PLSR-DA
with **unbalanced data**

matthieu.lesnoff@cirad.fr

Heliospir, Montpellier, 24-25 June 2025

Clément Grelet
CRA-W Gembloux
Quality and authentication of agricultural products Unit

Wallonie recherche CRA-W

➔ **BHB** (beta-hydroxybutyrate) dataset

- Biomarker of ketosis in milking cows

BHB concentration in blood

Usual
Blood sampling

2 classes

Low	≤1.2 mmolg/L	
High	>1.2 mmol/L	⇒ Disease

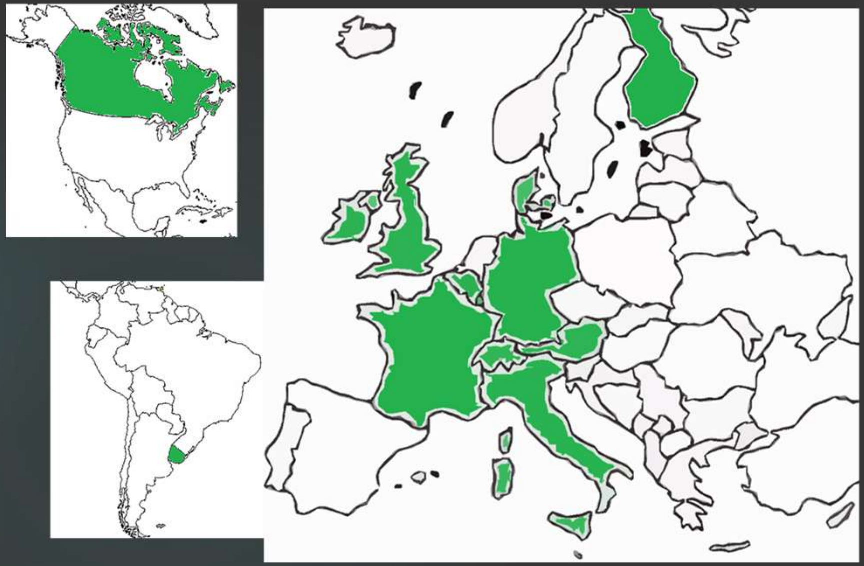
Possible alternative?

MIR spectrometry

on milk samples

European BHB consortium

Mark Crowe mark.crowe@ucd.ie; UCD, Ireland
Astrid Koeck koeck@zuchtdata.at; ZuchtData, Austria
Juergen Hummel jhummel@gwdg.de; University of Gottingen, Germany
Beat Bapst Beat.Bapst@qualitasag.ch; Qualitas, Switzerland
Valerie Wolf valerie.wolf@cel2590.fr; CEL25-90, France
Julie Leblois jleblois@awegroupe.be, Elévéo, Belgique



Wallonie recherche CRA-W

CONSEIL ÉLEVAGE 25-90

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

GE PLUS

UCD DUBLIN

ZUCHT DATA

elevéo by awé

64 herds N = 4,220 milk samples

Milk samples have been analyzed with 34 spectrometers,
in particular FT2, FT6000, FT+, and FT7 (Foss, Hillerød, Denmark),
and standard lactoscopes FT-MIR automatic (Delta Instruments, Drachten, the Netherlands).

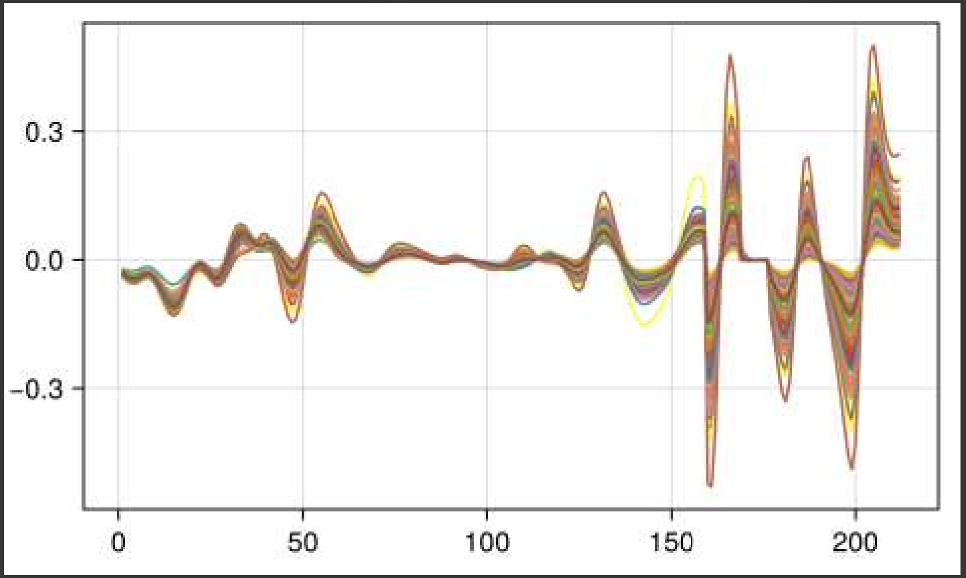
The MIR spectra from the different instruments were standardized to be merged into a common dataset

