Quantitative Economics 5 (2014), 495–530 1759-7331/20140495

Model averaging, asymptotic risk, and regressor groups

Bruce E. Hansen
University of Wisconsin

This paper examines the asymptotic risk of nested least-squares averaging estimators when the averaging weights are selected to minimize a penalized least-squares criterion. We find conditions under which the asymptotic risk of the averaging estimator is globally smaller than the unrestricted least-squares estimator. For the Mallows averaging estimator under homoskedastic errors, the condition takes the simple form that the regressors have been grouped into sets of four or larger. This condition is a direct extension of the classic theory of James–Stein shrinkage. This discovery suggests the practical rule that implementation of averaging estimators be restricted to models in which the regressors have been grouped in this manner. Our simulations show that this new recommendation results in substantial reduction in mean-squared error relative to averaging over all nested submodels. We illustrate the method with an application to the regression estimates of Fryer and Levitt (2013).

Keywords. Shrinkage, efficient estimation, averaging, risk.

JEL classification. C13.

1. Introduction

Model averaging is receiving growing attention in statistics and econometrics. Averaging is a smoothed extension of model selection, and substantially reduces risk relative to selection. The key issue is weight selection. A traditional approach to model selection is to minimize an information criterion that is an estimate of the risk of the selected estimator (e.g., Akaike (1973), Mallows (1973)). Similarly, averaging weights can be selected by minimizing an information criterion that is an estimate of the risk of the averaging estimator, as proposed in Hansen (2007). While the asymptotic nonparametric optimality of such estimators has been established, our understanding of the sampling distribution remains incomplete.

Following Hjort and Claeskens (2003), Schorfheide (2005), Saleh (2006), and Hansen (2013), this paper explores the asymptotic distribution and risk of nested averaging estimators in a local asymptotic framework where the coefficients are in a root-n neighborhood of 0. We derive the asymptotic distribution of a general class of averaging estimators that minimize a penalized least-squares criterion. We show that the asymptotic distribution can be written as a (nonlinear) function of the normal random vector

Bruce E. Hansen: behansen@wisc.edu
Research supported by the National Science Foundation. I thank a co-editor and three referees for excellent comments and suggestions.

DOI: 10.3982/QE332

that characterizes the unrestricted estimator. We then derive a representation for the asymptotic risk that is similar in form to those of shrinkage estimators (as presented, for example, in Lehmann and Casella (1998)). Using this representation, we derive sufficient conditions for the asymptotic risk of the averaging estimator to be globally smaller than the risk of the unrestricted estimator. We find that this condition is a simple generalization of the classic condition for shrinkage estimators. In particular, the Mallows averaging estimator of Hansen (2007) satisfies this condition under homoskedasticity if the regressors are grouped in sets of four or greater. This means that if we restrict attention to submodels that are differentiated by four or more regressors, we can guarantee that the Mallows averaging estimator will have reduced mean-squared error (MSE) relative to the least-squares estimator, regardless of the values of the coefficients or the distributions of the regressors or regression errors.

We find in a simple simulation experiment that this modified averaging estimator has substantially reduced risk relative to the standard averaging estimator as well as the least-squares estimator, and has much better risk performance than alternative methods such as the least absolute shrinkage and selection operator (Lasso) (Tibshirani (1996)), smoothed Akaike information criterion (AIC) (Burnham and Anderson (2002)), and approximate Bayesian model averaging (BMA).

The message from this analysis is simultaneously subtle yet profound. First, it reinforces our view that selection and averaging methods should be derived from rigorous theory, not from intuition or analogy. Second, it points to the need for careful examination of the submodels used for estimation. Rather than simply estimating every possible submodel, we should limit the number of submodels and enforce the constraint that the separation between each submodel be four coefficients or greater.

Nested model selection and averaging rests on the implicit assumption that the regressors are individually ordered, from "most relevant" to "least relevant." Similarly, our method requires that the regressors are groupwise ordered. In practice, it may be much easier to order regressors by groups rather than individually. For example, the difference between specifications may be whether or not all state dummy variables are included, which is a 50-member grouping. In this sense, our focus on groupwise ordering is somewhat attractive.

A limitation of our analysis is that it is critically confined to nested models. Nesting permits the application of Stein's lemma (Stein (1981)), which lies at the heart of our risk calculations. It would be greatly desirable to extend our results to the case of nonnested models, but it unclear how to do so.

This paper builds on an extensive literature. Stein (1956) first showed that a Gaussian estimator is inadmissible when the number of coefficients exceeds two. A feasible estimator with smaller risk than the Gaussian estimator was introduced by James and Stein (1961). Baranchick (1964) showed that a positive-part James–Stein estimator has even smaller risk. Efron and Morris (1973b) showed the close connection between Stein and empirical Bayes estimators. Akaike (1973), Mallows (1973), and Schwarz (1978) introduced information criteria suitable for model selection. Judge and Bock (1978) provided an extensive evaluation of the Stein-rule estimator in linear regression. Leamer (1978) proposed the method of Bayesian model averaging. Akaike (1979) proposed the expo-

nential AIC as an analog of Bayesian probability weights. Lehmann and Casella (1998) provided an excellent introduction to the theory of shrinkage estimation. Saleh (2006) is a recent review of statistical shrinkage methods.

The idea of grouping regressors for shrinkage has been investigated previously in the statistics literature, including Efron and Morris (1973a), Berger and Dey (1983), Dey and Berger (1983), and George (1986a, 1986b).

Model averaging is an extension of the idea of forecast combination, which was introduced by Bates and Granger (1969) and spawned a large literature. The idea of using Bayesian model averaging for forecast combination was pioneered by Min and Zellner (1993). Some excellent reviews include Clemen (1989), Diebold and Lopez (1996), Hendry and Clements (2004), Timmermann (2006), and Stock and Watson (2006). Related ideas are the empirical Bayes regressions of Knox, Stock, and Watson (2004), and the bagging method of Inoue and Kilian (2008).

Model averaging methods are receiving an explosion of interest in econometrics and statistics. Averaging methods for linear regression have been proposed by Buckland, Burnham, and Augustin (1997), Hjort and Claeskens (2003), Danilov and Magnus (2004), Hansen (2007), Hansen and Racine (2012), Liu (2012), and Liu and Okui (forthcoming). The theory has been further studied in Magnus (2002), Magnus, Powell, and Prüfer (2010), Wan, Zhang, and Zou (2010), Liang, Zou, Wan, and Zhang (2011), and McCloskey (2012). Averaging for instrumental variable and generalized method of moments estimation has been proposed by Kuersteiner and Okui (2010), Liao (2012), Lee and Zhou (2011), and DiTraglia (2013).

The remainder of the paper is organized as follows. Section 2 introduces the regression model and submodels. Section 3 introduces the submodel estimators. Section 4 presents the class of penalized weight criteria, and Section 5 rewrites the criteria using cumulative weights. Section 6 demonstrates the connection between the averaging estimator and James–Stein shrinkage. Section 7 presents the asymptotic distribution of the averaging estimator in the local asymptotic framework, and Section 8 calculates the asymptotic risk. Section 9 simplifies the conditions under bounded heteroskedasticity. Section 10 discusses weight selection under heteroskedasticity. Section 11 presents the results of simulation experiments. Section 12 is an empirical application to a regression example from Fryer and Levitt (2013). Section 13 presents a conclusion. Mathematical proofs are presented in the Appendix. Further simulation results are presented in a supplemental appendix, available on the journal website, http://qeconomics.org/supp/332/supplement.pdf. The replication codes for the simulation experiment and empirical application are also posted on the journal website, http://qeconomics.org/supp/332/code_and_data.zip.

2. Regression model

We have observations $\{y_i, \mathbf{x}_i : i = 1, \ldots, n\}$, where y_i is real-valued and $\mathbf{x}_i$ is $K \times 1$. The observations are assumed to satisfy the linear projection equation

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + e_i, \\ E(\mathbf{x}_i e_i) &= 0. \end{aligned} \tag{1}$$

In matrix notation, we write the equation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

We assume that the regressors can be partitioned into ordered groups as $\mathbf{x}_i = (\mathbf{x}_{1i}', \mathbf{x}_{2i}', \ldots, \mathbf{x}_{Mi}')'$, where $\mathbf{x}_{ji}$ is $k_j \times 1$ and the total number of regressors is $K = k_1 + \cdots + k_M$. We then consider M nested submodels, where the mth can be written as

$$\begin{aligned} y_i &= \sum_{j=1}^{m} \mathbf{x}_{mi}'\boldsymbol{\beta}_j + e_i(m) \\ &= \overline{\mathbf{x}}_{mi}'\overline{\boldsymbol{\beta}}_m + e_i(m). \end{aligned}$$

That is, the mth submodel includes the regressors $\mathbf{x}_{1i}$ through $\mathbf{x}_{mi}$ and excludes the remaining regressors.

In matrix notation,

$$\begin{aligned} \mathbf{y} &= \sum_{j=1}^{m} \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{e}(m) \\ &= \overline{\mathbf{X}}_m\overline{\boldsymbol{\beta}}_m + \mathbf{e}(m). \end{aligned} \tag{2}$$

Note that the mth submodel has

$$K_m = k_1 + \cdots + k_m \tag{3}$$

regressors and that the regressors $\mathbf{x}_{1i}$ are included in all models. Notationally, we allow $k_1 = 0$, in which case there is no $\mathbf{x}_{1i}$ and model 1 is the zero vector. Note that the error in equation (2) is not a projection error, as the coefficients are defined in the full model (1) and thus have common meaning across models. Another way of thinking about this is that equation (2) has omitted variables.

The ordering of the groups is important, as $\mathbf{x}_{1i}$ is included in all submodels, $\mathbf{x}_{2i}$ is included in all submodels except for model 1, and so on. Thus it is prudent for the user to construct the ordering so that the variables expected to be most relevant are included in the first groups, and those expected to be least relevant are included in the final groups. If the regressors have been standardized to have zero mean and common variance, then it would be ideal if the regressors are ordered so that their coefficients are descending in absolute value. The performance of our averaging estimator will depend on this ordering, in the sense that the efficiency gains will be greatest when the regressors have been so ordered. However, for all of our theoretical results, we do not impose any assumption on the ordering; it is not required to be "correct" in any sense.

3. Estimation

The unconstrained least-squares estimator of $\boldsymbol{\beta}$ in the full model is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

with residual $\widehat{e}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ or in vector notation as $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$.

The standard estimator of the coefficient vector $\boldsymbol{\beta}$ in the mth submodel is the least-squares estimate of $\mathbf{y}$ on the included regressors $\overline{\mathbf{X}}_m$. Notationally, let

$$\mathbf{S}_m = \begin{pmatrix} \mathbf{I}_{K_m} \\ \mathbf{0} \end{pmatrix} \tag{4}$$

be the $K \times K_m$ matrix that selects the included regressors; thus $\overline{\mathbf{X}}_m = \mathbf{X}\mathbf{S}_m$. The least-squares estimate of $\overline{\boldsymbol{\beta}}_m$ in the mth submodel is

$$\widehat{\overline{\boldsymbol{\beta}}}_m = \left(\overline{\mathbf{X}}_m'\overline{\mathbf{X}}_m\right)^{-1}\overline{\mathbf{X}}_m'\mathbf{y}$$

and that of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_m = \mathbf{S}_m\widehat{\overline{\boldsymbol{\beta}}}_m = \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{X}'\mathbf{X}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{X}'\mathbf{y}.$$

The corresponding residual is $\widehat{e}_{mi} = y_i - \overline{\mathbf{x}}_{mi}'\widehat{\overline{\boldsymbol{\beta}}}_m = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_m$, and in vector notation as $\widehat{\mathbf{e}}_m = \mathbf{y} - \overline{\mathbf{X}}_m\widehat{\overline{\boldsymbol{\beta}}}_m = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_m$. Note that since model M contains all regressors, then $\widehat{\boldsymbol{\beta}}_M = \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and $\widehat{\mathbf{e}}_M = \widehat{\mathbf{e}}$.

An averaging estimator of $\boldsymbol{\beta}$ is a weighted average of the submodel estimates. Let $\mathbf{w} = (w_1, w_2, \ldots, w_M)$ be a weight vector. We require that $w_m \geq 0$ and $\sum_{m=1}^M w_m = 1$, and thus is an element of the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H} = \left\{\mathbf{w}\colon w_m \geq 0,\ \sum_{m=1}^M w_m = 1\right\}. \tag{5}$$

An averaging estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^M w_m\widehat{\boldsymbol{\beta}}_m. \tag{6}$$

The residual from the averaging estimator is

$$\widehat{e}_i(w) = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^M w_m\widehat{e}_{mi}$$

or in vector notation

$$\widehat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^M w_m\widehat{\mathbf{e}}_m.$$

4. Penalized weight criteria

A general class of penalized least-squares criteria take the form

$$C_n(\mathbf{w}) = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2\sum_{m=1}^M w_m\widetilde{T}_m. \tag{7}$$

The leading term

$$\widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) = \sum_{i=1}^{n} \widehat{e}_i(w)^2 = \sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \sum_{i=1}^{n} \widehat{e}_{mi}\widehat{e}_{\ell i}$$

is the sum of squared residuals from the averaging estimator.

