Clustering and Classification

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Motivation/Definitions
- Classes of classification methods
- Linear Discriminant Analysis (LDA)
- K-means and K nearest neighbors (KNN)
- UNEQ
- SIMCA
- PLS Discriminant Analysis (PLS-DA)

EIGENVECTOR RESEARCH INCORPORATED

Motivation

- Often we want to know what "class" a particular sample belongs to:
  - Does this patient have liver disease?
  - Where did this oil come from?
  - Is this mushroom an amanita pantheria?
  - What chemical am I sensing?
- Many methods have been developed for classifying samples based on a multivariate response

EIGENVECTOR RESEARCH INCORPORATED

Definitions

Clustering: Identification of natural groupings (a.k.a. "classes") of samples without knowledge of their identity.

Classification: Using samples of known classes (or a model thereof) to identify the appropriate class of an unknown. "Supervised Classification" because we use known classes.

EIGENVECTOR RESEARCH INCORPORATED

Classification of Techniques

Parametric
Use information regarding the parent distribution:
LDA, QDA, SIMCA, UNEQ, PLS-DA

Non-Parametric
No use of information regarding parent distribution:
KNN, K-Means, ALLOC, PRIMA

Discriminating
Samples belong to one and only one class:
LDA, QDA, KNN, K-Means, ALLOC, PLS-DA

Modeling
Samples belong to one, none or several classes:
SIMCA, UNEQ, PRIMA

Probabilistic
Estimate degree of certainty of classification:
LDA, QDA, ALLOC, SIMCA, UNEQ, PLS-DA

Deterministic
Do not estimate degree of certainty:
KNN, K-Means, PRIMA

5

EIGENVECTOR RESEARCH INCORPORATED

Parametric vs. Nonparametric

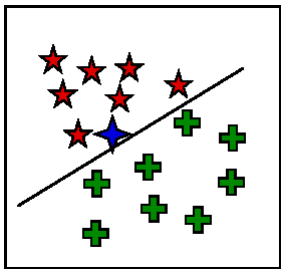
- To perfectly classify members, have data on all members of a class!
- Short of that, take representative sample
- Make assumptions about the distribution of the population to help make decisions
- Works well if assumptions are correct!
- If information is available about population distribution, it should be used

6

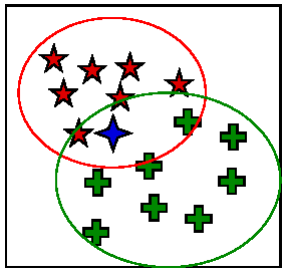
EIGENVECTOR RESEARCH INCORPORATED

Discrimination vs. Modeling

- Discrimination techniques emphasize differences between classes and try to set boundaries
- Modeling techniques emphasize similarities within classes



Discrimination



Modeling

EIGENVECTOR RESEARCH INCORPORATED

Probabilistic vs. Deterministic

- Probabilistic: classes have grey boundaries
- Deterministic: classes have sharp boundaries
- Probabilistic classifications often based on Bayes theorem:

$$\mathrm{P}(Q_q \mid X_k) = \frac{\mathrm{P}(X_k \mid Q_q)\mathrm{P}(Q_q)}{\sum_{l=1}^{r} \mathrm{P}(X_k \mid Q_l)\mathrm{P}(Q_l)}$$

The expression $\mathrm{P}(Q_q|X_k)$ is read: the probability of class Q_q given the data X_k

EIGENVECTOR RESEARCH INCORPORATED

Bayes Theorm Example

- Suppose 90% of our population of people are well (H_1), 10% have a special disease (H_2)
- Thus the prior probability of a person being well is: $P(H_1) = 0.90$
- Let T_{neg} = a negative test for illness, T_{pos} = a positive test for illness
- Probability of a negative test if they are well is $P(T_{neg}|H_1) = 0.95$, thus $P(T_{pos}|H_1) = 0.05$ (false positive or Type 2 error)
- Probability of a positive test if they are ill is $P(T_{pos}|H_2) = 0.50$, thus $P(T_{neg}|H_2) = 0.50$ (false negative or Type 1 error)
- So the probability they are well if test is negative is

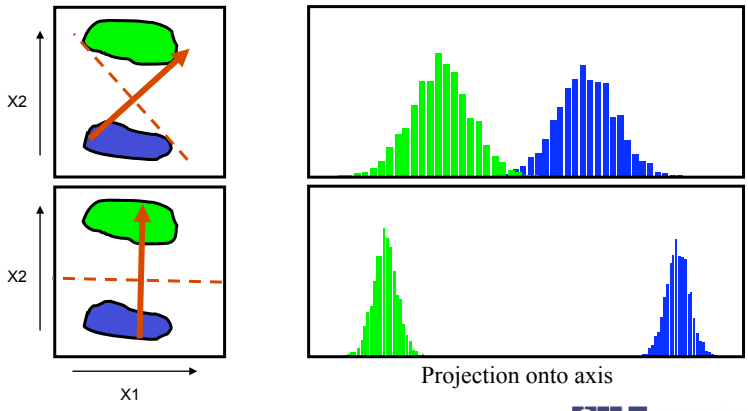
$$P(H_1 \mid T_{neg}) = \frac{P(T_{neg} \mid H_1)P(H_1)}{P(T_{neg} \mid H_1)P(H_1) + P(T_{neg} \mid H_2)P(H_2)} = 0.945$$

Linear Discriminant Analysis

- LDA is parametric, probabilistic and discriminating
- Example: two classes of vapors (polar and nonpolar) on SAWs coated with two polymers
- Want to determine axis to project data on that discriminates between classes
  - choose axis so individual distributions are narrow
  - choose axis so centers of distributions are far apart

Linear Discriminant Analysis

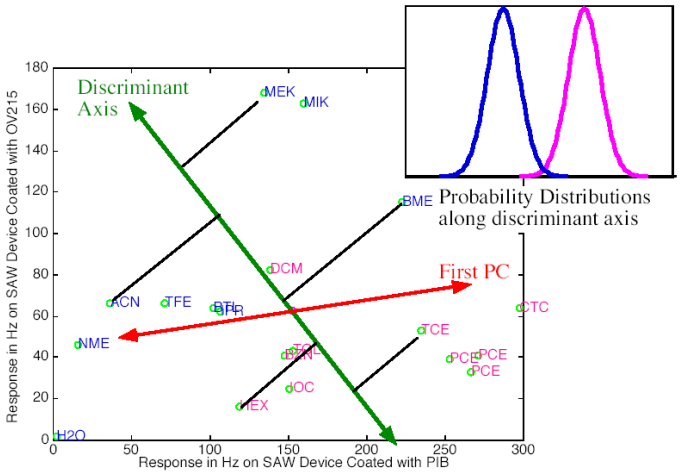
- LDA seeks axis (in n-D space) which maximizes ratio of between class to within class variance



