

Pattern Recognition and NMR Spectroscopy

A Tool for Investigation of Metabolic Responses to Toxins

Tim Ebbels*, George Tranter, Elaine Holmes and Jeremy Nicholson

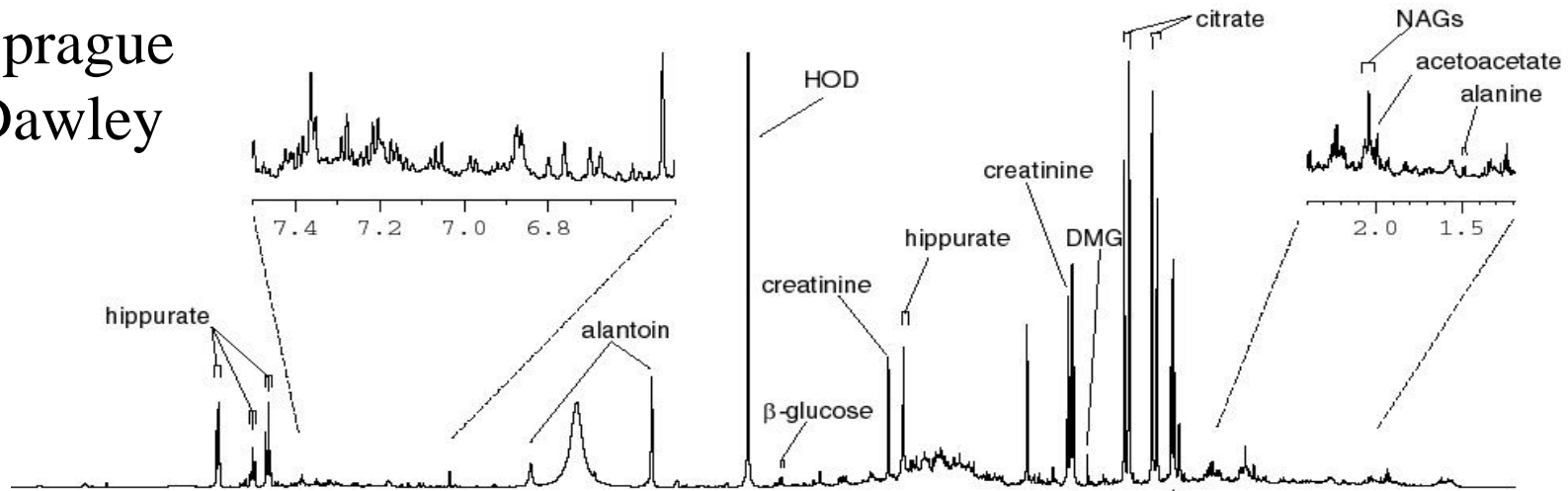
Biological Chemistry, Imperial College School of Medicine, SAF Building, Exhibition
Road, London SW7 2AZ. email: t.ebbels@ic.ac.uk

- NMR spectra of biofluids
- Pattern recognition methods
- PCA
- SIMCA
- Multi-dimensional Gaussian Class Modelling