The constants $\widetilde{T}_m$ are (possibly data-dependent) penalties satisfying $\widetilde{T}_1 < \widetilde{T}_2 < \cdots < \widetilde{T}_M$. For example, the Mallows averaging criterion (Hansen (2007)) sets $\widetilde{T}_m = s^2 K_m$, where K_m is the number of coefficients in model m as defined in (3), and

$$s^2 = \frac{1}{n-K}\widehat{\mathbf{e}}'\widehat{\mathbf{e}} \tag{8}$$

is the standard estimator of the unconditional error variance $\sigma^2 = E(e_i^2)$. Thus the Mallows averaging criteria is

$$C_n^{\text{Mallows}}(\mathbf{w}) = \widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w}) + 2s^2\sum_{m=1}^{M} w_m K_m. \tag{9}$$

Given the criterion (7), the selected weight vector $\widehat{\mathbf{w}}$ is the element of the unit simplex that minimizes (7):

$$\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_M) = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{H}} C_n(\mathbf{w}). \tag{10}$$

The averaging estimator (6) computed with the weights (10) is then

$$\widehat{\boldsymbol{\beta}}_A = \sum_{m=1}^{M} \widehat{w}_m \widehat{\boldsymbol{\beta}}_m. \tag{11}$$

For the weight vector $\widehat{\mathbf{w}}^{\text{MMA}}$ which minimizes the Mallows averaging criteria (9), Hansen's (2007) Mallows model averaging (MMA) estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{MMA}} = \sum_{m=1}^{M} \widehat{w}_m^{\text{MMA}} \widehat{\boldsymbol{\beta}}_m. \tag{12}$$

5. Cumulative weight criteria

It turns out that there is a convenient alternative representation of the averaging estimator (11) in terms of the cumulative weights:

$$w_m^* = w_1 + \cdots + w_m.$$

Set $\mathbf{w}^* = (w_1^*, \ldots, w_M^*)$. Notice that $\mathbf{w} \in \mathcal{H}$ is equivalent to $\mathbf{w}^* \in \mathcal{H}^*$, where

$$\mathcal{H}^* = \left\{\mathbf{w}^* : 0 \le w_1^* \le w_2^* \le \cdots \le w_M^* = 1\right\}.$$

Similarly, define the selected cumulative weights

$$\widehat{w}_m^* = \widehat{w}_1 + \cdots + \widehat{w}_m$$

and set

$$\widehat{\mathbf{w}}^* = \big(\widehat{w}_1^*, \ldots, \widehat{w}_M^*\big). \tag{13}$$

We can equivalently discuss averaging in terms of the weights $\mathbf{w} \in \mathcal{H}$ or cumulative weights $\mathbf{w}^* \in \mathcal{H}^*$. Notice that (6) is equivalent to

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \sum_{m=1}^{M-1} w_m^* (\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m) \tag{14}$$

and (11) is equivalent to

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \sum_{m=1}^{M-1} \widehat{w}_m^* (\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m). \tag{15}$$

The representation in terms of the cumulative weights $\mathbf{w}^*$ is convenient, as the penalized least-squares criterion (7) can be written as a simple function of $\mathbf{w}^*$. Define the marginal penalty for model $m+1$ as

$$\widetilde{t}_{m+1} = \widetilde{T}_{m+1} - \widetilde{T}_m.$$

Note, for example, that for the Mallows criterion, we have $\widetilde{t}_{m+1} = s^2 k_{m+1}$. Also, let $L_m = \widehat{\mathbf{e}}_m' \widehat{\mathbf{e}}_m$ denote the sum of squared residuals in the mth model.

LEMMA 1. *For the penalized criterion* (7),

$$C_n(\mathbf{w}) = C_n^*\big(\mathbf{w}^*\big) + L_M + 2\widetilde{T}_M,$$

where

$$C_n^*\big(\mathbf{w}^*\big) = \sum_{m=1}^{M-1} \big(w_m^{*2}(L_m - L_{m+1}) - 2w_m^* \widetilde{t}_{m+1}\big). \tag{16}$$

Hence

$$\widehat{\mathbf{w}}^* = \operatorname*{argmin}_{\mathbf{w}^* \in \mathcal{H}^*} C_n^*\big(\mathbf{w}^*\big). \tag{17}$$

Lemma 1 shows that $C_n(\mathbf{w})$ and $C_n^*(\mathbf{w}^*)$ are equivalent up to the term $L_M + 2\widetilde{T}_M$, which does not depend on the weight vector and thus the cumulative weights (13) are the minimizers of (16). We call $C_n^*(\mathbf{w}^*)$ the cumulative criterion. It is a simple function of $\mathbf{w}^*$, as it is quadratic with no cross-terms.

The representation (15)–(17) turns out to be useful because it facilitates an asymptotic distribution theory for the averaging estimator, as we show in Section 7.

6. James–Stein shrinkage

Consider the case of two submodels so $M = 2$ and for simplicity suppose $k_1 = 0$. In this case, write $w = w_1 = w_1^*$ and $\widetilde{t} = \widetilde{t}_2$ so that (16) equals

$$\begin{aligned} C_n^*(w) &= w^2(L_1 - L_2) - 2w\widetilde{t} \\ &= w^2\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} - 2w\widetilde{t}, \end{aligned}$$

where the second line uses the substitution $L_1 - L_2 = \mathbf{y}'\mathbf{y} - \widehat{\mathbf{e}}'\widehat{\mathbf{e}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$.

The solution (17) minimizes this function subject to the constraint $0 \le w \le 1$, and equals

$$\widehat{w} = \begin{cases} \dfrac{\widetilde{t}}{\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}}, & \text{if } \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \ge \widetilde{t}, \\ 1, & \text{otherwise.} \end{cases}$$

It follows that the averaging estimator (15) equals

$$\widehat{\boldsymbol{\beta}}_A = \left(1 - \frac{\widetilde{t}}{\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}}\right)_+ \widehat{\boldsymbol{\beta}}, \tag{18}$$

where $(a)_+ = a1_{a\ge 0}$ is the positive part operator.

This averaging estimator (18) is the classic James–Stein estimator with shrinkage parameter $\widetilde{t}$. That is, when there are two models, averaging estimators whose weights minimize penalized least-squares criteria of the form (7) are numerically identical to the James–Stein estimator. This is fascinating as it shows that averaging estimators are in the class of shrinkage estimators. Furthermore, note that the classic James–Stein recommendation was to set $\widetilde{t} = s^2(K_n - 2)$, while the Mallows criterion sets $\widetilde{t} = s^2K_n$. This is a modest difference for small K_n, and is quite minor when K_n is large.

7. Asymptotic distribution

We require that the least-squares estimator is asymptotically normal. The following condition is sufficient for our needs.

Assumption 1.

1. *Either* (a) $\{y_i, \mathbf{x}_i\}$ *is independent and identically distributed* (*i.i.d.*) *with finite fourth moments or* (b) $\{y_i, \mathbf{x}_i\}$ *is a strictly stationary and ergodic time series with finite* $q > 4$ *moments and* $E(e_i|\mathcal{F}_{i-1}) = 0$, *where* $\mathcal{F}_{i-1} = \sigma(\mathbf{x}_i, \mathbf{x}_{i-1}, \ldots; e_{i-1}, e_{i-2}, \ldots)$.
2. $\mathbf{Q} = E(\mathbf{x}_i\mathbf{x}_i') > 0$.

The following conditions are required to obtain an asymptotic distribution for the penalized weight criterion.

Assumption 2. *As* $n \to \infty$,

1. $\widetilde{T}_m \xrightarrow{p} T_m$ *for* $m = 1, \ldots, M$,
2. $n^{1/2}\boldsymbol{\beta}_m \to \boldsymbol{\delta}_m$ *for* $m = 2, \ldots, M$.

Assumption 1 states that the penalties in (7) converge in probability to constants. For example, in the case of Mallows averaging with i.i.d. data, $\widetilde{T}_m = s^2 K_m \xrightarrow{p} T_m = \sigma^2 K_m$.

Assumption 2 is a local asymptotic framework and it specifies the coefficients $\boldsymbol{\beta}_m$ to be in a local $n^{-1/2}$ neighborhood of 0. The coefficients $\boldsymbol{\beta}_1$ are not included in Assumption 2 since these variables are included in all models.

The local asymptotic framework is a technical device commonly seen in model selection and averaging asymptotic theory, for example, Hjort and Claeskens (2003) and Schorfheide (2005). It allows the application of asymptotic theory, for in a constant parameter model, the largest model will always dominate and the asymptotic analysis will not produce a useful approximation. Alternatively, Assumption 2 could be omitted and replaced by the assumption that the errors e_i are i.i.d. $\mathrm{N}(0, \sigma^2)$ with known σ^2, in which case the results described below are exact and finite sample, rather than asymptotic. The virtue of the local asymptotic framework of Assumption 2 is that it does not require i.i.d. normality and thus allows application to the wide variety of practical econometric applications. It is not a practical restriction.

THEOREM 1. *Under Assumptions 1 and 2, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}), \tag{19}$$

where

$$\begin{aligned} &\mathbf{V} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}, \\ &\boldsymbol{\Omega} = E\big(\mathbf{x}_i\mathbf{x}_i' e_i^2\big), \end{aligned} \tag{20}$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{Z} - \eta(\mathrm{Z} + \boldsymbol{\delta}), \tag{21}$$

where $\boldsymbol{\delta} = (\boldsymbol{\delta}_1', \boldsymbol{\delta}_2', \ldots, \boldsymbol{\delta}_M')'$ *with* $\boldsymbol{\delta}_1 = 0$,

$$\eta(\mathbf{x}) = \sum_{m=1}^{M-1} w_m^*(\mathbf{x})(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}\mathbf{x} \tag{22}$$

with

$$\mathbf{P}_m = \mathbf{S}_m\big(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\big)^{-1}\mathbf{S}_m', \tag{23}$$

$$\mathbf{w}^*(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{w}^* \in \mathcal{H}^*} C^*\big(\mathbf{w}^*, \mathbf{x}\big), \tag{24}$$

and

$$C^*\big(\mathbf{w}^*, \mathbf{x}\big) = \sum_{m=1}^{M-1} \big(w_m^{*2}\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}\mathbf{x} - 2w_m^* t_{m+1}\big) \tag{25}$$

with

$$t_{m+1} = T_{m+1} - T_m.$$

The function $\eta(\mathbf{x})$ is absolutely continuous, and $\mathbf{w}^*(\mathbf{x})$ satisfies the following characterization. For all $\mathbf{x}$, there exists an integer $J(\mathbf{x}) \le M$ and index set $\{m_1(\mathbf{x}), \ldots, m_J(\mathbf{x})\} \subset \{1, \ldots, M\}$ such that for $j = 1, \ldots, J(\mathbf{x})$,

$$w_\ell^*(\mathbf{x}) = \frac{T_{m_{j+1}(\mathbf{x})} - T_{m_j(\mathbf{x})}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}(\mathbf{x})} - \mathbf{P}_{m_j(\mathbf{x})})\mathbf{Q}\mathbf{x}}, \quad m_j(\mathbf{x}) \le \ell < m_{j+1}(\mathbf{x}), \tag{26}$$

and

$$w_\ell^*(\mathbf{x}) = 1, \quad m_J(\mathbf{x}) \le \ell \le M. \tag{27}$$

The index set $\{m_1(\mathbf{x}), \ldots, m_J(\mathbf{x})\}$ has the property that if $m_J(\mathbf{x}) < M$, then

$$\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J(\mathbf{x})})\mathbf{Q}\mathbf{x} \le T_M - T_{m_J(\mathbf{x})}. \tag{28}$$

The main contribution of Theorem 1 is (21), which is a representation of the asymptotic distribution of the averaging estimator as a (nonlinear) function of the limiting normal random vector Z. The characterization of this function in (22)–(28) will allow us to apply Stein's lemma to calculate the estimator's asymptotic risk. In addition, the asymptotic distribution (21) may be useful for alternative purposes such as inference.

The representation (21)–(22) shows that the weights are asymptotically random functions of the limiting distribution (19) plus the localizing parameters $\boldsymbol{\delta}$. The characterization of the weights in (26)–(28) shows that given the random variable $\mathbf{x} = \mathrm{Z} + \boldsymbol{\delta}$, there is a set of models $\{m_1(\mathbf{x}), \ldots, m_J(\mathbf{x})\}$ that receive positive weight and the remaining models receive zero weight. The set of models that receive positive weight is random (depends on Z), but largely influenced by the localizing parameters $\boldsymbol{\delta}$.

8. Asymptotic risk

The asymptotic trimmed risk or weighted mean-squared error (MSE) of an estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is

$$R(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} E \min\left\{ n(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Q}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \zeta \right\}. \tag{29}$$

While we use the matrix $\mathbf{Q}$ to weight the estimates, in principle other weight matrices could be used. The choice $\mathbf{Q}$ is particularly convenient for two reasons. One, it induces invariance to parameter scaling and rotation. Two, with this choice, the MSE function (29) plus σ^2 corresponds to out-of-sample mean-squared forecast error (under stationarity), which is a natural risk measure in time-series applications. The trimming in (29) conveniently avoids the need to establish uniform integrability.

When $\widetilde{\boldsymbol{\beta}}$ has an asymptotic distribution, that is, $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \psi$, then the asymptotic trimmed risk equals $E(\psi'\mathbf{Q}\psi)$ and is thus straightforward to calculate. For example,

the unrestricted least-squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ from (19) has the asymptotic distribution $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$. Thus its asymptotic trimmed risk is

$$R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta}) = E\big(\mathrm{Z}'\mathbf{Q}\mathrm{Z}\big) = \mathrm{tr}(\mathbf{QV}) = \mathrm{tr}\big(\mathbf{Q}^{-1}\boldsymbol{\Omega}\big).$$

We are now in a position to calculate the asymptotic risk of the averaging estimator. It will be convenient to define the matrices

$$\begin{aligned}\mathbf{A}_m &= \big(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\big)^{-1}\mathbf{S}_m'\boldsymbol{\Omega}\mathbf{S}_m \\ &= \big(E\big(\overline{\mathbf{x}}_{mi}\overline{\mathbf{x}}_{mi}'\big)\big)^{-1}E\big(\overline{\mathbf{x}}_{mi}\overline{\mathbf{x}}_{mi}'e_i^2\big)\end{aligned}$$

and the constants

$$D_m = \mathrm{tr}(\mathbf{A}_m) = E\big(\overline{\mathbf{x}}_{mi}'\mathbf{Q}_m^{-1}\overline{\mathbf{x}}_{mi}e_i^2\big), \tag{30}$$

where $\mathbf{Q}_m = E(\overline{\mathbf{x}}_{mi}\overline{\mathbf{x}}_{mi}')$. As we discuss in the next section, under conditional homoskedasticity, we have the simplifications $\boldsymbol{\Omega} = \mathbf{Q}\sigma^2$ and $D_m = \sigma^2 K_m$, so D_m is a measure of the number of coefficients adjusting for heteroskedasticity.

