

NIR Spectroscopy and Chemometrics in Food Analysis

Tom Fearn
UCL

Background

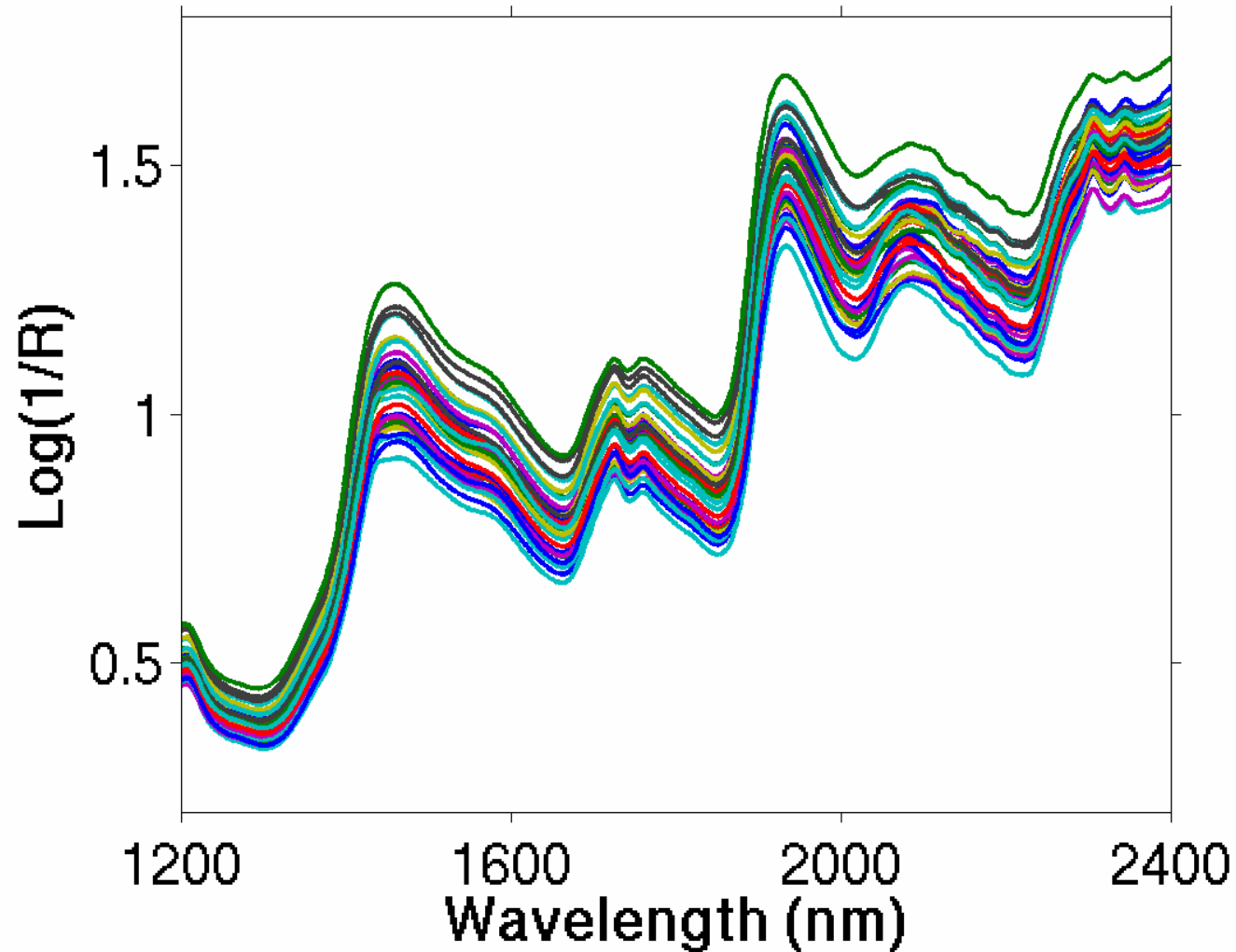
- Quantitative NIR spectroscopy and chemometrics have been closely related throughout the development of both, to their mutual benefit.
- Food analysis has always been an important application area for NIR
- I will discuss two examples
 - measuring the composition of biscuit doughs
 - classifying meat species

The first is quantitative the second qualitative

Example 1: Biscuit dough experiment

- Can we measure the composition of biscuit doughs using NIR spectroscopy?
- Experiment
 - 40 doughs with varying (and ‘known’) fat, flour, sugar, and water contents
 - measure NIR spectra of dough pieces
 - fit a prediction equation
 - test on 32 further doughs (test set)

Spectra of 40 biscuit doughs



Spectra with mean subtracted



The calibration problem

Want an equation to predict y from x with $p = 601$ variables in x and $n = 40$ training samples available

Some issues

- Linearity $\log(1/R)$
- Which way round? Take NIR as x
- Joint or separate? Both
- Pretreatment No
- Compositional data Ignore

Some solutions

- Select a few wavelengths
- Regression on factors
 - Principal components regression (PCR)
 - Partial least squares regression (PLSR)
- and many more . . .

