Sparse Principal Component Regression

Joseph Ryan G. Lansangan
University of the Philippines Diliman

Modeling of complex systems is commonly confronted with high dimensional set of independent variables. Similarly, econometric models are usually built using time series data that often exhibit nonstationarity due to the impact of some policies and other economic forces. In both cases, linear regression modeling may yield unstable least squares estimates of the regression coefficients. Principal component regression can provide solution, but in cases where the regressors are nonstationary or the dimension exceeds the sample size, principal components may yield simple averaging of the regressors and the resulting model is difficult to interpret due to biased estimates of the regression coefficients. As a solution, a sparsity constraint is added to the least squares criterion to induce sparsity.The components may then reflect the relative importance of each regressor in a sparse principal component regression (SPCR) model. Simulated and real data are used to illustrate and assess performance of the method. SPCR in many cases leads to better estimation and prediction than conventional principal component regression (PCR). SPCR is able to recognize relative importance of indicators from the sparse components as predictors. In summary, SPCR can be used in modeling high dimensional data, as an intervention strategy in regression with nonstationary time series data, and when there is a general problem of multicollinearity.

Keywords: sparsity, high dimensionality, multicollinearity, nonstationarity, sparse principal components

1. Introduction

Given the linear model, $y=X\beta+\in$ where $\in\sim(0, \sigma^2)$, the ordinary least squares (OLS) estimator of the regression coefficients is given by $\hat{\beta}=\left(X^TX\right)^{-1}X^Ty$ with variance $V\left(\hat{\beta}\right)=\left(X^TX\right)^{-1}\sigma^2$. If the columns of X are not orthogonal, then $(X^T X)$ is ill-conditioned. The minimum implication is that the variances of the estimates of the regression coefficients are inflated. The more dangerous

implication is that the matrix ($X^T X$) is singular and hence, the OLS estimate does not exist or is extremely unstable.

Also, when columns of X is non-orthogonal, i.e., the multicollinearity problem exists, ordinary least squares estimates will have inflated variance and cases of reverse signs among the estimated coefficients can be expected. One solution is to drop some variables that may possibly duplicate the effect of other variables. This may however create a void in a modeling framework where the dropped variable may serve as an indicator of a specific component of the framework. It may also lead to bias on the OLS estimates of some regression coefficients. Other solutions like principal components regression, ridge regression, and other shrinkage-type of estimators were proposed in the literature.

Principal components analysis provides a useful tool in modeling since it can create fewer orthogonal aggregates of the variables. In cases of non-stationary time series data however, there is a tendency for the first component to produce simple averaging of all variables. Regression on a component with equal weights for all standardized indicators can result to assignment of similar coefficients to all independent variables. Thus, the relative importance of the indicators in predicting the dependent variable will be masked with the first principal component alone.

As an alternative, components of the independent variables are derived via a sparsity-inducing constraint, and the resulting sparse components are then used as the regressors for the dependent variable. The performance of this regression method is assessed based on simulated data. Furthermore, the interpretability of the resulting estimates from a sparse principal component regression is illustrated in an endogenous growth model.

2. Dimensionality, Time Series and Multicollinearity

The vacuum in empirical investigation relating to the endogenous growth theory lies on the analysis at the national level using time series data. In order to minimize the possibility of missing out some important drivers of growth, more indicators or proximate indicators are postulated in the model. These high dimensional set of predictors are also collected/monitored over time, leading to high dimensional time series data.

In time series data of measurements of indicators that benefit from macroeconomic policies and national programs, natural drifting of the variables is expected, resulting to nonstationary behavior. Nonstationarity can easily cause the predictors in a model to exhibit non-orthogonality. In a linear model, non-orthogonality of the predictors causes the multicollinearity problem that generally results to instability of the least squares estimates of the regression coefficients. In many cases, the unstable estimates of the regression coefficients can lead to inverse signs relative to the theoretical direction of the relationships. One solution to this issue is to drop those duplicating variables, but doing so may void the dynamics being assessed according to some econometric framework.

Focusing on the variance inflation problem caused by multicollinearity, shrinkage estimators are also considered as alternative solutions. Constraints are added in the least squares objective function to produce nonsingular design matrix, alleviating variance inflation. The gain in precision is necessarily compensated by the propagation of bias in the parameter estimates. This can also complicate the interpretation of the relative contribution of the individual determinants towards the dependent variable.

Principal component regression has been proposed as a possible solution to the problem of multicollinearity. A subset of the orthogonal transformation of the independent variables can be used but with a discrepancy in the amount of information used between the raw individual predictors and the components. Principal component regression may work even if there are more predictors than the number of observations. However, nonstationarity in time series and bulkiness of the number of determinants may result to uniformity of component loadings and hence can complicate interpretations of results. Similarity of loadings in the component can imply that the individual determinants will have contributions to the dependent variable with similar magnitude. Thus, the relative importance of each determinant in the model might be difficult to understand.

Foucart (2000) suggested deleting components that are not significant. This will however introduce bias to the least squares estimates of the remaining coefficients. The magnitude of the eigenvalues associated with a component may also be considered as an alternative indicator of which component to retain but prediction of the dependent variable can potentially suffer. Jolliffe (1982) further cautioned about the misconception that principal components with small eigenvalues will rarely be of any use in regression. It was demonstrated that these components can be as important as those with large eigenvalues, thus, the search of a better way to aggregate high dimensional data or non-orthogonal predictors continues.

Instead of the usual maximum likelihood estimation for generalized linear regression, in the presence of ill-conditioned design matrix, Marx and Smith (1990) proposed an asymptotically biased principal component parameter estimation technique. The principal component regression for generalized linear regression is not always the best choice for model building, but depending on model orientation/specification, it can yield desirable variance properties with minimal bias.

