Simultaneous Dimension Reduction and Prediction Optimization: Method and Application to High Dimensional Data

Joseph Ryan Lansangan*
University of the Philippines, Quezon City, Philippines – jglansangan@up.edu.ph

Erniel Barrios
University of the Philippines, Quezon City, Philippines – ebbarrios@up.edu.ph

Abstract

A constrained optimization method is developed to address estimation problems when dealing with high dimensional input in regression. The method simultaneously considers dimension reduction (among the input variables) while maintaining relatively high predictive ability (in the fitted target variable). The method uses an alternating and iterative solution via soft thresholding, and yields fitted models with sparse regression coefficients. Results of the simulation studies show that the method may outperform other constrained regression methods in terms of predictive ability and selection of input variables. That is, the method selects a smaller set of input variables that both captures the dimensionality of the inputs (high retained variability from the inputs) and gives the most predictive model for the target variable (lowest squared prediction error). The method is applied to model cross-country quality of life (with emphasis on mortality). Environment, lifestyle, health care, health status, health policy, and morbidity indicators are considered as inputs. Results from the empirical data show that quality of life may be explained primarily by health condition of women, welfare of children, and government spending on health.

Keywords: high dimensional data; regression; variable selection; quality of life.

1. Introduction

Analyses of high dimensional data are abundant in many applications such as in genomics, bioinformatics, agriculture, astronomy, and business intelligence. However, the literature has been dominated by the assumption of smaller number of features (p) relative to the number of observations (n). Similarly, asymptotic theories may not be helpful as it assumes n approaching ∞ while p is fixed. These lead to difficulties in dealing with data having $p>>n$, i.e., data with a relatively larger number of features compared to the number of observations.

In the classical regression framework, it is assumed that $p \leq n$; otherwise, the design matrix is singular and therefore the parameters in the regression model are not uniquely estimable. As a solution, variables are dropped but at the expense of bias for the regression coefficients of the remaining variables. In time series data of indicators, e.g., those benefiting from macroeconomic policies, natural drifting of the variables as well as non-stationarity are expected resulting to ill-conditioning problem. Such problems can be mitigated commonly through using the growth rate (differencing) of the indicators instead of the original levels. Differencing, however, results to an alteration of the dependence structure in the data, thereby eliminating the effect of some important random shocks and possibly contaminating the relationship being investigated.

An alternative approach in modeling high dimensional data for purposes of dimension reduction and variable selection under a regression modeling framework is presented. The method provides a strategy for modeling high-order covariates and outputs in a regression-type problem, i.e., modeling multicollinear data (cross-sectional data) or nonstationary data (time series and/or spatio-temporal data). It further identifies key predictors among a large number of predictors (or equivalently, for a

small number of observations). Simulation studies and application to an empirical data are made to evaluate as well as demonstrate the use of the developed method.

2. Modeling High Dimensional Data

High dimensionality refers to either one of the following: the number of observations n is very large relative to the number of predictors p, or the number of predictors p is very large compared to the number of observations n. The higher the dimensionality (of p or of n), the more difficult it is to identify the best "representation" of the data which, in a general sense, is a "curse of dimensionality" (Bellman, 1957). Simultaneous testing of the p predictors becomes more and more inefficient as p gets larger and larger. Variable selection (and equivalently, observation clustering) becomes more and more difficult as p (or n) gets larger and larger. In regression modeling with very large p, identification of the most important set of predictors becomes more difficult since presence of too many predictors masks the importance of some, thereby leading to more potential problems of incorrect postulated model. The usefulness and interpretability of the identified "important" set of predictors may be problematic, or at least, doubtful.

Solutions to multicollinearity and singularity range from transformations, to variable selection methods, to modified estimation procedures; and issues were raised in using such solutions. Marx and Smith (1990) cite variable deletion, Stein estimation, ridge regression and principal component regression (PCR) as solutions to multicollinearity, and suggest that one-step adjustment to the maximum likelihood estimation (MLE) of the beta coefficients for ill-conditioned information matrix yields coefficients that are asymptotically biased. Garson (2012) noted that stepwise regression methods are even more affected by multicollinearity than regular methods since additional information is difficult to attain with the deletion of "unimportant" variables, and as such, the process of deletion sometimes introduces subjectivity. Garson (2012) further suggests that power and nonlinear transformations may cause over-fitting or even increase the level of multicollinearity.

