

A Semiparametric Generalized Ridge Estimator and Link with Model Averaging

Aman Ullah*, Alan T.K. Wan†, Huansha Wang‡, Xinyu Zhang§, Guohua Zou¶

July 5, 2014

Abstract

In recent years, the suggestion of combining models as an alternative to selecting a single model from a frequentist prospective has been advanced in a number of studies. In this paper, we propose a new semi-parametric estimator of regression coefficients, which is in the form of a feasible generalized ridge estimator by Hoerl and Kennard (1970b) but with different biasing factors. We prove that the generalized ridge estimator is algebraically identical to the model average estimator. Further, the biasing factors that determine the properties of both the generalized ridge and semi-parametric estimators are directly linked to the weights used in model averaging. These are interesting results for the interpretations and applications of both semi-parametric and ridge estimators. Furthermore, we demonstrate that these estimators based on model averaging weights can have properties superior to the well-known feasible generalized ridge estimator in a large region of the parameter space. Two empirical examples are presented.

1 Introduction

Ordinary least squares (OLS) is a widely used estimator for the coefficients of a linear regression model in econometrics and statistics (Schmidt (1976); Greene (2011)). It is shown here that the OLS estimator can also be obtained by estimating population moments (variances and covariances) of the economic variables involved in the regression by using empirical densities of their data sets. Further,

*Department of Economics, University of California, Riverside; email address: aman.ullah@ucr.edu.

†Department of Management Sciences, City University of Hong Kong; email address: msawan@cityu.edu.hk.

‡Department of Economics, University of California, Riverside; email address: huansha.wang@email.ucr.edu.

§Academy of Mathematics and Systems Science and Center of Forecasting Science, Chinese Academy of Sciences; email address: xinyu@amss.ac.cn.

¶Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Science, Capital Normal University; email address: ghzhou@amss.ac.cn.

we propose a new estimator of the regression coefficients by estimating population moments based on smooth kernel nonparametric density estimation. This proposed estimator, in contrast to the OLS estimator, is robust to multicollinearity, and we refer to this as the semi-parametric (SP) estimator of the regression coefficients. Although there are differences, this SP estimator turns out to be in the form of a generalized ridge regression (GRR) estimator developed by Hoerl and Kennard (1970b). Ridge regression (RR) (Hoerl and Kennard (1970a, b)) is a common shrinkage technique in linear regression when the covariates are highly collinear, and among the various ridge techniques, the GRR estimator is arguably the one that has attracted the most attention. The GRR estimator allows the biasing factor, which controls the amount of ridgeing, to be different for each coefficient; when the biasing factors are the same for all coefficients, the GRR estimator reduces to the ordinary RR estimator. However, since the biasing factors are unknown, the GRR estimator is not feasible. This is not the case for the SP estimator which is based on the information contained in the kernel density estimation of regressors, and hence the biasing factors are calculated using the data-based window-widths of regressors. Thus, the SP estimator, in contrast to the GRR estimator, is a feasible estimator. This SP estimator is compared with Hoerl and Kennard's (1970b) feasible GRR (FGRR) estimator based on the first step of a data-based iterative procedure for estimating the biasing factors. We note from Hemmerle and Carey (1983) that the FGRR estimator is more efficient than the estimator based on the closed form solution of Hoerl and Kennard's iterative method. For more details of GRR estimators, see Vinod and Ullah (1981) and Vinod, Ullah and Kadiyala (1981).

Yet another independently developed technique closely related to shrinkage estimation is model averaging, which is an alternative to model selection. While the process of model selection is an attempt to find a single best model for a given purpose, model averaging compromises across the competing models, and by so doing includes the uncertainty associated with the individual models in the estimation of parameter precision. Bayesian model averaging (BMA) has long been a popular statistical technique. In recent years, frequentist model averaging (FMA) has also been garnering interest. A major part of this literature is concerned with ways of weighting models. For BMA, models are usually weighted by their posterior model probabilities, whereas FMA weights can be based on scores of information criteria (e.g. Buckland, Burnham and Augustin (1997); Claeskens, Croux and van Kerckhoven (2006); Zhang and Liang (2011); Zhang, Wan and Zhou (2012)). Other FMA strategies that have been developed include adaptive regression by mixing by Yang (2001), Mallows model averaging (MMA) by Hansen (2007, 2008) (see also Wan, Zhang and Zou (2010)), optimal mean square error averaging by Liang, Zou, Wan and Zhang (2011), and Jackknife model averaging (JMA) by Hansen and Racine (2012) (see also Zhang, Wan and Zou (2013)). As well, Hjort and Claeskens (2003) introduced a local misspecification framework for studying the asymptotic properties of FMA estimators.

Given these two independent, but parallel, developments of research in ridge type shrinkage estimators and FMA estimators, the objective of this paper is to explore a link between them. An initial attempt in establishing this connection was made by Leamer and Chamberlain (1976), where a relationship between the ridge estimator and a model average estimator (which they called "search estimator") was noted. However, we emphasize that the ridge and model averaging estimators of Leamer and Chamberlain (1976) are respectively different from the ridge and model averaging estimators in our paper. More importantly, our results permit an exact connection between model averaging weights and ridge biasing factors, whereas their results do not. In addition, we propose a new SP ridge estimator and investigate its properties. The biasing factors of the SP estimator are also linked to the FMA weights. On the basis of these relationships, the selection of biasing factors in the GRR and SP estimators may be converted to the selection of weights in the FMA estimator. Our finding also implies that if the goal is to optimally mix the competing models based on a chosen criterion, e.g., Hansen's (2007) Mallows criterion, then there is always a GRR estimator that matches the performance of the resultant FMA estimator. We demonstrate via a Monte Carlo study that the GRR estimators with biasing factors derived from the weights used for Hansen's (2007) MMA and Hansen and Racine's (2012) JMA estimators perform well, in terms of risk, in a large region of parameter space.

This paper is organized as follows. In Section 2, we present the SP and GRR estimators of the regression coefficients. In Section 3, we derive the exact algebraic relationship between the biasing factors of the SP and GRR estimators and the weights in the FMA estimator. Section 4 presents asymptotically optimal procedures for choosing window-widths. Section 5 reports the results of a Monte Carlo study comparing the risks of the SP and FGRR estimators with biasing factors based on weights of the MMA and JMA estimators. Section 6 provides two empirical applications of the SP and GRR estimators using the equity premium data in Campbell and Thompson (2008) and the wage data from Wooldridge (2003). Section 7 offers some concluding remarks.

