Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Use of nonparametric regression in mode B type measurement model: A simulation study approach

Syed Jawad Ali Shah *, Qamruz Zaman

Department of Statistics, University of Peshawar, Peshawar, Pakistan

ARTICLE INFO

Article history:
Received 18 August 2019
Received in revised form
11 November 2019
Accepted 12 November 2019

Keywords:
Latent variable model
Nonparametric regression
PLS-Path model
Formative model

ABSTRACT

In the conventional PLS-path modeling, the relationship among latent variables (LVs) is estimated by fitting a simple/multiple linear regression lines. For this purpose, researchers have to assume that the endogenous LV is the linear function of exogenous LVs, which is rarely met in real data analysis. The statisticians have devised a non-linear model-fitting approach to overcome the issue of linearity, but for that purpose, one should assume some specific functional form like quadratic, cubic or some degree of a polynomial in advance. Hence, when the linearity assumption is violated, the only appropriate choice is to use the nonparametric regression approaches. This study is mainly focused on the estimation of the latent variable model by incorporating three nonparametric smoothing procedures: Kernel regression estimate, local polynomial estimate, and smoothing spline estimates. An algorithm for LV models is proposed and presented based on nonparametric regression approaches for the mode B type measurement model (i.e., formative model). From simulation studies, it was clearly concluded that nonparametric based LV modeling approaches perform well for large sample sizes (i.e., for sample size 100 and above) as compared to standard PLS-path modeling procedure. However, for small samples (less than 100 observations), the standard PLS-path modeling procedure was giving better results.

1. Introduction

The popularity of LV models is increasing day-by-day not only in the fields of social and behavioral sciences but also got a wide application in the disciplines of economics, medical and management sciences for studying the relationship among LVs as well as their associated manifest variables (MVs). In conventional PLS-path modeling, the relationship among LVs is estimated by applying the multiple linear regression. For this purpose, the researcher has to assume that the endogenous LV, denoted by "η" is the linear function of exogenous LVs, denoted by

$\xi_1, \xi_2, \dots, \xi_k$.

$$E\{\eta|\xi_1, \xi_2, \dots, \xi_k\} = \sum_{i=1}^{k} \xi_i \beta_i \qquad (1)$$

where β_i denotes the path coefficient and is interpreted just like regression coefficients.

Although these models have a beauty that the researcher can easily interpret the coefficient values in terms of significant contribution, these models are quite restrictive. Especially, the two assumptions: linearity and additivity, make it sometimes very impractical. The statisticians have devised a non-linear model-fitting approach to overcome the issue of linearity, but for that purpose, practitioners should assume some specific functional form like quadratic, cubic or some degree of the polynomial in advance. Hence, the only choice in the case of non-linearity is to use the nonparametric regression approaches.

In the literature, the estimation of regression function using a nonparametric regression approach has been studied for a long time. The most popular estimates for nonparametric regression function include kernel regression estimate, local polynomial regression estimate, and smoothing spline estimates. According to Kelava et al. (2017), the use of nonparametric regression in the context of LVs is a newly emerged research area. Recently, they estimated the LV model without specifying the underlying distributions. They adopted a two-step

* Corresponding Author.
Email Address: jawadalishah78@gmail.com (S. J. A. Shah)
https://doi.org/10.21833/ijaas.2020.01.011
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-0579-1935
2313-626X/© 2019 The Authors. Published by IASE.

procedure: In the first step, the measurement model is estimated by using a common factor model, while in the second step, nonparametric regression using smoothing splines estimates was used to analyze the relation among LVs. They did not study the other nonparametric estimates like local polynomial or kernel regression estimate etc. Hence, there is sufficient room left for research in adopting other estimation procedures like kernel regression estimate, or local polynomial etc. in fitting a latent variable model.

This research study is mainly focused on the estimation of the PLS-path model (having the mode B type measurement model) by incorporating the above mentioned nonparametric smoothing procedures. Before presenting the proposed procedure, a brief review of PLS-PM is presented in the next sections, followed by a review of nonparametric regression techniques. In the last sections of this article, the results of simulation studies, as well as an application to real-world data, are presented.

2. Summary of PLS-PM and nonparametric regression

2.1. PLS-path modeling

PLS-Path Modeling is a statistical modeling approach; in which, several blocks of variables are linked together to measure linear dependence relationships among them. The history of PLS-Path Modeling starts with the advent of NILES (Non-linear Iterative Least Squares) (Wold, 1966). Later on, it was re-named by Wold (1973) as NIPALS (Non-linear Iterative PArtial Least Squares), which later on extended to PLS-Path Modeling (Wold, 1982).

PLS-PM is comprised of two parts: First part is called "the inner model (or structural model)" while the second part is known as "the outer model (or the measurement model)" (Lohmöller, 1989). The "inner model" specifies the relationships among latent variables, while the "outer model" specifies the relationships between latent variables and their associated MVs. A simple PLS-PM is depicted in Fig. 1.