Principal component regression (PCR) as noted by Kosfeld and Lauridsen (2008) is best in cases with high multicollinearity. Rotation of axis of components in factor analysis may aid in interpretation and can yield superior model when the independent variables are moderately correlated. Using cointegrated time series data, Harris (1997) emphasized the advantage of requiring neither the normalization imposed by the triangular error correction model nor the specification of a finite-order vector autoregression.

A model with infinitely many parameters was proposed by Goldenshluger and Tsybakov (2001). To deal with this overparameterized model, an application of blockwise Stein's rule with "weakly" geometrically increasing blocks to the penalized least squares to fit the first N coefficients. Instead of conditional distribution of Y on X, Heland (1992) considered the joint distribution of Y and X then work on the fixed number of components.

Heij et al. (2007) compared forecasts of PCR and principal covariate regression (component loadings and regression coefficients are simultaneously estimated to minimize squared error quantities). With the number of factors chosen by BIC in principal covariate regression, its forecast outperform PCR forecast. Filzmoser and Croux (2002) also proposed an algorithm in constructing orthogonal predictor that will simultaneously maximize prediction of the dependent variable.

3. Sparse Principal Component Analysis

Principal Components Analysis (PCA) as a dimension reduction technique is used to detect possible structures existing among variables, particularly defining linear components that capture the information in the data contributed by the different variables (Jolliffe, 2002). The PCs are uncorrelated and hence characterization of the PCs is easily implemented. But one possible drawback of PCA is the difficulty in interpretation of the first few PCs. Several authors, e.g., Jolliffe and Uddin (2000), have used both cross-sectional models and time series models to assess and improve existing methods addressing this interpretability issue. Vines (2000) considered simplicity preserving linear transformations to generate simple and interpretable results. Sparsity has also become an issue on dimensionality reduction, this as one of the solutions to simplify interpretation of the PCs. Chipman and Gu (2005) addressed the problem by considering homogeneity constraints and sparsity constraints. Zou and Hastie (2005) further modified the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996).

Zou et al. (2006) used LASSO as a constraint to principal components extraction, thereby formulated the extraction as a regression problem, resulting in components with sparse loadings. The optimization problem, or the sparse principal component analysis (SPCA) criterion is given by, with $A_{pxk}=[\alpha_1,...,\alpha_k]$ and $B_{pxk}=[b_1,...,b_k]$,

$$(\hat{A},\hat{B}) = \arg \min_{A,B} \sum_{i=1}^{n} \left\| X_i - AB^T X_i \right\|^2 + \lambda \sum_{j=1}^{k} \left\| b_j \right\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| b_j \right\|_1 \text{ with } A^T A = I_k \quad (1)$$

where X_i are the data matrices for each observation i, with λ and $\lambda_{1,j}$ as penalizing constants chosen to facilitate existence of solution and convergence of the computational algorithm. The solution to the optimization problem in (1) is called sparse principal components (SPCs). Optimization is done through regression-

type criterion to derive SPCs in two stages – first is to perform an ordinary PCA, and second is to find sparse approximations of the first k vector of loadings of the PCs using the "naïve elastic net." But unlike PCA, the algorithm gives components that are correlated and loadings that are not orthogonal. Thus the total variance is not simply calculated as the sum of the predicted variances of the PCs. Lansangan and Barrios (2009) applied the technique to nonstationary time series and addressed the problem of finding values of the tuning parameters that will ensure convergence of the algorithm. As a modification to the algorithm, Leng and Wang (2009) proposed the adaptive LASSO (ALASSO) which improves efficiency of the estimation and selection results via a BIC-type tuning parameter selector.

4. Sparse Principal Component Regression

Principal component regression (PCR) where the principal components are extracted from the predictors prior to regression modeling is commonly used when such predictors are non-orthogonal. Unlike partial least squares estimation which defines components in relation to the response variable Y, PCR uses components which are constructed independently of Y. The method does not identify clearly whether to include only the first few PCs, which are associated to large eigenvalues, and hence large variability in X included in explaining Y. Jolliffe (1982) stresses out the importance of PCs with small eigenvalues. Hadi and Ling (1998) also present an example where only the last few PCs (those associated to small eigenvalues) were associated with the response variable.

If the predictors involved are collected over time, then there is a very high chance that the predictors exhibit nonstationarity that will result to averaging of the individual determinants in extracted components, as also showed by Lansangan and Barrios (2009). Such averaging will then mask the relative importance of some variables over the others. One econometric solution to this problem is to compute growth rate (differencing) of the indicators instead of the original levels in modeling. Differencing however, may result to an alteration of the dependence structure. Differencing generally filters low data frequencies and preserves the high frequencies, eliminating the effect of some important random shocks, possibly contaminating the econometric relationship being investigated. The principal components analysis intends to summarize many indicators on the same theme into few components in a form of linear combinations, and oftentimes, the first component is block (average of all indicators), the rest are contrasts, which are more difficult to interpret. Vines (2000) proposed an algorithm that will produce approximate principal components through 'simplicity preserving' linear transformations. Rousson and Gasser (2004) included a constraint in principal component extraction, resulting to suboptimal than ordinary principal components but with simpler, more interpretable components. Chipman and Gu (2005) imposed two constraints in principal component extraction, first (homogeneity

constraint) coefficients are constrained to equal a small number of values, second a sparsity constraint. Tibshirani (1996) minimized the residual sum of squares in a model subject to the sum of the absolute value of the coefficients being less than a constant (known as the "lasso" or elastic net). By extending the soft thresholding and lasso methods to generalized linear models, Klinger (2001) used penalized likelihood estimators for a large number of coefficients. The extension leads to an adaptive selection of model terms without substantial variance inflation.

