Journal of Econometrics 174 (2013) 82–94

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Model averaging by jackknife criterion in models with dependent data

Xinyu Zhang [a], Alan T.K. Wan [b,*], Guohua Zou [a]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[b] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 6 September 2011
Received in revised form
1 August 2012
Accepted 19 January 2013
Available online 9 February 2013

JEL classification:
C51
C52

Keywords:
Asymptotic optimality
Autocorrelation
Cross-validation
Lagged dependent variables
Model averaging
Squared error

ABSTRACT

The past decade witnessed a literature on model averaging by frequentist methods. For the most part, the asymptotic optimality of various existing frequentist model averaging estimators has been established under i.i.d. errors. Recently, Hansen and Racine [Hansen, B.E., Racine, J., 2012. Jackknife model averaging. Journal of Econometrics 167, 38–46] developed a jackknife model averaging (JMA) estimator, which has an important advantage over its competitors in that it achieves the lowest possible asymptotic squared error under heteroscedastic errors. In this paper, we broaden Hansen and Racine's scope of analysis to encompass models with (i) a non-diagonal error covariance structure, and (ii) lagged dependent variables, thus allowing for dependent data. We show that under these set-ups, the JMA estimator is asymptotically optimal by a criterion equivalent to that used by Hansen and Racine. A Monte Carlo study demonstrates the finite sample performance of the JMA estimator in a variety of model settings.

1. Introduction

Model averaging is an alternative to model selection. While model selection attempts to find a single best model for the given purpose, model averaging compromises across the competing models, thus providing a kind of insurance against selecting a very poor model. Model averaging has long been a popular approach within the Bayesian paradigm. In recent years, frequentist model averaging (FMA) has also made substantial grounds. Contributions to model averaging from a fully-fledged frequentist standpoint were made by Buckland et al. (1997), Yang (2001), Hjort and Claeskens (2003, 2006), Yuan and Yang (2005), Hansen (2007, 2008), Goldenshluger (2009), Schomaker et al. (2010), Wan et al. (2010), Liang et al. (2011), Zhang and Liang (2011), Zhang et al. (2012), among others. The majority of these studies emphasize model weights determination, inference based on model averaging, and asymptotic efficiency and finite sample properties of FMA estimators under a variety of model settings. Useful surveys of this rapidly expanding body of literature are given in Claeskens and Hjort (2008) and Wang et al. (2009). There is also an emerging empirical literature that employs FMA in applied settings (Kapetanios et al., 2008a,b; Pesaran et al., 2009; Wan and Zhang, 2009).

In a recent article, Hansen and Racine (2012) (hereafter referred to as HR, 2012) developed a jackknife model averaging (JMA) estimator that selects model weights by minimizing a cross-validation criterion. One major advantage of the JMA estimator is that the asymptotic optimality theory developed for it allows for heteroscedasticity in the errors, whereas those developed for other existing FMA schemes virtually all assume i.i.d. errors. HR (2012) showed that the JMA estimator has the smallest asymptotic expected squared errors relative to a large class of linear estimators constructed from a countable set of weights, including the least squares, ridge, Nadaraya–Watson and local polynomial kernel with fixed bandwidths, spline and some other nonparametric estimators. HR's (2012) Monte Carlo results also suggest that the JMA estimator is generally preferred to several other model selection and averaging estimators; in particular, when the errors are heteroscedastic, the JMA estimator significantly outperforms the Mallows model average (MMA) estimator developed by Hansen (2007) in mean squared error (MSE) terms in a large part of the parameter space. In view of these merits of the JMA estimator, more investigations into its properties are warranted.

Although HR's (2012) model set-up admits heteroscedastic errors, it rules out serial correlations in the errors. Their set-up also assumes complete exogeneity of regressors. An interesting question is whether the JMA estimator remains meritorious under other settings, particularly in models that admit dependent data. The

* Corresponding author. Tel.: +852 34427146; fax: +852 34420189.
E-mail addresses: msawan@cityu.edu.hk, Alan.Wan@cityu.edu.hk (A.T.K. Wan).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2013.01.004

current paper takes steps in this direction by enlarging HR's scope of analysis to include two other commonly encountered model settings, both involving dependent data. The first setting retains the regressor exogeneity assumption as in HR but admits a non-diagonal covariance structure in the errors, thus allowing for error processes such as ARMA and GARCH in addition to pure heteroscedastic and i.i.d. processes; it also nests the model of HR as a special case. The second setting allows for lagged dependent variables but assumes that the errors are i.i.d. Although neither of these two model settings is more general than the other, they both allow dependent data, a property not shared by the model examined in HR. We prove that the JMA estimator when applied to these model settings achieves an asymptotic optimality criterion equivalent to that under the model set-up of HR. Our theoretical analysis follows the approach of Wan et al. (2010) by allowing the model weights to be continuous. This is unlike the method of HR which follows that of Hansen (2007) by restricting the weights to a discrete set. We consider the extension from discrete to continuous weighting an advance as the latter has obvious appeal. It is instructive to point out that the conditions required for optimality by our method are neither stronger nor weaker than those required by HR's method. Like the latter method, our method also allows for an infinite number of models. Detailed comparisons of the technical conditions that underpin our theoretical results and those of HR are provided in Sections 2.2 and 2.3. For our second model setting that involves lagged dependent variables, we prove the asymptotic optimality of the JMA estimator using results in Ing and Wei (2003).

In a Monte Carlo study we also compare the finite sample performance of the JMA estimator with several other estimators, including the MMA, leave-one-out cross-validation, and AIC and BIC model selection estimators under the two model set-ups considered. Our Monte Carlo results suggest that under strictly exogenous regressors and ARMA and GARCH-type errors, the JMA estimator is frequently preferred to these alternative estimators. On the other hand, when the regressors are not strictly exogenous but contain lagged dependent variables, the JMA estimator has comparable efficiency to the MMA estimator. The latter estimator is known to exhibit performance superiority in many settings (Hansen, 2007, 2008).

The plan of this paper is as follows. In Section 2 we examine the JMA criterion and present results on the asymptotic optimality of the JMA estimator under a setting that assumes exogeneity of regressors but allows for both serial correlation and heteroscedasticity. While the main theorem in this section is applicable to general linear estimators, a special focus of discussion will be given to least squares estimation in the linear regression model. Section 3 examines the case of an infinite order linear autoregressive (AR) data generating process, and JMA being performed across models containing lagged dependent variables and possibly other regressors. Section 4 reports results of the Monte Carlo study. Section 5 concludes, and proofs of theorems are contained in Appendix.

2. Jackknife model averaging under a non-diagonal error covariance structure

2.1. Model framework and the jackknife criterion

We follow HR's (2012) notations as much as possible for readers' convenience. Wherever appropriate we point out the differences in the two set-ups. Consider the data generating process (DGP)

$$y_i = \mu_i + e_i = f(x_i) + e_i, \quad i = 1, \ldots, n, \tag{1}$$

with $x_i = (x_{i1}, x_{i2}, \ldots)$ being countably infinite, and $f(\cdot)$ a function with respect to x_i. Write $y = (y_1, \ldots, y_n)'$, $X = (x_1', \ldots, x_n')'$, $\mu = (\mu_1, \ldots, \mu_n)'$, and $e = (e_1, \ldots, e_n)'$. Further, assume that $E(e|X) = 0$ so that $\mu = E(y|X)$, and denote $Var(e|X) = \Omega$, where Ω is a positive definite symmetric matrix.

Let M_n be the number of candidate models in the model average, and $\{\widehat{\mu}^1, \ldots, \widehat{\mu}^{M_n}\}$ be a set of linear estimators of μ such that the mth estimator in the set, i.e., the estimator of μ in the mth model, may be written as $\widehat{\mu}^m = \mathbf{P}_m y$, where $\mathbf{P}_m$ is dependent on X but not on y. Many well-known estimators including the least squares, ridge, nearest neighbors, and spline are members of this class. Now, let $w = (w^1, \ldots, w^{M_n})'$ be a weight vector in the continuous set:

$$\mathcal{H}_n = \left\{ w \in [0, 1]^{M_n} : \sum_{m=1}^{M_n} w^m = 1 \right\}.$$

The model averaging estimator of μ is obtained by compromising across the linear estimators $\{\widehat{\mu}^1, \ldots, \widehat{\mu}^{M_n}\}$ in the model space. It has the form

$$\widehat{\mu}(w) = \sum_{m=1}^{M_n} w^m \widehat{\mu}^m = \sum_{m=1}^{M_n} w^m \mathbf{P}_m y \equiv \mathbf{P}(w) y. \tag{2}$$

The above set-up is the same as that of HR (2012) except for the following aspects. First, HR (2012) restricted Ω to be a diagonal matrix, but we permit Ω to be non-diagonal, thus allowing the errors to be both autocorrelated and heteroscedastic. This also allows y to be dependent when the design matrix X is assumed fixed. Second, although HR (2012) defined $\widehat{\mu}(w)$ as in (2), when proving the asymptotic optimality of the JMA estimator, they restricted $\mathcal{H}_n$ to the subset $\mathcal{H}_n^*$, which consists of the discrete weights w^m from the set $\{0, 1/N, 2/N, \ldots, 1\}$ for some positive integer N. We do not impose the same restriction in our analysis.

Denote $\tilde{\mu}^m$ as the estimator of μ when jackknife estimation based on the delete-one cross-validation is used. Write $\tilde{\mu}^m = (\tilde{\mu}_1^m, \ldots, \tilde{\mu}_n^m)'$, where $\tilde{\mu}_i^m$ is the estimator of μ_i obtained with the ith observation (y_i, x_i) removed from the sample. Thus, we can write $\tilde{\mu}^m = \tilde{\mathbf{P}}_m y$, where $\tilde{\mathbf{P}}_m$ has zeros on the diagonal and depends only on X. The model averaging estimator that smooths across the M_n jackknife estimators is thus

$$\tilde{\mu}(w) = \sum_{m=1}^{M_n} w^m \tilde{\mu}^m = \sum_{m=1}^{M_n} w^m \tilde{\mathbf{P}}_m y \equiv \tilde{\mathbf{P}}(w) y. \tag{3}$$

HR (2012) adopted the following squared error loss criterion for choosing the weight vector w:

$$CV_n(w) = \|y - \tilde{\mu}(w)\|^2, \tag{4}$$

where $\|a\|^2 = a'a$. Now, let $\widehat{w} = \arg\min_{w \in \mathcal{H}_n} CV_n(w)$ be the weight vector that minimizes $CV_n(w)$. The JMA estimator of μ is $\widehat{\mu}(\widehat{w})$. It is obtained by substituting $\widehat{w}$ for w in (2). Thus, the JMA estimator is a weighted average of the linear estimators $\widehat{\mu}^m$'s using $\widehat{w}$ as weight. It is different from the estimator $\tilde{\mu}(w)$ which combines the jackknife estimators $\tilde{\mu}^m$'s.

