Computational Statistics and Data Analysis 83 (2015) 275–286

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Testing predictor significance with ultra high dimensional multivariate responses

Yingying Ma [a], Wei Lan [b,*], Hansheng Wang [a]

[a] Guanghua School of Management, Peking University, Beijing, 100871, PR China

[b] Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, Sichuan, 610074, PR China

ARTICLE INFO

Article history:
Received 27 July 2013
Received in revised form 23 July 2014
Accepted 27 September 2014
Available online 18 October 2014

Keywords:
Hypotheses testing
Multivariate regression
Paid search advertising
Ultra high dimensional data

ABSTRACT

We consider here the problem of testing the effect of a subset of predictors for a regression model with predictor dimension fixed but ultra high dimensional responses. Because the response dimension is ultra high, the classical method of likelihood ratio test is no longer applicable. To solve the problem, we propose a novel solution, which decomposes the original problem into many testing problems with univariate responses. Subsequently, the usual residual sum of squares (RSS) type test statistics can be obtained. Those statistics are then integrated together across different responses to form an overall and powerful test statistic. Under the null hypothesis, the resulting test statistic is asymptotically standard normal after some appropriate standardization. Numerical studies are presented to demonstrate the finite sample performance of the test statistic and a real example about paid search advertising is analyzed for illustration purpose.

1. Introduction

We consider here a multivariate regression model with predictor dimension fixed but response dimension ultra high. We are particularly interested in testing statistical significance of a subset of predictors, after controlling for the effect of the others. Such a problem has been well studied for a standard linear regression problem with fixed dimensional predictors and univariate responses (Lehmann, 1998; Shao, 2003). In fact, partial F-test (Ravishanker and Dey, 2001; Chatterjee and Hadi, 2006) has been widely implemented in many standard statistical softwares (e.g., SAS, S-plus, R, etc.) and extensively used in practice. Despite its usefulness, this classical method cannot be used to deal with problems with ultra high dimensional data (Zhong and Chen, 2011). More specifically, if the predictor dimension is ultra high (i.e., much larger than the sample size), the classical partial F-test statistic is no longer computable and thus calls for a new testing procedure. To this end, much efforts have been devoted along this direction in the recent literature; see, for example, Goeman et al. (2004), Goeman et al. (2006), Goeman et al. (2011), Zhong and Chen (2011), and Lan et al. (2014).

It is remarkable that the aforementioned testing procedures can be useful for high dimensional data with univariate responses. They are not immediately applicable for multivariate responses. In the meanwhile, the problem of regression with multivariate responses is also an important subject for multivariate data analysis (Anderson, 1984; Johnson and Wichern, 2003). In a classical multivariate regression setup with normally distributed errors, an elegant likelihood ratio test has been well developed and its associated asymptotic distribution has been well studied, if both the predictor and response

* Corresponding author.
E-mail addresses: mayingying@pku.edu.cn (Y. Ma), facelw@gmail.com (W. Lan), hansheng@gsm.pku.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.csda.2014.09.020
0167-9473/

dimensions are fixed but the sample size goes to infinity (Anderson, 1984; Johnson and Wichern, 2003). Unfortunately, such a method cannot be extended to the situation with ultra high dimensional responses. In that case, the estimated covariance matrix for the multivariate residual is no longer positive definite. This makes the classical method of likelihood ratio test no longer applicable. As a result, we are theoretically motivated to fulfil this important gap.

It is worthwhile mentioning that this work is also empirically motivated. Consider for example the real data to be analyzed in Section 3.2. It is a dataset about paid search advertising. The response of interest is the total number of impressions generated by one particular keyword, which bids on Baidu, the largest search engine in China. Because the number of keywords involved is huge, the dimension of the multivariate response is thus ultra high. The predictor of interested here is a set of six dummy variables representing different days of the week. The practitioners are eager to know whether there exists a day of the week that has significant effect on certain individual keyword. To this end, the classical univariate F-test can be used to test for each individual keyword. Nevertheless, by doing so, the family wise Type I error gets inevitably inflated. To provide certain protection against the overall family wise error, an overall test involving every individual keyword is needed.

More specifically, we propose here a novel testing procedure to test the statistical significance of a subset of regression coefficients after controlling for the effects of the others. The proposed test is designed for the situation when the predictor dimension is fixed but the response dimension is ultra high. The proposed new test is simple to compute and asymptotically standard normal under mild conditions. The rest of this article is organized as follows. The model, notations, and technical conditions will be introduced in Section 2. The numerical experiments based on both simulated and real dataset are evaluated in Section 3. Section 4 concludes the article with a short discussion. All the technical details are relegated to the Appendix.