Wavelength selection

- Originally the standard method
- Largely replaced by PCR/PLSR
- Back in fashion using stochastic search methods: genetic algorithms, . . .
- Useful if we want to build a cheap instrument
- Forward selection can perform badly

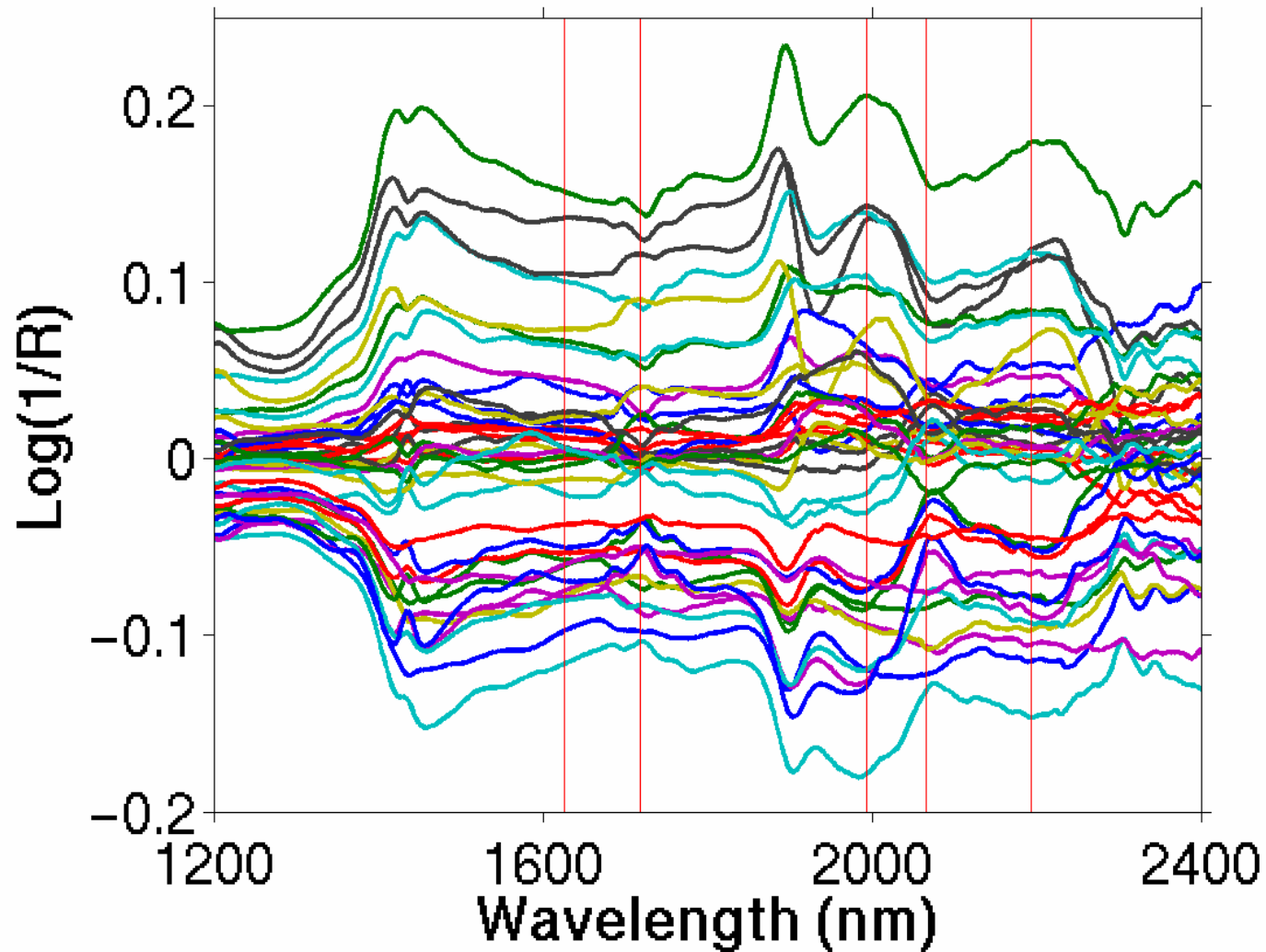
Some prediction results

	#w/l	RMSEP	BIAS	SEP
Fat	2 ⁺	0.29	-0.20	0.21
	5 [*]	0.26	-0.01	0.27
Sugar	3	1.62	1.13	1.18
	5 [*]	0.75	0.10	0.75
Flour	4	0.96	0.46	0.85
	5 [*]	0.69	-0.39	0.58
Water	3	0.68	0.51	0.46
	5 [*]	0.43	0.31	0.30

