Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Optimal Weight Choice for Frequentist Model Average Estimators

Hua LIANG, Guohua ZOU, Alan T. K. WAN, and Xinyu ZHANG

There has been increasing interest recently in model averaging within the frequentist paradigm. The main benefit of model averaging over model selection is that it incorporates rather than ignores the uncertainty inherent in the model selection process. One of the most important, yet challenging, aspects of model averaging is how to optimally combine estimates from different models. In this work, we suggest a procedure of weight choice for frequentist model average estimators that exhibits optimality properties with respect to the estimator's mean squared error (MSE). As a basis for demonstrating our idea, we consider averaging over a sequence of linear regression models. Building on this base, we develop a model weighting mechanism that involves minimizing the trace of an unbiased estimator of the model average estimator's MSE. We further obtain results that reflect the finite sample as well as asymptotic optimality of the proposed mechanism. A Monte Carlo study based on simulated and real data evaluates and compares the finite sample properties of this mechanism with those of existing methods. The extension of the proposed weight selection scheme to general likelihood models is also considered. This article has supplementary material online.

KEY WORDS: Asymptotic optimality; Finite sample property; Mallows criterion; Smoothed AIC; Smoothed BIC; Unbiased MSE estimate.

1. INTRODUCTION

It has been recognized that model selection neglects the uncertainty associated with the selection process, hence inference based on the final model can be seriously misleading (Hjort and Claeskens 2003). Traditional model selection procedures pick the best model that can explain the data at hand according to some model assessment criteria. The investigator then proceeds as if this model has been decided upon a priori. Conditional on the model chosen, statistical inference is typically conducted based on the corresponding conditional distribution of the parameter estimators. Standard errors conventionally estimated under such circumstances are well known to underreport variability (Hjort and Claeskens 2003; Danilov and Magnus 2004b). Model averaging, on the other hand, provides a coherent mechanism for accounting for this model uncertainty through combining parameter estimates across different models. When working with a distribution that is unconditional on the selected model, it incorporates rather than ignores the uncertainty inherent in the model selection process.

Model averaging has long been a popular technique among Bayesian statisticians. Reviews of the relevant Bayesian literature can be found in the works of Hoeting et al. (1999), Raftery, Madigan, and Hoeting (1997). One major criticism of Bayesian model averaging is that the procedure typically involves mixing a large number of priors regarding the unknowns, and it is unclear what the consequences will be when some of the priors are in conflict. Despite a growing frequentist model averaging literature, it would be fair to say that frequentists remain a distinct minority among those who advocate model averaging. However, this imbalance may soon be reconciled with some significant progress made in the frequentist literature in recent years. Buckland, Burnham, and Augustin (1997), for example, proposed a frequentist model weighting method according to values of a model selection criterion; Yang (2001, 2003) developed an adaptive regression by mixing method; Yuan and Yang (2005) further built on this method by proposing a model screening step prior to implementing adaptive regression by mixing; Hjort and Claeskens (2003) established a local misspecification framework for studying properties of post-selection and model average estimators; and Leung and Barron (2006) discussed a mixture least squares estimator with weights depending on the risk characteristics of the mixture estimator. The recent monograph of Claeskens and Hjort (2008) provided a useful summary of some of the progress that has been made in this area.

In a recent article, Hansen (2007) proposed a frequentist model average estimator with weights obtained by a minimization of the Mallows criterion (see also Shen and Huang 2006, who proposed a similar criterion). The justification of this method lies in the fact that the Mallows criterion is asymptotically equivalent to the squared error, so the model average estimator that minimizes the Mallows criterion also minimizes the squared error in large samples. Hansen's simulation results showed that the Mallows model average (MMA) estimator generally outperforms model average estimators based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) weights when the measures of variability obtained by using these estimators are compared in terms of the average squared error loss for the conditional mean prediction of the dependent variable.

Hansen's (2007) approach marks a significant step toward the development of optimal weight choice in the frequentist model

Hua Liang is Professor, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642 (E-mail: *hliang@bst.rochester.edu*). Guohua Zou is Professor (E-mail: *ghzou@amss.ac.cn*) and Xinyu Zhang is Assistant Professor (E-mail: *xinyu@amss.ac.cn*), Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. Alan T. K. Wan is Professor, Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong (E-mail: *msawan@cityu.edu.hk*). The authors thank the editor, the associate editor, and two referees for careful review of the manuscript and valuable suggestions. Thanks are also extended to Emmanuel Guerre, Hannes Leeb, Jan Magnus, Benedikt Pötcher, and Yuhong Yang for helpful comments. The first, second, and third authors' work was supported by NIH/NIADID grant AI59773 and NSF grants DMS 0806097 and DMS-1007167, National Natural Science Foundation of China grants 70625004 and 10721101, and Hong Kong Research Grant Council GRF 102709, respectively.

Journal of the American Statistical Association
Accepted for publication, Theory and Methods
DOI: 10.1198/jasa.2011.tm09478

average estimator. Now, let y be an $n \times 1$ vector, H be an $n \times P$ matrix of full column rank, H_p be an $n \times p$ ($\leq P$) matrix comprising the first p columns of H, and $(\omega_1, \omega_2, \ldots, \omega_P)'$ be a weight vector. The model average estimator considered by Hansen is of the form

$$\hat{\Theta}_{\text{mma}} = \sum_{p=1}^{P} \omega_p \begin{pmatrix} (H_p'H_p)^{-1}H_p'y \\ 0 \end{pmatrix}. \tag{1}$$

It is readily seen that $\hat{\Theta}_{\text{mma}}$ is the weighted sum of least squares estimators from a sequence of P strictly nested regression models, where the pth model uses the first p variables in H as regressors. One fundamental requirement on $\hat{\Theta}_{\text{mma}}$, therefore, is that the regressors be ordered prior to estimation. Asymptotic results on the MMA estimator developed by Hansen (2007) rely crucially on this assumption, which poses a strong limitation to his approach.

This article proposes a new method of weight choice for frequentist model average (FMA) estimators. As a basis for demonstrating our idea and to facilitate comparisons with existing FMA estimators, we adopt the linear regression model as our main analytical framework, although as we shall see, extensions to a more general likelihood framework are also feasible. Our setup assumes that the underlying model contains a set of focus regressors, whose inclusion in the model is mandatory on theoretical or other grounds irrespective of statistical significance, and a set of auxiliary regressors whose inclusion is viewed as optional. The model containing the focus regressors only is referred to as the narrow model; the extended models are those that contain the focus regressors and possibly some or all of the auxiliary regressors. This is the same setup used in a number of recent articles on pretesting (e.g., Magnus and Durbin 1999; Danilov and Magnus 2004a, 2004b), and it bears similarity to the local misspecification setup of Hjort and Claeskens (2003), which also distinguishes between narrow and extended models. The mandatory inclusion of focus regressors does *not* lead to any loss of generality since the focus regressors can in practice contain an intercept term only. Our approach to model weight selection is based on the mean squared error (MSE) properties of the combined estimator. Specifically, we derive an exact unbiased estimator of the MSE of the model average estimator, and propose selecting the model weights that minimize the trace of the MSE estimate. Unlike that of Hansen (2007), our approach does not require the regressors to be ordered, and in contrast to most previously proposed weight selection schemes, our criterion is based on analytical finite sample justifications. Our approach is similar in spirit to that advocated by Leung and Barron (2006), except that they focused on the risk bound of the combined estimator, whereas we provide an explicit weight choice criterion together with an analysis of the asymptotic and finite sample properties of the FMA estimator that results from the proposed weight choice method. Weighting schemes based on smoothed AIC (S-AIC) and smoothed BIC (S-BIC) (Buckland, Burnham, and Augustin 1997) are special cases of our proposed method. Our simulation results show that the estimator arising from the proposed weight selection method, which we label OPT estimator, frequently achieves smaller risk in terms of squared error loss than Hansen's (2007) MMA estimator and model average estimators based on S-AIC and S-BIC weights. While the bulk of our analysis focuses on the linear regression model, we also show that a similar weight choice mechanism may be crafted under a more general likelihood framework.

The presentation of this article goes as follows. Section 2 describes the model and estimators. In Section 3, we derive unbiased estimators of the finite sample MSEs of the FMA estimators, along with an investigation of the finite sample and asymptotic properties of the proposed criterion. Section 4 reports results of a Monte Carlo study based on simulated as well as real data. Section 5 discusses the generalization of our method to general parametric models, and our conclusions are presented in Section 6. Proofs of results are contained in the Appendix.

2. MODEL SETUP AND ESTIMATORS

Consider the linear regression model

$$y = X\beta + Z\gamma + \varepsilon, \qquad \varepsilon \sim \text{N}(0, \sigma^2 I_n), \tag{2}$$

where y ($n \times 1$) is a vector of observations, X ($n \times k$) and Z ($n \times m$) are nonrandom regressor matrices, and β ($k \times 1$) and γ ($m \times 1$) are parameter vectors. Additionally, we assume that $H \equiv (X : Z)$ has full column rank $k + m$. Here, X contains the focus regressors that must be included in the model, while Z contains the auxiliary or doubtful regressors whose inclusion in the model is optional. There is no loss of generality with this setup because X can contain no regressor other than an intercept term, or even be an empty matrix.

With m auxiliary regressors in Z, there are a maximum of 2^m extended models to choose between. Let N be the number of extended models embodied in the model selection/averaging process. If all extended models are considered, then $N = 2^m$. The case where only the full and narrow models are relevant corresponds to $N = 2$, and if we consider the nested setup as in the work of Hansen (2007), then $N = m + 1$. Traditional model selection procedures pick the best model that can explain the data at hand based on some model assessment criteria, while model averaging combines estimates obtained from different models. Although one can argue that model selection may be more attractive than model averaging when the true model is a candidate model, this should not be used as a criticism against the basic principle underlying model averaging. The reason is that there is always the possibility of selecting the wrong or even a very poor model. With model averaging, it is not necessary to assume that the true model is among the set of considered models, while such is not the case in typical model selection studies.

Under the above setup, the fully restricted (i.e., $\gamma = 0$) and unrestricted least squares estimators of β are $\hat{\beta}_r = (X'X)^{-1}X'y$ and $\hat{\beta}_u = \hat{\beta}_r - (X'X)^{-1}X'Z(Z'MZ)^{-1}Z'My$, respectively, where $M = I_n - X(X'X)^{-1}X'$, and the unrestricted least squares estimator of γ is $\hat{\gamma} = (Z'MZ)^{-1}Z'My$. Now, let $\theta = (Z'MZ)^{1/2}\gamma$ and $\hat{\theta} = (Z'MZ)^{1/2}\hat{\gamma}$. Note that $\hat{\theta} \sim \text{N}(\theta, \sigma^2 I_m)$; in particular, $\hat{\theta}$ has a covariance matrix that is a scalar multiple of an identity matrix. Hence, it is of analytical convenience to write $\hat{\beta}_u = \hat{\beta}_r - Q\hat{\theta}$, where $Q = (X'X)^{-1}X'Z(Z'MZ)^{-1/2}$. In conformity, the ith ($1 \leq i \leq 2^m$) partially restricted least squares estimator of β can be written as $\hat{\beta}_{(i)} = \hat{\beta}_r - QW_i\hat{\theta}$, where $W_i = I_m - P_i$, $P_i = (Z'MZ)^{-1/2}S_i\{S_i'(Z'MZ)^{-1}S_i\}^{-1}S_i'(Z'MZ)^{-1/2}$ is

an $m \times m$ symmetric idempotent matrix of rank $r_i \geq 0$, and S_i is an $m \times r_i$ selection matrix of rank r_i so that $S_i' = (I_{r_i} : 0)$ or a column permutation thereof (Danilov and Magnus 2004b). The ith partially restricted least squares estimator of γ is $\hat{\gamma}_{(i)} = (Z'MZ)^{-1/2} W_i \hat{\theta}$ under the restriction $S_i'\gamma = 0$. Further, let $\hat{\sigma}^2 = \|y - X\hat{\beta}_u - Z\hat{\gamma}\|^2/(n-k-m)$ be the estimator of σ^2 under the unrestricted model. Then, an FMA estimator of β in model (2) may be written as

$$\hat{\beta}_f = \sum_{i=1}^{N} \lambda_i \hat{\beta}_{(i)}, \tag{3}$$

with weights satisfying $\lambda_i \geq 0$ and $\sum_{i=1}^{N} \lambda_i = 1$.