The proposed optimization constraint in the choice components that will be used in the model is hoped to address nonstationarity, non-orthogonality, and high dimensionality of the data. The method, called sparse principal component regression (SPCR) uses sparse principal components (Zou et al., 2006; Leng and Wang, 2009) as inputs for the regression model. Since the SPCs are based solely on the first few PCs, but with the sparsity that comes in, it seems that sparse principal component regression (SPCR) may provide a solution to multicollinearity, high dimensionality, and the issue on components selection.

Theorem. Given the data on the response y(nx1) and the predictors X(nxp). Sparse principal component regression (SPCR) minimizes the following objective function for β(px1), A(pxk), and B(pxk):

$$||y - X\beta||^2 + ||X - XAB^T||^2 + \lambda_1||B||^2 + \lambda_2||B||_1 \text{ subject to } A^T A = I_k \quad (2)$$

Proof:

Let C = [β, AB^T] and D = [y, X]. Following Heij et al., (2007), (2) can be written as

$$||D - XC||^2 + \lambda_1||B||^2 + \lambda_2||B||_1 \quad (3)$$

The second and third term of (3) involves only the penalty for B to induce sparsity and does not contain any part of the data. Let f (C) = $||D-XC||^2$ which is minimized when tr[(D-XC)^T (D-XC)] is minimized. Following Heij et al. (2007), let X=USV^T be the singular value decomposition of X, where matrices U and V are such that U^{T}U=V^{T}V=I$_m$. So that we get min f (C) to be min tr(D-XC)^T (D-XC) = min $||U^TS - SVC^TC||^2$. Define $f^*(C) = tr\,(D^TUU^TD) - 2tr(D^TUSV^TC) + tr(C^TSV^2V^TC)$. The first term is independent of the choice of C (unknown quantities: regression coefficients and component loadings). Hence, $f(C)$ and $f^*(C)$ have the same minimum. From Eckart and Young (1936), $f(C)$ is minimized when C is composed of orthonormal eigenvectors associated with r largest eigenvalues of D^{T}D and X=DC. Since C = [β, AB^T] and D = [y, X], $tr[(D-XC)^T (D-XC)] = tr[(y-XC)^T (y-XC)] + tr[(X-XAB)^T(X-AB)]$, minimization of the first term leads to the OLS of the regression coefficients, and minimization of the second term leads to the usual principal components. Zou et al. (2006) proposed an algorithm of

SPCA that starts with ordinary PCA and proceeds iteratively to incorporate the sparsity constraints.

For the proposed procedure, the SPCA algorithm follows that of Zou et al. (2006) and adapts Leng and Wang (2009) criterion for the tuning parameter(s) selection ($\lambda_{1,j}$ *s*). Specifically, the tuning parameters are chosen so that BIC_λ defined as $BIC_\lambda = \frac{1}{n}\sum_{i=1}^{n}\left(\underline{Z} - \hat{b}_j'\underline{x}_i\right)^2 + df_\lambda * \frac{\ln(n)}{n}$ is the smallest over a range of $\lambda_{1,j}$ s where $\underline{Z}$ is the standardized PC scores and df_λ is the number of nonzero coefficients in the estimate β_j using λ as the tuning parameter.

5. Simulation Study

Data simulation adopted the strategy of Leng and Wang (2009). Let $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{id})' \in \Re^d$ be the d-dimensional data from the i^{th} unit i = 1,2,...,n. Without loss of generality, take $E(x_i)$ = 0 and $Cov(x_i)$=Σ for some positive definite matrix Σ, which can be represented as $\Sigma=\delta_1 b_1 b_1' + \delta_2 b_2 b_2' + ... + \delta_d b_d b_d'$ where δ_j is the j^{th} largest eigenvalue of Σ and $b_j = (b_{j1}, b_{j2}, ..., b_{jd})' \in \Re^d$ is its associated eigenvector. Here assume that $\delta_1 > \delta_2 > ... > \delta_d > 0$ and the first nonzero component of each b_j is positive. Under these conditions, $\{b_j\}$ are uniquely determined by Σ and are orthonormal. Thus, the j^{th} PC of $\underline{x}_i$ is given by $\underline{x}_i'b_j$, which accounts for a $\delta_j \Big/ \left(\sum_k \delta_k\right) * 100\%$ of the total variance $\sum_j Var(x_{ij})$. For the simulated normally distributed data matrix $\underline{X}$, different specifications of b_j and δ_j were considered. The multicollinearity problem (correlations among the $\underline{X}$) and the importance of the components (based on δ-specifications) are introduced and controlled, respectively, into the model.

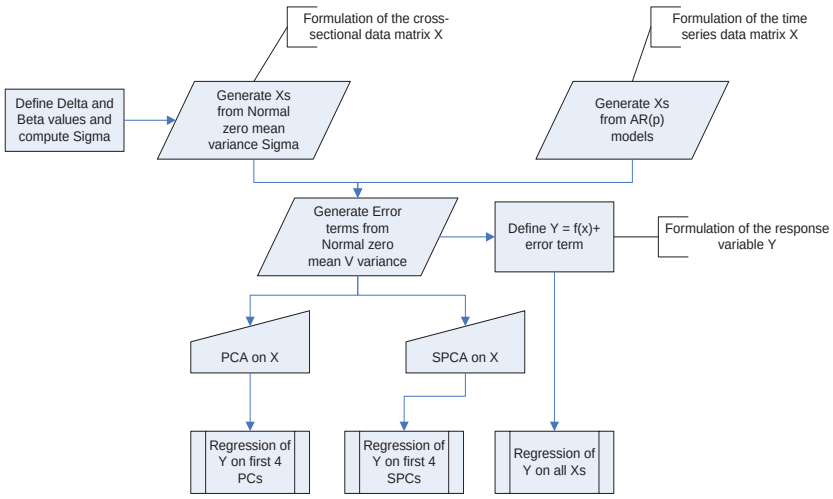
To emphasize the importance and/or non-importance of the input variables in the regression problem, the response variable y_i is taken as stochastic function of no more than half of the total $\underline{x}_i$, i.e., at most d/2 input variables are assumed to be significant predictors of the response variable. The error terms are generated from $N(0,\sigma^2)$, then added to the response to maintain control over the degree of fit of y_i on $\underline{x}_i$.

