Ridge estimation of a semiparametric regression model

Hongchang Hu

Department of Mathematics, Hubei Normal University, Huangshi 435002, China
College of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

Received 30 July 2003; received in revised form 4 June 2004

Abstract

Considered the semiparametric regression model

$$l_i = A_i^T X + s(t_i) + \Delta_i \quad (i = 1, 2, \ldots, n).$$

Firstly, ridge estimators of both parameters and nonparameters are attained without a restrained design matrix. Secondly, the ridge estimator will be compared with two steps estimation under a mean square error and some conditions in which the former excels the latter are given. Finally, the validity and feasibility of the method are illustrated by a simulating example.

© 2004 Published by Elsevier B.V.

Keywords: Semiparametric regression model; Ridge estimation; Two steps estimation; Mean square error

1. Introduction

Considered the semiparametric regression model

$$l_i = A_i^T X + s(t_i) + \Delta_i \quad (i = 1, 2, \ldots, n),$$

where $s_i = s(t_i)$ denotes the nonparametric signal of the observation and $l_i$ denotes a number related to the observation at $t_i$, $A_i \in R^p(n > p)$, $X = (x_1, \ldots, x_p)^T$ is a parameter vector with $p$ denoting...
the number of parameters or unknowns, and $\Lambda_i$ denotes the noise and is assumed to be independently $N(0, \sigma^2)$-distributed.

In vector notation, the data model is given by

$$L = AX + S + \Lambda,$$

where $L = (l_1, \ldots, l_n)^T$ and $S = (s_1, s_2, \ldots, s_n)^T$ correspond to $t = (t_1, \ldots, t_n)^T$ ($t_i \neq t_j$ for $i \neq j$), design matrix $A = (A_1, \ldots, A_n)^T$, without any restrained conditions, namely, $\text{rank} \ (A) < p$ or $\text{rank} \ (A) = p \ ($ill-conditioned or not$)$.

The model (1) has been used in the discussion of many methods, e.g., penalized least-squares (see [1]), smoothing splines (see [2]), piecewise polynomial (see [6]) and two steps estimation methods (see [5,7,8]). The essential thought of two steps estimation is the following: the first step, $S(t, X)$ is defined with supposition where $X$ is supposed to be known; the second step, the estimator of parametric $X$ is attained by a least-squares method; accordingly, $\hat{S}(t) = S(t, \hat{X})$ is gained. However, they all assume $\text{rank} \ (A) = p$. In fact, if full rank $A$ is an ill-conditioned matrix, then the results may not fulfil our wishes, or can even be false in some situations, especially for small samples. Many papers do not consider the case $\text{rank} \ (A) < p$, and few people investigate the situation that the design matrix $A$ is rank-deficient.

Although there are many results about ridge estimation of linear models (see [3,4,9]), to the best of my knowledge, nothing is known about a semiparametric regression model. It is noticeable that textual ridge estimation not only solves rank-deficient and ill-conditioned problems, but also offers a new method which can deal with (non)linear and semiparametric regression models for $\text{rank}(A) = p$ without ill-conditioning.

2. Ridge estimation method

In the following, one introduces ridge estimation method based on a two steps estimation process. In the first step, we assume that $X$ is known, and the nonparametric estimator of $S$ is defined by

$$S(t, X) = W(t, \lambda)(L - AX),$$

based on $\{l_i - A_i^T X, t_i\} (i = 1, \ldots, n)$, where $\lambda$ is an arbitrary parameter and $W(t, \lambda)$ is an $(n \times n)$ matrix. Depending on the particular choice of $W(t, \lambda)$, the two steps estimation process leads to different methods, such as wavelet estimate (see [8]), near neighbour estimation (see [5]), or kernel estimation (see [7]).

Substituting (2) into (1), we have

$$\tilde{L} = \tilde{A} X + \tilde{\Lambda},$$

where

$$\tilde{A} = (I - W) A, \quad \tilde{\Lambda} = (I - W) L, \quad \tilde{S} = \tilde{S} + (I - W) \Lambda, \quad \tilde{S} = (I - W) S.$$

(4)

Though (3) is a linear model, it is different from the generic one because the error $\tilde{\Lambda}$ is related to $S, t, X$ and $W$.

