Wheat moisture prediction using functional data analysis

Bassel MASRI Cyril DEVEDEUX

2/14/2021

Contents

1 Introduction

Commercial wheat products such as flower are sold on a weight basis. The grains, however, contain a certain amount of water (i.e. moisture) on which there are restrictions about what moisture contents are allowed. Such restrictions have become a useful standard of identity for grain and flour to make sure that purchasers are buying what they expect. Mastering the level of moisture in wheat grains would highly affect its lifespan as well as how long it could be stored. According to a study on grain moisture [ref.1], the ideal moisture of wheat should be between 13% and 17% On the other hand, determining the level of moisture requires expensive lab equipment and a significant time to obtain accurate results. To overcome such expenses, some data-driven methods approaches have been developed to predict the flour quality parameters from near infrared reflectance (NIR) spectroscopy of the wheat grains [ref.2] through neural networks. Our main contribution throughout this study is to use a functional data approach to determine the level of moisture based on the NIR spectroscopy of the wheat grains.

2 The problem and the data

2.1 Defining the research question

The spectral curves of wheat grains is easy to obtain, while chemical analysis to determine the level of any molecular structure of the wheat (including moisture) is time consuming and expensive. Therefore, to solve such problem, we will take a functional-data driven approach to regress on the moisture level of wheat grains using their equivalent NIR spectroscopy through a **scalar-on-functional regression analysis**. The purpose of regression analysis is to determine the form of dependence between the moisture level (i.e. the target variable Y) and the spectral curve (i.e. the functions X).

Mathematically, we would like to find a functional g such that

$$g : L^2 \to \mathbb{R} \text{ such that } Y = g(X)$$

Once the regression function is defined, it becomes easy to determine an approximate moisture level of a wheat grain sample from its spectral curve.

2.2 The data

The *Moisturespectrum* data included in the `R` package `fds` is a data set that consists of near-infrared reflectance spectra of 100 wheat samples, measured in 2 nm intervals from 1100to 2500nm. Their associated response variable, the samples' moisture content which is a scalar, is included in a different dataset called *Moisturevalues*.

Table 1: Table showing the first few rows and columns of the NIR spectra

	sample 1	sample 2	sample 3	sample 4	sample 5
1100	0.334	0.302	0.354	0.308	0.309
1102	0.333	0.301	0.354	0.307	0.308
1104	0.332	0.301	0.353	0.307	0.308
1106	0.332	0.300	0.352	0.306	0.307
1108	0.331	0.300	0.352	0.305	0.306

Figure 1 shows the spectral curves of 100 wheat samples densely packed together on the left, and their

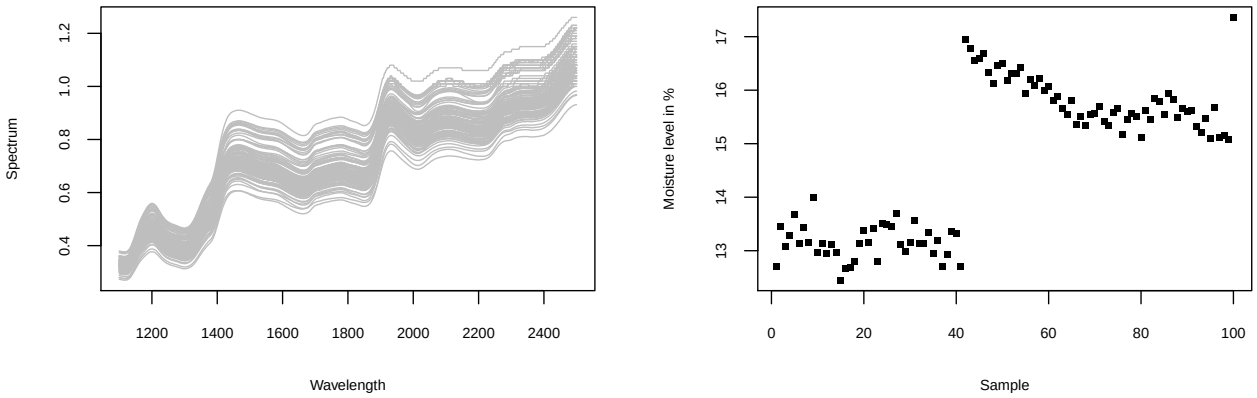


Figure 1: NIR spectroscopy of 100 wheat samples (on the left) and their moisture levels (on the right)

corresponding moisture level in percentage on the right. Table 1 displays the first few rows and columns of the NIR spectra values. The row names of the matrix represent the wavelength in nano-meters and the column names represent the sample number.

