Tips & tricks, (or rather perils and pitfalls) in metabolomic data analysis

Tim Ebbels

Imperial College London

Outline

Model validation & assessing model performance

Why use PLS for metabolic profiling?

Outline

Model validation & assessing model performance

Why use PLS for metabolic profiling?

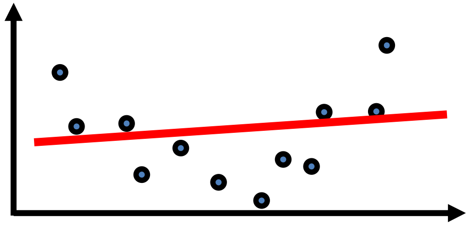
Model validation (1)

- Can we trust conclusions based on the model?
  - Statistical validation
  - Biological validation
- Statistical validation
  - Goodness of fit to data
  - Errors on model parameters
  - Goodness of *prediction* for new data
- Sometimes ignored, but a vital stage of modelling process

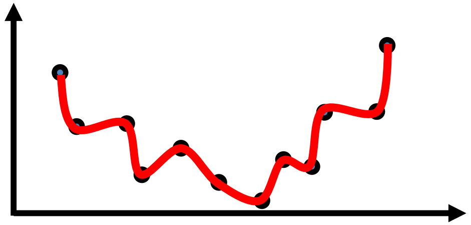
Overfitting

- A model should capture the important phenomena while ignoring random fluctuations (noise)
- Machine learning & multivariate stats: models are flexible & space is very large → prone to overfitting

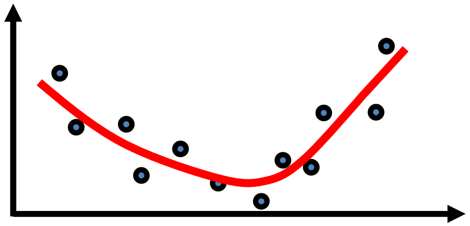
	Underfitting	Overfitting	Optimal fit
Fit to current data	✗	☑	☑
Fit to new data	✗	✗	☑
Model complexity	Too low	Too high	Just right



