SCHOOL OF STATISCS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Nonparametric Estimation of a Switching Regression Model

by

Ruffy S. Guilatco
University of the Philippines, Diliman

Erniel B. Barrios
University of the Philippines, Diliman

UPSS Working Paper No. 2014-04
May 2014

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

Nonparametric Estimation of a Switching Regression Model

Ruffy S. Guilatco
School of Statistics, University of the Philippines Diliman, Philippines
rsguilatco@gmail.com

Erniel B. Barrios
School of Statistics, University of the Philippines Diliman, Philippines
ebbarrios@upd.edu.ph

High dimensional data often exhibits multicollinearity leading to unstable estimates of regression coefficients. We postulate a switching regression model with high dimensional predictors. Principal components were extracted to mitigate the multicollinearity caused by high dimensional predictors. The principal components are specified in a nonparametric frameworkinto the switching regression model to mitigate the decline in predictive ability of the model due to lost information in using principal components instead of the original predictors. Simulation studies indicated that nonparametric principal component switching regression model yields better predictive ability than the parametric counterpart while mitigating the adverse effects of multicollinearity. The predictive ability of the model is also robust to the nature of switch (endogenous or exogenous) between the two regimes.

Key Words:high dimensional data, multicollinearity, principal components regression

1. Introduction

The work of impact evaluation has recently focused on measuring the impact of program interventions to make sure that inputs are carefully chosen to be aligned with the needs of the stakeholders. Impact evaluation provides policy-makers with necessary feedback for policy adjustment, that is, modification or cancellation of ineffective program interventions. Project beneficiaries are not randomly chosen, but are rather based on their inherent characteristics to qualify them to receive the interventions. A poverty alleviation program for example should necessarily be delivered among the poor members of the community. Thus, in measuring impact, it is necessary to choose samples that are not randomly selected, but rather a self-selection by the course of the circumstances they are in (e.g., the poor), resulting to sample selection bias. The impact of the program could have been different had the beneficiaries been in a different condition, i.e., a different nature of endowments are available to them.These so-called selection bias in impact evaluation are ideally addressed through a randomized social experiment. However, there are logistical and financial implications (to some extent, moral repercussions)in social experiments. For example, it is not right to deprive a poor community with poverty alleviation program because the community is not part of the "treatment" group. Alternatively, propensity score matching analysis conducts propensity score comparisons of the outcome of groups of subjects with similar inherent characteristics. However, this would still not correct for hidden bias since the outcomes of non-beneficiaries

may differ systematically from what the outcomes of beneficiaries would have been without the program, producing selection bias in estimated treatment effect(Rosenbaum and Rubin, 1983). Sample selection models would offer another alternative approach, for example,a switching regression model, evaluates social programs using the outcomes of non-beneficiaries to estimate the expected outcome of the beneficiaries had they not participated in the program.

In switching regression model, observations are classified into different regimes (usually two), a two-part model is then fitted (selection function and the regression equation in the regimes). The selection function involves a vector of characteristics that affect the individual's decision/qualification to participate in an intervention. The regression equation of the regimes contains a vector of individual characteristics that is thought to influence the outcome of interest, say, the measure of impact of the program intervention. Ordinary least squares (OLS) is a simple method used in estimating switching regression models, but (Maddala, 1983), pointed out that this is inappropriate because of consistency problems. Maximum likelihood estimation (MLE) is popularly used in switching regression model, estimates consistent and efficient, but the estimation can be cumbersome when the likelihood function is complicated because of the nature of the data.

Data obtained from programs usually involve measurement of various factors that are attributable to the success of the intervention, often resulting to a high dimensional setting. OLS and MLE do not consider the possible problems that can be encountered when modeling in high dimension. One immediate problem in the analysis using high dimensional data is multicollinearity, which occur when two or more independent (predictor) variables in a regression model are highly correlated. The presence of multicollinearity in the model inflates the variance of the estimated coefficients, i.e., regression coefficients are unstable (Curto and Pinto, 2007). In general, impact assessment data is vulnerableover the curse of dimensionality problem.

A common practice in analyzing high dimensional data is to first reduce the dimension before implementing the usual statistical methods. Principal Components Regression is a tool in modeling and dimension reduction (sequential), where a selection from the full set of principal components of the independent variables is used as the independent variables in regression analysis. Since there is loss of information by using only a subset of the principal components, classical methods of estimating the regression equation can result to a model with low predictive ability (Dunteman, 1989).