2 Semiparametric Estimator of Regression Coefficients

Let us consider a population multiple regression model

$$\begin{aligned} y &= x_1\beta_1 + \cdots + x_q\beta_q + u \\ &= x'\beta + u, \end{aligned} \tag{1}$$

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

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

$$\beta = [Exx']^{-1}Exy, \tag{2}$$

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 given by

$$\begin{aligned} Ex_j^2 &= \int_{x_j} x_j^2 f(x_j) dx_j, \quad j = 1, \dots, q, \\ \text{and } Ex_j x_{j'} &= \int_{x_j} \int_{x_{j'}} x_j x_{j'} f(x_j, x_{j'}) dx_j dx_{j'}, \quad j \neq j' = 1, \dots, q, \end{aligned} \quad (3)$$

respectively.

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

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

It is straightforward to note that

$$\begin{aligned} \hat{E}x_j^2 &= \int_{x_j} x_j^2 \hat{f}(x_j) dx_j = \int_{x_j} x_j^2 d\hat{F}(x_j) \\ &= \frac{1}{n} \sum_{i=1}^n x_{ij}^2 \end{aligned} \quad (5)$$

by using the empirical distribution of $\hat{F}(\cdot)$. The results for $\hat{E}x_j x_{j'}$ in (4) and $\hat{E}x_j y = \sum_{i=1}^n x_{ij} y_i / n$ follow similarly.

Using (4) and (5) in (2), we obtain, for all j and j' ,

$$\begin{aligned} \hat{\beta} &= (\hat{E}xx')^{-1} \hat{E}xy \\ &= (X'X)^{-1} X'Y, \end{aligned} \quad (6)$$

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 a smooth nonparametric kernel density instead of the empirical distribution function. This results in

$$\begin{aligned} \tilde{E}x_j^2 &= \int_{x_j} x_j^2 \tilde{f}(x_j) dx_j \\ &= \frac{1}{nh_j} \sum_{i=1}^n \int_{x_j} x_j^2 k\left(\frac{x_{ij} - x_j}{h_j}\right) dx_j \\ &= \frac{1}{n} \sum_{i=1}^n \int_{\Psi_{ij}} (x_{ij} - h_j \Psi_{ij})^2 k(\Psi_{ij}) d\Psi_{ij} \\ &= \frac{1}{n} \sum_{i=1}^n \int_{\Psi_{ij}} (x_{ij}^2 + h_j^2 \Psi_{ij}^2 - 2x_{ij} h_j \Psi_{ij}) k(\Psi_{ij}) d\Psi_{ij} \\ &= \frac{1}{n} \sum_{i=1}^n x_{ij}^2 + h_j^2 \mu_2, \end{aligned} \quad (7)$$

where $\tilde{f}(x_j) = \frac{1}{nh_j} \sum_{i=1}^n k(\frac{x_{ij}-x_j}{h_j})$ is a kernel density estimator, $\Psi_{ij} = \frac{x_{ij}-x_j}{h_j}$ 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_j is window-width. For implementation, h_j can be selected by biased cross-validation based on the Normal or Epanechnikov kernel as in Scott and Terrell (1987). For more details, see Pagan and Ullah (1999).

Similarly, it can be shown easily that

$$\begin{aligned}
\tilde{E}(x_j x_{j'}) &= \int_{x_j} \int_{x_{j'}} x_j x_{j'} \tilde{f}(x_j, x_{j'}) dx_j dx_{j'} \\
&= \frac{1}{nh_j h_{j'}} \sum_{i=1}^n \int_{x_j} \int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h_j}, \frac{x_{ij'}-x_{j'}}{h_{j'}}) dx_j dx_{j'} \\
&= \frac{1}{nh_j h_{j'}} \sum_{i=1}^n \int_{x_j} \int_{x_{j'}} x_j x_{j'} k(\frac{x_{ij}-x_j}{h_j}) k(\frac{x_{ij'}-x_{j'}}{h_{j'}}) dx_j dx_{j'} \\
&= \frac{1}{n} \sum_{i=1}^n \int_{\Psi_{ij}} \int_{\Psi_{ij'}} (x_{ij} - h_j \Psi_{ij})(x_{ij'} - h_{j'} \Psi_{ij'}) k(\Psi_{ij}) k(\Psi_{ij'}) d\Psi_{ij} d\Psi_{ij'} \\
&= \frac{1}{n} \sum_{i=1}^n x_{ij} x_{ij'}
\end{aligned} \tag{8}$$

and

$$\tilde{E}(x_j y) = \frac{1}{n} \sum_{i=1}^n x_{ij} y_i, \tag{9}$$

where the product kernels have been used without loss of generality and $\Psi_{ij'} = \frac{x_{ij'}-x_{j'}}{h_{j'}}$. Also, $\tilde{E}(x_j) = \frac{1}{n} \sum_{i=1}^n x_{ij} = \bar{x}_j$.

Thus, by using (7) to (9) in (2), we obtain the following new estimator of β :

$$\begin{aligned}
\tilde{\beta} &= (\tilde{E}xx')^{-1} \tilde{E}xy \\
&= (X'X + D)^{-1} X'Y,
\end{aligned} \tag{10}$$

where $D = \text{diag}(d_1, \dots, d_q)$ is a diagonal matrix with $d_j = nh_j^2 \mu_2$ as its j -th element ($j = 1, \dots, q$). We refer to $\tilde{\beta}$ as the SP estimator.

The estimators in (7) and (8) are based on kernel density estimation assuming that the continuous regressors have support in the entire Euclidean space. In this paper, we assume that all regressors satisfy this property. However, when the regressors have a bounded support, it is well-known that the kernel density estimator is asymptotically biased and one should use bias adjusted kernels instead; see Li and Racine (2007) and Darolles, Fan, Florens and Renault (2011). When the variables are discrete, the estimator in (8) remains the same, but the estimator of Ex_j^2 can be written as $\sum_i x_i^2 p(x_i) = \sum_i \sum_j x_i^2 I(x_j = x_i)/n = \sum_i x_i^2/n$, where $I(x_j = x_i) = 1$ if $x_j = x_i$ and 0 otherwise. In this case, the estimator in (10) reduces to the OLS estimator. On the other hand, when the regressor matrix contains

a mixture of discrete and continuous regressors, the estimator again has the form of (10), except that the matrix D is re-defined with its diagonal elements corresponding to the discrete variables set to zero. This can be explained by noting, for example, when x_1 is continuous and x_2 is discrete, that the estimator of

$$\begin{aligned}
E(x_1x_2) &= E_{x_2}[x_2E(x_1|x_2)] \\
&= \sum_i \int_{x_1} K((x_{i1} - x_1)/h_1) dx_1 E[(x_2 I(x_{i2} = x_2))/p(x_2)]/nh_1 \\
&= \sum_i \sum_j x_{i1}x_{j2} I(x_{i2} = x_{j2})/n \\
&= \sum_i x_{i1}x_{i2}/n.
\end{aligned}$$

Note that both the OLS and SP estimators are based on the population regression (1), where the regression coefficient vector depends on the population moments of the vector x and the scalar variable y . These moments are then estimated using sample data by two different methods. This leads to estimators of the regression coefficients in the sample linear regression model

$$Y = X\beta + U, \quad (11)$$

where the sample is drawn from the population linear regression model (1), and U is an $n \times 1$ vector of random errors with $EU = 0$ and $EUU' = \sigma^2 I_n$. By standard eigenvalue decomposition, we can write $X'X = G\Lambda G'$, where G is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_q)$.