underfit

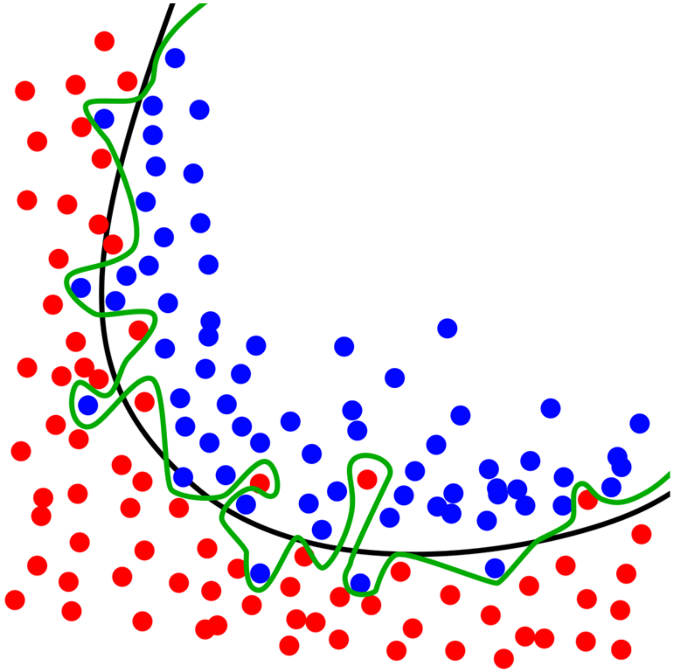


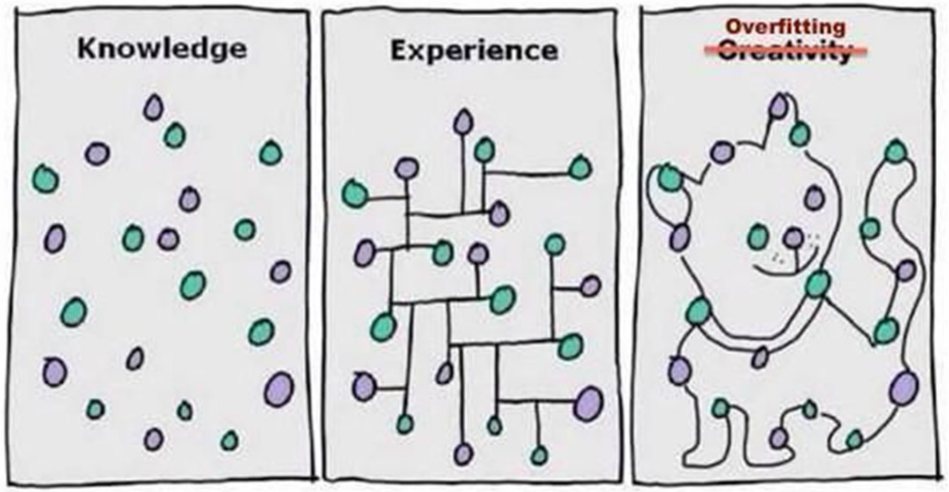
overfit



optimal

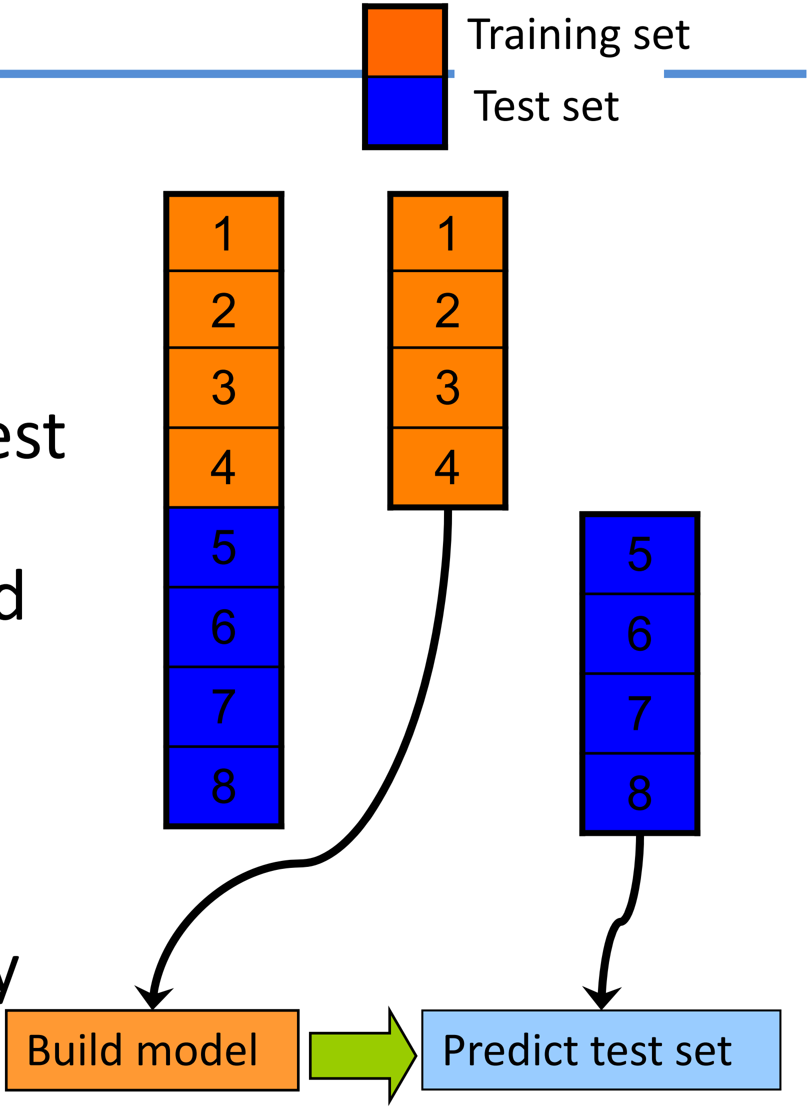
Overfitting





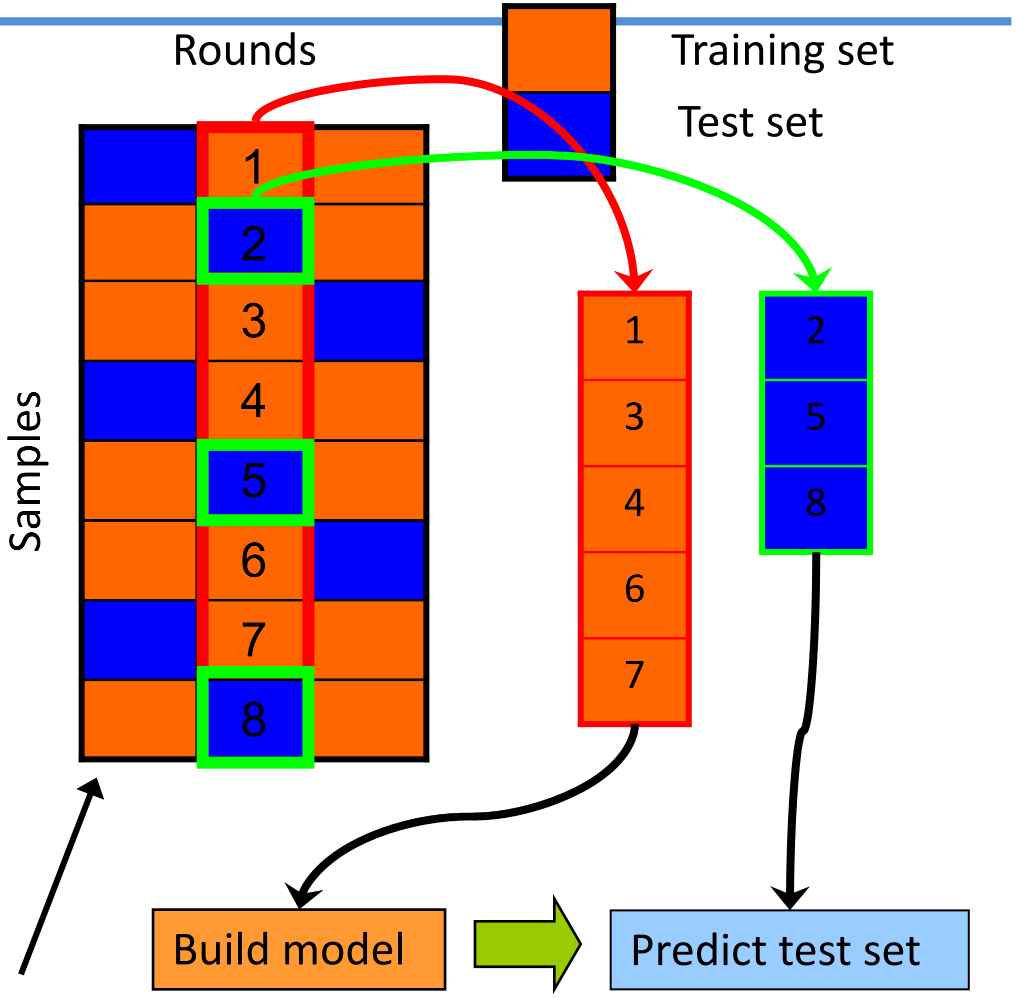
Model validation (2) - train/test

- *Split* the data
  - Training set - build the model
  - Test set - validate the model
  - Test set should be independent!
- Typically require >1/3 data in test set
- *All* model parameters optimised on training set
  - E.g. no. components, variables selected etc.
- Goodness of fit statistic on test data indicates predictive quality of the model



Model validation (3) - Cross-validation

- General principle:
  1. Remove some data
  2. Build model on remaining data
  3. Predict removed data
  4. Repeat until all samples removed once
- Compute predictions & residuals (e_{ik}) for each sample when left out
- Calculate PRESS and Q^2 from all residuals
- Can do this for X or Y



‘3-fold’ cross validatoin

Model validation (4) - R^2 & Q^2

- $R^2 \rightarrow$ how much of the total variance is explained by the model
- $R^2 = 1 – \text{RESS} / \text{TSS}$

where

RESS = Residual Error Sum of Squares

$= \text{Sum}(e_{ik}{}^2)$

and

TSS = Total Sum of Squares

$= \text{Sum}(x_{ik}{}^2)$

- $Q^2 \rightarrow$ how much variance is *predictable* by the model
- Or…how robust model is to removing data
- $Q^2 = 1 – \text{PRESS} / \text{TSS}$

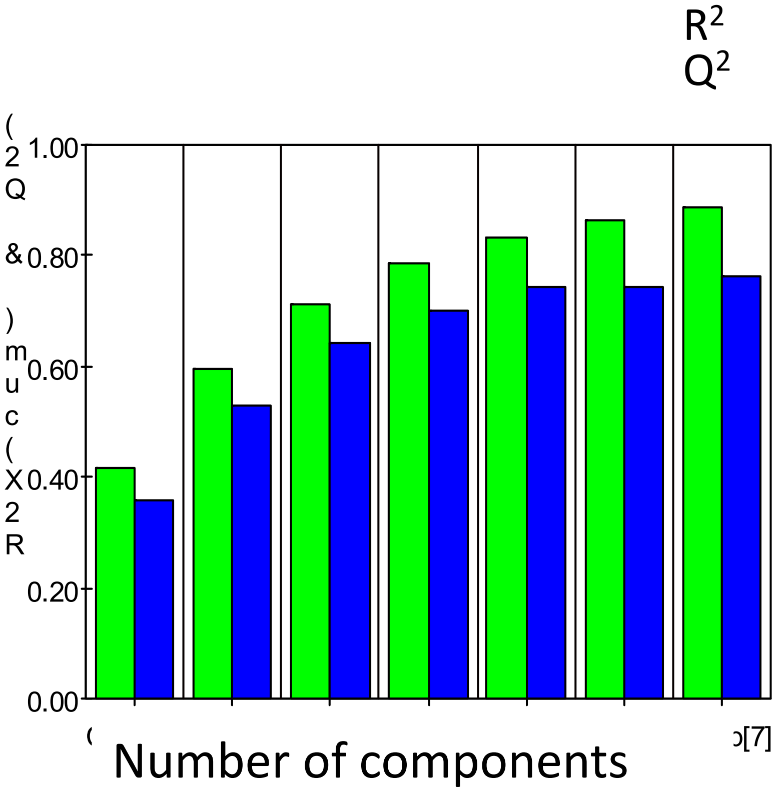
where

PRESS = *Predicted* Residual Error Sum of Squares

$= \text{Sum}(\hat{e}^2)$

Residual for a predicted sample

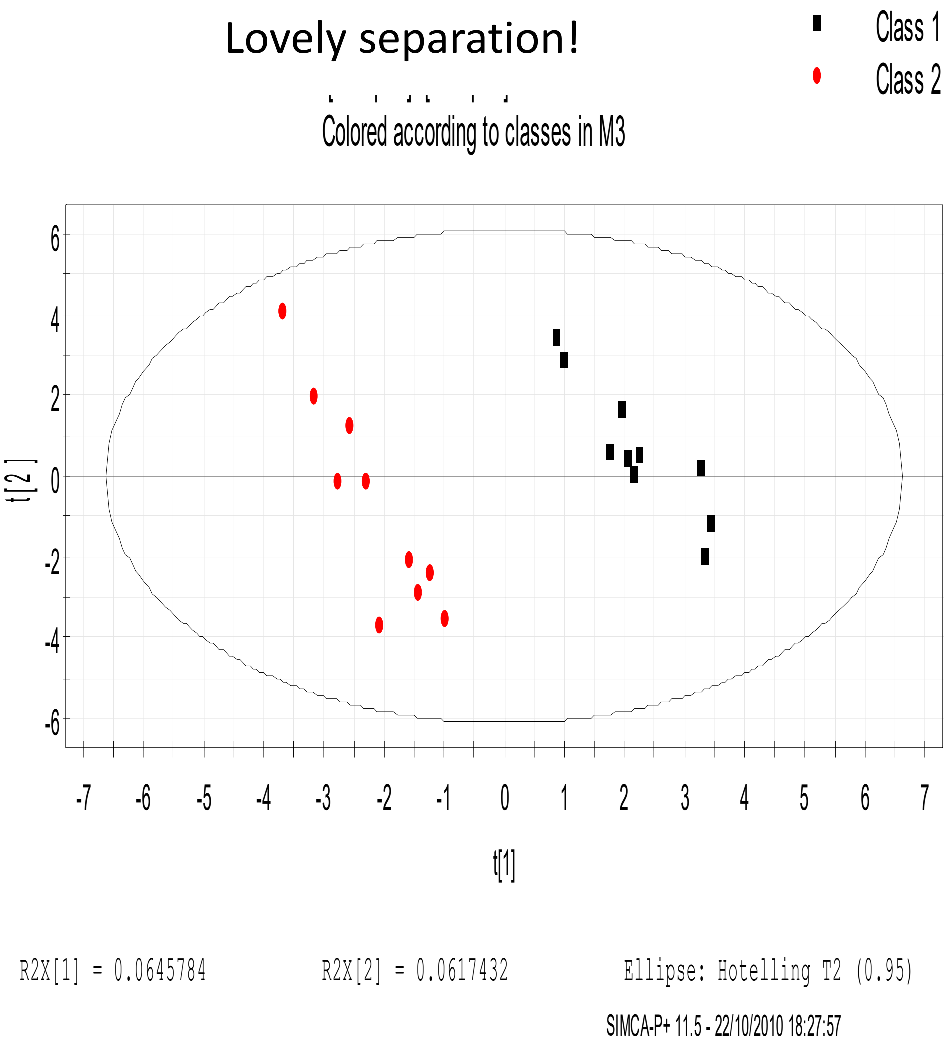
Cross-validation - R^2 and Q^2



- R^2 & Q^2 plot from SIMCA-P software
- R^2 rises with each component
- Q^2 rises, then reaches plateau or falls
- **Extra components are fitting structure which is unstable → noise**

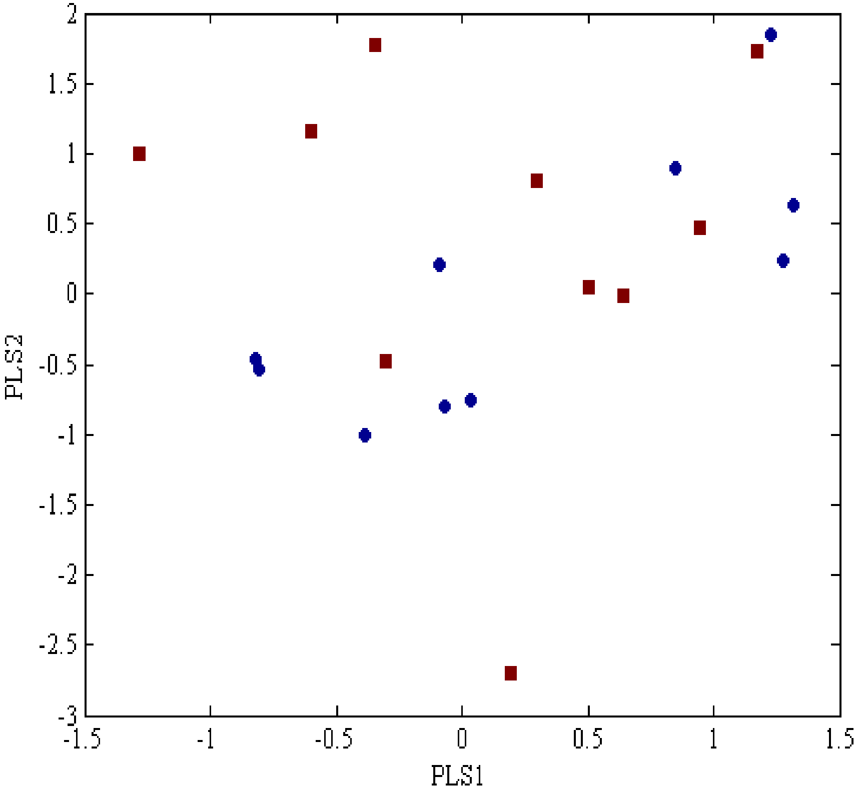
Cross-validated scores: PLS-DA

Lovely separation!



