K-Nearest Neighbor Method with Principal Component Analysis for Functional Nonparametric Regression

Shelan Saied Ismaeel1* Kurdistan M. Taher Omar1 Bo Wang2

1Department of Mathematics, Faculty of Science, University of Zakho, Zakho, Iraq.
2Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK
*Corresponding author: shelan.ismaeel@uoz.edu.kurd
E-mail addresses: kurdistan.taher@uoz.edu.krd, bo.wang@le.ac.uk

Received 6/7/2021, Accepted 23/8/2022, Published Online First 25/11/2022, Published 5/12/2022

Abstract:
This paper proposed a new method to study functional non-parametric regression data analysis with conditional expectation in the case that the covariates are functional and the Principal Component Analysis was utilized to de-correlate the multivariate response variables. It utilized the formula of the Nadaraya Watson estimator (K-Nearest Neighbour (KNN)) for prediction with different types of the semi-metrics, (which are based on Second Derivative and Functional Principal Component Analysis (FPCA)) for measuring the closeness between curves. Root Mean Square Errors is used for the implementation of this model which is then compared to the independent response method. R program is used for analysing data. Then, when the covariates are functional and the Principal Component Analysis was utilized to de-correlate the multivariate response variables model, results are more preferable than the independent response method. The models are demonstrated by both a simulation data and real data.

Keywords: Functional data analysis, K-Nearest Neighbour estimator, Multivariate response, Nonparametric regression, Principal Component Analysis.

Introduction
In recent years the issue of nonparametric functional regression has become a topic of growing interest, due to the sophistication in recent technological advances regarding collecting and storing data as curves. The functional data have become more combined in growing numbers of fields, such as biology, engineering, medical science, meteorology, psychology, statistics, among others. Ramsay and Silverman1 pioneered the area of functional data analysis which becomes popular, while case studies and applied problems with linear regression and multiple regression (parametric models) are pointed out by 2. The linear methods for regression with functional response and scalar input for more information see 3,4. In the situation when both the output and the input are functions to estimate functional multivariate data in functional multivariate linear regression method were examined by 5-7.

The story of nonparametric functional data began with 1 and the nonparametric functional regression models have then been the object of several researches.

The existing literature contains a considerable number of theoretical and experimental studies with different models on nonparametric functional data when the response is an independent response variable and covariate is functional. For instance, the functional Nadayara-Watson (NW) estimator approach, the functional k-nearest neighbour estimator method, the functional local linear estimator model, and distance-based local linear estimator model 8-11 showed the nonparametric models and the related theories for the situation when the output and the functional covariates are both functions.

In recent years, the multivariate nonparametric functional regression model was also presented with functional data, for example by 5,12-14 which examined the relationship between multiple scalar responses and functional predictors by using Gaussian basis function model. Chaouch and Laïb15 explained the issue of multivariate response model...
from functional covariates based on the L1-median regression estimation approach. Wang and Chen \(^{16}\) approached the Gaussian process regression with multivariate output and used principal component analysis to de-correlate multivariate response with functional and multivariate covariate variables. The nonparametric functional regression model for multivariate longitudinal data with multiple responses which is illustrated by different types of data for more details see \(^{17}\).

Omar and Wang \(^{18}\) expanded the independent response method to multivariate responses method with functional covariate in nonparametric functional regression which is applied with real data and simulated data then the new model results (multivariate responses model) are preferable than the independent response method. The paper proposed a new model to deal with multivariate responses variables and functional covariate. This paper used the K-Nearest Neighbour model with Principal component analysis to de-correlate the multivariate responses, and also utilizes the K-NN method for independent prediction regression. In the K-NN model, the semi-metrics as measure of closeness between the functional covariates was used which will be clarified in more details in the methodology.

The purpose of this study is to add some new results to the nonparametric regression of the conditional expectation when the covariates \(X\) is functional. \(Y\) is multivariate response. In the literature, the multivariate responses issue with principal component analysis from covariate function has not been studied before. The achievement of the presented method is compared with the independent output from covariate function in the nonparametric approaches.

