Computational Statistics and Data Analysis 79 (2014) 44–62

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Regularization and model selection for quantile varying coefficient model with categorical effect modifiers☆

Weihua Zhao [a], Riquan Zhang [b,c,*], Jicai Liu [b]

[a] School of Science, NanTong University, NanTong, 226007, PR China

[b] School of Finance and Statistics, East China Normal University, Shanghai, 200241, PR China

[c] Department of Mathematics, Shanxi Datong University, Datong, 037009, PR China

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 5 May 2014
Accepted 7 May 2014
Available online 27 May 2014

Keywords:
Quantile regression
Varying coefficient model
Variable selection
Fused Lasso
Categorical effect modifiers

ABSTRACT

A varying coefficient model with categorical effect modifiers is an effective modeling strategy when the data set includes categorical variables. With categorial predictors the number of parameters can become very large. This paper focuses on the model selection problem for varying coefficient model with categorical effect modifiers under the framework of quantile regression. After distinguishing between nominal and ordinal effect modifiers, a unified (adaptive-) Lasso-type regularization technique is proposed that allows for selection of covariates and fusion of categories of categorical effect modifiers, which can identify whether the coefficient functions are really varying with the level of a potentially effect modifying factor and provide a sparse model at different quantile levels. Moreover, the large sample properties are derived under appropriate conditions including a fixed bound on the number of parameters. The proposed methods are illustrated and investigated by extensive simulation studies and two real data evaluations.

1. Introduction

The varying coefficient model (VCM, Hastie and Tibshirani, 1993) is frequently used in statistical modeling due to its flexibility and interpretability. It has gained a lot of popularity during the past decade and it has been widely used to model complex dependency with data structures in various disciplines, such as finance, economics, medicine, ecology and biology. The standard varying coefficient model (VCM) has the form

$$Y = \beta_0(U_0) + X_1\beta_1(U_1) + \cdots + X_d\beta_d(U_d) + \epsilon, \tag{1}$$

where the functions $\beta_j(\cdot)$ $(j = 0, \ldots, d)$ are unknown, $X_1, \ldots, X_d$ are covariates, $U_0, \ldots, U_d$ are called index variables or effect modifiers, and ϵ is the error term. Note that the effect modifiers can but do not have to represent the same variable in model (1).

For continuous effect modifiers, the unknown functions $\beta_j(\cdot)$ are smooth and many papers have investigated the VCM by using local techniques or splines to approximate $\beta_j(\cdot)$s. Fan and Zhang (1999) proposed a two-step estimation procedure

☆ The research was supported in part by National Natural Science Foundation of China (11171112, 11101114, 11201190), National Statistical Science Research Major Program of China (2011LZ051), The 111 Project of China (B14019), Doctoral Fund of Ministry of Education of China (20130076110004), The Natural Science Project of Jiangsu Province Education Department (13KJB110024) and The Natural Science Fund of Nantong University (13ZY001).

* Corresponding author at: School of Finance and Statistics, East China Normal University, Shanghai, 200241, PR China.
E-mail address: zhangriquan@163.com (R. Zhang).

http://dx.doi.org/10.1016/j.csda.2014.05.003
0167-9473/

for the varying coefficient model when the coefficient functions have possibly different degrees of smoothness. Cai et al. (2000) developed a more efficient estimation procedure for varying coefficient models in the framework of generalized linear models. As special cases of VCM, time-varying coefficient models are particularly appealing in longitudinal studies, survival analysis and time series data. For more details, readers can refer to Hoover et al. (1998), Fan and Zhang (2000, 2008) and the references therein. With high-dimensional covariates in model (1), sparse modeling is often considered superior, owing to enhanced model predictability and interpretability, which is equivalent to determine if $\beta_j(\cdot) = 0$. Wang and Xia (2009) applied the KLASSO method, which combined the local polynomial smoothing and the group Lasso (Yuan and Lin, 2006), to select important covariates. Wang et al. (2008) conducted variable selection for VCM by using polynomial spline approximation and SCAD-penalization. On the other hand, it is also necessary to investigate which functions $\beta_j(U_j)$ actually vary over U_j. Borrowing the idea of KLASSO, Hu and Xia (2012) distinguished between the varying and non-varying coefficients by using difference penalty, while Leng (2009) distinguished between the varying and non-varying coefficients by applying the Cosso method (Lin and Zhang, 2006).