C. Grellet

The selected spectral area consisted of 212 wavelengths

- 968.1 – 1,577.5 cm^{-1}
- 1,731.8 – 1,762.6 cm^{-1}
- 1,781.9 – 1,808.9 cm^{-1}
- 2,831.0 – 2,966.0 cm^{-1}

to exclude areas not reproducible between instruments.



The spectra were pretreated by a first derivative

N = 4,220 samples

	BHB		
	Low	**High**	
Training (3/4)	2,939	256	48 herds
Test (1/4)	953	72	16 herds

⬇

~7-8%

➔ Unbalanced classes

Theoretical aspects

Usual PLSDA (simpler)

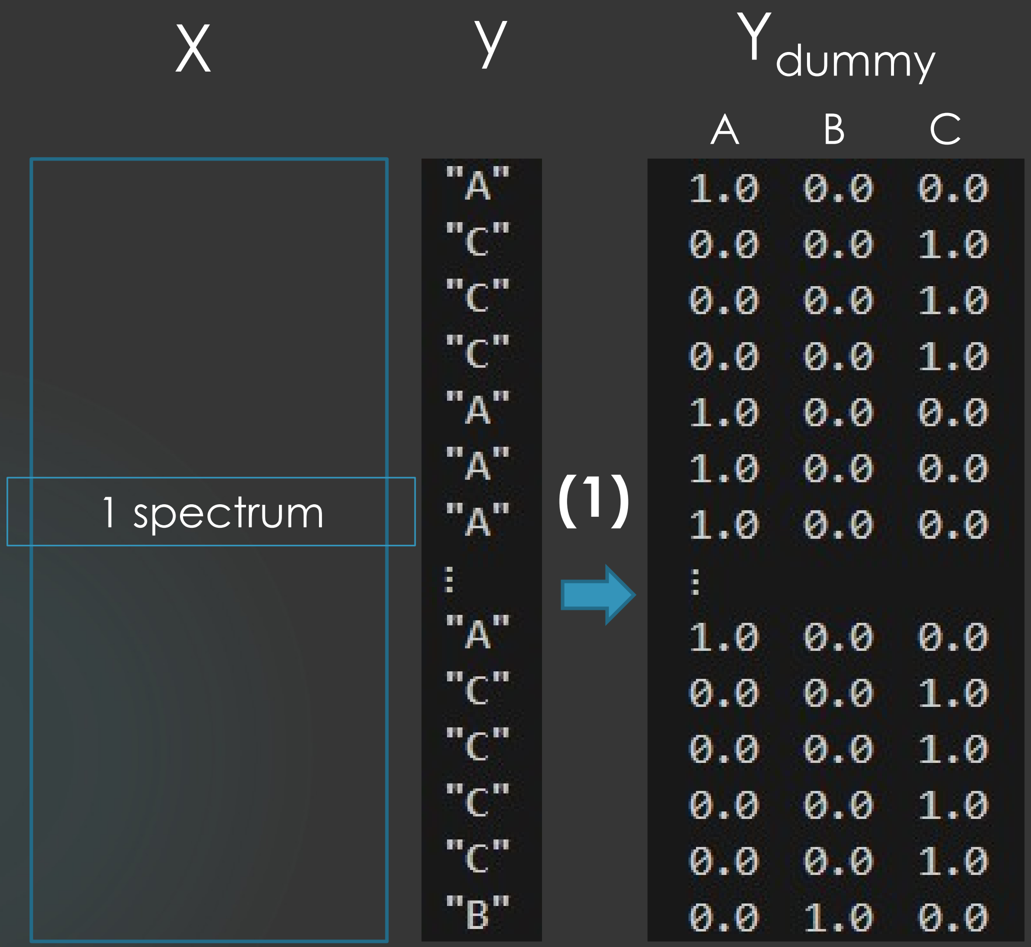
= PLSR-DA

= PLS-MLR-DA

Other methods not considered here

- PLS-LDA
- PLS-QDA
- PLS-KDEDA
- etc.