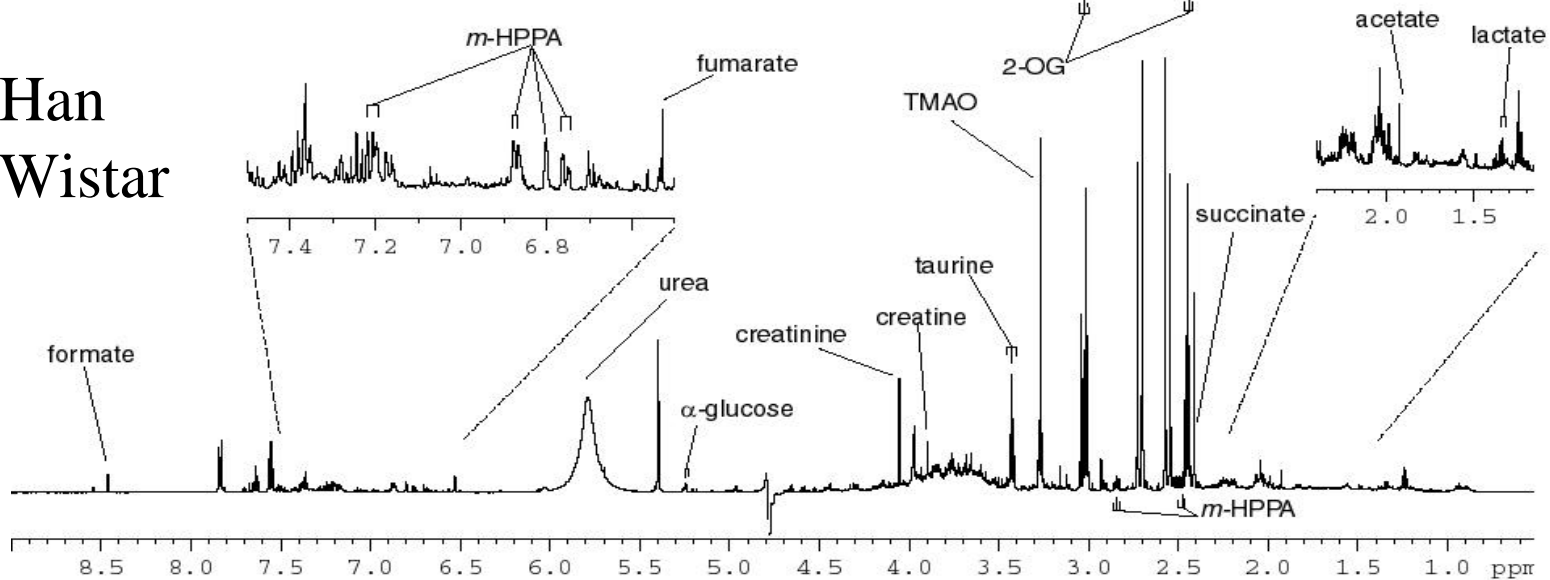
*Presenting author

^1H NMR Rat Urine Spectra

Sprague
Dawley



Han
Wistar

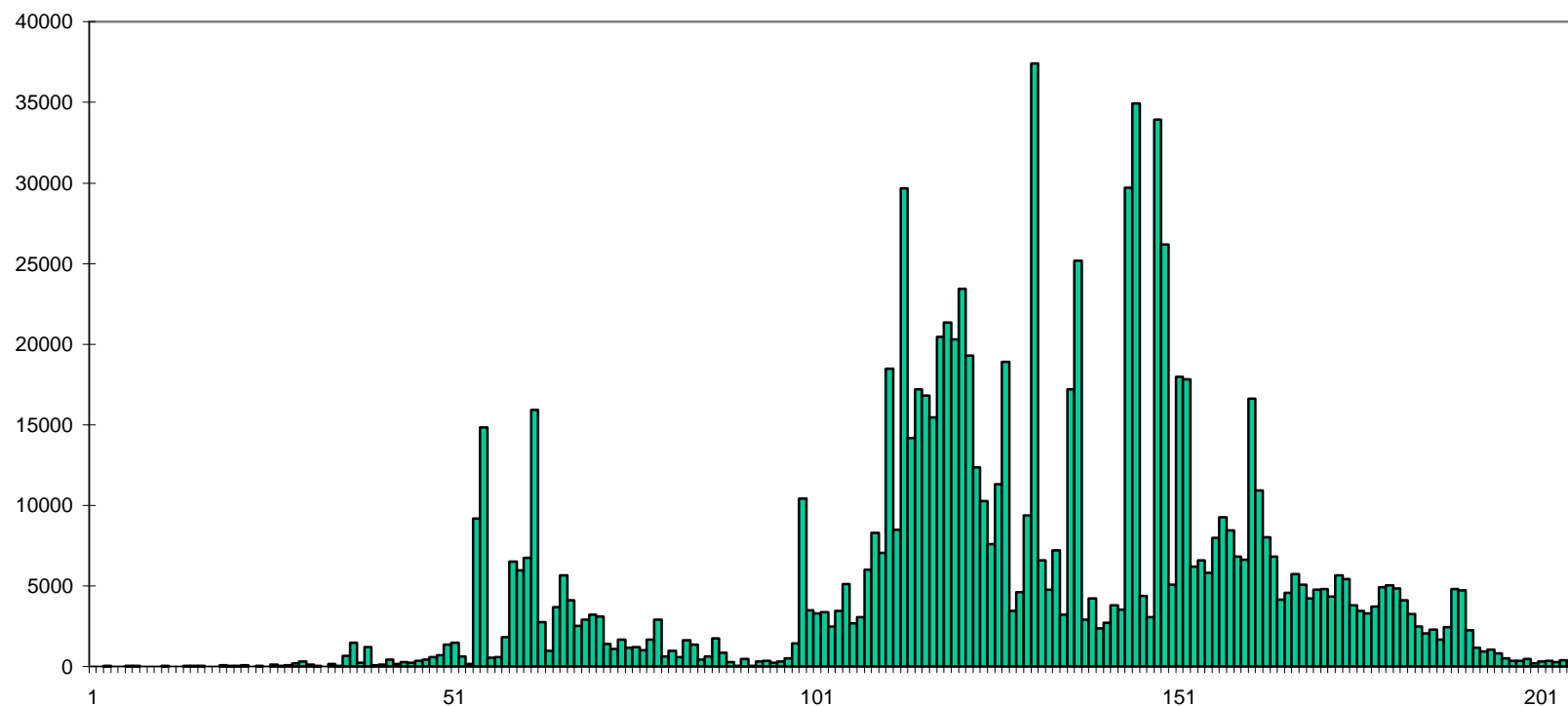


NMR spectra of biofluids

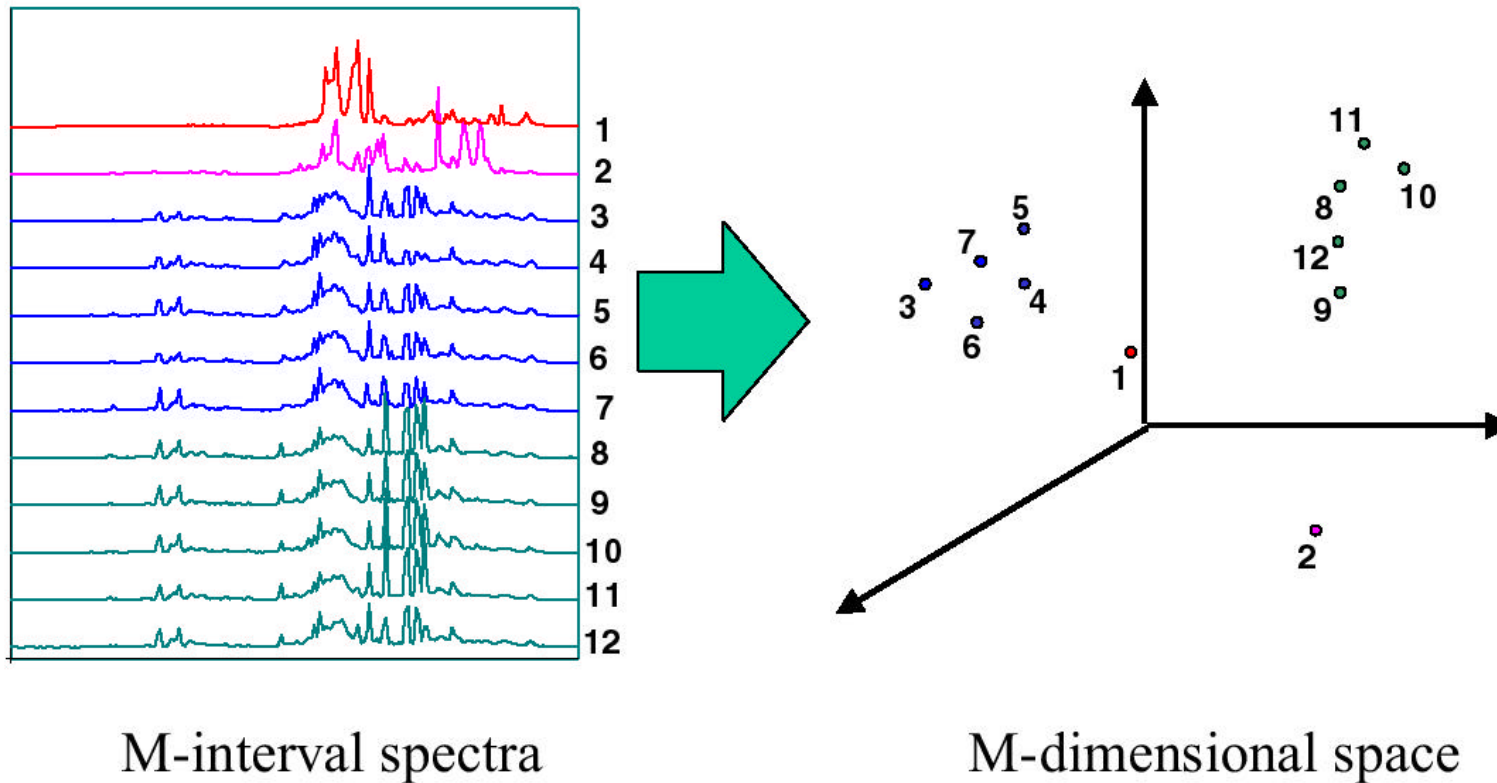
- Typical urine spectrum
 - ◆ high information content of low molecular weight species
 - ◆ correlated regions
 - ◆ high dimensionality
- Spectral information ➡ metabolic profile
 - ◆ Profile changes in response to toxic or disease processes
 - ◆ Following the pattern of changes gives insight into the site or mechanism of the process.
- Spectrum variables ➡ metabonomic space
 - ◆ Regions of this space will correspond to particular metabolic profiles
 - ◆ Structure of the space can be investigated using PR to see clustering / partitioning of the data into these different regions.
 - ◆ Unknown samples falling in different regions may be classified by PR methods.

Spectral pre-processing

- Data reduction (binning)
 - Natural resolution of NMR spectrum
 - Account for pH shifts
- Normalisation (total area = constant)
 - Account for differences in concentration
 - Only relative changes are detected
- Mean centre & scaling
- Exclusion of regions
 - eg poor solvent suppression



Representation of spectra in multidimensional space



Pattern recognition (PR)

- Purposes
 - ◆ Extract information from high dimensional data.
 - ✦ Take account of correlated variables.
 - ✦ Data can be interpreted more easily.
 - ◆ Uncover inherent patterns - differences and similarities in the data.
 - ◆ Find features which give rise to these patterns.
 - ◆ Build models capable of predicting class of unknown samples.
- Supervised versus unsupervised techniques
 - ◆ Unsupervised
 - ✦ Visualisation of relationship between different spectra in metabonomic space.
 - ✦ Methods choose their own classes based on inherent structure of the data.
 - ◆ Supervised
 - ✦ ‘training’ and ‘test’ data - usually split so that no animal in both.
 - ✦ Training data: toxic classes known. Test data: toxic classes not known.
 - ✦ Use model built with training data to classify the test spectra according to toxicity.

