Challenge 2018

matthieu.lesnoff@cirad.fr

Joint Research Unit
Mediterranean and tropical livestock systems
Montpellier

Chimiométrie 2018, 30-31 janvier 2018, Cnam, Paris

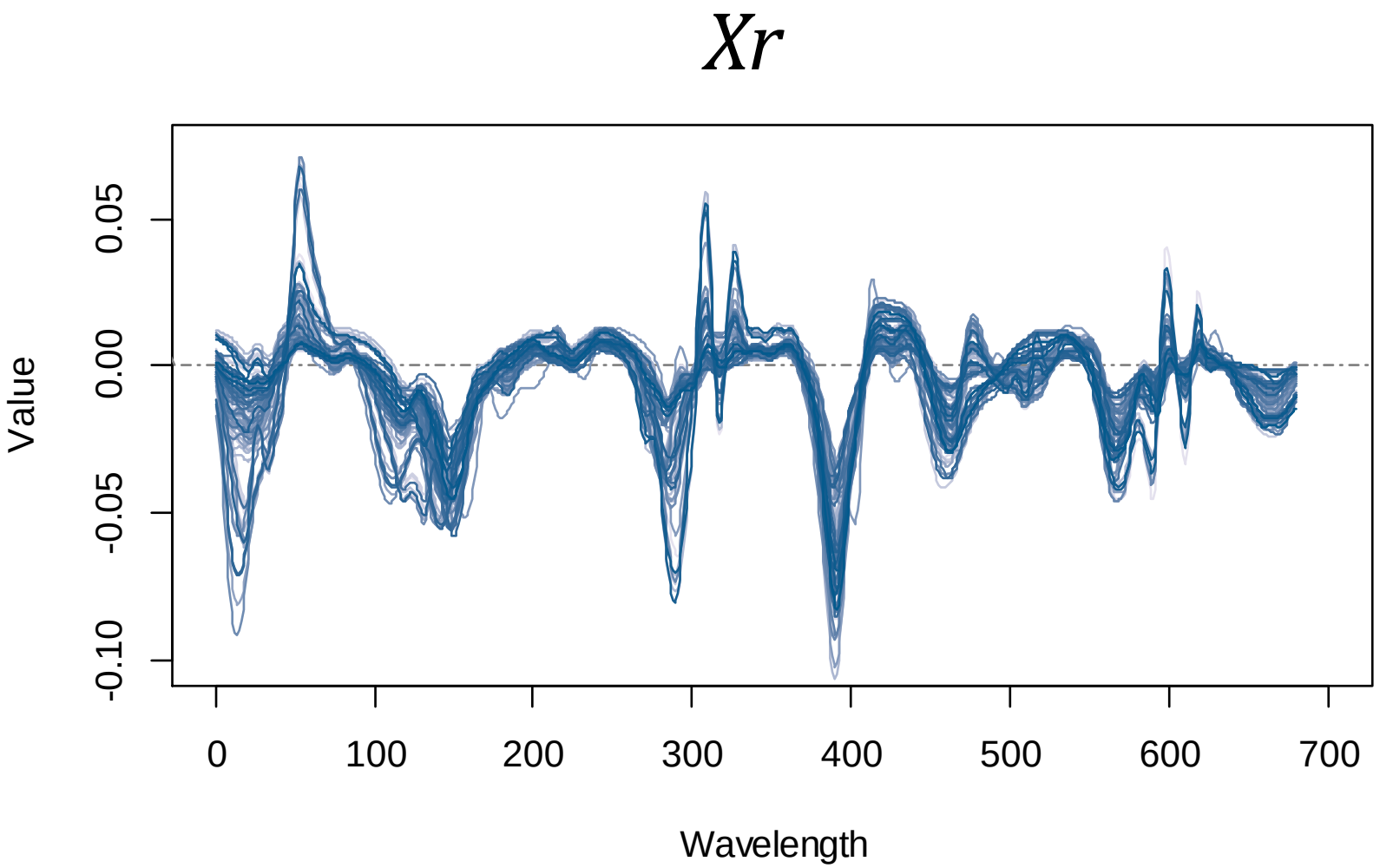
General approach

- **Data exploration**
- **Modelling & predictions**
- ***Pre-processing (1rst derivative + SNV)***