Two scenarios are presented, first, the case when the input data is time series, and second, when the input data is cross-sectional. Simulations are made on the premise that there are 4 inherent groupings of the input variables (or series). For the cross-sectional data, groupings are based on its variance-covariance matrix, i.e., the correlations among the input variables are "controlled" such that there are four distinct groupings (with high correlations within groups and low correlations across groups). For the time series data, groupings are based on the assumed model of the input series, i.e., each time series is an AR(p) model, hence the four distinct groupings rely on the specifications on the values of p. In all cases,

the number of time points (or number of units for cross-section data) is 40. The number of independent variables were varied from as low as 16 (p<n) to as many as 64 (p>n).

After generating Y and X, regression analysis is conducted, together with PCR, and the proposed SPCR. Note that for both PCR and SPCR, only the first 4 PCs and SPCs were included in the regression models. The existence of four natural groupings in the variation of the independent variables (or in the patterns of the time series) is expected to yield four important components, and hence PCR and SPCR are applied using the first four components extracted. A flow chart of the simulation and estimation procedures is presented in Figure 1.

Figure 1 Flow Chart of Simulations



5.1 Time sries data

Using all 16 independent variables (8 stationary series and 8 nonstationary series), the average of coefficient of determinations (R^2) for all datasets is expectedly high at 99.99%, with MAPE of 0.09%. Some 86% of the datasets yield 6 to 9 significant variables when all the 16 variables are used in the model. PCR yield a very low R^2, with a remarkable decline of about 93% when all the 16 individual variables (coming from the 4 components) served as regressors. MAPE is also relatively high at 13.30%. On the otherhand, SPCR yield very high R^2, with barely a decline of less than 1% when instead all the 16 independent variables are used as predictors. MAPE is also low at 6.65%, less than half of what has

been achieved in PCR. The R^2 of SPCR is better than that of PCR in 100% of the data. In PCR, 94% of the dataset also yield none of the four important PCs significant. In SPCR however, at least 1 of the four important SPCs turned out to be significant, with majority (83%) having 2 or 3 SPCs significant.

With 16 independent variables, all are stationary this time, the R^2 when all variables are used as predictor averages 85.70%, with 84% of the datasets having 4 to 8 significant variables. In majority of the datasets (61%), SPCR still exhibited better R^2 than PCR where the averages are at 36.76% and 30.52%, respectively. Many of the components are significant predictors in SPCR than in PCR. In SPCR, with 89% of the datasets yield 1-4 significant SPCs, while only 1-3 PCs are significant in PCR for 90% of the datasets.

If all the 16 indicators are stationary, PCR is comparable to SPCR. However, when all the indicators are nonstationary, SPCR is superior to PCR in 100% of the data based on R^2 values. Given all 16 independent variables, 5 to 9 significant variables are identified with average R^2 of 99.66% and MAPE of 8.40%. In SPCR, R^2 averages 94.82% with MAPE of 3.37% (even better than when all the 16 independent variables are included), 100% of the dataset also identified 1 to 4 important components as significant predictors. In contrast, PCR's average R^2 is only 36.48%, with MAPE at 13%, and only 1 or 2 PCs are significant predictors in 22% of the datasets.

Increasing the number of predictors (32) closer to the number of samples (40), a set with a combination of stationary and nonstationary time series, 100% of the datasets still exhibit superiority of SPCR over PCR in terms of R^2. With all predictors included, R^2 averages 99.42% with MAPE of 2.52% and 1 to 11 variables are significant in 92% of the datasets. In PCR, average R^2 is 2.01% with MAPE of 108% and none of the datasets exhibiting a significant PC as predictor in the model. In SPCR, R^2 averages 79.67% with MAPE of 14.78% and 96% of the datasets exhibited significance of 1 to 3 SPCs as predictor of the model. If all the predictors are stationary, PCR catches up with SPCR, with average R^2 at 28.38% and 33.95%, respectively. But when all the predictors are nonstationary, the advantage of SPCR over PCR is again emphasized.

With 40 predictors this time which is a combination of stationary and nonstationary time series, PCR yield an average R^2 of 72.98% and MAPE of 32.94% and most of the datasets identifying 1 to 2 PCs as significant predictors of the model. SPCR yield an average R^2 of 78.11% and MAPE of 22.91% with majority of the datasets identifying more significant SPCs (2 to 3) in the model. R^2 of SPCR is better than PCR in 70% of the datasets. When all the predictors are stationary, PCR yield an average R^2 of 16.58% with MAPE of 7.69% while SPCR have average R^2 of 26.31% with MAPE of 7.24%. Most of the datasets (66%) noted significance of 1-4 SPCs while only 45% of the datasets recognized significance of 1 to 2 PCs in the model. When all the predictors are nonstationary, SPCR again is superior to PCR in 100% of the datasets.

