Averaging a local-PLSR models to predict chemical compositions and nutritive values of forages from spectral near infrared data

Matthieu Lesnoff [a*, b, c]

a. SELMET, Univ Montpellier, CIRAD, INRAe, Institut Agro, Montpellier, France
b. CIRAD, UMR SELMET, Montpellier, France
c. ChemHouse Research Group, Montpellier, France

* Corresponding author: Selmet Joint Research Unit (Tropical and Mediterranean Animal Production Systems), Cirad, TA C-112 / A - Campus international de Baillarguet - 34398 Montpellier Cedex 5, France. Tel: +33 4 67 59 38 63, Fax: +33 4 67 59 38 25
Email: matthieu.lesnoff@cirad.fr

Short title: Averaging a local-PLSR models to predict NIR data

Abstract

Partial least squares regression (PLSR) is a reference method in chemometrics. In agronomy, it is used for instance to predict components of chemical composition (response variables $\boldsymbol{y}$) of vegetal materials from spectral near infrared (NIR) data $\boldsymbol{X}$ collected from spectrometers. The principle of PLSR is to reduce the dimension of the spectral data $\boldsymbol{X}$ by computing vectors that are then used as latent variables (LVs) in a multiple linear model. A difficulty is to determine the relevant dimensionality (number of LVs) of the model for the given available data. This step can also become time consuming when many different datasets have to be processed and/or the datasets are frequently updated. An alternative to determinate the relevant PLSR dimensionality is the ensemble learning method "PLSR-averaging". In the past, this method has been demonstrated to be efficient for complex biological materials such as mixed forages, and facilitates to automatize predictions (e.g. in user-friendly web interface platforms). This article presents the extension of the PLSR-averaging to a *k*-nearest neighbors locally weighted PLSR pipeline (kNN-LWPLSR). The kNN-LWPLSR pipeline has the advantage to account for non-linearity between **X** and **y** existing for instance in heterogeneous data (e.g. mixing of vegetal species, collection from different geographical areas, etc.). In the article, kNN-LWPLSR-averaging is applied to an extensive NIR database built to predict the chemical composition of European and tropical forages and feed. The main finding of the study was the overall superiority of the averaging compared to the usual kNN-LWPLSR. Averaging may therefore be recommended in local-PLSR pipelines to predict NIR forage data.

Key-words: locally weighted PLSR, *k*-nearest neighbors, model averaging, NIR data, forages

1 Introduction

Near-infrared spectroscopy (NIRS) is a fast and nondestructive analytical method used in many agronomic (or other) contexts, for instance to evaluate the nutritive quality of forages. Basically, spectral data $\boldsymbol{X}$ (matrix of n observations × p wavelengths) are collected on samples of the material to study (e.g. forages) using a spectrometer, and targeted response variables (e.g. chemical compositions) $\boldsymbol{Y}$ $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ (k vectors of n observations) are measured precisely in laboratory. Regression models of $\boldsymbol{Y}$ on $\boldsymbol{X}$ are then

fitted and used to predict the response variables from new spectral observations. Spectral data are known to be highly collinear and, in general, matrix $\boldsymbol{X}$ is ill-conditioned. Specific regression methods have to be implemented, in particular partial least squares regression (PLSR) [1–3]. The general principle of PLSR is to reduce the dimension of $\boldsymbol{X}$ to a limited number $a << p$ of orthogonal vectors $n \times 1$ maximizing the squared covariance with $\boldsymbol{Y}$ and referred to as scores. The scores are then used as regressor latent variables (LVs) in a multiple linear regression (MLR). PLSR is very efficient when the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is linear [4].

From several years, a general trend in the agronomic databases (e.g. in feed, food or soils researches) is to aggregate large numbers of samples of different natures or origins. This generates non-linearities in the data, such as curvatures and/or clustering, that can alter the PLSR predictions. Local-PLSR is an easy tool that can turn out non-linearity in the data [4–7]. The general principle is, for each new observation to predict, to do a pre-selection of k nearest neighbors of the observation (the kNN selection step) and then apply a PLSR to the neighborhood (i.e. the k neighbors). Many variants of local-PLSR pipelines can be built, depending essentially on the type of PLSR implemented, and on how are selected the neighborhood. One variant is the kNN-locally weighted PLSR (kNN-LWPLSR) described in Lesnoff et al. [8]. This pipeline consists in applying a locally weighted PLSR (LWPLSR) (instead of applying a PLSR as in more usual local-PLSRs) on the neighborhood. The particularity of LWPLSR is to weight (in the internal computations) each training observation depending on the distance between this observation and the observation to predict $\boldsymbol{x}_{\text{new}}$ (usual PLSR gives a uniform weight $1 / n$ to all the n training observations). Closer is the training observation to $\boldsymbol{x}_{\text{new}}$, higher is its weight in the iterative PLSR equations and therefore its importance in the prediction. kNN-LWPLSR has been demonstrated to be efficient for various data, including forages, as well as for regression as for discrimination [8,9].

