International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 39 العدد
Part 1 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Received	**2026/09/02**	**تم استلام الورقة العلمية في**
Accepted	**2026/09/22**	**تم قبول الورقة العلمية في**
Published	**2026/09/23**	**تم نشر الورقة العلمية في**

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA

Nawal Mohamed Othman Mustafa
Faculty of Science, Al-Marj, University of Benghazi - Libya
nawal.mustafa@uob.edu.ly

Abstract

The objective of this paper is to provide an empirical comparison between a linear dimensionality reduction technique, referred to as principal component analysis (PCA), and a nonlinear dimensionality reduction technique, that is, kernel principal component analysis (KPCA), used as a preprocessing step for nonparametric regression. The reason for this contrast is the curse of dimensionality, which is likely to worsen the performance of nonparametric regression with an increasing number of explanatory variables. However, although KPCA constitutes a nonlinear analog of PCA, this does not imply that KPCA would always necessarily provide better predictive performance than PCA, given that there are nonlinear terms involved in the regression function. Thus, we study the scenarios where PCA is good enough and those where KPCA is expected to bring some improvements in accuracy. Using Monte Carlo simulation, the present study was performed with varying sample sizes ($n = 20, 60, 150$), numbers of explanatory variables ($p = 5, 10$), and numbers of retained components ($q = 2, 4$) in the R environment. Three regression functions with increasingly complex nonlinearity were considered, with predictors and random errors drawn i.i.d. from normal distributions. After the dimensionality reduction, the Nadaraya–Watson kernel regression estimator was estimated using the components derived from PCA

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

and KPCA, where a Gaussian kernel was used in both steps. The predictive performance was evaluated using the Average Prediction Error (APE) and Mean Squared Prediction Error (MSPE) for 500 Monte Carlo replications for each simulation setting. The results show that the predictive performance improves with the number of retained components, from q = 2 to q = 4, and that larger sample sizes tend to decrease both APE and MSPE. For a fixed number of retained components, increasing the number of explanatory variables generally leads to higher prediction errors. The PCA vs. KPCA comparison does not show a consistent superiority of one over the other, but their relative performance depends on the sample size, predictor dimension, number of retained components, and the nonlinearity structure of the regression function.

Keywords: Dimensionality reduction, Principal Component Analysis, PCA, Kernel PCA, KPCA, Non-parametric regression, Curse of dimensionality, Nonlinearity.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 39 العدد
Part 1 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

تقليل الأبعاد الخطي وغير الخطي وأثره على الانحدار اللامعلمي: مقارنة باستخدام محاكاة مونت كارلو بين تحليل المكونات الرئيسية (PCA) وتحليل المكونات الرئيسية المعتمد على النواة(KPCA)

نوال محمد عثمان مصطفي
كلية العلوم المرج، جامعة بنغازي – ليبيا
nawal.mustafa@uob.edu.ly

الملخص

يهدف هذا البحث إلى إجراء مقارنة تجريبية بين تقنية خطية لخفض الأبعاد، تُعرف باسم "تحليل المكونات الرئيسية" (PCA)، وتقنية غير خطية لخفض الأبعاد، وهي "تحليل المكونات الرئيسية المعتمد على النواة" (KPCA)، وذلك لاستخدامهما كخطوة معالجة أولية في الانحدار غير المعلمي (nonparametric regression). ويعود سبب هذه المقارنة إلى "لعنة الأبعاد" (curse of dimensionality)، التي من المرجح أن تؤدي إلى تدهور أداء الانحدار غير المعلمي مع زيادة عدد المتغيرات التفسيرية. ومع أن تقنية KPCA تُعد نظيراً غير خطي لتقنية PCA، فإن هذا لا يعني بالضرورة أنها ستقدم دائماً أداءً تنبؤياً أفضل من PCA، حتى في ظل وجود حدود غير خطية في دالة الانحدار. لذا، ندرس السيناريوهات التي تكون فيها PCA كافية، وتلك التي يُتوقع فيها أن تحقق KPCA تحسينات في الدقة. أُجريت الدراسة باستخدام محاكاة "مونت كارلو" (Monte Carlo) مع تغيير أحجام العينات (n = 20, 60, 150)، وعدد المتغيرات التفسيرية (p = 5, 10)، وعدد المكونات المُحتفظ بها (q = 2, 4) في بيئة البرمجة R. وقد تم النظر في ثلاث دوال انحدار ذات درجات متفاوتة من التعقيد في عدم الخطية، مع سحب المتغيرات التنبؤية والأخطاء العشوائية بشكل مستقل ومتماثل التوزيع (i.i.d.) من توزيعات طبيعية. وبعد عملية خفض الأبعاد، تم تقدير نموذج انحدار النواة "نادارايا– واتسون" (Nadaraya–Watson) باستخدام المكونات المستمدة من PCA وKPCA، مع استخدام نواة غاوسية (Gaussian kernel) في كلتا الخطوتين. وجرى تقييم الأداء

