Journal of Statistical Planning and Inference 178 (2016) 37–60

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Empirical likelihood test for high-dimensional two-sample model

Gabriela Ciuperca *,[1], Zahraa Salloum [1]

Institut Camille Jordan, Université Lyon 1, France

ARTICLE INFO

Article history:
Received 26 December 2015
Received in revised form 21 April 2016
Accepted 13 May 2016
Available online 24 May 2016

MSC:
primary 62F03
62G10
secondary 62G10
62F05

Keywords:
Two-sample
High-dimension
Linear model
Empirical likelihood test

ABSTRACT

A non parametric method based on the empirical likelihood is proposed for detecting the change in the coefficients of high-dimensional linear model where the number of model variables may increase as the sample size increases. This amounts to testing the null hypothesis of no change against the alternative of one change in the regression coefficients. Based on the theoretical asymptotic behaviour of the empirical likelihood ratio statistic, we propose, for a fixed design, a simpler test statistic, easier to use in practice. The asymptotic normality of the proposed test statistic under the null hypothesis is proved, a result which is different from the χ^2 law for a model with a fixed variable number. Under alternative hypothesis, the test statistic diverges. Some Monte-Carlo simulations study the behaviour of the proposed test statistic.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The technology development and fast numerical techniques make it possible to consider and study statistical models with a large number of variables. High-dimensional model refers to a model with a number p of explanatory variables increasing to infinity as the number n of observations converges to infinity. When p diverges, traditional statistical methods may not work with this kind of growth dimensionality.

Most of the literature works on high-dimensional model use the LASSO (Least Absolute Shrinkage and Selection Operator) type methods, in order to automatically select the significant variables. The principle of these methods, introduced by Tibshirani (1996), is to optimize a penalized process, more precisely, a process with a L_1-type penalty. If the model contains outliers, the parameter estimators by the least squares method with LASSO penalty have a large error. An alternative method is then the penalized quantile method. Thereby, Dicker et al. (2014) consider a quantile model with seamless-L_0 penalty when the number p of explanatory variables is such that $p \to \infty$, $p/n \to 0$ as $n \to \infty$. For a general quantile regression, Wu and Liu (2009) propose the SCAD penalty, while, in the paper of Zou and Yuan (2008), a composite quantile regression is considered with an adaptive LASSO penalty. The case $p \to \infty$ is also considered in Fan and Peng (2004) for a non-concave penalized likelihood method, when $p^5/n \to \infty$. Concerning the group selection methods for high-dimensional models, the reader can find in Huang et al. (2012) a review of methods.

* Corresponding author.
E-mail addresses: Gabriela.Ciuperca@univ-lyon1.fr (G. Ciuperca), salloum@math.univ-lyon1.fr (Z. Salloum).
[1] Université de Lyon, Université Lyon 1, CNRS, UMR 5208, Institut Camille Jordan, Bat. Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne Cedex, France.

http://dx.doi.org/10.1016/j.jspi.2016.05.002
0378-3758/© 2016 Elsevier B.V. All rights reserved.

All these methods are based first on the principle of selecting (automatically) the significant variables. Then, the dependent variable is modelled only as a function of the significant variables, in order to have more accurate parameter estimators and a better adjustment for the dependent variable.

If the goal is to have the most accurate prediction and also robust, in the case of a model with outliers, one possibility is to consider the empirical likelihood (EL) method. But, for this type of method, in literature, most papers are devoted to the case of fixed p. For a high-dimensional linear regression model, we can refer first to paper Guo et al. (2013), when the design is deterministic. High-dimensional data are also studied in Liu et al. (2013), where EL method is considered for a sequence of i.i.d. random vectors with dimension p, when $p \to \infty$ as $n \to \infty$.

In this paper, we are interested in a change-point model, that is, a model which changes at some moment. The number p of explanatory variables varies with the number n of observations and p can converge to infinity if $n \to \infty$.

