A semiparametric generalized ridge estimator and link with model averaging

Aman Ullah, Alan T. K. Wan, Huansha Wang, Xinyu Zhang & Guohua Zou

ECONOMETRIC REVIEWS
2017, VOL. 36, NOS. 1–3, 370–384
http://dx.doi.org/10.1080/07474938.2015.1114564

Taylor & Francis
Taylor & Francis Group

A semiparametric generalized ridge estimator and link with model averaging

Aman Ullah [a], Alan T. K. Wan[b], Huansha Wang[a], Xinyu Zhang[c], and Guohua Zou[d]

[a]Department of Economics, University of California, Riverside, California, USA; [b]Department of Management Sciences, City University of Hong Kong, Hong Kong, China; [c]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China and School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China; [d]School of Mathematical Science, Capital Normal University, Beijing, China

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 article, we propose a new semiparametric 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 after reparameterization such that the regressors are orthogonal, 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 semiparametric estimators are directly linked to the weights used in model averaging. These are interesting results for the interpretations and applications of both semiparametric 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.

KEYWORDS

Biasing factors; mallows; ridge estimator; squared error loss; weight

JEL CLASSIFICATION

C13; C14

1. Introduction

Ordinary least squares (OLS) is a widely used estimator of the coefficients in 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, 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 semiparametric (SP) estimator of the regression coefficients. Although there are differences, this SP estimator turns out to be in the form of the 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 that controls the amount of ridging 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, because the biasing factors are unknown, the GRR estimator is not feasible. On the other hand, our SP estimator is based on the information contained in the kernel density estimation of regressors, and the biasing factors are calculated using the data-based window-widths of the regressors. Thus, in contrast to the GRR estimator, the SP 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

CONTACT Aman Ullah aman.ullah@ucr.edu Department of Economics, University of California, Riverside, CA, USA.

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 the GRR estimators, see Vinod and Ullah (1981) and Vinod et al. (1981). A related article by Cheng et al. (1997) has considered the incorporation of a ridging strategy in the local linear nonparametric estimator that alleviates the numerical instability issue in cases of sparse design density.

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 et al., 1997; Claeskens et al., 2006; Zhang and Liang, 2011; Zhang et al., 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 et al., 2010), optimal mean square error averaging by Liang et al. (2011), and Jackknife model averaging (JMA) by Hansen and Racine (2012) (see also Zhang et al., 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 article 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. We emphasize that the ridge and the search estimators considered by Leamer and Chamberlain (1976) are different from the ridge and model averaging estimators considered in this article. Most importantly, our results permit an exact connection between model averaging weights and ridge biasing factors, whereas their results do not allow the same. 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 article 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 the population multiple regression model as follows:

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

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

If we minimize $Eu^2 = E(y - x'\beta)^2$ with respect to β, and x is a random design vector of regressors, we obtain

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

where Exx' is a $q \times q$ moment matrix of q variables with the jth diagonal element and (j, j')th off diagonal elements given by

$$Ex_j^2 = \int_{x_j} x_j^2 f(x_j)dx_j,\ j = 1, \ldots, q, \tag{3}$$

and

$$Ex_jx_{j'} = \int_{x_j}\int_{x_{j'}} x_jx_{j'}f(x_j, x_{j'})dx_jdx_{j'}, \qquad j \neq j' = 1, \ldots, q,$$

respectively.

If the sample observations $\{y_i, x_{i1}, \ldots, x_{iq}\}$, $i = 1, \ldots, n$, are available, then the population averages in (3) can be estimated by their sample averages

$$\widehat{E}x_j^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \text{ and } \widehat{E}x_jx_{j'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'}. \tag{4}$$

It is straightforward to show that

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

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

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

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

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 OLS estimator.

Now, we consider the estimation of Ex_j^2 and $Ex_jx_{j'}$ by a smooth nonparametric kernel density instead of the empirical distribution function. This results in

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

where $\tilde{f}(x_j) = \frac{1}{nh_j}\sum_{i=1}^{n} k\left(\frac{x_{ij}-x_j}{h_j}\right)$ 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

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

and

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

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

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

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

where $D = diag(d_1, \ldots, d_q)$ is a diagonal matrix with $d_j = nh_j^2\mu_2$ as its jth element ($j = 1, \ldots, q$). 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 article, 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 et al. (2011). When the variables are discrete, and we consider an estimator of their distributions with $f(x_j) = 1/n$, then an estimator of Ex_j^2 is $\sum_i x_{ij}^2/n$, and that of $E(x_j x_{j'})$ is $\sum_i x_{ij}x_{ij'}/n$. 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 redefined with its diagonal elements corresponding to the discrete variables set to zero.

