REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 47–69

CHOICE OF SMOOTHING PARAMETER FOR KERNEL TYPE RIDGE ESTIMATORS IN SEMIPARAMETRIC REGRESSION MODELS

Authors: Ersin Yilmaz
– Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey
yilmazersin13@hotmail.com

Bahadir Yuzbasi
– Department of Econometrics, Inonu University, Malatya, Turkey
b.yzb@hotmail.com

Dursun Aydin
– Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey
duaydin@hotmail.com

Received: August 2017 Revised: July 2018 Accepted: September 2018

Abstract:

- This paper concerns kernel-type ridge estimators of parameters in a semiparametric model. These estimators are a generalization of the well-known Speckman's approach based on kernel smoothing method. The most important factor in achieving this smoothing method is the selection of the smoothing parameter. In the literature, many selection criteria for comparing regression models have been produced. We will focus on six selection criterion improved version of Akaike information criterion (AIC_c), generalized cross-validation (GCV), Mallows' C_p criterion, risk estimation using classical pilots (RECP), Bayes information criterion (BIC), and restricted maximum likelihood (REML). Real and simulated data sets are considered to illustrate the key ideas in the paper. Thus, suitable selection criterion are provided for optimum smoothing parameter selection.

Keywords:

- *semiparametric model; kernel smoothing; ridge type estimator; smoothing parameter generalized cross-validation.*

AMS Subject Classification:

- 62G08, 62J07, 65C60.

1. INTRODUCTION

Let us consider the following semiparametric regression model:

$$y_i = \mathbf{x}_i\beta + f(t_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where the y_i's are observations, the $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ are known p-vectors with $p < n$ and t_i's have bounded support, say the unit interval and have been reordered so that $t_i \leq t_2 \leq \cdots \leq t_n$. $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{\mathsf{T}}$ is an unknown p-dimensional vector of parameters, $f(\cdot)$ is unknown function and ε_i's are the random error terms assumed to be uncorrelated with mean zero and variance σ^2. Note that f symbolizes the smooth part of the model and assume that it shows the unparameterized functional relationship.

The model (1.1) is also called as a partially linear model, due to the connection with the classical linear model (see [8]). In matrix-vector form, the model (1.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\mathbf{y} = (y_1, ..., y_n)^{\mathsf{T}}$, $\mathbf{X} = (x_1, ..., x_n)^{\mathsf{T}}$, $\mathbf{f} = (f(t_1), f(t_2), ..., f(t_n))^{\mathsf{T}}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathsf{T}}$. The key idea is to estimate the unknown parameter vector $\boldsymbol{\beta}$, the nonparametric function $f(t)$ and the mean vector $\mu = \mathbf{X}\boldsymbol{\beta} + \mathbf{f}$ based on the data $y_i, \mathbf{x}_i, t_i$. Note that semiparametric models have received a considerable attention in the past two decades. One of the most important reasons for this is that these models are more flexible than the standard linear model because they combine both parametric and nonparametric components. In this context, a number of authors have studied the model (1.1), including Green and Silverman [12], Speckman [30], Eubank *et al.* [9], Schimek [28], Liang [21], Aydin *et al.* [3], Ahmed [1] and among others.

In many regression problems, there is a perfect or exact relationship between the columns of $\mathbf{X}$. In this case, multicollinearity is a serious problem which can dramatically influence the effectiveness of a regression model. The multicollinearity results in large variances and covariances of the parameter estimates and may lead to lack of statistical significance of individual parameters even though the overall model may be significant. For the purposes of the paper, we will employ the kernel type ridge regression procedure that is designed to deal with multicollinearity in semiparametric regression.

Concerning the collinear data, Gibbons [11] introduced a simulation study of ridge estimators for parametric linear models. Kibria [18] proposed some new estimators based on generalized ridge regression approach and considered some methods to estimate ridge parameter. For the linear regression models Muniz and Kibria [23] reviewed and proposed some estimators based on Kibria [18]. Key references for semiparametric regression based on kernel smoothing are Robinson [26] and Speckman [30]. It should be noted that Robinson [26] introduced an estimator for parametric part of a semiparametric model when nonparametric component is stochastic and of arbitrary dimension. Speckman [30] discussed two estimation method, one related to partial smoothing spline and the other modified by partial residual, in estimating the components of a semiparametric model and examined the asymptotic behaviours for both methods. Chen [6] studied the parametric component of the partial linear model. Foucart [10] used the ridge estimators on partial linear models for combat multicollinearity. Ridge estimation of a semiparametric regression model and a comparison of this ridge estimation with two steps estimation are introduced by Hu [15].

Roozbeh *et al.* [27], Yuzbasi and Ahmed [36] and Yuzbasi *et al.* [37] proposed a semiparametric ridge regression estimator for partially linear models. More recently, semiparametric regression models based on different selection methods were studied and compared by Aydin [2]. Lastly, the pretest and shrinkage ridge regression estimators based on smoothing spline approach for partially linear models was studied by Yuzbasi [35] and modified estimators in semiparametric regression models based on right censored data is studied by Aydin and Yilmaz [4].

The main difference of our study is that we consider various kernel type ridge estimators to estimate the components of a semiparametric regression model with collinear data. The most important issue in this problem is to determine an amount of smoothing. In order to specify an optimum smoothing parameter we use six different selection criteria under simulated and real data settings. The basic idea is to find a useful selection criteria that provides a good estimation of the model (1.1) based on multicollinear data. Due to smoothing parameter selection criteria, we provide a comparison of the different ridge type estimators. To the best of our knowledge, the studies in the literature often address the problem of comparing different ridge type estimators and the selection of ridge parameter, but such a study that includes kernel type ridge estimators based on different selection criteria has not yet been conducted. This paper is organized as follows. Estimation based on kernel smoothing is examined in Section 2. In Section 3, the kernel type ridge estimators in semiparametric models are discussed. Statistical properties of the ridge type estimators are examined in Section 4. Section 5 reviews six different smoothing parameter selection methods. Section 6 compares these methods via a real example. In Section 7, a simulation study is given. Finally, concluding remarks are presented in Section 8. Supplemental technical materials are relegated to the Appendix.

2. ESTIMATION BASED ON KERNEL SMOOTHING

First we consider the nonparametric estimation of the unknown regression function $f(t)$ in (1.1). For convenience, we assume that $\boldsymbol{\beta}$ in equation (1.1) is known. In this case, the relationship between $y_i - \mathbf{X}_i\boldsymbol{\beta}$ and t_i can be denoted by

$$(y_i - \mathbf{X}_i\boldsymbol{\beta}) = f(t_i) + \varepsilon_i, \quad i = 1, ..., n. \tag{2.1}$$

Equation (2.1) can be considered as equivalent to the nonparametric part of a semiparametric model. As expressed in the study of Speckman [30], this leads to the Nadaraya-Watson estimator proposed by Nadaraya [24] and Watson [34], and this is also referred to as the kernel estimator:

$$\hat{f}_\lambda(t) = \sum_{i=1}^{n} w_{i\lambda}(t_i)(y_i - \mathbf{X}_i\boldsymbol{\beta}) = \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.2}$$

where λ is a smoothing parameter (or bandwidth) and $\mathbf{W}_\lambda$ is a kernel smoother matrix with j-th entries $w_{i\lambda}$, given by

$$w_{i_\lambda}(t_i) = K\left(\frac{t - t_i}{\lambda}\right) \Bigg/ \sum_{i=1}^{n} K\left(\frac{t - t_i}{\lambda}\right) = K(u_i)/\sum_i K(u_i). \tag{2.3}$$

As shown in (2.1), kernel smoothing (or regression) uses the appropriate weights $w_{i\lambda}(t)$ to estimate $f(t)$. The weights given to the observations t_i are directed by the kernel function $K(u)$ with a smoothing parameter λ, which controls the size of the neighborhood around t [31]. Note that $K(u)$ in (2.3) is a kernel or weight function such that $\int K(u)du = 1$, and $K(u) = K(-u)$. The kernel function is selected to give most weight to observations close to and least weight to observations far from t.