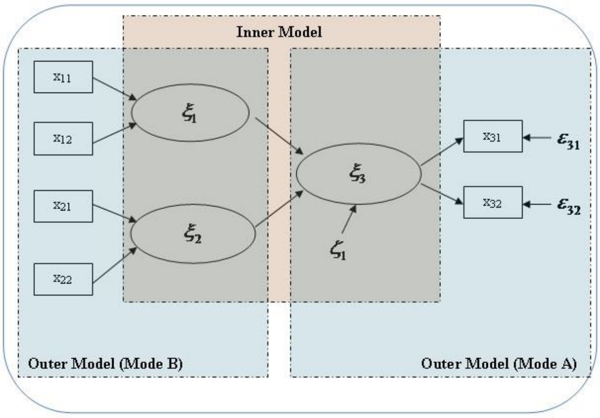

Fig. 1: PLS-path model

The PLS-PM algorithm suggested by Lohmöller (1989) comprises of following steps:

Step I: Initialization: To initialize the algorithm, any arbitrary numbers are chosen as weights to approximate the LV scores $\hat{\xi}$ or $\hat{\eta}$ by computing the linear combination of associated MVs. In simple words, each LV is constructed as a weighted sum of their associated MVs, and generally, the weights are all taken as equal to one (Monecke and Leisch, 2012). However, in the second and next iterations, the weights calculated at step number 4 are utilized.

Step II: Inner approximation: In this step, each LV is estimated by taking the weighted sum of other linked LVs. Now, the values of the weights, are depending on any of the three weighting schemes:

(i) centroid weighting scheme (Wold, 1982): Which utilizes the sign of the correlations between LVs (i.e., -1 or +1).

(ii) factor weighting scheme (Lohmöller, 1989): It takes the correlation values instead of their signs.

(iii) path weighting scheme (Lohmöller, 1989): Also known as a structural scheme, in which regression coefficients are taken as weights instead of correlation coefficients.

Step III: Outer approximation: In step I, all the weights were taken as "one" or any arbitrary number, but in this step, these weights are recalculated on the basis of estimated values of LVs obtained in step II according to the type of measurement model.

Step IV: Estimation of LV scores: The outer weights computed in **step III** are now used to estimate the LV scores by taking the weighted sum of their associated MVs.

Step V: Repeating the steps until convergence occurs: The process of inner approximation and outer approximation is repeated (i.e., loop of Step II to IV) until and unless the relative change between two consecutive iterations of all the outer weights become smaller than a prefixed threshold value or tolerance value (usually taken as 10^{-5}).

Step VI: Computing the path coefficients, loading coefficients, and total effects: Once the LV scores are finalized (after convergence of outer weights values), the path coefficients can be estimated by fitting multiple linear regression for each endogenous LV involved in the inner model.

2.2. Nonparametric regression

A major drawback of the classical parametric approach is that the observed data may fail to follow a specific parametric model and the incorrect modeling assumption may lead to seriously flawed statistical conclusions. The idea of nonparametric regression is to use models of the form:

$$Y_i = m(X_i) + \epsilon_i \tag{2}$$

where (X_i), some class of regression function, and ϵ is an independent and identically distributed random variable with zero mean and unit variance. The nonparametric regression does not impose any functional form assumption and estimates the relationship by a smooth curve. The most commonly used nonparametric regression techniques are kernel regression estimate, local polynomial regression estimate, splines smoothing estimate. The detail discussion related to all these nonparametric estimation methods are available in some excellent books like Härdle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Györfi et al. (2002), and Härdle et al. (2004) are few of them. However, a brief review of each is presented here.

2.2.1. Kernel regression estimate

Consider the simple case, that is, one predictor and one response variable, and the neighborhood points of X_0 be bounded in the interval $X_0 \pm h$, where “h” is called as bandwidth and always a positive real number. Then the nonparametric estimator of m(X) is given by:

$$\widehat{m}(X) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - X_0}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - X_0}{h}\right)}$$

which is known as "local constant" or "Nadarya-Watson" estimator. The smoothing parameter “h” (technically called bandwidth) is adjusted for the degree of smoothness. Here “K(.)” is Kernel function. There are various forms of Kernel function are available in the literature, and these might neither affect the estimates of regression function nor the form of density. For example, Uniform Kernel function may be expressed as:

$$K(x) = 1, \qquad x \in \left[\frac{-1}{2}, \frac{1}{2}\right]$$

The choice of “h” is usually done by trial and error, or by cross-validation. The level of smoothness depends on the value of “h”, i.e., smaller the value of “h”, the wigglier curve (wavy) will be, while a larger value of “h” produces a smooth curve.

2.2.2. Local linear estimate

NW estimator is a local constant approximation where the local constant is achieved by taking the average of Y values for all values of X lies in the interval $X_0 \pm h$. Another procedure, which fits a linear regression line locally (i.e., through the points lying in the same neighborhood), then this leads to a nonparametric technique known as Local Linear (LL) estimator. It's worthy to mention here that, if smoothing is increased i.e. when “h” approaches to infinity, the LL estimator and the parametric OLS estimator will be equal, but remember it is only true for a linear relationship.

2.2.3. Local polynomial estimate

To further improve the estimation, a local quadratic or cubic or polynomial of any order can be fitted rather than a local linear regression line. If "p" denotes the order of the local polynomial, then the local polynomial at p=0 will be equivalent to the NW estimator, while p=1 and p=2 will be exactly equal to Local Linear (LL) and local quadratic estimators respectively.

