SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Simultaneous Dimension Reduction and Variable Selection in Modeling High Dimensional Data

by

Joseph Ryan G. Lansangan
Erniel B. Barrios

UPSS Working Paper No. 2016-09
November 2016

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

Abstract

High dimensional predictors in regression analysis is often associated with multicollinearity along with other estimation problems. These problems can be mitigated through a constrained optimization method that simultaneously induces dimension reduction and variable selection that also maintains a high level of predictive ability of the fitted model. Simulation studies show that the method may outperform sparse principal component regression, least absolute shrinkage and selection operator, and elastic net procedures in terms of predictive ability and optimal selection of inputs. Furthermore, the method yields reduced models with smaller prediction errors than the estimated full models from the principal component regression or the principal covariance regression.

Keywords: high dimensionality, regression modeling, dimension reduction, variable selection, latent factors, sparsity, soft thresholding, sparse principal component analysis

1. Introduction

Large volumes of data that may come from different sources are available from genetic sequences, multi-point and multi-feature image data, transactional details, business processes, and even marketing campaigns. Analyses of these data are crucial in a wide spectrum of applications such as in genomics, bioinformatics, agriculture, astronomy, and business intelligence. The data are processed and summarized into useful information for strategic decision-making. However, the literature has been dominated by the assumption of smaller number of features (p) relative to the number of observations (n). Asymptotic theories, therefore, 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 regression analysis, multicollinearity may result to ill-conditioning and/or near-singularity of the associated design matrix, resulting to unstable estimates (inflated standard errors). Similarly, classical regression framework assumes $p \leq n$; otherwise, the design matrix is singular and therefore the parameters in the regression model are not uniquely estimable. Non-orthogonality of the predictors in a linear model causes the ill-conditioning problem, and as a solution, those duplicating 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 is expected resulting to similar ill-conditioning problem. For non-stationary time series, the ill-conditioning problem can be mitigated through the use of growth rate (differencing) of the indicators instead of the original levels. Differencing, however, results to an alteration of the dependence structure since it generally filters low frequencies and preserves high frequencies 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).

2. Modeling High Dimensional Data

In high dimensional data where the number of predictors p is very large compared to the number of observations n, the best "representation" of the data is usually difficult to achieve. Simultaneous testing of the p predictors becomes more and more inefficient as p gets larger. Variable selection (and equivalently, observation clustering) becomes more difficult as p (or n) gets larger. In regression modeling with very large p, the identification of the most important set of predictors becomes challenging since presence of too many predictors masks the importance of some, thereby leading to more potential problems of model misspecification. The usefulness and interpretability of the identified "important" set of predictors may be problematic, or at least, doubtful.

Given $\underline{y}_{nx1}$, a vector of observations from a dependent variable and $\underline{X}_{n\times p} = \left[\underline{x}_1, \ldots, \underline{x}_n\right]^T$, a matrix of observations on p variables for the n subjects. The hypothesized model takes the form $\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$, with $\underline{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)^T$ and $\underline{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)^T$. For $i = 1, 2, \ldots, n$, assume that the error terms ε_i are independent and each follows a Gaussian distribution with mean zero and constant variance $\sigma^2 > 0$. The ordinary least squares (OLS) regression estimator of $\underline{\beta}$ is $\underline{\hat{\beta}} = \left(\underline{X}^T\underline{X}\right)^{-1}\underline{X}^T\underline{y}$ is optimal (Gauss-Markov Theorem) provided that $n > p$.

When $p \gg n$, the estimator $\underline{\hat{\beta}}$ is not unique since high dimensionality of the data matrix leads to the singularity of the Gram matrix $\underline{X}^T\underline{X}$ (Chatterjee and Hadi, 2006; Draper and Smith, 1998). Similarly, the estimator $\underline{\hat{\beta}}$ is unstable, i.e., the estimators for $\underline{\beta}$ may not be reliable since the standard errors are also based on the Gram matrix (Draper and Smith, 1998). Thus, tests and confidence bounds that use the standard errors and the estimated variance-covariance matrix of the error terms (which is also based on the Gram matrix) are invalid. Even when $p < n$ but there are high correlations among the independent variables, tests and confidence bounds based on the ill-conditioned Gram matrix $\underline{X}^T\underline{X}$ are also invalid (Draper and Smith, 1998). In general, the OLS estimator $\underline{\hat{\beta}}$ are no longer optimal in the presence of multicollinearity and/or when $p \gg n$.

Solutions to multicollinearity and singularity range from transformations, to variable selection or stepwise regression methods, to modified estimation procedures; and issues were raised in using such solutions. However, Garson (2012) suggests that power and nonlinear transformations may cause over-fitting or even increase the level of multicollinearity. Garson (2012) also 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.

The use of principal components in regression (principal component regression or PCR), is proposed as a possible solution to the problem of multicollinearity (Jolliffe, 1998). 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 in modeling, 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 lead to biased residual variance estimates. Foucart (2000) proposed to discard principal components based on partial correlation coefficients aside from tests of significance (of the components in regression) and magnitude of eigenvalues (of the independent variables), while Hwang

and Nettleton (2003) provide an alternative approach of selecting a subset of components in PCR that minimizes MSE of the beta-coefficients.

On the other hand, De Jong and Kiers (1992) introduce the principal covariates regression (PCovR) which simultaneously minimizes the least squares regression residuals and the transformation residuals on the independent variables. PCovR is viewed as a one-step approach to PCR. Similarly, George and Oman (1996) proposed a multiple-shrinkage estimator on the regression coefficients to overcome the influence of multicollinearity on PCR. In the multivariate regression framework, Izenman (1975) and Reinsel et al. (1998) discussed applications of reduced-rank regression (RRR), wherein a restriction on the rank of the regression coefficient matrix is considered.

Focusing on the variance inflation problem caused by multicollinearity, shrinkage estimators and regularization techniques are considered as solutions (see for example Filzmoser and Croux, 2002; Goldenshluger and Tsybakov, 2001; Klinger, 2001; Zou and Hastie, 2005). Constraints are added in the least squares objective function to produce non-singular design matrix to alleviate variance inflation. Similarly, a penalty on the optimization framework is introduced. The gain in precision is necessarily compensated by the propagation of bias in the parameter estimates. This, however, complicates the interpretation of the relative contribution of the individual determinants toward the dependent variable.

One of the most commonly used regularization techniques is ridge regression (Hoerl and Kennard, 1970), which introduce bias on the $\underline{\beta}$ parameter estimates to stabilize the variance. Ridge regression however depends on the choice of the ridge parameter which tends to be subjective in nature. Accordingly, McDonald and Galarneau (1975) suggest methods of specifying the ridge parameter, which are essentially based on the variance component, the correlations among the inputs, and the regression coefficients. Lee (1987) provides different methods to optimize the choice of the ridge parameter.

Variances of the regression coefficients, however, 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. Recently, Candes and Tao (2007) consider a penalty, called the Dantzig Selector, which is similar to that of the ℓ_1 norm. This selector is well-aligned with sparsity considerations – identifying which parameters are "truly" non-zero. As a modification to RRR, Chen and Huang (2012) proposed a method, called sparse reduced-rank regression (SRRR), which introduces sparsity constraint on the RRR estimation via a group-LASSO type penalty.

Sparsity therefore, considers discarding unimportant variables and leaving a relatively smaller-spaced and more informative set of predictors. Finding sufficient data transformation that effectively reduces the dimension of the data without significant loss of information remains to be essential in achieving sparsity. As such, Cook (2007) identifies sufficient reduction definitions, depending mainly on the conditional distributions. Other methods to achieve sparsity in both the non-linear and the non-parametric models are present in the literature – some of which use general additive models and Bayesian approaches, see Ravikumar et al. (2007) and Chipman et al. (2010) for example.