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

التنبؤي باستخدام مقياسي "متوسط خطأ التنبؤ" (APE) و"متوسط مربع خطأ التنبؤ" (MSPE) عبر 500 تكرار لمحاكاة مونت كارلو لكل إعداد من إعدادات المحاكاة. وتُظهر النتائج أن الأداء التنبؤي يتحسن مع زيادة عدد المكونات المُحتفظ بها (من q = 2 إلى q = 4)، وأن أحجام العينات الأكبر تميل إلى خفض كل من APE وMSPE. وعند تثبيت عدد المكونات المُحتفظ بها، تؤدي زيادة عدد المتغيرات التفسيرية عموماً إلى ارتفاع أخطاء التنبؤ. كما أظهرت المقارنة بين PCA وKPCA عدم وجود تفوق ثابت لأي منهما على الآخر؛ إذ يعتمد الأداء النسبي لكل منهما على حجم العينة، وأبعاد المتغيرات التنبؤية، وعدد المكونات المُحتفظ بها، وبنية عدم الخطية في دالة الانحدار.

الكلمات المفتاحية :اختزال الأبعاد، تحليل المكونات الرئيسية(PCA) ، تحليل المكونات الرئيسية بالنواة(KPCA) ، الانحدار اللامعلمي، لعنة الأبعاد، اللاخطية

1.Introduction

The "curse of dimensionality" poses a challenge for high-dimensional data analysis, particularly for methods such as nonparametric regression. Estimation and prediction can become less accurate as the number of predictors or explanatory variables grows. Thus, dimensionality reduction is one of the key techniques for overcoming these issues **(**Ghojogh et al., 2021; Eckstein et al., 2023**).** Even though PCA is a widely used linear dimensionality reduction method, its linearity may restrict its ability to explore complex nonlinear structures (Jolliffe, 2002; Marukatat, 2023). KPCA extends PCA through the use of the kernel trick and can be regarded as capturing nonlinear features in a transformed feature space (Schölkopf et al., 1998; Marukatat, 2023**)**.Hence, PCA/KPCA-based dimensionality reduction might influence the predictive performance of nonparametric regression (Eckstein et al., 2023**).** KPCA, however, does not necessarily outperform PCA. The crucial question then is whether nonlinear dimensionality reduction with more sophisticated techniques will yield additional predictive gains for various sample sizes, numbers of predictors, and degrees of nonlinearity. We address this question using a Monte Carlo simulation study, thereby allowing for systematic manipulation of

http://www.doi.org/10.62341/istj-vol39-1-AN53

these factors. We investigate multiple sample sizes, numbers of explanatory variables, numbers of retained components, and three regression functions of increasing nonlinearity. The resulting component scores obtained from dimensionality reduction are then used in the Nadaraya–Watson kernel regression estimator. We evaluate predictive performance using the Average Prediction Error (APE) and Mean Squared Prediction Error (MSPE) over 500 Monte Carlo replications. The goal is to identify when PCA and KPCA yield similar or different predictive performance in nonparametric regression.