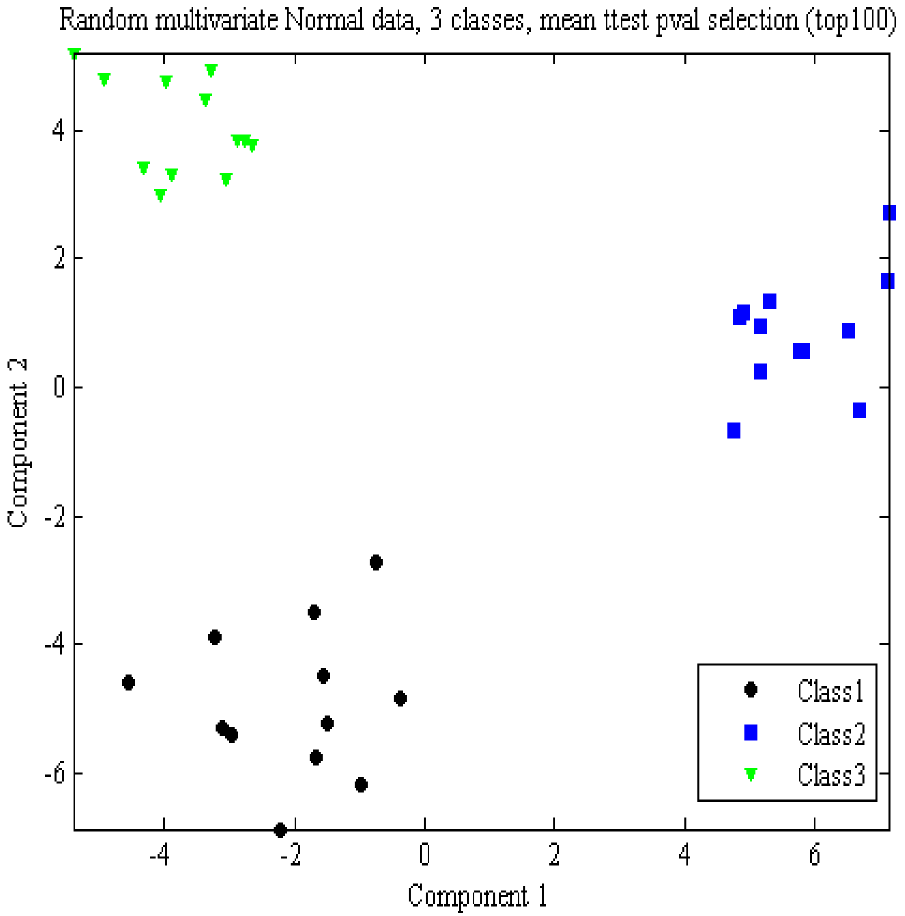
Random data! (20x100)

Cross validated scores:



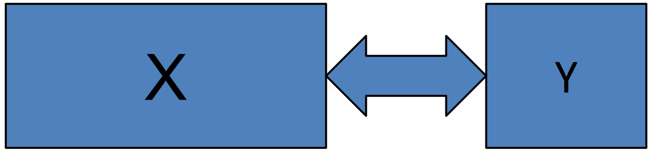
Variable selection & unsupervised methods

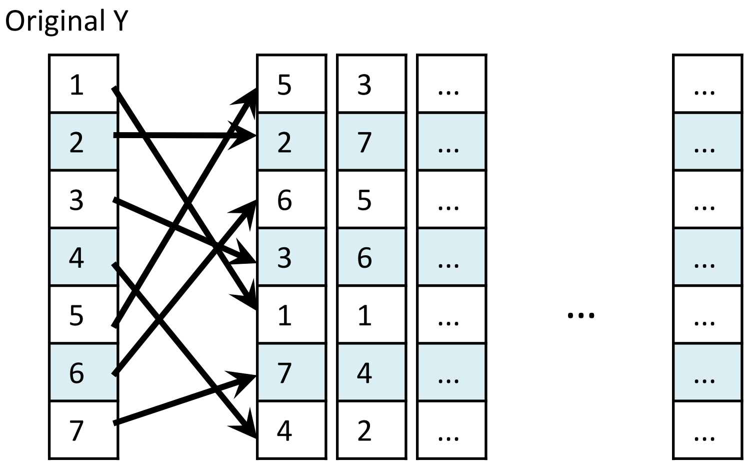
- Be suspicious of unsupervised methods (e.g. PCA) if data are pre-filtered
- Presence of clusters is not surprising!
- CV scores will not help...
  - Unless CV includes filtering step



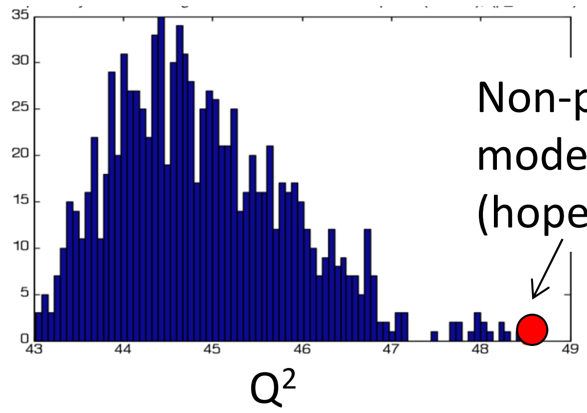
Model Significance (How good is my Q^2?)

- Permutation test:
  - Does order of Y make a difference?
  - If similar quality model with permuted Y then original model must be weak
- CV-ANOVA
  - Is regression model significantly better than constant model?
  - F-test on CV residuals





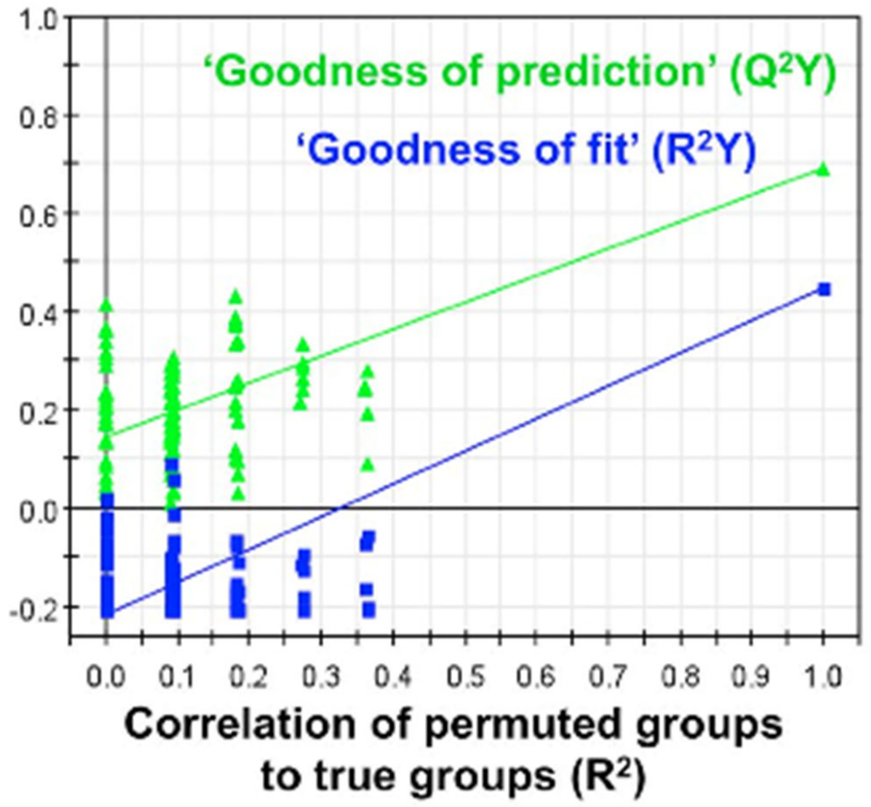
1	5	3	...	...	...
2	2	7	...		...
3	6	5	...		...
4	3	6	...		...
5	1	1	...	...	...
6	7	4	...		...
7	4	2	...		...



Non-permuted model (hopefully!)

Permutation tests

- Similarity to the original model depends on permutation
- Plot correlation of permuted Y to original Y vs. model performance
- Look for improvement of original model over performance at zero correlation



Model performance & confusion matrix

		Actual		Total
		+	-	
Predicted	+	TP	FP	P’
	-	FN	TN	N’
Total		P	N	

Sensitivity = TP/(TP + FN) = probability of detecting positive

Specificity = TN/(FP + TN) = probability of detecting negative

Positive Predictive Value (PPV) = TP/(TP + FP) = probability sample is positive *given* that it is predicted positive

→ Isn’t this what you want?!

PPV is great, what's the problem?

