PLSR-DA with unbalanced data

Illustration of the problems and two simple solutions

matthieu.lesnoff@cirad.fr
Cirad, UMR Selmet

https://github.com/mlesnoff/Jchemo.jl

PLSDA many methods

- PLSR-DA = PLS-MLR-DA = “usual” PLSDA
- Probabilistic $\Rightarrow$ PLS-L/Q/KDE-DA
- Etc.

This presentation $\Rightarrow$ only on this method

PLSR-DA

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

Step 1

PLS2

$\{X\ ,\ Y_{\text{dummy}}\}$

Scores T_X

Step 2

Multivariate

MLR

$\{T_X\ ,\ Y_{\text{dummy}}\}$

$\widehat{Y}_{\text{dummy}}$

PLS2-R

4

$\hat{Y}_{dummy}$

Membership probability estimates

not bounded in [0, 1]

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
⋮		
0.413478	0.245917	0.340604
0.343138	0.317523	0.339339
0.316082	0.348387	0.335531
0.29393	0.348038	0.358033
0.345968	0.386519	0.267513
0.294213	0.290373	0.415413

Highest prediction → $\hat{y}$

Case of unbalanced data

Iris data

- X = 4 quantitative variables

  sepal length/width, petal length/width

- y = categorical variable with 3 classes

  setosa, versicolor, virginica



N_{tot} = 150 observations

Training

Balanced

```
"setosa"     => 30
"versicolor" => 30
"virginica"  => 30
```

Unbalanced

```
"setosa"     => 30
"versicolor" => 30
"virginica"  =>  4
```

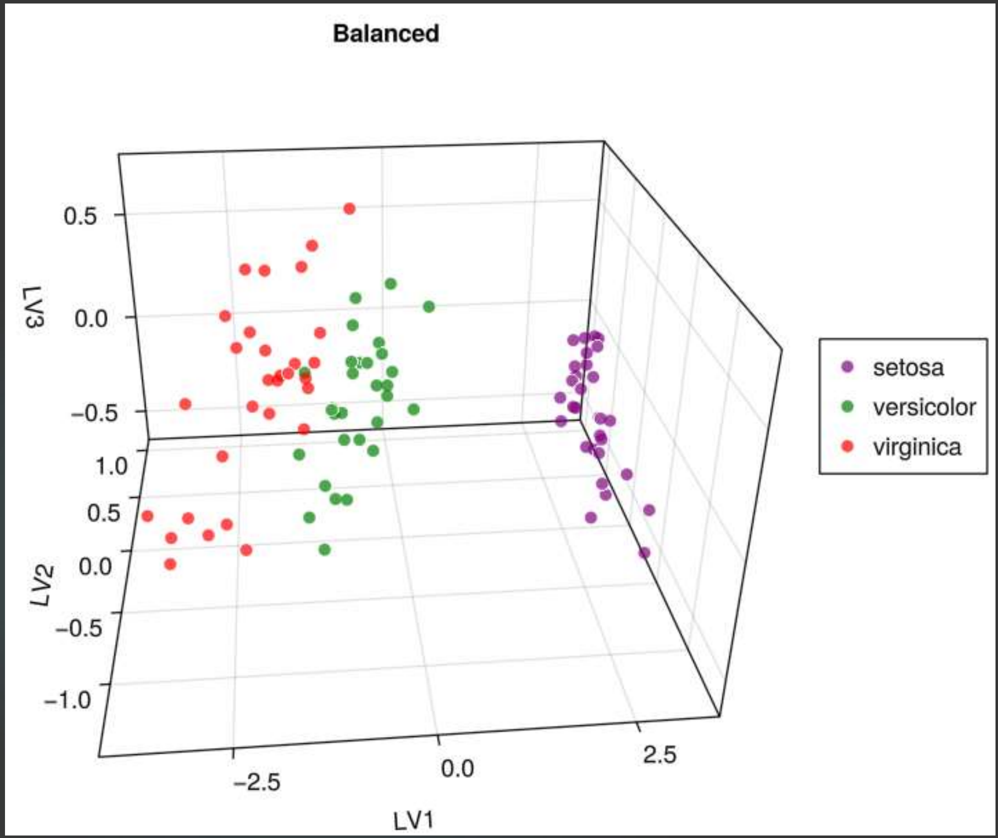
Class more difficult to predict

