Institute of Food Research

OPLS: an ideal tool for interpreting PLS regression models?

1st July 2008

Henri Tapp

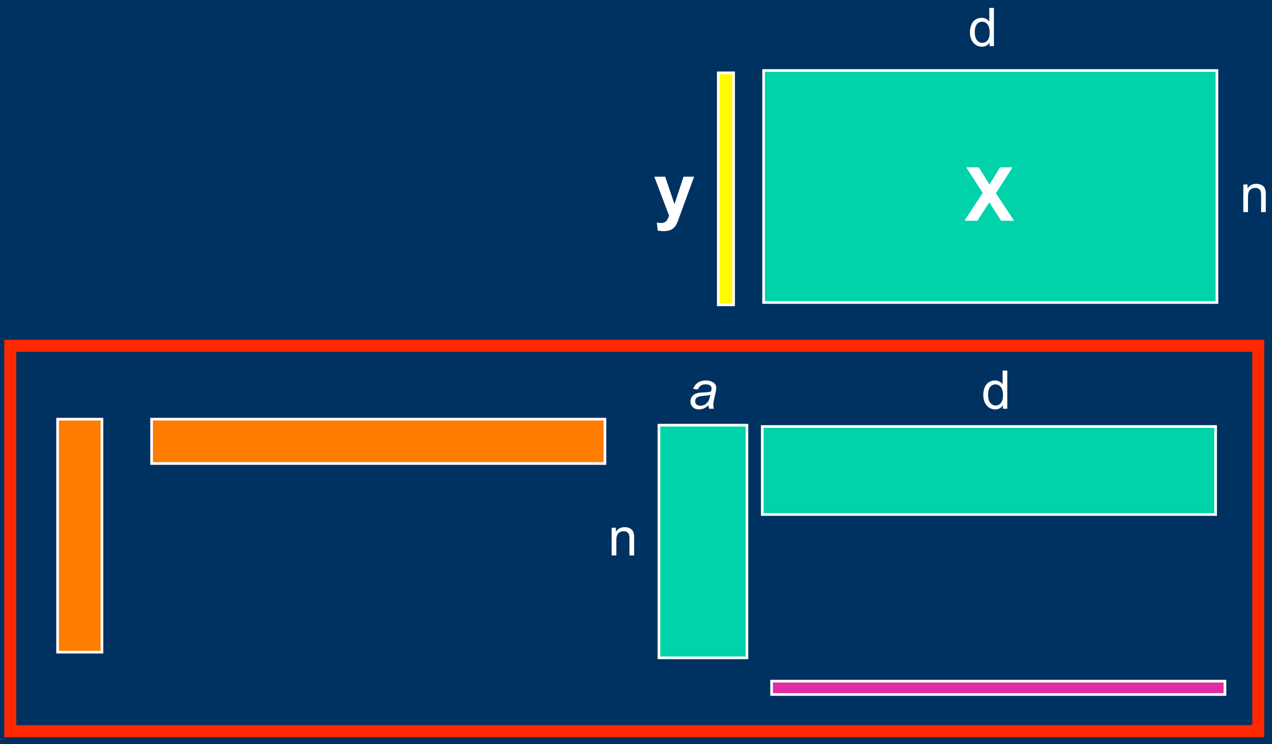
Kate Kemsley

Contents

- Introduction
- Why use PLS instead of OPLS
  - Same predictive performance
  - PLS faster
  - Can use 1st PLS vector for ‘OPLS interpretation’
- Look at the changes in X
- Conclusions

Introduction

- OPLS: PLS with integrated OSC filter



- PLS regression – predict vector **y** from matrix **X**
- Model dimensionality defined by no. PLS factors, *a*
- Estimate *a* using cross -validation
- Interpret regression model in terms of original variates

$$\mathbf{X}_{opls} = \mathbf{X} - \mathbf{T}_o \mathbf{P}_o^T \qquad \mathbf{X} = \mathbf{T}\,\mathbf{P}^T + \mathbf{E}_{pls} \qquad \mathbf{X} = \mathbf{T}_p \mathbf{P}_p^T + \mathbf{T}_o \mathbf{P}_o^T + \mathbf{E}_{opls}$$

- Tweak 1: modify X_{opls}, e.g. rescale, derivatise
- Tweak 2: reduce amount filtered based on PCA of $T_oP_o^T$

Introduction

- OPLS: two properties

For y vector case and **no** *tweaks*

1. Predictions using *q* OPLS filter factors and *p* PLS regression factors are identical to predictions from a PLS model using *a* regression factors, where *a* = *q* + *p*

