Massive spectral data analysis for plant breeding using parSketch-PLSDA method: Discrimination of sunflower genotypes

Maxime Ryckewaert, Maxime Metz, Daphné Héran, Pierre George, Bruno Grèzes-Besset, Reza Akbarinia, Jean Michel Roger, Ryad Bendoula

Massive spectral data analysis for plant breeding using parSketch-PLSDA method: discrimination of sunflower genotypes

Maxime Ryckewaert[a,b], Maxime Metz[a,b], Daphné Héran[a], Pierre George[c], Bruno Grèzes-Besset[c], Reza Akbarinia[d], Jean-Michel Roger[a,b], Ryad Bendoula[a]

[a] *ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*
[b] *ChemHouse Research Group, Montpellier, France*
[c] *Innolea, 6 chemin des Panedautes, 31700 Mondonville, France*
[d] *Inria & LIRMM, Univ Montpellier, France*

Abstract

In precision agriculture and plant breeding, the amount of data tends to increase. This massive data is becoming more and more complex, leading to difficulties in managing and analysing it. Optical instruments such as NIR Spectroscopy or hyperspectral imaging are gradually expanding directly in the field, increasing the amount of spectral database. Using these tools allows access to non-destructive and rapid measurements to classify new varieties according to breeding objectives. Processing this massive amount of spectral data is challenging. In a context of genotype discrimination, we propose to apply a method called parSketch-PLSDA to analyse such a massive amount of spectral data. ParSketch-PLSDA is a combination of an indexing strategy (parSketch) and the reference method (PLSDA) for predicting classes from multivariate data. For this purpose, a spectral database was formed by collecting 1,300,000 spectra generated from hyperspectral images of leaves of four different sunflower genotypes. ParSketch-PLSDA is compared to a PLSDA. Both methods use the same set of calibration and test. The prediction model obtained by PLSDA has a classification error close to 23% on average across all genotypes. ParSketch-PLSDA method outperforms PLSDA by greatly improving prediction qualities by 10%. Indeed, the model built with ParSketch-PLSDA has the ability to take into account non-linearities among data sets. These results are encouraging and allow us to anticipate the future bottleneck related to the generation of a large amount of data from phenotyping.

Keywords: Spectroscopy, Massive data, Digital Agriculture, Precision Agriculture, Chemometrics

1. Introduction

In recent years, precision agriculture and plant breeding have tended to increase the quantity and complexity of phenotyping related data (Mahlein, 2016; Tripodi et al., 2018; Awada et al., 2018). Managing and analysing huge amounts of data are identified as a future bottleneck in phenotyping (Tripodi et al., 2018). Indeed, over the last few years, high throughput phenotyping (HTP) platforms in the laboratory or directly in the field have been flourishing (Chawade et al., 2019; Shakoor et al., 2017). These platforms provide a monitoring of one or more phenotypic traits of the vegetation. This information can be obtained at different spatial, spectral or temporal scales depending on the studied level, which could be vegetation organ, individual or even population (Dhondt et al., 2013; Mahlein, 2016; Mutka et al., 2016). The higher the spatial, spectral or temporal resolution, the larger the amount of data.

Spectroscopy in the visible and near-infrared range (VIS-NIR) has proven to be relevant for providing useful information for vegetation monitoring. Several plant phenotyping issues can be tackled with high spectral resolution measurements such as biochemical variable access (Vigneau et al., 2011; Jay et al., 2017), disease (Lu et al., 2018; Yu et al., 2018) or stress detection (Behmann et al., 2014; Christensen et al., 2005).

From a technological point of view, spectral acquisitions directly in the field have been made possible thanks to spectrometer miniaturisation (Yan and Siesler, 2018; Beć et al., 2020) or hyperspectral imager evolution (Mishra et al., 2020; Fiorani and Schurr, 2013). Associated with mobile vectors (such as UAV, tractor, pedestrian), these tools become HTP instruments and generate a large amount of spectral data. However, simple computations on this amount of data such as outlier detection or the use of pre-processing become difficult to perform and very time consuming (Szymańska, 2018). Processing this massive amount of spectral data is challenging.

In chemometrics, most popular methods as Partial Least Square (PLS) (Wold et al., 2001) are based on an assumption of linear relationship between spectral data and specific variables (Mark and Workman, 2007). These methods are popular because of their good predictive performances and low computation time. Conversely, using these methods may not provide good prediction models when relationships between spectra and variable of interest are non-linear.

When dealing with a large amount of spectral data, complex structures and non-linear relationships can arise which may compromise linear regression approaches (Dardenne et al., 2000). In practice, using a linear classification or regression to predict a complex database would lead to degraded results (Bertran et al., 1999; Ni et al., 2014). As a consequence, linear methods are challenged on large amounts of data. Furthermore, some methods, called local methods, exploit data based on a restricted neighbourhood of individuals which greatly improves prediction quality (Dardenne et al., 2000; Pérez-Marín et al., 2007; Davrieux et al., 2016; Naes et al., 1990). These methods can be used to overcome non-linearity problems under the assumption that with a restricted neighbourhood, the relationship between spectra and variables becomes linear. The parSketch-PLSDA method has recently been proposed to implement a local approach to a large volume of data (Metz et al., 2020). Therefore, parSketch-PLSDA can be used to address the complex analysis of large amount of spectral data from phenotyping.

In this paper, we propose to study the use of the parSketch-PLSDA method to exploit a large amount of spectral data, generated from hyperspectral images of leaves of four different sunflower genotypes. Additionally, we compare this method with a reference method in an application of discrimination of different sunflower varieties.

