On Optimal Weight Choice in a Frequentist Model Average Estimator

HUA LIANG*, GUOHUA ZOU†, ALAN T.K. WAN* AND XINYU ZHANG†
* DEPARTMENT OF BIOSTATISTICS AND COMPUTATIONAL BIOLOGY, UNIVERSITY OF ROCHESTER MEDICAL CENTER, ROCHESTER, NY 14642, USA.
† ACADEMY OF SYSTEM SCIENCE, CHINESE ACADEMY OF SCIENCES, BEIJING 100190, CHINA
* DEPARTMENT OF MANAGEMENT SCIENCES, CITY UNIVERSITY OF HONG KONG, KOWLOON, HONG KONG

Abstract

The paradigm shift from the traditional approaches of model selection to model averaging is gaining ground in several disciplines, partly because of an obvious shortcoming of the former methods in that they tend to underreport variability and the benefit offered by the latter in that it helps account for the uncertainty inherent in the model selection process. Developing appropriate weight selection methods is an important and challenging problem for model average estimators. The model average estimator in this paper is from a linear regression model and we propose selecting the weights by minimizing the trace of an unbiased estimate of the model average estimator’ MSE. Asymptotic and analytic finite sample properties of the method are discussed. Simulations are carried out to study the behaviour of the model average estimator that results from the proposed weight selection method and comparisons are drawn with other weight choice schemes.

Key Words and Phrases: asymptotic optimality; finite sample properties; Mallows criterion; Smoothed AIC; Smoothed BIC; trace; Unbiased MSE estimate

Short title: Frequentist Model Averaging

1Correspondence to Professor Hua Liang, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, U.S.A., email: hliang@bst.rochester.edu
INTRODUCTION

It has long been recognized that as model selection neglects the uncertainty associated with the selection process, inference based on the final model can be seriously misleading (Hjort and Claeskens, 2003). Model averaging, on the other hand, provides a coherent mechanism for accounting for this model uncertainty. Bayesian model averaging has been widely studied and used in a variety of contexts. For a discussion of Bayesian model averaging, the readers are referred to Raftery, Madigan, and Hoeting (1997) or Hoeting, Madigan, Raftery and Volinsky (1999). 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. Although there is a growing frequentist literature on model averaging, it would be fair to say that frequentists, until recently, are a distinct minority among those who advocate model averaging, which has been dominated by the Bayesian approach. 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 and discussed properties of 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. 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 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 \( (\omega_1, \omega_2, ..., \omega_p) \) be a weight vector. The model average estimator considered by Hansen is of the form

\[
\hat{\Theta}_m = \sum_{p=1}^{P} \omega_p \begin{pmatrix} (H'_p H_p)^{-1} H'_p y \\ 0 \end{pmatrix}.
\] (1)

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

This paper develops a new method of weight choice for the frequentist model average (FMA) least squares estimator in a linear regression. Our set-up is a model containing a set of focus regressors, i.e., those we want in the model on theoretical or other grounds irrespective of statistical significance, and (possibly) a set of auxiliary regressors, i.e., those variables whose inclusion may be considered as optional. The model with the focus regressors only may be referred to as the narrow model; the extended models are those that contain some or all of the auxiliary regressors in addition to the focus regressors. With \( m \) auxiliary regressors, our set-up allows \( 2^m \) extended models to choose between. This is the same set-up used in a number of recent papers on pretesting (e.g., Magnus and Durbin, 1999; Danilov and Magnus, 2004, 2004a), and it bears similarity to the local misspecification set-up of Hjort and Claeskens (2003) which also distinguishes between narrow and extended models. The mandatory inclusion of focus regressors in the narrow model does not lead to any loss of generality since the focus regressors can in practice contain only an intercept term. 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 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 analytic finite sample justifications. Our approach is similar in spirit to that advocated by Leung and Barron (2006), except that Leung and Barron (2006) focuses on the risk bound of the combined estimator, whereas we provide an explicit weight choice criterion and 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 shall label as 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.

The presentation of this paper goes as follows. Section 2 describes the model and estimators. In Section 3, we derive an unbiased estimator of the finite sample MSE of the FMA estimator along with an investigation of the finite sample and asymptotic properties of the proposed criterion. Section 4 reports results of a large scale simulation study and Section 5 concludes. Proofs of theorems are contained in an appendix.

2 MODEL SET-UP AND ESTIMATORS

Consider the linear regression model,

\[ y = X\beta + Z\gamma + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \]  

(2)

where \( y(n \times 1) \) is a vector of observations, \( X(n \times k) \) and \( Z(n \times m) \) are non-random regressor matrices, and \( \beta(k \times 1) \) and \( \gamma(m \times 1) \) are parameter vectors. Additionally, we assume that \( (X, Z) \) has full column rank \( k + m \). Here, \( X \) contains the variables that must be included in the model, while \( Z \) contains the optional or doubtful variables included in the model for a variety of reasons.
such as that of producing a “better” estimator of $\beta$. There is no loss of generality with this set-up since the common situation of $X$ containing no regressor other than an intercept term is a special case of the set-up described here.

With $m$ auxiliary regressors in $Z$ there are $2^m$ models to choose between, leading to $2^m$ possible estimators of $\beta$. The fully restricted (i.e., $\gamma = 0$) and unrestricted estimators of $\beta$ are $\hat{\beta}_r = (X'X)^{-1}X'y$ and $\hat{\beta}_u = \hat{\beta}_r - Q\hat{\theta}$ respectively, where $Q = (X'X)^{-1}X'ZD$, $D = (Z'MZ)^{-1/2}$, $M = I_n - X(X'X)^{-1}X'$ and $\hat{\theta} = DZ'My$. Clearly, $\hat{\theta} \sim N(\theta, \sigma^2 I_m)$ with $\theta = D^{-1}\gamma$. The $i^{th}(1 \leq i \leq 2^m)$ partially restricted estimator of $\beta$ can be written as $\hat{\beta}_{(i)} = \hat{\beta}_r - QW_i\hat{\theta}$, where $W_i = I_m - P_i$, $P_i = DS_i (S_i^TD^2S_i)^{-1}S_i'D$ 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 $S_i^r = (I_{r_i} : 0)$ or a column permutation thereof (Danilov and Magnus, 2004). Further, let $\hat{\gamma} = D^2Z'My$ be the least squares estimator of $\gamma$ and $\hat{\sigma}^2 = (y - X\hat{\beta}_u - Z\hat{\gamma})'(y - X\hat{\beta}_u - Z\hat{\gamma})/(n - k - m)$ the least squares estimator of $\sigma^2$ under the unrestricted model. The $i^{th}$ partially restricted least squares estimator of $\gamma$ is $\hat{\gamma}_{(i)} = DW_i\hat{\theta}$ under the restriction $S_i^r\gamma = 0$.

Traditional model selection procedures pick the best model that can explain the data at hand based on some model assessment criteria. The investigator then proceeds as if this model has been decided upon \textit{a priori}. Inference conditional on the model chosen among several will clearly underestimate variability (Hjort and Claeskens, 2003, Danilov and Magnus, 2004). Model averaging overcomes this problem through combining estimators obtained from different models. A frequentist model average estimator of $\beta$ in (2) may be written as

$$
\hat{\beta}_f = \sum_{i=1}^{2^m} \lambda_i \hat{\beta}_{(i)},
$$

with weights satisfying $\lambda_i \geq 0$ and $\sum_{i=1}^{2^m} \lambda_i = 1$.

