Computational Statistics and Data Analysis 124 (2018) 132–150

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variable selection for high dimensional Gaussian copula regression model: An adaptive hypothesis testing procedure

Yong He [a], Xinsheng Zhang [b], Liwen Zhang [c,*]

[a] School of Statistics, Shandong University of Finance and Economics, Jinan, 250014, China
[b] School of Management, Fudan University, Shanghai, 200433, China
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 10 October 2016
Received in revised form 24 February 2018
Accepted 3 March 2018
Available online 14 March 2018

Keywords:
Gaussian copula regression
Variable selection
Multiple testing
FDR/FDV

ABSTRACT

In this paper we consider the variable selection problem for high dimensional Gaussian copula regression model. We transform the variable selection problem into a multiple testing problem. Compared to the existing methods depending on regularization or a stepwise algorithm, our method avoids the ambiguous relationship between the regularized parameter and the number of false discovered variables or the decision of a stopping rule. We exploit nonparametric rank-based correlation coefficient estimators to construct our test statistics which achieve robustness and adaptivity to the unknown monotone marginal transformations. We show that our multiple testing procedure can control the false discovery rate (FDR) or the average number of falsely discovered variables (FDV) asymptotically. We also propose a screening multiple testing procedure to deal with the extremely high dimensional setting. Besides theoretical analysis, we also conduct numerical simulations to compare the variable selection performance of our method with some state-of-the-art methods. The proposed method is also applied on a communities and crime unnormalized data set to illustrate its empirical usefulness.

1. Introduction

Linear regression has been widely used to find the relationship between a response and certain covariates by statisticians and this topic almost occupies the central position in statistical methodologies. The classical low dimensional setting has been well studied and the least square estimator or ridge regression estimator has been shown to enjoy certain optimality as the sample size tends to infinity while the dimensionality is fixed. Besides, the past decades have witnessed a huge amount of literature dealing with the high dimensional setting since the appearance of the Lasso estimator (Tibshirani, 1996). In the high dimensional setting, it is a fundamental problem to study the sparse pattern of the regression coefficients. The existing methods mainly fall into three categories: parameter estimation, pure feature selection and hypothesis testing. Regularization based approaches have been studied by vast majority of researchers since these methods can achieve parameter estimation and feature selection simultaneously. The basic idea is to solve a minimization problem combining a loss function (e.g. minus log likelihood) and different penalty functions on the regression coefficient vector. For regularization methods it is crucial to choose appropriate penalty functions to obtain nice properties of the estimators. In addition to the Lasso estimator, representative estimators of regularization methods include SCAD estimator (Fan and Li, 2001), Adaptive Lasso estimator (Zou, 2006) and MCP estimator (Zhang, 2010). Pure feature selection methods focus on the selection of

* Corresponding author.
E-mail address: liwenzhang10@fudan.edu.cn (L. Zhang).

https://doi.org/10.1016/j.csda.2018.03.003
0167-9473/

variables and the parameter estimation is excluded in the pursuit. Representative works include the least angle regression selection (LARS) (Efron et al., 2004), the compressive sampling matching pursuit (CoSaMP) (Needell and Tropp, 2009) and the forward Lasso adaptive shrinkage (FLASH) (Radchenko et al., 2011). Compared to regularization methods, pure feature selection methods bypass the choice of tuning parameters and have nice theoretical properties as well as appealing computational feasibility. However, investigation of stopping rule which is crucial to decide the final model is lacked in these works. Hypothesis testing methods aim to formulate the variable selection problem into testing multiple hypotheses that the regression coefficients are zero. Representative works include Javanmard and Montanari (2014, 2013), Van de Geer et al. (2014) and Liu and Luo (2014). The former three works construct a "de-biased" estimator based on the Lasso estimator and apply it into the hypothesis testing. It is not investigated whether the test statistics can be used for FDR control in their works. Liu and Luo (2014) constructed the test statistics based on bias-corrected sample covariances of residuals and verified that the proposed procedure can control FDR and FDV under suitable conditions.