Here, we concentrate on random weights. Specifically, we let λ_i depend on $\hat{\theta}$ and $\hat{\sigma}^2$. This consideration is motivated by the following observation. Let q_i be the number of regressors in the ith partially restricted model; then the AIC of the ith model is $\mathrm{AIC}(i) = n\log(\hat{\sigma}_i^2) + 2(q_i + 1)$, where $\hat{\sigma}_i^2 = \|y - X\hat{\beta}_{(i)} - Z\hat{\gamma}_{(i)}\|^2/n$ is the maximum likelihood estimator of σ^2 under the ith model. Note that we can write $\hat{\sigma}_i^2 = (n-k-m)\hat{\sigma}^2/n + \hat{\theta}' P_i \hat{\theta}/n$. Putting the latter expression of $\hat{\sigma}_i^2$ in the AIC expression of the ith model, we observe that the AIC depends on the data only through $\hat{\theta}$ and $\hat{\sigma}^2$. Motivated by this, we consider weights $\lambda_i = \lambda_i(\hat{\theta}, \hat{\sigma}^2)$ that only depend on $\hat{\theta}$ and $\hat{\sigma}^2$. Writing $W = \sum_{i=1}^{N} \lambda_i W_i$, we can rewrite the FMA estimator as $\hat{\beta}_f = \hat{\beta}_r - QW\hat{\theta}$.

Note that the model considered by Hansen (2007) makes no distinction between focus and auxiliary regressors, and may be stated using our notations as $y = H\Theta + u$, where $\Theta = (\beta', \gamma')'$ is a vector of coefficients, $u = B\upsilon + \varepsilon$, B is a set of omitted regressors, and υ is the corresponding coefficient vector. Hansen (2007) concentrated on the estimation of $\mu^* = H\Theta + B\upsilon$, and evaluated the performance of $\hat{\Theta}_{\mathrm{mma}}$, the MMA estimator of Θ, in terms of the criterion $R^* = E\|H\hat{\Theta}_{\mathrm{mma}} - \mu^*\|^2$, which is the risk of the estimator of the prediction vector.

3. UNBIASED ESTIMATION OF MSE AND OPTIMAL WEIGHT CHOICE

In this section we consider the choice of weights in (3). Our weight choice method is based on an MSE minimizing estimation objective that is designed to provide MSE improvements over other FMA estimators, especially in finite samples. Here, the MSE of $\hat{\beta}_f$ is defined as the matrix $E\{(\hat{\beta}_f - \beta)(\hat{\beta}_f - \beta)'\}$, with its diagonal elements being the MSEs of the estimators for the individual components of β. The trace of the MSE matrix is equal to the expected squared error loss function $E\|\hat{\beta}_f - \beta\|^2$, known also as the risk of the estimator under squared error loss, or the weak MSE accuracy measure (Wallace 1972).

To make our concept operational, we first derive an unbiased estimator of the MSE of $\hat{\beta}_f$. The optimal model average weights are then obtained by minimizing the trace of the MSE estimate. With appropriate modifications we also generalize the method to the derivation of the MSE estimator of the predictor $\hat{\mu}_f = H\hat{\Theta}_f$, where $\hat{\Theta}_f = (\hat{\beta}_f', \hat{\gamma}_f')'$, and $\hat{\gamma}_f = \sum_{i=1}^{N} \lambda_i(\hat{\theta}, \hat{\sigma}^2)\hat{\gamma}_{(i)}$ is the FMA estimator of γ corresponding to $\hat{\beta}_f$.

3.1 An Unbiased Estimator of the MSE of $\hat{\beta}_f$

Our principal result is stated in the following theorem:

Theorem 1. Under model (2), assuming that $\lambda_i(\hat{\theta}, \hat{\sigma}^2)$, $i = 1, \ldots, N$, are continuous functions with piecewise continuous partial derivatives with respect to $\hat{\theta}$, and provided that the expectations of $(\partial\lambda_i(\hat{\theta}, \hat{\sigma}^2)/\partial\hat{\theta})\hat{\theta}'$ and Ψ (defined below) exist, an unbiased estimator of the MSE of the FMA estimator $\hat{\beta}_f$ is given by

$$\widehat{\mathrm{MSE}}(\hat{\beta}_f) = \hat{\sigma}^2(X'X)^{-1} - \hat{\sigma}^2 QQ' + \{Q(I_m - W)\hat{\theta}\}^{\otimes 2} + \Psi(\hat{\theta}, \hat{\sigma}^2) + \{\Psi(\hat{\theta}, \hat{\sigma}^2)\}', \tag{4}$$

where $A^{\otimes 2} = AA'$ for any vector or matrix A, and

$$\Psi(\hat{\theta}, \hat{\sigma}^2) = \{(n-k-m)/2\}(\hat{\sigma}^2)^{-(n-k-m)/2+1} \times \int_0^{\hat{\sigma}^2} t^{(n-k-m)/2-1} \Psi_1(\hat{\theta}, t)\, dt, \tag{5}$$

with

$$\Psi_1(\hat{\theta}, t) = Q\left\{W + \sum_{i=1}^{N} (\partial\lambda_i(\hat{\theta}, t)/\partial\hat{\theta})\hat{\theta}' W_i\right\} Q'. \tag{6}$$

Proof. See the Appendix.

The unbiased estimator of the MSE of $\hat{\beta}_f$ provides a basis for measuring the estimator's precision that can be justified in finite samples. The goal of our criterion is to choose λ_i's in (3) that minimize the trace of $\widehat{\mathrm{MSE}}(\hat{\beta}_f)$. Now, from (4), it is straightforward to show that the trace of $\widehat{\mathrm{MSE}}(\hat{\beta}_f)$ is

$$\widehat{R}(\hat{\beta}_f) = \hat{\sigma}^2 \operatorname{tr}(X'X)^{-1} - \hat{\sigma}^2 \operatorname{tr}(QQ') + \{\hat{\theta}'(I_m - W)Q'\}^{\otimes 2} + 2\operatorname{tr}\{\Psi(\hat{\theta}, \hat{\sigma}^2)\}. \tag{7}$$

However, the practical application of (7) is limited to some degree by the complexity of the term $\Psi(\hat{\theta}, \hat{\sigma}^2)$, which is cumbersome to calculate. To get around this problem we replace $\Psi(\hat{\theta}, \hat{\sigma}^2)$ by an approximate quantity. Note that under the conditions of Theorem 1, $E_{\hat{\sigma}^2}\{\Psi(\hat{\theta}, \hat{\sigma}^2)\} = \sigma^2 E_{\hat{\sigma}^2}\{\Psi_1(\hat{\theta}, \hat{\sigma}^2)\}$, where $E_{\hat{\sigma}^2}$ denotes expectation only with respect to $\hat{\sigma}^2$ [see (A.4) and its proof in the Appendix for details]. Thus, it appears reasonable to replace $\Psi(\hat{\theta}, \hat{\sigma}^2)$ by $\hat{\sigma}^2\Psi_1(\hat{\theta}, \hat{\sigma}^2)$ in (7). This results in the following approximate risk quantity:

$$\widehat{R}_a(\hat{\beta}_f) = \hat{\sigma}^2 \operatorname{tr}(X'X)^{-1} - \hat{\sigma}^2 \operatorname{tr}(QQ') + \{\hat{\theta}'(I_m - W)Q'\}^{\otimes 2} + 2\hat{\sigma}^2 \operatorname{tr}\{\Psi_1(\hat{\theta}, \hat{\sigma}^2)\}. \tag{8}$$

The online supplementary material provides results which show that the values produced by $\hat{\sigma}^2\Psi_1(\hat{\theta}, \hat{\sigma}^2)$ typically accord with those of $\Psi(\hat{\theta}, \hat{\sigma}^2)$ very closely.

We note that if the FMA estimator $\hat{\beta}_f$ is based on the S-AIC weights, then $\lambda_i = e^{-q_i}(\hat{\sigma}_i^2)^{-n/2}/\sum_{j=1}^{N} e^{-q_j}(\hat{\sigma}_j^2)^{-n/2}$. If the S-BIC weights are used, then $\lambda_i = n^{-q_i/2}(\hat{\sigma}_i^2)^{-n/2}/\sum_{j=1}^{N} n^{-q_j/2} \times (\hat{\sigma}_j^2)^{-n/2}$. Also, if one uses the smoothed residual mean squares (S-RMS) weights (Bates and Granger 1969), then $\lambda_i = (n-q_i)(\hat{\sigma}_i^2)^{-1}/\sum_{j=1}^{N}(n-q_j)(\hat{\sigma}_j^2)^{-1}$. A natural generalization of these weights is found in the following class of random weights:

$$\lambda_i(\hat{\theta}, \hat{\sigma}^2) = \frac{a^{q_i}(n-q_i)^b(\hat{\sigma}_i^2)^c}{\sum_{j=1}^{N} a^{q_j}(n-q_j)^b(\hat{\sigma}_j^2)^c}, \tag{9}$$

where a (> 0), b (≥ 0), and c (≤ 0) are constants. The S-AIC weights are obtained by setting $a = e^{-1}$, $b = 0$, and $c = -n/2$; the S-BIC weights result when $a = n^{-1/2}$, $b = 0$, and $c = -n/2$; and when $a = b = 1$ and $c = -1$, $\lambda_i(\hat{\theta}, \hat{\sigma}^2)$ reduces to the S-RMS weights. Further, by setting $a = 1, b = 2$, and $c = -1$, we obtain weights that correspond to the smoothed generalized cross-validation of Craven and Wahba (1979), which is almost identical to the average prediction MSE criterion due apparently to J. M. Tukey (see Wetherill et al. 1986, p. 243; and Leeb and Pötscher 2008, p. 898). With slight modifications, weights corresponding to the smoothed version of Hurvich and Tsai's (1989) bias-corrected AIC can also be written as a special case of (9).

Now, define $L = (l_{ij})$ and $G = (g_{ij})$, where $l_{ij} = \hat{\theta}'(I_m - W_i)Q'Q(I_m - W_j)\hat{\theta}$ and $g_{ij} = (\hat{\sigma}_j^2)^{-1}\hat{\theta}'W_iQ'Q(I_m - W_j)\hat{\theta}$ for $i, j = 1, \ldots, N$. Additionally, let g and ϕ each be an $N \times 1$ vector with g consisting of the diagonal elements of G and the ith element of ϕ being $\operatorname{tr}(QW_iQ')$, $i = 1, \ldots, N$. Recognizing that

$$\partial\lambda_i(\hat{\theta}, \hat{\sigma}^2)/\partial\hat{\theta} = (2/n)c\lambda_i(\hat{\theta}, \hat{\sigma}^2)\Bigg\{(\hat{\sigma}_i^2)^{-1}(I_m - W_i) - \sum_{j=1}^{N}\lambda_j(\hat{\theta}, \hat{\sigma}^2)(\hat{\sigma}_j^2)^{-1}(I_m - W_j)\Bigg\}\hat{\theta}, \quad (10)$$

putting (10) in $\Psi_1(\hat{\theta}, \hat{\sigma}^2)$ given by (6) and using (8), we have

$$\begin{aligned}\widehat{R}_a(\hat{\beta}_f) &= \hat{\sigma}^2\operatorname{tr}(X'X)^{-1} - \hat{\sigma}^2\operatorname{tr}(QQ') + \lambda'(a,b,c)L\lambda(a,b,c)\\ &\quad - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)G\lambda(a,b,c) + 2\hat{\sigma}^2\lambda'(a,b,c)\phi\\ &\quad + (4/n)c\hat{\sigma}^2\lambda'(a,b,c)g, \end{aligned} \quad (11)$$

where $\lambda(a, b, c)$ is an $N \times 1$ vector comprising $\lambda_i(\hat{\theta}, \hat{\sigma}^2)$, $i = 1, \ldots, N$. We use a, b, and c, defined in Equation (9), as arguments of λ to emphasize the role of these parameters in the optimal weight choice. The primary goal is to select the appropriate values of a, b, and c in $\lambda(a, b, c)$ that minimize (11). Let $\lambda(a^*, b^*, c^*)$ be such a vector. We call the estimator $\hat{\beta}_f$ corresponding to $\lambda(a^*, b^*, c^*)$ the optimal FMA estimator (labeled as OPT estimator hereafter).

Before proceeding further we want to draw readers' attention to the following special case. Suppose we set $c = 0$ in (11); then the weights λ_i's in (9) reduce to deterministic weights. For this special case, if we consider mixing only $\hat{\beta}_u$ and $\hat{\beta}_r$ (i.e., all the regressors in Z are either in or out), then minimizing (11) with respect to (a, b) leads to the estimator

$$\hat{\beta}_{js} = \left\{1 - \frac{\hat{\sigma}^2\operatorname{tr}(Q'Q)}{\|\hat{\beta}_u - \hat{\beta}_r\|^2}\right\}\hat{\beta}_u + \frac{\hat{\sigma}^2\operatorname{tr}(Q'Q)}{\|\hat{\beta}_u - \hat{\beta}_r\|^2}\hat{\beta}_r, \quad (12)$$

which turns out to be the James–Stein type estimator studied by Kim and White (2001). In view of the fact that it is obtained from minimizing (11), $\hat{\beta}_{js}$ is also an optimal estimator by our criterion if one restricts attention to the subspace of $c = 0$. Clearly, the OPT estimator that minimizes (11) (regardless of the value of c) has optimal properties among a broader class of estimators. In this sense, $\hat{\beta}_{js}$ is suboptimal with respect to our criterion when compared to the OPT estimator.