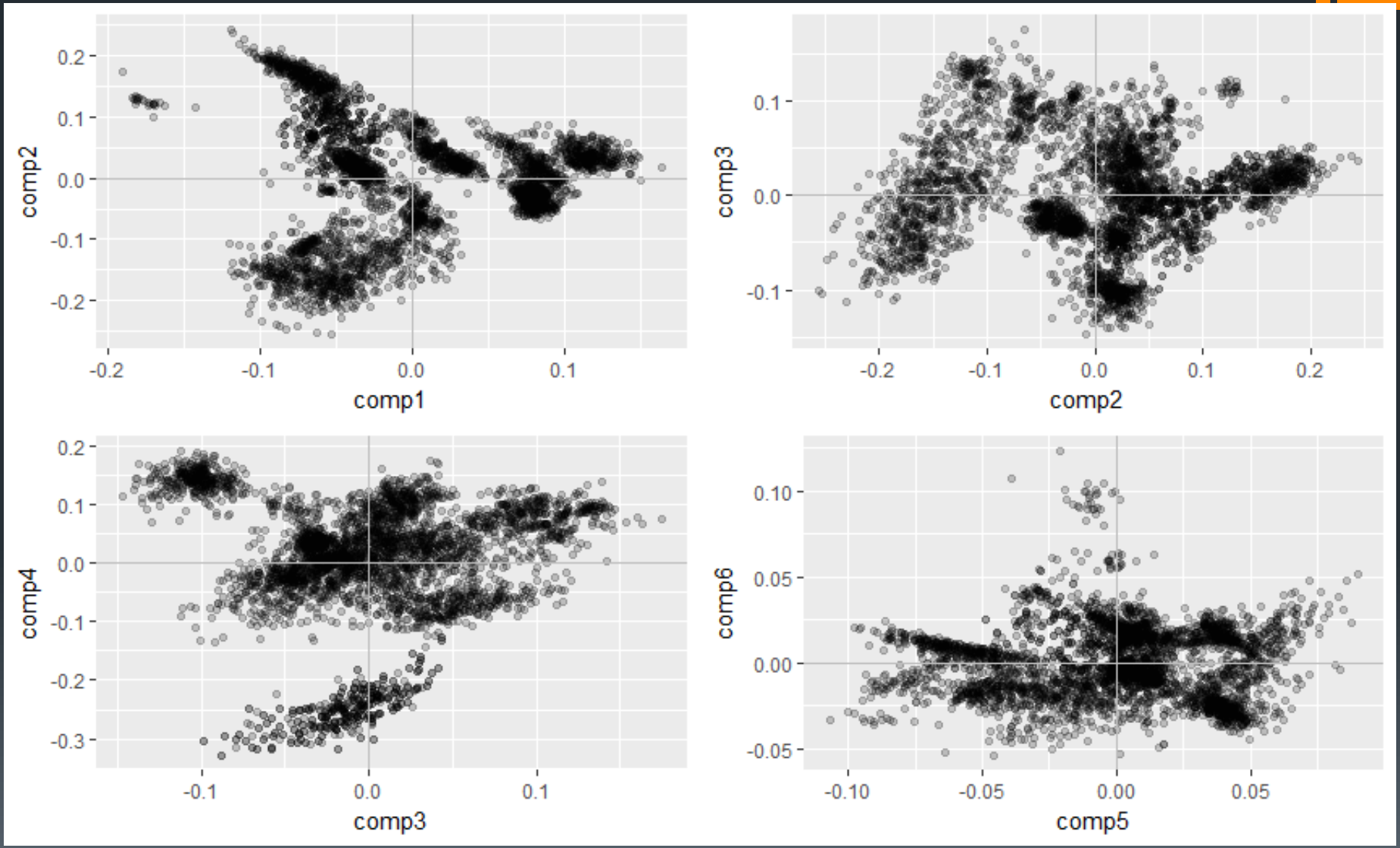
Notations

- **CAL set (Reference, Training)** Xr yr
- **VAL set (Test)** Xu

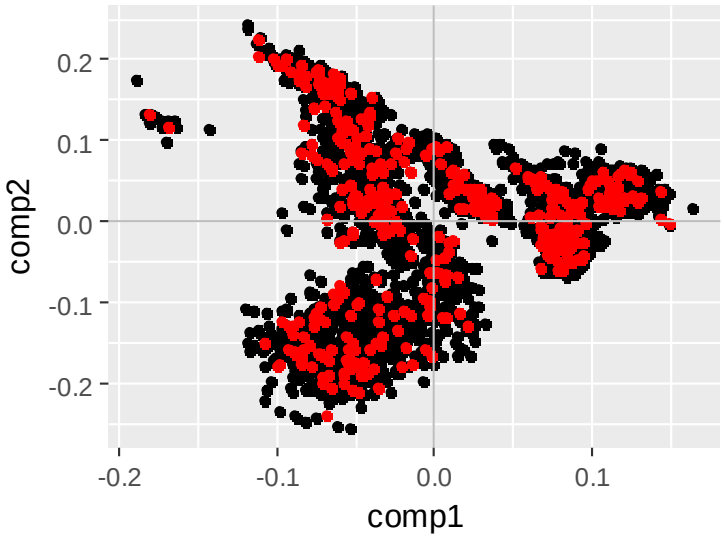
  yu Unknown

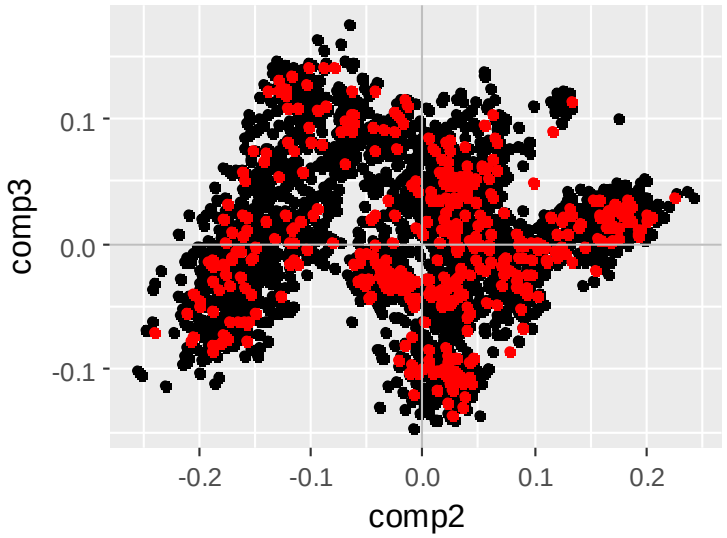

Xr
