THE CHALLENGES OF VALIDATING MULTIVARIATE METHODS FOR PATTERN RECOGNITION

Richard Brereton

richard.brereton1@gmail.com

Pattern Recognition

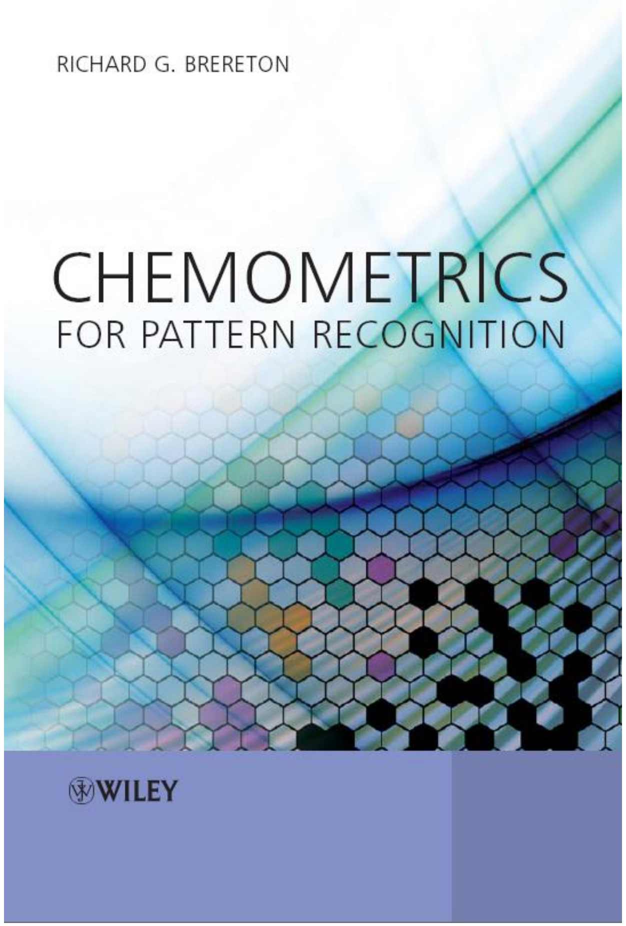
- Many definitions
- Most modern definitions involve classification
- Not just classification algorithms
  - Is there enough evidence to be able to group samples?
  - Are there outliers?
  - Are there unsuspected subgroups?
  - What are the most diagnostic variables / features / markers?
  - Is the method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

Etc.