EIGENVECTOR RESEARCH INCORPORATED

11

LDA Example



EIGENVECTOR RESEARCH INCORPORATED

12

LDA Assumptions

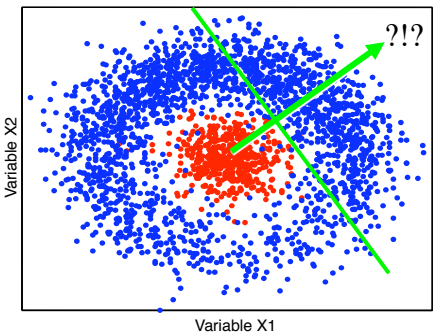
- Boundaries between groups are midpoints between centroids of adjacent groups
- Equivalent to assuming distributions identical
- Often not the case
- In Quadratic Discriminant Analysis (QDA) distributions not assumed equal, but model is much more complex

EIGENVECTOR RESEARCH INCORPORATED

Non-Linear Donut Problem

- Question: How does LDA handle a problem in which one class is "inside" the other?
- Non-linear problem! Discrimination vector points from center of donut, outward – but you can't draw a [straight] tangent line which closes off the inside!

(more on this problem later)



EIGENVECTOR RESEARCH INCORPORATED

Problems with LDA

- Discriminating axis is calculated to maximize ratio of between to within group variance by maximizing the probability P(k|x)

$$\max(P(k \mid x)) = \max[-0.5(x-\bar{x}_k)^{\mathrm{T}} C_k^{-1}(x-\bar{x}_k) - 0.5\ln|C_k| + \ln\pi_k]$$

- $(x-\bar{x}_k)^T C_k^{-1}(x-\bar{x}_k)$ is the squared Mahalanobis distance
- Problem is calculating the covariance matrix inverse C_k^{-1} which may not exist if data is collinear.

The Collinearity Problem

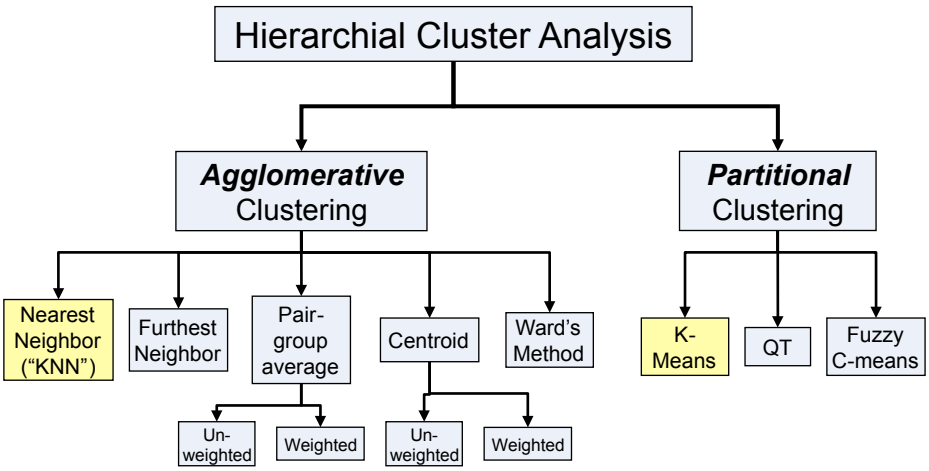
- Collinearity is a problem in many applications in analytical chemistry, particularly spectroscopy
- Solution: choose subset of variables that form an independent set
- Problems:
  - How to choose? Often very many combinations, *e.g.* 83,218,600,080 ways to choose 5 from 400
  - Lose multivariate advantage: signal averaging, outlier detection
  - Compare to regression methods: MLR vs. PCR

Cluster Analysis

- Implies ***unsupervised*** learning
  - Object groupings are **NOT** known *a priori*
  - Objects are grouped based only on their data
- **Agglomerative Clustering:** Start with each object as it's own cluster, then ***combine*** these into larger clusters
  - Ex. Nearest-Neighbor ("KNN")
- **Partitional Clustering:** Start with all objects in one cluster, then ***separate*** them into smaller clusters
  - Ex. K-means

EIGENVECTOR RESEARCH INCORPORATED

Cluster Analysis Methods



EIGENVECTOR RESEARCH INCORPORATED

k-Nearest Neighbor

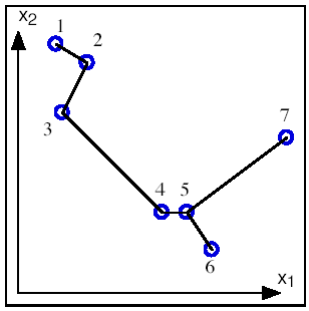
- KNN is non-parametric, discriminating, deterministic and very simple
- The distance between samples is calculated and the nearest samples are found
- Used as both a clustering and classification method

19

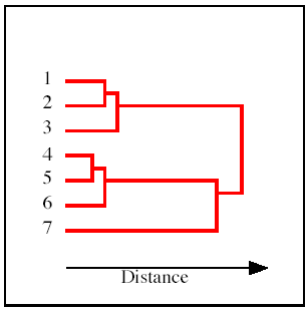
EIGENVECTOR RESEARCH INCORPORATED

KNN Clustering

- Closest samples are linked together to form groups, then groups are linked
- Results are often displayed as a dendrogram



Samples connected to nearest neighbors

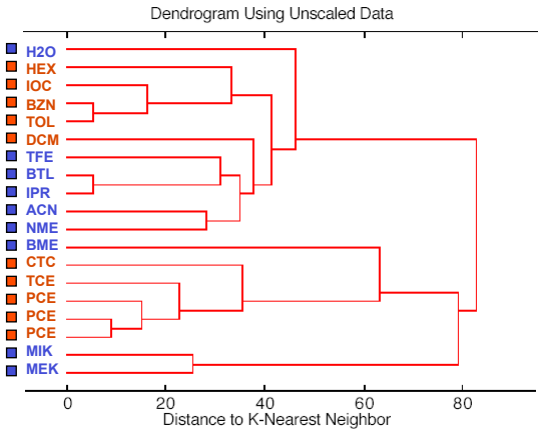


Resulting dendrogram

20

EIGENVECTOR RESEARCH INCORPORATED

KNN Dendrogram



EIGENVECTOR RESEARCH INCORPORATED

Distances in KNN

Distance can be defined several ways

- Simplest is Euclidean distance

$$d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$$

- Can also use distance on PC scores

$$d_{ij} = [(\mathbf{t}_i - \mathbf{t}_j)^T(\mathbf{t}_i - \mathbf{t}_j)]^{1/2}$$

