Journal of Multivariate Analysis 136 (2015) 56–74

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Shrinkage ridge estimators in semiparametric regression models

Mahdi Roozbeh

Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, P.O. Box 35195-363, Semnan, Iran

ARTICLE INFO

Article history:
Received 8 November 2013
Available online 19 January 2015

AMS subject classifications:
primary 62G08
secondary 62J05
62J07

Keywords:
Generalized restricted ridge estimator
Kernel smoothing
Linear restriction
Multicollinearity
Positive-rule shrinkage
Semiparametric regression model
Stein-type shrinkage

ABSTRACT

In this paper, ridge and non-ridge type shrinkage estimators and their positive parts are defined in the semiparametric regression model when the errors are dependent and some non-stochastic linear restrictions are imposed under a multicollinearity setting. The exact risk expressions in addition to biases are derived for the estimators under study and the region of optimality of each estimator is exactly determined. Also, necessary and sufficient conditions, for the superiority of the ridge type estimator over its counterpart, for selecting the ridge parameter k are obtained. Lastly, a simulation study and real data analysis are performed to illustrate the efficiency of proposed estimators based on the minimum risk criterion. In this regard, kernel smoothing and modified cross-validation methods for estimating the non-parametric function are used.

1. Introduction

Let $(y_1, \boldsymbol{x}_1, t_1), \ldots, (y_n, \boldsymbol{x}_n, t_n)$ be observations that follow the semiparametric regression model (SRM)

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f(t_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is an unknown p-dimensional parameter vector, the t_i's are known and non-random in some bounded domain $D \in \mathbb{R}$, $f(t_i)$ is an unknown smooth function and ϵ_i's are independent and identically distributed random errors with mean 0, variance σ^2, which are independent of $(\boldsymbol{x}_i, t_i)$. Semiparametric regression models are more flexible than standard linear models since they have a parametric and a nonparametric component. They can be a suitable choice when one suspects that the response y linearly depends on x, but that it is nonlinearly related to t.

Surveys regarding the estimation and application of the model (1.1) can be found in the monograph of Härdle et al. [13]. Bunea [10] suggested a consistent covariate selection technique in an SRM through penalized least squares criterion. He showed that the selected estimator of the linear part is asymptotically normal. For bandwidth selection in the context of kernel-based estimation in model (1.1), Li and Palta [23] and Li et al. [24] used cross-validation criteria for optimal bandwidth selection. Raheem et al. [30] considered absolute penalty and shrinkage estimators in PLMs where the vector of coefficients $\boldsymbol{\beta}$ in the linear part can be partitioned as $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, $\boldsymbol{\beta}_1$ is the coefficient vector of the main effects, and $\boldsymbol{\beta}_2$ is the vector of the nuisance effects.

E-mail addresses: m.roozbeh.stat@gmail.com, mahdi.roozbeh@profs.semnan.ac.ir.

http://dx.doi.org/10.1016/j.jmva.2015.01.002
0047-259X/

Now, consider a semiparametric regression model in the presence of multicollinearity. The existence of multicollinearity may lead to wide confidence intervals for the individual parameters or linear combination of the parameters and may produce estimates with wrong signs. For our purpose we only employ the ridge regression concept due to Horel and Kennard [17], to combat multicollinearity. There are a lot of works adopting ridge regression methodology to overcome the multicollinearity problem. To mention a few recent researches in full-parametric regression, see [32,31,29,1,12,28,14,15,19,21,22,2,20,27]. The main focus of this approach is to develop necessary tools for computing the risk function of regression coefficient in a semiparametric regression model based on the eigenvalues of design matrix. We are also seeking a new estimator for shrinkage parameter by making use of the existing ones in the literature. It will be shown that the new estimator performs better than all the others not only for the regression coefficient, but even for the non-parametric component as well.

The study is organized as follows: In Section 2, Stein-type shrinkage as well as its positive part are defined for the regression coefficient, while their biases and risks are driven in Section 3 and detailed analysis is incorporated to compare the performance of the proposed estimators for different values of the ridge parameter. In Section 4, the least/most values of the ridge parameter are identified for which the ridge estimators dominate each other. Section 5 contains the simulation studies and a real data example related to the hedonic prices of housing attributes to demonstrate the performance of the proposed estimators, numerically.

