Computational Statistics and Data Analysis 126 (2018) 78–91

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Balanced estimation for high-dimensional measurement error models

Zemin Zheng [a], Yang Li [a], Chongxiu Yu [b,c], Gaorong Li [c,*]

[a] Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, PR China
[b] College of Applied Sciences, Beijing University of Technology, Beijing 100124, PR China
[c] Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, PR China

ARTICLE INFO

Article history:
Received 7 November 2017
Received in revised form 2 April 2018
Accepted 21 April 2018
Available online 4 May 2018

Keywords:
Balanced estimation
Measurement errors
High dimensionality
Model selection
Nearest positive semi-definite projection
Combined L_1 and concave regularization

ABSTRACT

Noisy and missing data are often encountered in real applications such that the observed covariates contain measurement errors. Despite the rapid progress of model selection with contaminated covariates in high dimensions, methodology that enjoys virtues in all aspects of prediction, variable selection, and computation remains largely unexplored. In this paper, we propose a new method called as the balanced estimation for high-dimensional error-in-variables regression to achieve an ideal balance between prediction and variable selection under both additive and multiplicative measurement errors. It combines the strengths of the nearest positive semi-definite projection and the combined L_1 and concave regularization, and thus can be efficiently solved through the coordinate optimization algorithm. We also provide theoretical guarantees for the proposed methodology by establishing the oracle prediction and estimation error bounds equivalent to those for Lasso with the clean data set, as well as an explicit and asymptotically vanishing bound on the false sign rate that controls overfitting, a serious problem under measurement errors. Our numerical studies show that the amelioration of variable selection will in turn improve the prediction and estimation performance under measurement errors.

1. Introduction

Due to the highly developed technologies, high-dimensional statistical learning has been widely applied in various domains including economics, genomics, molecular biology, machine learning, and so on. To facilitate statistical inference in high dimensions, many powerful methods based on sparse modeling via regularization were proposed, including Tibshirani (1996), Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Candès and Tao (2007), Fan et al. (2009), Zhang (2010), Li et al. (2011), Sun and Zhang (2012), Fan and Lv (2014), Kong et al. (2016), Weng et al. (2017), Li et al. (2017), and among many others. While much of the existing literature focuses on the clean data setting where covariates are fully observed without measurement errors, we often encounter noisy or missing data in real applications such as surveys, votes, and sensor networks. Under such circumstances, naively applying the methods designed for clean data sets to analyze corrupted data will result in inconsistent and unstable estimates, thus leading to incorrect conclusions, especially in the high-dimensional settings (Liang and Li, 2009; Rosenbaum and Tsybakov, 2010). Therefore, it is important to develop approaches for model selection and estimation under measurement errors.

To alleviate the impacts of measurement errors in high dimensions, various methods were proposed for variable selection and parameter estimation with noisy or missing data in recent years. For instance, Liang and Li (2009) suggested to minimize

* Corresponding author.
E-mail address: ligaorong@gmail.com (G.R. Li).

https://doi.org/10.1016/j.csda.2018.04.009
0167-9473/

the penalized least squares after accounting for the covariance structure of the additive errors for partially linear models. Similarly, Ma and Li (2010) incorporated the error information in penalized estimating equations for general parametric and semi-parametric measurement error models. Furthermore, a Lasso-type estimator was proposed in Loh and Wainwright (2012) for high-dimensional sparse regression by replacing the corrupted Gram matrix with unbiased estimate of the true Gram matrix. However, it is challenging to solve the associated optimization problems since the negative likelihood functions often become nonconvex after incorporating the information of measurement errors. To address this issue, Datta and Zou (2017) approximated the unbiased Gram matrix estimate by the nearest positive semi-definite matrix and proposed the convex conditioned Lasso (CoCoLasso), which enjoyed the virtues of convex optimization and nice estimation accuracy in high-dimensional error-in-variables regression. For more discussions on methodology and applications of measurement error models, please refer to Fuller (1987), Carroll et al. (1995), Little and Rubin (2014), Li et al. (2016), and among many others.