1.1 Linearity and Nonlinearity of Data

Data linearity means that there is a structure in the data such that the relationships among variables can be explained by linear relationships. On the other hand, data nonlinearity refers to situations in which these relationships cannot be fully represented by linear relationships and may show curved or complex patterns in the original feature space. Here, the terms *linear data* and *nonlinear data* represent the nature of the predictor variables, but do not necessarily indicate whether the regression function is linear or nonlinear. This makes an important distinction because PCA and KPCA are followed by component scores, which are then used as predictors in the Nadaraya–Watson nonparametric regression estimator. While PCA provides a linear method for dimensionality reduction (Jolliffe, 2002)**,** KPCA provides a nonlinear approach that can capture more complex structures in the predictor space (Schölkopf et al., 1998; Girard & Saracco, 2014)**.**

1.2 Dimensionality Reduction

Dimensionality reduction refers to the process of mapping data from a high-dimensional space to a low-dimensional space, such that the information or statistics of the original data are well preserved in the reduced-dimensional space. In doing so, one attempts to attain a succinct representation of the data that, on the one hand, allows for simpler subsequent statistical analyses and, on the other hand, reduces the complexity effect induced by extreme dimensionalities (Ghojogh et al., 2021)**.**

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Denote the original data vector as

$$X = \left(X_1, X_2, \cdots, X_P\right)^T \in R_P$$

$$Z = \left(Z_1, Z_2, \cdots, Z_q\right)^T \in R_q, q < p$$

Then the general transformation:p

$$Z = f\left(x\right) \tag{1.1}$$

Then if W is a linear transformation, we have.: $z = W^T x$
Transformation matrix:- $W \epsilon R^{p \times q}$

$$x \epsilon R^p \qquad z \epsilon R^q \qquad q < p$$

We start with **p** original variables and transform them into **q** components, with the number of components being less than the number of original variables.
Hence, dimensionality reduction may be viewed as a process of transforming an original data vector in **p**-dimensional space into a new vector in **q**-dimensional space, where **q** is less than **p**, and the new vector retains the most important information/statistical feature of the original data (Ghojogh et al., 2021)**.** The mapping may be linear, e.g., **PCA** (Jolliffe, 2002)**,** or nonlinear, e.g., **KPCA** (Schölkopf et al., 1998)**.**

1.3 Principal Components Analysis (PCA)

It is a linear approach to dimensionality reduction that considers the transformation of the original set of correlated variables into a new set of uncorrelated variables, called the principal components. These components are new variables obtained as linear combinations of the original correlated variables, with the aim of reducing the number of variables and the dimensionality while retaining, as far as possible, the information or variance present in the data (Jolliffe, 2002).
The principal component PC_j is mathematically defined as follows

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

$$PC_{j=}a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{\mathrm{jp}}\mathrm{x_p} \qquad \textbf{1.2}$$

where**:**

$X_1, \cdots, X_{\mathrm{P}}$. Original variables

The principal components are ranked according to their variances. The first principal component has the largest variance and therefore accounts for the largest proportion of the total variance, followed by the second component which has the second largest variance and so on. The emerging components are mutually uncorrelated (Jolliffe, 2002**)**. PCA can be effectively used to treat multicollinearity and to reduce the dimensionality of data (Jolliffe, 2002; Hastie et al., 2009). Principal component coefficients j .

1.4 Kernel Principal Component Analysis (KPCA)

Kernel Principal Component Analysis (KPCA) represents a nonlinear technique designed to implement the methodology of traditional principal component analysis for data with nonlinear structures and correlations (Schölkopf et al., 1998; Marukatat, 2023). The essence of KPCA is the transformation of the initial dataset into a new, higher-dimensional feature space through the application of a nonlinear function and the performance of PCA. This procedure can be carried out without explicitly constructing the nonlinear transformation by using the so-called kernel trick, which employs a matrix of similarities between data points (Schölkopf et al., 1998; Marukatat, 2023)**.** KPCA proves to be very useful in cases where there are nonlinear patterns in the data that cannot be captured by traditional principal component analysis. Such a technique becomes particularly important in cases of dimensionality reduction where there are nonlinear relationships between different variables (Schölkopf et al., 1998; Marukatat, 2023; Wu et al., 2024)**.** In nonparametric regression models, KPCA can be applied as a preliminary stage for reducing the number of explanatory variables before building the regression model (Rosipal et al., 2001)**.**