Pattern Recognition Methods

- Unsupervised
 - ◆ Principle Components Analysis (PCA)
 - ◆ Hierarchical cluster analysis (HCA)
 - ◆ Non-linear maps
 - ◆ Kohonen networks
 - ◆ Rule induction
 - ◆ Probabilistic methods (eg Autoclass).
- Supervised
 - ◆ Discriminant analysis
 - ◆ Neural Networks (Back propagation)
 - ◆ SIMCA*
 - ◆ K-Nearest Neighbour (KNN)
 - ◆ Probabilistic methods (eg Multi-dimensional Gaussian Class Modelling*)
 - ◆ Rule induction
 - ◆ Regression techniques: MLR, PLS, PCR.

PR in screening: questions

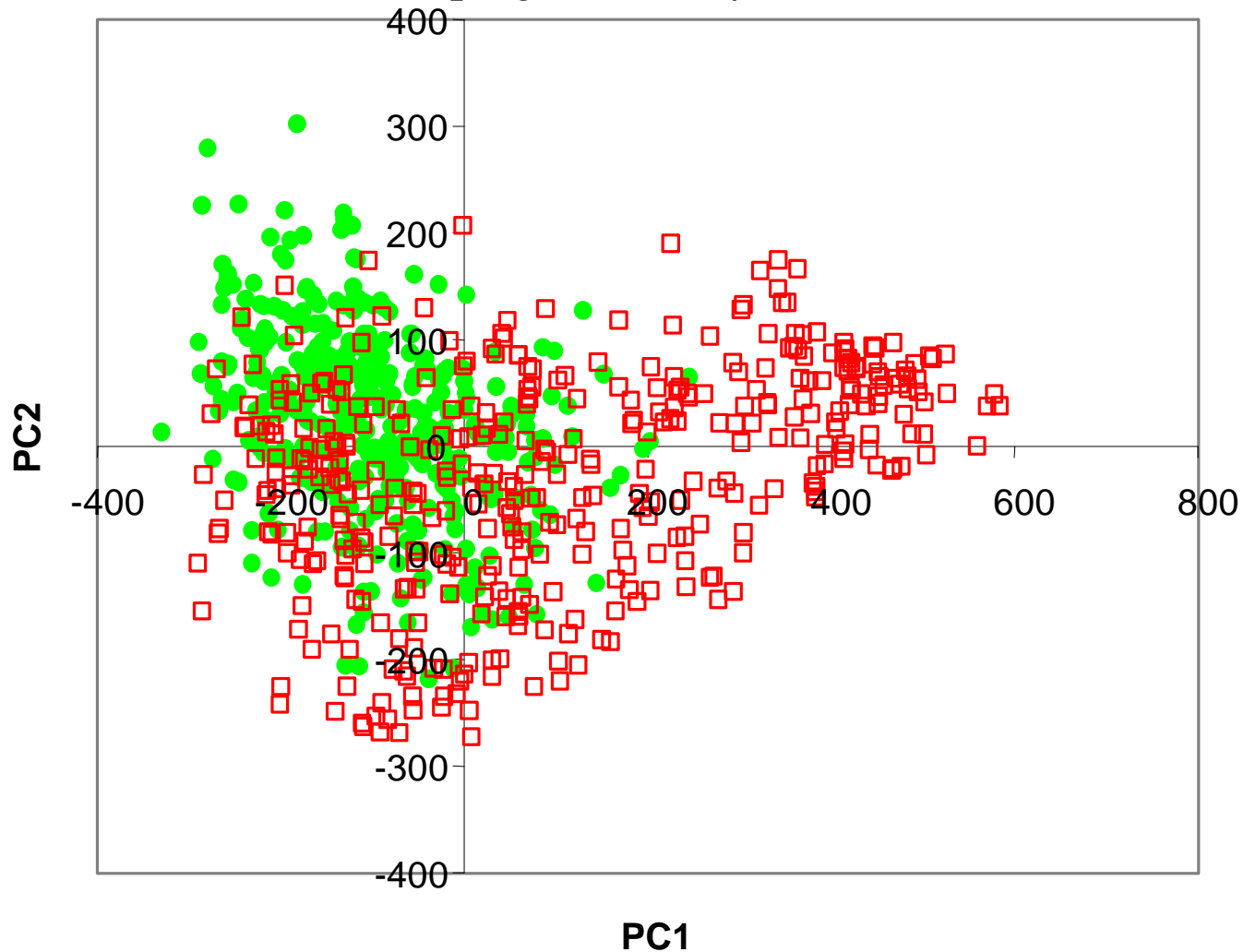
- **Outliers**
 - ◆ Does the sample fit into a predefined class within a database?
 - ◆ If not, possible causes: unmodelled mechanism of toxicity, disease, genetics, diet, physiological stress, bacterial contamination.

- **Classification**
 - ◆ Which class of toxicity / disease within the database does the sample fit?
 - ◆ eg organ and mechanism of toxicity.
 - ◆ Larger data sets ➡ quantitative, time-resolved, eg predict equivalent dose level.

- **Identification of biomarkers**
 - ◆ Identify spectral regions most important in making the classification.
 - ◆ Molecules indicative of particular type of toxicity or disease.
 - ◆ Information on mechanism of toxicity or disease.