THEOREM 2. *Under Assumptions 1 and 2,*

$$\begin{aligned}R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta}) &= \mathrm{tr}\big(\mathbf{Q}^{-1}\boldsymbol{\Omega}\big), \\ R(\widehat{\boldsymbol{\beta}}_{A}, \boldsymbol{\beta}) &= \mathrm{tr}\big(\mathbf{Q}^{-1}\boldsymbol{\Omega}\big) - E\big(q(\mathrm{Z} + \boldsymbol{\delta})\big),\end{aligned} \tag{31}$$

where

$$\begin{aligned}q(\mathbf{x}) = &\sum_{j=1}^{J-1} \frac{(T_{m_{j+1}} - T_{m_j})[2(D_{m_{j+1}} - D_{m_j}) - (T_{m_{j+1}} - T_{m_j})]}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}} \\ &- 4\sum_{j=1}^{J-1} \frac{(T_{m_{j+1}} - T_{m_j})}{(\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x})^2} \\ &\times \mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\boldsymbol{\Omega}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x} \\ &+ \big[2(D_M - D_{m_J}) - \mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}\big]\mathbf{1}_{(m_J < M)}.\end{aligned} \tag{32}$$

Equations (31)–(32) give an expression for the asymptotic risk of the averaging estimator. We now use this expression to show that its risk is smaller than the unrestricted least-squares estimator under a mild condition.

Let $\lambda_{\max}(\mathbf{A})$ denotes the largest eigenvalue of a symmetric matrix $\mathbf{A}$ and define

$$\overline{\lambda} = \lambda_{\max}\big(\mathbf{Q}^{-1/2}\boldsymbol{\Omega}\mathbf{Q}^{-1/2}\big) \tag{33}$$

and

$$d_m = D_m - D_{m-1}. \tag{34}$$

ASSUMPTION 3. *For all* $m \geq 2$,

(a) $d_m > 2\overline{\lambda}$,

(b) $0 < t_m \leq 2(d_m - 2\overline{\lambda})$.

THEOREM 3. *Under Assumptions* 1, 2, *and* 3,

$$R(\widehat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta}). \tag{35}$$

Theorem 3 states that the averaging estimator has smaller asymptotic trimmed risk than the unrestricted estimator. This holds regardless of the values of the coefficients $\boldsymbol{\beta}$ or other characteristics of the data distribution. Notice in particular that the result does not depend on the ordering of the regression groups. That is, (35) holds even if the groups have been improperly ordered. This shows that in a quite generic sense, the averaging estimator $\widehat{\boldsymbol{\beta}}_A$ dominates the least-squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$.

The key condition is Assumption 3. Condition (a) states that the number of coefficients in each regression group, adjusted for heteroskedasticity, is sufficiently large. Condition (b) states that the marginal penalties t_m are all positive, but not too large. Notice that condition (a) is necessary for condition (b) to be feasible.

Notice that Assumption 3 does not impose any conditions on the regressors $\mathbf{x}_{1i}$ that are included in all models. The set of such regressors can have any dimension, including 0, and still satisfy Assumption 3.

9. BOUNDED HETEROSKEDASTICITY

We can simplify Assumption 3 in a leading case of interest. Define the conditional variance function

$$E\left(e_i^2 | \mathbf{x}_i = \mathbf{x}\right) = \sigma^2(\mathbf{x}),$$

its maximal and minimal values

$$\overline{\sigma}^2 = \sup_{\mathbf{x}} \sigma^2(\mathbf{x}),$$

$$\underline{\sigma}^2 = \inf_{\mathbf{x}} \sigma^2(\mathbf{x}),$$

and the "variance ratio"

$$r = \frac{\overline{\sigma}^2}{\underline{\sigma}^2}.$$

In the leading case of conditional homoskedasticity, $E(e_i^2 | \mathbf{x}_i) = \sigma^2$, and then $\overline{\sigma}^2 = \underline{\sigma}^2$ and $r = 1$. The deviation of r from 1 measures the degree of heteroskedasticity.

We now give a sufficient condition for Assumption 3.

ASSUMPTION 4. $r < \infty$ *and for all* $m \geq 2$,

(a) $k_m > 2r$,

(b) $0 < t_m \leq 2(k_m - 2r)\underline{\sigma}^2$.

LEMMA 2. *Assumption 4 implies Assumption 3.*

COROLLARY 1. *Under Assumptions 1, 2, and 4, the averaging estimator (11) satisfies* $R(\widehat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})$.

Assumption 4(a) states that the number of coefficients in each regression group is larger than twice the variance ratio r.

In the context of two models ($M = 2$) and homoskedasticity, Assumption 4 is identical to the conditions required for a Stein-type estimator to be minimax—to have globally smaller risk than the unrestricted estimator. Assumption 4 extends these conditions to the case of multiple models, that each regressor "group" $\mathbf{x}_{ji}$ has three or more regressors and the marginal penalties satisfy $0 < t_m \leq 2(k_m - 2)\sigma^2$.

In the important special case of the Mallows averaging estimator (12), a sufficient condition for the asymptotic marginal penalties $t_m = \sigma^2 k_m$ to satisfy Assumption 4 is $k_m \geq 4r/(2 - r)$.

ASSUMPTION 5. *$r < 2$ and for all $m \geq 2$, $k_m \geq 4r/(2 - r)$.*

COROLLARY 2. *Under Assumptions 1, 2, and 5, the Mallows averaging estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}$ *satisfies* $R(\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})$.

Corollary 2 is a powerful and important result. It shows that for regression models, the Mallows averaging estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}$ of Hansen (2007) globally dominates the least-squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ as long as the degree of heteroskedasticity is not too large and the regressor groupings each have a sufficiently large number of regressors. Corollary 2 shows that this modification guarantees that the estimator is a strict improvement on least squares.

A particularly important case of interest is conditional homoskedasticity $E(e_i^2|\mathbf{x}_i) = \sigma^2$.

COROLLARY 3. *Suppose Assumptions 1 and 2 hold,* $E(e_i^2|\mathbf{x}_i) = \sigma^2$, *and* $k_m \geq 4$ *for all* $m \geq 2$. *Then the Mallows averaging estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}$ *satisfies* $R(\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})$.

Corollary 3 is our clearest statement of the gains from regressor grouping. It shows that for homoskedastic regressions, a sufficient condition for the global dominance of the Mallows averaging estimator over the least-squares estimator is that each regressor grouping has four or more regressors. This is a simple modification of the averaging estimator. In Section 11, we show that this modification results in significant improvements in finite sample mean-squared error.

As a final remark, we note that for the case $M = 2$ under homoskedasticity, the condition $k_m \geq 4$ is both necessary and sufficient for $R(\widehat{\boldsymbol{\beta}}_{\mathrm{MMA}}, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})$, as the inequality $0 < t \leq 2(k-2)\sigma^2$ is both necessary and sufficient for the James–Stein estimator to be minimax, and this condition is violated when $k = 3$ and $t = 3\sigma^2$. However, when $M > 2$, the necessity of $k_m \geq 4$ is unclear.

10. Heteroskedasticity

When heteroskedasticity is present, it may be desirable to use alternative penalties. We describe two possible approaches and their properties.

First, the upper bound in Assumption 4(b) suggests that the penalties could be based on estimates of the smallest conditional variance $\underline{\sigma}^2$ rather than the unconditional variance σ^2. Let $\widehat{\underline{\sigma}^2}$ be an estimate of $\underline{\sigma}^2$. For example, $\widehat{\underline{\sigma}^2} = \min_{1\le i\le n} \widehat{\sigma}^2(\mathbf{x}_i)$, where $\widehat{\sigma}^2(\mathbf{x})$ is a Nadaraya–Watson estimator of $\sigma^2(\mathbf{x}) = E(e_i^2|\mathbf{x}_i = \mathbf{x})$ from a standard kernel regression of the squared residuals $\widehat{e}_i^2$. Setting $\widetilde{T}_m = \widehat{\underline{\sigma}^2} K_m$, then $\widetilde{t}_m \xrightarrow{p} t_m = \underline{\sigma}^2 k_m$ and Assumption 4(b) is satisfied if $k_m \ge 4r$.

Alternatively, we can set the penalty $\widetilde{T}_m$ to be a consistent estimate of D_m defined in (30), so that the marginal penalty $\widetilde{t}_m$ is a consistent estimator of d_m defined in (34). An example is the heteroskedasticity-robust Mallows criterion of Liu and Okui (forthcoming), which sets the penalties to equal

$$\widetilde{T}_m^{\mathrm{LO}} = \frac{n}{n-K}\operatorname{tr}\big(\widehat{\mathbf{Q}}_m^{-1}\widehat{\boldsymbol{\Omega}}_m\big), \tag{36}$$

where

$$\widehat{\mathbf{Q}}_m = \frac{1}{n}\overline{\mathbf{X}}_m'\overline{\mathbf{X}}_m,$$

$$\widehat{\boldsymbol{\Omega}}_m = \frac{1}{n}\sum_{i=1}^n \overline{\mathbf{x}}_{mi}\overline{\mathbf{x}}_{mi}'\widehat{e}_i^2.$$

The penalties $\widetilde{T}_m^{\mathrm{LO}}$ are moment estimators of D_m with a degree-of-freedom adjustment $n/(n-K)$, which Liu and Okui suggest is useful in finite samples. It follows that $\widetilde{t}_m^{\mathrm{LO}} \xrightarrow{p} t_m = d_m \ge \underline{\sigma}^2 k_m$, the final inequality shown in (51) in the Appendix. It follows that Assumption 4(b) is satisfied if $k_m \ge 4r$.

Assumption 6.

(a) $\widetilde{T}_m = \widehat{\underline{\sigma}^2} K_m$, *where* $\widehat{\underline{\sigma}^2} \xrightarrow{p} \underline{\sigma}^2$, *or* $\widetilde{T}_m = \widetilde{T}_m^{\mathrm{LO}}$ *from* (36).

(b) *For all* $m \ge 2$, $k_m \ge 4r$.

Corollary 4. *Under Assumptions 1, 2, and 6, the averaging estimator* $\widehat{\boldsymbol{\beta}}_A$ *satisfies* $R(\widehat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) < R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})$.

The condition $k_m \ge 4r$ in Assumption 6 is considerably less restrictive than the condition $k_m \ge 4r/(2-r)$ in Assumption 5. This suggests that averaging estimators using these modified penalties should have broader robustness properties than the Mallows averaging estimator.

Closely related to the Liu–Okui estimator is the jackknife model averaging (JMA) estimator of Hansen and Racine (2012). This is an averaging estimator (11), where the weights are selected to minimize the cross-validation criterion

$$\mathrm{JMA}_n(\mathbf{w}) = \sum_{m=1}^M\sum_{\ell=1}^M w_m w_\ell \sum_{i=1}^n \tilde{e}_{mi}\tilde{e}_{\ell i},$$

where $\tilde{e}_{mi} = y_i - \overline{\mathbf{x}}_{mi}'\widehat{\boldsymbol{\beta}}_{-i,m}$ is the leave-one-out prediction residual, which is easily computed using the algebraic equivalence $\tilde{e}_{mi} = \widehat{e}_{mi}/(1 - h_{mi})$ with $h_{mi} = \overline{\mathbf{x}}_{mi}'(\overline{\mathbf{X}}_m'\overline{\mathbf{X}}_m)^{-1}\overline{\mathbf{x}}_{mi}$. The jackknife criterion $\mathrm{JMA}_n(\mathbf{w})$ is close to the Liu–Okui heteroskedasticity-robust Mallows criterion and thus has similar MSE properties under conditional heteroskedasticity.

11. Finite sample simulations

We now use simulation[1] to investigate the finite sample performance of the averaging estimators. We explore both cross-section and time-series settings.

11.1 *Cross-section regression*

The cross-section model is similar to that used in Hansen (2007). The data are generated by the linear regression

$$y_i = \beta_0 + \sum_{j=1}^{M} \beta_j x_{ji} + e_i \tag{37}$$

with $E(e_i|\mathbf{x}_i) = 0$ and $E(e_i^2) = 1$. We set $x_{ji} \sim \mathrm{N}(0, 1)$. For the results reported here, we set $e_i \sim \mathrm{N}(0, 1)$ and $M = 12$, though we discuss sensitivity to these assumptions below. We vary the sample size n among $\{50, 150, 500, 1000\}$.

The coefficients are set as $\beta_0 = 0$ and $\beta_j = cj^{-\alpha}$ for $j \geq 1$ with $\alpha \geq 0$. Higher values of α mean that the coefficients β_j decline more quickly to zero as j increases. Lower values of α mean that the coefficients β_j are of relatively similar magnitude. Thus α controls the trade-off between bias and parsimony, a key issue in model selection. We vary α among $\{0, 1, 2, 3\}$. Notice that in contrast to the asymptotic theory, we will treat the coefficients as fixed when we vary the sample size, so that the experiments reported here correspond to actual econometric practice.

The coefficient c is selected to vary the population $R^2 = \sum_{j=1}^{M} \beta_j^2/(1 + \sum_{j=1}^{M} \beta_j^2)$ on a 19-point grid in $[0.00, 0.90]$.

The estimators $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are assessed by finite sample mean-squared error

$$\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

We calculate $\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta})$ by simulation, averaging across 10,000 independent replications. We also normalize the MSE by the MSE of the unconstrained ordinary least-squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$. Thus a reported MSE below 1 indicates that the estimator has smaller MSE than unconstrained ordinary least squares (OLS), and a reported MSE above 1 indicates that the estimator has larger MSE than unconstrained OLS.

The estimates were constructed from $M + 1$ separate regressions of the form

$$y_i = \widehat{\beta}_0(m) + \sum_{j=1}^{m} \widehat{\beta}_j(m) x_{ji} + \widehat{e}_i(m)$$

[1]The R code used for the simulation is available on the journal website.

for $m = 0, \ldots, M$ and then $\widehat{\boldsymbol{\beta}}(m) = (\widehat{\beta}_0(m), \widehat{\beta}_1(m), \widehat{\beta}_2(m), \ldots, \widehat{\beta}_m(m), 0, \ldots, 0)$ was set. These are nested regression models ranging from intercept only to unconstrained.