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

KPCA is mathematically defined as follows:

$$x = (x_1, x_2, \cdots, x_n)^T \qquad x_i \epsilon R^p$$

Data are transformed into a feature space through $\emptyset: R^p \rightarrow f$

and using the kernel function

$$K(x_i, x_j) = \langle \emptyset(x_i), \emptyset(x_j) \rangle$$

The kernel matrix **K** is constructed and then centered according to the relation:

$$K_C = HKH \quad , H = I_n - \frac{1}{n} 11^T$$

Next, the eigenvalue problem is

$$K_C \propto_j = \Lambda_j \propto_j, \qquad \Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_n$$

The first **q** principal components are used to obtain the reduced data

$$Z_{KPCA} = (Z_1, Z_2, \cdots, Z_q), \quad q < p$$

Where Z_j represents the principal nonlinear components, and **q** is the number of retained components.

1.5 Nonparametric Regression Models

Nonparametric regression is a method for estimating the conditional expectation of the dependent variable in a regression model, without assuming a parametric form such as a linear or polynomial function. It can

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

approximate more complicated linear and nonlinear functions that may appear in the data. This is because these approaches rely on the data (as opposed to the model) to determine the general shape of the relationship. These approaches are data-driven (Wasserman, 2006; Härdle, 1990).

Typically, a nonparametric regression model can be written as:

$$y = g(x) + \varepsilon \tag{1.3}$$

$g(x)$ The unknown regression function which are estimated from the data, and is ε Random error limit.

Some of the most important methods for estimating the nonparametric regression function include Nadaraya–Watson kernel estimation, local linear regression, spline, and LOESS methods (Wasserman, 2006; Härdle, 1990).

1.6 Nadaraya–Watson Kernel Regression Estimator

The NW estimator is one of the most widely used nonparametric regression estimators based on kernel methods. It estimates the regression function $g(x) = E(Y/X = x)$ without specifying a parametric form for the relationship between the dependent variable and the independent variables. The kernel function is used to weight the observations in the sample such that observations closer to the point xx receive greater weights than those farther away (Girard & Saracco, 2014; Shahzad et al., 2023)

the Nadaraya–Watson estimator is given by the formula:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) y_i}{\sum_{j=1}^{n} \left(\frac{x - X_i}{h}\right)} \tag{1.4}$$

Kernel Function $\boldsymbol{K}(\cdot)$

$\boldsymbol{h}$ The smoothing parameter that controls the degree of smoothing

$\boldsymbol{x_i}$ The value of the explanatory variable for observation **i.**

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

y_i The value of the dependent variable for observation i
n Sample size

Main concept: The NW estimator allows us to calculate the value of $\hat{g}(x)$hat$g(x)$ at x as a weighted average of theY_i, where the weights are a function of the distance between x_i and x.(Girard & Saracco, 2014; Härdle, 1990)

Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right) \quad (1.5)$$

1.7 Bandwidth Selection by Cross-Validation

The bandwidth (h) is an essential smoothing parameter in the Nadaraya–Watson estimator. In this research, the bandwidth was determined independently for each iteration of the Monte Carlo simulation using a pre-defined candidate grid. It is noteworthy that the optimal bandwidth was selected based on the leave-one-out cross-validation (LOOCV) criterion using only the training data. The selected bandwidth was then used to estimate the regression function on the independent test dataset. Cross-validation provides a data-driven approach for selecting the smoothing parameter, which is crucial for balancing the bias–variance trade-off in nonparametric estimation. The use of an independent test dataset ensures that the evaluation of the model is not influenced by information from the test data (Girard & Saracco, 2014; Härdle, 1990; Wasserman, 2006).

bandwidth selection also ensures that the evaluation of the estimator is not affected by information from the test data.

