Computational Statistics and Data Analysis 112 (2017) 242–256

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous dimension reduction and variable selection in modeling high dimensional data

Joseph Ryan G. Lansangan, Erniel B. Barrios *

School of Statistics, University of the Philippines Diliman, Philippines

HIGHLIGHTS

- Dimension reduction and variable selection are integrated in an objective function.
- Existence of the solution to the constrained objective function is established.
- Solution is via optimizing predictive ability vis-à-vis selection of predictors.
- Smaller prediction errors are observed even under non-high dimensional settings.

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 17 March 2017
Accepted 18 March 2017
Available online 28 March 2017

Keywords:
High dimensionality
Regression modeling
Dimension reduction
Variable selection
Latent factors
Sparsity
Soft thresholding
Sparse principal component analysis

ABSTRACT

High dimensional predictors in regression analysis are 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.

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 $p \gg n$, i.e., data with a relatively larger number of features compared to the number of observations.

In regression analysis, multicollinearity may result in ill-conditioning and/or near-singularity of the associated design matrix, resulting in unstable estimates (inflated standard errors). Similarly, classical regression framework assumes $p \leq n$;

* Corresponding author.
E-mail address: ebbarrios@up.edu.ph (E.B. Barrios).

http://dx.doi.org/10.1016/j.csda.2017.03.015

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 in 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 in 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 $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$. 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 $\hat{\underline{\beta}} = \left(\underline{X}^T\underline{X}\right)^{-1}\underline{X}^T\underline{y}$ is optimal (Gauss–Markov Theorem) provided that $n > p$.

When $p \gg n$, the estimator $\hat{\underline{\beta}}$ 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 $\hat{\underline{\beta}}$ 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 $\hat{\underline{\beta}}$ 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, 1982). 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 and Velu (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 introduces 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.

Partial least squares (PLS) regression, introduced by Wold (1966), is also used as an alternative in the presence of ill-conditioning or when $p \gg n$. PLS regression projects the response and predictor variables to new spaces via a latent variable approach, with model optimization done by finding the dimensions in the predictor space that explains the most variation in the response space. PLS regression in high dimensional data may however result in linear combinations of the original predictors that are difficult to interpret, as PLS regression is not particularly tailored for feature selection (Chun and Keles, 2010). Chun and Keles (2010) then developed a sparse partial least squares (SPLS) approach to simultaneously achieve good predictive ability and variable selection. They provided an efficient implementation of SPLS regression using a LARS (least angle regression) algorithm (Chun and Keles, 2010).

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 in components with sparse loadings. 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 ith 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 jth variable, where $j = 1, 2, \ldots, p$. Thus, $\underline{X} = \left(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\right)^T = \left(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p\right)$ 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. Let $\underline{A} = \left[\underline{\alpha}_1, \ldots, \underline{\alpha}_k\right]$ and $\underline{B} = \left[\underline{b}_1, \ldots, \underline{b}_k\right]$ be $p \times k$ matrices, where $k < p$ and such that $\underline{A}^T\underline{A} = \underline{I}_k$. The sparse principal component analysis (SPCA) criterion is then given by

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

where λ and $\lambda_{1,j}$ are penalizing constants chosen to ensure existence of solution and convergence of the computational algorithm. Here, $\|\cdot\|_1$ is the ℓ_1 norm and $\|\cdot\|$ 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 correct the grouping effect (i.e., strongly correlated predictors tend to be in or out of the model together). Unlike PCA, the solution (and its corresponding algorithm) yields 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 accounts 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.

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). 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). With λ and $\underline{\lambda}_1 = \left(\lambda_{1,1}, \lambda_{1,2}, \ldots, \lambda_{1,k}\right)$ as some constants (specifically, the tuning parameters), the SPCA criterion (Zou et al., 2006) then minimizes

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

Now, consider regressing $\underline{y} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$ 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

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

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

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

Optimization of Eq. (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 $\underline{A}^T\underline{A} = \underline{I}_k$,

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

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 λ, $\underline{\lambda}_1$, $\underline{m} = (m_1, m_2)$ and the choice for k, find the values $\hat{\underline{A}}$, $\hat{\underline{B}}$ and $\hat{\underline{\beta}}^*$ for which

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

Note that the first two terms in Eq. (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 Eq. (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 Eq. (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 $\underline{\theta} = \underline{B\beta}$ 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 Eq. (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 labeled 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 Eq. (6). To facilitate initialization of the SVD, we separate β_0 from the rest of the β's.

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

Proof. Given $\underline{\beta}^*$ and $\underline{B}$, Eq. (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 $\underline{X}^T\underline{XB} = \underline{WEZ}^T$, then for a fixed m_2, the solution for $\underline{A}$ is given by $\hat{\underline{A}} = \underline{WZ}^T$.