The article is organized as follows. Section 1 contains the model of estimation. Section 2 proposes the competence of the presented method through a simulation instance. Real data examples are presented in Section 4. Finally, a general conclusion is supplied in Section 5.

**Methodology**

Let \((X_1, Y_1), \ldots, (X_n, Y_n)\), be \(n\) pairs that are independently and identically distributed as \((X, Y)\) and valued in \(f \times R^q\), where \((f, d)\) is a semi-metric space and \((X(t))\) is a covariate function (taking values from infinite dimensions).

Let \(\mathcal{Y} = (y_1, \ldots, y_q)\) be a multivariate response variables in \(R^q\). In this work, the issue of nonlinear regression method is

\[ Y = \mathcal{Y} = \mathcal{Y}(X(t)) + \varepsilon \]

Suppose \(\hat{\mu}\) be the sample mean and \(\hat{\Sigma}\) be the sample covariance matrix of

\[ Y = (y_1, \ldots, y_q)^T = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nn} \end{bmatrix} \]

respectively, and have the eigenvalue-(normalised) eigenvector pairs \((\lambda_1, e_1), \ldots, (\lambda_q, e_q)\) where \(\lambda_1 \geq \lambda_2 \geq \ldots \lambda_q \geq 0\). The principal scores are given by the principal component analysis:

\[ \mathcal{Y} = (\mathcal{Y} - \bar{\mathcal{Y}}) \psi \]

where \(\psi = (e_1, \ldots, e_q)\) and \(\bar{\mathcal{Y}} = (\bar{y}_1, \ldots, \bar{y}_q)^T\) is an \(n \times q\) matrix. Letting \(\mathcal{y}_l = (\mathcal{y}_{1l}, \mathcal{y}_{2l}, \ldots, \mathcal{y}_{nl})\) be the \(l\)th column of \(\mathcal{Y}\), then \(\mathcal{y}_{1l}, \ldots, \mathcal{y}_{ql}\) are samples of \(q\) uncorrelated random variables. Then, the nonparametric regression function \(\hat{r}(\cdot)\) can be proposed by the connection between \(\mathcal{y}_{il}\) and \(X_i(t)\), that is, for \(l = 1, \ldots, q\) and \(i = 1, \ldots, n\)

\[ \mathcal{y}_{il} = r_i(X_i(t)) + e_{il} \]

where \(e_{il} \sim N(0, \sigma^2)\). Via K-Nearest Neighbour \(K - NN\) model, let’s predict \(r_i(\cdot)\): assume that

\[ r_i(\cdot) = \frac{\sum_{l=1}^{n} \mathcal{y}_{il} K(h^{-1}d(X_i(t), X_l))}{\sum_{l=1}^{n} K(h^{-1}d(X_i(t), X_l))} \]

where \(K\) is a kernel and \(h\) is a bandwidth (depending on \(n\)). The KNN estimator to determine optimal bandwidth of neighbours \(K_{opt}\) is defined by

\[ h_{K_{opt}} = \arg\min_h GCV(k), \]

where

\[ GCV(k) = \sum_{i=1}^{n} \left( \mathcal{y}_{il} - \hat{y}_k^{(-1)}(X_i) \right)^2 \]

with

\[ \hat{r}_k^{(-1)}(X) = \frac{\sum_{j=1, j \neq i}^{n} \mathcal{y}_{ij} K\left(d(X_j, X)/h_k(X)\right)}{\sum_{j=1, j \neq i}^{n} K\left(d(X_j, X)/h_k(X)\right)}. \]

In this work, we fixed the semi-metric \(d\) as the measure of closeness and the kernel function \(K\). Using the same number of neighbour for any curve provides a global choice, and \(h_{K_{opt}}\) relies on \(X\) (the bandwidth \(h_{K_{opt}}\) is such that only the \(k_{opt}\) nearest neighbours of \(X\) are taken into account) but \(k_{opt}\) is the same for any curve \(X\), for more details see \(^{8,11}\).

In the literature, different types of semi-metrics have been introduced. In our numerical instances, we used the semi-metrics based on Second Derivative and Functional Principal Component Analysis (FPCA); see for more details\(^{9,19}\).