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{g}_{-i,h}(x_i)]^2 \quad (1.6)$$

Were $\hat{g}_{-i,h}(x_i)$ the **Nadaraya–Watson Estimator**

At x_i after excluding observation i.

$\hat{h}=\arg\min_{h\in H} CV(h)$

where H represents the set of candidate values for the bandwidth.

2. The Simulation Study

We performed a Monte Carlo simulation to investigate the quality of PCA and KPCA for dimensionality reduction in nonparametric regression (based on the Nadaraya–Watson estimator). The simulation included the sample sizes (n = 20, 60, and 150), the number of explanatory variables (p = 5 and 10), and the dimensions retained (q = 2 and 4). The explanatory variables were drawn from a standard normal distribution, and the response was defined by three regression functions of increasing functional complexity: low, moderate, and high. The first of these functions has linear and nonlinear terms, while the second and third functions add more nonlinear terms and interactions. Separate training and testing samples were produced for each replication using the Gaussian kernel in the Nadaraya–Watson estimator. A separate bandwidth was selected for each transformed dataset according to the chosen data-based rule. We did 500 Monte Carlo replications for each experimental condition and evaluated the prediction results of PCA and KPCA using the mean absolute prediction error (APE) and the mean squared prediction error (MSPE). The experimental factors and levels are shown in Table 1.

Table 1. Experimental factors and levels used in the simulation study

Factor	Levels/ Values
Sample size (n)	**20, 60, 150**
Number of explanatory variables (p)	**5, 10**
Number of retained components (q)	**2, 4**

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Factor	Levels/ Values
Regression-function complexity	**Low** $y = 2X_1 + X_2 + 0.25\sin(X_3) + \varepsilon$
	Moderate $y = 2X_1 + X_2 + 0.50\sin(X_3) + 0.25X_4^{\,2} + \varepsilon$
	High $y = 2X_1 + X_2 + \sin(X_3) + 0.50X_4^{\,2} + 0.50X_4X_5 + \varepsilon$
Dimensionality-reduction methods	PCA, KPCA
Nonparametric estimator	Nadaraya–Watson **(NW)**
Kernel function	Gaussian kernel
Monte Carlo (MC)	500 per cell
Performance measures	APE, MSPE

2.1 Numerical Summary of the Simulation Results

The numerical results of the simulation study are summarized in Tables 2–4 for various sample sizes, numbers of predictor variables, degrees of nonlinear complexity, and numbers of retained components for the PCA- and KPCA-based procedures. The predictive performance of the nonparametric regression models using PCA and KPCA was evaluated using APE and MSPE. Lower values of APE and MSPE indicate better predictive performance. The three regression functions were designed to represent increasing levels of nonlinearity. Function 1 represents a relatively low level of nonlinearity through a sine component. In Function 2, a higher level of nonlinear complexity is introduced through a quadratic component. By adding a sine term, a quadratic term, and an interaction term with a standard normal random variable, Function 3 represents the most complex nonlinear setting. This design allows the investigation of the performance of PCA- and KPCA-based procedures in progressively more complex nonlinear regression structure.

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Table 2. Simulation Results for PCA-NW and KPCA-NW Based on APE and MSPE for the First Regression Function

			PCA- NW		KPCA - NW	
	n	q	P=5	P=10	P=5	P=10
APE	20	2	1.724134	1.889653	1.721093	1.876401
		4	1.504201	1.810818	1.498840	1.799078
	60	2	1.669203	1.825926	1.669174	1.823429
		4	1.367513	1.743424	1.366115	1.731987
	150	2	1.643569	1.831344	1.637560	1.825482
		4	1.283915	1.708263	1.272465	1.700849
MSPE	20	2	4.800962	5.604425	4.784029	5.533999
		4	3.675905	5.188543	3.659194	5.110255
	60	2	4.472533	5.236786	4.482732	5.222691
		4	3.050095	4.804736	3.045615	4.743302
	150	2	4.326632	5.307970	4.312532	5.276963
		4	2.704650	4.630134	2.674425	4.587068

