Local PLSR averaging

matthieu.lesnoff@cirad.fr

Chemometricum 2023 Padova, Italy 27-30 June

Montpellier, France

SELMET cirad INRAE L'INSTITUT agro

The unit | Projet scientifique | Produits et publications | Enseignement et formation

The SELMET research unit is an international group focused on livestock production in Mediterranean and tropical areas. The unit has approximately 100 permanent and contractual staff drawn from CIRAD, INRAE and the Institut Agro Montpellier. SELMET offers research, teaching and expertise to support sustainable transitions in livestock activities.

NIR Laboratory

- Spectrometry ➔ Foss, Bruker, Asd
- Chemical analyses

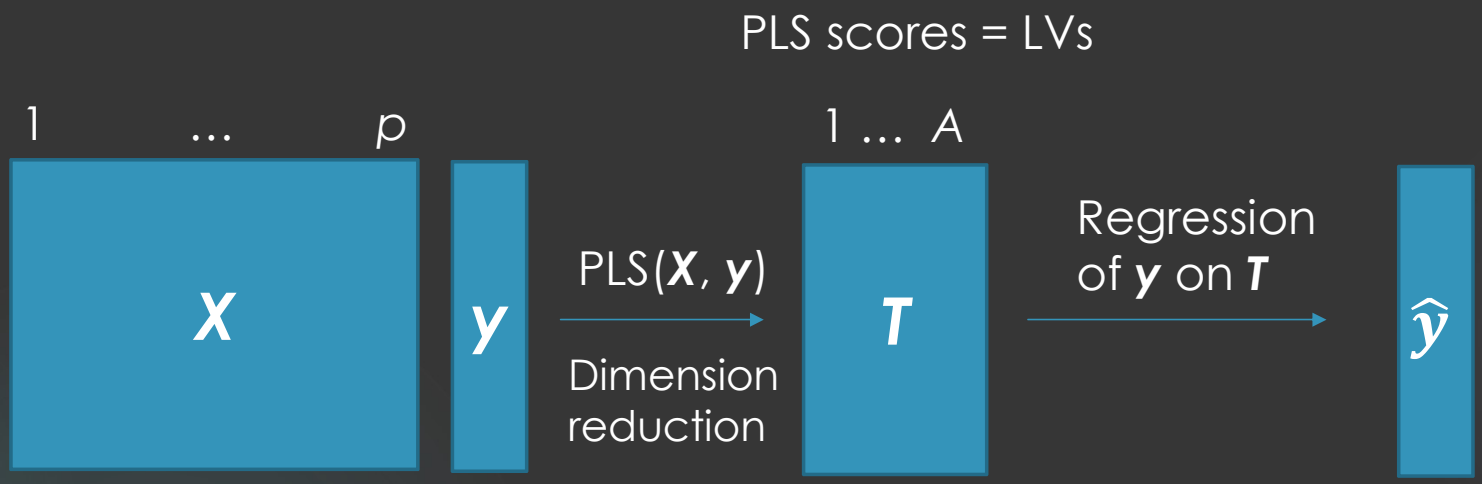
Compositions

- Forages
- Feed
- Feces (digestibility)

European and tropical

1. PLSR averaging
2. Extension to a method of **local PLSR**
3. Illustration: **forages and feed** dataset

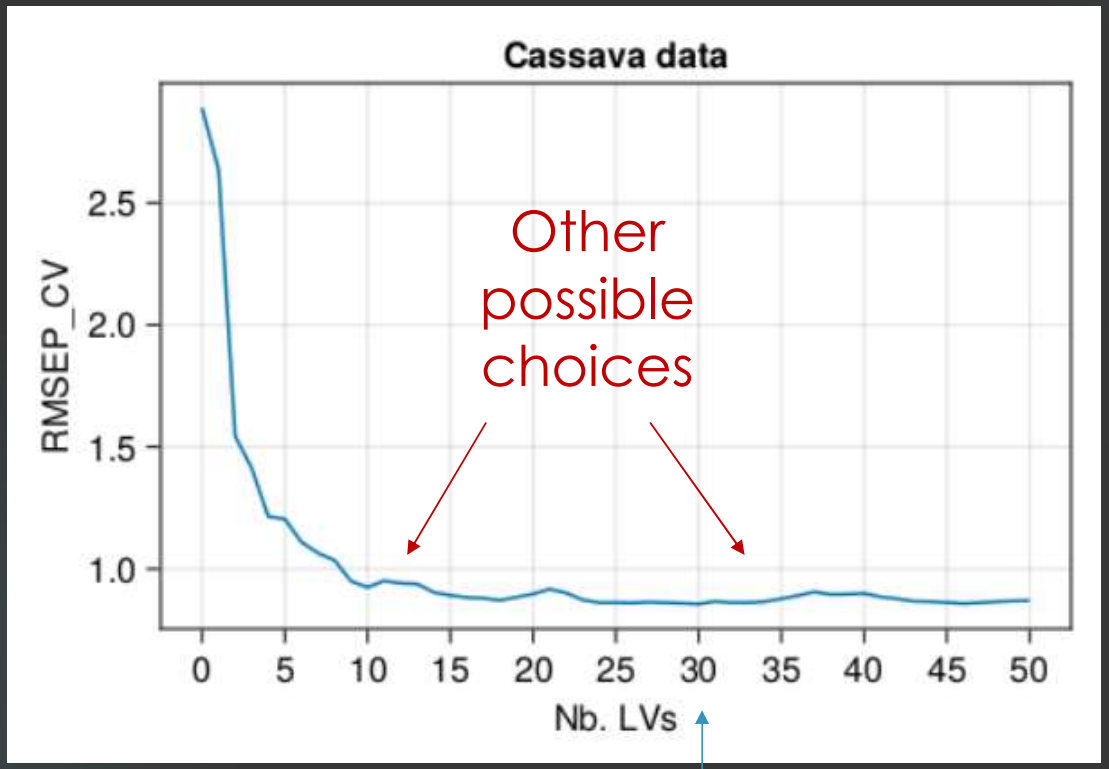
1. PLSR



Required step: Tuning A = nb. LVs

Which value for A?