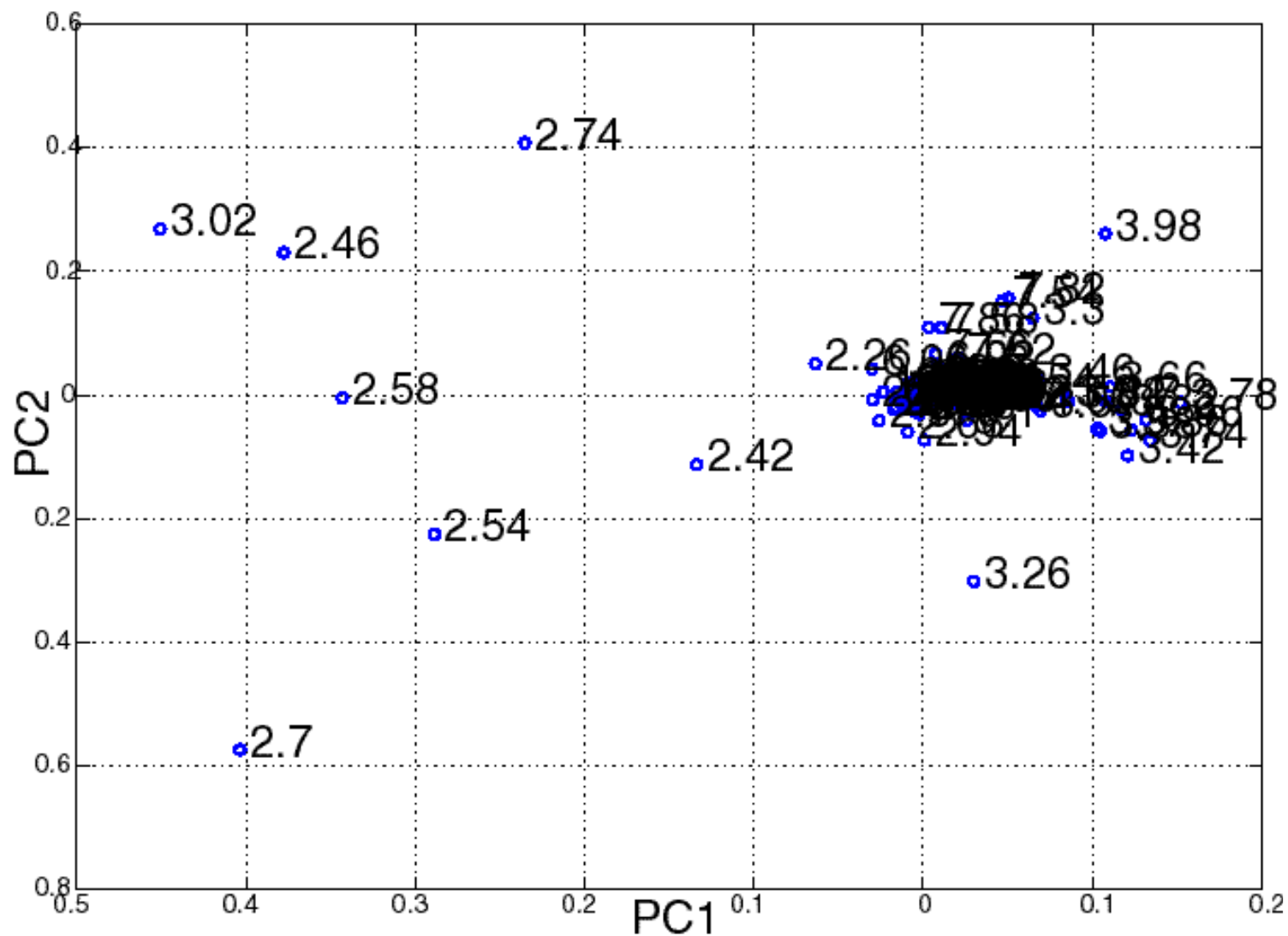
PCA: Scores

- Scores plot showing different regions mapped by 2 strains of control rats: Han-Wistar (■) and Sprague-Dawley (○).



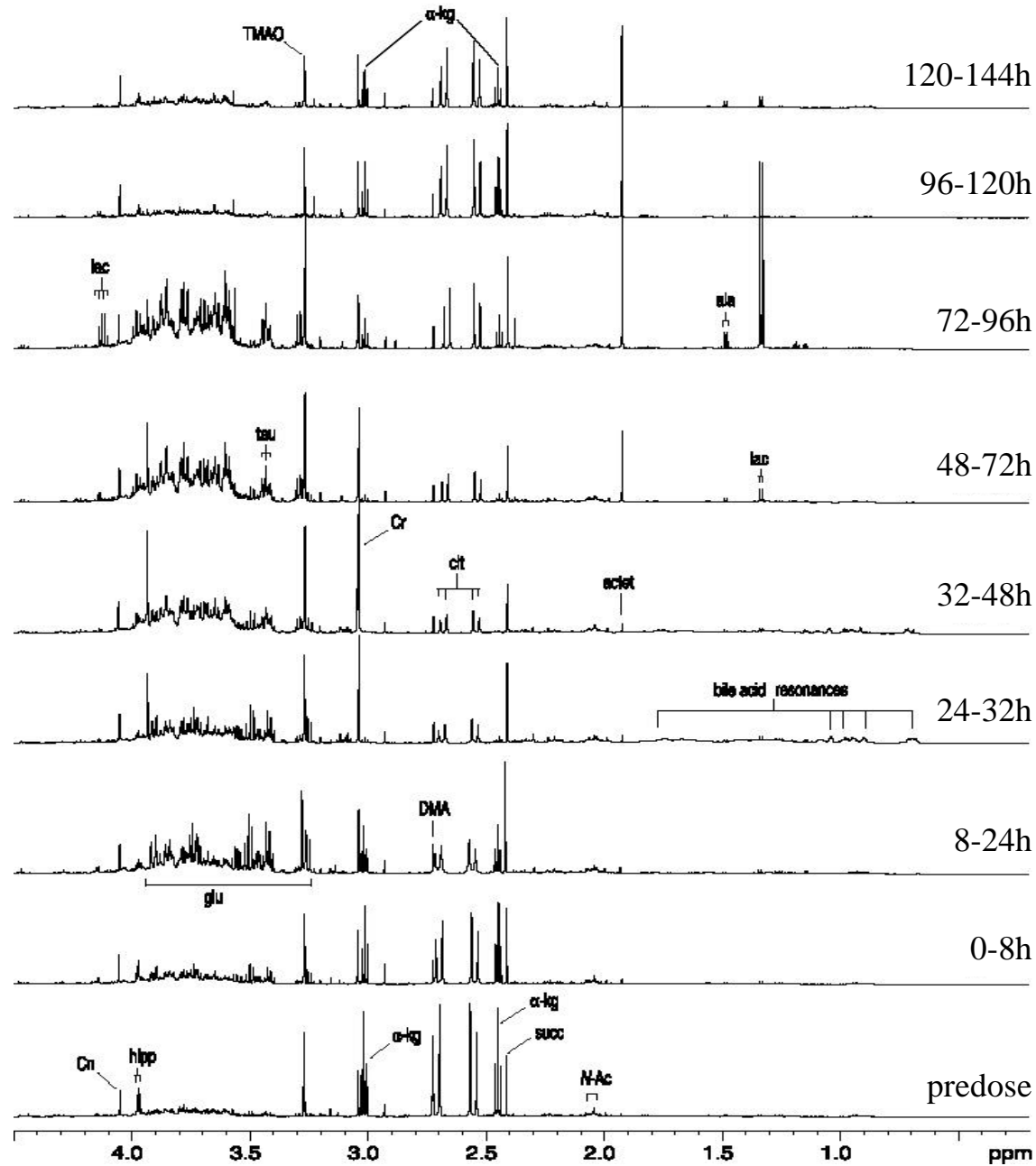
PCA: Loadings

- Identify regions causing separation between classes.



Time profiles

600 MHz ^1H NMR spectra of urine from rats treated with ANIT (@ 200 mg/kg).

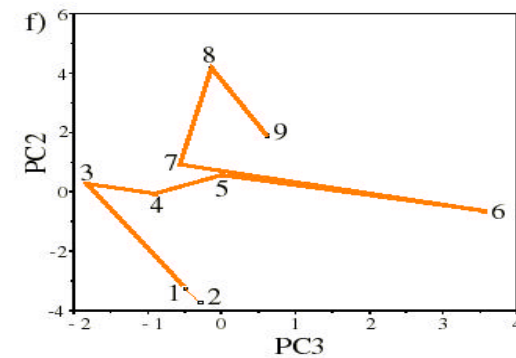
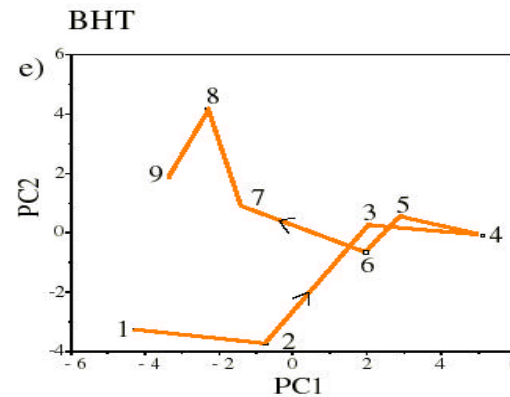
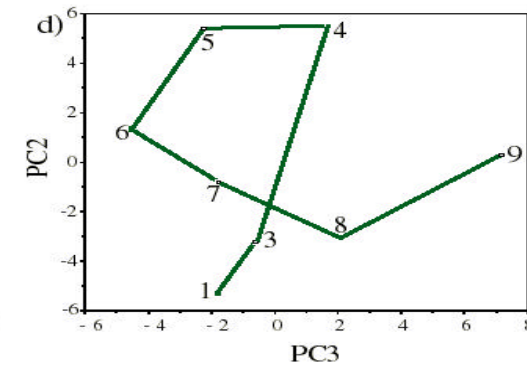
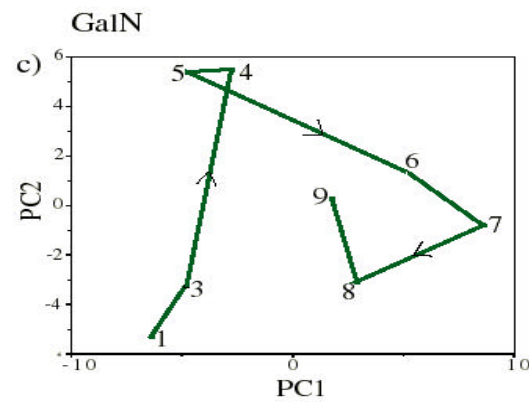
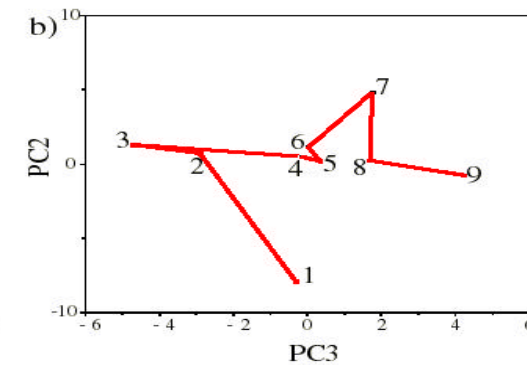
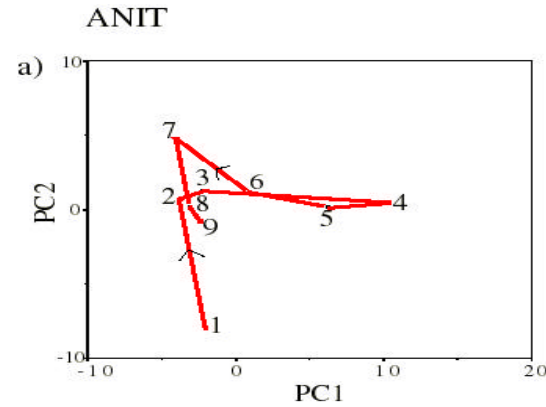