Using the matrix and vector form of the model (1.2), we can obtain the following partial residuals in matrix form:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \hat{\mathbf{f}} = (\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}, \tag{2.4}$$

where $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}$ and $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$. Thus, we obtain a transformed set of data based on kernel residuals. Considering these partial residuals for the vector $\boldsymbol{\beta}$ yields the following weighted least squares (WLS) criterion:

$$\mathrm{WLS}(\boldsymbol{\beta}; \boldsymbol{\lambda}) = ((\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))^{\mathsf{T}} ((\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) = \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)' \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right). \tag{2.5}$$

In analogy with ordinary least squares, the solution to the criterion $\mathrm{WLS}(\boldsymbol{\beta}; \lambda)$ given in equation (2.5) is easily seen to be

$$\hat{\boldsymbol{\beta}}_p = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}' \tilde{\mathbf{y}}. \tag{2.6}$$

Moreover, according to the equation (2.3) updating the steps for $f(t)$ simplifies to

$$\hat{f}_\lambda(t) = \sum_{i=1}^{n} K\left(\frac{t - t_i}{\lambda}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{t - t_i}{\lambda}\right) \left(y_i - X_i\hat{\boldsymbol{\beta}}_p\right). \tag{2.7}$$

Equation (2.7) can also be written in a matrix form as

$$\hat{\mathbf{f}}_p = W_\lambda \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p\right). \tag{2.8}$$

Our estimate of μ_p is then

$$\mu_p = \mathbf{X}\hat{\boldsymbol{\beta}}_p + \hat{\mathbf{f}}_p = \mathbf{X}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p)\right)$$

for

$$\mathbf{H}_p = \mathbf{W}_\lambda + \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}' (\mathbf{I} - \mathbf{W}_\lambda). \tag{2.9}$$

Equations (2.6) and (2.8) are hierarchical in the sense that the adjustment is made for t. Adjusting for $\mathbf{X}$ first would produce a different estimator. One advantage of $\hat{\beta}_p$ is that, there is no iteration in calculation of $\hat{\beta}_p$ even if a non-linear smoother is used. As a result the approach requires only a standard regression routine if a computation of the $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ has been done with smoother matrix $\mathbf{W}_\lambda$.

3. KERNEL TYPE RIDGE ESTIMATORS IN SEMIPARAMETRIC MODELS

Ridge regression has been proposed by Hoerl and Kennard [13], [14] as a solution to the multicollinearity problem. It is well known that a ridge estimator provides a slight

improvement on the estimations of partial regression coefficients when the column vectors of the matrix $\mathbf{X}$ in a linear model $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are highly correlated. Generally, the linear model can be written in centered and scaled form. For notational convenience, we do not consider an explicitly centered and scaled model here.Then, the ridge estimate of $\boldsymbol{\beta}$ for some $k > 0$ can be written as

$$\hat{\boldsymbol{\beta}}_r(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \tag{3.1}$$

where $\mathbf{I}$ is $p \times p$ identity matrix and k is the shrinkage parameter, whose value is specified by the researcher. When $k = 0$ the ridge estimate corresponds to the least squares estimate. To fit the model (1.1) to data, we can use ridge regression that shrinks the regression coefficients by imposing a penalty on their size. This procedure can be related to the idea of hints due to Speckman [30], where the parameter vector β is obtained by minimizing the penalized residual sum of squares criterion

$$\text{PRSS}(\boldsymbol{\beta};\lambda) = \sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta}\right)^2 + k\sum_{j=1}^{n}\beta_j^2 = \sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{n}(0 - k\beta_j)^2, \tag{3.2}$$

where $k \geq 0$ is the shrinkage parameter that controls the magnitude of the penalty term. The basic idea is to recast the linear regression problem as a linear smoother problem for another data set. This means that if artificial data having response value zero are introduced, then a fitting procedure can be forced to shrink the coefficients toward zero.

In matrix and vector form, equation (3.2) can be rewritten as

$$\text{PRSS}\left(\boldsymbol{\beta};\boldsymbol{\lambda}\right) = \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) + k\|0 - \boldsymbol{\beta}\|^2. \tag{3.3}$$

The main objective is to find parameter vector $\boldsymbol{\beta}$ such that equation (3.3) is as small as possible. The following theorem gives the estimates.

Theorem 3.1. *Let* $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ *where* $\tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{f}} + \boldsymbol{\varepsilon}^*$, $\tilde{\boldsymbol{f}} = (\boldsymbol{I} - \boldsymbol{W}_\lambda)\,\boldsymbol{f}$ *and* $\boldsymbol{\varepsilon}^* = (\boldsymbol{I} - \boldsymbol{W}_\lambda)\,\boldsymbol{\varepsilon}$. *Also,* $\tilde{\boldsymbol{X}}$ *is a* $n \times p$ *matrix and* $\tilde{\boldsymbol{y}}$ *is a* $n \times 1$ *vector, as defined in* (2.8)*, respectively. If* $\boldsymbol{W}_\lambda$ *is an arbitrary smoother matrix then the ridge regression estimates may be computed by augmenting data*

$$\boldsymbol{X}_A = \begin{bmatrix}\tilde{\boldsymbol{X}} \\ \sqrt{k}\boldsymbol{I}_p\end{bmatrix} \qquad \text{and} \qquad \tilde{\boldsymbol{y}}_A = \begin{bmatrix}\tilde{\boldsymbol{y}} \\ 0\end{bmatrix}.$$

The kernel type ridge estimator for $\boldsymbol{\beta}$ *is indicated by* $\hat{\boldsymbol{\beta}}_R(k)$ *and given by*

$$\hat{\boldsymbol{\beta}}(k) = \left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}. \tag{3.4}$$

Proof of the Theorem 3.1 is given in Appendix A.1.

As in discussion Theorem 3.1, $\hat{\boldsymbol{\beta}}_R(k)$ is the ridge type estimator of the vector $\boldsymbol{\beta}$ in the model (1.2). When $k = 0$, the ridge estimate reduces to a Speckman estimate problem in the equation (2.8). Also, it is seen that there is a formal similarity between the equation (3.3) and ridge estimator of the linear regression model. Combining equations (3.3) and (3.4) we obtain the estimator of $\mathbf{f}$ as

$$\hat{\mathbf{f}}_R(k) = \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R(k)\right). \tag{3.5}$$

Thus the estimator (3.5) is defined as the kernel type ridge estimator for the unknown function $\mathbf{f}$ in the model (1.2).

4. FURTHER PROPERTIES OF THE ESTIMATORS

It is easily seen that equation (3.4) is identical to

$$\hat{\boldsymbol{\beta}}_R(k) = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta}_p = \left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}_p, \tag{4.1}$$

where $\hat{\boldsymbol{\beta}}_p$ is the Speckman estimate, as defined in (2.8). Using the fact $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, the equation (3.4) also becomes

$$\hat{\boldsymbol{\beta}}_R(k) = \left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}_p = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}' + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{4.2}$$

It appears from (4.2) that the ridge type estimator is clearly biased, since

$$\left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} \neq \mathbf{I}_p.$$

Hoerl and Kennard [13], [14] used this interpretation as a basis for the definition of the $\hat{\boldsymbol{\beta}}_R(k)$ with $k \geq 0$, the shrinkage parameter that controls the size of coefficients. Also, equation (4.2) can be viewed as the Speckman estimator for $k = 0$.

Using the abbreviation

$$\mathbf{G}_k = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}. \tag{4.3}$$

Moments of the kernel type ridge estimator can be obtained as follows:

$$E\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \mathbf{G}_k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right) = \boldsymbol{\beta} - k\boldsymbol{G}_k\boldsymbol{\beta} + \mathbf{G}_k\tilde{\mathbf{X}}\mathbf{f}, \tag{4.4}$$

$$\text{Bias}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\mathbf{G}_k\boldsymbol{\beta}, \tag{4.5}$$

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \sigma^2\mathbf{G}_k\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right). \tag{4.6}$$

The implementation details of Equations (4.4)–(4.6) are given in Appendix A.2. It should be noted that in practice $\boldsymbol{\beta}$ and σ^2 stated in equations above are replaced by their estimated values.

4.1. Estimating the error variance

The error variance σ^2 is usually unknown. In practice, σ^2 needs to be estimated. In a general semiparametric regression model, the estimate of variance σ^2 can be found by the residual sum of squares

$$\begin{aligned} \text{RSS} &= \left(\mathbf{y} - \hat{\mathbf{y}}'\right)'\left(\mathbf{y} - \hat{\mathbf{y}}'\right) \qquad \text{where} \quad \{\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\} \\ &= \left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right)\right)'\left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right)\right). \end{aligned}$$

Substituting $\hat{\mathbf{y}} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right) = \mathbf{H}_\lambda \mathbf{y}$, we obtain

$$\text{RSS} = (\mathbf{y} - \mathbf{H}_\lambda \mathbf{y})' (\mathbf{y} - \mathbf{H}_\lambda \mathbf{y}) = \| (\mathbf{I} - \mathbf{H}_\lambda)\, \mathbf{y} \|_2^2, \tag{4.7}$$

where $\mathbf{H}_\lambda$ is called the smoother matrix which depends on $\lambda > 0$. Note that the matrix $\mathbf{H}_\lambda$ is used to estimate the fitted values of the model in (1.2) and is expressed as

$$\mathbf{H}_\lambda = \mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\, \tilde{\mathbf{X}} \mathbf{G}_k \tilde{\mathbf{X}}'. \tag{4.8}$$

Furthermore, the expected value of RSS is

$$E(\text{RSS}) = \sigma^2 \left[n - \text{tr}(2\mathbf{H}_\lambda - \mathbf{H}_\lambda^2) \right] + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda) E(\mathbf{y}),$$

where the first term measures the variance, while the second term measures bias, respectively. Detailed implementations of the equation (4.7) and $E(\text{RSS})$ are given in Appendix A.3.