Tuning PLSR models consists in determining the dimensionality, say a, i.e. the number of LVs that is used in the MLR. Different strategies have been addressed in the literature to guide the determination of an optimal dimensionality [10–15]. One of the most common is the cross-validation (CV) that searches the value a that minimizes the CV-error curve. These strategies face to several difficulties. Firstly, despite attempts of automated procedures, they often require case-by-case decisions based on expertise. Secondly, to find a unique optimal dimensionality is in general difficult when the data contain heterogeneity, which is often the case with complex biological materials such as forages (mixing of stems, leaves, different stages of development and geographical areas, etc.). Finally, it can also become time consuming when many datasets have to be processed (e.g. different models for a set of chemical compositions) and/or when the datasets are periodically updated with new training observations. As an alternative, the ensemble learning method "PLSR-averaging" [16–19] is a tentative to bypass the determination of a. The method consists in averaging the predictions of PLSR models with dimensionalities $r = 0, 1, 2 \ldots A$ LVs, where A is given a priori, voluntary larger than the expected values of "optimal" a. In a recent study on six datasets of forage [19], PLSR-averaging was shown more efficient than the usual procedure where the PLSR dimensionality a was determined by CV.

In the past, PLSR-averaging was proposed on a simple local-PLSR pipeline [18]. The present article proposes the extension to the pipeline kNN-LWPLSR [8]. The article is organized as follows. Theoretical points on kNN-LWPLSR and PLSR-averaging are firstly presented. Then, the kNN-LWPLSR pipeline with averaging, say kNN-LWPLSR-

AVG, is applied to an extensive NIR database (Cirad, Selmet research unit) on chemical composition of European and tropical forages and feed. The predictive performances of the pipeline are compared to usual kNN-LWPLSR where the optimal number of LVs a is determined by CV.

2 Theory

2.1 kNN-LWPLSR

LWPLSR and kNN-LWPLSR have been described in detail in Lesnoff et al. [8]. LWPLSR [20–22] is a particular case of weighted PLSR (WPLSR). In WPLSR, a $n \times 1$ vector of weights $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots \delta_n\}$ is inserted into the PLSR algorithm, in two steps: (a) the PLS scores (LVs) are computed by maximizing weighted (instead of unweighted) covariances and (b) the prediction equation is computed by weighted least-squares (WLS, instead of ordinary LS) on the scores. The specificity of LWPLSR is that $\boldsymbol{\delta}$ is computed from a decreasing function, say f, of the dissimilarities (e.g. distances) between the n training observations and $\boldsymbol{x}_{\text{new}}$ the new observation to predict. This is the same principle as for the well-known locally weighted regression [23,24]. kNN-LWPLSR adds a preliminary step to LWPLSR: a neighborhood is selected around $\boldsymbol{x}_{\text{new}}$ (kNN selection) on which is then applied LWPLSR.

This article used the same kNN-LWPLSR pipeline as in Lesnoff et al. [8]. This pipeline is built on top of a fast PLSR algorithm [25] and consists in the following steps. A global PLSR (i.e. over all the training observation) is fitted and Mahalanobis distances between the training observations and $\boldsymbol{x}_{\text{new}}$ are computed in this global score space (other spaces, e.g. PCA or kernel-PCA/PLS score spaces, and other types of distances or dissimilarities can be used). Theses distances are used for the kNN selection and to compute the weights $\boldsymbol{\delta}$ within the neighborhood. The weight function f has a negative exponential shape that can be tuned by a scalar parameter h [8,21]: lower is h, sharper is function f and more the closest neighbors of $\boldsymbol{x}_{\text{new}}$ have importance in the LWPLSR model fitting. The case $h = \infty$ is the unweighted situation corresponding to the common local-PLSR (kNN-PLSR).