Time

PCA: Metabolic trajectories

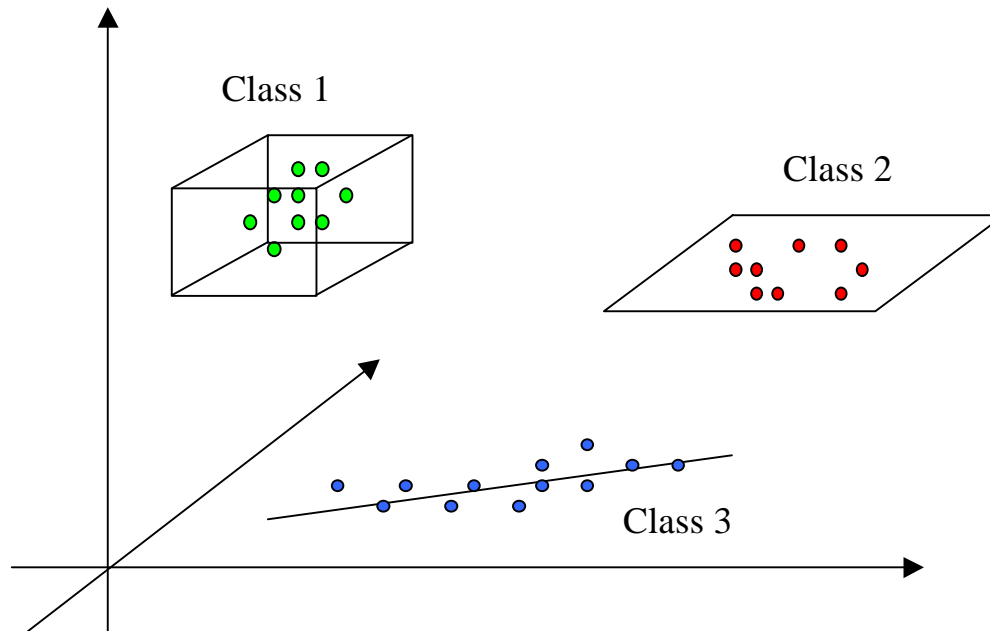
- Trajectory over time, showing changes in metabolic profile.

- Rats treated with 3 different model hepatotoxins.



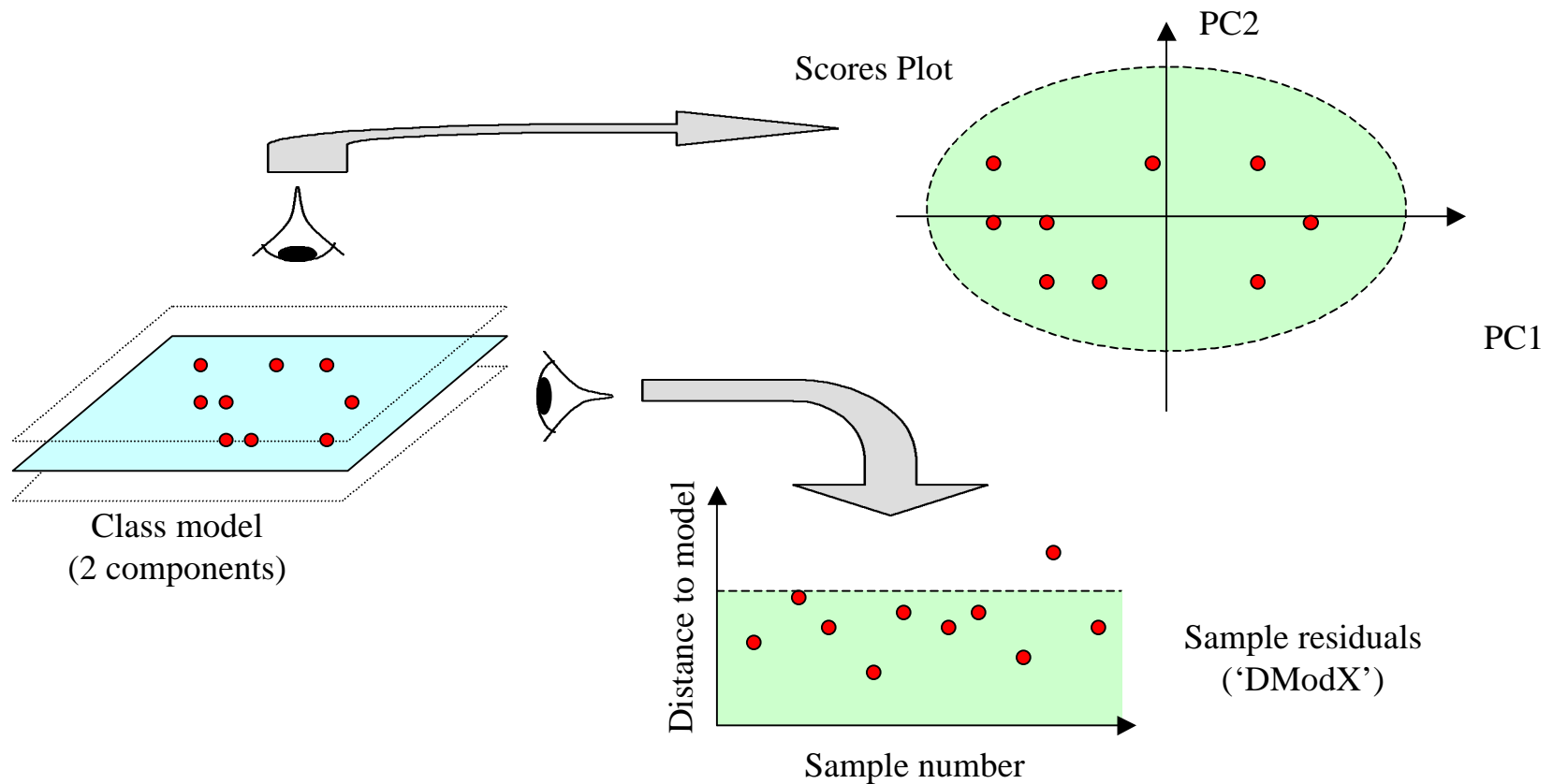
SIMCA

- SIMCA: Soft Independent Modelling of Class Analogy.
- Construct a separate PCA model for each class.
- Define hyper-volumes enclosing each class.