3.2 An Unbiased Estimator of the MSE of $\hat{\mu}_f$

With some efforts the preceding framework of finding optimal weights may be generalized to encompass the estimation of the vector $\mu = H\Theta$, which is the conditional mean prediction of the dependent variable. Denote the FMA estimator of μ as

$$\hat{\mu}_f = H\hat{\Theta}_f = H\sum_{i=1}^{N}\lambda_i(\hat{\theta}, \hat{\sigma}^2)\hat{\Theta}_{(i)} \equiv \sum_{i=1}^{N}\lambda_i(\hat{\theta}, \hat{\sigma}^2)\hat{\mu}_{(i)}, \quad (13)$$

where $\hat{\mu}_{(i)} = H\hat{\Theta}_{(i)}$, and $\hat{\Theta}_{(i)}$ is the ith partially restricted least squares estimator of Θ. Note that when there is no mandatory regressor, that is, $k = 0$, β does not exist, and thus there is no FMA estimator of β; on the other hand, the average of $\hat{\mu}_{(i)}$ exists whether or not there are focus regressors in the model. In this case, $\hat{\Theta}_{(i)}$ reduces to $\hat{\gamma}_{(i)}$, and $\hat{\theta} = (Z'Z)^{-1/2}Z'y$.

Theorem 2. Under model (2) and the same conditions as in Theorem 1, an unbiased estimator of the MSE of $\hat{\mu}_f$ is given by

$$\begin{aligned}&\widehat{\operatorname{MSE}}(\hat{\mu}_f)\\ &\quad = \hat{\sigma}^2X(X'X)^{-1}X' + \varphi(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\\ &\qquad - \varphi\big(\hat{\theta}, \hat{\sigma}^2, XQ, \{Z(Z'MZ)^{-1/2}\}'\big)\\ &\qquad - \varphi\big(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, (XQ)'\big)\\ &\qquad + \varphi\big(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, \{Z(Z'MZ)^{-1/2}\}'\big), \end{aligned} \quad (14)$$

where

$$\begin{aligned}\varphi(\hat{\theta}, \hat{\sigma}^2, D_1, D_2) &= -\hat{\sigma}^2D_1D_2 + D_1\{(I_m - W)\hat{\theta}\}^{\otimes 2}D_2\\ &\quad + D_1\Xi(\hat{\theta}, \hat{\sigma}^2)D_2\\ &\quad + D_1\{\Xi(\hat{\theta}, \hat{\sigma}^2)\}'D_2, \end{aligned} \quad (15)$$

$$\Xi(\hat{\theta}, \hat{\sigma}^2) = \{(n - k - m)/2\}(\hat{\sigma}^2)^{-(n-k-m)/2+1} \times \int_0^{\hat{\sigma}^2} t^{(n-k-m)/2-1}\Xi_1(\hat{\theta}, t)\,dt, \quad (16)$$

and

$$\Xi_1(\hat{\theta}, t) = W + \sum_{i=1}^{N}\frac{\partial\lambda_i(\hat{\theta}, t)}{\partial\hat{\theta}}\hat{\theta}'W_i. \quad (17)$$

Proof. See the Appendix.

The trace of $\widehat{\operatorname{MSE}}(\hat{\mu}_f)$ is

$$\begin{aligned}&\hat{R}(\hat{\mu}_f)\\ &\quad = k\hat{\sigma}^2 + \operatorname{tr}\big\{\varphi(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\big\}\\ &\qquad - 2\operatorname{tr}\big\{\varphi\big(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, (XQ)'\big)\big\}\\ &\qquad + \operatorname{tr}\big[\varphi\big(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, \{Z(Z'MZ)^{-1/2}\}'\big)\big]. \end{aligned} \quad (18)$$

Note that all terms except the last in (14) and (18) vanish when no regressor is considered mandatory.

To overcome the computational difficulties associated with (18), and noting that $E_{\hat{\sigma}^2}\{\Xi(\hat{\theta}, \hat{\sigma}^2)\} = \sigma^2E_{\hat{\sigma}^2}\{\Xi_1(\hat{\theta}, \hat{\sigma}^2)\}$ under the conditions of Theorem 1, we replace $\Xi(\hat{\theta}, \hat{\sigma}^2)$ by $\hat{\sigma}^2\Xi_1(\hat{\theta}, \hat{\sigma}^2)$ in (15). This is analogous to the substitution of $\Psi(\hat{\theta}, \hat{\sigma}^2)$ by $\hat{\sigma}^2\Psi_1(\hat{\theta}, \hat{\sigma}^2)$ in (8). Following this substitution we

obtain the following approximately unbiased estimator of the trace of the MSE of $\hat{\mu}_f$:

$$\begin{aligned}\hat{R}_a(\hat{\mu}_f) &= k\hat{\sigma}^2 + \operatorname{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\} \\ &\quad - 2\operatorname{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, (XQ)')\} \\ &\quad + \operatorname{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, Z(Z'MZ)^{-1/2}, (Z(Z'MZ)^{-1/2})')\}, \quad (19)\end{aligned}$$

where $\varphi_1(\hat{\theta}, \hat{\sigma}^2, D_1, D_2)$ has the same expression as $\varphi(\hat{\theta}, \hat{\sigma}^2, D_1, D_2)$, except that $\Xi(\hat{\theta}, \hat{\sigma}^2)$ in $\varphi(\cdot)$ is replaced everywhere by $\hat{\sigma}^2\Xi_1(\hat{\theta}, \hat{\sigma}^2)$.

Again, let the weight $\lambda_i(\hat{\theta}, \hat{\sigma}^2)$ be of the form (9). Note that P_j is symmetric idempotent and $\{XQ - Z(Z'MZ)^{-1/2}\}'\{XQ - Z(Z'MZ)^{-1/2}\} = I_m$. Putting (10) in (19), and after performing some manipulations, we obtain

$$\begin{aligned}\hat{R}_a(\hat{\mu}_f) &= (k-m)\hat{\sigma}^2 + \lambda'(a,b,c)\bar{L}\lambda(a,b,c) + 2\hat{\sigma}^2\lambda'(a,b,c)\bar{\phi} \\ &\quad - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)\bar{G}\lambda(a,b,c), \quad (20)\end{aligned}$$

where $\bar{L} = (\bar{l}_{ij})$ with $\bar{l}_{ij} = \hat{\theta}'P_iP_j\hat{\theta}$, $\bar{G} = (\bar{g}_{ij})$ with $\bar{g}_{ij} = (\hat{\sigma}_j^2)^{-1}\hat{\theta}'(P_j - P_iP_j)\hat{\theta}$, and $\bar{\phi}$ is an $N \times 1$ vector with $\bar{\phi}_i = m - r_i$ being its ith element. Equation (20) provides an alternative selection criterion for the weight vector $\lambda(a,b,c)$ when the prediction vector rather than the coefficient vector is the main subject of interest. The optimal weight vector is the λ vector that minimizes (20). We call the estimator $\hat{\mu}_f$ corresponding to this optimal weight choice the OPT estimator of μ.

3.3 Asymptotic Optimality of the OPT Estimator

Our foregoing discussion centers on the finite sample justification of the OPT estimators. Here, we enlarge the optimality consideration to large sample situations. All the convergence results to be presented are with respect to the sample size approaching infinity. For specification, we consider the estimation of the prediction vector μ. Define $L_n(\lambda(a,b,c)) = \|\hat{\mu}_f(\lambda(a,b,c)) - \mu\|^2$—the squared error loss, and $R_n(\lambda(a,b,c)) = E\{L_n(\lambda(a,b,c))\}$—the risk under the squared error loss. Denote $\mathcal{D} = \{(a,b,c)|a > 0, b \geq 0, -\bar{c} \leq c \leq 0\}$, where $\bar{c}$ is a positive constant. Let the set $\mathcal{U}$ index the "unbiased" models, which are models that nest the true model as a special case. Further, we consider the following subset of $\mathcal{D}$:

$$\mathcal{D}_0 = \left\{(a,b,c) \in \mathcal{D} \Big| \sum_{\tau\in\mathcal{U}} \lambda_\tau(a,b,c) \leq 1 - \rho\right\},$$

where ρ is a constant on the interval $(0, 1]$. The exclusion of $\rho = 0$ from $\mathcal{D}_0$ means that the sum of weights assigned to biased models in the model average has a nonzero lower bound, and this rules out the case where all models forming the model average are unbiased. The reason for excluding $\rho = 0$ is that for the case of $\mathcal{U} \neq \emptyset$, condition (22) (defined below), a key condition for the asymptotic optimality result of (23) to hold, cannot hold true when $\rho = 0$. The restriction of $\mathcal{D}$ to $\mathcal{D}_0$ is therefore a technical artifact of our proof technique. Fortunately, the loss of generality due to this restriction is fairly minimal. It is conceivably far more common for a model average to comprise some biased models than unbiased models alone. The latter model combining case is atypical, and will arise when the true model is the narrow model. Moreover, it is obvious that if $\rho = 0$ were allowed, then $\mathcal{D}_0$ would be identical to $\mathcal{D}$; now, as ρ can take on any positive value close to zero, the set $\mathcal{D}_0$ is in fact very close to $\mathcal{D}$.

Building on this framework, we define a nonrandom weight set

$$\mathcal{W} = \left\{w = (w_1, \ldots, w_N)' \Big| w_i \geq 0, \sum_{i=1}^N w_i = 1, \sum_{\tau\in\mathcal{U}} w_\tau \leq 1 - \rho\right\}.$$

Let $\zeta_n = \max_{1\leq i\leq N} E\|\hat{\mu}_{(i)} - \mu\|^2$ be the maximum risk based on a single submodel, and $\xi_n = \inf_{w\in\mathcal{W}} R_n(w)$. Denote $(\hat{a}, \hat{b}, \hat{c})$ as the value of (a,b,c) that minimizes $\hat{R}_a(\hat{\mu}_f(\lambda(a,b,c)))$ with $(a,b,c) \in \mathcal{D}_0$. Assume also that $(\hat{a}, \hat{b}, \hat{c})$ belongs in $\mathcal{D}_0$.

Theorem 3. When $n \to \infty$, provided that the conditions

$$\mu'\mu = O(n) \quad (21)$$

and

$$\xi_n^{-2}\zeta_n \to 0 \quad (22)$$

are satisfied, then

$$\frac{L_n(\lambda(\hat{a}, \hat{b}, \hat{c}))}{\inf_{(a,b,c)\in\mathcal{D}_0} L_n(\lambda(a,b,c))} \xrightarrow{p} 1. \quad (23)$$

Proof. See the Appendix.

Theorem 3 states that subject to the fulfillment of conditions (21) and (22), the large sample squared error associated with the OPT estimator converges in probability to the smallest achievable squared error of any FMA estimator in the form of (13) based on model weights given in (9), with values of (a,b,c) restricted to the subset $\mathcal{D}_0$.

The following discusses the relevance of the subset $\mathcal{D}_0$ and conditions (21) and (22). First, note that condition (21) is a common condition concerning the sum of $\mu_j^2, j = 1, \ldots, n$ (e.g., Shao 1997). Under (21), we have

$$\begin{aligned}\hat{\sigma}_i^2 &= (y'My - \hat{\theta}'W_i\hat{\theta})/n \leq y'y/n \\ &= (\mu'\mu + \varepsilon'\varepsilon + 2\mu'\varepsilon)/n = O_P(1) \quad (24)\end{aligned}$$

for any submodel i. Discounting the cases of $\mathcal{U}$ being an empty set and $\mathcal{U} = \{1, \ldots, N\}$, by the result that for any $\tau \in \mathcal{U}$, $\hat{\sigma}_\tau^2 \xrightarrow{p} \sigma^2 > 0$, we have, for any submodel $\tau \in \mathcal{U}$ and $i \notin \mathcal{U}$,

$$\hat{\sigma}_i^2/\hat{\sigma}_\tau^2 = O_P(1). \quad (25)$$

Combining (9), (25) and the restriction of $c \leq 0$, it can be seen that for any submodels $\tau \in \mathcal{U}$, $i \notin \mathcal{U}$, and any $(a,b,c) \in \mathcal{D}$, we have $\lambda_\tau(a,b,c)/\lambda_i(a,b,c) = O_P(1)$, and thus

$$\begin{aligned}&\sum_{\tau\in\mathcal{U}} \lambda_\tau(a,b,c) \Big/ \left(1 - \sum_{\tau\in\mathcal{U}} \lambda_\tau(a,b,c)\right) \\ &\quad \leq \sum_{\tau\in\mathcal{U}} \lambda_\tau(a,b,c)/\lambda_i(a,b,c) \\ &\quad = O_P(1).\end{aligned}$$

Consequently, $\sum_{\tau\in\mathcal{U}}\lambda_\tau(a,b,c)$ cannot tend to 1 in probability. This means that in large samples, the model weighting scheme stipulated by (9) cannot give rise to zero weight for all biased models. In other words, this weighting scheme implies that even with large samples, at least one of the models that form the FMA estimator must be a biased model. This is why the subset $\mathcal{D}_0$ for which $\sum_{\tau\in\mathcal{U}}\lambda_\tau(a,b,c)\le 1-\rho$ is relevant. Obviously, when ρ is very small, $\mathcal{D}_0$ is very close to $\mathcal{D}$.