Let \(X_1, \ldots, X_n\) be \(n\) curves and \(X = \{X(t)\}; t \in \tau\). Semi-metric based on FPCA is determined as

\[ d_q^{FPCA}(X_i, X_j) = \sqrt{\sum_{k=1}^{q} \left( \int (X_i(t) - X_j(t))v_k(t) \right)^2 dt}, \]

where \(v_1, \ldots, v_q\) are the orthonormal eigenfunctions of the covariance function \(\Gamma_X(s, t) = E(X(s)X(t))\) connected with the largest \(q\) eigenvalues; see for
more details. This kind of semi-metric is suitable for rough curves. Semi-metric built on derivatives is determined in more details.

\[
d_{q}^{deriv}(X_i, X_j) = \sqrt{\int \left( X_i^{(q)}(t) - X_j^{(q)}(t) \right)^2 dt}
\]

where \( X^{(q)} \) is the \( q \)th derivatives of \( X \) with regard to \( t \), which is computed using the B-spline approximation of the curves in exercise and see more details.

Suppose \( X^* \) is a test point and \( Y^* \) the corresponding response point. Therefore, the mean prediction \( \tilde{\mu} \) and variances of the scores then can be obtained by nonparametric functional regression and presented by \( \tilde{Y}_i^* \) and \( \tilde{\sigma}_i^* \) for \( i = 1, \ldots, m \). Thus the m-dimensional response \( Y^* \) for the predictive mean and variance are given by

\[
E(Y^*) = \tilde{\mu} + \nu \tilde{\gamma}
\]

and

\[
Var(Y^*) = Var(\tilde{\mu}) + \nu \Sigma^*(\nu)^T,
\]

where \( \gamma^* = (\tilde{Y}_1^*, \ldots, \tilde{Y}_m^*)^T, \Sigma^* = \text{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_m^2), \) and \( \text{Var}(\tilde{\mu}) = \Sigma / n \).

**Simulation Study**

The goal of this part is to verify the theoretical outcomes over the simulated data, which contains the sample of size \( n = 215 \). The outcomes get from the new model are compared with Independent response method. Using R program for analysing data.

Using the nonparametric functional regression:

\[
\tilde{Y}_i = r(X_i) + \varepsilon_i, \quad i = 1, \ldots, n = 215
\]

First of all, generating the curves:

\[
X_i(t_j) = \cos(t_j) + h_i(t_j - 0.5)^2 + g_i, \quad i = 1, \ldots, n
\]

when \( 0 = t_1 < t_2 < \ldots < t_{100} = 1 \) are equally spaced points and \( h_i, g_i \) are independently taken from a normal distribution \( h_i \sim N(0, (1)^2) \) and \( g_i \sim N(0, (1)^2) \). Figure 1 shows the 215 curves from one replication.

![Figure 1. 215 curves from one replication.](image)

The simulation of nonparametric functional regression model to compute the response variables which are calculated by two functional operators for building a regression operator \( r \) and they are expressed as

\[
\begin{align*}
\{r_1(X_i(t)) = \int_0^t X_i'(t) dt,
\{r_2(X_i(t)) = \int_0^t |X_i'(t)| \ log|X_i'(t)| dt,
\end{align*}
\]

Then compute the corresponding responses:

\[
\begin{align*}
\tilde{Y}_1 &= r_1(X_i) + \varepsilon_{i1}, \quad i = 1, 2, \ldots, n, \\
\tilde{Y}_2 &= r_2(X_i) + \varepsilon_{i2}, \quad i = 1, 2, \ldots, n.
\end{align*}
\]

where the error \( \varepsilon_i = (\tilde{Y}_{1i} - \tilde{Y}_{2i}) \sim N(0, \Sigma), \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \). Taking two different values \( \rho = (0, 0.9, 0.1) \) for correlations between the two response variables when the correlation 0.9 it means very strong correlation between them and 0.1 mentioned very week correlation, and using \( \sigma_1, \sigma_2 = 1, 2 \). Then the simulated sample is divided into two samples: first sample \( n = 160, \) from this sample we construct the model, and for testing sample \( n = 55, \) which is utilized to test the execution of the approach.