To estimate the switching regression model in a high dimensional setting, we propose to use principal components regression in each regime, and to mitigate the lost information due to selection of a subset of principal components, we relax the structural form of the models in each regime by postulating nonparametric functions. The goal of the nonparametric approach in regression is to estimate the function, rather than estimating the parameters (Fox, 2000), hoping that the lost information from the selection of principal components can be recovered by allowing the model to be as flexible as possible.In this approach, better fit of the regression model can be expected in each of the regimes, without the restrictions from the assumptions on the structure of the data.

2. Selection Bias and Switching Regression Models

Sample selection bias is caused by choosing non-random data for statistical analysis. The bias exists due to a flaw in the sample selection process, where a subset of the data is systematically excluded due to a particular attribute.Exclusion of the subset can influence the statistical significance of the test, or produce sampling distribution that is a distorted version of the population distribution.

Heckman (1979) considered the bias resulting from using non-randomly selected samples to estimate behavioural relationships as an ordinary specification bias that arises because of a missing data problem. An estimation procedure that addresses the problem with specification error was proposed.

Other solutions for sample selection bias in the estimation includes the propensity score matching (PSM), where propensity scoreisdefined as the probability of a sample being assigned to a particular group given a set of observed characteristics. Dehejia and Wahba (2002) showed that PSM methods for nonexperimental causal studies were able to alleviate bias due to systematic differences between the treated and comparison groups.PSM compare the outcome of groups of subjects with similar inherent characteristics, therefore, cannot correct for hidden bias (Rosenbaum and Rubin, 1983), alternatively, unobservable selection bias is addressed in sample selection models.

The basic idea of a sample selection model is that the outcome variable y, is observed only if some criteriondefined by a variable z is met. Switching regression model (SRM) is an example of selection model, usually represented by models in two stages: (1) the selection equation and; (2) the regression equations in each of the regimes. (Maddala, 1983) described a SRM as follows:Suppose the observations on the dependent variable y can be classified into two regimes, say y_1 and y_2, where $y_1 = \beta_1 X_1 + u_1$ iff $\gamma Z + u > 0$ and $y_2 = \beta_2 X_2 + u_2$ iff $\gamma Z + u \leq 0$. X_1, X_2 and Z are sets of explanatory variables; β_1, β_2 and γ

are sets of parameters; u_1, u_2 and u are error terms that are assumed to be normally distributed with mean vector **0** andcovariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1u} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2u} \\ \sigma_{1u} & \sigma_{2u} & 1 \end{bmatrix}$. If$\sigma_{1u} = \sigma_{2u} =$ 0, then the model is said to have exogenous switching, i.e.,the nature of switch is exogenous when the separation of observations on a dependent variable y is determined by a selection equation with predictors outside of the structure of the regression model. On the other hand, if $\sigma_{1u} \neq 0$ or $\sigma_{2u} \neq 0$, then the model is said to have endogenous switching and the observations are classified into different regimes using characteristics that are part of the model structure. The set of predictors in the selection equation overlap with the set of independent variables in the regression equation of the switching regimes, (Maddala, 1983). Furthermore, there are two types of SRM, model with known or unknown sample separation.

Given estimated parameters of the switching regression model, the following expectations can be calculated, giving important insights about the the model:

$$Xb_1 = E[y_1|X_1] = X_1\beta_1$$
$$Xb_2 = E[y_2|X_2] = X_2\beta_2$$
$$yc_{1\_1} = E[y_1|I = 1, X_1] = X_1\beta_1 + \sigma_1\rho_1 f(\gamma Z)/F(\gamma Z)$$
$$yc_{2\_1} = E[y_2|I = 1, X_2] = X_2\beta_2 + \sigma_2\rho_2 f(\gamma Z)/F(\gamma Z)$$
$$yc_{1\_2} = E[y_1|I = 0, X_1] = X_1\beta_1 - \sigma_1\rho_1 f(\gamma Z)/\{1 - F(\gamma Z)\}$$
$$yc_{2\_2} = E[y_2|I = 0, X_2] = X_2\beta_2 - \sigma_2\rho_2 f(\gamma Z)/\{1 - F(\gamma Z)\}$$

where

$Xb_1$is the unconditional linear prediction for the regression equation in regime 1

$Xb_2$is the unconditional linear prediction for the regression equation in regime 2

$yc_{1_1}$is the expected value of the dependent variable in regime 1 conditional on the dependent variable being observed

$yc_{2_1}$is the expected value of the dependent variable in regime 1 conditional on the dependent variable not being observed