2.2. PLSR-averaging

Let $\boldsymbol{x}_{\text{new}}$ be a new observation to predict, and $\hat{y}_{new,A}$ the prediction returned by the PLSR model having a number of A of LVs. The PLSR-averaging model prediction is defined by:

$$\hat{y}_{new,avg[A]} = w_0\hat{y}_{new,0} + w_1\hat{y}_{new,1} + \cdots + w_A\hat{y}_{new,A} \qquad (1)$$

where w_r ($r = 0, \ldots, A$) is the weight (bounded between 0 and 1) of the model with r LVs, with the constraint:

$$\sum_{r=1}^{A} w_r = 1 \qquad (2)$$

In Eq.1, the particular case $\hat{y}_{new,0}$ is the simple mean of $\boldsymbol{y}$. Vector $\boldsymbol{w} = \{w_0, w_1, \ldots, w_A\}$ represents the pattern of weights. The shape of this pattern is specific to a given averaging method. In practice, weight w_r should quantify the level of confidence that can be awarded to the PLSR model with r LVs, relatively to the other dimensionalities. Vector $\boldsymbol{w}$ can be for instance estimated from the predictive performance of each of the $A + 1$ PLSR models (e.g. estimated by CV, Akaike criterion or other indicators)[19] or from more integrated Bayesian approaches[26,27]. Another approach, close to model averaging, is the stacking[26],

where weights w_r ($r = 0, \ldots, A$) are estimated from a meta-regression model, and are not anymore bounded to [0, 1] nor constrained to sum to 1.

2.3 kNN-LWPLSR-AVG

The proposed kNN-LWPLSR-AVG pipeline consists simply in chaining kNN-LWPLSR and PLSR-averaging (i.e. at the PLSR prediction step, averaging is done instead of using a single dimensionality a).

In a previous study on forages, Lesnoff et al. [19] showed that the uniform weighting of the dimensionalities (i.e. $w_r = 1 / (A + 1)$; $r = 0, \ldots, A$) was in general as efficient as the other weighting patterns on such data. Likewise, stacking did not show improvements in the prediction performances. For these reasons and for simplicity, the present article only considered uniform weighting. It has also the advantage to be much faster to compute than other patterns. Fast computation time of PLSR-averaging is important when implementing pipelines of local-PLSR, since one model per new observation to predict ($\boldsymbol{x}_{new}$) has to be fitted.

3. Material and Methods

3.1 Datasets

The method is assessed on a NIR database built by Cirad (Selmet research unit) on European and tropical forages and feed. The absorbance spectra were collected on dried and grounded materials using Foss Instruments 5000 or 6500 models in the spectral range 1100 to 2498 nm (2 nm steps). The database objective is to predict twelve variables of chemical composition (Table 1). The total size of the dataset was $N = 18813$ observations but, depending on the response variables, available data ranged from $N = 2020$ observations to $N = 18055$ observations (Table 1).

After preliminary exploration, a preprocessing was applied to the spectra consisting to a standard normal variate (SNV) transformation followed by a Savitzky-Golay 2nd derivation (polynomial of order 3, and window of 21 spectral points). Examples of raw and preprocessed spectra are presented in Fig.1. A projection of the data in the three first PCA scores (Fig.2) clearly illustrates large spectral heterogeneity in the dataset.

Table 1: Variables of forage and feed chemical composition to be predicted (*N* = total number of observations).

Abbreviation	Unit	Variable	*N*	Mean	Min	Max	Std
DM	%	Dry matter	18055	92.2	61.4	99.9	2.5
ASH	%DM	Mineral matter	17639	9.7	0.04	99.9	8.3
CP	%DM	Crude protein	16640	14.9	0.20	98.4	12.3
EE	%DM	Crude fat	7143	5.7	0.08	69.4	7.2
CF	%DM	Crude fiber	13232	25.9	0.06	77.7	11.9
NDF	%DM	Neutral detergent fiber	10677	52.0	0.60	95.1	18.1
ADF	%DM	Acid detergent fiber	11360	33.3	0.14	89.8	13.5
ADL	%DM	Acid detergent lignin	11072	8.3	0.01	50.8	6.7
DMDCELL	%DM	DM enzymatic digestibility	9187	52.5	6.98	99.9	17.9
OMDCELL	%DM	OM enzymatic digestibility	8857	50.4	6.18	99.6	17.9
STARCH	%DM	Starch	2020	31.5	0.04	88.9	20.7
SUGARS	%DM	Total sugars	2123	8.0	0.05	75.6	9.9