- Sensitivity, specificity are properties of the model – do not change with class sizes
- PPV is dependent on prior proportion of +/-
- E.g. Same model, different class sizes:

		Actual		Total
		+	-	
Pred	+	45	5	50
	-	5	45	50
Total		50	50	

		Actual		Total
		+	-	
Pred	+	9	9	18
	-	1	81	82
Total		10	90	

Se = Sp = 45/(45+5) = 90%
PPV = 45/(45+5) = 90%

Se = 9/10 = Sp = 81/90 = 90%
PPV = 9/(9+9) = 50%

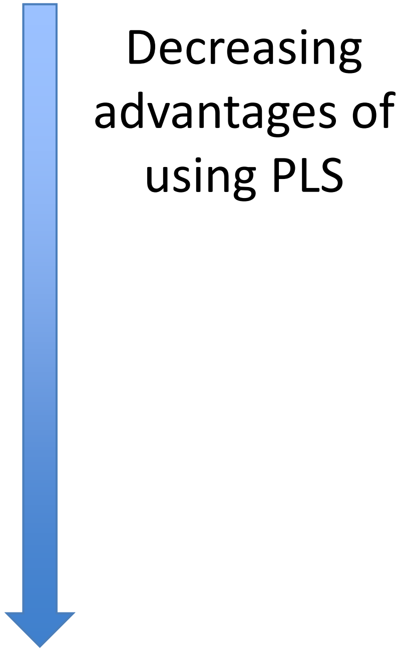
Outline

Model validation & assessing model performance

Why use PLS for metabolic profiling?

Why use PLS?

- To generate a model that is predictive of some parameter (or parameters) Y?
- To find out if two groups are different, taking into account the global metabolic profile?
- To find out what are the metabolic differences between two groups?

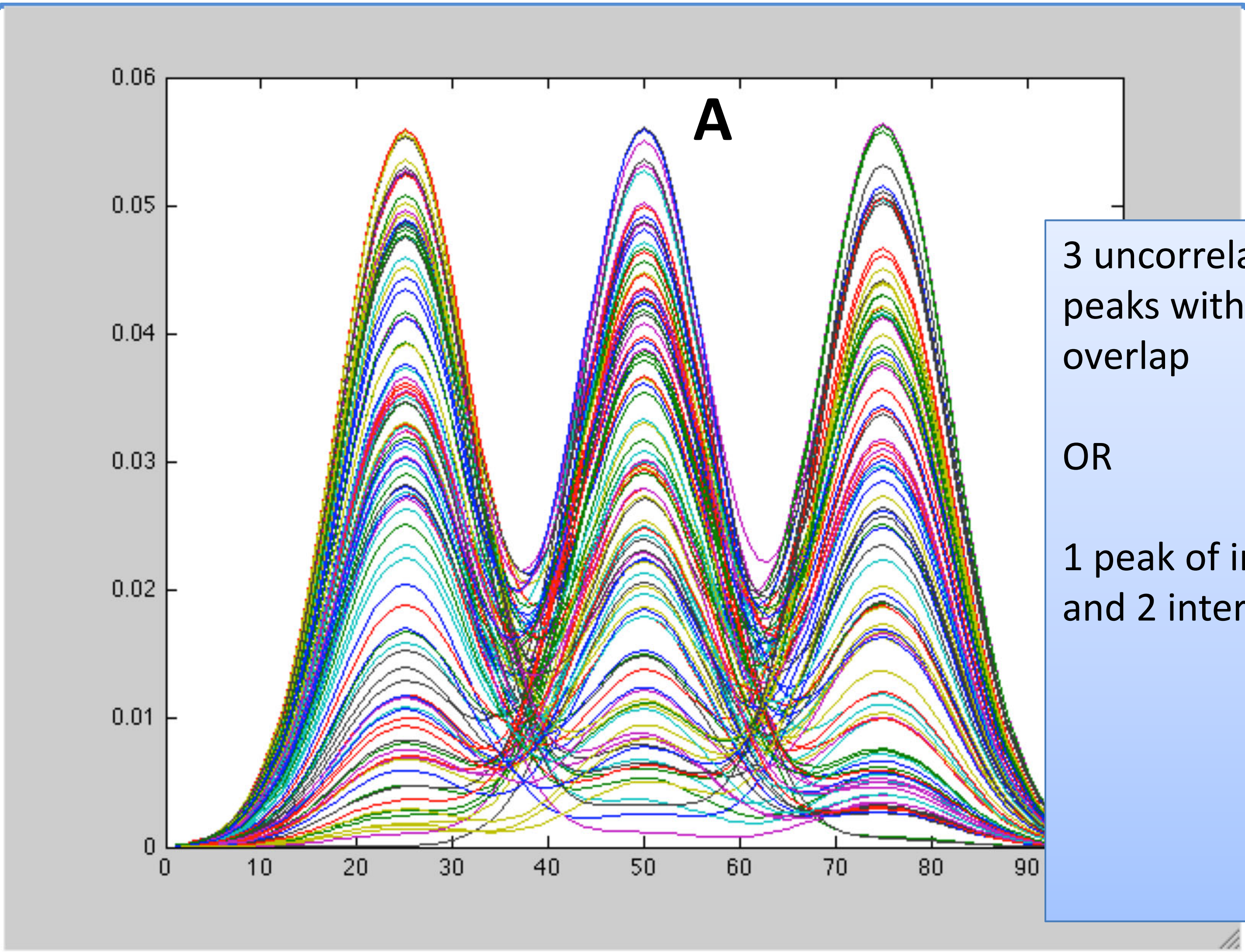


Decreasing advantages of using PLS

Main advantages of PLS are that it can model

1. **distributed correlations &**
2. **overlapping interferences**

Initial data set: no noise



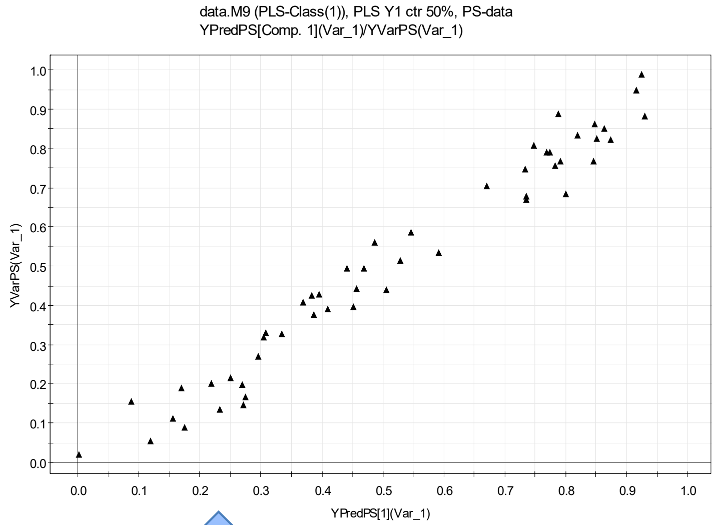
3 uncorrelated peaks with minor overlap

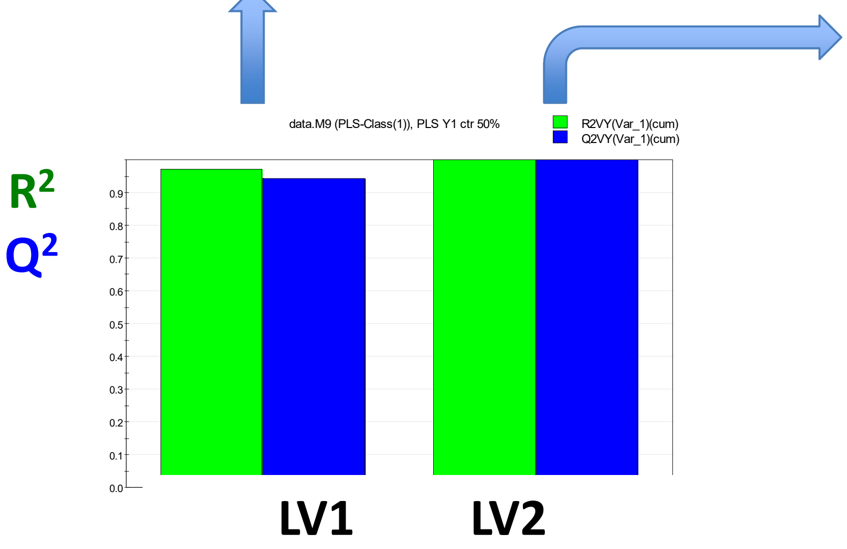
OR

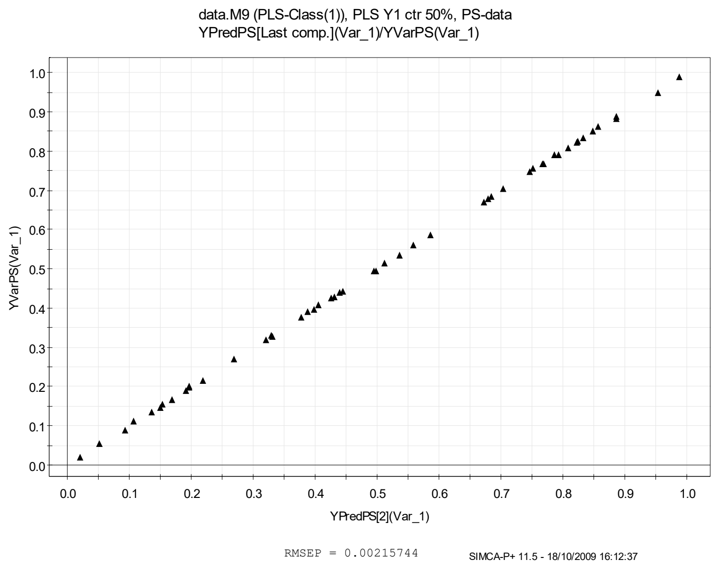
1 peak of interest and 2 interferences

PLS prediction of the concentration of A

- 1 latent variable ("component") does most of the work (Q2~0.95, RMSEP ~0.05)
- 2 LV (Q2=0.99), RMSEP ~0.002


data.M9 (PLS-Class(1)), PLS Y1 ctr 50%, PS-data
YPredPS[Comp. 1](Var_1)/YVarPS(Var_1)