Value
0.05
0.00
-0.05
-0.10
0
100
200
300
400
500
600
700
Wavelength

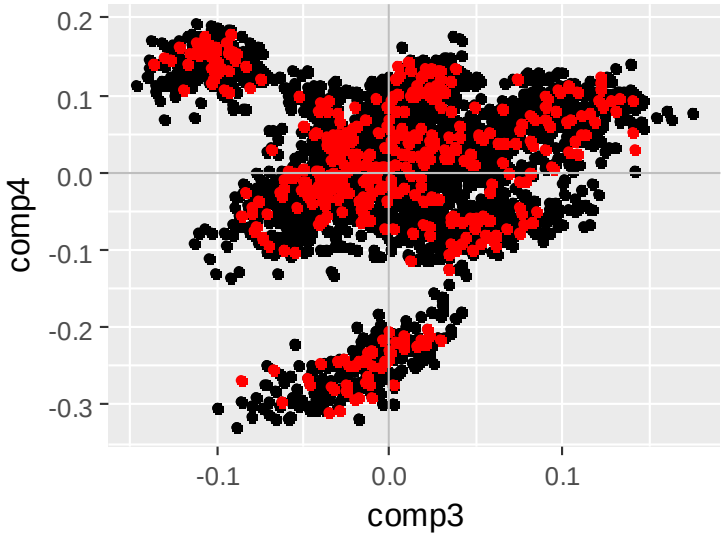
Global PLSR *Xr yr*

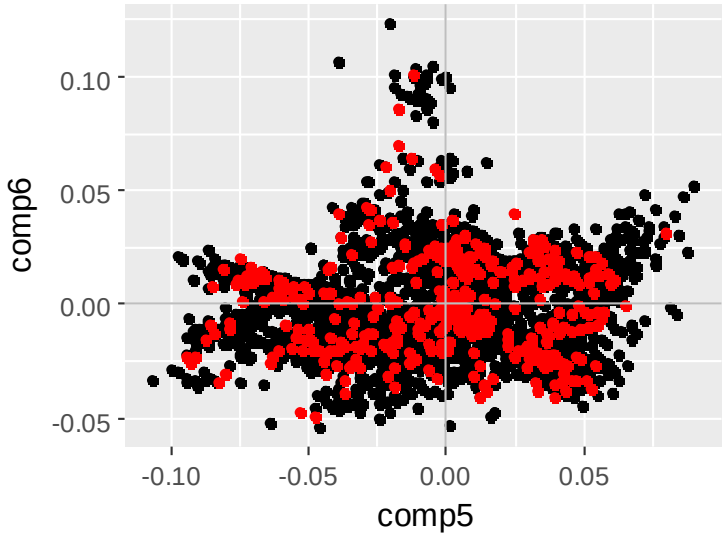


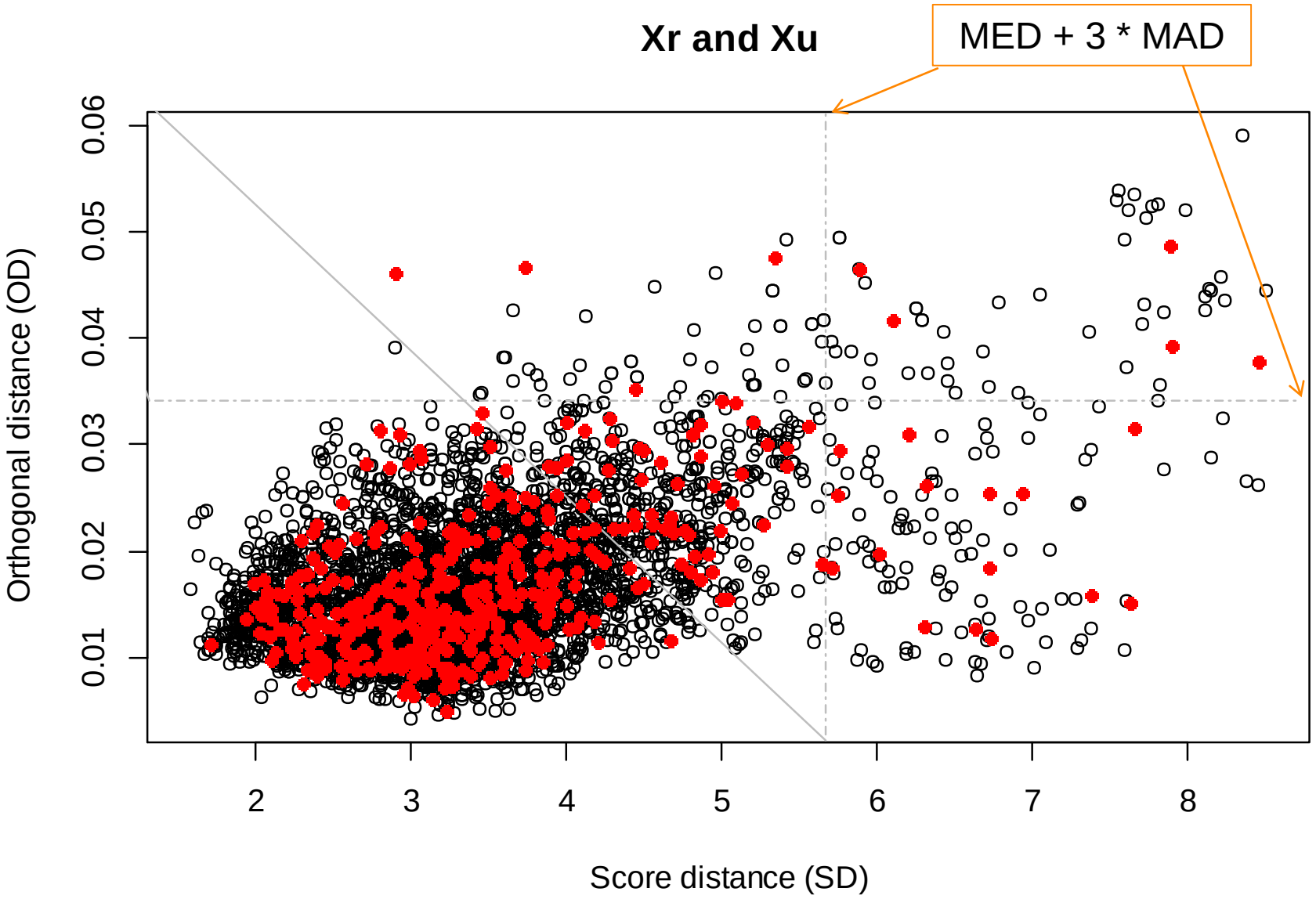