Methods mentioned above are all under the assumption of linear relationship between the predictors and the response, which is often too restrictive and unrealistic in real application. To deal with the nonlinear relationship between the covariates and the response, statisticians have built the additive regression model (Meier et al., 2009; Ravikumar et al., 2009; Yuan and Zhou, 2015), single index model (Foster et al., 2013; Luo and Ghosal, 2015; Zhu and Zhu, 2009), copula regression model (Masarotto et al., 2012; Pitt et al., 2006; Cai and Zhang, 2018) and so on. The models mentioned can all be included into the following model:

$$f_{\lambda_0}(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j f_{\lambda_j}(X_j) + \epsilon, \tag{1.1}$$

where Y is response variable, $X_1 \ldots, X_p$ are predictors, $f_{\lambda_j}(\cdot)$ are univariate functions and λ_j is the parameter associated with f_{λ_j}. Regression Model (1.1) has been widely used in econometrics, criminology, natural language processing and computational biology. Noh et al. (2013) investigated a plug-in approach for estimating a regression function based on copulas in the low dimensional setting. Cai and Zhang (2018) considered adaptive estimation and statistical inference for high dimensional sparse copula regression. In the present paper, we also consider high dimensional sparse Gaussian copula regression model. Different from the work of Cai and Zhang (2018) which falls into the regularization category, we formulate the variable selection for Gaussian copula regression model into a multiple testing problem. Specifically, suppose we have $(Y, \boldsymbol{X}^\top)^\top = (Y, X_1, \ldots, X_p)^\top \in \mathbb{R}^d$ where Y is the response variable, X_i's are the covariates and $d = p + 1$. We say $(Y, \boldsymbol{X}^\top)^\top$ satisfies a Gaussian copula regression model if and only if: **(I)** there exists a set of strictly increasing functions $f = \{f_0, f_1, \ldots, f_p\}$ such that the marginally transformed random vectors $\widetilde{\boldsymbol{Z}} = (\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top \triangleq (f_0(Y), f_1(X_1), \ldots, f_p(X_p))^\top \sim N(\boldsymbol{0}, \Sigma)$. **(II)** assume $\widetilde{\boldsymbol{Z}}_i = (\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i)^\top, i = 1, \ldots, n$ are n independent observations of $\widetilde{\boldsymbol{Z}}$, $(\widetilde{Y}_i, \widetilde{\boldsymbol{X}}_i)^\top$ satisfy

$$\widetilde{Y}_i = \widetilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. In this paper we transform the variable selection problem for high dimensional Gaussian copula regression model into the following multiple hypothesis tests

$$H_{0i} : \beta_i = 0, \quad \text{vs. } H_{1i} : \beta_i \neq 0, \ i = 1, \ldots, p. \tag{1.2}$$

We observe that the coefficient β_i is proportional to the covariance between the residual from the original linear regression model and the residual from the inverse regression model by viewing $\widetilde{X}_i$ as response variable under the Gaussian copula regression model. Thus we propose to construct our test statistics based on the bias-corrected sample covariances of residuals as in Liu (2013) and Liu and Luo (2014). The fundamental difference between the Gaussian copula regression model and the conventional linear regression model considered in Liu and Luo (2014) lies in that one can only observe the data set from $(Y, \boldsymbol{X}^\top)^\top$ rather than from $(\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top$, which poses great challenge to obtain the sample residuals and the theoretical property of the multiple testing procedure. To obtain good estimators for the residuals, we should first obtain decent estimators of the regression coefficient vectors. In this paper we exploit rank-based Kendall's tau to extract the correlation information of $(\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top$. Noting that the Kendall's tau based correlation matrix estimator is a sufficient statistic for estimating regression coefficients by Lasso or Dantzig selector, we thus achieve adaptive estimation of regression coefficients without knowledge of the marginal transformations. As is shown in Cai and Zhang (2018), the estimator obtained by Lasso is rate optimal under regularity conditions. In light of the work of Bickel et al. (2009), Dantzig selector can also achieve the rate optimality under some regularity conditions. After we obtain estimators of the regression coefficients, we then proceed to estimate the transformation functions. Then we can transform data set from $(Y, \boldsymbol{X}^\top)^\top$ into estimated Gaussian data from $(\widetilde{Y}, \widetilde{\boldsymbol{X}}^\top)^\top$, after which we can get the estimators for the residuals. We show that the test statistics based on the sample covariances of our estimated residuals are asymptotically normal distributed. Given the sparsity of the precision matrix of the predictors, we show that our testing procedure performs well from the FDR/FDV control point of view. FDV control is abbreviated for the control of the average number of falsely discovered variables. As is pointed out by Liu and Luo (2014), the FDV control is more suitable in the high dimensional regression setting because it is less conservative and gives an intuition on the quality of variable selection. In the extremely high dimensional setting (where dimension p can be an exponential function of sample size n), the idea of screening was initially proposed by Fan and Lv (2008) to reduce dimensionality from high