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

$$\begin{aligned}
\frac{\partial}{\partial\beta_0} f_{X,Y}\left(\underline{\beta}^* | \underline{A}, \underline{B}, \lambda, \underline{\Lambda}\right) &= \frac{\partial}{\partial\beta_0}\left[n\beta_0^2 + 2\beta_0\underline{1}^T\left(\underline{XB\beta} - \underline{y}\right)\right] \\
&\Rightarrow \beta_0 = \frac{1}{n}\underline{1}^T\left(\underline{y} - \underline{XB\beta}\right) \\
\frac{\partial}{\partial\underline{\beta}} f_{X,Y}\left(\underline{\beta}^* | \underline{A}, \underline{B}, \lambda, \underline{\Lambda}\right) &= \frac{\partial}{\partial\underline{\beta}}\left[-2tr\left(\underline{XB\beta y}^T\right) + tr\left(\underline{XB\beta\beta}^T\underline{B}^T\underline{X}^T\right) + 2\beta_0\underline{1}^T\underline{XB\beta}\right] \\
&\Rightarrow \hat{\underline{\beta}} = \left(\underline{B}^T\underline{X}^T\underline{XB}\right)^{-1}\underline{B}^T\underline{X}^T\underline{y}_C.
\end{aligned}$$

Thus,

$$\begin{aligned}
\beta_0 &= \frac{1}{n}\underline{1}^T\left(\underline{y} - \underline{XB\beta}\right) \\
&\Rightarrow \hat{\beta}_0 = \bar{y}. \quad \blacksquare
\end{aligned}$$

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

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

$$\begin{aligned}
f_{X,Y}\left(\underline{B} | \underline{A}, \underline{\beta}^*, \underline{m}, \lambda, \underline{\lambda}_1\right) = {}& -2m_1 tr\left(\underline{XB\beta y}^T\right) + m_1 tr\left(\underline{XB\beta\beta}^T\underline{B}^T\underline{X}^T\right) + 2m_1\beta_0\underline{1}^T\underline{XB\beta} \\
& - 2m_2 tr\left(\underline{XBA}^T\underline{X}^T\right) + m_2 tr\left(\underline{B}^T\underline{X}^T\underline{XB}\right) + \lambda tr\left(\underline{B}^T\underline{B}\right) + \underline{1}^T\underline{W} \otimes \underline{B1}
\end{aligned}$$

where $\underline{W}_{p\times k} = \left\{\lambda_{1,j} sign\left(b_{ij}\right)\right\} = \begin{bmatrix} \lambda_{1,1} sign\left(b_{11}\right) & \lambda_{1,2} sign\left(b_{12}\right) & \cdots & \lambda_{1,k} sign\left(b_{1k}\right) \\ \lambda_{1,1} sign\left(b_{21}\right) & \lambda_{1,2} sign\left(b_{22}\right) & \cdots & \lambda_{1,k} sign\left(b_{2k}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1,1} sign\left(b_{p1}\right) & \lambda_{1,2} sign\left(b_{p2}\right) & \cdots & \lambda_{1,k} sign\left(b_{pk}\right) \end{bmatrix}$, 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,

$$\begin{aligned} \frac{\partial f_{X,Y}\left(\underline{B}|\underline{A}, \underline{\beta}^*, m, \lambda, \underline{\lambda}_1\right)}{\partial \underline{b}_j} &= \frac{\partial}{\partial \underline{b}_j}\left\{-2\left(m_1\beta_j \underline{y}^T \underline{X}\underline{b}_j - 0.5 m_1 \beta_j^2 \underline{b}_j^T \underline{X}^T \underline{X}\underline{b}_j - m_1\beta_0\beta_j \underline{1}^T \underline{X}\underline{b}_j\right)\right. \\ &\quad \left. - 2\left(m_2 \underline{a}_j^T \underline{X}^T \underline{X}\underline{b}_j - 0.5 m_2 \underline{b}_j^T \underline{X}^T \underline{X}\underline{b}_j\right) + \lambda \underline{b}_j^T \underline{b}_j + \underline{1}^T \underline{w}_j \otimes \underline{b}_j\right\}, \\ &\quad \text{where } \underline{w}_j \text{ is the } j\text{th column of } \underline{W} \\ &\Rightarrow \underline{b}_j = m_1\beta_j \underline{D}\underline{X}^T \underline{y}_C + m_2 \underline{D}\underline{X}^T \underline{X}\underline{a}_j - 0.5\lambda_{1,j} \underline{D} sign(\underline{b}_j), \end{aligned}$$

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 ith row of $\underline{D}$, and taking $\underline{D} = \left\{d_{ij}\right\}$,

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

In Eq. (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 jth column of $\underline{A}$) is given by