Hence, similar to ordinary least squares regression,estimation of the error variance can be defined by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{\text{tr}\,(\mathbf{I} - \mathbf{H}_\lambda)^2} = \frac{\| (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \|_2^2}{n - p}, \tag{4.9}$$

where $\text{tr}\,(\mathbf{I} - \mathbf{H}_\lambda)^2 = n - \text{tr}\left(2\mathbf{H}_\lambda - \mathbf{H}_\lambda' \mathbf{H}_\lambda\right) = n - p$ is the residual degrees of freedom. From equation (4.9) see that the degrees of freedom for RSS is also known as the number of total observations minus total number of the parameters in the model.

To show that $\hat{\sigma}^2$ is biased or unbiased for σ^2, $E(\hat{\sigma}^2)$ is found as

$$E(\hat{\sigma^2}) = \frac{1}{n-p} E\left(\| (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \|_2^2 \right) = \frac{1}{n-p} E(\text{RSS}).$$

The expected value of $E(\text{RSS})$ implies that the estimator of σ^2 in equation (4.9) has a positive bias. However, it should be noted that the (4.9) yields asymptotically negligible bias. Considering this point of view, it is noteworthy that $\hat{\sigma}^2$ is equivalent to mean square error (MSE) which is a widely used criterion for measuring the quality of estimation (see Speckman [30]).

4.2. Measuring the risk and performance efficiency

This section investigates the superiority of a biased estimator $\hat{\boldsymbol{\beta}}_{R1}(k)$ with respect to any other biased estimator $\hat{\boldsymbol{\beta}}_{R2}(k)$. It is well known that ridge type estimators are biased and need to measure the loss of information. Generally, the expected loss of a vector $\hat{\boldsymbol{\beta}}_R(k)$ estimator is measured by risk (i.e., the bias-variance decomposition). Our task is now to approximate the risk in the models in (1.1) or (1.2). Such approximations have the advantage of being simpler to optimize the practical selection of smoothing parameters. For convenience, we will work with the scalar valued mean dispersion error.

Definition 4.1. The risk is closely related to the matrix valued mean dispersion error (MDE) of an estimator $\hat{\boldsymbol{\beta}}_R(k)$ of $\boldsymbol{\beta}$. The scalar valued version of the MDE matrix is specified as

$$\text{SMDE}\left(\hat{\boldsymbol{\beta}}_R(k),\boldsymbol{\beta}\right)=E\left(\hat{\boldsymbol{\beta}}_R(k)-\boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}}_R(k)-\boldsymbol{\beta}\right)=\text{tr}\left(\text{MDE}\left(\hat{\boldsymbol{\beta}}_R(k)-\boldsymbol{\beta}\right)\right).$$

Lemma 4.1. *Consider different estimators $\hat{\beta}_{jR}(k)$ of β_j.The mean dispersion error (*MDE*) of these estimators is the sum of the covariance matrix and the squared bias:*

$$E\left(\parallel\hat{\boldsymbol{\beta}}_R(k)-\boldsymbol{\beta}\parallel^2\right)=\sum_{j=1}^{k}E\left(\hat{\beta}_{jR}(k)-\beta_j\right)^2=\text{tr}\left[\text{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right]+\left[\text{Bias}\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right]^2.$$

Note that $\text{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right)$ *is the covariance matrix of* $\hat{\boldsymbol{\beta}}_R(k)$ *and its trace can be illustrated as* $\text{tr}\left(\sum_{j=1}^{p}\text{Var}\left(\hat{\boldsymbol{\beta}}_{jR}(k)\right)\right)$.

For the proof, see Appendix A.4.

Applying the equations (4.4), (4.5) and (4.6), we obtain

$$E\left[\left(\hat{\boldsymbol{\beta}}_R(k)-\boldsymbol{\beta}\right)^2\right]=\sigma^2\mathbf{G}_k\tilde{\mathbf{X}}'\left(\mathbf{I}-\mathbf{W}_\lambda\right)^2\tilde{\mathbf{X}}\mathbf{G}_k+\mathbf{G}_k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta}\right)\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta}\right)'\mathbf{G}_k. \tag{4.10}$$

As stated in Definition 4.1, the MDE matrix decomposes into a sum of the squared bias and covariance of the estimator. Also, it can be interpreted as the mean Euclidean distance between the vectors $\hat{\boldsymbol{\beta}}_R(k)$ and $\boldsymbol{\beta}$. Thus, from Definition 4.1, the MDE matrix is written as

$$\text{MDE}\left(\hat{\boldsymbol{\beta}}_R(k),\boldsymbol{\beta}\right)=\mathbf{G}_k\left(\sigma^2\tilde{\mathbf{X}}'\left(\mathbf{I}-\mathbf{W}_\lambda\right)^2\tilde{\mathbf{X}}+\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta}\right)\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta}\right)'\right)\mathbf{G}_k. \tag{4.11}$$

As in Definition 4.1, the scalar valued version of the MDE matrix in (4.11) is given by

$$\begin{aligned}\text{SMDE}\left(\hat{\boldsymbol{\beta}}_R(k),\boldsymbol{\beta}\right)&=\text{tr}\{\text{MDE}\left(\hat{\boldsymbol{\beta}}_R(k),\boldsymbol{\beta}\right)\}\\&=\text{tr}\{\mathbf{G}_k\left(\sigma^2\tilde{\mathbf{X}}'(\mathbf{I}-\mathbf{W}_\lambda)^2\tilde{\mathbf{X}}+(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta})(\tilde{\mathbf{X}}'\tilde{\mathbf{f}}-k\boldsymbol{\beta})'\right)\mathbf{G}_k\}.\end{aligned} \tag{4.12}$$

Hence, we can compare the quality of two estimators by looking at the ratio of their SMDE in (4.12). This ratio gives the following definition concerning the superiority of any two estimators.

Definition 4.2. The relative efficiency of an estimator $\hat{\boldsymbol{\beta}}_{R1}(k)$ compared to another estimator $\hat{\boldsymbol{\beta}}_{R2}(k)$ is obtained by the ratio,

$$\text{RE}\left(\hat{\boldsymbol{\beta}}_{R1}(k),\hat{\boldsymbol{\beta}}_{R2}(k)\right)=\frac{R\left(\hat{\boldsymbol{\beta}}_{R2}(k),\boldsymbol{\beta}\right)}{R\left(\hat{\boldsymbol{\beta}}_{R1}(k),\boldsymbol{\beta}\right)}=\frac{\text{SMDE}\left(\hat{\boldsymbol{\beta}}_{R2}(k)\right)}{\text{SMDE}\left(\hat{\boldsymbol{\beta}}_{R1}(k)\right)}, \tag{4.13}$$

where $R(\cdot)$ denotes the scalar risk that is equivalent to the equation (4.12). $\hat{\boldsymbol{\beta}}_{R2}(k)$ is said to be more efficient than $\hat{\boldsymbol{\beta}}_{R1}(k)$ if $\text{RE}\left(\hat{\boldsymbol{\beta}}_{R1}(k),\hat{\boldsymbol{\beta}}_{R2}(k)\right)<1$.

5. CHOOSING THE SMOOTHING PARAMETER

The main idea of this paper is how to select the smoothing parameter expressed in a penalized residual sum of squares criterion (3.3). Our task is to select an optimum value of the λ. In practice, this can be achieved by using smoothing parameter selection criteria. A reasonable value of λ can be chosen to minimize the mentioned criteria. Examples of the most widely used selection methods are summarized as follows:

GCV Criterion: The generalized cross validation (GCV) score is specified by (see Craven and Wahba, [7])

$$\mathrm{GCV}(\lambda) = n^{-1} \parallel (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \parallel^2 \Big/ \left[n^{-1} \operatorname{tr}(\mathbf{I} - \mathbf{H}_\lambda)\right]^2,$$

where $\mathbf{H}_\lambda$, as is defined in (4.8), is the smoother matrix based on λ.

C_p Criterion: This criterion proposed by Mallows [22] is aimed to provide an estimate of the MSE in (4.9) scaled by σ^2, and given as

$$\mathrm{C}_p(\lambda) = \frac{1}{n}\{\parallel (\mathbf{H}_\lambda - \mathbf{I})\mathbf{y} \parallel^2 + 2\sigma^2 \operatorname{tr}(\mathbf{H}_\lambda) - \sigma^2\} = \frac{1}{n}\{\parallel \mathbf{y} - \hat{\mathbf{f}}_\lambda \parallel^2 + 2\sigma^2 \operatorname{tr}(\mathbf{H}_\lambda) - \sigma^2)\}.$$

If σ^2 is unknown, in practice an estimation for σ^2 can be provided by

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\hat{\lambda}_p} = \parallel (\mathbf{H}_{\hat{\lambda}_p} - \mathbf{I})\mathbf{y} \parallel^2 \Big/ \operatorname{tr}\left(\mathbf{I} - \mathbf{H}_{\lambda_p}\right),$$

where $\hat{\lambda}$ is an estimate of λ pre-chosen with any of the selection criterion (for example GCV). For details, see Liang [21], Mallows [22] and Wahba [33].

