Signal Processing 131 (2017) 1–7

Contents lists available at ScienceDirect

Signal Processing

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

Short communication

High-dimensional variable selection in regression and classification with missing data

Qi Gao, Thomas C.M. Lee *

Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form
5 July 2016
Accepted 11 July 2016
Available online 15 July 2016

Keywords:
Adaptive lasso
Logistic regression
Low rank recovery
Matrix completion

ABSTRACT

Variable selection for high-dimensional data problems, including both regression and classification, has been a subject of intense research activities in recent years. Many promising solutions have been proposed. However, less attention has been given to the case when some of the data are missing. This paper proposes a general approach to high-dimensional variable selection with the presence of missing data when the missing fraction can be relatively large (e.g., 50%). Both regression and classification are considered. The proposed approach iterates between two major steps: the first step uses matrix completion to impute the missing data while the second step applies adaptive lasso to the imputed data to select the significant variables. Methods are provided for choosing all the involved tuning parameters. As fast algorithms and software are widely available for matrix completion and adaptive lasso, the proposed approach is fast and straightforward to implement. Results from numerical experiments and applications to two real data sets are presented to demonstrate the efficiency and effectiveness of the approach.

1. Introduction

High-dimensional data are encountered frequently nowadays in diverse fields of study and applications including biomedical research, finance, machine learning and signal processing. With rapid development of technology and increasing complexity of the contemporary scientific problems, both the number of instances and variables have been growing unprecedentedly. In particular, variable selection in the high-dimensional setting has drawn great attention and become a fertile field of research in recent years. Many methods have been proposed to solve this problem for both regression and classification; e.g., see [1,3,4,8,13,24,28,29].

In this paper we are interested in high-dimensional variable selection with the presence of missing data while the missing fraction can be fairly large (e.g., 50%). Although missing data occur frequently in various signal processing and statistical applications [25–27,30], there is not much work considering this problem in the existing literature, especially for classification. Some notable exceptions include [24] where the expectation–maximization (EM) algorithm is applied to estimate the inverse covariance matrix in sparse linear regression with missing data. However, like many other EM based solutions for complex problems, this EM-algorithm is computationally intensive and the missing fraction is usually relatively small. In [16] an algorithm is developed based on projected gradient descent for the cases of noisy and/or missing data in high-dimensional sparse linear regression, but again the missing fraction they study is relatively small.

In order to handle both high-dimensionality and high missing fraction, in this paper a new procedure that combines matrix completion techniques and adaptive lasso is developed for solving this variable selection problem. The proposed procedure is straightforward to implement, computationally fast, and capable of producing promising empirical results.

The rest of our paper is organized as follows. Section 2 provides a detailed description of the problem of interest, and presents the proposed method for variable selection in high-dimensional regression with missing data. The case for classification using logistic regression is presented in Section 3. The empirical performances of the proposed methods are evaluated by simulation experiments and a real data application in Sections 4 and 5 respectively. Lastly, concluding remarks are given in Section 6.

2. Variable selection for high-dimensional regression with missing data

We first illustrate our methodology with high-dimensional regression. Suppose observed are p predictors, denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_p$. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the response vector of n observations, $\mathbf{X}$ be the corresponding $n \times p$ design matrix and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ be the regression coefficients. We allow the

[*] Corresponding author.
E-mail addresses: qigao@ucdavis.edu (Q. Gao), tcmlee@ucdavis.edu (T.C.M. Lee).

http://dx.doi.org/10.1016/j.sigpro.2016.07.014
0165-1684/© 2016 Elsevier B.V. All rights reserved.

possibility of $p \gg n$, the so-called "large p small n" situation. With a linear regression model we have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ is the vector of random errors and $\mathbf{I}_n$ represents the identity matrix of size n. Since $p \gg n$, a sparsity assumption has to be imposed; i.e., many components of $\boldsymbol{\beta}$ are exactly zero or negligibly small. Let s denote the number of significant predictors, in practice we usually have $s < O(p)$ so that the full model can be well approximated by a much smaller submodel. With this assumption, an important problem is to determine which predictors are significant, and to estimate their corresponding regression coefficients.

This paper considers a more challenging and frequently encountered version of this problem: it allows some entries of $\mathbf{X}$ that are missing completely at random. Our proposed solution to this problem consists of two major stages:

1. Use a procedure to impute the missing entries of $\mathbf{X}$; denote the imputed $\mathbf{X}$ as $\hat{\mathbf{X}}$.
2. Given $\mathbf{y}$ and $\hat{\mathbf{X}}$, apply a procedure to select the significant predictors as well as estimating their coefficients.

Since we would like to allow the case when $p \gg n$ with a large n, the choices of the above procedures are limited. After extensive methodological consideration and numerical investigation, we propose using matrix completion for the first stage and adaptive lasso for the second. A full description of our choices is given below. As to be seen, the overall method is very fast and straightforward to implement. We stress that in our proposal the two stages are *not* executed independently, as the choice of the tuning parameters in Stage 1 requires information from Stage 2, and vice versa.

2.1. Missing value imputation using matrix completion

This subsection discusses our method for imputing the missing entries of $\mathbf{X}$. For clarity, we occasionally use a subscript to denote the size of a matrix; e.g., $\mathbf{X}_{n\times p}$.