2. Materials and methods

2.1. Biological material

Four sunflower genotypes (called A, B, C and D) were grown in a greenhouse at INRAE France in well-watered conditions by using a flood and drain system refilling water every 48 hours. All pots used the same potting soil (Pot Clay coarse, Floradur, Floragard). Water and lighting conditions were similar for each pot with a day-night cycle of 16h/8h. The temperature was 25°C with a relative humidity in the greenhouse ranging from 50% to 60%. The greenhouse was equipped with multispectral lighting (450 nm, 560 nm, 660 nm, 730 nm and 6000°K) controlled by Herbro automaton (GreenHouseKeeper) with Photosynthetically Active Radiation (PAR) set at 400 µmol/m2/s.

For the four selected genotypes, two potted plants (called P1 and P2) of each were grown. Four leaves were collected at the upper and middle parts of each plant, except for the genotype D where only two leaves of each plant were collected. Leaf petioles were immediately wrapped with water soaked paper before measurements. In total, 28 leaves were then collected and measured.

2.2. Spectral acquisitions

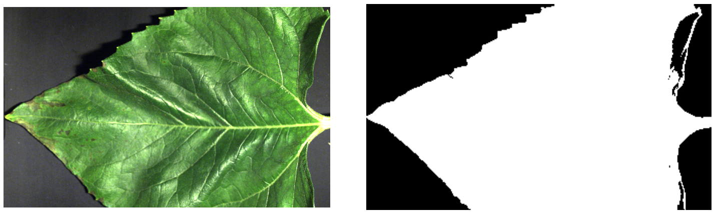
Spectral data of the prepared leaf samples were acquired in the reflectance mode by using a laboratory-based line scanning Hyperspectral Imaging System (HIS). The HIS system was composed of a linear halogen light (Haloline, Osram, 150 W), a translation rail (Linear Unit LES 4, Iselautomation, Germany), and a detection system. The sample was placed on a translation rail, synchronised with the acquisition software (NEO Hyspex, Norsk Elektro Optikk AS) which can record images when sample was scanned under the hyperspectral camera (NEO Hyspex VNIR-1600 with 30cm-objective, Norsk Elektro Optikk AS, Skedsmokorest, Norway). Spectral data were acquired in the $400 - 1000$ nm wavelength range with 3.7 nm intervals.

For each sample, the reflected light intensity ($I_s(\lambda)$) was measured at each wavelength . Dark current image ($I_b(\lambda)$) was also recorded for each measure. A white reference (SRS99, Spectralon ®) was used as a reference

($I_o(\lambda)$) to standardize images from non-uniformities of all components of the instrumentation (light source, lens, detector). From these measurements, reflectance ($R_s(\lambda)$) was calculated for each sample:

$$R_s(\lambda) = \frac{I_s(\lambda) - I_b(\lambda)}{I_0(\lambda) - I_b(\lambda)} \tag{1}$$

For all hyperspectral images, vegetation pixels were selected to form a spectral data set. This selection was made by a threshold procedure (Fig. 1). Indeed, vegetation and background pixels were easily identified by comparing their reflectance value at 800 nm to a threshold defined here at 30%. Leaf spectra were collected from the 28 hyperspectral images, representing more than 1,300,000 spectra.



(a)

(b)

Figure 1: Pixel selection from a hyperspectral image of a sunflower leaf (a) image, (b) mask based on threshold values

2.3. Data analysis

Two methods were used to compare their ability to discriminate sunflower genotypes. Both methods were applied to a similar data set, called test set, built out of the spectra database. Calculations were performed with the R software (version 3.6.1 (Core Team, 2013)) and rnirs package (https://

github.com/mlesnoff/rnirs) was used for classical discrimination methods (PLSDA).

2.3.1. PLSDA method

The Partial Least Squares for Discrimination Analysis (PLSDA) (Barker and Rayens, 2003) was used as reference method for classification. This method consisted of building models between multivariate data and a vector coding different classes (here, the four genotypes).

Multivariate data was represented by a matrix $\mathbf{X}$ of size $(n,\ p)$ where n was the observation number and p the variable number. The n observations were identified by their corresponding class in the vector y of size $(n,1)$ where values ranged from 1 to q, where q was the class number. The first step was to transform y into a dummy matrix $\mathbf{Y}$ of size $(n,\ q)$ also called disjunctive table.

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{bmatrix} \rightarrow \mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

An example of a dummy matrix is given in equation 2 with nine observations belonging to three classes. The matrix $\mathbf{Y}$ contains binary values (0,1) where each column corresponded to a class. For a given observation, the class-corresponding column has a value of 1 while other columns were equal to 0.

Then, a Partial Least Square (PLS) model (Wold et al., 2001) was applied between $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{Y}$ being multidimensional, the algorithm PLS2 adapted to the prediction of several responses was used. Finally, a linear discriminant analysis (LDA) (Fisher, 1936) was applied between the PLS2 scores and $\mathbf{Y}$.

2.3.2. ParSketch-PLSDA method

The other strategy was to apply the parSketch-PLSDA method (Metz et al., 2020), an extension of the K-Nearest Neighbours (KNN)-PLSDA method for massive data processing (Lesnoff et al., 2020). ParSketch-PLSDA was used to combine an indexation strategy (parSketch) and the PLSDA. An approximation of the neighbourhood was defined for each spectrum to be classified. This neighbourhood was then used to compute a PLSDA model and to predict which class belong new spectra.

ParSketch was performed in three steps: dimension reduction, grid creation, neighbourhood approximation. Three method parameters (v, s, m) were defined, corresponding to these three steps, and are described below.