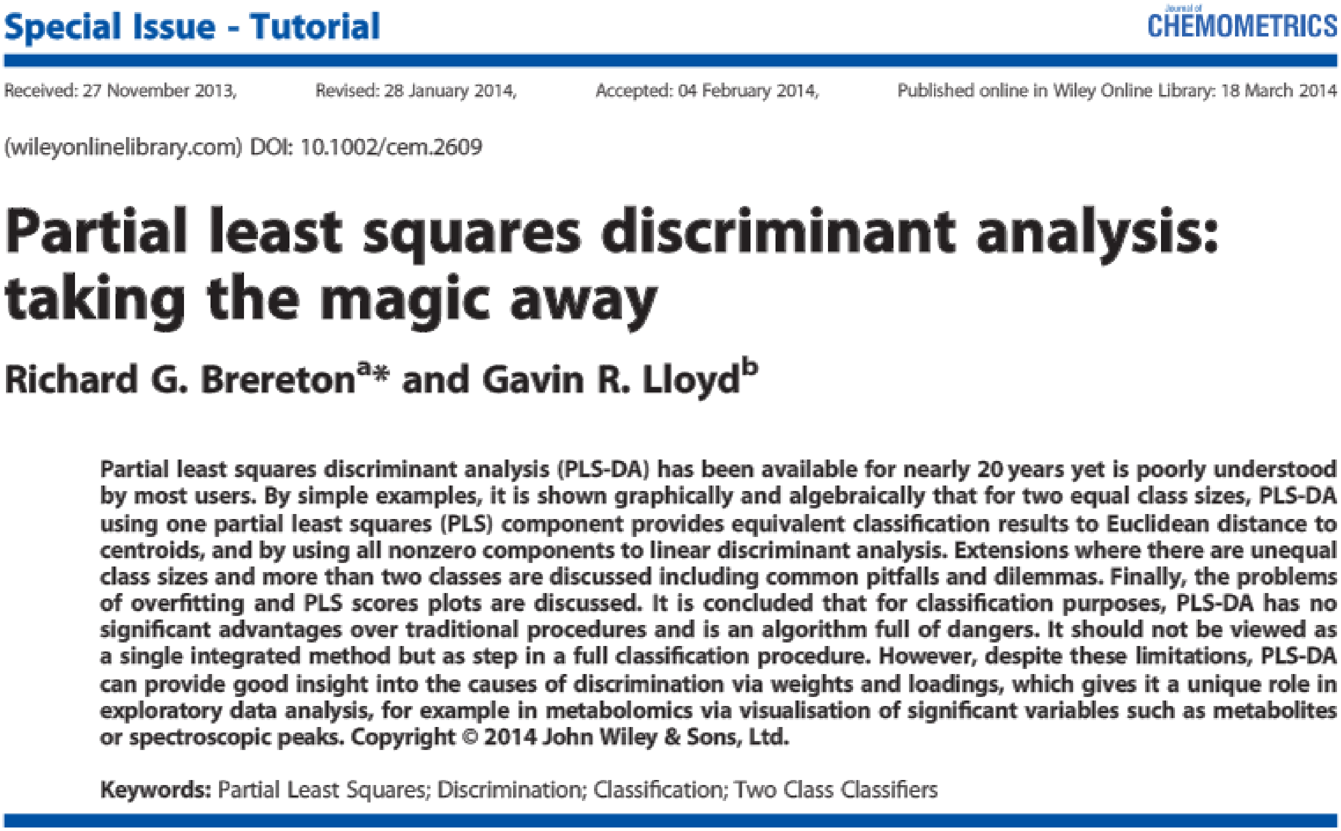
Pattern Recognition

Book

Chemometrics for Pattern Recognition, Wiley, 2009



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

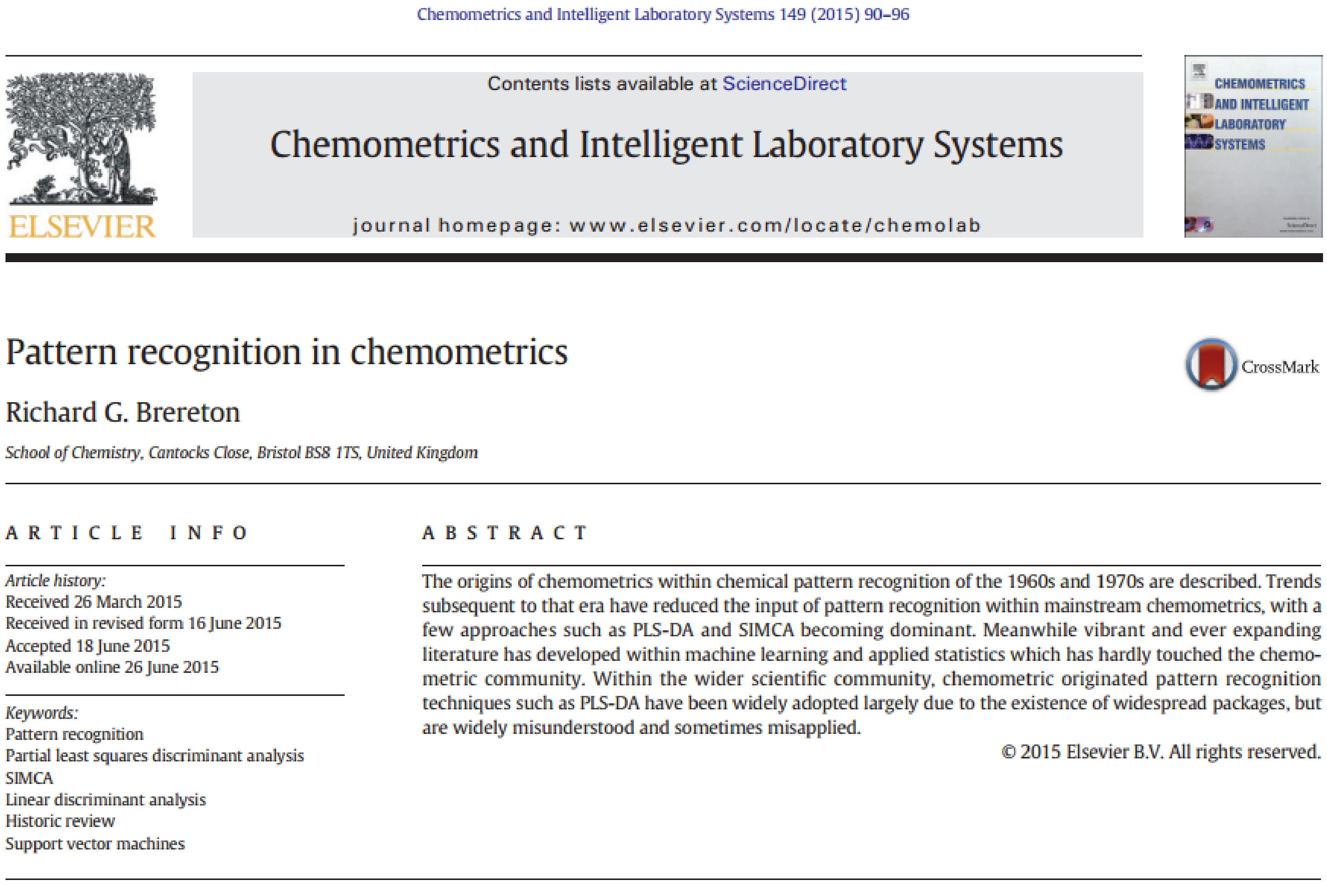
Partial least squares discriminant analysis: taking the magic away

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



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.

Validation

- How well does a method perform?
- Is there adequate information in a dataset to support a hypothesis?

Traditionally easy

- Often known answers, eg traditional taxonomy such as Fisher’s Iris data.