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

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 Eq. (7) through $sign(b_{ij})$. Soft thresholding is then implemented to isolate b_{ij} in Eq. (7). The estimated value of b_{ij} via soft thresholding is given by

$$\begin{aligned} \hat{\underline{b}}_j &= \begin{cases} \underline{w}_j - 0.5\lambda_{1,j} sign(w_{ij}), & \text{if } \left|\underline{w}_j\right| > 0.5\lambda_{1,j}\underline{d} \\ 0, & \text{if } \left|\underline{w}_j\right| \le 0.5\lambda_{1,j}\underline{d} \end{cases} \\ &= sign(\underline{w}_j) \otimes \left(\left|\underline{w}_j\right| - 0.5\lambda_{1,j}\underline{d}\right)_+, \end{aligned} \tag{9}$$

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 $\underline{d} = \{d_{ii};\ i = 1, 2, \ldots, p\}$ 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 $\underline{X}$, $\underline{X} = \underline{USV}^T$

(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 $\underline{\beta}^* = \left(\beta_0, \underline{\beta}\right)$ as

$$\beta_0 = \bar{y}, \quad \text{and}$$
$$\underline{\beta} = \left(\underline{B}^T \underline{X}^T \underline{X}\underline{B}\right)^{-1} \underline{B}^T \underline{X}^T \underline{y}_C, \quad \text{where } \underline{y}_C = \underline{y} - \bar{y}\underline{1}.$$

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

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

$\underline{d} = \{\, d_{ii};\ i = 1, 2, \ldots, p\}$ from the diagonals of $\underline{D}$, and
$sign^*\left(b_{ij}\right)$ 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 $\underline{w}_j^* = \left(w_{1j}^*, w_{2j}^*, \ldots, w_{pj}^*\right)$ such that for $i = 1,\ 2,\ \ldots,\ p$ and a given j,
$w_{ij}^* = m_1\beta_j\underline{D}^{*(i)}\underline{X}^T\underline{y}_C + m_2\underline{D}^{*(i)}\underline{X}^T\underline{Xa}_j - 0.5\lambda_{1,j}abs\left(\sum_{l\neq i} d_{il}^*sign^*(b_{lj})\right),$
$\underline{D}^* = \left(\underline{X}^T\underline{X}\right)^{-1}\left(m_1\beta_j^2 + m_2\right)^{-1},$
$\underline{d}^* = \left\{ d_{ii}^*;\ i = 1, 2, \ldots, p\right\}$ from the diagonals of $\underline{D}^*$, and
$sign^*\left(b_{ij}\right)$ is evaluated for b_{ij}'s from the previous iteration

Option4 (LaNS4): "High Sparsity Method"

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

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

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

(vi) Compute for $\underline{\beta} = \left(\underline{B}^T\underline{X}^T\underline{XB}\right)^{-1}\underline{B}^T\underline{X}^T\underline{y}_C$
(vii) Solve for the coefficients of the individual $\underline{X}$'s as $\underline{\theta} = \underline{B\beta}$
(viii) Solve for the SVD of $\underline{X}^T\underline{XB}$, say $\underline{X}^T\underline{XB} = \underline{WEZ}^T$, and take $\underline{\hat{A}} = \underline{WZ}^T$
(ix) Repeat steps (v)–(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 labeled 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 Eq. (10) whereas LaNS3 iteratively utilizes Eq. (11) to come-up with sparse sets of $\underline{b}_j$'s. Both Eqs. (10) and (11) are similar to the soft thresholding operator (on ℓ_1 norm) of the form $b = sign(w)\ (|w| - t)_+$, where b is approximated by a function (or a constant) w and a tuning parameter t that regulates the thresholding. In Eq. (10), the operator involves the function $\sum_{l\neq i} d_{il}sign(b_{lj})$ which may be negative thereby resulting in "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 Eq. (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 $\lambda_{1,j}\underline{d}$), changes 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 adjust 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 are compared to the different criteria and/or procedures for regression modeling, dimension reduction, and variable selection.

Given the general optimization problem in Eq. (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 in 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 Eq. (6), with $m_1 = m_2 = 1$, and when optimization is approached in two stages – first by finding $\left(\hat{\underline{A}}, \hat{\underline{B}}\right) = \arg\min_{\underline{A},\underline{B}} \left\{ \left\| \underline{X} - \underline{X}\underline{B}\underline{A}^T \right\|_F^2 + \lambda \left\| \underline{B}^T \right\|_F^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \underline{b}_j \right\|_1 \right\}$ 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 in further sparsity of the independent variables.