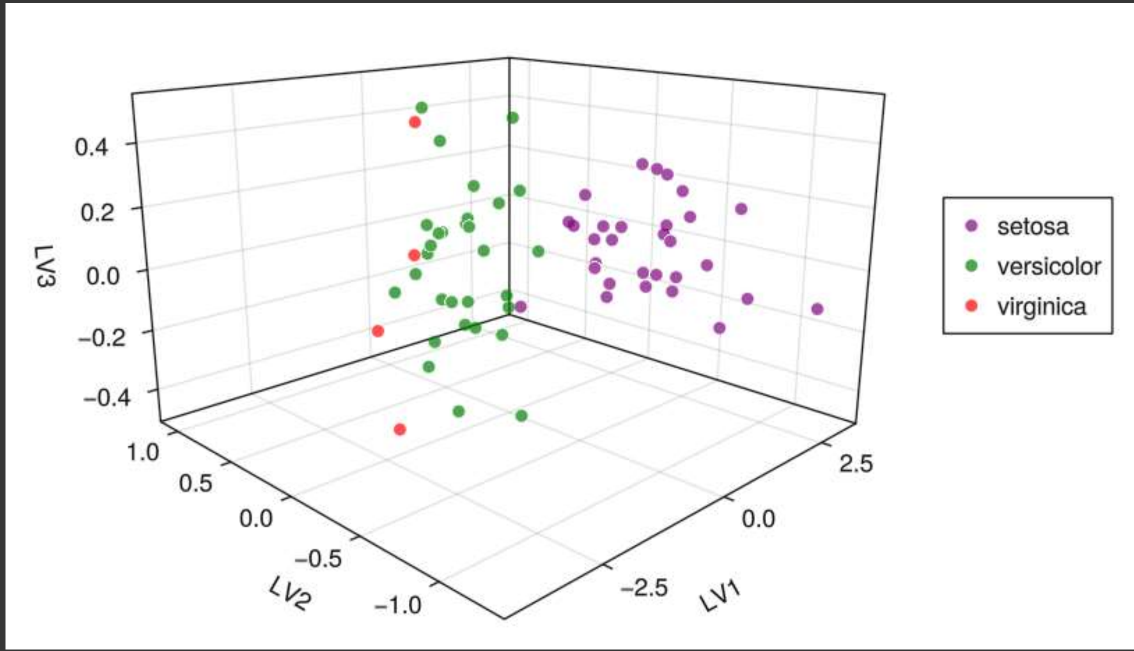
Two problems in PLSR-DA

Test

```
"setosa"     => 20
"versicolor" => 20
"virginica"  => 20
```

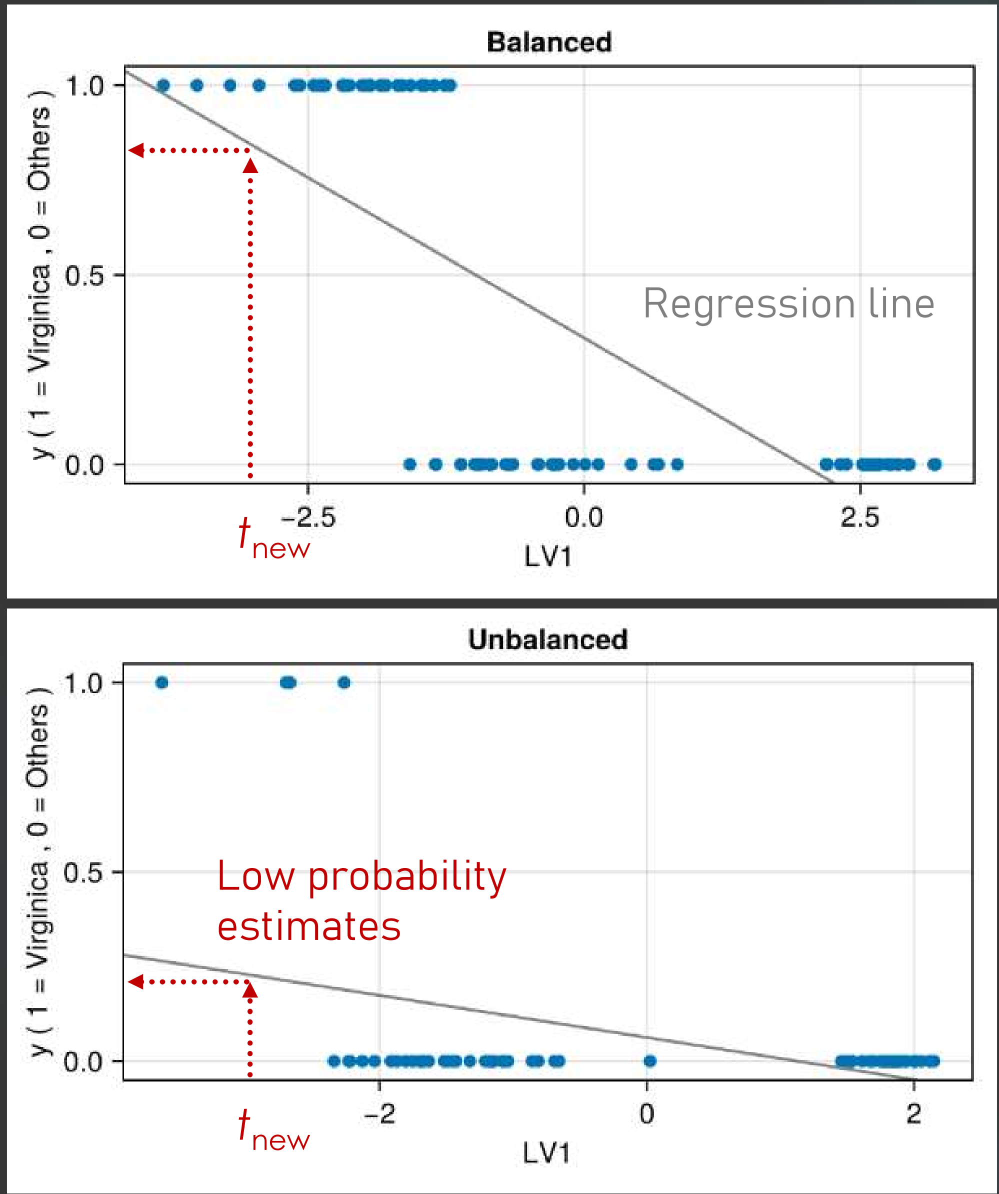
(1) PLS2 $\{X, Y_{\mathrm{dummy}}\}$





$\Rightarrow$ Masking effects

(2) MLR regression



⇒ Small class Almost never chosen as prediction

PLSR-DA Unbalanced training (N_i = 30, 30, 4)

Results on Test

y String	pred_setosa Int64	pred_versicolor Int64	pred_virginica Int64
setosa	20	0	0
versicolor	0	20	0
virginica	0	20	$\boxed{0}$

But solutions exists!

ex : WPLS

y String	pred_setosa Int64	pred_versicolor Int64	pred_virginica Int64
setosa	20	0	0
versicolor	0	16	4
virginica	0	3	$\boxed{17}$