   *PLS and OPLS have **same** performance*

2. The 1st PLS latent vector is unaltered during OPLS 'filtering'.

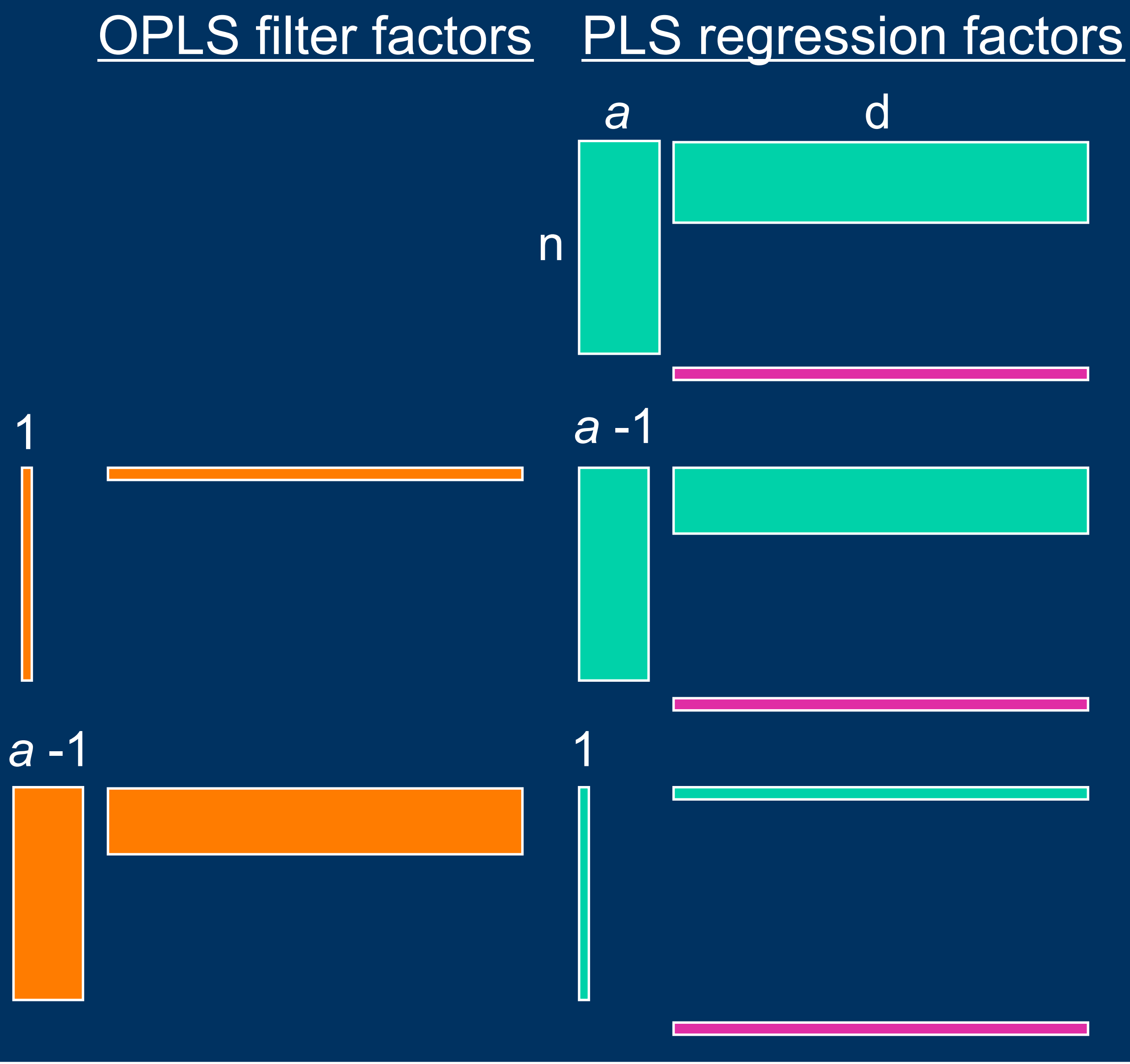
   1st PLS vector can be interpreted as regression coefficients for OPLS filtered **X**

Main conclusion: build models using PLS instead of OPLS

Why use PLS instead of OPLS?