Evidently, sparsity is associated with dimensionality reduction – with sparsity as 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. The extension leads to an adaptive selection of model terms without substantial variance inflation. Zou et al. (2006) developed sparse principal component analysis (SPCA) and the resulting sparse PCs can be used in regression analysis, i.e. sparse principal component regression (SPCR), and this is subsequently explored in this paper.

3. Dimension Reduction and Variable Selection

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) criterion, with $\underline{X}_{n\times p} = \left[\underline{x}_1, \ldots, \underline{x}_n\right]^T$, $\underline{A}_{p\times k} = \left[\underline{\alpha}_1, \ldots, \underline{\alpha}_k\right]$ and $\underline{B}_{p\times k} = \left[\underline{b}_1, \ldots, b_k\right]$, is given by

$$\left(\hat{\underline{A}}, \hat{\underline{B}}\right) = \underset{\underline{A},\underline{B}}{\operatorname{argmin}} \left\{\sum_{i=1}^{n} \left\|\underline{x}_i - \underline{A}\underline{B}^T\right\|^2 + \lambda \sum_{j=1}^{k} \left\|\underline{b}_j\right\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\|\underline{b}_j\right\|_1\right\} \text{ with } \underline{A}^T\underline{A} = \underline{I}_k \qquad (1)$$

where the $\underline{x}_i$'s are centered data vectors (with respect to the j^{th} variable, $j = 1, 2, \ldots, p$) for each observation i, λ and $\lambda_{1,j}$ are penalizing constants chosen to ensure existence of solution and convergence of the computational algorithm. Here, $\left\|\cdot\right\|_1$ is the ℓ_1 norm and $\left\|\cdot\right\|$ is the Euclidean norm, i.e., $\left\|\underline{w}\right\|_1 = \sum\left|w_j\right|$ and $\left\|\underline{w}\right\| = \sqrt{\sum w_j^2}$. Optimization is done through a regression-type criterion to derive SPCs in two stages: (1) perform ordinary PCA, and (2) find sparse approximations of the first k vector of loadings of the PCs using the "naive elastic net" estimation (Zou and Hastie, 2005), a penalized least squares method to overshrink regression parameters (i.e., solutions to Eq. 1) and corrects the grouping effect (i.e., strongly correlated predictors tend to be in or out the model together). Unlike PCA, the solution (and its corresponding algorithm) yield components that are correlated and loadings that are not orthogonal (Zou et al., 2006). Thus, the total predicted variance is not just the sum of the predicted variances of the SPCs, it also account for the correlations of SPCs. To generate SPCs via an alternating method, Zou et al. (2006) developed an algorithm that uses a heuristic/numerical approach. The algorithm also implements the QR-decomposition to estimate the adjusted variances explained by the SPCs.

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 computed 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 $p \gg n$.

SPCR uses the first few SPCs as inputs in the regression problem. In contrast, we developed a framework that combines both the construction of SPCs and the estimation of regression parameters as a one-time optimization problem. Thus, 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 $\underline{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T \in \mathbb{R}^p$ be the p-dimensional realization from the i^{th} subject, where $i = 1, 2, \ldots, n$. Equivalently, let $\underline{X}_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)^T \in \mathbb{R}^n$ 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. Recall that the singular value decomposition (SVD) of $\underline{X}$ is $\underline{X} = \underline{USV}^T$, 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 $\underline{S}$ (i.e., the first q diagonal entries of $\underline{S}$ arranged in descending order). With $rank\left(\hat{\underline{X}}\right) = q$ and $q < p$, $\hat{\underline{X}}$ 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 $f\left(\underline{A}, \underline{B}\right) = \left\|\underline{X} - \underline{XBA}^T\right\|_F^2$, where $\|\cdot\|_F^2$ is the squared Frobenius norm, and imposing the following constraints: orthonormality of $\underline{A}$ for identifiability; and restrictions on $\underline{B}$ to adjust component loadings. Note that $\underline{B}$ represents the component loadings which define the transformed (linear) combinations of $\underline{X}$, and $\underline{XB}$ having a reduced dimension $n \times k$. 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 (specifically, the tuning parameters). Then the SPCA criterion (Zou et al, 2006) minimizes

$$f_X\left(\underline{A}, \underline{B}, \lambda, \underline{\lambda}_1\right) = \left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 \text{ subject to } \underline{A}^T\underline{A} = \underline{I}_k, \quad (2)$$

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 $\underline{XB}$. Under the regular (no-intercept) regression problem with $\underline{\beta}$ as the $k \times 1$ vector of (regression) parameters, the ordinary least squares (OLS) approach to estimating $\underline{\beta}$ is equivalent to minimizing the squared norm

$$f_Y\left(\underline{\beta}\right) = \left\|\underline{y} - \underline{XB\beta}\right\|^2 . \quad (3)$$

Combining equations (2) and (3), the objective function is to minimize, subject to $\underline{A}^T\underline{A} = \underline{I}_k$,

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}, \lambda, \underline{\lambda}_1\right) = \left\|\underline{y} - \underline{XB\beta}\right\|^2 + \left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 . \quad (4)$$

Optimization of equation (4) 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 $\underline{A}^T\underline{A} = \underline{I}_k$,

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}^*, \lambda, \underline{\lambda}_1\right) = \left\|\underline{y} - \left[\underline{1}\ \underline{XB}\right]\underline{\beta}^*\right\|^2 + \left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 . \quad (5)$$

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 minimizing, subject to $\underline{A}^T\underline{A} = \underline{I}_k$,

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}^*, \lambda, \underline{\lambda}_1, \underline{m}\right) = m_1\left\|\underline{y} - \left[\underline{1}\ \underline{XB}\right]\underline{\beta}^*\right\|^2 + m_2\left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1 ,$$

i.e., given the tuning parameters $\lambda, \underline{\lambda}_1$ and $\underline{m} = (m_1, m_2)$, find the values $\hat{\underline{A}}$, $\hat{\underline{B}}$ and $\hat{\underline{\beta}}^*$ for which

$$\left(\hat{\underline{A}}, \hat{\underline{B}}, \hat{\underline{\beta}}^*\right) = \underset{\underline{A}, \underline{B}, \underline{\beta}^*}{\operatorname{argmin}}\left\{m_1\left\|\underline{y} - \left[\underline{1}\ \underline{XB}\right]\underline{\beta}^*\right\|^2 + m_2\left\|\underline{X} - \underline{XBA}^T\right\|_F^2 + \lambda\left\|\underline{B}^T\right\|_F^2 + \sum_{j=1}^k \lambda_{1,j}\left\|\underline{b}_j\right\|_1\right\} . \quad (6)$$

Note that the first two terms in equation (6) are equivalent to setting upper bounds (as some function of the tuning parameters $\underline{m}$) for the loss due to the use of predicted values for $\underline{y}$ and for the dimension reduction in $\underline{X}$, i.e., the closer the bound is to 0, the smaller the loss. In the same sense, the

larger the bound, the higher the tolerance level is for the loss due to prediction and/or dimension reduction. Thus, the tuning parameters may be set such that one is related to the other. Intuitively, m_1 and m_2 may be considered as weighting parameters that set the significance of either prediction in $\underline{y}$ or dimension reduction in $\underline{X}$. The remaining terms in equation (6) are similar to the SPCA criterion via the elastic net as used by Zou et al. (2006).

The terms in the penalized optimization in equation (6) 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. 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 (6) is referred to as the *LaNS criterion*. 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. Theorems 1 and 2 exhibit existence of $\underline{A}$, $\underline{B}$ and $\underline{\beta}^*$ in equation (6). To facilitate initialization of the SVD, we separate β_0 from the rest of the β's.

