Journal of Multivariate Analysis 145 (2016) 22–36

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Inference for biased models: A quasi-instrumental variable approach

Lu Lin[a,b,1], Lixing Zhu[c,d,*,1], Yujie Gai[e]

[a] Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan, China
[b] School of Mathematics, Shandong University, Jinan, China
[c] School of Statistics,Beijing Normal University, China
[d] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
[e] Central University of Finance and Economics, Beijing, China

ARTICLE INFO

Article history:
Received 27 February 2015
Available online 8 December 2015

AMS subject classifications:
62C05
62F10
62F12
62G05

Keywords:
High-dimensional regression
Non-sparse structure
Instrumental variable
Re-modeling
Bias correction
Dantzig selector

ABSTRACT

For linear regression models who are not exactly sparse in the sense that the coefficients of the insignificant variables are not exactly zero, the working models obtained by a variable selection are often biased. Even in sparse cases, after a variable selection, when some significant variables are missing, the working models are biased as well. Thus, under such situations, root-n consistent estimation and accurate prediction could not be expected. In this paper, a novel remodeling method is proposed to produce an unbiased model when quasi-instrumental variables are introduced. The root-n estimation consistency and the asymptotic normality can be achieved, and the prediction accuracy can be promoted as well. The performance of the new method is examined through simulation studies.

1. Introduction

The research described here is particularly motivated by variable selection, but not limited to this area. Consider the linear regression model:

$$Y = \beta^{\top} X + \varepsilon. \tag{1.1}$$

The model is seen as the full model that contains all possibly relevant predictors $X^{(1)}, \ldots, X^{(p)}$ in the predictor vector X, implying $\mathrm{E}(\varepsilon|X) = 0$. Here the dimension p of X is large and even larger than the size n of an available sample. As is well known, the root-n estimation consistency and the asymptotic normality play very important role in further data analyses such as confidence region and prediction interval constructions. However, when there are too many predictors, that is, when p is too large, even $p > n$, the root-n estimation consistency is often impossible, and prediction could be inaccurate. Variable selection is necessary for us to use those "kept predictors" in a working model such that the data analyses can

[*] Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

[1] The first two authors are in charge of the methodology development and material organization.

http://dx.doi.org/10.1016/j.jmva.2015.11.011

go forward. As such, variable selection is usually used to remove other predictors out from the full model. We will call them the "removed predictors". Without loss of generality, suppose that the first q predictors, $X^{(1)}, \ldots, X^{(q)}$, are kept in a working model whereas the last $p-q$ predictors are removed from the full model (1.1) via a variable selection approach. X is then partitioned to be $X = (Z^\top, U^\top)^\top$, where $Z = (X^{(1)}, \ldots, X^{(q)})^\top$ and $U = (X^{(q+1)}, \ldots, X^{(p)})^\top$. Correspondingly, β is partitioned as $\beta = (\theta^\top, \gamma^\top)^\top$. Thus, the working model is the following post-selection model:

$$Y = \theta^\top Z + \eta, \tag{1.2}$$

where η is a new error term. Here working model means that after selection, the model is used to describe the relationship between the predictor vector Z and the response Y. Note that in this modeling, the error η can be rewritten as $\eta = \gamma^\top U + \varepsilon$ which in effect contains all the removed predictors of U. When the full model is sparse such that $\gamma \equiv 0$, there are a great number of research works available in the literature to obtain root-n consistent estimation and oracle property, see e.g. the LASSO and the adaptive LASSO [11,14], the SCAD [3,5], the Dantzig selector [2] and its relevant development, and the MCP [12].

However, when Z is correlated with some components of U, the following may occur:

$$\mathrm{E}(\eta|Z) \neq 0. \tag{1.3}$$

It implies that model (1.2) could be biased. This problem could appear in two scenarios. First, in most cases, the full model cannot be exactly sparse. Many "insignificant predictors" with "small but not nonzero" coefficients are removed in selection procedures. However, Leeb and Pötscher [9] showed that when some coefficients are of order $n^{-1/2}$, the conventional model selection consistency may go wrong. Zhang and Huang [13] also considered the model that contains many small coefficients. Under the condition that controls the magnitudes of the small coefficients, although rate consistency was obtained in some sense, the root-n consistency cannot be ensured. Further, a more practical issue is that even when the model is sparse, any variable selection method would miss some "significant predictors" and may cause the working model to be biased. Because of the model bias, it is difficult to construct the confidence region for the coefficients in the working model, and the prediction accuracy may be deteriorated. These observations motivate us to consider how to consistently estimate θ when the coefficients associated with the removed predictors are of slower rate than $n^{-1/2}$ and $\mathrm{E}(\eta|Z)$ is not asymptotically negligible. To the best of our knowledge, none of existing results handle bias correction in the literature.

Of primary interest in the present paper is to correct the model bias for the working model. To this end, our idea is to introduce quasi-instrumental variables for bias correction. It is worthwhile to point out that here we call quasi-instrumental variables because they are not really instrumental variables in the literature which are given in advance. The quasi-instrumental variables in effect need to be determined by ourselves. The main idea is motivated from the fact that those removed predictors may contain some information about the response and the kept predictors in the working model. We then determine some quasi-instrumental variables as functions of these removed predictors. It will be seen that the use of quasi-instrumental variables makes the re-constructed model to be an unbiased partially linear model, which is different in structure from the full model. This partially linear modeling is of course different from the classical partially linear modeling in which the predictors in nonparametric component are given.

We should emphasize the following three points for our study.

- First, if the number of kept predictors is a fixed value, our partially linear modeling can be directly applied. In other words, practically, our method is always feasible.
- Second, for those post-selection models, the number of kept predictors is often random. Thus we should assume the model identifiability for us to do further statistical analyses. As an example, we introduce some regularity conditions under which the working model selected by the Dantzig selector (DS) can be identifiable as the sample size goes to infinity. The same idea can be applied to other variable selection methods such as the LASSO, the SCAD or the MCP with different conditions accordingly. It is worth pointing out that the condition allows small coefficients to tend to zero at a slower rate than $n^{(c-1)/2}$ for a constant $0 < c < 1$. This shows the importance of bias correction and remodeling suggested in the present paper because as was commented on the results from [9,13], under this setting, existing methods cannot ensure the root-n estimation consistency and asymptotic normality.
- Third, compared with the commonly used estimations for linear models, our method may be more computational intensive. However, to avoid the risk of possible unreliable modeling and analysis caused by bias, the cost is worthwhile to pay.

This work may be a first attempt to achieve the root-n estimation consistency and then reliable further statistical analyses. There are several issues that deserve further investigations. The current version of this paper is an updated version of an early manuscript by Lin et al. [10].

The rest of the paper is organized as follows. In Section 2, when any conventional selection procedures such as the Dantzig selector is used, identifiability conditions are presented, which are particularly designed for the working model when the model size is random, a bias-corrected working modeling is proposed and a method about constructing quasi-instrumental variables is suggested. In Section 3, the estimation and prediction procedures for the new working model are given and the asymptotic properties of the resulting estimation are obtained. In Section 4, a method about how to construct low dimensional nonparametric function is introduced and an approximate algorithm for constructing quasi-instrumental variables is proposed for the case where the dimension of the related nonparametric estimation is relatively large. Simulation