$yc_{2_2}$is the expected value of the dependent variable in regime 2 conditional on the dependent variable being observed

$yc_{1_2}$is the expected value of the dependent variable in regime 2 conditional on the dependent variable not being observed

There werenumerous applications of the switching regression, example, Throst (1977) examined the expenditures on housing services in owner-occupied and rental housing. The housing-demand model determined the individual decision to own or rent a house and the amount spent on housing services. Lee (1978) constructed the union-nonunion model, which investigated the joint determination of the extent of unionism and the effects of unions on

wage rates. The tendency to join a union depends on the net wage gains. The model showed the interdependence between the wage-gain equation and the union-membership equation.Cai et. al. (2008) used an endogenous switching regression model to examine how farmers' characteristics affect their decisions to join the contract farming and their performance with or without the contract. They also compared farmers' expected performance under the contract and without the contract.

3. Methodology

Consider the data usually generated from the conduct of impact assessment studies. This will contain two important features: high dimensional predictors and sample selection bias. High dimensionality of predictors arises from the complex dynamics in which inputs are translated towards the expected outputs or outcomes of the intervention or program. Selection bias is accrued due to the inherent characteristics of the beneficiaries that entitles them to benefit/not benefit from the intervention.

Principal components analysis is first applied on the high dimensional predictors, keeping only those few components accounting for most of the total variance contained in the predictors. Selection only of few principal components would result to information loss in a linear model fitted on the principal components. Following Umali and Barrios (2014), the switching regression models are postulated as nonparametric functions of the principal components. The lost information from selection of only a few principal components can be recovered by postulating a flexible nonparametric structure.

In lieu of the original variables, the principal components are used in specifying the nonparametric switching regression as follows:

Under regime 1,

$$y_{1i} = f_{10} + f_{11}(c_{1i}) + f_{12}(c_{2i}) + \cdots + f_{1j}(c_{ji}) + \varepsilon_{1i} \quad (1)$$

Under regime 2:

$$y_{2i} = f_{20} + f_{21}(c_{1i}) + f_{22}(c_{2i}) + \cdots + f_{2k}(c_{ki}) + \varepsilon_{2i} \quad (2)$$

Furthermore, the switch is defined as:

$$I_i^* = \gamma Z_i + u_i \quad (3)$$

$$\text{for } i = 1,2,\ldots,n \quad \text{and } j,k < p$$

where

I_i^* is a latent variable that determines the regime of unit *i*

$Z_i$is a vector characteristics that influences the choice of regime

$c_{1i}, c_{2i}, \ldots, c_{ji}$ are elements of vectors of j selected principal components

y_{1i}andy_{2i} are dependent variables observed under regimes 1 and 2

$f_1 = \{f_{10}, f_{11}, .., f_{1j}\}$ and$f_2 = \{f_{20}, f_{21}, .., f_{2j}\}$are series of smooth functions, and

$$(\varepsilon_1, \varepsilon_2, u)' \sim N(\mathbf{0}, \mathbf{\Sigma}) \text{with} \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1\varepsilon} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2\varepsilon} \\ \sigma_{1\varepsilon} & \sigma_{2\varepsilon} & 1 \end{bmatrix}$$

The observed dichotomous realization I_i of the latent variable I_i^* of whether the individual *i*belongs to a particular regime has the following form:

$$I_i = 1 \quad \text{if} I_i^* > 0$$
$$I_i = 0 \quad \text{otherwise}$$

Only the principal components that contribute to about 70% of the total variation will be included in the regression models. Hence, there is an expected loss of information. Modeling using the parametric approach will lead to deterioration of the predictive ability of the model. Thus, a more flexible form of modeling will be used, that is, the nonparametric regression on the retained principal components in each of the regimes.

Suppose that $\boldsymbol{X}$can be represented by the principal components $\boldsymbol{C}$ of the original variables, then we fit the regression models in Equations (1) and (2). The use of principal components is beneficial not only in dimension reduction, but also in ensuring uncorrelated predictors in the model. Investing on the orthogonality of the components in$\boldsymbol{C}$, we can use the backfitting algorithm (Hastie and Tibshirani, 1990), since Equations (1) and (2) suffices the assumptions of an additive model. Each term in the model accounts for the individual, non-overlapping contribution of the components (c_{1i}, c_{2i}, …,c_{ji}).