+ Projection of the Test set Xu



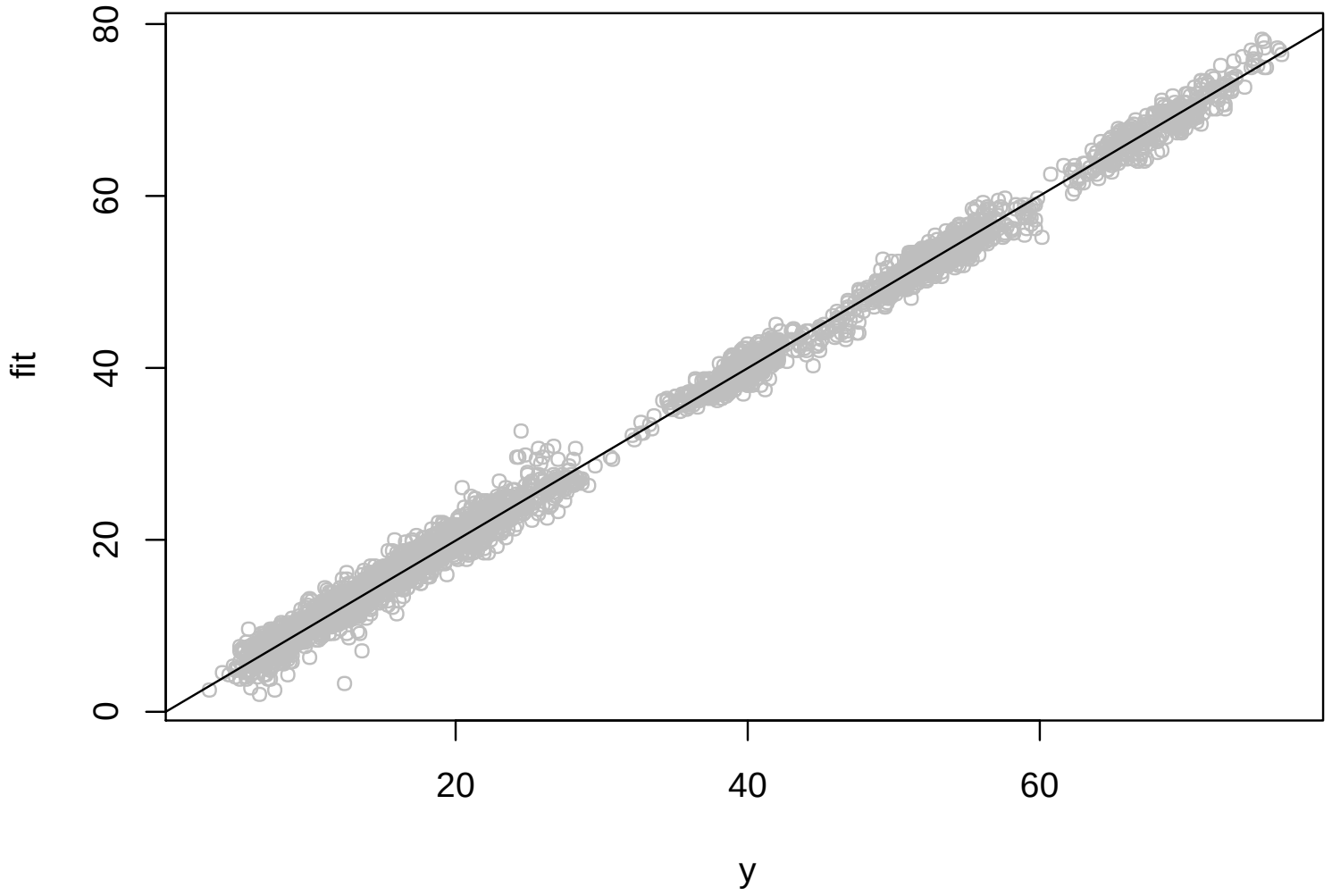






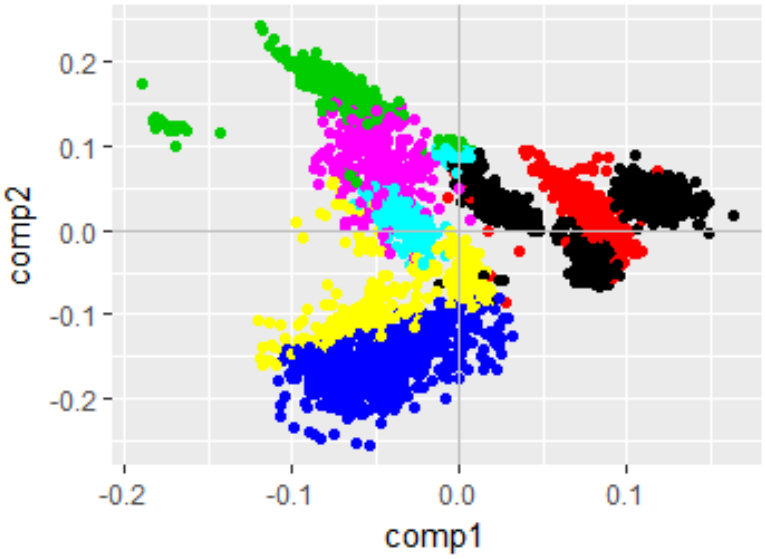

Xr and Xu
MED + 3 * MAD
Orthogonal distance (OD)
0.01
0.02
0.03
0.04
0.05
0.06
2
3
4
5
6
7
8
Score distance (SD)