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

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

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 $EUU' = \sigma^2 I_n$ conditional on X. By standard eigenvalue decomposition, we can write $X'X = G\Lambda G'$, where G is an orthogonal matrix and $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_q)$.

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

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

where $K = diag(k_1, k_2, \ldots, k_q)$ is a diagonal matrix with $k_j \geq 0, j = 1, \ldots, q$. The $k_j's$ are the biasing factors controlling the amount of ridging in $\hat{\beta}(K)$. When $k_1 = k_2 = \cdots = k_q = k$, $\hat{\beta}(K)$ 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 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

$$Y = Z\alpha + U. \tag{13}$$

Correspondingly, the GRR estimator of α is

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

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

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

Hence

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

That is, the trace of the mean squared error 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 submodels

$$Y = Z_s\alpha_s + U, \; s = 1, 2, \ldots, S, \tag{17}$$

where Z_s is a submatrix 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

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

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,

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

where $w = (w_1, w_2, \ldots, w_S)'$ 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 submodels.

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 sth submodel. For example, if the regressor matrix of the sth submodel 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 $w^* = (w_1^*, \ldots, w_q^*)'$.

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 $\lambda's$ depend on the data, and $w^{*\prime}s$ 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 submodels 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

$$\hat{w}_1^* = \sum_{s=1}^{3} \hat{w}_s I(1 \in \Psi_s) = \hat{w}_1 + \hat{w}_3 = 0.8$$

and

$$\hat{w}_2^* = \sum_{s=1}^{3} \hat{w}_s I(2 \in \Psi_s) = \hat{w}_2 + \hat{w}_3 = 0.5.$$

Then

$$\hat{k}_1 = \hat{w}_1^{*-1}\lambda_1 - \lambda_1 = 0.25$$

and

$$\hat{k}_2 = \hat{w}_2^{*-1}\lambda_2 - \lambda_2 = 1.5.$$

Equation (24) also shows that when $k_1 = k_2 = \cdots = k_q = 0$ 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 et al., 2010, and Magnus et al., 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 $\tilde{\alpha} = (\Lambda + GDG')^{-1}Z'y$, 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 jth 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, although $\tilde{\beta}(D_z) = G'^{-1}\hat{\alpha} = (X'X + GD_zG')^{-1}X'Y$ is identical to the GRR estimator except for the replacement of D_z by K, 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 the exact risk of the SP estimator and prediction

From (10) and (11),

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

which yields

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

Throughout Section 4, the results on risk are obtained conditional on X (fixed design). Therefore, by taking expectations on both sides of (26), we can write

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

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

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

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

where $\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - q)$ is an unbiased estimator of σ^2. This results in the following unbiased estimator of $R(h)$:

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

This expression can be used to find an optimal h. On the other hand, we note that

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

Therefore, it can be verified that

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

is an unbiased estimator of $R(h)$ up to a term $tr((X'X)^{-1})$ that 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

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

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

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

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 $\widehat{R}(h)$ or $\widetilde{R}_1(h)$ is asymptotically optimal. Further, we refer $\tilde{\beta}(h)$ based on $\widehat{R}(h)$ as asymptotically optimal semiparametric (ASOP), and based on $\widetilde{R}_1(h)$ as AOSP$_1$.

4.1.1. The choice of optimal bandwidth based on the Mallows criterion

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

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

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 defined above:

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

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

$$\hat{h} = \text{argmin}_{h \in H} \widetilde{R}_1(h). \tag{35}$$

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

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

where Φ is a positive definite matrix, and

$$\xi \to \infty, \xi^{-2}\mu'\mu = o(1), \tag{37}$$

which can be verified by (36) and assumptions $h^* = h_1^2 = \cdots = h_q^2$, $h^* \to 0$, and $\inf_{h \in H} h^{*2} n^{1/2} \to \infty$.

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

$$\frac{L(\hat{h})}{\inf_{h\in H} L(h)} \to^{p} 1. \tag{38}$$

Proof of (38). Observe that

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

and

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

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

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

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

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

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

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

and

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

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

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

and

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

we need only to show that

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

and

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

Equations (46) and (47) are implied by condition (36). The proof of (38) thus follows. In addition, by the above proof, we also have

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

4.1.2. The choice of optimal bandwidth based on an unbiased estimator of R(h)

We restrict h to a set $H \subset R^q$. Thus, the selected h is

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

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

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

and

$$n^{1/2}\tilde{\xi} \to \infty. \tag{49}$$

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

$$\frac{\widetilde{L}(\tilde{h})}{\inf_{h\in H} \widetilde{L}(h)} \to^p 1. \tag{50}$$