From Hoerl and Kennard (1970a, b), the GRR estimator of β is

$$\hat{\beta}(K) = (X'X + GKG')^{-1}X'Y, \quad (12)$$

where $K = \text{diag}(k_1, k_2, \dots, k_q)$ is a diagonal matrix with $k_j \geq 0$, $j = 1, \dots, q$. The k_j 's are the biasing factors controlling the amount of ridging in $\hat{\beta}(K)$. When $k_1 = k_2 = \dots = k_q = k$, $\hat{\beta}(K)$ is commonly called the ordinary ridge regression estimator. We note that the SP estimator in (10) is in the form of the GRR estimator but these two estimators are not exactly the same. However, one may define an alternative SP-type estimator by equating the diagonal matrix D to the diagonal of the matrix GKG' . Thus, the elements of D can be determined from the biasing factors of the GRR estimator. Of course, if $K = kI$, then $D = K$ and the SP estimator is identical to the GRR estimator.

Define $Z = XG$ and $\alpha = G'\beta$. Then $Z'Z = \Lambda$ and model (11) may be reparameterized as

$$Y = Z\alpha + U. \quad (13)$$

Correspondingly, the GRR estimator of α is

$$\hat{\alpha}(K) = (Z'Z + K)^{-1}Z'Y = (\Lambda + K)^{-1}Z'Y = BZ'Y, \quad (14)$$

where $B = (\Lambda + K)^{-1}$ is a diagonal matrix. It is straightforward to show that

$$\hat{\alpha}(K) = G'\hat{\beta}(K). \quad (15)$$

Hence

$$E(\hat{\alpha}(K) - \alpha)'(\hat{\alpha}(K) - \alpha) = E(\hat{\beta}(K) - \beta)'(\hat{\beta}(K) - \beta). \quad (16)$$

That is, the trace of the MSE matrix (or equivalently, the risk under squared error loss) of the GRR estimator of α is the same as that of β , and the matrix K that minimizes the risk of $\hat{\alpha}(K)$ also minimizes that of $\hat{\beta}(K)$. It is well-known that the GRR estimator in (12) can be derived by minimizing $u'u$ with respect to β subject to the restriction that $\beta'GKG'\beta$ is bounded. Similarly, the SP estimator in (10), derived from using smooth kernel density estimators of moments, also results from minimizing $u'u$ with respect to β subject to a bounded restriction of $\beta'D\beta$. Note that both the GRR and SP estimators are robust to multicollinearity, a property not shared by the OLS estimator derived using empirical density estimation of moments. In Sections 4 and 5 we will show that the proposed SP and GRR estimators have superior performance to the OLS estimator in risk under squared error loss sense.

3 Connection between SP and Ridge Estimators and Model Averaging

To examine the connection between the SP and GRR estimators and model averaging, let us consider an averaging scheme across the sub-models

$$Y = Z_s\alpha_s + U, \quad s = 1, 2, \dots, S, \quad (17)$$

where Z_s is a sub-matrix containing $q_s \leq q$ columns of Z , and α_s is the corresponding coefficient vector.

Least squares estimation of the models in (17) yields the OLS estimators

$$\hat{\alpha}_s = (Z_s'Z_s)^{-1}Z_s'Y. \quad (18)$$

Let us write $\alpha_s = A_s\alpha$, where $A_s = (I_{q_s} : 0_{q_s \times (q-q_s)})$ (or its column permutation) is a $q_s \times q$ selection matrix. Conformably, we write $Z_s = ZA_s'$.

The model averaging (MA) estimator of α ,

$$\hat{\alpha}(w) = \sum_{s=1}^S w_s A_s' \hat{\alpha}_s, \quad (19)$$

where $w = (w_1, w_2, \dots, w_S)'$ is the weight vector with $w_s \geq 0$ and $\sum_{s=1}^S w_s = 1$, is formed by a weighted combination of coefficient estimators across the S sub-models.

We can equivalently write $\hat{\alpha}(w)$ in (19) as

$$\begin{aligned}\hat{\alpha}(w) &= \sum_{s=1}^S w_s A_s' (A_s Z' Z A_s')^{-1} A_s Z' Y \\ &= CZ'Y,\end{aligned}\tag{20}$$

where

$$\begin{aligned}C &= \sum_{s=1}^S [w_s A_s' (A_s Z' Z A_s')^{-1} A_s] \\ &= \begin{pmatrix} w_1^* \lambda_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_q^* \lambda_q^{-1} \end{pmatrix}\end{aligned}\tag{21}$$

and

$$w_j^* = \sum_{s=1}^S w_s I(j \in \Psi_s),\tag{22}$$

with $I(\cdot)$ being an indicator function that takes on 1 if $j \in \Psi_s$ and 0 otherwise, and Ψ_s being a set comprising the column indices of Z included in the s -th sub-model. For example, if the regressor matrix of the s -th sub-model comprises the first, second and fourth columns of Z , then $\Psi_s = \{1, 2, 4\}$. In view of the relationship between w_j^* and w_s , we can write (20) as

$$\hat{\alpha}(w^*) = CZ'Y = \hat{\alpha}(w)\tag{23}$$

where $w^* = (w_1^*, \dots, w_q^*)'$.

Comparing equations (14) and (20), we notice an algebraic similarity between the GRR estimator $\hat{\alpha}(K) = BZ'Y$ and the MA estimator $\hat{\alpha}(w^*) = CZ'Y$. Clearly, $\hat{\alpha}(K) = \hat{\alpha}(w^*)$ if $B = C$, or more explicitly,

$$\begin{aligned}w_1^* \lambda_1^{-1} &= (\lambda_1 + k_1)^{-1} \\ &\vdots \\ &\vdots \\ w_q^* \lambda_q^{-1} &= (\lambda_q + k_q)^{-1}.\end{aligned}\tag{24}$$

This is the essence of the algebraic equivalence between the GRR and MA estimators. Note that λ 's depend on the data, and w 's can be determined by the MA weights w 's derived under a given criterion. Subsequently, the biasing factors k 's of the GRR estimator in (12) can be obtained from (24).