3 Approach and results

In this section, we will go through the details of our analysis steps we took in order to solve the problem. We start with an exploratory analysis, then move on to the modeling aspect along with the results and discussions.

3.1 Exploratory data analysis

As with all statistical modeling, data exploration is always the first step prior to any model development. Therefore, our analysis begins with a suitable exploratory analysis for functional data where we investigate a suitable basis function to smooth the data then compute the point wise statistics such as mean, standard deviation and their 95% confidence intervals.

3.1.1 Manual smoothing

The first task is to smooth the functions using a basis. The data do not show any periodicity as we can see in Figure 1. Therefore, the optimal basis function would be B splines. At first, we choose 25 basis and explore the plot produced in Figure 2. We add to the latter the mean function (the black curve) as well as the standard deviation plotted in red. We also compute and plot the 95% confidence interval in dashed green.

Interestingly, the smoothed data follows the classic two standard deviation rule surprisingly well with nearly all of the curves falling between the green lines. Figure 3 shows the point-wise sample standard deviation which gives us an idea about the variability of curves at any point t. Indeed, we notice more variability in the early samples than in the last samples. This may be explained by the fact that the moisture level in those samples are quite different.

The point-wise sample standard deviation gives no information on how the values of the curves at point t relate to those at point s. Therefore we compute the sample covariance function $\hat{c}(t, s)$ and we plot its perspective and its contour plot as seen in Figure 4 and Figure 5

The 3D perspective plot shows that the variation in the spectrum is higher in the first samples. Indeed, this is confirmed in the contour plot in Figure 5 in the highlighted red circle.