2. The proposed estimators

Consider the following semiparametric regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}(t) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is an $n \times p$ matrix, $\boldsymbol{f}(t) = (f(t_1), \ldots, f(t_n))'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. We assume that in general, $\boldsymbol{\epsilon}$ is a vector of disturbances, which is distributed as a multivariate normal, $N_n(\boldsymbol{0}, \sigma^2\boldsymbol{V})$, where $\boldsymbol{V}$ is a symmetric, positive definite known matrix and σ^2 is an unknown parameter.

In this paper we confine ourselves to the semiparametric kernel smoothing estimator of $\boldsymbol{\beta}$, which attains the usual parametric convergence rate $n^{1/2}$ without under smoothing the nonparametric component $f(\cdot)$ [34]. Assume that $(y_i, \boldsymbol{x}_i, t_i)$, $i = 1, \ldots, n$ satisfy model (1.1). Since $E(\varepsilon_i) = 0$, we have $f(t_i) = E(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$ for $i = 1, \ldots, n$. Hence, if we know $\boldsymbol{\beta}$, a natural nonparametric estimator of $f(\cdot)$ is

$$\hat{f}(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t)(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{2.2}$$

where the positive weight functions $W_{ni}(\cdot)$ satisfy three conditions below:

(i) $\max_{1\le i\le n} \sum_{j=1}^{n} W_{ni}(t_j) = O(1)$,
(ii) $\max_{1\le i,j\le n} W_{ni}(t_j) = O(n^{-2/3})$,
(iii) $\max_{1\le i\le n} \sum_{j=1}^{n} W_{ni}(t_j) I(|t_i - t_j| > c_n) = O(d_n)$,

where I is the indicator function, c_n satisfies $\limsup_{n\to\infty} nc_n^3 < \infty$, and d_n satisfies $\limsup_{n\to\infty} nd_n^3 < \infty$.

The above assumptions guarantee the existence of $\hat{f}(t, \boldsymbol{\beta})$ at the optimal convergence rate $n^{-4/5}$, in semiparametric regression models with probability one. See Müller [26] for more details.

To estimate $\boldsymbol{\beta}$, we use the generalized least squares estimator (GLSE) given by

$$\hat{\boldsymbol{\beta}}_G = argmin_{\boldsymbol{\beta}}\, SS(\boldsymbol{\beta}) = \boldsymbol{C}^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{V}^{-1}\widetilde{\boldsymbol{y}}, \quad \boldsymbol{C} = \widetilde{\boldsymbol{X}}'\boldsymbol{V}^{-1}\widetilde{\boldsymbol{X}}, \tag{2.3}$$

where $SS(\boldsymbol{\beta}) = (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})$, $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_n)'$, $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n)'$, $\widetilde{y}_i = y_i - \sum_{j=1}^{n} W_{nj}(t_i)y_j$ and $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \sum_{j=1}^{n} W_{nj}(t_i)\boldsymbol{x}_j$ for $i = 1, \ldots, n$.

In this section, we will discuss about a biased estimation technique under multicollinearity. Simultaneously, we assume that $\boldsymbol{\beta}$ satisfies a linear non stochastic constraint, i.e.,

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}, \tag{2.4}$$

where $\boldsymbol{H}$ is a $q \times p$ non zero matrix with rank $q < p$ and $\boldsymbol{h}$ is a $q \times 1$ vector. In this paper, we refer restricted semiparametric regression model (RSRM) to (2.1).

For the RSRM, one generally adopts the well-known generalized restricted estimator (GRE)

$$\hat{\boldsymbol{\beta}}_{GR} = \hat{\boldsymbol{\beta}}_G + \boldsymbol{C}^{-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}')^{-1}(\boldsymbol{h} - \boldsymbol{H}\hat{\boldsymbol{\beta}}_G). \tag{2.5}$$

The GRE is widely applied as an unbiased estimator. In practice, the researchers often encounter the problem of multicollinearity. That is, $\boldsymbol{C}$ is always ill-conditioned due to linear relationship among the regressors of $\widetilde{\boldsymbol{X}}$ matrix. Therefore, the unknown coefficients, which are estimated by GLSE, are usually unstable and give misleading information. To overcome this problem, many studies on the general linear model without linear restriction have been made. In fact, the coefficient parameter $\boldsymbol{\beta}$ can be regarded as a vector in p dimensions space. If there exists multicollinearity in $\boldsymbol{C}$, the $\hat{\boldsymbol{\beta}}_{GR}$ would be badly apart