Nonparametric counterfactual analysis

Counterfactual analysis facilitates the evaluation to what would have happened to beneficiaries in the absence of the intervention, and impact is estimated by comparing counterfactual outcomes of those observed under the intervention.

Using the fitted regression equations in (1) and (2), the predicted values to be used in the counterfactual analysis are derived using the following:

xb_{1i}^* = the unconditional prediction for the regressionequation in regime 1

xb_{2i}^*= the unconditional prediction for the regression equation in regime 2

$yc_{1\_1i}^*$ = the predicted value of the dependent variable in regime 1 conditional on the dependent variable being in regime 1

$yc_{2\_1i}^*$= the predicted value of the dependent variable in regime 2 conditional on the dependent variable being in regime 1

$yc_{2\_2i}^*$= the predicted value of the dependent variable in regime 2 conditional on the

dependent variable being in regime 2

$yc_{1\_2i}^{*}$ = the predicted value of the dependent variable in regime 1 conditional on the dependent variable being in regime 2

Based on above estimates, indicators can be constructed to compare the measure of the outcome of interest between the two regimes.

i. $\pi = xb_{1i}^{*} - xb_{2i}^{*}$

Here, π is equal to the sample unit$i$'s (irrespective of his/her choice of regime) expected measure of the outcome of interest under regime 1 minus his/her expected measure under regime 2. The mean of π is the individuals benefit from choosing to be in regime 1.

ii. $\pi_1 = yc_{1\_1i}^{*} - yc_{2\_1i}^{*}$

π_1 is the predicted value of the dependent variable for sample individuals (that belong to regime 1) under regime 1 minus his/her predicted value under regime 2. The mean of π_1 is the benefit of sample individuals that belong to regime 1 from choosing to be in regime 1.

iii. $\pi_2 = yc_{1\_2i}^{*} - yc_{2\_2i}^{*}$

π_2 is the predicted value of the dependent variable for sample individuals (that belong to regime 2) under regime 1 minus his/her expected value under regime 2. The mean of π_2 is the benefit of sample individuals that belong to regime 2 if they choose to be in regime 1.

4. Simulation Studies

A simulation study is conducted using the combination of five factors (see Table 1) which include the form of switch, the dimensionality of data, functional form of the data-generating model, level of multicollinearity, and the fit of the data-generating model.

In Table 1, the predictive ability or fit of the data-generating model is only considered when the dependent variable is a linear combination of the independent variables in low dimensional data. A model with high R^2 is expected in high dimensional data since R^2 is greatly affected by the number of predictors in the model. In addition, R^2 will be not be considered when the dependent variable is a nonlinear combination of the independent variables (R^2 is a measure of linear relationship Y on the Xs).

Table 1. Simulation settings

Nature of switch	Data Dimensionality	Functional form of thedata-generating model	Level of multicollinearity	Fit of the data-generating model
endogenous	high dimensional	linear	strong	
			weak	
			none	

		nonlinear	strong	
			weak	
			none	
	low dimensional	linear	strong	good
				poor
			weak	good
				poor
			none	good
				poor
		nonlinear	strong	
			weak	
			none	
exogenous	high dimensional	linear	strong	
			weak	
			none	
		nonlinear	strong	
			weak	
			none	
	low dimensional	linear	strong	good
				poor
			weak	good
				poor
			none	good
				poor
		nonlinear	strong	
			weak	
			none	

The simulated data

We consider endogenous switch if there is a correlation between the error terms of the selection equation and regression equation in the regimes. Hence, the predictor in the selection equation is related to the independent variables in the regression equation in the regimes. Conversely, exogenous switch is simulated by considering independent error terms, implying that the classification of the observations into the two regimes is outside of the model structure. Data is said to have high dimension when $n \leq p$, where n is the number of observation and p is the number of independent variables. The data is high dimension only in the independent variables (Xs) in the regression equation of the switching regime and not

in the predictors of the selection equation (Zs). Datasets with multicollinearity are generated by taking X_{j+1} as a function of X_j and a constant k that determines the level of correlation between independent variables. We used the exponential function for the nonlinear functional form of the data-generating model. The fit of the data-generating model is adjusted by multiplying a constant m in the error term. Higher values of m result to poor fit (may also be interpreted as misspecification error) of the data-generating model.