Denote $\tilde{e}^m = y - \tilde{\mu}^m$ and $\tilde{e} = (\tilde{e}^1, \ldots, \tilde{e}^{M_n})$. Then we can write

$$CV_n(w) = w'\tilde{e}'\tilde{e}w, \tag{5}$$

a quadratic function of w. Thus, the minimization of $CV_n(w)$ with respect to w is a quadratic programming problem. Numerous software packages are available for obtaining a solution to this problem (e.g., Matlab and R), and they generally work effectively and efficiently even when M_n is large; for example, when $n = 200$ and $M_n = 100$, it takes only 0.15 s to obtain the solution to (5) by Matlab.

A referee pointed out that one could consider block cross-validation as an alternative to delete-one cross-validation. Although traditionally the choice of block lengths has been an issue, recent advances in automated methods (e.g., Politis and White, 2004; Patton et al., 2009) have made the selection of optimal block length practically feasible. Racine (1997) also showed that the amount of calculations needed for deleting a block can

be the same as that for deleting a single observation. However, the difficulty with block cross-validation here is that the estimator of μ resulting from a data driven method of block length selection is generally a non-linear function of y. This latter feature deviates from our basic analytical framework and poses technical challenges to subsequent asymptotic analysis. Developing a model averaging scheme based on block cross-validation with a valid asymptotic theory remains an interesting point of departure for future research.

2.2. Asymptotic optimality

Unless otherwise stated, all limiting processes discussed in this and subsequent sections are with respect to $n \to \infty$. To evaluate the asymptotic efficiency of the JMA estimator $\widehat{\mu}(\widehat{w})$, we consider the following squared error loss and associated risk criteria:

$$L_n(w) = \|\widehat{\mu}(w) - \mu\|^2 \tag{6}$$

and

$$R_n(w) = E\{L_n(w)|X\} = \|\mathbf{A}(w)\mu\|^2 + \mathrm{tr}\{\mathbf{P}(w)\Omega\mathbf{P}(w)\}, \tag{7}$$

where $\mathbf{A}(w) = I_n - \mathbf{P}(w)$. Our objective is to demonstrate that the JMA estimator $\widehat{\mu}(\widehat{w})$ satisfies the optimality condition

$$\text{(OPT)}: \frac{L_n(\widehat{w})}{\inf_{w\in\mathcal{H}_n} L_n(w)} \xrightarrow{p} 1.$$

This optimality criterion is nearly the same as the one considered by HR (2012) – the only difference being that we do not restrict the weights to lie in the discrete set $\mathcal{H}_n^*$. When (OPT) is satisfied, the JMA estimator is said to be optimal in the sense that its squared errors are asymptotically identical to those of the infeasible best possible model averaging estimator.

Now, write $\tilde{\mathbf{A}}(w) = I_n - \tilde{\mathbf{P}}(w)$. Let $\tilde{L}_n(w)$ and $\tilde{R}_n(w)$ be respectively the jackknife squared error loss and risk obtained by replacing $\widehat{\mu}(w)$ by $\tilde{\mu}(w)$, $\mathbf{A}(w)$ by $\tilde{\mathbf{A}}(w)$, and $\mathbf{P}(w)$ by $\tilde{\mathbf{P}}(w)$ in (6) and (7). Let $\xi_n = \inf_{w\in\mathcal{H}_n} R_n(w)$, $\tilde{\Omega} = \Omega - diag(\Omega_{11}, \ldots, \Omega_{nn})$, Ω_{ii} be the ith diagonal element of Ω, w_m^o be a weight vector with the mth element taking on the value of unity and other elements zeros, and $\mathcal{S}(B)$ denote the largest singular value of a matrix B. The following theorem, which extends Theorem 1 of HR (2012), gives the conditions under which the JMA estimator satisfies the (OPT) criterion under a model set-up that permits a non-diagonal error covariance structure.

Theorem 2.1. *If*

$$\overline{\lim_{n\to\infty}} \max_{1\le m\le M_n} \mathcal{S}(\mathbf{P}_m) < \infty, \quad a.s. \tag{8}$$

$$\overline{\lim_{n\to\infty}} \max_{1\le m\le M_n} \mathcal{S}(\tilde{\mathbf{P}}_m) < \infty, \quad a.s. \tag{9}$$

$$\sup_{w\in\mathcal{H}_n} \left|\tilde{R}_n(w)/R_n(w) - 1\right| \xrightarrow{a.s.} 0, \tag{10}$$

$$e|X \sim \mathcal{N}(0, \Omega), \tag{11}$$

$$\mathcal{S}(\Omega) \le \bar{C} < \infty, \quad a.s., \text{ where } \bar{C} \text{ is a constant}, \tag{12}$$

$$M_n\xi_n^{-2G}\sum_{m=1}^{M_n}\left(R_n(w_m^o)\right)^G \xrightarrow{a.s.} 0, \quad \text{for some constant } G \ge 1, \tag{13}$$

and

$$\sup_{w\in\mathcal{H}_n} \left|tr\left(\tilde{\mathbf{P}}(w)\tilde{\Omega}\right)/\tilde{R}_n(w)\right| \xrightarrow{a.s.} 0, \tag{14}$$

then $\widehat{\mu}(\widehat{w})$ satisfies the (OPT) asymptotic optimality criterion.

Remark 1. Conditions (8) and (9) correspond to conditions (A.3) and (A.4) of HR (2012) respectively. Condition (10), which is standard in cross-validation analysis, is almost the same as condition (A.5) of HR (2012), except that the continuous set $\mathcal{H}_n$ is used here in place of the discrete set $\mathcal{H}_n^*$ in HR.

Remark 2. By condition (11), the asymptotic optimality of the JMA estimator applies only to Gaussian regressions. From a generality point of view, this is an obvious limitation of our results; on the other hand, the Gaussian regression model is the most widely used model in practice, and it also plays a central part in the majority of the model selection and averaging literature (e.g., Shibata, 1981; Hurvich and Tsai, 1989; Zhang, 1992; Danilov and Magnus, 2004; Leung and Barron, 2006; Bunea et al., 2007). More importantly, condition (11) can be removed for proving JMA's asymptotic optimality when the coefficients in each model are estimated by least squares (see Section 2.3 for details). Condition (12) corresponds to the moment bound condition (A.2) of HR (2012); it requires the largest singular value of the correlation matrix Ω to be finite as the sample size increases.

Remark 3. Condition (13), adopted from Wan et al. (2010), is also required for proving the asymptotic optimality of the Mallows model averaging estimator. The implications of this condition are thoroughly discussed in Wan et al. (2010). Most importantly, under our set-up, M_n, the number of candidate models, is allowed to be infinite, although condition (13) places a restriction on the rate at which M_n increases with n. It is worth noting that if $M_n/n = O(1)$ (see Assumption 4 of Kuersteiner and Okui, 2010), then condition (13) is implied by the weaker condition $\xi_n^{-2}R_n(w_m^o) = O(n^{-\tilde{c}})$ for any m, with $\tilde{c}$ being a positive constant. Wan et al. (2010) gave an example[1] under Hansen's (2007) nested model set-up where condition (13) is implied by the condition

$$M_n = O(n^v), \tag{15}$$

with v being a positive constant (see Example 2 of Wan et al., 2010, for details). Conditions similar to (15) can be found in Shibata (1980) and Newey (1997). When the DGP is a linear model (see Section 2.3), condition (13) may be replaced by (21) or (25). Note that although HR's (2012) Theorem 1 does not require condition (13), it does require an alternative condition (A.6) which also has implications on the rate at which M_n grows with n. Their condition (A.6) may be viewed as a counterpart to our condition (13). It is straightforward to show that HR's condition (A.6) implies

$$M_n^N / \sup_{w\in\mathcal{H}_n} R_n(w)^{N+1} \xrightarrow{a.s.} 0, \tag{16}$$

which is neither strictly stronger nor weaker than (13). For example, suppose that $\sup_{w\in\mathcal{H}_n} R_n(w) \sim n$ and $\xi_n \sim n^{3/5}$; if $M_n \sim n^2$, then (13) holds but HR's condition (A.6) does not hold. On the other hand, suppose that $M_n \sim n^{1/3}$ and $\xi_n \sim n^{1/3}$; if $\min_{m\in\{1,\ldots,M_n\}} R_n(w_m^o) \sim n^{3/4}$, then HR's condition (A.6) holds but our condition (13) does not hold.

To reiterate, both HR's and our optimality theories allow M_n to be infinite, but both require conditions on the rate of increase of M_n, and neither their nor our conditions are strictly stronger or weaker than the other's.

Remark 4. Condition (14) is related to the correlation pattern of e. This condition is not required when the underlying DGP is a linear model.

Remark 5. The technical conditions that underlie Theorem 2.1 simplify substantially when $e_1, \ldots, e_n$ are independent such that Ω is a diagonal matrix. For this special case, the asymptotic

[1] In that example, $R_n(w_m^o)$ is assumed to be of order $nm^{-a} + m$ with $a > 0$, the same order as the risk of the series estimator given in Theorem 1 of Newey (1997).

optimality (OPT) criterion holds for the JMA estimator $\widehat{\mu}(\widehat{w})$ provided that conditions (8), (9), (10) and (13) are satisfied,

$$\sup_i E(e_i^{4G}|x_i) < \infty \quad \text{a.s.} \tag{17}$$

and

$$\inf_i \sigma_i^2 \geq \underline{\sigma}^2 > 0, \quad \text{a.s.} \tag{18}$$

where $\underline{\sigma}$ is a constant. The above result may be compared with Theorem 1 of HR (2012). In particular, conditions (17) and (18) are the same as conditions (A.2) and (A.1) of HR (2012) respectively. The major difference between this result and Theorem 1 of HR (2012) is that our result permits the model weights to be continuous, whereas Theorem 1 of HR (2012) constrains the weights to the discrete set $\mathcal{H}_n^*$. The proof of the above result is given in Appendix.

