A new permutation and Lasso-based interval selection technique

Rosa Arboretti Giancristofaro [*] Riccardo Ceccato [†] Luca Pegoraro [‡]
Luigi Salmaso [§]

Abstract

Near-infrared (NIR) spectroscopy is an analytical technique used to determine chemical and physical features of a sample. The sample is illuminated with near-infrared light and its properties, such as absorbance or reflectance, are measured at different wavelengths within the near-infrared region of the electromagnetic spectrum. A calibration model is then adopted to use information from the obtained spectral data to predict the chemical or physical feature of interest. Given that hundreds of wavelengths are commonly taken into consideration, it is fundamentally important to be able to distinguish between informative wavelengths and those providing only irrelevant or redundant information. Each wavelength corresponds to an independent variable to be included in the calibration model, so we are interested in identifying an appropriate feature selection approach. Rather than considering the commonly-used filter, wrapper or embedded methods, such as the Chi-squared test, Lasso regression or step-wise selection, in this paper we focus on a different family of feature selection techniques, namely interval selection methods. These methods are often used to select groups of consecutive wavelengths in the field of NIR spectroscopy due to the continuous nature of spectral data. As such it makes more sense for practitioners to select small informative regions of spectral points rather than a single point.

In this paper, we propose a new interval selection technique called Permutation and Lasso-based Interval Selection (PLIS), based on the adoption of Lasso Regression and permutation tests. The performance of this solution is then evaluated by means of a simulation study and a toy example coming from a real industrial problem.

Key Words: interval selection, NIR, permutation, Lasso

1. Introduction

Near-infrared (NIR) spectroscopy is an analytical technique widely used in food and pharmaceutical industries to determine chemical and physical properties of a product, to identify a component concentration, or for quality control [1]. It uses the near-infrared region of the electromagnetic spectrum, ranging from 750 nm to 2500 nm [2]. Using this technique the sample is illuminated by NIR radiation that can be absorbed, transmitted or reflected. NIR spectra are then collected, showing the amount of interaction between the light and the sample as a function of the wavelength. When several samples are analyzed and the related properties of interest are measured, the gathered spectral data are commonly pre-processed and used to train a calibration model which can be adopted to predict properties of new samples.

Suppose we are measuring the absorbance of a sample at several wavelengths. Each wavelength corresponds to an independent variable that will be used by the calibration model to predict the variable of interest [3–5]. Commonly the model will have to deal with hundreds of potentially highly correlated variables, among which only a few are really informative. For this reason, several authors have suggested that appropriate feature selection

[*]Civil, Environmental and Architectural Engineering, University of Padova, Padova, Italy
[†]Department of Management and Engineering, University of Padova, Vicenza, Italy
[‡]Department of Management and Engineering, University of Padova, Vicenza, Italy
[§]Department of Management and Engineering, University of Padova, Vicenza, Italy

techniques should be used to improve the predictive performances of the calibration model [3,4,6] by removing redundant or uninformative variables.

Additionally, having to deal with a high number of wavelengths means that simpler regression methods, such as linear regression, cannot be adopted because the sample size N is generally much lower than the number of variables V. To have $N \geq V$ would mean gathering NIR spectra from hundreds of samples. Although in NIR spectroscopy this problem is commonly solved by adopting Partial Least Squares (PLS) regression [7,8], i.e. a method performing a dimensionality reduction and making it possible to address problems where $N << V$, the use of feature selection techniques should not be overlooked.

Feature selection methods are commonly grouped into essentially three categories - filter methods, wrapper methods and embedded methods [9,10]. However, an additional category is considered particularly in the field of NIR spectroscopy: interval selection methods. These techniques are intended to select small regions of wavelengths, i.e. groups of adjacent variables, rather than single variables. Practitioners in this field are more interested in identifying regions of informative wavelengths because of the continuous nature of spectra and the high correlation between variables measured at consecutive wavelengths. Indeed several interval selection methods have been proposed in the literature, such as interval PLS (iPLS) [11], interval VISSA (iVISSA) [12], and interval Random Frog (iRF) [13].