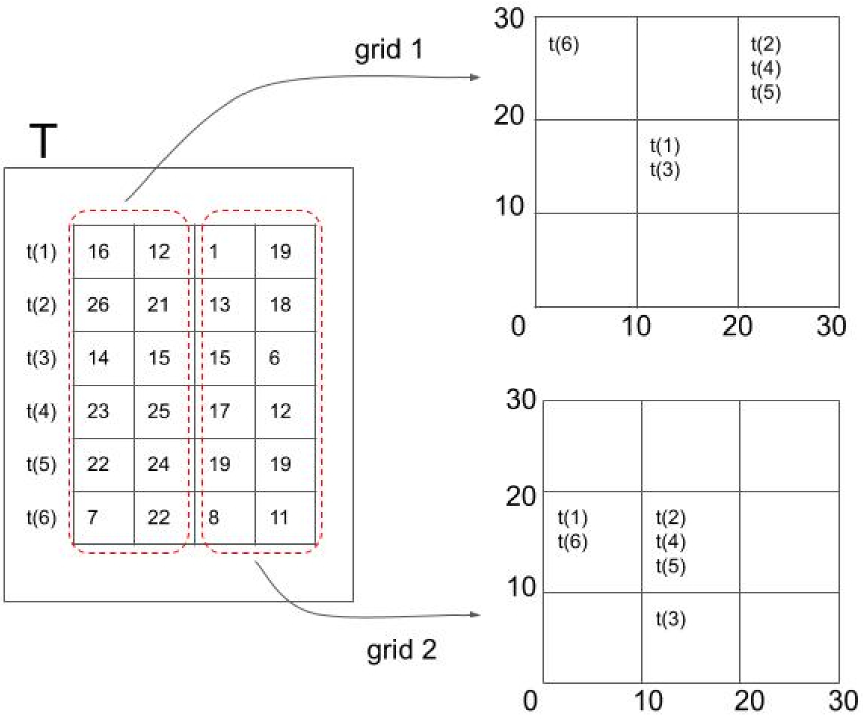
First, a dimension reduction was achieved by calculating the matrix $\mathbf{T}$ corresponding to the sketch of the matrix $\mathbf{X}$ as follows:

$$\mathbf{T} = \mathbf{X}\mathbf{P} \tag{3}$$

Where $\mathbf{P}$ was a matrix of size (p,v) containing values of -1 or 1 according to a random selection. The first parameter of ParSketch (v), corresponding to the column number of $\mathbf{P}$ was then defined. The higher the value of v

the better the approximation of the neighbourhood. However, the larger the value of v the longer the parSketch method computation time.

The second step corresponding to the grid creation process (see Fig. 2) was to segment the space (2d) formed by adjacent pairs of **T** columns. The number of segments (s) is the second parSketch parameter. The higher the value of s the better the approximation of the nearest neighbours. However, the greater the value of s, the smaller the number of neighbours.



t(1)	16	12	1	19
t(2)	26	21	13	18
t(3)	14	15	15	6
t(4)	23	25	17	12
t(5)	22	24	19	19
t(6)	7	22	8	11

Figure 2: An illustrated example of grid creation with a segment number $s = 3$

The last step to configure parSketch was to define the minimal number m of grids returned in the neighbour's search. This step corresponded to the neighbourhood approximation for the grid search process (see Fig. 3). The higher the value of m the better the approximation of the nearest neighbours. Observations present in the same cell for at least m grid number are selected as neighbours of the individual to be predicted. However, the greater the value of m, the smaller the number of neighbours returned by parSketch method.

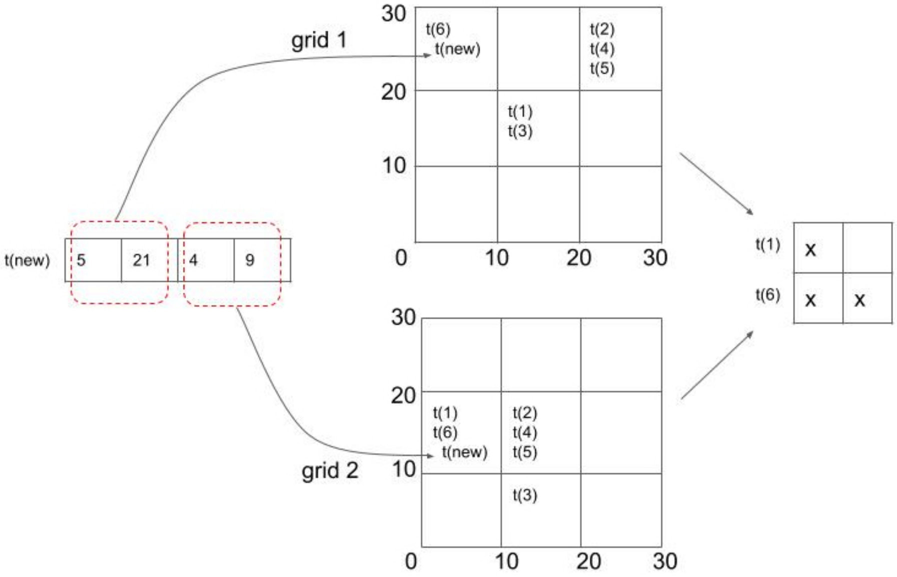


Figure 3: An illustrated example of grid search. For a new measure $t(new)$ and a m value equals to 2, $t(6)$ will be returned as a neighbour because it is present in two grids next to $t(new)$, whereas $t(1)$ will not be considered as a neighbour

2.4. Evaluation strategies and method parameterization

The data set was divided into two independent data sets: a calibration set and an independent test set . The calibration set was formed with the 14 images acquired on P1 plants and corresponding to about 650,000 spectra. The PLSDA model and parSketch-PLSDA method were both calibrated using all spectra of this calibration set.