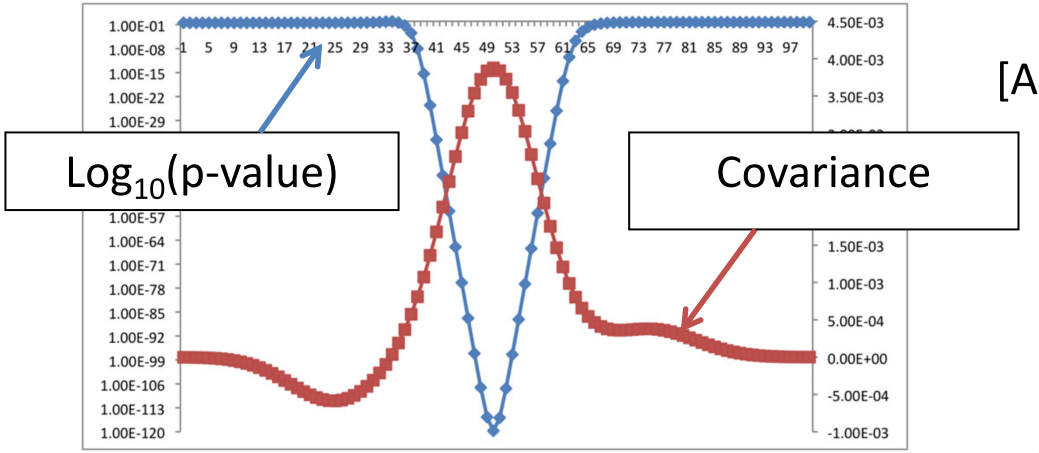

data.M9 (PLS-Class(1)), PLS Y1 ctr 50%


data.M9 (PLS-Class(1)), PLS Y1 ctr 50%, PS-data
YPredPS[Last comp.](Var_1)/YVarPS(Var_1)

RMSEP = 0.00215744

SIMCA-P+ 11.5 - 18/10/2009 16:12:37

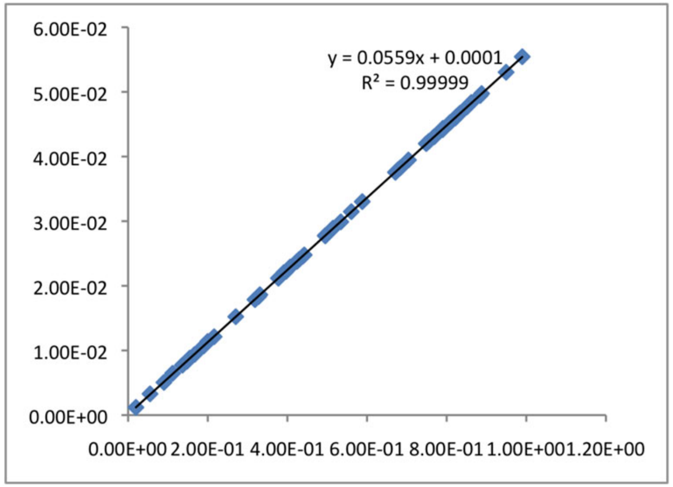
Comparison to univariate analysis



Log_{10}(p-value)

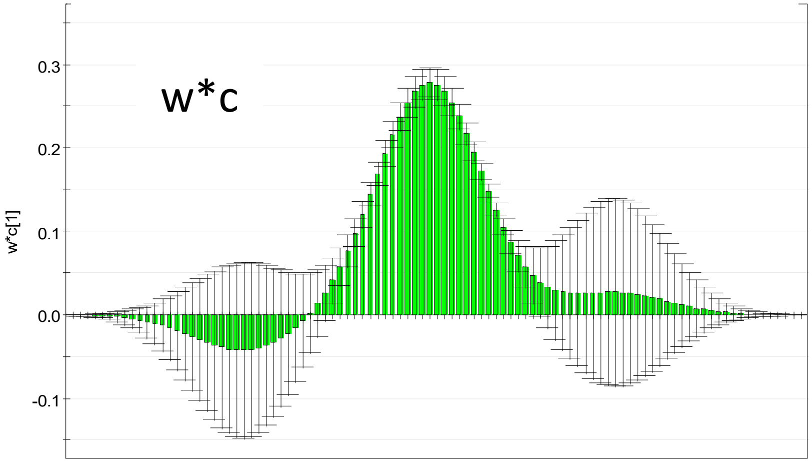
Covariance

[A]



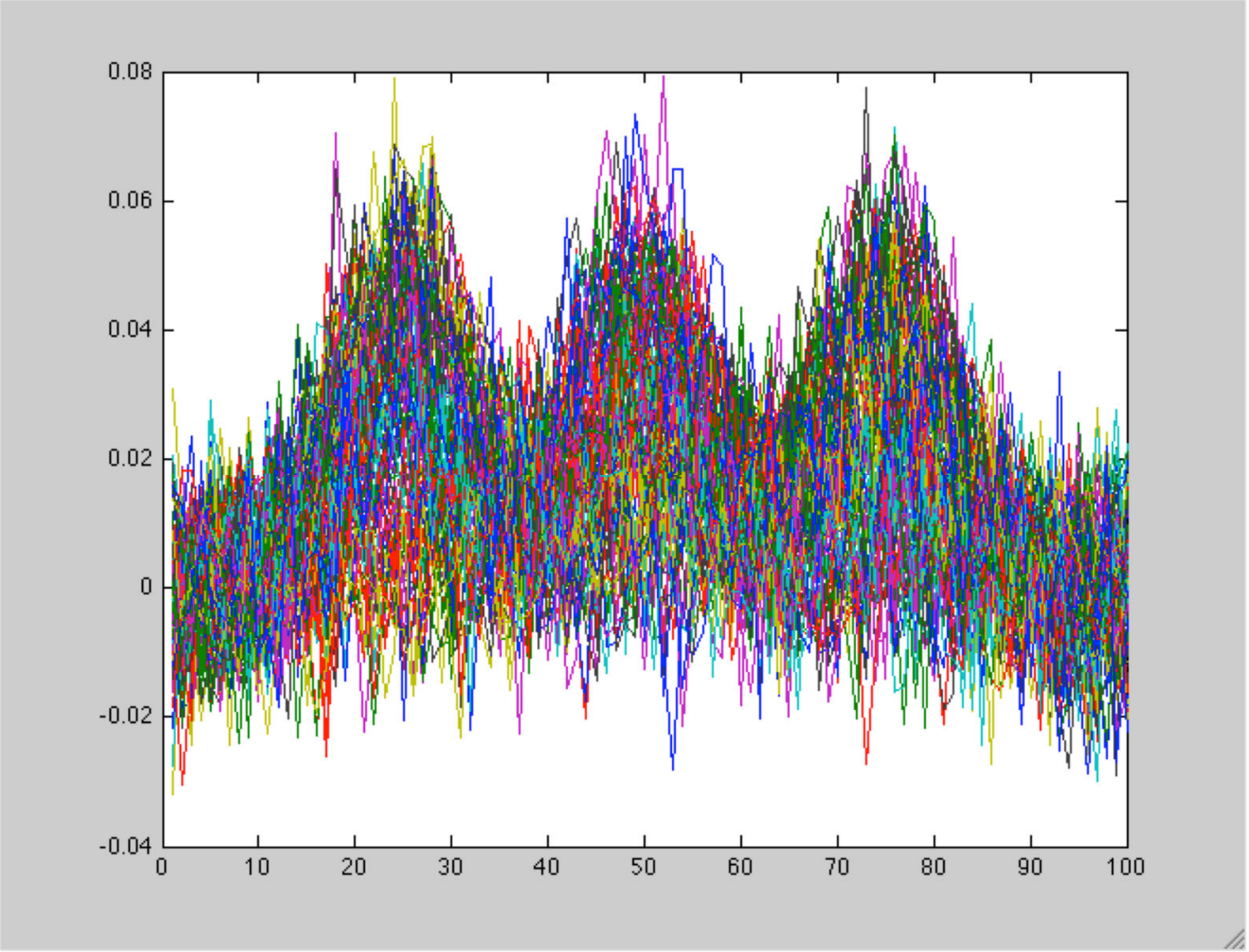
Intensity of most significant variable

Both (univariate) correlation and PLS pick centre variable as best predictor

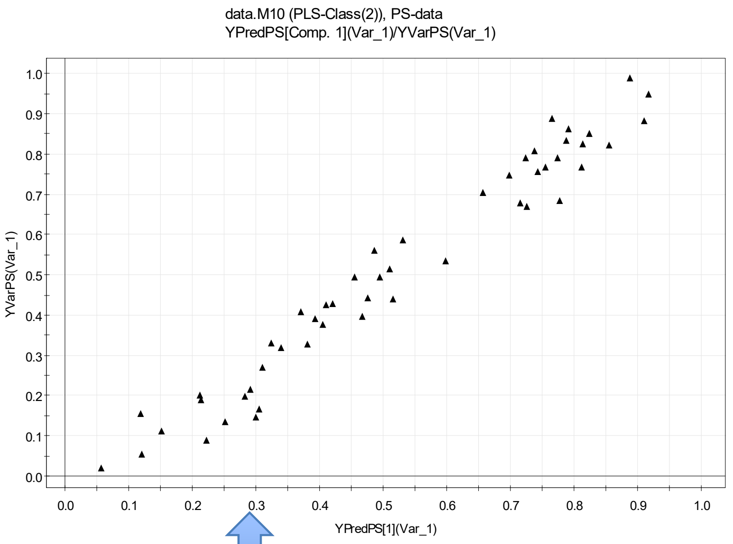


PLS model regression weights (1 LV)

What happens when there is lots of noise?