The aim of this paper is to introduce a new interval selection technique, namely Permutation and Lasso-based Interval Selection (PLIS), based on the use of Lasso regression and permutation techniques, and to use this technique to explore the impact of interval selection on the predictive performances of calibration models. A simulation study is conducted and a toy example from a real industrial application is studied.

2. Permutation and Lasso-based Interval Selection (PLIS)

The key idea in Permutation and Lasso-based Interval Selection (PLIS) is to adopt a variable clustering algorithm, namely ClustOfVar [14], to group wavelengths, and then apply Lasso regression and permutation tests to select the most informative regions.

The adoption of ClustOfVar allows us to define clusters of variables without constraints on the data type. The resulting clusters are as homogeneous as possible, using the definition of homogeneous cluster provided by Chavent et al. [14], i.e. the variables of a cluster are strongly related to a central quantitative synthetic variable. Indeed, given a partition P_K in K clusters G_k, ClustOfVar aims to maximize the homogeneity criterion $H^{P_K} = \sum_{k=1}^{K} H(G_k) = \sum_{k=1}^{K} \left[\sum_{x_j \in G_k} \rho^2_{v_k,x_j} + \sum_{z_j \in G_k} \eta^2_{v_k|z_j} \right]$, where v_k is the related synthetic variable, $\rho^2_{v_k,x_j}$ is the squared Pearson correlation coefficient between v_k and the quantitative variable x_j, and $\eta^2_{v_k|z_j}$ is the correlation ratio between v_k and the qualitative variable z_j. In other words, for each cluster G_k it aims to maximize the sum of two quantities measuring the relationship between the synthetic variable and the quantitative variables and the link between the synthetic variable and the qualitative variables. It is worth noting that when data are strictly quantitative, as in the case of spectral data, the second quantity is ignored.

Let us now define what a synthetic variable is. Given a cluster G_k, the synthetic variable v_k is computed as $v_k = \text{argmax}_v \left\{ \sum_{x_j \in G_k} \rho^2_{v,x_j} + \sum_{z_j \in G_k} \eta^2_{v|z_j} \right\}$, In other words, it is the variable with the strongest link to all the other variables. The solution to this maximization problem is provided by the first principal component of PCAMIX. This particular principal component method is able to deal with mixed data [15]. Additionally, its empirical variance $V(v_k) = \sum_{x_j \in G_k} \rho^2_{x_j,v_k}$ is equal to λ_k, the first eigenvalue of PCAMIX.

Chavent et al. [14] claim that this empirical variance is equal to $H(G_k)$, so that the

homogeneity criterion H^{P_K} can be written as $H^{P_K} = \sum_{k=1}^{K} \lambda_k$.

To maximize H^{P_K}, ClustOfVar adopts a hierarchical clustering algorithm. Given V variables, this algorithm involves the following steps:

1. Create a partition P_V in V clusters.
2. For $i = 1, \ldots, V-2$: aggregate the pair of clusters G_l and G_m with the smallest dissimilarity $D(G_l, G_m) = H(G_l) + H(G_m) - H(G_l \cup G_m)$, measuring the loss of homogeneity after merging the two clusters, and a new partition P_{V-i} is generated.
3. When $i = V - 1$, stop.

At the end of the procedure we are able to achieve the most homogeneous partition possible in K clusters.

The first phase of PLIS essentially uses ClustOfVar considering a set of Q possible K values, so that Q partitions P_q are provided. To identify the best partition, for each P_q we apply Lasso regression on the related synthetic variables and retrieve the cross-validation error using repeated 5-fold cross-validation. We then use these errors and apply a ranking procedure based on permutation tests [16], i.e. highly flexible nonparametric tests that do not require any strict assumption on the data, particularly in relation to underlying distributions and their size [17]. Let us briefly describe this ranking procedure. After performing all possible comparisons between the Q partitions, $Q(Q-1)$ p-values are achieved and gathered in a matrix $\mathbf{p}_{Q \times Q}$ where each cell (l, m), $l \neq m$ contains the p-value related to the comparison between P_l and P_m, while each cell (l, l) is equal to 1. The steps are:

1. Create a matrix $\mathbf{s}$ with $s_{l,m} = 1$ if $p_{l,m} \leq \alpha/2$ and $s_{l,m} = 0$ if $p_{l,m} > \alpha/2$, where α is the desired significance level.
2. Compute the vector $\left\{r_u^l = 1 + \#\left[(Q - \sum_{m=1}^{Q} s_{l,m}) > (Q - \sum_{m=1}^{Q} s_{z,m})\right], z = 1, \ldots, Q, z \neq l\right\}$, $l = 1, \ldots, Q$, where $\#$ means number of times
3. Calculate the vector $\left\{r_d^m = 1 + \sum_{l=1}^{p} s_{l,m}\right\}$, $m = 1, \ldots, Q$,
4. Compute the vector $\mathbf{r}$ whose elements are $\left\{r^l = 1 + \#\left[\frac{(r_u^l + r_d^l)}{2} > \frac{(r_u^m + r_d^m)}{2}\right], m = 1, \ldots, Q, l \neq m\right\}$, $l = 1, \ldots, Q$.

The final partition P_K is the one occupying the first place in the achieved ranking. From this partition we then select only the k^* clusters for which the corresponding synthetic variables had a non-null coefficient in the previous Lasso regression model. In this way, we manage to select groups of informative variables and address the interval selection problem.

3. Simulation study

To evaluate the performances of this new proposal we performed a simulation study. In this study we adopted an R [18] package, namely hsdar, developed by Lehnert et al. (2018) [19] with the aim of allowing users to analyze and simulate hyperspectral data. In particular, the PROSPECT function allowed us to simulate reflectance spectra using the leaf reflectance model introduced by Jacquemoud and Baret [20] which links reflectance to a structure parameter, a pigment concentration, and water content. In particular, we took advantage of a version of this model proposed by Féret et al. [21], which involves multiple important pigments, such as chlorophyll, which have an impact on the optical properties of a leaf.

We considered 5 possible values of chlorophyll concentration (Cab), i.e. 10, 20, 30, 40 and 50 μg/cm^2, and for each of them r spectra were simulated, adding noise to each

spectrum by randomly generating an error term e for each wavelength, drawn from a Normal distribution with mean equal to 0 and standard deviation σ. It is worth noting that only 201 wavelengths were chosen inside the default interval $[400, 2500]$nm. The differences between spectra are therefore due to two main sources: chlorophyll concentration and noise. Given the strong relationship between the simulated reflectance and the chlorophyll concentration, we tried to build a calibration model to predict Cab beginning with the simulated spectra. Before training the calibration model, we also applied PLIS to see how the performances of the model change when using an informative subset of the 201 wavelengths rather then all of them.

Four different settings were therefore investigated:

- S_1: $r = 5$ and $\sigma = 0.005$ (i.e. moderate sample size and low amount of noise)
- S_2: $r = 5$ and $\sigma = 0.01$ (i.e. moderate sample size and moderate amount of noise)
- S_3: $r = 3$ and $\sigma = 0.005$ (i.e. small sample size and low amount of noise)
- S_4: $r = 3$ and $\sigma = 0.01$ (i.e. small sample size and moderate amount of noise)

In other words, we tried to vary the amount of noise and the sample size to see the impact of these two aspects on the capability of the PLIS procedure to improve the prediction error of the calibration model. 1000 simulation runs were performed. During each run, simulated data were divided into training and test sets (using a 67/33 ratio) and a ridge regression model and a Partial-Least Squares regression model were applied, firstly using all the explanatory variables and then only the ones selected by PLIS. At the end, the related Mean Absolute Prediction Errors (MAPEs) on the test set were retrieved and used for the final evaluation of the proposed interval selection technique. Indeed, a well performing interval selection method should allow us to substantially reduce the number of variables (i.e. wavelengths), but also improve the predictive performances of the chosen calibration model.