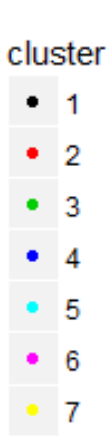
K-Fold cross validation Global PLSR Xr yr

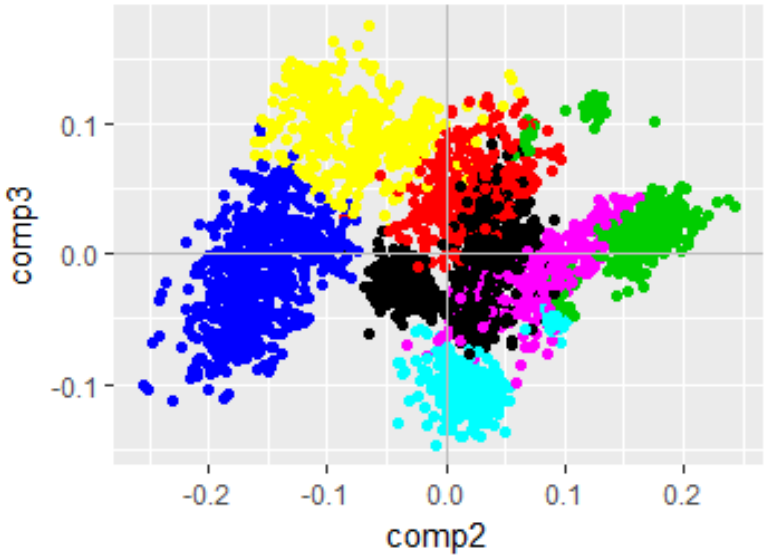


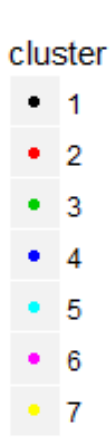
Unsupervised clustering of the Reference set