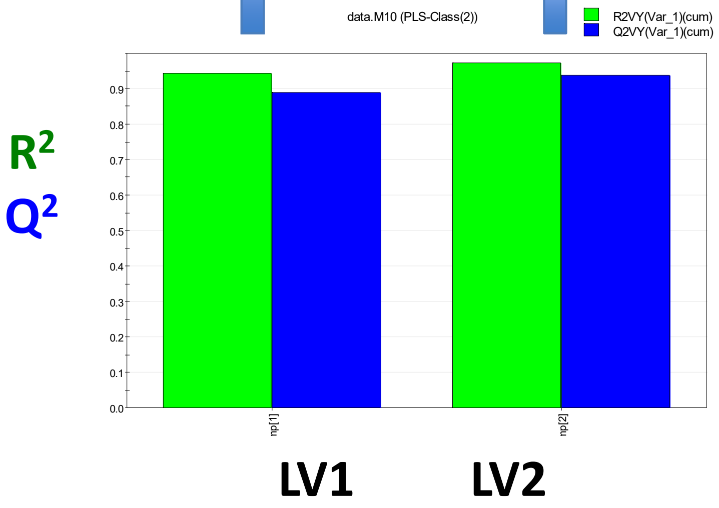


PLS prediction of the concentration of A with a low signal to noise dataset

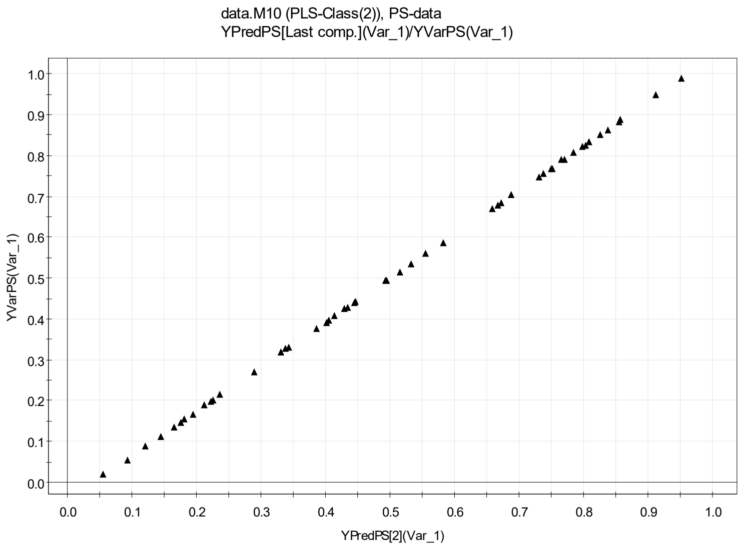


RMSEP = 0.0624908

SIMCA-P+ 11.5



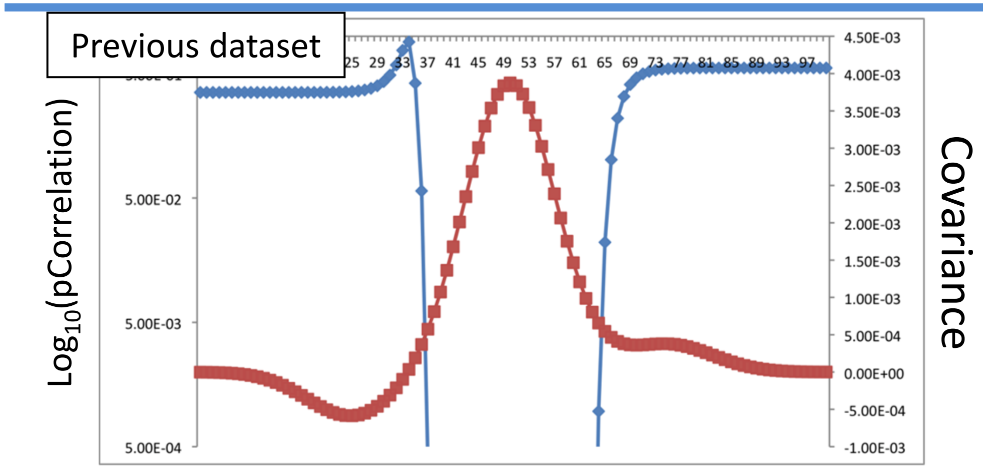
- 1 latent variable (“component”) **STILL** does most of the work (Q2~0.89, RMSEP ~0.06)
- 2 LV (Q2=0.94), RMSEP ~0.02)
- Still great overall prediction



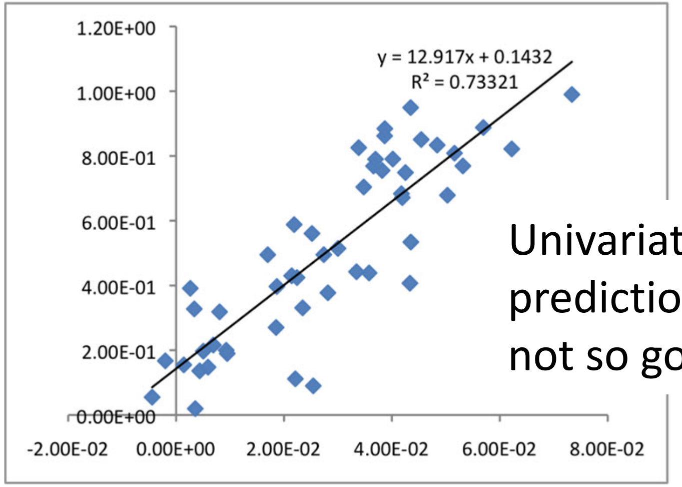
RMSEP = 0.0211616

SIMCA-P+ 11.5 - 18/10/2009 16:09:43

Comparison to univariate analysis with a low signal-to-noise dataset

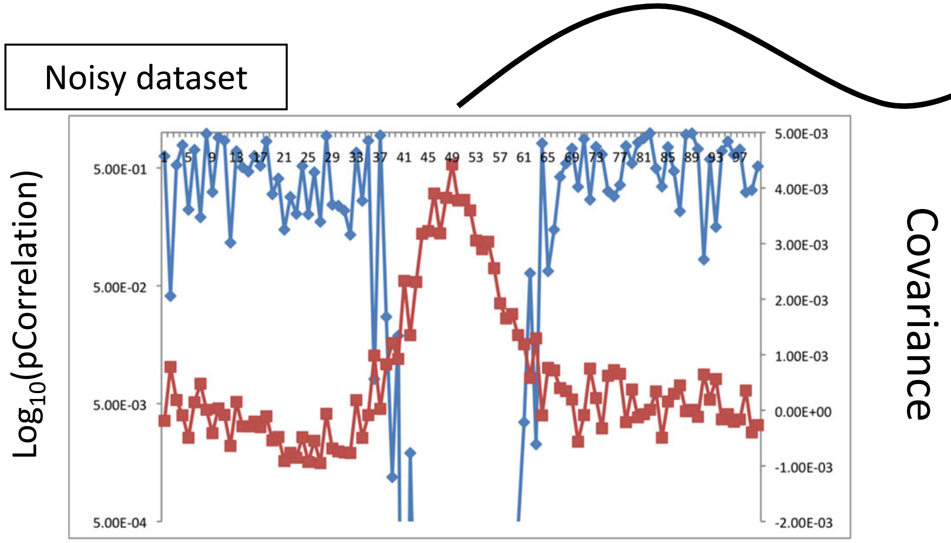

Previous dataset

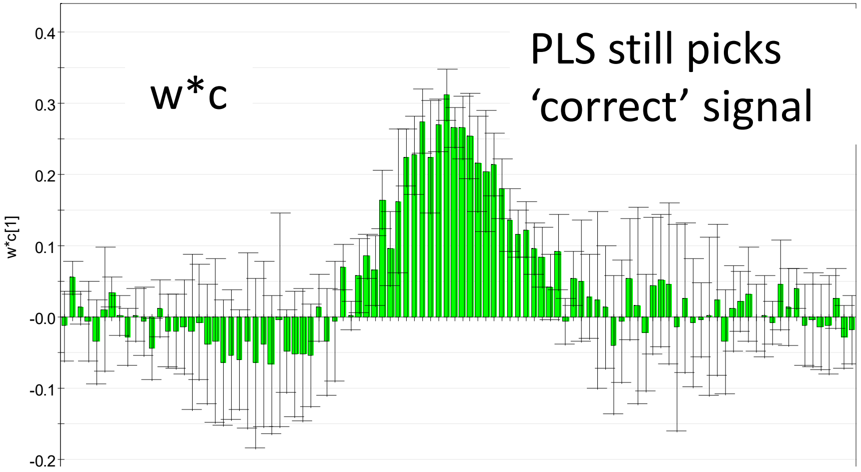
[A]