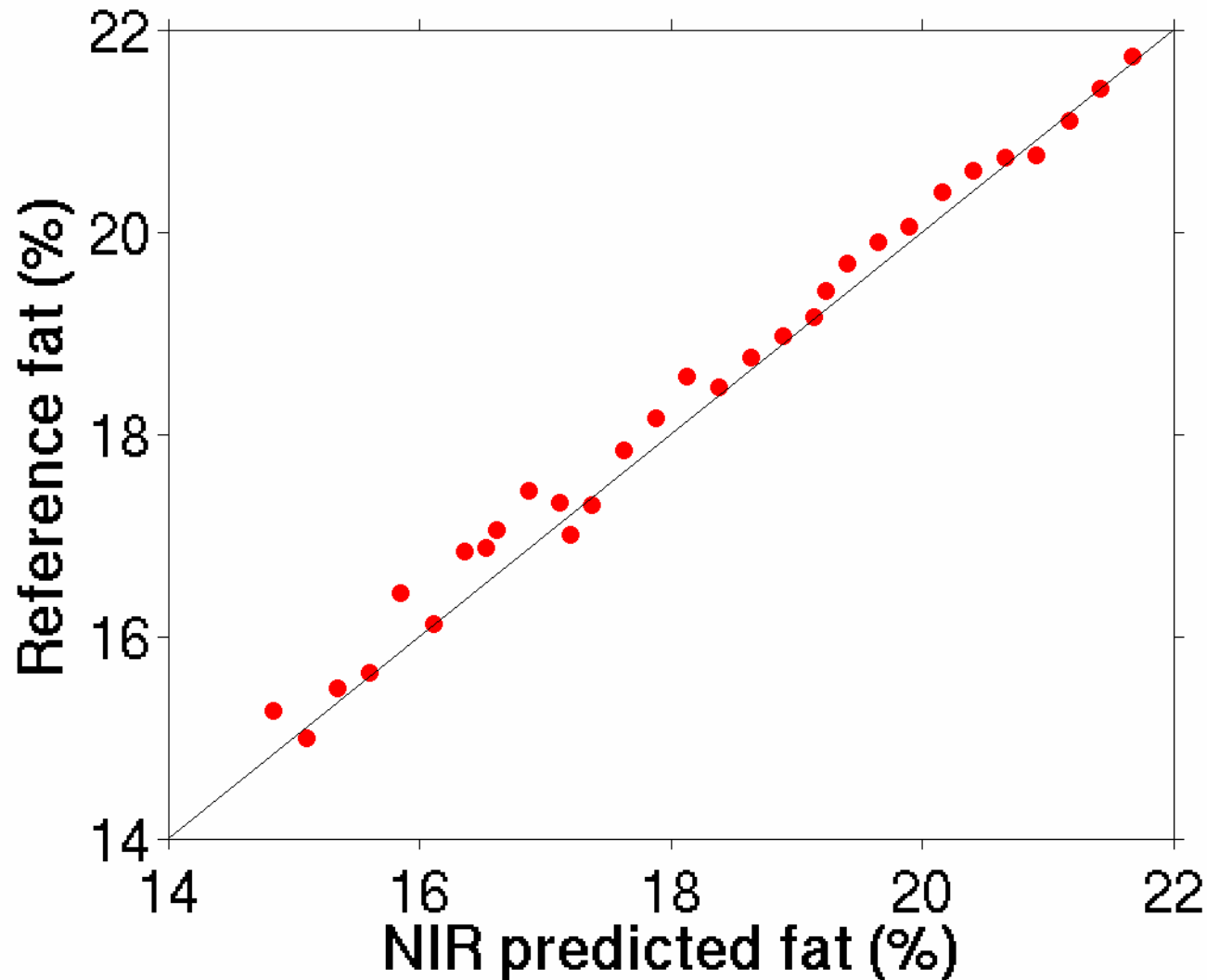
+ 1550, 1734 nm

* 1626, 1718, 1994, 2066, 2194 nm

Five selected wavelengths



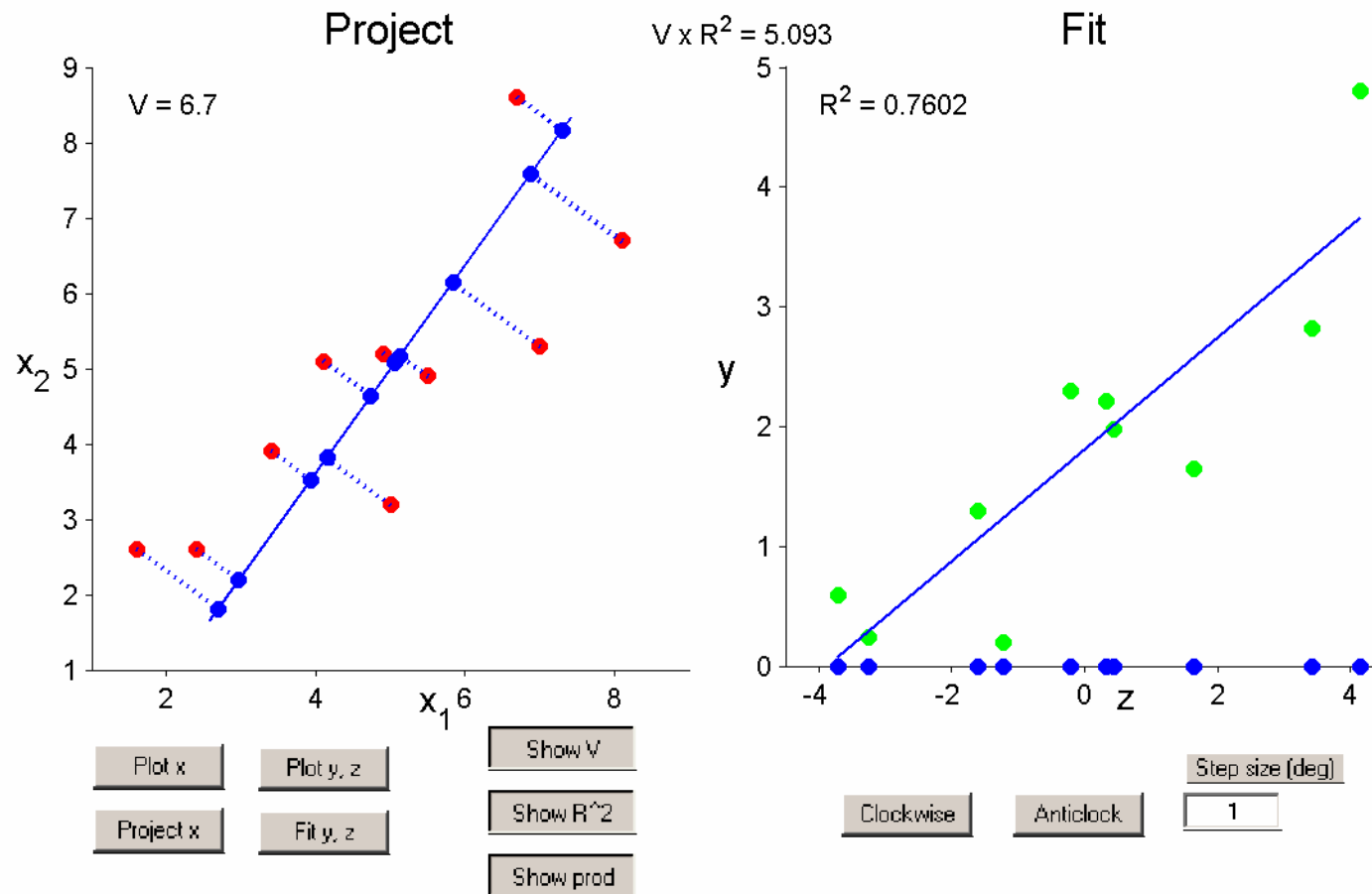
Predictions for fat, 2 wavelengths



Regression on factors

- PCR (UK) and PLSR (Scandinavia) first used in this context in early 1980s
- Both work by constructing new variables as linear combinations of the spectral data and regressing reference data on these new variables