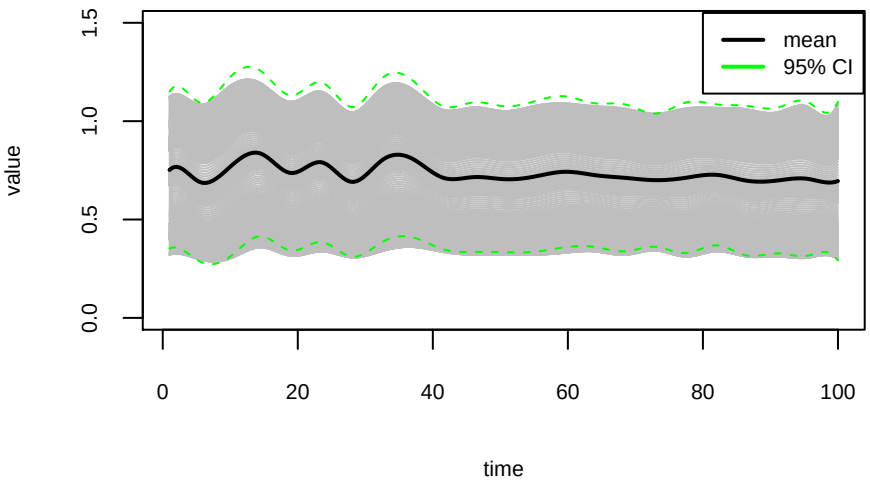


Figure 2: Smoothed curves using 25 B splines

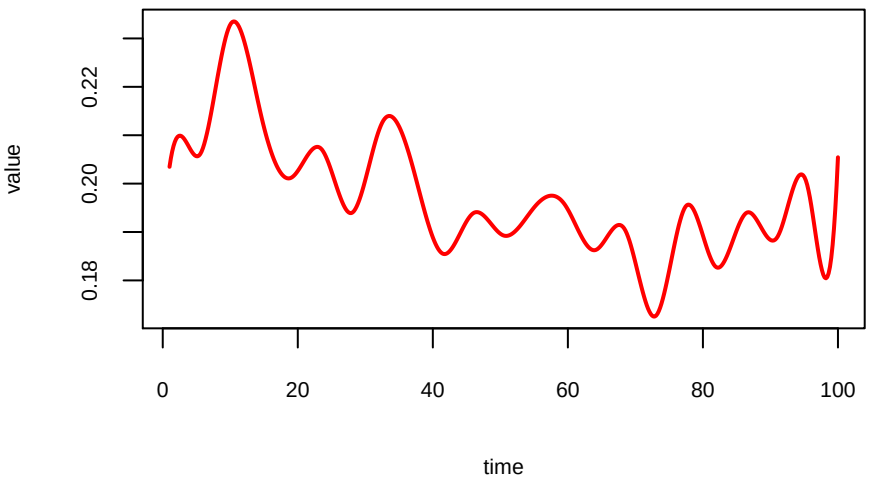


Figure 3: Standard deviation of the curves

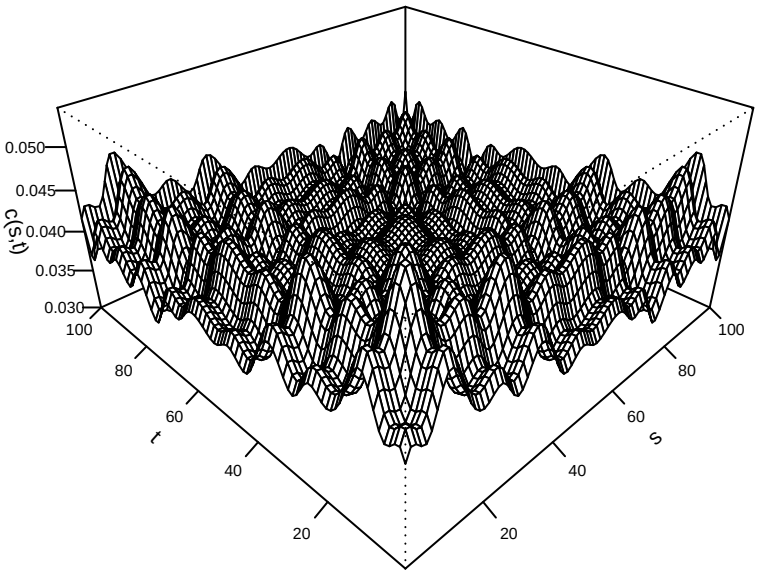


Figure 4: Perspective plot of the covariance function

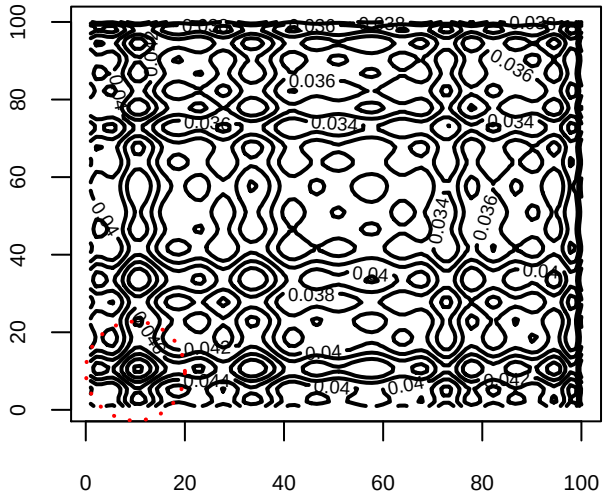


Figure 5: Contour plot of the covariance function

3.1.2 Penalized smoothing

Typically, when the raw data curves exhibit a substantial level of noise, the functional objects constructed using manual basis expansion smoothing (i.e $M = 25$ in our case) will inherit this variability, and thus resulting in *wiggly* curves. To avoid amplifying said variability, we will perform smoothing using a penalized approach.

To choose the tuning parameter λ, generalized cross-validation (GCV) is employed. The aim is to minimize the penalized sum of squares with respect to the tuning parameter λ.