- **Same predictive performance**
- Avoid overfitting when *a* = 1
- PLS already standard chemometric tool
- **PLS faster** (e.g. SIMPLS)
- Simpler: 1 step versus 2 steps
- **Can use 1st PLS vector for 'OPLS interpretation' …**
- Have choice of post-processing methods for further analysis

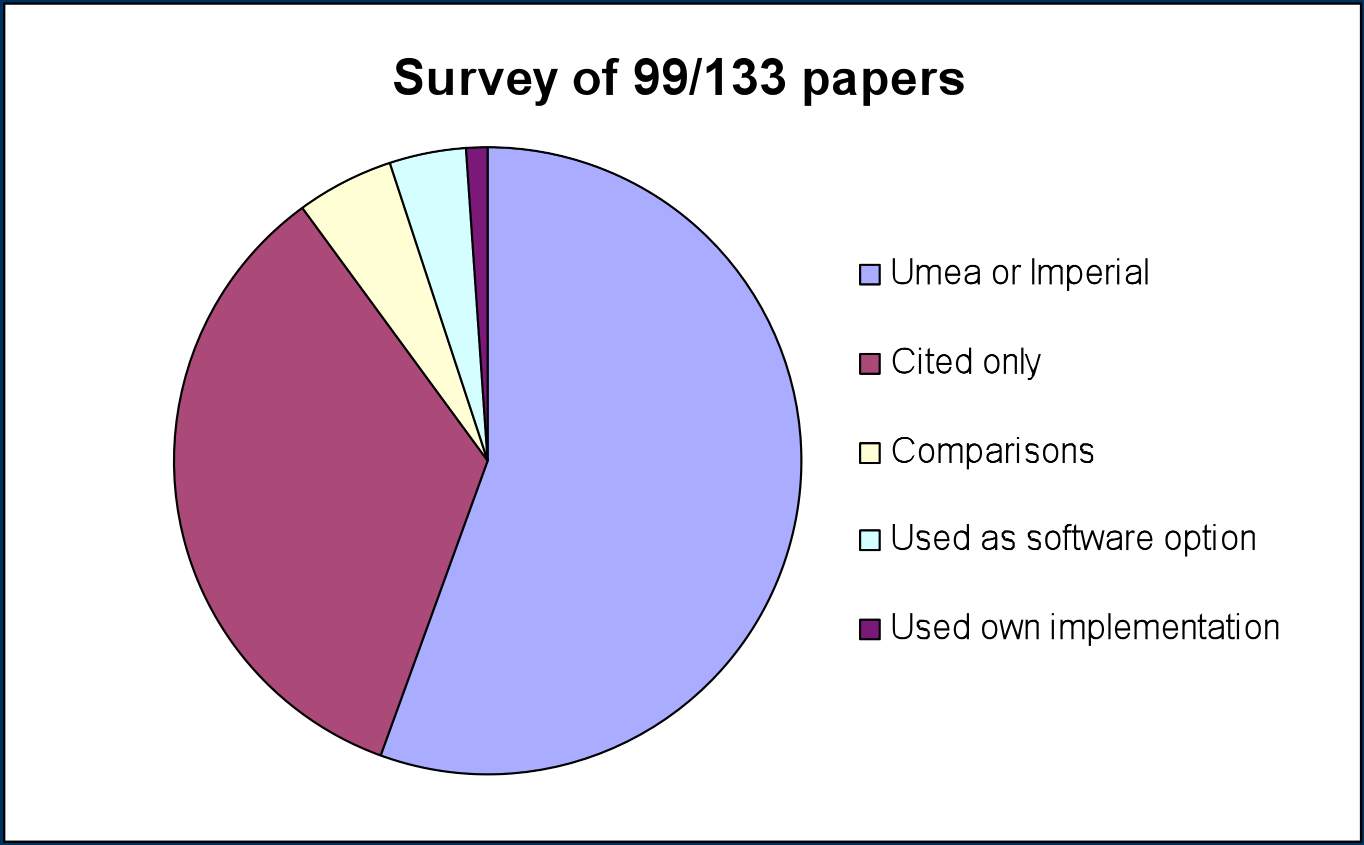
Why use PLS instead of OPLS?
- Same predictive performance



- Have choice in partitioning model into filter and regression components
- Implications when interpreting **X** in terms of ‘orthogonal’, ‘predictive’, and ‘residual’ parts
- Convenient to use *a* - 1 filter factors
- Alternative interpretation: model split into ‘univariate weights’ and ‘multivariate advantage’

Why use PLS instead of OPLS?

- *Same predictive performance:* who knew?



Literature survey of OPLS

133 papers citing original OPLS paper or 2 subsequent ones on O2PLS (inc. original)

99 surveyed in detail

70/133 by Umea or Imperial (53%)

55/99 in survey (56%)

Reasonably representative

- Bruwer *et al* 2007 Ind. Eng. Chem. Res. **46** 864

Why use PLS instead of OPLS?

- *Same predictive performance:* who knew?

Good papers

- Ergon 2007 J Chemom. **21** 537 ★ ★ ★ ★ ★
- Ergon 2005 J Chemom. **19** 1 ★ ★ ★ ★ ★
- Yu and MacGregor 2004 CILS **73** 199 ★ ★ ★ ★ ★

Message sometimes unclear

- Whelehan et al 2006 CILS **84** 82
- Want et al 2007 J Proteome Research **6** 459
- Rezzi et al 2007 J Proteome Research **6** 513
- Wagner et al 2007 Anal. Chem **79** 2918

Tweaks can obscure equivalence

- Thennadil and Martin 2005 J Chemom **19** 77

Possible misinterpretation

- Samp et al 2003 J Inst. Brew. **109** 16

Why use PLS instead of OPLS?

