Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
Published Online First: Suppl. November 2022 **2022, 19(6): 1612-1617** **E-ISSN: 2411-7986**

DOI: https://dx.doi.org/10.21123/bsj.2022.6476

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

***Shelan Saied Ismaeel*[1*]** ***Kurdistan M.Taher Omar*[1]** ***Bo Wang*[2]**

[1]Department of Mathematics, Faculty of Science, University of Zakho, Zakho, Iraq.
[2]Department 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 measureing 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 stimator, 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 Silverman[1] 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 exerimental 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 ifunctional 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 Laïb [15] 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 detials 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 ultizes 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, $\mathcal{Y}$ 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 $(X_1, \mathcal{Y}_1), \ldots, (X_n, \mathcal{Y}_n)$, be n pairs that are independently and identically distributed as $(X, \mathcal{Y})$ and valued in $f \times R^q$, where (f, d) is d a semi-metric space and $(X(t))$ is a covariate function (taking vaules from infinite dimensions).

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

$$\mathcal{Y} = r(X(t)) + \varepsilon \qquad 1$$

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

$$Y = (y_1, \ldots, y_n)^t = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{bmatrix}$$ 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 :

$$\boldsymbol{\gamma} = (\mathcal{Y} - \vartheta)\varphi. \qquad 2$$

where $\varphi = (e_1, \ldots, e_q)$ and $\vartheta = (\hat{\mu}, \ldots, \hat{\mu})^T$ is an $n \times q$ matrix. Letting $\boldsymbol{\gamma}_l = (\gamma_{1l}, \gamma_{2l}, \ldots, \gamma_{nl})$ be the lth column of $\boldsymbol{\gamma}$, then $\gamma_1, \ldots, \gamma_q$ are samples of q uncorrelated random variables. Then, the nonparametric regression function $r(.)$ can be proposed by the connection between $\boldsymbol{\gamma}_{il}$ and $X_i(t)$, that is, for $l = 1, \ldots, q$ and $i = 1, \ldots, n$

$$\boldsymbol{\gamma}_{il} = r_l(X_i(t)) + e_{il} \qquad 3$$

where $e_{il} \sim N(0, \sigma_l^2)$. Via K-Nearest Neighbour $K - NN$ model, let's predict $r_l(.)$: assume that

$$r_l(.) = \frac{\sum_{i=1}^{n} \boldsymbol{\gamma}_{il} K\left(h^{-1} d(\chi, X_i)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(\chi, X_i)\right)}.$$

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

$$h_{kopt} = arg\min_{h} GCV(k),$$

where

$$GCV(k) = \sum_{i=1}^{n} \left(\boldsymbol{\gamma}_{li} - m_{KNN}^{(-i)}(X_i)\right)^2$$

with

$$r_{KNN}^{(-1)}(\mathbf{x}) = \frac{\sum_{j=1, j\neq i}^{n} \boldsymbol{\gamma}_{li} K\left(d(X_j, \mathbf{x})/h_k(x)\right)}{\sum_{j=1, j\neq i}^{n} K\left(d(X_j, \mathbf{x})/h_k(x)\right)}.$$

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_{kopt} relies on (X) (the bandwidth h_{kopt} 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

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

where $\upsilon_1, \ldots, \upsilon_q$ 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 [9]. This kind of semi-metric is suitable for rough curves.

Semi-metric built on derivatives is determined in more details [8, 11]

$$d_q^{deriv}(X_i, X_j) = \sqrt{\int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt}.$$

where $X^{(q)}$ is the qth derivatives of X with regard to t, which is computed using the B-spline approximation of the curves in exercise and see [8] for more details.

