Journal of Quantitative Economics, Vol. 11 Nos.1&2 (Combined), January-July 2013

A CLASS OF SEMIPARAMETRIC ORDINARY RIDGE ESTIMATORS OF REGRESSION COEFFICIENTS

HUANSHA WANG[1]

Abstract

In this article, a class of easy-to-implement semiparametric ordinary ridge estimators of regression coefficients is introduced. Their properties are investigated and simulation results are provided to investigate their behaviors when the error variances are small and relatively large, respectively. The semiparametric estimators outperform the Hoerl, Kennard and Baldwin (1975) estimator in the sense that they give less risk (total mean squared error). An empirical application is also presented.

Keywords: Semiparametric, Ridge, Kernel Density Estimation.

JEL Classification: C13, C14, C51.

1 Introduction

In the ordinary least squares (OLS) estimation, if the prediction vectors are not orthogonal, there is a high probability that the OLS estimators may be unsatisfactory. In particular, the estimated coefficients tend to be abnormally large in absolute value and sometimes have the wrong sign. To fix this problem, the ridge estimation was introduced by Hoerl (1962) and Hoerl and Kennard (1970). By adding an extra increment to the original X′X, the ridge estimation circumvents the non-orthogonality problem. Compared to the methods of principal components, computation of 2^p regressions, some subset of all regressions using fractional factorials, a branch and bound technique, ridge estimation "gives an insight into the structure of the factor space and the sensitivity of the results to the particular set of data at hand" (Hoerl and Kennard 1970). In addition, most multiple regression models suffer multicollinearity to some degree. Thus, under these circumstances, introducing the ridge regression could gain, in the sense that the ridge estimators will outperform the OLS estimator and alleviate the multicollinearity problem.

As one can expect, the choice of the increment to X′X is significant in the implementation of the ridge estimation. In the ordinary ridge estimation, we usually use kI, where the I being the identity matrix, to denote it. Many different choices of the biasing parameter k have been proposed in the literature, such as Hoerl, Kennard and Baldwin (1975), see Vinod and Ullah (1981).

[1] Address for Correspondence: Department of Economics, University of California, Riverside, California, USA; Email: huansha.wang@email.ucr.edu.

The author gratefully thanks the valuable corrections and constructive suggestions on the subject matter made by the referee.

In this paper, a class of semiparametric ordinary ridge estimators is proposed. Starting from kernel density estimator of the regressors, these estimators bear more information than the OLS estimator and both simulation and empirical application results in the following content show the usefulness of these easy-to-implement estimators. The properties of these estimators are also investigated. The rest of this paper is arranged as follows. Section 2 introduces the class of ordinary semiparametric (OSP) estimators. Section 3 develops the approximate and exact unbiased mean squared errors (MSE) of the OSP estimators, and proposed the choices of window-width by minimizing them. Section 4 provides some simulation results comparing this class of OSP estimators with the OLS estimator and the estimator proposed by Hoerl, Kennard and Baldwin (1975). Section 5 gives the results for one empirical application. The last section concludes.

2. Semiparametric Estimator of Regression Coefficients

Consider a population multiple regression model

$$y = x_1\beta_1 + \cdots + x_q\beta_q + u \qquad \text{... (1)}$$
$$= x'\beta + u$$

where y is a scalar dependent variable, $\left[x = x_1, \cdots, x_q\right]'$ is a vector of q regressors, β is an unknown vector of regression coefficients, and u is a disturbance with Eu = 0 and V(u)= σ^2.