3 steps

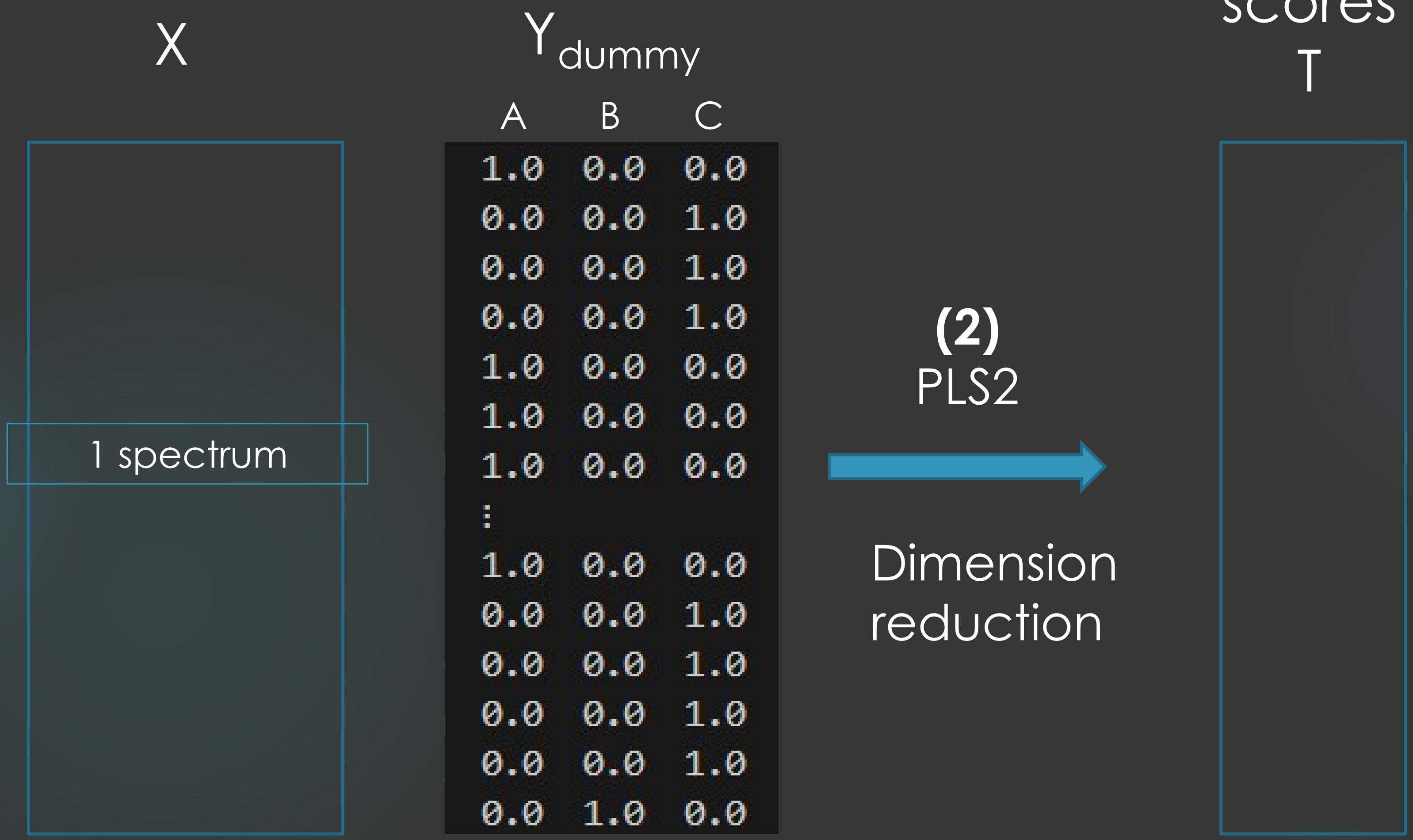


y

```
"A"
"C"
"C"
"C"
"A"
"A"
"A"
⋮
"A"
"C"
"C"
"C"
"C"
"B"
```

Y_{dummy}

```
 A    B    C
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
1.0  0.0  0.0
1.0  0.0  0.0
1.0  0.0  0.0
⋮
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  1.0  0.0
```



X

1 spectrum

Y_{dummy}

```
 A    B    C
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
1.0  0.0  0.0
1.0  0.0  0.0
1.0  0.0  0.0
⋮
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  1.0  0.0
```

(2)
PLS2

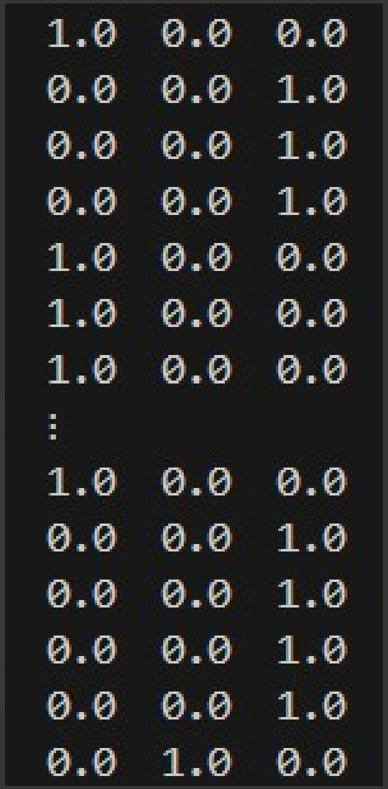
Dimension reduction

PLS scores T

PLS scores T

Y_{dummy}

A B C



```