2.3. The special case of linear regression

In this subsection, we focus on the special case of a linear regression model. By least squares estimation,

$$\mathbf{P}_m = X^m(X^{m\prime}X^m)^{-1}X^{m\prime}, \tag{19}$$

where X^m is the regressor matrix of the mth candidate model. It is assumed that X^m is of full column rank. Now, it is straightforward to show that

$$\tilde{\mathbf{P}}_m = \mathbf{D}_m(\mathbf{P}_m - I_n) + I_n, \tag{20}$$

where $\mathbf{D}_m$ is a diagonal matrix with $(1 - h_{ii}^m)^{-1}$ being its ith element, h_{ii}^m the ith diagonal element of $\mathbf{P}_m$, and I_n is an $n \times n$ identity matrix. Let $\mathbf{X} = \left[X^1, \ldots, X^{M_n}\right]$, $r = \text{rank}(\mathbf{X})$, $\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^- \mathbf{X}'$, and k_m be the number of regressors in X^m. The following theorem relates to the asymptotic optimality of $\widehat{\mu}(\widehat{w})$.

Theorem 2.2. *Provided that condition (12) holds,*

$$\xi_n^{-2} \sum_{m=1}^{M_n} R_n(w_m^o) \xrightarrow{a.s.} 0, \tag{21}$$

$$r\xi_n^{-1} \xrightarrow{a.s.} 0, \tag{22}$$

$$\mu'\mu n^{-1} = O(1), \quad a.s. \tag{23}$$

and

$$h_{ii}^m \leq \Lambda k_m n^{-1}, \quad a.s. \tag{24}$$

where Λ is a constant, then $\widehat{\mu}(\widehat{w})$ satisfies the (OPT) asymptotic optimality criterion.

Remark 6. Condition (21) is similar to condition (13), but with M_n removed and G set to 1. When $G = 1$, condition (21) is clearly weaker than (13). Assigning unity to G is justified because the elements in e are *not* required to be normal here. It is worth mentioning that condition (21) may be replaced by

$$r\mu'\mathbf{P}\mu\xi_n^{-2} \xrightarrow{a.s.} 0, \tag{25}$$

which carries no obvious intuitive meaning but is sometimes weaker than condition (21). For instance, in Example 1 of Wan et al. (2010),[2] a sufficient condition for (25) to hold is $s_n = O(n^{3/17})$, but (21) does not hold under the same condition. For (21) to hold, a sufficient condition is $s_n = O(\log n)$, which is considerably stronger than $s_n = O(n^{3/17})$. The proof relevant to this remark is simple, and is available upon request from the authors.

Remark 7. Condition (22) places a constraint on the growth rate of the number of regressors. HR (2012) did not use the same restriction; instead they imposed a restriction on the growth rate of the largest number of models of any given dimension (see condition (A.8) in HR, 2012). These two restrictions are based on different perspectives, and neither of them is strictly stronger nor weaker than the other. To illustrate the latter point, consider again Example 1 of Wan et al. (2010); when $s_n \sim n^{3/10}$, condition (22) holds but the same is not true for condition (A.8) of HR (2012) (the proof is simple and is available on request); on the other hand, when the models are nested, condition (A.8) of HR (2012) holds as $\xi_n \to \infty$, but in order for our condition (22) to hold, the growth rate in the number of regressors in the largest model must be slower than $\xi_n \to \infty$.

Remark 8. Condition (23), which concerns the average of μ_i^2, is quite common and reasonable. A similar condition can be found in Shao (1997) and Wan et al. (2010). Condition (24) is commonly used in studies of asymptotic optimality of cross-validation methods (e.g., Li, 1987; Andrews, 1991).

3. Jackknife model averaging in the presence of lagged dependent variables

While the preceding analysis extends the work of HR to allow for a non-diagonal error covariance matrix, it retains the assumption of strict exogeneity of regressors as in HR (2012). As such, the theory developed in the last section does not allow for dynamic models that include lagged values of the dependent variable as regressors. In this section, we take some steps in developing an asymptotic optimality theory for the JMA estimator in models where some columns of the regressor matrix contain lagged values of y. Our techniques rely on applying results in mathematical statistics developed by Ing and Wei (2003). In order for their results to be applicable, it is necessary to assume that y_i follows the stationary AR(∞) process

$$y_i + \sum_{l=1}^{\infty} a_l y_{i-l} = e_i, \quad i = \ldots, -1, 0, 1, \ldots \tag{26}$$

where i is an index of time, e_i's are i.i.d., each with a mean of zero and variance of σ^2, $\sum_{l=1}^{\infty} |a_l| < \infty$, and $1 + \sum_{l=1}^{\infty} a_l z^l$ is bounded away from zero for $|z| \leq 1$. Let $\mu_i = E(y_i|y_{i-1}, y_{i-2}, \ldots)$.

Notwithstanding (26), we allow the inclusion of "extraneous regressors" in addition to lagged values of y in the regressor matrix of the candidate models. For analytical convenience, we assume that all models regardless of lag orders use the same number of observations. Let $r_1(\geq 0)$ be the maximal lag order in any candidate model, and observations of y_i be available for $i = -r_1 + 1, \ldots, n$, so that $y = (y_1, \ldots, y_n)'$ is the vector of observations of the dependent variable in *every* candidate model. Define $y^{(-j)} = (y_{1-j}, \ldots, y_{n-j})'$, $j = 1, \ldots, r_1$, write $Y_L = (y^{(-1)}, \ldots, y^{(-r_1)})$, and let X^* be an $n \times r_2$ matrix containing observations of $r_2(\geq 0)$ extraneous regressors. The regressor matrix in the full model (i.e., the model that contains all regressors) is therefore $\mathbf{X} = (Y_L, X^*)$, and it is assumed that $\mathbf{X}$ has (full column) rank $r = r_1 + r_2$. The regressor matrix X^m of the mth candidate model is formed by combining columns in $\mathbf{X}$. Thus, our framework allows for pure AR as well as ARX models, and nested as well as non-nested AR processes in the lagged component of the model. When n increases, we allow r_1 to increase but assume that r_2 is fixed.

As before, we focus our interest on $\mu = (\mu_1, \ldots, \mu_n)'$, and apply the delete-one cross-validation method discussed previously

[2] In that example, the DGP is $y_i = \mu_i + e_i = f(x_i) + e_i = \sum_{s=1}^{\infty} s^{-7/12} \cos((s-1)x_i)/s + e_i, i = 1, \ldots, n$, where $x_i = 2\pi(i-1)/n$ and e_i's are i.i.d. $N(0, \sigma^2)$. The model averaging estimator compromises across models that comprise subsets of the first s_n ($s_n < S$) regressors, with S being the largest integer that is not greater than $n/2$.

to choose the weight vector $\widehat{w}$ in the JMA estimator $\widehat{\mu}(\widehat{w})$. Now, let $V_n(w) = \|\mathbf{A}(w)\mu\|^2 + \sigma^2 \mathrm{tr}\left(\mathbf{P}(w)\mathbf{P}'(w)\right)$, $\tilde{V}_n(w) = \|\tilde{\mathbf{A}}(w)\mu\|^2 + \sigma^2 \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\tilde{\mathbf{P}}'(w)\right)$, $\xi_n^* = \inf_{w\in\mathcal{H}_n} V_n(w)$, $\tilde{\xi}_n^* = \inf_{w\in\mathcal{H}_n} \tilde{V}_n(w)$, $\mathbf{M} = I_n - Y_L(Y_L'Y_L)^{-1}Y_L'$, and $F_i(\cdot)$ be the distribution function of e_i, where $\mathbf{A}(w)$, $\mathbf{P}(w)$, $\tilde{\mathbf{A}}(w)$, and $\tilde{\mathbf{P}}(w)$ are defined as in Section 2. The asymptotic optimality of the JMA estimator is given in the following theorem:

Theorem 3.1. *The JMA estimator $\widehat{\mu}(\widehat{w})$ under the analytical framework described in this section satisfies the (OPT) asymptotic optimality criterion provided that the following conditions hold:*

(C.1). $r\xi_n^{*-1} = o_p(1)$, $\mu'\mu n^{-1} = O_p(1)$, $nk_m^{-1}h_{ii}^m = O_p(1)$, *and* $r\mu'\mathbf{P}\mu\xi_n^{*-2} = o_p(1)$;

(C.2). $n^{-1/2}X^{*\prime}e \xrightarrow{d} N(0, \Delta)$, *where* Δ *is a positive definite matrix,* $\mathcal{S}(n^{-1}X^{*\prime}X^*) = O_p(1)$, *and* $\mathcal{S}((n^{-1}X^{*\prime}\mathbf{M}X^*)^{-1}) = O_p(1)$;

(C.3). *there exist some positive constants* α_1, α_2 *and* α_3 *such that* $|F_i(d_1) - F_i(d_2)| \le \alpha_3|d_1 - d_2|^{\alpha_1}$ *for all* i *when* $|d_1 - d_2| \le \alpha_2$; *and*

(C.4). *either* $r_1^{6+\alpha_4} = O(n)$ *for some* $\alpha_4 > 0$ *and* $\sup_{-\infty<i<\infty} Ee_i^4 < \infty$, *or* $r_1^{2+\alpha_4} = O(n)$ *for some* $\alpha_4 > 0$ *and* $\sup_{-\infty<i<\infty} Ee_i^s < \infty$ *for all* s.