If we minimize $Eu^2 = E\left(y - x'\beta\right)^2$ with respect to β, we obtain

$$\beta = \left[Exx'\right]^{-1} Exy \qquad \text{... (2)}$$

Where Exx' is a $q \times q$ moment matrix of q variables with the j-th diagonal element and j, j'-th off diagonal elements, respectively, given by

$$Ex_j^2 = \int_{x_j} x_j^2 f(x_j)dx_j, \quad j = 1, \cdots, q, \qquad \text{... (3)}$$

$$Ex_j x_{j'} = \int_{x_j}\int_{x_{j'}} x_j x_{j'} f(x_j, x_{j'})dx_j dx_{j'}, \quad j \neq j' = 1, \cdots, q.$$

Suppose we have the sample observations $\left\{y_i, x_{i1}, \cdots, x_{iq}\right\}$, $i = 1, \cdots, n$. Then the population averages in (3) can be estimated by their sample averages as

$$\hat{E}x_j^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2,\ \hat{E}x_j x_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'}\ . \qquad \text{... (4)}$$

The result $\hat{E}x_j y = \sum_{i=1}^{n} x_{ij}y_i / n$ follows similarly.

Therefore we get[2]

---

[2] With the intercept, the expression for the $\hat{\beta}_{q-1}$ for the (q-1) estimators could be defined as $(X'_{q-1}X_{q-1} + nh^2\mu_2 I - \bar{X}_{q-1}\bar{X}'_{q-1})^{-1}(X'_{q-1}Y - \bar{X}_{q-1}\bar{Y}')$ where $\bar{X}_{q-1} = (\bar{X}_2, \cdots, \bar{X}_q)$; and $\hat{\beta}_1 = \bar{Y} - \bar{X}_{q-1}\hat{\beta}_{q-1}$.

$$\hat{\beta} = (\hat{E}xx')^{-1}\hat{E}xy \quad \text{... (5)}$$
$$= (X'X)^{-1}X'Y$$

where X is an $n \times q$ matrix of observations on q variables, Y is an $n \times 1$ vector of n observations and $\hat{\beta}$ is the well known ordinary least squares (OLS) estimator.

Now we consider the estimation of Ex_j^2 and $Ex_j x_{j'}$ by using a smooth nonparametric kernel density estimation instead of empirical distribution function. In this case,

$$\tilde{E}x_j^2 = \int_{x_j} x_j^2 \tilde{f}(x_j)dx_j \quad \text{... (6)}$$
$$= \frac{1}{nh}\sum_{i=1}^{n}\int_{x_j} x_j^2 k(\frac{x_{ij}-x_j}{h})dx_j$$
$$= \frac{1}{n}\sum_{i=1}^{n}\int_{\psi_{ij}} (x_{ij}^2 + h^2\psi_{ij}^2 - 2x_{ij}h\psi_{ij})k(\psi_{ij})d\psi_{ij}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + h^2\mu_2$$

where $\tilde{f}(x_j) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x_{ij}-x_j}{h})$ is a kernel density estimator, $\psi_{ij} = \frac{x_{ij}-x_j}{h}$ is a transformed variable, $\mu_2 = \int v^2 k(v)dv > 0$ is the second moment of kernel function, $k(\psi_{ij})$ is a symmetric second order kernel, and h is window-width. For implementation, kernel is chosen as normal or Epanechnikov quadratic function, see Pagan and Ullah (1999).

Similarly, it can easily be shown that

$$\tilde{E}(x_j x_{j'}) = \int_{x_j}\int_{x_{j'}} x_j x_{j'} \tilde{f}(x_j, x_{j'})dx_j dx_{j'} \quad \text{... (7)}$$
$$= \frac{1}{nh^2}\sum_{i=1}^{n}\int_{x_j}\int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h})k(\frac{x_{ij'}-x_{j'}}{h})dx_j dx_{j'}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\int_{\psi_{ij}}\int_{\psi_{ij'}} (x_{ij}-h\psi_{ij})(x_{ij'}-h\psi_{ij'})k(\psi_{ij})k(\psi_{ij'})d\psi_{ij}d\psi_{ij'}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'},$$

and

$$\tilde{E}(x_j y) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}y_i \quad \text{... (8)}$$

Where we have used kernels without any loss of generality and $\psi_{ij'} = \frac{x_{ij'} - x_{j'}}{h}$. Also, $\tilde{E}(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j$.