5. Results and Discussion

We present results of the simulation study and the assessment of the performance of the proposed nonparametric estimation of the switching regression model as compared to the parametric counterpart over various data simulation scenarios. Five factors were considered in the simulation study, including the nature of switch in the model specification of the switching regression, dimensionality of data, functional form of the data-generating model, level of multicollinearity, and the fit of the data-generating model (presence of misspecification error). A method is superior in the simulated dataset if the computed mean absolute percentage error (MAPE) is smaller. In addition, the proportion of datasets where the proposed method is superior is also presented.

Effect of the nature of switch

Models with endogenous switching were generated by setting the correlation between the error terms of the selection equation and regression equation in the regimes nonzero. Also, the independent variables in the regression equations were included in the specification of the selection equation. On the other hand, models with exogenous switching were generated by setting the correlation between the error terms of the selection equation and regression equation in the regimes to zero.

The summary statistics of the MAPE generated by the parametric and nonparametric PCR on the two regimes considering the effect of the form of switch are presented in Table 2. Generally, summary statistics in datasets with endogenous switching are similar to the datasets with exogenous switching. This implies that the performance of the parametric and nonparametric PCR is not significantly affected whether the data is grouped using a characteristic outside or within the structure of the model. In both nature of switch, the nonparametric PCR is superior to the parametric PCR. Almost all simulated datasets in regime 1 have smaller MAPE in using the nonparametric PCR. On the other hand, there is only a small number of cases where the parametric PCR in regime 2 is superior.

Table 2. Comparison of the MAPE generated from the parametric and nonparametric PCR (by nature of switch)

Statistics	Nature of switch	Regime 1		Regime 2	
		PAR	NPAR	PAR	NPAR
Average	endogenous	31.15	18.48	15.95	12.24
	exogenous	26.33	15.68	17.24	11.14
Minimum	endogenous	0.92	0.26	1.42	0.55
	exogenous	1.04	0.57	1.63	0.67
Maximum	endogenous	1301.67	799.80	1466.98	564.51
	exogenous	1302.62	661.62	3775.31	650.90
Median	endogenous	12.38	11.18	13.04	10.63
	exogenous	10.24	8.90	12.23	10.17
Proportion with smaller MAPE	endogenous	1.87	98.13	13.80	86.20
	exogenous	1.47	98.53	10.20	89.80

Effect of the data dimensionality

High dimensional data is defined in this study as data with at least 30 independent variables and low dimensional if the number of predictors is at most five. From Table 3, the maximum value of the MAPE can be very high in datasets with high dimension. This is the effect of dimension-reduction in PCR, since some information in the original raw data were lost. However, it is very clear that using the nonparametric approach of estimating the regression equations on the principal components, the fit of the model is improved, producing smaller values for MAPE compared to the parametric estimation. In low dimensional datasets, the performance of the parametric PCR is somehow comparable to that of the nonparametric PCR. But still, all statistics are better in favor of the nonparametric PCR. Generally, the nonparametric PCR performs better in high dimensional data, suggesting that advantages in using the nonparametric PCR in low dimensional datasets might be marginal and the use of ordinary least squares can be more appropriate.

Table 3. Comparison of the MAPE generated from the parametric and nonparametric PCR (by dimensionality of the data)

Statistics	Dimensionality of data	Regime 1		Regime 2	
		PAR	NPAR	PAR	NPAR
Average	high dimensional	52.01	24.93	19.51	8.74
	low dimensional	13.22	11.84	14.65	13.66
Minimum	high dimensional	0.92	0.26	1.42	0.55
	low dimensional	1.04	0.97	2.41	2.25

Maximum	high dimensional	1302.62	799.80	3775.31	650.90
	low dimensional	37.83	33.76	42.96	33.44
Median	high dimensional	8.52	6.22	8.36	5.97
	low dimensional	12.10	10.90	13.80	12.91
Proportion with smaller MAPE	high dimensional	0.50	99.50	3.00	97.00
	low dimensional	2.44	97.56	18.00	82.00

Effect of the functional form of the data-generating model

There are two forms considered in this study: (1) y is linearly associated and generated by the Xs, and; (2) y is generated by exponential functions of the Xs. The nonparametric PCR is better than the parametric PCR in both linear and nonlinear form of the data-generating model based on the summary statistics of MAPE presented in Table 4. Although, the MAPE values from the two methods are comparable, majority of the datasets exhibited that nonparametric PCR is superior. The parametric and nonparametric PCR performed better when the data is generated by a nonlinear model.