Looking at the first scenario (i.e. S_1), it is possible to see how the adoption of PLIS positively impacts the predictive performances of both considered calibration models (see Figure 1). Using PLIS, the average MAPE moves from 24.66% to 22.87% when ridge regression is adopted, while when PLS is applied, it goes from 14.33% to 10.46%. Additionally, on average, 112 wavelengths out of 201 were selected by our proposal, therefore data dimensionality is substantially reduced.

Increasing the amount of noise (i.e. S_2), the performances of the calibration models appear to worsen. However, the use of PLIS still seems to improve their predictive performances (see Figure 2). The average MAPEs when using ridge regression and PLS regression are equal to 26.20% and 15.80% respectively. On the other hand, when focusing on the selected informative regions, the average MAPE values become 24.87% and 12.10%. The average number of selected variables is also similar to the number achieved under S_1, i.e. 117.

Reducing the sample size, it appears that the observed MAPE values reasonably tend to increase, but the impact of interval selection also looks different. Under S_3, the achieved MAPE values are: 26.17% when using ridge regression and no interval selection; 17.97% when using PLS regression and no interval selection; 25.51% when using ridge regression and PLIS; and 13.72% when using PLS regression and PLIS. In other words, it appears that the predictive performances of the ridge regression model are only slightly improved by the use of interval selection (see Figure 3). The phenomenon appears to be even more evident under the fourth scenario (see Figure 4), where higher noise is present. Using PLIS together with ridge regression, MAPE moves from 27.48% to 27.09%, while with PLS it

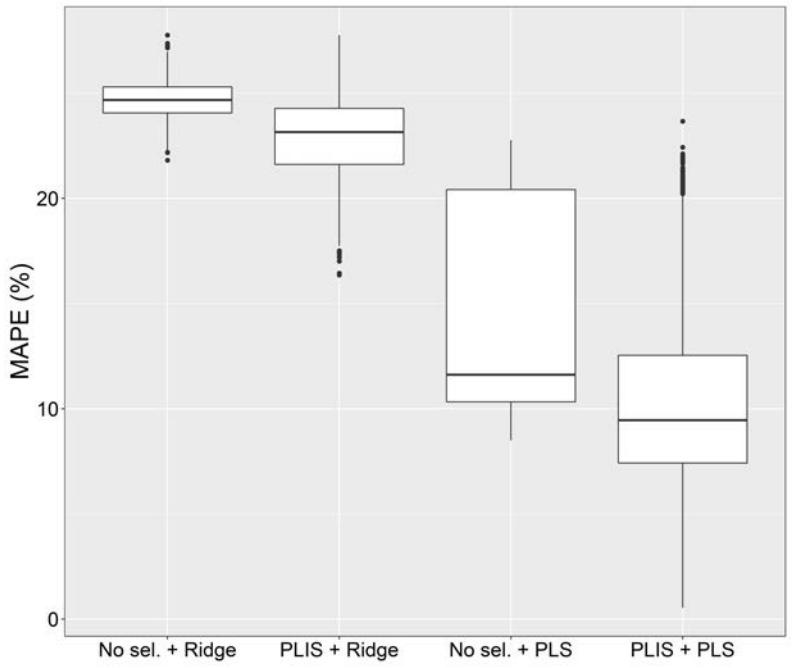

Figure 1: MAPE under S_1.