Thus, using (6) to (8) in (2), a new semiparametric estimator of β can be introduced as

$$\begin{aligned} \tilde{\beta}(h) &= (\tilde{E}xx')^{-1}\tilde{E}xy \qquad \text{... (9)} \\ &= (X'X + nh^2\mu_2 I)^{-1}X'Y \\ &= (X'X + D)^{-1}X'Y \end{aligned}$$

where $D = nh^2\mu_2 I$ is a diagonal matrix. We refer this estimator as the ordinary semiparametric (OSP) estimator.

We note that both OLS and OSP estimators are obtained by first considering the population regression (1), in which the regression coefficient vector depends on the population moments of vector x and scalar variable y, and then estimating these moments by two different methods with the help of sample data. These are the estimators of the regression coefficients in the sample linear regression model

$$Y = X\beta + U \qquad \text{... (10)}$$

where the sample is drawn from the population linear regression model (1), and U is an $n \times 1$ vector or fandom errors with $EU = 0$ and $EUU' = \sigma^2 I_n$.

The class of ordinary ridge estimator due to Hoerl and Kennard (1970a) is defined as

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y \qquad \text{... (11)}$$

where k is an unknown parameter. An operational ordinary ridge estimator, from Hoerl, Kennard and Baldwin (1975) is defined with $k = qs^2/\hat{\beta}'\hat{\beta}$ and $s^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-q}$. We refer this estimator as the HKB estimator in the following content.

3. Unbiased Estimation of MSE and Optimal Window-Width Choice

In this section, the choices of window-width h are considered. These are based on the minimization of the approximate total MSE, risk, and the unbiased estimator of the exact MSE of $\tilde{\beta}(h) = E[\tilde{\beta}(h) - \beta)'(\tilde{\beta}(h) - \beta)]$. Also we determine the choice of h based o the minimization of the total MSE of the predictor of y, which is $MSE(\tilde{\mu}(h)) = E[\tilde{\beta}(h) - \beta)'X'X(\tilde{\beta}(h) - \beta)]$, where $\tilde{\mu}(h) = X\tilde{\beta}(h)$.

3.1 An Approximate Estimator of the MSE of $\tilde{\beta}(h)$

Theorem 1. *Under the conditions A1-A6 in the appendix, with* $A \equiv n\mu_2(X'X)^{-1}$, *approximate MSE (AMSE) of* $\tilde{\beta}(h)$ is

$$\text{AMSE}(\tilde{\beta}(h)) = \sigma^2 tr(X'X)^{-1} - 2h^2\sigma^2 \frac{trA^2}{n\mu_2} + h^4[2\beta'A^2\beta + 3\sigma^2 \frac{trA^3}{n\mu_2}].$$

Proof: See the Appendix.

Remark 1: By minimizing the AMSE with respect to h^2, the first order condition gives the optimal choice of h^2 as

$$h^2 = \frac{\sigma^2 trA^2}{2n\mu_2\beta'A^2\beta + 3\sigma^2 trA^3}. \quad \text{... (12)}$$

From the AMSE, we could observe that $\text{AMSE} - \text{MSE}(\hat{\beta}) = -2h^2\sigma^2 \frac{trA^2}{n\mu_2} + h^4(2\beta'A^2\beta + 3\sigma^2 trA^3)$, thus, as long as $0 \le h^2 \le \frac{2\sigma^2 trA^2}{2n\mu_2\beta'A^2\beta + 3\sigma^2 trA^3}$, $\tilde{\beta}(h)$ will outperform the OLS estimator in the sense that it generates smaller mean squared error.

Also, we consider an approximation of h^2 in (12) as

$$h_1^2 = \frac{\sigma^2 trA^2}{2n\mu_2\beta'A^2\beta}, \quad \text{... (13)}$$

and thus, this $\tilde{\beta}(h_1)$ is referred as AOSP1.

3.2 An Exact Unbiased Estimator of the MSE of $\tilde{\beta}(h)$

Theorem 2. *Under the same conditions as in Theorem 1, the exact unbiased estimator of MSE of* $\tilde{\beta}(h)$ *is given by*

$$\widehat{\text{MSE}}(\tilde{\beta}(h)) = s^2 tr(XD^2X') + (nh^2\mu_2)^2[\hat{\beta}'D^2\hat{\beta} - s^2 trD^2(X'X)^{-1}]$$

where $D = (X'X + n\mu_2 h^2 I)^{-1}$.