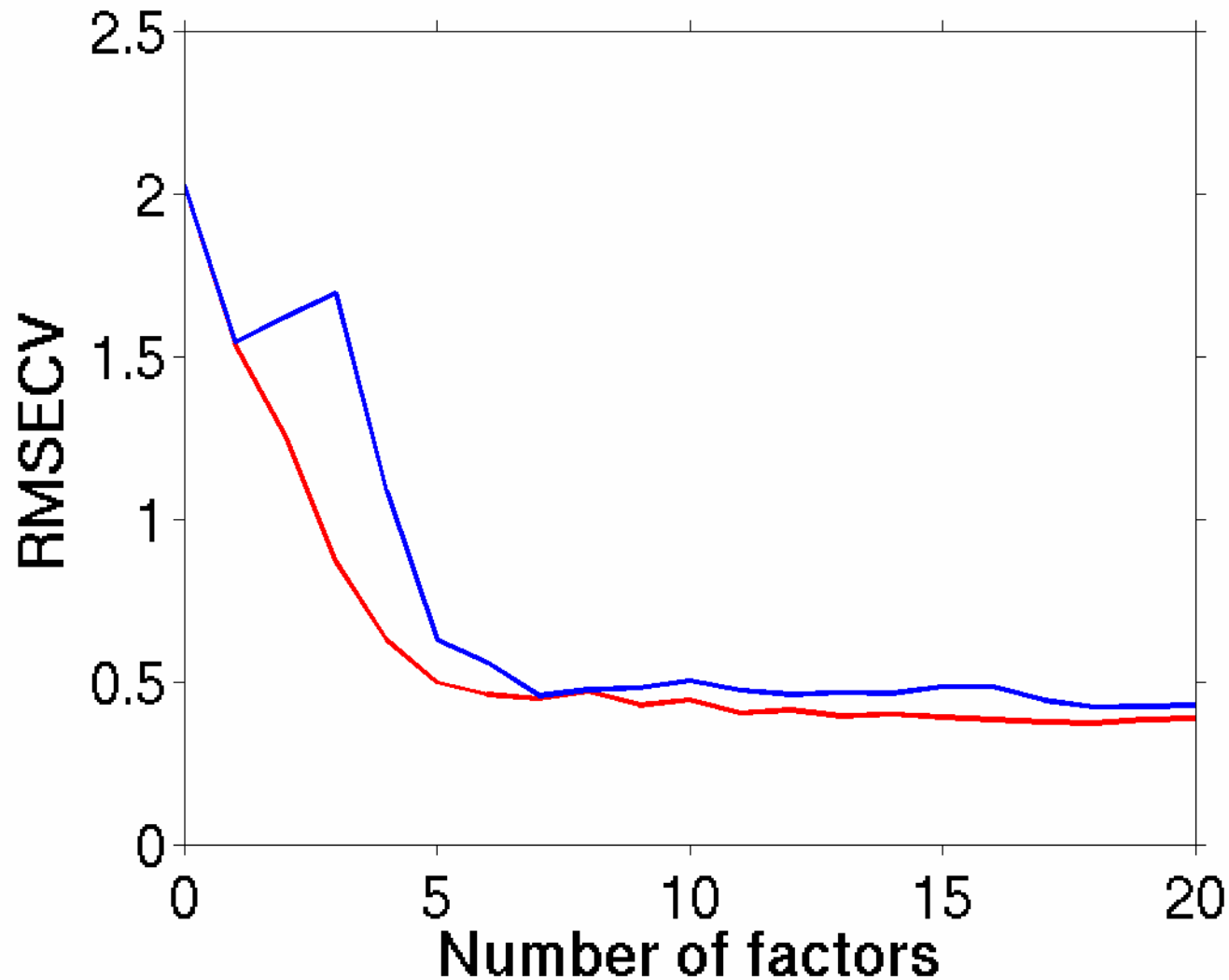
PCR and PLSR



PCR and PLSR

- In practice construct several factors and use multiple linear regression on the scores
- Use cross-validation to decide on number of factors