Two simple alternatives for correction

1. Sub-sampling the training within the large classes
   - $N_A \sim N_B \gg N_C \Rightarrow n_A \sim n_B \sim N_C$

2. Weighting the classes in the PLS2-R $\Rightarrow$ WPLSR
   - $w_A = w_B = w_C = 1/3$

Illustration

Congress Chimiometrie2024, Nantes – Challenge data

- *X* = VIS-NIR spectra

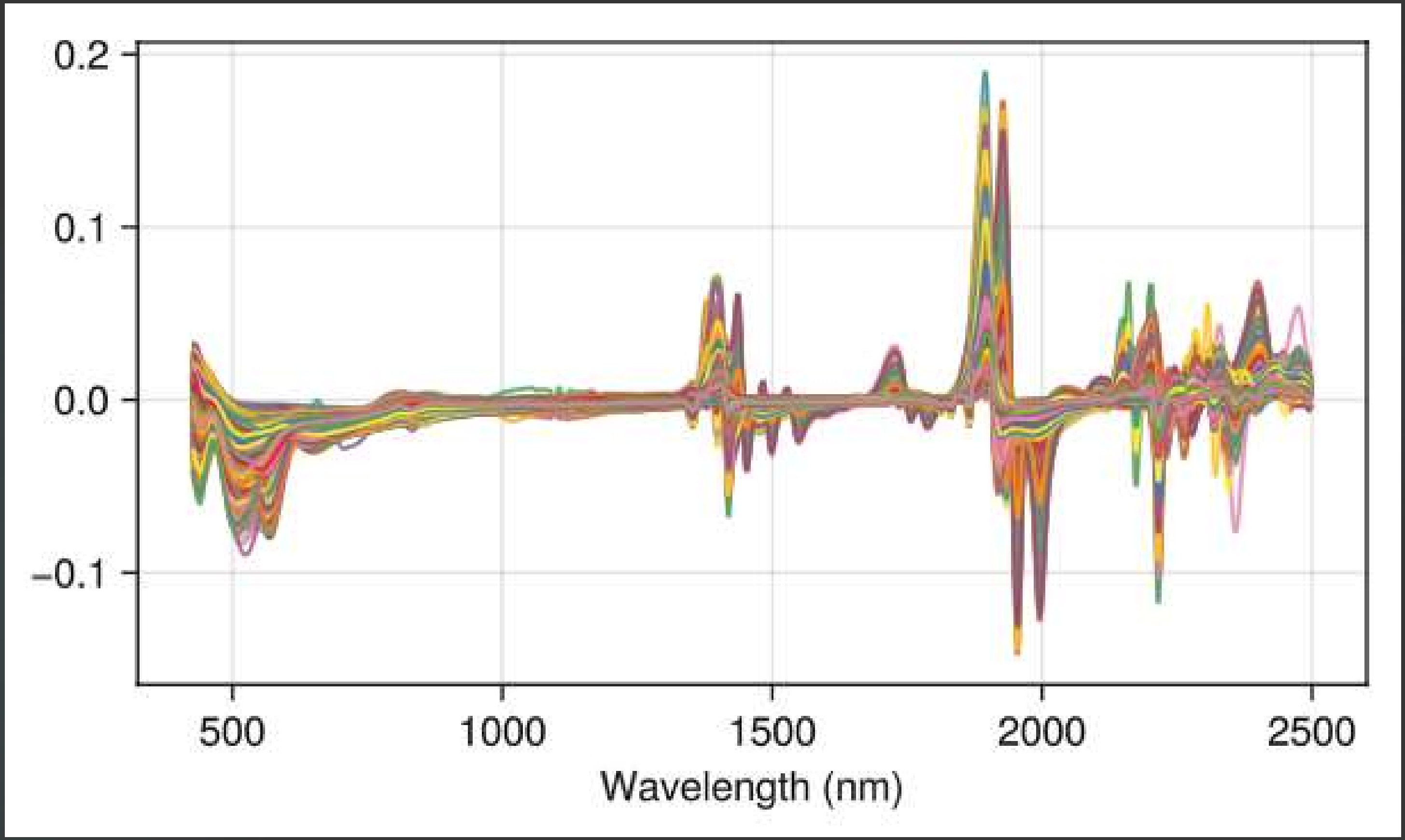
  Soils (Belgium)   FOSS 428-2498 nm (step 2 nm)

- *y* = categorical variable with 2 classes

  mineral >> organic

- *N* = 19,036 observations

SavGol deriv2 + SNV



Data for this illustration (sub-sampling in N)