Here, we concentrate on random weights with $\lambda_i$ depending 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 $i^{th}$ partially restricted model. Then the AIC of the $i^{th}$ model is $AIC(i) = n \log(\hat{\sigma}_i^2) + 2(q_i + 1)$, where $\hat{\sigma}_i^2 = (y - X\hat{\beta}_{(i)} - Z\hat{\gamma}_{(i)})'(y - X\hat{\beta}_{(i)} - Z\hat{\gamma}_{(i)})/n$ is the maximum likelihood estimator of $\sigma^2$ under the $i^{th}$ 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 \( i^{th} \) model, we observe that the AIC depends on the data only through \( \hat{\theta} \) and \( \hat{\sigma}^2 \). In light of this we write the weights as \( \lambda_i = \lambda_i(\hat{\theta}, \hat{\sigma}^2) \). Now, writing 
\[
W = \sum_{i=1}^{m} \lambda_i W_i,
\]
we can re-write 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 in our notation as 
\[
y = H_\Theta + u, \quad H = (X, Z), \quad \Theta \text{ is a vector of coefficients}, \quad u = B\upsilon + \epsilon, B \text{ is a set of omitted regressors and } \upsilon \text{ the corresponding coefficient vector.}
\]
Hansen (2007) concentrated on the estimation of 
\[
\mu^* = H_\Theta + B\upsilon,
\]
and evaluated the performance of \( \hat{\Theta}_m \), the MMA estimator of \( \Theta \), in terms of the criterion 
\[
L = E\left[ (H\hat{\Theta}_m - \mu^*)(H\hat{\Theta}_m - \mu^*)' \right],
\]
which is the risk of the estimator of the prediction vector.

3 DERIVATION 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 frequentist model average estimators, especially in finite samples. To this end, we will first derive an unbiased estimator of the MSE of \( \hat{\beta}_f \) and the optimal model average weights according to our criterion are those that minimize the trace of the MSE estimate. With some appropriate modifications we then generalize the method to the derivation of the MSE estimator of the predictor 
\[
\hat{\mu}_f = H_\hat{\Theta}_f, \quad \text{where } \hat{\Theta}_f = (\hat{\beta}_f', \hat{\gamma}_f')' \quad \text{and } \hat{\gamma}_f = \sum_{i=1}^{m} \lambda_i(\hat{\theta}, \hat{\sigma}^2) \hat{\gamma}_i \text{ is the estimator of } \gamma \text{ corresponding to } \hat{\beta}_f.
\]

3.1 An unbiased estimator of the MSE of \( \hat{\beta}_f \)

Our principal result is given in the following theorem:

**THEOREM 1** Under model (2), an unbiased estimator of the MSE of the FMA estimator \( \hat{\beta}_f \) is given by

\[
\hat{\text{MSE}}(\hat{\beta}_f) = \hat{\sigma}^2 (X'X)^{-1} - \hat{\sigma}^2 QQ' + \left\{ Q(I_m - W)\hat{\theta} \right\} \otimes^2 + \Psi(\hat{\theta}, \hat{\sigma}^2) + \left\{ \Psi(\hat{\theta}, \hat{\sigma}^2) \right\}' \quad (4)
\]

where

\[
\left\{ Q(I_m - W)\hat{\theta} \right\} \otimes^2 = \left\{ Q(I_m - W)\hat{\theta} \right\} \left\{ Q(I_m - W)\hat{\theta} \right\}' \text{, and}
\]

\[
\Psi(\hat{\theta}, \hat{\sigma}^2) = \left( (n - k - m)/2 \right) (\hat{\sigma}^2)^{-(n-k-m)/2+1} \int_0^{\hat{\sigma}^2} t^{(n-k-m)/2-1} \Psi_1(\hat{\theta}, t) dt \quad (5)
\]
with
\[\Psi_1(\hat{\theta}, t) = Q \left\{ W + \sum_{i=1}^{2m} \left( \partial \lambda_i(\hat{\theta}, t) / \partial \hat{\theta} \right) \hat{\theta}' W_i \right\} Q'. \tag{6}\]

**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 \(\lambda_i's\) in (3) that minimize the trace of \(\hat{MSE}(\hat{\beta}_f)\). The trace MSE accuracy measure is equivalent to the risk of the estimator under a squared error loss function, also known as the weak MSE criterion (Wallace, 1972). Now, from (4), it is straightforward to show that the trace of \(\hat{MSE}(\hat{\beta}_f)\) is
\[\hat{R}(\hat{\beta}_f) = \hat{\sigma}^2 \text{tr}(X'X)^{-1} - \hat{\sigma}^2 \text{tr}(QQ') + \hat{\theta}'(I_m - W)Q'(I_m - W)\hat{\theta} + 2\hat{\sigma}^2 \text{tr} \left\{ \Psi(\hat{\theta}, \hat{\sigma}^2) \right\}. \tag{7}\]
One problem with (7), however, is that the relative complexity of the term \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) precludes its implementation in any straightforward way to obtain a weight choice for \(\hat{\beta}_f\). One way to avoid the cumbersome calculations of (7) caused by the term \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) is to replace \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) by an approximate quantity. This leads to the following theorem:

**Theorem 2** Under model (2), an approximately unbiased estimator of the trace of the MSE of the FMA estimator \(\hat{\beta}_f\) is given by
\[\hat{R}_a(\hat{\beta}_f) = \hat{\sigma}^2 \text{tr}(X'X)^{-1} - \hat{\sigma}^2 \text{tr}(QQ') + \hat{\theta}'(I_m - W)Q'(I_m - W)\hat{\theta} + 2\hat{\sigma}^2 \text{tr} \left\{ \Psi_1(\hat{\theta}, \hat{\sigma}^2) \right\}, \tag{8}\]
where \(\Psi_1(\hat{\theta}, \hat{\sigma}^2) = Q \left\{ W + \sum_{i=1}^{2m} \left( \partial \lambda_i(\hat{\theta}, \hat{\sigma}^2) / \partial \hat{\theta} \right) \hat{\theta}' W_i \right\} Q'. \]

Formula (8) is obtained by replacing \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) by \(\hat{\sigma}^2 \Psi_1(\hat{\theta}, \hat{\sigma}^2)\) in (7). The justification for this lies in the fact that under some mild conditions, \(E_{\hat{\sigma}^2} \{ \Psi(\hat{\theta}, \hat{\sigma}^2) \} = \sigma^2 E_{\hat{\sigma}^2} \{ \Psi_1(\hat{\theta}, \hat{\sigma}^2) \}\) (see the proof of Theorem 1 for details). Our numerical results (not reported here but available upon request) show that the values produced by \(\hat{\sigma}^2 \Psi_1(\hat{\theta}, \hat{\sigma}^2)\) typically accord with those of \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) closely.