Remark 9. Condition (C.1) is analogous to the conditions described in (22)–(25) being fulfilled in probability. The first two parts of condition (C.2) hold when $\{X_i^*e_i\}$ is a stationary and ergodic martingale difference sequence with finite fourth moments, and $n^{-1}X^{*\prime}X^*$ converges to a positive definite matrix in probability, respectively. For the third part, note that by decomposing $\mathbf{M} = P_{\mathbf{M}}'\Lambda_{\mathbf{M}}P_{\mathbf{M}}$, we have $n^{-1}X^{*\prime}\mathbf{M}X^* = n^{-1}X^{*\prime}P_{\mathbf{M}}'\Lambda_{\mathbf{M}}P_{\mathbf{M}}X^*$, where $P_{\mathbf{M}}$ is an orthogonal matrix, and $\Lambda_{\mathbf{M}} = diag(1, \ldots, 1, 0, \ldots, 0)$ is a diagonal matrix with $n - r_1$ elements of unity and remaining elements of zeros since $\mathbf{M}$ is symmetric and idempotent. Now, let X^o be a matrix taking on the first $n - r_1$ rows of $P_{\mathbf{M}}X^*$. When $(n - r_1)^{-1}X^{o\prime}X^o$ converges to a positive definite matrix in probability, by condition (C.4), $n^{-1}X^{*\prime}\mathbf{M}X^* = n^{-1}(n-r_1)(n-r_1)^{-1}X^{o\prime}X^o$ converges to the same matrix and thus the third part of condition (C.2) also holds.

Remark 10. Condition (C.3) is the same as condition (K.2) of Ing and Wei (2003), who studied the properties of the least squares predictor in an increasing order setting when the observations are generated by an AR(∞) process. As discussed in Ing and Wei (2003), this is a mild condition easily fulfilled by any distribution with a bounded density. Condition (C.4) is a restatement of the conditions in Theorem 2 of Ing and Wei (2003). For (C.4) to hold, a tradeoff between the rate of increase of r_1 and the existence of higher moments is needed.

4. A Monte Carlo study

This section reports results of a Monte Carlo study undertaken to compare the performance of the JMA estimator with (1) the AIC model selection estimator (AIC), (2) the BIC model selection estimator (BIC), (3) the leave-one-out cross-validation model selection estimator (CV), and (4) the MMA (model averaging) estimator of Hansen (2007). Our study is based on four sampling designs ranging from linear and non-linear exogenous set-ups to a dynamic model set-up. To conserve space and facilitate discussion, we report the Monte Carlo results in graphical form with the captions shown in the top left diagram of each set of graphs. The details and results of the Monte Carlo study are discussed below.

Design 1: Our first Monte Carlo design is based on the same set-up as that of HR (2012), except that HR focused on a pure heteroscedastic error process, whereas we consider an error process that is both autocorrelated and heteroscedastic. Specifically, the DGP being examined is

$$y_i = \mu_i + e_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + e_i, \quad i = 1, \ldots, n,$$

where $x_{i1} = 1$, and observations of all other x_{ij}'s are generated from the $N(0, 1)$ distribution and are independent; $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 0.5$; the error process is given by $e_i = e_{i,1} + e_{i,2}$, $e_{i,1}$'s are independent observations from the $N(0, x_{i2}^2)$ distribution, and $e_{i,2}$ follows an AR(1) process with an autocorrelation coefficient ψ set to either 0.5 or 0.9. The sample size varies at 25, 50, 75, and 100. The number of approximating models is determined by $M_n = INT(3n^{1/3})$, where the function $INT(A)$ returns the nearest integer from A. This results in $M_n = 9$, 11, 13 and 14, for $n = 25$, 50, 75, and 100 respectively.

The approximating models are $y_i = \sum_{j=1}^{m} \theta_j x_{ij} + e_i$, $i = 1, \ldots, n$, $m = 1, \ldots, M_n$. Following HR (2012), we let $\tilde{R}^2 = c^2/(1 + c^2)$ vary between 0.1 and 0.8. We use the following mean squared error (MSE) measure to assess the accuracy of estimators:

$$\sum_{d=1}^{D} \left\|\widehat{\mu}(w)^{(d)} - \mu^{(d)}\right\|^2 / D, \tag{27}$$

where $D = 10000$ is the number of simulation trials. To facilitate comparisons, the MSEs of all estimators are normalized by the MSE of the infeasible optimal least squares estimator.

The results of the simulations are depicted in Figs. 1 and 2. We find that the JMA estimator is almost always the best estimator among those considered, although when $\tilde{R}^2$ is very large, the MMA estimator can sometimes be marginally preferred to the JMA estimator. As shown in Fig. 1, when $\psi = 0.5$, the MSE difference in favor of the JMA estimator over other estimators is very noticeable for small to moderate $\tilde{R}^2$. When $\psi = 0.9$, the general pattern of results is similar, except that the gap in MSE between the JMA and other estimators is smaller than when $\psi = 0.5$. In the overwhelming majority of cases, either of the two model averaging estimators is preferred to any of the three model selection estimators, although occasionally small to moderate reductions in MSE can be made with model selection; for example, the BIC estimator can sometimes yield a smaller MSE than the MMA estimator when $\tilde{R}^2$ is small. Of the three model selection estimators, the AIC estimator frequently has the worst performance, and it performs especially poorly when $\tilde{R}^2$ is small. By and large, our results accord with those of HR (2012), and our findings reinforce their conclusion that significant efficiency gains can be made with the JMA method.

Design 2: This is based on the same set-up as Design 1, except that the data generating process of e_i here is assumed to be the GARCH(p, q) process: $e_i = \sigma_i v_i$; $\sigma_i^2 = 0.5 + \Sigma_{j=1}^{p} a_j\sigma_{i-j}^2 + \Sigma_{j=1}^{q} b_j e_{i-j}^2$; and $v_i \sim$ *i.i.d.* $N(0, 1)$. We consider two GARCH specifications: GARCH(1, 1) with $a_1 = 0.1$ and $b_1 = 0.8$; and GARCH(2, 2) with $a_1 = 0.2$, $a_2 = 0.3$, $b_1 = 0.2$ and $b_2 = 0.2$. The results are summarized in Figs. 3 and 4.

Broadly speaking, the conclusions are similar to those found under Design 1. In particular, the JMA estimator frequently yields the most accurate estimates followed by the MMA estimator, and both model averaging estimators enjoy significantly smaller MSEs than the model selection estimators over a large portion of the $\tilde{R}^2$ space. With the GARCH specifications, we find evidence of the BIC model selection estimator outperforming the JMA estimator when $\tilde{R}^2$ is small (say, $\tilde{R}^2 \le 0.2$), and a much smaller difference in MSE between the JMA and MMA estimators. For the GARCH(2, 2)

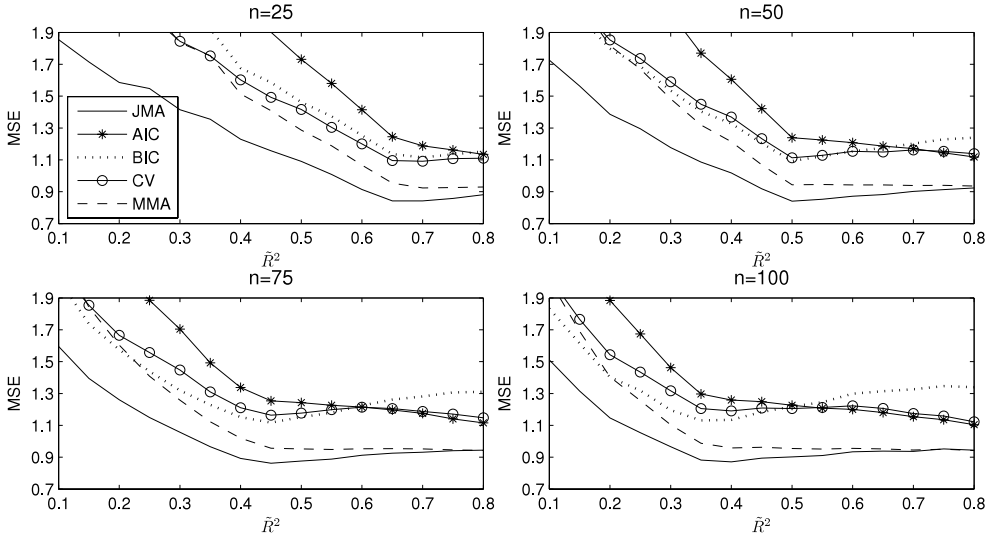

Fig. 1. Results for Monte Carlo Design 1 with $\psi = 0.5$.

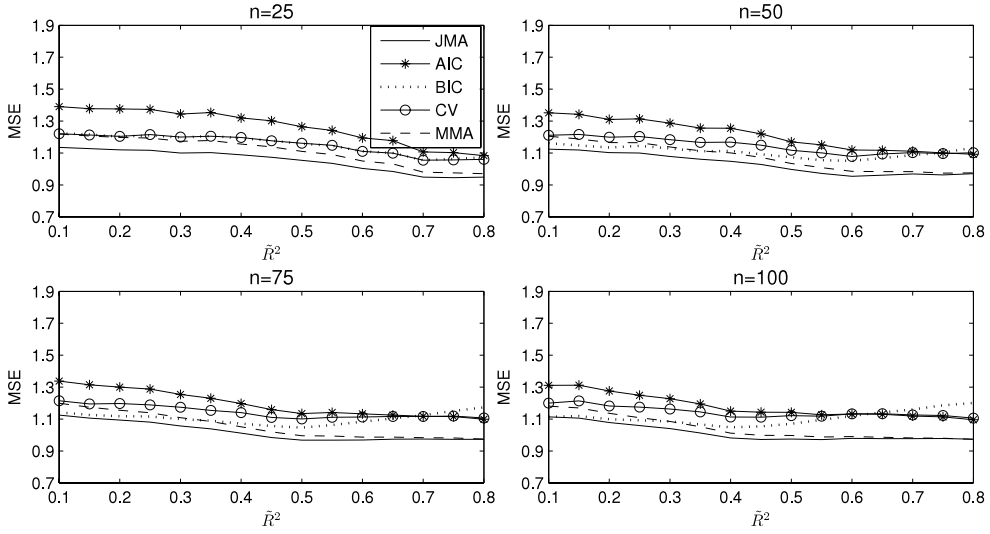

Fig. 2. Results for Monte Carlo Design 1 with $\psi = 0.9$.

specification, the MSEs of the two estimators almost coincide everywhere, with the JMA estimator having only a slight edge in most cases. Under the GARCH(1, 1) specification, moderate gains can still be made by using the JMA estimator in lieu of the MMA estimator.