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

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

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

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 Eq. (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 $m_2 = \frac{\alpha}{\|\underline{X}^*\|^2}$ (and is also equivalent to PCR); (b) an adjusted set of PCovR estimates at $\alpha \in (0, 1)$ when $(m_1, m_2) = g\left(\frac{1-\alpha}{\|\underline{y}\|_F^2}, \frac{\alpha}{\|\underline{X}\|_F^2}\right)$; 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: $V_1 \sim N\left(300,\ 300^2\right)$; $V_2 \sim N\left(300,\ 290^2\right)$, 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 $X_1,\ X_2,\ \ldots,\ X_{1000}$ are each derived as

$$\begin{aligned} X_j &= V_1 + \varepsilon^{(j)}, && \text{for } j = 1,\ 2,\ \ldots,\ 10 \\ X_j &= V_2 + \varepsilon^{(j)}, && \text{for } j = 11,\ 12,\ \ldots,\ 20 \\ X_j &= V_3 + \varepsilon^{(j)}, && \text{for } j = 21,\ 22,\ \ldots,\ 1000 \end{aligned}$$

where the $\varepsilon^{(j)}$'s are independent and such that $\varepsilon^{(j)} \sim N\,(0,\ 10)$ for $j = 1,\ 2,\ \ldots,\ 1000$. Given the formulation of the independent variables, the pairs of variables from any of the sets $C_1 = (X_1,\ X_2,\ \ldots,\ X_{10})$, $C_2 = (X_{11},\ X_{12},\ \ldots,\ X_{20})$ and $C_3 = (X_{21},\ X_{22},\ \ldots,\ X_K)$ 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 $X_1,\ X_2,\ \ldots,\ X_{1000}$. Given the values on the ith observation, $X_{i1}, X_{i2}, \ldots, X_{i1000}$, the dependent variable Y_i is computed as

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

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

Note that the parameters $\beta_1,\ \beta_2,\ \ldots,\ \beta_{1000}$ are specified to control for the relative contributions of the independent variables $X_1,\ X_2,\ \ldots,\ X_{1000}$ to the dependent variable Y. Depending on the scenario, those simulated to be "important

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	1000	55%	35%	10%
2	NHD	40	35%	10%	55%
	HD	1000	35%	10%	55%

predictors" have coefficients ranging from 1.25 to 2.75, and those "non-important predictors" have coefficients ranging from 0.35 to 0.45. Accordingly, 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 $X_1, X_2, \ldots, X_{1000}$ 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 $X_1, X_2, \ldots, X_{40}$ are considered. Across all scenarios, a total of 100 observations are considered. Specifications of the different simulation settings are summarized in Table 1.

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_is in C_j.

6.1. Comparison of different methods

The different LaNS options are then 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.

For the different simulation studies, the values of m_1 and m_2 as well as λ in the LaNS criterion are set to 1. The values for $\lambda_{1,1}$'s and $\lambda_{1,2}$'s vary across scenarios, ranging from 0.05 to 10^6. Also, 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, and EN, 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 give 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 in 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 $SSPE = \sum \left(y_i - \hat{y}_i\right)^2$, 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 $BIC = \frac{MSPE}{\mathrm{Var}(\underline{y})} + NNZ\frac{\log(n)}{n}$, where $MSPE = \frac{1}{n}SSPE$, $\mathrm{Var}\left(\underline{y}\right)$ 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. All comparisons and/or assessments will be based on training (in-sample) and test (out-sample) data sets.

Table 2
Summary of SSPE and BIC in-sample measures 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	40.00	172,451.7	1.844	172,451.7	1.844	10.00	10.00	20.00
PCR	40.00	536,405.4	1.849	172,451.7	1.844	10.00	10.00	20.00
PCovR(0.85)	40.00	430,419.0	1.848	172,451.7	1.844	10.00	10.00	20.00
PCovR(0.5)	40.00	263,959.8	1.846	172,451.7	1.844	10.00	10.00	20.00
PCovR(0.15)	40.00	184,155.2	1.844	172,451.7	1.844	10.00	10.00	20.00
LaNS1	40.00	175,841.6	1.844	172,451.7	1.844	10.00	10.00	20.00
LaNS2	24.80	234,021.2	1.145	202,241.1	1.145	8.20	8.20	8.40
LaNS3	17.40	453,589.1	0.807	270,043.2	0.805	7.20	6.40	3.80
LaNS4	11.80	544,990.8	0.551	380,829.8	0.548	4.80	3.80	3.20
SPCR	22.00	82,620,374.0	2.081	25,352,435.3	1.314	6.00	10.00	6.00
LASSO	11.60	3,665,814.8	0.580	447,888.5	0.540	7.40	4.20	0.00
EN(0.01)	11.80	5,010,559.2	0.606	454,926.8	0.549	8.00	3.80	0.00
EN(100)	11.80	17,344,710.2	0.766	12,560,660.8	0.706	10.00	0.00	1.80

Table 3
Summary of SSPE and BIC out-sample measures under the NHD case.