Theorem 1: *The constrained minimization of equation (6) has a solution for $\underline{A}$ when $\underline{B}$ and $\underline{\beta}^*$ are known, given by $\underline{\hat{A}} = \underline{W}\underline{Z}^T$, where $\underline{W}$ and $\underline{Z}$ are derived from the SVD of $\underline{X}^T\underline{XB}$, i.e., $\underline{X}^T\underline{XB} = \underline{W}\underline{E}\underline{Z}^T$. Also, when $\underline{A}$ and $\underline{B}$ are known, the constrained minimization of equation (6) has a solution for $\underline{\beta}^*$, given by $\hat{\beta}_0 = \overline{y}$ and $\underline{\hat{\beta}} = \left(\underline{B}^T\underline{X}^T\underline{XB}\right)^{-1}\underline{B}^T\underline{X}^T\underline{y}_C$, where $\underline{y}_C = \underline{y} - \overline{y}\underline{1}$.*

Proof:

Given $\underline{\beta}^*$ and $\underline{B}$, equation (6) reduces to minimizing $m_2\left\|\underline{X} - \underline{XBA}^T\right\|_F^2$ subject to $\underline{A}^T\underline{A} = \underline{I}_k$. From the Reduced Procrustes Rotation Theorem (RPRT) of Zou et al. (2006), if the SVD of $\underline{X}^T\underline{XB}$ is given by $\underline{X}^T\underline{XB} = \underline{W}\underline{E}\underline{Z}^T$, then for a fixed m_2, the solution for $\underline{A}$ is given by $\underline{\hat{A}} = \underline{W}\underline{Z}^T$.

If $\underline{A}$ and $\underline{B}$ are given, then the constrained optimization of equation (6) reduces to minimizing $m_1\left\|\underline{y} - [\underline{1}\ \underline{XB}]\underline{\beta}^*\right\|^2$. Given m_1, this is equivalent to minimizing $\left\|\underline{y} - [\underline{1}\ \underline{XB}]\underline{\beta}^*\right\|^2$. Getting the partial derivatives with respect to β_0 and $\underline{\beta}$, and equating the derivatives to zeros,

$$\frac{\partial}{\partial\beta_0} f_{X,Y}\left(\underline{\beta}^* \mid \underline{A}, \underline{B}, \lambda, \underline{\Lambda}\right) = \frac{\partial}{\partial\beta_0}\left[n\beta_0^{\,2} + 2\beta_0\underline{1}^T\left(\underline{X}\,\underline{B}\underline{\beta} - \underline{y}\right)\right]$$

$$\Rightarrow \quad \beta_0 = \frac{1}{n}\underline{1}^T\left(\underline{y} - \underline{X}\,\underline{B}\underline{\beta}\right)$$

$$\frac{\partial}{\partial\underline{\beta}} f_{X,Y}\left(\underline{\beta}^* \mid \underline{A}, \underline{B}, \lambda, \underline{\Lambda}\right) = \frac{\partial}{\partial\underline{\beta}}\left[-2tr\left(\underline{X}\,\underline{B}\underline{\beta}\underline{y}^T\right) + tr\left(\underline{X}\,\underline{B}\underline{\beta}\underline{\beta}^T\underline{B}^T\underline{X}^T\right) + 2\beta_0\underline{1}^T\underline{X}\,\underline{B}\underline{\beta}\right]$$

$$\Rightarrow \quad \underline{\hat{\beta}} = \left(\underline{B}^T\underline{X}^T\underline{X}\,\underline{B}\right)^{-1}\underline{B}^T\underline{X}^T\underline{y}_C$$

Thus,

$$\beta_0 = \frac{1}{n}\underline{1}^T\left(\underline{y} - \underline{X}\,\underline{B}\underline{\beta}\right)$$

$$\Rightarrow \quad \hat{\beta}_0 = \bar{y} \quad . \qquad \blacksquare$$

Theorem 2: *The constrained minimization of equation (6) has an iterative solution for* $\underline{B}$ *when* $\underline{A}$ *and* $\underline{\beta}^*$ *are known.*

Proof:

Given $\underline{A}$ and $\underline{\beta}^*$, equation (6) reduces to minimizing

$$f_{X,Y}\left(\underline{B} \mid \underline{A}, \underline{\beta}^*, \underline{m}, \lambda, \underline{\lambda}_1\right) = -2m_1 tr\left(\underline{X}\,\underline{B}\underline{\beta}\underline{y}^T\right) + m_1 tr\left(\underline{X}\,\underline{B}\underline{\beta}\underline{\beta}^T\underline{B}^T\underline{X}^T\right) + 2m_1\beta_0\underline{1}^T\underline{X}\,\underline{B}\underline{\beta}$$
$$-2m_2 tr\left(\underline{X}\,\underline{B}\underline{A}^T\underline{X}^T\right) + m_2 tr\left(\underline{B}^T\underline{X}^T\underline{X}\,\underline{B}\right) + \lambda tr\left(\underline{B}^T\underline{B}\right) + \underline{1}^T\underline{W} \otimes \underline{B}\underline{1}$$

where $\underline{W}_{p\times k} = \left\{\lambda_{1,j} sign\left(b_{ij}\right)\right\} = \begin{bmatrix} \lambda_{1,1}sign(b_{11}) & \lambda_{1,2}sign(b_{12}) & \cdots & \lambda_{1,k}sign(b_{1k}) \\ \lambda_{1,1}sign(b_{21}) & \lambda_{1,2}sign(b_{22}) & \cdots & \lambda_{1,k}sign(b_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1,1}sign(b_{p1}) & \lambda_{1,2}sign(b_{p2}) & \cdots & \lambda_{1,k}sign(b_{pk}) \end{bmatrix}$, and $\otimes$ is the element-wise product operator. For each column $\underline{b}_j$ of $\underline{B}$, $j = 1, 2, \ldots, k$, we calculate the partial derivatives and equate these to zeros,

$$\frac{\partial\, f_{X,Y}\left(\underline{B} \mid \underline{A}, \underline{\beta}^*, m, \lambda, \underline{\lambda}_1\right)}{\partial \underline{b}_j} = \frac{\partial}{\partial \underline{b}_j}\left\{-2\left(m_1\beta_j\underline{y}^T\underline{X}\,\underline{b}_j - 0.5m_1\beta_j^2\underline{b}_j^T\underline{X}^T\underline{X}\,\underline{b}_j - m_1\beta_0\beta_j\underline{1}^T\underline{X}\,\underline{b}_j\right)\right.$$
$$\left. -2\left(m_2\underline{a}_j^T\underline{X}^T\underline{X}\,\underline{b}_j - 0.5m_2\underline{b}_j^T\underline{X}^T\underline{X}\,\underline{b}_j\right) + \lambda\underline{b}_j^T\underline{b}_j + \underline{1}^T\underline{w}_j \otimes \underline{b}_j\right\},$$
where $\underline{w}_j$ is the j^{th} column of $\underline{W}$

$$\Rightarrow \quad \underline{b}_j = m_1\beta_j\underline{D}\underline{X}^T\underline{y}_C + m_2\underline{D}\underline{X}^T\underline{X}\,\underline{a}_j - 0.5\lambda_{1,j}\underline{D}sign(\underline{b}_j) \ ,$$
where $\underline{D} = \left(m_1\beta_j^2\underline{X}^T\underline{X} + m_2\underline{X}^T\underline{X} + \lambda\underline{I}_p\right)^{-1}$

Given j, with $\underline{b}_j = \left(b_{1j}, b_{2j}, \ldots, b_{pj}\right)^T = \left\{b_{ij}\right\}$ for $i = 1, 2, \ldots, p$, and $\underline{D}^{(i)}$ as the i^{th} row of $\underline{D}$, and taking $\underline{D} = \left\{d_{ij}\right\}$,

$$b_{ij} = m_1\beta_j\underline{D}^{(i)}\underline{X}^T\underline{y}_C + m_2\underline{D}^{(i)}\underline{X}^T\underline{X}\,\underline{a}_j - 0.5\lambda_{1,j}\sum_{l\neq i} d_{il}\,sign(b_{lj}) - 0.5\lambda_{1,j}d_{ii}\,sign(b_{ij}) \ . \qquad (7)$$