Proof: See the Appendix.

3.3 An Exact Unbiased Estimator of the MSE of $\tilde{\mu}(h)$

Theorem 3. *Under the same conditions as in Theorem 1, the exact unbiased estimator of MSE of* $\tilde{\mu}(h)$ *is given by*

$$\widehat{\text{MSE}}(\tilde{\mu}(h)) = s^2 tr(XD'X')^2 + (nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2 trD'X'XD(X'X)^{-1}]$$

where $D = (X'X + n\mu_2 h^2 I)^{-1}$.

Proof: See the Appendix.

Remark 2: The expression for the exact unbiased estimator of the MSE of $\tilde{\beta}(h)$ and $\tilde{\mu}(h)$ are nonlinear; thus in implementation, there's no closed form solution for the optimal window-width h; we will use the constraint optimization function built in R *version 2.13.1*. to approximate

the optimal h. We refer these two estimators of $\tilde{\beta}(h)$ from the two $h's$ as the exact ordinary semiparametric (EOSP) estimator and an alternative exact ordinary semiparametric (EOSP1) estimators, respectively.

Another asymptotic optimal semiparametric under Mallows criterion is also considered in the simulation (see Hansen (2007) for reference on the Mallows criterion), where the optimal h^2 is obtained through minimize $(Y - X\tilde{\beta}(h))'(Y - X\tilde{\beta}(h)) + 2s^2 tr[(X'X + nh^2\mu_2 I)^{-1}X']$. We refer this estimator as the Mallows ordinary semiparametric (MOSP) estimator.

4. Simulation

Our Monte Carlo experiments are based on two DGP's.

DGP1: $y_i = \sum_{j=1}^{q} \theta_j x_{ij} + e_i,$, x_{ij} are *iid* N(0,1). The errors e_i are uncorrelated with $x's$ so it is set to be *iid* N(0,1) and N(0,25) respectively. This model is from Hansen (2007) with the values of θ_j considered here as $0.7071j^{-3/2}$. Further, sample sizes taken are n=50, q=11 and n=150, q=16.

DGP2: in DGP2, parameters and function are set as in DGP 1, but x_{2i} is set to be the sum of x_{3i} to x_{50i} plus an error which follows N(0,1) so that this DGP incorporates near-perfect collinearity.

Under both DGP 1 and DGP 2, 1000 simulations are done. Further, normal kernel is selected, $K(\varphi) = (2\pi)^{-1/2} \exp[-\frac{1}{2}\varphi^2]$, and thus $\mu_2 = 1$. Performance of the estimators is evaluated in terms of total MSE (risk), $E(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$, where $\hat{\beta}$ is an estimator. Simulation results are reported in Table 1.

Table 1: Risk for Each Estimator

DGP	*Estimators*	*σ = 1*		*σ = 5*	
		n = 50	*n = 150*	*n = 50*	*n = 150*
1	OLS	0.3098	0.1181	7.0873	3.0761
	HKB	0.2097	0.1019	2.1995	1.1095
	AOSP	0.2604	0.1079	5.4799	2.3175
	AOSP1	0.2178	1.1030	2.8807	1.2573
	EOSP	0.2221	0.1041	1.2837	0.7022
	EOSP1	0.2126	0.1035	1.0036	0.6585
	MOSP	0.2126	0.1035	1.0037	0.6585
2	OLS	0.3069	0.1248	6.9182	3.0682
	HKB	0.2037	0.1107	2.0970	1.0557
	AOSP	0.2574	0.1155	5.3411	2.3276
	AOSP1	0.2120	0.1140	2.6461	1.2229
	EOSP	0.2145	0.1136	1.1760	0.6735
	EOSP1	0.2044	0.1140	0.9950	0.6291
	MOSP	0.2044	0.1140	0.9950	0.6291

From Table 1, we could observe that under different DGP settings, all classes of ordinary ridge estimators beat the OLS estimator. Compare column 4 with column 6 for example, as error variance increases, more gains are obtained through the usage of the semiparametric ordinary ridge estimators, especially MOSP and EOSP/EOSP1 estimators since much smaller risks are obtained. And another interesting, yet not surprising phenomenon we observe, is that the MOSP and EOSP1 generate very similar risks under different DGP's. This is due to the fact that both criteria are unbiased estimators of the MSE of $\tilde{\mu}(h)$.