- Same predictive performance: who knew?

Good paper

“The **separation provided by OPLS–DA is particularly impressive** and warrants further investigation in other proteomic studies.” - Whelehan *et al* 2006 CILS **84** 82

M

“Like PLS-DA, O-PLS-DA is a supervised pattern recognition technique, but has **improved predictive quality** because the structured noise is modeled separately.” - Want *et al* 2007 J Proteome Research **6** 459

“The O-PLS-DA method provides a **prediction similar to that of PLS-DA**, but the interpretation of the models is improved because the structured noise is modeled separately from the variation common to the X and Y matrices.” – Rezzi *et al* 2007 J Proteome Research **6** 513

Tweaks can obscure equivalence

- Thennadil and Martin 2005 J Chemom **19** 77

“In OPLS, the group discrimination is forced to the first component, and thus **classification results improved enormously** as shown in Figure 4A and 4B.” - Wagner et al 2007 Anal. Chem **79** 2918

Possib

Why use PLS instead of OPLS?
- PLS faster

Can use SIMPLS rather than NIPALS

- de Jong 1993 CILS **18** 251

Is speed important?

YES - for model validation

OPLS survey

34 OPLS-DA

9 PLS-DA with 'OPLS interpretation'

Westerhuis et al 2008 Metabolomics **4** 81 ★ ★ ★ ★ ★
Assessment of PLSDA cross validation

Need to evaluate lots of sub models:

- Estimation of performance (Use double-cross validation)
- Significance of summary stats. (Use y-scrambling)

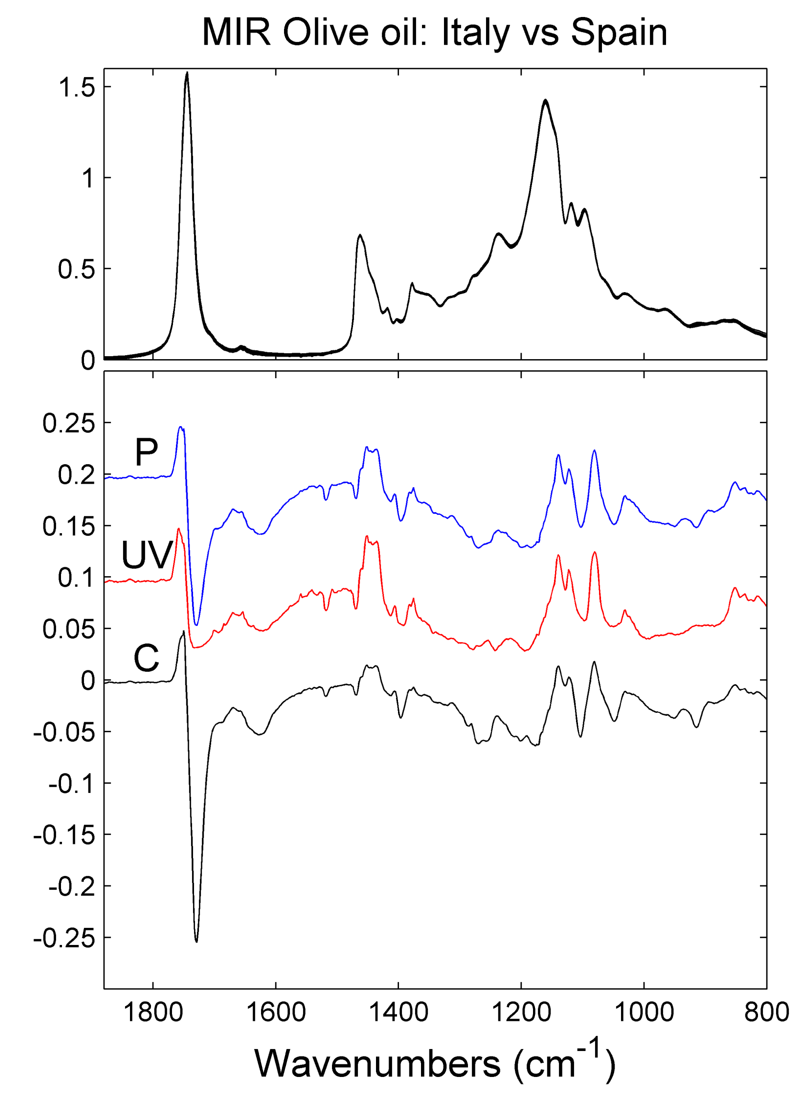
Are PLS regression coefficients still important?