Univariate prediction not so good

Intensity of most significant variable


Noisy dataset

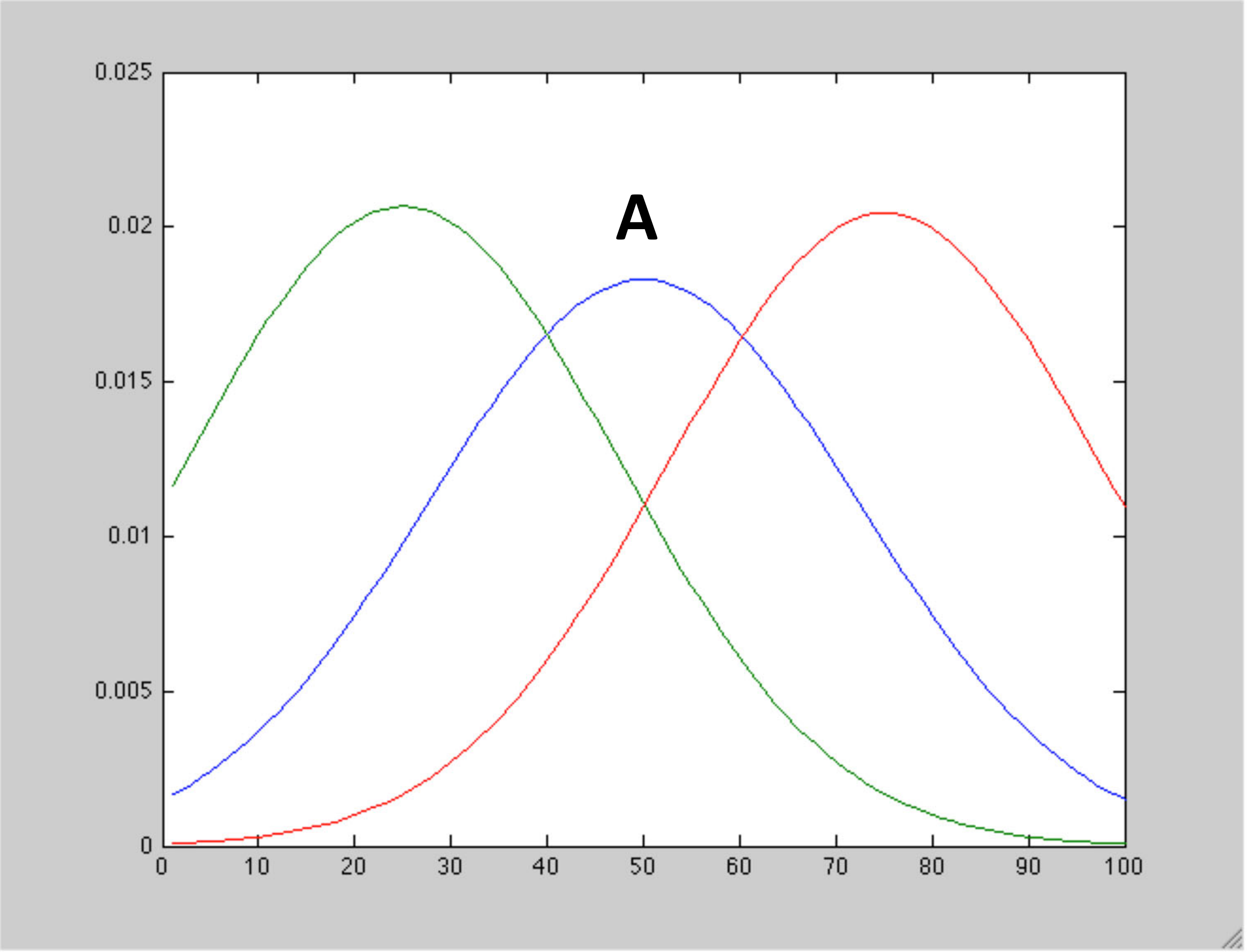


w*c

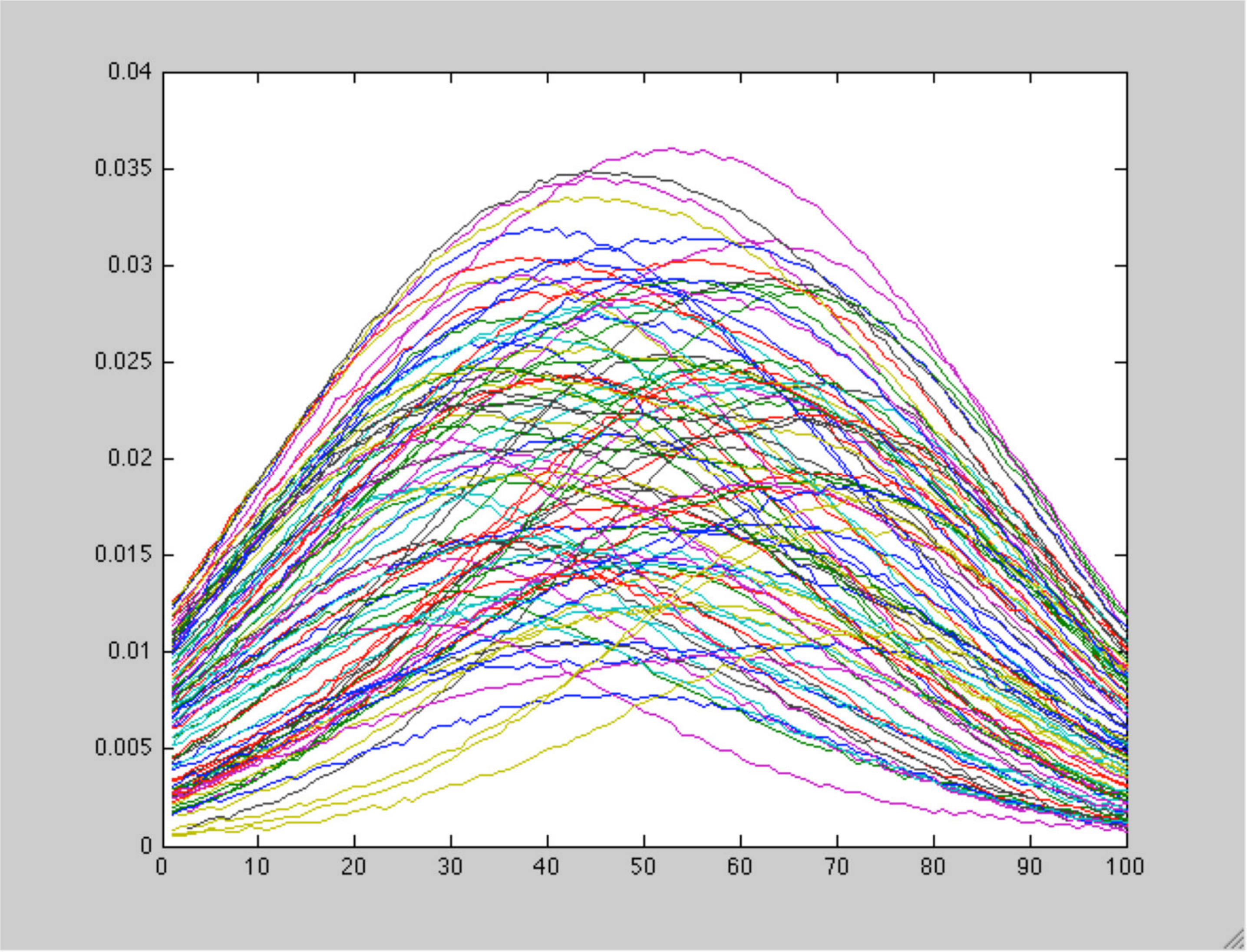
PLS still picks 'correct' signal

PLS model regression weights (1 LV)

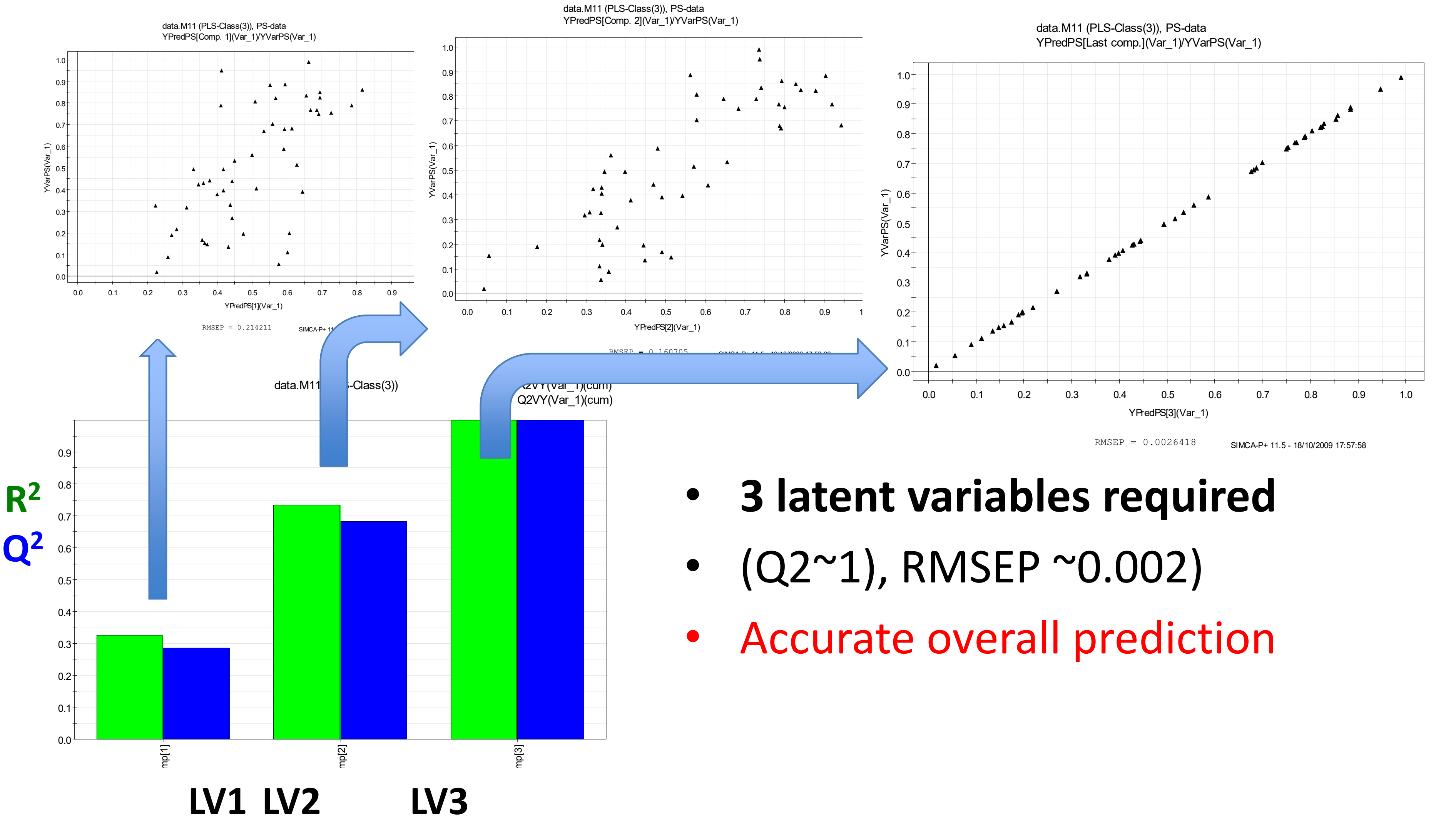
What happens when signals are highly overlapped?