Table.3. Simulation Results for PCA-NW and KPCA-NW Based on APE and MSPE for the Second Regression Function

			PCA- NW		KPCA- NW	
	n	q	P=5	P=10	P=5	P=10
APE	20	2	1.754907	1.970355	1.777289	1.933477
		4	1.551641	1.886962	1.539739	1.882509
	60	2	1.685346	1.894389	1.676968	1.871948
		4	1.383692	1.776812	1.385999	1.754758
	150	2	1.671854	1.859146	1.657039	1.860023
		4	1.290333	1.731564	1.287295	1.721585
MSPE	20	2	4.979955	6.074015	5.041138	5.906055
		4	3.927310	5.672134	3.891466	5.587880
	60	2	4.579079	5.653865	4.548073	5.534018
		4	3.149086	5.012116	3.186189	4.900296
	150	2	4.471107	5.449218	4.423745	5.475484
		4	2.738905	4.774627	2.752792	4.705272

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Table.4. Simulation Results for PCA-NW and KPCA-NW Based on APE and MSPE for the Third Regression Function

			PCA- NW		KPCA- NW	
	n	**q**	**P=5**	**P=10**	**P=5**	**P=10**
APE	**20**	**2**	1.908638	2.089225	1.906783	2.086253
		4	1.651329	2.000517	1.654512	1.988914
	60	**2**	1.834033	2.026954	1.831731	2.026521
		4	1.477152	1.908567	1.481301	1.905427
	150	**2**	1.785734	2.003819	1.798056	1.999396
		4	1.383573	1.870827	1.387414	1.869878
MSPE	**20**	**2**	5.935967	6.954988	5.922424	6.938250
		4	4.538794	6.413349	4.544010	6.340027
	60	**2**	5.479347	6.549449	5.474549	6.539569
		4	3.648038	5.851980	3.691311	5.825927
	150	**2**	5.210328	6.409777	5.285423	6.382664
		4	3.213091	5.632822	3.271465	5.623250

Summary and Discussion of the Simulation Results

Results for the Three Regression Functions (Tables 2 to 4)

KPCA-NW also resulted in slightly smaller prediction errors than PCA-NW in most cases across the three regression functions, although the numerical differences were generally small and varied across the regression functions and dimensional settings. For the first regression function (Table 1.2), KPCA-NW produced slightly lower numerical errors in almost every case. For example, at n = 20, q = 2, and p=5, the APE was 1.721093 for KPCA-NW compared with **1.724134** for PCA-NW, while the MSPE was 4.784029 compared with 4.800962, respectively. The only exception was the MSPE at n = 60, q = 2, and p = 5, where PCA-NW yielded a slightly lower value **(4.472533)** than KPCA-NW **(4.482732).**For the second regression function (Table 1.3), KPCA-NW generally produced slightly lower numerical errors at p = 10, although an exception was observed at n = 150 and q=2, where PCA-NW yielded a slightly lower MSPE **(5.449218)** than KPCA-NW (5.475484). At p = 5, the results were mixed, with the two methods alternating in their

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

numerical performance across different sample sizes and numbers of retained components. For the third regression function (Table 1.4), the numerical differences between the two procedures were generally small. At p=10, KPCA-NW produced slightly lower errors across the considered settings, although the differences were minor. At p=5, PCA-NW was slightly better in several cases, particularly for q=4 and n=150. For example, at n=150 and q=4, the MSPE was 3.213091 for PCA-NW compared with 3.271465 for KPCA-NW.