Optimization based on formula (8) is difficult due to the complexity of the expression even though (8) can be readily used as a basis for comparison with other weighting schemes. In what
follows, we consider a general class of random weights in the form of

\[
\lambda_i(\hat{\theta}, \hat{\sigma}^2) = \frac{a^0(n - q_i)^b(\hat{\sigma}_i^2)^c}{\sum_{i=1}^{2m} a^0(n - q_i)^b(\hat{\sigma}_i^2)^c}, \tag{9}
\]

where \(a > 0\), \(b \geq 0\) and \(c \leq 0\) are constants. The weights based on (9) generalize in a natural way many of the weight choices commonly adopted in practice. The S-AIC weights correspond to \(a = e^{-1}, b = 0\) and \(c = -n/2\). The S-BIC weights result when \(a = n^{-1/2}, b = 0\) and \(c = -n/2\).

With slight modifications, the smoothed version of Hurvich and Tsai’s (1989) bias corrected AIC can be written as a special case of (9). In addition, the smoothed residual mean squares (RMS) weights (Bates and Granger, 1969) correspond to \(a = b = 1\) and \(c = -1\); in this case,

\[
\lambda_i(\hat{\theta}, \hat{\sigma}^2) = \frac{(n - q_i)^2(\hat{\sigma}_i^2)^{-1}}{\sum_{i=1}^{2m} (n - q_i)^2(\hat{\sigma}_i^2)^{-1}}.
\]

The smoothed average forecast mean squared errors weights almost correspond to \(a = 1, b = 2\) and \(c = -1\); in this case,

\[
\lambda_i(\hat{\theta}, \hat{\sigma}^2) = \frac{(n - q_i)^2(\hat{\sigma}_i^2)^{-1}}{\sum_{i=1}^{2m} (n - q_i)^2(\hat{\sigma}_i^2)^{-1}} \approx \frac{\left\{\frac{n}{(n - q_i)(n - q_i - 1)}\hat{\sigma}_i^2\right\}^{-1}}{\sum_{i=1}^{2m} \left\{\frac{n}{(n - q_i)(n - q_i - 1)}\hat{\sigma}_i^2\right\}^{-1}}.
\]

Let \(\lambda\) be a \(2^m \times 1\) vector comprising \(\lambda_i(\hat{\theta}, \hat{\sigma}^2), i = 1, \ldots, 2^m\). 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 \(g_{ij} = (\hat{\sigma}^2)^{-1}\hat{\theta}'W_iQ'Q(I_m - W_j)\hat{\theta}\), for \(i, j = 1, \ldots, 2^m\). Additionally, let \(g\) and \(\phi\) each be a \(2^m \times 1\) vector with \(g\) consisting of the diagonal elements of \(G\) and the \(i^{th}\) element of \(\phi\) being \(tr(QW_iQ')\), \(i = 1, \ldots, 2^m\). Recognizing that

\[
\partial \lambda_i(\hat{\theta}, \hat{\sigma}^2) / \partial \hat{\theta} = (2/n)c \lambda_i(\hat{\theta}, \hat{\sigma}^2) \{ (\hat{\sigma}^2)^{-1} (I_m - W_i) \\
- \sum_{j=1}^{2^m} \lambda_j(\hat{\theta}, \hat{\sigma}^2)(\hat{\sigma}_j^2)^{-1} (I_m - W_j) \}\hat{\theta}, \tag{10}
\]

putting (10) in \(\Psi(\hat{\theta}, \hat{\sigma}^2)\) and using (8), we have

\[
\widehat{R}_a(\hat{\beta}_f) = \hat{\sigma}^2 tr(X'X)^{-1} - \hat{\sigma}^2 tr(QQ') + X'L\lambda - (4/n)c \hat{\sigma}^2 \lambda'G\lambda
+ 2\hat{\sigma}^2 X'\phi + (4/n)c \hat{\sigma}^2 \lambda'g. \tag{11}
\]
The essential problem here is to select an appropriate weight vector $\lambda$ that minimizes (11). Let $\lambda^*$ be such a vector. We call the estimator $\hat{\beta}_f$ corresponding to $\lambda^*$ the optimal frequentist model average estimator (labeled as OPT estimator hereafter).

Before proceeding it is of some interest to mention a special case that arises as follows. Suppose we set $c = 0$ in (11). For this special case, if one considers just mixing $\hat{\beta}_u$ and $\hat{\beta}_r$ (i.e., all the regressors in $Z$ are either in or out), then minimizing (11) with respect to $\lambda$ with $c = 0$ leads to the estimator

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

which turns out to be the James–Stein type estimator studied in Kim and White (2001). In view of our theoretical results this estimator also has some optimal properties if one restricts attention to the sub-space of $c = 0$. Clearly, the OPT estimator that minimizes (11) (regardless of the value of $c$) has optimal properties among a broader class than the James–Stein type estimator in (12).

### 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 $H\Theta$. Denote the FMA estimator of $H\Theta$ as $\hat{\mu}_f = H \hat{\Theta}_f = H \sum_{i=1}^{m} \lambda_i(\hat{\theta}, \hat{\sigma}^2) \hat{\Theta}(i)$, where $\hat{\Theta}(i)$ is the $i^{th}$ partially restricted estimator of $\Theta$.

**Theorem 3** Under model (2), an unbiased estimator of the MSE of $\hat{\mu}_f$ is:

$$\text{MSE}(\hat{\mu}_f) = \hat{\sigma}^2 X'(X'X)^{-1}X' + \varphi\left(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)'\right) - \varphi\left(\hat{\theta}, \hat{\sigma}^2, XQ, (ZD)'\right)$$

$$- \varphi\left(\hat{\theta}, \hat{\sigma}^2, ZD, (XQ)'\right) + \varphi\left(\hat{\theta}, \hat{\sigma}^2, ZD, (ZD)'\right),$$

where

$$\varphi(\hat{\theta}, \hat{\sigma}^2, C_1, C_2) = -\hat{\sigma}^2 C_1C_2 + C_1\left((I_m - W)\hat{\theta}\right) \otimes C_2 + C_1 \Xi(\hat{\theta}, \hat{\sigma}^2)C_2$$

$$+ C_1 \left\{\Xi(\hat{\theta}, \hat{\sigma}^2)\right\}'C_2,$$

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

and

$$\Xi_1(\hat{\theta}, t) = W + \sum_{i=1}^{m} \frac{\partial \lambda_i(\hat{\theta}, t)}{\partial \hat{\theta}} \hat{\theta}' W_i.$$
Proof: see the Appendix.

