

PLSR-DA

with unbalanced data

matthieu.lesnoff@cirad.fr
Heliospir, Montpellier, 24-25 June 2025



ChemProject



Clément Grelet

CRA-W Gembloux

Quality and authentication of agricultural products Unit



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

- Biomarker of ketosis in milking cows

BHB concentration in blood

3



Usual
Blood sampling

2 classes

Low $\leq 1.2 \text{ mmol/L}$

High $> 1.2 \text{ mmol/L} \Rightarrow \text{Disease}$

Possible alternative?

MIR spectrometry
on milk samples

European BHB consortium

4

Mark Crowe mark.crowe@ucd.ie; UCD, Ireland

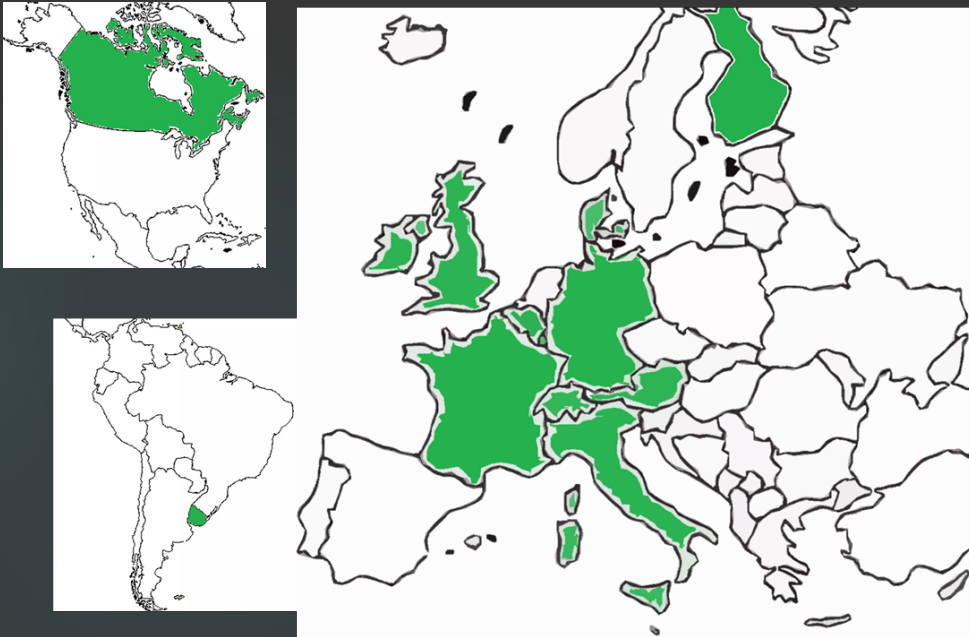
Astrid Koeck koeck@zuchtdata.at; ZuchtData, Austria

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



64 herds N = 4,220 milk samples (MIR)