$\mathrm{AIC_c}$ Criterion: Notice that the classical Akaike information criterion tends to overfit when the sample size is relatively small. Hurvich *et al.* [16] suggested an improved version, called $\mathrm{AIC_c}$, which is defined by

$$\mathrm{AIC_c}(\lambda) = 1 + \log\left[\parallel (\mathbf{H}_\lambda - \mathbf{I})\mathbf{y} \parallel^2 \Big/ n\right] + \left[2\{\operatorname{tr}(\mathbf{H}_\lambda) + 1\} \Big/ n - \operatorname{tr}(\mathbf{H}_\lambda) - 2\right].$$

BIC Criterion: Schwarz [29] improved the Bayesian information criterion (BIC) by using Bayes estimators. Thus, the BIC is also called Schwarz Information Criterion (SIC). The criterion is expressed as

$$\mathrm{BIC}(\lambda) = 1/n \parallel (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \parallel^2 + (\log(n)/n) \operatorname{tr}(\mathbf{H}_\lambda).$$

RECP Criterion: Risk estimation criteria (RECP) measures the distance between $\mathbf{f}$ and $\hat{\mathbf{f}}_\lambda$. By direct calculation, the $\mathrm{RECP}(\hat{\lambda}_p)$ score is defined as

$$\mathrm{RECP}(\lambda_p) = 1/n\{\parallel (\mathbf{H}_\lambda - \mathbf{I})\hat{\mathbf{f}}_{\lambda_p} \parallel^2 + \hat{\sigma}^2_{\lambda_p} \operatorname{tr}(\mathbf{H}_\lambda \mathbf{H}_\lambda^{\mathsf{T}})\} = 1/nE \parallel \mathbf{f} - \hat{\mathbf{f}}_{\lambda_p} \parallel^2,$$

where $\hat{\sigma}^2_{\lambda_p}$ and $\hat{\mathbf{f}}_{\lambda_p}$ are the appropriate *pilot estimates* for σ^2 and $\mathbf{f}$, respectively. The pilot λ_p selected by classical methods is used for computation of the pilot estimates (see Lee [19], [20]).

REML Criterion: The restricted maximum likelihood (REML) criterion motivates treating λ as a variance parameter. The REML and GCV have a similar form and provide identical values. Moreover, the derivatives of both the REML and the GCV with respect to λ can be determined quite naturally in a common form (see Reiss and Ogden [25]). The REML score can be specified as

$$\text{REML}(\lambda) = \| (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \|^2 \Big/ n - \text{tr}(\mathbf{H}_\lambda).$$

5.1. Comparisons of computational times

In this paper, we discuss different parameter selection techniques proposed in the literature. Generally, they differ in the amount of computational time as well as a priori information required. The four selection methods GCV, AIC_c, BIC, and REML need approximately the same computational time for finding their corresponding smoothing parameter λ, as their computations only require one numerical minimization problem. From computational perspective, a causing difficulty term is $\text{tr}(\mathbf{H}_\lambda)$, which takes $O(n^2)$ operations to assess directly, for each set of smoothing parameters. Compared to these four methods, both C_p and RECP require a longer computation time, as they need an estimate of parameter λ pre-chosen with a selection criterion, such as GCV. So, there are two numerical minimization in computations of C_p and RECP. However, it should be noted that some calculations are unnecessary for these two numerical minimizations. For this reason, when careful programming is made, the overall calculation time will not be doubled.

6. REAL DATA EXAMPLE

In this study, to illustrate how ridge type kernel method works on real data, power plant data has been used. The power plant dataset includes 500 data points collected from a Combined Cycle Power Plant. The goal is to predict the net hourly electrical energy output (EP) of the plant from the features consisting of hourly average ambient variables such as temperature (T), ambient pressure (AP), relative humidity (RH) and exhaust vacuum (V).

Tufekci [32] has used the dataset for prediction of electrical power output of a base load operated combined cycle power plant using machine learning methods. Also, Kaya *et al.* [17] have used this data in their study called "Local and Global Learning Methods for Predicting Power of a Combined Gas and Steam Turbine".

In order to explain the variables clearly, their intervals and units are defined as follows: T, AP, RH, V and EP lie in the range 1.81–37.11 Celsius, 992.89–1033.30 milibar, 25.56%–100.16%, 25.36–81.56 cm-Hg, and 420.26–495.76 MW, respectively. The averages are taken from various locations around the plant. Also, ambient variables are recorded every second.

Scatterplot matrix and Correlogram of these variables are shown in Figures 1–2, respectively. According to Figure 1, V seems to have a curvilinear structure according to response variable EP. In this context, this variable breaches the linearity assumption of the classical

regression model. Therefore, V will compose a nonparametric part of the semiparametric regression model. Other variables have considerable linear structure; consequently, T, AP, RH variables will be the parametric component of the semiparametric model.

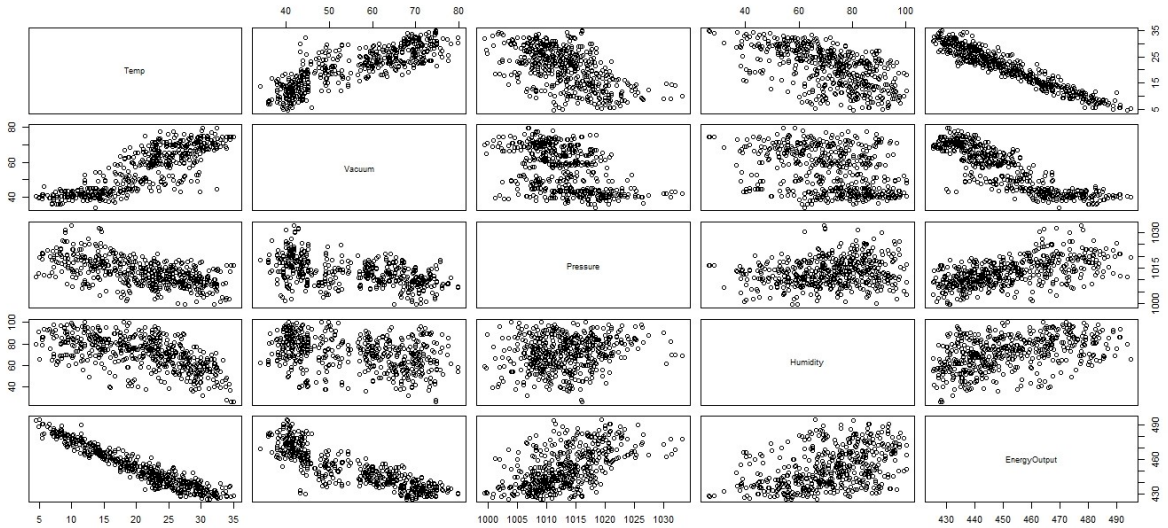

Figure 1: Scatterplot matrix of power plant data.

Thus, the semiparametric regression model in equation (1.1) can be defined the following way:

$$EP_i = \beta_1(T_i 1) + \beta_2(AP_i 2) + \beta_3(RH_i 3) + f(V_i 4) + \varepsilon_i, \quad i = 1, ..., 500. \tag{6.1}$$

Collinearity can be checked by simply calculating the correlations of the predictors in the model (6.1). Let $\mathbf{X}$ be a 500×4 matrix of the levels of the predictors in our real data example. A very simple measure of multicollinearity is inspection of the Correlogram given in Figure 2. It can be seen that several predictors have strong relationships with each other.

The eigenvalues of the $\mathbf{X}'\mathbf{X}$ for power plant data are $\lambda_1 = 0.01$, $\lambda_2 = 1613$, $\lambda_3 = 3481$, $\lambda_4 = 138950$, respectively. As is known, small eigenvalues indicate a bad condition in the data and maybe a collinearity problem. In order to determine the existence of multicollinearity, a condition index might be used. Condition Index (CI) is commonly used as an overall collinearity measure (Belsley *et al.*, [5]). If the value of CI exceeds 30, then we conclude that there is a strong multicollinearity in the data. This index is calculated as follows:

$$CI = \left[\lambda_{max}(\mathbf{X}'\mathbf{X})/\lambda_{min}(\mathbf{X}'\mathbf{X})\right]^{1/2} = 3723.10.$$

The value of $CI = 3723.10$ is an indication of potential multicollinearity problems. To combat with the collinearity, researchers use the ridge regression estimators given in (3.1). The illustration here will be based on kernel type ridge estimators given in (3.4) and (3.5). The parameter are chosen by minimizing the $\mathrm{AIC_c}, \mathrm{GCV}, \mathrm{BIC}, \mathrm{REML}, \mathrm{C}_p$ and RECP criteria, respectively. Also, the tuning parameter k is chosen with the generalized ridge regression estimator suggested by Hoerl and Kennard [13], [14]. The outcomes are given in Table 1.