In equation (7), b_{ij} is a function of the $b_{lj}'s$ for $l = 1, 2, \ldots, p$ through the constant $\underline{D}^{(i)}sign(\underline{b}_j) = \sum_{l=1}^{p} d_{il}\,sign(b_{lj})$. Suppose the vector $sign(\underline{b}_j)$ is approximated by its "previous iteration" value $sign^*(\underline{b}_j)$. Then a solution for the b_{ij}'s using the previous iteration through $sign^*(\underline{b}_j)$ (where the initial value of $\underline{b}_j$ is the j^{th} column of $\underline{A}$) is given by

$$\hat{b}_{ij} = m_1\beta_j\underline{D}^{(i)}\underline{X}^T\underline{y}_C + m_2\underline{D}^{(i)}\underline{X}^T\underline{X}\,\underline{a}_j - 0.5\lambda_{1,j}\sum_{l\neq i} d_{il}\,sign^*(b_{lj}) - 0.5\lambda_{1,j}d_{ii}\,sign^*(b_{ij}). \qquad (8)$$

Alternatively, consider only the approximation of the $b_{lj}'s$, $l = 1, 2, \ldots, i-1, i+1, \ldots, p$, through the corresponding past iteration values, and maintain the current value of b_{ij} on the right hand side of equation (7) through $sign(b_{ij})$. Soft thresholding is then implemented to or isolate b_{ij} in equation (7). The estimated value of b_{ij} via soft thresholding is given by

$$\hat{\underline{b}}_j = \begin{cases} \underline{w}_j - 0.5\lambda_{1,j}sign(w_{ij}) & , if \ \left|\underline{w}_j\right| > 0.5\lambda_{1,j}\underline{d} \\ 0 & , if \ \left|\underline{w}_j\right| \le 0.5\lambda_{1,j}\underline{d} \end{cases}$$

$$= \; sign(\underline{w}_j) \otimes \left(\left| \underline{w}_j \right| - 0.5\lambda_{1,j} \underline{d} \right)_+ \; , \tag{9}$$

where $w_{ij} = m_1 \beta_j \underline{D}^{(i)} \underline{X}^T \underline{y}_C + m_2 \underline{D}^{(i)} \underline{X}^T \underline{X} \underline{a}_j - 0.5\lambda_{1,j} \sum_{l \neq i} d_{il} sign^*(b_{lj})$, and the vector $\underline{d} = \{ d_{ii}; i = 1,2,...,p \}$ contains the diagonal elements of $\underline{D} = \left(m_1 \beta_j^2 \underline{X}^T \underline{X} + m_2 \underline{X}^T \underline{X} + \lambda \underline{I}_p \right)^{-1}$ ■

4. The LaNS Algorithm

Minimization of the LaNS criterion can be solved using the *LaNS algorithm* discussed below. The tuning parameters (as well as the parameter k for the number of PCs for dimension reduction) must be specified at commencement of the algorithm.

(i) Get the SVD of $\underline{X}$, $\underline{X} = \underline{U}\underline{S}\underline{V}^T$

(ii) Let $\underline{A} = \underline{V}_{(k)}$, i.e., the first k columns of the loadings vector $\underline{V}$

(iii) Initialize $\underline{B}$ as $\underline{B} = \underline{A}$

(iv) Compute for $\underline{\beta}^* = (\beta_0, \underline{\beta})$ as

$\beta_0 = \bar{y}$, and

$\underline{\beta} = \left(\underline{B}^T \underline{X}^T \underline{X} \underline{B} \right)^{-1} \underline{B}^T \underline{X}^T \underline{y}_C$, where $\underline{y}_C = \underline{y} - \bar{y}\underline{1}$

(v) Option1 (LaNS1): "Near Sparsity Method" to update $\underline{B}$, for each $\underline{b}_j$

$\underline{b}_j = m_1 \beta_j \underline{D} \underline{X}^T \underline{y} + m_2 \underline{D} \underline{X}^T \underline{X} \underline{a}_j - 0.5\lambda_{1,j} \underline{D} sign^*(\underline{b}_j)$,

where $\underline{D} = \left(m_1 \beta_j^2 \underline{X}^T \underline{X} + m_2 \underline{X}^T \underline{X} + \lambda \underline{I}_p \right)^{-1}$, and

$sign^*(\underline{b}_j)$ is evaluated for $\underline{b}_j$'s from the previous iteration

Option2 (LaNS2): "Low-Moderate Sparsity Method"

$$\underline{b}_j = sign(\underline{w}_j) \otimes \left(\left| \underline{w}_j \right| - 0.5\lambda_{1,j} \underline{d} \right)_+ \; , \tag{10}$$

where $\underline{w}_j = (w_{1j}, w_{2j}, ..., w_{pj})$ such that for $i = 1,2,...,p$ and a given j,

$w_{ij} = m_1 \beta_j \underline{D}^{(i)} \underline{X}^T \underline{y}_C + m_2 \underline{D}^{(i)} \underline{X}^T \underline{X} \underline{a}_j - 0.5\lambda_{1,j} \sum_{l \neq i} d_{il} sign^*(b_{lj})$,

$\underline{d} = \{ d_{ii}; i = 1,2,...,p \}$ from the diagonals of $\underline{D}$, and

$sign^*(b_{ij})$ is evaluated for b_{ij}'s from the previous iteration

Option3 (LaNS3): "Low-Moderate Sparsity Method"

$$\underline{b}_j = sign(\underline{w}_j^*) \otimes \left(\left| \underline{w}_j^* \right| - 0.5\lambda_{1,j} \underline{d}^* \right)_+ \; , \tag{11}$$

where $\underline{w}_j^* = (w_{1j}^*, w_{2j}^*, ..., w_{pj}^*)$ such that for $i = 1,2,...,p$ and a given j,

$w_{ij}^* = m_1 \beta_j \underline{D}^{*(i)} \underline{X}^T \underline{y}_C + m_2 \underline{D}^{*(i)} \underline{X}^T \underline{X} \underline{a}_j - 0.5\lambda_{1,j} abs\left(\sum_{l \neq i} d_{il}^* sign^*(b_{lj}) \right)$,

$\underline{D}^* = \left(\underline{X}^T \underline{X} \right)^{-1} \left(m_1 \beta_j^2 + m_2 \right)^{-1}$,

$\underline{d}^* = \{ d_{ii}^*; i = 1,2,...,p \}$ from the diagonals of $\underline{D}^*$, and

$sign^*(b_{ij})$ is evaluated for b_{ij}'s from the previous iteration

Option4 (LaNS4): "High Sparsity Method"

$$\underline{b}_j = sign(\underline{v}_j) \otimes \left(\left| \underline{v}_j \right| - 0.5\lambda_{1,j} \underline{d} \right)_+ , \quad (12)$$

where $\underline{v}_j = m_1 \beta_j \underline{D}^* \underline{X}^T \underline{y}_C + m_2 \underline{D}^* \underline{X}^T \underline{X} \underline{a}_j$, and

$$\underline{D}^* = \left(\underline{X}^T \underline{X} \right)^{-1} \left(m_1 \beta_j^2 + m_2 \right)^{-1}$$

(vi) Compute for $\underline{\beta} = \left(\underline{B}^T \underline{X}^T \underline{X} \underline{B} \right)^{-1} \underline{B}^T \underline{X}^T \underline{y}_C$

(vii) Solve for the coefficients of the individual $\underline{X}$'s as $\underline{\theta} = \underline{B} \underline{\beta}$

(viii) Solve for the SVD of $\underline{X}^T \underline{X} \underline{B}$, say $\underline{X}^T \underline{X} \underline{B} = \underline{W} \underline{E} \underline{Z}^T$, and take $\underline{\hat{A}} = \underline{W} \underline{Z}^T$

(ix) Repeat steps (v) to (viii), until convergence.