Perhaps the most well-known approach for handling missing data problems is the EM-algorithm and its variants. However, these model-based methods are not practical for the present problem as they can be computationally demanding and hard to incorporate large missing fraction, especially under the "large p small n" situation. Therefore, we take a different path and assume that $\mathbf{X}$ has a low rank structure. That is, $\mathbf{X}$ can be well approximated by the product of two matrices $\mathbf{V}$ and $\mathbf{G}$ such that $\mathbf{X}_{n\times p} \approx \mathbf{V}_{n\times r}\mathbf{G}_{r\times p}$ with the rank r of $\mathbf{X}$ satisfies $r \ll \min(n, p)$. In practice, $r < \sqrt{\min(n, p)}$ when $p > \sqrt{n}$. Under this assumption, many fast matrix completion algorithms can be applied to quickly impute the missing entries of $\mathbf{X}$; e.g., see [6,10,12,15,17–19].

In this paper we use the `softImpute-ALS` procedure proposed by [10] which is a relatively new matrix completion algorithm and has the advantage of fast computation compared with other methods. This procedure completes large matrices efficiently by combining two popular methodologies: nuclear-norm-regularized matrix approximation and maximum-margin matrix factorization. Let Ω be the index set of the non-missing entries of $\mathbf{X}$; i.e. $\Omega = \{(i, j)\colon \text{the } ij\text{th entry } X_{ij} \text{ is observed}\}$. Also let $P_\Omega(\mathbf{X})$ be the projection of the $n \times p$ matrix which preserves the non-missing entries of $\mathbf{X}$ and replaces the missing entries with 0. Similarly, $P_\Omega^\perp$ denotes the projection onto the complement of Ω. Then `softImpute-ALS` solves the following minimization problem:

$$\underset{\mathbf{A},\mathbf{B}}{\text{minimize}}\left\{ \|P_\Omega(\mathbf{X} - \mathbf{A}\mathbf{B}^T)\|_F^2 + \eta(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2) \right\}, \tag{1}$$

where $\mathbf{A}_{n\times r}$ and $\mathbf{B}_{r\times p}^T$ are each of rank at most $r \le \min(n, p)$, and $\|\cdot\|_F$ is the Frobenius norm. Denote the minimizers as $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. The imputed $\mathbf{X}$ is then obtained as $\hat{\mathbf{X}} = P_\Omega(\mathbf{X}) + P_\Omega^\perp(\hat{\mathbf{A}}\hat{\mathbf{B}}^T)$. In all our numerical work we use the `R` package `softImpute` to implement this algorithm. A rank r for $\mathbf{X}$ is specified and η is fixed as a small number (0.5) so as not to over-penalize. Next we discuss the choice of r.

Methods are available for selecting the rank r for matrix completion methods, such as the BIC criterion of [18]. However, these methods are suboptimal for our problem as the information in $\mathbf{y}$ is not utilized in the selection. In the context of multivariate response regression models, the Rank Selection Criterion (RSC) is introduced by [2] for selecting the rank of the coefficient matrix estimates. Inspired by this, we propose choosing r as the minimizer of the following modified RSC:

$$\text{RSC}(r) = \|\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\lambda,r}\|_F^2 + \mu r, \tag{2}$$

where $\hat{\boldsymbol{\beta}}_{\lambda,r}$ is an estimate for $\boldsymbol{\beta}$ (to be discussed below). The quantity μ is a tuning parameter and by theoretical consideration the following lower bound is derived by [2]: $\mu^{1/2} > \hat{\sigma}\{\text{ncol}(\hat{\boldsymbol{\beta}})^{1/2} + \text{rank}(\hat{\boldsymbol{\beta}})^{1/2}\}$ where $\hat{\sigma}$ is an estimate for the random noise standard deviation σ which can be estimated by the standard deviation of the residuals $\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\lambda,r}$ and ncol stands for the number of columns. We use the smallest μ that satisfies this bound; i.e., $\mu = 4\hat{\sigma}^2$. We use the RSC in [2] as a starting point to develop a method for choosing r for the problem that we consider, and it turns out this modified RSC possesses very good empirical properties as shown later in numerical experiments. Now we are ready to discuss the problem of selecting significant predictors given an imputed $\hat{\mathbf{X}}$.

2.2. Variable selection with adaptive lasso

As mentioned in the introduction, many methods have been developed for simultaneous variable selection and parameter estimation for the high-dimensional regression problem. Virtually any of these methods could be used here for our problem, but we recommend using the adaptive lasso of [31] for the following reasons. First, its theoretical properties are well studied and it has been shown for example by [11] to perform extremely well for high-dimensional problems and satisfy the oracle properties defined by [7]. Therefore, we believe that it is a reliable option for our high-dimensional variable selection problem. Second, fast algorithms and software exist for computing its solutions; e.g., the *R* package `glmnet` of [20]. And lastly, it can be straightforwardly extended to generalized linear models to handle classification problems.

The adaptive lasso estimate $\hat{\boldsymbol{\beta}}_{\lambda,r}$ for $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_{\lambda,r} = \underset{\boldsymbol{\beta}}{\text{argmin}}\left(\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_F^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j| \right), \tag{3}$$

where λ is a tuning parameter and $\hat{w}_1, \ldots, \hat{w}_p$ are pre-set weights. We follow [31] and use ridge regression to obtain these weights; i.e., for all j we set $\hat{w}_j = 1/|\hat{\beta}_{j,\text{ridge}}|$, where $\hat{\beta}_{j,\text{ridge}}$ is the ridge regression estimate of β_j.

To calculate (3), we need to choose λ, and to do so we use the Extended Bayesian Information Criterion (EBIC) of [3]. In brief EBIC is a modified version of BIC that is tailored to "large p small n" problems. By taking into account the complexity of the enlarged model space in addition to the number of the parameters, EBIC is able to control the false discovery rate. It is also shown to be consistent in [3] under some regularity conditions. For the current problem, EBIC is