5. An Empirical Application

5.1 Forecasting Excess Stock Returns

The data is the same as in Campbell and Thompson (2008). In the monthly data from January 1950 to December 2005 the total sample size is equal to 672. The dependent variable Y is the excess stock returns, which is defined as the difference between the monthly stock returns and the risk-free rate. We consider 12 regressors, default yield spread, treasury bill rate, new equity expansion, term spread, dividend price ratio, earnings price ratio, long term yield, book-to-market ratio, inflation, return on equity, lagged dependent variable, smoothed earnings price ratio and the sub-models are nested. The independent variables are ordered in according to their correlation with the dependent variable. Thus, 12 candidate sub-models are generated with regressors $\{x_1\},\{x_1,x_2\},\cdots,\{x_1,x_2,\cdots,x_{12}\}$, respectively.

The in-sample estimation periods T1 are set to be 144, 180, 216, and 336 respectively. We define the out-of-sample R^2 as

$$R^2 = 1 - \frac{\sum_{t=T1}^{T-1}(Y_{t+1} - \hat{Y}_{t+1})^2}{\sum_{t=T1}^{T-1}(Y_{t+1} - \bar{Y}_{t+1})^2}$$

where $\hat{Y}_{t+1}$, $\bar{Y}_{t+1}$ are the one-period-ahead prediction and historical average, respectively, using the sample of size T1.

5.2 Forecasting Results

The out-of-sample R^2 are reported in Table 2.

From Campbell and Thompson (2008), we know that when no restriction is put on the sign coefficients and return forecasts, the OLS estimator will generate forecasts that cannot beat the historical average. But considering the positivity restriction they tend to show that the restricted OLS estimators beat the historical average. However, with the ordinary ridge estimators and no constraints imposed on the signs, we could still generate forecasts beating the historical average for most of the cases. This may be due to the fact that the ordinary ridge estimators are restricted to be bounded, which makes them stable. The reason to use the ridge estimator is that, compared to Campbell and Thompson (2008), where only one independent variable is considered in each structure to forecast the equity premium, in this application, more than one explanatory variables are included, multicollinearity, if not perfect, exists and needs to be taken care of.

Table 2. Out-of-Sample R^2

Estimator	*T 1 = 144*	*T 1 = 180*	*T 1 = 200*	*T 1 = 216*
OLS	-0.0625	-0.0123	-0.0340	-0.0479
HKB	0.0021	0.0532	0.0351	0.0190
AOSP1	-0.0475	0.0051	-0.0153	-0.0322
EOSP1	0.0485	0.0664	0.0071	-0.0325
MOSP	0.0485	0.0664	0.0071	-0.0325

6. Concluding Remarks

In this article, a class of semiparametric ordinary ridge estimators is proposed. Through the kernel density estimation, we are able to derive the estimator and obtain the estimator of regression coefficients in the ridge form. The properties of the estimators have also been investigated. Easy to implement, this class of estimators outperforms both the OLS estimator and the ordinary ridge estimator proposed by Hoerl, Kennard and Baldwin (1975) in both simulation and empirical applications. Also, one of the semiparametric estimator proposed, the EOSP1 estimator, generates almost the same result as the Mallows ordinary ridge estimator, due to the fact that both estimators are obtained through the minimization of unbiased estimators of MSE of the predictor of y, $\mu(h)$. This is an interesting result and we expect to see more applications of our estimator in the future research.