The test set was formed with the other 14 images acquired on P2 plants (independent from P1). 1000 spectra were randomly selected in each image totalling 14,000 spectra for the test set. Spectra that could not be predicted by parSketch due to lack of neighbours were removed from the test set. In the end, the same test set were used for parSketch-PLSDA and for PLSDA.

For both methods, validation steps were performed on the calibration set in order to minimize overfitting.

To build the PLSDA model, the cross-validation step consisted of splitting the calibration data set into different blocks in order to calculate calibration and validation errors. This approach, also called k-fold validation (Wold, 1978; Camacho and Ferrer, 2012) was carried out with five blocks repeated three times. Validation errors were then computed and led to the number of latent variables to be retained.

For parSketch-PLSDA, a parametrisation step was performed to configure the three parSketch parameters. This step was performed by analyzing distributions of returned neighbours according to two parSketch parameters: number of segments s and the common minimum grids m. Here, the number of random vectors v was set to a value of 20. Afterwards, a PLSDA model was established. The number of latent variables was optimized for a subset of the calibration set, called the validation set. This validation set was formed with four images of the calibration set by randomly selecting 1000 spectra in each image.

In order to compare both methods, confusion matrices were obtained and percentages of precision and recall were calculated according to the following equations:

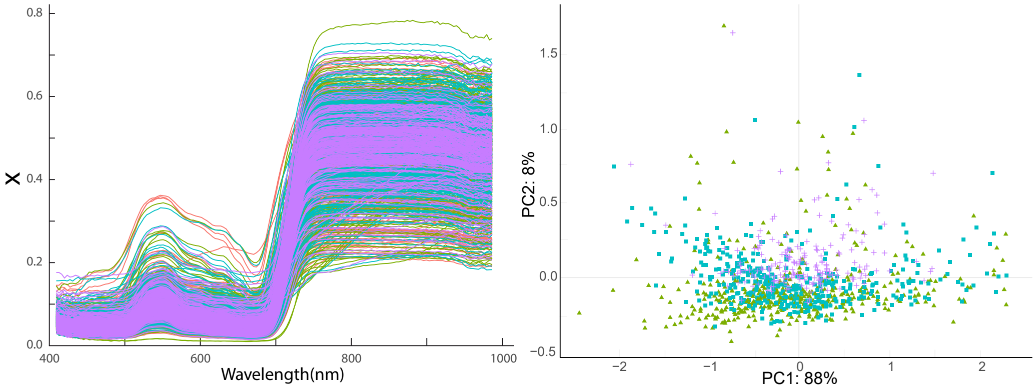
$$\text{Precision} = \frac{tp}{tp + fp} \tag{4}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{5}$$

Where tp, fp and fn corresponded to true positives, false positives and false negatives respectively. On the one hand, for a given class, precision value assessed the predictive quality of the model based on the proportion of well-classified observations among all observations that were classified in the same corresponding class. On the other hand, recall, also called sensitivity, evaluated the number of well-classified observations compared to the total number of observations of the given class. These two criteria are complementary to evaluate the model performances. These two figures of merit were expressed as percentages. The higher the values, the better the model performance.

3. Results and discussion

3.1. Data visualization



(a)

(b)

Figure 4: All data set including Genotype A (red), B (green), C (cyan) and D (violet): (a) spectra, (b) score plot of the first two principal components

Reflectance spectra shown in Fig. 4a correspond to 1000 spectra per class randomly selected among all the data set. These spectra correspond to vegetation spectra (Xu et al., 2019) : specific hollows at 450 nm and 650 nm related to chlorophyll content, anthocyanin content at 550 nm; the red-edge towards 780 nm and a plateau in the near-infrared between 780 nm and 1000 nm. Besides, the main observed variability in the spectrum plot corresponds to an additive effect due to the scattering effect of the structure of the leaves. However, the number of spectra is too large to be able to describe difference between classes.

A principal component analysis was applied to these spectra. Figure 4b shows the score plot of the two first components. The first component represents 88% of the spectra variability and 8% for the second component. On these two components, scores are uniformly distributed without any evident distinction between genotypes. The exploratory study of the spectra shows that there are no outliers and that there is no distinct group on the first two components.

3.2. Model calibration

3.2.1. PLSDA

Figure 5 shows the cross-validated error rate curve for PLSDA applied to all spectra of the calibration data set. The behaviour of the curve decreases continuously according to the latent variable (LV) number.

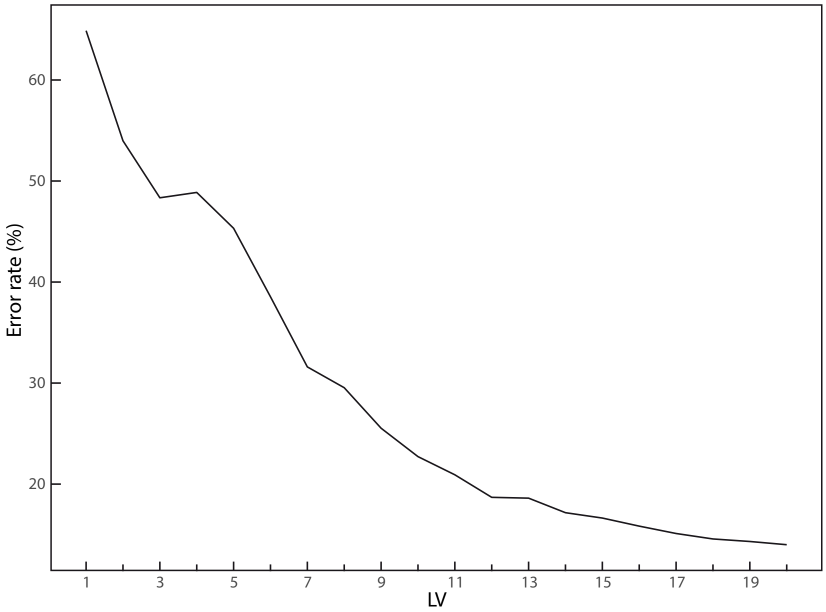


Figure 5: Evolution of the cross-validated error rate as a function of latent variables (LV) for the PLSDA applied to all spectra of the calibration data set