LaNS1 suggests adjusting the regression coefficients which are not necessarily sparse, hence the label "Near Sparsity Method." LaNS2 gives sparse solutions, with the choice of λ having little effect on the optimization process and so λ may be set to zero. In the context of soft thresholding to achieve sparse solutions, $0.5\lambda_{1,j} \sum_{l \neq i} d_{il}^* sign^*(b_{lj})$ in LaNS2 must always be positive to directionally shrink the soft thresholding operator's value toward zero. Hence, LaNS3 takes the absolute value of $\sum_{l \neq i} d_{il}^* sign^*(b_{lj})$ (since the $\lambda_{1,j}$'s are always positive). Unlike LaNS1, both LaNS2 and LaNS3 give sparse solutions and hence the label "Low-Moderate Sparsity Method." Finally, LaNS4 is formulated so as to attain more sparse solutions, if not faster convergence, by maintaining the magnitude of the thresholding across the b_{ij}'s on each iteration, and therefore is labelled as "High Sparsity Method."

Note that LaNS3 is a "re-scaling" of LaNS2 so that the sparsity may be achieved in a potentially faster manner. LaNS2 iteratively uses equation (10) whereas LaNS3 iteratively utilizes equation (11) to come-up with sparse sets of $\underline{b}_j$'s. Both equations (10) and (11) are similar to the soft thresholding operator (on ℓ_1 norm) of the form $b = sign(w)\left(|w| - t \right)_+$, where b is approximated by a function (or a constant) w and a tuning parameter t that regulates the thresholding. In equation (10), the operator involves the function $\sum_{l \neq i} d_{il} sign(b_{lj})$ which may be negative thereby resulting to "bloating" rather than "deflating" of the $\underline{b}_j$'s toward zero at certain iterations. This "shortcoming" is addressed in LaNS3 by taking the absolute value of $\sum_{l \neq i} d_{il} sign(b_{lj})$, as shown in equation (11).

The value of $\sum_{l \neq i} d_{il} sign(b_{lj})$ in LaNS2, or its absolute value in LaNS3, together with the other soft thresholding parameter (i.e., the term defined by $\lambda_{1,j} \underline{d}$), change in every iteration, thus, possibly leading to a slower rate of deflation of the $\underline{b}_j$'s to zero for either LaNS2 or LaNS3. LaNS4 resolves the slow decay by fixing the parameters (i.e., using $\lambda_{1,j}$ instead of $\lambda_{1,j} \underline{d}$) for soft thresholding on the $\underline{b}_j$'s, and by re-specifying the w_{ij}'s to v_{ij}'s across the iterations. LaNS4, therefore, applies a "true" ℓ_1 norm soft thresholding operator on $\underline{b}_j$'s that may give fast convergence and/or sparser coefficients.

5. Comparison of LaNS Algorithm with Other Procedures

The different options in the LaNS algorithm adjusts sparsity vis-à-vis minimizing squared prediction error, thereby potentially identifying a small or sparse set of independent variables that has

the "best" representation of the entire set (of independent variables), and which remains highly predictive of the dependent variable. As such, the LaNS criterion and the different LaNS options is compared to the different criteria and/or procedures for regression modeling, dimension reduction, and variable selection.

Given the general optimization problem in equation (6), with $m_1 = m_2 = 1$ and $\underline{B} = \underline{A}$, and when the optimization is made through a two-stage process – first, by finding the solution $\hat{\underline{A}} = \arg\min_{\underline{A}} \left\{ \left\| \underline{X} - \underline{X}\,\underline{A}\underline{A}^T \right\|_F^2 + \lambda \left\| \underline{A}^T \right\|_F^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \underline{a}_j \right\|_1 \right\}$ subject to $\underline{A}^T \underline{A} = \underline{I}_k$, and second, by finding the solution $\hat{\underline{\beta}}^* = \arg\min_{\underline{\beta}^*} \left\{ \left\| \underline{y} - \left[\underline{1} \;\; \underline{X}\hat{\underline{A}} \right] \underline{\beta}^* \right\|^2 \right\}$ – then this simplifies to PCR. LaNS1 is viewed as an adjustment of the coefficients of the PCs derived from PCA to deflate some of the coefficients of independent variables relative to its predictive ability. Therefore, LaNS1 may result to the PCR of all the PCs, or worst, to the OLS of all the independent variables when too much "effective" adjustment is made. That is, for a large number of iterations, regardless of the values of the tuning parameters λ and $\lambda_{1,j}$'s, LaNS1 eventually yields the OLS regression coefficients. This suggests that under LaNS1, when the number of iteration is increased, the estimation of the regression coefficients is dominated by the predictive ability constraint, more than by the dimension reduction constraint.

Given equation (6), with $m_1 = m_2 = 1$, and when optimization is approached in two stages – first, by finding the solution $\left(\hat{\underline{A}}, \hat{\underline{B}}\right) = \arg\min_{\underline{A},\underline{B}} \left\{ \left\| \underline{X} - \underline{X}\,\underline{B}\underline{A}^T \right\|_F^2 + \lambda \left\| \underline{B}^T \right\|_F^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \underline{b}_j \right\|_1 \right\}$ subject to $\underline{A}^T \underline{A} = \underline{I}_k$, and second, by finding the solution $\hat{\underline{\beta}}^* = \left\| \underline{y} - \left[\underline{1} \;\; \underline{X}\,\underline{B} \right] \underline{\beta}^* \right\|^2$ – then method is equivalent to SPCR. LaNS2 is viewed as a modification of the loadings of the SPCs derived from SPCA. Specifically, LaNS2 fine-tunes the already-sparse loadings by simultaneously optimizing prediction of the dependent variable, resulting to further sparsity of the independent variables.

Recall that in principal covariates regression (PCovR), the objective function is to minimize

$$f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}^*\right) = \frac{(1-\alpha)}{\left\|\underline{y}\right\|^2} \left\| \underline{y} - \left(\underline{1} \,,\, \underline{X}^*\,\underline{B} \right) \underline{\beta}^* \right\|^2 + \frac{\alpha}{\left\| \underline{X}^* \right\|^2} \left\| \underline{X}^* - \underline{X}^*\,\underline{B}\underline{A}^T \right\|_F^2$$

$$\text{subject to } \underline{A}^T \underline{A} = \underline{I}_k \text{ and } \alpha \in (0,1),$$

where $\underline{X}^*$ is the scaled $\underline{X}$ having zero mean and unit variance (for each of the independent variable). Thus, the LaNS optimization problem in equation (6) may derive the PCovR when the $\underline{X}$'s are rescaled, the λ and $\lambda_{1,j}$'s are all set to zero, and with specific values of m_1 and m_2. Minimizing $f_{X,Y}\left(\underline{A}, \underline{B}, \underline{\beta}^*\right) = m_1 \left\| \underline{y} - \left(\underline{1} \,,\, \underline{X}^*\,\underline{B} \right) \underline{\beta}^* \right\|^2 + m_2 \left\| \underline{X}^* - \underline{X}^*\,\underline{B}\underline{A}^T \right\|_F^2$ subject to $\underline{A}^T \underline{A} = \underline{I}_k$ yields at the very first iteration: (a) the same set of PCovR estimates at $\alpha = 1$ when $m_2 = \frac{\alpha}{\left\| \underline{X}^* \right\|^2}$ (and is also equivalent to PCR); (b) an adjusted set of PCovR estimates at $\alpha \in (0,1)$ when $(m_1, m_2) = g\left(\frac{1-\alpha}{\left\|\underline{y}\right\|_F^2}, \frac{\alpha}{\left\|\underline{X}\right\|_F^2} \right)$; and (c) a

nearly-similar set of PCovR estimates at $\alpha = 0$ when $m_1 \to \infty$ and is also equivalent to the Reduced-Rank Regression (RRR).