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

$$
\hat{R}(\hat{\mu}_f) = k\hat{\sigma}^2 + \text{tr}\left\{\varphi(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\right\} - 2\text{tr}\left\{\varphi(\hat{\theta}, \hat{\sigma}^2, ZD, (XQ)')\right\} \\
+ \text{tr}\left\{\varphi(\hat{\theta}, \hat{\sigma}^2, ZD, (ZD)')\right\}
$$

(17)

Define $\Xi_1(\hat{\theta}, \hat{\sigma}^2) = W + \sum_{i=1}^{2m} \frac{\partial \lambda_i(\hat{\theta}, \hat{\sigma}^2)}{\partial \theta} \hat{\theta}' W_i$. Note that $E_{\hat{\sigma}^2} \left\{\Xi(\hat{\theta}, \hat{\sigma}^2)\right\} = \sigma^2 E_{\hat{\sigma}^2} \left\{\Xi_1(\hat{\theta}, \hat{\sigma}^2)\right\}$. To overcome the computational difficulties associated with (17), we replace $\Xi(\hat{\theta}, \hat{\sigma}^2)$ by $\hat{\sigma}^2 \Xi_1(\hat{\theta}, \hat{\sigma}^2)$ in (14). 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). After this substitution we obtain:

Theorem 4 Under model (2), an approximately unbiased estimator of the trace of the MSE of $\hat{\mu}_f$ is:

$$
\hat{R}_a(\hat{\mu}_f) = k\hat{\sigma}^2 + \text{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\} - 2\text{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, ZD, (XQ)')\} \\
+ \text{tr}\{\varphi_1(\hat{\theta}, \hat{\sigma}^2, ZD, (ZD)')\},
$$

(18)

where $\varphi_1(\hat{\theta}, \hat{\sigma}^2, C_1, C_2)$ has the same expression as $\varphi(\hat{\theta}, \hat{\sigma}^2, C_1, C_2)$, except that $\Xi(\hat{\theta}, \hat{\sigma}^2)$ is replaced everywhere by $\hat{\sigma}^2 \Xi_1(\hat{\theta}, \hat{\sigma}^2)$.

Again, as in (9), write $\lambda_i(\hat{\theta}, \hat{\sigma}^2) = a^0(n - q_i)^b(\hat{\sigma}^2)^c / \sum_{i=1}^{2m} a^0(n - q_i)^b(\hat{\sigma}^2)^c$. Noting that $P_j$ is symmetric idempotent and $(XQ - ZD)'(XQ - ZD) = I_m$, putting (10) in (18) and after performing some manipulations, we obtain

$$
\hat{R}_a(\hat{\mu}_f) = (k - m)\hat{\sigma}^2 + \lambda' \hat{\lambda} + 2\hat{\sigma}^2 \lambda' \hat{\phi} - (4/n)\hat{\sigma}^2 \lambda' \hat{G} \lambda,
$$

(19)

where $\hat{L} = (\hat{I}_{ij}), \hat{\phi} = (\hat{\phi}_i), \hat{G} = (\hat{g}_{ij}), \hat{I}_{ij} = \hat{\theta}' P_j \hat{\theta}, \hat{g}_{ij} = (\hat{\sigma}^2)^{-1} \hat{\theta}' (P_j - P_i) \hat{\theta}$, and $\hat{\phi}_i = m - r_i$.

Equation (19) provides an alternative selection criterion for the weight vector $\lambda$ when the prediction vector rather than the coefficient vector is the main subject of interest. The optimal weight vector $\lambda$ is the one that minimizes (19).

3.3 Asymptotic optimality of the OPT estimator

Our foregoing discussion centers on the finite sample justification of the OPT estimator. Here we enlarge the optimality consideration to large sample situations. Consider the general weight
expression in (9) and denote it as \( \lambda_i(a, b, c) \). Let \( \lambda(a, b, c) \) be the weight vector comprising \( \lambda_i(a, b, c), i = 1, \ldots, 2^m \). Now, consider the estimation of \( \mu = H \Theta \). Define \( L_n(\lambda(a, b, c)) = (\hat{\mu}_f - \mu)'(\hat{\mu}_f - \mu) \) the squared error loss, and \( R_n(\lambda(a, b, c)) = E(L_n) \) the corresponding risk. Denote \( D = \{(a, b, c)| a > 0, b \geq 0, -\bar{c} \leq c \leq 0\} \), where \( \bar{c} \) is a positive constant that can take on any value. Further, let \( \xi_n = \max_{1 \leq i \leq 2^m} \{E(H\hat{\Theta}_{(i)} - \mu)'(H\hat{\Theta}_{(i)} - \mu)\} \) be the maximum risk based on a single sub-model, \( \xi_n = \inf_{(a, b, c) \in D} R_n(\lambda(a, b, c)) \) and \((\hat{a}, \hat{b}, \hat{c}) = \arg \min_{(a, b, c) \in D} \hat{R}_n(\hat{\mu}_f)\).

**Theorem 5** When \( n \to \infty \), provided that the condition

\[
\xi_n^{-2} \zeta_n \to 0
\]

is satisfied, then

\[
\frac{L_n(\lambda(\hat{a}, \hat{b}, \hat{c}))}{\inf_{(a, b, c) \in D} L_n(\lambda(a, b, c))} \to^p 1.
\]

**Proof:** See the Appendix.

Theorem 5 states that subject to the fulfillment of a convergence condition, the large sample squared error associated with the OPT estimator converges in probability to the smallest achievable squared error of any estimator. This result is of obvious theoretical appeal but its validity depends crucially on the assumption \( \xi_n^{-2} \zeta_n \to 0 \) as \( n \to \infty \). In order for this assumption to hold, it is essential that \( \xi_n \to \infty \) as \( n \to \infty \). The latter condition, which simply means there is no finite approximately model for which the bias is zero, is usually fulfilled in practice. To explain, assume that \( n^{-1} \sum_{i=1}^n \mu_i^2 = O(1) \), then we have

\[
\hat{\sigma}_i^2 = (y'My - \hat{\theta}'W_i\hat{\theta})/n \leq y'y/n = (\mu'\mu + \varepsilon'\varepsilon + 2\mu'\varepsilon)/n = O_P(1)
\]

for any sub-model. Now, suppose that there exists a sub-model for which the bias is zero and let the \( \tau^{th} \) model be such a model. Further, let \( \ell_\tau \) a unit vector in which the \( \tau^{th} \) element is 1 and all other elements are 0. Then it can then be shown using (22) that for any \((a, b, c) \in D\), the weight vector \( \lambda(a, b, c) \) cannot tend to \( \ell_\tau \) in probability, otherwise \( \lambda_\tau(a, b, c) \to^p 1, \lambda_i(a, b, c) \to^p 0 \ (i \neq \tau) \) and consequently

\[
\frac{\lambda_i(a, b, c)}{\lambda_\tau(a, b, c)} = \frac{a^q(n - q_i)^b(\hat{\sigma}_i^2)^c}{a^q(n - q_\tau)^b(\hat{\sigma}_\tau^2)^c} \to^p 0.
\]
The last expression rules out the possibility of \( c \) being zero. Now, if \(-\bar{c} \leq c < 0\) and given that \( \hat{\sigma}_1^2 \) is a consistent estimator of \( \sigma^2 \), we must have \( \hat{\sigma}_1^2 \to_p \infty \). But this contradicts (22) which holds for estimators of \( \sigma^2 \) in all sub-models. This contradiction in turn implies that the weights in \( D \) cannot result in a model with no bias. Thus, the condition \( \xi_n \to \infty \) is in fact a very mild condition that is likely to be fulfilled in practice.

Now, if \( \xi_n \to \infty \) is true then \( \xi_n^{-2} \zeta_n \to 0 \) holds as long as \( \zeta_n \) converges to infinity at a rate slower than that of \( \xi_n^2 \) to infinity. Note that the convergence rate of \( \zeta_n \to \infty \) decreases when there is a reduced number of sub-models in the FMA estimator. Now, by removing some of the very poor models prior to combining, one can then reduce the number of sub-models in the OPT estimator. It seems desirable to have a model screening step that serves this purpose. The model screening step developed by Yuan and Yang (2005) may be useful in this regard.