2.2.4. Splines smoothing regression

A spline is defined as a piecewise polynomial having pieces connected by a sequence of knots $\varphi_1 < \varphi_2 < \ldots < \varphi_k$ such that these pieces are joining smoothly at these knots. The Spline may be linear or of any degree. A spline of degree "d" is generally expressed as:

$$S(x) = \sum_{j=0}^{d} \beta_j x^j + \sum_{j=1}^{k} \lambda_j (x - \varphi_j)_+^d$$

which is a power series and where,

$$(x - \varphi_j)_+ = \begin{cases} x - \varphi_j, & x > \varphi_j \\ 0, & otherwise \end{cases}$$

Hence, if d=1, then the linear spline will be of the form:

$$S(x) = \beta_0 + \beta_1 x + \lambda (x - \varphi)_+$$

3. Methodology

3.1. The proposed procedure for using nonparametric regression in PLS-Path modeling

The existing procedure of PLS-path modeling consists of six steps, which are already illustrated in subsection 1.1. To fit the LV model using the PLS-path modeling approach, the linearity pattern among LVs is assumed, which may not be fulfilled at every situation (as discussed in the Introduction section). In this section, a fully nonparametric algorithm for LV models is proposed by modifying the existing methodology of the PLS-path modeling approach. The modification is done in two places:

1. A nonparametric weighting scheme is proposed based on LOESS (Sen, 1968) approach, i.e. similar to path weighting scheme (Lohmöller, 1989), i.e., the median of slopes for local linear lines are taken as weights.
2. After finalizing the LV scores, the nonparametric regression smoothers (kernel smoothing or local polynomial regression or splines smoothing regression) is adopted to estimate the relationship among LVs instead of fitting simple/multiple linear regression.

3.2. Simulation study

In literature, Monte Carlo simulation is extensively used to empirically assess the performance of statistical procedures under certain conditions, like the size of the model, sample size etc. In LV models literature, most of the studies are designed under the guidelines provided in Paxton et al. (2001). In this section, three simulation studies are designed to investigate the performance of the proposed nonparametric LV modeling algorithm for a formative model (Mode B) keeping in view the guidelines of Paxton et al. (2001). The R programming language is used to code the program for the proposed algorithm (with a certain level of modifications in the "plspm (version 0.4.9)" package (Sanchez et al., 2015).

The three simulation studies are designed (Ranging from simple to complex) small to large sample sizes (i.e., seven different sample sizes starting from 20, 30, 50, 100, 200, 300, and 500). The numbers of replications are fixed at 500. The following models were fitted on each data set: Conventional PLS-path modeling, and proposed NP-based LV modeling with three different smoother approaches i.e., kernel smoothing, local polynomial smoothing (degree=0, degree=1, and degree=2) and spline smoothing. The consistency threshold is fixed at 0.00001. The performance of a model can be judged by considering how much the predicted values are closer to observed values. Two different consistency criteria MAE (Mean Absolute Error) and RMSE (Root Mean Square Error), are used to compare the performance of nonparametric-based path modeling and the existent PLS-PM approach. The predicted values of LV scores are determined by a 10-fold cross-validation approach for each sample size at each iteration. The simulation results are presented by tabular form (the amount of MAE and RMSE) in section 4.

3.2.1. Simulation study 1

The simplest model is considered in this first simulation study, by taking one endogenous LV and one exogenous latent variable having two MVs associated with each. The path coefficient and loading values are fixed at 0.7, as these were taken by many researchers for assessing the performance of PLS-path modeling. The specified model is depicted in Fig. 2.

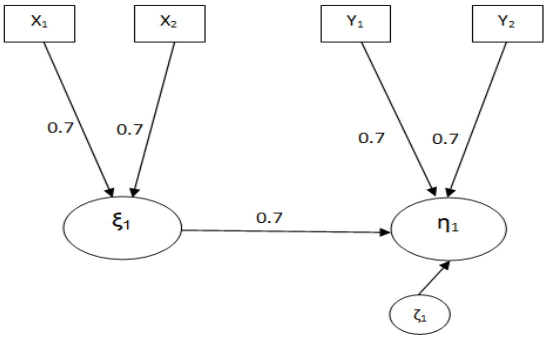

Fig. 2: Specification of the LV model for simulation study 1

Using this specification, 3500 datasets are generated with seven different sample sizes starting from 20, 30, 50, 100, 200, 300, and 500, i.e. 500 replications for each sample size produces 500 X 7= 3500 datasets. The unit value (i.e., 1) is used as an initial approximation for weights. Further, different skewness values (-3, -4) and kurtosis values (5, 6) are applied to generate non-normal data for each MV. The path weighting scheme is applied in conventional PLS-path modeling while the LOESS approach is incorporated for nonparametric-based LV Modeling approaches. The results of MAE and RMSE for each standardized parameter estimate are presented in Table 1.