6. Simulation Studies

The performance of LaNS is further evaluated through simulation studies. Assume that the data come from 3 latent factors V_1, V_2 and V_3. Suppose: $V_1 \sim N(300, 300^2)$; $V_2 \sim N(300, 290^2)$, independent from V_1, and; $V_3 = 0.9 * V_1 - 0.15 * V_2 + \omega$, $\omega \sim N(0,10)$. The latent factor V_1 gives the most information (having high variability), closely followed by V_2, then by V_3. V_1 and V_2 are independent, suggesting that both give different (uncorrelated) yet important information. In contrast, V_3 is a function of V_1 and V_2, and thus V_3 is also as important and that it carries further information coming from V_1 and V_2.

The independent variables $X_1, X_2, ..., X_{1000}$ are each derived as

$$X_j = V_1 + \varepsilon^{(j)}, \quad \text{for} \quad j = 1, 2, ..., 10$$
$$X_j = V_2 + \varepsilon^{(j)}, \quad \text{for} \quad j = 11, 12, ..., 20$$
$$X_j = V_3 + \varepsilon^{(j)}, \quad \text{for} \quad j = 21, 22, ..., 1000$$

where the $\varepsilon^{(j)}$'s are independent and such that $\varepsilon^{(j)} \sim N(0,10)$ for $j = 1, 2, ..., 1000$. Given the formulation of the independent variables, the pairs of variables from any of the sets $C_1 = (X_1, X_2, ..., X_{10})$, $C_2 = (X_{11}, X_{12}, ..., X_{20})$ and $C_3 = (X_{21}, X_{22}, ..., X_K)$ where K=40 for the non-high dimensional (NHD) case and K=1000 for the high dimensional (HD) case, are correlated within sets (i.e., when two variables come from the same set) and across combinations of sets C_1 and C_3 or of sets C_2 and C_3 (i.e., when one variable comes from C_3 and the other comes from either C_1 or C_2).

The dependent variable Y is then computed from $X_1, X_2, ..., X_{1000}$. Given the values on the i^{th} observation, $X_{i1}, X_{i2}, ..., X_{i1000}$, the dependent variable Y_i is computed as

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{1000} X_{i1000} + \varepsilon_i$, $\varepsilon_i \sim N(0, 50^2)$ for the HD case; and

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{40} X_{i40} + \varepsilon_i$, $\varepsilon_i \sim N(0, 50^2)$ for the NHD case.

Note that the parameters $\beta_1, \beta_2, ..., \beta_{1000}$ are specified to control for the relative contributions of the independent variables $X_1, X_2, ..., X_{1000}$ 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), all the variables $X_1, X_2, ..., X_{1000}$ 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 variables $X_1, X_2, ..., X_{40}$ are considered. Across all scenarios, a total of 100 observations are considered. Specifications of the different simulation settings are summarized in Table 1.

Table 1. Summary of Scenarios

Scenario		p	% Contribution of C_1 on Y	% Contribution of C_2 on Y	% Contribution of C_3 on Y
1	NHD	40	55%	35%	10%

	HD	1,000	55%	35%	10%
2	NHD	40	35%	10%	55%
	HD	1,000	35%	10%	55%

Scenario 1 (for both NHDs and HDs) is 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 remaining independent variables derived from V_3 are implicitly derived from V_1 or V_2, and therefore these independent variables still have minimal yet important impact on Y. In contrast, Scenario 2 (NHD and HD) is formulated such that the independent variables most predictive of Y come from a large (or very large) set of independent variables derived from V_3. Thus, the independent variables derived from V_1 or V_2 are as important yet less contributing to Y. Note here that the "percent contribution of C_j on Y" is the proportion of (the magnitude of) Y that is derived from the set of $X_i s$ in C_j.

6.1 Comparison of Different Methods

The different LaNS options are compared to various regression methods that address multicollinearity or mitigate the issues associated with high dimensional inputs. For the non-high dimensional cases (NHDs), the fitted full model from LaNS is compared to those of ordinary least squares regression (OLS), principal component regression (PCR), and principal covariates regression (PCovR); for models with sparse coefficients, LaNS is compared to OLS, sparse principal component regression (SPCR), regression with LASSO, and regression with elastic net (EN); and the ordinary least squares regression model using selected variables from LaNS are compared to the ordinary least squares regression models using the corresponding selected variables from SPCR, LASSO or EN (note that PCR and PCovR do not give sparse solutions, hence the full models are the same as the reduced models). 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 data is simulated so that the third latent factor is as important as the first and second latent factors, so that either the third latent factor may already be explained by the first and second latent factors or vice-versa. Similarly, the independent variables with most contribution to the dependent variable may come from any two latent factors. Thus, identification of only two (out of the three) latent factors may already be sufficient. To facilitate comparisons, for LaNS, PCR, and SPCR that require a specification on the number of dimensions, the parameter for the number of dimensions is set at 2. And for methods that achieve sparsity like LaNS, SPCR, LASSO, EN, and SPCR, the number of non-zero coefficients of the $\underline{X}'s$ in a sparse solution is set at a maximum of 20.

PCovR α values are set at 0.85, 0.50, and 0.15, respectively. Higher values for α gives results leaning toward the direction of PCR having the most dimension reduction. In contrast, lower values for α yield solutions leaning toward the direction of RRR with fitted models that are focused on predictive ability. For models derived using EN, tuning parameters are set at either 0.01 or 100. A tuning parameter of 100 for EN gives heavier penalty on the ℓ_2 norm constraint, while a tuning parameter of 0.01 for EN almost ignores the ℓ_2 norm constraint, resulting to solutions similar to that of the LASSO.

The models are assessed based on their predictive ability through the sum of squared prediction error (SSPE), computed as $SSPE = \sum(y_i - \hat{y}_i)^2$, where y_i and $\hat{y}_i$ are the true and predicted values of the dependent variable, respectively. 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 Schwarz (1978) and Zou et al. (2007), the BIC-type criterion is defined as $BIC = \frac{MSPE}{Var(\underline{y})} + NNZ\,\frac{\log(n)}{n}$, where $MSPE = \frac{1}{n}SSPE$, $Var\left(\underline{y}\right)$ is the variance of the dependent variable, and NNZ is the number of nonzero coefficients of $\underline{X}$. 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.

6.2 Scenario 1

The data for Scenario 1 is generated from a structure where 55%, 35%, and 10% of the dependent variable is explained by predictors from the first latent factor, predictors from the second latent factor, and predictors from the third latent factor, respectively. Summary of the results are in Table 2 and Table 3 for the NHD and HD cases, respectively, and discussions of which follow. Note that 5 replicates were generated under the HD case and 7 replicates under the NHD case.

Table 2. Summary of SSPE and BIC measures for the different Models under the NHD Case

	Number of Variables in Model	Using Model		After OLS		No. of variables from Component		
		SSPE	BIC	SSPE	BIC	From C_1	From C_2	From C_3
OLS	40	139812.3	1.844	139812.3	1.844	10	10	20
PCR	40	561325.6	1.849	139812.3	1.844	10	10	20
PCovR(0.85)	40	442344.6	1.848	139812.3	1.844	10	10	20
PCovR(0.5)	40	247981.8	1.845	139812.3	1.844	10	10	20
PCovR(0.15)	40	166311.9	1.844	139812.3	1.844	10	10	20
LaNS1	40	150312.0	1.844	139812.3	1.844	10	10	20
LaNS2	22.14	434255.0	1.023	178214.8	1.022	8.143	8.429	5.571
LaNS3	16.86	1842303.0	0.784	246962.4	0.779	6.857	6.286	4.333
LaNS4	10.71	749158.2	0.501	336778.6	0.498	5.286	2.857	2.571
SPCR	12.86	74300154.7	1.541	23461308	0.886	6.571	3.143	3.143
LASSO	10.71	6609285.7	0.577	460546.7	0.499	6.429	4.286	0
EN(0.01)	10.71	8515472.3	0.602	464796	0.499	6.571	4.143	0
EN(100)	10.71	25218588.3	0.829	13608213	0.668	9.857	0.143	0.714

For the NHD case under the full model, clearly the LaNS1 generates a relatively similar model to that of the OLS in terms of predictive capability (comparing SSPEs and BICs of OLS vs LaNS1). As suggested in the formulation of LaNS, LaNS1 yields OLS estimates when sparsity is not of the main interest. The fitted models from PCR have the lowest predictive ability on the average even when all independent variables are used in the model. PCovR dominates PCR, with SSPEs and BICs

for PCovR (at different settings) lower than those of the 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).