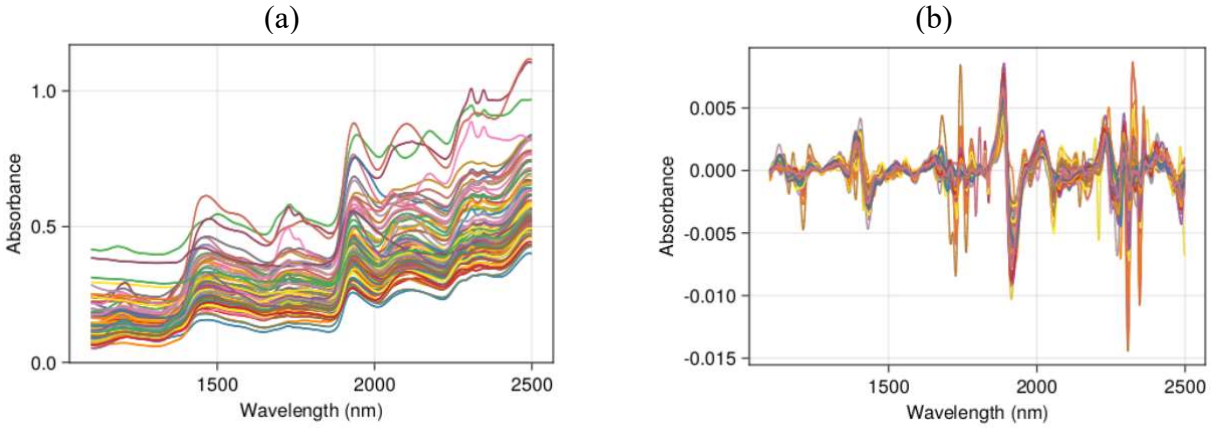

Fig.1: Example of 100 spectra of the forage and feed dataset: a) raw, b) preprocessed.

3.2 Predictive performances of the models

The performances of the models were compared by computing prediction errors rates on test sets. The procedure was the same for each given response variable (chemical composition) and each of the compared models, as follows. Let us consider response variable j:

- The total available data for this variable (size $N_{tot,j}$) is randomly divided to a training set, say *Train* (size $N_{train,j}$ = 80% * $N_{tot,j}$), and a test set, say *Test* (size $N_{test,j}$ = 20% * $N_{tot,j}$).
- Then, *Train* is randomly divided to a calibration set, say *Cal* (size $N_{cal,j}$ = 80% * $N_{train,j}$), and a validation set, say *Val* (size $N_{val,j}$ = 20% * $N_{train,j}$),
- Finally, *Tot* = *Train* + *Test*, and *Train* = *Cal* + *Val*.
- The model is tuned by exhaustive grid-search over *Train*:
  - All the K combinations of the predefined values of the model parameters are listed.
  - For each combination, the model is fitted on *Cal* and the performance of the combination is evaluated by the root mean squared prediction error rate computed on *Val* ($RMSEP_{Val}$).
- The optimal combination (lower $RMSEP_{Val}$ within the K combinations) is selected. The optimal model (i.e. with the selected combination) is then re-fitted on *Train* and used to predict *Test* and compute $RMSEP_{Test}$ that is used as final estimate of generalization error.

Due to the random splitting (*Train* vs. *Test*, and *Cal* vs. *Val*), $RMSEP_{Test}$ is subjected to sampling variation. To avoid misleading comparative results between the targeted methods, it is important to replicate this process to get the distribution of the results. In this article, the process was replicated n_{rep} = 30 times. The $RMSEP_{Test}$ distributions were summarized by means and standard errors.