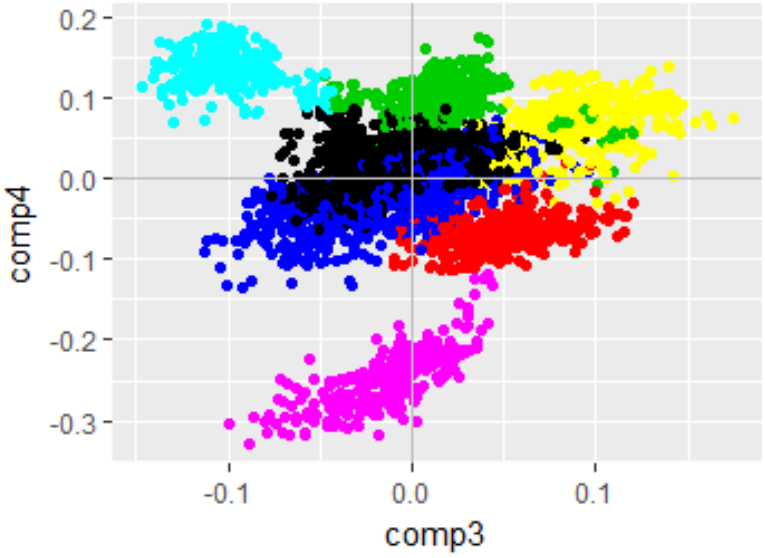
K-means on global PLSR scores Tr

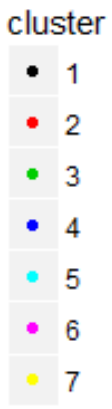


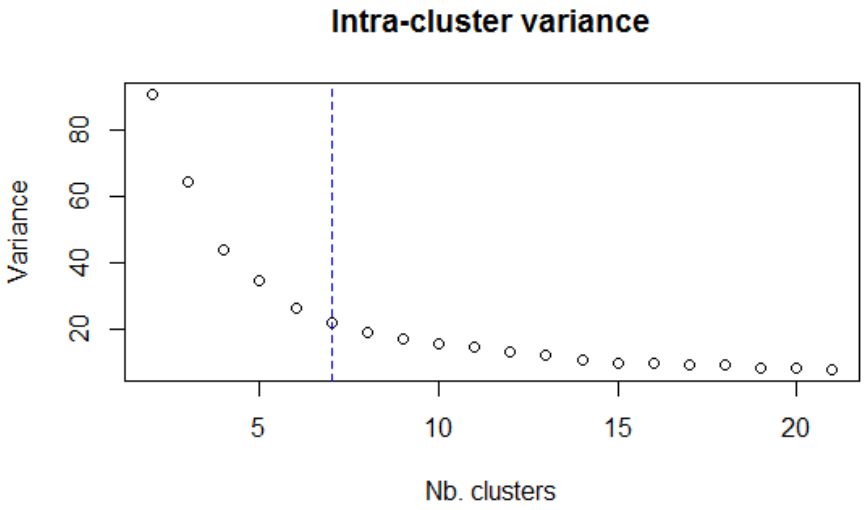












Proportions in the Reference

cluster	nb.r	p.r
1	637	0.163
2	372	0.095
3	452	0.116
4	395	0.101
5	437	0.112
6	440	0.113
7	1175	**0.301**

cluster	nb.r	Proportions in the Reference p.r	nb.u	Proportions in the Test p.u
1	637	0.163	69	0.161
2	372	0.095	42	0.098
3	452	0.116	50	0.117
4	395	0.101	41	0.096
5	437	0.112	50	0.117
6	440	0.113	49	0.114
7	1175	**0.301**	128	**0.298**

Three prediction approaches

1. LPLSR

A local PLSR model (e.g. Centner and Massart 1998)

- For each spectrum xu:
  - neighborhood $Xr[u]$
    (Mahalanobis D calculated after PLSR data compression)
  - Fitting of a PLSR model on $Xr[u]$ $yr[u]$ ➔ prediction yu

2. LPLSR-S

Same as above but Xr , Xu are replaced by global scores $\boldsymbol{Tr}$, $\boldsymbol{Tu}$ (calculated by a preliminary global PLSR)

- Faster

3. simcaPLSR

A "simca" PLSR model (e.g. Vinzi et al, 2005)