From these $M + 1$ regressions, the following estimates were compared.

1. OLS: Ordinary least squares $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \widehat{\boldsymbol{\beta}}(M)$.

2. MMA: The Mallows model averaging estimator using all $M + 1$ models.

3. MMA_4: The Mallows averaging estimator, grouping regressors in sets of four. For $M = 12$, this includes models $m = \{0, 4, 8, 12\}$.

4. Stein: A James–Stein estimator, shrinking the unconstrained least-squares estimator $\widehat{\boldsymbol{\beta}}(M)$ toward the intercept-only model $\widehat{\boldsymbol{\beta}}(0)$.

5. Lasso: Least-angle regression (Tibshirani (1996)) with penalty λ selected to minimize fivefold cross-validation.

6. BMA: Approximate Bayesian model averaging or smoothed Bayesian information criterion.

7. SAIC: Smoothed Akaike information criterion (Akaike (1979), Buckland, Burnham, and Augustin (1997), and Burnham and Anderson (2002)).

Methods 4–7 are alternative averaging and shrinkage methods that are included for comparison.

As discussed in Section 6, the Stein estimator is equivalent to the MMA estimator restricted to two submodels. Thus by comparing MMA_4 with the Stein estimator, we illustrate the gains by averaging over more than two models.

The Lasso is a popular method for regression shrinkage that does not require regressor ordering. We use the glmnet function in R with all default settings.

BMA and SAIC are two popular model averaging methods. The estimators take the form (11) with the weights $\widehat{w}_m$ proportional to $\exp(-\frac{1}{2}\mathrm{BIC}_m)$ (where BIC denotes the Bayesian information criterion) and $\exp(-\frac{1}{2}\mathrm{AIC}_m)$, respectively, with

$$\mathrm{BIC}_m = n \log\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_{mi}^2\right) + \log(n) K_m,$$

$$\mathrm{AIC}_m = n \log\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_{mi}^2\right) + 2K_m.$$

The results are reported graphically in Figures 1–4. Each figure corresponds to a single value of α, and each figure has four panels, for $n = 50$, 150, 500, and 1000. Each panel plots the normalized MSE of the estimators as a function of the population R^2. To reduce the clutter in the figures, the SAIC method is not displayed here, but is displayed in the plots in the supplemental appendix available on the journal website. (In most cases, SAIC performs quite similarly to MMA.)

From the results, we can see some clear trends. First, both the MMA_4 and the Stein estimators globally have reduced risk relative to OLS (their normalized MSEs are everywhere less than 1), but the MMA estimator has risk that is greater than OLS for some

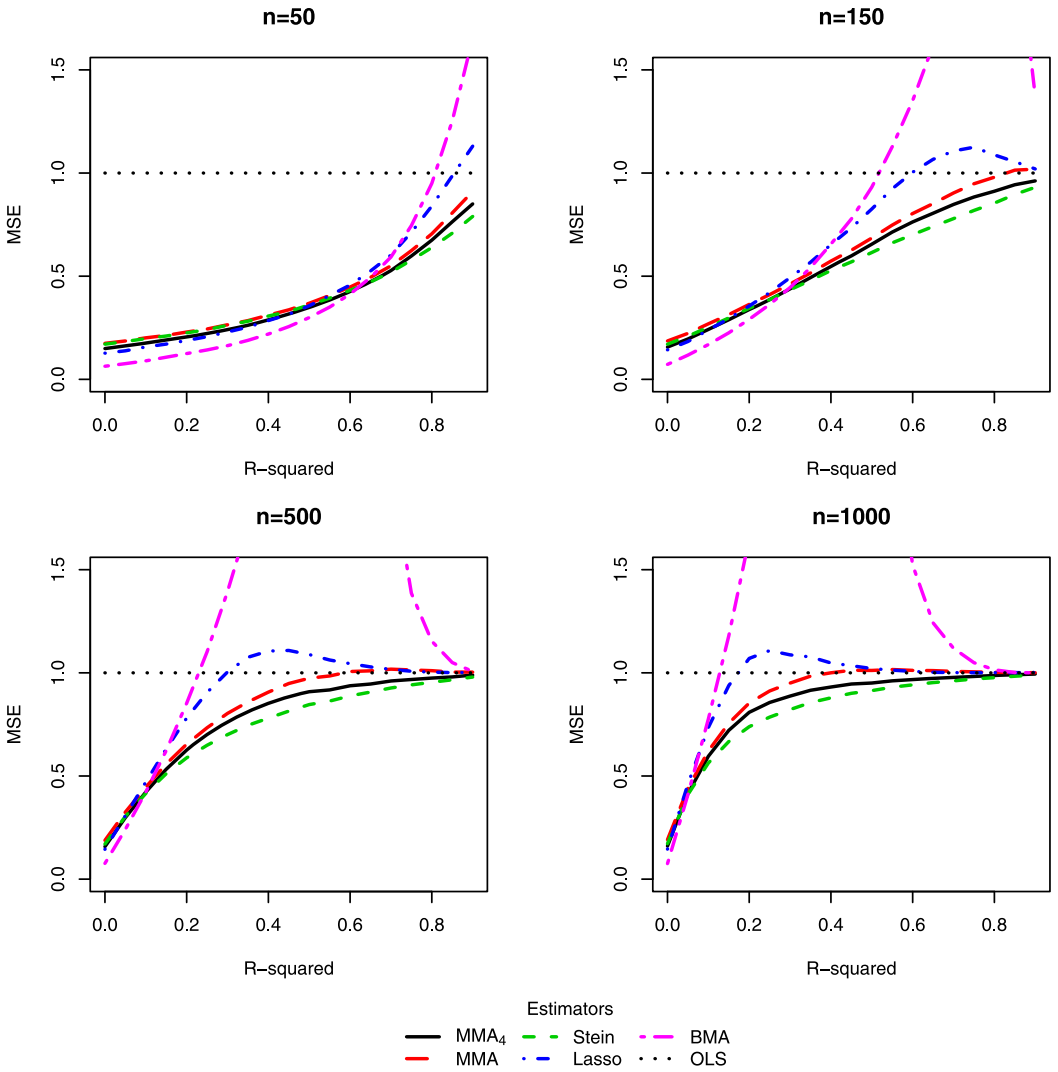

FIGURE 1. $\alpha = 0$.

parameter values. This confirms a strong prediction of the asymptotic theory. Second, for most parameter values, the MMA_4 estimator dominates the Stein estimator. This is especially the case for large values of α and for small sample sizes. This is because the MMA_4 estimator is able to exploit the ordering of the regressors, while the Stein estimator treats all symmetrically. For larger α, the differences in MSE are quite substantial. Third, for most parameter values, the MMA_4 estimator dominates the MMA estimator. This is especially the case for large sample sizes. In summary, by grouping the regressors in sets of four before applying Mallows averaging, MSE is reduced for most parameter settings, and the averaging estimator globally dominates OLS.

The results also show that the MMA_4 method compares quite well relative to the alternative averaging and shrinkage methods. In particular, the BMA method has er-

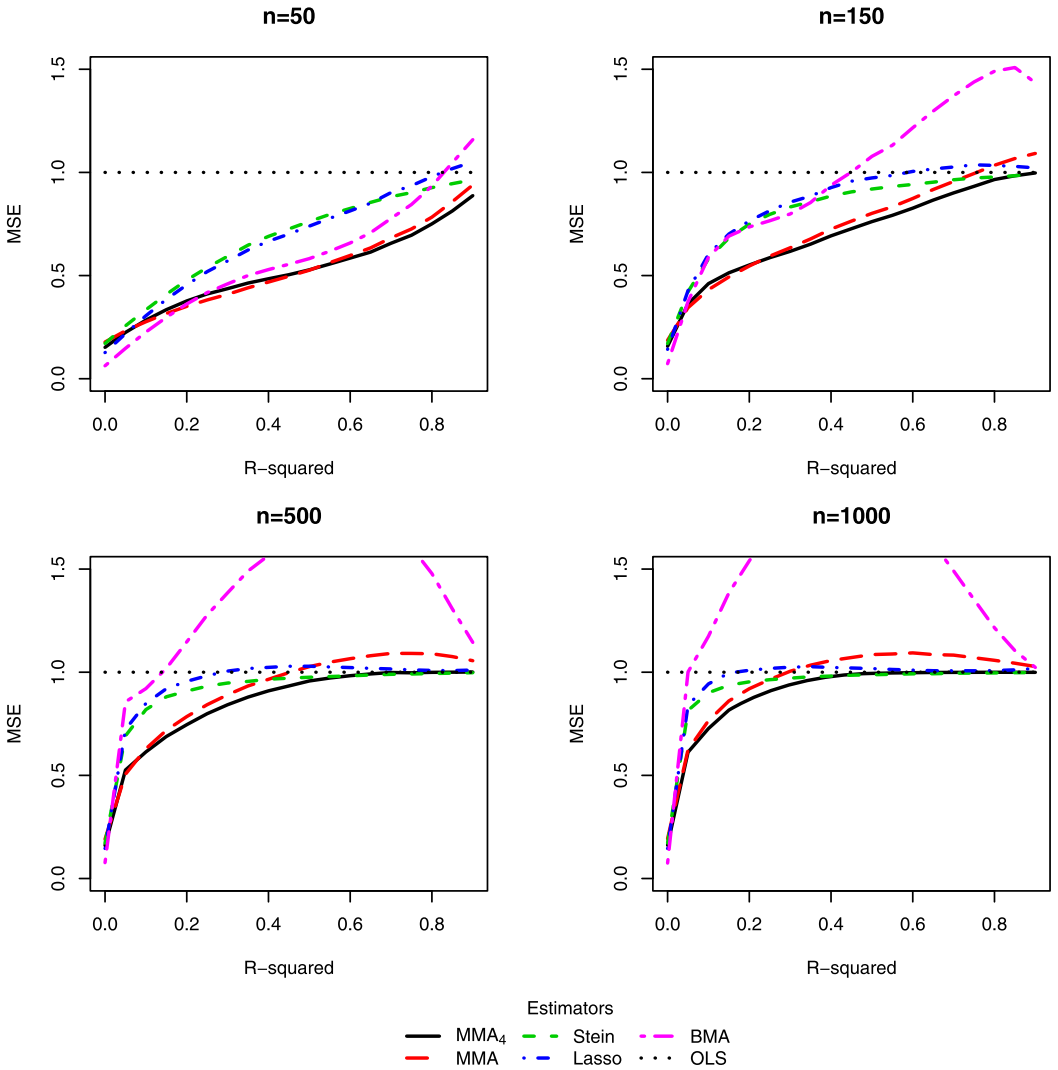

FIGURE 2. $\alpha = 1$.

ratic performance, with some parameterizations leading to extremely high MSE, and this poor performance gets worse with larger sample sizes. MMA_4 also generally has lower MSE than the Lasso and SAIC methods for most parameterizations and sample sizes.

To explore the sensitivity of the simulation results to the design, we varied some of the assumptions. We summarize the results here for brevity: graphs of the results can be found in the supplemental appendix available on the journal website. First, we sampled the error e_i from a skewed nonnormal distribution, and there was no change in the results. Second, we sampled the error from the heteroskedastic distribution $e_i \sim \text{N}(0, (1 + x_{2i}^2)/2)$, and there was no change in the results. Third, we introduced correlation between the regressors. The performance of several of the shrinkage and averaging

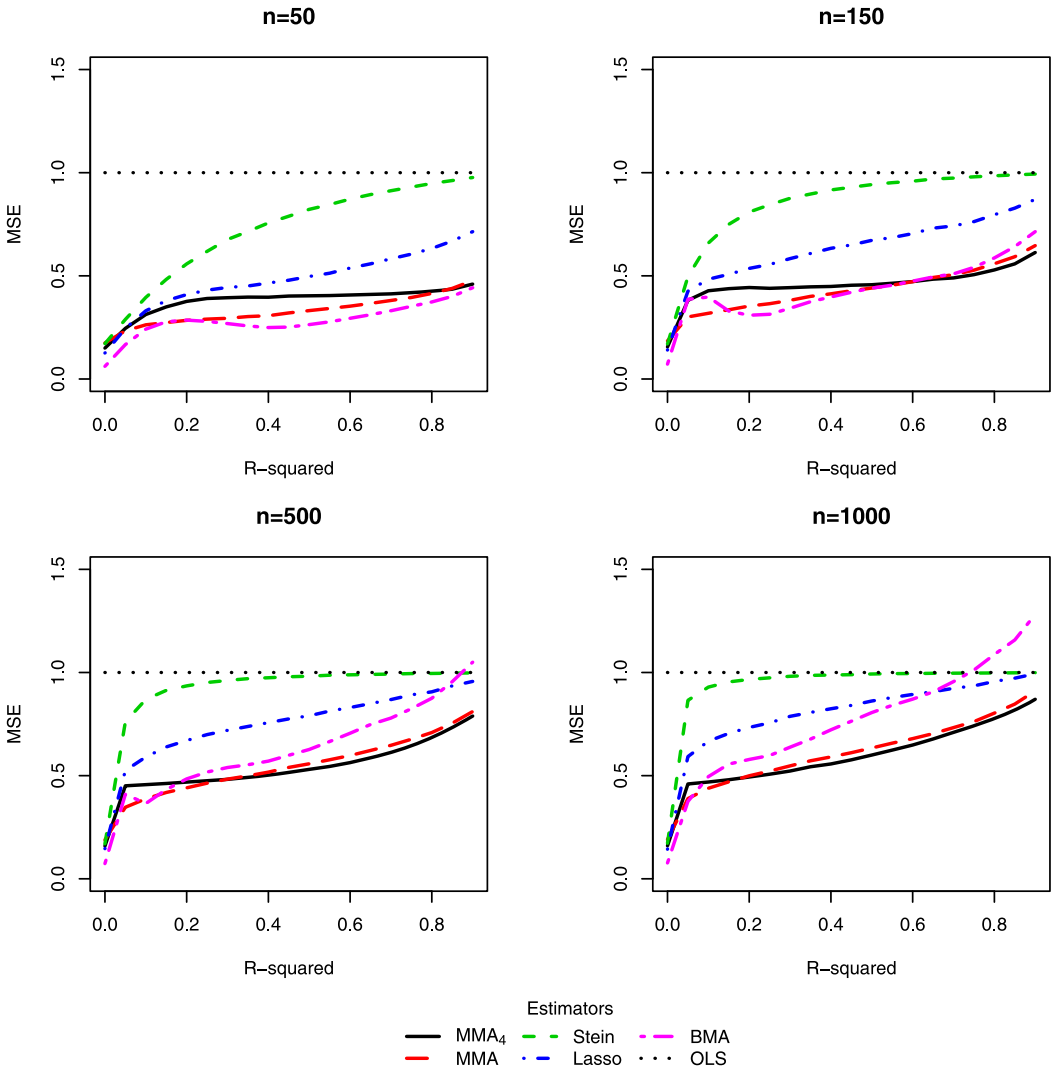

FIGURE 3. $\alpha = 2$.

estimators greatly improved (relative to OLS), especially for small α, but otherwise the qualitative results were unchanged. Fourth, we increased the number of regressors to $M = 24$. Again, the performance of the shrinkage and averaging estimators greatly improved relative to OLS, but otherwise the qualitative results were unchanged.