As a simple illustration, suppose that $q = 2$ in model (11) and the data observations are such that $\lambda_1 = 1$ and $\lambda_2 = 1.5$. In this case, the model average is a combination of $S = 3$ candidate models including the full model. The two sub-models contain the first and second regressors respectively, while the full model contains both regressors. Now, suppose that the weights assigned to the three models are $\hat{w}_1 = 0.5$, $\hat{w}_2 = 0.2$ and $\hat{w}_3 = 0.3$ respectively. By (22), we have

$$\begin{aligned}\hat{w}_1^* &= \sum_{s=1}^3 \hat{w}_s I(1 \in \Psi_s) = \hat{w}_1 + \hat{w}_3 = 0.8 \quad \text{and} \\ \hat{w}_2^* &= \sum_{s=1}^3 \hat{w}_s I(2 \in \Psi_s) = \hat{w}_2 + \hat{w}_3 = 0.5.\end{aligned}$$

Then

$$\begin{aligned}\hat{k}_1 &= \hat{w}_1^{*-1} \lambda_1 - \lambda_1 = 0.25 \quad \text{and} \\ \hat{k}_2 &= \hat{w}_2^{*-1} \lambda_2 - \lambda_2 = 1.5.\end{aligned}$$

Equation (24) also shows that when $k_1 = k_2 = \dots = k_q = 0$ such that the GRR estimator reduces to the OLS estimator, the MA estimator reduces to the OLS estimator in the full model. It should be mentioned that although (22) allows unique w_j^* to be determined from the given values of w_j^* s, the converse need not to be true. Thus, while one can obtain unique GRR biasing parameters from the MA weights using (24), the reverse derivation of unique MA weights from the GRR biasing parameters is not always feasible.

Note that the connection between model averaging and ridge estimators has been established on the basis of the orthogonal model. If we apply model averaging to the original regressors X directly, we cannot write the resulting model averaging estimator as a GRR estimator (see (12)), especially since $X'X + GKG'$ is not a diagonal matrix. It is only through orthogonalization that the GRR estimator (14) and model averaging estimator (20) have a common structure, i.e., a diagonal matrix multiplied by $Z'Y$. Due to the convenience it offers, orthogonalization is commonly used in the ridge literature (see Vinod and Ullah (1981)). It has also been used in recent model averaging studies (e.g., Magnus, Powell and Prufer (2010) and Magnus, Wan and Zhang (2011)).

It is also instructive to note that if model averaging is applied to the original regressors, no direct connection can be established for the SP estimator in (12) and the model averaging estimator since $X'X + D$ is not a diagonal matrix. Additionally, the estimator for the orthogonal model is $\tilde{\alpha} = (\Lambda + GKG')^{-1}Z'y$, for which no algebraic relationship with the model averaging estimator is apparent. However, if we write model (1) as $y = x'GG'\beta + U = z'\alpha + U$, with $z' = x'G$ and $\alpha = G'\beta$, then by using the technique of moments based on kernel density estimation with respect to (7) and (8), we can obtain $\hat{\alpha} = (Z'Z + D_z)^{-1}Z'Y = (\Lambda + D_z)^{-1}Z'Y$, where D_z is identical to D in (10) except that h_j , the window-width for the j -th variable x_j , is replaced by the window-width h_{jz} used for the density

estimation of the j -th variable z_j . Thus, there is a direct linkage between the SP estimator applied to the transformed population model and the model averaging estimator. However, $\tilde{\beta}(D_z) = G'^{-1}\hat{\alpha} = (X'X + GD_zG')^{-1}X'Y$, which is identical to the GRR estimator except for the replacement of D_z by K , but it is not the same as the SP estimator $(X'X + D)^{-1}X'Y$ unless $X'X + D = X'X + GD_zG'$, i.e., they are identical only when $D = GD_zG'$. Although not reported here, our simulation results show that these two different looking SP estimators yield similar risk performance. Furthermore, as D and K are diagonal matrices, the optimal choice of K will uniquely determine the optimal choice of D_z ; in other words, k_j uniquely determines h_j .

4 Asymptotically Optimal Selection of Window-Width in $\tilde{\beta}$

4.1 Unbiased Estimator of Exact Risk of SP Estimator and Prediction

From (10) and (11),

$$\tilde{\beta} - \beta = (X'X + D)^{-1}(X'U - D\beta), \quad (25)$$

which yields

$$(\tilde{\beta} - \beta)'(\tilde{\beta} - \beta) = \beta'D(X'X + D)^{-2}D\beta + U'X(X'X + D)^{-2}X'U - 2\beta'D(X'X + D)^{-2}X'U. \quad (26)$$

Therefore, by taking expectations on both sides of (26), we can write

$$R(h) = R(\tilde{\beta}) = \beta'A_1\beta + \sigma^2\text{tr}A_2, \quad (27)$$

where $A_1 = D(X'X + D)^{-2}D$, $A_2 = (X'X + D)^{-2}X'X$ and $h = (h_1^2, \dots, h_q^2)'$.

Now, note that an unbiased estimator of $\beta'A_1\beta$ is

$$\hat{\beta}'A_1\hat{\beta} - \hat{\sigma}^2\text{tr}(A_1(X'X)^{-1}), \quad (28)$$

where $\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - q)$ is an unbiased estimator of σ^2 . Thus, an unbiased estimator of $R(h)$ is

$$\hat{R}^*(h) = \hat{\beta}'A_1\hat{\beta} + \hat{\sigma}^2\text{tr}(A_2 - A_1(X'X)^{-1}). \quad (29)$$

This expression can be used to find an optimal h . However we note that

$$\text{tr}(A_2 - A_1(X'X)^{-1}) = 2\text{tr}((X'X + D)^{-1}) - \text{tr}((X'X)^{-1}). \quad (30)$$

Therefore, it can be verified that,

$$\begin{aligned} \hat{R}(h) &= \hat{\beta}'A_1\hat{\beta} + 2\hat{\sigma}^2\text{tr}((X'X + D)^{-1}) \\ &= (\tilde{\beta} - \hat{\beta})'(\tilde{\beta} - \hat{\beta}) + 2\hat{\sigma}^2\text{tr}((X'X + D)^{-1}) \end{aligned} \quad (31)$$

is an unbiased estimator of $R(h)$ up to a term $tr((X'X)^{-1})$ which does not depend on h . Thus the optimization of h based on (31) is the same as that obtained from (29).