Table 1: Simulation results for the specified simplest LV model involving two latent variables

	Sample size	PLS-PM	Kernel	Local Polynomial Spline			Spline
				Degree=0	Degree=1	Degree=2	
MAE	20	0.7750	0.9266	0.7881	0.8795	1.8003	0.8113
	30	0.7580	0.8868	0.7703	0.8223	0.9242	0.7882
	50	0.7545	0.8767	0.7542	0.7735	0.8421	0.7541
	100	0.7533	0.8140	0.7475	0.7658	0.7831	0.7527
	200	0.7467	0.7963	0.7390	0.7492	0.7658	0.7383
	300	0.7446	0.7957	0.7360	0.7452	0.7544	0.7339
	500	0.7402	0.7783	0.7346	0.7435	0.7452	0.7248
RMSE	20	1.0030	1.2067	1.0349	1.3023	3.1764	1.0793
	30	0.9997	1.1598	1.0276	1.2409	1.6388	1.0001
	50	0.9977	1.1019	0.9965	1.0627	1.4146	0.9938
	100	0.9947	1.0760	0.9886	1.0422	1.1154	0.9901
	200	0.9877	1.0493	0.9832	0.9893	1.0514	0.9824
	300	0.9862	1.0341	0.9818	0.9874	1.0202	0.9807
	500	0.9770	1.0041	0.9720	0.9804	0.9836	0.9722

3.2.2. Simulation study 2

Another model which is more complex than model 1 is considered in this simulation study, by taking one endogenous LV and two exogenous LVs having three MVs associated with each. The path coefficient and loading values are fixed at 0.6, as these were taken by many researchers for assessing the performance of PLS-path modeling (Paxton et al., 2001). The specified model is depicted in Fig. 3.

Using these specifications, 3500 datasets are generated with seven different sample sizes starting from 20, 30, 50, 100, 200, 300, and 500, i.e. 500 replications for each sample size produces 500 X 7=

3500 datasets. The unit value (i.e., 1) is used as an initial approximation for weights. Further, different skewness values (-3, -4, -5) and kurtosis values (5, 6, 7) are applied to generate non-normal data for each associated MV. The path weighting scheme is applied in conventional PLS-path modeling while the LOESS approach is incorporated for nonparametric-based LV modeling approaches. The results of MAE and RMSE for each standardized parameter estimate are presented in Table 2.

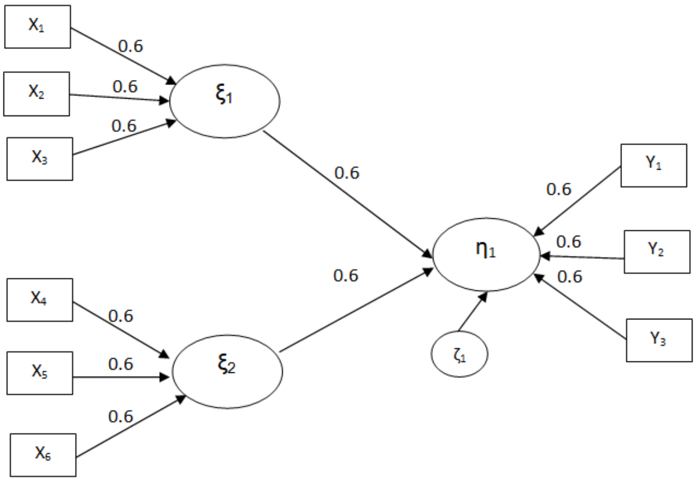

Fig. 3: Specification of the LV model for simulation study 2

Table 2: Simulation results for the specified LV model involving three latent variables

	Sample Size	PLS-PM	Kernel	Local Polynomial			Spline
				Degree=0	Degree=1	Degree=2	
MAE	20	0.6666	0.7969	0.6995	0.7898	1.3279	0.6761
	30	0.6657	0.7601	0.6992	0.7553	1.3043	0.6696
	50	0.6638	0.7067	0.6923	0.7202	0.8355	0.6669
	100	0.6624	0.6994	0.6806	0.6978	0.8246	0.6622
	200	0.6617	0.6955	0.6607	0.6818	0.7369	0.6586
	300	0.6543	0.6571	0.6524	0.6712	0.6722	0.6505
	500	0.6426	0.6509	0.6414	0.6499	0.6551	0.6381
RMSE	20	0.8985	1.1740	0.9394	1.2320	2.6228	0.9016
	30	0.8947	1.1942	0.9370	1.1647	2.2319	0.8986
	50	0.8930	1.0284	0.9358	1.0443	1.4006	0.8957
	100	0.8923	0.9671	0.9154	0.9721	1.2518	0.8935
	200	0.8918	0.9374	0.8938	0.9647	1.0048	0.8820
	300	0.8883	0.9049	0.8864	0.9577	0.9913	0.8789
	500	0.8809	0.9005	0.8768	0.9099	0.9428	0.8686

3.2.3. Simulation study 3

Another more complex model is considered in this simulation study, by taking two endogenous LV and three exogenous LVs having three MVs associated with each. Here, to make it more complex, the loading coefficient and structural path coefficients are also not fixed but varied to become more representative for real-world models. The loading coefficients are taken as 0.7, 0.6 and 0.5, while structural path coefficients are fixed at 0.5 and 0.6 for both endogenous LVs. The specified model is depicted in Fig. 4.

Using these specifications, 3500 datasets are generated with seven different sample sizes starting from 20, 30, 50, 100, 200, 300, and 500, i.e. 500 replications for each sample size produces 500 X 7= 3500 datasets. The unit value (i.e., 1) is used as an initial approximation for weights. Further, different Skewness values (-3, -4, -5) and kurtosis values (5, 6, 7) are applied to generate non-normal data for each associated MV. The path weighting scheme is applied in conventional PLS-path modeling while the LOESS approach is incorporated for nonparametric LV modeling approaches. Here, the model involves two endogenous variables, so the prediction performance of these two LVs are tabulated in Table 3 and Table 4. While the results of overall prediction performance in terms of MAE and RMSE are presented in Table 5.