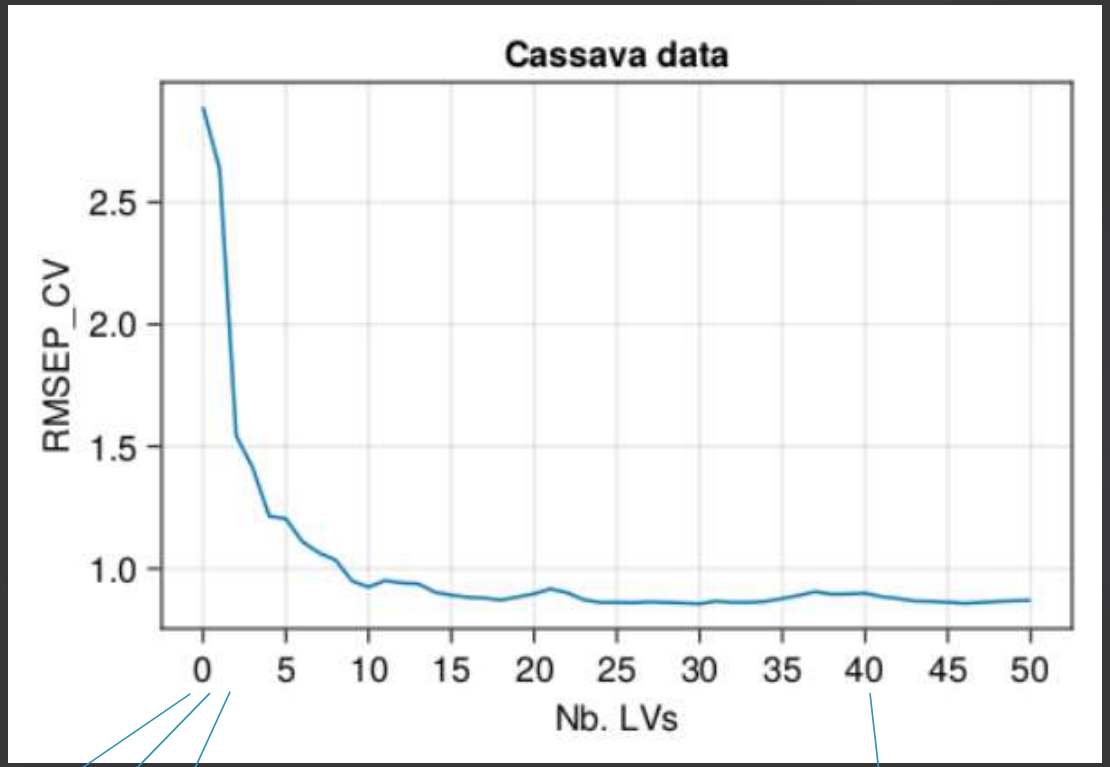
One usual tuning approach: **Cross-validation**



a_{opt} = 30 LVs

Prediction

$\Rightarrow \hat{y}_{[a\ opt]}$



Prediction with $\hat{y}_{[0]}$ **0 LV** $\hat{y}_{[1]}$ **1 LV** $\hat{y}_{[2]}$ **2 LVs** … $\hat{y}_{[A]}$ **40 LVs**

Average

$$\hat{y}_{avg,[A]} = \frac{\hat{y}_{[0]} + \hat{y}_{[1]} + \cdots + \hat{y}_{[A-1]} + \hat{y}_{[A]}}{A+1}$$

Final prediction

= PLSR averaging with uniform weighting

More elaborated weighting

$$\hat{y}_{avg,[A]} = \theta_0\,\hat{y}_{[0]} + \theta_1\hat{y}_{[1]} + \cdots + \theta_A\hat{y}_{[A]}$$

$$\sum_{a=0}^{A} \theta_a = 1$$

~ Confidence given to model [a]

- Probability (Bayes)
- Model performance
  - RMSEP-CV
  - AIC
  - etc.

