Bandwidth Selection Problem in Nonparametric Functional Regression

Daniela Kuruczová [1] | *Masaryk University, Brno, Czech Republic*
Jan Koláček [2] | *Masaryk University, Brno, Czech Republic*

Abstract

The focus of this paper is the nonparametric regression where the predictor is a functional random variable, and the response is a scalar. Functional kernel regression belongs to popular nonparametric methods used for this purpose. The two key problems in functional kernel regression are choosing an optimal smoothing parameter and selecting an appropriate semimetric as a distance measure. The former is the focus of this paper – several data-driven methods for optimal bandwidth selection are described and discussed. The performance of these methods is illustrated in a real data application. A conclusion is drawn that local bandwidth selection methods are more appropriate in the functional setting.

Keywords

Functional data, nonparametric regression, kernel methods, bandwidth selection

JEL code

C14

INTRODUCTION

Functional data analysis is a relatively recent topic in statistics. It is based on the concept of a functional random variable, which is a random variable taking values in infinite-dimensional space. Natural examples of a functional random variable are random curves and surfaces, but the concept itself also covers more complex objects (Ramsay and Silverman, 1997; Ferraty and Vieu, 2006). Realizations of a functional random variable are called functional data, which are treated as members of an infinite-dimensional space. Typical datasets suitable for this approach are usually high-dimensional and multicollinear, violating assumptions of most traditional regression methods. Therefore, a functional approach brings new options to analyse such data. This paper focuses on a nonparametric functional regression, namely, kernel functional regression with a scalar response. The functional kernel regression is based on kernel smoothing, one of the most popular nonparametric techniques. The main advantages of kernel methods are their simplicity of expression and ease of implementation.

The move from a real random variable to a functional one brings several challenges. One of the most important ones is the choice of a distance measure. In contrast to the finite-dimensional setting, the selection

[1] Department of Mathematics and Statistics, Faculty of Science, Kotlářská 2, 611 37 Brno, Czech Republic. E-mail: xkuruczovad@math.muni.cz.
[2] Department of Mathematics and Statistics, Faculty of Science, Kotlářská 2, 611 37 Brno, Czech Republic. E-mail: kolacek@math.muni.cz.

of distance measure is not straightforward and has a noticeable impact on the properties of the estimates. Ferraty and Vieu (2006) propose the use of semimetrics as a suitable distance measure and introduce three classes of semimetrics: based on derivatives, functional principal component analysis, and partial least squares regression. The use of semimetrics instead of metrics appears to bring good practical results, but poses theoretical challenges as the theory of semimetric spaces is less straightforward compared to the theory of metric spaces. In practice, it also seems that the choice of an appropriate semimetric for the data is crucial. Derivative-based semimetrics appear to be suitable for relatively smooth data whereas the other two kinds of semimetrics work well for rougher or noisy data (Benhenni et al., 2007).

Just as in a finite-dimensional setting, kernel regression estimators depend on a smoothing parameter that controls the smoothness of the estimated curve. The choice of an optimal smoothing parameter is a crucial factor in the quality of kernel estimates. Unlike the finite case, it is not possible to rely on visualization tools when selecting optimal bandwidth. Thus, automatic (data-driven) selection procedures are useful for many practical situations. Successful approaches to bandwidth selection in the finite dimensional setting can be transferred to the functional setting. Most of these procedures are based on the mean squared error (*MSE*) estimation and further minimization of such an estimate. A widely known and commonly used method for bandwidth selection is leave-one-out cross validation (Ferraty and Vieu, 2002). A method of penalizing functions is similar, but lesser known, see e.g. Härdle (1992). It is based on the idea of penalizing the biased estimate of the *MSE* and can be considered a generalization of the leave-one-out cross validation. In this paper, use of the penalizing functions method in the functional setting is proposed. Both approaches mentioned above are global methods, i.e. the value of smoothing parameter is fixed for all data points. However, local methods determine a custom value of the smoothing parameter for each data point. The local version of cross-validation based on the number of nearest neighbors is proposed by Benhenni et al. (2007). One of the latest results in this area is the paper by Chagny and Roche (2016) where an adaptive, data-driven, local bandwidth selection rule was derived.