5

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

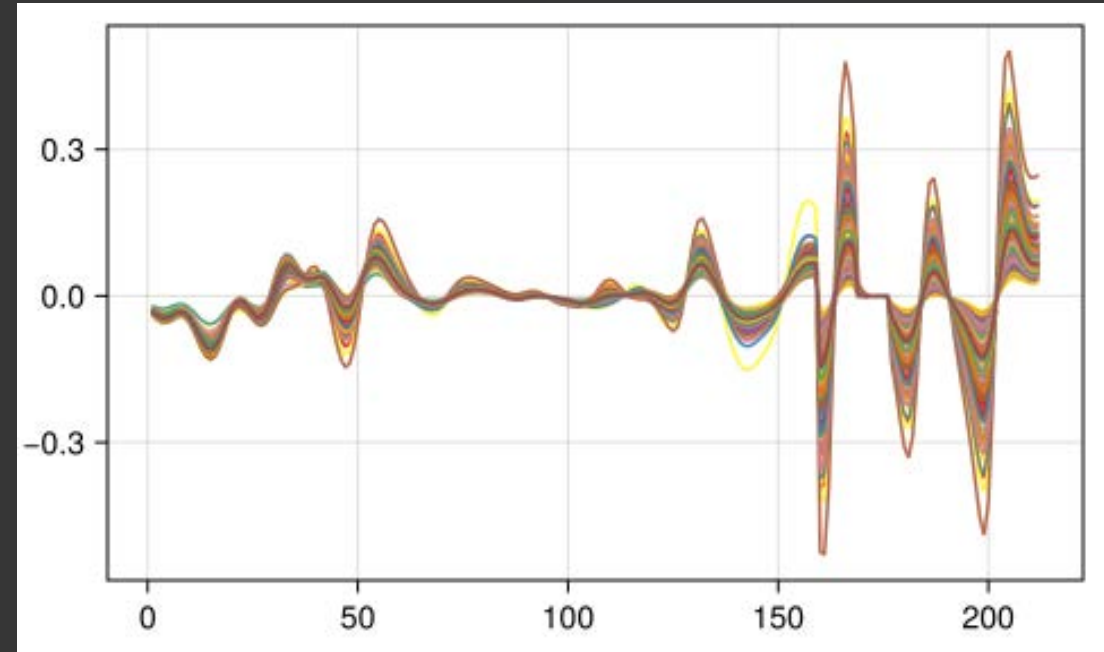
C. Grellet

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

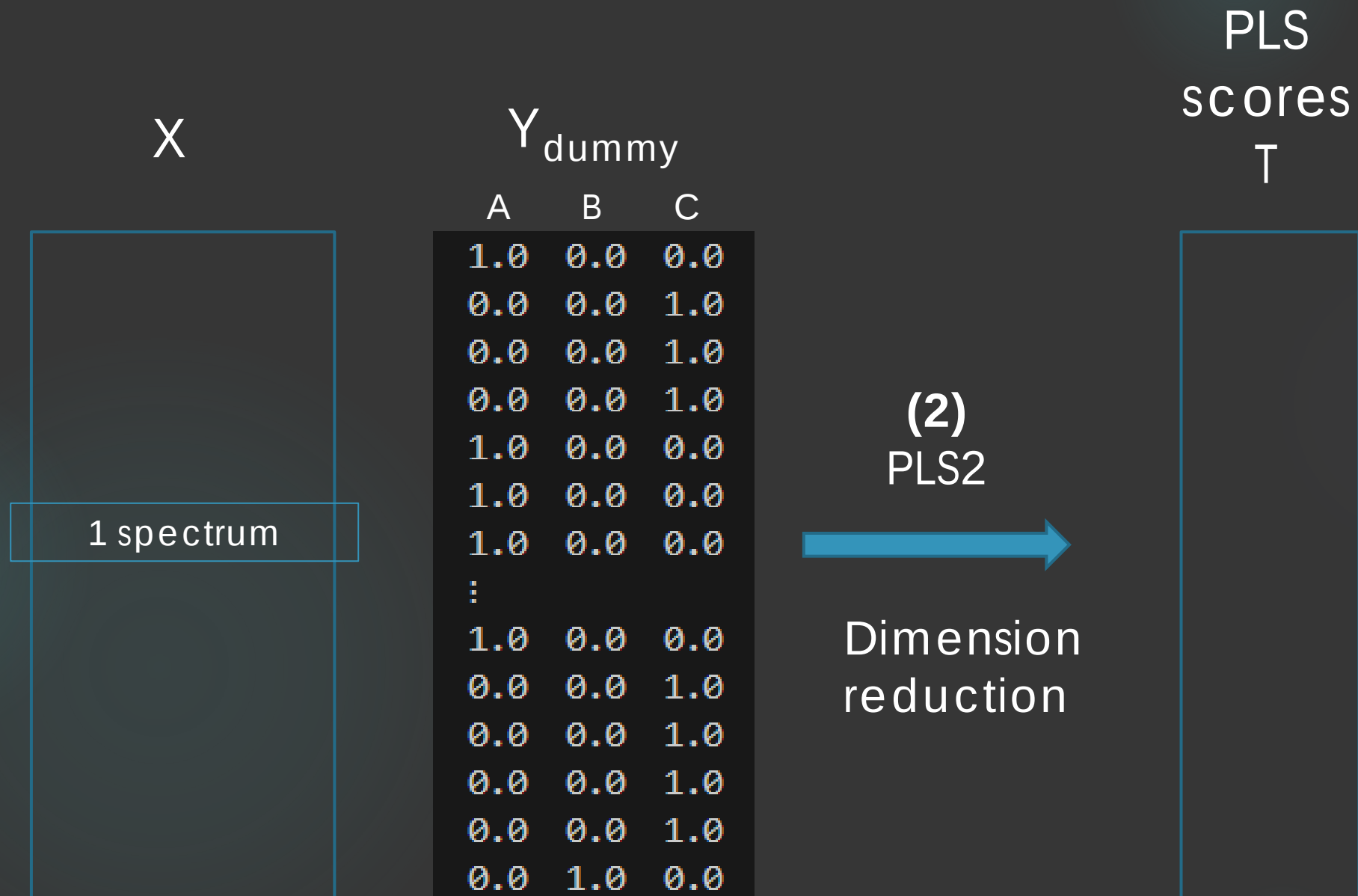
Other methods not considered here

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