In both methods (Multivariate Response with Principal Component Analysis (M-P) and Independent Response method (I-R)) which used the quadratic kernel function and semi-metric built on the second derivative (\( q = 2 \)) for measure of closeness between curves. For the calculation of estimation execution, utilizing the root mean square error (RMSE) between the estimated values and the true values.

The pursuance of the proposed method (M-P) is discussed with that of I-R method where the two responses are determined independently and without taking into account the correlation between responses. The average of the RMSEs is presented in Table 1, after 20 iterations. Table 1 shows that, in both situations the Multivariate Response model considerably progresses the outcomes compared with each Independent Response (I-R) model. Then, it is clear that from Table 1, even no correlation between the components of the response variables the Multivariate method is more appropriate for prediction than the independent model.

**Real Application:**

In this part of the article, testing the presented method on two different kinds of real data sets, Tecator data and Soil data. The importance of applying two types of real data for conferring the proposed model outcomes is better than the models in the literature. The R program is used to analyse data in our study.
Table 1. The average RMSEs for the simulated study.

| case  | components | M-P | I-R |
|-------|------------|-----|-----|
| I     | 1          | 1.07| 1.08|
|       | 2          | 3.77| 3.81|
| II    | 1          | 1.06| 1.05|
|       | 2          | 4.305| 4.44|

Table 2. The RMSEs for the Fat, Water and Protein content.

| Responses | M-P | I-R |
|-----------|-----|-----|
| Fat       | 1.72| 1.879|
| Water     | 1.83| 2.10|
| Protein   | 1.42| 1.57|

Tecator data

This data is extremely common in the society of nonparametricians because various implementations have been done on it and by different models. Spectrometric Data arrives from the quality control problem and can be found at http://lib.stat.cmu.edu/datasets/tecator.

The objective of Tecator data is to permit the exposure of the proportion of the specific chemical meaning because the examination by chemistry procedure would take more time and be more costly. This instance works out when the response variable is scalar and covariates are function. Indeed, the correlation coefficients between 3-contents (Fat, Water, and Protein contents) are given by $\rho_{\text{Fat, Water}} = -0.988$, $\rho_{\text{Fat, Protein}} = -0.86$ and $\rho_{\text{Water, Protein}} = 0.82$. The three variables in meat (Fat, Water, and Protein) are strongly correlated so it will be more appropriate to estimate these contents together rather than each one, individually.

We divide the original sample into two sub-samples. The first 160 sample units are used for training sample and the second sample includes the last 55 for testing sample. Same as the simulation example, the RMSE is then computed for the methods as $\text{MSE} = \left(\frac{1}{55}\sum_{i=1}^{55} S_{e_{i}}\right)^{1/2}$.

Runing the funopare.knn.gcv function in R structure for estimation for independent response model, and it is valid on the site of Nonparametric Functional Data Analysis (NFDA). Also using the semi-metric based on the second derivative ($q=2$) for both models (independent response and Multivariate response vectors by principal component analysis). Table 2 is reported to discuss the capacity of the methods, taking 10 times randomly 55 testing sample curves then taking the average of 10 times.

Table 2 concludes that the M-P method notably progress the estimation accuracy for the Fat, Water and Protein compared to I-R method.

Soil Data

Rinnan and Rinnan analysed this data set originaly, after that took a sample of these data and utilized Gaussian process regression with multivariate response on two components soil organic matter (SOM) and ergosterol concentration (EC).

Table 3. The RMSEs for the SOM and EC.

| Responses | M-P | I-R |
|-----------|-----|-----|
| SOM       | 1.08| 3.61|
| EC        | 6.23| 46.11|

The soil data samples were obtained from a long-term field experiment at a subarctic fell in Abisiko, northern Sweden. The number of samples is 108, and the wave-length intervals of 400-2500 nm (visible and near infrared spectrum) which was scanned at 2 nm intervals with an INR spectrophotometer; for more detail see. Two component variables, Soil Organic Matter (SOM) was weighted as loss on ignition at 550 0C, and Eergosterol Concentration (EC) was defined through HPLC. As the functional covariates were smooth, the semimetric built on second derivative was adopted in our instase. To know the efficacy of the study, leave-one-out cross validation was undertaken, that is, each of the 108 samples was left as test data while the rest data were utilised for model training. Table 3 presents that, same as the previous example the Root Mean Square Errors is computed as the measure of efficiency for the comparison of two approaches (M-P model and I-R model).