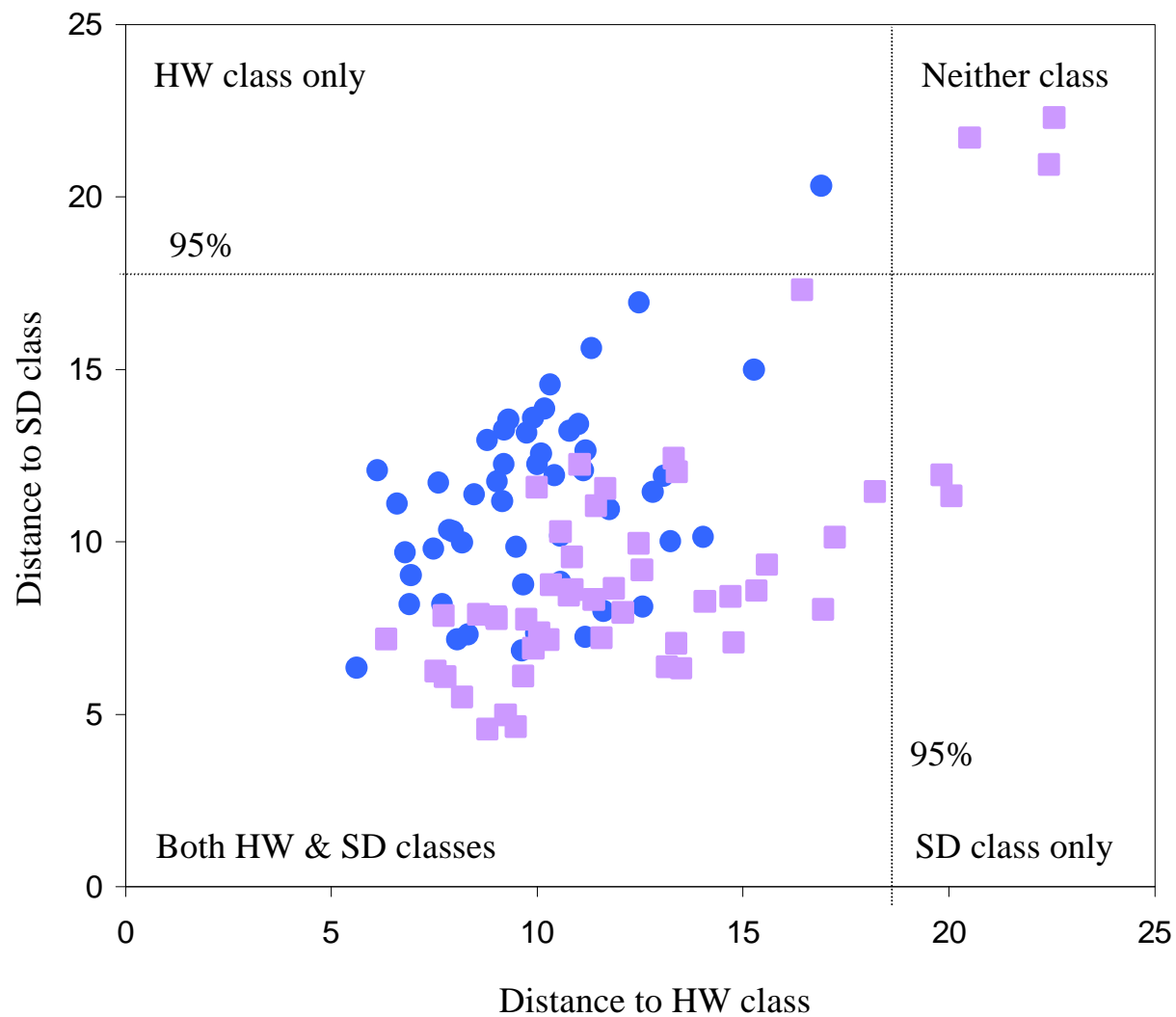


SIMCA: Class boundaries

- Define boundaries in terms of standard deviation of points from the model.



SIMCA: Coomans Plot

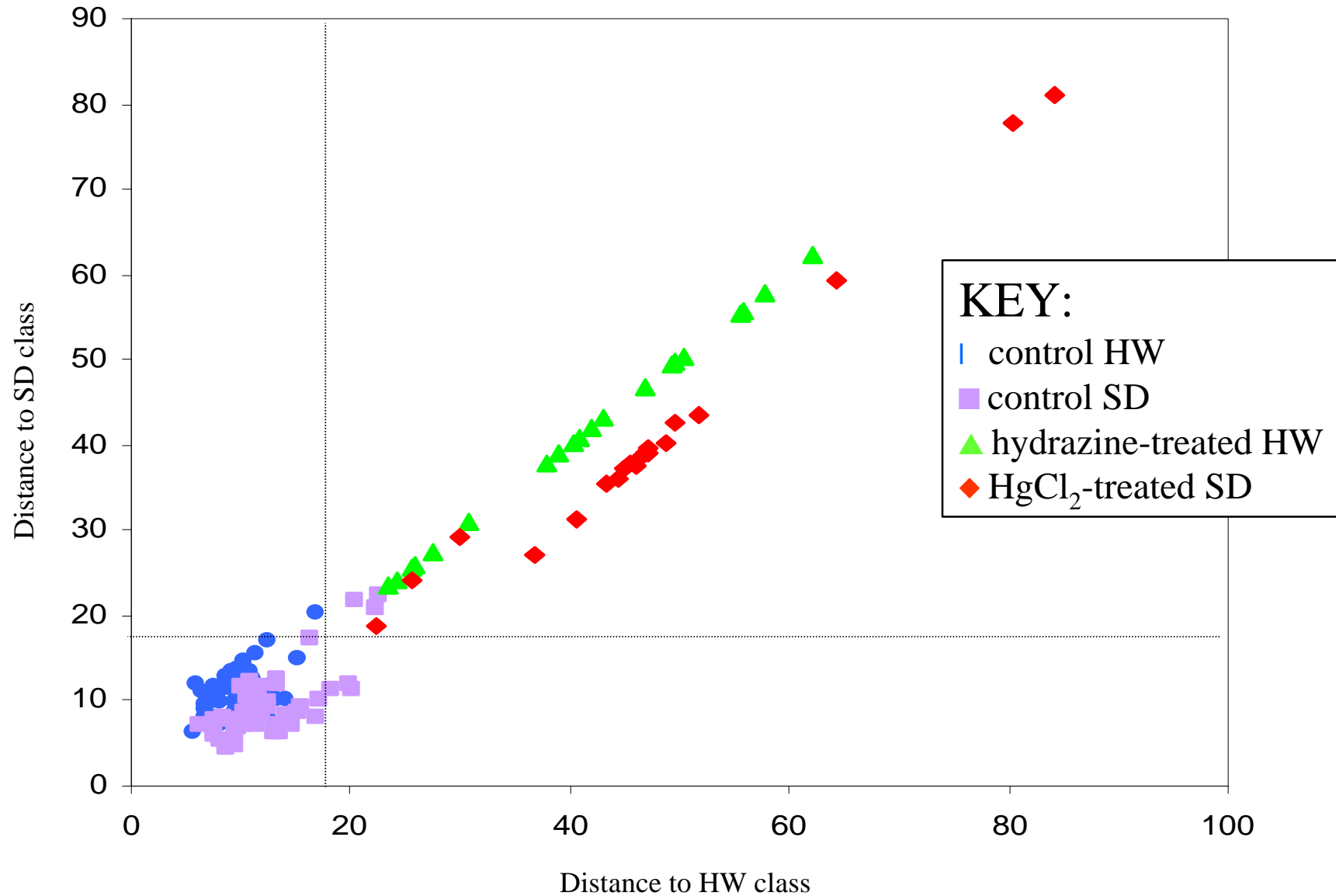


- Shows classification of control urine samples from HW (●) and SD (■) rats.