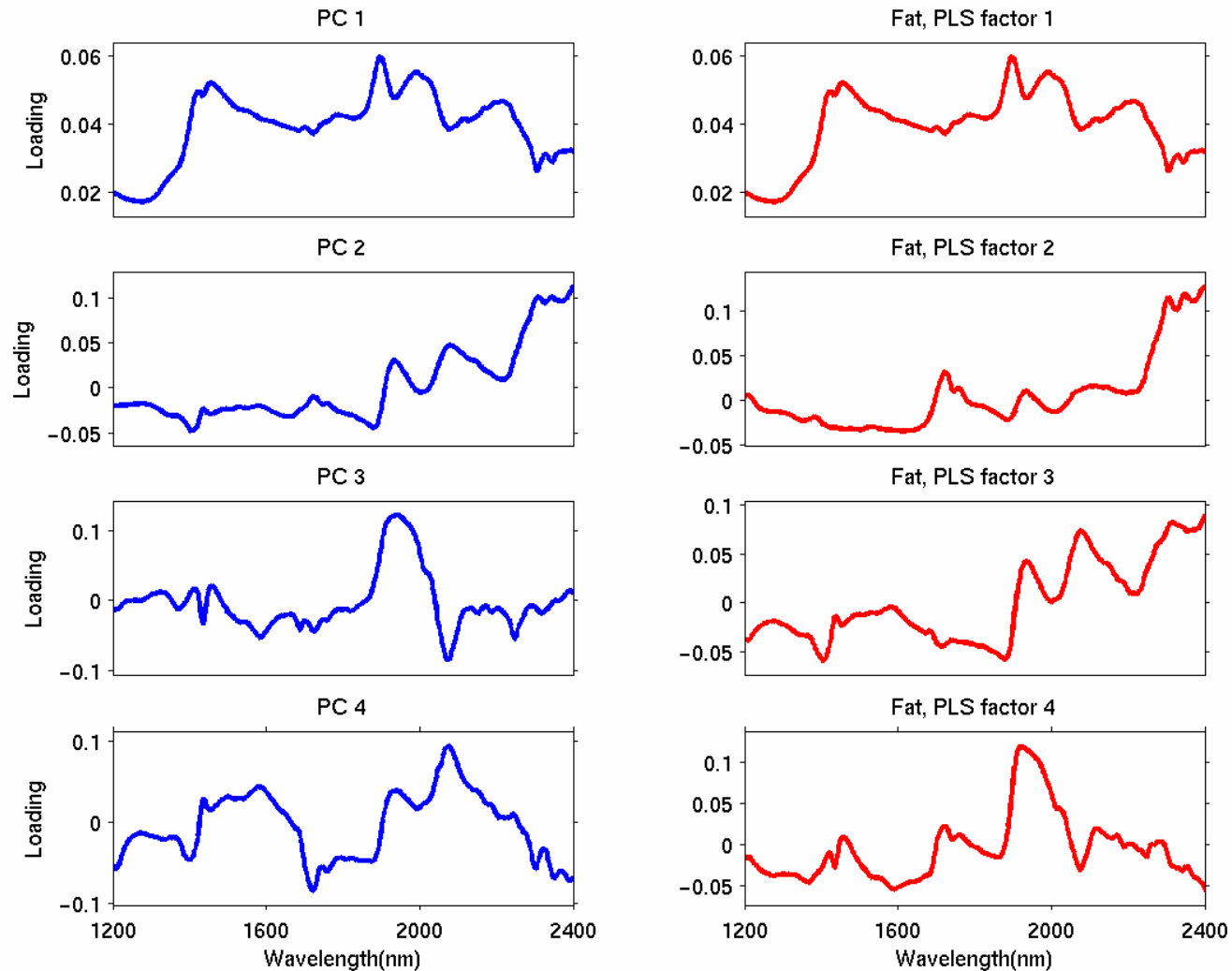
CV for fat, PCR and PLSR



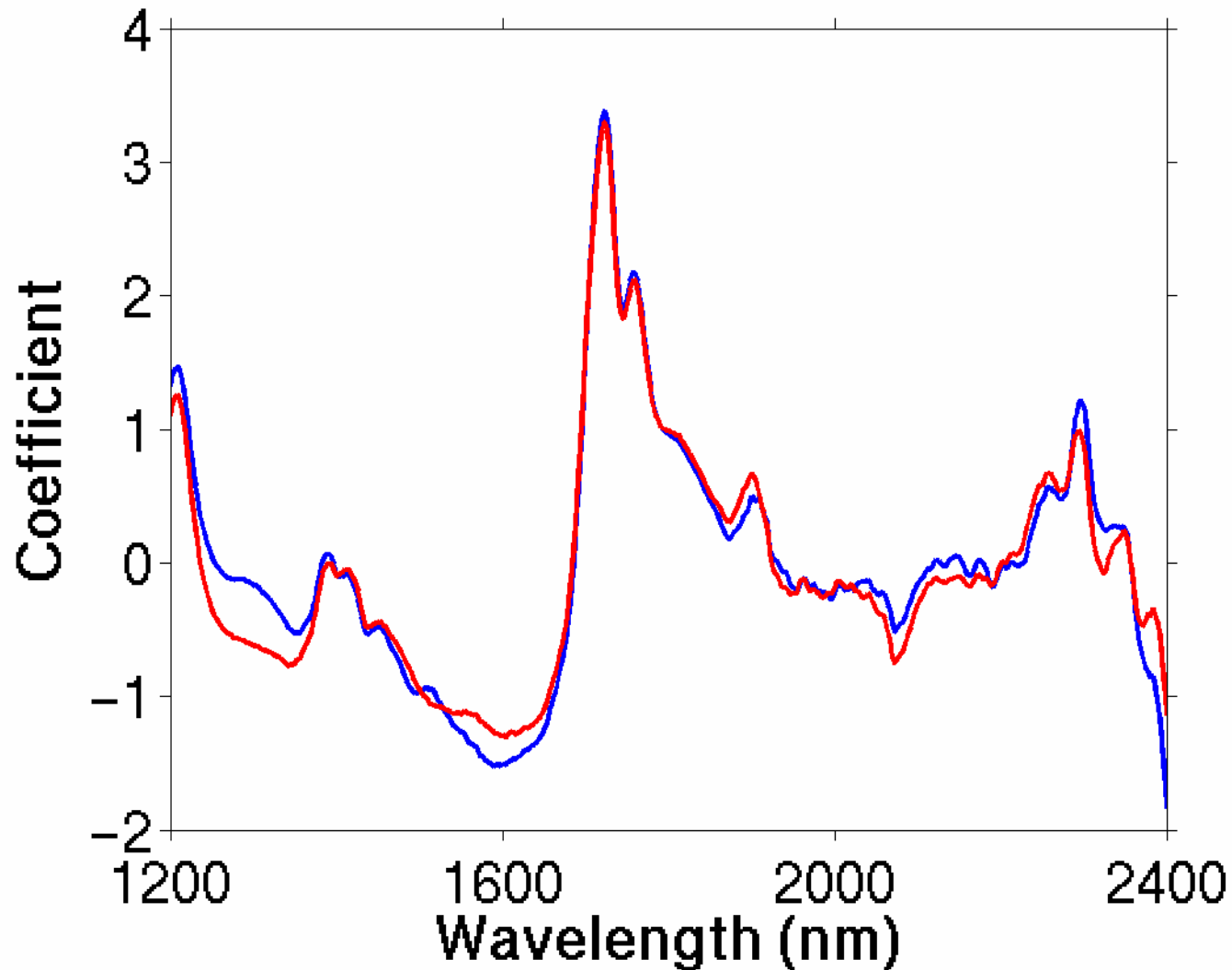
Plots of loadings, coefficients

- One of the nice things about PCR and PLSR is that you can plot things that look like spectra
- Chemometrics software is generally pretty good at this
- The plots help the spectroscopist to understand what is going on

