Journal of Multivariate Analysis 136 (2015) 26–40

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Optimal partial ridge estimation in restricted semiparametric regression models

Morteza Amini [a,b], Mahdi Roozbeh [c,*]

[a] Department of Statistics, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, P.O. Box 14155-6455, Tehran, Iran

[b] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran

[c] Department of Statistics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box 35195-363, Semnan, Iran

ARTICLE INFO

Article history:
Received 13 June 2014
Available online 17 January 2015

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

Keywords:
Generalized restricted ridge estimator
Kernel smoothing
Linear restriction
Multicollinearity
Semiparametric regression model

ABSTRACT

This paper is concerned with the ridge estimation of the parameter vector $\boldsymbol{\beta}$ in partial linear regression model $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + f(t_i) + \epsilon_i,\ 1 \le i \le n$, with correlated errors, that is, when $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, with a positive definite matrix $\boldsymbol{V}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, under the linear constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, for a given matrix $\boldsymbol{R}$ and a given vector $\boldsymbol{r}$. The partial residual estimation method is used to estimate $\boldsymbol{\beta}$ and the function $f(\cdot)$. Under appropriate assumptions, the asymptotic bias and variance of the proposed estimators are obtained. A generalized cross validation (GCV) criterion is proposed for selecting the optimal ridge parameter and the bandwidth of the kernel smoother. An extension of the GCV theorem is established to prove the convergence of the GCV mean. The theoretical results are illustrated by a real data example and a simulation study.

1. Introduction

Semiparametric regression models or partial linear models are suitable models, when a suitable link function of the mean response is assumed to have a linear parametric relationship to some explanatory variables and its relation to other variables has an unknown form. Let $(y_1, x_1, t_1), \ldots, (y_n, x_n, t_n)$ be observations that follow the semiparametric regression model

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

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 design points which belong to some bounded domain $D \in \mathbb{R}$, $f(t)$ is an unknown smooth function and ϵ_i's are random errors which are assumed to be independent of $(\boldsymbol{x}_i, t_i)$.

This model is first considered by Engle et al. [3] to study the effect of weather on electricity demand, in which they assumed that the mean relationship between temperature and electricity usage was unknown while other related factors such as income and price were parameterized linearly. Surveys regarding the estimation and application of the model (1.1) can be found in the monograph of Härdle et al. [7]. Speckman [16] studied partial residual estimation of $\boldsymbol{\beta}$ and $f(\cdot)$ in (1.1) and obtained asymptotic bias and variance of the estimators. He showed that these estimators are less biased compared to the

* Corresponding author.
E-mail addresses: morteza.amini@ut.ac.ir (M. Amini), m.roozbeh.stat@gmail.com, mahdi.roozbeh@profs.semnan.ac.ir (M. Roozbeh).

http://dx.doi.org/10.1016/j.jmva.2015.01.005
0047-259X/

partial smoothing spline estimators. You and Chen [22] considered the problem of estimation in model (1.1) with serially correlated errors, obtained the semiparametric generalized least squares estimator of the parametric component and studied the asymptotic properties of the estimator. You et al. [23] developed statistical inference for the model (1.1) for both heteroscedastic and/or correlated errors. The general assumption $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, with a positive definite matrix $\boldsymbol{V}$ is assumed. For bandwidth selection in the context of kernel-based estimation in model (1.1), Li et al. [11] used cross-validation criteria for optimal bandwidth selection.

Consider a semiparametric regression model in the presence of multicollinearity, or an overfitting caused by a large number of variables, which is often called "large p small n problem". 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. A severe multicollinearity or overfitting may also lead to the singularity of the matrix $\boldsymbol{X}'\boldsymbol{X}$. In the partial residual estimation, the singularity problem may be intensified, when we consider $\widetilde{\boldsymbol{X}}'\boldsymbol{X}$ in place of $\boldsymbol{X}'\boldsymbol{X}$, where $\widetilde{\boldsymbol{X}} = (\boldsymbol{I}_n - \boldsymbol{K})\boldsymbol{X}$ is the partial residual adjusted design matrix and $\boldsymbol{K}$ is the smoother matrix. An efficient approach to combat such singularity and ill-posed problems is the ridge estimation (see [9]). Hu [10] developed the ridge estimator of the parametric and nonparametric parts in a semiparametric regression model.