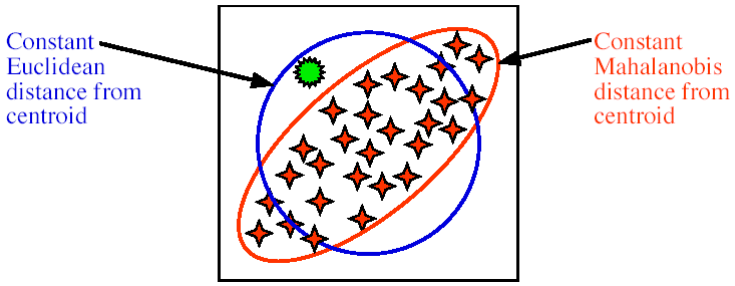
- Or Mahalanobis distance

$$d_{ij} = [(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$$

EIGENVECTOR RESEARCH INCORPORATED

Mahalanobis Distance

- Mahalanobis distance measure accounts for fact that changes in some directions are less likely (and therefore more significant) than changes in other directions

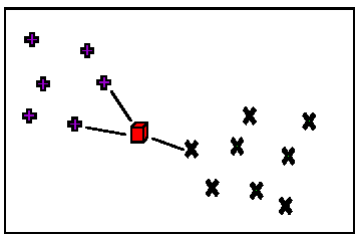


EIGENVECTOR RESEARCH INCORPORATED

23

Classification in KNN

- Classification of unknowns can be done using a voting method.
- Locate an *odd* number of closest samples to an unknown. The group assignment that is most represented is assumed to be correct for unknown.



unknown (▣) is near to 2 ✚ samples and 1 ✖ so is presumed to be a ✚

EIGENVECTOR RESEARCH INCORPORATED

24

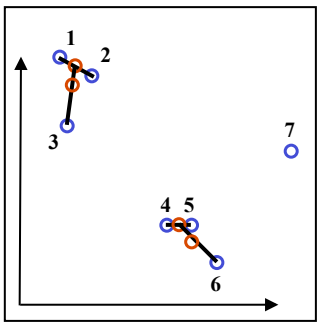
Agglomerative Clustering Methods

Method	Distance Between Existing Clusters	Linkage Rule
Nearest Neighbor	Minimum of pair-wise distances between any two objects in each cluster	join 2 nearest clusters
Furthest Neighbor	Maximum of pair-wise distances between any two objects in each cluster	join 2 nearest clusters
Pair-Group Average	Average distance between all pairs of objects in each cluster	join 2 nearest clusters
Centroid or Mean	Distance between centroid or mean of each cluster	join 2 nearest clusters
Ward's Method	N/A	Join clusters such that the resulting within-cluster variance (with respect to centroids) is minimized

EIGENVECTOR RESEARCH INCORPORATED

k-Means Agglomerative Clustering

- Samples are paired with another sample or a cluster one-at-a -time
- Position of each cluster is mean of all samples in cluster.
- Recalculation of distance can take a long time with lots of samples



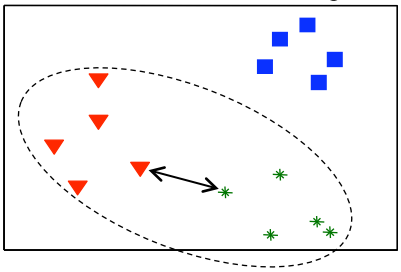
EIGENVECTOR RESEARCH INCORPORATED

KNN vs. K-Means

Two clusters are grouped together when…

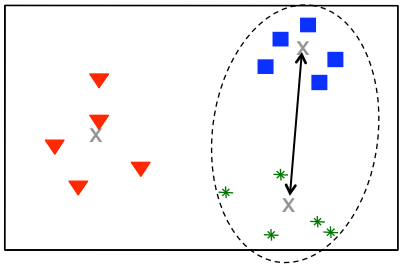
KNN

…two of their members are the closest of all dissimilar samples



K-Means

…the cluster means are the closest of all cluster means



X = cluster mean

Note: these rules apply even when one of the "groups" is a single sample in a group of its own.

EIGENVECTOR RESEARCH INCORPORATED

27

k-Means Partitional Clustering

- Choose k samples as cluster "targets"
  - random selection of samples
  - "pure samples": choose samples on outside of data (furthest from all other samples)
- Classify all samples into one of those k clusters.
- Calculate mean of each cluster's samples
- Repeat classification and cluster means until no samples are re-classed after mean recalculation.
- Much faster, but dependent on initial guess of samples

EIGENVECTOR RESEARCH INCORPORATED

28

TOF-SIMS of Time Released Drug Delivery System

- Multilayer drug beads serve as controlled-release delivery system
- TOF-SIMS taken of cross-section of bead
- Evaluate integrity of layers, distribution of ingredients
- Thanks to Physical Electronics and Anna Belu for the data

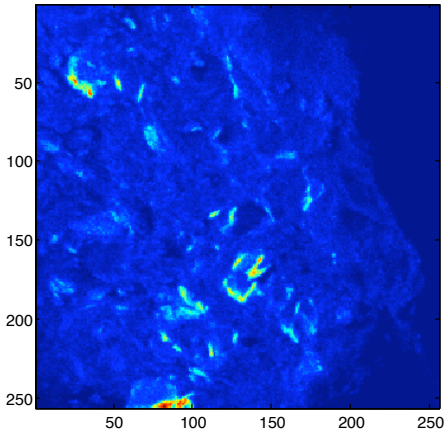
Reference: A.M. Belu, M.C. Davies, J.M. Newton and N. Patel, "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems, Anal. Chem., 72(22), pps 5625-5638, 2000

29

EIGENVECTOR RESEARCH INCORPORATED

Imaging Mass Spec

- Image is 256x256x90
- The mass spectrum was 41945 mass channels selected and binned into 93 channels
- Image of total ion count
  - false color

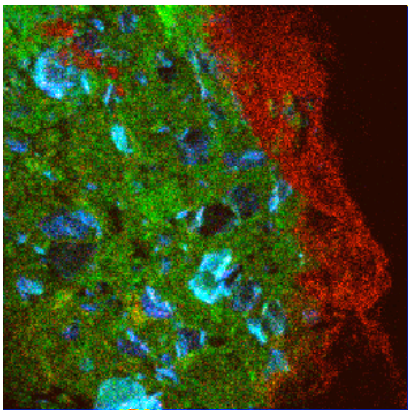


30

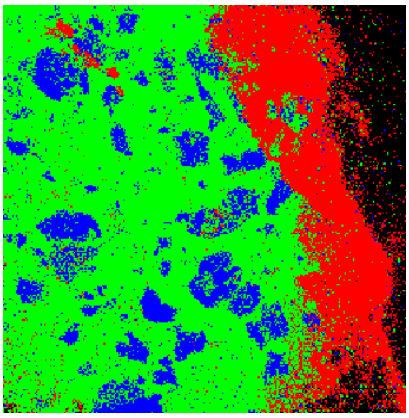
EIGENVECTOR RESEARCH INCORPORATED

Avicel Pure-Sample k-Means Clustering

False-color MCR Results



Pure Pixel Clusters



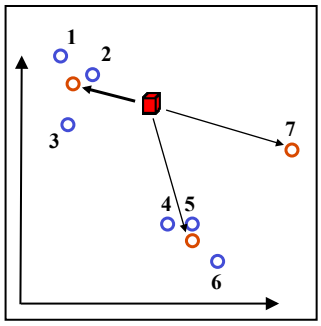
(3 clusters)