Design 3: Our third Monte Carlo design is based on the non-linear model:

$$y_i = \mu_i + e_i = cg(x_i) + e_i, \quad i = 1, \ldots, n,$$

where c is the constant defined in Design 1 for controlling $\tilde{R}^2$, $x_i \sim U(-2, 2)$, e_i follows the same AR(1) process as in Design 1 with ψ set to either 0.5 or 0.9, and $n = 25$, 50, 75 and 100. We set $g(x_i)$ to $\sin(5\pi x_i)$ for generating y_i. In the Monte Carlo replications, we estimate $g(\cdot)$ by cubic B-splines with the number of knots chosen from $\{0, 1, \ldots, INT(n^{1/3})\}$. This results in $M_n = 1 + INT(n^{1/3})$ approximating models. With the chosen values of n, we have $M_n =$ 4, 5, 5 and 6.

The results, presented in Figs. 5 and 6, are similar to those obtained under Design 1. Again, when $\psi = 0.5$, the JMA estimator is seen to be the best estimator with the MMA being a close second in a large region of the $\tilde{R}^2$ space. When $\psi = 0.9$, the JMA estimator continues to outperform other estimators frequently but by smaller margins. When $\tilde{R}^2$ is small, the AIC estimator is habitually the worst performing estimator, but when $\tilde{R}^2$ is large and $n \geq 50$, the BIC estimator usually yields the least accurate estimates. On the other hand, when $\tilde{R}^2$ is small and $n \geq 50$, the BIC model selection estimator can outperform the JMA estimator.

Design 4: Our last experimental design considers a dynamic linear model. We use the following ARMA(1, 1) data generating process of y_i to represent the AR(∞) process described in Section 3:

$$y_i = \tilde{a} y_{i-1} + e_i + 0.5 e_{i-1}, \tag{28}$$

where $e_i \sim N(0, 1)$, and $\tilde{a}$ is determined by the value of $R^2 = (var(y_i) - var(e_i))/var(y_i)$. We assume that the candidate models follow the AR specification: $y_i = \tilde{b} + e_i$, $y_i = \tilde{b} + \sum_{j=1}^{1} \theta_j y_{i-j} + e_i, \ldots, y_i = \tilde{b} + \sum_{j=1}^{M_n} \theta_j y_{i-j} + e_i$, where $\tilde{b}$ is an intercept and $\tilde{M}_n = INT(3n^{1/3})$, with n being the number of observations *available for estimation* in each model. Note that M_n, the number of candidate models, equals $\tilde{M}_n + 1$ under this set-up. We let n vary at $n = 15$, 25, 50 and 75. There are 8, 10, 12 and 14 candidate models when n is set to 15, 25, 50 and 75, respectively.

We gauge the different estimators' accuracy in predicting y_{n+1}, i.e., the one-step-ahead forecast from the last observation of the

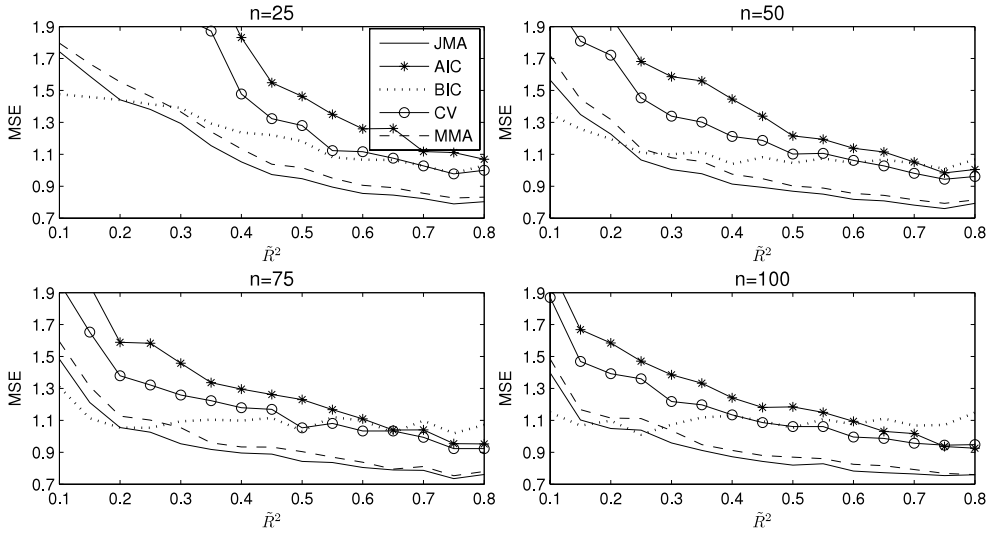

Fig. 3. Results for Monte Carlo Design 2 with $a_1 = 0.1$ and $b_1 = 0.8$.

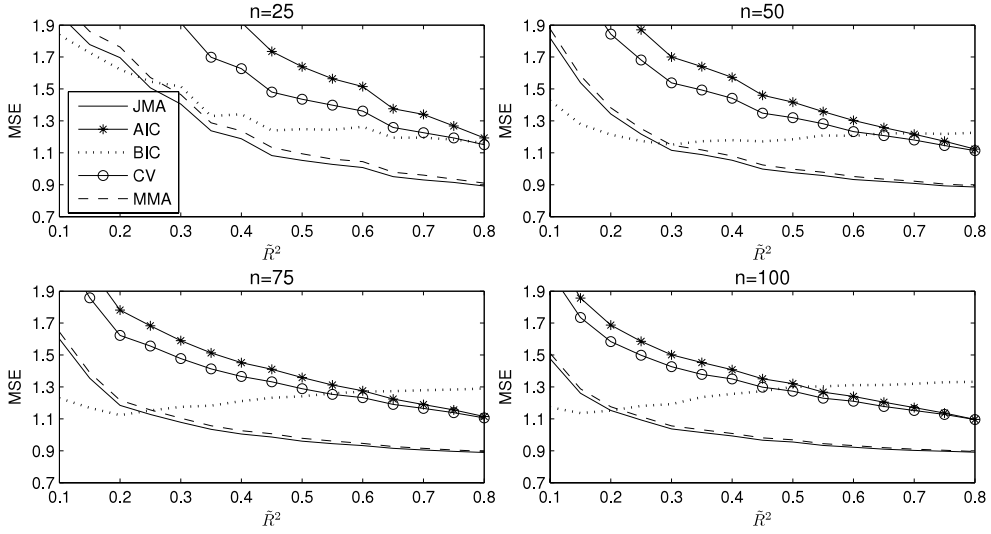

Fig. 4. Results for Monte Carlo Design 2 with $a_1 = 0.2$, $a_2 = 0.3$, $b_1 = 0.2$ and $b_2 = 0.2$.

sample. The mean squared forecast error (MSFE) of the estimator $\widehat{y}_{n+1}$ is $E\,\|\widehat{y}_{n+1} - y_{n+1}\|^2$. Following Hansen (2008), we evaluate the estimators based on a modified MSFE (mMSFE) measure obtained by subtracting $var(e_{n+1}) = 1$ from the MSFE. Again, our experiment is based on $D = 10000$ simulation trials. Fig. 7 reports the results. In all cases, the JMA and MMA estimators have smaller mMSFE than the three model selection estimators, with the superiority being more marked when n is small. When $n = 15$, the JMA estimator has a slight edge over the MMA estimator; when n is larger (say 25), the two estimators have virtually the same mMSFE everywhere in the R^2 space. The (relative) performance of the BIC estimator changes from being worse to better than the CV estimator when n increases from 15 to 25, while the AIC estimator is the worst estimator for all sample sizes.

5. Concluding remarks

This paper extends the work of HR (2012) on the properties of the JMA estimator to model settings with either strictly exogenous regressors and a non-diagonal error covariance structure, or lagged dependent regressors and i.i.d. errors. Our investigation substantially broadens HR's (2012) scope of analysis which assumes strict exogeneity of regressors, allowing for heteroscedasticity but excluding autocorrelation. We show that under these extended set-ups, the JMA estimator remains asymptotically optimal in the sense of achieving the smallest possible squared errors, in addition to performing very favorably in finite samples compared with several other model averaging and selection estimators.

Throughout the paper we assume a continuous weight set for model averaging. This differs from the approach of HR (2012) which restricts the weights to a discrete set. We consider the extension from discrete to continuous weighting an advance as the latter has obvious appeals. It is worth reiterating that continuous weighting allows an infinite number of candidate models, and while it places constraints on the rate of expansion of the number of models, these constraints are not stronger or weaker than those imposed under HR's discrete weighting regime for the same purpose and a strict ordering is not possible. In practice, when there are many candidate models, one strategy is to apply model screening procedures to remove the very poor models prior to combining (e.g., Yuan and Yang, 2005). These procedures have proved to be quite useful and they ease computational

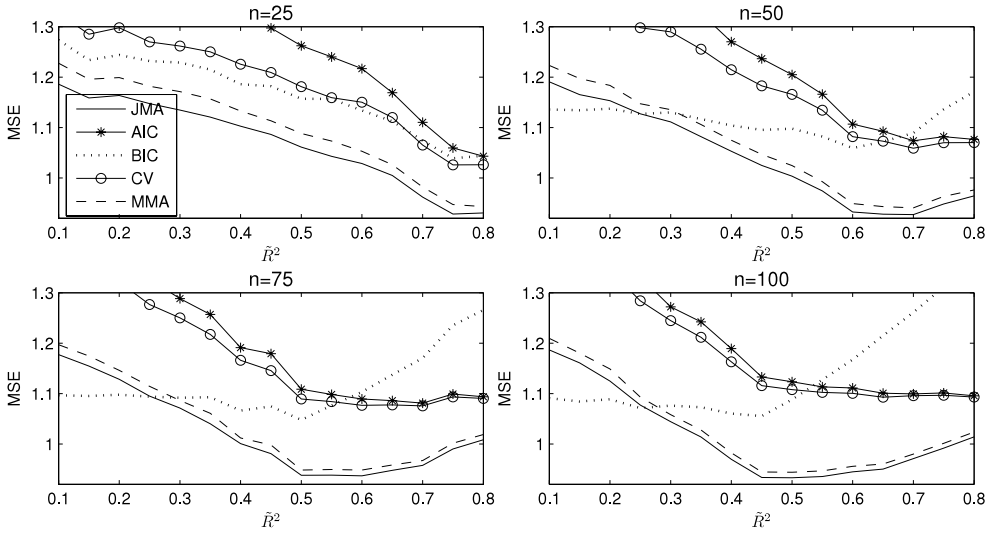

Fig. 5. Results for Monte Carlo Design 3 with $\psi = 0.5$.