Across all three regression functions, the prediction errors generally decreased as the sample size **n** and the number of retained components **q** increased, while they increased as the original predictor dimension increased from p = 5 to p=10. The reduction in error when increasing **q** from 2 to 4 suggests that retaining additional components preserved information that was relevant to prediction under the simulation settings. The higher errors observed for p=10 are consistent with the greater difficulty of nonparametric estimation in higher-dimensional predictor spaces, even after dimensionality reduction. The observed numerical differences between KPCA-NW and PCA-NW at p=10 may be related to the ability of KPCA to represent nonlinear structure in the predictors. However, because these differences were generally small and varied across the regression functions and simulation settings, they should be interpreted cautiously and should not be regarded as evidence of a universally superior method without additional formal statistical comparisons.

General Discussion

Sample size (n) effect. The APE and MSPE decrease with increasing n as expected because the kernel-based estimation is more accurate based on more observations. For instance, in Table 1.2 the MSPE of KPCA-NW for q = 4 and p = 5 decreases from 3.659194 (n = 20) to 2.674425 (n = 150). The effect is strongest for q = 4, for q = 2 and p = 10 the errors are almost not affected by n and are even not monotone in a few cases (e.g. PCA-NW in Table 1.2).

Number of components (q). The prediction errors decrease substantially as q increases from 2 to 4 in each table because the larger number of components preserves more information among the explanatory variables. The decrease is more pronounced at p = 5 than at p = 10. For instance, in Table 1.2 (n = 150), the MSPE of PCA-NW decreases from 4.326632 to 2.704650 for p = 5, but only from 5.307970 to 4.630134 for p = 10. For a fixed number of components, a decreasing fraction of the total variability is explained as the dimension gets higher.(p) Dimensionality (p). Errors are larger at p=10 than at p=5 for all cases and both methods. This is consistent with the curse of dimensionality diminishing the utility of kernel smoothers as the number of predictors grows.

Conclusion

Simulation results show that the two methods have comparable prediction performance, with **KPCA-NW** outperforming in most cases (particularly when p = 10. **PCA-NW** also gives competitive and sometimes slightly better results for p = 5. For each method, accuracy improves with sample size and number of selected components, but decreases as the number of explanatory variables increases. KPCA-NW is therefore appropriate when the predictors have a nonlinear structure or the dimension is high, and PCA-NW is sufficient for simple (low-dimensional) data.

References

Eckstein, S., Iske, A., & Trabs, M. (2023). Dimensionality reduction and Wasserstein stability for kernel regression. *Journal of Machine Learning Research, 24*(334), 1–35.

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). *Sufficient dimension reduction for high-dimensional regression and low-dimensional embedding: Tutorial and survey*. arXiv. https://arxiv.org/abs/2110.09620

Girard, S., & Saracco, J. (2014). An introduction to dimension reduction in nonparametric kernel regression. *ESAIM: Proceedings and Surveys, 66*, 167–196. https://doi.org/10.1051/eas/1466012

Abstract Linear and Nonlinear Dimensionality Reduction and Their Impact on Nonparametric Regression: A Monte Carlo Comparison of PCA and KPCA
http://www.doi.org/10.62341/istj-vol39-1-AN53

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Marukatat, S. (2023). Tutorial on PCA and approximate PCA and approximate kernel PCA. *Artificial Intelligence Review, 56*(6), 5445–5477.

Rosipal, R., Girolami, M., Trejo, L. J., & Cichocki, A. (2001). Kernel PCA for feature extraction and de-noising in nonlinear regression. *Neural Computing & Applications, 10*(3), 231–243.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Shahzad, U., Ahmad, I., Almanjahie, I. M., Al-Noor, N. H., & Hanif, M. (2023). Adaptive Nadaraya-Watson kernel regression estimators utilizing some non-traditional and robust measures: A numerical application of British food data. *Hacettepe Journal of Mathematics and Statistics, 52*(5), 1425–1437.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer.

Wu, Y., Chen, A., Xu, Y., Pan, G., & Zhu, W. (2024). Modeling mortality with kernel principal component analysis (KPCA) method. *Annals of Actuarial Science, 18*(3),626–643. https://doi.org/10.1017/S1748499524000277