- Most urine samples map within the class boundaries for both strains.

- 3 samples (SD) not recognised as control

SIMCA: Coomans plot with toxin treated spectra.



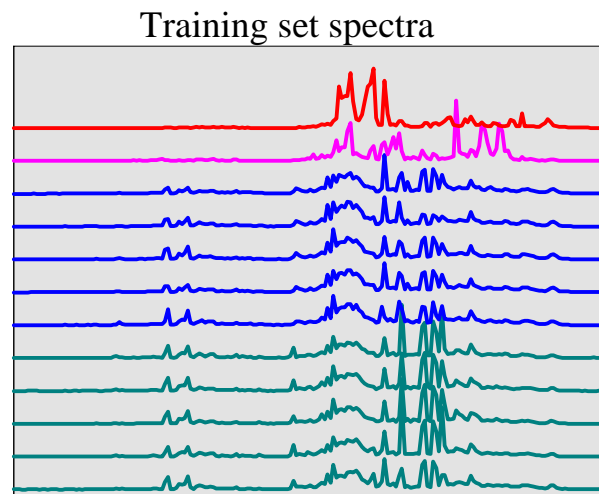
Multi-dimensional Gaussian Class Modelling (MGCM)

- Probability distribution for each class - ‘probability cloud’
- Each point in the space (ie each spectrum) has a certain probability p_i of belonging to class i .
- Build each cloud from summing many small clouds, each defined by a spectrum in the training set.

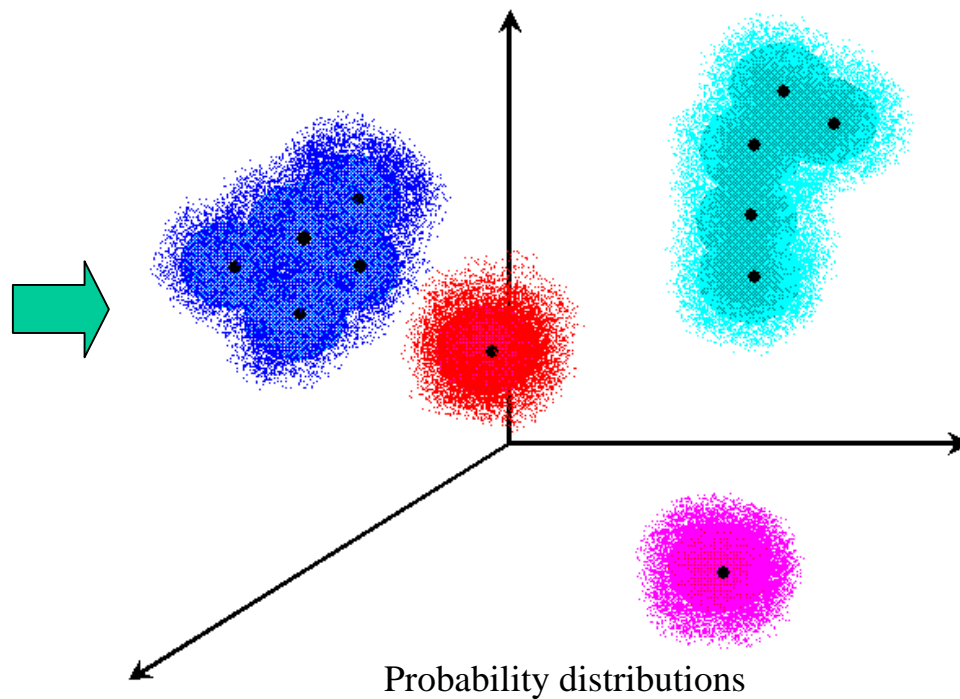
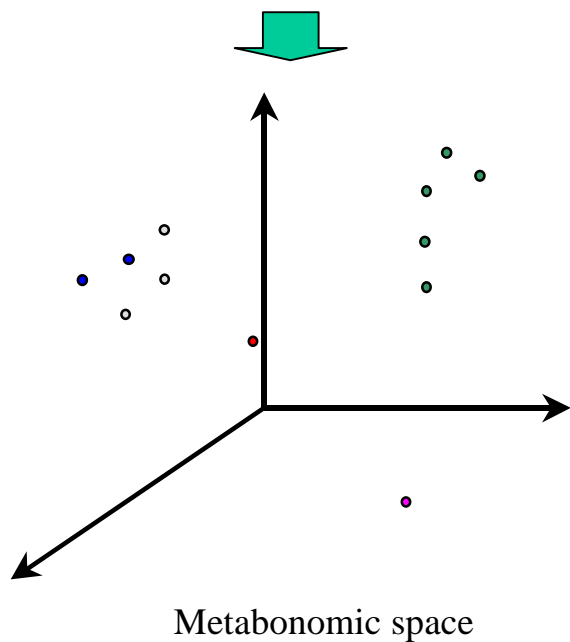
- Unknown samples: p_i 's give the probability of belonging to each class. Classification can then be assigned to the class with the highest probability (or more complex criteria).

- Implementation: probabilistic neural network.

MGCM: Probability distributions



Distributions constructed from Gaussians centred on training set data points.