As a final important robustness check, we investigated the sensitivity of the results to the *ordering* of the regressors. The MMA, MMA_4, BMA, and SAIC methods all depend on the ordering of the regressors, and the above results are constructed using the correct ordering (ordering the regressors by the magnitude of the true coefficients). To investigate the sensitivity of the MMA_4 method to this knowledge, we *reverse* the order of the regressors and then implement the MMA_4 method. This should be most unfavorable to nested model averaging, as the regressors are ordered from smallest to largest coeffi-

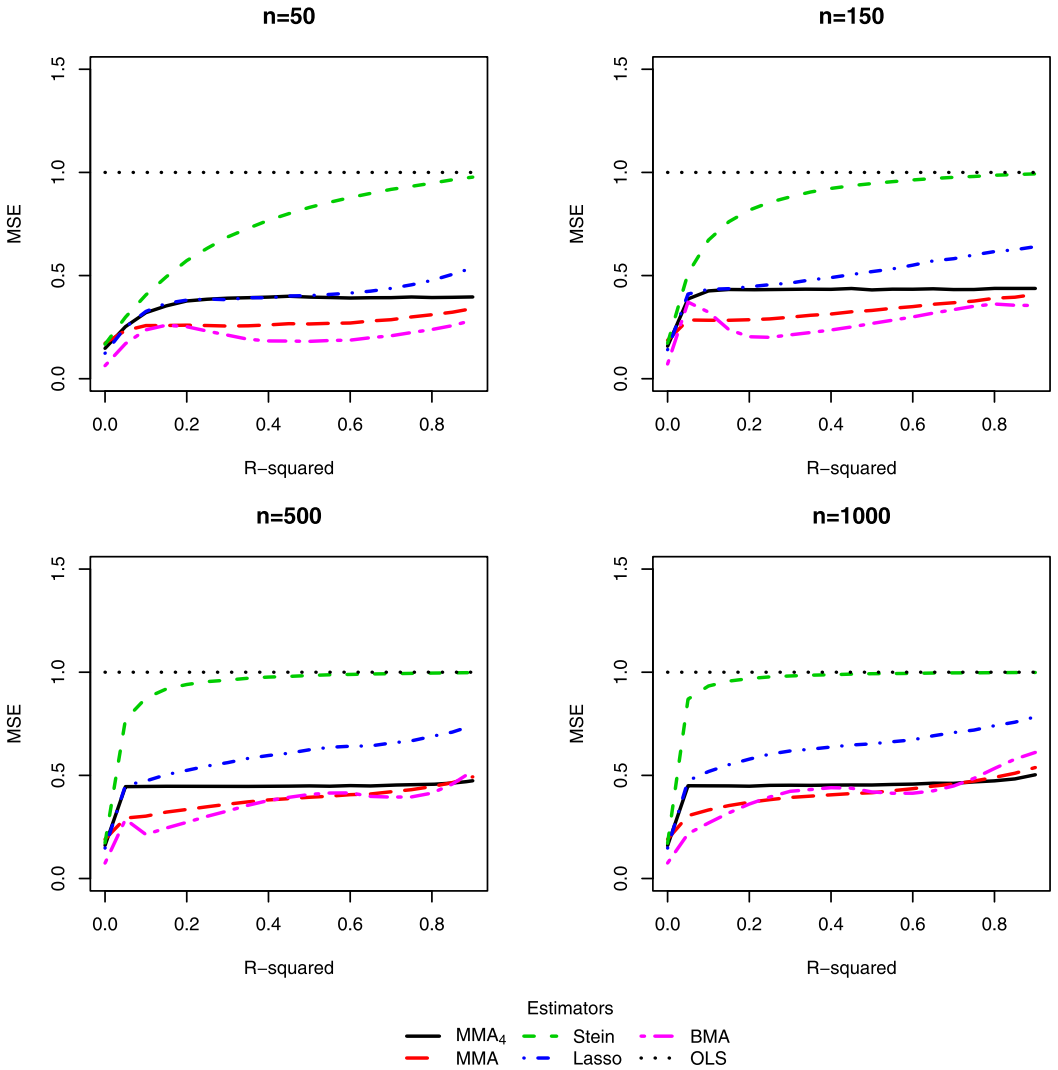

FIGURE 4. $\alpha = 3$.

cients. The results are quite interesting. In nearly all cases (except $\alpha = 0$, where reversing the order has no effect), the reversed-order MMA_4 method has nearly identical performance to the Stein estimator. Thus its MSE is better than OLS, but not as good as the MMA_4 method with the correct order. This confirms the asymptotic prediction that regardless of the ordering, MMA_4 will have lower MSE than OLS. The reason MMA_4 has nearly identical MSE with the Stein estimator is because in this context the MMA_4 criterion typically puts all weight on only two models—the intercept-only model and the full model. The other models are irrelevant and thus receive zero weight. What this shows is that the cost of misordering regressors is not too severe, and the MMA_4 estimator has improved efficient relative to OLS.

Overall, the simulation results confirm the predictions from the asymptotic theory. Averaging can greatly reduce estimation error relative to unconstrained estimation if the averaging weights are selected by minimizing a penalized least-squares criterion and the regressors are grouped in sets of four or larger.

11.2 *Time-series regression*

Our second simulation experiment explores the performance of the method in a simple time-series autoregression. The model is an AR(M),

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_M y_{t-M} + e_t$$

with $\beta_0 = 0$, e_t i.i.d. N$(0, \sigma^2)$, $\sigma^2 = 1$, and the coefficients set to be monotonically decreasing as

$$\beta_j = \theta \frac{1 - j/(M+1)}{\sum_{\ell} (1 - \ell/(M+1))}.$$

The parameter θ controls the magnitude of the coefficients and is varied from 0 to 0.95 in steps of 0.05.

As in the previous experiment, we set $M = 12$ and $n = \{50, 150, 400, 1000\}$. We apply the same set of estimators and again calculate the MSE of the coefficient estimates normalized by the MSE of the OLS estimator.

The results are displayed in Figure 5. The findings are quite similar to the cross-section results, with the exception that the MMA_4 method does not uniformly dominate OLS, as for small samples and large θ (high persistence) the ordering is reversed. We suspect that this is a consequence of the fact that as θ approaches 1, the autoregression approaches nonstationarity and the asymptotic approximations become unreliable.

12. Testing for racial differences in the mental ability of young children

To illustrate the method, we apply the grouped MMA estimator to the regression analysis of Fryer and Levitt (2013). Their goal was to assess whether there are measurable differences across races in the mental ability of very young children, especially after controlling for birth, demographic, and socioeconomic factors. The answer to this question helps shed light on the extent to which ability is genetic versus environmental.

Fryer and Levitt use two data sets, but primarily focus on the Early Childhood Longitudinal Study Birth Cohort (ECLS-B), which is a sample of over 10,000 children born in 2001 and includes two waves of mental functioning tests, the first when most of the children were between 8 and 12 months of age, and the second when the children were close to 2 years old.

For each test wave, Fryer and Levitt estimated six nested regressions of children's test scores on racial categories plus varying sets of controls. These control groups are as follows:

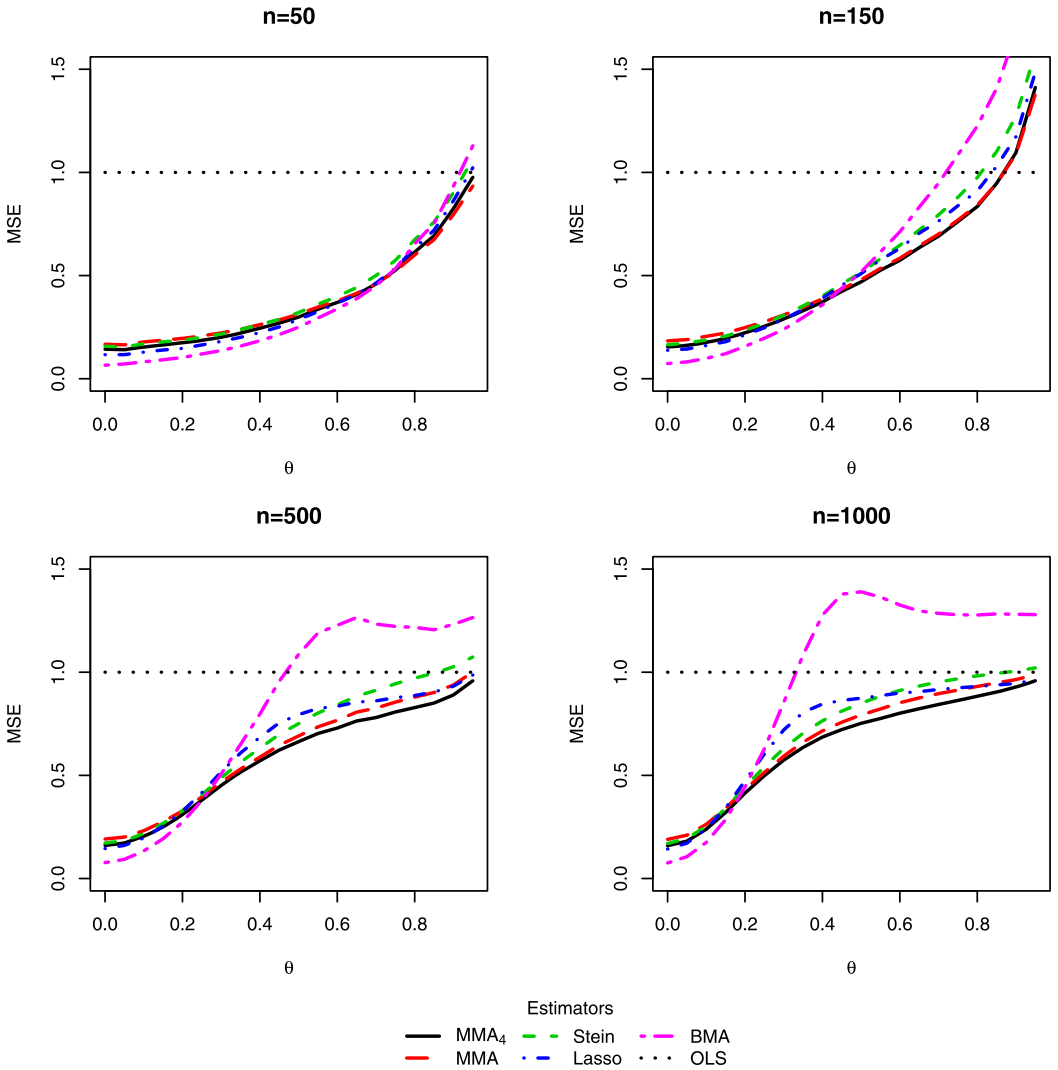

FIGURE 5. Autoregression.

Model 1. No controls beyond racial groups.

Model 2. Adds interviewer fixed effects.

Model 3. Adds age and gender of child (age in eight categories for first wave, in three categories for second).

Model 4. Adds socioeconomic status quintiles (four categories).

Model 5. Adds home environment variables (number of siblings as dummy categories, family configuration, region, mother's age as fifth order polynomial, parent-as-teacher measure as fifth order polynomial).

Model 6. Adds prenatal variables (child's birth wight in four categories, number of days premature in twelve categories, singleton birth, twin birth).

TABLE 1. Mallows model averaging estimates using Fryer–Levitt (2013) regressions.

	Models						
	1	2	3	4	5	6	MMA
	First wave (children approximately 9 months of age)						
Black	−0.054	−0.073	−0.063	−0.036	−0.015	0.015	**0.013**
Hispanic	−0.018	0.001	−0.037	−0.005	−0.004	−0.007	**−0.007**
Asian	−0.007	0.026	−0.018	−0.017	−0.014	−0.010	**−0.010**
Other	−0.026	−0.024	−0.034	−0.016	−0.010	−0.001	**−0.002**
$\widehat{w}_m$	**0.031**	**0.000**	**0.000**	**0.000**	**0.000**	**0.969**	
	Second wave (children approximately 2 years of age)						
Black	−0.383	−0.402	−0.402	−0.250	−0.228	−0.213	**−0.229**
Hispanic	−0.423	−0.405	−0.422	−0.253	−0.247	−0.249	**−0.265**
Asian	−0.217	−0.277	−0.285	−0.301	−0.297	−0.296	**−0.289**
Other	−0.246	−0.249	−0.248	−0.152	−0.138	−0.136	**−0.146**
$\widehat{w}_m$	**0.093**	**0.000**	**0.000**	**0.000**	**0.000**	**0.907**	

Note: Data are from Fryer and Levitt (2013). The dependent variable is normalized to have a mean of 0 and a standard deviation of 1. Non-Hispanic whites are in the omitted race category. Estimation is by weighted least squares. The number of observations is 8871. As in Fryer and Levitt, observations with missing test scores, race, interviewer identification, or sampling weight are excluded, and for other covariates an indicator variable for missing values is included.

