. So take care!

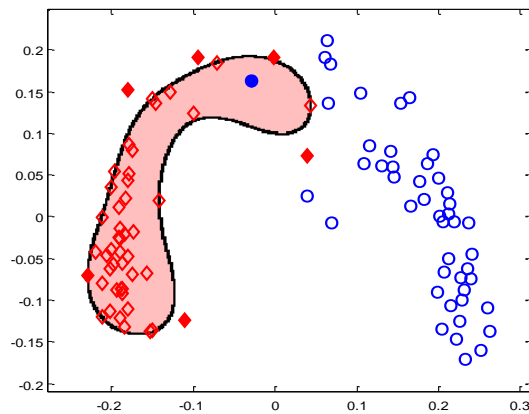
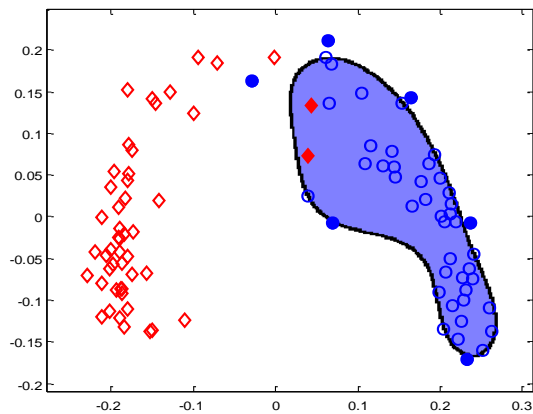
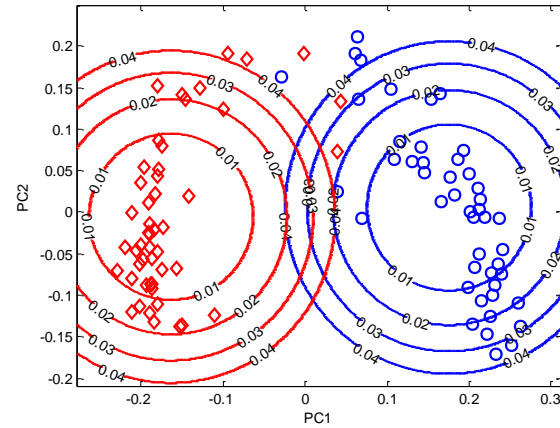
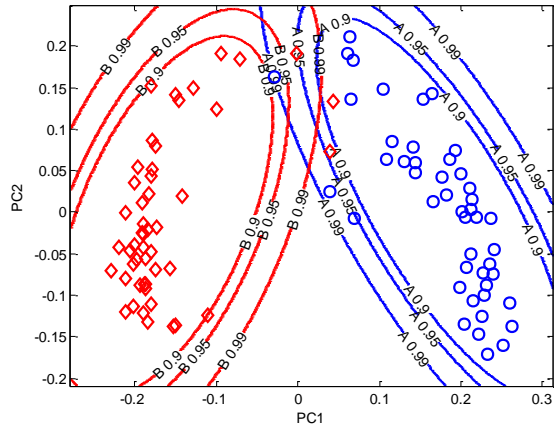
One Class Classifiers

- **Forms a boundary around a class**
 - Usually at a certain percentage probability
 - For example 99% means that for a training set group we expect 99 out of 100 samples to be within that boundary.
 - Often depends on samples being normally distributed
- **Common Approaches**
 - Quadratic Discriminant Analysis
 - Support Vector Domain Description
 - Incorporated into SIMCA

One Class Classifiers

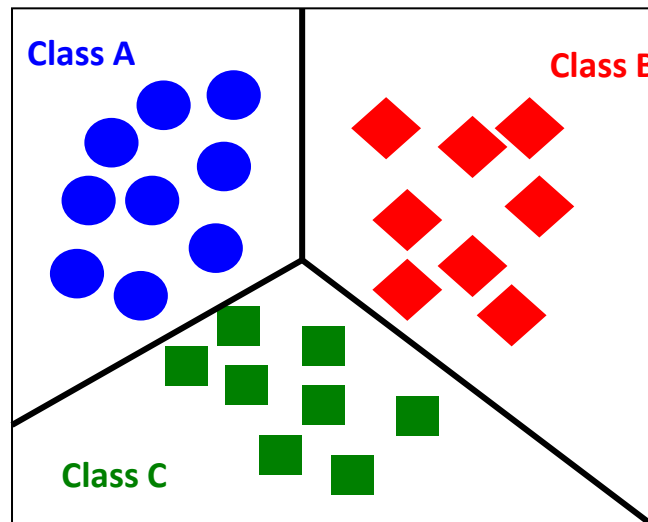


One Class Classifiers



Multiclass Classifiers

- Extension of two class classifiers
 - Simple for some approaches such as LDA or QDA
 - Difficult and often misapplied for approaches such as PLS-DA