As an aside, it is of interest to mention that if there exist constants κ_1 and κ_2 such that $0<\kappa_1\le\kappa_2<\infty$, and $\kappa_1\le\hat{\sigma}_i^2/\sigma^2\le\kappa_2$ with probability 1 for any $i\in\{1,\ldots,N\}$ (see, e.g., Yang 2003 and Yuan and Yang 2005 for similar conditions), and constants $\bar{a}_1$, $\bar{a}_2$, and $\bar{b}$ such that $0<\bar{a}_1\le a\le\bar{a}_2<\infty$, $0\le b\le\bar{b}<\infty$, then $\sum_{\tau\in\mathcal{U}}\lambda_\tau(a,b,c)\le 1-\rho$ is always true with probability 1, provided that $\mathcal{U}$ is not the set $\{1,\ldots,N\}$. The proof is available from the online supplementary material. It is instructive to note that $\sum_{\tau\in\mathcal{U}}\lambda_\tau(a,b,c)\le 1-\rho$ automatically holds for the case of $\mathcal{U}$ being empty.

The validity of Theorem 3 also depends on condition (22). A necessary condition for (22) to hold is

$$\xi_n\to\infty. \tag{26}$$

Condition (26) is in fact very mild, and thus likely to be fulfilled in practice because $\mathcal{W}$ does not include any weight vector that assigns nonzero weights only to the unbiased models. Now, if $\xi_n\to\infty$ holds, then $\xi_n^{-2}\zeta_n\to 0$ also holds as long as ζ_n tends to infinity at a rate slower than that of ξ_n^2 to infinity. One can expect the rate of $\zeta_n\to\infty$ to reduce if some of the very poor models that are associated with large risks are removed at the outset. Thus, it seems desirable to combine over an optimal subset rather than the full set of models. The model screening step developed by Yuan and Yang (2005) may be useful in this regard.

The following simple example sheds further light on condition (22). Consider Equation (2) with $\beta=1$, $\gamma=0.1$, $X=1_n$ being an $n\times 1$ vector of ones, and $Z=(\cos(\frac{2\pi}{n}),\cos(\frac{4\pi}{n}),\ldots,\cos(\frac{2n\pi}{n}))'$. Under this setup, there is only one restricted model in addition to the unrestricted model as candidates for model combination. We show in the online supplementary material that when $n\ge 200\sigma^2$, $\zeta_n=0.005n+\sigma^2$ and $\xi_n\ge\rho^2\zeta_n$. Condition (22) therefore holds. The online supplementary material provides more theoretical examples concerning the relevance of condition (22).

An attempt is also made to verify condition (22) by simulations based on a model setup similar to that of Example 1 in Section 4. We consider the case of three auxiliary regressors, and let $\theta=(5,8,7,c_3(1,0,1))'$. The importance of the auxiliary regressors relative to the focus regressors is measured by the ratio $\alpha=\mathrm{var}(\sum_{j=4}^{6}\theta_j x_{ji})/\mathrm{var}(\sum_{j=1}^{3}\theta_j x_{ji})$. We set c_3 to values that correspond to $\alpha=0.1$, 0.5, and 0.9, and ρ in the weight set $\mathcal{W}$ to 0.3. The simulation results as shown in Figure 1 indicate that $\xi_n^{-2}\zeta_n$ converges to 0 from above as the sample size increases, and an increase in α has the effect of speeding up the convergence of $\xi_n^{-2}\zeta_n$ to zero, ceteris paribus.

It is instructive to mention that the restriction of $\rho>0$ is needed only for condition (22) to hold; once (21) and (22) are established, our subsequent steps for proving (23) do not explicitly require $\rho>0$. We suspect that (23) also holds without having to invoke the assumption of $\rho>0$ as an underlying condition. However, if the nonzero restriction on ρ is relaxed, the technical challenge for establishing (23) will be formidable since one then has to bypass condition (22). The development of such a proof technique is left for future research.

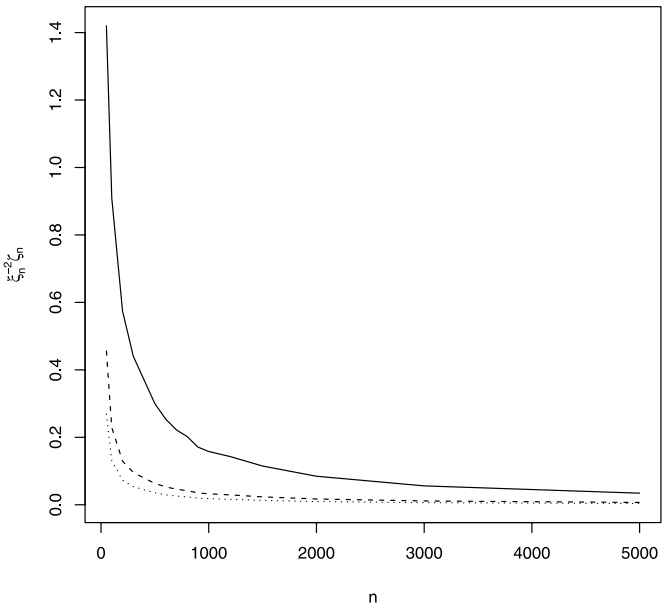

Figure 1. $\xi_n^{-2}\zeta_n$ versus n for different α values (solid line: $\alpha=0.1$; dashed line: $\alpha=0.5$; and dotted line: $\alpha=0.9$).

4. SIMULATION STUDIES

In this section, we conduct simulation experiments to compare the finite sample performance of the OPT estimator with the MMA estimator and the model average estimators based on the S-AIC and S-BIC weights (hereafter referred to as the S-AIC and S-BIC estimators, respectively). Example 1 is based on Equation (2), with the MMA estimator obtained using a regressor ordering pattern decided a priori, and all other FMA estimators combined across 2^m submodels. Example 2, which is based on the same setting as one of the experiments in the article by Hansen (2007), examines the performance of the OPT estimator when it combines models in the same manner as the MMA estimator. The purpose is to evaluate the OPT estimator relative to the MMA estimator when both are considered on a platform that is supposed to favor the MMA estimator. In Example 3, we further demonstrate the advantages of the OPT estimator over the MMA estimator using a real dataset taken from the work of Danilov and Magnus (2004a). The main objective is to examine the extent to which different patterns of regressor ordering affect the properties of the MMA estimator. The results of Example 3 show that the performance of the MMA estimator is highly sensitive to the pattern of regressor ordering, thus illustrating the viability of the OPT estimator as an alternative.

It also appears important to emphasize that the implementation of the OPT method is based on the set $\mathcal{D}$, not $\mathcal{D}_0$. The latter subset is only relevant for the optimality theory, not for empirical application of the OPT method. In fact, $\mathcal{D}_0$ cannot be determined in practice because one cannot know which of the candidate models are unbiased when the true model is unknown. However, as mentioned before, this does not render the

theoretical results of the previous section irrelevant because ρ can take on any positive value close to zero; hence $\mathcal{D}_0$ is generally very close to $\mathcal{D}$. In all cases of our simulations, we set the value of $\bar{c}$ to $n/2$ so that $\mathcal{D}$ can encompass the S-AIC and S-BIC weights.

Example 1. The data are generated from the model

$$y_i = \sum_{j=1}^{10} \theta_j x_{ji} + e_i,$$

where $x_{1i} = 1, x_{ji} \sim \mathrm{N}(0, 1)$ for $j = 2, \ldots, 10$, and $e_i \sim \mathrm{N}(0, 1)$, $i = 1, \ldots, n$. The sample size varies between n = 30, 80, 150, and 300. The error term e_i is independent of x_{ji}'s, and all x_{ji}'s are independent of one another. Arbitrarily, we let x_{1i}, x_{2i}, and x_{3i} be the focus regressors, and consider all other regressors as auxiliary. The parameters are given by $\theta = (\theta_1, \ldots, \theta_{10})' = (1, c_1(3, 4, c_2(0.5, 0.6, 0, 1, 0.4, 0.3, 0.8)))'$. Let $\alpha = \mathrm{var}(\sum_{j=4}^{10} \theta_j x_{ji}) / \mathrm{var}(\sum_{j=1}^{3} \theta_j x_{ji})$, which may be written as $\alpha = c_1^2 c_2^2 (0.5^2 + 0.6^2 + 0^2 + 1^2 + 0.4^2 + 0.3^2 + 0.8^2)/(c_1^2(3^2 + 4^2)) = c_2^2/10$ in the present context. Note that α measures the importance of the auxiliary regressors relative to the focus regressors; the larger the value of c_2^2 (and hence α), the greater the importance of the auxiliary regressors. We set α to 0.1 and 0.9. The population $R^2 = 25c_1^2(1 + \alpha)/(1 + 25c_1^2(1 + \alpha))$ is controlled by the parameter c_1, where $25c_1^2(1+\alpha) = \mathrm{var}(\sum_{j=1}^{10} \theta_j x_{ji})$ is the variance of the linear combination of all regressors, focus and auxiliary. We set R^2 in the range of $[0.1, 0.9]$. With seven auxiliary regressors, the OPT, S-AIC, and S-BIC estimators average estimates across 2^7 models. In computing the MMA estimator, we order the regressors as $x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}$, and $x_{10,i}$. The MMA estimator is then obtained by applying Equation (1) with the model weights derived by minimizing the Mallows criterion [see Hansen 2007, equation (11)]. Our simulation experiment is based on 2000 replications.

We begin by discussing the results when the estimators are evaluated in terms of risk under the loss function

$$L^{(1)} = \left\| \hat{\mu} - \sum_{j=1}^{10} \theta_j x_j \right\|^2,$$

that is, the predictive loss of $\hat{\mu}$, where $x_j = (x_{j1}, \ldots, x_{jn})'$. With the predictive loss as the penalty function, we obtain the OPT estimator by selecting the λ that minimizes (20). Results of risk comparisons are given in Figures 2 and 3. As in the articleby Hansen (2007), we normalize the risk by dividing by the risk of the infeasible optimal least squares estimator, that is, the risk of the best-fitting model among the 2^7 models. Figures 2 and 3 reveal that the OPT estimator generally has better risk performance than the other three estimators no matter the value of n and α. Exceptions occur when R^2 is very large or small. For example, when R^2 is near 0.1 and n is small, the MMA estimator typically achieves the lowest risk; when R^2 is near 0.9 and n is large, the S-AIC and S-BIC estimators can be superior to both the MMA and OPT estimators.

Next we consider the efficiency of the estimators of the coefficients of the focus regressors. Evaluation is based on the loss function

$$L^{(2)} = \sum_{j=1}^{3} (\hat{\theta}_j - \theta_j)^2.$$

In this case we compute the weight vector of the OPT estimator by minimizing (11). Because the MMA approach does not distinguish between focus and auxiliary regressors, it makes no sense to include the MMA estimator in the evaluation, and thus

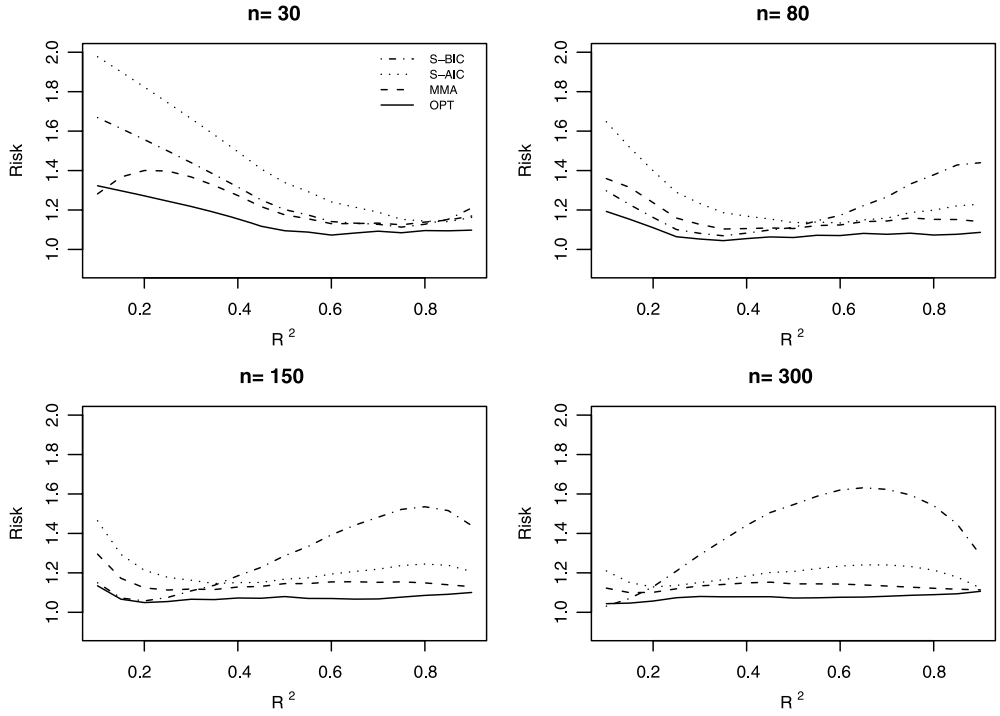

Figure 2. Results for Example 1: risk comparisons under $L^{(1)}$ loss when $\alpha = 0.1$.