The restricted models are widely applicable in the problem of general hypothesis testing specially the generalized likelihood ratio (GLR) tests in regression models. Akdenïz and Tabakan [1] developed the restricted ridge estimators in semiparametric regression models. The problem of restricted ridge partial residual estimation in a semiparametric regression model with correlated errors is studied by Roozbeh et al. [15]. They derived the asymptotic distributional bias and the risk of the estimators under the balanced loss function. The feasible restricted ridge estimation in a semiparametric regression model with correlated errors is considered by Roozbeh and Arashi [14] using kernel smoothing and cross validation methods. They obtained the necessary and sufficient conditions for the superiority of the ridge type estimator over non-ridge type estimator.

The main focus of this paper is to study the asymptotic properties of the restricted ridge partial residual estimators of $\boldsymbol{\beta}$ and $f(\cdot)$ in model (1.1) with correlated errors. The estimation method is presented in Section 2. Section 3 is devoted to obtaining the asymptotic bias and variance of the proposed estimators. To select the optimal ridge parameter and the bandwidth of the kernel smoother, a generalized cross validation criterion is proposed in Section 4. An extension of the GCV theorem of Golub et al. [6] is established to prove the convergence of the expectation of the GCV criterion. Finally, in Section 5, the theoretical results are applied to analyze the Canadian crime rate data set. The optimal selection of the ridge parameter and the bandwidth is demonstrated in a simulation case study.

2. Restricted ridge partial residual estimation

The semiparametric regression model (1.1) can be rewritten in the matrix form as

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is the $n \times p$ fixed known design matrix, $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. Also, suppose that $t_1, \ldots, t_n$ have bounded domain $D \subset \mathbb{R}^k$.

We assume that in general, the error term $\boldsymbol{\epsilon}$ satisfies $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\boldsymbol{V}$, where σ^2 is unknown parameter and $\boldsymbol{V}$ is a symmetric positive definite known matrix.

To estimate $\boldsymbol{\beta}$ and $f(t)$ for a point $t \in D$, first consider the simplified model

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.2}$$

obtained from (2.5) with $\boldsymbol{\beta} = \boldsymbol{0}$. The linear smoother of $f(t)$ in (2.2) is $\hat{f}(t) = \boldsymbol{k}(t)\boldsymbol{y}$, with $\boldsymbol{k}(t) = (K_{n\omega}(t, t_1), \ldots, K_{n\omega}(t, t_n))$, where $K_{n\omega}(\cdot)$ is a kernel function of order m with bandwidth parameter ω.

If the kernel function $K_{n\omega}(\cdot)$ is of order m, according to [16], there exist bounded functions h_1 and h_2, such that for each $t \in D$,

$$\mathrm{E}\big(\boldsymbol{k}(t)\boldsymbol{y} - f(t)\big) = \omega^m h_1(t) f^{(m)}(t) + o(\omega^m) \tag{2.3}$$

and

$$\mathrm{Var}\big(\boldsymbol{k}(t)\boldsymbol{y}\big) = \sigma^2(n\omega)^{-1} h_2(t)(1 + o(1)), \tag{2.4}$$

where $f^{(m)}(t)$ is the mth derivative of $f(t)$.

To estimate the parameters of the model (2.1), we first remove the non-parametric effect, apparently. Assuming $\boldsymbol{\beta}$ to be known, a natural nonparametric estimator of $f(\cdot)$ is $\hat{f}(t) = \boldsymbol{k}(t)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Replacing $f(t)$ by $\hat{f}(t)$ in (2.1), the model is simplified to

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.5}$$

where $\widetilde{\boldsymbol{y}} = \big(\boldsymbol{I}_n - \boldsymbol{K}\big)\boldsymbol{y}$, $\widetilde{\boldsymbol{X}} = \big(\boldsymbol{I}_n - \boldsymbol{K}\big)\boldsymbol{X}$ and $\boldsymbol{K}$ is the smoother matrix with i, jth component $K_{n\omega}(t_i, t_j)$.

We can estimate the linear parameter $\boldsymbol{\beta}$ in (2.1) under the assumption $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, by minimizing the generalized sum of squared errors

$$SS(\boldsymbol{\beta}) = (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}). \tag{2.6}$$