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

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

and

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

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

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

and

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

Hence we obtain (50). In addition, by using the above proof and the formula (51), we can also obtain that

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

Here, we consider a simple case with $h^* = h_1^2 = \cdots = h_q^2$. It can be easily verified that, if $h^* = o(1)$ and $(n^{1/2}h^*)^{-1} = O(1)$, then from (27), $R(h) = O(h^{*2})$. Hence, if $h^{*2} = O(n^{-1})$, then (19) is not

satisfied even though $R(h) \to 0$ as $n \to \infty$. However, if $h^{*2} = O(n^{-a})$ with $0 < a < 1/2$, then (19) is satisfied and $R(h) \to 0$.

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 MMA and 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 $(Y - Z\hat{\alpha}(w))'(Y - Z\hat{\alpha}(w)) + 2\hat{\sigma}^2 tr(ZCZ')$, 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 $\hat{g}_s = (\hat{g}_{1s}, \ldots, \hat{g}_{ns})'$, $\hat{g}_{is} = x_i^{s'}(X_{-i}^{s'} X_{-i}^{s})^{-1} X_{-i}^{s'} Y_{-i}$, and X_{-i}^{s} and Y_{-i} being, respectively, the matrices X^s (the regressor matrix of the sth submodel) and Y with the ith 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 $\hat{\alpha}_j = \hat{\sigma}^2/\hat{\alpha}_{j,f}^2$, with $\hat{\alpha}_{j,f}$ being the jth element of $\hat{\alpha}_f$, the asymptotically optimal GRR (AOGRR) estimator, with $k_j's$ obtained by directly minimizing the Mallows criterion $(Y - Z\hat{\alpha}(K))'(Y - Z\hat{\alpha}(K)) + 2\hat{\sigma}^2 tr(ZBZ')$ as a function of K, the AOSP estimator in Section 4.1.2 based on the optimization of risk of $\tilde{\beta}(h)$, and the asymptotically optimal SP (AOSP_1) estimator, with the window-widths obtained by minimizing the Mallows criterion $(Y - Z\hat{\alpha}(D))'(Y - Z\hat{\alpha}(D)) + 2\hat{\sigma}^2 tr(ZB_1Z')$ as a function of D (see Section 4.1.1), $\hat{\alpha}(D) = (\Lambda + G'DG)^{-1}Z'Y = B_1Z'Y$, and $B_1 = (\Lambda + G'DG)^{-1}$. Note that $\hat{\alpha}(D)$ is the SP estimator. Recall from our discussion in Section 4.1 that optimization under the Mallows criterion is equivalent to optimization with respect to unbiased estimator of the predictive risk of $\tilde{\beta}(h)$. When implementing the GRRM, AOGRR, AOSP, and AOSP_1 estimators, we used the routine "constrOptim" for optimization subject to linear inequality constraints in conjunction with the routine "solve.QP" that implements Goldfarb and Idnani's (1982, 1983) dual method for constrained optimization available in the packages "stat" and "quadprog," respectively, in R (version 2.13.1); when executing these routines, we only restricted the bandwidth to be positive and imposed no other restrictions on the parameters. When computing the AOGRR estimate, we used $k's$ from the FGRR method as the initial value.

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

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

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

Our analysis is based on 1,000 replications. We adopt the Gaussian kernel $K(\phi) = (2\pi)^{-1/2} \exp[-\frac{1}{2}\phi^2]$, resulting in $\mu_2 = 1$. Following Scott and Terrell (1987), we compute the window-widths of the SP estimator using a biased cross-validation procedure that is based on a (slightly) biased estimator of the mean integrated squared error of the density estimator. Scott and Terrell (1987) showed that using this biased estimator instead of the usual unbiased estimator often results in large gains in asymptotic efficiency, especially when the density is reasonably smooth. Although we do not report the results here, in our simulation experiments, we have also found that this biased cross-validation procedure

Table 1. Risks of estimators.