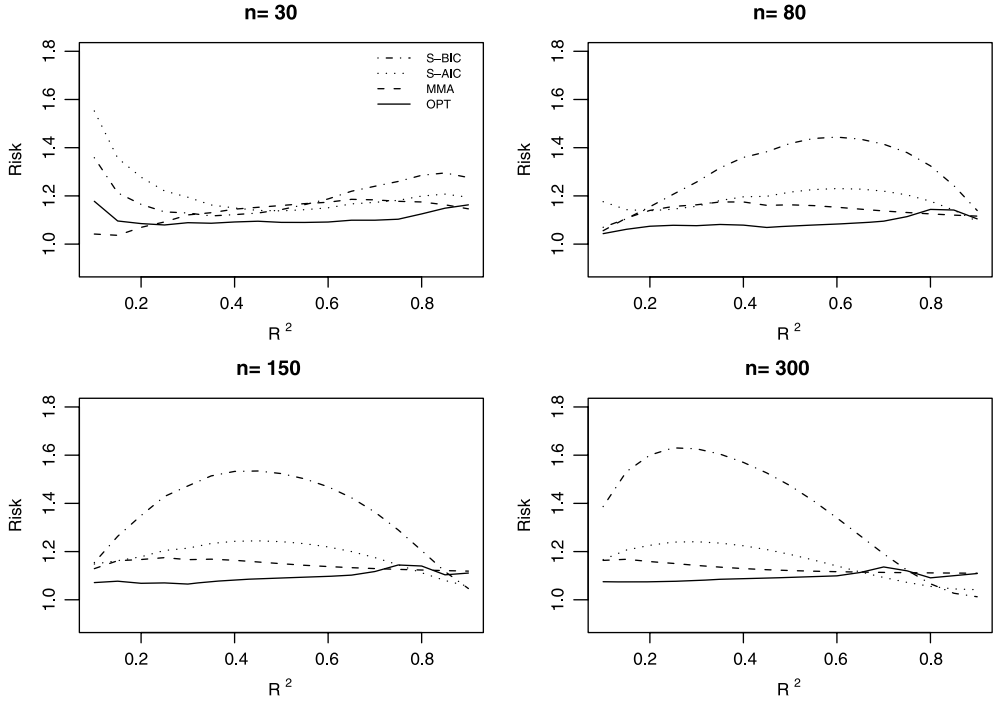

Figure 3. Results for Example 1: risk comparisons under $L^{(1)}$ loss when $\alpha = 0.9$.

we compare only the OPT, S-AIC, and S-BIC estimators when interest focuses on the estimation of the coefficients of the focus regressors. Figure 4 provides a selection of results. Again, in each case the risk is normalized by dividing by the risk of the infeasible optimal least squares estimator. From Figure 4, we observe that when n and α are both small, the S-BIC estimator has the best performance while the OPT estimator has the worst; however, when n is large or α is large, except when R^2 is very large or very small, the OPT estimator is the best while the S-BIC is the worst. Results of cases not shown here are available in the online supplementary material; in general, they depict very similar characteristics to those shown in Figure 4.

Example 2. This example is based on the same setting as in the article by Hansen (2007), that is, $y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i$, $x_{1i} = 1$, all remaining x_{ji}'s are $N(0, 1)$, e_i is distributed as

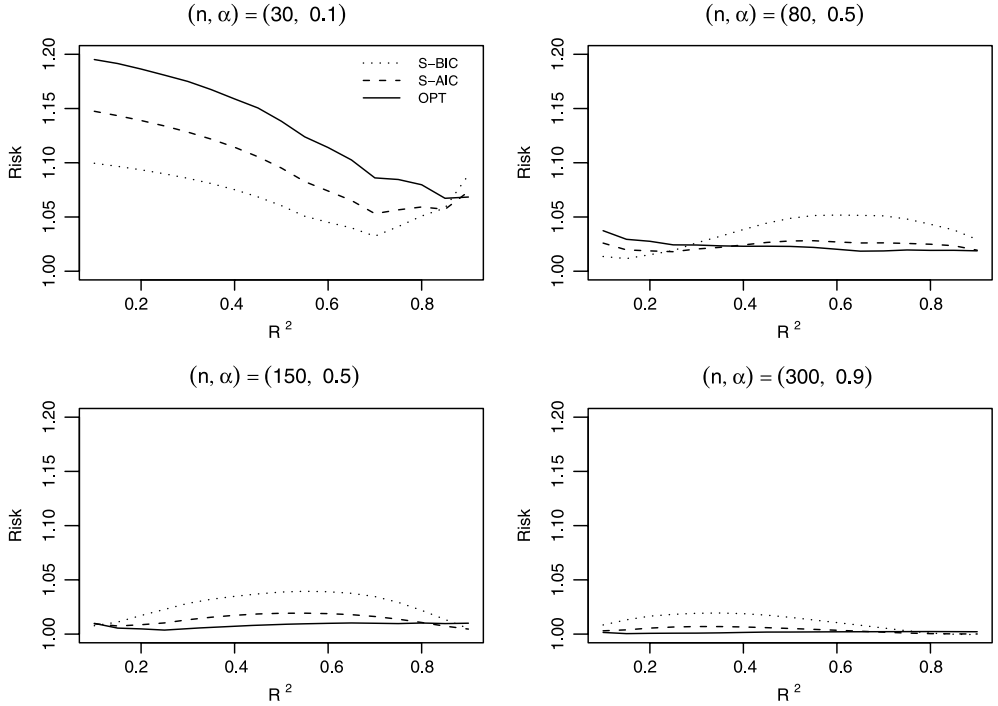

Figure 4. Results for Example 1: risk comparisons under $L^{(2)}$ loss.

N(0, 1), independent of x_{ji}'s, all x_{ji}'s are independent of one another, $\theta_j = c_3\sqrt{2\alpha_1}j^{-\alpha_1-1/2}$, and the population $R^2 = c_3^2/(1+c_3^2)$ is controlled by c_3. Sample size varies between $n = 50, 150, 400$, and 1000, α_1 is set to 0.5, 1.0, and 1.5, and R^2 is set in the range of $[0.1, 0.9]$. The total number of regressors P in the regression is determined by $P = 3n^{1/3}$. Like Hansen (2007), we consider P nested approximating submodels with the pth submodel comprising the first p regressors. All four model average estimators combine estimates across these P submodels. Note that although the OPT, S-AIC, and S-BIC can potentially combine estimates from all candidate models, we only consider P nested models here—the purpose is to evaluate the OPT estimator when all estimators are considered on a platform that is supposed to favor the MMA estimator. As in the article by Hansen (2007), evaluation is based on the loss function

$$L^{(3)} = \left\| \hat{\mu} - \sum_{j=1}^{\infty} \theta_j x_j \right\|^2.$$

Results for four different cases are depicted in Figure 5. Again, in each case the risk is normalized by dividing by the risk of the infeasible optimal least squares estimator. It is seen from the figures that the MMA estimator habitually yields better estimates than the S-AIC and S-BIC estimators—these results are in accord with those observed by Hansen (2007). What is more striking is that the OPT estimator is found to be superior to the MMA estimator in a large region of the parameter space, and this superiority is most marked when n is large. This result is particularly encouraging given that the experiment has been performed under the same setting as Hansen's (2007), where it was shown that the MMA estimator performs best. Results of the cases not depicted here have characteristics similar to those shown in Figure 5. Readers may refer to the online supplementary material for details.

Example 3. The third design is based on a model from Pearson and Timmermann (1994), who considered the predictability of excess returns for the Standard and Poor 500 index. The same model has been used to illustrate the consequences of ignoring pre-testing on forecasts by Danilov and Magnus (2004a). The model is expressed in the following equation:

$$y_t = \beta_1 + \beta_2 PI_{t-2} + \beta_3 DI3_{t-1} + \beta_4 SPREAD_{t-1} + \gamma_1 YSP_{t-1} + \gamma_2 DIP_{t-1} + \gamma_3 PER_{t-1} + \gamma_4 DLEAD_{t-2} + \varepsilon_t, \quad (27)$$

where y_t is excess returns, PI_{t-2} is annual inflation rate (lagged two periods), $DI3_{t-1}$ is change in 3-month T-bill rate (lagged one period), $SPREAD_{t-1}$ is credit spread (lagged one period), YSP_{t-1} is dividend yield on SP500 portfolio (lagged one period), DIP_{t-1} is annual change in industrial production (lagged one period), PER_{t-1} is price-earnings ratio (lagged one period), and $DLEAD_{t-2}$ is annual change in leading business cycle indicator (lagged two periods). The data contain 46 annual observations on each of the variables described above over the period 1956–2001. The data and their sources were given in the work of Danilov and Magnus (2004a). Specifically, Danilov and Magnus (2004a) were uncertain whether the last four regressors, namely, YSP_{t-1}, DIP_{t-1}, PER_{t-1}, and $DLEAD_{t-2}$, should be included. The regressors PI_{t-2}, $DI3_{t-1}$, $SPREAD_{t-1}$ and the intercept are focus regressors that are required to be in the model. Danilov and Magnus (2004a) reported estimates from a (forward) stepwise model selection procedure which discarded all auxiliary regressors but YSP_{t-1}.

Alternative approaches could be the use of the OPT and MMA estimators described above. The OPT estimator takes average across $2^4 = 16$ models. Again, for the MMA scheme, one need not distinguish between focus and auxiliary regressors, but must first order the regressors. Since the model contains eight regressors including the intercept term, there are

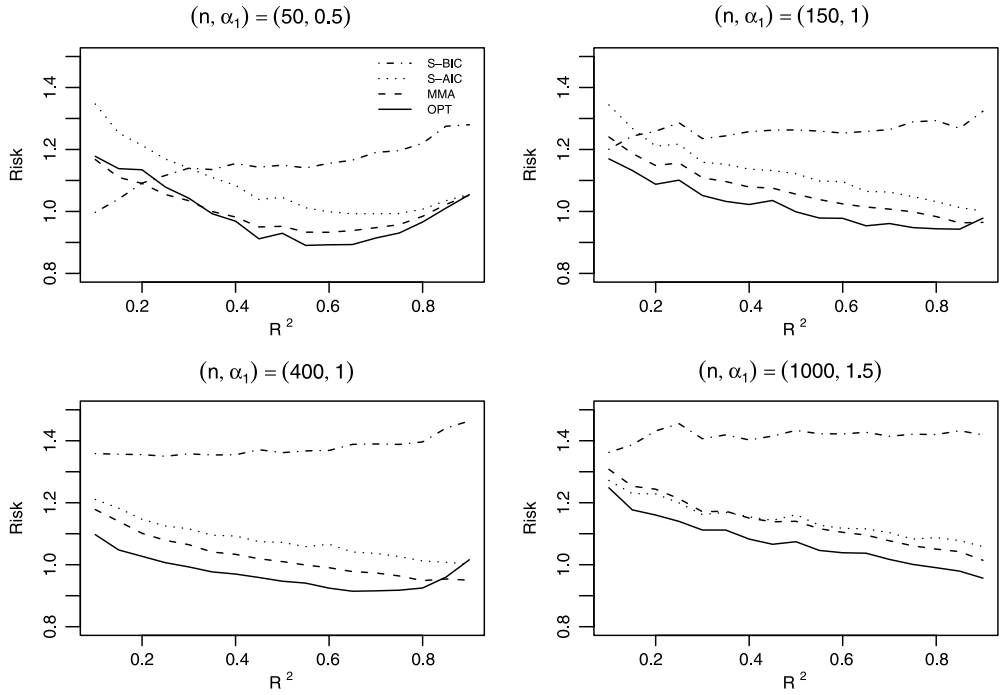

Figure 5. Results for Example 2: risk comparisons under $L^{(3)}$ loss.

$8! = 40{,}320$ possible ways to order the regressors. After ordering, the MMA scheme averages over eight models obtained by adding the eight regressors one at a time to the regression model. While in practice the regressors are ordered based on some preconceived notions of the investigator, for this discussion we consider all 40,320 possible ordering sequences to give a comprehensive picture of the performance of the estimator in all cases. To increase the realism of the simulation, the values of the dependent variable in each round of the simulations are obtained by drawing 46 random disturbances with replacements from the residuals of the least squares estimation of (27). Denote the lth such sample of disturbances as e_l^*, and values of the dependent variable in the lth sample are generated using $Y_l^* = H\Theta + e_l^*$. The experiment uses the least squares estimates of the coefficients in (27) as the true parameter vector Θ. A total of 100 samples are drawn, and the OPT predictor $\hat{\mu}_f = H\hat{\Theta}_f$ and the MMA predictor $\hat{\mu}_{\text{mma}} = H\hat{\Theta}_{\text{mma}}$ are computed. There are altogether 40,320 $\hat{\mu}_{\text{mma}}$'s depending on the ordering of regressors. It should be noted that no estimators are exactly the same among these 40,320 $\hat{\mu}_{\text{mma}}$'s. For each estimator of $H\Theta$, the risk under the squared error loss is calculated.