Similarly it can be shown that an unbiased estimator of the predictive risk of $\tilde{\mu} = X\tilde{\beta}$, $E((\tilde{\mu} - \mu)'(\tilde{\mu} - \mu)) = \beta' A_3 \beta + \sigma^2 tr(A_4) = R_1(h)$, is

$$\tilde{R}_1^*(h) = \hat{\beta}' A_3 \hat{\beta} + \hat{\sigma}^2 tr(A_4 - A_3(X'X)^{-1}) \quad (32)$$

where $\mu = X\beta$, $A_3 = X'(X(X'X + D)^{-1}X' - I)^2X$ and $A_4 = ((X'X + D)^{-1}X'X)^2$. Further, the minimization of $\tilde{R}_1^*(h)$ with respect to h is the same as the minimization of Mallows criterion

$$\tilde{R}_1(h) = (\tilde{\mu} - Y)'(\tilde{\mu} - Y) + 2\hat{\sigma}^2 tr((X'X + D)^{-1}X'X),$$

which is an unbiased estimator of $R_1(h)$ up to a term unrelated to h .

In the following subsections we show that h obtained by minimizing $\hat{R}(h)$ or $\tilde{R}_1(h)$ is asymptotically optimal. Further, we refer $\tilde{\beta}(h)$ based on $\tilde{R}(h)$ as AOSP, and based on $\tilde{R}_1(h)$ as AOSP₁.

4.1.1 Asymptotically Optimal h Using $\tilde{R}_1(h)$ (Mallows Criterion)

Let $P(h) = X(X'X + D)^{-1}X'$. Then from 4.1:

$$\tilde{\mu}(h) = X\tilde{\beta} = P(h)Y. \quad (33)$$

The squared error loss function is $L(h) = (\tilde{\mu}(h) - \mu)'(\tilde{\mu}(h) - \mu)$ and the corresponding risk is $R_1(h) = E(L(h))$. We consider the choice of h by a minimization of the following Mallows criterion from above as:

$$\tilde{R}_1(h) = (\tilde{\mu}(h) - Y)'(\tilde{\mu}(h) - Y) + 2\hat{\sigma}^2 tr(P(h)). \quad (34)$$

When minimizing $\tilde{R}_1(h)$, we restrict h to the set $H \subset R^q$. Thus, the selected h is

$$\hat{h} = \operatorname{argmin}_{h \in H} \tilde{R}_1(h). \quad (35)$$

Let $\xi = \inf_{h \in H} R_1(h)$. We assume that

$$\mu'\mu = O(n), X'U = O_p(n^{1/2}) \text{ and } n^{-1}X'X \rightarrow \Phi, \quad (36)$$

where Φ is a positive definite matrix, and

$$\xi \rightarrow \infty, \xi^{-2}\mu'\mu = o(1). \quad (37)$$

By using conditions (36)-(37), and the proof steps of Theorem 2.2 of Zhang, Wan and Zou (2013), we obtain the following asymptotic optimality property:

$$\frac{L(\hat{h})}{\inf_{h \in H} L(h)} \xrightarrow{p} 1. \quad (38)$$

Proof of (38). Observe that

$$\begin{aligned}\tilde{R}_1(h) &= (\tilde{\mu}(h) - Y)'(\tilde{\mu}(h) - Y) + 2\hat{\sigma}^2 \text{tr}(P(h)) \\ &= L(h) + U'U - 2U'P(h)U - 2\mu'P(h)U + 2\mu'U + 2\hat{\sigma}^2 \text{tr}(P(h))\end{aligned}\tag{39}$$

and

$$\begin{aligned}R_1(h) &= (P(h)\mu - \mu)'(P(h)\mu - \mu) + \sigma^2 \text{tr}(P^2(h)) \\ &= L(h) - U'P^2(h)U - 2(P(h)\mu - \mu)'P(h)U + \sigma^2 \text{tr}(P^2(h)).\end{aligned}$$

Hence to prove (38), it suffices to show that

$$\sup_{h \in H} \frac{|\mu'P(h)U|}{R_1(h)} = o_p(1),\tag{40}$$

$$\sup_{h \in H} \frac{|U'P(h)U|}{R_1(h)} = o_p(1),\tag{41}$$

$$\sup_{h \in H} \frac{|\hat{\sigma}^2 \text{tr}(P(h))|}{R_1(h)} = o_p(1),\tag{42}$$

$$\sup_{h \in H} \frac{|U'P^2(h)U|}{R_1(h)} = o_p(1),\tag{43}$$

$$\sup_{h \in H} \frac{|(P(h)\mu - \mu)'P(h)U|}{R_1(h)} = o_p(1)\tag{44}$$

and

$$\sup_{h \in H} \frac{|tr(P^2(h))|}{R_1(h)} = o_p(1).\tag{45}$$

Let $\lambda(A)$ be the largest eigenvalue of the matrix A . From condition (37) and the following formulae:

$$\begin{aligned}\sup_{h \in H} \lambda(P(h)) &\leq \lambda(X(X'X)^{-1}X') = 1, \\ \sup_{h \in H} tr(P(h)) &\leq tr(X(X'X)^{-1}X') = q, \\ \sup_{h \in H} U'P(h)U &\leq U'X(X'X)^{-1}X'U, \\ (\mu'P(h)U)^2 &\leq \mu'\mu U'P^2(h)U \leq \mu'\mu \lambda(P(h))U'P(h)U\end{aligned}$$

and

$$((P(h)\mu - \mu)'P(h)U)^2 \leq (P(h)\mu - \mu)'(P(h)\mu - \mu)U'P^2(h)U \leq R_1(h)U'P(h)U,$$

we need only to show that

$$U'X(X'X)^{-1}X'U = O_p(1),\tag{46}$$

and

$$\hat{\sigma}^2 = O_p(1). \quad (47)$$

Equations (46) and (47) are implied by condition (36). The proof of (38) thus follows. If in addition $\{L(\hat{h}) - \xi\}\xi^{-1}$ is uniformly integrable, then

$$\frac{R_1(\hat{h})}{\inf_{h \in H} R_1(h)} \rightarrow^p 1$$

4.1.2 Asymptotic Optimal h Using $\hat{R}(h)$

We restrict h in a set $H \subset R^q$. So the selected h is

$$\tilde{h} = \arg \min_{h \in H} \hat{R}(h).$$

Let $\tilde{L}(h) = (\tilde{\beta}(h) - \beta)'(\tilde{\beta}(h) - \beta)$ be the squared loss function and $\tilde{\xi} = \inf_{h \in H} R(h)$, and $\bar{h} = \max(h_1^2, \dots, h_q^2)$. We assume the following conditions

$$\bar{h} \rightarrow 0, \quad X'U = O_p(n^{1/2}), \quad \text{and } n^{-1}X'X \rightarrow \Phi \text{ where } \Phi \text{ is a positive definite matrix,} \quad (48)$$

$$n^{1/2}\tilde{\xi} \rightarrow \infty. \quad (49)$$

By using the conditions (48)-(49), we can obtain the following asymptotic optimality

$$\frac{\tilde{L}(\tilde{h})}{\inf_{h \in H} \tilde{L}(h)} \rightarrow^p 1. \quad (50)$$