Loadings for PCR and PLSR for fat



Coefficients for PCR(7) and PLSR(5) for fat



Some prediction results

		RMSEP	BIAS	SEP
Fat	PCR ⁺	0.35	-0.22	0.28
	PLS*	0.41	-0.33	0.24
Sugar	PCR	0.82	0.25	0.79
	PLS	0.74	0.13	0.73
Flour	PCR	0.66	-0.12	0.65
	PLS	0.60	-0.04	0.60
Water	PCR	0.32	0.10	0.31
	PLS	0.33	0.15	0.29

+ PCR with 7 factors

* PLS with 5 factors

Which is best?

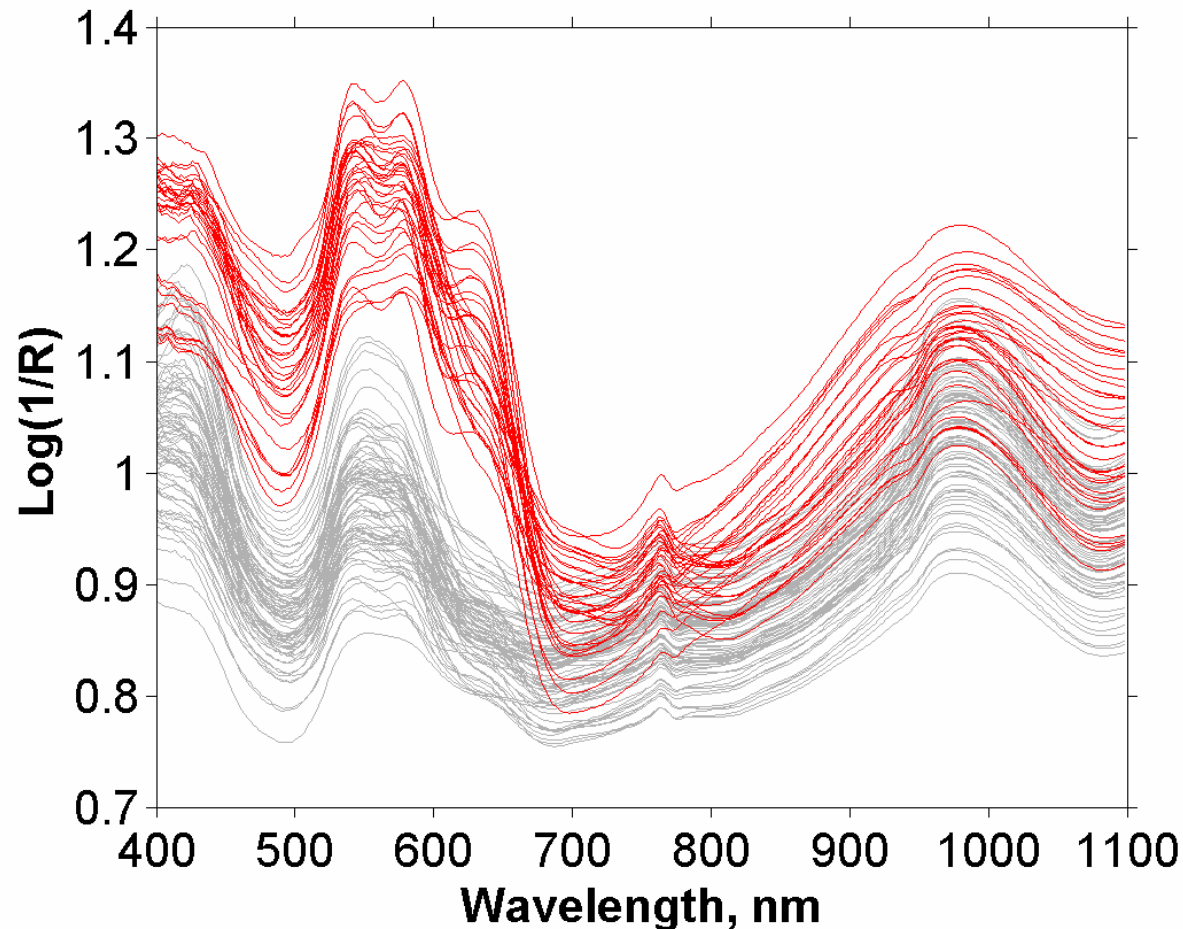
- No approach is a clear winner
- Very little to choose between PCR/PLSR
- Wavelength selection better for
 - simple equation
 - building a cheap instrument
- Factor methods
 - possibly easier to use
 - may give better diagnostics