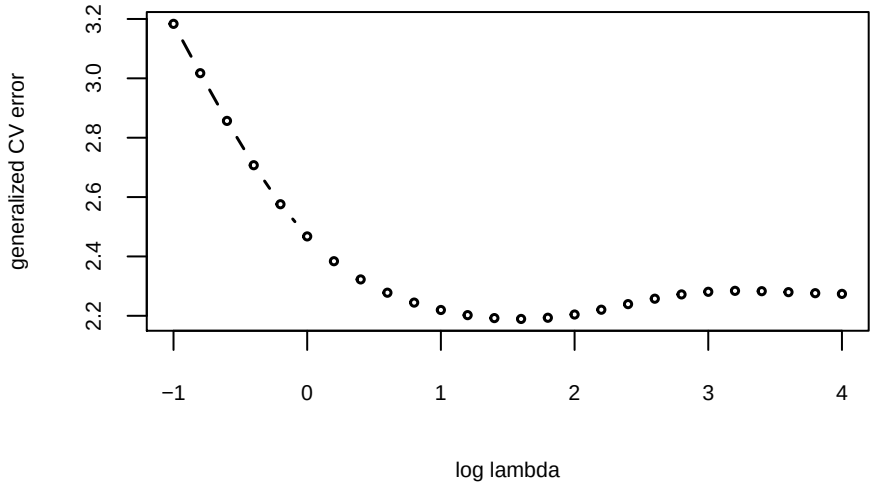


Figure 6: Values of the tuning parameter using generalized cross validation

Generalized cross validation shows an optimal smoothing parameter $\lambda = 39.8$ when choosing a range for λ between $[10^{-1}, 10^4]$ and a number of basis of 150 knowing that we only have 100 samples. The cross validation curve with respect to the values of λ is shown in Figure 6. Using the minimum value of λ, we penalize the basis expansion and repeat the smoothing task which yields the curves in the bottom plot in Figure 8.

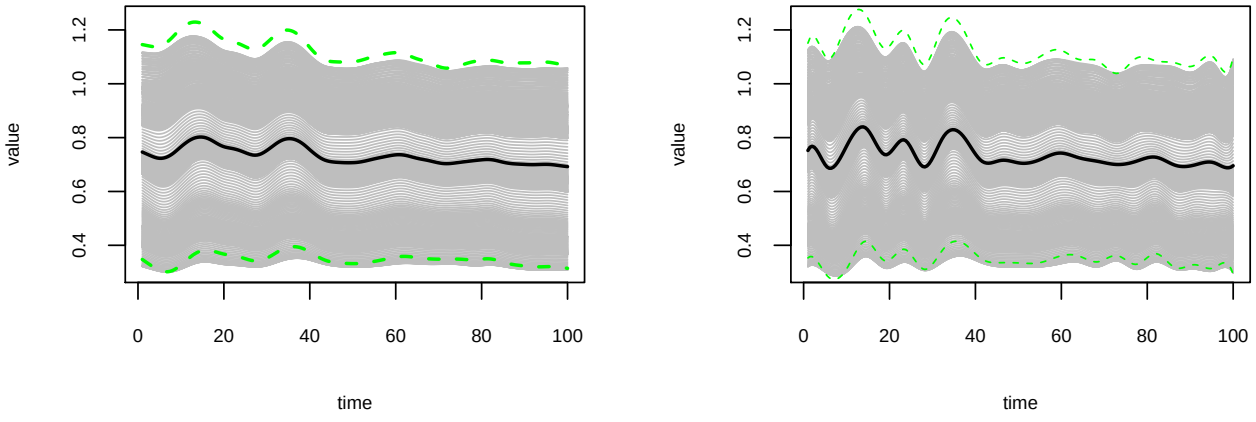


Figure 7: Smoothed curves without penalization (right) vs with penalization (left)

Indeed, the results indicate that penalizing the smoothed basis reduces unnecessary variance significantly. The mean curve of the penalized smoothing seen in the top plot of Figure 8 is much less noisy, showing peaks where the relative difference between the curves actually matters, and a smooth plateau where the difference is insignificant.

3.1.3 Functional principal components analysis

One of the most useful tools in functional data analysis is the principal component analysis. Estimated functional principal components, EFPC's, are related to the sample covariance function $\hat{c}(t,s)$. Similar to usual multivariate statistics, we try to reduce the dimensionality of the feature space using only a few functions $\hat{\nu}_j$ such that the centered functions $X_n - \bar{X}_N$ are represented as follows :

$$X_n(t) - \bar{X}_N(t) \approx \sum_{j=1}^{p} \xi_{nj}\hat{\nu}_j(t)$$

Where p is a much smaller dimension than M the number of basis. Note that ξ_{nj} are the scores of the components. The principal component analysis will be applied on the functions produced after penalized smoothing observed in the top plot of Figure 8.