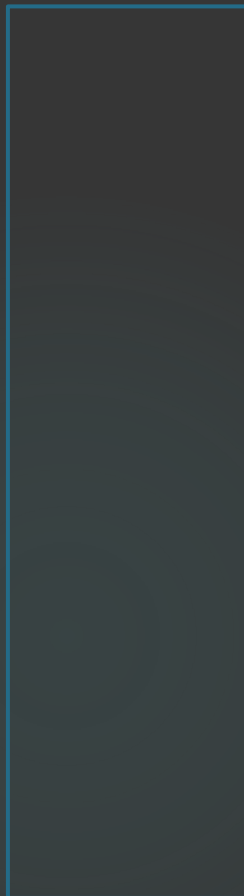
X	y	Y _{dummy}		
		A	B	C
1 spectrum	"A"	1.0	0.0	0.0
	"C"	0.0	0.0	1.0
	"C"	0.0	0.0	1.0
	"C"	0.0	0.0	1.0
	"A"	1.0	0.0	0.0
	"A"	1.0	0.0	0.0
	"A"	1.0	0.0	0.0
	⋮	⋮		
	"A"	1.0	0.0	0.0
	"C"	0.0	0.0	1.0
	"C"	0.0	0.0	1.0
	"C"	0.0	0.0	1.0
	"C"	0.0	0.0	1.0
	"B"	0.0	1.0	0.0

(1)