Proof of (50). Observe that, with $\tilde{\beta}(h) = B(h)\hat{\beta}$ and $B(h) = (X'X + D)^{-1}X'X$, from (31)

$$\begin{aligned} \hat{R}(h) &= (B(h)\hat{\beta} - \hat{\beta})'(B(h)\hat{\beta} - \hat{\beta}) + 2\hat{\sigma}^2 \text{tr}(B(h)(X'X)^{-1}) \\ &= \tilde{L}(h) + (\hat{\beta} - \beta)'(\hat{\beta} - \beta) - 2\hat{\beta}'B'(h)(\hat{\beta} - \beta) + 2\beta'(\hat{\beta} - \beta) + 2\hat{\sigma}^2 \text{tr}(B(h)(X'X)^{-1}) \\ &\equiv \tilde{L}(h) + \Xi_1(h), \end{aligned} \quad (51)$$

and

$$\begin{aligned} R(h) &= (B(h)\beta - \beta)'(B(h)\beta - \beta) + \sigma^2 \text{tr}(B'(h)B(h)(X'X)^{-1}) \\ &= \tilde{L}(h) + (\hat{\beta} - \beta)'B'(h)B(h)(\hat{\beta} - \beta) \\ &\quad - 2\hat{\beta}'B'(h)B(h)(\hat{\beta} - \beta) + 2\beta'B(h)(\hat{\beta} - \beta) + \sigma^2 \text{tr}(B'(h)B(h)(X'X)^{-1}) \\ &\equiv \tilde{L}(h) + \Xi_2(h). \end{aligned} \quad (52)$$

From the condition (48), we have $\hat{\beta} - \beta = O_p(n^{-1/2})$, which, together with the conditions (48)-(49) leads to

$$\sup_{h \in H} \frac{|\Xi_1(h)|}{R(h)} = o_p(1) \quad (53)$$

and

$$\sup_{h \in H} \frac{|\Xi_2(h)|}{R(h)} = o_p(1). \quad (54)$$

Hence, we can obtain (50). From (50), and an additional condition that $(\tilde{L}(\tilde{h}) - \xi_n)\xi_n^{-1}$ is uniformly integrable, we further have

$$\frac{R(\tilde{h})}{\inf_{h \in H} R(h)} \rightarrow^p 1. \quad (55)$$

5 A Monte Carlo Study

The purpose of this section is to demonstrate via a Monte Carlo study the finite sample properties of GRR estimators with biasing factors obtained based on model weights of the Mallows MA (MMA) and Jackknife MA (JMA) estimators. As mentioned previously, these MA estimators were proposed by Hansen (2007) and Hansen and Racine (2012). We denote the corresponding GRR estimators as GRRM and GRRJ estimators respectively.

The weights of the MMA estimator are obtained by minimizing the quadratic form $(Y - Z\hat{\alpha}(w))'(Y - Z\hat{\alpha}(w)) + 2\hat{\sigma}^2 \text{tr}(ZCZ')$, where $\hat{\sigma}^2 = (Y - Z\hat{\alpha}_f)'(Y - Z\hat{\alpha}_f)/(n - q)$ and $\hat{\alpha}_f$ is the OLS estimator of α in the full model. On the other hand, the weights of the JMA estimator are determined by minimizing the leave-one-out least squares cross-validation function $CV_n(w) = (Y - \hat{g}(w))'(Y - \hat{g}(w))/n$, where $\hat{g}(w) = \sum_{s=1}^S w_s \hat{g}_s$, with $\hat{g}_s = (\hat{g}_{1s}, \dots, \hat{g}_{ns})'$, $\hat{g}_{is} = x_i^{s'}(X_{-i}^{s'}X_{-i}^s)^{-1}X_{-i}^{s'}Y_{-i}$, and X_{-i}^s and Y_{-i} being respectively the matrices X^s (the regressor matrix of the s -th submodel) and Y with the i -th element deleted. Following Hansen (2007), we assume that the candidate models in the model average are nested.

Our interest is focused on the risk performance under squared error loss of estimators in terms of the β space in the original model. For purposes of comparisons, we also evaluate the risks of the OLS estimator, the FGRR estimator $\hat{\alpha}_j$, where $\hat{\alpha}_j = \hat{\sigma}^2/\hat{\alpha}_{j,f}^2$ with $\hat{\alpha}_{j,f}$ being the j -th element of $\hat{\alpha}_f$, the asymptotically optimal GRR (AOGRR) estimator, with k'_j 's obtained by directly minimizing the Mallows criterion $(Y - Z\hat{\alpha}(K))'(Y - Z\hat{\alpha}(K)) + 2\hat{\sigma}^2 \text{tr}(ZBZ')$ as a function of K , and the asymptotically optimal SP (AOSP₁) estimator, with window-widths obtained by minimizing the Mallows criterion (Section 4.1.1) $(Y - Z\hat{\alpha}(D))'(Y - Z\hat{\alpha}(D)) + 2\hat{\sigma}^2 \text{tr}(ZB_1Z')$ as a function of D , where $\hat{\alpha}(D) = (\Lambda + G'DG)^{-1}Z'Y = B_1Z'Y$ is the SP estimator and $B_1 = (\Lambda + G'DG)^{-1}$. When implementing the GRRM, AOGRR, AOSP and AOSP₁ estimators, we made use of a constrained optimization routine available in R.version. 2.13.1. We used k'_j 's from the FGRR method as the initial values for computing the AOGRR estimator. In Section 4.1 we have shown that the optimization under the Mallows criterion is equivalent to the optimization with respect to unbiased estimator of the predictive risk of $\tilde{\beta}(h)$. Therefore, in our

simulation, we also consider the AOSP estimator in Section 4.1.2 based on the optimization of risk of $\tilde{\beta}(h)$.