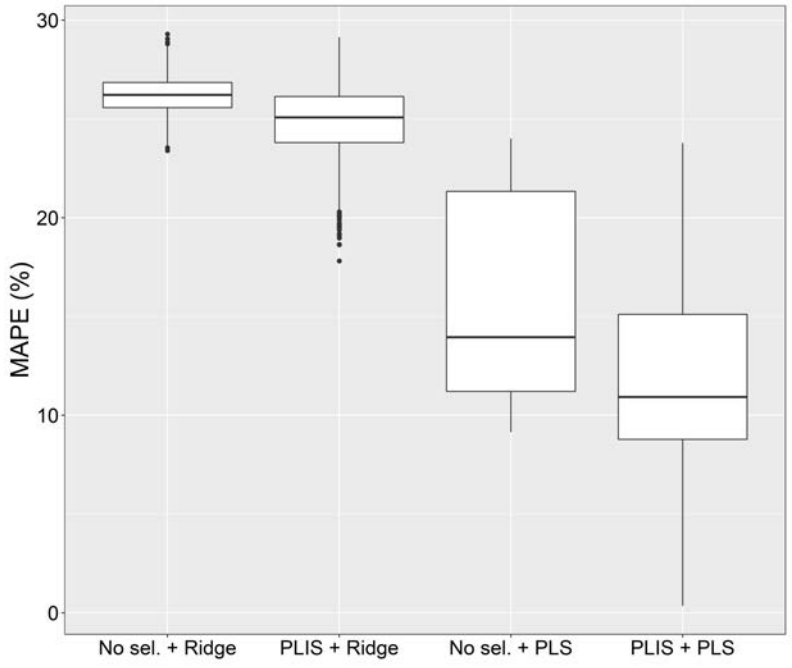

Figure 2: MAPE under S_2.

goes from 18.92% to 14.88%. Additionally, for both S_3 and S_4 the average number of selected variables increases (see Figure 5).

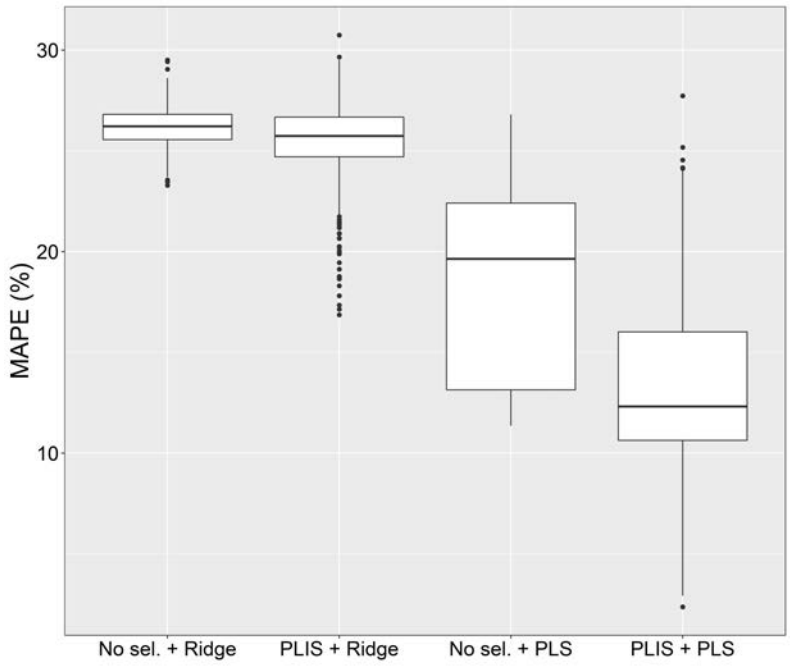

Figure 3: MAPE under S_3.

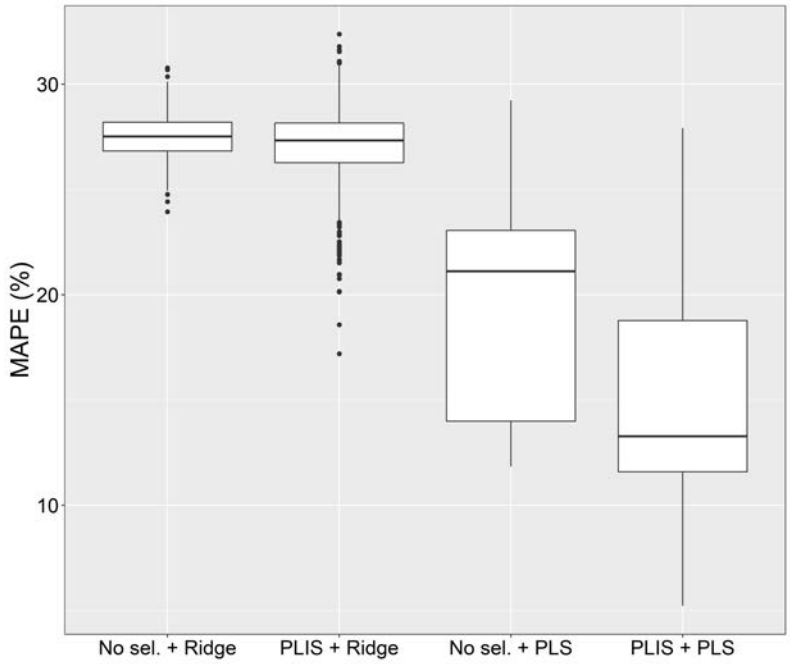

Figure 4: MAPE under S_4.