2. The methodology

2.1. The model and notations

Let (Y_i, X_i) be the observation collected from the ith subject with $1 \le i \le n$, where $Y_i = (Y_{i1}, \ldots, Y_{im})^\top \in \mathbb{R}^m$ is the m-dimensional multivariate responses and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor. Throughout the rest of this article, we assume that the predictor dimension p is fixed but the response dimension m is ultra high, i.e., $m \gg n$. This makes our work clearly different from the existing literatures, where they all assume $m = 1$ but $p \gg n$ (Goeman et al., 2004, 2006, 2011; Zhong and Chen, 2011; Lan et al., 2014). Without loss of generality, we assume that the component of X_i has been appropriately sorted, so that it can be decomposed as $X_i = (X_{ia}^\top, X_{ib}^\top)^\top \in \mathbb{R}^p$, where $X_{ia} = (X_{i1}, \ldots, X_{iq})^\top \in \mathbb{R}^q$ is a set of predictors need to be controlled. In contrast, $X_{ib} = (X_{i(q+1)}, \ldots, X_{ip})^\top \in \mathbb{R}^{p-q}$ collects all the predictors that need to be tested. To establish the relationship between Y_i and X_i, we assume a standard multivariate linear regression model as follows,

$$Y_i = BX_i + \varepsilon_i = B_a X_{ia} + B_b X_{ib} + \varepsilon_i, \tag{2.1}$$

where $B = (B_a, B_b) = (\beta_1, \ldots, \beta_m)^\top \in \mathbb{R}^{m\times p}$ is the coefficient matrix with $\beta_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})^\top \in \mathbb{R}^p$. Furthermore, $B_a = (\beta_{1a}, \ldots, \beta_{ma})^\top \in \mathbb{R}^{m\times q}$ and $B_b = (\beta_{1b}, \ldots, \beta_{mb})^\top \in \mathbb{R}^{m\times(p-q)}$, where $\beta_{ka} = (\beta_{kj} : j \le q)^\top \in \mathbb{R}^q$ and $\beta_{kb} = (\beta_{kj} : q < j \le p)^\top \in \mathbb{R}^{p-q}$ for each $1 \le k \le m$. In addition, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^\top \in \mathbb{R}^m$ is the residual vector and ε_{ik} is the random noise associated with the k-th response. We are then interested in testing the following statistical hypotheses

$$H_0 : B_b = 0 \quad \text{vs.} \quad H_1 : B_b \ne 0. \tag{2.2}$$

Intuitively, under the null hypothesis of (2.2), X_{ib} should be irrelevant for Y_i after controlling for the effect of X_{ia}.

2.2. Likelihood ratio test

For illustration purpose, we first consider the situation with $m \ll n$. Assume that $\varepsilon_i \in \mathbb{R}^m$ follows a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\sigma_{k_1k_2}) \in \mathbb{R}^{m\times m}$. Then the negative two times log-likelihood function can be expressed as

$$\mathcal{L}(B, \Sigma) = \sum_{i=1}^{n} \left\{ (Y_i - BX_i)^\top \Sigma^{-1} (Y_i - BX_i) \right\} + n \log|\Sigma|,$$

where the constants independent of B and Σ are omitted. This leads to the following maximum likelihood estimator (MLE) as $\hat{B}^\top = (n^{-1}\sum_{i=1}^n X_i X_i^\top)^{-1}(n^{-1}\sum_{i=1}^n X_i Y_i^\top)$ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (Y_i - \hat{B}X_i)(Y_i - \hat{B}X_i)^\top$. Subsequently, the corresponding minimum negative log-likelihood function is given by $2^{-1}n\log|\hat{\Sigma}|$. Similarly, under the null hypothesis of (2.2) that $B_b = 0$, the MLE for B_a and Σ are given by $\hat{B}_a^\top = (n^{-1}\sum_{i=1}^n X_{ia}X_{ia}^\top)^{-1}(n^{-1}\sum_{i=1}^n X_{ia}Y_i^\top)$ and $\hat{\Sigma}_a = n^{-1}\sum_{i=1}^n (Y_i - \hat{B}_a X_{ia})(Y_i - \hat{B}_a X_{ia})^\top$, respectively. The corresponding minimum negative likelihood, under the null hypothesis, becomes $\ell_0 = 2^{-1}n\log|\hat{\Sigma}_a|$. Accordingly, the likelihood ratio test statistic becomes $-n\log(|\hat{\Sigma}|/|\hat{\Sigma}_a|)$. The LR test statistic is then adjusted by $-\{n - p - 0.5(m - p + q)\}\log(|\hat{\Sigma}|/|\hat{\Sigma}_a|)$, which is asymptotically distributed as a chi-squared distribution with $m(p - q)$ degrees of freedom (Anderson, 1984; Johnson and Wichern, 2003).