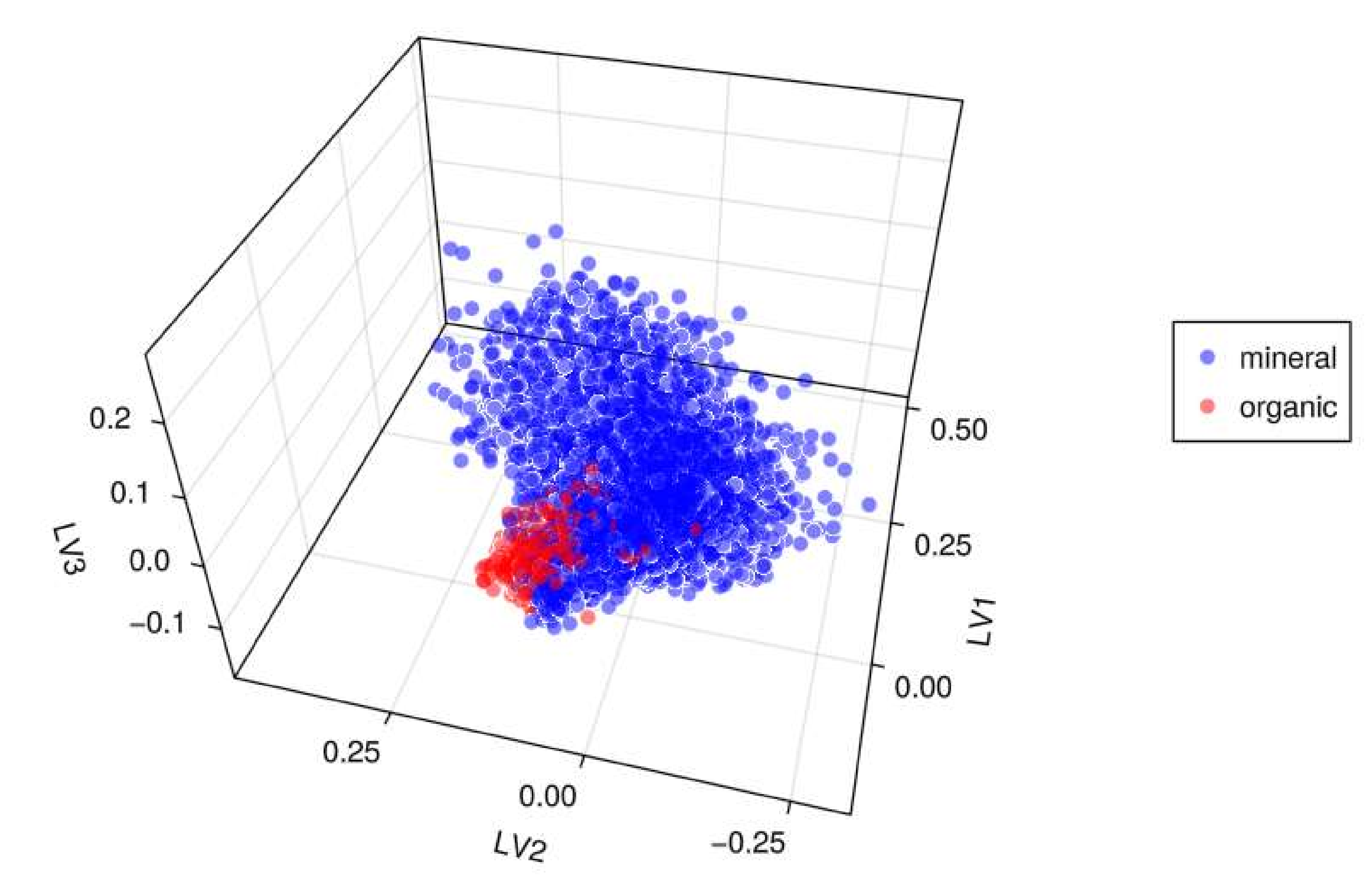
Training

"mineral"	=>	**17000**
"organic"	=>	**1000**

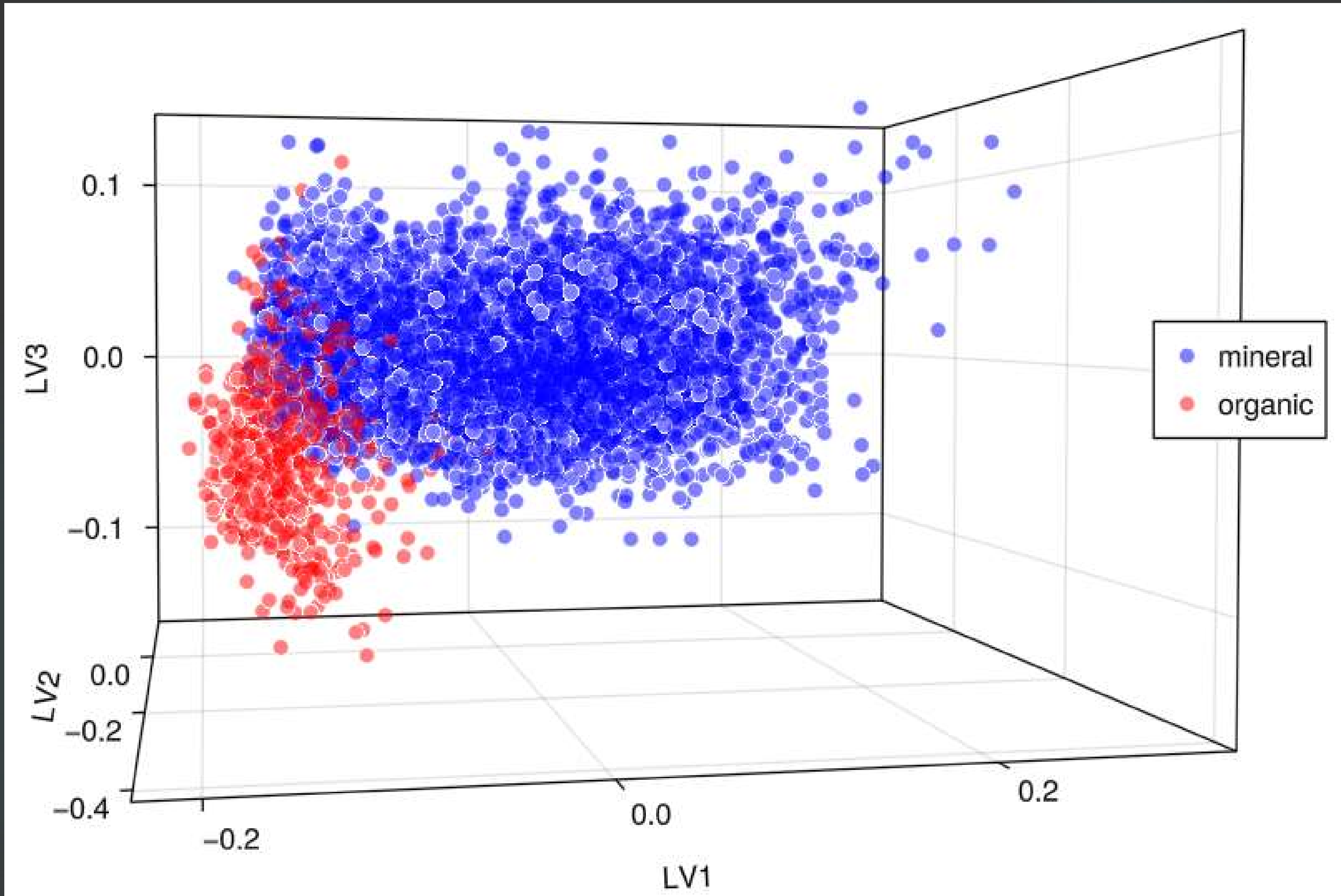
Test

"mineral"	=>	**200**
"organic"	=>	**200**

Naïve PCA



Naïve PLS



Prediction of Test $N_{test} = 400$

- PLSR-DA 20 LVs

Results on Test

Naïve approach

Nb.

y String	pred_mineral Int64	pred_organic Int64
mineral	200	0
organic	89	111

Row %s

levels String	pred_mineral Float64	pred_organic Float64
mineral	100.0	$\boxed{0.0}$
organic	$\boxed{44.5}$	55.5

Too optimistic

Bad prediction

Too optimistic

ERR = .22

Mean ERR = .22

Alternative 1 Subsampling within mineral

$\Rightarrow$ New Training $n_{\text{mineral}} = N_{\text{organic}} = 1000$

Nb.

y String	pred_mineral Int64	pred_organic Int64
mineral	184	16
organic	12	188

Row %s

levels String	pred_mineral Float64	pred_organic Float64
mineral	92.0	8.0
organic	6.0	94.0

ERR = .07

Mean ERR = .07

Alternative 2 WPLSR-DA

Nb.

y String	pred_mineral Int64	pred_organic Int64
mineral	185	15
organic	14	186

Row %s

levels String	pred_mineral Float64	pred_organic Float64
mineral	92.5	7.5
organic	7.0	93.0

ERR = .075

Mean ERR = .075

More extreme case

Training

```
"mineral"      =>  17000
"organic"      =>    100
```

Test

```
"mineral"      =>    200
"organic"      =>    200
```

Error rates by class

	Naive	1) Subsampling	2) WPLSR-DA
mineral	.00	.09	.07
organic	.98	.08	.07

Jchemo.jl

https://github.com/mlesnoff/Jchemo.jl

Chemometrics and machine learning on high-dimensional data with Julia

docs stable | docs dev | CI passing | repo status Active

```
model = plsrda(nlv = 20)     # WPLSRDA (default: prior = :unif)

# Naive model
# plsrda(nlv = 20, prior = :prop)

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

errp(pred, ytest)          # global ERRP
merrp(pred, ytest)         # mean ERRP
conf(pred, ytest)          # confusion
```