New data set (low noise)



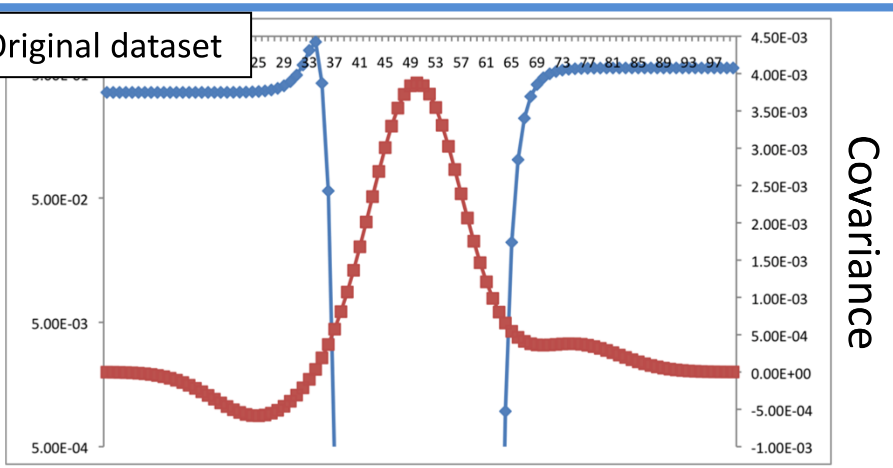
PLS prediction of the concentration of A with a poorly-resolved dataset



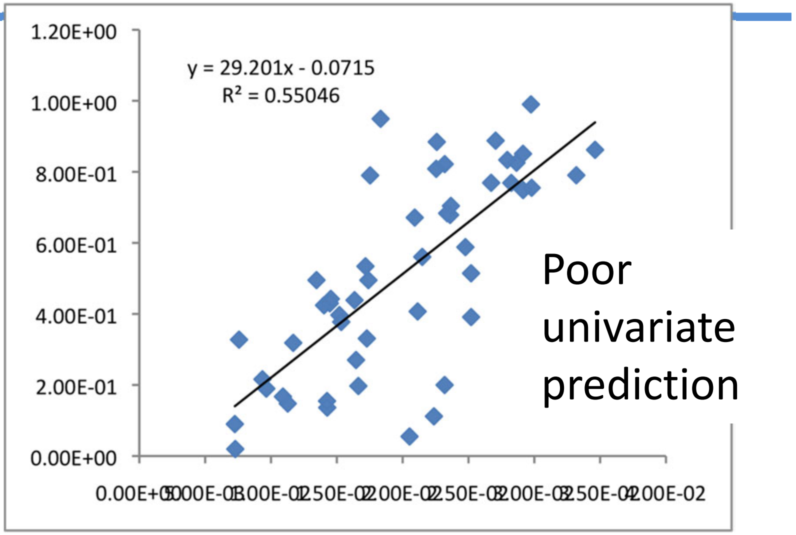
- **3 latent variables required**
- (Q2~1), RMSEP ~0.002)
- Accurate overall prediction

Comparison to univariate analysis with a high overlap dataset

Original dataset



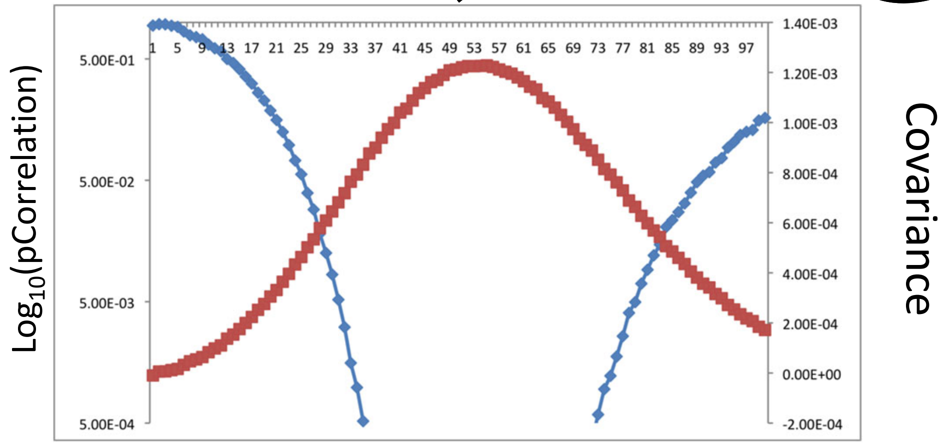
[A]

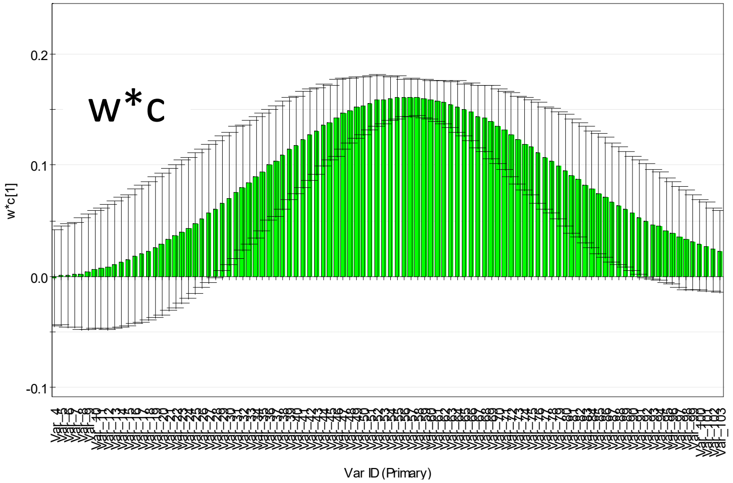


Poor univariate prediction

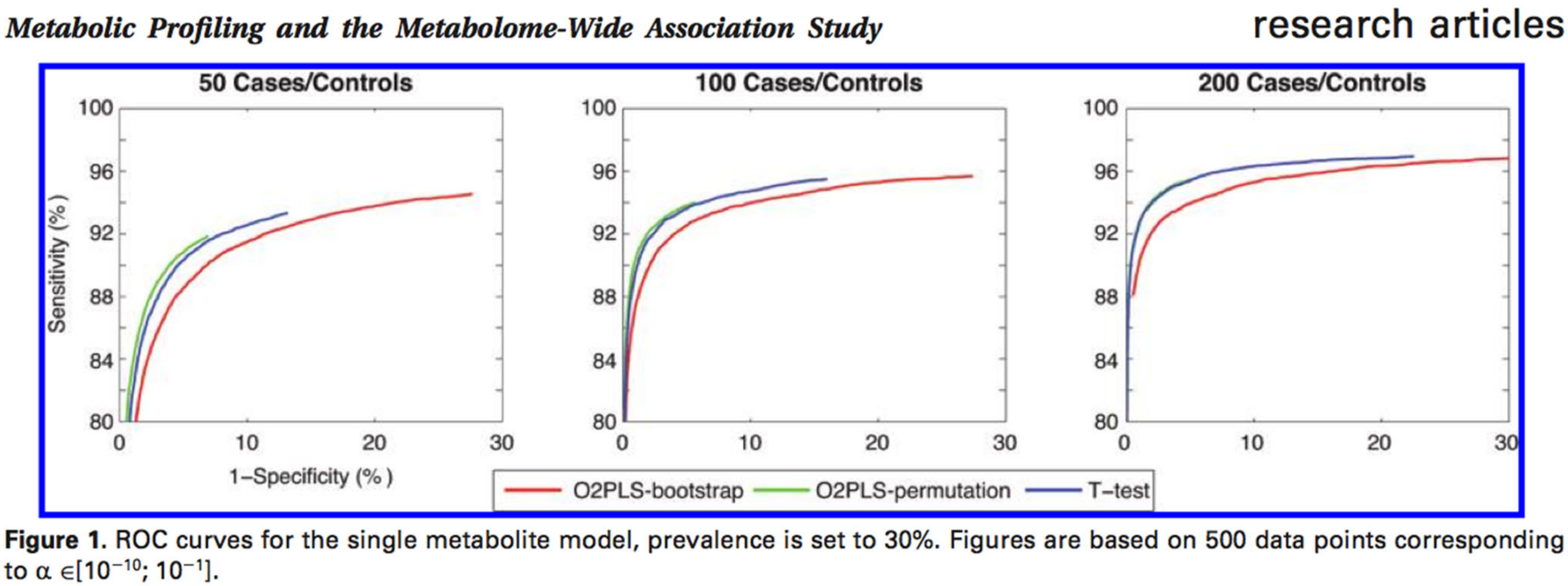
Intensity of most significant variable

Overlapped dataset





PLS model regression weights (1 LV)



Metabolic Profiling and the Metabolome-Wide Association Study research articles

Figure 1. ROC curves for the single metabolite model, prevalence is set to 30%. Figures are based on 500 data points corresponding to $\alpha \in [10^{-10}; 10^{-1}]$.

In practice PLS may not be better at telling exactly **which** metabolites are different between e.g. two groups than standard univariate approaches

Chadeau-Hyam, Ebbels et al. J Proteome Res 2010

What types of spurious variation can PLS **not** cope with?

- High random noise in datasets with few samples to variables (e.g. > 1:100)
- Outliers (though these are easily flagged)
- Random variation in a few variables with very high intensity (scaling)
- Large variation in global intensity (normalisation problem)
- Experimental bias correlated to the parameter of interest

Summary

A mixed bag:

- Model validation is a *must*
- Be aware of variable selection, multiple testing, class imbalance, permutation tests
- PLS has advantages in coping with distributed, correlated responses and interferences