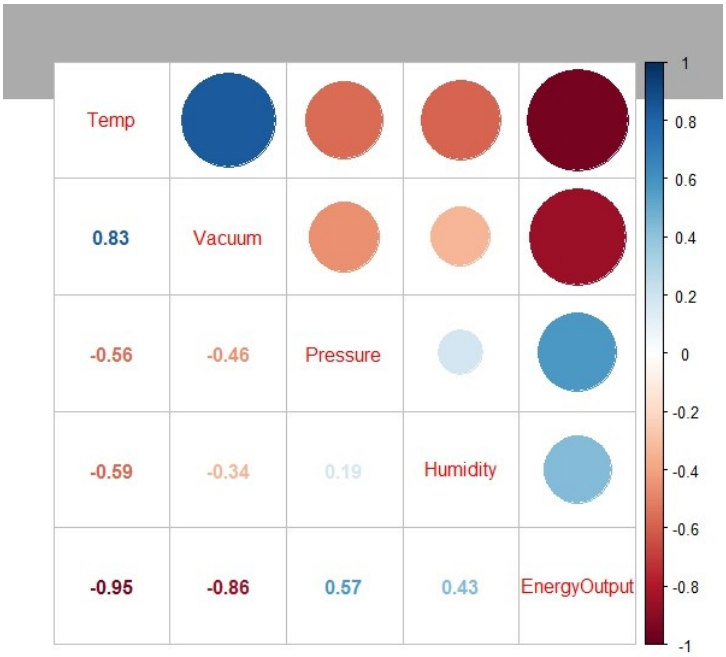

Figure 2: Correlogram for power plant data: Red colour indicates a negative correlation between the variables, while blue colour denotes a positive correlation. Size of the circle and intensity of the colour shows the strength of the relationships between variables.

As denoted in Table 1, slope parameters estimated with $\mathrm{AIC_c}$ and RECP are very similar and likewise BIC and GCV. The SMDE values, variances and bias values of the semiparametric model have been obtained from six different selection methods. The SMDE, bias and variance values calculated by RECP criterion smaller than other methods. This is indicated in bold. In this situation, it is obvious that the RECP criterion has a more convincing performance for the selection of the parameter λ and that C_p method does not perform well under study.

Table 1: Estimated coefficients of parametric component of the model.

	$\hat{\beta}_{\mathrm{AIC_c}}$	$\hat{\beta}_{\mathrm{BIC}}$	$\hat{\beta}_{\mathrm{GCV}}$	$\hat{\beta}_{\mathrm{REML}}$	$\hat{\beta}_{\mathrm{RECP}}$	$\hat{\beta}_{\mathrm{C}_p}$
T	−2.21931	−2.1932	−2.1924	−2.1928	−2.20437	−2.20721
AP	0.5024	0.4925	0.5003	0.5023	0.48276	0.47746
RH	−0.1660	−0.1678	−0.1657	−0.1659	−0.16813	−0.16862
SMDE	425.2455	571.9484	565.9700	672.4861	408.2248	696.5099
Bias	19.8500	23.2257	23.1069	25.1973	19.69050	25.67990
Variance	31.2230	32.7831	32.3670	37.9560	20.50790	37.0527

The smooth curves in Figure 3 are the graph of $\hat{y} = \hat{f}(V)$, different nonparametric estimates of the effect of V variable on EP. For smoothed curves the MISE values given in (7.2) are 24.8788, 26.0882, 26.0683, 26.3782, 24.1794 and 26.7481, respectively. Here, all of the selection methods have shown almost the same performance except the C_p criterion. Thus, we can say that the C_p does not provide a good empirical approximation.

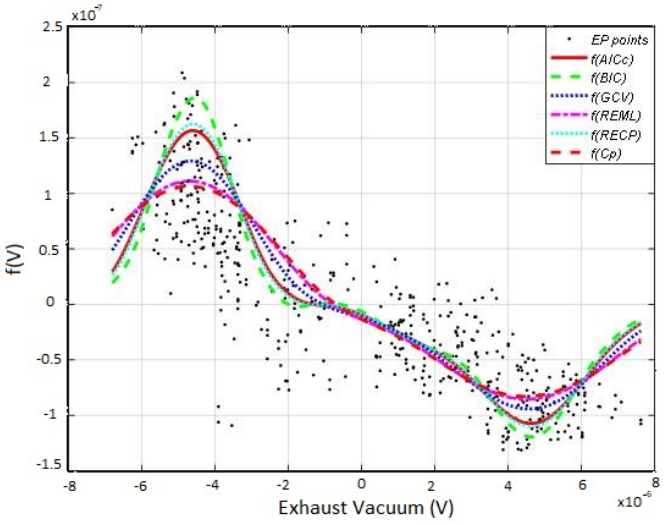

Figure 3: The smoothed curves for different kernel type ridge estimators based on $\mathrm{AIC_c}, \mathrm{BIC}, \mathrm{GCV}, \mathrm{REML}, \mathrm{RECP}$ and C_p methods, respectively.

7. MONTE CARLO SIMULATION STUDY

In this section, a Monte Carlo simulation study are carried out to compare the performance of the six selection methods expressed in Section 5. In the study, we simulate the response variable for samples of size $n = 50, 100$ and 200 with 103 iteration from the following model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta x_{i4} + f(t_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{7.1}$$

where $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ which the values of $\sigma = 0.5$ and 1, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' = (5, 4, 3, 2)'$, x_1, x_2, x_3 and x_4 are the correlated random variables, from the normal distribution. In here, three correlation (ρ) levels are considered as: 0.85, 0.95 and 0.99. Finally, the function f is represented by

$$f(t_i) = \sqrt{t_i(1 - t_i)} \sin(2\pi/t_i) \qquad \text{with} \quad t_i = (1 - 0.5)/n.$$

It should be emphasized that we investigate three correlation levels, as stated above. If $\rho = 0.85$, for instance, this allows us to obtain about the same correlation levels between all pairs of variables. They are displayed in Table 2 for detecting correlations between the explanatory variables. Note that the outcomes from correlated data based on $\rho = 0.95$, and 0.99 are not reported here, because of space limitations.

Table 2: Correlation matrix for $\rho = 0.85$ level.

X	x_1	x_2	x_3	x_4
x_1	1.00	0.83	0.83	0.82
x_2	0.83	1.00	0.83	0.86
x_3	0.83	0.83	1.00	0.84
x_4	0.82	0.86	0.84	1.00

7.1. Evaluating the parametric part

The focus of the study is to estimate the parametric and nonparametric components of the semiparametric model. Additionally, the study is illustrating behaviors and performances of the selection methods with small, medium and large samples under multicollinear data sets. For each of the data sets, 1000 estimates of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are obtained. These estimates are formed through a parametric component of the semiparametric regression model. The following tables and figures summarize the results of the simulation study.

There are four panels in Figure 4. In each panel, "AIC1, AIC2 and AIC3" denote the parametric biases of $\hat{\boldsymbol{\beta}}$ from semiparametric regression using ridge type kernel smoothing based on a smoothing parameter selected by improved $\mathrm{AIC_c}$ method for $n = 50$, 100 and 200, respectively; similarly, "BIC1, BIC2 and BIC3" denote the case using BIC method for the sample sizes; "GCV1, GCV2 and GCV3" denote the case for GCV method; "R1, R2 and R3" denote REML method; "P1, P2 and P3" denote the RECP method; "Cp1, Cp2 and Cp3" illustrate Mallows' C_p method . The ordinate indicates the scale of the biases of regression coefficients.

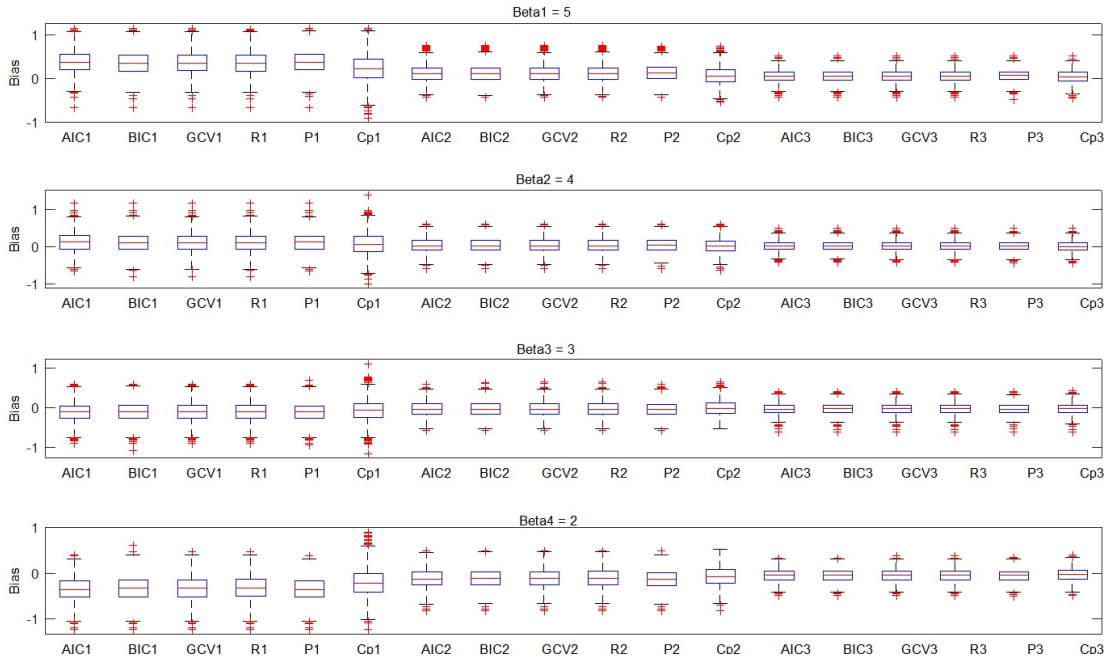

Figure 4: Boxplots of the estimates ($n = 50$, 100 and 200) obtained from semiparametric model for $\rho = 0.95$ and $\sigma = 1$. Panels indicate the boxplots of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and $\hat{\beta}_4$.