1 FUNCTIONAL KERNEL REGRESSION

Kernel methods for finite data can be easily extended to functional data. Ferraty and Vieu (2006) demonstrate the application of kernel methods for regression and classification on several datasets from various fields, e.g. economics (electricity consumption dataset), chemistry (spectrometric dataset), and phoneme recognition.

In the case of functional data, the kernel function serves the same purpose as in a finite case – to assign weights to observations. The most notable difference is the use of asymmetrical kernels, because the measure of distance (semimetric) is always non-negative.

If we consider a pair of random variables $(X,Y) \in E \times R$, where E is a semimetric space and R are real numbers, the functional regression operator with scalar response can be defined as:

$$Y = r(X) + \varepsilon = E(Y \mid X) + \varepsilon, \tag{1}$$

where ε is the random error with an expected value of zero and finite variance, and the regression operator r is expressed as the conditional mean of the finite random variable Y given the functional random variable X.

Let $\{(Y_i, X_i), i = 1,..., n\}$ be an observed data sample with the same distribution as the pair *(X, Y),* then for the observed data the formula is obtained:

$$Y_i = r(X_i) + \varepsilon_i, \tag{2}$$

with independent, identically distributed errors ε_i that are independent of $X_i, i = 1,..., n$.

One of the options for the functional kernel estimate of the regression operator r is a simple extension of the finite case – Nadaraya-Watson estimator (Nadaraya, 1964):

$$\hat{r}(x,h) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(x,X_i)}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h}\right)} = \sum_{i=1}^{n} W_{i,h}(x) Y_i. \tag{3}$$

The function K is an asymmetrical kernel function, assuming that:

$$\int_R K(x)dx = 1,$$

where d denotes a semimetric and the parameter $h>0$ is called bandwidth. It can be shown that under certain assumptions (most notably, in the case of continuity of the regression operator r) this estimate converges almost completely to the actual regression operator (Ferraty and Vieu, 2006).

2 BANDWIDTH SELECTION

The functional setting brings several obstacles to the problem of optimal bandwidth selection. One of the very useful tools, visualization, does not transfer well to the infinite-dimensional setting. Another issue is the sparsity of data in certain areas of the functional space (Benhenni et al., 2007).

The problem of bandwidth selection is closely tied to estimation of the mean squared error (*MSE*) of the regression operator estimate in fixed data point x:

$$MSE(\hat{r},h,x) = E[(\hat{r}(x,h) - r(x))^2], \tag{4}$$

where the optimal bandwidth value h can be defined as the value of h that minimizes the mean squared error.

An exact *MSE* formula for the case of separable Banach space was derived by Ferraty et al. (2007) and for semimetric space by Geenens (2015). In both cases the final formula is based on the decomposition of *MSE* into bias and variance terms. The decomposition shows the expected influence of parameter h – with increasing values of h the bias increases and the variance decreases. Compared to the finite case, the influence of h in the variance term is less straightforward. It is moderated through a notion called the small ball probability:

$$\varphi_x(h) = P(d(X,x) \le h), \tag{5}$$

which expresses the probability of functional random variable occurring within a ball centered around x with a radius of h. As h decreases, the small ball probability decreases, so lower values of h still lead to higher variance, and vice versa.

The minimization problem leads to the usual variance-bias trade-off known from the finite case. However, in the functional setting, the *MSE* formula contains several expressions unknown in practice, so it cannot be directly used for bandwidth selection.

2.1 Global bandwidth selection

One possible approach to bandwidth selection is to select one bandwidth value for the whole dataset. Two methods, both using minimalization of a biased *MSE* estimate obtained through cross-validation, will be presented.