- Often sample to variable ratios large.

Traditional problems : uncertain hypothesis

We do not know whether the underlying hypothesis is correct or not.

- Traditionally algorithms are benchmarked against known outcomes
- For example, comparing the proportion of misclassifications using different algorithms
- Sometimes simulations are used, but do these have the correct correlation structure or uncertainties / ambiguities of the relevant datasets?

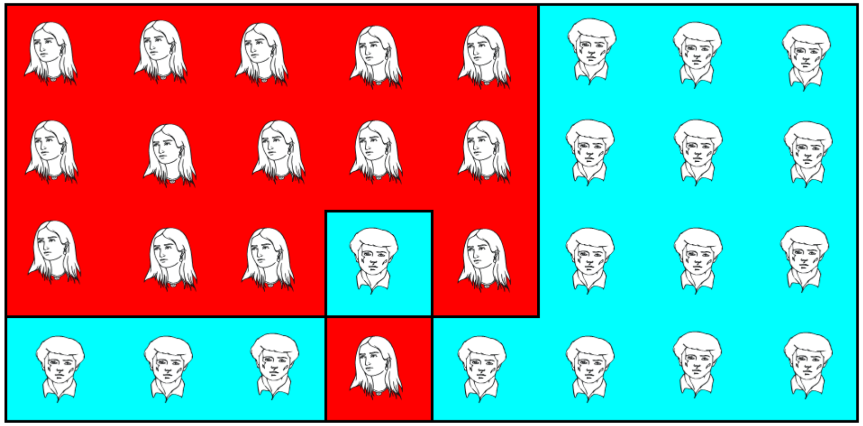
- Is the classification method used to determine whether a hypothesis is correct, or to determine the provenance of an unknown sample?