More the model is performant, higher is θ_a

One variant: Stacking

θ_a are estimated from a meta-regression model

$$\sum_{a=0}^{A} \theta_a \neq 1$$

PLSR $\Rightarrow$ PLSR-AVG

Natural extension

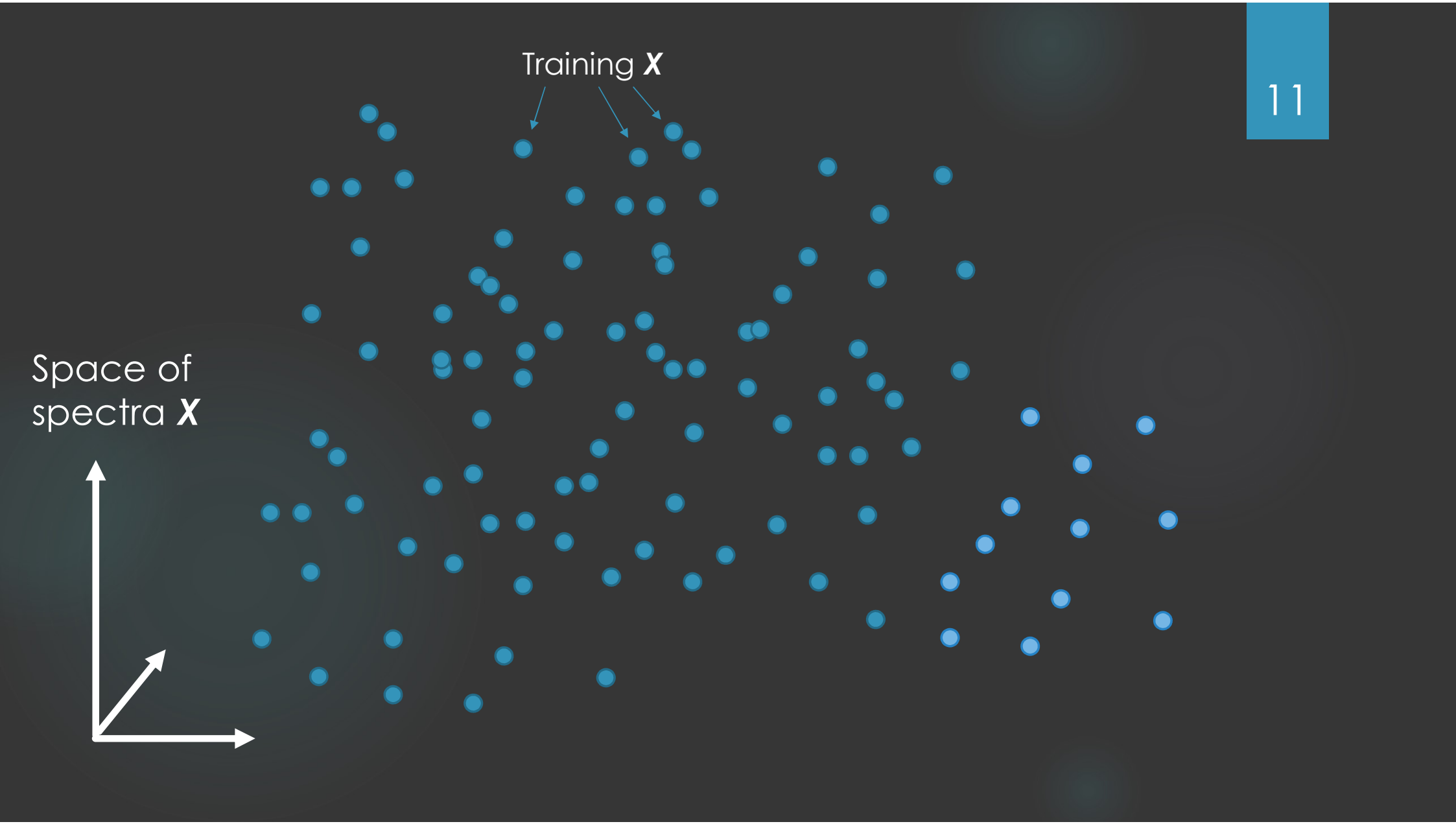
Local PLSR $\Rightarrow$ Local PLSR-AVG

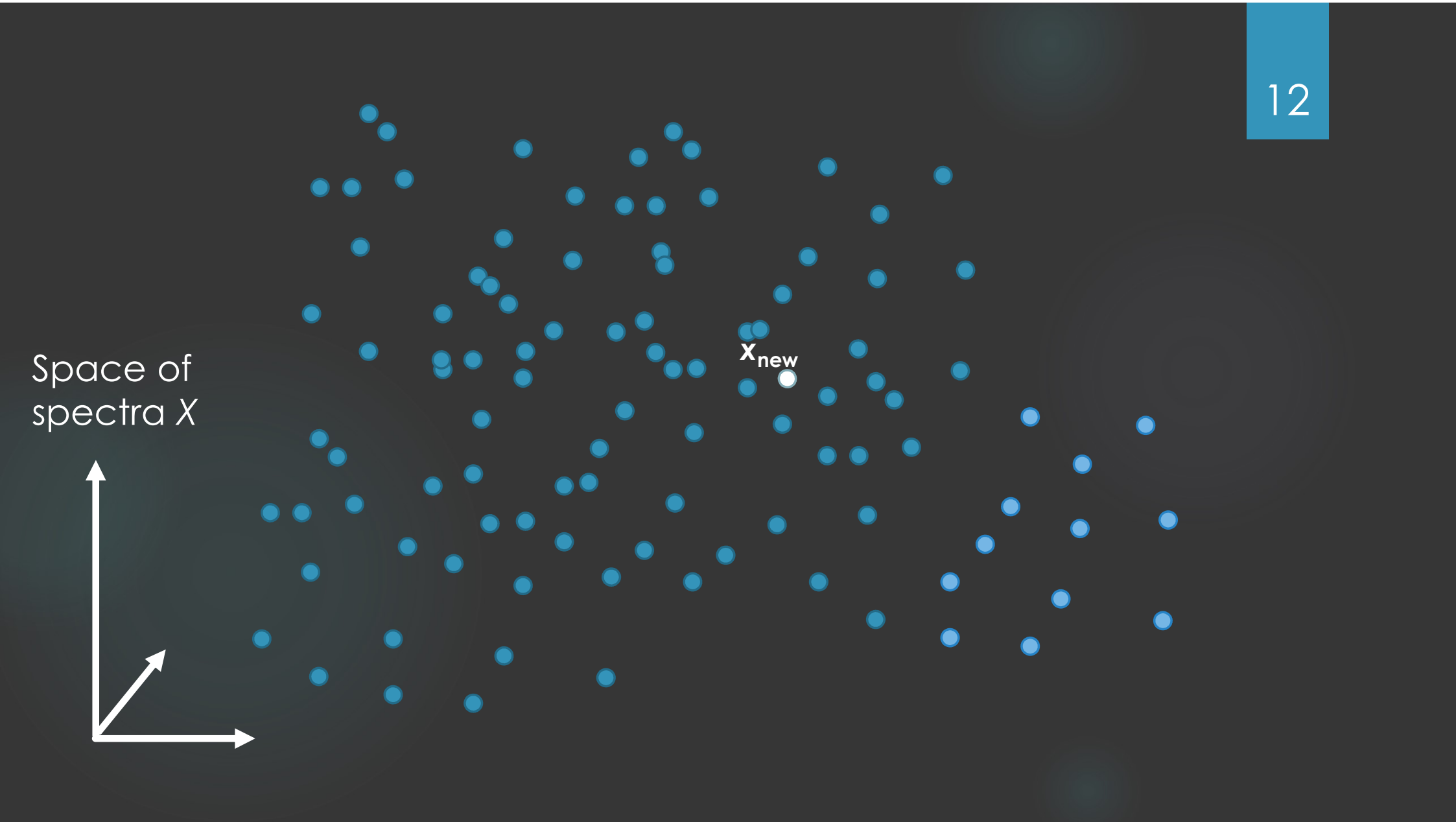
2. One local PLSR pipeline (*): kNN-LWPLSR

