Neurocomputing 000 (2017) 1–14

Contents lists available at ScienceDirect

Neurocomputing

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

Sparse estimation based on square root nonconvex optimization in high-dimensional data

He Jiang[a,b,*]

[a] School of Statistics, JiangXi University of Finance and Economics, Nanchang, JiangXi, 330013, China
[b] Applied Statistics Research Center, Nanchang, Jiangxi, 330013, China

ARTICLE INFO

Article history:
Received 6 May 2017
Revised 20 November 2017
Accepted 11 December 2017
Available online xxx

Communicated by Dacheng Tao

Keywords:
High-dimensional data
Sparse estimation
Square root loss function
Nonconvex penalty

ABSTRACT

Variable selection plays a dominant role in building forecast models when high-dimensional data appears. However, how to select important variables from a large number of candidate variables efficiently and accurately poses a critical challenge to researchers from various scientific fields including machine learning, genetics, medicine, and finance. In this paper, a novel approach for sparse estimation is proposed. This approach combines the advantages of the square root loss function and nonconvex penalty to obtain an interpretable model with high forecasting accuracy. In particular, the square root loss function facilitates the choice of regularization parameters based on the noise level that is critically difficult to estimate as the number of variables increases; the nonconvex penalty is shown to be superior over the convex penalty in terms of selection consistency especially when the number of variables exceeds the sample size. In computation, a fast and simple-to-implement algorithm is developed with a theoretical guarantee of its convergence. Furthermore, an accelerated gradient method is utilized to further speed up the convergence and the proposed algorithm is proved to scale well to high-dimensional data. Simulation examples with diverse sample sizes, dimensions, correlation coefficients, noise levels, and real data examples focusing on the inbred mouse microarray gene selection problem are exhibited to demonstrate the efficiency and efficacy of this novel approach compared with other existing competitors.

1. Introduction

The arrival of high-dimensional data has presented many difficulties to researchers when establishing forecast models. It is well known that a good forecast model can not only fully utilize data resources in a reasonable way, but also avoid model redundancy. Variable selection is a crucial way to extract vital data information in that important variables are often masked by noise. Thus, it is of great importance to build a more parsimonious and interpretable model using few important variables. Another reason why researchers carry out the variable selection task is the overwhelmingly expensive computational cost (CC) to build the forecasting model employing all the variables. However, it is critically challenging in high-dimensional data analysis especially when the dimensionality d and the sample size n satisfy $d = \mathcal{O}(\exp(n^{\xi}))$ for some $\xi \in (0, 1/2)$ [1]. For instance, in microarray gene expression data analysis, biologists devote themselves to investigating the most efficient way to select few cancer-related genes from hundreds of thousands of candidate genes. Obviously, the performance of ordinary least squares (OLS) regression is not satisfactory in solving this problem since it provides large expected prediction error. Furthermore, traditional methods of best subset selection, such as the Akaike information criterion (AIC) [2], Bayesian information criterion (BIC) [3], and Mallow's Cp [4] that use "ℓ_0 type" penalty, are nondeterministic polynomial (NP) hard. Shrinkage methods based on convex optimization are advocated to be applied as a remedy to reduce the expected prediction error and expensive CC. These methods include, but are not limited to, the ridge regression [5], the nonnegative garrote [6], the least absolute shrinkage and selection operator (LASSO) [7], the bridge regression [8], the least-angle regression (LARS) [9], the elastic net [10], and the adaptive LASSO [11]. However, there are still issues regarding selection inconsistency and computational inefficiency when one applies shrinkage methods to high-dimensional data [12]. In particular, the high correlation between important variables and nuisance variables causes nuisance variables and response variables to be highly correlated. Therefore, variables that are pretty close to each other always result in a coherent design. Taking spectrum analysis as an example, when the resolution on frequency is required to be high, variables associated with frequencies that are

* Corresponding author at: School of Statistics, JiangXi University of Finance and Economics, Nanchang, JiangXi, 330013, China.
E-mail address: jiangsky2005@aliyun.com

https://doi.org/10.1016/j.neucom.2017.12.025
0925-2312/

close to each other necessarily demonstrate high correlation [13]. Under this circumstance, LASSO fails completely in high coherent design showing selection inconsistency. Although nonconvex penalties, including smoothly clipped absolute deviation (SCAD) [14], minimax concave penalty (MCP) [15], and hard ridge (HR) [13], are able to tackle these difficulties and boost the accuracy, they are computationally intractable because of their nonconvex penalty form. Therefore, a novel variable selection procedure is urgently needed to obtain forecasting accuracy and selection consistency simultaneously with less computational time.