EIGENVECTOR RESEARCH INCORPORATED

31

k-Means Classification

- Done as single-nearest-neighbor classification (k=1) using cluster means from clustering on calibration data as samples.



unknown (■) is closest to, and therefore presumed to be in the 1/2/3 cluster

EIGENVECTOR RESEARCH INCORPORATED

32

Agglomerative Clustering Variations

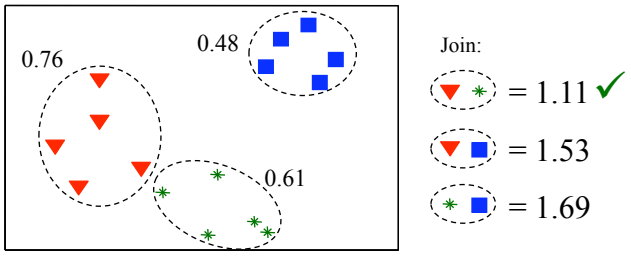
- Distance between groups can be defined as distance between centroid (K-Means) or furthest points
- KNN Classification can be done using number of closest samples but also *distance* to each of those samples
- KNN Distance can be used to estimate certainty of assignment

33

EIGENVECTOR RESEARCH INCORPORATED

Example: Ward's Clustering Method

Group sample into cluster or two clusters together when that association causes the minimum change in sum squared deviation from the cluster mean.

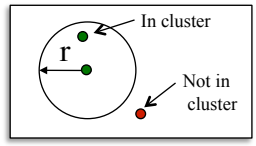


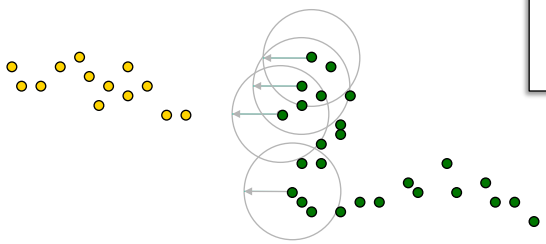
34

EIGENVECTOR RESEARCH INCORPORATED

Example: Density Based Scan (DBSCAN)

- Agglomerative clustering: Connect samples which are within a specified distance.
- Works well for unusual shaped clusters.





```
See: dbscan.m    (help dbscan)
```

EIGENVECTOR RESEARCH INCORPORATED

35

Partitional Clustering Variations

- K-means (Mean of each cluster)
  - User selects *number of clusters*
  - Membership based on distance from cluster center
- QT (Quality Threshold)
  - User selects *maximum cluster diameter*
  - Membership based on distance from cluster center
- Fuzzy C-means
  - User selects *number of clusters*
  - Every object has some "degree of membership" to each cluster

EIGENVECTOR RESEARCH INCORPORATED

36

Advantages and Problems with KNN and k-Means

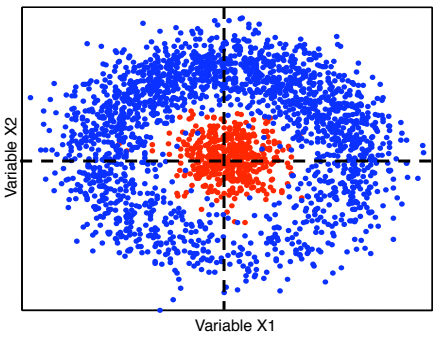
- Although easy to update, KNN and k-means classification "models" are highly sensitive to the calibration data supplied.
- Can classify with non-linear behavior if sufficient sampling is achieved.
- Does NOT explicitly take "density" of samples into account (Except Ward's method)

EIGENVECTOR RESEARCH INCORPORATED

Non-Linear Donut Problem

- Question: How does KNN handle a problem in which one class is "inside" the other?
- Non-linear problem but KNN does just fine!
- k-Means, however, does not do so well (mean of the outer circle is same as inner circle)

(yet more on this later)



EIGENVECTOR RESEARCH INCORPORATED

UNEQ

- A parametric, probabilistic, modeling technique for UNEQually dispersed classes
- Each class is modeled by its centroid and its covariance.
- A generalized distance is calculated from the class centroid to each sample:

$$\tilde{d}(k, M_q) = \left\{ d^2 \frac{(n_q - p - 2)}{n_q - (p / n_q)} \right\}^{\frac{1}{2}}$$

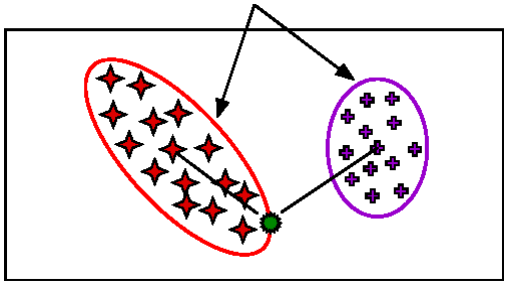
d^2 is the Mahalanobis distance, p is the number of variables, n_q is the number of objects

$$d^2 = \left(x_k - x_q\right)^T C_p^{-1} \left(x_k - x_q\right)$$

EIGENVECTOR RESEARCH INCORPORATED

Classification with UNEQ

95% confidence limit on class boundaries



EIGENVECTOR RESEARCH INCORPORATED

Problems with UNEQ

- Limited by number of variables that can be used due to collinearity problem
- Consider: PCA produces new orthogonal variables – could use UNEQ approach on PCA scores!
- Leads to Soft Independent Modeling of Class Analogy (SIMCA)

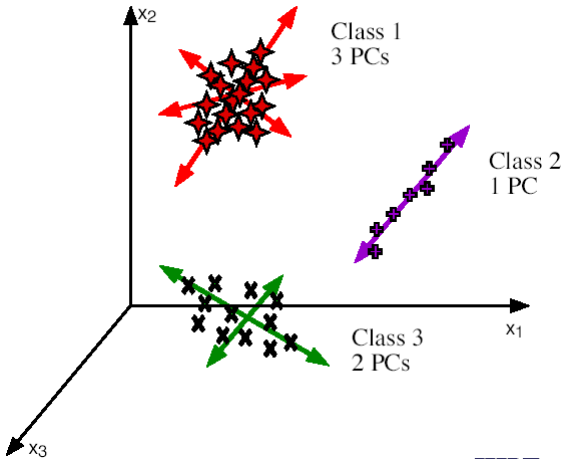
EIGENVECTOR RESEARCH INCORPORATED

SIMCA

- SIMCA is a parametric, probabilistic, modeling technique
- Each class is described by an independent PCA model
- New samples are compared with the existing PCA models to determine if they belong to each class
- Samples can belong to one, none or several classes