2.1.1 General cross-validation

The general cross-validation function (also called leave-one-out cross-validation) is based on the cross-validation function:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_{-i}(X_i, h))^2, \tag{6}$$

where:

$$\hat{r}_{-i}(X_i, h) = \frac{\sum_{\substack{j=1 \\ i \neq j}}^{n} Y_i K\left(\frac{d(X_i, X_j)}{h}\right)}{\sum_{\substack{j=1 \\ i \neq j}}^{n} K\left(\frac{d(X_i, X_j)}{h}\right)} \tag{7}$$

is the regression operator estimate at point X_i based on data that does not contain X_i, the well-known leave-one-out regression estimator. This method was proposed in a functional setting by Ferraty and Vieu (2002).

2.1.2 Penalizing function method

The penalizing function method uses another means to avoid the bias caused by using the same dataset for creating the regression operator estimate and estimating response values – a penalizing function Ξ. The penalizing function penalizes small values of h to avoid undersmoothing. It was proposed by Härdle (1992) for the case of finite dimension. The same idea can be adapted to an infinite-dimensional setting. Thus, the error function to be minimized takes the form:

$$PF(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}(X_i, h))^2 \Xi(W_{i,h}(X_i)). \tag{8}$$

Several examples of penalizing functions are given in Härdle (1992) and Koláček (2002). One of the best-known examples of a penalizing function is Akaike's information criterion, which is also used in the section concerning practical application of the method.

2.2 Local bandwidth selection

Another option is to select the bandwidth value for each data point separately. Two methods based on different approaches will be presented.

2.2.1 Nearest neighbors method

This method, also known as k nearest neighbors method, is based on a local cross-validation (Benhenni et al., 2007) function minimized separately for each data point, where:

$$LCV_x(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_{-i}(X_i, h))^2 w_{n,x}(X_i). \tag{9}$$

The local cross-validation function uses the leave-one-out regression estimate along with a weight function dependent on data point x. If the weight function does not depend on x, the function becomes a global cross-validation function.

The weight function for the nearest neighbors method is defined as:

$$w_{n,x}(X_i) = \begin{cases} 1 & d(X_i, x) < h \\ 0 & otherwise \end{cases}. \tag{10}$$

For each value of h, the sum of weights:

$$\sum_{i=1}^{n} w_{n,x}(X_i) = k,$$

is an integer that expresses the number of nearest neighbors around the curve x that were used to make the estimate. When the local cross-validation function is minimized with respect to h for each data point x, we get the value of h_x indirectly expressing an optimal number of curves to use for the estimate in x.

2.2.2 Chagny-Roche method

Unlike previously mentioned methods, this method presented by Chagny and Roche (2016) is not based on cross-validation, but rather on the actual estimate of the mean squared error. This method was derived in the classical L^2 metric space. Assuming that K is a type I Kernel, as defined by Ferraty and Vieu (2006), and the regression operator r is Hölder continuous with an exponent of β, then we can determine for $h>0$ the upper bound of the mean squared error as:

$$MSE(\hat{r}, x, h) \leq C\left(h^{2\beta} + \frac{\sigma^2}{n\varphi_x(h)}\right), \tag{11}$$

where C is a positive constant depending only on constants from type I Kernel definition. The first part of the expression corresponds to the bias term of the *MSE* and the second part corresponds to the variance term. Instead of the *MSE,* the upper-bound expression is minimized with respect to h.

As all parts of the expression cannot be determined in practice (namely, the parameter β and the small ball probability), an estimate of the expression is used:

$$ChR_x(h) = \hat{A}(h, x) + \hat{V}(h, x), \tag{12}$$

where $\hat{A}(h, x)$ approximates the bias term and $\hat{V}(h, x)$ is an empirical estimate of the variance. See the original paper for exact formulas (Chagny and Roche, 2016).

3 APPLICATION TO REAL DATA

3.1 Quality of estimates

In the context of kernel methods, the empirical version of the *MSE* (also called mean squared prediction error, Ferraty and Vieu, 2006) is the most commonly used tool to compare the quality of various estimates:

$$MSPE(\hat{r}, h) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}(x_i, h) - y_i)^2. \tag{13}$$

Due to the squared difference between the estimated and the actual value, the empirical *MSE* is more sensitive to large differences. This is a desired property as large differences in regression estimates tend to be more problematic than smaller ones. Furthermore, all methods for optimal bandwidth selection are based on the theoretical *MSE*, so using the empirical version of the *MSE* is natural for their comparison.