YES – to assess the model stability in the presence of ALL the systematic variability

Why model using PLS instead of OPLS?

- Can use 1st PLS vector for ‘OPLS interpretation’

- Convenient to use
- Relative weights variate subset independent
- Potential for updating
- 1st vector dependant on scaling
- Determines which variates look interesting



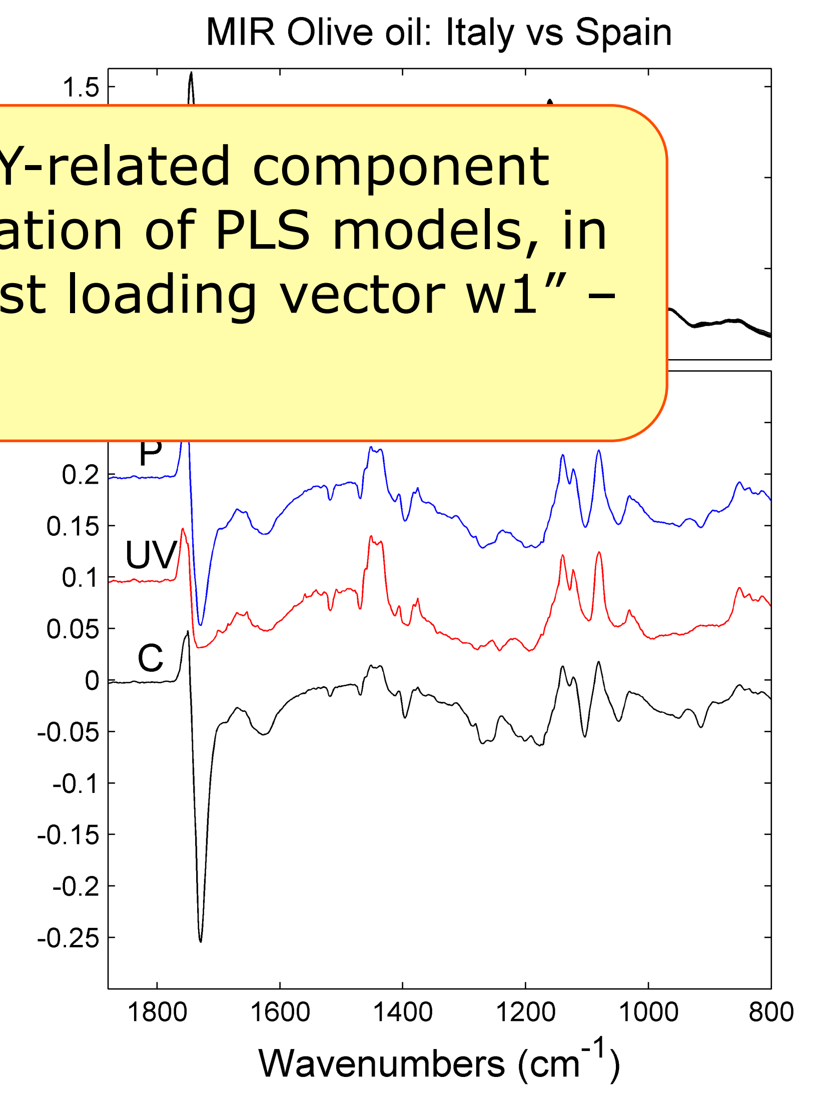
Why model using PLS instead of OPLS?

- Can use 1st PLS vector for 'OPLS interpretation'

- Convenient to use

> "It is now **known** that there exists only one Y-related component for a single Y-variable and that the interpretation of PLS models, in the single Y case, should be based on the first loading vector w1" – Jonsson *et al* (2005) Analyst **130** 701

- Potential for updating
- 1st vector dependant on scaling
- Determines which variates look interesting



Why model using PLS instead of OPLS?

- Can use 1st PLS vector for 'OPLS interpretation'