Comparison of methods

There is a large and very misleading literature comparing methods - beware

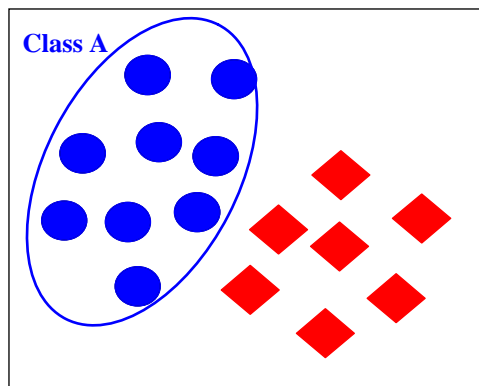
- For example there will be claims that method A is better than methods B, C and D
- The method will be claimed to be better as judged by the difference in one or more performance indicator such as %CC (percent correctly classified), usually on a test set and on one or more carefully chosen datasets.

Comparison of methods

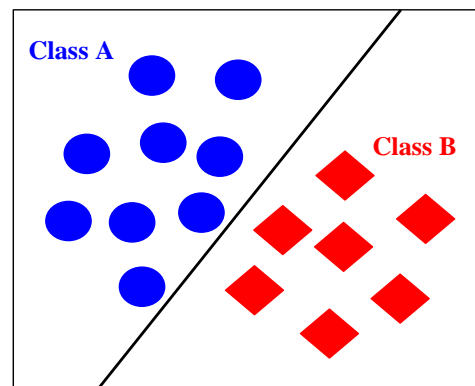
- There is strong pressure eg to get PhDs, get grants, get papers, or even conference presentations
- Often a method that isn't "better" is regarded as a waste of time, no more grants, papers or PhDs
- Hence there are ever more claims of improved methods in the literature and at conferences.
- Beware.

Comparison of methods

- It is often not possible to compare methods directly.
- Example
 - One class classifiers (eg SIMCA, Support Vector Data Description, certain types of QDA)
 - Two class classifiers (eg LDA, PLS-DA, Euclidean Distance)



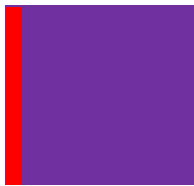
One class



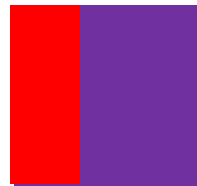
Two class

Traditional problems : comparison of methods

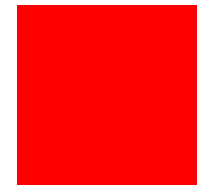
- Preprocessing can radically change the performance of a method
- Example
 - PLS-DA is the same as EDC (Euclidean Distance to Centroids) if only one PLS component is used
 - PLS-DA is the same as LDA if all components used
 - So we can't say "we have used PLS-DA" without qualifying this



1 component
PLS-DA=EDC



Several components
Intermediate



All non-zero components
PLS-DA=LDA

Traditional problems : comparison of methods

- Should we use PLS-DA as opposed to statistical methods?
 - The statistical properties eg for LDA (linear discriminant analysis) and EDC (Euclidean distance to centroids) are well known and well established.
 - A traditional limitation of LDA is that Mahalanobis distance cannot be calculated if number of variables $>$ number of samples, but this is not so, just use the sum of squares of standardised non-zero PCs
 - So why use PLS-DA? And why compare to LDA because PLS-DA could be the same as PLS-DA.

Traditional problems : comparison of methods

- Many other choices of parameters for some methods
- Eg PLS-DA
 - Data transformation
 - Type of centring
 - Acceptance criteria
 - Number of components
- Etc.
- Other methods very little choice

Often the choice of parameters has as much or more influence than the choice of classification algorithm

Traditional problems : comparison of methods

- How to view this
- View the classifier just as one step in a series, just like addition and multiplication but a little more complicated
- Focus as much on the data preparation step and decision making as on the algorithm
- We probably have access to all the algorithms we need, resist trying to invent new ones.

It is often unwise to compare different approaches directly, and if done, one needs to understand all steps.

The pragmatic approach is to use several quite incompatible methods and simply come to a consensus.

Conclusions

- Historical origins in UK agriculture of the 1920s-30s.
- Chemometrics developed in the 1960s-70s
- Rapid and easy computing power important
- Multivariate advantage
- The nature of the problem has changed since the 1970s
- Answer often not known for certain in advance
- Classifiers are often not comparable
- Too much emphasis on named methods and on comparisons
- Much historic software and literature based in 1970s problems