Suppose $X^\star$ is a test point and $\mathcal{Y}^\star$ the corresponding response point. Therefore, the mean prediction $\hat{\mu}$ and variances of the scores then can be obtained by nonparametric functional regression and presented by $\boldsymbol{\gamma}_l^\star$ and $\hat{\sigma}_l^{\star 2}$ for $l = 1, \ldots, m$. Thus the m-dimensional response $Y^\star$ for the predictive mean and variance are given by

$$\mathbb{E}(\mathcal{Y}^\star) = \hat{\mu} + v\hat{\boldsymbol{\gamma}}$$

and

$$Var(\mathcal{Y}^\star) = Var(\hat{\mu}) + v\Sigma^\star(v)^T,$$

where $\boldsymbol{\gamma}^\star = (\hat{\boldsymbol{\gamma}}_1^\star, \ldots, \hat{\boldsymbol{\gamma}}_m^\star)^T, \Sigma^\star = diag(\hat{\sigma}_1^{\star 2}, \ldots, \hat{\sigma}_m^{\star 2}), and\ Var(\hat{\mu}) = \frac{\hat{\Sigma}}{n}$.

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:

$$\mathcal{Y}_i = r(X_i) + \varepsilon_i \quad i = 1, \ldots, n = 215$$

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.

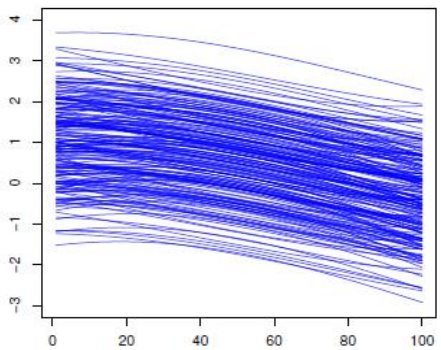

Figure 1. 215 curves from one replication.

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

$$\begin{cases} r_1(X_i(t)) = \int_0^\pi X_i'(t)dt, \\ r_2(X_i(t)) = \int_0^\pi |X_i'(t)| \quad log|X_i'(t)|dt, \end{cases}$$

Then compute the corresponding responses:

$$\mathcal{Y}_{1i} = r_1(X_i) + \varepsilon_{1i}, \quad i = 1,2,\ldots,n,$$
$$\mathcal{Y}_{2i} = r_2(X_i) + \varepsilon_{2i}, \quad i = 1,2,\ldots,n.$$

where the error $\varepsilon_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N(0, \sum), \Sigma = \begin{bmatrix} (\sigma_1)^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & (\sigma_2)^2 \end{bmatrix}$. Taking two different values $\rho = (0.9, 0.1)$ for correlations between the two response variables when the correlation 0.9 it means very strong correlation between them and 0.1 mensioned 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 Table1, 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.

		RMSE	
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.

	RMSE	
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 [9,10,11]. 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 for 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 [8, 20] 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 $\rho_{Fat,Water} = -0.988$, $\rho_{Fat,Protein} = -0.86$ and $\rho_{Water,Protein} = 0.82$. The three varibles 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 $MSE = (\frac{1}{55}\sum_{i=161}^{215} Se_i)^{1/2}$.

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.

Table2 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 [21] analysed this data set originaly, after that [16] 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.