		$\sigma = 1$		$\sigma = 5$	
DGP	Estimators	$n = 50$	$n = 150$	$n = 50$	$n = 150$
1	OLS	0.0267	0.0076	0.6762	0.1869
	FGRR	0.0213	0.0071	0.3430	0.1051
	GRRM	0.0369	0.0238	0.1126	0.0454
	GRRJ	0.0369	0.0238	0.1085	0.0450
	AOGRR	0.0228	0.0078	0.2683	0.0850
	SP	0.0173	0.0062	0.3506	0.1292
	$AOSP_1$	0.0150	0.0045	0.2462	0.0762
	AOSP	0.0143	0.0044	0.2329	0.0746
2	OLS	0.0270	0.0075	0.6813	0.1944
	FGRR	0.0222	0.0070	0.3829	0.1096
	GRRM	0.0357	0.0232	0.1079	0.0446
	GRRJ	0.0357	0.0232	0.1027	0.0442
	AOGRR	0.0232	0.0077	0.2698	0.0886
	SP	0.0173	0.0062	0.3458	0.1337
	$AOSP_1$	0.0148	0.0044	0.2427	0.0791
	AOSP	0.0145	0.0044	0.2295	0.0777

for window-width choice habitually improves over the naive and Akaike information criterion (AIC) cross-validation procedures in terms of estimator's risk. However, it should be noted that the optimal window-width for density estimation is not necessarily the optimal window-width for the SP estimator of the regression coefficients. In our simulations, the $AOSP_1$ and AOSP estimates are computed based on the window-widths selected by the two criteria described in Section 4.

The simulation results reported in Table 1 show that, although the SP, $AOSP_1$, 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 the 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 the result of the biasing factors chosen for the FGRR estimator being optimal (see Vinod et al., 1981, p. 363, and Hoerl and Kennard, 1970b, p. 63). Further, we observe that the AOSP estimator usually has a slight edge over the $AOSP_1$ estimator. This can be explained by noting that our comparison is in terms of estimator's risk; the slight disadvantage of the $AOSP_1$ estimator is likely the result of its bandwidth being selected based on the minimization of predictive risk ($\widetilde{R}_1(h)$) instead of the estimator's risk ($\widetilde{R}(h)$) as in the AOSP estimator. If the comparison is in terms of predictive risk, then the reverse will likely be observed.

6. Empirical applications

This section considers two empirical data applications of the proposed methods. In these applications, we use the proposed methods for forecasting excess stock returns and wages.

6.1. Forecasting excess stock returns

The data used in this application are taken from Campbell and Thompson (2008). Lu and Su (2015) and Jin et al. (2014) also used the same data in their studies. The dataset contains $n = 672$ monthly observations of Y, the excess returns of the S&P 500 Index from January 1950 to December 2005. Thus, Y is the difference between the S&P 500 returns and the risk-free rate. Data observations are also available for the following twelve explanatory variables over the same time period, ordered by their magnitudes of correlations with Y: default yield spread (x_1), treasury bill rate (x_2), new equity expansion (x_3), term spread (x_4), dividend price ratio (x_5), earnings price ratio (x_6), long term yield (x_7),

Table 2. Out-of-sample R^2.

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

book-to-market ratio (x_8), inflation (x_9), returns on equity (x_{10}), the one-period lag of excess returns (x_{11}), and smoothed earnings price ratio (x_{12}). Our model average thus contains the following 13 nested models: $\{1\}, \{1, x_1\}, \{1, x_1, x_2\} \ldots, \{1, x_1, x_2, \ldots, x_{12}\}$.

Our estimation is based on $n_1 = 144$, 180, 216, 336, and 456 observations, and we use the remaining $n - n_1$ observations for out-of-sample forecast accuracy assessment purposes. We evaluate forecast accuracy using the following out-of-sample R^2 measure:

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

where $\widehat{Y}_p$ is the prediction of Y_p based on a given forecast method, and $\bar{Y}$ is the average of the values of Y across the n_1 observations. This measure represents the relative difference in squared error predictive risks. Although not considered here, alternative measures based on the squares of the correlation between Y and $\widehat{Y}$, developed by Doksum and Samarov (1995) and Yao and Ullah (2013), can also be used. The out-of-sample R^2 is negative (positive) when $\widehat{Y}$ yields a larger (smaller) sum of squared one-period ahead forecast errors compared with that obtained based on $\bar{Y}$. Table 2 reports the out-of-sample R^2 for six estimators considered in Section 5. We report the results for $AOSP_1$ and not those for AOSP because our evaluation here is in terms of predictive risk—recall that the $AOSP_1$ is based on a minimization of predictive risk, whereas AOSP is derived on the basis of minimizing the estimator's risk. The results show that except when $n_1 = 180$, OLS forecasts are inferior to forecasts based on the historical average. This is consistent with the findings of Welch and Goyal (2008), who used the same data in their study, that the historical mean gives better forecasts when no restrictions are imposed. In all but one case, the FGRR, AOGRR, and $AOSP_1$ estimators are also inferior to the historical average in terms of prediction accuracy. On the other hand, the GRRM and GRRJ model averaging estimators result in positive out-of-sample R^2, and thus are superior to the historial average, in the large majority of cases, with the GRRJ estimator being the slightly superior estimator of the two.