To sum up, when the total sample size is particularly low (i.e. 12 in our study) and data are noisy, PLIS capability to substantially improve the performances of the calibration model appears to depend on the nature of this model. However, the combined use of the widely-used PLS regression and PLIS is a reasonably good solution under these circum-

stances as well. Overall, the proposed interval selection procedure appears to be a useful method for application in fields such as NIR spectroscopy, where spectral data are used to predict a chemical or physical property of interest, to improve the predictive ability of calibration models.

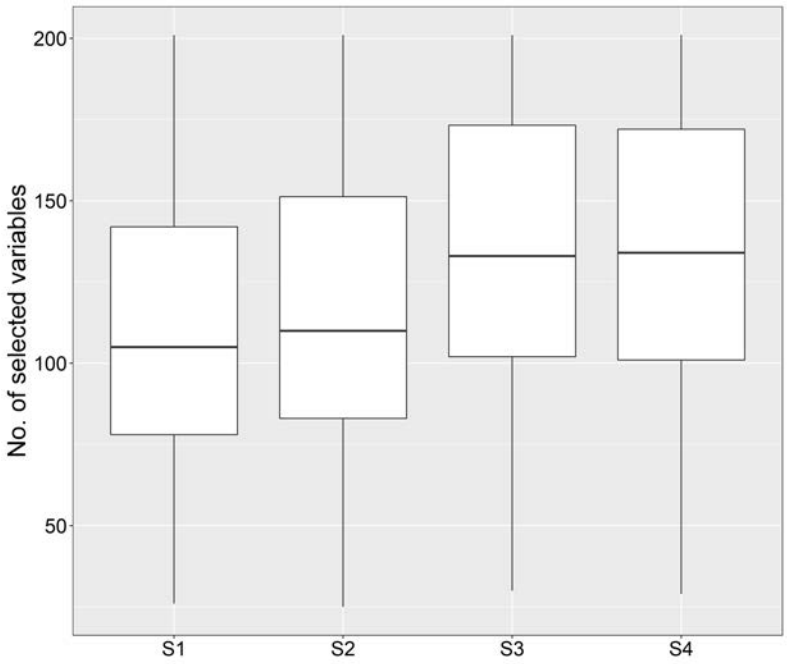

Figure 5: Selected variables under different scenarios.

4. A toy example

We then considered a toy example, exploiting data from a real industrial problem in which Near-Infrared Spectroscopy was adopted.
The original problem involved several different products, each characterized by a specific combination of 6 different chemical components. Each product was analyzed multiple times, so that r spectra was recorded for each sample. The objective of the analysis was to be able to predict the formula composition using an observed NIR spectrum.

Of the original V_{orig} wavelengths, only 86 were considered in our toy example. Additionally, we did not use all r spectra for each combination of components, but only the first one. 72 pre-treated spectra were thus available to be used to predict the amount of a single chemical component Y_1. We adopted PLIS to select the k^* most informative intervals of wavelengths and then considered two different possible calibration models: a ridge regression model and a Partial-Least Squares regression model. The performances of these models in predicting Y_1 in an appropriate test set were used as evaluation criteria for the proposed interval selection technique.

Firstly, out of the 72 spectra, only 54 were used for interval selection and calibration model training. The remaining 18 were considered a test set. We thus applied a ridge regression model and a Partial-Least Squares regression model on the whole dataset, conducting a grid search to tune the regularization parameter λ, as required by the first method, and the number of components, as required by PLS. The final considered values of these parameters were the ones minimizing the Root Mean Squared Error achieved using 5-fold

cross-validation. By evaluating the prediction error on the test set of the tuned models, we saw that PLS regression appears to outperform ridge regression (see Table 1).