Simulating 64 predictors (p>n) which is a combination of stationary and nonstationary time series, PCR yield an average R^2 of 2.48%, MAPE of 13.98%, and none of the datasets indicating significance of the 4 PCs as predictors of the model. SPCR yield an average R^2 of 99.86% with very low MAPE at 0.49%. All the datasets indicated at least 2 SPCs as significant predictors of the model. With all predictors stationary, 78% of the datasets proved superiority of SPCR over PCR. PCR has low R^2 of 10.62% and MAPE of 4.02%. SPCR also yield low R^2 (but higher than in PCR) at 23.09%, with MAPE at 3.74%. In SPCR, 68% of the datasets yield 1 to 3 SPCs as significant predictors of the model, while only 17% of the datasets yield 1 to 2 significant PCs in PCR. With all nonstationary predictors, SPCR is superior over PCR in 100% of the datasets based on R^2 values. R^2 in PCR average 23.57% with MAPE of 12.70% and 95% of the datasets indicating none of the important PCs significant. In SPCR, average R^2 is 97.58%, MAPE is 2.01% and 100% of the datasets correctly indicating 2 to 4 SPCs as significant predictors of the model.

5.2 Cross-section data

For the cross-section data, four groups of predictors are created based on the variance-covariance matrix generated. With 16 predictors, a model including all raw predictors yields an average R^2 of 86.93% and MAPE of 174%. Also, 88% of the datasets indicated significance of 1 to 9 predictors in the model. In PCR, average R^2 is 13.40% with MAPE of 507% and only 24% of the datasets recognized significance of 1 to 3 PCs as predictors of the model. In SPCR, average R^2 is 70.55% and MAPE is 201%. In all datasets, 1 to 4 SPCs are noted as significant predictors of the model.

Increasing the number of predictors to 32, the original model has average R^2 of 98.14% and MAPE of 154%. In PCR, average R^2 is only 23.01% with MAPE of 267% and 1 to 2 PCs are significant predictors in 21% of the datasets. In SPCR, average R^2 is 70.07% and MAPE is 286%. Also, 100% of the datasets noted 1 to 4 SPCs as significant predictors of the model. As more indicators are added in the simulations, SPCR becomes even more superior over PCR in terms of R^2 and MAPE.

5.3 Summary

For cross sectional data, SPCR is better than PCR in terms of R^2. SPCR also has lower MAPE relative to PCR at least 52% of the time. If p<n, PCR may still be comparable to SPCR, but as p increases further until it is larger than n, SPCR becomes more advantageous than PCR.

For all stationary series predictors, MAPE in SPCR is lower than in PCR 52% of the time (about 75% at most). For combination of stationary and nonstationary time series as predictors, SPCR again is better than PCR 76% of the time based on MAPE. For the case of p=16 or p=64, SPCR is better than PCR in terms of

MAPE 100% of the time. For p=32, SPCR is better 97% of the time using MAPE. For nonstationary time series, SPCR is better 100% of the time than PCR using MAPE.

For time series data, SPCR is far better than PCR for all cases (stationary–nonstationary series combination, stationary series, and nonstationary series) in terms of retained information and/or predictive power. Noticeably, in terms of R^2, SPCR is always better than PCR for the combination and the nonstationary series data sets 100% of the time (except when p=40, with only 70%), while for data with all stationary series, SPCR is better than PCR 60-78% of the time.

Unlike the first few PCs when used as the predictor variables for *y*, the first few SPCs retain much information coming from the original independent variables. For the cross-sectional data, the stationary-nonstationary combination time series data, and the nonstationary time series data, there is only at the most about 29% decrease in R-square. Such characteristic, however, remains ambiguous for the stationary cases – where the drop in R-square or adjusted R-square is at least 57%.

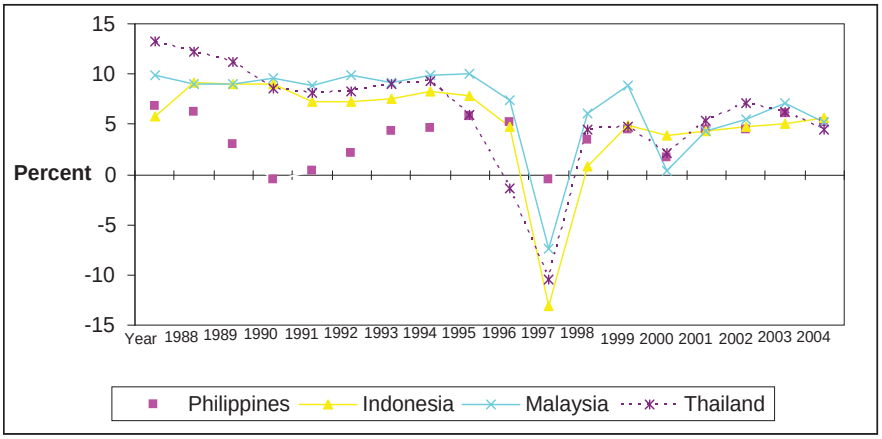
6. Applications to Endogenous Growth Models

Using data for Philippines, Indonesia, Malaysia, and Thailand from 1988 to 2004, real GDP growth was regressed on the following determinants: X_1= logarithm of GDP; X_2= population growth; X_3= percent share of agriculture to total GDP; X_4= percent change in food prices; X_5= percentage of total government expenditure to GDP; X_6= percent growth in export; X_7= percent growth in import; X_8= external debt as a proportion of GNI; and X_9= debt servicing as a percentage of exports. The growth rates from 1998 to 2004 of the four countries are plotted in Figure 2. The Philippines exhibited a different growth pattern from the other three countries prior to the 1997 Asian financial crisis. While the Philippines had the least reaction to the crisis, the growth patterns of all four countries exhibit similar behavior in the recovery period from the crisis.

