Computational Statistics and Data Analysis 102 (2016) 85–97

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Feature screening for generalized varying coefficient models with application to dichotomous responses

Xiaochao Xia [a,b,*], Hu Yang [a], Jialiang Li [c,d,e]

[a] College of Mathematics and Statistics, Chongqing University, China
[b] College of Science, Huazhong Agricultural University, China
[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore
[d] Duke-NUS Graduate Medical School, Singapore
[e] Singapore Eye Research Institute, Singapore

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form 21 March 2016
Accepted 23 April 2016
Available online 3 May 2016

Keywords:
Generalized varying coefficient model
Variable screening
High dimensional data
Sure screening property
Ranking consistency

ABSTRACT

Generalized varying coefficient model (GVCM) is an important extension of generalized linear model and varying coefficient model. It has been widely applied in many areas. This paper mainly considers the variable screening problem with dichotomous response data under GVCM, where a spline approximation is employed to estimate the coefficient function for each covariate. Two screening procedures based on marginal maximum likelihood estimation and marginal likelihood ratio statistics are studied. The sure independence screening property and the ranking consistency of these two approaches are established under some technical conditions. Some refined algorithms are presented to control the false selection rate. Extensive numerical studies are conducted to evaluate the performance of the proposed methodology.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid development of information collection and data storage, nowadays many areas face the challenge of analyzing enormously large datasets. Translating to the regression framework, we may possibly encounter large number of covariates measured from relatively small number of observations. In the motivating example to be studied in Section 5 of this study, we consider a breast cancer dataset where expression levels for 24,481 gene probes are included. For such a dataset, discovering a small set of relevant cancer genes and using them to construct meaningful models for classification and prediction is very important. In this paper, we are interested in the situation where the number of covariates p and the sample size n satisfy $\log p = O(n^r)$ for some $r > 0$. This is well known as the "large p, small n" problem and is also referred to the ultrahigh dimensional data problem in the literature (Fan and Lv, 2008). We note that in genetic studies dichotomous responses (e.g. disease-present or disease-absent status) such as in the breast cancer example may be more common than continuous responses.

Plenty of efforts on how to select the important covariates and estimate their effects simultaneously have been devoted over the past two decades. Most earlier works that incorporate the framework of penalized likelihood or penalized least squares and various flexible penalty principles were documented in the literature, such as LASSO by Tibshirani (1996), SCAD and other folded-concave penalty by Fan and Li (2001), Elastic net by Zou and Hastie (2005), adaptive LASSO by Zou (2006),

* Corresponding author at: College of Mathematics and Statistics, Chongqing University, China.
E-mail address: xxc@cqu.edu.cn (X. Xia).

http://dx.doi.org/10.1016/j.csda.2016.04.008
0167-9473/© 2016 Elsevier B.V. All rights reserved.

group LASSO by Yuan and Lin (2006) and MCP by Zhang (2010). However, most of these fundamental works are studied in either parametric models or the case where p is less than n. Modeling the regression coefficients as smooth functions may be more meaningful in practice. Varying-coefficient model (VCM) inherits many merits of familiar parametric and nonparametric models. There exist a lot of penalty-based methods adapted to the varying-coefficient models in the high-dimensional data analysis where p diverges with but not larger than n, see Wang and Xia (2009), Wang et al. (2008), Hu and Xia (2011) and Wei et al. (2011), among others. Whereas, for the ultrahigh dimensional data, these methods may not perform well in terms of computational expediency, statistical accuracy and algorithmic stability (Fan and Lv, 2008). Thus, necessary procedures are demanded and the relevant properties are also needed to be studied systematically. More recently, Fan et al. (2014) considered the variable screening problem under ultrahigh dimensional VCM. Liu et al. (2014) proposed an approach called the conditional correlation learning to screen important variables under VCM. Cheng et al. (2014) studied the same problem via a two-stage method for functional data analysis. However, to the best of our knowledge, when the number of covariates is much larger than the sample size, the problem on how to select underlying variables under GVCM, especially under the binary response model, has not been studied.