Before applying PLIS, we decided to try 17 possible values of K within $[2, \ldots, 18]$ and set $\alpha = 0.05$. The interval selection procedure led us to select 10 clusters of variables for a total of 38 different wavelengths. A substantial reduction in the number of wavelengths to be considered in future applications would be possible using PLIS (see Figure 6). By again applying ridge regression and PLS regression on the subset of selected variables, it can be seen that the error metrics tend to substantially decrease with ridge regression (see Table 1). When PLS regression is used, the performances are almost identical under the two scenarios.

To sum up, from this toy example it emerges that the use of the proposed interval selection technique can benefit or at least preserve the predictive performances of the considered calibration models and strongly reduce the computational burden in future applications.

Table 1: Prediction errors on the test set.

Interval selection	Model	RMSE	MAE	MAPE
No	Ridge	0.0187	0.0156	56.6%
No	PLS	0.0094	0.0067	24.0%
Yes	Ridge	0.0093	0.0072	22.2%
Yes	PLS	0.0093	0.0074	23.7%

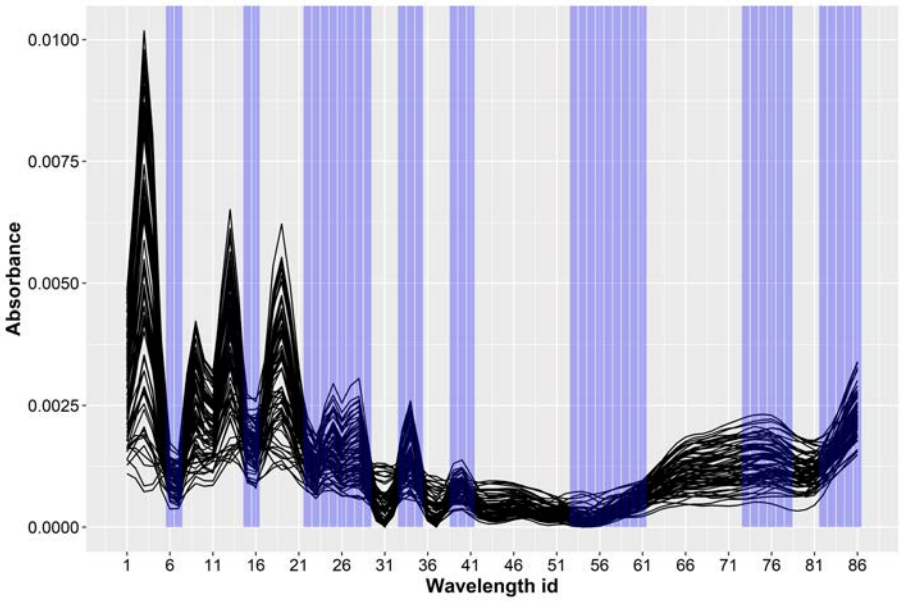

Figure 6: Spectra and selected wavelengths.

5. Conclusions

In this paper we proposed a new interval selection technique, i.e. a new method for selecting informative groups of variables. The method, called Permutation and Lasso-based Interval Selection (PLIS), starts by using the ClustOfVar algorithm [14] to accurately define a partition of V variables into K clusters. Given a set of possible partitions, the optimal one is identified taking advantage of Lasso regression and a permutation-based ranking proce-

dure [16]. The informative clusters of variables are then extracted from this partition using the coefficients of the previously trained Lasso regression model.

A simulation study was performed and a toy example was analyzed using PLIS. Both studies demonstrated the usefulness of the proposed procedure. Focusing only on the variables selected by PLIS does indeed appear to provide us with better performing calibration models in NIR spectroscopy problems. Additionally, PLIS appears to be able to greatly reduce the computational and operational burden by substantially reducing the number of regions and single wavelengths at which optical properties such as absorbance or reflectance need to be recorded.