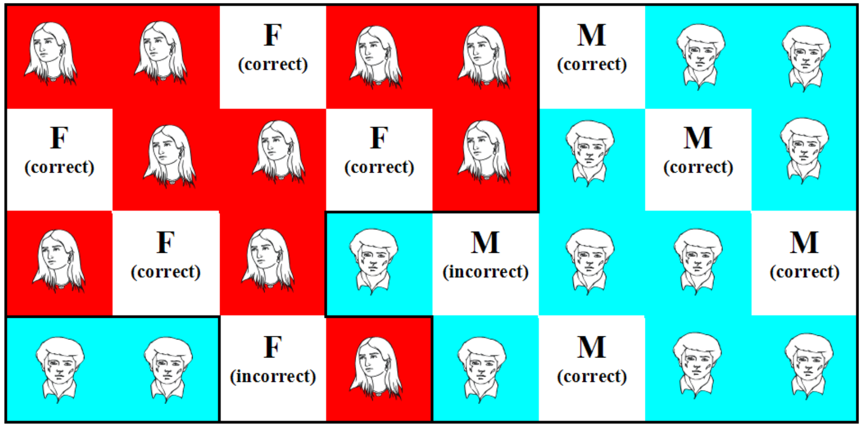
Traditional problems : uncertain hypothesis

- Use test sets for hypothesis testing.
  - An autopredictive model can, if a boundary is complex enough, perfectly predict any dataset. Classification methods can all be viewed as determining a boundary between or around groups.
  - A test set model allows this hypothesis to be tested out. A test set is a series of samples that is left out.
  - Seating plan – to determine gender, and test the hypothesis that we are in an audience in a traditional society where people of the same gender groups together
  - Rules

    1. For each unknown place look at the gender of the nearest neighbours both horizontally and vertically. For places in the middle of the seating plan this will be four, for places at the edges 3, and in the corners just 2.
    2. Look at the gender of the nearest neighbours and assign the empty seat to the majority gender.