EIGENVECTOR RESEARCH INCORPORATED

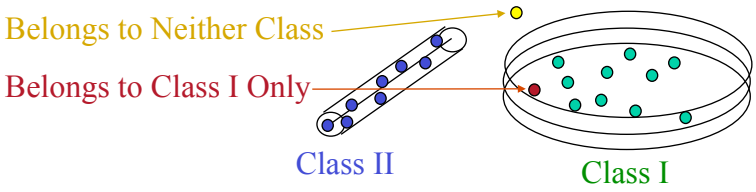
A SIMCA Model

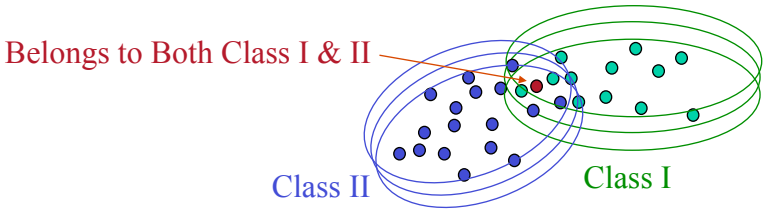


EIGENVECTOR RESEARCH INCORPORATED

43

SIMCA Class Assignment





EIGENVECTOR RESEARCH INCORPORATED

44

SIMCA Summary

- A SIMCA model is a collection of PCA models, one for each class
- Each PCA model consists of
  - a vector describing the class mean
  - a vector describing the variance scaling, if any
  - some number of principal component vectors
  - the statistical limits on Q
  - the statistical limits on Hotelling's T^2
- When a SIMCA model is applied, new samples are compared with all the class models
- Samples belong to one, none or several classes, based on distance on Q and T^2 and confidence limits

EIGENVECTOR RESEARCH INCORPORATED

More on SIMCA Class Assignment

- For a given class model, Q and Hotelling's T^2 can have statistical limits put onto them. Falling inside both limits implies the sample is a member of the class.
- Can also determine class assignment based on just one or the other statistic (Q is often most sensitive of the two)
- The confidence levels for the observed value of Q and T^2 can also be calculated to determine probability for the given class.

EIGENVECTOR RESEARCH INCORPORATED

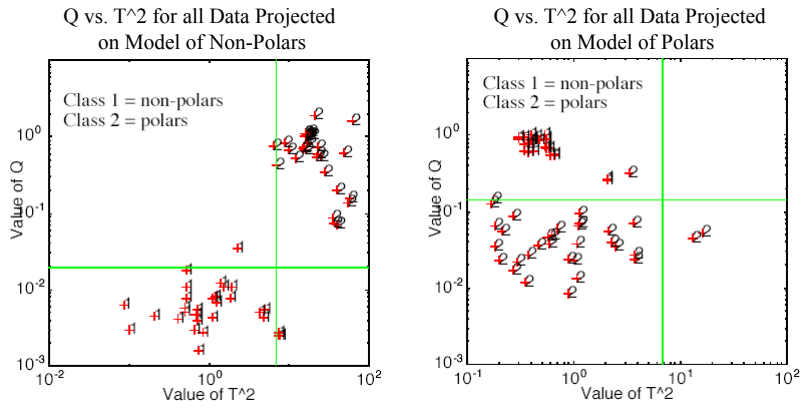
SIMCA Example: SAW Sensors

- Data consists of responses of 13 SAW sensors with different coatings exposed to 19 analytes
- Goal: develop SIMCA model that discriminated polar vapors from non-polar
- Preprocessing: normalize responses to vectors of unit length (attempt to make response independent of concentration)
- Develop model on 3/4 of the data, test on remaining 1/4

47

EIGENVECTOR RESEARCH INCORPORATED

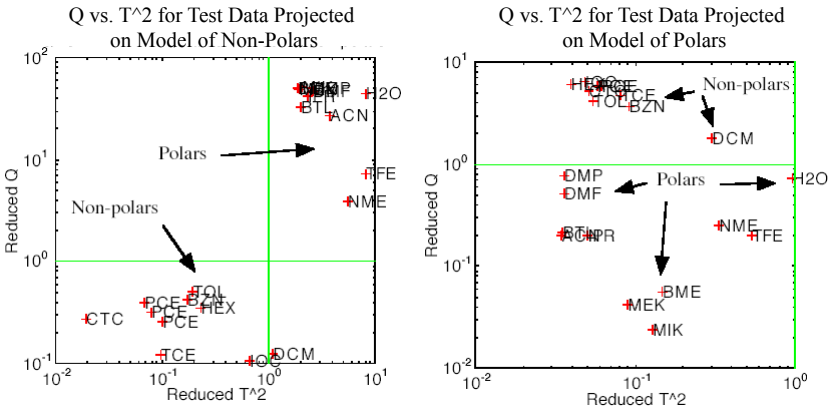
Individual PCA Models on Training Data



48

EIGENVECTOR RESEARCH INCORPORATED

Individual PCA Models on Test Data



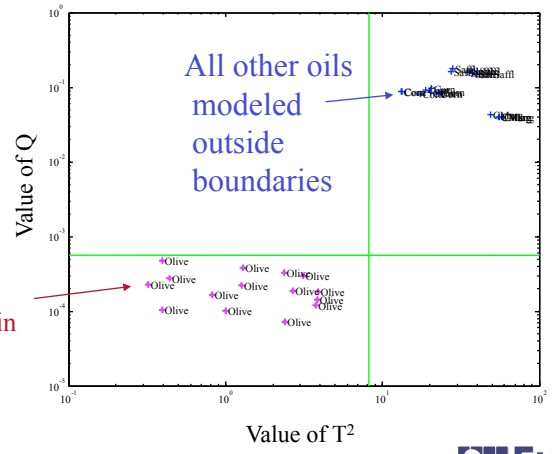
EIGENVECTOR RESEARCH INCORPORATED

SIMCA Example: FTIR of Edible Oils

- Use FT-IR spectra and pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.
- Calibration Data consists of corn oil, olive oil, safflower oil, and corn margarine
- Test data consists of new samples of all calibration oils plus corn oil in olive oil (5, 10, 20, 30 & 40%), almond oil, peanut oil, and sesame oil.
- Used Multiplicative Scatter Correction (MSC) to correct for baseline and scaling variations and mean-centering to each individual class.