	Using model		After OLS		%Diff SSPE vs. OLS		
	SSPE	BIC	SSPE	BIC	Min	Median	Max
OLS	157,469.8	1.844	157,469.8	1.844	0	0	0
PCR	531,351.0	1.848	157,469.8	1.844	184.9	228.0	333.0
PCovR(0.85)	452,938.7	1.847	157,469.8	1.844	140.4	198.2	248.3
PCovR(0.5)	352,461.0	1.846	157,469.8	1.844	100.1	133.3	152.5
PCovR(0.15)	373,737.3	1.846	157,469.8	1.844	119.0	132.2	174.6
LaNS1	247,855.4	1.845	157,469.8	1.844	1.7	3.6	183.0
LaNS2	347,074.6	1.146	198,016.3	1.144	38.6	80.7	106.7
LaNS3	611,697.9	0.808	277,975.6	0.804	100.9	126.3	127.7
LaNS4	825,343.0	0.553	449,455.0	0.549	24.6	80.2	142.4
SPCR	92,078,518.6	2.072	25,458,160.9	0.822	33.6	51,847.6	65,883.1
LASSO	4,335,587.2	0.585	432,798.9	0.539	390.2	1010.1	1338.3
EN(0.01)	5,863,502.4	0.613	431,567.0	0.548	574.3	1273.3	1971.3
EN(100)	19,854,482.2	0.773	13,466,158.4	0.699	31.7	51.0	53.3

6.2. Scenario 1

The data for Scenario 1 is generated from a structure where 55%, 35%, and 10% of the dependent variable are 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 is in Tables 2 and 3 for the NHD case, and Tables 4 and 5 for the HD case, discussions of which follow. Note that 5 replicates were generated under the HD case and 5 replicates under the NHD case.

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, both for in-samples and out-samples). 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 25 independent variables and LaNS3 having about 18 on the average. Both LaNS2 and LaNS3 select independent variables coming from all three latent factors. Both LaNS2 and LaNS3 are as good (in terms of SSPE, both in-sample and out-sample) 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 12 independent variables (on the average) coming from all three latent factors and gives the fitted model with highest predictive ability (in- and out-samples), unlike EN(100) which includes almost always all 10 independent variables from the first latent factor and always none from the second latent factor, and unlike EN(0.01) and LASSO which include variables always only 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 from the different methods, and implementing OLS using only the selected variables as predictors, those variables identified by LaNS4 give better or at-par 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

Table 4
Summary of SSPE and BIC in-sample measures 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)	1000.00	700,865.8	46.061	–	–	10.00	10.00	980.00
PCovR(0.5)	1000.00	241,798.6	46.055	–	–	10.00	10.00	980.00
PCovR(0.15)	1000.00	21,697.3	46.052	–	–	10.00	10.00	980.00
LaNS2/LaNS3	31.75	1,163,550.4	1.462	1,163,550.4	1.477	9.25	10.00	12.50
LaNS4	11.20	2,714,623.0	0.549	430,333.8	0.522	7.20	4.00	0.00
SPCR	15.40	72,291,356.9	1.651	14,256,434.2	0.904	8.10	3.00	4.30
LASSO	11.20	4,308,076.8	0.572	395,217.2	0.521	6.00	4.90	0.30
EN(0.01)	11.20	6,926,885.5	0.606	426,881.4	0.522	6.30	4.80	0.10
EN(100)	11.20	38,863,952.1	1.042	14,940,324.8	0.714	9.00	0.00	2.20

Table 5
Summary of SSPE and BIC out-sample measures under the HD case.

	Using model		After OLS		%Diff SSPE vs. OLS		
	SSPE	BIC	SSPE	BIC	Min	Median	Max
PCovR(0.85)	899,615.7	46.063	–	–	–	–	–
PCovR(0.5)	721,383.3	46.061	–	–	–	–	–
PCovR(0.15)	598,508.4	46.059	–	–	–	–	–
LaNS2/LaNS3	1,371,412.4	1.480	256,418.2	1.465	277.3	315.9	914.0
LaNS4	2,942,086.9	0.552	496,252.6	0.522	10.7	155.1	1670.2
SPCR	84,811,786.3	1.728	12,935,722.9	0.853	68.4	809.0	34,885.0
LASSO	4,738,308.8	0.574	476,889.2	0.522	74.6	760.8	1710.7
EN(0.01)	7,287,817.8	0.605	495,588.4	0.522	82.6	1112.8	2289.1
EN(100)	39,205,873.7	0.988	11,514,602.9	0.655	39.6	226.2	445.0

independent variables and at the same time best explains the dependent variable. If identification of a smaller set while maximizing potential dimensionality is of interest, then LaNS4 appears to be a better option than LASSO or EN.

The prediction error of LaNS4 (except for few cases, hence the inflation of SSPE in in-sample and out-sample measures) is comparable to that of OLS using the selected variables from LaNS4 (see %Diff SSPE vs. OLS in Table 3). 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.

For the HD case, note that LaNS2 and LaNS3 were grouped (referred to as LaNS2/3 this point onwards) as the resulting models are sparse but not as sparse as LaNS4. For those with sparse solutions, LaNS2/3 and LaNS4 clearly dominate any of SPCR, LASSO or EN. LaNS4 is even better than PCovR(0.85) in terms of predictive ability. SPCR is the least performing among those with sparse solutions. Interestingly, LaNS2/3 yield solutions that are optimal at the OLS scale, that is, the estimated model by LaNS2/3 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 SPCR or EN(100), and at-par with LASSO or EN(0.01). 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 are 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 5 replicates under the NHD case. Summary of the results is presented in Tables 6 and 7 for the NHD case, and Tables 8 and 9 for the HD case.