Traditional problems : uncertain hypothesis





Left : autoprediction

Right : training and test sets

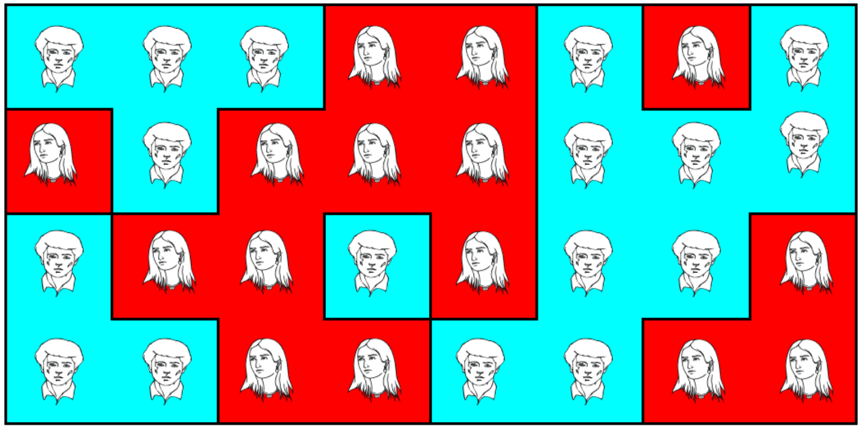
The traditional society audience

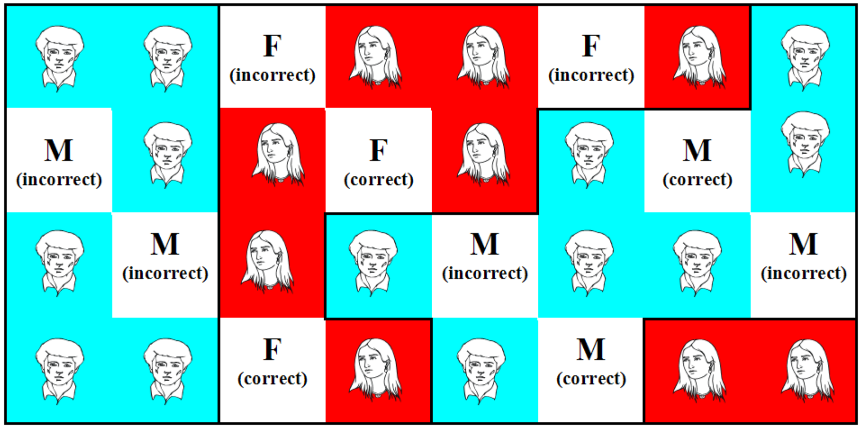
Our rules predict 8 out of 10 of the test set correctly

80% correctly classified (%CC)

High predictive ability – so hypothesis probably right

Traditional problems : uncertain hypothesis





Left : autoprediction

Right : training and test sets

The family grouping

Our rules predict 4 out of 10 of the test set correctly

40% correctly classified (%CC)

A “random” model will be close to 50% if two groups

Traditional problems : uncertain hypothesis

- Is the classification method used to determine whether a hypothesis is correct or to determine the provenance of an unknown sample?

○ In the previous example, the %correctly classified says nothing about the model or classifier

○ It does not assume any specific distribution

○ It mainly asks about the underlying hypothesis

Therefore performance abilities may equally well just be about hypothesis testing

Traditional problems : uncertain hypothesis

- No easy way out of this for comparing methods
- Can test against simulations
  - But there is no way to obtain a perfect simulation
  - Always unusual features eg mislabelled sample, inhomogeneous groups, problems finding representative training sets, outliers, new future features in a process.
- Huge literature comparing classification ability of different approaches

Traditional problems : uncertain hypothesis

- Null datasets
  - Randomly generated
  - Still need to ensure some of the original features such as scale and shape are introduced
  - A random dataset of 100 samples, 50 of group A and 50 of group B, should give approximately 50% correctly classified if the method is unbiassed
  - Follow the proposed method through and see what happens to the null dataset, compare to the real dataset.

Traditional problems : uncertain hypothesis

- Permutations
  - Randomly generated classifier
  - Sometimes called “Monte Carlo” methods
  - Keep the experimental data (often called the “X” block) constant, and change the classifier randomly (often called the “y” or “c” block).
  - Often do this several times
  - See if the performance on the unpermuted dataset is significantly better than an ensemble of permuted data, usually using an unparametric test

Compare classifier to control datasets either via permutations or null datasets : this is rarely done

Traditional problems : optimisation

Optimisation is often confused with validation

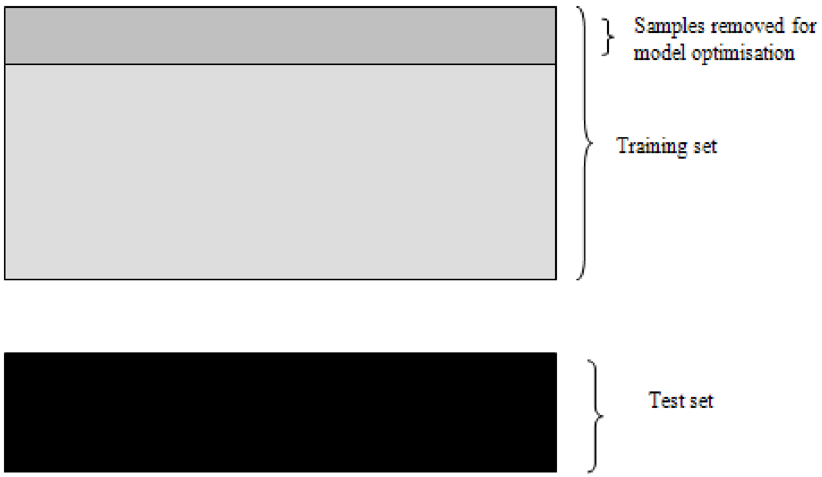
- ○ Traditionally they are confused.
- ○ Need to optimise
  - ○ eg how many components in a PC or PLS model?
  - ○ what are the most appropriate penalty errors and / or radius for a RBF in support vector methods