4. Results and discussion

4.1. Discussion of simulation results

The results for the simplest model involving two LVs (one endogenous and one exogenous LV) presented in Table 1, showed that the amount of MAE and RMSE reduces as the sample size increases for all approaches. Further, by comparing the results row-wise, it can be concluded that a sample size of 20 and 30, the conventional PLS-PM approach gives better prediction performance (MAE= 0.7750, 0.7580 and RMSE= 1.0030, 0.9997), the smallest amount as compare to Kernel or local polynomial or spline-based approaches. But as the sample size increases, the local polynomial at degree=0 (i.e., constant local line approach) and spline-smoothers give better results. At sample size 100 and above, the spline-smoother gives more stable and better results than other approaches. However, this is applicable only in this case, when the model consists of two LVs having total of four indicators.

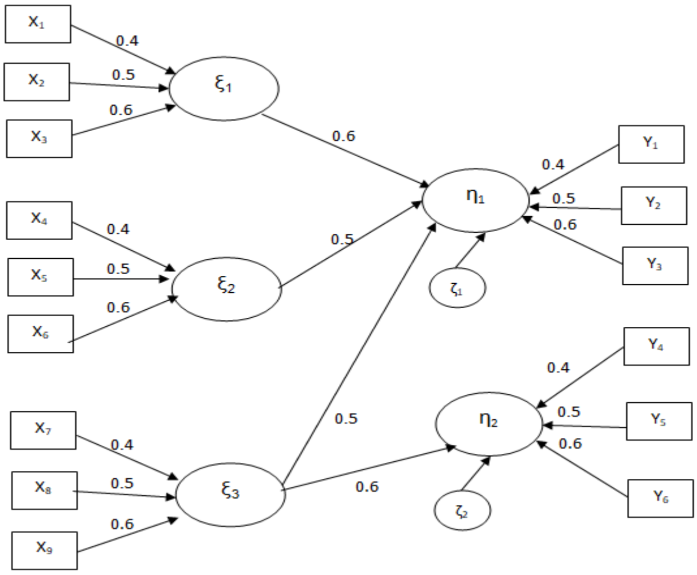

Fig. 4: Specification of the LV model for simulation study 3

Table 3: Prediction performance of the proposed NP-LV model for η_1 based on ξ_1 and ξ_2

	Sample Size	PLS-PM	Kernel	Local Polynomial			Spline
				Degree=0	Degree=1	Degree=2	
MAE	20	0.7326	0.8199	0.7425	0.8512	1.3620	0.7412
	30	0.7307	0.8094	0.7418	0.8316	1.1018	0.7437
	50	0.7205	0.7507	0.7304	0.8197	1.0420	0.7314
	100	0.7172	0.7448	0.7229	0.7592	0.8859	0.7252
	200	0.7154	0.7356	0.7147	0.7258	0.7511	0.7137
	300	0.7122	0.7290	0.7097	0.7245	0.7452	0.7066
	500	0.7107	0.7164	0.7053	0.7163	0.7431	0.7028
RMSE	20	0.9524	1.0570	0.9719	1.2945	2.5159	0.9863
	30	0.9510	1.0233	0.9655	1.2534	2.1165	0.9840
	50	0.9491	0.9977	0.9520	1.1273	2.1352	0.9541
	100	0.9448	0.9913	0.9494	1.0311	1.5203	0.9493
	200	0.9420	0.9879	0.9405	0.9464	0.9944	0.9396
	300	0.9406	0.9652	0.9371	0.9450	0.9823	0.9356
	500	0.9397	0.9500	0.9321	0.9415	0.9801	0.9303

Table 4: Prediction performance of the proposed NP-LV model for η_2 based on ξ_2 and ξ_3

	Sample Size	PLS-PM	Kernel	Local Polynomial			Spline
				Degree=0	Degree=1	Degree=2	
MAE	20	0.7338	0.7427	0.7512	0.7632	1.2611	0.7660
	30	0.7283	0.7405	0.7480	0.7618	1.0205	0.7347
	50	0.7203	0.7389	0.7340	0.7578	0.9826	0.7255
	100	0.7171	0.7347	0.7192	0.7395	0.8411	0.7193
	200	0.7163	0.7260	0.7139	0.7227	0.8339	0.7129
	300	0.7116	0.7252	0.7095	0.7186	0.7339	0.7022
	500	0.7109	0.7241	0.7016	0.7133	0.7216	0.6974
RMSE	20	0.9565	0.9977	1.0033	1.1709	2.3676	1.1432
	30	0.9546	0.9901	0.9977	1.1425	1.9049	0.9597
	50	0.9507	0.9861	0.9609	1.0452	1.1557	0.9536
	100	0.9484	0.9825	0.9512	0.9654	1.0753	0.9495
	200	0.9467	0.9557	0.9428	0.9527	1.0062	0.9323
	300	0.9413	0.9446	0.9345	0.9487	0.9559	0.9266
	500	0.9399	0.9417	0.9222	0.9417	0.9531	0.9201

The results tabulated in Table 2 for a model involving three LVs (one endogenous and two exogenous LV) showed that the amount of MAE and RMSE reduces as the sample size increases for all approaches. Further, by comparing the results row-wise, it can be concluded that a sample size of 20, 30, and 50 the conventional PLS-PM approach gives better prediction performance (MAE= 0.6666, 0.66657, 0.6638 and RMSE= 0.8985, 0.8947, 8930), the smallest amount as compare to Kernel or local polynomial or spline-based approaches. But as the sample size increases, the local polynomial at degree=0 (i.e., constant local line approach) and spline-smoothers give better results. At sample size 100 and above, the spline-smoother gives more stable and better results than other approaches.