6.2. Forecasting wages

This application example uses a cross-sectional sample of n=526 observations from the U.S. Current Population Survey for the year 1976 given in Wooldridge (2003). The dependent variable is lwage, the logarithm of average hourly earnings. We consider the following ten explanatory variables, ordered according to their magnitudes of correlations with the dependent variable: profocc (=1 if in a professional occupation), educ (=years of education), tenure (=years with current employer), gender (=1 if female), servocc (=1 if in a service occupation), marital status (=1 if married), trade (=1 if employed in wholesale or retail trade), SMSA (=1 if living in a standard metropolitan statistical area), servocc (=1 if in a service occupation), and clerkocc (=1 if in a clerical occupation). Thus, our model average comprises across 11 models nested in the same manner as described in the last data example. Our estimation is based on $n_1 = 100$, 200, 300, and 400 observations, and we use the remaining $n - n_1$ observations for out-of-sample forecast accuracy assessment purposes.

Table 3 reports the out-of-sample R^2 values for the same six estimators as in the last example. The results show that all six estimators yield more accurate forecasts than the historical average, but the GRRJ and GRRM forecasts are inferior to the OLS, FGRR, AOGRR, and $AOSP_1$ forecasts. The latter is the exact

Table 3. Out-of-sample R^2.

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
GRRM	0.3964	0.3366	0.3390	0.3644
GRRJ	0.3877	0.3357	0.3375	0.3627
AOGRR	0.4509	0.4418	0.4642	0.4390
$AOSP_1$	0.4550	0.4477	0.4664	0.4470

opposite to the results observed under the last example, where it is found that the two model averaging estimators generally yield the best forecasts. This is perhaps not surprising given that the variance noise level for the current model is relatively low ($R^2 = 0.509$) compared to the model of the last example ($R^2 = 0.097$)—recall from our simulation findings in Section 5 that the GRRM and GRRJ estimators usually outperform other estimators when there is a large error variance associated with the model, but are outperformed by the nonaveraging estimators when the model is relatively stable in variance. For the current model, of the two model averaging estimators, the GRRM estimator has a slight advantage over the GRRJ estimator.

7. Conclusions

We have proposed a new SP estimator of regression coefficients in the form of the GRR estimator of Hoerl and Kennard (1970b). Unlike the GRR estimator, the biasing factors in our SP estimator can be easily determined by the window-width and the second moment of the kernel function used in density estimation. We have also considered methods of window-width selection based on the Mallows criterion and minimization of the estimator's risk. Moreover, we have shown that the GRR estimator is a model averaging estimator when the regressors are orthogonal, and there is an exact algebraic relationship between the biasing factors of the 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 averaging estimator. This is an interesting finding useful for interpretations and future applications of the SP and GRR estimators. Our Monte Carlo results have shown 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 region of the parameter space.

Acknowledgements

The authors are thankful to Essie Maasoumi, the associate editor, referees, and participants of a seminar in memory of T. D. Dwivedi, Concordia University, Montreal, especially John Galbraith, for helpful comments. Also, they are thankful to Najrin Khanom for her computational help.

Funding

Aman Ullah's work was supported by the Academic Senate, UCR. Alan Wan's work was supported by a General Research Grant from the Hong Kong Research Grants Council (Grant no. 9042086). Xinyu Zhang's and Guohua Zou's work was supported by the National Natural Science Foundation of China (Grant nos. 71522004, 11471324 and 11271355 for Zhang, and Grant no. 11331011 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.
Cheng, M. Y., Hall, P., Titterington, D. (1997). On the shrinkage of local linear curve estimators. *Statistics and Computing* 7:11–17.
Darolles, S., Fan, Y., Florens, J., Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* 79:1541–1565.
Doksum, K., Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Annals of Statistics* 23:1443–1473.
Goldfarb, D., Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In: Hennart, J. P., ed. *Numerical Analysis*. Berlin: Springer-Verlag, pp. 226–239.
Goldfarb, D., Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* 27(1):1–33.
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., 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. (2014). Robustify financial time series forecasting. *Econometric Reviews* 33:575–605.
Leamer, E. E., Chamberlain, G. (1976). A Bayesian interpretation of pretesting. *Journal of the Royal Statistical Society (Series B)* 13(38):85–94.
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. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics*. 188:40–58.
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.
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. *Sankyā (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., 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.
Yao, F., Ullah, A. (2013). A nonparametric R2 test for the presence of the relevant variables. *Journal of Statistical Planning and Inference* 143(9):1527–1547.
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.