In the second step, with minimal condition

$$V^T V + \beta \tilde{X}^T \tilde{X} = \min \ (V = \tilde{A} \tilde{X} - \tilde{L}),$$

(5)
we obtain its solution, namely
\[
\hat{X} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{L}.
\] (6)

Combining (6) and (2), we obtain the ridge estimator
\[
\hat{S} = S(t, \hat{X}) = W(t, \hat{X})(L - A\hat{X}).
\] (7)

Since there is a formal resemblance between (6) and the ridge estimator of the linear model, we call it a ridge estimator of the semiparametric model.

In the above method, it is crucial to choose the weight vector \(W\) and balance parameter \(\beta\), so we discuss their choices in the sequel.

In the estimator \(\hat{S}\), we may choose the weight function \(W(t, \hat{X}) = (w_1(t, \hat{X}), \ldots, w_n(t, \hat{X}))\) (see [8]),
\[
w_i(t, \hat{X}) = \int_{I_i} \lambda^{-1} \sum_{k \in \mathbb{Z}} \phi(\lambda^{-1} t - k) \phi(\lambda^{-1} u - k) \, du,
\]
where \(I_i = [u_{i-1}, u_i]\), \(u_0 = 0\), \(u_n = 1\), \(u_i = (t_i + t_{i+1})/2\), \(i = 1, \ldots, n - 1\), where \(\phi\) is a scaling function and \(\lambda\) is called the bandwidth. We choose \(\phi(x) = I(0 \leq x < 1)\) and \(\lambda = 2^{-m}\), where \(m\) depends on \(n\). The optimal bandwidth is selected to minimize the average squared error (ASE), that is \(\text{ASE}(\hat{S}) = \frac{1}{n} \sum_{j=1}^{n} (\hat{s}(t_j) - s(t_j))^2\), where \([t_j, j = 1, \ldots, n]\) are grid points.

To choose the balance parameter \(\beta\), we consider the case \(\text{rank}(\tilde{A}) < p\). First, we choose some \(\{\beta_h, h = 1, \ldots, k\}\) such that \(\text{rank}(\tilde{A}^T \tilde{A} + \beta_h I) = p\). Second, we find a minimal \(\beta_{k_0}\) from \(\{\beta_h, h = 1, \ldots, k\}\), such that \(\beta_{k_0}\) satisfies
\[
(\tilde{A}^T \tilde{A} - \tilde{L})^T (\tilde{A}^T \tilde{L} - \tilde{L}) + \beta_{k_0} \tilde{X}^T \hat{X} = \min.
\]

**Remark 1.** Let \(s(t) \equiv 0\). Then Eq. (6) becomes ridge estimator of a linear model. However, there exists an essential difference: general ridge estimation of a linear model usually solves rank-deficiency and ill-conditioning. However, in this paper we do not restrict ourselves to a design matrix.

**Remark 2.** Let \(\beta = 0\), and assume that \(\tilde{A}^T \tilde{A}\) is of full rank. Then ridge estimation becomes two steps estimation of semiparametric model. And if \(\tilde{A}^T \tilde{A}\) is a rank-deficient matrix, then the two steps method will fail. Thus, it can be seen that ridge estimation excels in two steps estimation.

### 3. Comparison of ridge estimation with two steps estimation

From the above-mentioned data, we know that ridge estimation of semiparametric model has avoided the shortage of the two steps estimation. Moreover, we will show that (using a mean square error) the former is superior to the latter.

If \(\text{rank}(A) < p\), then \(\text{rank}(\tilde{A}) < p\), so it is well known that
\[
\text{MSE}(\hat{X}) < \text{MSE}(\hat{X}_{TS}) = +\infty \quad (\text{for } \forall \beta > 0),
\]
where \(\hat{X}_{TS}\) denotes the estimator by the two steps method. Hence, in the sequel, we assume \(\text{rank}(A) = p\).
Theorem 1. Let \( \text{rank}(A) = p \), and let there exist a matrix \( W \) such that \( \text{rank}(\tilde{A}) = p \).