From these as well as from Tables 3-5 results, spline-smoothing outperforms in case of large samples.

Table 5: Simulation results for the overall prediction performance of the specified LV model involving five latent variables

	Sample Size	PLS-PM	Kernel	Local Polynomial			Spline
				Degree=0	Degree=1	Degree=2	
MAE	20	0.7332	0.7813	0.7469	0.8072	1.3116	0.7536
	30	0.7295	0.7750	0.7449	0.7967	1.0612	0.7392
	50	0.7204	0.7448	0.7322	0.7888	1.0123	0.7285
	100	0.7172	0.7398	0.7211	0.7494	0.8635	0.7223
	200	0.7159	0.7308	0.7143	0.7243	0.7925	0.7133
	300	0.7119	0.7271	0.7096	0.7216	0.7396	0.7044
	500	0.7108	0.7203	0.7035	0.7148	0.7324	0.7001
RMSE	20	0.9545	1.0274	0.9876	1.2327	2.4418	1.0648
	30	0.9528	1.0067	0.9816	1.1980	2.0107	0.9719
	50	0.9499	0.9919	0.9565	1.0863	1.6455	0.9539
	100	0.9466	0.9869	0.9503	0.9983	1.2978	0.9494
	200	0.9444	0.9718	0.9417	0.9496	1.0003	0.9360
	300	0.9410	0.9549	0.9358	0.9469	0.9691	0.9311
	500	0.9398	0.9459	0.9272	0.9416	0.9666	0.9252

The results tabulated in Tables 3, Tables 4 and Tables 5 for a complex model involving five LVs (two endogenous and three exogenous LV) showed that the amount of MAE and RMSE reduces as the sample size increases for all approaches. Further, by comparing the results row-wise, it can be concluded that at sample size up to 100, the conventional PLS-PM approach gives better prediction performance (MAE= 0.7332, 0.7295, 0.7204, 0.7172 and RMSE= 0.9545, 0.9528, 0.9499, 0.9466), the smallest amount as compare to Kernel or local polynomial or spline-based approaches.

But as the sample size increases, the local polynomial at degree=0 (i.e., constant local line approach) and spline-smoothers give better results. At sample size 100 and above, the spline-smoother gives more stable and better results than other approaches. Hence, from all these simulation results, spline-smoothing outperforms in the case of large samples.

4.2. Application of proposed procedure on real data set: Offense model

In this section, the proposed nonparametric-based path modeling is applied on a real data set "Offense". The data set "Offense" contains the offense statistics of American's National Football League (NFL) for the season 2010-11. The "offense" data set is freely available in "plspm" package in R, or it can be downloaded from www.teamrankings.com. The data set contains 32 observations on 17 manifest variables. These 17 MVs are associated with five latent variables:

- Rushing Quality (includes three MVs: Rush1 to Rush3),
- Passing Quality (includes three MVs: Pass1 to Pass3),
- Special Teams and others (includes two MVs: Spec1 to Spec2),
- Scoring success (includes three MVs: Scor1 to Scor3),
- Offense performance (includes six MVs: Offen1 to Offen6).

For further details on each MV, see Sanchez and Trinchera (2012). The full structural and measurement model for the offense model is sketched in Fig. 5.

There are three exogenous LV involve in this model (i.e., Special, Rushing and Passing) while the Scoring and Offense LVs are depending on one or more than one LVs. Suppose this model is fitted by the PSL-path modeling technique and factor scores are computed. To study the relationship pattern among these LVs, the scatterplot of each endogenous LV vs exogenous LV is sketched and depicted in Fig. 6.

From these scatterplots, it is evident that one of the plots don't exhibit a linear pattern between endogenous and exogenous LVs, i.e., scoring vs. Special. So, it is clearly an indication of a violation of the linearity assumption. Hence, the only choice in the case of nonlinearity is to use the nonparametric regression approaches. The proposed procedure is applied to the "offense" data set and the factor scores for Scoring are predicted using PLS-PM, local polynomial (degree=0) and spline approaches. The performance of NP-LV models is assessed via MAE and RMSE, computed through a one-leave-one-out cross-validation approach. The MAE and RMSE amounts, as well as predicted factor scores for initial twenty observations, are tabulated in Table 6.

The predicted factor scores shown in Table 6 are obtained by applying the conventional PLS-PM approach and two nonparametric Local polynomials (degree=0) and spline smoothing indicate that the predictions will not same. For example, the predicted factor scores for the fifth observation are -0.5149, -0.3932 and -0.7007. The reason is that: In the conventional PLS-PM approach, simple or multiple linear regression lines are globally fitted while using nonparametric approaches local lines or curves are fitted. Further, the prediction performance measures (MAE and RMSE) also indicate that spline smoothing and local polynomial are giving better performance for the prediction of factor scores of Scoring.