It is well known that the estimation and variable selection based on the least squares or likelihood method may be unstable and suffer from poor performance when the error distribution in (2) has heavy tails and/or there are some outliers in the data set. Quantile regression (Koenker and Bassett, 1978), one important robust regression method, has received much attention as it can provide richer information than the classic mean regression. For a complete review on quantile regression, see Koenker (2005). The variable selection approach based on quantile regression has also been paid much attention recently. Wu and Liu (2009) discussed the variable selection of quantile regression with adaptive-Lasso and SCAD penalties. Belloni and Chernozhukov (2011) derived a nice error bound on the estimation error and the number of selected variables of the quantile regression with the Lasso penalty. Wang et al. (2012) further investigated nonconvex penalized quantile regression for analyzing ultra-high dimensional data. On the other hand, Kato (2011) focused on the group selection problem for high dimensional sparse quantile regression model, Bang and Jhun (2012) proposed the regularized simultaneous multiple quantile regression by using adaptive sup-norm penalty. In addition, Li et al. (2010), Kyung et al. (2010) and Alhamzawi et al. (2012) proposed the quantile regularized estimation methods from the perspective of the Bayesian method. For the quantile varying coefficient model with continuous effect modifiers, there have been several articles for related research. For example, by using the B-spline approximation for varying coefficient functions, Hohsuk et al. (2012) and Tang et al. (2013) studied variable selection of VCM in quantile regression, the former focused on the independent data while the latter considered longitudinal data; Zhao et al. (2013) developed the variable selection for quantile VCM based on the kernel estimation method.

However, in practice, we often encounter that the effect modifiers are categorical variables. More specially, with categorical effect modifiers $U_j \in \{1, \ldots, k_j\}$, the varying functions have the form $\beta_j(U_j) = \sum_{r=1}^{k_j} \beta_{jr}\mathbb{I}(U_j = r)$, where $\mathbb{I}(\cdot)$ denotes the indicator function and $\beta_{j1}, \ldots, \beta_{jk_j}$ represent parameters. Then, model (1) can be expressed as

$$Y = \sum_{r=1}^{k_0} \beta_{0r}\mathbb{I}(U_0 = r) + \sum_{j=1}^{d} X_j \sum_{r=1}^{k_j} \beta_{jr}\mathbb{I}(U_j = r) + \epsilon. \tag{2}$$

The prominent advantage of model (2) is the interaction considered between categorical effect modifiers and covariates, which can provide more information for the response variable.

Note that there are a total of $q = \sum_{j=0}^{d} k_j$ parameters need to estimate in model (2). For notation simplicity, we denote the total coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_d^T)^T$, where $\boldsymbol{\beta}_j^T = (\beta_{j1}, \ldots, \beta_{jk_j})$. In many situations, however, the number of parameters has to be reduced in order to stabilize estimation of parameters and/or to facilitate interpretation. That means, one wants to determine which predictors are influential, and if they are influential, which categories have to be distinguished. Therefore, regularization techniques are needed. For that purpose, recently, Gertheiss and Tutz (2012) proposed a penalty approach that accounts for both variable selection with respect to predictors X_j and investigate whether the functions $\beta_j(\cdot)$ are (partially) constant or not, i.e., to decide if some of the parameters β_{jr} and β_{js} are equal for fixed j. But they treated the case of Gaussian responses only. Oelker et al. (2012) further investigated it in the framework of generalized linear model. Except the above two papers, little work has been done to discuss the regularization and model selection methods for VCM with categorical effect modifiers. Due to the advantages of quantile regression, in this paper, we study the estimation and variable selection of the VCM with categorical effect modifiers based on the quantile regression method. Besides the proposed procedure is a robust approach, it can provide the complete description of the conditional response distribution, which can uncover different structural relationships and interaction effect between covariates, categorical variables and response at the upper or lower tails.

The rest of this paper is organized as follows. In Section 2, we first describe a penalized check loss function criterion to achieve our goals. Then some computational aspects for the proposed penalized estimation are discussed and the cross-validation is recommended to select the tuning parameters. To achieve the consistent in terms of variable selection and the identification of relevant coefficients' difference, we investigate an adaptive version QR penalized estimation and its large sample properties are derived under appropriate conditions including a fixed bound on the number of parameters in Section 3. In Section 4, the proposed methods are examined by two simulation studies together with some comparisons with the least squares based method (Gertheiss and Tutz, 2012). Two real data examples are used to further illustrate the