Example 2: Classifying meat species

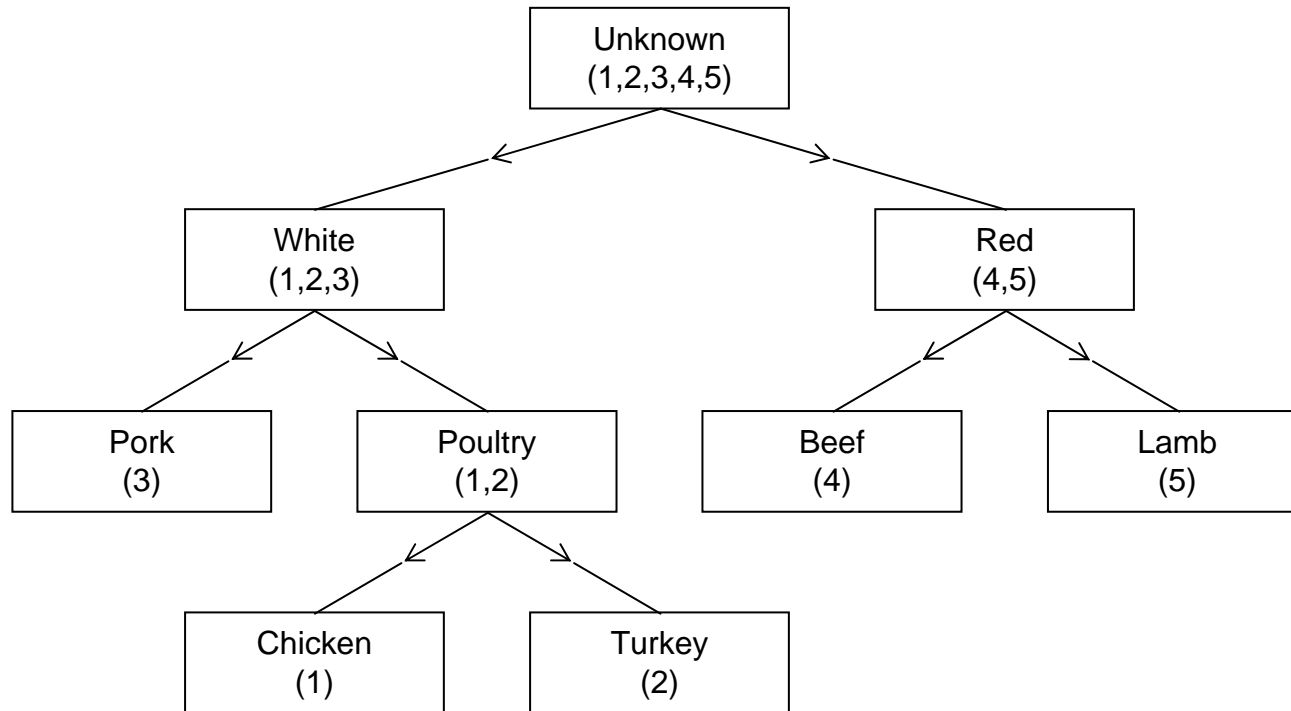
- Can we discriminate between meat species using visible/NIR spectroscopy?
- Experiment
 - 115 samples of homogenised chicken, turkey, pork, beef and lamb,
 - measure spectra in the range 400-1100nm
 - derive a classification rule
 - test on a further 115 samples

Spectra of training set ($n = 115$)

Grey = chicken, turkey, pork; Red = beef, lamb



Hierarchical structure of five-group problem

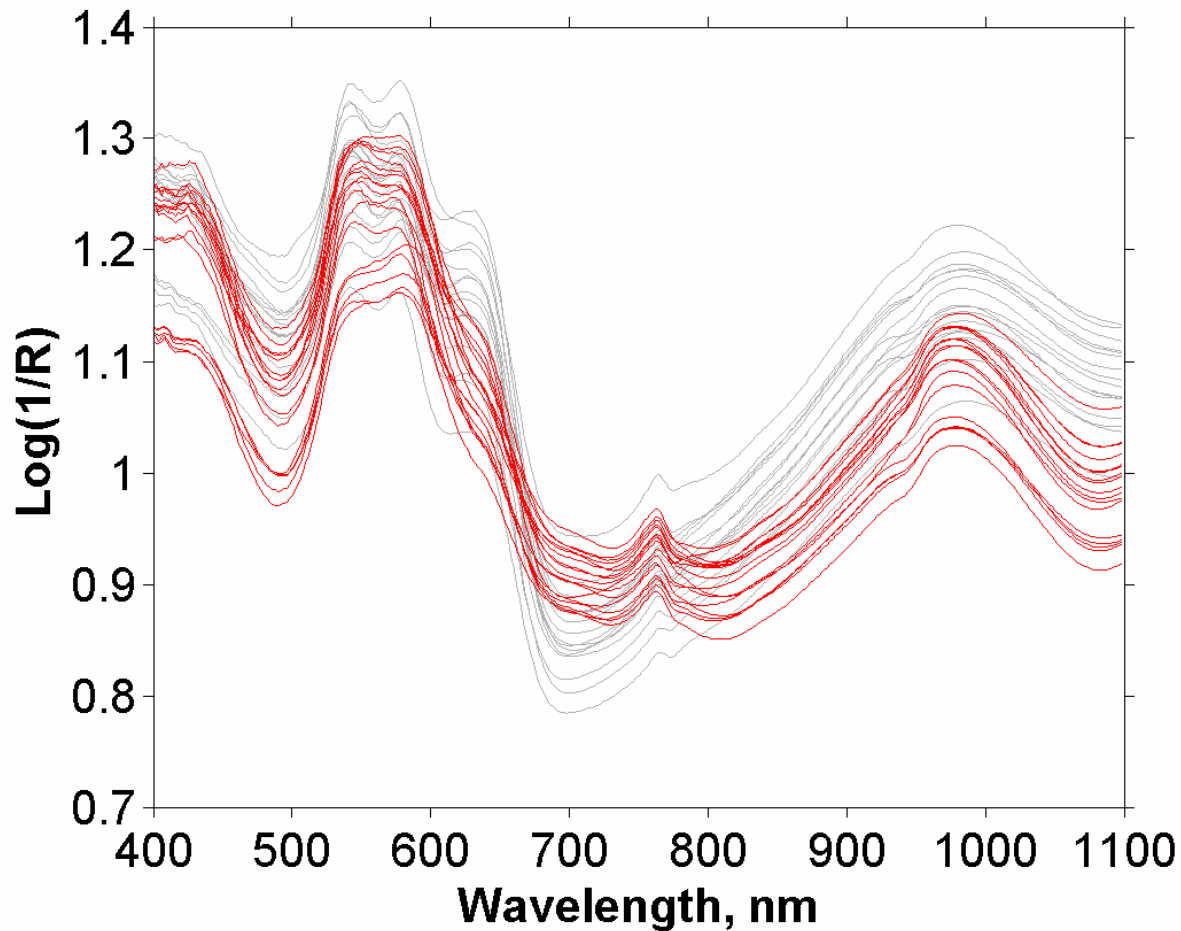


Methods for deriving classifiers (1)