The use of principal components in regression (principal component regression or PCR), is proposed as a possible solution to the problem of multicollinearity (Jolliffe, 2002). PCR, as noted by Kosfeld and Lauridsen (2008), may work for cases with highly multicollinear independent variables since PCR reduces the variability of the regression coefficients estimates but at the expense of its bias. Fewer components may be used (based on eigenvalues and/or tests of significance), but with discrepancy in the amount of information between the raw individual predictors and the PCs. Foucart (2000) also notes that deleting components that are not significant may introduce bias to the least squares estimates of the remaining coefficients and may reduce the unbiased residual variance.

Focusing on the variance inflation problem caused by multicollinearity, shrinkage estimators are considered as solutions (see for example Filzmoser and Croux, 2002; Goldenshluger and Tsybakov, 2001; Klinger, 2001; Zou and Hastie, 2005). Similarly, it is common to implement a regularization technique, i.e., by introducing a penalty on the optimization framework. One of the most commonly used regularization techniques is the ridge regression (Hoerl and Kennard, 1970), which introduce bias on the parameters to stabilize the variance. Ridge regression, however, depends on the choice of the ridge parameter which tends to be subjective in nature. The propagation of bias in the parameter estimates also complicates the interpretation of the relative contribution of the individual determinants toward the dependent variable. Also, variances of the regression coefficients remain to be potentially large even with the introduction of the ℓ_2 norm penalty in ridge regression modeling. Thus as a new direction, Tibshirani (1996) introduces a regularized method, called the least absolute shrinkage and selection operator (LASSO), which considers a penalty under the ℓ_1 norm. The method generally leads to sparse solutions, i.e., those "less significant" parameters tend to be nearly-zero or exactly zero.

Evidently, sparsity is one of the key solutions to the ease of interpretation of (linear) combinations of variables. For instance, Chipman and Gu (2005) address the interpretability problem by considering homogeneity constraints and sparsity constraints. Zou and Hastie (2005) introduce the elastic net (EN) penalty as a modification of the LASSO by Tibshirani (1996). Klinger (2001) uses penalized likelihood estimators for a large number of coefficients to extend soft thresholding and LASSO methods on generalized linear models. Zou et al. (2006) developed sparse principal component analysis (SPCA) and the resulting sparse PCs can be used in regression analysis, the method is subsequently explored in this paper.

3. Dimension Reduction and Variable Selection – The LaNS Framework

Zou et al. (2006) use the LASSO and ridge-type constraints to principal components extraction. The extraction is formulated as a regression problem and optimization results to components with sparse loadings. The sparse principal component analysis (SPCA) uses both ℓ_1 and ℓ_2 penalties to come up with the sparse principal components (SPCs). Optimization is done through a regression-type criterion to derive the SPCs in two stages. Sparse principal component regression (SPCR) uses SPCs as predictors in the model. With the sparsity that comes in under this two-step procedure (SPCA first on the data matrix $\underline{X}$, then regression on the response $\underline{y}$ using the derived SPCs), SPCR provides a solution to multicollinearity and to the issue on components selection. Although there is little known properties and advantages of using SPCR over PCR, SPCR may be the more logical option for cases when $p \gg n$.

The developed framework on the other hand combines both the construction of SPCs and the estimation of regression parameters as a one-time optimization problem. That is, the framework considers a simultaneous approach for addressing issues on high dimensionality and/or multicollinearity in the regression problem while optimizing captured information among the original input variables and minimizing the error on prediction of the dependent variable using the sparse components.