(1) If \( \lambda^2_i \varepsilon^2_i > \sigma^2 a_{ii} + \eta^2_i \) and \( b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i > 0 \), then

\[
\text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \quad \text{for} \quad 0 < \beta \leq (b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i)/(\varepsilon_i^2 \lambda^2_i - \eta^2_i - \sigma^2 a_{ii}).
\]

(2) If \( \lambda^2_i \varepsilon^2_i < \sigma^2 a_{ii} + \eta^2_i \) and \( b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i > 0 \), then

\[
\text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \text{ for arbitrary } \beta > 0.
\]

(3) If \( \lambda^2_i \varepsilon^2_i < \sigma^2 a_{ii} + \eta^2_i \) and \( b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i < 0 \), then

\[
\text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \quad \text{for} \quad \beta \geq (b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i)/(\varepsilon_i^2 \lambda^2_i - \eta^2_i - \sigma^2 a_{ii}).
\]

(4) If \( \lambda^2_i \varepsilon^2_i \geq \sigma^2 a_{ii} + \eta^2_i \) and \( b_{ii} \lambda^2_i + \sigma^2 \lambda_i a_{ii} + \eta^2_i \lambda_i < 0 \), then

\[
\text{MSE}(\hat{X}) > \text{MSE}(\hat{X}_{TS}) \text{ for arbitrary } \beta > 0.
\]

All denotations will be defined in the course of the proof.

Proof. Considering (3), (4) and (6), we obtain

\[
E \hat{X} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T E \tilde{L} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T (\tilde{A} X + \tilde{S}),
\]

(8)

\[
D \hat{X} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T (D \tilde{L})(\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T
\]

\[
= \sigma^2 (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T (I - W)(I - W)^T ((\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T)^T.
\]

(9)

Using the definition of mean square error and (8), (9), we obtain

\[
\text{MSE}(\hat{X}) = E\|\hat{X} - X\|^2 = E(\hat{X} - X)^T (\hat{X} - X) = \text{tr}(D \hat{X}) + \text{tr}(E \hat{X} - X)(E \hat{X} - X)^T
\]

\[
= \text{tr}(D \hat{X}) + \text{tr}\left\{(\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T (I - W)(I - W)^T ((\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T - I)^T\right\}
\]

\[
+ 2\text{tr}\left\{(\tilde{A}^T \tilde{A} + 2\beta I)^{-1} (\tilde{A}^T \tilde{A} - I) X (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{S})^T\right\}
\]

\[
+ \text{tr}\left\{(\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{S})(\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{S})^T\right\}
\]

\[
=: U_1 + U_2 + 2U_3 + U_4.
\]

(10)

Because \( \text{rank}(\tilde{A}) = p \), \( \tilde{A}^T \tilde{A} \) is a positive-definite matrix. So there exists an orthogonal matrix \( G \) in the sense that

\[
\tilde{A}^T \tilde{A} = G \text{diag}(\lambda_1, \ldots, \lambda_p) G^T =: GAG^T \quad (\lambda_i > 0).
\]

(11)

From (9)–(11) we have

\[
U_1 = \sigma^2 \text{tr}(G^{-T}(A + 2\beta I)^{-1} G^{-1}. G^{-T}(A + 2\beta I)^{-T} G^{-1}. \tilde{A}^T (I - W)(I - W)^T \tilde{A})
\]

\[
= \sigma^2 \sum_{i=1}^{p} \frac{a_{ii}}{(2\beta + \lambda_i)^2},
\]

(12)

where \( A_w = G^T \tilde{A}^T (I - W)(I - W)^T \tilde{A} G = (a_{ij})_{p \times p} \).
From (10) and (11) we obtain
\[ U_2 = \text{tr}\{G((A + 2\beta I)^{-1}A - I)G^TXX^TG((A + 2\beta I)^{-1}A - I)^T G^T}\]  
\[ = \text{tr}\{((A + 2\beta I)^{-1}A - I)^2G^TXX^TG\} = \sum_{i=1}^{p} \frac{4\beta^2\tilde{\epsilon}_i^2}{(2\beta + \lambda_i)^2}, \]  \tag{13}

\[ U_3 = \text{tr}\{G((A + 2\beta I)^{-1}A - I)G^T\tilde{X}^T\tilde{A}G^{-1}(A + 2\beta I)^{-T}G^{-1}\} \]
\[ = \text{tr}\{((A + 2\beta I)^{-1}A - I)(A + 2\beta I)^{-T}G^T\tilde{X}^T\tilde{A}G\} = -\sum_{i=1}^{p} \frac{2\beta b_{ii}}{(2\beta + \lambda_i)^2}, \]  \tag{14}

\[ U_4 = \text{tr}\{(G^{-T}(A + 2\beta I)^{-1}G^{-1}\tilde{A}^T\tilde{S})(G^{-T}(A + 2\beta I)^{-1}G^{-1}\tilde{A}^T\tilde{S})^T\} \]
\[ = \text{tr}\{(A + 2\beta I)^{-2}G^T\tilde{A}^T\tilde{S}\tilde{S}^T\tilde{A}G\} = \sum_{i=1}^{p} \frac{\eta_i^2}{(2\beta + \lambda_i)^2}, \]  \tag{15}

where \( \xi = (\xi_1, \ldots, \xi_p)^T = G^T X, B = G^T \tilde{X}^T \tilde{A}G = (b_{ij})_{p \times p}, \eta = (\eta_1, \ldots, \eta_p) = \tilde{S}^T \tilde{A}G. \)