EIGENVECTOR RESEARCH INCORPORATED

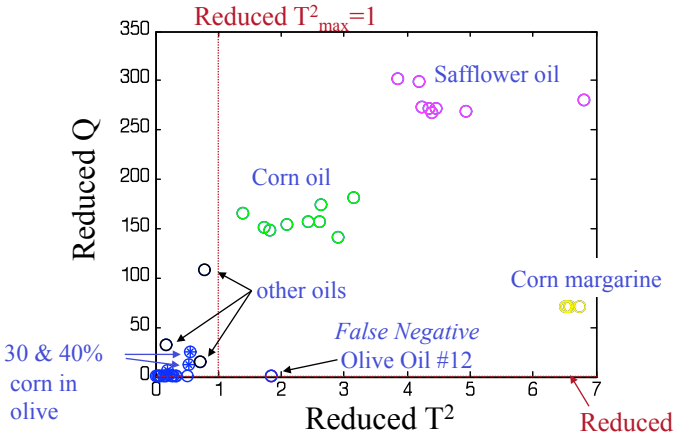
All Calibration Samples Projected onto Olive Oil Model



EIGENVECTOR RESEARCH INCORPORATED

51

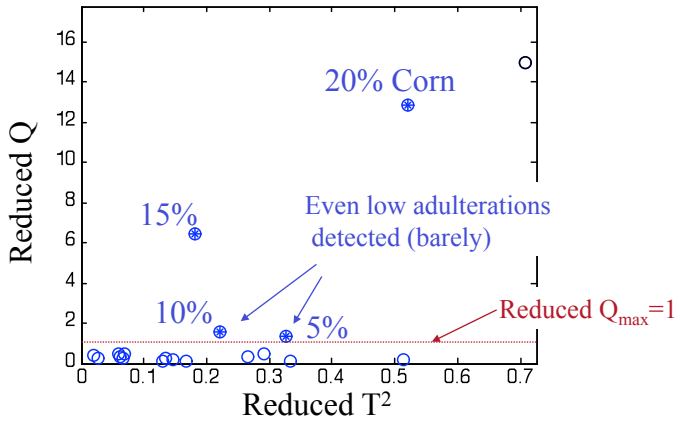
Test Set Projected onto Olive Oil Model



EIGENVECTOR RESEARCH INCORPORATED

52

Test Set Projected onto Olive Oil Model

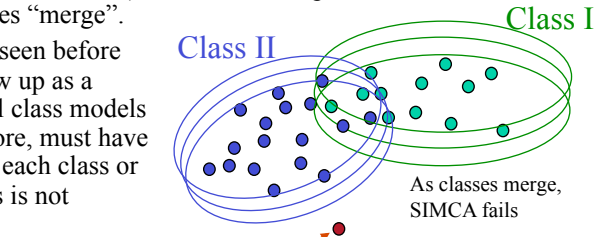


EIGENVECTOR RESEARCH INCORPORATED

53

Problems with SIMCA

- SIMCA is forced to account for all variance in a class, whether or not it is unique to that class (as with PCR vs PLS)
- If the between-class variation is smaller than the within-class variation (or if too many PCs are used) the model false positive rate will increase as classes "merge".
- A new class, not seen before will usually show up as a "negative" on all class models (high Q) Therefore, must have PCA models for each class or unexpected class is not alarmed.

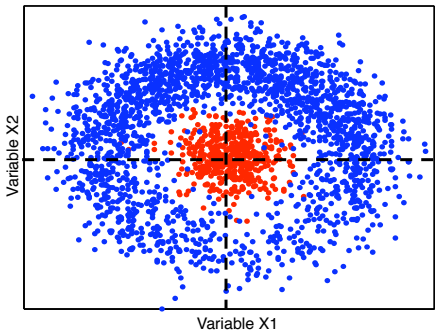


EIGENVECTOR RESEARCH INCORPORATED

54

Non-Linear Donut Problem (again)

- Question: How does SIMCA handle a problem in which one class is "inside" the other?
- Inside class works fine (=PCA with T^2 limit)
- Outside class NOT modelable (except when including inside class)
- Do by "exclusion" – to be outside class, it must be in the combined class but not in inside class.



EIGENVECTOR RESEARCH INCORPORATED

55

Partial Least Squares Discriminate Analysis (PLS-DA)

- PLS-DA is parametric, probabilistic and modeling
- Exactly as with LDA, we want to determine axis to project data on that discriminates between classes
  - choose axis so individual distributions are narrow
  - choose axis so centers of distributions are far apart
- Determine axes from factor-based model of data therefore more stable with high collinearity.
- Will automatically attempt to identify directions of interest!

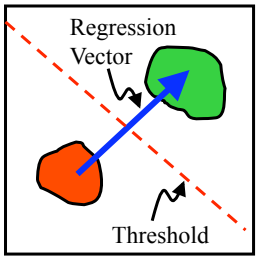
EIGENVECTOR RESEARCH INCORPORATED

56

Partial Least Squares Discriminate Analysis (PLS-DA)

- Use logicals (0,1) in Y-block to indicate if sample belongs to a class or not.
- Develop PLS model to predict class block
- Thresholds must be set between 0 and 1 to indicate if new samples are a member of each class...

  Can use Bayes theorem to set threshold and include prior probability of each class

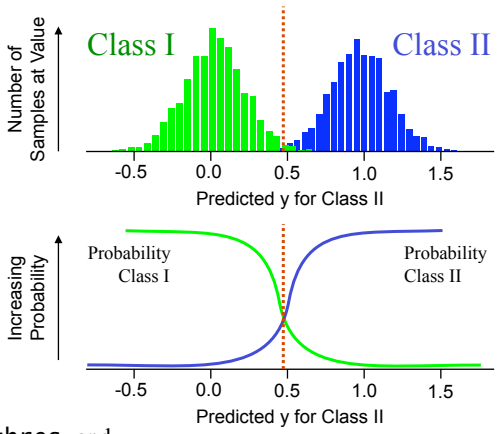


EIGENVECTOR RESEARCH INCORPORATED

57

Thresholds in PLS-DA

Observed distribution of predictions can be handled in a straight-forward Bayesian way



see `plsdthres` and `discrimprob` functions

EIGENVECTOR RESEARCH INCORPORATED

58

Why PLSDA?

- PCA-based models (like SIMCA) capture variance within the data set, whether or not that variance is useful for separating classes.
- PLS-DA tends to capture variance which is useful in separating classes and ignoring variance within a class. (goal: maximize between-group variance while minimizing within-group variance)
- The result is a model which is generally superior at separating classes (but requires knowledge of classes being separated)

EIGENVECTOR RESEARCH INCORPORATED

Example: Identify Pinot Noir Wine According to its Region of Origin

We have the following bottles of wine:

Pacific Northwest	17
California	9
France	12

How shall we distinguish the wines?

Wine is mostly water and alcohol.

Need to look at trace differences to distinguish between wines of different types and/or origins.