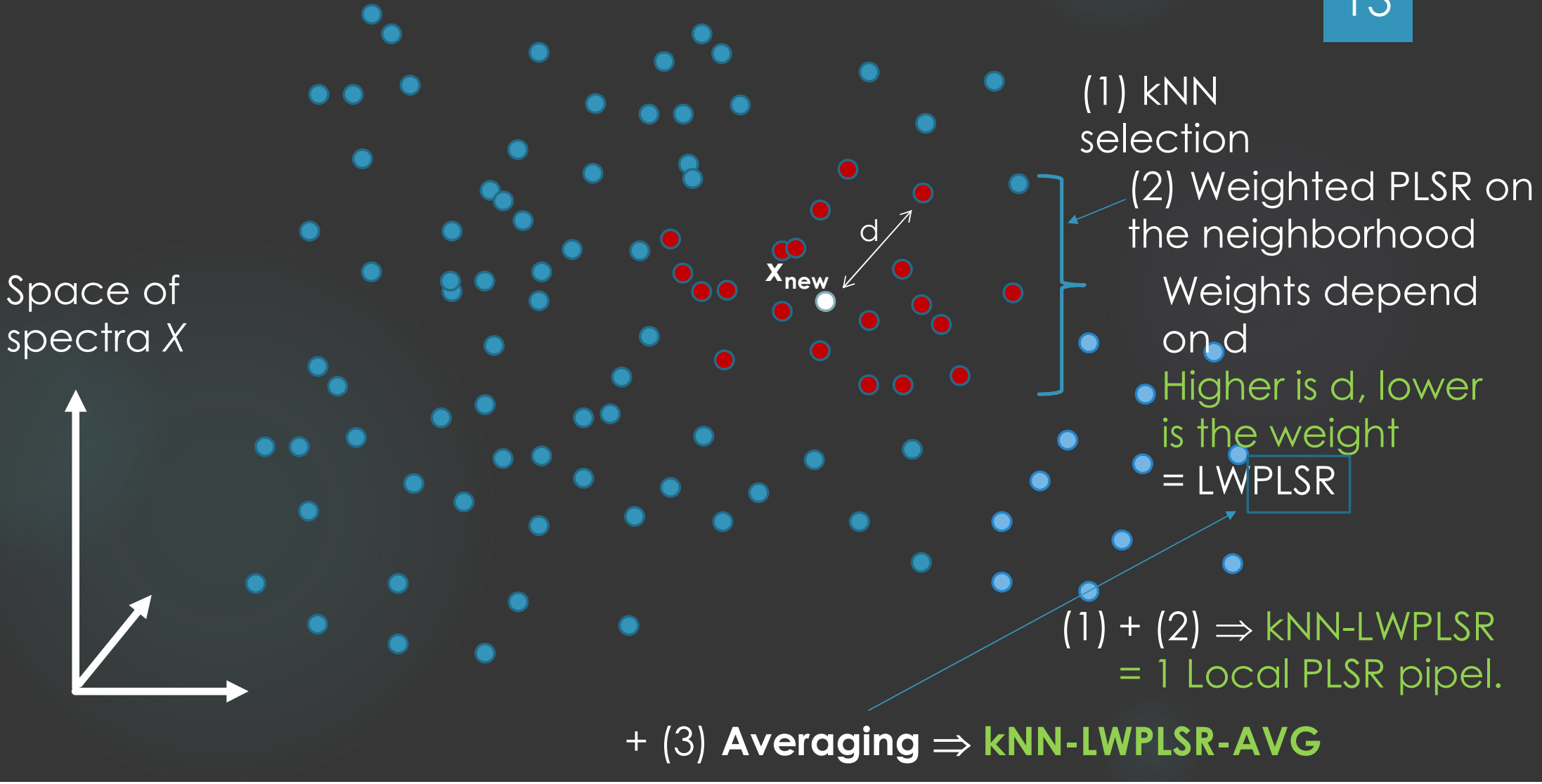
k nearest neighbors - locally weighted PLSR

