Journal of Multivariate Analysis 144 (2016) 25–37

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Testing covariates in high dimension linear regression with latent factors

Wei Lan [a], Yue Ding [b], Zheng Fang [c], Kuangnan Fang [d,*]

[a] School of Statistics and center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, China
[b] School of Economics and Management, Southwest Jiaotong University, Chengdu, China
[c] Business School, Sichuan University, Chengdu, China
[d] School of Economics and MOE Key Laboratory of Econometrics, Xiamen University, Xiamen, China

ARTICLE INFO

Article history:
Received 17 November 2014
Available online 3 November 2015

AMS subject classifications:
62F03

Keywords:
Approximate factor model
Global significance testing
High dimension regression
Individual effect testing

ABSTRACT

We propose here both F-test and z-test (or t-test) for testing global significance and individual effect of each single predictor respectively in high dimension regression model when the explanatory variables follow a latent factor structure (Wang, 2012). Under the null hypothesis, together with fairly mild conditions on the explanatory variables and latent factors, we show that the proposed F-test and t-test are asymptotically distributed as weighted chi-square and standard normal distribution respectively. That leads to quite different test statistics and inference procedures, as compared with that of Zhong and Chen (2011) when the explanatory variables are weakly dependent. Moreover, based on the p-value of each predictor, the method of Storey et al. (2004) can be used to implement the multiple testing procedure, and we can achieve consistent model selection as long as we can select the threshold value appropriately. All the results are further supported by extensive Monte Carlo simulation studies. The practical utility of the two proposed tests are illustrated via a real data example for index funds tracking in China stock market.

1. Introduction

Traditional F-test and z-test (or t-test) are commonly used to detect the relationship between a response variable $Y_i \in \mathbb{R}^1$ and a set of explanatory variables $X_i \in \mathbb{R}^p$ in a linear regression model when the number of explanatory variables p is fixed. By contrast, when p is diverging and much larger than the sample size n, classical statistical inferences (F test and z-test) were not applicable since the resulting ordinary least square (OLS) estimator is no longer computable. To fix the issue, there is a large stream of papers intending to extend the traditional F-test and z-test (or t-test) to accommodate high dimensional settings; see, for example, [22,9,21,12].

The aforementioned testing procedures are quite useful for high dimensional data analyses. However, their applicability is heavily relying on one critical assumption, i.e., the explanatory variables are weakly dependent such that $tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}$, where $\Sigma = \mathrm{cov}(X_i) \in \mathbb{R}^{p\times p}$. For more detailed illustrations for such assumption, we refer to [22,23]. It is remarkable that such assumption is violated if the explanatory variables X_i admit a latent factor structure, which is usually encountered in real practice [6,19]. Specifically, we consider the following data generation process $X_i = \gamma Z_i + \widetilde{X}_i$, where each element of the common factors $Z_i \in \mathbb{R}^d$ and random errors $\widetilde{X}_i$ are all independently generated from a standard

* Corresponding author.
E-mail addresses: facelw@gmail.com (W. Lan), dingyue-300@163.com (Y. Ding), 149281891@qq.com (Z. Fang), xmufkn@xmu.edu.cn (K. Fang).

http://dx.doi.org/10.1016/j.jmva.2015.10.013
0047-259X/

normal distribution, with $d > 0$ is the finite number of common factors. Moreover, the factor loadings $\gamma \in \mathbb{R}^{p\times d}$ satisfy $p^{-1}\gamma^\top\gamma \to I_d$, where I_d represents the identity matrix of dimension d. In this setting, one can verify that $tr(\Sigma^4) = tr(\gamma\gamma^\top)^4\{1+o(1)\} = tr(\gamma^\top\gamma)^4\{1+o(1)\} = p^4 tr(I_d)\{1+o(1)\} = dp^4\{1+o(1)\}$, and $tr(\Sigma^2) = dp^2\{1+o(1)\}$. As a result, we can have $tr(\Sigma^4)/tr^2(\Sigma^2) \to 1/d \neq 0$, which violates condition (2.8) of [22], and condition (C1) of [12]. Consequently, how to construct testing procedures for this special types of explanatory variables is a problem of theoretical demand.