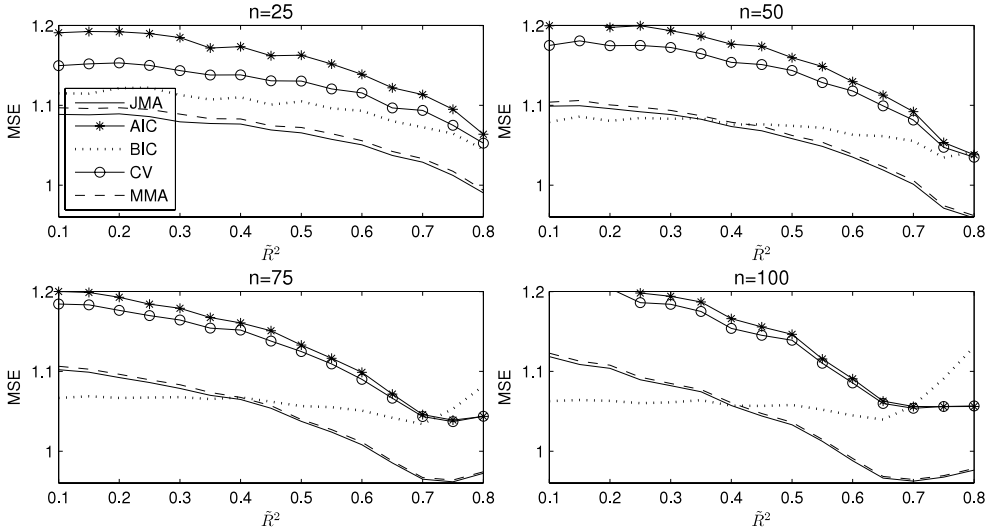

Fig. 6. Results for Monte Carlo Design 3 with $\psi = 0.9$.

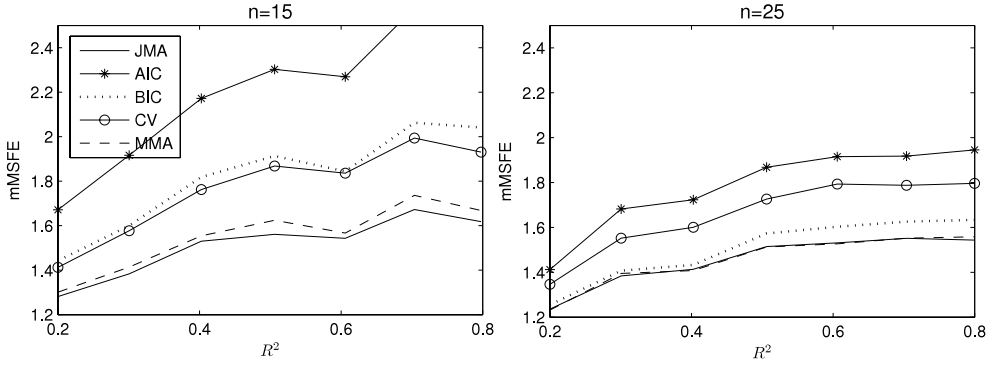

Fig. 7. Results for Monte Carlo Design 4.

burden significantly. It remains for future research to consider model screening in conjunction with JMA. Also, both our and HR's studies emphasize the efficiency gains of using JMA for point estimation. Little is known about the implications of JMA for inference. To deal with the latter aspect, knowledge of the distribution of the JMA estimator is required. Liu (2011) recently derived some results on the asymptotic distribution of the JMA estimator assuming a linear model under local misspecification (Hjort and Claeskens, 2003). Further analysis under a more general model setting would be in order.

Acknowledgments

Zhang's work was supported by the National Natural Science Foundation of China (Grant nos. 71101141 and 70933003), Science Foundation of the Chinese Academy of Sciences, and NCMIS. Wan's work was supported by a General Research Fund from the Hong Kong Research Grants Council (Grant no. CityU-102709), and Zou's work was supported by the National Natural Science Foundation of China (Grant nos. 70625004 and 11021161) and the Hundred Talents Program of the Chinese Academy of Sciences. We thank the editor Han Hong, the associate editor, and three anonymous referees for comments and suggestions on an earlier version of this paper. All remaining errors are our own.

Appendix. Proofs of results

Proof of Theorem 2.1. Let $\tilde{\xi}_n = \inf_{w\in\mathcal{H}_n} \tilde{R}_n(w)$. If $\sup_{w\in\mathcal{H}_n} \left|\frac{\tilde{R}_n(w)}{R_n(w)} - 1\right| \le 1$, then

$$\begin{aligned}
\tilde{\xi}_n^{2G}\left\{\sum_{m=1}^{M_n}\left(\tilde{R}_n(w_m^o)\right)^G\right\}^{-1} &= \left\{\inf_{w\in\mathcal{H}_n}\left(R_n(w)\frac{\tilde{R}_n(w)}{R_n(w)}\right)\right\}^{2G} \\
&\quad\times\left\{\sum_{m=1}^{M_n}\left(R_n(w_m^o)\right)^G\left(\frac{\tilde{R}_n(w_m^o)}{R_n(w_m^o)}\right)^G\right\}^{-1} \\
&\ge \left\{\inf_{w\in\mathcal{H}_n}\frac{\tilde{R}_n(w)}{R_n(w)}\right\}^{2G}\left\{\max_{1\le m\le M_n}\frac{\tilde{R}_n(w_m^o)}{R_n(w_m^o)}\right\}^{-G} \\
&\quad\times\xi_n^{2G}\left\{\sum_{m=1}^{M_n}\left(R_n(w_m^o)\right)^G\right\}^{-1} \\
&\ge \left\{1+\inf_{w\in\mathcal{H}_n}\left(\frac{\tilde{R}_n(w)}{R_n(w)}-1\right)\right\}^{2G} \\
&\quad\times\left\{\sup_{w\in\mathcal{H}_n}\left(\frac{\tilde{R}_n(w)}{R_n(w)}-1\right)+1\right\}^{-G}\xi_n^{2G}\left\{\sum_{m=1}^{M_n}\left(R_n(w_m^o)\right)^G\right\}^{-1} \\
&\ge \left\{1-\sup_{w\in\mathcal{H}_n}\left|1-\frac{\tilde{R}_n(w)}{R_n(w)}\right|\right\}^{2G}\left\{\sup_{w\in\mathcal{H}_n}\left|\frac{\tilde{R}_n(w)}{R_n(w)}-1\right|+1\right\}^{-G} \\
&\quad\times\xi_n^{2G}\left\{\sum_{m=1}^{M_n}\left(R_n(w_m^o)\right)^G\right\}^{-1}. \qquad (A.1)
\end{aligned}$$

Using conditions (10) and (13) in (A.1), we have

$$M_n\tilde{\xi}_n^{-2G}\sum_{m=1}^{M_n}\left(\tilde{R}_n(w_m^o)\right)^G \xrightarrow{a.s.} 0. \qquad (A.2)$$

We next show that

$$\tilde{L}_n(\widehat{w})/\inf_{w\in\mathcal{H}_n}\tilde{L}_n(w) \xrightarrow{p} 1 \qquad (A.3)$$

by using (A.2), together with conditions (9), (11), (12) and (14).

Observe that

$$\begin{aligned}
CV_n(w) &= \tilde{L}_n(w) + \|e\|^2 + 2\mu'\tilde{\mathbf{A}}'(w)e \\
&\quad - 2\left\{e'\tilde{\mathbf{P}}(w)e - \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\right)\right\} - 2\mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\right), \qquad (A.4)
\end{aligned}$$

and $\|e\|^2$ is independent of w. Hence by condition (14), to prove (A.3), it suffices to show that

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)e\right|/\tilde{R}_n(w) \xrightarrow{p} 0, \qquad (A.5)$$

$$\sup_{w\in\mathcal{H}_n}\left|e'\tilde{\mathbf{P}}(w)e - \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\right)\right|/\tilde{R}_n(w) \xrightarrow{p} 0, \qquad (A.6)$$

and

$$\sup_{w\in\mathcal{H}_n}\left|\tilde{L}_n(w)/\tilde{R}_n(w) - 1\right| \xrightarrow{p} 0. \qquad (A.7)$$