Lesnoff et al. 2020 J. Chemometrics
https://doi.org/10.1002/cem.3209

(*) Many other pipelines are possible

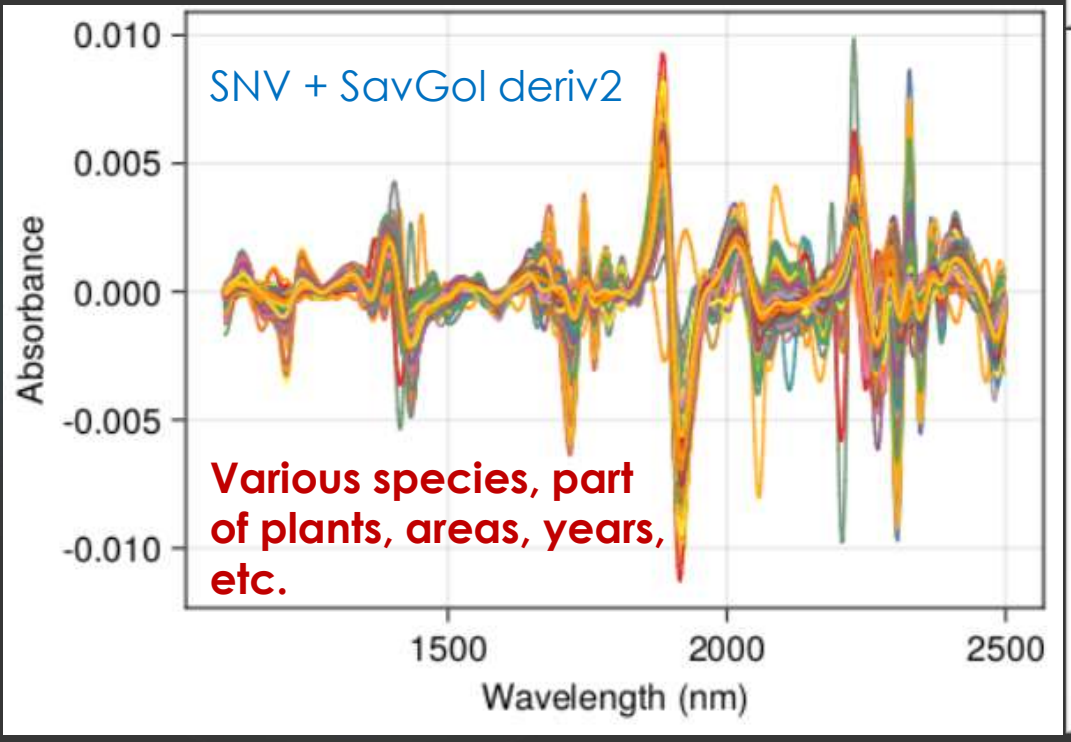

Training **X**
Space of spectra **X**


Space of
spectra X
x_{new}

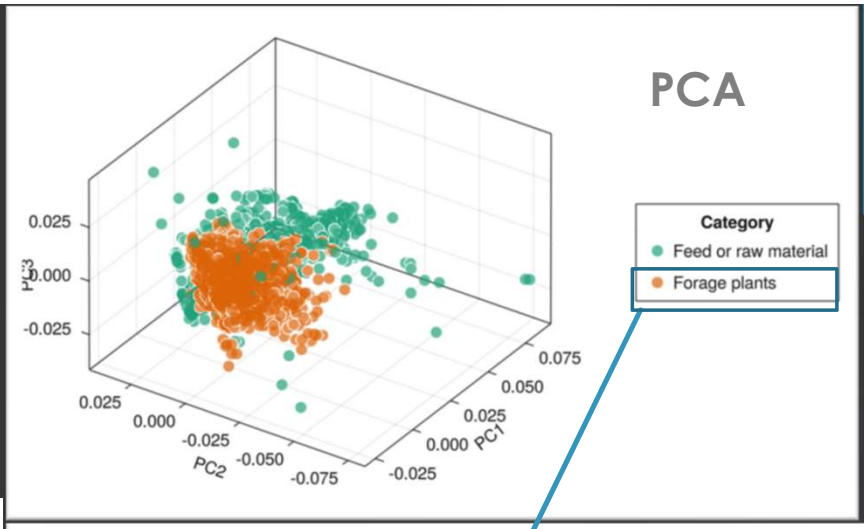

Space of spectra X
d
x_{new}
(1) kNN selection
(2) Weighted PLSR on the neighborhood
Weights depend on d
Higher is d, lower is the weight
= LWPLSR
(1) + (2) ⇒ kNN-LWPLSR
= 1 Local PLSR pipel.
+ (3) **Averaging** ⇒ **kNN-LWPLSR-AVG**

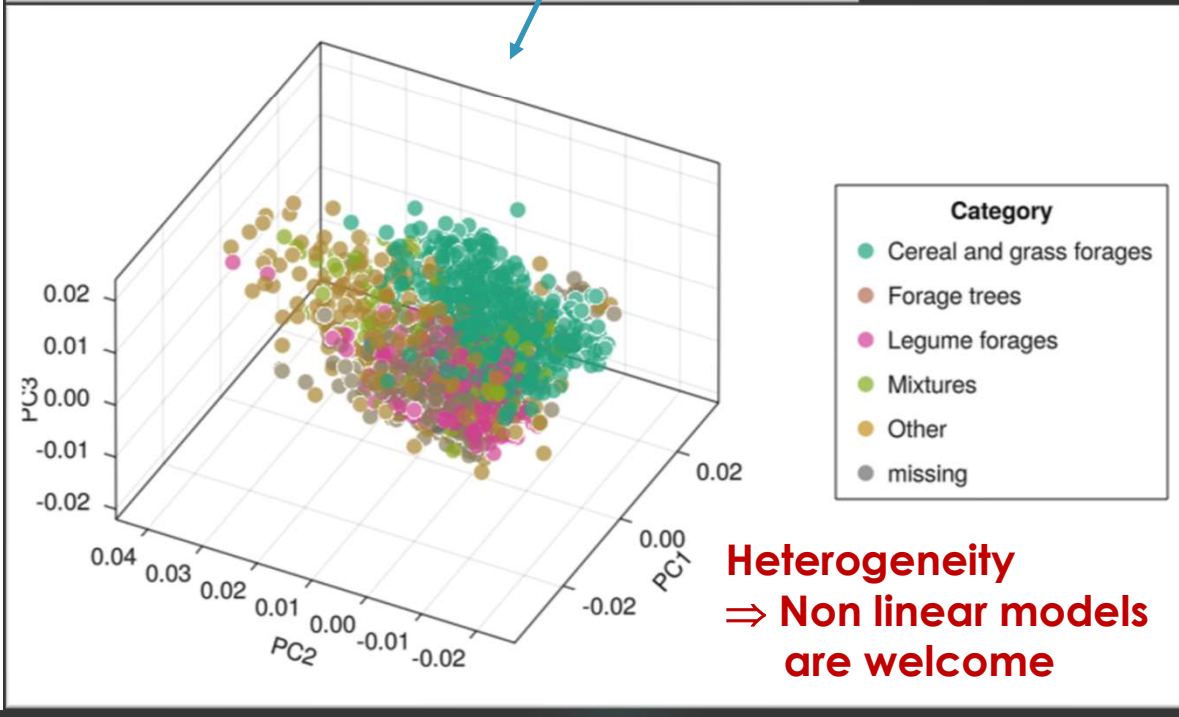
3. Application

Forages and feed (European and tropical)
Chemical composition
NIR Foss 1100-2498 nm step = 2 nm



PCA





Heterogeneity
⇒ Non linear models are welcome

12 variables to predict (chemical composition)

Abbreviation	Unit	Variable	Available *N*
DM	%	Dry matter	17714
ASH	%DM	Mineral matter	17299
CP	%DM	Crude protein	16295
EE	%DM	Crude fat	7004
CF	%DM	Crude fiber	13006
NDF	%DM	Neutral detergent fiber	10414
ADF	%DM	Acid detergent fiber	11097
ADL	%DM	Acid detergent lignin	10809
DMDCELL	%DM	DM enzymatic digestibility	9043
OMDCELL	%DM	OM enzymatic digestibility	8713
STARCH	%DM	Starch	1919
SUGARS	%DM	Total sugars	2014

