Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

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 xxxx

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
0167-9473/

Please cite this article in press as: Lansangan, J.R.G., Barrios, E.B., Simultaneous dimension reduction and variable selection in modeling high dimensional data. Computational Statistics and Data Analysis (2017), 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 $\underset{\sim}{y}_{nx1}$, a vector of observations from a dependent variable and $\underset{\sim}{X}_{n\times p} = \left[\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\right]^T$, a matrix of observations on p variables for the n subjects. The hypothesized model takes the form $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, with $\underset{\sim}{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)^T$ and $\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$. 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 $\underset{\sim}{\beta}$ is $\hat{\underset{\sim}{\beta}} = \left(\underset{\sim}{X}^T\underset{\sim}{X}\right)^{-1}\underset{\sim}{X}^T\underset{\sim}{y}$ is optimal (Gauss–Markov Theorem) provided that $n > p$.

When $p \gg n$, the estimator $\hat{\underset{\sim}{\beta}}$ is not unique since high dimensionality of the data matrix leads to the singularity of the Gram matrix $\underset{\sim}{X}^T\underset{\sim}{X}$ (Chatterjee and Hadi, 2006; Draper and Smith, 1998). Similarly, the estimator $\hat{\underset{\sim}{\beta}}$ is unstable, i.e., the estimators for $\underset{\sim}{\beta}$ 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 $\underset{\sim}{X}^T\underset{\sim}{X}$ are also invalid (Draper and Smith, 1998). In general, the OLS estimator $\hat{\underset{\sim}{\beta}}$ 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,

Please cite this article in press as: Lansangan, J.R.G., Barrios, E.B., Simultaneous dimension reduction and variable selection in modeling high dimensional data. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.03.015