LaNS2 and LaNS3 offer sparse solutions for which BIC values remain lower than LaNS1, with LaNS2 having about 23 independent variables and LaNS3 having about 18 on the average. Both LaNS2 and LaNS3 select independent variables coming from all three latent factors. Except for one of the replicates (and thus the inflated SSPE of LaNS3), both LaNS2 and LaNS3 are as good as those of PCovR(0.50) or PCoVR(0.85). LaNS4, as formulated, gives the most sparse solution among the LaNS options. Among those methods yielding relatively similar number of variables, LaNS4 identifies independent variables from all three latent factors and gives the fitted model with highest predictive ability, unlike EN(100) which include almost always all 10 independent variables from the first latent factor, and unlike EN(.01) and LASSO which include variables always from both the first and second latent factors. EN(100) tends to select the set of independent variables that is highly correlated with the dependent variable, while SPCR tends to select the set of independent variables with the most variation.

Comparing the selected variables using different methods, and implementing OLS using only the selected variables as predictors, those variables identified by LaNS4 give better prediction than those identified by any of SPCR, LASSO, EN(0.01), or EN(100). Similarly, LaNS4 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 LaNS4 appears to be a better option than LASSO or EN.

The prediction error of LaNS4 (except for one case, hence the inflation of SSPE) is comparable to that of OLS using the selected variables from LaNS4. That is, LaNS4 may have estimated coefficients that are either nearly the same as that of first finding the best set and then fitting the regression model (i.e., the two-step approach), or are different from that of the two-step approach but have nearly the same predictive ability. Such may also be inferred under LaNS2 or LaNS3. These results suggest that the one-step approach of LaNS in dimension reduction and variable selection may already be sufficient for finding the set of predictors that are most predictive of the response.

Table 3. Summary of SSPE and BIC measures for the different Models under the HD Case

	Number of Variables in Model	Using Model		After OLS		No. of variables from Component		
		SSPE	BIC	SSPE	BIC	From C_1	From C_2	From C_3
PCovR(0.85)	1000	771869.3	46.061					
PCovR(0.5)	1000	266164.8	46.055					
PCovR(0.15)	1000	23872.0	46.052					
LaNS2	27.80	2704327.1	1.301	2704327.1	1.311	9.20	8.60	10.00
LaNS4	11.20	817456.4	0.518	441988.3	0.521	7.80	3.40	0
SPCR	16.40	83056632.4	1.738	9968154.8	0.874	10.00	0	6.400
LASSO	11.00	3752915.4	0.552	541586.4	0.513	6.80	4.20	0
EN(0.01)	11.00	5463551.8	0.581	553258.3	0.522	6.80	4.40	0
EN(100)	11.00	28052031.6	0.849	12885700.8	0.661	9.60	1.20	0.20

For the HD case, note that LaNS3 is not included in the model as resulting models are sparse but still have more than 30 independent variables. For those with sparse solutions, LaNS2 and LaNS4 clearly dominate any of SPCR, LASSO or EN. LaNS4 is even comparable with PCovR(0.85) in terms of predictive ability. SPCR is the least performing among those with sparse solutions. Interestingly, LaNS2 yield solutions that are optimal at the OLS scale, that is, the estimated model by LaNS2 with the few independent variables is the same as the OLS model for these independent variables (assuming selected a priori).

Comparing the selected variables using different methods, and implementing OLS using only the selected variables as predictors, those variables identified by LaNS4 give better prediction than those identified by any of SPCR, LASSO, EN(0.01), or EN(100). Again, if identification of a smaller set is of interest, then LaNS4 may be a better option than SPCR, LASSO, or EN.

6.3 Scenario 2

For Scenario 2, simulated data for both NHD and HD cases are based on a structure where 35%, 10%, and 55% of the dependent variable is explained by the predictors derived from the first latent factor, by the predictors derived from the second latent factor, and by those derived from the third latent factor, respectively. Note that 5 replicates were generated under the HD case and 7 replicates under the NHD case. Summary of the results are presented in Table 4 and Table 5.

Table 4. Summary of SSPE and BIC measures for the different Models under the NHD Case

	Number of Variables in Model	Using Model		After OLS		No. of variables from Component		
		SSPE	BIC	SSPE	BIC	From C_1	From C_2	From C_3
OLS	40	140159.1	1.846	140159.1	1.846	10	10	20
PCR	40	381022.2	1.851	140159.1	1.846	10	10	20
PCovR(0.85)	40	298638.1	1.849	140159.1	1.846	10	10	20
PCovR(0.5)	40	183798.3	1.847	140159.1	1.846	10	10	20
PCovR(0.15)	40	145291.8	1.846	140159.1	1.846	10	10	20
LaNS1	40	146242.1	1.846	140159.1	1.846	10	10	20
LaNS2	25.43	250183.7	1.176	160757.6	1.175	7.43	6.14	12.00
LaNS3	16.71	906468.0	0.784	273686.2	0.776	5.43	4.29	7.00
LaNS4	10.43	1215649.3	0.510	565086.6	0.494	1.00	0.86	8.57
SPCR	13.14	39210674.3	1.569	6340676.2	0.757	9.00	1.86	3.57
LASSO	10.14	5485865.8	0.607	519485.1	0.480	5.57	0	4.57
EN(0.01)	10.43	8930454.6	0.704	526144.4	0.493	6.00	0	4.43
EN(100)	10.43	37199143.7	1.391	578611.8	0.494	7.14	0	3.29

For the NHD case, LaNS1 generates non-zero coefficients for all independent variables, with the fitted model being comparable with PCovR(0.15) and better than those of OLS, PCR, PCovR(0.85), and PCovR(0.50). Noticeably, LaNS1 and PCovR(0.15) have on the average nearly the same predictive ability as that of the OLS. Most of the identified variables for LaNS2, LaNS3 and LaNS4 come from the third latent factor. LaNS4, on the other hand, gives the smallest prediction error among all other models with the same sparsity level (SPCR, LASSO and EN). LaNS4 maintains a representation mainly from the third latent factor – which is hypothetically the case since the variables from the third latent factor have the most contribution to the dependent variable. This however is the opposite for SPCR, LASSO and EN, as the identified variables come mostly from the first latent factors. In addition, when considering a pre-process of selecting the independent variables for the

OLS, the LaNS procedure gives far better results than any of SPCR, LASSO, or EN. SPCR gives the highest prediction error indicating that variable selection via the SPCA may not give the best set of highly predictive independent variables.

Table 5. Summary of SSPE and BIC measures for the different Models under the HD Case

	Number of Variables in Model	Using Model		After OLS		No. of variables from Component		
		SSPE	BIC	SSPE	BIC	From C_1	From C_2	From C_3
PCovR(0.85)	1000	244706.8	46.057					
PCovR(0.5)	1000	65497.4	46.053					
PCovR(0.15)	1000	4687.0	46.052					
LaNS2	31.80	402541.6	1.465	188364.1	1.469	9.20	10.00	14.40
LaNS4	12.00	1278161.0	0.577	564079.8	0.564	3.20	0	9
SPCR	16.80	48748322.6	1.728	585724.0	0.785	10.00	0	6.80
LASSO	12.00	3355146.4	0.617	505289.9	0.563	6.00	0	6.00
EN(0.01)	12.00	6797982.6	0.681	505546.8	0.563	6.40	0	6.00
EN(100)	12.00	44554051.0	1.429	50395718.0	1.543	9.40	0	2.60

For the HD case, LaNS2 and LaNS4 give sparse models with only about 32 and 12 independent variables, respectively. LaNS2 identifies more independent variables and thus yields better prediction than any of the more sparse models (SPCR, LASSO, EN). Evidently for HD, the "best" models must have a relatively large number of predictors.