It is also noteworthy that the above testing problems are also empirically motivated. For example, consider the problem of index fund tracking of reproducing the performance of a stock market index. In this particular application, the response of interest is the return on some specific market index, say Shanghai composite index in China stock market, while the explanatory variables can be the return of all the stocks in China stock market. Therefore, the number of explanatory variables may be very large compared with the number of observations; see Section 3.2 of real data analysis for details. For these types of explanatory variables, we cannot expect that the returns across different stocks are weakly dependent. In fact, it has long been recognized empirically and theoretically that there should exist some latent common factors that influence all stock returns [16,4,7,5]. To this end, it is quite natural and reasonable to assume that the explanatory variables X_i follow a latent factor structure so that the condition $tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}$ is violated.

Motivated by the theoretical and practical demand, we intend to construct some testing procedures for the regression coefficients when the explanatory variables admit a latent factor structure [19]. We develop both F-test and z-test (or t-test) for testing global significance and effect of each single predictor respectively in high dimension regression model. Specifically, we revisit the test statistic of [12] used for testing global significance of regression coefficients for weakly dependent explanatory variables, and show that the resulting test statistic is asymptotic weighted chi-square when the explanatory variables follow an approximate factor model under some mild conditions. That leads to quite different test statistics and inference procedures, as compared with that of [22,12], when the explanatory variables are weakly dependent. In addition, after controlling for the latent common effect of the explanatory variables, the remaining factor profiled predictors are weakly dependent [19]. As a consequence, the univariate regression [7] can be used to assess the significance of each variable. Based on the p-value of each predictor, we can then apply the method of [17] to control the false discovery rate (FDR), and the method can achieve consistent model selection as long as we can set the nominal level appropriately. Extensive simulation results and an empirical example on index fund tracking in China stock market confirmed the usefulness of the proposed method.

The remainder of the paper is organized as follows. Section 2 introduces global significance testing, and individual effect testing with FDR control together with their theoretical properties. Numerical studies, including simulation and a real data analysis, are reported in Section 3. Section 4 concludes the article with a short discussion and all the technical details are provided in the Appendix.

2. The methodology

2.1. Model and notations

Let (Y_i, X_i) be the observation collected at ith unit for $1 \le i \le n$, where $Y_i \in \mathbb{R}^1$ is the response value, $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ be the p-dimensional explanatory variables with mean $\mathbf{0}$ and covariance matrix $\Sigma = (\sigma_{j_1j_2}) \in \mathbb{R}^{p\times p}$. Unless explicitly stated otherwise, we hereafter assume that $p \gg n$ and n tends to infinity for asymptotic behavior. In addition, we assume that all the explanatory variables have been appropriately standardized such that $E(X_{ij}) = 0$, and $\sigma_{jj} = 1$ for every $1 \le j \le p$. To establish the relationship between Y_i and X_i, we consider the following linear regression model,

$$Y_i = X_i^\top\beta + \varepsilon_i, \tag{2.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown vector of regression coefficients, ε_i is the random noise that is independent of X_i, distributed with mean 0 and finite variance $\sigma^2 < \infty$. For notation convenience, define $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be a vector of response variable, $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ be the design matrix with the jth column $\mathbb{X}_j = (X_{1j}, \ldots, X_{nj})^\top \in \mathbb{R}^n$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.

Since the traditional F-test and z-test (or t-test) are no longer applicable when p is diverging and much larger than the sample size n, there is a large stream of papers intending to extend the traditional F-test and z-test (or t-test) to accommodate high dimensional settings; see, for example, [22,9,21,12]. For the statistical validity of the aforementioned tests, appropriate technical conditions have to be assumed. Among all the conditions, Zhang and Zhang [21] and Lan et al. [12] require that

$$\lambda_{\max}(\Sigma) < \infty, \tag{2.2}$$

where $\lambda_{\max}(A)$ represents for the largest eigenvalues of any arbitrary matrix A. In contrast, Zhong and Chen [22] replaced condition (2.2) by

$$tr(\Sigma^4) = o\big\{tr^2(\Sigma^2)\big\}. \tag{2.3}$$

We find that both (2.2) and (2.3) are sensible if Σ is not highly singular, this should happen if the predictors are weakly correlated. Unfortunately, conditions (2.2) and (2.3) are violated if the explanatory variables X_i are highly correlated that