EIGENVECTOR RESEARCH INCORPORATED

Wine Data X-Block

17 Trace metals concentrations as determined by Atomic Emission Spectroscopy (ppm):

Cd	Mo	Mn	Ni	Cu	Al	Ba
Cr	Sr	Pb	B	Mg	Si	Na
Ca	P	K				

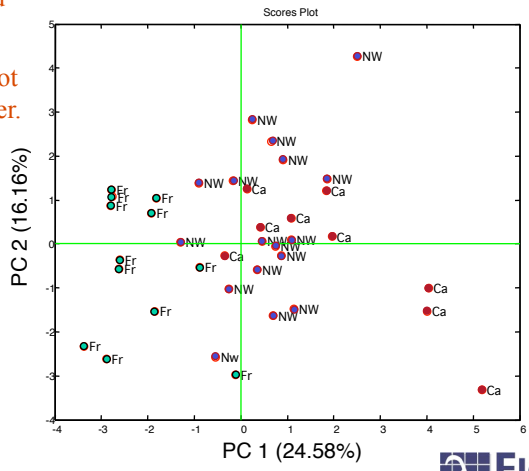
Preprocessed with autoscaling

EIGENVECTOR RESEARCH INCORPORATED

PCA Scores Plot for X-Block of All 38 Samples

Not very good resolution!

SIMCA does not work well either.



This suggests that the *major* sources of variance in the data set are not due to differences in the region.

VERY COMMON PROBLEM!

EIGENVECTOR RESEARCH INCORPORATED

Try PLS-1 with the Y-Block Representing the Origin of the Wine

We could represent the region of origin using numbers:

Pacific Northwest	1
California	2
France	3

Then do a PLS-1

EIGENVECTOR RESEARCH INCORPORATED

A Bad Idea!

- Such a system implies that California wine is somehow in between Pacific Northwest and French wines
- We need to ask the questions:
  - Is it a Pacific Northwest wine? Yes or No?
  - Is it a California wine? Yes or No?
  - Is it a French wine? Yes or No?

EIGENVECTOR RESEARCH INCORPORATED

A Better Way of Expressing the Y-Block:

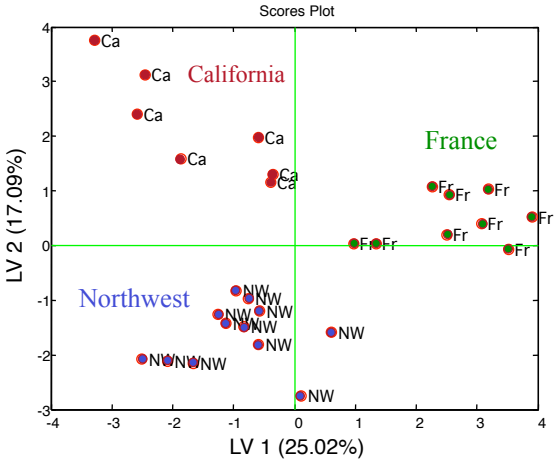
Sample	NW	Ca	Fr
1	1	0	0
2	0	1	0
3	0	0	1

NW wine
Not French wine
Not Calif. wine

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot LV2 vs LV1

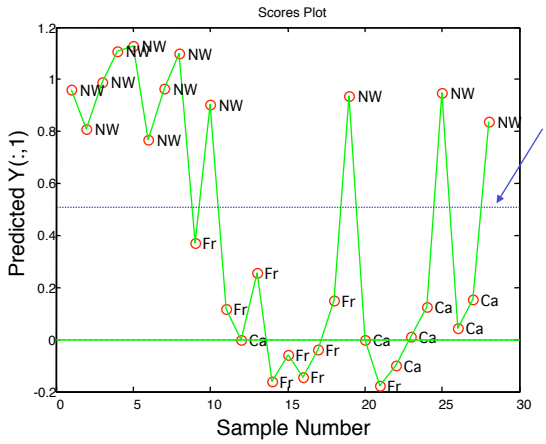
Much Better Separation



LV 1+2
~42.11%

EIGENVECTOR RESEARCH INCORPORATED

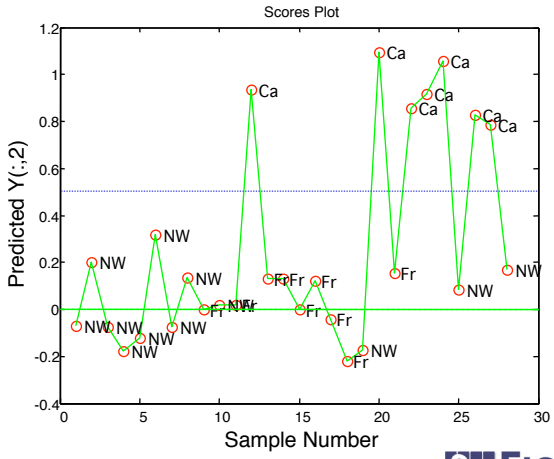
Estimation of Learning Set Northwest Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

67

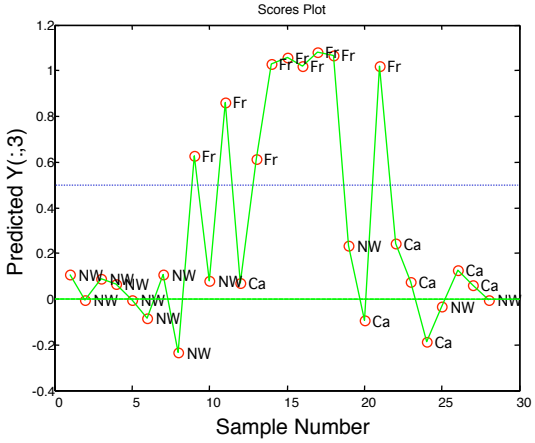
Estimation of Learning Set California Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

68

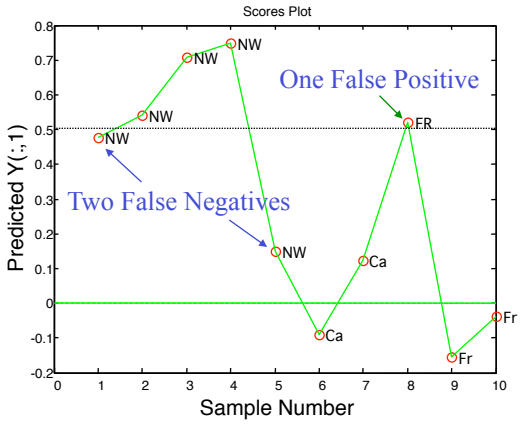
Estimation of Learning Set
French Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

69

Prediction of Test Set
Northwest Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

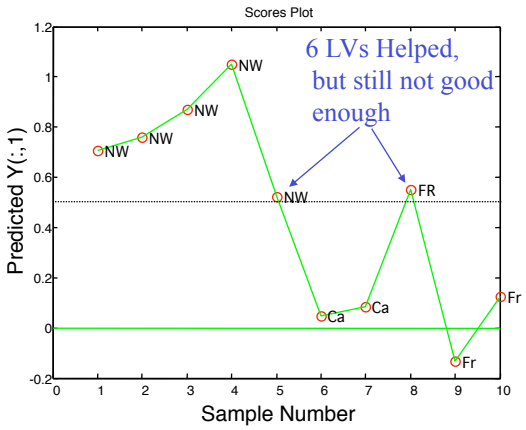
70

What's the Problem?