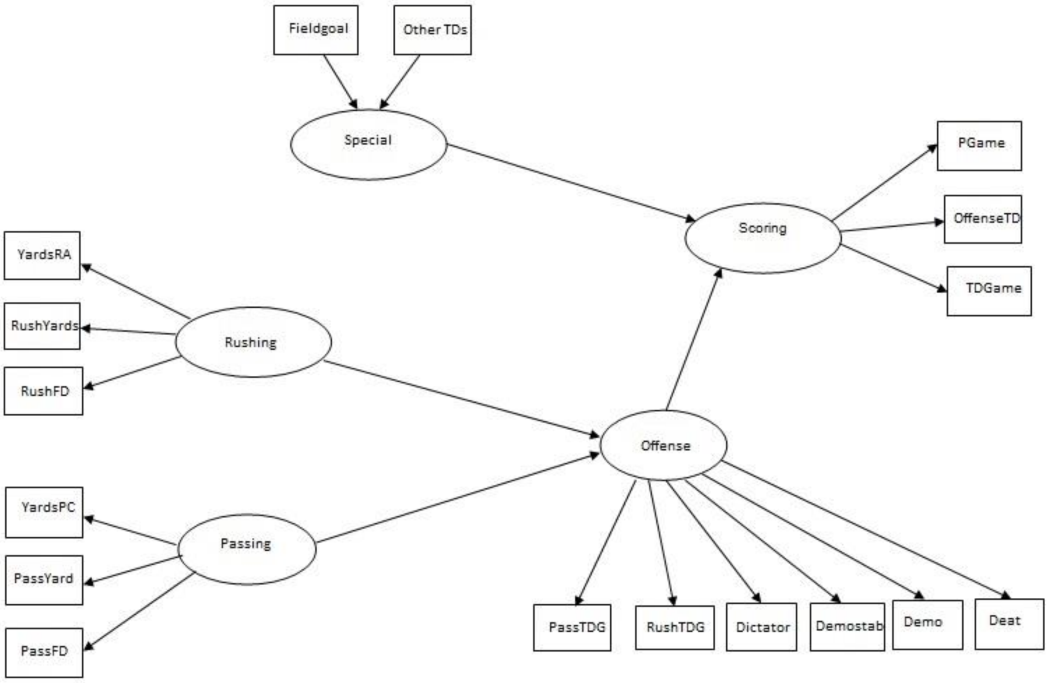

Fig. 5: LV model sketch of offence model

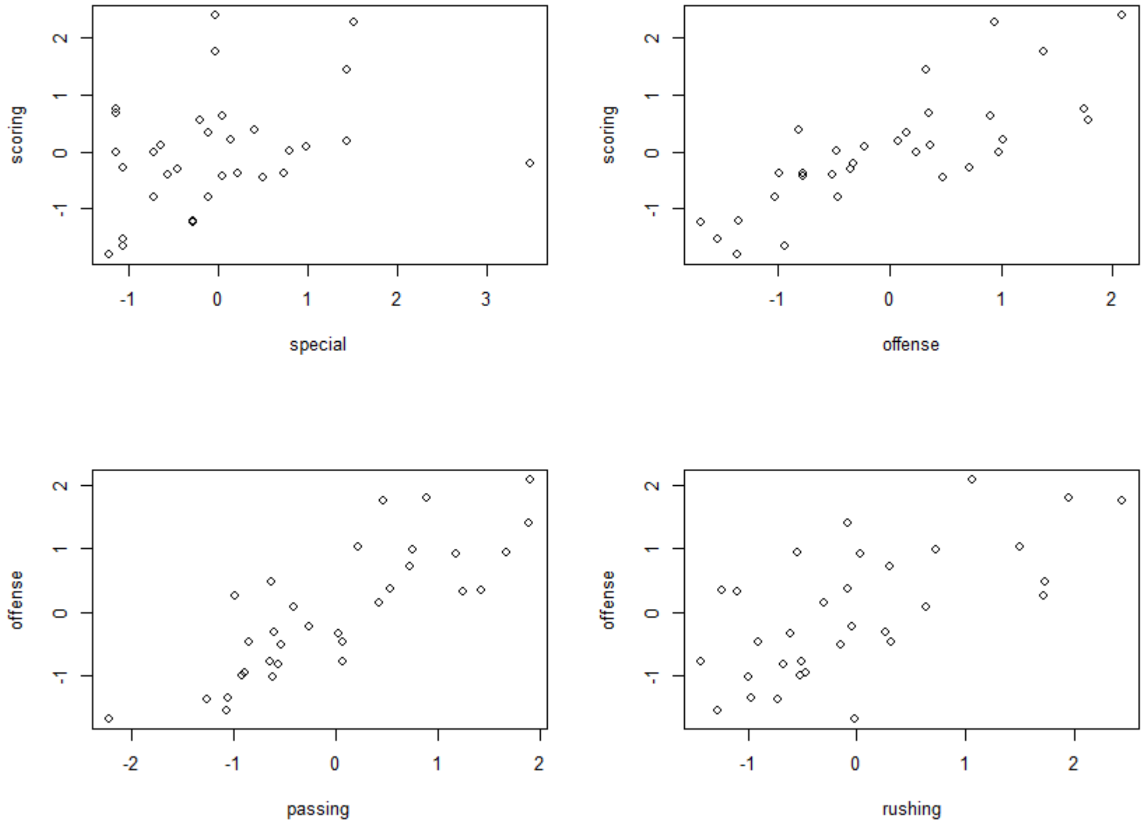

Fig. 6: Scatter plots for LVs of Offense model

5. Conclusion

In this study, an algorithm based on nonparametric regression is proposed for LV path modeling having measurement models of Formative type (Mode B). Three approaches: Kernel regression, local polynomial regression and spline smoothers are implemented to get the relationship among LVs and finally to get the predicted factor scores of endogenous LVs. The performance of the proposed procedure is assessed by conducting a variety of simulation designs (simple to complex) and results are computed through computing MAE and RMSE. Although simulation results give a clear indication that the conventional PLS-PM approach is performing well at small sample sizes, while nonparametric-based proposed procedure outperforms in case of large sample size (i.e., sample size 100 and above). The literature also recommends that nonparametric regression should be used for large sample sizes. But, when the linearity assumption is violated the only choice is to use nonparametric regression, otherwise, prediction results will be over or under-estimated. In the future, the current research can be extended by introducing the interaction effects in the model.

Table 6: Predicted Factor Scores and prediction performance measures for Scoring LV

	Observation No.	PLS-PM	Local Polynomial (degree=0)	Spline
Predicted Factor Scores	1	-0.0564	0.0198	-0.0569
	2	-0.0182	-0.0331	0.0733
	3	0.3329	0.1653	0.3718
	4	0.5334	0.2534	0.5143
	5	-0.5149	-0.3932	-0.7007
	6	2.1805	2.0000	2.1931
	7	0.0195	-0.0303	0.1484
	8	-0.2706	-0.0840	-0.3720
	9	-0.1484	-0.0417	-0.1744
	10	0.0958	0.0422	0.2744
	11	0.3555	0.1248	0.3338
	12	0.2925	0.1748	0.2646
	13	-0.0175	-0.0676	0.0719
	14	0.0567	0.1170	0.0888
	15	-0.1080	-0.0659	-0.0334
	16	-0.1286	0.0510	-0.2547
	17	-0.0095	0.0928	-0.0293
	18	-0.6127	0.1122	-0.6226
	19	-0.1009	-0.1253	0.0086
	20	-0.0709	-0.1015	0.0059
	MAE	0.8587	0.8256	0.8563
	RMSE	1.0749	1.0554	1.0722

List of symbols

d	*Degree of spline*
h	*Bandwidth*
p	*Order of local polynomial*
X	*Manifest variable associated with exogenous LV*
Y	*Manifest variable associated with endogenous LV*
β	*Path coefficients*
ϵ	*Disturbance term in regression model*
φ	*Knot position in spline*
η	*Exogenous Latent Variable*
ξ	*Endogenous Latent Variable*

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Fan J and Gijbels I (1996). Local polynomial modelling and its applications. Chapman and Hall, London, UK.