Although CoCoLasso achieved computational advantages over most of the existing methods, sparse modeling via Lasso penalty tends to induce biases and yields a larger model than the true one to minimize the prediction errors (Bickel et al., 2009; Hastie et al., 2009). In the clean data setting, the problems of biases and overfitting can be significantly reduced by further applying a least squares refitting on the support of Lasso estimate. However, a refitting based on the corrupted data will induce more biases as the true covariates are unknown. This loss of information was shown to cause overfitting and make the estimate very unstable in high dimensions (Rosenbaum and Tsybakov, 2010). In this paper, we propose a new method called as balanced estimation to ameliorate the issues of biases and overfitting under measurement errors by combining the strengths of the nearest positive semi-definite projection in Datta and Zou (2017) and the combined L_1 and concave regularization in Fan and Lv (2014). The proposed balanced estimator utilizes the oracle prediction property of Lasso penalty and refines the selected model by an additional concave regularization with contaminated covariates. In fact, the idea of combining the strengths of different penalties were also explored in the clean data setting. See, for example, Zou and Hastie (2005), Liu and Wu (2007), Zou and Zhang (2009) and Fan and Lv (2014).

The major contributions of this paper are twofold. First, we develop a new method for high-dimensional error-in-variables regression that enjoys nice estimation accuracy and successfully prevents overfitting under both additive and multiplicative measurement errors. It utilizes the combined L_1 and concave penalties to achieve an ideal balance between prediction and variable selection based on a convex estimate of the negative log-likelihood function, which can be efficiently solved through the path-following coordinate optimization algorithm. The amelioration of overfitting can significantly improve the model interpretability. Second, we provide theoretical guarantees for the proposed method by establishing oracle prediction and estimation error bounds equivalent to those in Bickel et al. (2009) for Lasso with the clean data set. Moreover, an explicit and asymptotically vanishing bound on the false sign rate is proved for our method, which is generally not shared by the CoCoLasso estimate. Our numerical studies show that the amelioration of variable selection will in turn improve the prediction and estimation performance under measurement errors.

The remainder of this paper is organized as follows. Section 2 presents the problem setup and the new methodology of balanced estimation. Theoretical properties including bounds on the oracle estimation errors and the number of falsely discovered signs are established in Section 3. We provide numerical studies in Section 4. Section 5 concludes with discussions and possible future work. The proofs and additional technical details are provided in the Appendix.

Notations. For a vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$, denote $\|\mathbf{x}\|_q = \left(\sum_{j=1}^p |x_j|^q\right)^{1/q}$ the L_q-norm for $q \in (0, \infty)$, $\|\mathbf{x}\|_0 = |\{j : x_j \neq 0\}|$, and $\|\mathbf{x}\|_\infty = \max_{1\le i\le p}|x_i|$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, denote $\|\mathbf{A}\|_1 = \max_{1\le j\le q}\sum_{i=1}^p |a_{ij}|$, $\|\mathbf{A}\|_\infty = \max_{1\le i\le p}\sum_{j=1}^q |a_{ij}|$, $\|\mathbf{A}\|_{\max} = \max_{i,j}|a_{ij}|$, $\|\mathbf{A}\|_2 = \Lambda_{\max}(\mathbf{A}^\top\mathbf{A})^{1/2}$ the matrix L_1-norm, L_∞-norm, elementwise maximum norm, and spectral norm, respectively, where $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ are the minimum and maximum eigenvalues of the given matrix $\mathbf{A}$.

2. Balanced estimation under measurement errors

2.1. Model setting

In the clean data setting, high-dimensional linear regression model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the n-dimensional response vector, $\mathbf{X} = (x_{ij})_{n\times p}$ is the fixed design matrix with p covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is an n-dimensional error vector independent of $\mathbf{X}$ with $\mathbf{I}_n$ denoting the $n \times n$ identity matrix.[1] The columns of $\mathbf{X}$ are assumed to have a common L_2 norm $n^{1/2}$, matching that of the constant covariate $\mathbf{1}$ for the intercept.

[1] The Gaussian distribution is assumed for simplicity, and similar theoretical results hold when the tail probability of the error vector decays exponentially.