4 SIMULATION EXPERIMENTS AND RESULTS

In this section, we report results of a large scale simulation study that compares the performance of the OPT estimator with the MMA estimator and the model average estimators based on S-AIC and S-BIC (hereafter we simply refer to the latter estimators as the S-AIC and S-BIC estimators). Example 1 is based on the model set-up given in equation (2), with the MMA estimator obtained using an ordering pattern of regressors decided \textit{apriori}, while all other estimators average across \( 2^m \) sub-models. Example 2 is based on a real data set taken from Danilov and Magnus (2004a). The main objective of Example 2 is to examine the extent to which different patterns of ordering the regressors affect the properties of the MMA estimator. Among other things, the results of Example 2 illustrate the advantage of the OPT estimator over the MMA estimators when the ordering patterns of regressors can result in markedly different risk performance of the MMA estimator. Example 3 follows the same setting as in Hansen (2007), where only nested sub-models are considered. In this case all FMA estimators combine estimates obtained from the nested sub-models only. The main purpose of Example 3 is to illustrate the performance of the OPT estimator when it combines models in the same manner as that of the MMA estimator.
Example 1 The data are generated from the model:

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

where \( x_1 = 1, x_j \sim N(0, 1) \) for \( j = 2, \ldots, 10 \), and \( e_i \sim N(0, 1) \). The error term \( e_i \) is independent of \( x'_{ji}s \), and all \( x'_{ji}s \) are independent of one another. Arbitrarily, we let \( x_1, x_2 \) and \( x_3 \) be the focus regressors and the others be auxiliary regressors. 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 \( \alpha = \frac{\text{var}(\sum_{j=4}^{10} \theta_j x_{ji})}{\text{var}(\sum_{j=1}^{3} \theta_j x_{ji})}, \) i.e., \( \alpha = \frac{c_1^2 c_2^2 (0.5^2 + 0.6^2 + 0^2 + 1^2 + 0.4^2 + 0.3^2 + 0.8^2)/(c_1^2 (3^2 + 4^2))}{c_2^2/10}. \) Note that the larger the value of \( c_2^2 \) (and hence \( \alpha \)), the higher the relative importance of the auxiliary regressors in the regression. The population \( R^2 = 25c_1^2 (1 + \alpha)/(1 + 25c_1^2 (1 + \alpha)) \) is controlled by the parameter \( c_1 \), where \( 25c_1^2 (1 + \alpha) = \text{var}(\sum_{j=1}^{10} \theta_j x_{ji}) \) is the variance of the linear combination of all regressors, focus and auxiliary. Sample size varies between \( n = 30, 80, 150 \) and \( 300 \), \( \alpha \) is set to 0.1, 0.5 and 0.9, and \( R^2 \) is set in the range of \([0.1, 0.9]\). With 7 auxiliary regressors, the OPT, S-AIC and S-BIC estimators average estimates across \( 2^7 \) models. In computing Hansen’s (2007) MMA estimator, we order the regressors as \( x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 \) and \( x_{10} \). The MMA estimator is then obtained by applying equation (1) with the model weights derived by minimizing the Mallows criterion (see equation (11) in Hansen, 2007).

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

\[ L^{(1)} = \left( \hat{y} - \sum_{j=1}^{10} \theta_j x_j \right)' \left( \hat{y} - \sum_{j=1}^{10} \theta_j x_j \right), \]

i.e., the predictive loss of \( y \). With the predictive loss as the penalty function, we obtain the OPT estimator by selecting the \( \lambda \) vector that minimizes (19). Our simulation experiment is based on 2000 replications. Results of risk comparisons are given in Figures 1-3. As in Hansen (2007), we normalize the risk by dividing by the risk of the infeasible optimal least squares estimator, i.e., the risk of the best-fitting model among the \( 2^7 \) models. Figures 1-3 reveal that the OPT estimator generally has better risk performance than the other three estimators no matter the value of \( n \) or \( \alpha \). 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 are found to 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 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 $\alpha$ are both small, the S-BIC estimator has the best while the OPT estimator has the worst performance; however, when $n$ is large or $\alpha$ is large, except when $R^2$ is very large or very small, the OPT estimator is the best estimator while the S-BIC is the worst. Results of cases not depicted here indicate very similar characteristics to Figure 4 and are available on request from the authors.

**Example 2** The second 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:

$$y_t = \beta_1 + \beta_2 PI_{t-2} + \beta_3 DI_{3t-1} + \beta_4 SPREAD_{t-1} + \gamma_1 YSP_{t-1} + \gamma_2 DIP_{t-1} + \gamma_3 PER_{t-1} + \gamma_4 DLEAD_{t-2} + \epsilon_t, \quad (23)$$

where $y_t$ is excess returns, $PI_{t-2}$ is annual inflation rate (lagged two periods), $DI_{3t-1}$ 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 contains 46 annual observations on each of the variables described above over the period 1956 - 2001. The data and their sources are given in 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 $PI_{t-2}$, $DJ3_{t-1}$, $SPREAD_{t-1}$ and the constant 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 needs not distinguish between focus and auxiliary regressors but must first order the regressors. Since the model contains 8 regressors including the constant term, there are $8! = 40320$ possible ways to order the regressors. After ordering, the MMA scheme averages over 8 models obtained by adding the 8 regressors one at a time to the regression model, whereas in practice the regressors are ordered based on some pre-conceived notions of the investigator. For this discussion we consider all 40320 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 dependent variables in each round of the Monte Carlo simulations are obtained by drawing 46 random disturbances with replacement from the residuals of the OLS estimation of (23). Denote the $l^{th}$ such sample as $e_l^*$ and the $l^{th}$ Monte Carlo sample is generated using $Y_l^* = H\Theta + e_l^*$. The experiment uses the OLS estimates of the coefficients in (23) as the true parameter vector $\Theta$. A total of $l = 100$ Monte Carlo samples are drawn and the OPT predictor $\hat{\mu}_f = H\hat{\Theta}_f$ and the MMA predictor $\hat{\mu}_m = H\hat{\Theta}_m$ are computed. There are altogether 40320 $\hat{\mu}_m^l$ $s$ depending on the ordering of regressors. It should be noted that no estimators are exactly the same among these 40320 $\hat{\mu}_m^l$.s. For each estimator of $H\Theta$, the risk under 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 $\hat{\mu}_m$ 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 $\hat{\mu}_m$, obtained by averaging across the risks of all
40320 \tilde{\mu}_m s, is 0.0788. Of the 40320 patterns of ordering considered, the MMA estimator results in higher risk than the OPT estimator in 31839 out of 40320 or 79% of cases. Clearly, the extent to which the ordering pattern of regressors affects the risk behaviour 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 in 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 data set there is 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.