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
1.0  0.0  0.0
1.0  0.0  0.0
1.0  0.0  0.0
⋮
1.0  0.0  0.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  0.0  1.0
0.0  1.0  0.0
```

(3) MLR → Regression model

$$\mathbf{(2)} + \mathbf{(3)} \quad = \quad \text{PLSR2} \ \{\ X\ ,\ Y_{\text{dummy}}\ \}$$

Prediction

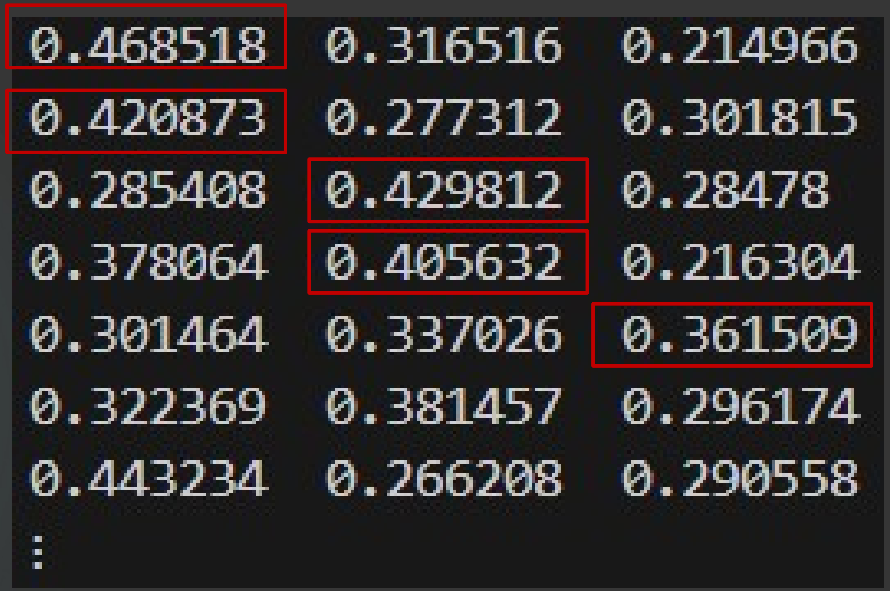
$\hat{Y}_{\text{dummy, new}}$

$\hat{y}_{\text{new}}$

X_{new}

Fitted PLSR2 model {X, Ydummy }

A B C



```
0.468518  0.316516  0.214966
0.420873  0.277312  0.301815
0.285408  0.429812  0.28478
0.378064  0.405632  0.216304
0.301464  0.337026  0.361509
0.322369  0.381457  0.296174
0.443234  0.266208  0.290558
⋮
```