A high value of LV number generally shows the complex structure of a data set. This is expected with spectral measurements on vegetation (Metz et al., 2020). With 16 LVs, an error rate with a value close to 12% is obtained. However, after 16 LVs the predictive performance gain is very small. Consequently, the PLSDA model is set to 16 LVs.

3.2.2. ParSketch-PLSDA

ParSketch parameters s and m are studied according to the statistical distribution of the number of returned neighbours (Fig. 6).

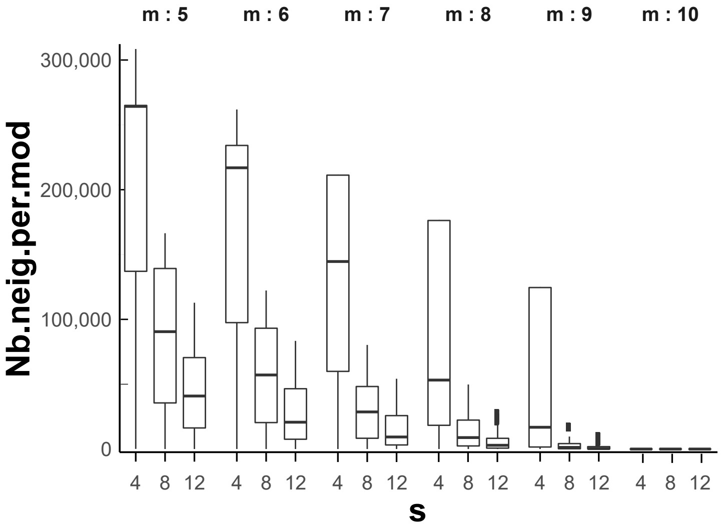


Figure 6: Distribution of the returned neighbours according to parSketch parameters s and m (number of segments and common minimum grids). Here, v (number of random vectors) parameter is fixed to 20

The number of neighbours decreases to a value close to zero when parameter values (s,m) increase. Indeed, on the one hand, an increase of number of segments s, the number of returned neighbours will be lower. And on the other hand, by increasing the minimum number of common grids m, the risk of not having neighbours is high. This global trend is expected (Metz et al., 2020).

By contrast, when parameter values are low, the number of returned neighbours is high (close to 300 000 neighbours by individual to be predicted). This situation is not desirable, as it may cause problems related to computation time constraints. As a result, parameters m and s must be chosen to have a sufficient number of neighbours, neither too much nor too little. As several values are possible, four combinations of the parSketch parameters are selected (Table 1) to compare their model performances.

Table 1: Combinations of the selected parSketch parameters and the corresponding median number of returned neighbours

Combination	m	v	s	Median neighbour number
(a)	9	20	8	1246
(b)	8	20	12	2903
(c)	7	20	12	9303
(d)	6	20	12	27030

Table 1 shows the retained values of the three parSketch parameters for these four combinations. The combination **(a)** was selected because the median number of neighbours is 1246. This low number of neighbours enables to quickly calibrate PLSDA models but it could be insufficient to have a good predictive quality. Indeed, a low median number of neighbours means that a large amount of observations do not have neighbour at all. For the combinations **(b)** and **(c)**, higher numbers of neighbours are returned, with median values of 2903 and 9303 respectively. Finally, the highest median number of neighbours returned by parSketch is chosen with the combination

(d) with a value equal to 27030. In this case, constraints in computation time might appear. Moreover, the linear relationship between spectra and class variable of a small neighbourhood might be lost.

Validation error curves for the four retained combinations for parSketch-PLSDA are shown in the figure 7.