Compared models

I. kNN-LWPLSR — Usual CV tuning
⇒ Optimize Nb. LVs in the local PLSR models

II. kNN-LWPLSR-AVG — Averaging over 0-20 LVs

For each variable

Available data *N* ⇒ Train + 20% Test (random)

- Train ⇒ Cal + 20% Val (random)
  Tuning

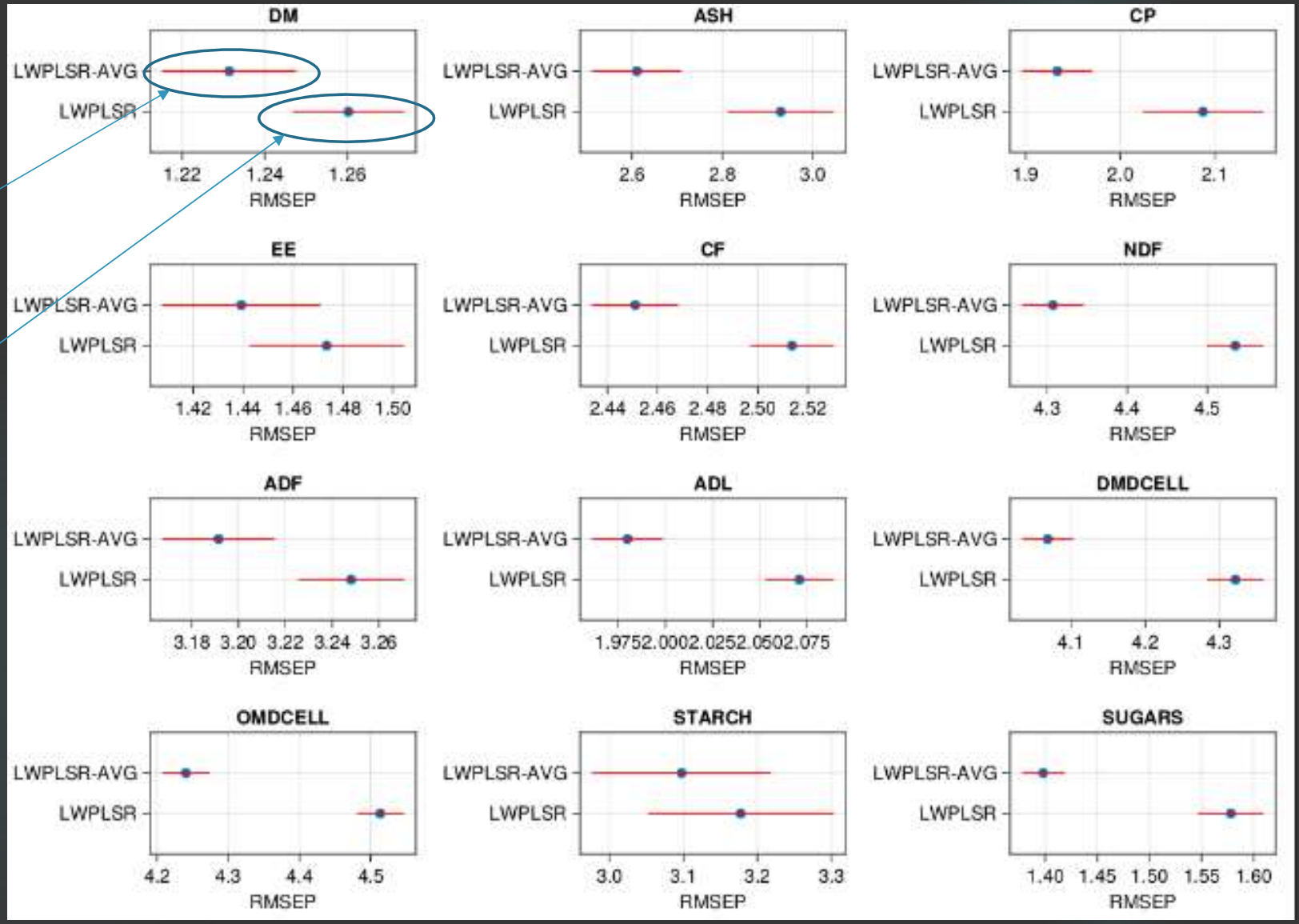
Test error ($RMSEP_{test}$)