A
A
B
B
C
…

- **(2) :** PLS2 $\{ X , Y_{\text{dummy}} \}$ ➔ T
- **(3) :** MLR $\{ T , Y_{\text{dummy}} \}$ ➔ $\hat{y}$

Bias if classes unbalanced

Dominant class is favored in the predictions

Simple approach to decrease the bias

- Weighting the PLS

	Usual PLSR	Weighted PLSR
Means	$1'X$	$1' \mathbf{D} X$
Covariances	$T'y$	$T' \mathbf{D} y$
MLR $\hat{\beta}$	$(T'T)^{-1}T'y$	$(T' \mathbf{D} T)^{-1} T \mathbf{D}' y$

$$\underset{n \times n}{\mathbf{D}} = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

such as $\underbrace{\Sigma_{\mathrm{i}} w_{i,\,A}}_{\text{weight class A}} = \underbrace{\Sigma_{\mathrm{i}} w_{i,\,B}}_{\text{weight class B}} = \underbrace{\Sigma_{\mathrm{i}} w_{i,\,C}}_{\text{weight class C}} = 1/3$

BHB dataset

Replicated K-Fold CV
K = 3, nrep = 50

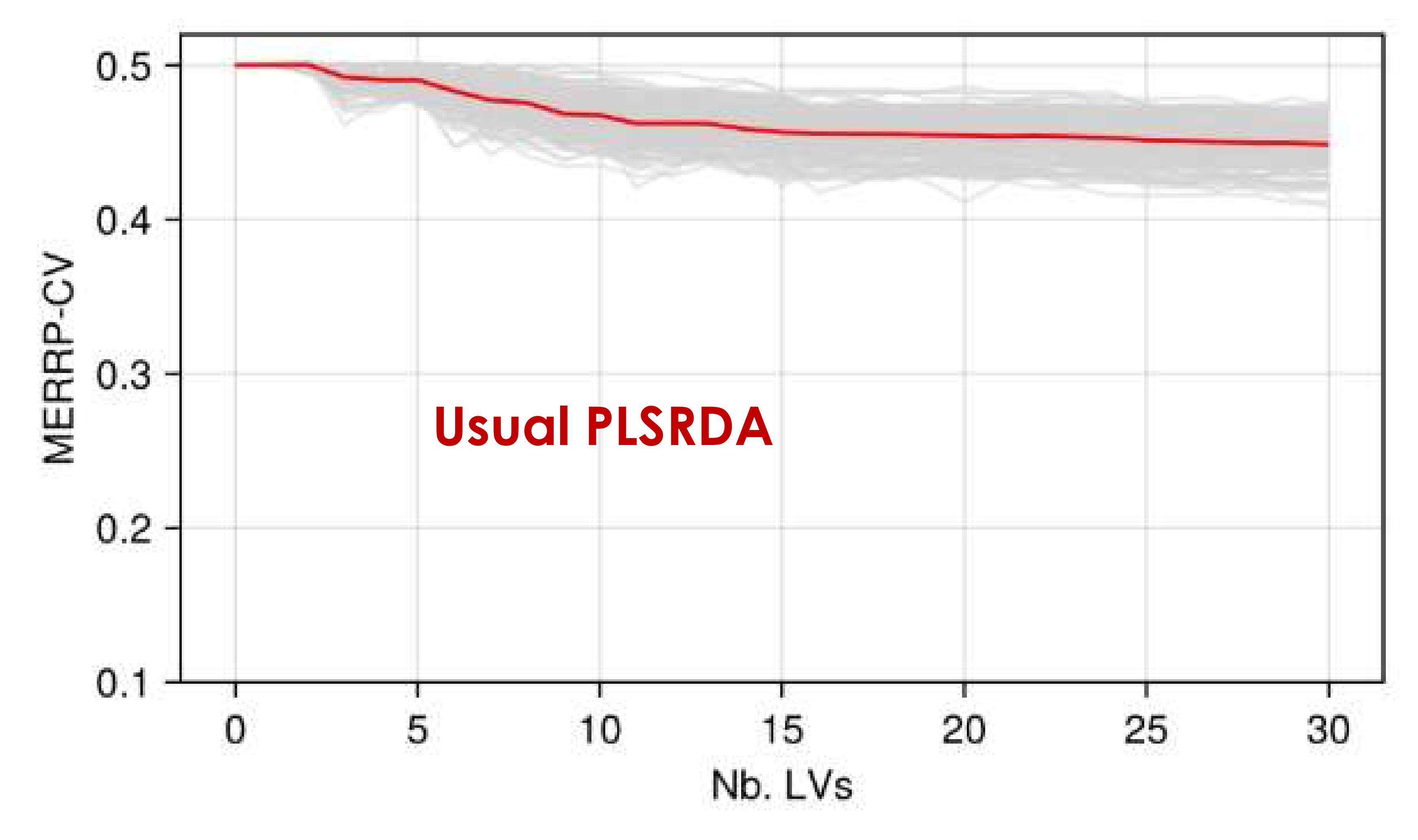
Generalization error

	BHB	
	Low	**High**
Training (3/4)	2,939	256
Test (1/4)	953	72

When **validation / test sets** are unbalanced

$$Mean\ ERRP = \frac{\sum_{i=1}^{G} ERR_i}{G}$$

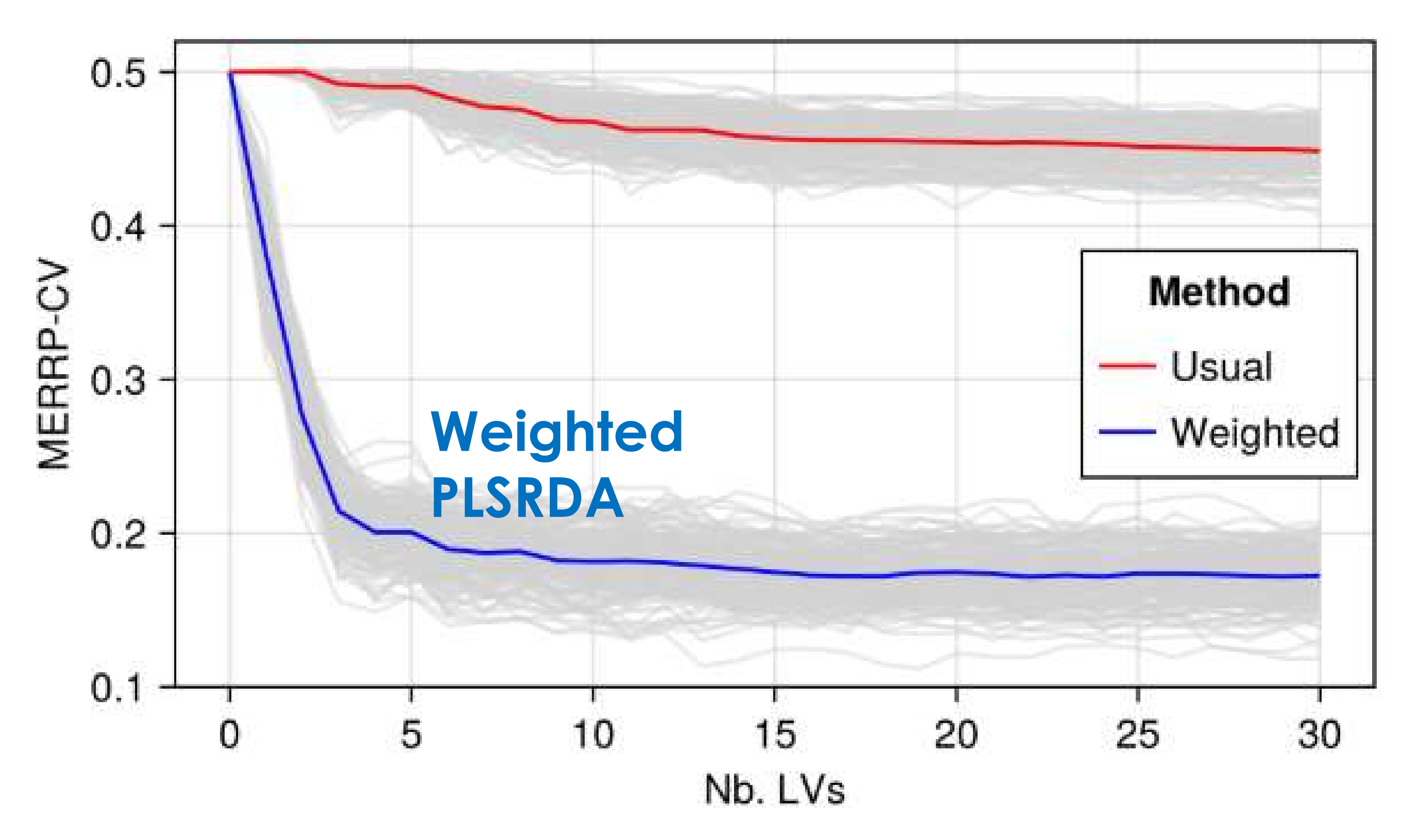
Nb. classes


MERRP-CV
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
Nb. LVs
Usual PLSRDA

Test set results
Usual PLSRDA with 15 LVs

	Predictions		
Observed	Low	High	
Low	100	0	Row %
High	92	8	Row %

Mean ERRP = 46%



Weighted PLSRDA

Test set results
Weighted PLSDA with 15 LVs

Predictions

	Low	High	
Observed			
Low	87	13	Row %
High	18	82	Row %

Mean ERRP = 16%

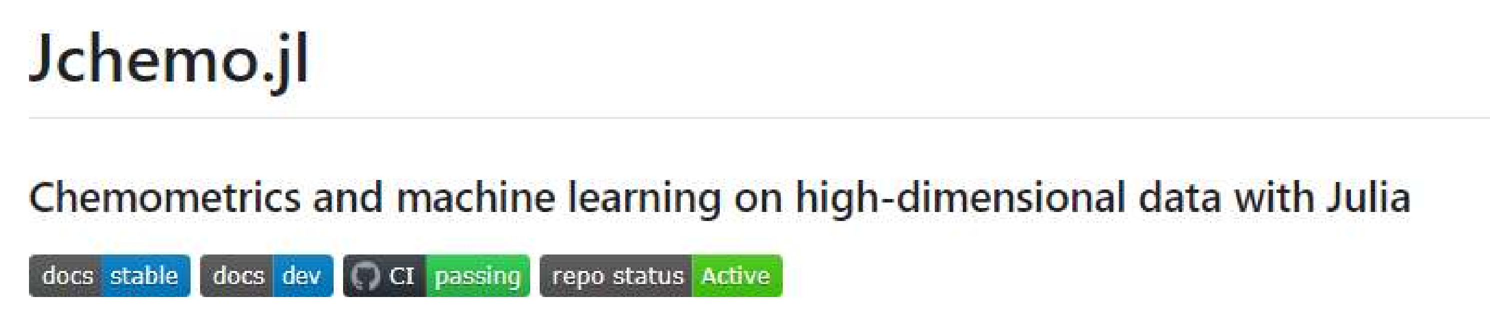
Conclusions

- PLSRDA: simple, very fast (can manage very large DB)
  - But only recommended when few classes (2-3)
  - And highly biased when unbalanced classes

- Easy solutions to remove the bias
  - Weighting: very performant (should be the default)
  - Sub-sampling the classes to balance the training (but loss of information)

- Other PLSDA methods (e.g. probabilistic) less sensitive but can also be weighted

- Weighted PLSRDA in practice, see next slides



Jchemo.jl

Chemometrics and machine learning on high-dimensional data with Julia

```
model = plsrda(nlv = 20, prior = :unif)
```

Weighting

```
fit!(model, Xtrain, ytrain)
pred = predict(model, Xtest).pred
```

Also possible with package **rchemo**

Brandolini-Bunlon M. et al.

https://cran.r-project.org/web/packages/rchemo

(But: a weight vector has to be specified manually before the model fitting)

Thank you!

Hélio SPIR

cirad

SELMET

ChemProject