Our Monte Carlo experiments are based on following data generating processes (DGP's):

DGP1: $y_i = \sum_{j=1}^q \theta_j x_{ij} + e_i$, $i = 1, \dots, n$, with x_{ij} being *iid* $N(0, 1)$, e_i being *iid* $N(0, 1)$ and $N(0, 25)$ and are uncorrelated with x 's. The same DGP was considered by Hansen (2007) in his Monte Carlo study. We let $\theta_j = 0.7071j^{-3/2}$, and consider the following pairs of $(n, q) = (50, 11)$ and $(150, 16)$. To facilitate interpretation of the SP estimates, without loss of generality, we assume the DGP contains no intercept.

DGP2: The set-up is the same as DGP1, except x_{i2} is taken to be the sum of x_{i3}, \dots, x_{i50} plus an $N(0, 1)$ distributed error term. The regressors are thus nearly perfectly correlated.

Our analysis is based on 100 replications. We adopt the Gaussian kernel with $K(\phi) = (2\pi)^{-1/2} \exp[-\frac{1}{2}\phi^2]$, resulting in $\mu_2 = 1$. Following Scott and Terrell (1987), we compute the window-widths of the SP estimator using biased cross-validation procedure which is based on a lightly biased estimate of mean integrated squared error of the density estimator. They show large gains in asymptotic efficiency, especially when the density is sufficiently smooth, compared to the least squares cross-validation procedure based on an unbiased estimator of the mean integrated squared errors. Although we do not report here, in our simulations too we have found that the biased cross-validation procedure performing, in risks sense, better than the naive and AIC cross-validation window-widths. However, the optimal window-width for density estimation may not be the same or optimal for the SP estimator of regression coefficients. With this in view we have also provided AOSP₁ and AOSP estimators based on the window-widths using Mallows criterion($\tilde{R}_1(h)$) and $\hat{R}(h)$ respectively, see Section 4.

Table 1: Risk for Each Estimator					
DGP	Estimators	$\sigma = 1$		$\sigma = 5$	
		$n = 50$	$n = 150$	$n = 50$	$n = 150$
1	OLS	0.0251	0.0076	0.6775	0.1802
	FGRR	0.0201	0.0071	0.3374	0.0974
	GRRM	0.0377	0.0236	0.1065	0.0435
	GRRJ	0.0373	0.0236	0.0980	0.0432
	AOGRR	0.0217	0.0079	0.2591	0.0764
	SP	0.0167	0.0064	0.3514	0.1248
	AOSP ₁	0.0135	0.0046	0.2220	0.0713
	AOSP	0.0126	0.0045	0.2170	0.0706
2	OLS	0.0253	0.0076	0.6846	0.1807
	FGRR	0.0198	0.0072	0.3388	0.0997
	GRRM	0.0341	0.0228	0.1066	0.0426
	GRRJ	0.0335	0.0228	0.0979	0.0424
	AOGRR	0.0212	0.0080	0.2592	0.0797
	SP	0.0166	0.0064	0.3488	0.1247
	AOSP ₁	0.0140	0.0046	0.2251	0.0704
	AOSP	0.0136	0.0045	0.2157	0.0701

The simulation results reported in Table 1 show that although the SP, AOSP₁ and AOSP estimators behave well when the error variance is small, the GRRM and GRRJ are clearly the preferred estimators when the error variance is large, and often by a large margin. This finding is consistent with our intuition that the large variance associated with the true model makes it difficult to identify the best model, thus making model averaging, which shields against choosing a bad model, a more viable strategy. It is also apparent from Table 1 that FGRR and AOGRR estimators yield similar risk performance. This is perhaps attributable to the fact that the biasing factors chosen for the FGRR estimator are optimal in risk sense. See Vinod, Ullah, and Kadiyala (1981, p.363) and Hoerl and Kennard (1970b, p.63). Further, we observe that values of AOSP are smaller compared to AOSP₁. This may be due to h used in the AOSP₁ estimator is based on minimizing predictive risk ($\tilde{R}_1(h)$) instead of estimator's risk ($\tilde{R}(h)$). By comparing DGP 1 and DGP 2, we notice that when the error variance is large, all estimators in DGP 2 deliver larger risk deductions compared to DGP 1.

6 Empirical Applications

This section considers two empirical applications of the proposed methods. The first application uses the methods as forecasting devices for excess stock returns while the second considers wage forecasts.

6.1 Forecasting Excess Stock Returns

The data for this example are taken from Campbell and Thompson (2008). The same data set was also used by Jin, Su and Ullah (2012) and Lu and Su (2012) in their studies. This dataset contains $n = 672$ monthly observations between January 1950 and December 2005 of Y , the monthly excess stock returns of S&P 500 Index, defined as the difference between the monthly stock returns and the risk-free rate. In addition, data observations over the same period are also provided for the following twelve regressors variables, ordered by the magnitude of their correlations with Y , as: 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, the one-period lag of excess returns and smoothed earnings price ratio. We order these 12 regressors by the magnitude of their correlations with Y . Our model average thus contains the following 13 nested models: $\{1\}, \{1, x_1\}, \{1, x_1, x_2\}, \dots, \{1, x_1, x_2, \dots, x_{12}\}$.

Our estimation is based on $n_1 = 144, 180, 216, 336$ and 456 observations and we use the remaining $n_2 = n - n_1$ observations for out-of-sample forecast evaluation purpose. We measure forecast accuracy based on the out-of-sample R^2 defined as follows:

$$R^2 = 1 - \frac{\sum_{t=n_1}^{n-1} (Y_{t+1} - \hat{Y}_{t+1})^2}{\sum_{t=n_1}^{n-1} (Y_{t+1} - \bar{Y}_{t+1})^2},$$

where \hat{Y}_p is the prediction of Y_p based on a given forecast method and \bar{Y} is the average of Y across the sample of the n_1 observations used for estimating the model. The out-of-sample R^2 is thus negative (positive) when the forecast method yields a larger (smaller) sum of squared forecast errors than does \bar{Y} . Table 2 reports the out-of-sample R^2 based on the six estimators considered in Section 5 and the selected n_1 values. The results show that except when $n_1 = 180$, the OLS forecasts are inferior to forecasts based on the historical average. This is consistent with the findings of Welch and Goyal (2008) for this data set, that the historical mean gives better forecasts when no restrictions are imposed. In all but one case, the FGRR, AOGRR and AOSP₁ estimators are also inferior to the historical average in terms of prediction accuracy. On the other hand, the GRRJ and GRRM model averaging estimators result in positive out-of-sample R^2 in the large majority of cases, with GRRJ being the slightly better estimator of the two. The result based on AOSP estimator is not presented here because it does not perform as well as AOSP₁. This may be because our evaluation here is based on predictive risk instead of risk of $\tilde{\beta}(h)$.