For the case of non-stochastic X, following steps similar to the proof of Eq. (A.1) in Wan et al. (2010), and using Chebyshev's inequality, Theorem 2 of Whittle (1960) and (11), we have, for any $\delta > 0$,

$$\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)e\right|/\tilde{R}_n(w) > \delta\right\} \\
&\le \Pr\left\{\sup_{w\in\mathcal{H}_n}\sum_{m=1}^{M_n} w^m\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\
&= \Pr\left\{\max_{1\le m\le M_n}\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\
&= \Pr\left\{\left\{\left|\mu'\tilde{\mathbf{A}}'(w_1^o)e\right| > \delta\tilde{\xi}_n\right\}\cup\cdots\cup\left\{\left|\mu'\tilde{\mathbf{A}}'(w_{M_n}^o)e\right| > \delta\tilde{\xi}_n\right\}\right\} \\
&\le \sum_{m=1}^{M_n}\Pr\left\{\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\
&\le \delta^{-2G}\tilde{\xi}_n^{-2G}\sum_{m=1}^{M_n}E\left(\mu'\tilde{\mathbf{A}}'(w_m^o)\Omega^{1/2}\Omega^{-1/2}e\right)^{2G} \\
&\le C_1\delta^{-2G}\tilde{\xi}_n^{-2G}\sum_{m=1}^{M_n}\left\|\Omega^{1/2}\tilde{\mathbf{A}}(w_m^o)\mu\right\|^{2G} \\
&\le C_1\delta^{-2G}\tilde{\xi}_n^{-2G}\mathcal{S}^G(\Omega)\sum_{m=1}^{M_n}\left\|\tilde{\mathbf{A}}(w_m^o)\mu\right\|^{2G}, \qquad (A.8)
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{H}_n}\left|e'\tilde{\mathbf{P}}(w)e - \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\right)\right|/\tilde{R}_n(w) > \delta\right\} \\
&\le \sum_{m=1}^{M_n}\Pr\left\{\left|e'\tilde{\mathbf{P}}(w_m^o)e - \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\right)\right| > \delta\tilde{\xi}_n\right\} \\
&\le \sum_{m=1}^{M_n}\delta^{-2G}\tilde{\xi}_n^{-2G}E\left[e'\Omega^{-1/2}\Omega^{1/2}\tilde{\mathbf{P}}(w_m^o)\right. \\
&\quad\left.\times\ \Omega^{1/2}\Omega^{-1/2}e - \mathrm{tr}\left(\Omega^{1/2}\tilde{\mathbf{P}}(w)\Omega^{1/2}\right)\right]^{2G} \\
&\le C_2\delta^{-2G}\tilde{\xi}_n^{-2G}\sum_{m=1}^{M_n}\left[\mathrm{tr}\left(\Omega^{1/2}\tilde{\mathbf{P}}(w_m^o)\Omega\tilde{\mathbf{P}}'(w_m^o)\Omega^{1/2}\right)\right]^G \\
&\le C_2\delta^{-2G}\tilde{\xi}_n^{-2G}\mathcal{S}^G(\Omega)\sum_{m=1}^{M_n}\left[\mathrm{tr}\left(\tilde{\mathbf{P}}(w_m^o)\Omega\tilde{\mathbf{P}}'(w_m^o)\right)\right]^G, \qquad (A.9)
\end{aligned}$$

where C_1 and C_2 are two positive constants.

On the other hand, when X is random, by the dominated convergence theorem, conditions (12) and (A.2), the results in (A.5) and (A.6) are implied by (A.8) and (A.9) respectively.

In addition, note that

$$\begin{aligned}
&\left|\tilde{L}_n(w) - \tilde{R}_n(w)\right| \\
&= \left|\|\tilde{\mathbf{P}}(w)e\|^2 - \mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\tilde{\mathbf{P}}'(w)\right) - 2\mu'\tilde{\mathbf{A}}'(w)\tilde{\mathbf{P}}(w)e\right|. \qquad (A.10)
\end{aligned}$$

Hence to prove (A.7), it suffices to show that

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)\tilde{\mathbf{P}}(w)e\right|/\tilde{R}_n(w) \xrightarrow{p} 0 \qquad (A.11)$$

and

$$\sup_{w\in\mathcal{H}_n}\left|\|\tilde{\mathbf{P}}(w)e\|^2-\mathrm{tr}\left(\tilde{\mathbf{P}}(w)\Omega\tilde{\mathbf{P}}'(w)\right)\right|/\tilde{R}_n(w)\xrightarrow{p}0. \tag{A.12}$$

By (A.2), and conditions (9), (11) and (12), (A.11) and (A.12) can be proved along the lines of proving Eqs. (A.4) and (A.5) in Wan et al. (2010). This completes the proof of (A.3).

Likewise, using the technique in deriving (A.7) and condition (8), the following result can also be shown:

$$\sup_{w\in\mathcal{H}_n}|L_n(w)/R_n(w)-1|\xrightarrow{p}0. \tag{A.13}$$

Define

$$V_n(\widehat{w})=\|\mathbf{A}(\widehat{w})\mu\|^2+\mathrm{tr}\left(\mathbf{P}(\widehat{w})\Omega\mathbf{P}'(\widehat{w})\right) \tag{A.14}$$

and

$$\tilde{V}_n(\widehat{w})=\|\tilde{\mathbf{A}}(\widehat{w})\mu\|^2+\sigma^2\mathrm{tr}\left(\tilde{\mathbf{P}}(\widehat{w})\Omega\tilde{\mathbf{P}}'(\widehat{w})\right). \tag{A.15}$$

Now, upon using (10), (A.3), (A.7) and (A.13) in

$$\begin{aligned}
\frac{L_n(\widehat{w})}{\inf_{w\in\mathcal{H}_n}L_n(w)}-1&=\sup_{w\in\mathcal{H}_n}\left(\frac{L_n(\widehat{w})}{L_n(w)}-1\right)\\
&=\sup_{w\in\mathcal{H}_n}\left(\frac{L_n(\widehat{w})}{V_n(\widehat{w})}\frac{R_n(w)}{L_n(w)}\frac{\tilde{R}_n(w)}{R_n(w)}\frac{V_n(\widehat{w})}{\tilde{V}_n(\widehat{w})}\right.\\
&\quad\left.\times\frac{\tilde{V}_n(\widehat{w})}{\tilde{L}_n(\widehat{w})}\frac{\tilde{L}_n(w)}{\tilde{R}_n(w)}\frac{\tilde{L}_n(\widehat{w})}{\tilde{L}_n(w)}-1\right)\\
&\le\sup_{w\in\mathcal{H}_n}\left(\frac{L_n(w)}{R_n(w)}\right)\sup_{w\in\mathcal{H}_n}\left(\frac{R_n(w)}{L_n(w)}\right)\\
&\quad\times\sup_{w\in\mathcal{H}_n}\left(\frac{\tilde{R}_n(w)}{R_n(w)}\right)\sup_{w\in\mathcal{H}_n}\left(\frac{R_n(w)}{\tilde{R}_n(w)}\right)\\
&\quad\times\sup_{w\in\mathcal{H}_n}\left(\frac{\tilde{R}_n(w)}{\tilde{L}_n(w)}\right)\sup_{w\in\mathcal{H}_n}\left(\frac{\tilde{L}_n(w)}{\tilde{R}_n(w)}\right)\frac{\tilde{L}_n(\widehat{w})}{\inf_{w\in\mathcal{H}_n}\tilde{L}_n(w)}-1,
\end{aligned} \tag{A.16}$$

we obtain the desired (OPT) criterion. □

Proof of result under Remark 5 – asymptotic optimality for the special case where Ω is a diagonal matrix

First, note that in the case of a diagonal Ω, condition (12) is implied by (17). Second, when Ω is diagonal and positive definite (implied by condition (18)), it is straightforward to see that the elements of $\Omega^{-1/2}e|X$ are independent. We require this independence condition in order to utilize Theorem 2 of Whittle (1960). In the proof of Theorem 2.1, the only reason for using the normality condition (11) is to make the elements of $\Omega^{-1/2}e|X$ independent. Finally, when Ω is diagonal, $\tilde{\Omega}$ is a zero matrix. Hence condition (14) is satisfied. The OPT criterion thus follows from Theorem 2.1. □

Proof of Theorem 2.2. It is well-known that the following equalities are satisfied for any square matrices B_1 and B_2 with identical dimensions (see, for example, Li, 1987):

$$\mathcal{S}(B_1+B_2)\le\mathcal{S}(B_1)+\mathcal{S}(B_2)\quad\text{and}\quad\mathcal{S}(B_1B_2)\le\mathcal{S}(B_1)\mathcal{S}(B_2). \tag{A.17}$$

Now, let $h^*=\max_{1\le m\le M_n}\max_{1\le i\le n}h_{ii}^m$ and $\tilde{h}=h^*/(1-h^*)$. Then by (24), we have

$$h^*=O(rn^{-1}),\qquad\tilde{h}=O(rn^{-1})\quad\text{a.s.} \tag{A.18}$$

and

$$\begin{aligned}
h^*&\le\Lambda\max\{k_1,\ldots,k_{M_n}\}n^{-1}\xi_n^{-1}\left(\mu'\mathbf{A}_m\mu+\mathrm{tr}[\mathbf{P}_m\Omega\mathbf{P}_m]\right)\\
&\le\Lambda rn^{-1}\xi_n^{-1}\left(\mu'\mathbf{A}_m\mu+\mathrm{tr}[\mathbf{P}_m\Omega\mathbf{P}_m]\right)\\
&\le\Lambda r\xi_n^{-1}\left(\mu'\mu+\mathcal{S}(\Omega)k_m\right)n^{-1},\quad\text{a.s. for } 1\le m\le M_n.
\end{aligned}$$