- PLSR data compression of Xr , Xu ➔ score matrices Tr , Tu
- Unsupervised clustering (K-means) of Tr ➔ G clusters
- Calibrating a PLSR model in each cluster g : $Xr[g]$, $yr[g]$
- Assignment of tu to the closest cluster g_u (Euclidean distance to the center) ➔ xu assigned to cluster g_u
- Prediction yu by the PLSR model fitted over $Xr[g_u]$, $Xr[g_u]$

a) Ranking of the three approaches

➔ Naive random K-Fold cross-validation on Xr yr K = 5

model	RMSE.CV
LPLSR	**0.713**
LPLSR-S	0.787
simcaPLSR	0.814

*(GLOBAL PLSR 15 comp. RMSE.CV = **1.60***
*30 comp. RMSE.CV = **1.22)***

a) Ranking of the three approaches

➔ Naive random K-Fold cross-validation on Xr yr

model	RMSE.CV	Running.time.sec
LPLSR	**0.713**	383
LPLSR-S	0.787	162
simcaPLSR	0.814	**22**

*(GLOBAL PLSR 15 comp. RMSE.CV = **1.60***
*30 comp. RMSE.CV = **1.22)***

b) Finer calibration of the LPLSR model

Find the best combination (in average) “nb. neighbors” × “nb. local components”

Monte Carlo cross-validation within Xr yr with the objective to “mimic” as possible the data set VAL