The root mean square errors is computed as measure of efficiency for the compare two methods (M-P model and I-R model) as stated in Table 3.

The proposed M-P model presented significantly improves the efficiency of the prediction for both SOM and EC in comparison with the I-R model.

Conclusion:

This study presented a new model for nonparametric regression analysis where the covariate is functional and uses Principal Component Analysis to de-correlate the multivariate response variables. It utilized the formula of the Nadaraya Watson estimator (K-Nearest Neighbour (KNN)) for prediction. It is presented that the results obtained from a new model supplies better estimations when compared with the outcomes from the independent
output method. The evaluation of closeness between the functional covariates is through the semi-metrics. The use of the M-P model and I-R model is clarified through some numerical examples. The results obtained from the covariate is functional and uses Principal Component Analysis to de-correlate the multivariate response variables model were significantly improved than the Independent Response model for both study simulation and real data.

Acknowledgement:
The authors would like to thank all the participants of this study for their help and cooperation.

Authors' declaration:
- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Zakho.

Authors' contributions statement:
Sh. S. I.: She did the conception, and design of paper. K. M. T. O.: The paper was my idea. She did the acquisition and analysis of the data. B. W.: He did analysis, interpretation, revision and proofreading.

References:
1. Ramsay JO, Silverman BW . Functional Data Analysis. 2nd Ed., Springer, New York. 2005. https://link.springer.com/book/10.1007/b988888
2. Cao G, Wang S, Wang L. Estimation and inference for functional linear regression models with partially varying regression coefficients. Stat. 2020; 9(1): e286.
3. Klepsch J, Klüppelberg C. An innovations algorithm for the prediction of functional linear processes. J Multivar Anal. 2017 Mar 1; 155: 252-71.
4. Chiu J M, Müller H G, Wang J L. Functional response models. Stat Sin. 2004: 675–693.
5. Wang B, Xu A. Gaussian process methods for nonparametric functional regression with mixed predictors. Comput Stat Data Anal. 2019 Mar 1; 131: 80-90.
6. Görecki T, Krzyśko M, Waszak L, Wołyński W. Selected statistical methods of data analysis for multivariate functional data. Stat Pap. 2018 Mar; 59(1): 153-82.
7. Alheety MI. New versions of liu-type estimator in weighted and non-weighted mixed regression model. Baghdad Sci J. 2020 Mar 18; 17(1 (Suppl.)): 0361-.
كفاءة نموذج kNN باستخدام تحليل المكون الرئيسي الدالي في الانحدار الدالي اللامعملي

شيلان سعيد إسماعيل1 كردنستن محمد طاهر1 بو وينك2

1قسم الرياضيات، الكلية العلوم، جامعة زاخو، زاخو، العراق
2قسم الرياضيات، ليستر، جامعة ليستر، LE1 7RH، بريطانيا.

الخلاصة:

يهدف البحث إلى تحسين القدرة التنبؤية لمقدرات الانحدار الدالي اللامعملي من خلال مجموعة المنهجيات المقترحة تطبيقاً ونظرياً وهي استخدام تحليل المكون الرئيسي للإلتزام بين متغيرات متعددة الاستجابة ونموذج التقدير Nadaraya-Watson (k- nearest neighbour (KNN)). تم تقييم مقدار الارتباط بين المتغيرات المستقلة مع الاستجابة المستقلة. تم استخدام برامج R لتحليل البيانات. تم استخدام خوارزميات فك الارتباط من خلال التطبيق على مثالين، الأول حقيقي والثاني تجريبياً. النتائج تشير إلى أن الاستجابة اللامعملي متعددة المتغيرات أكثر كفاءة من تطبيق التحليل الا لمعملي لفكي المتغير لكل استجابة بشكل مستقل.

الكلمات المفتاحية: الانحدار اللامعملي، تحليل البيانات الوظيفية، الاستجابة متعددة المتغيرات، تحليل المكونات الأساسية، خمن KNN