Appendix

Let $f = f(x)$ denote the continuous density function of a random variable X at point x, and $x_1, x_2, \cdots, x_n$ be the observations from f. As in section 2, kernel density estimator $\tilde{f}(x_j) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x_{ij} - x_j}{h})$, where $k(\cdot)$ is the kernel function. In the population $Y = X\beta + U$, Y is a scalar dependent variable, $X = [X_1, \cdots, x_q]'$ is a vector of q regressors, β is an unknown vector of regression coefficients, U is an $n \times 1$ vector of random errors. We make the following assumptions following Pagan and Ullah (1999).

A1. The observations $x_1, x_2, \cdots, x_n$ are independent and identically distributed (i.i.d.).

A2. The kernel $k(\cdot)$ is a symmetric function around zero satisfying

(i) $\int k(v)dv = 1$,

(ii) $\int v^2 k(v)dv = \mu_2 \neq 0$,

(iii) $\int k^2(v)dv < \infty$.

A3. The second order derivatives of f are continuous and bounded in some neighborhood of x.

A4. $h = h_n \to 0$ as $n \to \infty$.

A5. $nh_n \to \infty$ as $n \to \infty$.

A6. $EU = 0$ and $EUU' = \sigma^2 I_n$.

Proof of Theorem 1.

Under the above assumptions, since

$$\begin{aligned}\tilde{\beta}(h) &= (X'X + nh^2\mu_2 I)^{-1}X'Y \\ &= (X'X + nh^2\mu_2 I)^{-1}X'(X\beta + U) \\ &= (X'X + nh^2\mu_2 I)^{-1}((X'X + nh^2\mu_2 I - nh^2\mu^2 I)\beta + U),\end{aligned}$$

we have

$$\begin{aligned}\tilde{\beta}(h) - \beta &= (X'X + nh^2\mu_2 I)^{-1}(X'U - nh^2\mu_2\beta) \\ &= [I + (X'X)^{-1}nh^2\mu_2 I]^{-1}[(X'X)^{-1}X'U - (X'X)^{-1}nh^2\mu_2\beta].\end{aligned}$$

Let $A = n\mu_2(X'X)^{-1}$, then $A = A'$, since

$$\tilde{\beta}(h) - \beta = (I + h^2 A)^{-1}[(X'X)^{-1}X'U - h^2 A\beta].$$

Since the window-width is small, or $h^2 \to 0$, we expand at 1, get a geometric series at the right hand side as

$$(I + h^2 A)^{-1} = I - h^2 A + h^4 A^2 + O(h^6).$$

Thus

$$\tilde{\beta}(h)-\beta \simeq [I+h^2A+h^4A^2][(X'X)^{-1}X'U-h^2A\beta].$$

$$= (X'X)^{-1}X'U-h^2A\beta-h^2A(X'X)^{-1}X'U+h^4A^2\beta+h^4A^2(X'X)^{-1}X'U,$$

and

$$\text{Bias} = E(\tilde{\beta}(h)-\beta) = h^4A^2\beta-h^2A\beta,$$

$$V(\tilde{\beta}(h)-\beta) = E[(\tilde{\beta}(h)-\beta)(\tilde{\beta}(h)-\beta)']$$

$$= \sigma^2(X'X)^{-1}-h^2\sigma^2(X'X)^{-1}A'+h^4\sigma^2(X'X)^{-1}A'^2-h^2\sigma^2A(X'X)^{-1}$$

$$+h^4A\beta\beta'A'+h^4\sigma^2A(X'X)^{-1}A'+h^4\sigma^2A^2(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}-2h^2\sigma^2\frac{A^2}{n\mu_2}+h^4[A\beta\beta'A'+3\sigma^2\frac{A^3}{n\mu_2}].$$

And hence the AMSE of $\tilde{\beta}(h)$ is

$$\text{AMSE} = V(\tilde{\beta}(h)-\beta)+(\text{Bias})^2$$

$$= V(\tilde{\beta}(h)-\beta)+(h^4A^2\beta-h^2A\beta)(h^4A^2\beta-h^2A\beta)'$$

$$\simeq V(\tilde{\beta}(h)-\beta)+h^4A\beta\beta'A'$$

$$= \sigma^2(X'X)^{-1}-2h^2\sigma^2\frac{A^2}{n\mu_2}+h^4[2A\beta\beta'A'+3\sigma^2\frac{A^3}{n\mu_2}].$$