Most of the variables included in LaNS4 are from the third latent factor, whereas SPCR, LASSO and EN identify independent variables from mainly the first latent factor. Using the sets of selected variables for OLS, the fitted model from LaNS4 gives a relatively higher prediction error compared to those of LASSO or EN(0.01) but relatively lower prediction error compared to SPCR and EN(100). These suggests that the fitted model from LaNS4 is better than the fitted models from SPCR, LASSO and EN; and that the OLS of the selected variables from LaNS4 is comparable with the OLS of those selected by any of SPCR, LASSO or EN.

7. Illustration: Quality of Life Across Countries

While quality of life is a multidimensional phenomenon, we focused in this example on aspects of mortality (adult, infant, maternal, by specific causes, etc.) and explored the interplay between various factors affecting mortality at the macro level. Using variables from WHO website (WHO, 2016) at country level, we analyzed data in the vicinity of 2010 (census year for most countries) to identify a quality of life index based on mortality indicators. Initial screening for redundancy and data quality (specifically, missing data values), leads to 97 variables. Using sparse principal component analysis (Zou et al, 2006), the first component accounts for 56.05% of the variance, with 35 non-zero loadings. With negative loadings for mortality indicators, the component aptly labelled as Quality of Life Index (QoLI), was re-scaled so that values will range between 0-100 for ease of interpretation. High values of QoLI was observed for Israel, Japan, Republic of Korea, France, and Netherland. On the other hand, low values of QoLI was observed for Turkmenistan, Somalia, Central African Republic, Afghanistan, and Uzbekistan.

Given QoLI (y), we aim to identify its determinants, this will help various government in prioritizing limited resources and focus on those that really have impact on the quality of life of their

people. Due to high dimensionality of possible predictors (related to: environment, lifestyle, health status, health policy, health care, and morbidity) we simultaneously reduce dimension of the predictors while we choose the important variables using the LaNS algorithm. We used the following tuning parameters: $m_1 = 1$; $m_2 = 1$; $k = 1$; $\lambda = 1$. For $\lambda_{1,j}$, any value from 15 to 255 are all feasible to induce sparsity and the resulting determinants are meaningful.

Possible determinants were again screened for redundancy of variables and data quality, this results to 106 variables. The algorithm was able to identify 7 determinants. Recall the objective function simultaneously reduce the dimension (sparsity) while selecting the variables that best explain the dependent variable QoLI. The final regression model is fitted using the 7 determinants, this accounts for 71% of the total variation in QoLI, reasonably high given that these were chosen from a total of 106 predictors. The estimated regression coefficients are given in Table 6. This indicates that countries should focus on immunization and care for women to improve mortality-related quality of life.

Table 6. Regression Coefficient of Subset Determinants of QoLI

Variable	Reference Year	Estimate	p-value
Intercept		88.7771	<0.0001
Outdoor air pollution (Annual PM10 [ug/m3])	2004	-0.0556	0.1433
Consumption of alcohol per capita among ages 15+ (liters)	2000	-1.6565	0.0967
Population using improved sanitation facilities (%)	2011	0.0396	0.5730
Percentage women with raised age-standardized BP	2008	-1.8685	<0.0001
Hib immunization coverage among 1-year-olds (%)	2000	0.1883	<0.0001
Polio immunization coverage among 1-year-olds (%)	2000	0.3783	0.0043
Per capita government expenditure on health (US$)	2009	0.0035	0.0326

Quality of life index is explained primarily by health condition of women (measured by blood pressure), welfare of children (measured in terms of immunization coverage), and the amount of spending by the government on health.

8. 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. The regression estimation under LaNS is not directly implemented on all the independent variables (from the full model), but rather on a smaller set of transformed independent variables via the modified SPCs. 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. This characteristic suggests that grouping effect, i.e., selection of independent variables or inputs that explains the same factor/dimensionality is avoided by the LaNS procedure. Also, the LaNS procedure tends to select inputs coming from the most "predictive" subset (as identified by a latent dimension),

followed by those from the next most "predictive" subset (as identified by another latent dimension), and so on.

References:

Candes, E. and T. Tao (2007). The Dantzig Selector: Statistical estimation when p is much larger than n, The Annals of Statistics, Vol. 35, No. 6, pp. 2313-2351

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

Chen, L. and J. Huang (2012). Sparse Reduced-rank Regression for Simultaneous Dimension Reduction and Variable Selection, Journal of the American Statistical Association, 107(500):1533-1545

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

Chipman, H., George, E. and R. McCulloch (2010). BART: Bayesian Additive Regression Trees, Annals of Applied Statistics, Vol. 4, No. 1, pp. 266-298

Cook R. D. (2007). Fisher lecture: dimension reduction in regression (with discussion). Statist. Sci., Vol. 22, pp. 1–43

De Jong, S. and H.A.L. Kiers (1992). Principal Covariates Regression, Chemometrics and Intelligent Laboratory Systems, Vol. 14, pp. 155-164

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

George, E.I. and S.D. Oman (1996). Multiple-Shrinkage Principal Component Regression, The Statistician, Vol. 45, No. 1, pp. 111-124

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

Helland, I. (1992). Maximum Likelihood Regression on Relevant Components, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 54, No. 2, pp. 637-647

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

Hsieh, W. (2004), From PCA to Nonlinear PCA, AMS Short Course on Artificial Intelligence Methods in Atmospheric and Oceanic Sciences: Neural Networks, Fuzzy Logic, and Genetic Algorithms, pp. 10-11

Hwang, J. and D. Nettleton (2003). Principal Components Regression with Data-chosen Components and Related Methods, Technometrics, Vol. 45, pp. 70-79

Izenman, A.J. (1975). Reduced-rank Regression for the Multivariate Linear Model, Journal of Multivariate Analysis, Vol. 5, pp. 248-264

Jolliffe, I. (1982). A Note on the Use of Principal Components in Regression, Journal of Applied Statistics, Vol. 31, No. 3, pp. 300-303

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

Jolliffe, I. and M. Uddin (2000). The Simplified Component Technique: An Alternative to Rotated Principal Components, Journal of Computational and Graphical Statistics, Vol. 9, pp. 689-710

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

Lee, T.S. (1987). Algorithm AS 223: Optimum Ridge Parameter Selection, Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 36, No. 1, pp. 112-118

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

McDonald, G.C. and D.I. Galarneau (1975). Journal of the American Statistical Association, Vol 70, No 550, pp. 407-416

Podesta, F. (2002). Recent Developments in Quantitative Comparative Methodology: The case of Pooled Time Series Cross-Section Analysis, DSS Papers, Soc 3-02, 44 pp.

Ravikumar P., Liu H., Lafferty J., and L. Wasserman (2007). SpAM: sparse additive models. In Platt J., Koller D., Singer Y., Roweis S., eds., Advances in neural information processing systems, Cambridge, MA: MIT Press, Vol. 20, pp. 1201–1208.

Reinsel, G.C. and P.R. Velu (1998). Multivariate Reduced-rank Regression: Theory and Applications, New York: Springer.

Rousson, V. and T. Gasser (2004). Simple component analysis, Applied Statistics, 53(4):539-555.

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.

Vines, S.K (2000), Simple Principal Components, Applied Statistics, 49(4):441-451.

Wold, H. (1975). Path models with latent variables: The NIPALS approach, In Blalock, H., Aganbegian, A., Borodkin, F., Boudon, R., and Capecchi, V., eds., Quantitative sociology: International perspectives on mathematical and statistical modeling, New York: Academic, pp. 307-357

WHO (2016), WHO Website available at: http://www.who.int/gho/en/

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. and T. Hastie (2012), elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA, R package version 1.1, http://CRAN.R-project.org/package=elasticnet

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