The key findings are presented as follows. The risk of $\hat{\mu}_f$ is 0.0749, while the risk of $\hat{\mu}_{\text{mma}}$ ranges from a minimum of 0.0583 to a maximum of 0.0893, depending on the pattern in which the regressors are ordered. The "average risk" of $\hat{\mu}_{\text{mma}}$, obtained by taking an average of the risks of all 40,320 $\hat{\mu}_{\text{mma}}$'s, is 0.0788. Of the 40,320 patterns of ordering considered, the MMA estimator results in higher risk than the OPT estimator in 31,839 out of 40,320 or 79% of cases. Clearly, the extent to which the ordering pattern of regressors affects the risk behavior of the MMA estimator is a notable feature of this study. It also points to the narrow scope of the preceding Experiment 1, and the simulation experiment considered by Hansen (2007). These experiments examined only one pattern of ordering regressors for the MMA estimator. Results of the current experiment show that for the current dataset there is a clear tendency for the OPT estimator to provide better estimates than the MMA estimator in most cases. The OPT estimator has worse performance than the best MMA estimator, but this is more than compensated for by a substantial reduction in risk of the OPT estimator over the MMA estimator in the majority of cases considered.

5. EXTENSIONS TO GENERAL PARAMETRIC MODELS

The preceding analysis focusing on the linear regression model can be extended to model combination in general parametric models. This is accomplished by utilizing the local misspecification framework developed by (Hjort and Claeskens 2003).

Assume that the observations $y_1, \ldots, y_n$ are iid, and generated from the density

$$f_{\text{true}}(y) = f(y, \delta, \gamma) = f(y, \delta, \gamma_0 + \theta/\sqrt{n}), \tag{28}$$

where δ is a $k \times 1$ unknown vector, γ_0 is an $m \times 1$ known vector, and θ is an $m \times 1$ unknown vector representing the degree of the departure from the narrow model. As in Section 2, a candidate model always contains all k parameters in δ, and potentially some or all of the m parameters in γ associated with auxiliary regressors whose inclusion in the model is uncertain. The parameter of interest is $\mu = \mu(\delta, \gamma) = \mu(\delta, \gamma_0 + \theta/\sqrt{n})$. Altogether there are 2^m submodels mapping onto the set $S \subset \{1, \ldots, m\}$; note that $\theta_j = 0$ for $j \in S^{\text{C}}$ with S^{C} being the complement of S. We consider model combinations over N ($\leq 2^m$) of these submodels. Let $\hat{\delta}_S$ and $\hat{\gamma}_S$ be submodel estimators based on maximum likelihood (ML) in the model that employs γ_j's with $j \in S$. The ML estimator of μ is then $\hat{\mu}_S = \mu(\hat{\delta}_S, \hat{\gamma}_S, \gamma_{0,S^{\text{C}}})$. Based on this framework, the FMA estimator of μ may be written as

$$\hat{\mu} = \sum_S \tilde{c}(S|D_n)\hat{\mu}_S, \tag{29}$$

where $\tilde{c}$ is a weight that depends on $D_n \equiv \hat{\theta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0)$. Note that D_n is analogous to $\hat{\theta}$ in Section 2.

To study the choice of $\tilde{c}$, we first present some notations. Denote by J_{full} the $(k+m) \times (k+m)$ information matrix of the full model evaluated at the null point (δ, γ_0). That is,

$$J_{\text{full}} = \text{var}_0 \begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix},$$

with inverse

$$J_{\text{full}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

where $U(y) = \partial \log f(y, \delta, \gamma_0)/\partial\delta$ and $V(y) = \partial \log f(y, \delta, \gamma_0)/\partial\gamma$ are the score functions.

Also, let π_S be the projection matrix mapping the vector $v = (v_1, \ldots, v_m)'$ to its subvector $\pi_S v = v_S$ that consists of v_j with $j \in S$. Denote $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, $K_S = (\pi_S K^{-1} \pi_S')^{-1}$, $H_S = K^{-1/2}\pi_S' K_S \pi_S K^{-1/2}$, and $\omega = J_{10}J_{00}^{-1}\,\partial\mu/\partial\delta - \partial\mu/\partial\gamma$ with the partial derivatives evaluated at the null point (δ, γ_0). Further, we define H_ϕ as the null matrix of size $m \times m$, where ϕ is the empty set.

From (Hjort and Claeskens 2003), we have

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial\mu}{\partial\delta}\right)' J_{00}^{-1} M + \omega'\{\theta - \hat{\theta}(D)\}, \tag{30}$$

where $\hat{\theta}(D) = K^{1/2}\{\sum_S \tilde{c}(S|D)H_S\}K^{-1/2}D$, $D \sim \text{N}_m(\theta, K)$ is the limiting variable of D_n in distribution, and $M \sim \text{N}_k(0, J_{00})$ is independent of D.

If we denote $Z = K^{-1/2}D$, then $Z \sim \text{N}(a^*, I)$ with $a^* = K^{-1/2}\theta$, and we can write $\hat{\theta}(D) = K^{1/2}\{\sum_S c^*(S|Z)H_S\}Z \equiv K^{1/2}\hat{a}^*(Z)$. Thus, $\Lambda = (\frac{\partial\mu}{\partial\delta})' J_{00}^{-1} M + \omega' K^{1/2}\{a^* - \hat{a}^*(Z)\}$. From this analysis we can show that the asymptotic risk of $\hat{\mu}$ is

$$R_a(\hat{\mu}) = E(\Lambda^2) = \tau_0^2 + E\{\omega' K^{1/2}\hat{a}^*(Z) - \omega' K^{1/2} a^*\}^2, \tag{31}$$

with $\tau_0^2 = (\frac{\partial\mu}{\partial\delta})' J_{00}^{-1} (\frac{\partial\mu}{\partial\delta})$.

Assume (for the time being) that τ_0, ω, K, and $\widehat{a}^*(Z)$ are known, and let the weight $c^*(S|z)$ be a continuous function. Further, assume that the piecewise continuous partial derivatives of $c^*(S|z)$ with respect to z exist as do their expectations. Then, by a similar proof technique to that used for Theorem 1, we can derive the following unbiased estimator of the asymptotic risk of $\hat{\mu}$:

$$\widetilde{R}_a(\hat{\mu}) = \tau_0^2 - \omega' K \omega + \omega' K^{1/2}(\hat{a}^*(Z) - Z)(\hat{a}^*(Z) - Z)' K^{1/2}\omega + 2\omega' K^{1/2} \frac{\partial \hat{a}^*(Z)}{\partial Z'} K^{1/2}\omega. \tag{32}$$

Note that $\widetilde{R}_a(\hat{\mu})$ depends on τ_0^2, ω, K, and $\hat{a}^*(Z)$, which are actually unknown, and dependent on δ and/or Z. Also, the first and second terms on the right side of Equation (32) are independent on the model. Thus, we suggest a criterion of weight selection that minimizes the following quantity:

$$\hat{R}_a(\hat{\mu}) = \left\{\hat{\omega}'\hat{K}^{1/2}(\hat{a}^*(Z_n) - Z_n)\right\}^2 + 2\hat{\omega}'\hat{K}^{1/2}\frac{\partial \hat{a}^*(Z_n)}{\partial Z'}\hat{K}^{1/2}\hat{\omega}, \quad (33)$$

which is obtained by neglecting the first two terms on the right side of (32), and replacing J_{full} and δ in the same equation by their estimators $\hat{J}_{\text{full}}$ and $\hat{\delta}$, respectively, and Z by the approximation $Z_n = \hat{K}^{-1/2}D_n$.

Consider the simplest special case where there are only the full and narrow models. Then (32) reduces to

$$\hat{R}_a(\hat{\mu}) = (c_{\text{full}} - 1)^2\left(\hat{\omega}'\hat{K}^{1/2}Z_n\right)^2 + 2\,c_{\text{full}}\hat{\omega}'\hat{K}\hat{\omega}. \quad (34)$$

It is easily shown that (34) is minimized at

$$c_{\text{full}} = 1 - \frac{\hat{\omega}'\hat{K}\hat{\omega}}{(\hat{\omega}'\hat{K}^{1/2}Z_n)^2}, \quad (35)$$

and thus

$$c_{\text{narrow}} = 1 - c_{\text{full}} = \frac{\hat{\omega}'\hat{K}\hat{\omega}}{(\hat{\omega}'\hat{K}^{1/2}Z_n)^2}. \quad (36)$$

These weights are very close to but not exactly the same as the weights chosen by (Hjort and Claeskens 2003). This slight difference in weights is due to the fact that Hjort and Claeskens's (2003) criterion minimizes the asymptotic risk itself, whereas our criterion seeks weights that minimize the unbiased estimator of the risk. In fact, when θ is the only unknown quantity, the corresponding weights in the work of (Hjort and Claeskens 2003) may be obtained by taking the expectations of the numerators and denominators separately in the weights (35) and (36), by noting that $E(\omega'K^{1/2}Z)^2 = \omega'K\omega + (\omega'K^{1/2}a^*)^2$. Work in progress evaluates the empirical performance of the weight choice criterion based on $\hat{R}_a(\hat{\mu})$ in (33).

We end this section by noting that the above results developed for iid y_i's can be extended to the case of regression models under some mild regularity conditions. Assume that y_i's are generated from the density

$$f_{i,\text{true}}(y|x_i) = f(y|x_i, \delta, \gamma) = f(y|x_i, \delta, \gamma_0 + \theta/\sqrt{n}),$$

where δ typically comprises a $k \times 1$ vector of regression coefficients β and a scalar parameter σ. The matrix

$$J_{n,\text{full}} = \frac{1}{n}\sum_{i=1}^{n}\text{var}_0\begin{pmatrix}\partial \log f(y_i|x_i,\delta,\gamma_0)/\partial\delta \\ \partial \log f(y_i|x_i,\delta,\gamma_0)/\partial\gamma\end{pmatrix} = \begin{pmatrix}J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11}\end{pmatrix}$$

is assumed to converge to a suitable positive definite matrix J_{full} as n tends to infinity. The extension to the regression models is accomplished by replacing $\hat{J}_{\text{full}}$ in $\hat{R}_a(\hat{\mu})$ of Equation (33) with a suitable estimate of $J_{n,\text{full}}$.

6. CONCLUDING REMARKS

There has been a quickening of interest in frequentist model averaging in recent years. This article suggests a new approach to select model weights for a linear regression FMA estimator. The proposed estimator has been shown to be quite promising, and yields improved estimator performance over the estimators developed in the literature in a wide variety of circumstances. Among the known FMA estimators, Hansen's (2007) MMA estimator has considerable appeal, but to implement this estimator the regressors must be ordered at the outset. One practical issue addressed in our investigation is how the various patterns of ordering regressors affect the *finite sample* performance of the MMA estimator. The simulation results presented here suggest that the way the regressors are ordered is indeed a major determinant of the finite sample behavior of the MMA estimator. For the experiments considered, the risks of the MMA estimators corresponding to different patterns of regressor ordering can differ markedly. The OPT estimator requires no such prior ordering of regressors, and is supported by both asymptotic as well as analytic finite sample justifications. Another feature of our analytical framework is that it nests other weights such as those based on S-AIC and S-BIC as special cases. Note also that our proposed criteria permit comparisons of different weighting schemes. While the bulk of our analysis emphasizes the normal linear regression model, a similar weight choice mechanism is also developed for model averaging in general likelihood models.

Admittedly, if model averaging is performed over the full set or a large subset of extended models, the computational burden quickly increases as m increases. In this regard, the model screening prior to combining approach advocated by Yuan and Yang (2005), or the orthogonalization method developed recently by Magnus, Powell, and Prüfer (2010), may be desirable alternatives to direct computation. Recently, Hansen (2008) extended the idea of Mallows model averaging to forecast combinations. It would be interesting to further extend the OPT approach to an out-of-sample forecasting setting. We end by reiterating that although model combination captures the uncertainty inherent in model selection, from a statistical inference viewpoint, model combination itself does not guarantee that subsequent inference would be on sound footing; in order for post-model averaging to produce the "correct" inference, one has to work with distribution of the model average estimator. Studies of the distributional properties of the FMA estimators are of recent vintage. Readers are referred to the work of Pötscher (2006), who investigated the distributional properties of a special case of the FMA estimator discussed by Leung and Barron (2006). It remains a challenging endeavor to derive the full distribution of the OPT estimator discussed here.