Hence, using (10) and (12)–(15) we have
\[ \text{MSE} \left( \hat{X} \right) = \sigma^2 \sum_{i=1}^{p} \frac{a_{ii}}{(2\beta + \lambda_i)^2} + \sum_{i=1}^{p} \frac{4\beta^2\tilde{\epsilon}_i^2}{(2\beta + \lambda_i)^2} - \sum_{i=1}^{p} \frac{2\beta b_{ii}}{(2\beta + \lambda_i)^2} + \sum_{i=1}^{p} \frac{\eta_i^2}{(2\beta + \lambda_i)^2}. \]  \tag{16}

Setting \( \beta = 0 \) in (16) we have
\[ \text{MSE} \left( \hat{X}_{TS} \right) = \sigma^2 \sum_{i=1}^{p} \frac{a_{ii}}{\lambda_i^2} + \sum_{i=1}^{p} \frac{\eta_i^2}{\lambda_i^2}. \]  \tag{17}

Moreover, by (16) and (17) we have
\[ \text{MSE} \left( \hat{X} \right) - \text{MSE} \left( \hat{X}_{TS} \right) = 4 \sum_{i=1}^{p} \beta(\lambda_i(\lambda_i + 2\beta))^{-2}(\tilde{\epsilon}_i^2 \lambda_i^2 - \eta_i^2 - \sigma^2 a_{ii}) \beta \]
\[ - (b_{ii} \lambda_i^2 - \eta_i^2 \lambda_i + \sigma^2 a_{ii} \lambda_i)). \]

Using this formula and the supposed conditions the theorem is immediately proved. \( \square \)

**Remark 3.** The theorem not only shows that the ridge estimation excels in the two steps estimation under some conditions, but it is also a theoretic warning to adopt the ridge estimation or two steps estimation in case \( \tilde{A}^T \tilde{A} \) is of full rank.

**Corollary 1.** Assume \( \text{rank}(A) = p \), and there exists a matrix \( W \) such that \( \text{rank}(I - W) = n \). Then \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \) for \( 0 < \beta \leq \min_{1 \leq i \leq p} \{b_{ii}\tilde{\epsilon}_i^{-2}\} \).

**Proof.** Since \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) + \sum_{i=1}^{p} \left( \frac{4\beta}{(2\beta + \lambda_i)^2}(\tilde{\epsilon}_i^2 \beta - b_{ii}) \right) \), the result easily follows. \( \square \)
Corollary 2. Let \( s(t) \equiv 0 \) and \( \text{rank}(A) = p \).

1. If \( \xi_i^2 > \sigma^2 \), then \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{LS}) \) for \( 0 < \beta < \frac{\lambda_i \sigma^2}{4(\xi_i^2 - \sigma^2)} \) and,
2. If \( \xi_i^2 \leq \sigma^2 \), then \( \text{MSE}(\hat{X}) > \text{MSE}(\hat{X}_{LS}) \) for \( \forall \beta > 0 \),

where the symbol \( \xi_i \) is the same as in Theorem 1 and \( \lambda_i > 0 \) is defined in the corresponding proof. \( \hat{X}_{LS} \) denotes the estimation of \( X \) in a linear model by a least-squares method.

Proof. Since \( \text{rank}(A) = p \), \( A^T A \) is a positive-definite matrix. So there exists an orthogonal matrix \( G \) such that

\[
A^T A = G \text{diag}(\lambda_1, \ldots, \lambda_p) G^T =: \tilde{G} \tilde{G}^T.
\]

From the proof of Theorem 1, we immediately obtain

\[
\text{MSE}(\hat{X}) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(2\beta + \lambda_i)^2} + \sum_{i=1}^{p} \frac{4\beta^2 \xi_i^2}{(2\beta + \lambda_i)^2},
\]

(18)

\[
\text{MSE}(\hat{X}_{LS}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}.
\]

(19)

Using (18) and (19), we get

\[
\text{MSE}(\hat{X}) - \text{MSE}(\hat{X}_{LS}) = 4\beta \sum_{i=1}^{p} \frac{\lambda_i^{-1} (2\beta + \lambda_i)^{-2} ((\xi_i^2 - \sigma^2) \beta - \sigma^2 \lambda_i)}{\lambda_i \xi_i^2 - \sigma^2}.
\]

By this formula and supposed conditions, the corollary is immediately proved. \( \square \)

For convenience of application, a simple result is given in the following theorem:

Theorem 2. Let \( \text{rank}(A) = p \) and \( \text{rank}(\tilde{A}) = p \), \( S - S(t, X) = 0 \).