- Traditionally methods like cross validation can be used for both.
  - ○ Samples are left out and they predictive error is used both to determine the optimum model and the model’s performance.

Traditional problems : optimisation

Solution is to separate validation and optimisation

- Well established in areas like neural networks
- Not too well established in traditional chemometric approaches



Traditional problems : representative samples

There are often unrepresentative samples in the training or test sets.

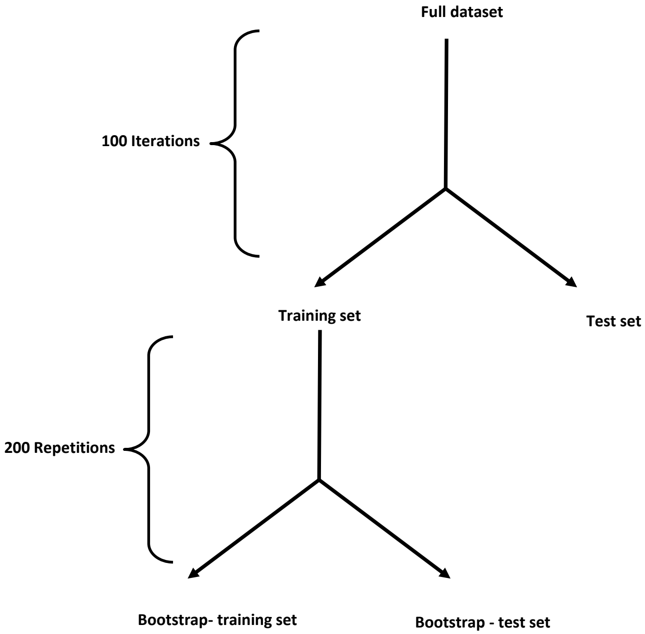
- ○ Often people compare methods with closely similar error rates, for example method A may result in 92% Correctly Classified, and method B 96%, so method B is seen to be better.
- ○ Unless there really is a big difference (eg linear methods for classes that can only be separated using quadratic classifiers) the difference in performance is usually quite small
- ○ These differences may depend on a very small number of samples.
- ○ There will always be outliers, or mislabelled samples, or subgroups.
- ○ Most data is not multinormal, and many classifiers assume this.

Traditional problems : representative samples

- To overcome this, generate all subsets of data many times over
  - Both optimisation and testing can be done by repeated regeneration of data.
  - Optimisation : Bootstrap involves repeated resampling to determine optimum number of components
  - Repeated generation of test set also.