For the NHD case, LaNS1 generates non-zero coefficients for all independent variables, with the fitted model comparable with PCovR (at different settings) and PCR. Noticeably, LaNS1 and PCovR(0.15) have on the average nearly the same predictive ability (LaNS1 better in in-sample, and relatively poor in out-sample). 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

Table 6
Summary of SSPE and BIC in-sample measures 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	40.00	150,690.0	1.845	150,690.0	1.845	10.00	10.00	20.00
PCR	40.00	342,863.1	1.849	150,690.0	1.845	10.00	10.00	20.00
PCovR(0.85)	40.00	275,190.3	1.848	150,690.0	1.845	10.00	10.00	20.00
PCovR(0.5)	40.00	184,273.4	1.846	150,690.0	1.845	10.00	10.00	20.00
PCovR(0.15)	40.00	154,084.4	1.845	150,690.0	1.845	10.00	10.00	20.00
LaNS1	40.00	150,690.2	1.845	150,690.0	1.845	10.00	10.00	20.00
LaNS2	27.40	245,955.3	1.266	158,361.8	1.265	7.00	6.20	14.20
LaNS3	20.00	268,663.9	0.925	207,316.1	0.925	6.00	4.40	9.60
LaNS4	14.80	860,264.8	0.700	261,370.6	0.687	2.40	1.60	10.80
SPCR	18.00	56,227,513.0	1.939	501,602.9	0.839	10.00	0.00	8.00
LASSO	14.20	2,158,710.0	0.693	397,509.0	0.662	7.20	0.40	6.60
EN(0.01)	14.80	5,829,492.7	0.787	404,635.7	0.690	7.60	0.20	7.00
EN(100)	14.80	38,978,181.2	1.436	464,950.9	0.691	9.80	0.00	5.00

Table 7
Summary of SSPE and BIC out-sample measures under the NHD case.

	Using model		After OLS		%Diff SSPE vs. OLS		
	SSPE	BIC	SSPE	BIC	Min	Median	Max
OLS	154,131.7	1.846	154,131.7	1.846	0	0	0
PCR	342,863.1	1.850	154,131.7	1.846	104.0	121.8	145.4
PCovR(0.85)	328,592.3	1.850	154,131.7	1.846	84.1	129.9	137.8
PCovR(0.5)	317,739.1	1.850	154,131.7	1.846	78.0	101.1	143.1
PCovR(0.15)	369,297.4	1.851	154,131.7	1.846	102.4	129.9	204.2
LaNS1	487,533.1	1.854	154,131.7	1.846	130.4	165.5	457.6
LaNS2	398,942.4	1.271	210,364.1	1.267	59.8	77.1	119.8
LaNS3	382,536.7	0.930	243,497.6	0.927	33.3	57.9	86.3
LaNS4	1,112,254.7	0.707	418,084.7	0.692	30.7	90.8	699.1
SPCR	46,433,664.6	1.937	606,418.9	0.843	4262.0	8619.0	12,148.2
LASSO	1,935,928.0	0.698	456,309.9	0.665	30.6	100.8	706.0
EN(0.01)	4,845,111.8	0.792	470,861.4	0.693	23.2	255.7	2070.1
EN(100)	32,063,990.2	1.438	516,386.6	0.694	3841.9	6529.1	7611.0

Table 8
Summary of SSPE and BIC in-sample measures 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)	1000.00	228,044.9	46.057	–	–	10.00	10.00	980.00
PCovR(0.5)	1000.00	60,793.7	46.053	–	–	10.00	10.00	980.00
PCovR(0.15)	1000.00	4337.0	46.052	–	–	10.00	10.00	980.00
LaNS2/3	32.20	418,762.6	1.492	192,150.0	1.487	8.80	10.00	13.40
LaNS4	14.40	914,850.3	0.670	597,226.1	0.677	4.80	0	9.60
SPCR	20.00	49,897,762.3	2.030	454,282.3	0.931	10.00	0	10.00
LASSO	14.40	5,440,164.8	0.771	352,627.4	0.662	6.00	0.50	7.90
EN(0.01)	14.40	8,414,747.7	0.844	427,720.3	0.673	7.30	0.20	6.90
EN(100)	14.40	38,099,563.5	1.456	43,966,501.5	1.594	9.22	0	4.56

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 procedures give far better results than any of SPCR, LASSO, or EN (in- and out-samples). SPCR gives the highest prediction error indicating that variable selection via the SPCA may not give the best set of highly predictive independent variables.