Example 3 This example is based on the same setting as in Hansen (2007), that is

\[ y_i = \sum_{j=1}^{\infty} \theta_j x_{ji} + e_i, \quad x_1 = 1, \] all remaining \( x'_j s \) are \( N(0, 1) \), \( e_i \) is distributed as \( N(0, 1) \) and independent of \( x'_j s \) and all \( x'_j s \) are independent of one another, \( \theta_j = c_3 \sqrt{2\alpha_2} j^{-\alpha_2 - 1/2} \) 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 \( 1000, \) \( \alpha_2 \) 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 sub-models with the \( p^{th} \) sub-model comprising the first \( p \) regressors. All four model average estimators combine estimates across these \( P \) sub-models. Note that even though the OPT, S-AIC and S-BIC can potentially combine estimates from up to \( 2^P \) models, we only consider \( P \) nested models here - the purpose is to evaluate the the OPT estimator when all estimators are considered on a platform that is supposed to favour the MMA estimator. As in Hansen (2007), evaluation is based on the loss function

\[ L^{(3)} = \left( \hat{y} - \sum_{j=1}^{\infty} \theta_j x_j \right)' \left( \hat{y} - \sum_{j=1}^{\infty} \theta_j x_j \right). \]

We carried out this simulation experiment with 2000 replications. 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 is shown that the MMA estimator performs best. Results of the cases not depicted here have characteristics similar to those shown in Figure 5 and are available on request from the authors.

5 CONCLUDING REMARKS

There has been a quickening of interest in frequentist model averaging in the past few years. This article has suggested a new approach to select model weights for a linear regression FMA estimator. The proposed estimator has been shown to be quite promising and yield 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 appeals, 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 the regressors would affect the finite sample performance of the MMA estimator. The simulation results presented here suggest that the way by which the regressors are ordered is indeed a major determinant of the finite sample behaviour of the MMA estimator. For the experiment considered, the risks of the MMA estimators corresponding to different patterns of regressor ordering can differ markedly. Unfortunately, there is no obvious way to order the regressors in many practical situations. It is also reasonably unlikely that the investigator, in a given application of the linear regression model, should have prior knowledge of which pattern of ordering would ultimately attain the lowest possible risk. The OPT estimator requires no such prior ordering of regressors and has both asymptotic as well as analytic finite sample justifications. Another strong feature of our analytical framework is that it nests other weighting schemes such as the S-AIC and S-BIC as special cases.

Although our set-up assumes normal errors, similar weight choice methods could be developed for models with non-normal errors. This is an issue that certainly warrants further research.
mittedly, our approach requires combining estimates across $2^m$ sub-models. This may not be an issue when $m$ is small but the computational burden will quickly increase 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 in Magnus, Powell and Prüfer’s (2008) may be useful alternatives to direct computation. Recently, Hansen (2008) extended the idea of Mallows model averaging to forecast combinations. It remains a challenging endeavour to extend the OPT approach to an out-of-sample forecasting setting.

REFERENCES


APPENDIX: PROOFS OF THEOREMS

\textbf{A.1 Proof of Theorem 1}

Note that the MSE of $\hat{\beta}_f$ may be written as

\[
MSE(\hat{\beta}_f) = E \{ (\hat{\beta}_f - \beta)(\hat{\beta}_f - \beta)' \} = \sigma^2 (X'X)^{-1} + Q E \{ (W\hat{\theta} - \theta)(W\hat{\theta} - \theta)' \} Q'. \]  \tag{A.1}

\[
\text{where } \sigma^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - y_j)^2}{n(n-1)} \text{ and } Q = \sum_{i=1}^{n} w_i^{-1} x_i x_i'.
\]

\[
\text{Further, } \hat{\theta} = \hat{\beta}_f - \hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_i.
\]

\[
\text{and } \hat{\beta}_i = (X_i'X_i)^{-1} X_i' y_i - \hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \left( (X_i'X_i)^{-1} X_i' y_i - \hat{\beta} \right).
\]

\[
\text{Then, } \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i - \hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \left( (X_i'X_i)^{-1} X_i' y_i - \hat{\beta} \right) - \hat{\beta}.
\]

\[
\text{This gives } \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \left( (X_i'X_i)^{-1} X_i' y_i - \hat{\beta} \right) + \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_i.
\]

\[
\text{Hence, } \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_i = \hat{\beta}_f. \]
So, the MSE of $\hat{\beta}_f$ depends crucially on the MSE of $W\hat{\theta}$, and determining $\lambda$ that minimizes the MSE of $\hat{\beta}_f$ is the same as determining $\lambda$ such that the MSE of $W\hat{\theta}$ is minimized. By the definition of $W$, we can write

$$E \left\{ (W\hat{\theta} - \theta)(W\hat{\theta} - \theta)' \right\} = E \left\{ (W - I_m)\hat{\theta} \left[ (W - I_m)\hat{\theta}' \right]' \right\} + \sum_{i=1}^{2m} E \left\{ \lambda_i(\hat{\theta}, \hat{\sigma}^2)(\hat{\theta} - \theta)\hat{\theta}' \right\} (W_i - I_m)$$

$$+ \sum_{i=1}^{2m} (W_i - I_m) E \left\{ \lambda_i(\hat{\theta}, \hat{\sigma}^2)\hat{\theta}(\hat{\theta} - \theta)' \right\} + E \left\{ (\hat{\theta} - \theta)(\hat{\theta} - \theta)' \right\}.$$

Assuming that $\lambda_i(\hat{\theta}, \hat{\sigma}^2)$ are continuous functions with piecewise continuous partial derivatives with respect to $\hat{\theta}$, and noting that $\hat{\theta}$ and $\hat{\sigma}^2$ are independent, we can show using integration by parts that

$$E \left\{ \lambda_i(\hat{\theta}, \hat{\sigma}^2)(\hat{\theta} - \theta)\hat{\theta}' \right\} = \sigma^2 E \left\{ \lambda_i(\hat{\theta}, \hat{\sigma}^2)I_m + \left( \partial \lambda_i(\hat{\theta}, \hat{\sigma}^2) / \partial \hat{\theta} \right) \hat{\theta}' \right\},$$

providing that the expectation on the right-hand side exists. Hence

$$Q \left\{ \sum_{i=1}^{2m} E \left\{ \lambda_i(\hat{\theta}, \hat{\sigma}^2)(\hat{\theta} - \theta)\hat{\theta}' \right\} (W_i - I_m) \right\} Q'$$

$$= \sigma^2 Q \left\{ E \left( W + \sum_{i=1}^{2m} \left( \partial \lambda_i(\hat{\theta}, \hat{\sigma}^2) / \partial \hat{\theta} \right) \hat{\theta}' W_i \right) - I_m \right\} Q'$$

$$= \sigma^2 E \left\{ \Psi(\hat{\theta}, \hat{\sigma}^2) \right\} - \sigma^2 QQ'.$$  \(\text{(A.2)}\)

Equation (A.2) indicates that the MSE of $\hat{\beta}_f$ can be unbiasedly estimated if $\sigma^2$ known. Now, assume that $\partial \lambda_i(\hat{\theta}, \hat{\sigma}^2) / \partial \hat{\theta}$ is a continuous function in $\hat{\sigma}^2$, then we can again show by integration by parts that

$$E_{\hat{\sigma}^2} \left\{ \Psi(\hat{\theta}, \hat{\sigma}^2) \right\} = \sigma^2 E_{\hat{\sigma}^2} \left\{ \Psi_1(\hat{\theta}, \hat{\sigma}^2) \right\},$$

where $\Psi(\hat{\theta}, \hat{\sigma}^2)$ is defined in (5). The verification of this formula is easy and hence omitted. The approach used to derive $\Psi(\hat{\theta}, \hat{\sigma}^2)$ is similar to that adopted by Carter, Srivastava, Srivastava and Ullah (1990), or Giles and Srivastava (1991). Full details of derivation are available upon request from the authors. Taking these results together and replacing $\sigma^2$ by $\hat{\sigma}^2$ in (A.1) lead to the unbiased MSE estimator of $\hat{\beta}_f$ in (4).
A.2 Proof of Theorem 3