Table 4. Comparison of the MAPE generated from the parametric and nonparametric PCR (by functional form of the data-generating model)

Statistics	**Functional form of the data-generating model**	**Regime 1**		**Regime 2**	
		PAR	**NPAR**	**PAR**	**NPAR**
Average	linear	42.04	23.73	21.03	13.90
	nonlinear	8.79	7.10	9.94	8.38
Minimum	linear	1.04	0.97	2.41	2.25
	nonlinear	0.92	0.26	1.42	0.55
Maximum	linear	1302.62	799.80	3775.31	650.90
	nonlinear	26.15	23.41	25.88	22.79
Median	linear	13.49	12.10	14.76	11.61
	nonlinear	8.16	6.00	9.80	7.46
Proportion with smaller MAPE	linear	2.50	97.50	15.17	84.83
	nonlinear	0.42	99.58	7.25	92.75

Effect of the level of multicollinearity

Three levels of multicollinearity were considered in the simulation study. For datasets with strong multicollinearity, the independent variables are generated as a function of other

independent variables with high degree of association in the data-generating model. Similar approach of simulation is considered for data with weak multicollinearity, except that low degree of association is assumed instead. For datasets with no multicollinearity, data were generated by uncorrelated independent variables.

From Table 5, the nonparametric PCR is superior in almost all simulated datasets with different levels of multicollinearity. In addition, the nonparametric PCR has better summary statistics of the MAPE generated. Both methods are more effective when applied to datasets with strong multicollinearity. This result can be due to the fact that by summarizing the information from the variables using PCA, in the process, the initial PCs extracted account for large variation in the data, resulting to new set of independent variables that provide a good summary of the raw data.

Table 5. Comparison of the MAPE generated from the parametric and nonparametric PCR (by level of multicollinearity)

Statistics	**Level of multicollinearity**	**Regime 1**		**Regime 2**	
		PAR	**NPAR**	**PAR**	**NPAR**
Average	strong	10.51	9.67	11.24	10.57
	weak	36.01	19.68	20.16	11.79
	none	39.69	21.88	18.38	12.72
Minimum	strong	1.04	0.97	2.00	1.52
	weak	0.92	0.29	1.42	0.55
	none	0.98	0.26	1.45	0.58
Maximum	strong	34.42	33.76	34.39	33.44
	weak	1301.67	381.28	3775.31	564.51
	none	1302.62	799.80	1080.52	650.90
Median	strong	9.35	8.15	9.65	9.05
	weak	12.06	10.43	13.09	10.22
	none	12.78	11.01	14.60	11.33
Proportion with smaller MAPE	strong	3.90	96.10	17.20	82.80
	weak	0.20	99.80	10.00	90.00
	none	0.90	99.10	8.80	91.20

Effect of the fit of the data-generating model (misspecification error)

The effect of the fit of the data-generating model (presence or absence of misspecification error) is assessed only for settings with low dimensional data and when the dependent variable is a linear combination of the independent variables. High dimensional data tend to

produce models with high R^2 since the coefficient of determination is greatly affected by the number of independent variables in the model. Furthermore, R^2 is a measure of linear relationship of Y on the Xs and is not applicable in cases when Y has a nonlinear association with the Xs. The effect of misspecification error is simulated by multiplying the error terms with a constant that will consequently affect the fit of the data-generating model.

Table 6 shows that for datasets generated from a model with good fit, the MAPE using the two methods are almost similar. But, the nonparametric PCR is always superior over the parametric PCR. Furthermore, both procedures resulted to smaller MAPE when implemented in datasets generated from a model with good fit, since all statistics of MAPE are smaller as compared to the MAPE produced by the procedures in datasets from a model with poor fit. However, there are more than 20 per cent of datasets generated by a model with good fit where the parametric PCR is superior over the nonparametric PCR. On the other hand, nonparametric PCR clearly outperformed the parametric counterpart in cases where the data are generated from a model with poor fit. This result corroborates one of the advantages of nonparametric regression, that is, the procedure allows the data to search for the appropriate functional form of the relationships between variables, without relying on assumed form of the structural relationship.

Table 6. Comparison of the MAPE generated from the parametric and nonparametric PCR (by fit of the data-generating model)

Statistics	**Fit of the data-generating model**	**Regime 1**		**Regime 2**	
		PAR	**NPAR**	**PAR**	**NPAR**
Average	good	10.65	9.65	11.09	10.67
	poor	17.13	15.29	19.40	17.96
Minimum	good	1.04	0.97	2.41	2.25
	poor	4.59	3.25	8.97	7.73
Maximum	good	24.09	20.18	25.46	22.98
	poor	37.83	33.76	42.96	33.44
Median	good	10.43	9.65	10.92	10.69
	poor	16.33	14.35	18.77	17.40
Proportion with smaller MAPE	good	6.17	93.83	37.83	62.17
	poor	0.67	99.33	7.67	92.33

Improvement in the prediction of the superior method

In general, larger improvements in using the nonparametric over the parametric PCR are seen

when the data is generated from a model with weak to no multicollinearity as seen in Table 7 and 8. There are some instances when the parametric PCR is superior; however, the improvement in predictive ability is still larger in cases when the nonparametric PCR is superior.