Let $\underline{x}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)^T \in \mathbb{R}^p$ be the p-dimensional response from the i^{th} subject, where $i = 1, 2, \dots, n$. Equivalently, let $\underline{X}_j = \left(x_{1j}, x_{2j}, \dots, x_{nj}\right)^T \in \mathbb{R}^n$ be the n-dimensional observation on the j^{th} variable, where $j = 1, 2, \dots, p$. Thus, $\underline{X} = \left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n\right)^T = \left(\underline{X}_1, \underline{X}_2, \dots, \underline{X}_p\right)$ is the $n \times p$ matrix of observed values for the p (original) variables over the n subjects, $\underline{X}_j's$ are assumed to be centered. The singular value decomposition (SVD) of $\underline{X}$ is $\underline{X} = \underline{U}\underline{S}\underline{V}^T$, where $\underline{U}$ is $n \times n$ and $\underline{V}$ is $p \times p$ for which $\underline{U}^T\underline{U} = \underline{I}_n$ and $\underline{V}^T\underline{V} = \underline{I}_p$, and $\underline{S}$ is $n \times p$ rectangular diagonal matrix. Thus, an approximation of $\underline{X}$ is given by $\hat{\underline{X}} = \underline{U}_q\underline{S}_q\underline{V}_q^T$, where $\underline{U}_q$ and $\underline{V}_q$ are the first q columns of $\underline{U}$ and $\underline{V}$, respectively, and $\underline{S}_q$ is the $q \times q$ diagonal matrix of the singular values in $\underline{S}$. $\hat{\underline{X}}$ becomes a low-rank approximation of $\underline{X}$ (Eckart and Young, 1936). Let $\underline{A}$ and $\underline{B}$ be $p \times k$ matrices, where $k < p$ and such that $\underline{A}^T\underline{A} = \underline{I}_k$, then a generalized solution $\hat{\underline{X}}$ for an approximation of $\underline{X}$ can be based on the minimization of the function $f\left(\underline{A}, \underline{B}\right) = \left\|\underline{X} - \underline{X}\underline{B}\underline{A}^T\right\|_F^2$, where $\|\cdot\|_F^2$ is the squared Frobenius norm, and imposing orthonormality of $\underline{A}$ for identifiability and restrictions on $\underline{B}$ to adjust for component loadings. In the case that $\underline{B} = \underline{A}$, the solution for the optimization problem is the set of first k PCs derived from the PCA of $\underline{X}$ (Zou et al., 2006).

Let λ and $\underline{\lambda}_1 = \left(\lambda_{1,1}, \lambda_{1,2}, \dots, \lambda_{1,k}\right)$ be some constants, then the SPCA criterion (Zou et al., 2006) minimizes $f_X\left(\underline{A}, \underline{B}, \lambda, \underline{\lambda}_1\right) = \left\|\underline{X} - \underline{X}\underline{B}\underline{A}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1$ subject to $\underline{A}^T\underline{A} = \underline{I}_k$, where $\underline{B} = \left[\underline{b}_1, \underline{b}_2, \dots, \underline{b}_k\right]$. Now, consider regressing $\underline{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$ on the transformed $\underline{X}$, i.e., on the set of k (with $k \leq p$) linear transformations of $\underline{X}\underline{B}$. Under the regular (no-intercept) regression

problem with $\underline{\beta}$ as the $k \times 1$ vector of (regression) parameters, the objective function is to minimize, subject to $\underline{A}^T\underline{A} = \underline{I}_k$,

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}, \lambda, \underline{\lambda}_1\right) = \left\|\underline{y} - \underline{XB}\underline{\beta}\right\|^2 + \left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 . \quad (1)$$

Optimization of equation (1) simultaneously minimizes the loss due to dimension-reduction in $\underline{X}$ and on using a fitted regression for $\underline{y}$. If an intercept is included, and with $\underline{\beta}^* = \left[\beta_0, \underline{\beta}^T\right]^T$, then the optimization problem becomes minimizing, subject to $\underline{A}^T\underline{A} = \underline{I}_k$,

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}^*, \lambda, \underline{\lambda}_1\right) = \left\|\underline{y} - [1\ \underline{XB}]\underline{\beta}^*\right\|^2 + \left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 . \quad (2)$$

Suppose the optimization problem is constrained further on the loss due to dimension reduction of $\underline{X}$ and on the loss due to regression for $\underline{y}$. Then the generalized optimization problem becomes, given the tuning parameters $\lambda, \underline{\lambda}_1$ and $\underline{m} = (m_1, m_2)$, finding the values $\hat{\underline{A}}$, $\hat{\underline{B}}$ and $\hat{\underline{\beta}}^*$ for which

$$\left(\hat{\underline{A}}, \hat{\underline{B}}, \hat{\underline{\beta}}^*\right) = \underset{\underline{A},\underline{B},\underline{\beta}^*}{\operatorname{argmin}}\left\{m_1\left\|\underline{y} - [1\ \underline{XB}]\underline{\beta}^*\right\|^2 + m_2\left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1\right\}. \quad (3)$$