Since statistical techniques in high-dimension are fairly recent, there are not many papers in literature that address the change-point problem in a high-dimensional model. Lung-Yut-Fong et al. (2012) proposes an approach for detection of a change-point in high-volume network traffic. The asymptotic distribution of the test statistic proposed in Lung-Yut-Fong et al. (2012), under the null hypothesis that there is no change-point, is the *argsup* of a Brownian Bridge. There are some papers where LASSO type methods are used. Lee et al. (2015) consider a possible change-point in a high-dimensional regression with Gaussian errors. The main result of the article is to show that the sparsity property is maintained, even if there is a change in the model. There is no hypothesis test to decide the presence or absence of change in model. In Ciuperca (2014), LASSO-type and adaptive LASSO estimators are studied, while in Ciuperca (2013) quantile model with SCAD penalty is considered. These last two papers consider models with p fixed. In order to choose the change-point number, a model selection criterion is also proposed by Ciuperca (2014).

To the authors' knowledge, the EL technique has not yet been addressed in a high-dimensional two-sample model, that makes the interest of this work. We study the asymptotic behaviour of the empirical likelihood ratio test statistic when the design is deterministic.

We consider a first linear model:

$$Y_i = \mathbf{X}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Consider now a second linear model which changes at observation k. It is called two-phase model, or model with one change-point:

$$Y_i = \begin{cases} \mathbf{X}_i^t \boldsymbol{\beta} + \varepsilon_i, & 1 \le i \le k, \\ \mathbf{X}_i^t \boldsymbol{\beta}_2 + \varepsilon_i, & k < i \le n, \end{cases} \tag{1.2}$$

where $\mathbf{X}_i$ is a $p \times 1$ vector of p explanatory variables, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_2$ are $p \times 1$ vectors of unknown parameters and ε_i designates the model error. The parameter $\boldsymbol{\beta}$ of the first phase of (1.2) coincides with that of (1.1). For models (1.1) and (1.2), Y_i is observation i of the response variable. The errors ε_i are supposed independent identically distributed (i.i.d), with mean zero and finite variance σ^2.

We assume that the number p of explanatory variables $\mathbf{X}_i$ depends on the sample size n: $p = p_n$, such that $p_n \to \infty$ as $n \to \infty$. The change-point k of (1.2) also depends on n. The change in model (1.2) takes place far enough from the first observation and sufficiently previous to the last observation. So, we suppose that $\lim_{n\to\infty} k/n \in (0, 1)$.

In this paper, for given k, we use the empirical likelihood method to construct the confidence region for $\boldsymbol{\beta} - \boldsymbol{\beta}_2$, or equivalent to test the null hypothesis of no change in model (1.2). Under null hypothesis, the model has the form (1.1), that is

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}. \tag{1.3}$$

The alternative hypothesis assumes that one change occurs in the regression parameters, that is

$$H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}. \tag{1.4}$$

The paper is organized as follows. In Section 2 we first present the EL method for the two-sample model. Some notations used throughout the paper are defined and needed assumptions for the theoretical study are also announced. In Section 3, we construct an empirical likelihood ratio test statistic and we study its asymptotic behaviour. The asymptotic distribution under H_0 of the test statistic is obtained, while, under H_1, this statistic diverges. Next, in Section 4, we analyse the empirical size and the empirical power by means of simulations, which confirm the performance of proposed test. A new critical value is also proposed in order to improve the empirical size. The proofs of the main results are given in the Appendix followed by some Lemmas and their proofs.

2. Preliminaries

In this section, we introduce the EL method for the two-sample model. Notations and assumptions are also given.

Under null hypothesis H_0, that is model (1.1), let $\boldsymbol{\beta}^{\mathbf{0}}$ denote the true value of the parameter $\boldsymbol{\beta}$. Under alternative hypothesis H_1, that is model (1.2), the true values of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_2$, respectively, are $\boldsymbol{\beta}^0$, $\boldsymbol{\beta}_2^0$.

In order to define the profile empirical likelihood (under H_0 and under H_1), we introduce the following random p-vector, for all $\boldsymbol{\beta} \in \mathbb{R}^p$ and $i = 1, \ldots, n$:

$$\mathbf{z}_i(\boldsymbol{\beta}) \equiv \mathbf{X}_i(Y_i - \mathbf{X}_i^t \boldsymbol{\beta}).$$

Download English Version:

https://daneshyari.com/en/article/1147558

Download Persian Version:

https://daneshyari.com/article/1147558

Daneshyari.com