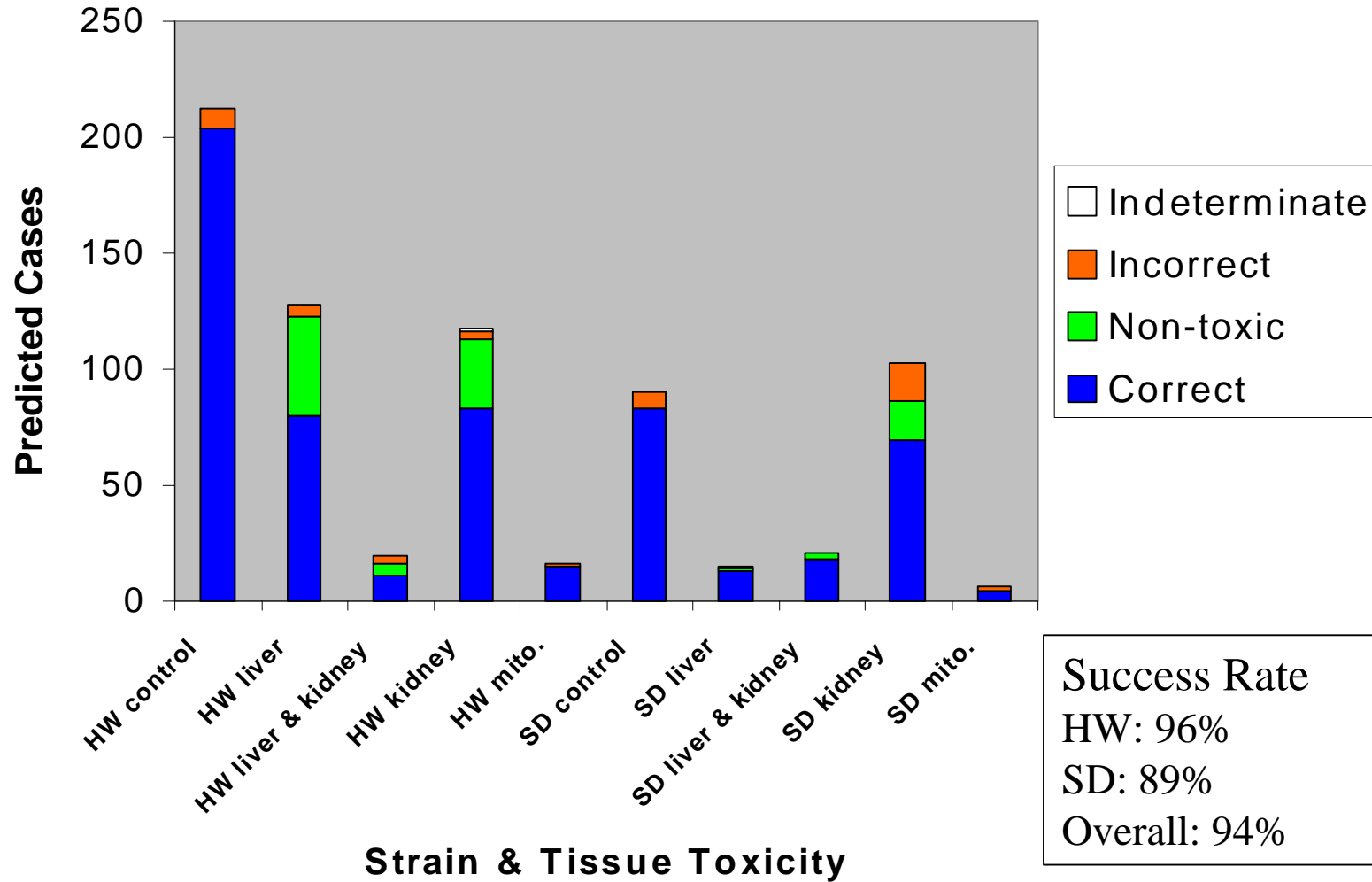
Features of MGCM

- Classification is ‘fuzzy’ - although one best class can be chosen, all the p_i 's give information on how well the sample is classified.
- Samples may be member of several classes, eg compounds exhibiting more than one type of toxicity.
- Classification ‘unknown’ if all p values small - this detects outliers in the test data. Perhaps a new class?
- Disjoint and embedded classes can be modelled easily.
- Prior information easily included (Bayesian method).
 - ◆ Eg weighting against some classes or accounting for differences in the number of samples in each class in the training set.
- Sensitivity of classification to any particular input variable may be calculated - like PCA loadings.
- No outlier detection in the training data.

Example - Toxicity Data

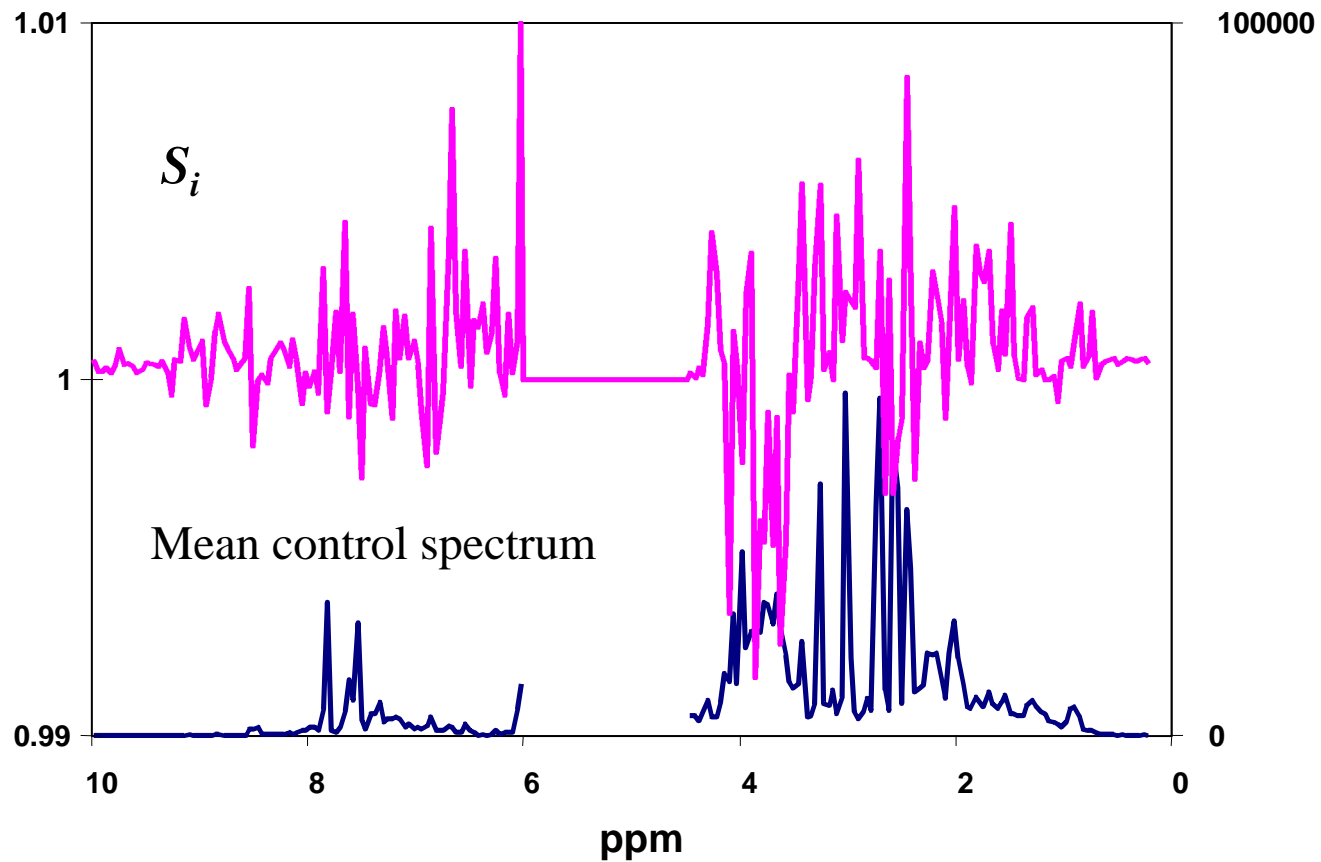
- Data set: 11 Toxins
- 2 rat strains (Han-Wistar & Sprague-Dawley)
- 1310 urine spectra.
- 550 control samples, 760 dosed samples.
- Training set: 583, test set: 727. No animal in both sets.
- Range scaling on each variable.

Classification of Validation Set by Strain & Tissue Toxicity



MGCM: Sensitivity ratio

$$S_i = \frac{\text{success rate with all variables}}{\text{success rate with } i\text{'th variable left out}}$$



Summary

- NMR spectra of biofluids provide a high degree of information on the metabolic state of an organism
- PR methods can extract this information reliably for visualisation, and building predictive models.
- NMR and PR together can be used to model toxic / disease processes, predict class of unknown samples and identify biomarkers.
- SIMCA
 - ◆ Build a PCA model - a hyper volume for each class
 - ◆ Detect outliers, predict unknown samples & identify biomarkers.
- Multi-dimensional Gaussian Class Modelling
 - ◆ Build probability distributions from training set.
 - ◆ Fuzzy classification - gives most likely class for unknown.
 - ◆ Takes account of full dimensionality of data.