It is readily seen that

\[
\text{MSE}(\hat{\mu}_f) = E \left\{ (\hat{\mu}_f - H\hat{\Theta})(\hat{\mu}_f - H\hat{\Theta})' \right\}
\]

\[
= H \begin{pmatrix}
E(\hat{\beta}_f - \beta)(\hat{\beta}_f - \beta)'
& E(\hat{\gamma}_f - \gamma)(\hat{\beta}_f - \beta)'

E(\hat{\beta}_f - \beta)(\hat{\gamma}_f - \gamma)'
& E(\hat{\gamma}_f - \gamma)(\hat{\gamma}_f - \gamma)'
\end{pmatrix} H'.
\]  \hspace{1cm} (A.3)

It is also straightforward to show that

\[
E \left\{ (\hat{\gamma}_f - \gamma)(\hat{\gamma}_f - \gamma)' \right\} = DE \left\{ (W\hat{\theta} - \theta)(W\hat{\theta} - \theta)' \right\} D'
\]  \hspace{1cm} (A.4)

and

\[
E \left\{ (\hat{\gamma}_f - \gamma)(\hat{\beta}_f - \beta)' \right\} = -DE \left\{ (W\hat{\theta} - \theta)(W\hat{\theta} - \theta)' \right\} Q'.
\]  \hspace{1cm} (A.5)

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

\[
E \left\{ (W\hat{\theta} - \theta)(W\hat{\theta} - \theta)' \right\} = E \left\{ \varphi(\hat{\theta}, \hat{\sigma}^2, I_m, I_m) \right\} .
\]  \hspace{1cm} (A.6)

Equation (13) is obtained by using (A.6) in (A.4) and (A.5) and substituting the resultant expressions in (A.3).

A.3 Proof of Theorem 5

First, we prove that the sum of the first three terms of \(R_a(\hat{\mu}_f)\) in formula (19), i.e., \(\Pi(\lambda(a, b, c)) \equiv (k - m)\hat{\sigma}^2 + \lambda'\hat{L}\lambda + 2\hat{\sigma}^2\lambda'\phi\), is equal to \(C_a(\lambda(a, b, c))\) plus a term that is not related to the weight vector \(\lambda(a, b, c)\). Let \(V_i = [0_{r_x k} : S_i']\), then the restriction \(S_i'\gamma = 0\) reduces to \(V_i\Theta = 0\), under which the restricted estimator of \(\Theta\) is

\[
\hat{\Theta}_{(i)} = (H'\hat{H})^{-1}H'y - (H'\hat{H})^{-1}V_i'(V_i(H'\hat{H})^{-1}V_i')^{-1}V_i(H'\hat{H})^{-1}H'y.
\]  \hspace{1cm} (A.7)

It is well known that

\[
(H'\hat{H})^{-1} = \begin{pmatrix}
(X'X)^{-1} + QQ' & -QD \\
-DQ' & D^2
\end{pmatrix}.
\]  \hspace{1cm} (A.8)
from which we have
\[ S_i'(D^2S_j) = V_i(H'H)^{-1}V'_j, \] (A.9)
and
\[ S_i'(D^2Z'My) = S_i'(-D^2Z'X(X'X)^{-1}X'y + D^2Z'y) = V_i(H'H)^{-1}H'y. \] (A.10)

By the definitions of \( P_i \) and \( \hat{\theta} \),
\[ \hat{\theta}'P_iP_i\hat{\theta} = y'MZD^2S_i(S_i'D^2S_i)^{-1}S_i'D^2S_j(S_j'D^2S_j)^{-1}S_j'D^2Z'My. \] (A.11)

Combining (A.7), (A.9), (A.10) and (A.11), we get
\[ \bar{l}_{ij} = y'H(H'H)^{-1}V_i'(V_i(H'H)^{-1}V'_i)^{-1}V_j(H'H)^{-1}H' \]
\[ \times H(H'H)^{-1}V_j'(V_j(H'H)^{-1}V'_j)^{-1}V_j(H'H)^{-1}H'y \]
\[ = (H(H'H)^{-1}H'y - H\hat{\Theta}_{(i)})'(H(H'H)^{-1}H'y - H\hat{\Theta}_{(j)}) \]
\[ = (y - H\hat{\Theta}_{(i)})'(y - H\hat{\Theta}_{(j)}) - (y - H(H'H)^{-1}H'y)'(y - H(H'H)^{-1}H'y) \]
\[ = (y - H\hat{\Theta}_{(i)})'(y - H\hat{\Theta}_{(j)}) - (n - k - m)\hat{\sigma}^2. \] (A.12)

Define \( \bar{e} = (y - H\hat{\Theta}_{(1)}, ..., y - H\hat{\Theta}_{(2m)}) \). From (A.12), we get
\[ \lambda'\bar{L}\lambda = \lambda'\bar{e}\bar{e} - (n - k - m)\hat{\sigma}^2 \]
\[ = (y - H\hat{\Theta})'(y - H\hat{\Theta}) - (n - k - m)\hat{\sigma}^2 \]
\[ = C_n(\lambda(a, b, c)) - 2\hat{\sigma}^2\lambda'K - (n - k - m)\hat{\sigma}^2. \] (A.13)

By the definition of \( \bar{\phi} \),
\[ \lambda'\bar{\phi} = \lambda'K - k. \] (A.14)

Thus, from (A.13) and (A.14), we have
\[ \Pi(\lambda(a, b, c)) = C_n(\lambda(a, b, c)) - n\hat{\sigma}^2. \] (A.15)

From condition (20) and the fact that \( \varepsilon \sim N(0, \sigma^2 I_n) \), it can be shown that (for saving space, we omit the detailed proofs of following two results that are available on request for the authors)
\[ \sup_{(a,b,c)\in \mathcal{D}} \left| \frac{L_n(\lambda(a, b, c))}{R_n(\lambda(a, b, c))} - 1 \right| \rightarrow^p 0, \] (A.16)
and
\[
\sup_{(a,b,c)\in D} \left| C_n(\lambda(a, b, c)) - L_n(\lambda(a, b, c)) \right| R_n(\lambda(a, b, c)) - \|\varepsilon\|^2 \rightarrow^p 0. \tag{A.17}
\]

Now, similarly to the proof of Theorem 2 in Li (1987), for completing the proof, we need only to consider the last term of \( \hat{R}_n(\hat{\mu}_f) \) in formula (19); that is, we only need to prove
\[
\sup_{(a,b,c)\in D} \left| 4c\hat{\sigma}^2 \lambda' \hat{G} \lambda/n \right| R_n(\lambda(a, b, c)) \rightarrow^p 0. \tag{A.18}
\]