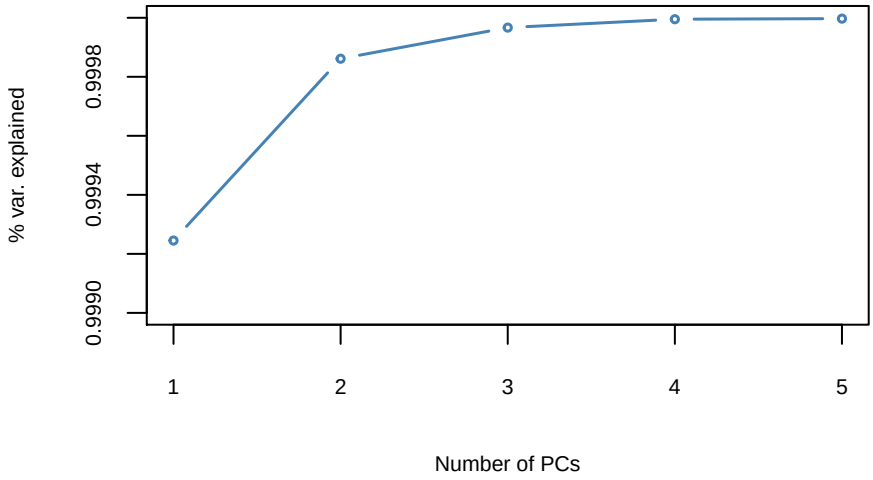


Figure 8: The percentage of variability explained vs FPCs on smoothed curves using penalization

Table 2: Table showing the cumulated variance of the first 5 PC

	FPC1	FPC2	FPC3	FPC4	FPC5
Cumulated Var.	0.9992452	0.9998613	0.9999667	0.9999951	0.9999971

Figure 8 and Table 2 indicate that we are able to capture more than 99% of the variability using only the first principal component. This is not surprising due to the fact that the plots are very similar and densely packed together as we can observe in the Figure 1.

We visualize the plot of the first 3 components in Figure 9. The first principal component (in black line) indicates that almost all samples' variance is captured equivalently. This is coherent with our previous findings regarding curve similarities. There will be no need to further explore the curves for the next steps.

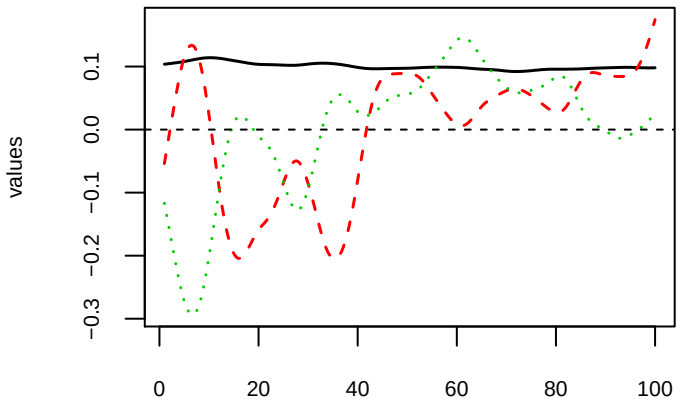


Figure 9: plot of the first three FPCs

3.2 Modeling

3.2.1 Model definition and evaluation

Functional regression models can be identified using three main categories based on the nature of the target variable and the predictors. These categories are; scalar-on-function, function-on-scalar and function-on-function. Our exploratory data analysis has shown that the response variable is a scalar (i.e. a number on $\mathbb{R}$), which makes a scalar-on-function regression analysis an ideal candidate to solve the problem at hand [ref.3]. The regression model is defined as follows:

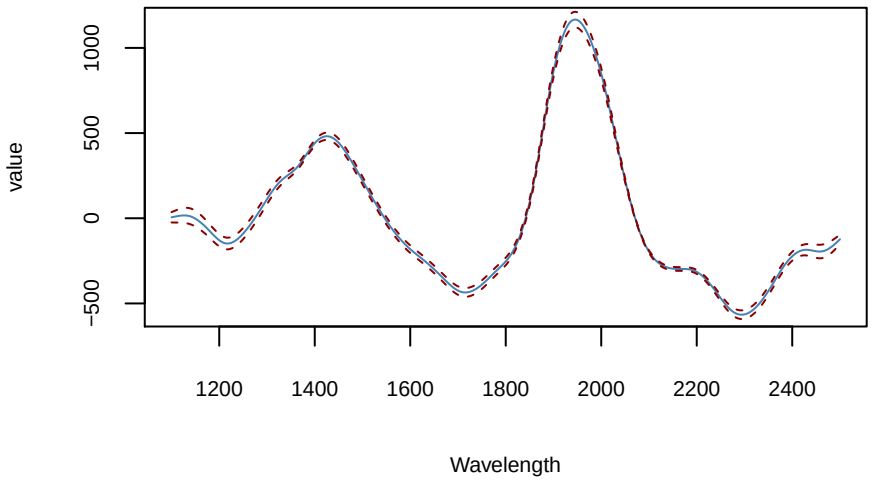
$$Y_i = \int \beta(s) X_i(s) ds + \varepsilon_i$$

In order to evaluate the model, we will use the R^2 as our main metric. Further more, we will display the diagnostic plots and analyze them to further confirm (or reject) the model.