Table 1 lists the OLS estimates of the coefficients on the racial groups from these six models. The top section of the table displays the estimates for the first test wave (children approximately 9 months of age) and the bottom section displays the estimates for the second test wave (children approximately 2 years of age). Fryer and Levitt use these estimates to make the point that for infants, the mean differences across racial groups is small and diminishes after controlling for covariates, yet the differences are meaningful for toddlers. Tables of this form are commonly seen in empirical economics.

Yet there is an inherent ambiguity about how to concisely summarize the findings from these six regression estimates. Which is the best estimate? What number is the best summary? The richer models have more controls (so less omitted variable bias) yet have higher variances due to larger number of estimated parameters. The conventional estimates do not give one concise summary estimate.

The method of model averaging (MMA) gives a single estimate, averaging across the six regression estimates, and our theory shows this estimator has reduced risk (mean-squared error) relative to ordinary least-squares estimation of the full model. Thus MMA is a concrete way to concisely summarize the point estimates. We computed the MMA estimates for the Fryer–Levitt estimates and include these estimates in the seventh column of the table. We also report the MMA weights $\widehat{w}_m$ at the bottom of each column. The MMA estimates are the weighted average of the individual model estimates using these weights, and these weights are selected to minimize the Mallows averaging criterion.

We can see that the MMA criterion puts a weight of 97% on the full estimates for the first wave and a weight of 91% on the full model for the second wave, so that for both regressions, the MMA estimates are close to the full regression estimates. The fact that MMA puts most of the weight on the full model is largely a consequence of the way that Fryer and Levitt ordered their regressors. Some of the most important regressors

(birth weight and days premature) are only included in the full model, and thus it is not surprising that MMA wants to give a high weight to this model.

To rectify this situation, we redefined and reordered the regressors. We treated age, birth weight, number of days premature, and the number of siblings as continuous variables (rather than grouping them into categories) and included powers up to order 5 (as done for the parent-as-teacher measure and mother's age), but only included the higher powers in the largest models. We estimated eight nested models. The first two are the same as Fryer and Levitt; the remaining six are as follows:

Model 3′. Adds gender of child, number of siblings, mother's age, parent-as-teacher measure, birth weight, and number of days premature.

Model 4′. Adds socioeconomic status quintiles, family configuration, region, singleton birth, twin birth.

Model 5′. Adds second power of continuous variables (number of siblings, mother's age, parent-as-teacher measure, birth weight, and number of days premature).

Model 6′. Adds third power of continuous variables.

Model 7′. Adds fourth power of continuous variables.

Model 8′. Adds fifth power of continuous variables.

Table 2 reports the OLS estimates from these eight nested model, plus the MMA estimates in the final column and the MMA weights $\widehat{w}_m$ at the bottom. What is noticeable is that the MMA method distributes weights differently across the models. For the first wave data set (children aged 9 months), 53% of the weight is put on the full model, 21% on the fifth model (the one with quadratic terms), 14% on the seventh model, and 11% on the third model. For the second wave data set (2-year-old children), MMA puts no weight on the full model, 51% on the sixth model, and the remainder is split between models 1, 4, 5, and 7.

TABLE 2. Mallows model averaging estimates: reordered regressions.

	Models								
	1	2	3′	4′	5′	6′	7′	8′	MMA
	First wave (children approximately 9 months of age)								
Black	−0.054	−0.073	0.006	−0.023	0.023	0.024	0.022	0.020	**0.018**
Hispanic	−0.018	0.001	−0.014	−0.002	−0.003	−0.003	−0.003	−0.003	**−0.005**
Asian	−0.007	0.026	0.008	0.002	−0.010	−0.012	−0.013	−0.012	**−0.010**
Other	−0.026	−0.024	−0.002	0.010	−0.000	0.000	−0.001	−0.006	**−0.001**
$\widehat{w}_m$	**0.015**	**0.000**	**0.108**	**0.000**	**0.208**	**0.000**	**0.139**	**0.529**	
	Second wave (children approximately 2 years of age)								
Black	−0.383	−0.402	−0.284	−0.208	−0.209	−0.206	−0.207	−0.207	**−0.220**
Hispanic	−0.423	−0.405	−0.350	−0.251	−0.253	−0.253	−0.253	−0.253	**−0.265**
Asian	−0.217	−0.277	−0.263	−0.291	−0.297	−0.296	−0.298	−0.298	**−0.290**
Other	−0.246	−0.249	−0.181	−0.129	−0.132	−0.131	−0.132	−0.132	**−0.139**
$\widehat{w}_m$	**0.072**	**0.000**	**0.000**	**0.149**	**0.093**	**0.571**	**0.115**	**0.000**	

The MMA estimate of the controlled 9-month test score difference between non-Hispanic whites and blacks is 0.02 standard deviation units (blacks outscore whites, but by an infinitesimal amount). For 2-year-olds, the MMA estimate is that, on average, whites outscore blacks by 0.22 standard deviation units. Our theory suggests that these are the best estimates of the key parameters of interest.

13. Conclusion

This paper has extended our understanding of model selection and combination. We examine averaging weights selected by minimizing a penalized criteria and find that such averaging estimators have reduced risk relative to unconstrained estimation if the regressors are grouped in sets of four or larger so that the Stein shrinkage effect holds. The simulation shows that the gains are substantial and hold in finite samples.

While the theory of this paper has focused on the context of least-squares regression, we believe that the concepts can be extended to other contexts including panel data and the generalized method of moments.

The theory also is confined to the context of nested models. While it would be greatly desirable to extend the analysis to include nonnested models, it is not clear how this could be accomplished.

Another unexplored issue is inference. The asymptotic distributions of selection and averaging estimators are nonstandard (at least in the local asymptotic framework used here). This is routinely ignored in applications involving postselection estimators, but is difficult to avoid when using averaging estimators. This is a challenging topic and quite important for future investigation.

Appendix

Proof of Lemma 1. Note that $L_j \geq L_{j+1}$ and $\widehat{e}_j'\widehat{e}_m = L_{\max(j,m)}$ by the properties of the least-squares residuals. The penalized criterion is then

$$C_n(\mathbf{w}) = \sum_{j=1}^{M}\sum_{m=1}^{M} w_j w_m L_{\max(j,m)} + 2\sum_{m=1}^{M} w_m \widetilde{T}_m. \tag{38}$$

The first term in (38) can be rewritten as

$$\begin{aligned}
&w_1^2 L_1 + \big(w_2^2 + 2w_2w_1\big)L_2 + \big(w_3^2 + 2w_3(w_1+w_2)\big)L_3 + \cdots \\
&\quad + \big(w_M^2 + 2w_M(w_1 + \cdots + w_{M-1})\big)L_M \\
&= w_1^2 L_1 + \big((w_1+w_2)^2 - w_1^2\big)L_2 + \big((w_1+w_2+w_3)^2 - (w_1+w_2)^2\big)L_3 + \cdots \\
&\quad + \big((w_1 + \cdots + w_M)^2 - (w_1 + \cdots + w_{M-1})^2\big)L_M \\
&= w_1^2(L_1 - L_2) + (w_1+w_2)^2(L_2 - L_3) + \cdots \\
&\quad + (w_1 + \cdots + w_{M-1})^2(L_{M-1} - L_M) + (w_1 + \cdots + w_M)^2 L_M \\
&= \sum_{m=1}^{M-1} w_m^{*2}(L_m - L_{m+1}) + L_M.
\end{aligned} \tag{39}$$

The second term in (38) is

$$\begin{aligned} 2\sum_{m=1}^{M} w_m \widetilde{T}_m &= 2w_1(\widetilde{T}_1 - \widetilde{T}_2) + 2(w_1+w_2)(\widetilde{T}_2 - \widetilde{T}_3) + \cdots \\ &\quad + 2(w_1 + \cdots + w_{M-1})(\widetilde{T}_{M-1} - \widetilde{T}_M) + (w_1 + \cdots + w_M)\widetilde{T}_M \qquad (40) \\ &= -2\sum_{m=1}^{M-1} w_m^*(\widetilde{T}_{m+1} - \widetilde{T}_m) + 2\widetilde{T}_M. \end{aligned}$$

Summing (39) and (40), we find $C_n^*(\mathbf{w}^*) + L_M + 2\widetilde{T}_M$ with $C_n^*(\mathbf{w}^*)$ as defined in (16). □

LEMMA 3. *For* $\mathbf{P}_m$ *defined in* (23),

$$\begin{aligned} &\mathbf{P}_m\mathbf{Q}\mathbf{P}_\ell = \mathbf{P}_{\min(m,\ell)}, \\ &(\mathbf{P}_\ell - \mathbf{P}_m)\mathbf{Q}(\mathbf{P}_\ell - \mathbf{P}_m) = \mathbf{P}_\ell - \mathbf{P}_m, \\ &(\mathbf{P}_j - \mathbf{P}_\ell)\mathbf{Q}(\mathbf{P}_\ell - \mathbf{P}_m) = \mathbf{0}, \end{aligned}$$

the second and third equalities for $m < \ell < j$.

PROOF. Suppose $m \le \ell$. Recall the definition for $\mathbf{S}_m$ given in (4). Since the models are nested, $\mathbf{S}_m$ is in the range space of $\mathbf{S}_\ell$ and, therefore, $\mathbf{S}_m = \mathbf{S}_\ell\mathbf{G}$ for some matrix $\mathbf{G}$. Then

$$\begin{aligned} \mathbf{P}_m\mathbf{Q}\mathbf{P}_\ell &= \mathbf{S}_m\big(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\big)^{-1}\mathbf{S}_m'\mathbf{Q}\mathbf{S}_\ell\big(\mathbf{S}_\ell'\mathbf{Q}\mathbf{S}_\ell\big)^{-1}\mathbf{S}_\ell' \\ &= \mathbf{S}_m\big(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\big)^{-1}\mathbf{G}'\mathbf{S}_\ell' \\ &= \mathbf{S}_m\big(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\big)^{-1}\mathbf{S}_m' \\ &= \mathbf{P}_m, \end{aligned}$$

as claimed. Next,

$$\begin{aligned} (\mathbf{P}_\ell - \mathbf{P}_m)\mathbf{Q}(\mathbf{P}_\ell - \mathbf{P}_m) &= \mathbf{P}_\ell\mathbf{Q}\mathbf{P}_\ell - \mathbf{P}_m\mathbf{Q}\mathbf{P}_\ell - \mathbf{P}_\ell\mathbf{Q}\mathbf{P}_m + \mathbf{P}_m\mathbf{Q}\mathbf{P}_m \\ &= \mathbf{P}_\ell - \mathbf{P}_m - \mathbf{P}_m + \mathbf{P}_m \\ &= \mathbf{P}_\ell - \mathbf{P}_m \end{aligned}$$

and, similarly,

$$\begin{aligned} (\mathbf{P}_j - \mathbf{P}_\ell)\mathbf{Q}(\mathbf{P}_\ell - \mathbf{P}_m) &= \mathbf{P}_j\mathbf{Q}\mathbf{P}_\ell - \mathbf{P}_\ell\mathbf{Q}\mathbf{P}_\ell - \mathbf{P}_j\mathbf{Q}\mathbf{P}_m + \mathbf{P}_\ell\mathbf{Q}\mathbf{P}_m \\ &= \mathbf{P}_\ell - \mathbf{P}_\ell - \mathbf{P}_m + \mathbf{P}_m \\ &= \mathbf{0}. \end{aligned}$$

□

PROOF OF THEOREM 1. Since x_{1i} is included in all models, all the centered submodel estimates $\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}$ are invariant to $\boldsymbol{\beta}_1$, the coefficient on x_{1i}, and thus so are both the

least-squares and averaging estimators. Hence, without loss of generality, we set $\boldsymbol{\beta}_1 = 0$. Combined with Assumption 2, this yields $n^{1/2}\boldsymbol{\beta} \to \boldsymbol{\delta}$ as $n \to \infty$.

The organization of the argument is as follows. We first derive the joint asymptotic distribution of the least-squares estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$, the differenced submodel estimates $\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m$, and the differenced sum of squared errors $L_m - L_{m+1}$. We then derive the asymptotic distribution of the cumulative criterion $C_n^*(\mathbf{w}^*)$, its minimizer $\widehat{\mathbf{w}}^*$, and the averaging estimator. After that, we characterize the minimization problem (24).

Assumption 1 is sufficient to imply that

$$\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q} \tag{41}$$

and

$$\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{e} \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Omega}) \tag{42}$$

with $\boldsymbol{\Omega}$ defined in (20).

Combining (41), (42), the assumption $\mathbf{Q} > 0$ and the continuous mapping theorem,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}) &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{e}\right) \\ &\xrightarrow{d} \mathrm{N}(0, \mathbf{V}) \\ &\equiv \mathrm{Z}\end{aligned} \tag{43}$$

which is (19). The condition $n^{1/2}\boldsymbol{\beta} \to \boldsymbol{\delta}$ allows us also to deduce that

$$\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}) + \boldsymbol{\delta} \xrightarrow{d} \mathrm{Z} + \boldsymbol{\delta}. \tag{44}$$

Since

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_m &= \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{X}'\mathbf{X}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{X}'\mathbf{y} \\ &= \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{X}'\mathbf{X}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\end{aligned}$$

it follows from (44) that

$$\begin{aligned}\sqrt{n}\widehat{\boldsymbol{\beta}}_m &= \mathbf{S}_m\left(\mathbf{S}_m'\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \\ &\xrightarrow{d} \mathbf{S}_m\left(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\right)^{-1}\mathbf{S}_m'\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta}) \\ &= \mathbf{P}_m\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta})\end{aligned}$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m) \xrightarrow{d} (\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta}). \tag{45}$$

Since $\widehat{\mathbf{e}}_{m+1}'\widehat{\mathbf{e}}_m = \widehat{\mathbf{e}}_{m+1}'\widehat{\mathbf{e}}_{m+1}$ and $\widehat{\mathbf{e}}_m - \widehat{\mathbf{e}}_{m+1} = \mathbf{X}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m)$, we calculate that

$$\begin{aligned} L_m - L_{m+1} &= \widehat{\mathbf{e}}_m'\widehat{\mathbf{e}}_m - \widehat{\mathbf{e}}_{m+1}'\widehat{\mathbf{e}}_{m+1} \\ &= (\widehat{\mathbf{e}}_m - \widehat{\mathbf{e}}_{m+1})'(\widehat{\mathbf{e}}_m - \widehat{\mathbf{e}}_{m+1}) \\ &= (\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m) \\ &= \sqrt{n}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m)'\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)\sqrt{n}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m). \end{aligned}$$

Applying (41) and (45), this converges in distribution to

$$\begin{aligned} &(\mathrm{Z} + \boldsymbol{\delta})'\mathbf{Q}(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta}) \\ &\quad = (\mathrm{Z} + \boldsymbol{\delta})'\mathbf{Q}(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta}), \end{aligned} \tag{46}$$

the equality by Lemma 3.