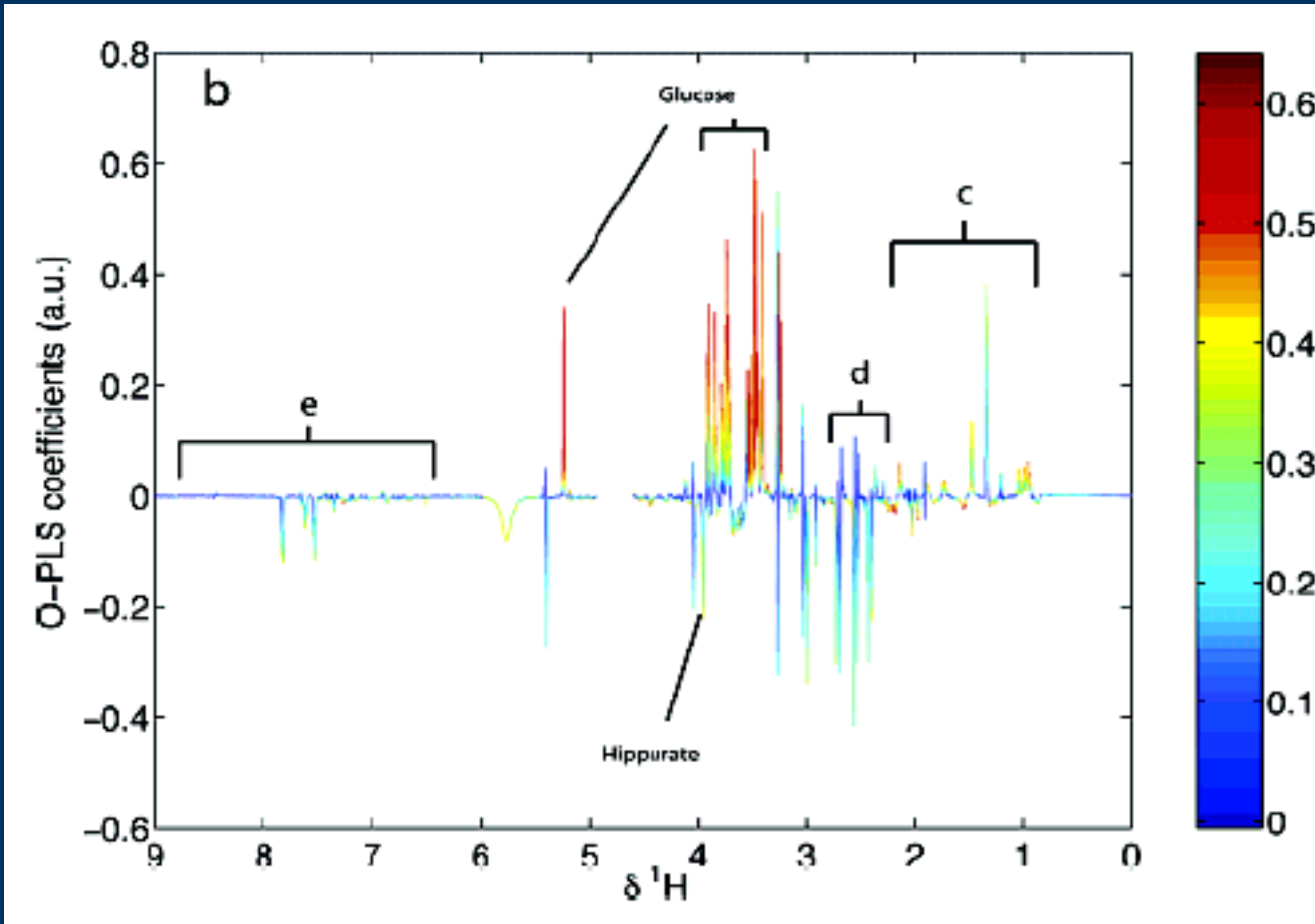
What to use?

1. Model using UV scaling – correlation based weights
2. Plot covariance – 'back scaling'
3. Colour code by correlation
4. Look for variates which are **both** high in correlation and covariance

Isn't this just a univariate based analysis?

"Differential metabogram"

- Martin et al 2007 J Proteome Research **6** 1471



Holmes *et al* 2006 J Proteome Research **5** 1313

Why model using PLS instead of OPLS?

- Can use 1st PLS vector for 'OPLS interpretation'

What to use?