Estimator	$n_1 = 144$	$n_1 = 180$	$n_1 = 216$	$n_1 = 336$	$n_1 = 456$
OLS	-0.0390	0.0062	-0.0434	-0.0425	-0.0208
FGRR	-0.0375	-0.0369	-0.0398	-0.0610	-0.0621
GRRM	0.0408	0.0895	0.0564	0.0103	-0.0003
GRRJ	0.0692	0.1079	0.0701	0.0180	0.0020
AOGRR	-0.0375	-0.0369	-0.0398	-0.0610	-0.0621
AOSP ₁	-0.0302	0.0195	-0.0271	-0.0170	-0.0148

6.2 Forecasting Wages

We use the data given in Wooldridge (2003) containing a cross sectional sample of 526 observations from the U.S. Current Population Survey from year 1976. The dependent variable of interest is the logarithm of average hourly earnings. We consider the following ten regressors, ordered according to their correlation with the dependent variable: professional occupation, education, tenure, female, service occupation, married, trade, SMSA, services, and clerk occupation based on their correlations with the dependent variable. We consider model averages based on 11 nested models in the same manner described in the last example, and $n_1 = 100, 200, 300, 400..$

Table 3 reports the out-of-sample R^2 for the six methods.

Estimator	$n_1 = 100$	$n_1 = 200$	$n_1 = 300$	$n_1 = 400$
OLS	0.4516	0.4465	0.4656	0.4450
FGRR	0.4514	0.4440	0.4658	0.4410
AOGRR	0.4509	0.4418	0.4642	0.4390
GRRM	0.3964	0.3366	0.3390	0.3644
GRRJ	0.3877	0.3357	0.3375	0.3627
AOSP ₁	0.4550	0.4477	0.4664	0.4470

Table 3 reports the out-of-sample R^2 for the six methods. It is apparent from the results that all six estimators yield more accurate forecasts than the historical average. However, the advantage of the GRRJ and GRRM estimators observed in the last example does not extend to the present case, where it is found that the FGRR, AOGRR and AOSP₁ estimators all result in more accurate forecasts than the two model averaging estimators. This can be explained by noting that $R^2 = .509$ for the wage data is much higher than that for the equity premium data, which is .097. Thus the standard deviation of

errors in the wage data is much smaller compared to the standard deviation of error for the equity premium data, and our simulations suggest that for the small standard deviation case the GRRM and GRRJ estimators are outperformed by other estimators considered. This is also the reason why in the equity premium data, the GRRM and GRRJ estimators prevail since the standard deviation for this data set is much higher. Of the two model averaging estimators, the GRRM estimator is slightly preferred to the GRRJ estimator.

7 Conclusions

We have proposed a new SP estimator of regression coefficients which is in the form of the GRR estimator of Hoerl and Kennard (1970b). However, in contrast to the GRR, the biasing factors in our SP estimator are easily implemented by the window-width and the second moment of the kernel function used in the kernel density estimation. The selection of window-width that minimizes Mallows criterion(predictive risk) as well as estimator's risk are also proposed. We also show that the GRR estimator is in fact a model average estimator, and there is an algebraic relationship between the biasing factors of GRR and SP estimators and the model average weights. Naturally, the SP and GRR estimators that select the biasing factors based on this relationship have the same properties as the corresponding model average estimator. This is an interesting finding for the future application and interpretations of the SP and GRR estimators. Our Monte Carlo results demonstrate that some of the recently introduced weight choice strategies for model averaging can result in more accurate estimators than the well-known FGRR and OLS estimators over a wide range of parameter space.

Acknowledgements

The authors are thankful to Essie Maasoumi, associate editor, referees and participants of a seminar in memory of T.D. Dwivedi, Concordia University, Montreal, especially John Galbraith, for helpful comments. Aman Ullah's work was supported by the Academic Senate, UCR. Alan Wan's work was supported by a Strategic Research Grant from the City University of Hong Kong (Grant no. 7008134). Xinyu Zhang's and Guohua Zou's work was supported by the National Natural Science Foundation of China (Grant nos. 71101141 and 11271355 for Zhang, and Grant nos. 11331011 and 70933003 for Zou). The usual disclaimer applies.

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? *Review of Financial Studies* 21(4):1509-1531.
- Darolles, S., Fan, Y., Florens, J., Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* 79: 1541-1565.
- Fan, J., Li, R., (2001). Variable selection via nonconcave penalizedlikelihood 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). Least squares model averaging. *Econometrica* 75(4):1175-1189.
- Hansen, B.E. (2008). Least squares forecast averaging. *Journal of Econometrics* 146(2): 342-350.
- 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.
- Jin, S., Su, L., Ullah, A. (2013). Robustify financial time series forecasting. *Econometric Reviews*, forthcoming.
- Leamer, E.E., Chamberlain, G. (1976). A Bayesian interpretation of pretesting. *Journal of the Royal Statistical Society: Series B* 13(38):85-94.
- Li, K.C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics* 14(3):1101-1112.

- Li, K.C. (1987). Asymptotic optimality for C_p, C_L , cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics* 15(3):958-975.
- Li, Q., Racine, J.S. (2007). *Nonparametric Econometrics: Theory and Practice*. New Jersey: Princeton University Press.
- 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.
- Lu, X., Su, L. (2012). Jackknife model averaging for quantile regressions. Working paper, School of Economics, Singapore Management University.
- Magnus, J.R., Powell, O., Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* 154(2):139-153.
- Magnus, J.R., Wan, A.T.K., Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis* 55(3):1331-1341.
- 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). Evaluation of the mean squared error of certain generalized ridge estimators using confluent hypergeometric functions. *Sankhyā:Series B* 43(3):360-383.
- Wan, A.T.K., Zhang, X., Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156(2):277-283.
- Welch, I. and Goyal, A. (2008) A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21: 1455-1508.
- Wooldridge, J.M. (2003). *Introductory Econometrics: A Modern Approach*. Kentucky: Thompson South-Western.
- Yang, Y. (2001). Adaptive regression by mixing, *Journal of the American Statistical Association* 96(454):574-586.
- Yuan, Z., Yang, Y. (2005). Combining linear regression models: when and how? *Journal of the American Statistical Association* 100(472):1202-1204.

Zhang, X., Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* 39(1):174-200.

Zhang, X., Wan, A.T.K., Zhou, S.Z. (2012). Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business and Economic Statistics* 30(1):132-143.

Zhang, X., Wan, A.T.K., Zou, G. (2013). Model averaging by Jackknife criterion in models with dependent data. *Journal of Econometrics* 174(2): 82-94.