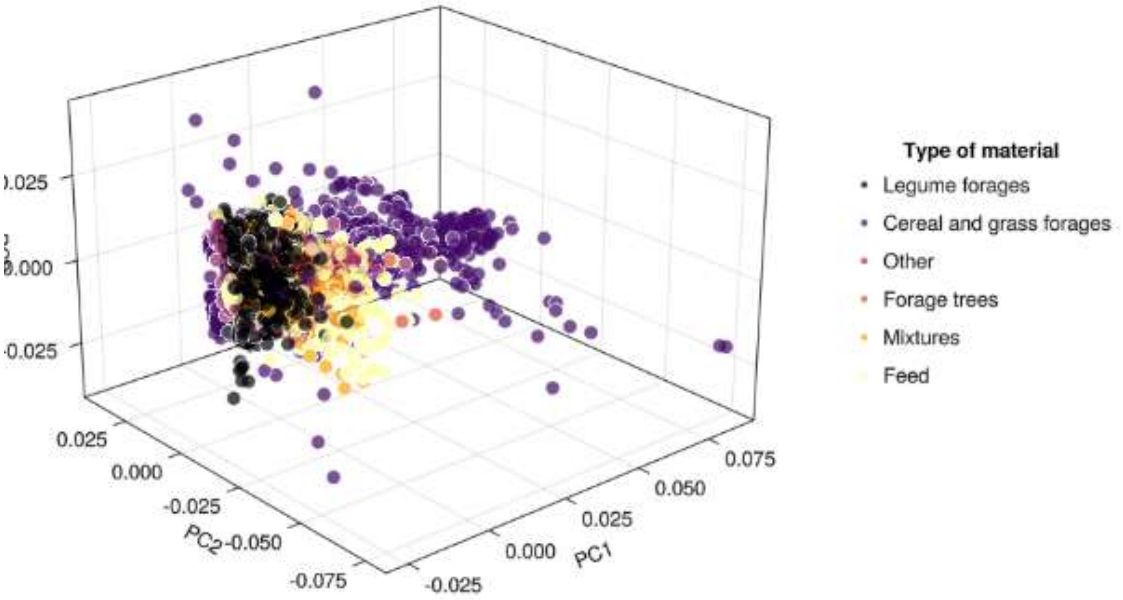

Fig.2: Projection of the observations (preprocessed forage spectra; $N = 18813$) on the three first PCA scores space.

3.3 The compared models

The two compared models where the usual kNN-LWPLSR in which the optimal number of LVs a is determined by CV, and the kNN-LWPLSR-AVG pipeline where a PLSR-averaging is used instead of determining a.

All the computations were implemented with the package Jchemo [28]. This package is written in the free and fast language Julia [29] (https://julialang.org).

kNN-LWPLSR

The preliminary global PLS score space was of dimensionality $nlv_{dis} = 25$ scores, from which were computed for each new observation to predict $\boldsymbol{x}_{new}$ the neighborhoods (kNN selections based on Mahalanobis distances) and the weights $\boldsymbol{\delta}$. The grid-search was undertaken over the combinations of the following values of parameters:

- $h = \{1, 2.5, 5\}$ Shape factor for the weighting function f
- $k = \{300, 500, 1000\}$ Neighborhood size (nb. neighbors)
- $nlv = \{0, 1, \ldots, 20\}$ Nb. LVs a (PLSR on the neighborhood)

that consisted in 315 combinations.

kNN-LWPLSR-AVG

The same values as above were considered for h and k. For each $\boldsymbol{x}_{new}$, the prediction was the average of the predictions of the models (LWPLSRs on the neighborhood) with $nlv = 0, \ldots, 20$ LVs ($A = 20$), i.e. the average of 21 predictions.

kNN-LWPLSR-AVG with omnibus strategy

Some applications require to process many training datasets and response variables. This is for instance the case of platforms of automated predictions when there is frequent updating of the training data. In such a situation, PLSR-averaging predictions can be facilitated in practice by using an "omnibus" strategy [19]. This strategy consists in defining one unique model, i.e. a set of *a priori* parameters values, that is applied blindly to all the training sets and response variables, instead of trying to optimize the model for each

dataset and variable. The defined omnibus model is expected (ideally) to have an acceptable efficiency for most of the datasets/variables to predict, even if not specifically optimal.

In this article, an omnibus strategy was applied to kNN-LWPLSR-AVG and compared the two above methods (kNN-LWPLSR and kNN-LWPLSR-AVG) optimized by grid-search. From preliminary studies on forages (sorghum and other data not considered in this article [8,19]), the omnibus parameters were set to $h = 1$ and $k = 1000$ neighbors.

4. Results

The performances of models kNN-LWPLSR and kNN-LWPLSR-AVG optimized by grid-search for the $n_{rep} = 30$ replications are summarized in Fig.3. For all the response variables, PLSR-averaging improved the predictions. After averaging, the mean $RMSEP_{Test}$ decreased (in relative values) from 1.7% (ADF) to 11.4% (SUGARS).

Fig.4 compares kNN-LWPLSR-AVG optimized by grid-search vs. with the omnibus parameters $h = 1$ and $k = 1000$ neighbors. The two strategies showed very close efficiencies: the relative difference between the mean $RMSEP_{test}$ were always lower than 2.5% in absolute value. The omnibus parameters provided lower vs. higher error rates depending on the response variables.