The primary purpose of this paper is to study such a problem. We consider the following model

$$g\{\mu(\mathbf{X}, T)\} = \boldsymbol{\beta}^{\tau}(T)\mathbf{X}, \tag{1.1}$$

where g is a known link function, $\mu(\mathbf{X}, T) = \mathbb{E}(Y|\mathbf{X}, T)$, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\tau}$ is a p-vector of covariates, $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^{\tau}$ is a p-vector of unknown coefficient functions, T is a univariate index variable and Y is the response. An intercept can be incorporated if necessary. Model (1.1) is more flexible than the well-known class of generalized linear model (GLM) (McCullagh and Nelder, 1989) because it retains linear structure in $\mathbf{X}$ and admits nonlinear interaction between $\mathbf{X}$ and T, which allows more complicated dependence structures. Apparently, when an identity link is used, model (1.1) reduces to the ordinary VCM. The related works include, for instance, Cai et al. (2000) who developed a kernel-based approach to rectify the computational difficulty, Zhang and Peng (2010) for simultaneous confidence band and hypothesis test, and Lian (2012) for high-dimensional variable selection under the framework of quasi-likelihood.

The contributions of this paper can be summarized as follows. First, for model (1.1), we develop a marginal screening strategy to identify the active variables, which are believed to be important to the response variable, from the p candidates. The proposed methods can be used to handle the ultrahigh dimensional continuous response and the ultrahigh dimensional binary response, where the latter is main focus of this study. Second, we prove that the proposed approaches enjoy the sure screening property and the ranking consistency under proper technical conditions. Third, to facilitate the estimation and improve its accuracy we present some refined algorithms, which are stable and fast in computation, according to our numerical studies. The approaches and results obtained in this paper can be regarded an extension of Fan and Song (2010) and Fan et al. (2014). However, the extension is non-trivial both theoretically and empirically. Despite most screening studies do not require very high estimation efficiency, the successful implementation of functional estimation for a general exponential family response is still not easy (Zhang and Peng, 2010). The problem is quite remarkable in the situation that the outcome space is crystalized at two atoms, which largely reduces the continuity and variability for estimating infinite dimensional functions.

The paper is organized as follows. In Section 2, a screening method based on *maximum marginal likelihood estimator (MMLE)* is proposed and the corresponding theoretical properties including model selection consistency as well as ranking consistency are established. In Section 3, an extensive screening based on *marginal likelihood ratio statistics (MLRS)* is discussed. In Section 4, an iterative algorithm, specific implementation details and a greedy version are provided. Numerical results are presented in Section 5. A conclusion is given in Section 6. All the theoretical proofs are relegated to the Appendix A.

2. Methodology and results

Suppose that $\{(\mathbf{x}_i, t_i, y_i), i = 1, \ldots, n\}$ is the sample data, independently and identically generated from model (1.1). Assume that the functional coefficient vector $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \beta_2(\cdot), \ldots, \beta_{p_n}(\cdot))^{\tau}$ is sparse and $\mathbb{E}X_j = 0$ and $\mathbb{E}X_j^2 = 1$. We use p_n to emphasize the dependence of the dimension p on the sample size n. We start with some necessary notations for GLM. A random variable Y is said to come from an exponential family if its probability density function takes the following canonical form

$$f_Y(y, \theta) = \exp\{y\theta - b(\theta) + c(y)\}, \tag{2.1}$$

for some known functions $b(\cdot)$, $c(\cdot)$ and unknown canonical parameter θ. Throughout this paper, we ignore the dispersion parameter, as our focus is on the mean effect. The mean of Y is $b'(\theta)$, the first derivative of $b(\theta)$ with respect to θ. Under model (1.1), we have

$$\mu(\mathbf{x}, t) = b'(\theta(\mathbf{x}, t)) = g^{-1}\big(\boldsymbol{\beta}^{\tau}(t)\mathbf{x}\big). \tag{2.2}$$

If g is taken as the canonical link, i.e., $g^{-1} = b'$, then $\theta(\mathbf{x}, t) = \sum_{j=1}^{p} \beta_j(t)x_j$. In the following, we present a screening method based on the maximum marginal likelihood estimator (MMLE).

Download English Version:

https://daneshyari.com/en/article/6868977

Download Persian Version:

https://daneshyari.com/article/6868977

Daneshyari.com