The augmented Dickey-Fuller (ADF) unit root test yield almost all independent variables nonstationary at 5% level. The possible multicollinearity problem postulated earlier is confirmed, with condition numbers ranging from 135,649 to as much as 1,269,738, all indicative that the problem can seriously ruin the stability of the least squares estimates of the regression coefficients.

Three components were extracted from the predictors for each country. In ordinary principal components extraction, the percentage of the total variance explained by the three components ranges from 77% to 87%. Using sparse principal component extraction criterion, the variance explained by three components ranges from 70% to 79%. The gain in sparsity of the components is paid off by the slightly lower proportion of variance explained by the sparse principal components. The component loading based on standardized determinants are summarized in Appendix 1. Because of the averaging effect of nonstationarity of the determinants, the principal components are difficult to interpret. From the

Figure 2 GDP Growth Rates of Philippines, Indonesia, Malaysia and Thailand



sparse component loadings however, it is easy to interpret that the first component represents the effect of foreign trade on local prices of food in Indonesia, and the structure of the macro-economy for Malaysia, Thailand and Philippines.

Philippines

The ordinary regression on all predictors yield an R^2 of 89%, much lower in principal component regression at 54% and 41% in sparse principal component regression. The estimates of the regression coefficients are given in Table 1. The ordinary least squares of directly regressing growth of GDP to all nine determinants, only the logarithm of GDP yield a significant coefficient. In the principal component regression (PCR), population change (–), percent change in food prices (–), percent of government expenditures to GDP(–), growth in export(+), growth in import(+), and proportion of external debt to GNI (–) yield smaller standard errors. Most of the signs are consistent with theory but a few like percent of government expenditures to GDP seems to exhibit reverse signs.

From the coefficients of sparse principal component regression (SPCR), percent change in food prices (–) and growth in import (+) yield smaller standard errors. Higher food prices hit the local consumers while importation (of food) can possibly facilitate growth of the Philippine economy where the endogenous growth driver may have focused on consumer demand. Consumer spending tends to fuel growth, hence, prices of food commodities and importation can help mitigate the market in favor of the consumers.

Table 1 Regression Coefficients for Philippines

Variable	SPCR		PCR		OLS		
	Coeffi-cient	Std Error	Coeffi-cient	Std Error	Coeffi-cient	Std Eror	p-value
Intercept	13.1164	11.8298	29.4288	10.9058	-139.7940	71.0266	0.0897
X1	-0.6548	0.8608	0.2747	0.5119	17.9528	7.4445	0.0467
X2	-0.7095	0.9327	-4.3514	1.6043	-8.2364	8.0844	0.3422
X3	-0.0381	0.0501	-0.0364	0.0382	-1.1937	0.5542	0.0682
X4	-0.1247	0.0405	-0.1414	0.0525	0.1489	0.1369	0.3131
X5	--	--	-0.7851	0.2146	-0.4295	0.4955	0.4148
X6	0.0030	0.0113	0.0226	0.0145	0.0729	0.0578	0.2473
X7	0.1366	0.0492	0.0743	0.0195	0.1417	0.0757	0.1034
X8	-0.0030	0.0109	-0.0583	0.0212	0.1078	0.0730	0.1832
X9	-0.0546	0.0667	0.0307	0.0380	-0.0174	0.1592	0.9159

Malaysia

The coefficient of determination for ordinary regression on all predictors is 87%, much lower in principal component regression at 61% and 57% in sparse principal component regression. The estimates of the regression coefficients are given in Table 2. Only the external debt as a proportion to GNI yield significant coefficient in the direct regression of growth in GDP to all nine indicators. For the PCR, population change (+), percent change in food prices (–), growth in import (+), and external debt as a proportion to GDP(–) have lower standard errors. Population growth usually exhibits negative effect on growth, in PCR however, it yields positive sign.

The percent share of agriculture to GDP, percent share of external debt to GNI and percent of debt servicing to exports yield lower standard errors in PCR. The positive effect of share of agriculture to GDP can be explained by the possibly improving contribution of agriculture and not having the industries failing to expand significantly. As external debt increases relative to GNI, the country will be forced to finance the debt rather than fuel growth.

Indonesia

In the ordinary regression on all predictors, principal component regression, and in sparse principal component regression, the coefficient of determination is always high (97%, 92% and 84%, respectively). The estimates of the regression coefficients are given in Table 3. Only one factor (percent change in food prices) yield significant parameter estimate in the ordinary regression. PCR yield more determinants with smaller standard errors of estimates than the SPCR. Population

growth rates and percent of government expenditures to GDP were included in PCR, but SPCR included percent share of agriculture to total GDP instead. All the other variables including: percent change in food prices, growth rate of exports, growth rate of imports, and external debt relative to export yield smaller standard errors for both SPCR and PCR, also producing the same signs. Growth of export and import in Indonesia had bigger growth contribution as estimated in SPCR compared to PCR.

Table 2 Regression Coefficients for Malaysia

Variable	SPCR		PCR		OLS		
	Coeffi-cient	Std Error	Coeffi-cient	Std Error	Coeffi-cient	Std Eror	p-value
Intercept	27.4037	8.1140	5.2604	5.2370	-3.5435	80.1790	0.9660
X_1	-0.7760	0.6083	-0.4386	0.6480	3.3559	5.6440	0.5708
X_2	-	-	4.8039	1.3100	1.1398	2.3462	0.6419
X_3	0.1018	0.0798	0.0118	0.0715	0.4002	1.0336	0.7101
X_4	-0.1160	0.1579	-0.5189	0.1887	-0.7478	0.6370	0.2788
X_5	0.1350	0.1838	0.1640	0.1299	-0.6717	0.3679	0.1106
X_6	0.0129	0.0184	0.0087	0.0250	-0.2161	0.1746	0.2557
X_7	0.0167	0.0131	0.0906	0.0200	0.1381	0.1132	0.2620
X_8	-0.3620	0.1049	-0.1945	0.0472	-0.3501	0.1263	0.0276
X_9	0.0562	0.0396	-0.0152	0.0611	0.2788	0.4100	0.5184