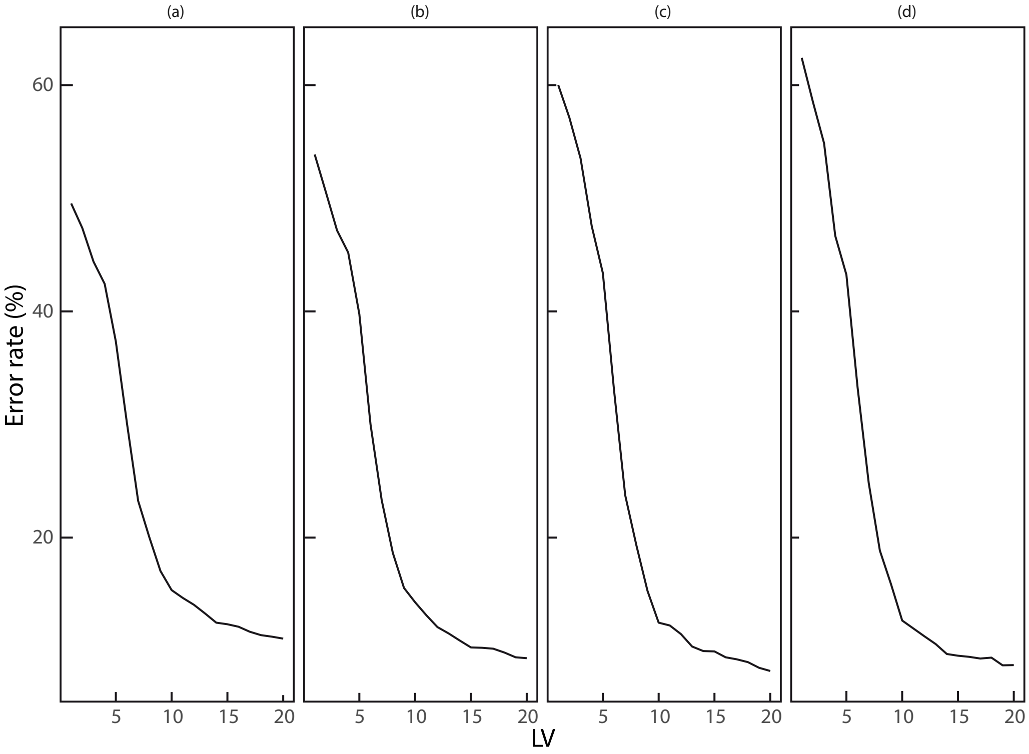


Figure 7: Evolution of the validation error rate as a function of latent variables for the four parameter combinations of parSketch-PLSDA

With higher error values, the combination **(a)** is less predictive than other parameter combinations. As expected, this combination having the smallest number of neighbours, the resultant model has poorer predictive performances than the three other ones.

The error curve obtained with the combination **(b)** reaches lower rates than the combination **(a)** curve. This means that the predictive capabilities of the model can be improved by slightly increasing the number of neighbours. The combination **(c)** has best predictive performance for the validation set with lowest values of classification error. The combination **(d)** has lower prediction quality than the combination **(c)** for a larger median number of neighbours per sample to be predicted (cf. Table 1).

The model with the lowest predictive quality has a high number of neighbours. This degradation reflects a non-linear aspect of the data set. By further increasing the number of returned neighbours, prediction qualities of parSketch will be close to the PLSDA method on the whole data set.

Finally, for this validation set the optimal parameter combination is the combination **(c)**. In this case, the number of latent variables is not easy to define. The number of latent variables is defined by a trade-off between the size of the model and the benefit of adding an extra dimension to the model. The number of latent variables chosen is therefore 16.

3.3. Model testing

The PLSDA model has been calibrated with all the spectra of the calibration set. Then this model is applied to the test set defined previously and its prediction performances are assessed in Table 2.

Table 2: Confusion matrix for PLSDA (16 LVs)

	A	**B**	**C**	**D**	**Recall (%)**
A	3120	121	339	250	**81**
B	238	2812	619	154	**74**
C	127	241	3463	75	**89**
D	283	173	270	1213	**63**
Precision(%)	**83**	**84**	**74**	**72**	

Precision and recall values are high for all classes A, B, C and D with values ranging from 72% to 84% for precision and from 63% to 89% for recall. Genotypes A and B have the highest precision values with values of 83% and 84%, respectively. This means that 83% of spectra classified in genotype A, actually belong to genotype A. Few other genotypes are found in this class. The same argumentation holds true for 84% of genotype B spectra. For recall, genotypes A and C have the best values with 81% and 89%, respectively. For these genotypes, spectra are mainly well-classified that is infrequently assigned to other classes.

The percentage missings from recall values correspond to the prediction error for each class. The prediction error of the whole data set, corresponding to the average error, is close to 23%. It is expected to have value for the test error slightly higher than the 12% observed during calibration (see Fig. 5). This means that the calibration set samples are representative of the test set despite their independence (as mentioned above, the test set corresponds to other plants of the same genotype).

Table 3: Confusion matrix for parSketch-PLSDA with combination**(c)** ($m = 7$, $v = 20$, $s = 12$)

	A	**B**	**C**	**D**	**Recall (%)**
A	3547	114	115	54	**93**
B	40	3256	473	54	**85**
C	58	211	3590	47	**92**
D	51	112	260	1516	**78**
Precision(%)	**96**	**88**	**81**	**91**	

Table 3 shows the parSketch-PLSDA prediction performance by giving percentages of precision and recall for each genotype. Genotypes A and D have the highest precision values with values of 96% and 91% respectively. Besides, genotypes A and C have the highest recall values with values of 93% and 92% respectively. Genotype D has low recall values with both methods (63% for PLSDA and 78% for parSketch-PLSDA). This is probably due to the under-representation in the data set which may degrade the model calibration. Indeed, only two images were acquired per plant of genotype D compared to four images for the other genotypes.

Finally, overall recall and precision values have increased by almost 10% with parSketch-PLSDA (83% and 87% respectively) compared to PLSDA (76% and 77% respectively). Consequently, the model prediction error decreases to a value of 13%. This implies that parSketch-PLSDA model performs better than the reference discriminant strategy. This improvement in the classification results demonstrates the advantage of using a limited number of neighbours to create a model.

As the methods used are locally linear, this improvement confirms the hypothesis that with a limited number of neighbours, the problem becomes linear. The prediction improvement obtained with parSketch-PLSDA method highlights the presence of non-linear relationships between spectra and a class variable in the whole data set. which can be encountered when building a large spectral database.

4. Conclusion

In this study, we compared the two classification strategies on the same calibration and test data sets.

For both methods, classification results are encouraging and confirm the interest of VIS-NIR spectroscopy for variety discrimination. Results showed that parSketch-PLSDA method outperforms PLSDA by improving prediction qualities by 10%. The use of the parSketch-PLSDA procedure in the exploitation of massive spectral data is confirmed and shows the interest of using a close neighbourhood of the spectra to be predicted.