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

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

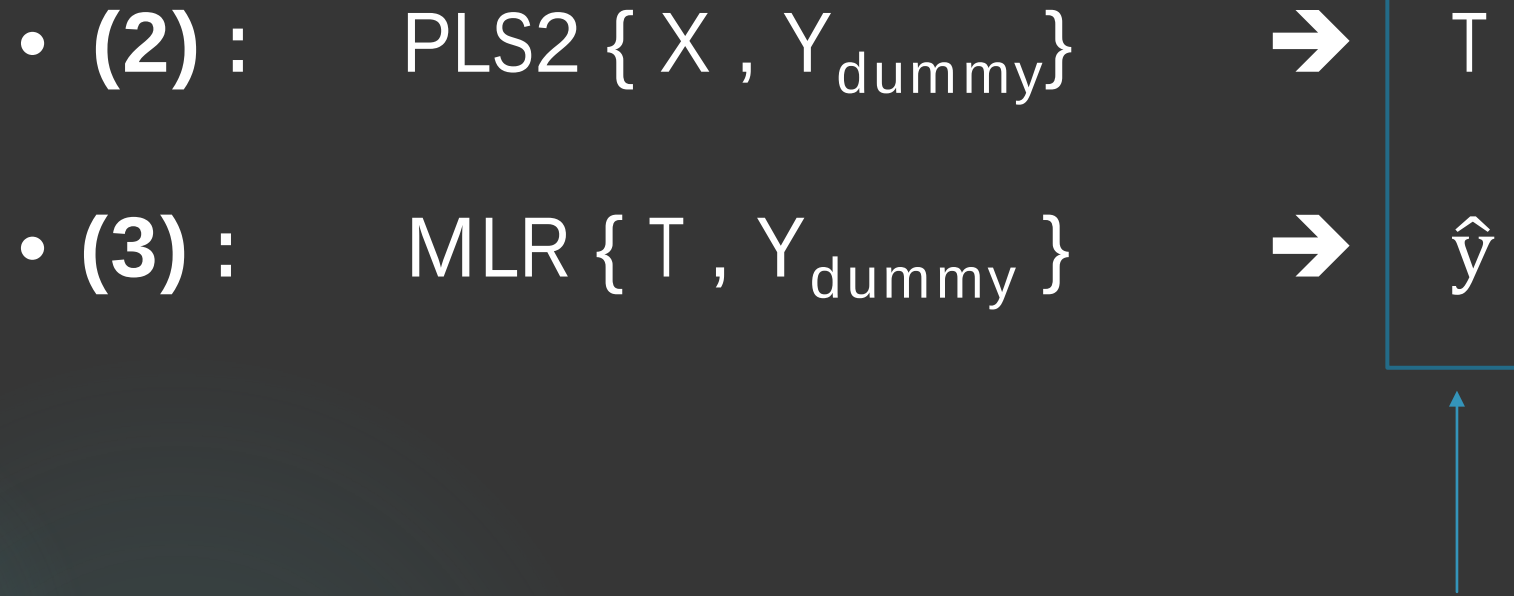
Prediction

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



Bias if classes unbalanced

Dominant class is favored in the predictions

Simple approach to decrease the bias

- **Weighting** the PLS

Usual PLSRWeighted PLSR

Means

$$\mathbf{1}'\mathbf{X}$$

$$\mathbf{1}' \mathbf{D} \mathbf{X}$$

Covariances

$$\mathbf{T}'\mathbf{y}$$

$$\mathbf{T}' \mathbf{D} \mathbf{y}$$

MLR $\hat{\beta}$