Table 3 Regression Coefficients for Indonesia

Variable	SPCR		PCR		OLS		
	Coeffi-cient	Std Error	Coeffi-cient	Std Error	Coeffi-cient	Std Eror	p-value
Intercept	4.8440	7.5642	-6.2624	3.6906	-37.4182	31.6032	0.2751
X_1	-0.6706	0.5282	0.1992	0.3987	2.6554	2.1286	0.2523
X_2	0.6204	0.4886	3.0847	0.3952	3.9090	2.6723	0.1869
X_3	0.1106	0.0871	0.0992	0.0622	-0.2323	0.4400	0.6138
X_4	-0.0724	0.0096	-0.1062	0.0129	-0.2394	0.0563	0.0038
X_5	0.3774	0.3917	0.4134	0.1941	-0.0263	0.3583	0.9436
X_6	0.1393	0.0186	0.0694	0.0154	-0.0744	0.0942	0.4554
X_7	0.0797	0.0106	0.0488	0.0057	0.0621	0.0514	0.2662
X_8	-0.0484	0.0064	-0.0585	0.0060	0.0177	0.0384	0.6595
X_9	0.0410	0.0323	-0.0027	0.0239	0.3692	0.1758	0.0739

Thailand

Relatively higher coefficient of determination is also observed in Thailand with 95% for ordinary regression on all predictors, 87% for principal component regression, and 78% for sparse principal component regression. The estimates of the regression coefficients are given in Table 4. There are more significant determinants estimated from OLS in the model for Thailand, including: growth of export, growth of import, and external debt as a proportion of GNI. There are also fewer parameters with small standard errors in SPCR compared to PCR.

Table 4 Regression Coefficients for Thailand

Variable	SPCR		PCR		OLS		
	Coeffi-cient	Std Error	Coefficient	Std Error	Coeffi-cient	Std Err	p-value
Intercept	31.8262	6.9061	22.9748	7.7001	45.0701	58.5108	0.4663
X_1	-1.1848	0.6313	0.8021	0.8725	-1.8135	4.3360	0.6883
X_2	1.9962	0.6061	0.6665	0.4766	7.3421	3.4664	0.0719
X_3	0.2779	0.0844	0.1181	0.0700	-0.7055	0.5610	0.2489
X_4	0.0312	0.0810	-0.2886	0.0855	-0.0655	0.2370	0.7903
X_5	-0.3539	0.1075	-1.0504	0.1173	-0.5182	0.6545	0.4546
X_6	-0.0942	0.0302	-0.0669	0.0205	-0.2358	0.0878	0.0313
X_7	0.0305	0.0093	0.0970	0.0113	0.2148	0.0604	0.0093
X_8	-0.1829	0.0396	-0.1377	0.0177	-0.1159	0.0449	0.0364
X_9	-0.0457	0.1187	-0.1701	0.0756	-0.1720	0.2338	0.4860

7. Conclusion

When $p<n$ in time series data, SPCR is generally advantageous over PCR based on R^2 and MAPE, at the least, they are comparable. SPCR is able to identify correctly that the important SPCs are significant predictors of the model. This is not the case for PCR where many times, the important PCs are not necessarily significant predictors of the model. This observation is true for cross sectional data and in time series data with either a combination of stationary and nonstationary predictors or all predictors are nonstationary. If all the predictors are stationary, PCR and SPCR are comparable. As the number of predictors becomes large (closer to the number of observations), SPCR is still better than PCR. In case $p>n$, and all the predictors are stationary, PCR and SPCR are comparable, but when the predictors are combination of stationary and nonstationary time series or all nonstationary time series, SPCR becomes more advantageous than PCR in terms of predictive ability in a linear regression model.

In cross-sectional data where p<n, SPCR have higher R^2 than PCR, but their MAPE are comparable. As p increases to surpass n, SPCR still yield higher R^2 than PCR and MAPE is also lower. In SPCR, the important SPCs are also correctly identified, not the case in PCR where in many datasets, important PCs, are not necessarily significant predictors of the model.

In an econometric model with determinants that exhibit nonstationary behavior, the multicollinearity problem can easily affect least squares estimation of the parameters. Aside from parameter estimates that are unstable, it can also yield signs that are reversed of what is expected. Principal component regression can help resolve the multicollinearity problem but it may yield models that are difficult to interpret because the first component usually averages all determinants if nonstationarity dominates the predictors. Sparsity in the components in sparse principal component regression can facilitate the interpretability of the resulting model as illustrated is some endogenous growth model for four ASEAN-member countries.

REFERENCES

CHIPMAN, H. and H. GU, 2005, Interpretable dimension reduction, *Journal of Applied Statistics* 32(9): 969-987.

ECKART, C. and G. YOUNG, 1936, The approximation of one matrix by another of lower rank, *Psychometrika1*(3): 211-218.

FILZMOSER, P. and C. CROUX, 2002, Dimension reduction of the explanatory variables in multiple linear regression, *Pliska Stud. Math.Bulgar* 29: 1-12.

FOUCART, T., 2000, A decision rule for discarding principal components in regression, *Journal of Statistical Planning and Inference* 89: 187-195.

GOLDENSHLUGER, A. and A. TSYBAKOV, 2001, Adaptive prediction and estimation in linear regression with infinitely many parameters, *The Annals of Statistics* 29(6): 1601-1619.

HADI, A. and R. LING, 1998, Some cautionary notes on the use of principal components regression, *The American Statistician* 52: 15-19.