In high dimensional datasets, the improvement in the fit of the model generated is very evident and can go as high as over 100%. On the other hand, in low dimensional datasets, the improvement in the predictive ability of the generated model is lower with differencesaround20%.

Table 7. Percent difference in MAPE between the methods in high dimensional datasets

Form of switch	**Functional form of the data generating model**	**Multicolli nearity**	**Regime 1**		**Regime 2**	
			PAR	**NPAR**	**PAR**	**NPAR**
endogenous	linear	strong	-	25.04	-	21.76
		weak	-	138.50	-	115.29
		none	74.87	150.02	-	123.65
	nonlinear	strong	3.44	15.37	2.97	11.66
		weak	-	107.97	-	85.06
		none	-	111.75	-	88.80
exogenous	linear	strong	-	22.77	-	23.48
		weak	60.37	145.59	-	142.95
		none	28.92	141.22	-	145.00
	nonlinear	strong	-	14.22	2.90	12.85
		weak	-	99.54	-	68.90
		none	-	100.45	-	80.10

Table 8. Per cent difference in the generated MAPE of the superior method in low dimensional datasets

Form of switch	**Functional form of the data-generating model**	**Multicolli nearity**	**Fit of the data-generating model**	**Regime 1**		**Regime 2**	
				PAR	**NPAR**	**PAR**	**NPAR**
endogenous	linear	strong	high	1.18	4.81	3.41	4.61
		strong	low	0.42	5.16	1.38	3.79
		weak	high	-	12.73	5.89	10.11

		weak	low	-	15.43	0.92	11.62
		none	high	2.43	12.31	5.33	9.61
		none	low	-	18.64	2.66	10.38
	nonlinear	strong		0.79	5.52	2.15	6.10
		weak		-	13.60	2.28	9.60
		none		-	16.58	1.60	12.69
exogenous	linear	strong	high	1.42	6.69	3.29	6.62
		strong	low	0.27	4.05	1.08	3.78
		weak	high	1.28	14.68	3.71	10.02
		weak	low	-	13.83	2.45	11.85
		none	high	0.82	14.77	5.49	12.02
		none	low	-	18.27	-	11.83
	nonlinear	strong		0.11	6.14	2.29	7.33
		weak		-	16.52	4.00	13.33
		none		-	17.27	2.94	13.94

6. Application in a Social Development Problem

The performance of the proposed procedure is assessed in an actual dataset. The Client Satisfaction Survey, commissioned by the World Bank in 2005, is used in the study. The aim of the survey was to develop a perception-based survey that will facilitate the verification of the effect of the outputs of the rural sector agencies on rural development in the Philippines, (Barrios, 2008).

The response variable considered is the summary information of the respondent's assessment of the rural development index. An index was extracted from the information obtained from the list of questions on the rural development status using PCA, see (Barrios, 2008) for further details. The independent variables considered in the illustration consist of household characteristics and respondent's assessment of living conditions. There are 30 independent variables in the dataset.

Furthermore, the grouping variable used is the Department of Land Reform (DLR) funding (either locally-funded or foreign funded areas). Here, the nature of switch is assumed to be exogenous, that is, the classification into the two regimes is most likely affected by outside factors, excluding the characteristics included in the model.

For this study, the respondents are grouped into two regimes–individuals from Department of Land Reform (DLR) locally-funded areas (regime 1) and individuals from DLR foreign-funded areas (regime 2). The response variable considered is the summary

information of the assessment of the rural development index by the respondents extracted using PCA. The independent variables included consist of household characteristics and a series of questions of respondent's self-assessment of living conditions.

We examined the value of the variance inflation factors (VIF) and the tolerance value (TOL) to assess the nature of multicollinearity in the data. There is a serious level of multicollinearity in the two regimes since there are several independent variables exhibited VIF greater than 10 or TOL less than 0.1.