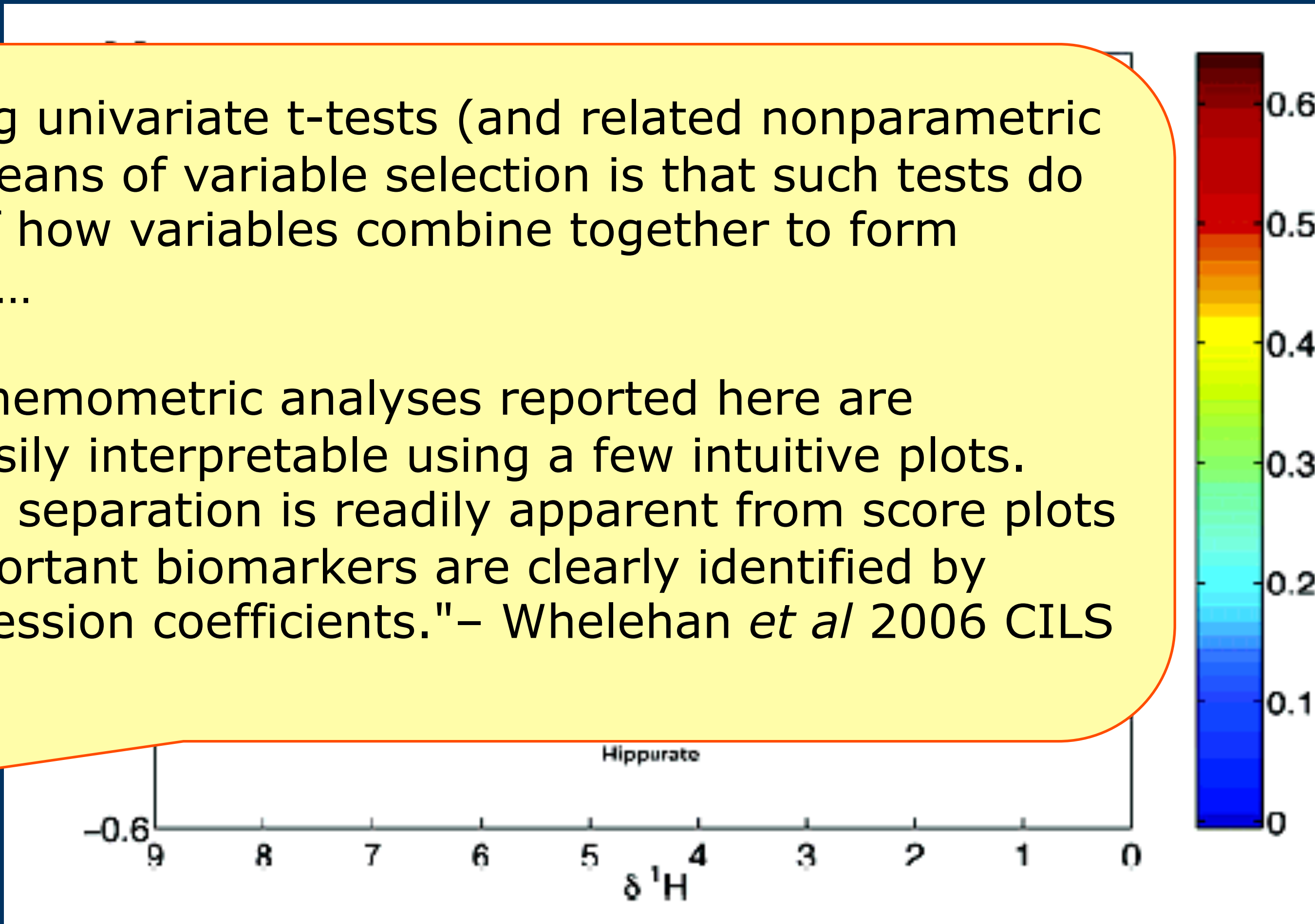
1. Model
corre
2. Plot c
scalir
3. Colou
4. Look
both
cova

Isn't this ju
based analysis?

Apparently not

"Differential metabogram"

- Martin et al 2007 J Proteome Research **6** 1471



"The danger of using univariate t-tests (and related nonparametric techniques) as a means of variable selection is that such tests do not take account of how variables combine together to form diagnostic patterns…

The results of the chemometric analyses reported here are transparent and easily interpretable using a few intuitive plots. The degree of class separation is readily apparent from score plots while the most important biomarkers are clearly identified by inspecting the regression coefficients."– Whelehan *et al* 2006 CILS **84** 82

Holmes *et al* 2006 J Proteome Research **5** 1313

Why model using PLS instead of OPLS?
- Can use 1st PLS vector for 'OPLS interpretation'