3.2.2 Functional principal components regression

We have seen in the previous section that it is possible to capture a high level of variance using only two functional principal components. Naturally, the easiest regression model we could fit is a Penalized Functional Regression (pfr) model on the principal components. We consider all 100 observations as our training set and use the `R` command pfr to regress Y on `fpc(X)` setting the number of components to 4.

Upon fitting the model using the above-mentioned settings, we obtain a high 97.33% R^2 score. Next, we produce a plot of the fitted coefficient along with its 95% confidence interval in the Figure below.



The coefficient plot, however, does not give much insight on how well the model predicts values. Therefore, we produce a plot of fitted against observed values to visualize the linear trend and get a better idea on the goodness of fit. In addition, we plot the residuals to make sure that there is no structure in its distribution and that it is centered around 0.

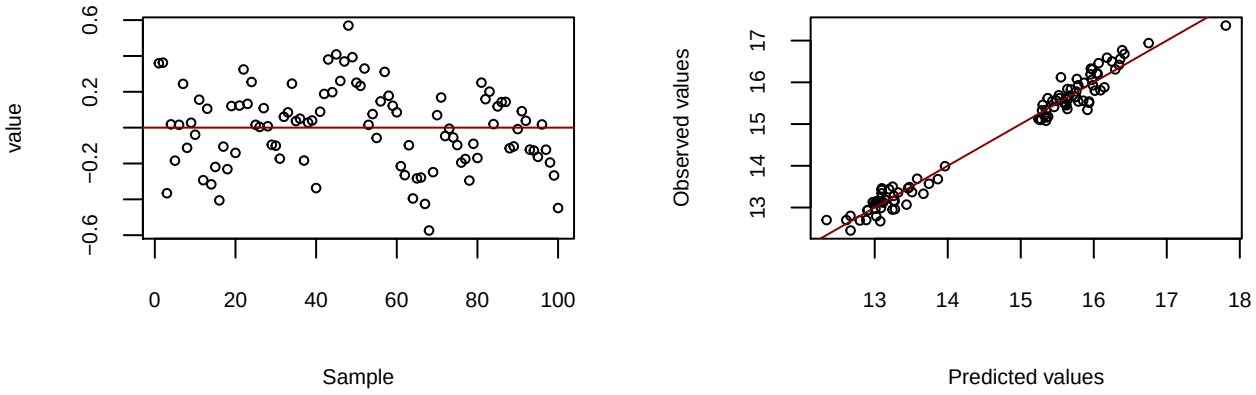


Figure 10: plot of residuals(left) and observed vs. predicted values (right)

The two diagnostic plots produced in Figure 10 show a nice linear trend between predicted and observed percentages of moisture in the wheat grains. The residuals plot shows a random structure and centered around 0 which further confirms a good fit.

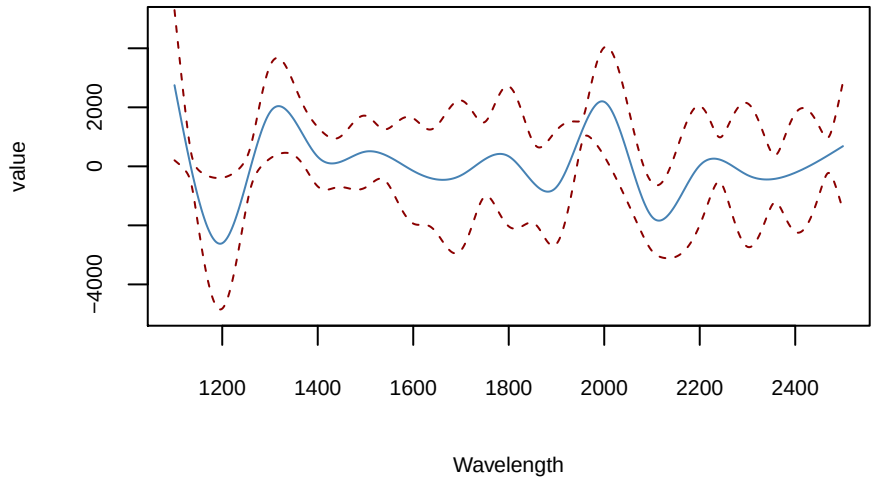
3.2.3 Regression using basis expansion approach