$$(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

$$(\mathbf{T}' \mathbf{D} \mathbf{T})^{-1}\mathbf{T}' \mathbf{D}' \mathbf{y}$$

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

$n \times n$

such as $\sum_i w_{i,A} = \sum_i w_{i,B} = \sum_i w_{i,C} = 1/3$



weight
class A



weight
class B



weight
class C

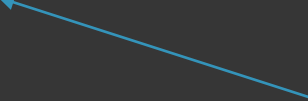
BHB dataset

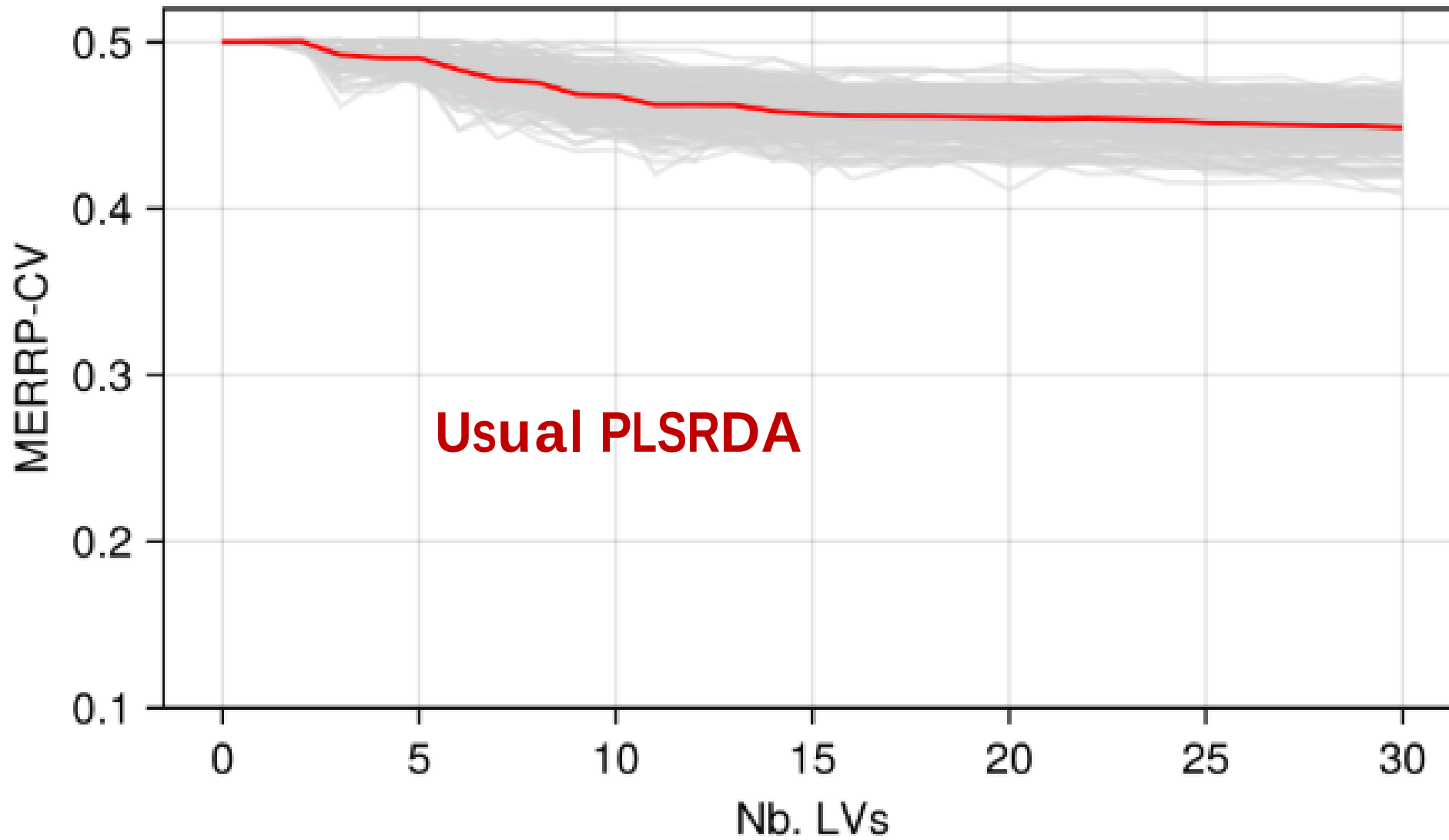
		BHB	
		Low	High
Replicated K-Fold CV K = 3, nrep = 50	→ Training (3/4)	2,939	256
Generalization error	→ Test (1/4)	953	72