To minimize the risk, we need

$$E(\tilde{\beta}(h)-\beta)'(\tilde{\beta}(h)-\beta) = \text{tr(AMSE)} \qquad \text{... (A1)}$$

$$= \sigma^2\text{tr}(X'X)^{-1}-2h^2\sigma^2\frac{\text{tr}A^2}{n\mu_2}+h^4[2\beta'A^2\beta+3\sigma^2\frac{\text{tr}A^3}{n\mu_2}].$$

The first order condition of equation (A1) gives $\frac{\partial \text{trAMSE}}{\partial h}=0$, and thus

$$h^2 = \frac{\sigma^2\text{tr}A^2}{2n\mu_2\beta'A^2\beta+3\sigma^2\text{tr}A^3}. \quad \blacksquare$$

Proof of Theorem 2.

Under condition A1-A6 above, let $D=(X'X+nh^2\mu_2 I)^{-1}$, so $D=D'$ and

$$\tilde{\beta}(h)-\beta = (X'X+nh^2\mu_2 I)^{-1}X'Y-\beta.$$

Thus MSE of $\tilde{\beta}(h)$ is

$$\text{MSE}(\tilde{\beta}(h)) = E[(\tilde{\beta}(h)-\beta)'(\tilde{\beta}(h)-\beta)]$$

$$= E[(U'X-nh^2\mu_2\beta')D'D(X'U-nh^2\mu_2\beta)]$$

$$= \sigma^2 tr(XD^2X') + (nh^2\mu_2)^2\beta'D^2\beta .$$

Since $\hat{E}(\hat{\beta}-\beta)'(\hat{\beta}-\beta) = s^2tr(D^2X'X) + h^2\hat{\beta}D^2\hat{\beta}$, where $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, we have

$$\begin{aligned} E(\hat{\beta}'D^2\hat{\beta}) &= E[(\hat{\beta}-\beta)'D^2(\hat{\beta}-\beta)] + \beta'D^2\beta \\ &= E[(\hat{\beta}-\beta)'(X'X)^{\frac{1}{2}}(X'X)^{-\frac{1}{2}}D'D(X'X)^{-\frac{1}{2}}(X'X)^{\frac{1}{2}}(\hat{\beta}-\beta)] + \beta'D^2\beta \\ &= E[Z'CZ] + \beta'D^2\beta \\ &= \sigma^2 trC + \beta'D^2\beta \end{aligned}$$

where $Z \equiv (X'X)^{\frac{1}{2}}(\hat{\beta}-\beta)$, $C \equiv (X'X)^{-\frac{1}{2}}D'D(X'X)^{-\frac{1}{2}}$.

Thus the unbiased estimator of $\beta'D^2\beta$ is $\hat{\beta}'D^2\hat{\beta} - \sigma^2 trC$, and the unbiased estimator of the MSE of $\tilde{\beta}(h)$ is $s^2tr(XD^2X') + (nh^2\mu_2)^2[\hat{\beta}'D^2\hat{\beta} - s^2trD^2(X'X)^{-1}]$. ■

Proof of Theorem 3.

Under the same condition with Theorem 2,

$$\begin{aligned} MSE(\tilde{\mu}(h)) &= E[(\tilde{\mu}(h)-\mu)'(\tilde{\mu}(h)-\mu)] \\ &= E[(\tilde{\beta}(h)-\beta)'X'X(\tilde{\beta}(h)-\beta)] \\ &= E[(U'X - nh^2\mu_2\beta')D'X'XD(X'U - nh^2\mu_2\beta)] \\ &= E[U'XD'X'XDX'U + (nh^2\mu_2)^2\beta'D'X'XD\beta] \\ &= \sigma^2 tr[XD'X'XDX'] + (nh^2\mu_2)^2\beta'D'X'XD\beta . \end{aligned}$$

Following the same logic in the proof of Theorem 2, the unbiased estimator of $(nh^2\mu_2)^2\beta'D'X'XD\beta$ is $(nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2tr(D'X'XD(X'X)^{-1})]$.