In this study, there are 18 different configurations. Since it is hard to illustrate here all of these configurations, some of them are given in Figure 4 for correlation level $\rho = 0.95$ and $\sigma = 1$. As the sample size n gets larger, the range of estimates are getting narrower. That means that estimates from medium and large sized samples are more stable than those from small sized samples. If there is a correlation between the predictors, then the sample size has an effect on the quality of parametric estimates. We can say that kernel type ridge estimators work well for all samples. The key idea of the study is to compare the SMDEs for

the estimators computed with each one of the criteria. The values of SMDE are illustrated in Table 3. The criterion that has the smallest SMDE is the best one.

Table 3: Average SMDEs of the parameters based on 1000 Monte Carlo runs.

n	ρ	CI	σ	AIC_c	BIC	GCV	REML	RECP	C_p
50	0.85	21.62	0.5	0.0043	0.0040	0.0040	0.0040	0.0040	0.0045
		34.03	1.0	0.0050	0.0048	0.0047	0.0047	0.0045	0.0055
	0.95	47.70	0.5	0.0157	0.0155	0.0159	0.0155	0.0132	0.0168
		53.21	1.0	0.0211	0.0209	0.0201	0.0193	0.0199	0.0246
	0.99	98.15	0.5	0.1348	0.1444	0.1267	0.1047	0.1366	0.1195
		100.56	1.0	0.2356	0.2244	0.2457	0.2032	**0.1995**	0.2010
100	0.85	14.78	0.5	0.0007	0.0007	0.0007	0.0007	0.0007	0.0008
		30.16	1.0	0.0010	0.0009	0.0009	0.0009	0.0009	0.0018
	0.95	45.84	0.5	0.0043	0.0043	0.0043	0.0042	0.0038	0.0044
		68.12	1.0	0.0052	0.0051	0.0052	0.0051	0.0050	0.0064
	0.99	75.24	0.5	0.0272	0.0275	0.0278	0.0268	0.0258	0.0278
		91.01	1.0	0.0385	0.0384	0.0398	0.0381	**0.0383**	0.0399
200	0.85	24.42	0.5	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
		22.31	1.0	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	0.95	57.71	0.5	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011
		63.94	1.0	0.0014	0.0014	0.0015	0.0014	0.0014	0.0018
	0.99	99.41	0.5	0.0102	0.0110	0.0111	0.0113	0.0088	0.0136
		110.48	1.0	0.0152	0.0154	0.0156	0.0151	**0.0148**	0.0177

As discussed in real data, the values of CI presented in Table 3 are given to measure the extent of multicollinearity in the simulated data sets. It is readily seen that we have mostly multicollinear data sets. According to the same table, it is possible to see that the BIC, GCV, REML and RECP outperform AIC_c and C_p criteria for samples of size $n = 50$ and $\rho = 0.85$. Also, we see that the performances of five criteria, except C_p, behaviour quite similar in the medium and large sized samples generated by various scenarios. Notice, however, that RECP has a better performance under multi-collinear data sets especially for highly correlation levels. They are indicated in bold in Table 3. A very attractive component here is that as the sample size increase, the SMDE values decrease for all criteria based on correlation level of $\rho = 0.99$.

Table 4: Simulated bias of the slope parameters for $\rho = 0.99$ and $\sigma = 0.5$.

n	$\boldsymbol{\beta}$	AIC_c	BIC	GCV	REML	RECP	C_p
50	$\hat{\beta}_1$	0.0827	0.0811	0.0866	0.0852	0.0841	0.0933
	$\hat{\beta}_2$	0.0462	0.0458	0.0472	0.0492	0.0428	0.0444
	$\hat{\beta}_3$	0.0240	0.0234	0.0248	0.0242	0.0237	0.0320
	$\hat{\beta}_4$	0.0842	0.0980	0.0951	0.0985	0.0846	0.0874
100	$\hat{\beta}_1$	0.0547	0.0559	0.0558	0.0589	0.0563	0.0393
	$\hat{\beta}_2$	0.0186	0.0177	0.0176	0.0181	0.0127	0.0175
	$\hat{\beta}_3$	0,0138	0.0151	0.0151	0.0146	0.0107	0.0153
	$\hat{\beta}_4$	0.0311	0.0391	0.0390	0.0382	0.0283	0.0321
200	$\hat{\beta}_1$	0.0219	0.0226	0.0219	0.0222	0.0150	0.0164
	$\hat{\beta}_2$	0.0031	0.0037	0.0035	0.0033	0.0019	0.0036
	$\hat{\beta}_3$	0.0032	0.0032	0.0033	0.0032	0.0028	0.0034
	$\hat{\beta}_4$	0.0152	0.0169	0.0173	0.0168	0.0126	0.0177

Table 4 presents a checking of the bias of the slope parameters of the model (7.1). The number of parameters $p = 4$ and the parametric component of the model consists of real parameter vector $\boldsymbol{\beta} = (5, 4, 3, 2)^{\mathsf{T}}$. In general, sample sizes get larger, estimates obtained by six different kernel type estimators give small bias values, as expected. Among six kernel type ridge estimators, the one obtained by using RECP criterion provide the smallest bias of the estimation of real coefficients, especially for samples of size $n = 200$. Results that related to other correlation and sigma levels are similar. So, they are not reported here.

7.2. Measuring and comparing the efficiencies

In order to illustrate and compare the efficiency of the selection methods based on highly correlated data, a relative efficiency values are constructed from the SMDE ratios in (4.13). For each sample size the mentioned values are displayed in Figure 5. As can be seen from Figure 5, relative efficiency values of the RECP are better than others except for samples of size $n = 50$ and $\rho = 0.85$. This case shows that RECP is more efficient than the other selection methods, especially for all samples based on highly correlated data. Note also that outcomes from correlated data based on $\rho = 0.90$ are similar to the results displayed in Figure 5 under $\rho = 0.99$ and are not reported here.

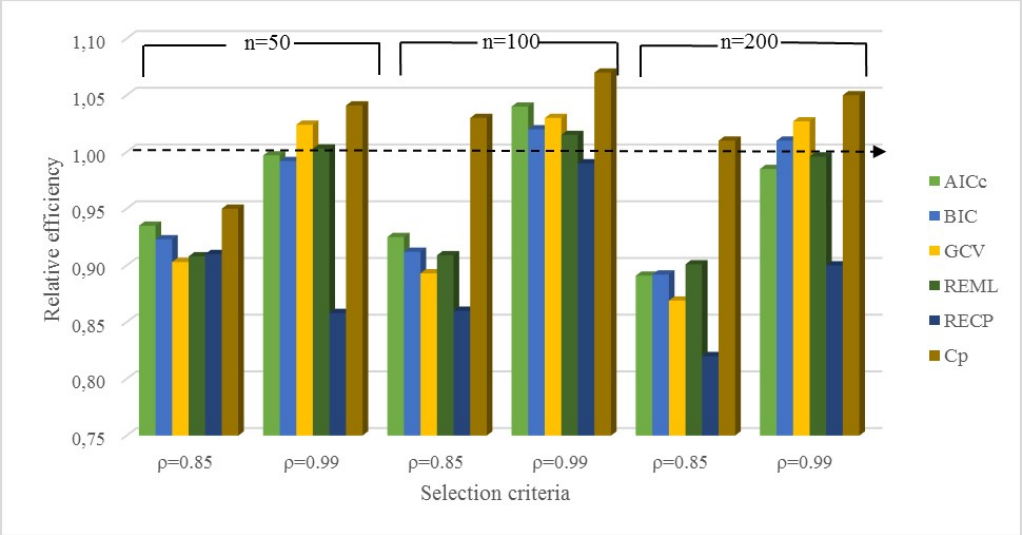

Figure 5: The column chart provides the averaged-relative efficiencies computed by the selection criteria.

Inspection of the relative efficiency values in Figure 5 also reveal that for $\rho = 0.85$, RECP criterion converges at 0.82, the highest rate when sample size is large. This indicates that under multicollinear data and noisy data, RECP criterion has the best performance among all other criteria, making it an ideal selection method for semiparametric regression based on ridge type kernel smoothing method. It can also be observed from Figure 5 in which four criteria, AIC_c, BIC, GCV and REML, perform similarly, and better than the C_p criterion.