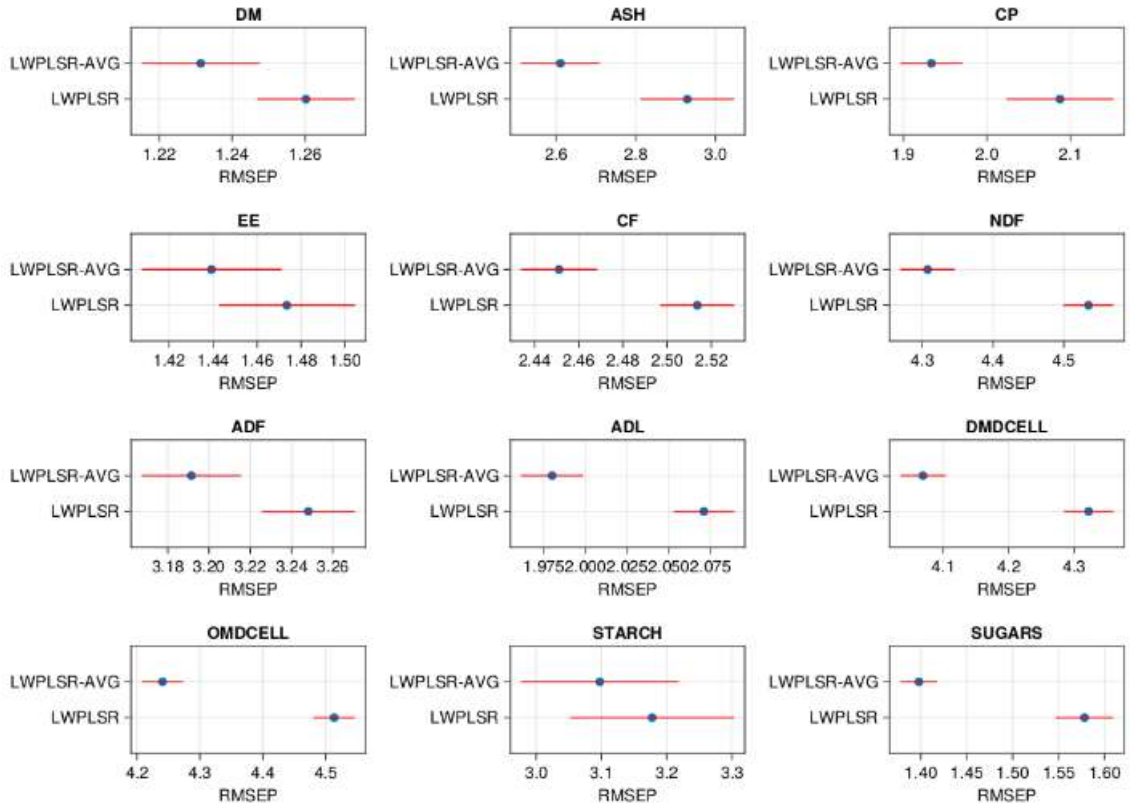

Figure 3: $RMSEP_{test}$ for kNN-LWPLSR and kNN-LWPLSR-AVG optimized by grid-search (averages ± standard errors computed on the $n_{rep} = 30$ replications).