References

[1] J. C. Bart, E. Gucciardi, and S. Cavallaro, "8 - quality assurance of biolubricants," in *Biolubricants* (J. C. Bart, E. Gucciardi, and S. Cavallaro, eds.), Woodhead Publishing Series in Energy, pp. 396 – 450, Woodhead Publishing, 2013.

[2] M.-Z. Zhu, B. Wen, H. Wu, J. Li, H. Lin, Q. Li, Y. Li, J. Huang, and Z. Liu, "The quality control of tea by near-infrared reflectance (nir) spectroscopy and chemometrics," *Journal of Spectroscopy*, vol. 2019, 2019.

[3] R. Balabin and S. Smirnov, "Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data," *Analytica Chimica Acta*, vol. 692, no. 1-2, pp. 63–72, 2011. cited By 269.

[4] Z. Xiaobo, Z. Jiewen, M. Povey, M. Holmes, and M. Hanpin, "Variables selection methods in near-infrared spectroscopy," *Analytica Chimica Acta*, vol. 667, no. 1-2, pp. 14–32, 2010. cited By 517.

[5] M. Manley and V. Baeten, "Chapter 3 - spectroscopic technique: Near infrared (nir) spectroscopy," in *Modern Techniques for Food Authentication (Second Edition)* (D.-W. Sun, ed.), pp. 51 – 102, Academic Press, second edition ed., 2018.

[6] C. Pasquini, "Near infrared spectroscopy: A mature analytical technique with new perspectives – a review," *Analytica Chimica Acta*, vol. 1026, pp. 8–36, 2018. cited By 120.

[7] D. Pirouz, "An overview of partial least squares," *SSRN Electronic Journal*, 10 2006.

[8] E. Vigneau, M. Devaux, E. Qannari, and P. Robert, "Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration," *Journal of Chemometrics*, vol. 11, no. 3, pp. 239–249, 1997. cited By 48.

[9] S. Zhang and Z. Zhao, "Feature selection filtering methods for emotion recognition in chinese speech signal," in *2008 9th international conference on signal processing*, pp. 1699–1702, IEEE, 2008.

[10] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2012.

[11] L. Nørgaard, A. Saudland, J. Wagner, J. Nielsen, L. Munck, and S. Engelsen, "Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy," *Applied Spectroscopy*, vol. 54, no. 3, pp. 413–419, 2000. cited By 884.

[12] B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren, and Y.-Z. Liang, "A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals," *Analyst*, vol. 140, no. 6, pp. 1876–1885, 2015. cited By 54.

[13] Y.-H. Yun, H.-D. Li, L. E. Wood, W. Fan, J.-J. Wang, D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 111, pp. 31–36, 2013. cited By 84.

[14] M. Chavent, V. Kuentz, B. Liquet, and L. Saracco, "Clustofvar: An r package for the clustering of variables," *arXiv preprint arXiv:1112.0295*, 2011.

[15] H. A. Kiers, "Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables," *Psychometrika*, vol. 56, no. 2, pp. 197–212, 1991.

[16] R. Arboretti, S. Bonnini, L. Corain, and L. Salmaso, "A permutation approach for ranking of multivariate populations," *Journal of Multivariate Analysis*, vol. 132, pp. 39 – 57, 2014.

[17] F. Pesarin and L. Salmaso, *Permutation tests for complex data: theory, applications and software*. Wiley, 2010.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[19] L. W. Lehnert, H. Meyer, W. A. Obermeier, B. Silva, B. Regeling, and J. Bendix, "Hyperspectral data analysis in r: The hsdar package," *arXiv preprint arXiv:1805.05090*, 2018.

[20] S. Jacquemoud and F. Baret, "Prospect: A model of leaf optical properties spectra," *Remote sensing of environment*, vol. 34, no. 2, pp. 75–91, 1990.

[21] J.-B. Féret, A. Gitelson, S. Noble, and S. Jacquemoud, "Prospect-d: Towards modeling leaf optical properties through a complete lifecycle," *Remote Sensing of Environment*, vol. 193, pp. 204–215, 2017.