HARRIS, D., 1997, Principal components analysis of cointegratedtime series, *Econometric Theory* 13: 529-557.

HEIJ, C., P. GROENEN, and D. VAN DIJK, 2007, Forecast comparison of principal component regression and principal covariate regression, *Computational Statistics and Data Analysis* 51: 3612-3625.

HELLAND, I., 1992, Maximum likelihood regression on relevant components, *Journal of the Royal Statistical Soc. Ser. B* 54(2): 637-647.

JOLLIFFE, I., 1982, A note on the use of principal components in regression, *Journal of Applied Statistics* 31(3): 300-303.

_____, 2002, *Principal Component Analysis*, 2nd ed. Springer-Verlag, New York.

JOLLIFFE, I. and M. UDDIN, 2000, The simplified component technique: An alternative to rotated principal components, *Journal of Computational and Graphical Statistics* 9: 689-710.

KLINGER, A., 2001, Inference in high dimensional generalized linear models based on soft thresholding, *Journal of the Royal Statistical Soc. Ser. B* 63(2): 377-392.

KOSFELD, R. and J. LAURIDSEN, 2008, Factor analysis regression. *Statistical Papers* 49: 653-667.

LANSANGAN, J.R. and E. BARRIOS, 2009, Principal components analysis of nonstationary time series data, *Statistics and Computing* 19: 173-187.

LENG, C. and H. WANG, 2009, On general adaptive sparse principal component analysis, *Journal of Computational and Graphical Statistics* 19: 173-187.

MARX, B. and E. SMITH, 1990, Principal component estimation for generalized linear regression, *Biometrika* 77: 23-31.

ROUSSON, V. and T. GASSER, 2004, Simple component analysis. *Applied Statistics* 53(4): 539-555.

TIBSHIRANI, R., 1996, Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Soc. Ser. B* 58(1): 267-288.

VINES, S., 2000, Simple principal components, *Applied Statistics* 49(4): 441-451.

ZOU, H. and T. HASTIE, 2005, Regularization and variable selection vie the Elastic net, *Journal of the Royal Statistical Soc. Ser. B* 67(2): 301-320.

ZOU, H., T. HASTIE and R. TIBSHIRANI, 2006, Sparse principal component analysis, *Journal of Computational and Graphical Statistics* 15(2): 265-286.

Appendix 1Loadings of Principal Components and Sparse Principal Components

Country	Variable	Loadings of SPC			Loadings of PC		
		SPC1	SPC2	SPC3	PC1	PC2	PC3
Indonesia	X_1	0	-0.47885	0	-0.01331	-0.50772	-0.01676
Indonesia	X_2	0	0.431075	0	0.368977	0.349509	-0.03786
Indonesia	X_3	0	0.559054	0	0.065038	0.545049	-0.06826
Indonesia	X_4	-0.46994	0	0	-0.4095	0.167894	0.501166
Indonesia	X_5	0	0	-1	-0.26896	0.162897	-0.74424
Indonesia	X_6	0.468516	0	0	0.454584	-0.01462	0.246501
Indonesia	X_7	0.562715	0	0	0.457652	-0.12847	0.087629
Indonesia	X_8	-0.49296	0	0	-0.43737	0.107254	0.324975
Indonesia	X_9	0	0.521862	0	0.10971	0.48994	0.120485
Malaysia	X_1	-0.47914	0	0	-0.43336	0.230426	-0.1346
Malaysia	X_2	0	0	0	0.21478	0.280563	-0.55336
Malaysia	X_3	0.584526	0	0	0.484191	-0.11373	0.182092
Malaysia	X_4	0	0.547691	0	0.212151	0.458223	0.441127
Malaysia	X_5	0	-0.82094	0	-0.02304	-0.65404	-0.09144
Malaysia	X_6	0.319056	0.143262	0	0.366782	0.321661	0.06257
Malaysia	X_7	0.41592	0	0	0.410367	-0.01591	-0.33881
Malaysia	X_8	0	0	1	-0.17723	-0.0127	0.522371
Malaysia	X_9	0.392392	-0.07463	0	0.389547	-0.3371	0.219049
Philippines	X_1	0.574885	0	0	0.453093	-0.05864	0.049089
Philippines	X_2	0.414215	0	0	0.385166	-0.15376	0.441013
Philippines	X_3	0.553056	0	0	0.445888	0.135822	0.154308
Philippines	X_4	0.210102	0.109767	0.247695	0.30937	0.327198	0.443377
Philippines	X_5	0	0	0	-0.23057	-0.18703	0.435428
Philippines	X_6	0.222073	-0.224	0	0.310635	-0.39131	-0.08869
Philippines	X_7	0.258089	0	-0.96884	0.328222	0.13449	-0.54423
Philippines	X_8	-0.17886	0.181415	0	-0.28796	0.402784	0.271997
Philippines	X_9	0	0.951244	0	0.107325	0.692089	-0.12278
Thailand	X_1	-0.31439	0.231765	0	-0.34054	0.424593	-0.28151
Thailand	X_2	0.46225	0	0	0.411762	0.078009	0.148661
Thailand	X_3	0.48984	0	0	0.421521	-0.16259	0.153601
Thailand	X_4	0	0.48101	0	0.206134	0.522163	0.367531
Thailand	X_5	-0.46343	0	0	-0.40793	-0.01004	0.307746
Thailand	X_6	0.253101	0	0.377781	0.321917	0.156391	0.368065
Thailand	X_7	0.351775	0	0	0.357461	-0.1481	-0.2237
Thailand	X_8	-0.21205	0	0.925895	-0.29181	0.003379	0.642233
Thailand	X_9	0	-0.84553	0	-0.10649	-0.68414	0.217469