Traditional problems : representative samples

- Typical approach



Traditional problems : representative samples

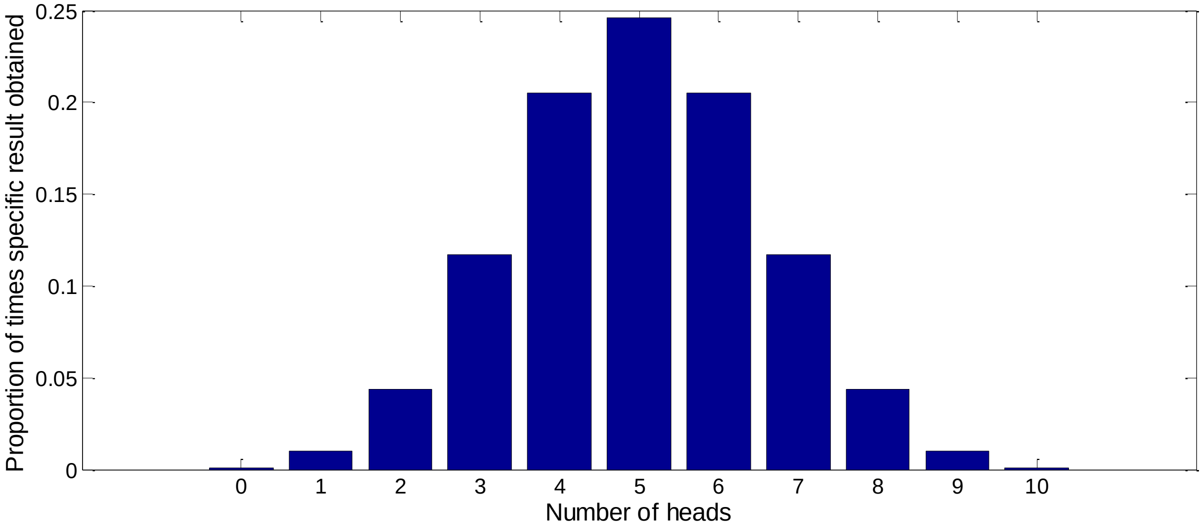
- Modern computing allows very intense methods
- Iterative approaches available both for optimisation and validation
- It is rarely possible to obtain a perfectly representative dataset (or what? Or all future trends?)

Use iterative and computationally intense methods to include / exclude samples many times and get an overview

Traditional problems : variable selection

Huge number of variables, but supervised variable selection can bias a method

- Example : toss of a coin 10 times

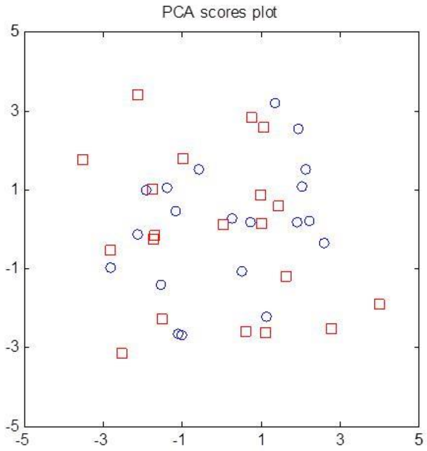


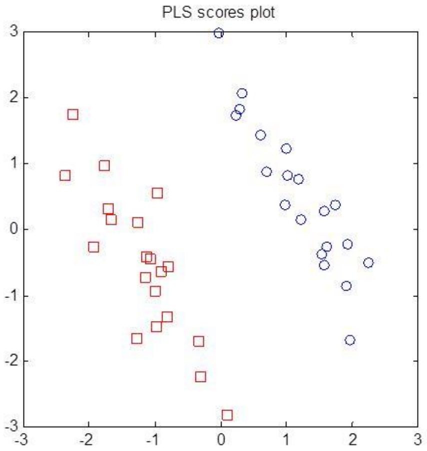
Traditional problems : variable selection

- Example : toss of a coin 10 times
  - Each time an unbiassed coin is tossed 10 times, there is a chance it will turn up 8 or 9 Heads
  - If this experiment is repeated 1000 times, there will be several cases of 8 or more Heads.
  - If we then selected those cases, approximately 45, then we could remove or reduce the influence of the remaining 955 sets of tosses
  - We would incorrectly conclude the coin is biassed
  - In practice this is what a method such as PLS-DA (Partial Least Squares Discriminant Analysis) does
  - This is especially serious when variable to sample ratios are large

Traditional problems : variable selection

A randomly generated 40 × 200 dataset, arbitrarily divided into two groups, graphs of the scores of the first two principal components and partial least squares components