We present in Table 9 the MAPE of the nonparametric and parametric PCR, as well as the MAPE of the OLS when applied to the data. It is very clear that there is a problem of using the ordinary least squares regression when applied in a high dimensional data that exhibits multicollinearity, as suggested by very large values of MAPE. When the PCR is applied in the regression regimes, there is a tremendous improvement on the fit of the PCR model when estimated using the nonparametric approach. For DLR locally-funded areas, the predictive ability of the PCR model improved by about 190% in using the nonparametric approach as compared to the parametric counterpart. On the other hand, the MAPE generated by the nonparametric PCR in DLR foreign-funded areas is smaller to the parametric PCR by 600%. The advantage of using the nonparametric over the parametric approaches is obvious in these instances. The generated model in regime 2 is adequate since the computed MAPE is less than 20%. On the other hand, the model generated in regime 1 might not be adequate since the computed MAPE is still 66.60%, but nevertheless, great improvement is achieved when nonparametric PCR is used over the parametric approach. These results further suggest that the nonparametric PCR was able to adequately model the relationship of the household characteristics and the respondent's self-assessment of living conditions to the respondent's assessment of the rural development status, but only in the case that the individuals belong to regime 2. This implies that programs in foreign assisted areas were more defined and were properly monitored compared to the programs in locally funded areas.

Table 9. MAPE generated by parametric and nonparametric PCR in comparison with the result of OLS

Regime 1			Regime 2		
MAPE using OLS	MAPE using PAR	MAPE using NPAR	MAPE using OLS	MAPE using PAR	MAPE using NPAR
133.41	192.40	66.40	31.54	58.70	8.22

When further applied to the counterfactual simulation, the results are clearly misleading if the

resultant model from the OLS or parametric PCR will be used. In Table 10, the measure of individuals' assessment of government's rural development projects if they are included in the first regime is improved by 18.13% if the nonparametric PCR is used over the parametric PCR. Furthermore, there is 47.52% improvement in the estimation of the measure of individuals' (under DLR-local funding) assessment of rural development projects, given they chose to be funded locally by DLR, if the nonparametric PCR is used. Similarly, there is a 44.42% improvement in the estimation of the measure of individuals' (under DLR-foreign funding) assessment of rural development projects, given they are placed under DLR-locally funded areas by using the nonparametric PCR.

Table 10. Counterfactual simulation for parametric and nonparametric PCR

Indicators	**OLS**	**Parametric PCR**	**Nonparametric PCR**	**Per cent difference**
π	-1.18	-0.68	-0.58	18.13
π_1	-1.57	-0.72	-1.37	47.53
π_2	-0.63	-0.64	-1.14	44.42

Using these results, we can say that the expected assessment to the government's rural development programs of the individuals under DLR-local funding is -0.58, a fairly low assessment since the highest index measure is 4.94.

7. Conclusions

Nonparametric principal component regime switching model is generally superior to the parametric counterpart. The nature of switch (endogenous or endogenous) does not affect the predictive ability of both the nonparametric and parametric methods. However, the nonparametric method is superior under high dimensionaldata and specially when the data-generating model in nonlinear. Furthermore, the nonparametric method produced better predictive ability when severe multicollinearity is present or there is misspecification error or that the linear model does not fit the data well.

REFERENCES

Barrios EB. Infrastructure and rural development: Household perceptions on rural development. Progress in Planning. 2008; 70(1).

Cai J, Ung L, Setboonsarng S, Leung P. Rice Contract Farming in Cambodia: Empowering Farmers to Move Beyond the Contract Toward Independence. ADBI Discussion Paper 109.

Curto J, Pinto J. New Multicollinearity Indicators in Linear Regression Models. International Statistical Review. 2007; 75(1): 114-21.

Dehejia R, Wahba S. Propensity score matching methods for non-experimental causal studies. The Review of Economics and Statistics. 2002; 84 (1): 151-61.

Dunteman J. Principal Component Analysis. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-069. Thousand Oaks, CA: Sage; 1989.

Fox J. Nonparametric Simple Regression: Smoothing Scatterplots. Thousand Oaks CA: Sage; 2000.

Hastie T, Tibshirani R. Generalized Additive Models, Chapman and Hall. 1990.

Heckman J. Sample Selection Bias as a Specification Error. Econometrica. 1979; 47(1): 153-161.

Maddala GS. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press; 1983.

Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1): 41-55.

Umali J., Barrios, E.Nonparametric Principal Components Regression. Communications in Statistics-Simulation and Computing. 2014,43(7): 1797–1810.