It would be interesting to test such methods on a larger number of genotypes. This increase in the spectral database can potentially lead to an increase in complexity hence reducing the data set quality. Therefore, it would be interesting, in perspective, to evaluate other methods dealing with non-linearity.

In the framework of plant breeding, hyperspectral imaging or field microspectrometers as tools for high-throughput plant phenotyping could be considered in real time with this method. In an applicative aspect, parSketch procedure is parallelisable, which shows the possibility of fast real-time prediction of a large amount of data. We used parSketch-PLSDA on spectral data for close-range plant phenotyping. Other applications to plant breeding (disease, biotic/abiotic stress) or other applications related to precision agriculture could be considered. More generally, this method can be applied to any other application in analytical chemistry or metabolomics.

Acknowledgement

This work was conducted within the OPTIPAG Project supported by the grant ANR-16-CE04-0010 from the French Agence Nationale de la Recherche.

References

Anne-Katrin Mahlein. Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping. *Plant disease*, 100(2):241–251, 2016. Publisher: Am Phytopath Society.

Pasquale Tripodi, Daniele Massa, Accursio Venezia, and Teodoro Cardi. Sensing technologies for precision phenotyping in vegetable crops: current status and future challenges. *Agronomy*, 8(4):57, 2018. Publisher: Multidisciplinary Digital Publishing Institute.

Lana Awada, Peter W. B. Phillips, and Stuart J. Smyth. The adoption of automated phenotyping by plant breeders. *Euphytica*, 214(8):148, August 2018. ISSN 0014-2336, 1573-5060. doi: 10.1007/s10681-018-2226-z. URL http://link.springer.com/10.1007/s10681-018-2226-z.

Aakash Chawade, Joost van Ham, Hanna Blomquist, Oscar Bagge, Erik Alexandersson, and Rodomiro Ortiz. High-throughput field-phenotyping tools for plant breeding and precision agriculture. *Agronomy*, 9(5):258, 2019. Publisher: Multidisciplinary Digital Publishing Institute.

Nadia Shakoor, Scott Lee, and Todd C. Mockler. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current opinion in plant biology*, 38:184–192, 2017. Publisher: Elsevier.

Stijn Dhondt, Nathalie Wuyts, and Dirk Inzé. Cell to whole-plant phenotyping: the best is yet to come. *Trends in Plant Science*, 18(8): 428–439, August 2013. ISSN 1360-1385. doi: 10.1016/j.tplants.2013.04.008. URL http://www.sciencedirect.com/science/article/pii/S1360138513000812.

Andrew M. Mutka, Sarah J. Fentress, Joel W. Sher, Jeffrey C. Berry, Chelsea Pretz, Dmitri A. Nusinow, and Rebecca Bart. Quantitative, image-based phenotyping methods provide insight into spatial and temporal dimensions of plant disease. *Plant Physiology*, page pp.00984.2016, July 2016. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.16.00984. URL http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.00984.

Nathalie Vigneau, Martin Ecarnot, Gilles Rabatel, and Pierre Roumet. Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in wheat. *Field Crops Research*, 122(1):25–31, April 2011. ISSN 03784290. doi: 10.1016/j.fcr.2011.02.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0378429011000451.

Sylvain Jay, Nathalie Gorretta, Julien Morel, Fabienne Maupas, Ryad Bendoula, Gilles Rabatel, Dan Dutartre, Alexis Comar, and Frédéric Baret.

Estimating leaf chlorophyll content in sugar beet canopies using millimeter-to centimeter-scale reflectance imagery. *Remote Sensing of Environment*, 198:173–186, 2017. Publisher: Elsevier.

Jinzhu Lu, Reza Ehsani, Yeyin Shi, Ana Isabel de Castro, and Shuang Wang. Detection of multi-tomato leaf diseases (late blight , target and bacterial spots) in different stages by using a spectral-based sensor. *Scientific Reports*, 8(1):2793, February 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-21191-6. URL https://www.nature.com/articles/s41598-018-21191-6. Number: 1 Publisher: Nature Publishing Group.

Kang Yu, Jonas Anderegg, Alexey Mikaberidze, Petteri Karisto, Fabio Mascher, Bruce A. McDonald, Achim Walter, and Andreas Hund. Hyperspectral Canopy Sensing of Wheat Septoria Tritici Blotch Disease. *Frontiers in Plant Science*, 9, 2018. ISSN 1664-462X. doi: 10.3389/fpls.2018.01195. URL https://www.frontiersin.org/articles/10.3389/fpls.2018.01195/full. Publisher: Frontiers.

Jan Behmann, Jörg Steinrücken, and Lutz Plümer. Detection of early plant stress responses in hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:98–111, 2014. URL http://www.sciencedirect.com/science/article/pii/S092427161400094X.

Lene K. Christensen, Shrinivasa K. Upadhyaya, Bernie Jahn, David C. Slaughter, Eunice Tan, and David Hills. Determining the Influence of Water Deficiency on NPK Stress Discrimination in Maize using Spectral and Spatial Information. *Precision Agriculture*, 6(6):539–550, December 2005. ISSN 1573-1618. doi: 10.1007/s11119-005-5643-7. URL https://doi.org/10.1007/s11119-005-5643-7.

Hui Yan and Heinz W. Siesler. Hand-held near-infrared spectrometers: State-of-the-art instrumentation and practical applications. *NIR news*, 29(7): 8–12, 2018. Publisher: SAGE Publications Sage UK: London, England.