From \( \hat{\sigma}_i^2 = ((n - k - m)\hat{\sigma}_i^2 + \hat{\theta}' P_i \hat{\theta})/n \), we have
\[
\hat{\sigma}_i^2 \left| \hat{g}_{ij} \right|/n = \frac{\hat{\sigma}_i^2 \left| \hat{\theta}' P_i \hat{\theta} - \hat{\theta}' P_i \hat{\theta} \right|}{(n - k - m)\hat{\sigma}_i^2 + \hat{\theta}' P_i \hat{\theta}}.
\]

When \( \hat{\theta}' P_i P_j \hat{\theta} \geq 0,
\[
\hat{\sigma}_i^2 \left| \hat{g}_{ij} \right|/n \leq \frac{\hat{\sigma}_i^2 \hat{\theta}' P_j \hat{\theta} + \hat{\sigma}_i^2 \hat{\theta}' P_i \hat{\theta}}{(n - k - m)\hat{\sigma}_i^2 + \hat{\theta}' P_i \hat{\theta}} \\
\leq \frac{\hat{\sigma}_i^2 + \hat{\theta}' P_i \hat{\theta}}{n - k - m} \\
= \frac{(y - H \hat{\Theta}_{(i)})'(y - H \hat{\Theta}_{(j)}) - (n - k - m)\hat{\sigma}_i^2}{n - k - m}.
\]

When \( \hat{\theta}' P_i P_j \hat{\theta} < 0,
\[
\hat{\sigma}_i^2 \left| \hat{g}_{ij} \right|/n \leq \frac{\hat{\sigma}_i^2 \hat{\theta}' P_j \hat{\theta} - \hat{\sigma}_i^2 \hat{\theta}' P_i \hat{\theta}}{(n - k - m)\hat{\sigma}_i^2 + \hat{\theta}' P_i \hat{\theta}} \\
\leq \frac{\hat{\sigma}_i^2 - \hat{\theta}' P_i \hat{\theta}}{n - k - m} \\
= \frac{(y - H \hat{\Theta}_{(i)})'(y - H \hat{\Theta}_{(j)}) - (n - k - m)\hat{\sigma}_i^2}{n - k - m} + 2\hat{\sigma}_i^2.
\]

Combining (A.19) and (A.20), we obtain
\[
\left| \hat{\sigma}_i^2 \lambda' \hat{G} \lambda/n \right| \leq \frac{\lambda' \varepsilon \varepsilon' \lambda}{n - k - m} + 2\hat{\sigma}_i^2, \text{ and hence}
\]

\[
\sup_{(a,b,c)\in D} \left| 4c\hat{\sigma}^2 \lambda' \hat{G} \lambda/n \right| R_n(\lambda(a, b, c)) \leq \sup_{(a,b,c)\in D} \frac{4 |c| C_n(\lambda(\lambda(a, b, c))) - 2\hat{\sigma}_i^2 \lambda' K}{n - k - m R_n(\lambda(a, b, c))} + \sup_{(a,b,c)\in D} \frac{4 |c| 2\hat{\sigma}_i^2}{R_n(\lambda(a, b, c))}
\]

\[
\leq \sup_{(a,b,c)\in D} \frac{4 |c| C_n(\lambda(\lambda(a, b, c)))}{n - k - m R_n(\lambda(a, b, c))} + \frac{4 |c| 2\hat{\sigma}_i^2}{R_n(\lambda(a, b, c))}.
\]

\[
\leq \frac{4}{n-k-m} \sup_{(a,b,c) \in D} |c| \left| \frac{L_n(\lambda(a,b,c))}{R_n(\lambda(a,b,c))} - 1 \right|
\]
\[
+ \frac{4}{n-k-m} \sup_{(a,b,c) \in D} |c| \left| \frac{n\sigma^2}{R_n(\lambda(a,b,c))} \right|
\]
\[
+ \frac{4}{n-k-m} \sup_{(a,b,c) \in D} |c| \left| \frac{C_n(\lambda(a,b,c)) - L_n(\lambda(a,b,c)) - \|\varepsilon\|^2}{R_n(\lambda(a,b,c))} \right|
\]
\[
+ 8 \sup_{(a,b,c) \in D} |c| \frac{\sigma^2}{R_n(\lambda(a,b,c))} + 8 \sup_{(a,b,c) \in D} |c| \frac{\|\hat{\sigma}^2 - \sigma^2\|}{\xi_n (n-k-m)}
\]
\[
+ \frac{n\nu\bar{c}}{n-k-m} \left( \frac{\|\varepsilon\|^2 - n\sigma^2}{\xi_n} \right)
\]
\[
+ \frac{n\nu\bar{c}}{n-k-m} \sup_{(a,b,c) \in D} \left| \frac{C_n(\lambda(a,b,c)) - L_n(\lambda(a,b,c)) - \|\varepsilon\|^2}{R_n(\lambda(a,b,c))} \right|
\]
\[
+ 8 \frac{\sigma^2\bar{c}}{\xi_n} + 8 \frac{\varepsilon, (I_n - H(H'H)^{-1}H')\varepsilon - (n-k-m)\sigma^2}{\xi_n (n-k-m)}. \tag{A.21}
\]

Further, by Chebyshev’s inequality and Theorem 2 of Whittle (1960) and noting that \( \varepsilon \sim N(0, \sigma^2 I_n) \), we have, for any \( \delta > 0 \),
\[
\Pr \left\{ \frac{\|\varepsilon\|^2 - n\sigma^2}{n-k-m} > \delta \right\} \leq \frac{E(\|\varepsilon\|^2 - n\sigma^2)}{\delta^2(n-k-m)^2} \leq \frac{C_1 n}{\delta^2(n-k-m)^2} \to 0, \tag{A.22}
\]
and
\[
\Pr \left\{ \frac{|\varepsilon, (I_n - H(H'H)^{-1}H')\varepsilon - (n-k-m)\sigma^2|}{n-k-m} > \delta \right\}
\]
\[
\leq \frac{E(|\varepsilon, (I_n - H(H'H)^{-1}H')\varepsilon - (n-k-m)\sigma^2|)}{\delta^2(n-k-m)^2}
\]
\[
\leq \frac{C_2 (n-k-m)}{\delta^2(n-k-m)^2} \to 0, \tag{A.23}
\]

where \( C_1 \) and \( C_2 \) are two constants. Finally, combining (A.16), (A.17), (A.22), (A.23) and (A.21), we obtain (A.18), and this completes the proof.
Figure 1: Simulation study: Risk comparison under the $L^{(1)}$-loss when $\alpha = 0.1$. 
Figure 2: Simulation study: Risk comparison under the $L^{(1)}$-loss when $\alpha = 0.5$. 

\[ \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{n} \sum_{j=1}^{n} \left[ y_j - \hat{f}(x_j) \right] \right)^2 \]
Figure 3: Simulation study: Risk comparison under the $L^{(1)}$-loss when $\alpha = 0.9$. 

\[ \text{Risk} \]

\[ n=30 \]

\[ n=80 \]

\[ n=150 \]

\[ n=300 \]
Figure 4: Simulation study: Risk comparison under the $L^{(2)}$-loss.
Figure 5: Simulation study: Risk comparison under the $L^{(3)}$-loss.