For the HD case, LaNS2/3 and LaNS4 give sparse models with only about 32 and 14 independent variables, respectively. LaNS2/3 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 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 on the average a relatively higher prediction error compared to those of SPCR, LASSO or EN.

Table 9
Summary of SSPE and BIC out-sample measures under the HD case.

	Using model		After OLS		%Diff SSPE vs. OLS		
	SSPE	BIC	SSPE	BIC	Min	Median	Max
PCovR(0.85)	383,556.7	46.060	–	–	–	–	–
PCovR(0.5)	353,664.4	46.060	–	–	–	–	–
PCovR(0.15)	345,410.5	46.059	–	–	–	–	–
LaNS2/3	87,328,317.2	3.405	179,101.3	1.487	36,986.7	54,546.9	59,116.2
LaNS4	1,078,192.3	0.687	620,695.7	0.677	10.8	89.4	167.1
SPCR	51,661,549.0	2.031	417,825.3	0.930	9784.4	12 190.9	14,644.8
LASSO	5,779,830.8	0.775	380,058.2	0.662	153.6	540.0	3127.5
EN(0.01)	8,833,396.3	0.848	436,585.6	0.673	140.0	785.1	4274.3
EN(100)	37,308,545.9	1.443	558,541.0	0.647	3764.3	7302.7	9756.5

Table 10
Comparison of different methods to predict QoLI.

Modeling method	No. of predictors	Remarks/identified determinants of QoLI	OLS Adj R^2	OLS SSPE	No. of Significant predictors at pval <0.2; and resulting Adj R^2 and SSPE using this set
LaNS4	8	Includes variables relating to Environment (1), Lifestyle (2), Health care (1), Health status (2), Health policy (1) and Morbidity (1)	0.7444	19,025.16	7: 0.7447; 19,183.12
LASSO	8	Includes variables relating to Lifestyle (3), Health status (2) and Health policy (3)	0.7239	20,551.33	6: 0.7229; 21,007.65
EN(0.01)	8	Includes variables relating to Lifestyle (1), Health status (3) and Health policy (4)	0.7160	21,142.50	4: 0.7166; 21,880.78
EN(100)	8	Includes variables relating to Health status (6) and Health policy (2)	0.6722	24,405.71	5: 0.6326; 28,106.63
SPCR	8	Includes variables relating to Health policy (8)	0.4796	38,742.16	5: 0.4749; 40,172.46
PCovR	–	No results due to singularity issues	–	–	–

7. Illustration: quality of life across countries

While quality of life is a multidimensional phenomenon, in this example we focused on the aspect 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 related to mortality. Using SPCA (Zou et al., 2006), the first sparse component (SPC) accounts for 56.05% of the variance, with 35 non-zero loadings. With negative loadings for mortality indicators, the SPC was aptly labeled as Quality of Life Index or QoLI, and was re-scaled so that values will range between 0 and 100 for ease of interpretation. High values of QoLI were observed for Israel, Japan, Republic of Korea, France, and Netherland. On the other hand, low values of QoLI were observed for Turkmenistan, Somalia, Central African Republic, Afghanistan, and Uzbekistan.

Given QoLI (y), we then aim to identify its determinants that may help various governments in prioritizing limited resources and focus on those that really have impact on the quality of life. Due to high dimensionality of possible predictors (106 variables related to environment, lifestyle, health status, health policy, health care, and morbidity; 117 countries analyzed), we simultaneously reduce the dimensionality of the predictors and choose the important variables using the LaNS algorithm. The tuning parameters were set as follows: $m_1 = 1$; $m_2 = 1$; $k = 2$; $\lambda = 1$. For $\lambda_{1,j}$, any values from 0.15 to 255 are all feasible to induce sparsity and the resulting determinants are meaningful.

Table 10 presents the results of modeling QoLI. It is evident that the data suffers from multicollinearity and/or ill-conditioning, and so PCoVR is not computationally possible. From the results of all other modeling approaches, QoLI may be explained primarily by health policies (the amount of spending by the government on health), health status (proportion with immunization) and lifestyle (sanitation, alcohol consumption). LaNS, however, also identifies the environmental (pollution, radiation), health care (average blood pressure, average BMI) and morbidity (percent prevalence of disorders, injuries or diseases) aspects. SPCR is the least successful in terms of the identification of predictors from diverse aspects (identifying variables only from one dimension), as well as in terms of predictive ability (having the lowest adjusted R^2). The OLS model via LaNS4 is the most predictive, having the highest adjusted R^2 value as well as the lowest SSPE value. Also, 7 out of the 8 determinants identified via LaNS4 are significant at 0.20 level, in contrast, only 4–6 determinants are significant from each of LASSO's, EN's, and SPCR's 8 determinants. Clearly, LaNS4 yields a model for QoLI that is most predictive and at the same time captures the most dimensionality of QoLI's determinants.

Table 11
OLS regression coefficients and VIFs of determinants of QoLI via LaNS.