	RMSE	
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 Abisko, northern Sweeden. The number of samples is 108, and the wave-length interval of 400-2500 nm (visible and near infrared spectrum) which was scanned at 2 nmintervals with an INR spectrophotometer; for more detail see [20]. Two component varibles, 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 rediction for both SOM and EC in comarison 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 uzed 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/b98888
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. Chiou 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. Górecki T, Krzyśko M, Waszak Ł, 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-.
8. Ferraty F, Vieu P. Nonparametric functional data analysis: theory and practice. SSBM. springer series in statistic. 2006 https://books.google.iq/books/about/Nonparametric_Functional_Data_Analysis.html?id=lMy6WPFZYFcC&printsec=frontcover&source=kp_read_button&hl=en&redir_esc=y.
9. Ferraty F, Vieu P. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. J Nonparametr Stat. 2004 Feb 1; 16(1-2): 111-25.
10. Chowdhury J, Chaudhuri P. Nonparametric depth and quantile regression for functional data. Bernoulli. 2019 Feb; 25(1): 395-423.
11. Ahmed MS, N'diaye M, Attouch MK, Dabo-Niang S. k-nearest neighbors prediction and classification for spatial data. arXiv preprint arXiv:1806.00385. 2018 Jun 1.
12. Dai W, Genton MG. Multivariate functional data visualization and outlier detection. J Comput Graph Stat . 2018 Oct 2; 27(4): 923-34.
13. Doori A. Hazard Rate Estimation Using Varying Kernel Function for Censored Data Type I. Baghdad Sci J. 2019 Sep 23; 16(3 (Suppl.)): 0793-.
14. Lam KK, Wang B. Robust non-parametric mortality and fertility modelling and forecasting: Gaussian process regression approaches. Forecasting 2021; 3(1): 207-227; https://doi.org/10.3390/forecast3010013
15. Chaouch M, Laïb N. Nonparametric multivariate L_1-median regression estimation with functional covariates. Electron. 2013 ; 7: 1553-1586.
16. Wang B, Chen T, Xu A. Gaussian process regression with functional covariates and multivariate response. Chemometr Intell Lab Syst. 2017; Apr; 163: 1-6.
17. Xiang D, Qiu P, Pu X. Nonparametric regression analysis of multivariate longitudinal data. Stat Sin. 2013 Apr 1: 769-89.
18. Omar KM, Wang B. Nonparametric regression method with functional covariates and multivariate response. Commun Stat Theory Methods. 2019 Jan 17; 48(2): 368-80.
19. Sugianto S, Rusdi M. Functional Data Analysis: An Initiative Approach for Hyperspectral Data. J Phys Conf Ser. 2019 Nov 1; 1363(1) : 012087.
20. Shang HL. Bayesian bandwidth estimation and semi-metric selection for a functional partial linear model with unknown error density. J Appl Stat. 2021 Mar 12; 48(4): 583-604.
21. Rinnan R, Rinnan Å. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. Soil Biol Biochem. 2007 Jul 1; 39(7): 1664-73.

Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
Published Online First: Suppl. November 2022 **2022, 19(6): 1612-1617** **E-ISSN: 2411-7986**

كفاءة نموذج kNN باستخدام تحليل المكون الرئيسي الدالي في الانحدار الدالي اللامعلمي

شيلان سعيد إسماعيل[1] كردستان محمد طاهر[1] بو وينك[2]

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

الخلاصة:

يهتم هذا البحث بتعزيز وتحسين القدرة التنبؤية لنماذج الانحدار الدالي اللا معلمي من خلال مجموعة المنهجيات المقترحة تطبيقاً و نظرياً وهي استخدام تحليل المكون الرئيسي الدالي لتقليل الارتباط بين متغيرات متعدد الاستجابة وتستند معادلة التقدير Nadaraya-Watson (k- nearest neighbour(KNN)) للتنبؤ باستخدام طريقتين وهي المشتقة الثانية وتحليل المكون الرئيسي الدالي من شبه المقياس لقياس المسافة بين المنحنيات. تم تقديرمطلق متوسط الاخطاء التربيعية للقيم المتوقعة لقياس كفاءة التنبؤ ومقارنتها مغ الاستجابة المستقلة. تم استخدام برنامج R لتحليل البيانات. عندما تكون المتغيرات المشتركة وظيفية وتم استخدام تحليل المكون الرئيسي لفك الارتباط . تم التطبيق على مثالين الاول حقيقي والمثال الثاني لبيانات مولدة تجريبياً. اثبتت النتائج ان استجابات اللا معلمية متعددة المتغيرات اكثر كفاءة من تطبيق التحليل اللا معلمي احادي المتغير لكل استجابة بشكلٍ مستقل.

الكلمات المفتاحية: الانحدار اللا معلمي، تحليل البيانات الوظيفية، الاستجابة متعددة المتغيرات، تحليل المكونات الأساسية, مخمن KNN