Using equations (16) and (46), and using Assumption 1, we find that

$$\begin{aligned} C_n^*(\mathbf{w}^*) &= \sum_{m=1}^{M-1}\left(w_m^{*2}(L_m - L_{m+1}) - 2w_m^*\widetilde{t}_{m+1}\right) \\ &\xrightarrow{d} C^*(\mathbf{w}^*, \mathrm{Z} + \boldsymbol{\delta}) \end{aligned}$$

with $C^*(\mathbf{w}^*, \mathbf{x})$ defined in (25). Since (17) is a convex minimization problem ($C_n^*(\mathbf{w}^*)$ is quadratic and $\mathcal{H}^*$ is convex), we can apply the argument of Kim and Pollard (1990) and deduce that $\widehat{\mathbf{w}}^* \xrightarrow{d} \mathbf{w}^*(\mathrm{Z} + \boldsymbol{\delta})$, where $\mathbf{w}^*(\mathbf{x})$ is defined in (24). Combined with (45), it follows that

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) &= \sum_{m=1}^{M-1}\widehat{w}_m^*\sqrt{n}(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m) \\ &\xrightarrow{d} \mathrm{Z} - \sum_{m=1}^{M-1} w_m^*(\mathrm{Z} + \boldsymbol{\delta})(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}(\mathrm{Z} + \boldsymbol{\delta}) \\ &= \mathrm{Z} - \eta(\mathrm{Z} + \boldsymbol{\delta}), \end{aligned} \tag{47}$$

which is (21).

We now consider the minimization problem (24). Note that this is a deterministic problem and the solution is a function of the argument $\mathbf{x}$. We now fix $\mathbf{x}$ and, to simplify notation, we omit dependence (of the weights and selected models) on $\mathbf{x}$.

Since $C^*(\mathbf{w}^*, \mathbf{x})$ is quadratic in $\mathbf{w}^*$, the unconstrained minimum is simply

$$\overline{w}_m^* = \frac{t_{m+1}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m+1} - \mathbf{P}_m)\mathbf{Q}\mathbf{x}}.$$

If $\overline{\mathbf{w}}^* \in \mathcal{H}^*$, then $\mathbf{w}^* = \overline{\mathbf{w}}^*$, which satisfies (26)–(27) with $J = M$. If $\overline{\mathbf{w}}^* \notin \mathcal{H}^*$, then $\mathbf{w}^*$ lies on the boundary of $\mathcal{H}^*$. The latter is the union of sets of binding constraints, each constraint corresponding to excluding a specific model m. Equivalently, each section of the boundary of $\mathcal{H}^*$ consists of the set of models $\{m_1, \ldots, m_J\}$ with positive weight. If a model m is

not in this set, it receives a weight of zero and $w_m^* = w_{m-1}^*$. Thus if $\{m_1, \ldots, m_J\}$ are the models with positive weight, then for $j = 1, \ldots, M-1$,

$$w_\ell^* = w_{m_j}^*, \quad m_j \le \ell < m_{j+1},$$

and

$$w_\ell^* = 1, \quad m_J \le \ell \le M.$$

Thus (25) can be written as

$$\begin{aligned}
&\sum_{j=1}^{J-1} \sum_{\ell=m_j}^{m_{j+1}-1} \bigl(w_\ell^{*2} \mathbf{x}'\mathbf{Q}(\mathbf{P}_{\ell+1} - \mathbf{P}_\ell)\mathbf{Q}\mathbf{x} - 2w_\ell^* t_{\ell+1}\bigr) \\
&\qquad + \sum_{\ell=m_J}^{M-1} \bigl(w_\ell^{*2} \mathbf{x}'\mathbf{Q}(\mathbf{P}_{\ell+1} - \mathbf{P}_\ell)\mathbf{Q}\mathbf{x} - 2w_\ell^* t_{\ell+1}\bigr) \\
&\quad = \sum_{j=1}^{J-1} \bigl(w_{m_j}^{*2} \mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x} - 2w_{m_j}^* (T_{m_{j+1}} - T_{m_j})\bigr) \\
&\qquad + \bigl(\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x} - 2(T_M - T_{m_J})\bigr),
\end{aligned}$$

which is minimized by

$$w_{m_j}^* = \frac{T_{m_{j+1}} - T_{m_j}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}. \tag{48}$$

This establishes (26)–(27).

It remains to show (28). Assume that $m_J < M$, which means that $w_{m_J}^* = 1$. Consider minimizing (25), allowing the models $\{m_1, \ldots, m_J, M\}$ to have positive weight (and setting the remaining models to have zero weight). We know (by assumption) that the solution puts positive weight on the models $\{m_1, \ldots, m_J\}$ and zero weight on model M. The constrained optimization problem can be written as

$$\min_{w^*, \lambda} C^*\bigl(\mathbf{w}^*, \mathbf{x}\bigr) - \lambda_1 w_{m_1}^* - \sum_{j=2}^{J} \lambda_j \bigl(w_{m_j}^* - w_{m_{j-1}}^*\bigr) - \lambda_{J+1}\bigl(1 - w_{m_J}^*\bigr),$$

where $\lambda_j \ge 0$ are Kuhn–Tucker multipliers enforcing the constraints $w_1^* \ge 0$, $w_{m_j}^* \ge w_{m_j-1}^*$, and $w_{m_J}^* \le 1$, respectively. The first-order condition for $w_{m_J}^*$ can be solved to find

$$w_{m_J}^* = \frac{T_M - T_{m_J} + \lambda_J - \lambda_{J+1}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}}.$$

Since m_J is a model with positive weight, then $w_J^* > w_{J-1}^*$ and $\lambda_J = 0$ (as the constraint is not binding). Also, since model M receives zero weight, then $w_{m_J}^* = 1$. Thus

$$1 = w_{m_J}^* = \frac{T_M - T_{m_J} - \lambda_{J+1}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}} \le \frac{T_M - T_{m_J}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}},$$

which implies (28) as desired. □

The proof of Theorem 2 will require the application of a famous result known as Stein's lemma. We use a version from Hansen (2013, Lemma 1). (See Stein (1981).)

LEMMA 4. *If* $\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}) \in \mathbb{R}^K$, $\mathbf{V} > 0$, *and* $\eta(\mathbf{x}) : \mathbb{R}^K \to \mathbb{R}^K$ *is absolutely continuous, then*

$$E\big(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{Q}\mathrm{Z}\big) = E\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{Q}\mathbf{V}\right).$$

Stein's lemma allows a simple calculation of the asymptotic risk for estimators with asymptotic distributions that take the form $\mathrm{Z} - \eta(\mathrm{Z}+\boldsymbol{\delta})$.

PROOF OF THEOREM 2. From (21) in Theorem 1, the averaging estimator has the asymptotic distribution

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{Z} - \eta(\mathrm{Z}+\boldsymbol{\delta}).$$

Thus

$$R(\widehat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) = E\big(\mathrm{Z}'\mathbf{Q}\mathrm{Z}\big) - 2E\big(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{Q}\mathrm{Z}\big) + E\big(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{Q}\eta(\mathrm{Z}+\boldsymbol{\delta})\big).$$

As discussed in the text, $E(\mathrm{Z}'\mathbf{Q}\mathrm{Z}) = \mathrm{tr}(\mathbf{Q}^{-1}\boldsymbol{\Omega})$, and by Lemma 4 and $\mathbf{Q}\mathbf{V} = \boldsymbol{\Omega}\mathbf{Q}^{-1}$,

$$E\big(\eta(\mathrm{Z}+\boldsymbol{\delta})'\mathbf{Q}\mathrm{Z}\big) = E\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}'}\eta(\mathrm{Z}+\boldsymbol{\delta})\boldsymbol{\Omega}\mathbf{Q}^{-1}\right).$$

Hence (31) holds with

$$q(\mathbf{x}) = 2\,\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{x}}\eta(\mathbf{x})'\boldsymbol{\Omega}\mathbf{Q}^{-1}\right) - \eta(\mathbf{x})'\mathbf{Q}\eta(\mathbf{x}). \tag{49}$$

We now show that (49) equals (32).

From (22) and (26)–(27), we see that

$$\begin{aligned}\eta(\mathbf{x}) &= \sum_{j=1}^{J-1}\sum_{\ell=m_j}^{m_{j+1}-1} w^*_{m_j}(\mathbf{P}_{\ell+1} - \mathbf{P}_\ell)\mathbf{Q}\mathbf{x} + \sum_{\ell=m_J}^{M-1}(\mathbf{P}_{\ell+1} - \mathbf{P}_\ell)\mathbf{Q}\mathbf{x}\mathbf{1}_{(m_J<M)} \\ &= \sum_{j=1}^{J-1}\frac{(T_{m_{j+1}} - T_{m_j})}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x} + (\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}\mathbf{1}_{(m_J<M)},\end{aligned}$$

where, for simplicity, we do not write J and m_j explicitly as functions of $\mathbf{x}$.

Using Lemma 3, we calculate that

$$\eta(\mathbf{x})'\mathbf{Q}\eta(\mathbf{x}) = \sum_{j=1}^{J-1}\left(\frac{(T_{m_{j+1}} - T_{m_j})^2}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}\right) + \mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}\mathbf{1}_{(m_J<M)}.$$

Next,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{x}}\eta(\mathbf{x})' &= \sum_{j=1}^{J-1}\frac{T_{m_{j+1}}-T_{m_j}}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\\ &\quad - 2\sum_{j=1}^{J-1}\frac{T_{m_{j+1}}-T_{m_j}}{(\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x})^2}\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\\ &\quad + \mathbf{Q}(\mathbf{P}_M-\mathbf{P}_{m_J})\mathbf{1}_{(m_J<M)}.\end{aligned}$$

We calculate that

$$\begin{aligned}\operatorname{tr}\bigl(\mathbf{Q}\mathbf{P}_m\boldsymbol{\Omega}\mathbf{Q}^{-1}\bigr) &= \operatorname{tr}(\mathbf{P}_m\boldsymbol{\Omega})\\ &= \operatorname{tr}\bigl(\mathbf{S}_m\bigl(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\bigr)^{-1}\mathbf{S}_m'\boldsymbol{\Omega}\bigr)\\ &= \operatorname{tr}\bigl(\bigl(\mathbf{S}_m'\mathbf{Q}\mathbf{S}_m\bigr)^{-1}\mathbf{S}_m'\boldsymbol{\Omega}\mathbf{S}_m\bigr)\\ &= D_m\end{aligned}$$

and, thus,

$$\operatorname{tr}\bigl(\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\boldsymbol{\Omega}\mathbf{Q}^{-1}\bigr) = D_{m_{j+1}}-D_{m_j}.$$

Hence,

$$\begin{aligned}&\operatorname{tr}\biggl(\frac{\partial}{\partial \mathbf{x}}\eta(\mathbf{x})'\boldsymbol{\Omega}\mathbf{Q}^{-1}\biggr)\\ &\quad = \sum_{j=1}^{J-1}\frac{(T_{m_{j+1}}-T_{m_j})(D_{m_{j+1}}-D_{m_j})}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}\\ &\qquad - 2\sum_{j=1}^{J-1}\frac{(T_{m_{j+1}}-T_{m_j})}{(\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x})^2}\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\boldsymbol{\Omega}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}\\ &\qquad + (D_M-D_{m_J})\mathbf{1}_{(m_J<M)}.\end{aligned}$$

Thus,

$$\begin{aligned}q(\mathbf{x}) &= \sum_{j=1}^{J-1}\frac{(T_{m_{j+1}}-T_{m_j})[2(D_{m_{j+1}}-D_{m_j})-(T_{m_{j+1}}-T_{m_j})]}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}\\ &\quad - 4\sum_{j=1}^{J-1}\frac{(T_{m_{j+1}}-T_{m_j})}{(\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x})^2}\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\boldsymbol{\Omega}(\mathbf{P}_{m_{j+1}}-\mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}\\ &\quad + \bigl[2(D_M-D_{m_J})-\mathbf{x}'\mathbf{Q}(\mathbf{P}_M-\mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}\bigr]\mathbf{1}_{(m_J<M)}.\end{aligned}$$

This is (32). □

PROOF OF THEOREM 3. The inequality $\mathbf{b}'\boldsymbol{\Omega}\mathbf{b} \le \mathbf{b}'\mathbf{b}\lambda_{\max}(\boldsymbol{\Omega})$ with $\mathbf{b} = \mathbf{Q}^{1/2}\mathbf{c}$ implies

$$\begin{aligned}\mathbf{c}'\boldsymbol{\Omega}\mathbf{c} &= \mathbf{c}'\mathbf{Q}^{1/2}\mathbf{Q}^{-1/2}\boldsymbol{\Omega}\mathbf{Q}^{-1/2}\mathbf{Q}^{1/2}\mathbf{c} \\ &\le \mathbf{c}'\mathbf{Q}\mathbf{c}\lambda_{\max}\big(\mathbf{Q}^{-1/2}\boldsymbol{\Omega}\mathbf{Q}^{-1/2}\big) \\ &= \sigma^2\mathbf{c}'\mathbf{Q}\mathbf{c}\overline{\lambda},\end{aligned}$$

where $\overline{\lambda}$ is defined in (33). Setting $\mathbf{c} = (\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}$, we find

$$\begin{aligned}&\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\boldsymbol{\Omega}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x} \\ &\quad\le \sigma^2\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}\overline{\lambda} \\ &\quad= \sigma^2\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}\overline{\lambda},\end{aligned}$$

where the final equality uses Lemma 3. Equation (28) shows that

$$\mathbf{x}'\mathbf{Q}(\mathbf{P}_M - \mathbf{P}_{m_J})\mathbf{Q}\mathbf{x}\mathbf{1}_{(m_J<M)} \le \sigma^2(T_M - T_{m_J})\mathbf{1}_{(m_J<M)}.$$