Thus, the unbiased estimator of $MSE(\tilde{\mu}(h))$ is $s^2tr(XD'X')^2 + (nh^2\mu_2)^2[\hat{\beta}'D'X'XD\hat{\beta} - s^2trD'X'XD(X'X)^{-1}]$. ■

References

Buckland, S.T., Burnham, K.P., Augustin, N.H. (1997), "Model Selection: An Integral Part of Inference", *Biometrics*, 53(2): 603-618.

Claeskens, G., Croux, C., van Kerckhoven, J. (2006), "Variable Selection for Logistic Regression Using A Prediction-Focused Information Criterion", *Biometrics*, 62(4): 972-979.

Campbell, J.Y., Thompson, S.B. (2008), "Predicting Excess Stock Returns Out of Sample: Can Anything Beat The Historical Average?", *The Review of Financial Studies*, 21(4): 1509-1531.

Fan, J., Li, R., (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties", *Journal of American Statistical Association*, 96(456): 1348-1360.

Greene, W.H. (2011), *Econometric Analysis*, New Jersey: Prentice Hall.

Hansen, B.E. (2007), "Notes and Comments Least Squares Model Averaging", *Econometrica*, 75(4): 1175-1189.

Hansen, B.E., Racine, J. (2012), "Jackknife Model Averaging", *Journal of Econometrics*, 167(1): 38-46.

Hemmerle, W.J., Carey M.B. (1983), "Some Properties of Generalized Ridge Estimators", *Communications in Statistics: Computation and Simulation*, 12(3): 239-253.

Hoerl, A.E. (1962), "Application of Ridge Analysis to Regression Problems", *Chemical Engineering Progress*, 58(3): 54-59.

Hoerl, A.E., Kennard. R.W., Baldwin, K.F. (1975), "Ridge Regression: Some Simulations",. *Communications in Statistics*, 4(2): 105-123.

Hoerl, A.E., Kennard, R.W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12(1): 55-67.

Hoerl, A.E., Kennard, R.W. (1970b), "Ridge Regression: Application to Nonorthogonal Problems", *Technometrics*, 12(1): 69-82.

Hjort, N.L., Claeskens, G. (2003), "Frequentist Model Average Estimators", *Journal of the American Statistical Association*, 98(464): 879-899.

Leamer, E.E., Chamberlain, G. (1976), "A Bayesian Interpretation of Pretesting", *Journal of Royal Statistical Society*, 13(38): 85-94.

Li, K.C. (1986), "Asymptotic Optimality of CL and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing", *The Annals of Statistics*, 14(3): 1101-1112.

Li, K.C. (1987), "Asymptotic Optimality for C_p,C_L, Cross-Ralidation and Generalized Cross-Validation: Discrete Index Set", *The Annals of Statistics*, 15(3): 958-975.

Pagan, A., Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

Schmidt, P. (1976), *Econometrics*, New York: CRC Press.

Scott, D.W., Terrell C.R. (1987), "Biased and Unbiased Cross-validation in Density Estimation", *Journal of American Statistical Association*, 82(400): 1131-1146

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

Vinod, H.D., Ullah, A. (1981), *Recent Advances in Regression Methods*, New York: Marcel Dekker.

Vinod, H.D., Ullah, A., Kadiyala, K. (1981), "A Family of Improved Shrinkage Factors for the Ordinary Ridge Estimator", *The Economic Studies Quarterly*, 32(2): 164-175.

Liang, H., Zou, G., Wan, A.T.K., Zhang, X. (2011), "Optimal Weight Choice for Frequentist Model Average Estimators", *Journal of the American Statistical Association*, 106(495): 1053-1066.

Wan, A.T.K., Zhang, X., Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion", *Journal of Econometrics*, 156(2): 277-283.

♦♦♦