APPENDIX: PROOFS OF THEOREMS

Proof of Theorem 1

The MSE of $\hat{\beta}_f$ may be written as

$$\text{MSE}(\hat{\beta}_f) = E\{(\hat{\beta}_f - \beta)^{\otimes 2}\} = \sigma^2(X'X)^{-1} + QE\{(W\hat{\theta} - \theta)^{\otimes 2}\}Q' \quad (A.1)$$

(see theorem 1 of Danilov and Magnus 2004b). By the definition of W, we can write

$$E\{(W\hat{\theta}-\theta)^{\otimes 2}\}$$
$$= E\big[\{(W-I_m)\hat{\theta}\}^{\otimes 2}\big] + \sum_{i=1}^{N} E\{\lambda_i(\hat{\theta},\hat{\sigma}^2)(\hat{\theta}-\theta)\hat{\theta}'\}(W_i - I_m)$$
$$+ \sum_{i=1}^{N} (W_i - I_m)E\{\lambda_i(\hat{\theta},\hat{\sigma}^2)\hat{\theta}(\hat{\theta}-\theta)'\} + E\{(\hat{\theta}-\theta)^{\otimes 2}\}. \quad \text{(A.2)}$$

Noting that $\hat{\theta}$ and $\hat{\sigma}^2$ are independent, and using the assumptions on $\lambda_i(\hat{\theta},\hat{\sigma}^2)$, it can be shown by Stein's Lemma (Stein 1981) that

$$E\{\lambda_i(\hat{\theta},\hat{\sigma}^2)(\hat{\theta}-\theta)\hat{\theta}'\} = \sigma^2 E\big[\lambda_i(\hat{\theta},\hat{\sigma}^2)I_m + \{\partial\lambda_i(\hat{\theta},\hat{\sigma}^2)/\partial\hat{\theta}\}\hat{\theta}'\big].$$

Hence,

$$Q\left[\sum_{i=1}^{N} E\{\lambda_i(\hat{\theta},\hat{\sigma}^2)(\hat{\theta}-\theta)\hat{\theta}'\}(W_i - I_m)\right]Q'$$
$$= \sigma^2 Q\left(E\left[W + \sum_{i=1}^{N}\{\partial\lambda_i(\hat{\theta},\hat{\sigma}^2)/\partial\hat{\theta}\}\hat{\theta}'W_i\right] - I_m\right)Q'$$
$$= \sigma^2 E\{\Psi_1(\hat{\theta},\hat{\sigma}^2)\} - \sigma^2 QQ'. \quad \text{(A.3)}$$

Further,

$$E_{\hat{\sigma}^2}\{\Psi(\hat{\theta},\hat{\sigma}^2)\} = \sigma^2 E_{\hat{\sigma}^2}\{\Psi_1(\hat{\theta},\hat{\sigma}^2)\}. \quad \text{(A.4)}$$

Equation (A.4) can be proven by noting that $(n-k-m)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k-m)$, resulting in

$$p(t) = \frac{(n-k-m)^{(n-k-m)/2}}{2^{(n-k-m)/2}\Gamma((n-k-m)/2)(\sigma^2)^{(n-k-m)/2}} \times t^{(n-k-m)/2-1}e^{-(n-k-m)t/(2\sigma^2)}$$
$$\equiv C_0 t^{(n-k-m)/2-1}e^{-(n-k-m)t/(2\sigma^2)}, \qquad t>0$$

as the density function of $\hat{\sigma}^2$. Thus,

$$E_{\hat{\sigma}^2}\{\Psi(\hat{\theta},\hat{\sigma}^2)\}$$
$$= \int_0^\infty \Psi(\hat{\theta},t)\cdot C_0 t^{(n-k-m)/2-1}e^{-(n-k-m)t/(2\sigma^2)}\,dt$$
$$= \{C_0(n-k-m)/2\} \times \int_0^\infty\int_0^t u^{(n-k-m)/2-1}\Psi_1(\hat{\theta},u)e^{-(n-k-m)t/(2\sigma^2)}\,du\,dt$$
$$= \{C_0(n-k-m)/2\} \times \int_0^\infty\int_u^\infty u^{(n-k-m)/2-1}\Psi_1(\hat{\theta},u)e^{-(n-k-m)t/(2\sigma^2)}\,dt\,du$$
$$= \sigma^2 C_0 \int_0^\infty u^{(n-k-m)/2-1}\Psi_1(\hat{\theta},u)e^{-(n-k-m)u/(2\sigma^2)}\,du$$
$$= \sigma^2 E_{\hat{\sigma}^2}\{\Psi_1(\hat{\theta},\hat{\sigma}^2)\}.$$

Taking (A.1)–(A.4) together and replacing σ^2 by $\hat{\sigma}^2$ in (A.1) lead to the unbiased MSE estimator of $\hat{\beta}_f$ in (4), and this proves Theorem 1. Note that the approach used to derive $\Psi(\hat{\theta},\hat{\sigma}^2)$ is similar to that adopted by Carter et al. (1990) and Giles and Srivastava (1991). The online supplementary material provides the details of the derivation.

Proof of Theorem 2

It is readily seen that

$$\mathrm{MSE}(\hat{\mu}_f) = E\{(\hat{\mu}_f - H\Theta)^{\otimes 2}\} = H\begin{pmatrix} E(\hat{\beta}_f-\beta)^{\otimes 2} & E(\hat{\beta}_f-\beta)(\hat{\gamma}_f-\gamma)' \\ E(\hat{\gamma}_f-\gamma)(\hat{\beta}_f-\beta)' & E(\hat{\gamma}_f-\gamma)^{\otimes 2}\end{pmatrix}H'. \quad \text{(A.5)}$$

It is also straightforward to show that

$$E\{(\hat{\gamma}_f-\gamma)^{\otimes 2}\} = DE\{(W\hat{\theta}-\theta)^{\otimes 2}\}D' \quad \text{(A.6)}$$

and

$$E\{(\hat{\gamma}_f-\gamma)(\hat{\beta}_f-\beta)'\} = -DE\{(W\hat{\theta}-\theta)^{\otimes 2}\}Q'. \quad \text{(A.7)}$$

Using the same arguments as in the proof of Theorem 1 concerning $\lambda_i(\hat{\theta},\hat{\sigma}^2)$ and $\partial\lambda_i(\hat{\theta},\hat{\sigma}^2)/\partial\hat{\theta}$, we can show that

$$E\{(W\hat{\theta}-\theta)^{\otimes 2}\} = E\{\varphi(\hat{\theta},\hat{\sigma}^2,I_m,I_m)\}. \quad \text{(A.8)}$$

Equation (14) is obtained by using (A.8) in (A.6) and (A.7) and substituting the resultant expressions in (A.5).

Proof of Theorem 3

We first show the relationship between our criterion $\hat{R}_a(\hat{\mu}_f(\lambda(a,b,c)))$ and the squared error loss $L_n(\lambda(a,b,c))$. Let $V_i = [0_{r_i\times k} : S_i']$; then the restriction $S_i'\gamma = 0$ may be equivalently written as $V_i\Theta = 0$. Correspondingly, the restricted least squares estimator of Θ is given by

$$\hat{\Theta}_{(i)} = (H'H)^{-1}H'y - (H'H)^{-1}V_i'\{V_i(H'H)^{-1}V_i'\}^{-1}V_i(H'H)^{-1}H'y. \quad \text{(A.9)}$$

It is well known (e.g., Rao 1973) that

$$(H'H)^{-1} = \begin{pmatrix} (X'X)^{-1}+QQ' & -Q(Z'MZ)^{-1/2} \\ -(Z'MZ)^{-1/2}Q' & (Z'MZ)^{-1}\end{pmatrix}, \quad \text{(A.10)}$$

by which we can write

$$S_i'(Z'MZ)^{-1}S_j = V_i(H'H)^{-1}V_j' \quad \text{(A.11)}$$

and

$$S_i'(Z'MZ)^{-1}Z'My = S_i'\{-(Z'MZ)^{-1}Z'X(X'X)^{-1}X'y + (Z'MZ)^{-1}Z'y\} = V_i(H'H)^{-1}H'y. \quad \text{(A.12)}$$

By the definitions of P_i and $\hat{\theta}$, we have

$$\hat{\theta}'P_iP_j\hat{\theta} = y'MZ(Z'MZ)^{-1}S_i(S_i'(Z'MZ)^{-1}S_i)^{-1}S_i'(Z'MZ)^{-1} \times S_j(S_j'(Z'MZ)^{-1}S_j)^{-1}S_j'(Z'MZ)^{-1}Z'My. \quad \text{(A.13)}$$

Combining (A.9), (A.11), (A.12), and (A.13), we obtain

$$\bar{l}_{ij} = \hat{\theta}'P_iP_j\hat{\theta} = y'H(H'H)^{-1}V_i'\{V_i(H'H)^{-1}V_i'\}^{-1}V_i(H'H)^{-1}H' \times H(H'H)^{-1}V_j'\{V_j(H'H)^{-1}V_j'\}^{-1}V_j(H'H)^{-1}H'y$$
$$= \{H(H'H)^{-1}H'y - \hat{\mu}_{(i)}\}'\{H(H'H)^{-1}H'y - \hat{\mu}_{(j)}\}$$
$$= (\hat{\mu}_{(i)}-y)'(\hat{\mu}_{(j)}-y) - \|H(H'H)^{-1}H'y - y\|^2$$
$$= (\hat{\mu}_{(i)}-y)'(\hat{\mu}_{(j)}-y) - (n-k-m)\hat{\sigma}^2. \quad \text{(A.14)}$$

It then follows that

$$\lambda'(a,b,c)\bar{L}\lambda(a,b,c) = \|\hat{\mu}_f(\lambda(a,b,c)) - y\|^2 - (n-k-m)\hat{\sigma}^2. \quad \text{(A.15)}$$

By the definition of $\bar{\phi}$, we see that

$$\lambda'(a,b,c)\bar{\phi} = \lambda'(a,b,c)q - k, \tag{A.16}$$

where $q = (q_1, \ldots, q_N)'$. Now, let H_i be the regressor matrix in the ith submodel, $T_i = H_i(H_i'H_i)^{-1}H_i'$, and $A_i = I_n - T_i$. For any weight vector w, define $T(w) = \sum_{i=1}^N w_i T_i$ and $A(w) = \sum_{i=1}^N w_i A_i$. We can define $T(\lambda(a,b,c))$ and $A(\lambda(a,b,c))$ in a similar way. Hence, it is clear that for $i = 1, \ldots, N$,

$$\hat{\mu}_{(i)} = T_i y, \tag{A.17}$$

and thus $\hat{\mu}_f(\lambda(a,b,c)) = T(\lambda(a,b,c))y$. From (A.15) and (A.16), we obtain

$$\begin{aligned}
&\hat{R}_a\big(\hat{\mu}_f(\lambda(a,b,c))\big)\\
&\quad = \big\|\hat{\mu}_f(\lambda(a,b,c)) - y\big\|^2 + 2\hat{\sigma}^2\lambda'(a,b,c)q\\
&\qquad - n\hat{\sigma}^2 - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)\bar{G}\lambda(a,b,c)\\
&\quad = \big\|\hat{\mu}_f(\lambda(a,b,c)) - \mu\big\|^2 - 2\big(\hat{\mu}_f(\lambda(a,b,c)) - \mu\big)'\varepsilon + \|\varepsilon\|^2\\
&\qquad + 2\hat{\sigma}^2\lambda'(a,b,c)q\\
&\qquad - n\hat{\sigma}^2 - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)\bar{G}\lambda(a,b,c)\\
&\quad = L_n(\lambda(a,b,c)) + 2\mu'A(\lambda(a,b,c))\varepsilon - 2\varepsilon'T(\lambda(a,b,c))\varepsilon\\
&\qquad + 2\hat{\sigma}^2\lambda'(a,b,c)q - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)\bar{G}\lambda(a,b,c)\\
&\qquad - n\hat{\sigma}^2 + \|\varepsilon\|^2\\
&\quad \equiv L_n(\lambda(a,b,c)) + t_n(a,b,c) - n\hat{\sigma}^2 + \|\varepsilon\|^2,
\end{aligned} \tag{A.18}$$

where the last two terms on the right side of (A.18) are unrelated to a, b, or c, and

$$\begin{aligned}
t_n(a,b,c) &= 2\mu'A(\lambda(a,b,c))\varepsilon - 2\varepsilon'T(\lambda(a,b,c))\varepsilon\\
&\quad + 2\hat{\sigma}^2\lambda'(a,b,c)q - (4/n)c\hat{\sigma}^2\lambda'(a,b,c)\bar{G}\lambda(a,b,c).
\end{aligned}$$

Therefore, we can write

$$(\hat{a},\hat{b},\hat{c}) = \arg\min_{(a,b,c)\in\mathcal{D}_0}\big\{L_n(\lambda(a,b,c)) + t_n(a,b,c)\big\}. \tag{A.19}$$

Consequently,

$$\begin{aligned}
&\inf_{(a,b,c)\in\mathcal{D}_0}\big\{L_n(\lambda(a,b,c)) + t_n(a,b,c)\big\}\\
&\qquad = L_n(\lambda(\hat{a},\hat{b},\hat{c})) + t_n(\hat{a},\hat{b},\hat{c}).
\end{aligned} \tag{A.20}$$

On the other hand, by noting that

$$\begin{aligned}
R_n(w) &= EL_n(w) = E\|T(w)y - \mu\|^2\\
&= \|T(w)\mu - \mu\|^2 + \sigma^2\operatorname{tr}\{T^2(w)\}\\
&= \|T(w)y - \mu\|^2 + \|T(w)\varepsilon\|^2\\
&\quad - 2(T(w)y - \mu)'T(w)\varepsilon + \sigma^2\operatorname{tr}\{T^2(w)\}\\
&= L_n(w) + \|T(w)\varepsilon\|^2\\
&\quad - 2(T(w)\mu + T(w)\varepsilon - \mu)'T(w)\varepsilon + \sigma^2\operatorname{tr}\{T^2(w)\}\\
&= L_n(w) + 2\mu'A(w)T(w)\varepsilon\\
&\quad - \|T(w)\varepsilon\|^2 + \sigma^2\operatorname{tr}\{T^2(w)\},
\end{aligned} \tag{A.21}$$

we have

$$L_n(w) = R_n(w) + u_n(w), \tag{A.22}$$

where $u_n(w) = \|T(w)\varepsilon\|^2 - \sigma^2\operatorname{tr}\{T^2(w)\} - 2\mu'A(w)T(w)\varepsilon$.