Determinant	Estimate	*p*-value	VIF
(Intercept)	6.5038	0.7754	–
UV radiation	0.0066	<0.0001	3.4429
Consumption of wine per capita among ages 15+(liters)	0.2446	0.0132	1.6153
Population using improved drinking-water sources (%)	0.5451	0.0002	3.1106
Percentage with raised age-standardized BP	−1.4891	<0.0001	1.4700
Hib immunization coverage among 1-year-olds (%)	0.1162	0.0162	2.1804
Polio immunization coverage among 1-year-olds (%)	0.2666	0.0580	2.5757
Per capita government expenditure on health (US$)	0.0089	<0.0001	3.7566
Prevalence of alcohol-use disorders among women ages 15+(%)	−2.3592	0.3458	1.7220

The OLS regression model using the 8 determinants identified via LaNS accounted for 74% of the total variation in QoLI, which is reasonably high given that these were chosen from the original set of 106 predictors, and given that these determinants also capture the most dimensionality (aspects) in QoLI. The estimated regression coefficients (and the intercept) as well as the variance inflation factors (or VIFs) are given in Table 11. Further regression diagnostics suggest that, at 0.05 level of significance, there are no issues of multicollinearity and nonlinearity, and that the error variances are uncorrelated and homoscedastic. All these demonstrate the validity and usefulness of the fitted linear model for QoLI using the 8 determinants derived via LaNS4.

8. Conclusions

The LaNS procedure estimates a model that is sparse while it also exhibits at-par to 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 achieved 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 $p \gg n$, the fitted LaNS models with sparse regression coefficients capture "representatives" from the different latent factors, as evident from the simulations and real data example. 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. Such properties of the LaNS procedure are evident from the different simulation settings and scenarios. Further study may be needed for cases with large number of latent factors, relatively lopsided ratio of variables to observations, different natures of the response and predictor variables, and nonlinear relationships.

References

Candes, E., Tao, T., 2007. The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35 (6), 2313–2351.

Chatterjee, S., Hadi, A., 2006. Regression Analysis by Example, fourth ed. In: Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., Hoboken, New Jersey.

Chen, L., Huang, J., 2012. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Amer. Statist. Assoc. 107 (500), 1533–1545.

Chipman, H., George, E., McCulloch, R., 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298.

Chipman, H., Gu, G., 2005. Interpretable dimension reduction. J. Appl. Stat. 32 (9), 969–987.

Chun, H., Keles, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (1), 3–25.

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

De Jong, S., Kiers, H.A.L., 1992. Principal covariates regression. Chemometr. Intell. Lab. Syst. 14, 155–164.

Draper, N., Smith, H., 1998. Applied Regression Analysis, third ed. In: Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., Hoboken, New Jersey.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 (3), 211–218.

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

Foucart, T., 2000. A decision rule for discarding principal components in regression. J. Statist. Plann. Inference 89 (1), 187–195.

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

George, E.I., Oman, S.D., 1996. Multiple-shrinkage principal component regression. Statistician 45 (1), 111–124.

Goldenshluger, A., Tsybakov, A., 2001. Adaptive prediction and estimation in linear regression with infinitely many parameters. Ann. Statist. 29 (6), 1601–1619.

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

Hwang, J., Nettleton, D., 2003. Principal components regression with data-chosen components and related methods. Technometrics 45, 70–79.

Izenman, A.J., 1975. Reduced-rank regression for the multivariate linear model. J. Multivariate Anal. 5, 248–264.

Jolliffe, I., 1982. A note on the use of principal components in regression. J. Appl. Stat. 31 (3), 300–303.

Klinger, A., 2001. Inference in high dimensional generalized linear models based on soft thresholding. J. Roy. Statist. Soc. 63 (2), 377–392.

Kosfeld, R., Lauridsen, J., 2008. Factor analysis regression. Statist. Papers 49 (4), 653–667.
Lee, T.S., 1987. Algorithm AS 223: Optimum ridge parameter selection. J. Roy. Statist. Soc. Ser. C 36 (1), 112–118.
McDonald, G.C., Galarneau, D.I., 1975. J. Amer. Statist. Assoc. 70 (550), 407–416.
Ravikumar, P., Liu, H., Lafferty, J., Wasserman, L., 2007. SpAM: sparse additive models. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, Vol. 20. MIT Press, Cambridge, MA, pp. 1201–1208.
Reinsel, G.C., Velu, P.R., 1998. Multivariate Reduced-Rank Regression: Theory and Applications. Springer, New York.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
WHO, 2016. WHO Website available at: http://www.who.int/gho/en/.
Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P.R. (Ed.), Multivariate Analysis. Academic Press, New York, pp. 391–420.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.
Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Statist. 15 (2), 265–286.
Zou, H., Hastie, T., Tibshirani, R., 2007. On the "Degrees of Freedom" of the LASSO. Ann. Statist. 35 (5), 2173–2192.