3.2 Implementation of methods

For practical application of the presented methods, a few issues must be considered. The first is the choice of the set of possible bandwidth values, as it is not feasible to test over the entire theoretical interval. The other is the sparsity of the functional data – a situation where one curve is more distant from all the other curves than the selected optimal bandwidth is much more likely to happen than in finitely dimensional datasets, especially with estimates for the testing dataset. How these issues were resolved in respective implementations will be briefly discussed.

The general cross-validation method is implemented in companion material to the publication by Ferraty and Vieu (2006) and by default uses the 5th to 50th percentiles of distances between curves in a learning dataset (i.e. the dataset with known response values) to determine the optimal bandwidth. If the distance of a curve from all other curves is greater than the optimal bandwidth (and therefore the estimate for this curve cannot be computed), the value of h is increased until all curves can be evaluated.

The penalizing functions method was implemented to correspond to the original general cross-validation function, and the choice of bandwidth interval and distant data treatment therefore remained the same.

The nearest neighbors method was again implemented by Ferraty and Vieu (2006). The method uses the bandwidth interval spanning from the ten nearest curves to 50% of the nearest curves of the training dataset. The step in the sequence is defined as the number of training curves divided by 100 and rounded up to the nearest integer. Since the value of h is mediated through the number of neighbors, the situation in which one curve is too distant is not possible.

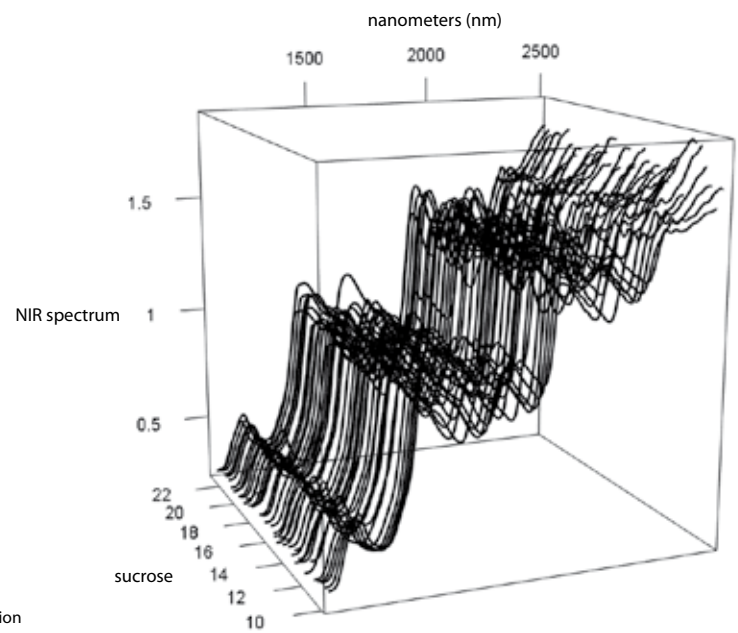
Using the original paper (Chagny and Roche, 2016), the Chagny-Roche method was implemented. The bandwidth interval was chosen in the same way as in the case of the general cross-validation – 5th to 50th percentiles of distances between curves, but separately for each curve. Thanks to this choice, distant curves are not possible.

To provide a fair comparison, a Nadaraya-Watson estimator with quadratic kernel was used for all bandwidth selection methods.

3.3 Spectrometric dataset – fat and sugar content in cookies

To compare the presented bandwidth selection methods to real data, we chose cookie dough spectrometric data (Osborne et al., 1984; Brown et al., 2001). The dataset contains a NIR (near-infrared spectroscopy) reflectance spectrum measured from 1 100 to 2 498 nanometres for each dough piece along with information about fat, sucrose, flour, and water content. Figure 1 illustrates the spectrometric curves that were used as the functional predictor.