Regression using basis expansion approach works is by expanding the function β using deterministic basis functions. This allows us to rewrite it as follows :

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{c}_k B_k(t)$$

Then, we could integrate it in the original regression equation defined above to solve the problem. One disadvantage of such approach is that the basis functions B_k and their number K are highly subjective to the user. In our case, since we used a B-spline basis function as an expansion of the regressor functions, we will do the same here. For $k = 15$, we get an R^2 score of 97.66 which is highly similar to the one produced by the principal components regression. Interestingly, the plot below shows a significantly different coefficient curve

compared to the one produced before. The confidence intervals are larger and the overall "roughness" of the curve is higher. This, again, is highly dependent on the user's choice of basis functions and their number.



To compare the results with the previous section, we produce the same residuals plot and the observed vs. predicted values plot as seen in Figure 11.

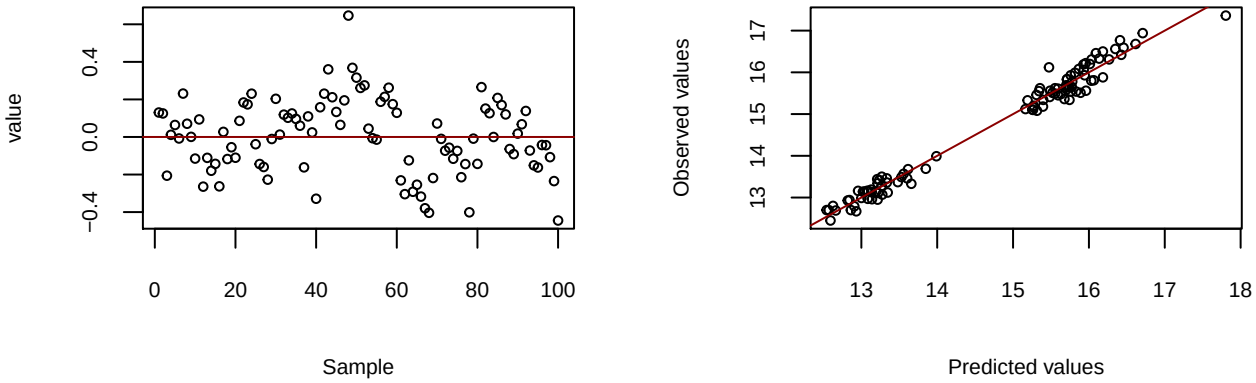


Figure 11: plot of residuals(left) and observed vs. predicted values (right)

4 Conclusion

A dimensionality reduction approach has always been useful to solve high dimensional multivariate data by projecting it onto a smaller, more manageable feature space. From a functional data approach, this has proven to give accurate results when solving a scalar-on-function regression to solve the problem in this study which is to be able to detect the level of moisture in wheat grains from its NIR spectra. Indeed, based on its R^2 estimate, the problem at hand has been solved which might help reduce the cost of chemical analyses in future works. On the other hand, a regression using basis expansion approach is more easily interpretable based on the fact that we do not reduce the dimension of the data. The downside, however, is that a careful choice of the basis expansion functions and their number has to be taken.

Throughout this study, we have considered only the R^2 as our primary model selection metric. Future works could explore different (if not multiple) evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). In addition, the evaluation has been done on the same data we used to fit the model which makes it a biased estimator. In order to obtain an unbiased metric, one would have to perform a train-validation split which was beyond the scope of this study.

5 References

[ref.1] Grain moisture – guidelines for measurement link https://projectblue.blob.core.windows.net/media/Default/Imported%20Publication%20Docs/Grain%20moisture%20%E2%80%93%20guidelines%20for%20measurement.pdf

[ref.2] Prediction of wheat quality parameters using near-infrared spectroscopy and artificial neural networks link https://www.researchgate.net/publication/227299821_Prediction_of_wheat_quality_parameters_using_near-infrared_spectroscopy_and_artificial_neural_networks

[ref.3] Kokoszka, P. and Reimherr, M. (2017). Introduction to Functional Data Analysis. Chapman and Hall/CRC.