Krzysztof B. Beć, Justyna Grabska, Heinz W. Siesler, and Christian W. Huck. Handheld near-infrared spectrometers: Where are we heading? *NIR news*, 31(3-4):28–35, 2020. Publisher: SAGE Publications Sage UK: London, England.

Puneet Mishra, Santosh Lohumi, Haris Ahmad Khan, and Alison Nordon. Close-range hyperspectral imaging of whole plants for digital phenotyping: Recent applications and illumination correction approaches. *Computers and Electronics in Agriculture*, 178:105780, November 2020. ISSN 01681699. doi: 10.1016/j.compag.2020.105780. URL https://linkinghub.elsevier.com/retrieve/pii/S016816992031869X.

Fabio Fiorani and Ulrich Schurr. Future Scenarios for Plant Phenotyping. *Annual Review of Plant Biology*, 64(1):267–291, April 2013. ISSN 1543-5008, 1545-2123. doi: 10.1146/annurev-arplant-050312-120137. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-arplant-050312-120137.

Ewa Szymańska. Modern data science for analytical chemical data – A comprehensive review. *Analytica Chimica Acta*, 1028:1–10, October 2018.

ISSN 0003-2670. doi: 10.1016/j.aca.2018.05.038. URL http://www.sciencedirect.com/science/article/pii/S0003267018306421.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

Howard Mark and Jerry Workman. *Chemometrics in spectroscopy*. Elsevier/Academic Press, Amsterdam, 2007. ISBN 978-0-12-374024-3. OCLC: 255877127.

Pierre Dardenne, George Sinnaeve, and Vincent Baeten. Multivariate Calibration and Chemometrics for near Infrared Spectroscopy: Which Method? *Journal of Near Infrared Spectroscopy*, 8(4):229–237, October 2000. ISSN 0967-0335, 1751-6552. doi: 10.1255/jnirs.283. URL http://journals.sagepub.com/doi/10.1255/jnirs.283.

E. Bertran, M. Blanco, S. Maspoch, M. C. Ortiz, M. S. Sánchez, and L. A. Sarabia. Handling intrinsic non-linearity in near-infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 49(2): 215–224, October 1999. ISSN 0169-7439. doi: 10.1016/S0169-7439(99) 00043-X. URL http://www.sciencedirect.com/science/article/pii/S016974399900043X.

Wangdong Ni, Lars Nørgaard, and Morten Mørup. Non-linear calibration models for near infrared spectroscopy. *Analytica Chimica Acta*, 813:1–14, February 2014. ISSN 0003-2670. doi: 10.1016/j.aca.2013.

12.002. URL http://www.sciencedirect.com/science/article/pii/S0003267013015158.

D. Pérez-Marín, A. Garrido-Varo, and J. E. Guerrero. Non-linear regression methods in NIRS quantitative analysis. *Talanta*, 72(1):28–42, April 2007. ISSN 0039-9140. doi: 10.1016/j.talanta.2006.10.036. URL http://www.sciencedirect.com/science/article/pii/S0039914006007119.

F. Davrieux, D. Dufour, P. Dardenne, J. Belalcazar, M. Pizarro, J. Luna, L. Londoño, A. Jaramillo, T. Sanchez, N. Morante, F. Calle, L.A. Becerra Lopez-Lavalle, and H. Ceballos. LOCAL Regression Algorithm Improves near Infrared Spectroscopy Predictions When the Target Constituent Evolves in Breeding Populations. *Journal of Near Infrared Spectroscopy*, 24(2):109–117, April 2016. ISSN 0967-0335, 1751-6552. doi: 10.1255/jnirs.1213. URL http://journals.sagepub.com/doi/10.1255/jnirs.1213.

Tormod. Naes, Tomas. Isaksson, and Bruce. Kowalski. Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, 62(7):664–673, April 1990. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac00206a003. URL https://pubs.acs.org/doi/abs/10.1021/ac00206a003.

Maxime Metz, Matthieu Lesnoff, Florent Abdelghafour, Reza Akbarinia, Florent Masseglia, and Jean-Michel Roger. A “big-data” algorithm for KNN-PLS. *Chemometrics and Intelligent Laboratory Systems*, 203:104076, August 2020. ISSN 0169-7439. doi: 10.1016/j.chemolab.2020.

104076. URL http://www.sciencedirect.com/science/article/pii/S0169743920301908.

R. Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. *Available*, 2013.

Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, March 2003. ISSN 0886-9383, 1099-128X. doi: 10.1002/cem.785. URL http://doi.wiley.com/10.1002/cem.785.

R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936. ISSN 2050-1439. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x.

Matthieu Lesnoff, Maxime Metz, and Jean-Michel Roger. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. *Journal of Chemometrics*, page 13, 2020.

Svante Wold. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, 20(4):397–405, 1978. ISSN 0040-1706. doi: 10.2307/1267639. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

José Camacho and Alberto Ferrer. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects. *Journal of Chemometrics*, 26(7):361–373, 2012. ISSN 1099-128X. doi: https://doi.org/10.1002/cem.2440. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.2440. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.2440.

Jun-Li Xu, Alexia Gobrecht, Daphné Héran, Nathalie Gorretta, Marie Coque, Aoife A. Gowen, Ryad Bendoula, and Da-Wen Sun. A polarized hyperspectral imaging system for in vivo detection: Multiple applications in sunflower leaf analysis. *Computers and Electronics in Agriculture*, 158:258–270, March 2019. ISSN 0168-1699. doi: 10.1016/j.compag.2019.02.008. URL http://www.sciencedirect.com/science/article/pii/S0168169918314455.