30 replications

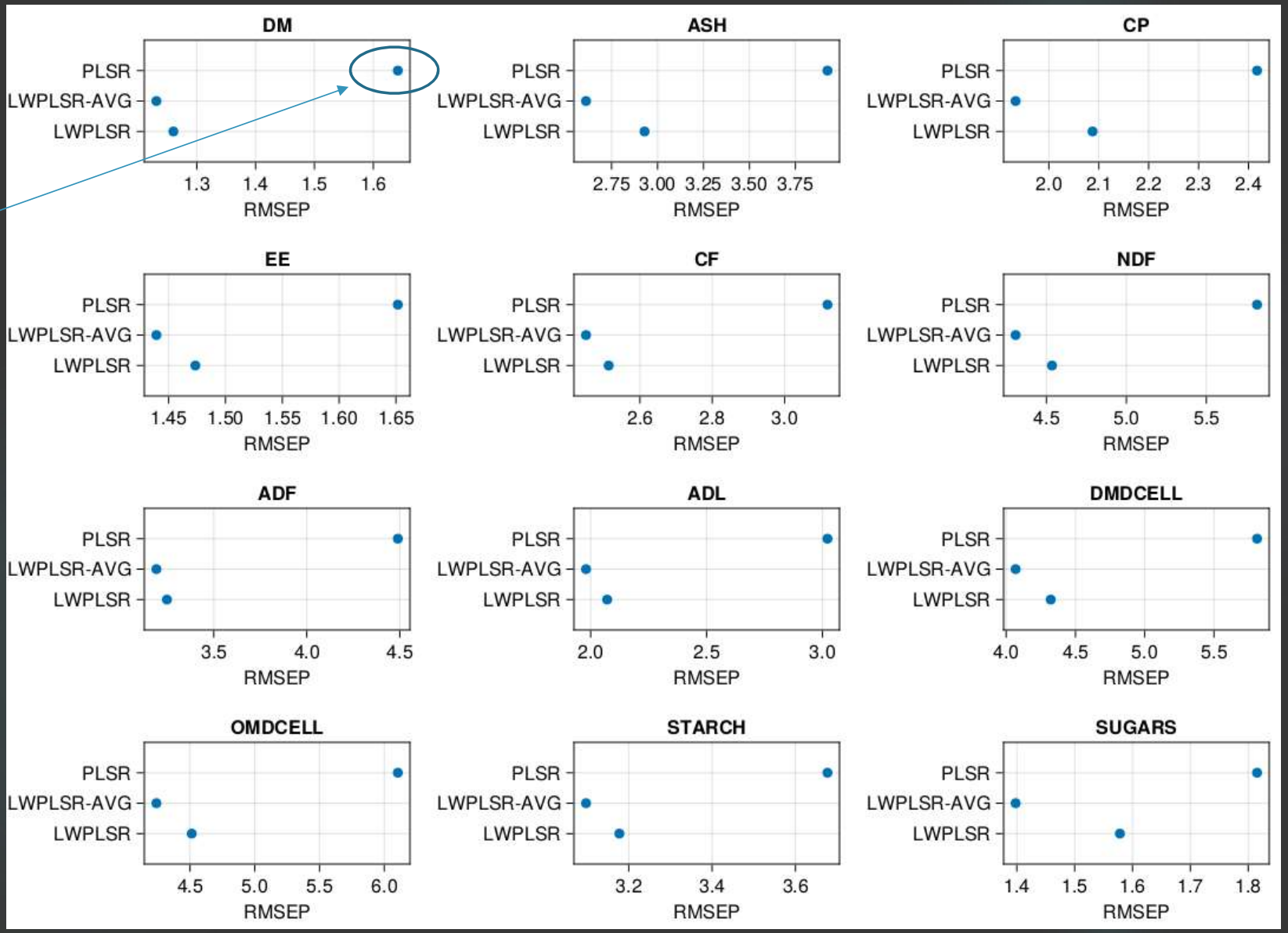
Test errors

Averaging 0-20 LVs

Usual CV No averaging



PLSR


DM
PLSR
LWPLSR-AVG
LWPLSR
1.3 1.4 1.5 1.6
RMSEP
ASH
PLSR
LWPLSR-AVG
LWPLSR
2.75 3.00 3.25 3.50 3.75
RMSEP
CP
PLSR
LWPLSR-AVG
LWPLSR
2.0 2.1 2.2 2.3 2.4
RMSEP
EE
PLSR
LWPLSR-AVG
LWPLSR
1.45 1.50 1.55 1.60 1.65
RMSEP
CF
PLSR
LWPLSR-AVG
LWPLSR
2.6 2.8 3.0
RMSEP
NDF
PLSR
LWPLSR-AVG
LWPLSR
4.5 5.0 5.5
RMSEP
ADF
PLSR
LWPLSR-AVG
LWPLSR
3.5 4.0 4.5
RMSEP
ADL
PLSR
LWPLSR-AVG
LWPLSR
2.0 2.5 3.0
RMSEP
DMDCELL
PLSR
LWPLSR-AVG
LWPLSR
4.0 4.5 5.0 5.5
RMSEP
OMDCELL
PLSR
LWPLSR-AVG
LWPLSR
4.5 5.0 5.5 6.0
RMSEP
STARCH
PLSR
LWPLSR-AVG
LWPLSR
3.2 3.4 3.6
RMSEP
SUGARS
PLSR
LWPLSR-AVG
LWPLSR
1.4 1.5 1.6 1.7 1.8
RMSEP

Averaging in PLSR is not very new …

- Shenk, J., Westerhaus, M., Berzaghi, P., **1997** Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy 5, 223. https://doi.org/10.1255/jnirs.115
- Zhang, M.H.; Xu, Q.S.; Massart, D.L. **2004** Averaged and Weighted Average Partial Least Squares. *Anal. Chim. Acta*, *504*, 279–289
- Silalahi, D.D.; Midi, H.; Arasan, J.; Mustafa, M.S.; Caliman, J.-P. **2020** Automated Fitting Process Using Robust Reliable Weighted Average on Near Infrared Spectral Data Analysis. *Symmetry*, *12*, 2099
- Lesnoff, M., Andueza, D., Barotin, C., Barre, P., Bonnal, L., Fernández Pierna, J.A., Picard, F., Vermeulen, P., Roger, J.-M., **2022** Averaging and Stacking Partial Least Squares Regression Models to Predict the Chemical Compositions and the Nutritive Values of Forages from Spectral Near Infrared Data. *Applied Sciences* 12, 7850

Six datasets of tropical and European forages
⇒ demonstrates **PLSR-AVG efficiency** **for forages data**

The present study confirmed the results

PLSR averaging belongs to a general frame:

"model averaging"

model 1 $\hat{y}_1$

model 2 $\hat{y}_2$

model 3 $\hat{y}_3$

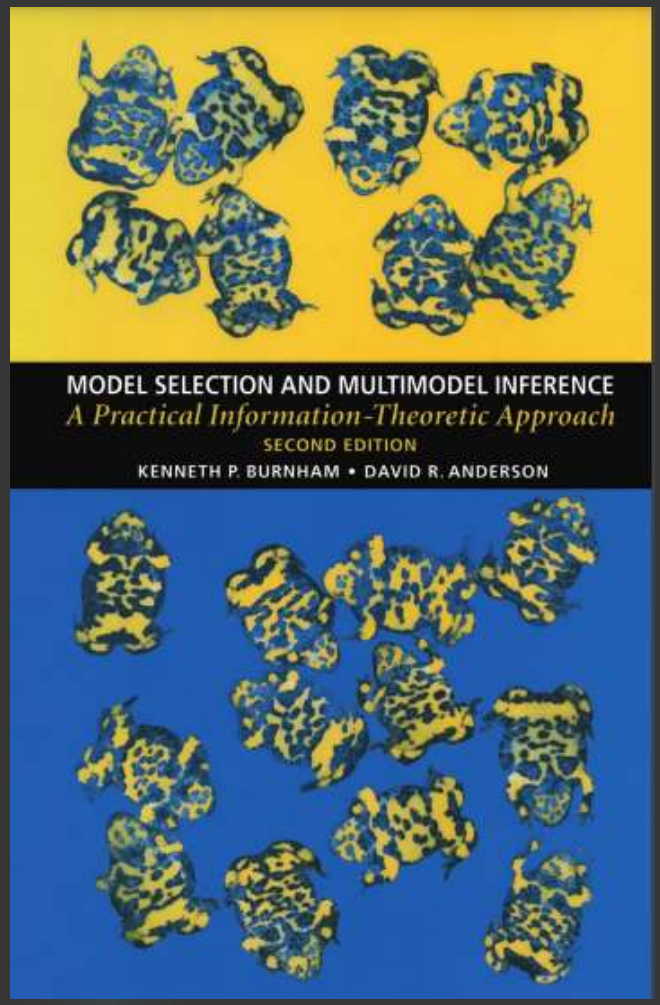
…

model *K* $\hat{y}_K$

$$\hat{y}_{avg} = \theta_0\ \hat{y}_{[0]} + \theta_1 \hat{y}_{[1]} + \cdots + \theta_A \hat{y}_{[A]}$$

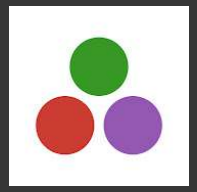
Often

Var($\hat{y}_{avg,i}$) < Var($\hat{y}_{[k],i}$)



Conclusions

- Averaging in local PLSR is in general beneficial
  - For heterogeneous materials such as mixed plants (information everywhere over the wavelengths)
  - Less efforts of tuning — interesting when **automatized prediction** platforms
  - Fast algorithm — if uniform weighting
- Can also be used for **(a)** multivariate Y (PLS2), and **(b)** kNN-LWPLSDA (but I observed less performance gains)



Available in package Jchemo

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

« **Dimension reduction, Regression and Discrimination for Chemometrics** »

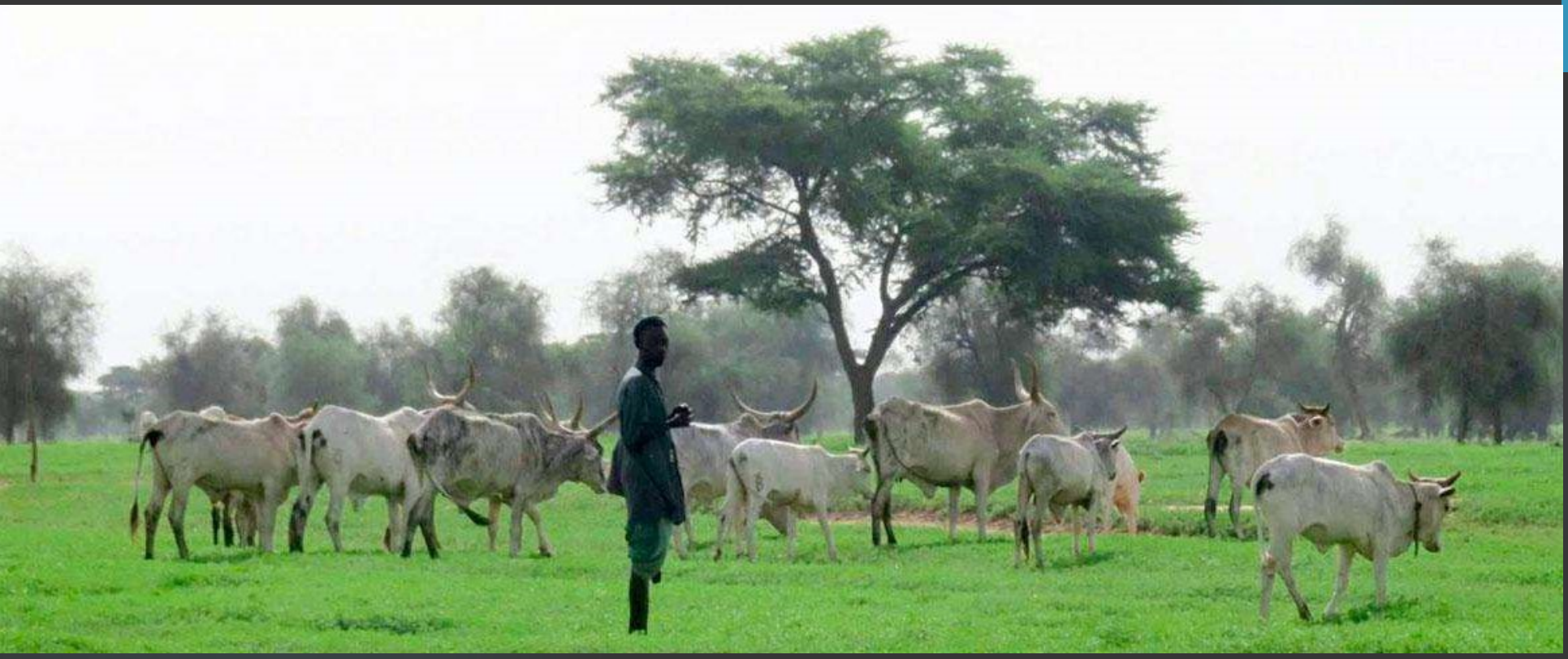
lwplsr
lwmlr
lwplsr
lwmlr_s
lwplsr_s
lwplsravg kNN-LWPLSR-AVG

Local regression pipelines

Exists also on R Cran Package rchemo

(slower)

Can be installed from the **official Julia repo**



Thank you for your attention