- Cross-validation showed that the Northwest Wines needed 6 LVs to achieve best separation and our PLS-2 model used only 4 LVs.
- Were all types of Northwest Wines represented in the Learning Set?
- Were there enough samples of all types in the Learning Set to really define the groups?
- Were all the samples labeled correctly?

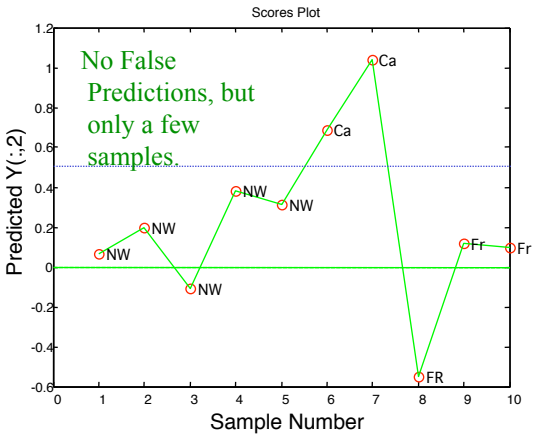
EIGENVECTOR RESEARCH INCORPORATED

Prediction of Test Set Northwest Wines Using 6 LVs



EIGENVECTOR RESEARCH INCORPORATED

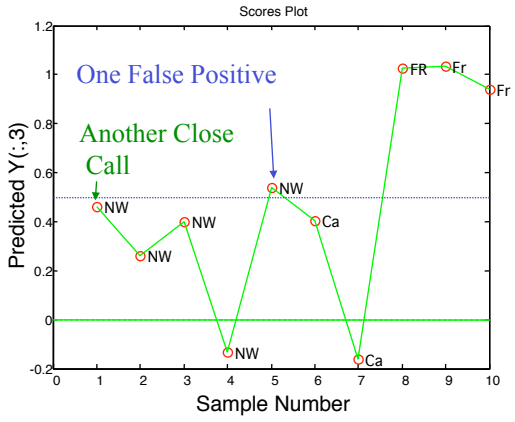
Prediction of Test Set California Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

73

Prediction of Test Set French Wines Using 4 LVs



EIGENVECTOR RESEARCH INCORPORATED

74

This is a Nice Preliminary Study

It demonstrates a great deal of promise for this approach

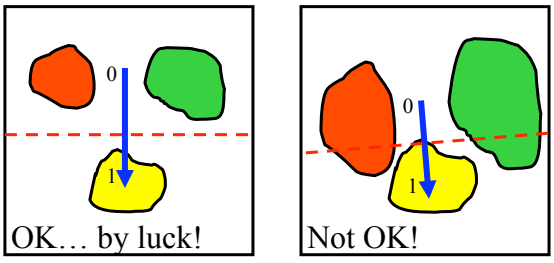
Because of the heterogeneous nature of Pinot Noir wines within a geographic origin, I want a lot more samples for both the Learning and Test Sets.

75

EIGENVECTOR RESEARCH INCORPORATED

Multiple-Class PLSDA models

Attempting to discriminate one group from several other groups with a single regression vector may not provide best separation.

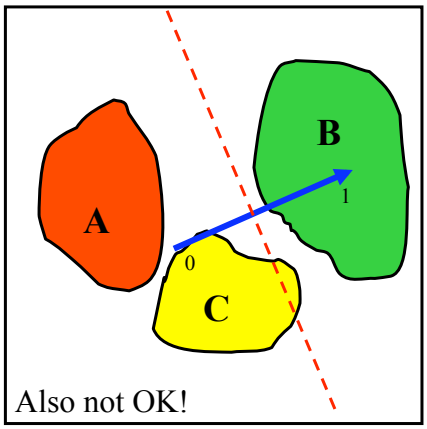


Regression vector
Threshold

76

EIGENVECTOR RESEARCH INCORPORATED

Multiple-Class PLSDA models



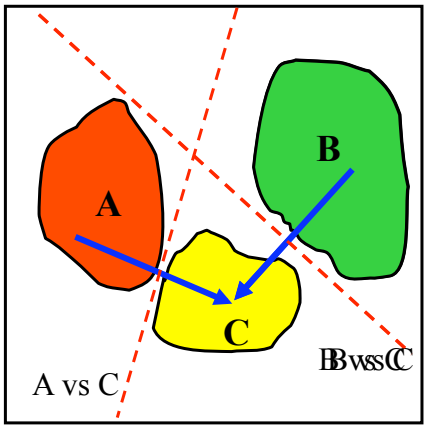
Even B versus A and C may be a problem because PLSDA will attempt to make A and C both zero

Regression vector
Threshold

EIGENVECTOR RESEARCH INCORPORATED

77

Multiple-Class PLSDA models



Splitting into multiple PLSDA models (one class vs one class) provides better results.

Regression vectors can be optimized for two-group separation and separate loadings can be selected for each class paring.

Use in a multi-block PLS-DA: Use prediction results from one-on-one models in a master PLS-DA model.

Regression vector
Threshold

EIGENVECTOR RESEARCH INCORPORATED

78

Problems with PLS-DA

- Regression method: Temptation to overfit is always there (cross-validation should be used). Sufficient sampling of all classes is necessary.
- Assumes linear (or approximate) plane can be drawn to separate classes (as does LDA).
- When you want to add new classes, they must be re-modeled against all other classes.

EIGENVECTOR RESEARCH INCORPORATED

Classification Preprocessing: Questions to consider…

- Does intensity matter? Do you care about aboslute signal level or just whether a particular covariance is there? Normalization is *common* in classification (particularly with quantitative analytical methods!)
- Are there extraneous sources of variation within your groups that might make them look more similar than they are? Consider "pre-whitening" such as with GLSW or OSC.
- Are there sources of variation between the groups which is not related to the group (systematic error)? Use baseline correction or calibration transfer to remove variations, or adjust your experimental design.

EIGENVECTOR RESEARCH INCORPORATED

Classification Summary

- KNN and the like are simple and unassuming but provide no warning when sample doesn't fit (i.e. unmodeled class gives false positive!). Works with non-linear systems but only with sufficient sampling. New classes added easily.
- SIMCA works well for identifying one class among many others without requiring models of other classes. If PCA works well for observing clusters, SIMCA will work well (and may even when PCA is cluttered). Unmodeled classes usually show up as true false, but no indicator that it is otherwise unusual. Easy to add new classes (just create new class model, add to others).

EIGENVECTOR RESEARCH INCORPORATED

Classification Summary (cont…)

- PLS-DA works well when all expected classes are known and can be superior when within-class differences are significant relative to between-class differences. Unmodeled classes detected by unusually high Q or T^2. New classes added by recalculating all models.
- Preprocessing: Consider the physics/chemistry – what do you expect to be different between the classes and how should that manifest itself? What other effects might be present – how can you remove those?
- Sample ID errors will cause problems with all methods!

EIGENVECTOR RESEARCH INCORPORATED