Figure 1 NIR reflectance spectra for the cookie dough sorted by their sucrose content



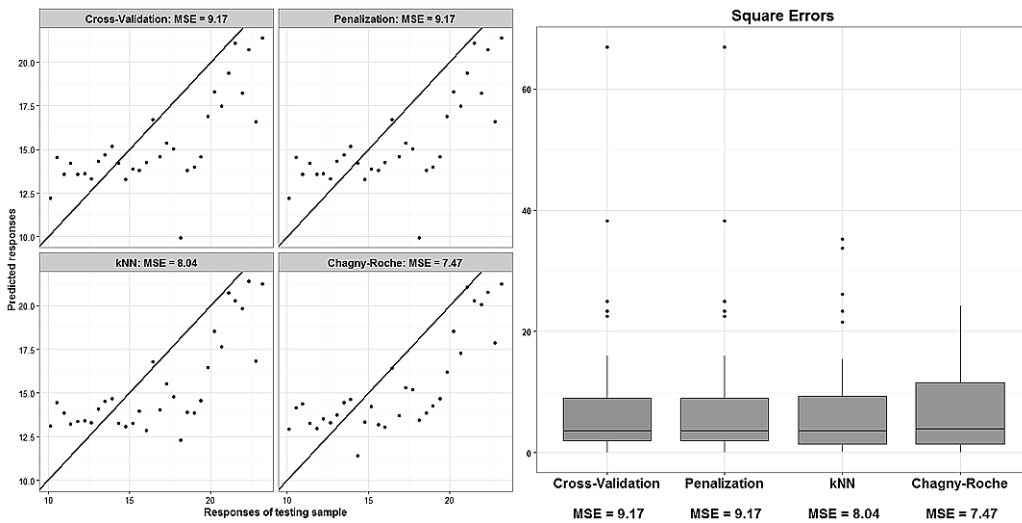
Source: Own construction

The fat and sucrose content were chosen as the two scalar responses. The semimetric based on the second derivative was used to measure the distance between curves as it tested as the most appropriate for this dataset. The dataset was divided into 40 training and 32 testing curves as per the original dataset description.

3.3.1 Sugar content

All four presented methods for bandwidth selection were used to find the optimal bandwidth parameter for the testing dataset, using the information from the training dataset. Afterward, predictions for the testing dataset were compared to the actual values, see Figure 2.

Figure 2 Performance of the four bandwidth selectors



Source: Own construction

The results show that both global selection methods perform the same. The methods based on local selection performed better than the global ones, and the Chagny-Roche method was the best method overall.

Figure 2 also illustrates the distribution of the squared error for all four methods. The boxplots show that prediction using global bandwidth selection methods was quite distant from the actual value in several cases. This phenomenon is less pronounced with the nearest neighbors method and almost non-existent with the Chagny-Roche method. The Chagny-Roche method is the only that does not use the cross-validation function, so the absence of outliers in its case might be attributed to its different mechanics.