7.3. Evaluating the nonparametric part

In order to measure the nonparametric component of the semiparametric model, 1000 estimates of function f are obtained for each selection criterion. Smoothness and appropriateness of curve estimates have been measured by using the mean of the integrated squared error (MISE) value:

$$\text{MISE} = \frac{1}{1000}\sum_{j=1}^{1000} \text{ISE}_j, \tag{7.2}$$

where ISE_j denotes the integrated square error for the sample j, given by

$$\text{ISE}_j = \int \left(f(t) - \hat{f}_j(t)\right)^2 dt \approx \frac{1}{n}\sum_{i=1}^{n}\left(f(t_i) - \hat{f}_j(t_i)\right)^2 \qquad \text{where} \quad t_i = \frac{i-0.5}{n},$$

where $f(t_i)$ value at t_i points to the appropriate function f. In our simulation study, because 18 different configurations are carried out, it is very hard to illustrate all of them. Therefore, only four different configurations will be presented in Figure 6. The left panels in the figure represent the smoothed curves together with a real function $f(t)$. In each graph, the smoothed curves, f(AICc), f(BIC), f(GCV), f(REML), f(Cp), respectively, are estimates of function $f(t)$ using ridge type kernel smoothing based on AIC_c, BIC, GCV, REML, RECP and C_p criteria. Also, the right panels of the Figure 6 denote the boxplots of the MISE values in (7.2) for each criterion.

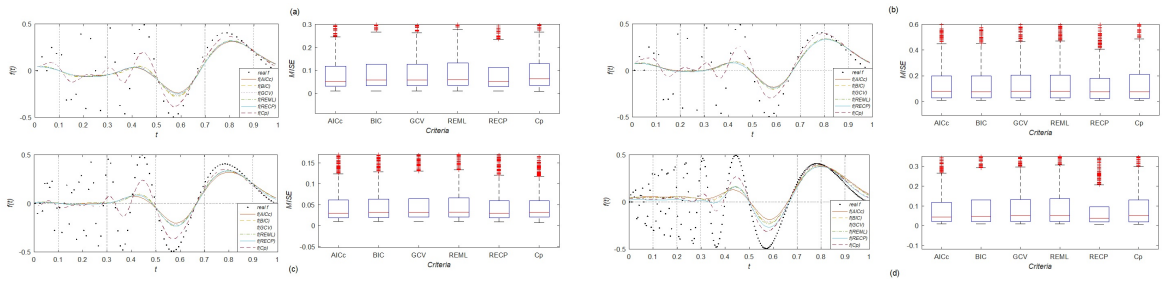

Figure 6: (a) $n = 50$, $\rho = 0.85$, $\sigma = 1$; (b) $n = 50$, $\rho = 0.99$, $\sigma = 0.5$; (c) $n = 100$, $\rho = 0.85$, $\sigma = 1$; (d) $n = 200$, $\rho = 0.99$, $\sigma = 1$.

In Figure 6 we see that the improvements in the MISE values mostly depend on the size of samples used in study. We also see that increasing the levels of correlation leads to poor performance in terms of MISE values, even if the sample sizes are the same. On the other hand, a visual inspection of the boxplots in all panels ((a) to (d)) denoted that RECP criteria maintain their dominance over the remaining selection methods, especially for large sized samples (say $n = 200$) based on data sets with $\rho = 0.99$ and $\sigma = 1$. On the contrary, the C_p criterion similar behaviors to others in terms of performance (see panels (a) and (c) of Figure 6). Notice, however, that the C_p yields poor estimates of the nonparametric component, compared to the estimates obtained by other methods, as in parametric cases.

8. CONCLUDING REMARKS

In this paper, for the parameters of the semiparametric model we proposed, kernel type ridge estimators minimize the penalized residual sum of squares method. Efficient computation of this method requires an optimum smoothing parameter λ. This optimum parameter is provided by means of AIC_c, BIC, GCV, REML, RECP and C_p criteria. Accordingly, we obtained six different estimators for the parametric and nonparametric components of the semiparametric model. We considered a real data example and simulated 1000 test observations to compare six different kernel type ridge estimators.

The empirical results confirmed that in the case of multicollinearity the kernel type ridge estimators based on AIC_c, BIC, GCV, REML and RECP, criteria have similar values of the SMDEs. The RECP, however, are superior to others in terms of SMDEs, especially when higher correlation levels are used. Throughout this discussion, the estimators based on C_p do not yield better performance in prediction of parametric and nonparametric components. On the other hand, although the REML criterion is more stable than AIC_c, GCV and RECP criteria, its performance is not good for all sample sizes and correlation levels. For the simulation studies, the findings of the numerical experiments are summarized in Tables 3–4 and Figures 4–6. We conclude the following statements from these tables and figures:

- For all the selection criteria, the SMDE, variance, and bias values of the slope parameters (or regression coefficients) start to decrease as the sample size n gets larger.
- For small sample sizes, as expected the bias values of slope parameter increase as the correlation and sigma levels increase.
- Also expected, when the lower correlation levels (i.e., $\rho = 0.85$) are used, the MISE values decreases for all selection criteria.
- Finally, when comparing the six selection methods, we see that the kernel type ridge estimators based on RECP method perform better than the others in terms of the SMDE, variance and bias values of the estimates for all sample sizes under collinear data.

A. APPENDIX: SUPPLEMENTAL TECHNICAL MATERIALS

A.1. Proof of Theorem 3.1

Consider data augmentation methods of penalized residual sum of squares fitting. Suppose that $\mathbf{W}_\lambda$ is symmetric smoother matrix. We wish to obtain the vector $\hat{\boldsymbol{\beta}}_R(k)$ that minimizes the penalized residual sum of squares criterion (3.3) by using augmented data sets of the form

$$\mathbf{X}_A = \begin{bmatrix} \tilde{\mathbf{X}}_{n\times p} \\ \left(\sqrt{k}\mathbf{I}\right)_p \end{bmatrix} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \ldots & \tilde{x}_{1p} \\ \tilde{x}_{21} & \tilde{x}_{22} & \ldots & \tilde{x}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \ldots & \tilde{x}_{np} \\ \sqrt{k} & 0 & \ldots & 0 \\ 0 & \sqrt{k} & \ldots & \vdots \\ \vdots & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & \sqrt{k} \end{bmatrix}_{((n+p)\times p)} \quad \text{and} \quad \mathbf{y}_A = \begin{bmatrix} \tilde{\mathbf{y}}_{n\times 1} \\ \mathbf{0}_p \end{bmatrix} = \begin{bmatrix} \tilde{y}_{11} \\ \tilde{y}_{21} \\ \vdots \\ \tilde{y}_{n1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{((n+p)\times 1)},$$

where $\sqrt{k}\mathbf{I}_p$ is a $p \times p$ new diagonal matrix with diagonal elements equal to the square root of the shrinkage parameter and $\mathbf{0}_p$ is $p \times 1$ new vector of zeros. Also, $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}$ and $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$ as defined in equation (2.8), are partial residuals.

Similar to the ordinary least squares, the kernel ridge type estimators can be conveniently obtained using an augmented data set. A researcher could use this information to construct a penalized least-squares estimator $\hat{\boldsymbol{\beta}}_R(k)$ of $\boldsymbol{\beta}$. The estimator can be derived by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}_A'\mathbf{X}_A\right)^{-1}\mathbf{X}_A'\mathbf{y}_A \\ &= \left(\begin{bmatrix} \tilde{\mathbf{X}}' & (\sqrt{k}\mathbf{I}_p)' \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{X}}' \\ (\sqrt{k}\mathbf{I}_p)' \end{bmatrix}\right)^{-1}\begin{bmatrix} \tilde{\mathbf{X}}' & (\sqrt{k}\mathbf{I}_p)' \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{y}} \\ \mathbf{0}_p \end{bmatrix} \\ &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \left(\sqrt{k}\mathbf{I}_p\right)^2\right)^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \left(\sqrt{k}\mathbf{I}_p\right)\mathbf{0}_p\right) \\ &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{y}}. \end{aligned}$$

Hence, as claimed, this confirms that the kernel type ridge type estimator of the unknown parameters in the models (1.1) or (1.2) is

$$\hat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}_A'\mathbf{X}_A\right)^{-1}\mathbf{X}_A'\mathbf{y}_A = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{A.1}$$

A.2. Derivation of the equations (4.4)–(4.6)

Using the definition of $\hat{\boldsymbol{\beta}}_R(k)$ ridge and our modeling assumption on the mean function $E(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\boldsymbol{\beta}$, we obtain:

$$\begin{aligned}
\left(\hat{\boldsymbol{\beta}}_R(k)\right) &= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right] = E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\mathbf{y}\right] \\
&= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}\right)\right] \\
&= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\boldsymbol{\varepsilon}\right] \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p - k\mathbf{I}_p\right)\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)\boldsymbol{\beta} - k\mathbf{I}_p\boldsymbol{\beta}\right] + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left[\mathbf{I}_p - k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\right]\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \boldsymbol{\beta} - k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}.
\end{aligned}$$

Equivalently, from (4.1), we obtain

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_R(k)\right) = E\left(\left(\mathbf{I}_p + k(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}_p\right) &= E\left[\left(\mathbf{I}_p + k(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\right)^{-1}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\mathbf{y}\right] \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right).
\end{aligned}$$

Hence, using the abbreviation in equation (4.3), as claimed before, it is obtained $E\left(\hat{\boldsymbol{\beta}}_R(k)\right)$, and Bias $\left(\hat{\boldsymbol{\beta}}_R(k)\right)$ in equations (4.4), (4.5), and (4.6), respectively. Also, we denote the variance property of an estimator $\hat{\boldsymbol{\beta}}_R(k)$ by covariance matrix:

$$\begin{aligned}
\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) &= E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)'\right] \\
&= E\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\,\boldsymbol{\varepsilon} \\
&\quad - \left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right] \\
&= E\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)'\Big) \\
&= E\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\right)\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\right)E(\boldsymbol{\varepsilon}^{\mathbf{2}}) \\
&= \sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}.
\end{aligned}$$

As a result, it can be expressed as the following result with abbreviation

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \sigma^2\mathbf{G}_k\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\tilde{\mathbf{X}}\mathbf{G}_k,$$

as claimed.