1. If \( \lambda_i \xi_i^2 > \sigma^2 \), then \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \) for \( 0 < \beta < \frac{\lambda_i \sigma^2}{\lambda_i \xi_i^2 - \sigma^2} \) and
2. If \( \lambda_i \xi_i^2 \leq \sigma^2 \), then \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) \) for arbitrary \( \beta > 0 \).

Proof. Since \( S - S(t, X) = 0 \), we have

\[
E \hat{X} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{L} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T \tilde{A} X,
\]

(20)

\[
D \hat{X} = (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T (D \tilde{L}) ((\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T)^T
= \sigma^2 (\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T ((\tilde{A}^T \tilde{A} + 2\beta I)^{-1} \tilde{A}^T)^T.
\]

(21)
From (20) and (21), we have
\[
\text{MSE}(\hat{X}) = E\|\hat{X} - X\|^2 = \text{tr}(D\hat{X}) + \text{tr}(E\hat{X} - X)(E\hat{X} - X)^T \\
= \sigma^2 \sum_{i=1}^{p} \frac{\hat{\lambda}_i}{(2\beta + \hat{\lambda}_i)^2} + \sum_{i=1}^{p} \frac{4\beta^2 \hat{\xi}_i^2}{(2\beta + \hat{\lambda}_i)^2},
\]
(22)
\[
\text{MSE}(\hat{X}_{LS}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\hat{\lambda}_i}.
\]
(23)

Using (22), (23) and supposed conditions, the proof is immediate. □

4. Numerical example

We will simulate a simple semiparametric model \( L = AX + S + \Delta \), where \( A = (a_{ij})_{100 \times 1}, a_{i1} = (\frac{i}{100})^2 (i = 1, 2, \ldots, 100), X = 3, S = (s(t_1), \ldots, s(t_{100}))^T, s(t_i) = 10t_i, t_i = \frac{i}{101} \), and the random errors \( \Delta \) are composed of 100 data, which are \( N(0, 1) \)-distributed.

Choose \( W = \text{diag}(0, 0.5, \ldots, 0.5)_1 + \text{diag}(0.5, \ldots, 0.5)_{-1} + \text{diag}(0.99, 0 \ldots, 0, 0.99) \), where \( \text{diag}(\cdot)_1 \) denotes upper sub-diagonal elements of matrix \( W \), \( \text{diag}(\cdot)_{-1} \) denotes lower sub-diagonal elements of the matrix \( W \). By calculation, we have
\[
\lambda_i^2 \hat{\eta}_i^2 - \sigma^2 a_{ii} - \eta_i^2 = 0.0352 \quad \text{and} \quad b_{ii} \lambda_i^2 + \sigma^2 \lambda_i a_{ii} + \lambda_i \eta_i^2 = 3.3563e - 0.04.
\]

Therefore, using Theorem 1 (1), we get \( \text{MSE}(\hat{X}) \leq \text{MSE}(\hat{X}_{TS}) (0 < \beta \leq 0.0095) \). In fact, setting \( \beta = 0.0045 \), we can validate the following relation
\[
\text{MSE}(\hat{X}) = 0.1448 < 0.1585 = \text{MSE}(\hat{X}_{TS}).
\]

From (6) we can also obtain \( \hat{X} = 3.0203 \). It closely approximates the true value of \( X \). Moreover, by the two steps estimation, we have \( \hat{X}_{TS} = 3.4510 \). Its error is quite large and in fact a wrong estimation.

The example not only validates Theorem 1 but also shows that the ridge estimation is superior to the two steps estimation. It can be seen that our method is successful. However, a further discussion of the choices of the weight function and balance parameter is needed so that we can find a good method to use in practical applications.

Acknowledgements

The author thanks the anonymous referees and the Editor, B. Sommeijer, for their very valuable discussions and suggestions, which led to a great improvement of the paper. The referees’ remarks about the choices of \( W \) and \( \beta \) have been incorporated.

References