Traditional problems : variable selection

- Randomness is not uniformity
  - A sequence HTHTHTHTHT is not random
  - A typical sequence might be HHTHTTTTHH
  - Some clumping
  - “Clever” algorithms take account of this clumping
  - A completely random dataset can look separable using PLS-DA, especially if there are lots of variables

Traditional problems : variable selection

- Often variable selection is important
- Noisy variables eg uniformative wavelengths or peaks
- These can dominate a dataset and so hugely degrade performance
- Usually though people use a supervised way of variable selection
- This can result in over-optimistic predictions, like the case of the unbiassed coin.

Traditional problems : variable selection

- How to avoid this
- Select variables on training set only
- The dilemma
- If generate hundreds of training sets, there will be different variables selected each time
- This will not only influence the variables in the model, but other parameters such as means and standard deviations and number of components

Traditional problems : variable selection

- Solutions
  - No universal solution
  - Can select variables on a consensual way, ie which are most frequently selected, or after an optimal model is chosen

Beware selection and weighting of variables can result in over-optimistic models, and use training sets only for this purpose

Traditional problems : comparison of methods

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

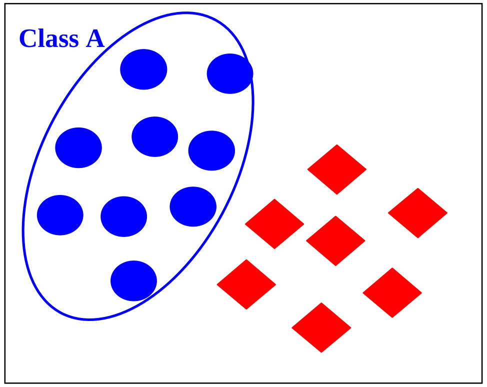
- For example there will be claims that method A is better than methods B, C and D
- The method will be claimed to be better as judged by the difference in one or more performance indicator such as %CC (percent correctly classified), usually on a test set and on one or more carefully chosen datasets.

Traditional problems : comparison of methods

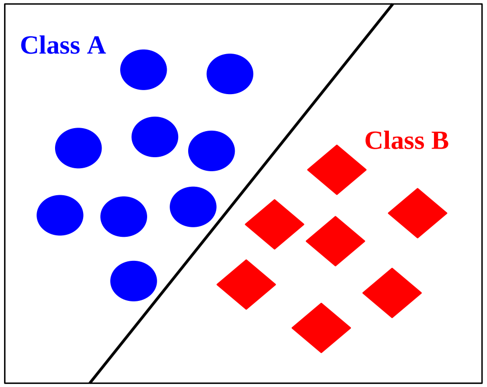
- There is strong pressure eg to get PhDs, get grants, get papers, or even conference presentations
- Often a method that isn’t “better” is regarded as a waste of time, no more grants, papers or PhDs
- Hence there are ever more claims of improved methods in the literature and at conferences.
- Beware.

Traditional problems : 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

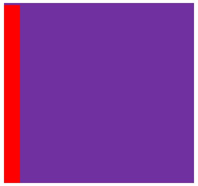
- Different types of verdict. How can you compare unlike with unlike?

	Group A (True Class)	Group B (True Class)
Group A (Predicted)		
Group B (Predicted)		

	Group A (True Class)	Group B (True Class)
Group A (Predicted)		
Group B (Predicted)		
Both groups (Predicted)		
Neither groups (Predicted)		

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

- 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

- Traditionally validation and comparison of classifiers was quite simple
  - Datasets quite straight-forward – known answers
  - Variables much fewer
  - Focus on algorithms
  - Most methods quite simple often based on straight-forward statistical choices, less decisions
- Modern problems much more complex
  - Take advantage of greater computer power
  - As there is more data, so computers get more powerful in parallel
  - Too much emphasis on algorithms, not enough on basics