A.3. The derivation of the smoother matrix and $E(\text{RSS})$

$$\begin{aligned}
\hat{\mathbf{y}} &= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k) = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R(k)\right) \\
&= \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left[\mathbf{y} - \mathbf{X}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right] \\
&= \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left[\mathbf{y} - \mathbf{X}\mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right] \\
&= \mathbf{X}(\mathbf{I}_p - \mathbf{W}_\lambda)\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y} + \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}(\mathbf{I}_p - \mathbf{W}_\lambda)\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y}\right) \\
&= \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y} + \mathbf{W}_\lambda\mathbf{y} - \mathbf{W}_\lambda\mathbf{H}\mathbf{y} = \mathbf{W}_\lambda\mathbf{y} + (\mathbf{I}_p - \mathbf{W}_\lambda)\,\mathbf{H}\mathbf{y} \\
&= \left[\mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\,\mathbf{H}\right]\mathbf{y} = \mathbf{H}_\lambda\mathbf{y},
\end{aligned}$$

where $\mathbf{H} = \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'$. Accordingly, the smoother matrix based on smoothing parameter λ is

$$\mathbf{H}_\lambda = \mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}',$$

as defined in the equation (4.8).

The expected value of the RSS in equation (4.9) can be given by

$$\begin{aligned}
E(\text{RSS}) &= E\left((\mathbf{y} - \mathbf{H}_\lambda)'(\mathbf{y} - \mathbf{H}_\lambda)\right) \\
&= E\left(\mathbf{y}'(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y}\right) = E\left(\mathbf{y}'(\mathbf{I} - \mathbf{H}_\lambda)^2\mathbf{y}\right) \\
&= \operatorname{tr}\left((\mathbf{I} - \mathbf{H}_\lambda)^2\sigma^2\mathbf{I}\right) + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)^2E(\mathbf{y}) \\
&= n\sigma^2\operatorname{tr}(\mathbf{H}_\lambda^2) - 2\sigma^2\operatorname{tr}(\mathbf{H}_\lambda) + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)^2E(\mathbf{y}) \\
&= \sigma^2\left[n - \operatorname{tr}(2\mathbf{H}_\lambda - \mathbf{H}_\lambda^2)\right] + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)E(\mathbf{y}).
\end{aligned}$$

A.4. Proof of Lemma 4.1

Since the MDE equals $\sum_{j=1}^{k} E\left(\hat{\beta}_{jR}(k) - \beta_j\right)^2$ it is sufficient to prove for a scalar $\hat{\boldsymbol{\beta}}_R(k)$

$$\begin{aligned}
E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right)^2\right] &= \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) + \operatorname{Bias}^2\left(\hat{\boldsymbol{\beta}}_R(k)\right) \\
&= E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right) + \left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)\right]^2 \\
&= E\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)^2 + \left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)^2 \\
&\quad + 2\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)'\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right) \\
&= E\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)^2 + E\left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)^2 \\
&= \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) + \operatorname{Bias}^2\left(\hat{\boldsymbol{\beta}}_R(k)\right).
\end{aligned}$$

This completes the proof of the Lemma 4.1.

REFERENCES

[1] Ahmed, S.E. (2014). *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*, Springer, New York.

[2] Aydin, D. (2014). Estimation of partially linear model with smoothing spline based on different selection methods: a comparative study, *Pakistan Journal of Statistics*, **30**(1), 35–56.

[3] Aydin, D.; Memeddeli, M. and Omay, R.E. (2013). Smoothing parameter selection for nonparametric regression using smoothing spline, *European Journal of Pure and Applied Mathematics*, **6**(2), 222–238.

[4] Aydin, D. and Yilmaz, E. (2018). Modified estimators in semiparametric regression models with right-censored data, *Journal of Statistical Computation and Simulation*, **88**(8), 1470–1498.

[5] Belsley, D.A.; Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

[6] Chen, H. (1988). Convergence rates for parametric components in a partially linear models, *The Annals of Statistics*, **16**(1), 136–146.

[7] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Num. Math.*, **31**(4), 377–403.

[8] Engle, R.; Granger, C.; Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales, *Journal of American Statistical Association*, **81**(394), 310–320.

[9] Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

[10] Foucart, T. (1999). Stability of the inverse correlation matrix, partial ridge regression, *Journal of Statistical Planning And Inferences*, **77**(1), 141–154.

[11] Gibbons, D.G. (1981). A simulation study of some ridge estimators, *Journal of the American Statistical Association*, **76**(373), 131–139.

[12] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Model*, Chapman and Hall, New York.

[13] Hoerl, A.E. and Kennard, R.W. (1970a). Ridge regression: biased estimation for non orthogonal problems, *Technometrics*, **12**(1), 55–67.

[14] Hoerl, A.E. and Kennard, R.W. (1970b). Ridge regression: applications to non orthogonal problems, *Technometrics*, **12**(1), 69–82.

[15] Hu, H. (2005). Ridge estimation of a semiparametric regression model, *Journal of Computational and Applied Mathematics*, **176**(1), 215–222.

[16] Hurvich, C.M.; Simonoff, J.S. and Tasi, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *J. R. Statist. Soc. B.*, **60**(2), 271–293.

[17] Kaya, H.; Tufekci, P. and Gurgen, S.F. (2012). Local and global learning methods for predicting power of a combined gas steam turbine, *Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE*, 13–18.

[18] Kibria, B.M. (2003). Performance of some new ridge regression estimators, *Communications in Statistics – Simulation and Computation*, **32**(2), 419–435.

[19] Lee, Thomas C.M. (2003). Smoothing parameter selection for smoothing splines: a simulation study, *Computational Statistics and Data Analysis*, **42**(1–2), 139–148.

[20] Lee, Thomas C.M. (2004). Improved smoothing spline regression by combining estimates of different smoothness, *Statistics and Probability Letters*, **67**(2), 133–140.

[21] Liang, H. (2006). Estimation partially linear models and numerical comparison, *Computational Statistics and Data Analysis*, **50**(3), 675–687.

[22] Mallows, C. (1973). Some comments on C_p, *Technometrics*, **15**(4), 661–675.

[23] Muniz, G. and Kibria, B.M.G. (2009). On some ridge regression estimators: an empirical comparisons, *Communications in Statistics – Simulation and Computation*, **38**(3), 621–630.

[24] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and Its Applications*, **9**(1), 141–142.

[25] Reiss, P.T. and Ogden, R.T. (2004). Smoothing parameter selection for a class of semiparametric linear models, *J. R. Statist. Soc. B*, **71**(2), 505–523.

[26] Robinson, M.P. (1988). Root-n-consistent semi-parametric regression, *Econometrica*, **56**(4), 931–954.

[27] Roozbeh, M.; Arashi, M. and Niroumanda, H.A. (2010). Semiparametric ridge regression approach in partially linear models, *Communications in Statistics – Simulation and Computation*, **39**(3), 449–460.

[28] Schimek, G. Michael (2000). *Smoothing and Regression: Approaches, Computation, and Application*, John Willey and Sons, USA.

[29] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

[30] Speckman, P. (1988). Kernel smoothing in partially linear model, *J. Royal Statist., Soc. B.*, **50**(3), 413–436.

[31] Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association*, **84**(405), 276–283.

[32] Tufekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power using machine learning methods, *International Journal of Electrical Power and Energy Systems*, **60**, 505–523.

[33] Wahba, G. (1990). *Spline Model for Observational Data*, SIAM, Philadelphia.

[34] Watson, G.S. (1964). Smooth regression analysis, *Sankhya, Series A*, **26**(4), 359–372.

[35] Yuzbasi, B. (2014). *Penalty and Non-Penalty Estimations Strategies for Linear and Partially Linear Models*, PhD Thesis, Inonu University, Malatya.

[36] Yuzbasi, B. and Ahmed, S.E. (2016). Shrinkage and penalized estimation in semi-parametric models with multicollinear data, *Journal of Statistical Computation and Simulation*, **86**(17), 3543–3561.

[37] Yuzbasi, B.; Ahmed, S.E. and Aydin, D. (2017). Ridge-type pretest and shrinkage estimations in partially linear models, *Statistical Papers*, **61**(2), 869–898. Doi: 10.1007/s00362-017-0967-8.