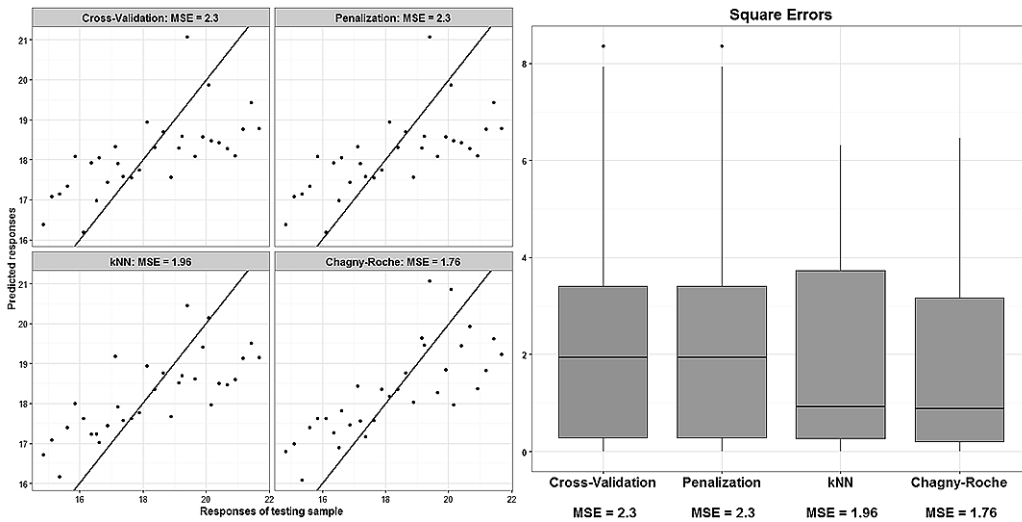
3.3.2 Fat content

The same procedure was applied to the data with fat content as the scalar response, see Figure 3 for the comparison.

The overall *MSE* was smaller, which was expected because the fat content had a lower variance than sugar content in this dataset. The results once again show that both local selection methods performed better than the global ones and that the Chagny-Roche method was the best method overall.

Boxplots on Figure 3 again illustrate that global bandwidth selection methods have more distant predictions than local methods, the superiority of the Chagny-Roche method being even more pronounced in this case.

Figure 3 Performance of the four bandwidth selectors



Source: Own construction

CONCLUSION

Bandwidth selection methods developed for a finite dimensional setting can be successfully transferred to the functional kernel regression. Besides methods already proven to work with an infinite setting (leave-one-out cross-validation, nearest neighbors method), the penalization functions method and a recent result in the field – the Chagny-Roche method – were successfully applied.

A summary of known methods applied to real data and their performance compared in terms of mean squared error for two different scalar responses was presented. According to the results, methods based on local bandwidth selection perform better. This is not a surprising result, as concerns about functional data sparsity appear to be justified. It is also consistent with previous findings by Benhenni et al. (2007), where the leave-one-out cross-validation and local nearest neighbors methods were compared using spectrometric data for predicting fat content in meat. An interesting finding is that the Chagny-Roche method, derived in a metric space setting, demonstrated the best performance even though a semimetric, not a metric, was used. We believe that local bandwidth selection methods should be further developed to obtain better methods for optimal bandwidth selection in an infinite-dimensional setting. We also should point out that finding the optimal bandwidth still leaves us with the problem of the optimal semimetric selection.

ACKNOWLEDGEMENT

This research was supported by Masaryk University, project GAČR GA15-06991S.

References

BENHENNI, K., FERRATY, F., RACHDI, M., VIEU, P. Local smoothing regression with functional data. *Computational Statistics*, 2007, 22(3), pp. 353–369.

BROWN, P. J., FEARN, T., VANNUCCI, M. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 2001, 96(454), pp. 398–408.

CHAGNY, G. AND ROCHE, A. Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, 2016, 146, pp. 105–118.

FERRATY, F. AND VIEU, P. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 2002, 17(4), pp. 545–564.

FERRATY, F. AND VIEU, P. *Nonparametric functional data analysis: theory and practice.* New York: Springer, 2006.

FERRATY, F., MAS, A., VIEU, P. Advances on nonparametric regression for functional variables. *Australian and New Zeland Journal of Statistics*, 2007, 49(3), pp. 1–20.

GEENENS, G. Moments, errors, asymptotic normality and large deviation principle in nonparametric functional regression. *Statistics & Probability Letters*, 2015, 107, pp. 369–377.

HÄRDLE, W. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press, 1992.

KOLÁČEK, J. Kernel Estimation of the Regression Function – Bandwidth Selection. *Datastat 01, Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis, Mathematica 11*, Brno: Masaryk University, 2002, 10, pp. 129–138.

NADARAYA, E. A. On estimating regression. *Theory of Probability & Its Applications*, 1964, 9(1), pp. 141–142.

OSBORNE, B. G., FEARN, T., MILLER, A. R., DOUGLAS, S. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 1984, 35(1), pp. 99–105.

RAMSAY, J. O. AND SILVERMAN, B. W. *The analysis of functional data.* New York: Springer, 1997.