Also, by the definition of infimum, there exist a series of nonnegative ϑ_n and sets $(a_n, b_n, c_n) \in \mathcal{D}_0$ such that $\vartheta_n \to 0$ when $n \to \infty$, and

$$\inf_{(a,b,c)\in\mathcal{D}_0} L_n(\lambda(a,b,c)) = L_n(\lambda(a_n,b_n,c_n)) - \vartheta_n. \tag{A.23}$$

Now, from (A.20), (A.22), and (A.23), it can be shown that for any $\delta > 0$,

$$\begin{aligned}
&\Pr\left\{\left|\frac{\inf_{(a,b,c)\in\mathcal{D}_0} L_n(\lambda(a,b,c))}{L_n(\lambda(\hat{a},\hat{b},\hat{c}))} - 1\right| > \delta\right\}\\
&\quad \le 2\Pr\left\{\frac{\sup_{(a,b,c)\in\mathcal{D}_0}|t_n(a,b,c)|}{\xi_n}\cdot\frac{1}{1+\inf_{w\in\mathcal{W}}u_n(w)/\xi_n} > \frac{\delta}{3}\right\}\\
&\qquad + \Pr\left\{\frac{\vartheta_n}{\xi_n}\cdot\frac{1}{1+\inf_{w\in\mathcal{W}}u_n(w)/\xi_n} > \frac{\delta}{3}\right\}\\
&\qquad + 3\Pr\left\{\frac{|\inf_{w\in\mathcal{W}}u_n(w)|}{\xi_n} \ge 1\right\}.
\end{aligned} \tag{A.24}$$

The detailed proof of (A.24) is available in the online supplementary material. Hence, to demonstrate the theorem, it suffices to prove that, as $n \to \infty$,

$$\sup_{(a,b,c)\in\mathcal{D}_0}|t_n(a,b,c)|/\xi_n \xrightarrow{p} 0, \tag{A.25}$$

$$\sup_{w\in\mathcal{W}}|u_n(w)|/\xi_n \xrightarrow{p} 0, \tag{A.26}$$

and

$$\vartheta_n/\xi_n \xrightarrow{p} 0. \tag{A.27}$$

Noting that $\vartheta_n \to 0$, the convergence described in Equation (A.27) is obvious from condition (22). By the same condition, together with Chebyshev's inequality, theorem 2 of the article by Whittle (1960), and the fact that $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, it can be shown that (see the online supplementary material for detailed proofs)

$$\sup_{w\in\mathcal{W}}|\mu'A(w)\varepsilon|/\xi_n \xrightarrow{p} 0, \tag{A.28}$$

$$\sup_{w\in\mathcal{W}}|\varepsilon'T(w)\varepsilon - \hat{\sigma}^2 w'q|/\xi_n \xrightarrow{p} 0, \tag{A.29}$$

$$\sup_{w\in\mathcal{W}}|\mu'A(w)T(w)\varepsilon|/\xi_n \xrightarrow{p} 0, \tag{A.30}$$

and

$$\sup_{w\in\mathcal{W}}\big|\|T(w)\varepsilon\|^2 - \sigma^2\operatorname{tr}\{T^2(w)\}\big|/\xi_n \xrightarrow{p} 0. \tag{A.31}$$

So, by (A.30), (A.31), and the definition of $u_n(w)$, we see that (A.26) is true. If we let

$$t_n(w,c) = 2\mu'A(w)\varepsilon - 2\varepsilon'T(w)\varepsilon + 2\hat{\sigma}^2 w'q - (4/n)c\hat{\sigma}^2 w'\bar{G}w, \tag{A.32}$$

then in order for (A.25) to hold, we need only to prove

$$\sup_{w\in\mathcal{W},-\bar{c}\le c\le 0}|t_n(w,c)|/\xi_n \xrightarrow{p} 0. \tag{A.33}$$

In light of (A.28) and (A.29), to complete the proof, we need only to consider the last term on the right side of (A.32). That is, to prove that (A.25) is true, it suffices to show

$$\sup_{w\in\mathcal{W},-\bar{c}\le c\le 0}|c\hat{\sigma}^2 w'\bar{G}w/n|/\xi_n \xrightarrow{p} 0. \tag{A.34}$$

From (A.14) and $\hat{\sigma}_i^2 = \{(n-k-m)\hat{\sigma}^2 + \hat{\theta}' P_i \hat{\theta}\}/n$, we see that for any $1 \le i, j \le N$,

$$
\begin{aligned}
|\hat{\sigma}^2 \bar{g}_{ij}|/n &= \frac{\hat{\sigma}^2 |\hat{\theta}' P_j \hat{\theta} - \hat{\theta}' P_i P_j \hat{\theta}|}{(n-k-m)\hat{\sigma}^2 + \hat{\theta}' P_j \hat{\theta}} \le \frac{|\hat{\theta}' P_j \hat{\theta} - \hat{\theta}' P_i P_j \hat{\theta}|}{n-k-m} \\
&\le \frac{\hat{\theta}' P_j \hat{\theta}}{n-k-m} + \frac{|\hat{\theta}' P_i P_j \hat{\theta}|}{n-k-m} \\
&= \frac{y' A_j y - (n-k-m)\hat{\sigma}^2}{n-k-m} + \frac{|y' A_i A_j y - (n-k-m)\hat{\sigma}^2|}{n-k-m} \\
&= \frac{y' A_j y - (n-k-m)\hat{\sigma}^2}{n-k-m} \\
&\quad + \frac{|y'(A_i A_j + A_j A_i) y/2 - (n-k-m)\hat{\sigma}^2|}{n-k-m} \\
&\le \frac{\mathcal{S}(A_i) y' y}{n-k-m} + \frac{\mathcal{S}(A_i A_j + A_j A_i) y' y/2}{n-k-m} \\
&\le \frac{\mathcal{S}(A_i) y' y}{n-k-m} + \frac{\mathcal{S}(A_i)\mathcal{S}(A_j) y' y}{n-k-m} \\
&= \frac{2\mu'\mu + 4\mu'\varepsilon + 2\varepsilon'\varepsilon}{n-k-m} \\
&= O_p(1), \qquad \text{(A.35)}
\end{aligned}
$$

where $\mathcal{S}(\cdot)$ denotes the largest singular value of a matrix. The last inequality in (A.35) results from $\mathcal{S}(U_1 U_2) \le \mathcal{S}(U_1)\mathcal{S}(U_2)$, and $\mathcal{S}(U_1 + U_2) \le \mathcal{S}(U_1) + \mathcal{S}(U_2)$ for any $n \times n$ matrices U_1 and U_2 (see Li 1987), while the last equality in (A.35) is obtained from condition (21).

The proof is completed by noting that condition (22) and (A.35) imply (A.34).

SUPPLEMENTARY MATERIALS

Proofs and Results: The supplemental materials contain detailed proofs and additional simulation results as follows (SupplementaryMaterial.pdf):

- Derivation of $\Psi(\hat{\theta}, \hat{\sigma}^2)$ in Theorem 1
- Numerical comparison of $\hat{\sigma}^2 \Psi_1(\hat{\theta}, \hat{\sigma}^2)$ and $\Psi(\hat{\theta}, \hat{\sigma}^2)$
- Proof of (A.28)
- Proof of (A.29)
- Proofs of (A.30) and (A.31)
- Proof of the result on the upper bound of $\sum_{\tau \in \mathcal{U}} \lambda_\tau(a, b, c)$ in Section 3.3
- Proofs of the results related to the simple example in Section 3.3
- Other examples in which condition (22) is satisfied
- Further results for simulation Example 1 in Section 4
- Further results for simulation Example 2 in Section 4.

[Received August 2009. Revised August 2010.]

REFERENCES

Bates, J. M., and Granger, C. W. J. (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468. [3]

Buckland, S., Burnham, K., and Augustin, N. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [1,2]

Carter, R. A. L., Srivastava, M. S., Srivastava, V. K., and Ullah, A. (1990), "Unbiased Estimation of the MSE Matrix of Stein-Rule Estimators, Confidence Ellipsoids and Hypothesis Testing," *Econometric Theory*, 6, 63–74. [12]

Claeskens, G., and Hjort, N. L. (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press. [1]

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403. [4]

Danilov, D., and Magnus, J. R. (2004a), "Forecast Accuracy After Pretesting With an Application to the Stock Market," *Journal of Forecasting*, 23, 251–274. [2,6,9]

——— (2004b), "On the Harm That Ignoring Pretesting Can Cause," *Journal of Econometrics*, 122, 27–46. [1-3,12]

Giles, D. E. A., and Srivastava, V. K. (1991), "An Unbiased Estimator of the Covariance Matrix of the Mixed Regression Estimator," *Journal of the American Statistical Association*, 86, 441–444. [12]

Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [1-3,6-11]

——— (2008), "Least Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350. [11]

Hjort, N., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [1,2,10,11]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [1]

Hurvich, C., and Tsai, C. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307. [4]

Kim, T., and White, H. (2001), "James–Stein-Type Estimators in Large Samples With Application to the Least Absolute Deviations Estimator," *Journal of the American Statistical Association*, 96, 697–705. [4]

Leeb, H., and Pötscher, B. (2008), "Model Selection," in *Handbook of Financial Time Series*, New York: Springer, pp. 889–925. [4]

Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [1,2,11]

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_l, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [14]

Magnus, J. R., and Durbin, J. (1999), "Estimation of Regression Coefficients of Interest When Other Regression Coefficients Are of No Interest," *Econometrica*, 67, 639–643. [2]

Magnus, J., Powell, O., and Prüfer, P. (2010), "A Comparison of Two Averaging Techniques With an Application to Growth Empirics," *Journal of Econometrics*, 154, 139–153. [11]

Pearson, M., and Timmermann, A. (1994), "Forecasting Stock Returns—An Examination of Market Trading in the Presence of Transaction Costs," *Journal of Forecasting*, 13, 335–367. [9]

Pötscher, B. M. (2006), "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding Its Estimation," in *Time Series and Related Topics. IMS Lecture Notes—Monograph Series*, Vol. 52, pp. 113–129. [11]

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [1]

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), Wiley. [12]

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264. [5]

Shen, X., and Huang, H.-C. (2006), "Optimal Model Assessment, Selection and Combination," *Journal of the American Statistical Association*, 101, 554–568. [1]

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [12]

Wallace, T. D. (1972), "Weaker Criteria and Tests for Linear Restrictions in Regression," *Econometrica*, 40, 689–698. [3]

Wetherill, G. B., Duncombe, P., Kenward, M., Köllerström, J., Paul, S. R., and Vowden, B. J. (1986), *Regression Analysis With Applications. Monographs on Statistics and Applied Probability*, London: Chapman & Hall. [4]

Whittle, P. (1960), "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, 5, 302–305. [13]

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588. [1]

——— (2003), "Regression With Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783–809. [1,6]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214. [1,6,11]

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Optimal Weight Choice for Frequentist Model Average Estimators
Author : Hua Liang, Guohua Zou, Alan T. K. Wan, Xinyu Zhang
Subject : Journal of the American Statistical Association, Vol.0, No.0, 2011, 1-14
Keywords: Asymptotic optimality, Finite sample property, Mallows criterion, Smoothed AIC, Smoothed BIC, Unbiased MSE estimate

THE LIST OF URI ADRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page: *http://www.e-publications.org/ims/support/urihelp.html*.

302 http://pubs.amstat.org [2:pp.1,1] Found
--- mailto:hliang@bst.rochester.edu [2:pp.1,1] Check skip
--- mailto:ghzou@amss.ac.cn [2:pp.1,1] Check skip
--- mailto:xinyu@amss.ac.cn [2:pp.1,1] Check skip
--- mailto:msawan@cityu.edu.hk [2:pp.1,1] Check skip
200 http://www.amstat.org [2:pp.1,1] OK
302 http://pubs.amstat.org/loi/jasa [2:pp.1,1] Found
404 http://dx.doi.org/10.1198/jasa.2011.tm09478 [2:pp.1,1] Not Found