Together,

$$\begin{aligned}q(\mathbf{x}) &\ge \sigma^4\sum_{j=1}^{J-1}\frac{(T_{m_{j+1}} - T_{m_j})[2(D_{m_{j+1}} - D_{m_j}) - (T_{m_{j+1}} - T_{m_j}) - 4\overline{\lambda}]}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}} \\ &\quad+ \sigma^2\big[2(D_M - D_{m_J}) - (T_M - T_{m_J})\big]\mathbf{1}_{(m_J<M)} \\ &= \sigma^4\sum_{j=1}^{J-1}\frac{(T_{m_{j+1}} - T_{m_j})}{\mathbf{x}'\mathbf{Q}(\mathbf{P}_{m_{j+1}} - \mathbf{P}_{m_j})\mathbf{Q}\mathbf{x}}\left[\left(\sum_{\ell=m_j}^{m_{j+1}-1}(2d_{\ell+1} - t_{\ell+1})\right) - 4\overline{\lambda}\right] \\ &\quad+ \sigma^2\sum_{\ell=m_J}^{M-1}(2d_{\ell+1} - t_{\ell+1})\mathbf{1}_{(m_J<M)} \\ &\ge 0.\end{aligned}$$

The final inequality holds since Assumption 3 implies $2d_{\ell+1} - t_{\ell+1} > 4\overline{\lambda}$ for all ℓ.

For any $\mathbf{x}$ such that $m_J(\mathbf{x}) < M$, we have the strict inequality

$$q(\mathbf{x}) \ge \sigma^2\sum_{\ell=m_J}^{M-1}(2d_{\ell+1} - t_{\ell+1}) \ge \sigma^2(M - m_J)\overline{\lambda} > 0.$$

Thus $q(\mathbf{x}) \ge 0$ for all $\mathbf{x}$ and $q(\mathbf{x}) > 0$ for some $\mathbf{x}$. Since Z has a continuous distribution, we deduce that $E(q(\mathrm{Z} + \boldsymbol{\delta})) > 0$. Thus

$$\begin{aligned}R(\widehat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) &= \operatorname{tr}\big(\mathbf{Q}^{-1}\boldsymbol{\Omega}\big) - E\big(q(\mathrm{Z} + \boldsymbol{\delta})\big) \\ &= R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta}) - E\big(q(\mathrm{Z} + \boldsymbol{\delta})\big) \\ &< R(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \boldsymbol{\beta})\end{aligned}$$

and (35) holds. □

PROOF OF LEMMA 2. We show below that

$$\overline{\lambda} \leq \overline{\sigma}^2 \tag{50}$$

and

$$d_m \geq \underline{\sigma}^2 k_m. \tag{51}$$

It follows from (51), Assumption 4(a), the definition of r, and (50) that

$$d_m \geq \underline{\sigma}^2 k_m > 2\underline{\sigma}^2 r = 2\overline{\sigma}^2 \geq 2\overline{\lambda}$$

and thus Assumption 3(a) is satisfied.

Similarly, it follows from Assumption 4(b), (50), and (51) that

$$0 < t_m \leq 2k_m \underline{\sigma}^2 - 4\overline{\sigma}^2 \leq 2d_m - 4\overline{\lambda}$$

and thus Assumption 3(b) is satisfied.

We now show (50) and (51). Using the property of the maximum eigenvalue and the bound $E(e_i^2|\mathbf{x}_i) = \sigma^2(\mathbf{x}_i) \leq \overline{\sigma}^2$,

$$\begin{aligned}
\overline{\lambda} &= \lambda_{\max}\bigl(\mathbf{Q}^{-1/2}\boldsymbol{\Omega}\mathbf{Q}^{-1/2}\bigr) \\
&= \max_{u'u=1} u'\mathbf{Q}^{-1/2}E\bigl(\mathbf{x}_i\mathbf{x}_i'e_i^2\bigr)\mathbf{Q}^{-1/2}u \\
&= \max_{u'u=1} E\bigl(\bigl(u'\mathbf{Q}^{-1/2}\mathbf{x}_i\bigr)^2 e_i^2\bigr) \\
&= \max_{u'u=1} E\bigl(\bigl(u'\mathbf{Q}^{-1/2}\mathbf{x}_i\bigr)^2 \sigma^2(\mathbf{x}_i)\bigr) \\
&\leq \max_{u'u=1} E\bigl(u'\mathbf{Q}^{-1/2}\mathbf{x}_i\bigr)^2 \overline{\sigma}^2 \\
&= \lambda_{\max}\bigl(\mathbf{Q}^{-1/2}E\bigl(\mathbf{x}_i\mathbf{x}_i'\bigr)\mathbf{Q}^{-1/2}\bigr)\overline{\sigma}^2 \\
&= \lambda_{\max}(\mathbf{I}_K)\overline{\sigma}^2 \\
&= \overline{\sigma}^2.
\end{aligned}$$

This establishes (50).

Similarly,

$$\begin{aligned}
d_m &= E\bigl(\bigl(\overline{\mathbf{x}}_{mi}'\mathbf{Q}_m^{-1}\overline{\mathbf{x}}_{mi} - \overline{\mathbf{x}}_{m-1i}'\mathbf{Q}_{m-1}^{-1}\overline{\mathbf{x}}_{m-1i}\bigr)e_i^2\bigr) \\
&\geq E\bigl(\overline{\mathbf{x}}_{mi}'\mathbf{Q}_m^{-1}\overline{\mathbf{x}}_{mi} - \overline{\mathbf{x}}_{m-1i}'\mathbf{Q}_{m-1}^{-1}\overline{\mathbf{x}}_{m-1i}\bigr)\underline{\sigma}^2 \\
&= \bigl(\operatorname{tr}\bigl(E\bigl(\overline{\mathbf{x}}_{mi}\overline{\mathbf{x}}_{mi}'\bigr)\mathbf{Q}_m^{-1}\bigr) - \operatorname{tr}\bigl(E\bigl(\overline{\mathbf{x}}_{m-1i}\overline{\mathbf{x}}_{m-1i}'\bigr)\mathbf{Q}_{m-1}^{-1}\bigr)\bigr)\underline{\sigma}^2 \\
&= (K_m - K_{m-1})\underline{\sigma}^2 \\
&= k_m\underline{\sigma}^2.
\end{aligned}$$

This establishes (51). □

References

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on Information Theory* (B. Petroc and F. Csake, eds.), 267–281, Akadémia Kiadó, Budapest. [495, 496]

Akaike, H. (1979), "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting." *Biometrika*, 66, 237–242. [496, 510]

Baranchick, A. (1964), "Multiple regression and estimation of the mean of a multivariate normal distribution." Technical Report 51, Department of Statistics, Stanford University. [496]

Bates, J. M. and C. M. W. Granger (1969), "The combination of forecasts." *Operational Research Quarterly*, 20, 451–468. [497]

Berger, J. O. and D. K. Dey (1983), "Combining coordinates in simultaneous estimation of normal means." *Journal of Statistical Planning and Inference*, 8, 143–160. [497]

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997), "Model selection: An integral part of inference." *Biometrics*, 53, 603–618. [497, 510]

Burnham, K. P. and D. R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York. [496, 510]

Clemen, R. T. (1989), "Combining forecasts: A review and annotated bibliography." *International Journal of Forecasting*, 5, 559–583. [497]

Danilov, D. and J. R. Magnus (2004), "On the harm that ignoring pretesting can cause." *Journal of Econometrics*, 122, 27–46. [497]

Dey, D. K. and J. O. Berger (1983), "On truncation of shrinkage estimators in simultaneous estimation of normal means." *Journal of the American Statistical Association*, 78, 865–869. [497]

Diebold, F. X. and J. A. Lopez (1996), "Forecast evaluation and combination." In *Handbook of Statistics*, Vol. 14 (G. S. Maddala and C. R. Rao, eds.), 241–268, North-Holland, Amsterdam. [497]

DiTraglia, F. J. (2013), "Using invalid instruments on purpose: Focused moment selection and averaging for GMM." Working paper, University of Pennsylvania. [497]

Efron, B. and C. Morris (1973a), "Combining possibly related estimation problems." *Journal of the Royal Statistical Society, Series B*, 35, 379–421. [497]

Efron, B. and C. Morris (1973b), "Stein's estimation rule and its competitors: An empirical Bayes approach." *Journal of the American Statistical Association*, 68, 117–130. [496]

Fryer, R. G., Jr. and S. D. Levitt (2013), "Testing for racial differences in the mental ability of young children." *American Economic Review*, 103, 981–1005. [495, 497, 515, 517]

George, E. I. (1986a), "Minimax multiple shrinkage estimation." *The Annals of Statistics*, 14, 188–205. [497]

George, E. I. (1986b), "Combining minimax shrinkage estimators." *Journal of the American Statistical Association*, 81, 437–445. [497]

Hansen, B. E. (2007), "Least squares model averaging." *Econometrica*, 75, 1175–1189. [495, 496, 497, 500, 507, 509]

Hansen, B. E. (2013), "Efficient shrinkage in parametric models." Working paper, University of Wisconsin. [495, 524]

Hansen, B. E. and J. S. Racine (2012), "Jackknife model averaging." *Journal of Econometrics*, 167, 38–46. [497, 508]

Hendry, D. F. and M. P. Clements (2004), "Pooling of forecasts." *Econometrics Journal*, 7, 1–31. [497]

Hjort, N. L. and G. Claeskens (2003), "Frequentist model average estimators." *Journal of the American Statistical Association*, 98, 879–899. [495, 497, 503]

Inoue, A. and L. Kilian (2008), "How useful is bagging in forecasting economic time series? A case study of U.S. consumer inflation." *Journal of the American Statistical Association*, 103, 511–522. [497]

James, W. and C. M. Stein (1961), "Estimation with quadratic loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 361–380, University of California Press, Berkeley. [496]

Judge, G. and M. E. Bock (1978), *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North-Holland, Amsterdam. [496]

Kim, J. and D. Pollard (1990), "Cube root asymptotics." *The Annals of Statistics*, 18, 191–219. [522]

Knox, T., J. H. Stock, and M. W. Watson (2004), "Empirical Bayes regression with many regressors." Working paper, Princeton University. [497]

Kuersteiner, G. and R. Okui (2010), "Constructing optimal instruments by first-stage prediction averaging." *Econometrica*, 78, 697–718. [497]

Leamer, E. E. (1978), *Specification Searches: ad hoc Inference With Nonexperimental Data*. Wiley, New York. [496]

Lee, Y. and Y. Zhou (2011), "Averaged instrumental variables estimator." Working paper, University of Michigan. [497]

Lehmann, E. L. and G. Casella (1998), *Theory of Point Estimation*, second edition. Springer, New York. [496, 497]

Liang, H., G. Zou, A. T. K. Wan, and X. Zhang (2011), "Optimal weight choice for frequentist model average estimators." *Journal of the American Statistical Association*, 106, 1053–1066. [497]

Liao, Z. (2012), "Adaptive GMM shrinkage estimation with consistent moment selection." Working paper, UCLA. [497]

Liu, C.-A. (2012), "A plug-in averaging estimator for regressions with heteroskedastic errors." Working paper, National University of Singapore. [497]

Liu, Q. and R. Okui (forthcoming), "Heteroskedasticity-robust C_p model averaging." *Econometrics Journal*. [497, 508]

Magnus, J. R. (2002), "Estimation of the mean of a univariate normal distribution with known variance." *Econometrics Journal*, 5, 225–236. [497]

Magnus, J. R., O. Powell, and P. Prüfer (2010), "A comparison of two model averaging techniques with an application to growth empirics." *Journal of Econometrics*, 154, 139–153. [497]

Mallows, C. L. (1973), "Some comments on C_p." *Technometrics*, 15, 661–675. [495, 496]

McCloskey, A. (2012), "Bonferroni-based size-correction for nonstandard testing problems." Working paper, Brown University. [497]

Min, C.-K. and A. Zellner (1993), "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates." *Journal of Econometrics*, 56, 89–118. [497]

Saleh, A. K. Md. E. (2006), *Theory of Preliminary Test and Stein-Type Estimation With Applications*. Wiley, Hoboken. [495, 497]

Schorfheide, F. (2005), "VAR forecasting under misspecification." *Journal of Econometrics*, 128, 99–136. [495, 503]

Schwarz, G. (1978), "Estimating the dimension of a model." *The Annals of Statistics*, 6, 461–464. [496]

Stein, C. M. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 197–206, University of California Press, Berkeley. [496]

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 9, 1135–1151. [496, 524]

Stock, J. H. and M. W. Watson (2006), "Forecasting with many predictors." In *Handbook of Economic Forecasting*, Vol. 1 (G. Elliott, C. W. J. Granger, and A. Timmermann, eds.), 515–554, North-Holland, Amsterdam. [497]

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [496, 510]

Timmermann, A. (2006), "Forecast combinations." In *Handbook of Economic Forecasting*, Vol. 1 (G. Elliott, C. W. J. Granger, and A. Timmermann, eds.), 135–196, North-Holland, Amsterdam. [497]

Wan, A. T. K., X. Zhang, and G. Zou (2010), "Least squares model averaging by Mallows criterion." *Journal of Econometrics*, 156, 277–283. [497]

Submitted November, 2012. Final version accepted August, 2013.