So far, build model using all the variates, then focus on interesting bits

Two PLS based suggestions

1. COVPROC

Hoskuldsson 2001 CILS **55** 23

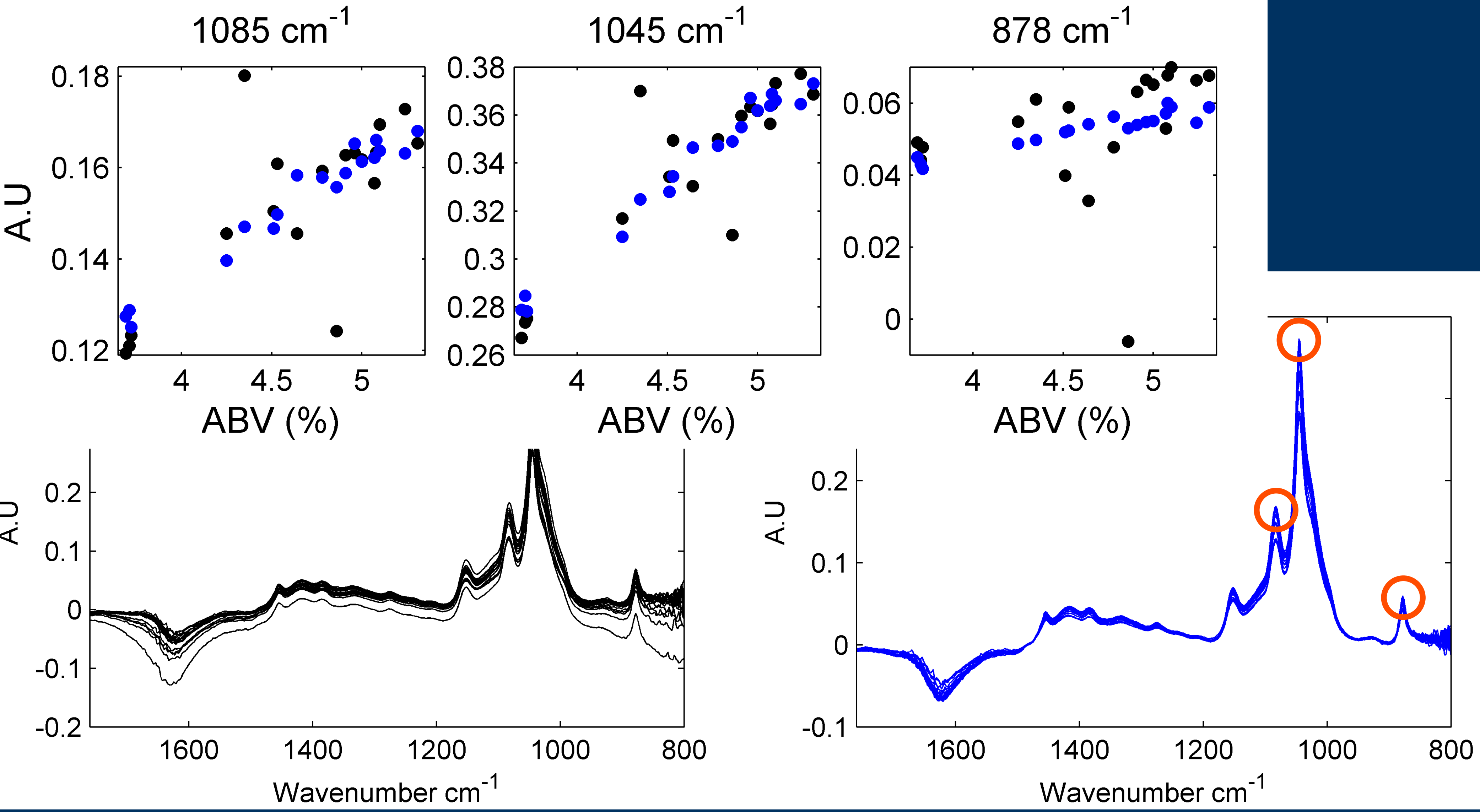
- adds variates based on magnitude of 1st PLS vector
- can appraise subset model performance
- complements ranked lists of variates, eg GSEA

2. Powered PLS

Indahl 2005 J Chemom. **19** 32

- Optimise the univariate weights
- Performs a restricted optimization of the weight vectors … that passes through the PLS1 weight vector solution.
- Also has variate subset selection properties

Look at the changes in X
– beer fermentation, variation in ABV (%)



Conclusions

- PLS is a tried and tested chemometric technique – don’t ditch it just yet
- There is no performance advantage over PLS
- OPLS explicitly splits PLS model into multivariate advantage and univariate weight
- Can look at impact of filter on individual variates – tangible representation of filter action
- OPLS better used for post processing rather than pre-filtering

And finally…

$\mathbf{y} = \mathbf{X}\,\beta$

Direct approach

$\beta = \mathbf{X}^{-1}\,\mathbf{y}$

Approximate $\mathbf{X}^{-1}$ by $\mathbf{X}^{T}$

$\mathbf{y}_{\text{new}} = \mathbf{x}_{\text{new-opls}} \times [\,\mathbf{X}^{T}\mathbf{y}\,]$

OPLS compensates for approximating the inverse of X by its transpose

Least-squares method of normal equations

$\beta = (\mathbf{X}^{T}\mathbf{X})^{-1} \times \mathbf{X}^{T}y$

$$\mathbf{y}_{\text{new}} = \mathbf{x}_{\text{new}} \times \quad [\,\underbrace{(\mathbf{X}^{T}\mathbf{X})^{-1}}_{\text{Multivariate advantage}} \quad \times \quad \underbrace{\mathbf{X}^{T}\mathbf{y}}_{\text{Univariate weight}}\,]$$

OPLS acts as regularised inverse of the covariance matrix