The terms in the penalized optimization in equation (3) are collectively considered as a "dimension reduction and variable selection penalty." Using the transformed independent variables $\underline{XB}$, this penalty on the regression of $\underline{y}$ yields a vector of coefficients $\underline{\theta} = \underline{B}\underline{\beta}$ of the (untransformed) individual $\underline{X}'s$, which then gives a linear combination of the $\underline{X}'s$ with possibly non-replete (or sparse) coefficients. That is, the penalty translates to a ***L**inear **a**nd **N**on-replete **S**election (***LaNS**) of the independent variables. Hereafter, the optimization problem in equation (3) is referred to as the *LaNS criterion* or the *LaNS optimization*. Accordingly, the equivalent bounds in the equation are referred to as the *LaNS penalty*, and solutions and models under this framework are labelled as *LaNS*. An alternating solution for $\underline{A}$, $\underline{B}$, and $\underline{\beta}^*$, given the values of $\underline{m} = (m_1, m_2)$, λ, and $\underline{\lambda}_1$, is used for the minimization of the LaNS criterion via the *LaNS algorithm.*

4. Equivalence and Simulations

The performance of LaNS is evaluated through simulation studies. Assume that the data come from 3 latent factors V_1, V_2 and V_3. V_1 and V_2 are Normal and independent and $V_3 = f(V_1, V_2) + \omega$ where ω is Normal. The latent factor V_1 gives the most information (having high variability), closely followed by V_2, then by V_3. The independent variables are derived as $X_j = V_k + \varepsilon^{(j)}$ where the $\varepsilon^{(j)}{}'s$ are independent and $j = 1, 2, \ldots, 1000$. The dependent variable Y is then computed from the $X_j{}'s$, with the beta-coefficients specified to control for the relative contributions of the independent variables $X_j{}'s$ to the dependent variable Y. Similarly, the relative contributions of the latent factors V_1, V_2 and V_3 to Y are controlled. For the high dimensional case (HD), the number of variables is set at 1,000, thus, all the variables $X_j{}'s$ are included in the computation of Y. For the non-high dimensional case (NHD), the number of variables is set at 40, so that only the first 40 $X_j{}'s$ are considered. The scenarios (for both NHD and HD) are formulated so that the independent variables most predictive of the dependent variable are relatively few, i.e., those independent variables are derived either from V_1 or V_2.

The different LaNS solutions are compared to various regression methods that address multicollinearity or mitigate the issues associated with high dimensional inputs. For the NHD, the fitted full model from LaNS is compared to those of OLS, PCR, and PCovR; for models with sparse coefficients, LaNS is compared to SPCR, regression with LASSO, and regression with EN; and the OLS regression model using selected variables from LaNS are compared to the OLS regression models using the corresponding selected variables from SPCR, LASSO or EN. For the high dimensional cases (HDs), on the other hand, LaNS is compared to PCR, PCovR, SPCR, LASSO, EN, and whenever possible, to the OLS of the corresponding reduced models.

The models are assessed based on their predictive ability through the sum of squared prediction error (SSPE). SSPE is equivalent to the residual sum of squares in a regression fit (Chatterjee and Hadi, 2006; Draper and Smith, 1998). Thus, SSPE measures how close the fitted values are to the original values, the lower the SSPE, the higher the predictive ability of the fitted model. Aside from prediction error, a BIC-type measure is also used to compare the different methods, following that of Schwarz (1978) and Zou et al. (2007). The BIC penalizes the measure of predictive ability of the model using the number of nonzero coefficients as well as number of observations. Thus, relative to BIC, the most suitable model is the most parsimonious, i.e., the model must have the smallest prediction error at the fewest number of predictors selected as possible, taking into consideration the inherent variability in the dependent variable. While SSPE is used to compare models with same number of predictors, BIC is used to compare competing models with varying numbers of predictors.

For both NHD and HD, the fitted model from PCR have lower predictive ability even when all independent variables are used in the model. Similar to PCR, LaNS generates a fitted model with non-zero coefficients for all the independent variables but with greater predictive ability (SSPE and BIC are smallest). PCovR dominates PCR, with SSPEs and BICs for PCovR lower than that of PCR. This may suggest an advantage in predictive ability of a one-step approach (PCovR) over a two-step approach (PCR) for dimension reduction and variable selection. Expectedly, PCovR(0.15) improves on predictive ability compared to PCovR(0.85) or PCovR(0.50).

LaNS offers sparse solutions for which BIC values remain lower, and for which the SSPEs are as good as that of PCovR(0.50) or PCoVR(0.15). LaNS selects independent variables coming from all three latent factors. Unlike EN(100) which include all 10 independent variables from the first latent factor, EN(.01) or LASSO which includes all but 1 from the first latent factor, and SPCR which identifies variables only from the first two latent factors, LaNS identifies independent variables from all three latent factors and gives the fitted model with highest predictive ability.