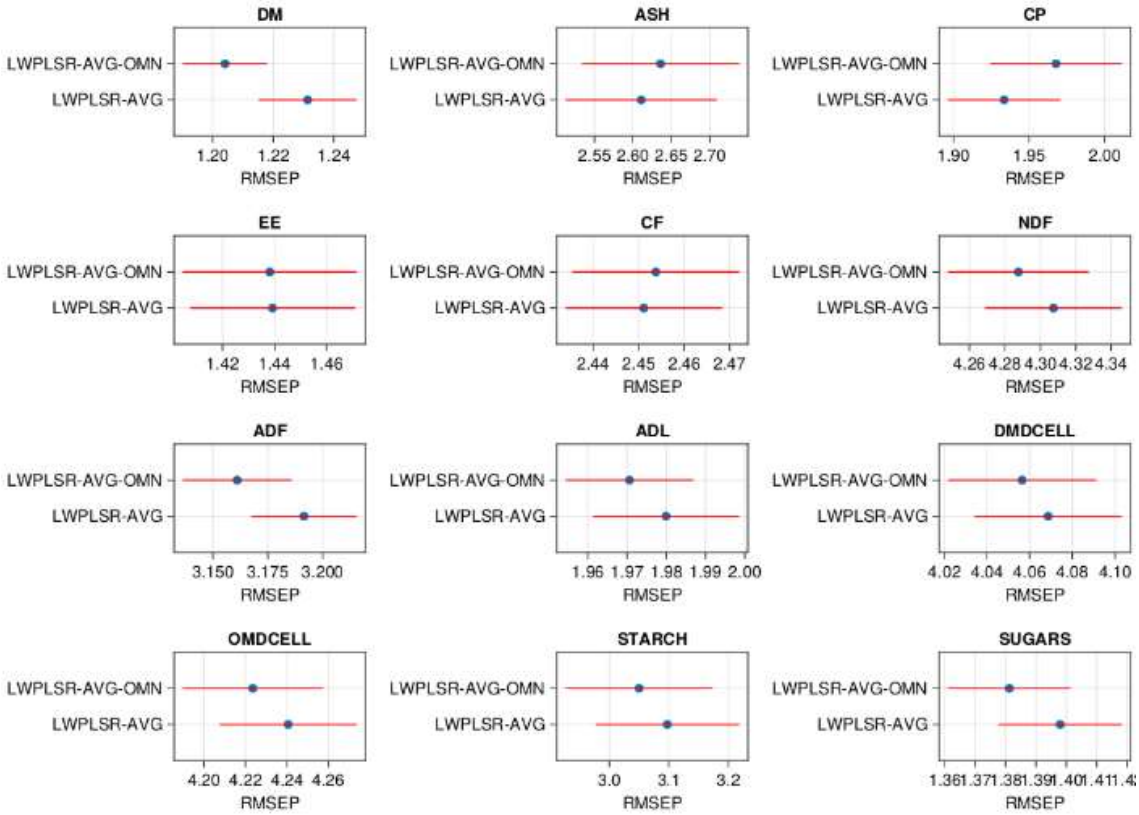

Figure 4: $RMSEP_{test}$ for kNN-LWPLSR-AVG optimized by grid-search *vs.* with the omnibus parameters $h = 1$ and $k = 100$ (averages ± standard errors computed on the $n_{rep} = 30$ replications).

5. Discussion and conclusions

PLSR-averaging improved the performances of the kNN-LWPLSR models, compared to the usual strategy that optimizes the number of LVs. The improvement was also observed for the omnibus strategy where unique values of parameters h and k were used for all the response variables. Since similar results were initially obtained for global PLSR on mixed forage data [19], this seems to confirm the overall superiority of PLSR-averaging for such materials. Mixed forage and feed data contain in general a high intrinsic complexity (different species, parts of plants, years, geographical areas, etc.), with information that can be, depending on the samples, distributed over all the spectral range. In such a case, averaging models with different dimensionality may generate certain gains of robustness, in addition to the statistical decrease of the prediction variances intrinsic to ensemble learning methods [26,30].

The present article focused on the uniform weighting, which is the simplest averaging approach. Although more elaborated weightings can be considered, they are not always more efficient [19] and, in the context of local-PLSR, can become too time consuming to compute. For instance, additional computations (not detailed in this article) on the study dataset showed that a weighting based on the CV-errors [19] did not improve the predictions. Moreover, a weighting based on AIC [19] required several days of computation to run the $n_{rep} = 30$ replications, which is not realistic in practice. The uniform weighting seems to be a good compromise for kNN-LWPLSR-AVG.

To reinforce the finding of this article, kNN-LWPLSR-AVG should be evaluated on other agronomic materials than mixed forages and feed, with the same level of heterogeneity or, conversely, with less complexity. In this last case, the method may be less advantageous.
In practice, such evaluations can be easily implemented, either by using already existing functions (e.g. in package Jchemo [28]) or by adapting an averaging in other pipelines of local-PLSR.

Acknowledgements
I thank Laurent Bonnal (Cirad, Montpellier) for the preparation of the spectral and chemical data.

References

1. Höskuldsson A. PLS regression methods. *Journal of Chemometrics*. 1988;**2**(3):211-228. doi:10.1002/cem.1180020306.

2. Wold H. Nonlinear iterative partial least squares (NIPALS) modeling: some current developments. In: *Multivariate Analysis II*. Wright State University, Dayton, Ohio, USA. June 19–24, 1972. New York : Academic Press: Krishnaiah , P. R.; 1973:383-407.

3. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001;**58**(2):109-130. doi:10.1016/S0169-7439(01)00155-1.

4. Dardenne P, Sinnaeve G, Baeten V. Multivariate Calibration and Chemometrics for near Infrared Spectroscopy: Which Method? *J Near Infrared Spectrosc, JNIRS*. 2000;**8**(4):229-237.

5. Clairotte M, Grinand C, Kouakoua E, et al. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma*. 2016;**276**:41-52. doi:10.1016/j.geoderma.2016.04.021.

6. Davrieux F, Dufour D, Dardenne P, et al. LOCAL regression algorithm improves near infrared spectroscopy predictions when the target constituent evolves in breeding populations. *Journal of Near Infrared Spectroscopy*. 2016;**24**(2):109. doi:10.1255/jnirs.1213.