When **validation / test sets** are unbalanced

$$\text{Mean ERRP} = \frac{\sum_{i=1}^G \text{ERR}_i}{G}$$

Nb. classes



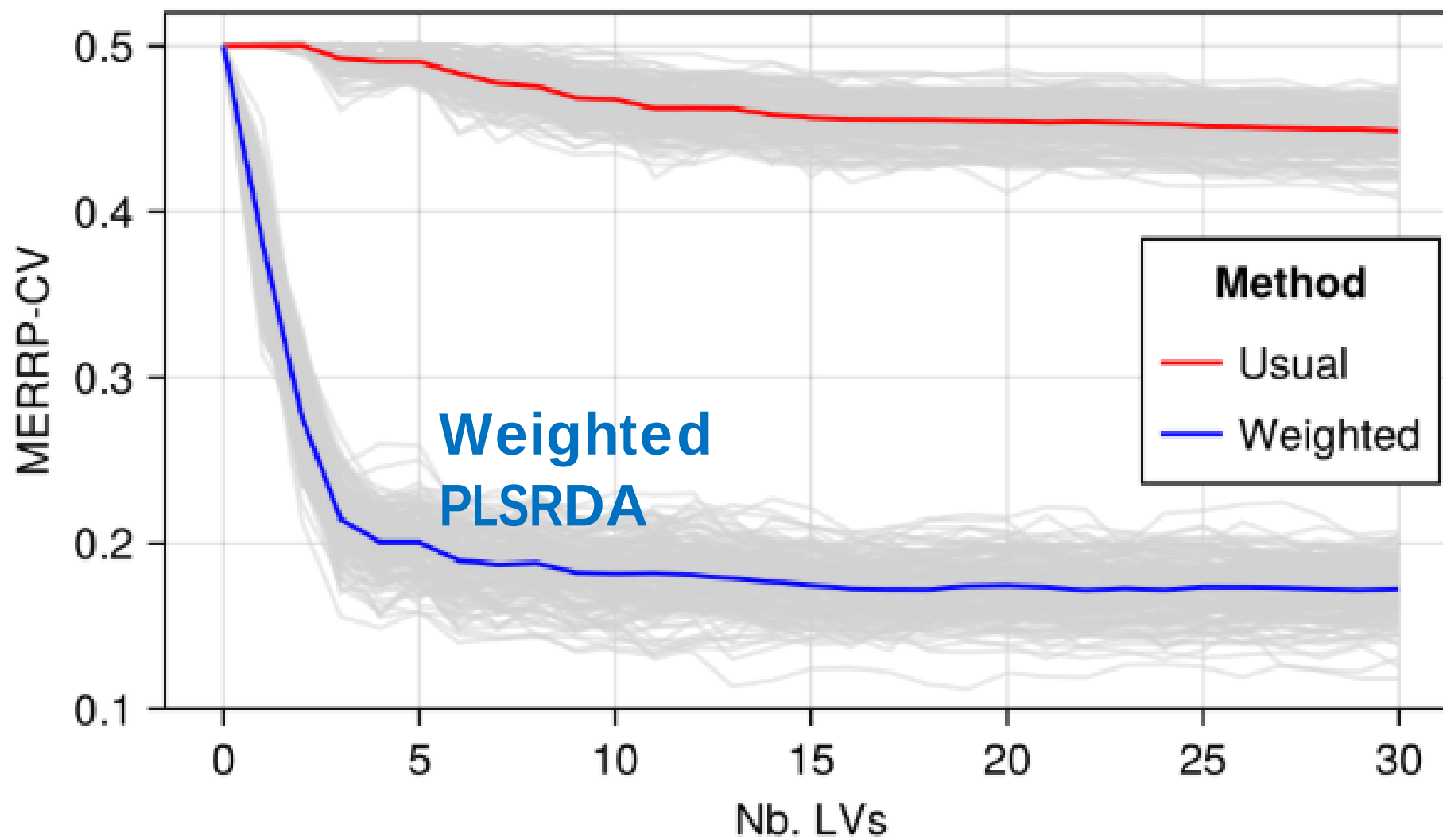


Test set results

Usual PLSRDA with 15 LVs

<u>Observed</u>	<u>Predictions</u>		
	Low	High	
Low	100	0	Row %
High	92	8	Row %

Mean ERRP = 46%



Test set results

Weighted PLSDA with 15 LVs

<u>Predictions</u>			
	Low	High	
<u>Observed</u>			
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

docs stable docs dev CI passing repo status Active



24

```
model = plsrd(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!