Györfi L, Kohler M, Krzyzak A, and Walk H (2002). A distribution-free theory of nonparametric regression. Springer Science and Business Media, New York, USA. https://doi.org/10.1007/b97848

Härdle W (1990). Applied nonparametric regression. Cambridge University Press, Cambridge, USA. https://doi.org/10.1017/CCOL0521382483

Härdle WK, Müller M, Sperlich S, and Werwatz A (2004). Nonparametric and semiparametric models: An introduction. Springer, New York, USA. https://doi.org/10.1007/978-3-642-17146-8

Kelava A, Kohler M, Krzyżak A, and Schaffland TF (2017). Nonparametric estimation of a latent variable model. Journal of Multivariate Analysis, 154: 112-134. https://doi.org/10.1016/j.jmva.2016.10.006

Lohmöller JB (1989). Latent variable path modeling with partial least squares. Springer-Verlag Berlin Heidelberg, Berlin, Germany. https://doi.org/10.1007/978-3-642-52512-4

Monecke A and Leisch F (2012). SEMPLS: Structural equation modeling using partial least squares. Journal of Statistical Software, 48(3): 1-32. https://doi.org/10.18637/jss.v048.i03

Paxton P, Curran PJ, Bollen KA, Kirby J, and Chen F (2001). Monte Carlo experiments: Design and implementation. Structural Equation Modeling, 8(2): 287-312. https://doi.org/10.1207/S15328007SEM0802_7

Sanchez G and Trinchera L (2012). PLS-PM: Partial least squares data analysis methods. Available online at: https://bit.ly/36rb1ko

Sanchez G, Trinchera L, and Russolillo G (2015). PLS-PM: An R package for partial least squares path modeling. Available online at: https://bit.ly/36G5Jlx

Sen PK (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324): 1379-1389. https://doi.org/10.1080/01621459.1968.10480934

Wand MP and Jones MC (1995). Kernel smoothing. Chapman and Hall, London, UK. https://doi.org/10.1007/978-1-4899-4493-1

Wold H (1966). Estimation of principal components and related models by iterative least squares. In: Krishnaiaah PR (Ed.), Multivariate analysis: 391-420. Academic Press, New York, USA.

Wold H (1973). Nonlinear iterative partial least squares (NIPALS) modeling: Some current. In: Krishnaiah PR (Ed.), Multivariate analysis: In the international symposium on multivariate analysis: 383-407. Academic Press, New York, USA. https://doi.org/10.1016/B978-0-12-426653-7.50032-6

Wold H (1982). Soft modeling: the basic design and some extensions. In: Jőreskog KG and Wold H (Ed.), Systems under indirect observations: Causality, structure, prediction-part 2: 1-54. North-Holland Publisher, Amsterdam, Netherlands.