Comparing the selected variables using different methods, and implementing OLS using only the selected variables as predictors, those variables identified by LaNS give better prediction than those identified by any of SPCR, LASSO, EN(0.01), or EN(100). Similarly, LaNS provides a smaller set of predictors that already represents the entire set of independent variables and at the same time best explains the dependent variable. If identification of a smaller set is of interest, then LaNS appears to be a better option than LASSO – although both give relatively the same set of independent variables, those identified by LaNS4 has a slightly better predictive ability than those identified by LASSO.

5. Application to Real Data

World Health Organization (WHO) provides data on mortality and related indicators for years around 2010 (census year for most countries). A quality of life index (QoLI) was constructed based on morality. A total of 106 variables from environment, lifestyle, health care, health status, health policy, and morbidity indicators are then used to model QoLI. The LaNS algorithm was used to come up with a final model.

The final model accounts for 71% of the total variation in QoLI, with identified determinants as follows: outdoor air pollution, UV radiation, consumption on alcohol, improved sanitation, female blood pressure, immunization coverage among 1-year-olds, and expenditure on health. The standardized coefficients suggest that the quality of life is primarily explained by health condition of women (measured by blood pressure), welfare of children (measured by immunization coverage), and the government spending on health.

6. Conclusions

The LaNS procedure estimates a model that is sparse while it also exhibits optimal predictive ability, addressing multicollinearity issues and/or ill-conditioning in regression analysis with high dimensional predictors. Dimension reduction is implemented such that prediction error is minimized, thus, the selected variables (with non-zero estimates of regression coefficients) become the "best" predictors for the dependent variable, i.e., the fitted model is the most optimal for both the dimensionality of the inputs and the prediction of the dependent variable. The LaNS procedure is capable of fitting models with independent variables potentially coming from different latent factors. For both $n>p$ and $p>>n$, the fitted LaNS models with sparse regression coefficients capture "representatives" from the different latent factors, as evident from the simulations.

References

Bellman, E., (1957). Dynamic programming. Princeton University Press

Chatterjee, S. and A. Hadi (2006). Regression Analysis by Example, 4th ed., Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., Hoboken, New Jersey

Chipman, H. and G. Gu (2005). Interpretable Dimension Reduction, Journal of Applied Statistics, Vol.32, No. 9, pp. 969-987

Draper, N. and H. Smith (1998). Applied Regression Analysis, 3rd ed., Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., Hoboken, New Jersey

Eckart, C. and G. Young (1936). The approximation of one matrix by another of lower rank. Psychometrika, Vol. 1, No. 3, pp. 211-218

Filzmoser, P. and C. Croux (2002). A projection algorithm for regression with collinearity. In K. Jajuga, A. Sokolowski, and H.-H. Bock, editors, Classification, Clustering, and Data Analysis, Springer-Verlag, Berlin, pp. 227-234

Foucart, T. (2000). A decision rule for discarding principal components in regression, Journal of Statistical Planning and Inference, Vol. 89, No. 1, pp. 187-195

Garson, G.D. (2012). Multiple Regression. Asheboro, NC: Statistical Associates Publishers

Goldenshluger, A. and A. Tsybakov (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters, The Annals of Statistics, Vol. 29, No. 6, pp. 1601- 1619

Hoerl A.E. and R.W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics Vol. 12, pp. 55–82

Jolliffe, I. (2002). Principal Component Analysis, 2nd ed. (New York: Springer-Verlag)

Kosfeld, R. and J. Lauridsen (2008). Factor analysis regression, Statistical Papers, Vol. 49, No. 4, pp. 653-667

Klinger, A. (2001). Inference in High Dimensional Generalized Linear Models Based on Soft Thresholding, Journal of the Royal Statistical Society, Vol. 63, No. 2, pp. 377-392

Marx, D. and P. Smith (1990). Principal Component Estimation for Generalized Linear Regression, Biometrica, Vol 77, No. 1, March 1990, pp. 23-31

Schwarz, G. (1978). Estimating the Dimension of a Model, The Annals of Statistics, 6(2):461-464.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the LASSO, Journal of the Royal Statistical Society, Ser. B, 58(1):267-288.

Zou, H. and T. Hastie (2005), Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society Ser. B, 67(2):301-320.

Zou, H., Hastie, T. and R. Tibshirani (2006), Sparse Principal Component Analysis, Journal of Computational and Graphical Statistics, Vol. 15, No. 2, pp. 265-286

Zou, H., Hastie, T. and R. Tibshirani (2007), On the "Degrees of Freedom" of the LASSO, The Annals of Statistics, Vol. 35, No. 5, pp. 2173-2192