Within-cluster sampling representative of the proportions in VAL data set

Random replications of

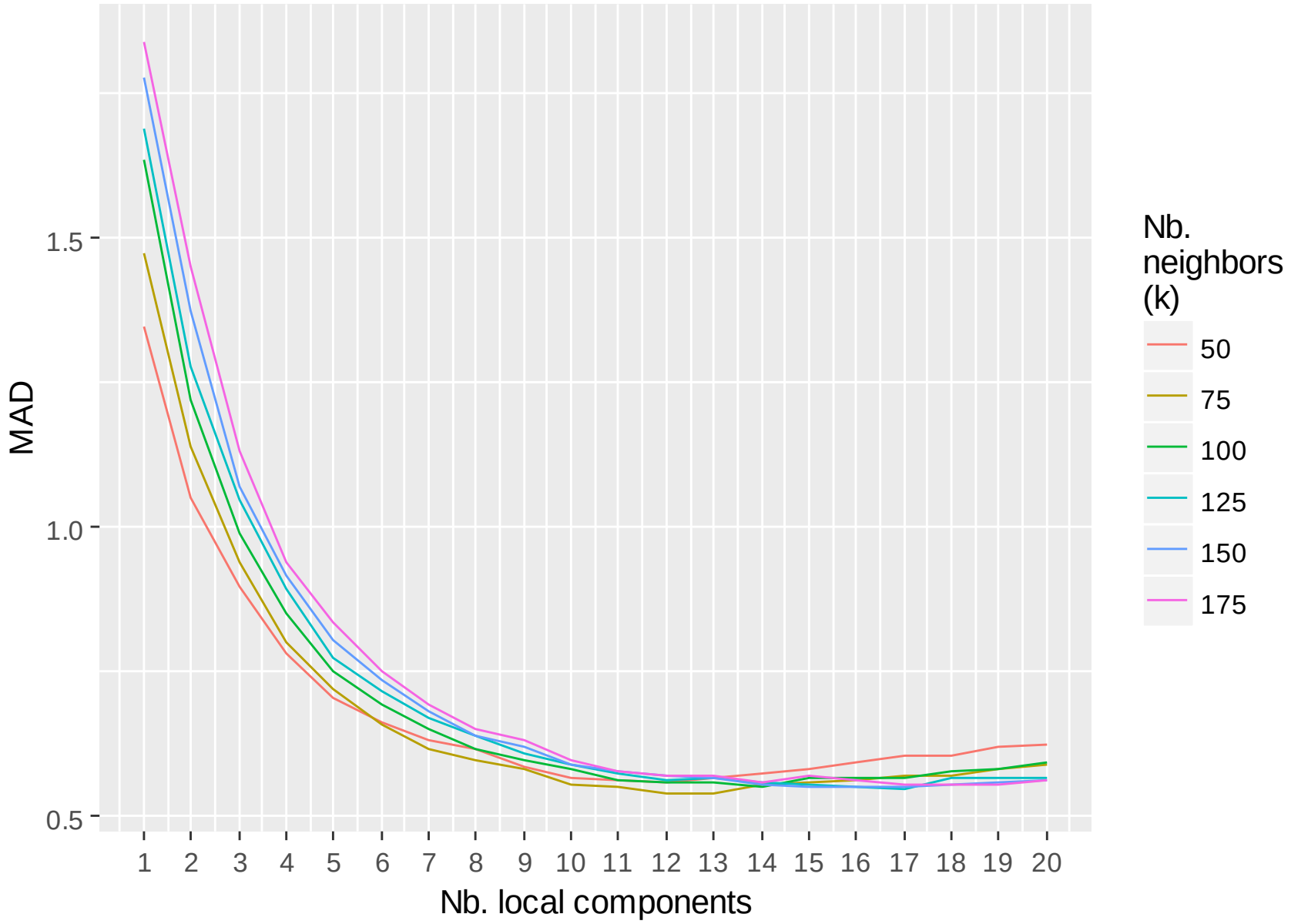
length s = 430

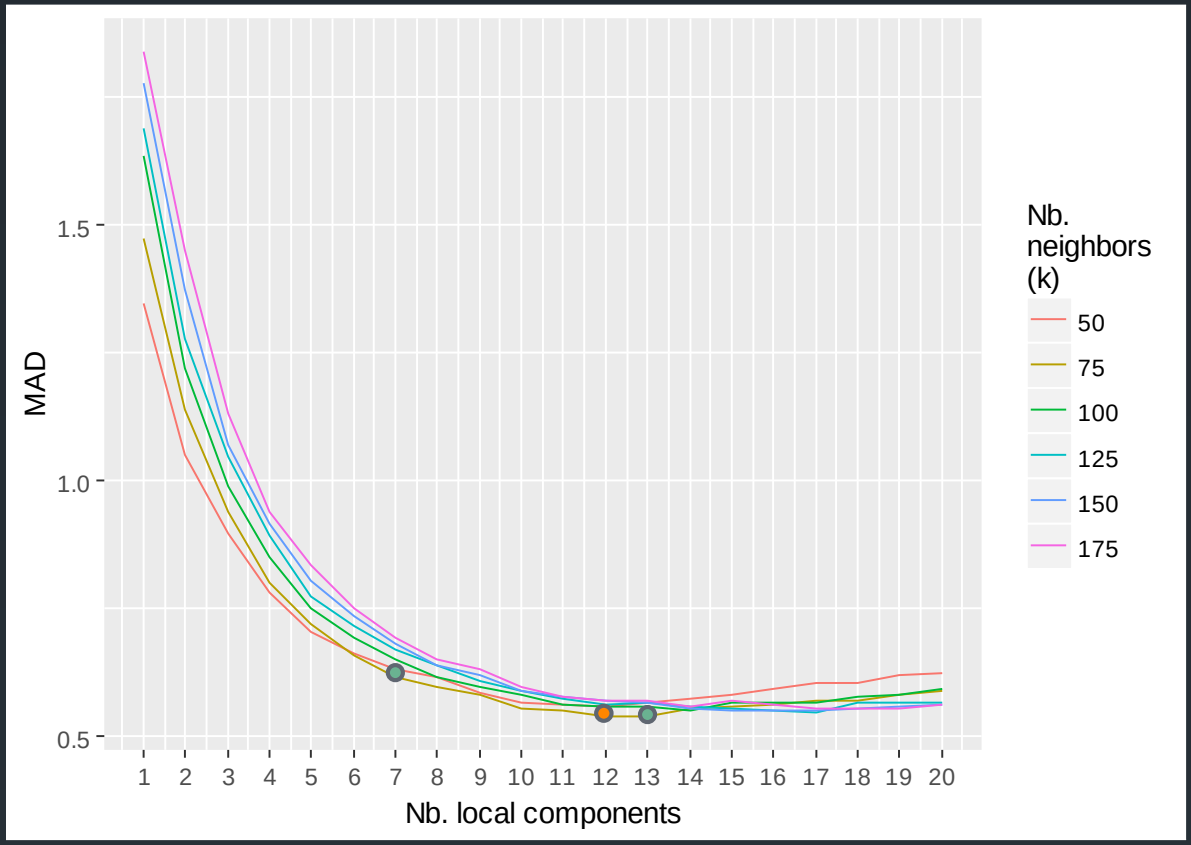
Sample of an index vector = s

“Test” $Xr[s]$ $yr[s]$

“Calibration” $Xr[-s]$ $yr[-s]$

cluster	nr	pr	nu	pu
1	637	0.163	69	0.161
2	372	0.095	42	0.098
3	452	0.116	50	0.117
4	395	0.101	41	0.096
5	437	0.112	50	0.117
6	440	0.113	49	0.114
7	1175	**0.301**	128	**0.298**


MAD
1.5
1.0
0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Nb. local components
Nb. neighbors (k)
50
75
100
125
150
175



k	ncomp	MAD	RMSE
75	7	0.61	0.78
75	**12**	**0.54**	0.73
75	**13**	**0.53**	0.71
125	14	0.56	0.71
150	14	0.55	**0.70**

Package `nirs` available at http:/livtools.cirad.fr/nirs

Uses CRAN packages

- **`pls`** (*Mevik, Wehrens & Liland, 2016*): global PLSR algorithms
- **`Rfast`** (*Papadakis et al. 2017*): fast calculations (distances)

THANKS!