2. Related works and contribution of this paper

Much progress has been made over the last decade in sparse estimation techniques such as sure independent screening (SIS). SIS is shown to be successful in removing nuisance variables in a supervised manner roughly. It is usually used without hesitation only when variables of the model are independent of each other or the collinearity of the model is low. However, in high-dimensional settings, there is high probability that variables are highly correlated with each other. Thus, SIS cannot select all the important variables. Forward regression (FR) [16] and the orthogonal greedy algorithm (OGA) [17] are two popular variable selection methods that are proposed in recent decades. Both of them are proved to be selection consistency theoretically. Furthermore, FR is also shown to be successful in interaction selection in ultra-high-dimensional data [18]. In [19], partial residual sure independent screening (PRSIS) was proposed and its application was studied in ultra-high-dimensional longitudinal data. Sparse estimation has also been widely applied in the machine learning community. In [20], a novel approach was advocated using multiview locality-sensitive sparse encoding in image-based 3D human pose recovery. A maximum-margin space encoding algorithm was derived in [21] to learn sets of overcomplete dictionaries. A sparse patch alignment framework was proposed in [22] to extract features for image clustering. The sparse coding extreme learning machine (SCELM) was investigated in [23] to overcome the drawbacks of the traditional extreme learning machine (ELM) in classification problems. A multimodal sparse coding technique was studied in [24] to predict image clicks for improving the performance of a text-based image search. A new method was proposed in [25] for indoor scene classification by embedding semantic information in the weighted hypergraph learning. Robust extreme multi-label learning method is studied in [26] to handle tail labels. A theoretical analysis of estimation performance is also provided. To solve the class imbalance problem, a novel cost-sensitive feature selection method optimizing F-measures instead of accuracy is proposed in [27]. A novel way to handle the incomplete views in multi-view learning is investigated in [28] by exploiting the connections between multiple views.

In addition, the design of an appropriate strategy to select tuning parameters plays an important role in boosting selection consistency. However, this is challenging because accurately estimating the noise level is difficult in high-dimensional data [29,30]. Square root LASSO (SRL) was advocated in [31] and avoids estimating the noise level based on a square root loss function. This facilitates parameter tuning work. The data experiments revealed that SRL is superior to LASSO in terms of forecasting accuracy. The connection between SRL and sparse iterative covariance-based estimation (SPICE) [32] was shown in [33] via designing a covariance-matching-technique-based approach following the idea of SRL in an array-processing application. Furthermore, SRL is extended to the group version by studying group SRL (GSRL) [34] to select groups of variables. Square root convex optimization was studied in [35] and established an oracle inequality with leading constant 1 that directly reveals the benefits of square root loss theoretically. However, to the best of the authors' knowledge, there is no research work on sparse estimation using both the square root loss function and nonconvex penalty in the literature. The main contributions of this paper are as follows:

- A sparse estimation method called square root nonconvex optimization (SRNO) is derived using the square error loss function and nonconvex penalty.
- An efficient and simple-to-implement algorithm for SRNO is designed.
- The convergence of the proposed algorithm is shown theoretically.
- Simulation data examples and high-dimensional gene expression data are considered in the numerical experiments to reveal the advantages of SRNO over other existing methods in terms of forecasting accuracy , selection consistency and computational efficiency.

The rest of this paper is organized as follows. Section 3 represents the general framework of SRNO. Section 4 designs a fast and simple-to-implement algorithm with theoretical guarantee of its convergence and optimality. In Section 5, simulation and real data analysis are shown to exhibit the excellent prediction and selection accuracy.

3. Methodology

Without loss of generality, assume both $\boldsymbol{X}$ and $\boldsymbol{y}$ have been centered such that no intercept term exists in the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d] \in \mathbb{R}^{n\times d}$ is the design matrix consisting of n samples and d variables, $\boldsymbol{\beta}^* \in \mathbb{R}^d$ indicates the true coefficient vector, and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is the noise term with the noise level and identity matrix given by $\sigma^2 > 0$ and $\boldsymbol{I}$, respectively. To select the important variables, a general penalized regression problem is considered as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{2n} + \mathcal{P}(\boldsymbol{\beta}; \lambda) \right\}, \tag{1}$$

where $\mathcal{P}(\boldsymbol{\beta}; \lambda)$ denotes the penalty function that is used to introduce sparsity to the linear model and $\lambda > 0$ represents the regularization parameter that is going to be tuned. The LASSO [7], which is a very famous and powerful variable selection approach, uses an ℓ_1 penalty function $\mathcal{P}(\boldsymbol{\beta}; \lambda) = \lambda\|\boldsymbol{\beta}\|_1$. It is computationally efficient in big data computation because it is convex. However, the selection consistency of the LASSO estimator can only be guaranteed under some regularity assumptions, such as the irrepresentative condition [36], restricted eigenvalue assumption [37], and so on. Actually, this is a universal drawback for all the convex penalty functions. The poor performance of ℓ_1 penalization in variable selection motivates the researchers to apply nonconvex penalties including folded concave penalties such as the SCAD penalty, MCP, and HR penalty. Specifically, the SCAD penalty is given as

$$\mathcal{P}(\boldsymbol{\beta}; \lambda) = \sum_{j=1}^{d} p(\beta_j; \lambda),$$

$$\text{where } p(t; \lambda) = \int_0^{|t|} \left\{ \lambda 1_{\{z\le\lambda\}} + \frac{a\lambda - z}{a-1} 1_{\{z>\lambda\}} \right\}$$

for given constant $a = 3.7$. The MCP is given as

Please cite this article as: H. Jiang, Sparse estimation based on square root nonconvex optimization in high-dimensional data, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.12.025