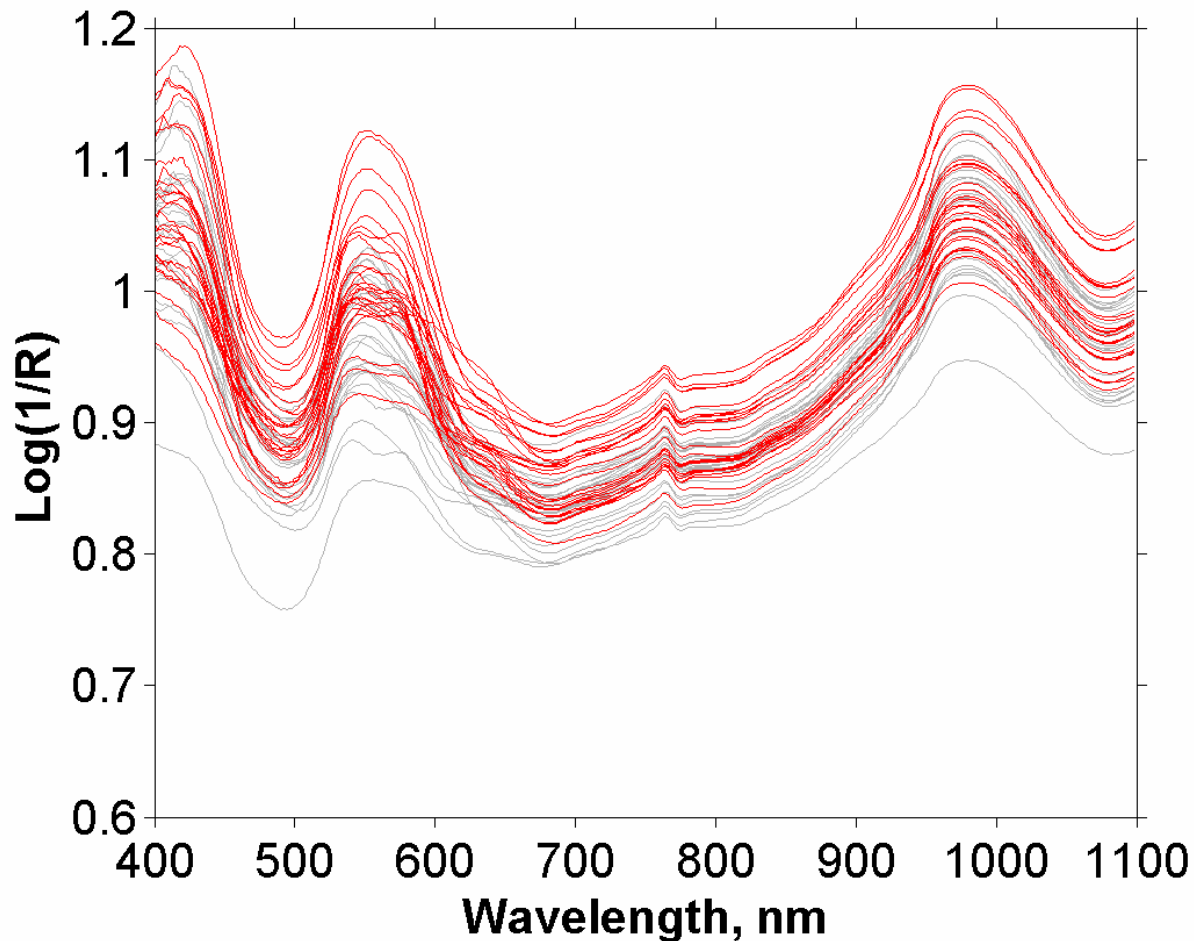
PCDA (Principal Component Discriminant Analysis)

- PCA on spectra for both groups together
- Linear discriminant analysis (LDA) on scores
 - equivalent to regression on group indicator variable
- Result is a discriminant score that is a linear function of the spectral data
- Gives 100% correct results (training *and* test sets) on all splits except chicken v turkey

Spectra of beef (grey) and lamb (red) in training set



Spectra of chicken (grey) and turkey (red) in training set



Methods for deriving classifiers (2)

SIMCA (Soft Independent Modelling of Class Analogies)

- PCA on spectra for each group separately
- Decide on PCA ‘model’ for each group, ie choose the number of components
- To compare an unknown with a particular group, measure two distances
 - Q is the distance of the spectrum from its projection onto the PC model for that group
 - T is the distance from the group centre, using scores

SIMCA (continued)

- Compare an unknown X with each group, judging the distances against thresholds
- Possible conclusions (for two groups)
 - (1) X is consistent with group A only
 - (2) X is consistent with both groups
 - (3) X is consistent with neither group
- In cases 2 and 3 one might force a choice of the nearer group, combining Q and T eg as $\sqrt{(Q^2 + T^2)}$
- Forcing a choice, the results on the test set for chicken v turkey are 53 correct out of 55

SIMCA v PCDA/PLSDA?

- SIMCA is
 - easy to extend when new groups come along
 - attractive because it can say it is not sure
- but
 - perhaps rather harder to ‘tune’ than PCDA/PLSDA

A reference

A hierarchical discriminant analysis for species identification in raw meat by visible and near infrared spectroscopy, T. Arnalds, J. McElhinney, T. Fearn and G. Downey, *J. Near Infrared Spectrosc.* **12**, 183-188 (2004).