7. Tran H, Salgado P, Tillard E, Dardenne P, Nguyen XT, Lecomte P. “Global” and “local” predictions of dairy diet nutritional quality using near infrared reflectance spectroscopy. *Journal of Dairy Science*. 2010;**93**(10):4961-4975. doi:10.3168/jds.2008-1893.

8. Lesnoff M, Metz M, Roger J-M. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics*. 2020;**n/a**(n/a):e3209. doi:10.1002/cem.3209.

9. Fernández Pierna JA, Laborde A, Lakhal L, et al. The applicability of vibrational spectroscopy and multivariate analysis for the characterization of animal feed where the reference values do not follow a normal distribution: A new chemometric challenge posed at the 'Chimiométrie 2019' congress. *Chemometrics and Intelligent Laboratory Systems*. 2020;**202**:104026. doi:10.1016/j.chemolab.2020.104026.

10. Gowen AA, Downey G, Esquerre C, O'Donnell CP. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics*. 2011;**25**(7):375-381. doi:https://doi.org/10.1002/cem.1349.

11. Kalivas JH. Multivariate Calibration, an Overview. *Analytical Letters*. 2005;**38**(14):2259-2279. doi:10.1080/00032710500315904.

12. van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*. 1994;**25**(2):313-323. doi:10.1016/0169-7439(94)85050-X.

13. Westad F, Marini F. Validation of chemometric models – A tutorial. *Analytica Chimica Acta*. 2015;**893**:14-24. doi:10.1016/j.aca.2015.06.056.

14. Wiklund S, Nilsson D, Eriksson L, Sjöström M, Wold S, Faber K. A randomization test for PLS component selection. *Journal of Chemometrics*. 2007;**21**(10-11):427-439. doi:10.1002/cem.1086.

15. Lesnoff M, Roger J-M, Rutledge DN. Monte Carlo methods for estimating Mallows's Cp and AIC criteria for PLSR models. Illustration on agronomic spectroscopic NIR data. *Journal of Chemometrics*. 2021;**n/a**(n/a):e3369. doi:10.1002/cem.3369.

16. Silalahi DD, Midi H, Arasan J, Mustafa MS, Caliman J-P. Automated Fitting Process Using Robust Reliable Weighted Average on Near Infrared Spectral Data Analysis. *Symmetry*. 2020;**12**(12):2099. doi:10.3390/sym12122099.

17. Zhang MH, Xu QS, Massart DL. Averaged and weighted average partial least squares. *Analytica Chimica Acta*. 2004;**504**(2):279-289. doi:10.1016/j.aca.2003.10.056.

18. Shenk J, Westerhaus M, Berzaghi P. Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*. 1997;**5**(1):223. doi:10.1255/jnirs.115.

19. Lesnoff M, Andueza D, Barotin C, et al. 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*. 2022;**12**(15):7850. doi:10.3390/app12157850.

20. Schaal S, Atkeson CG, Vijayakumar S. Scalable Techniques from Nonparametric Statistics for Real Time Robot Learning. *Applied Intelligence*. 2002;**17**(1):49-60. doi:10.1023/A:1015727715131.

21. Kim S, Kano M, Nakagawa H, Hasebe S. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics*. 2011;**421**(2):269-274. doi:10.1016/j.ijpharm.2011.10.007.

22. Sicard E, Sabatier R. Theoretical framework for local PLS1 regression, and application to a rainfall data set. *Computational Statistics & Data Analysis*. 2006;**51**(2):1393-1410. doi:10.1016/j.csda.2006.05.002.

23. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*. 1979;**74**(368):829. doi:10.2307/2286407.

24. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*. 1988;**83**(403):596-610. doi:10.1080/01621459.1988.10478639.

25. Dayal BS, MacGregor JF. Improved PLS algorithms. *Journal of Chemometrics*. 1997;**11**(1):73-85. doi:10.1002/(SICI)1099-128X(199701)11:1<73::AID-CEM435>3.0.CO;2-#.

26. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, New York; 2009.

27. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York, NY, USA: Springer; 2002.

28. Lesnoff M. Jchemo: Julia package for machine learning, with focus on chemometrics and high-dimensional data. 2021. https://github.com/mlesnoff/Jchemo.

29. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A fresh approach to numerical computing. *SIAM Rev*. 2017;**59**(1):65-98. doi:10.1137/141000671.

30. Burnham KP, Anderson DR. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*. 2004;**33**(2):261-304. doi:10.1177/0049124104268644.