These results, together with conditions (12), (22) and (23), imply that

$$h^*\to0\quad\text{and}\quad\tilde{h}\to0.\quad\text{a.s.} \tag{A.19}$$

Denote $\mathbf{Q}_m$ as an $n\times n$ diagonal matrix with $\mathbf{Q}_{m,ii}=h_{ii}^m/(1-h_{ii}^m)$. By (20), we have $\tilde{\mathbf{P}}_m=\mathbf{P}_m-\mathbf{Q}_m\mathbf{A}_m$, and $\tilde{\mathbf{A}}_m=\mathbf{A}_m+\mathbf{Q}_m\mathbf{A}_m$. Let $\mathbf{Q}(w)=\sum_{m=1}^{M_n}w^m\mathbf{Q}_m$, $\mathbf{T}_m=\mathbf{Q}_m\mathbf{P}_m$, and $\mathbf{T}(w)=\sum_{m=1}^{M_n}w^m\mathbf{T}_m$. By (A.17), (A.19), and the idempotent symmetric property of $\mathbf{P}_m$s and $\mathbf{A}_m$s, we obtain the following results which hold almost surely for any weight $w\in\mathcal{H}_n$ (as $n\to\infty$):

$$\mathrm{tr}\left[\mathbf{P}(w)\Omega\mathbf{P}(w)\right]\le\mathcal{S}(\mathbf{P}(w))\mathrm{tr}\left[\mathbf{P}(w)\Omega\right]\le\mathrm{tr}\left[\mathbf{P}(w)\Omega\right], \tag{A.20}$$

$$\begin{aligned}
\mathrm{tr}\left[\mathbf{P}(w)\Omega\right]&=\sum_{m=1}^{M_n}w^m\mathrm{tr}\left[\mathbf{P}_m\Omega\right]\\
&\le\mathcal{S}(\Omega)\max\{\mathrm{rank}\mathbf{P}_1,\ldots,\mathrm{rank}\mathbf{P}_{M_n}\}\le\mathcal{S}(\Omega)r,
\end{aligned} \tag{A.21}$$

$$\begin{aligned}
\mathrm{tr}\left[\tilde{\mathbf{P}}(w)\Omega\tilde{\mathbf{P}}'(w)\right]&\le\mathrm{tr}\left[\mathbf{P}(w)\Omega\mathbf{P}(w)\right]+\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{Q}(w)\right]\\
&\quad+\mathrm{tr}\left[\mathbf{T}(w)\Omega\mathbf{T}'(w)\right]+2\left|\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{P}(w)\right]\right|\\
&\quad+2\left|\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{T}'(w)\right]\right|+2\left|\mathrm{tr}\left[\mathbf{T}(w)\Omega\mathbf{P}(w)\right]\right|,
\end{aligned} \tag{A.22}$$

$$\begin{aligned}
\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{Q}(w)\right]&\le\mathcal{S}(\mathbf{Q}(w))\mathrm{tr}\left[\mathbf{Q}(w)\Omega\right]\le\tilde{h}\mathrm{tr}\left[\mathbf{Q}(w)\Omega\right]\\
&\le\tilde{h}\mathcal{S}(\Omega)\sum_{m=1}^{M_n}w^m\mathrm{tr}\mathbf{Q}_m\le\tilde{h}(1-h^*)^{-1}\mathcal{S}(\Omega)\sum_{m=1}^{M_n}w^m\mathrm{tr}\mathbf{P}_m\\
&\le\tilde{h}(1-h^*)^{-1}\mathcal{S}(\Omega)\max\{\mathrm{tr}\mathbf{P}_1,\ldots,\mathrm{tr}\mathbf{P}_{M_n}\}\\
&\le\tilde{h}(1-h^*)^{-1}\mathcal{S}(\Omega)r,
\end{aligned} \tag{A.23}$$

$$\begin{aligned}
\left|\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{P}(w)\right]\right|&=\left|\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{P}(w)+\mathbf{P}(w)\Omega\mathbf{Q}(w)\right]\right|/2\\
&\le\mathcal{S}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)+\mathbf{P}(w)\Omega\mathbf{Q}(w)\right)\\
&\quad\times\mathrm{rank}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)+\mathbf{P}(w)\Omega\mathbf{Q}(w)\right)/2\\
&\le2\mathcal{S}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)\right)\mathrm{rank}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)\right)\\
&=2\mathcal{S}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)\right)\mathrm{rank}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)\mathbf{P}\right)\\
&\le2\mathcal{S}\left(\mathbf{Q}(w)\right)\mathcal{S}\left(\Omega\right)\mathcal{S}\left(\mathbf{P}(w)\right)\mathrm{rank}\left(\mathbf{Q}(w)\Omega\mathbf{P}(w)\mathbf{P}\right)\\
&\le2\tilde{h}\mathcal{S}\left(\Omega\right)r,
\end{aligned} \tag{A.24}$$

$$\begin{aligned}
\left|\mathrm{tr}\left[\mathbf{T}(w)\Omega\mathbf{P}(w)\right]\right|&\le2\mathcal{S}(\mathbf{T}(w)\Omega\mathbf{P}(w))\mathrm{rank}(\mathbf{T}(w)\Omega\mathbf{P}(w))\\
&\le2\mathcal{S}(\Omega)\mathcal{S}(\mathbf{P}(w))\mathcal{S}(\mathbf{T}(w))r\\
&\le2\mathcal{S}(\Omega)\sum_{m=1}^{M_n}w^m\mathcal{S}(\mathbf{Q}_m\mathbf{P}_m)r\le2\mathcal{S}(\Omega)\tilde{h}r,
\end{aligned} \tag{A.25}$$

$$\begin{aligned}
\left|\mathrm{tr}\left[\mathbf{Q}(w)\Omega\mathbf{T}'(w)\right]\right|&\le2\mathcal{S}(\mathbf{Q}(w)\Omega\mathbf{T}'(w))\mathrm{rank}(\mathbf{Q}(w)\Omega\mathbf{T}'(w))\\
&\le2\mathcal{S}(\Omega)\tilde{h}^2r,
\end{aligned} \tag{A.26}$$

$$\begin{aligned}
\mathrm{tr}\left[\mathbf{T}(w)\Omega\mathbf{T}'(w)\right]&\le\mathcal{S}(\mathbf{T}(w)\Omega\mathbf{T}'(w))\mathrm{rank}(\mathbf{T}(w)\Omega\mathbf{T}(w))\\
&\le\mathcal{S}(\Omega)\tilde{h}^2r,
\end{aligned} \tag{A.27}$$

$$\begin{aligned}
|\mu'\mathbf{A}(w)\mathbf{P}(w)e|/R_n(w)&\le\left(e'\mathbf{P}(w)\mathbf{P}(w)e\|\mathbf{A}(w)\mu\|^2/R_n^2(w)\right)^{1/2}\\
&\le\left(e'\mathbf{P}(w)\mathbf{P}(w)e/R_n(w)\right)^{1/2},
\end{aligned} \tag{A.28}$$

$$\begin{aligned}
|\mu'\tilde{\mathbf{A}}'(w)\tilde{\mathbf{P}}(w)e|/\tilde{R}_n(w)&\le\left(e'\tilde{\mathbf{P}}'(w)\tilde{\mathbf{P}}(w)e\|\tilde{\mathbf{A}}(w)\mu\|^2/\tilde{R}_n^2(w)\right)^{1/2}\\
&\le\left(e'\tilde{\mathbf{P}}'(w)\tilde{\mathbf{P}}(w)e/\tilde{R}_n(w)\right)^{1/2},
\end{aligned} \tag{A.29}$$

$$e'\tilde{\mathbf{P}}'(w)\tilde{\mathbf{P}}(w)e \le e'\mathbf{P}(w)\mathbf{P}(w)e + e'\mathbf{Q}(w)\mathbf{Q}(w)e + e'\mathbf{T}'(w)\mathbf{T}(w)e + 2\left|e'\mathbf{P}(w)\mathbf{T}(w)e\right| + 2\left|e'\mathbf{P}(w)\mathbf{Q}(w)e\right| + 2\left|e'\mathbf{Q}(w)\mathbf{T}(w)e\right|, \tag{A.30}$$

$$\left|e'\mathbf{P}(w)\mathbf{Q}(w)e\right|/R_n(w) \le \left(\left[e'\mathbf{Q}(w)\mathbf{Q}(w)e/R_n(w)\right]\left[e'\mathbf{P}(w)\mathbf{P}(w)e/R_n(w)\right]\right)^{1/2}, \tag{A.31}$$

$$\left|e'\mathbf{P}(w)\mathbf{T}(w)e\right|/R_n(w) \le \left(\left[e'\mathbf{T}'(w)\mathbf{T}(w)e/R_n(w)\right]\left[e'\mathbf{P}(w)\mathbf{P}(w)e/R_n(w)\right]\right)^{1/2}, \tag{A.32}$$

$$\left|e'\mathbf{Q}(w)\mathbf{T}(w)e\right|/R_n(w) \le \left(\left[e'\mathbf{T}'(w)\mathbf{T}(w)e/R_n(w)\right]\left[e'\mathbf{Q}(w)\mathbf{Q}(w)e/R_n(w)\right]\right)^{1/2}, \tag{A.33}$$

$$e'\mathbf{Q}(w)\mathbf{Q}(w)e = \sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m e'\mathbf{Q}_t\mathbf{Q}_m e \le \sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m e'e\,\delta(\mathbf{Q}_t\mathbf{Q}_m) \le \tilde{h}^2 e'e, \tag{A.34}$$

$$e'\mathbf{P}(w)\mathbf{P}(w)e \le \delta(P(w))e'\mathbf{P}(w)e \le e'\mathbf{P}(w)e, \tag{A.35}$$

$$\begin{aligned} e'\mathbf{T}'(w)\mathbf{T}(w)e &= e'\sum_{m=1}^{M_n} w^m\mathbf{P}_m\mathbf{Q}_m \sum_{m=1}^{M_n} w^m\mathbf{Q}_m\mathbf{P}_m e \\ &= \sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m e'\mathbf{P}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{P}_m e \\ &= \sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m e'(\mathbf{P}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{P}_m + \mathbf{P}_m\mathbf{Q}_m\mathbf{Q}_t\mathbf{P}_t)e/2 \\ &\le \sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m \delta\left(\mathbf{P}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{P}_m\right) e'e \\ &\le \tilde{h}^2 e'e, \end{aligned} \tag{A.36}$$

$$\begin{aligned} \left|e'\tilde{\mathbf{P}}(w)e\right| &\le \left|e'\sum_{m=1}^{M_n} w_m\mathbf{Q}_m\mathbf{A}_m e\right| + e'\mathbf{P}(w)e \\ &\le \sum_{m=1}^{M_n} w^m\left|e'\mathbf{Q}_m\mathbf{A}_m e\right| + e'\mathbf{P}(w)e \\ &\le \sum_{m=1}^{M_n} w^m\delta\left(\mathbf{Q}_m\mathbf{A}_m\right)e'e + e'\mathbf{P}(w)e \\ &\le \tilde{h}e'e + e'\mathbf{P}(w)e, \end{aligned} \tag{A.37}$$

$$\begin{aligned} &\left|\mu'\tilde{\mathbf{A}}'(w)\tilde{\mathbf{A}}(w)\mu - \mu'\mathbf{A}(w)\mathbf{A}(w)\mu\right|/R_n(w) \\ &= \left|\sum_{t=1}^{M_n}\sum_{m=1}^{M_n}\left\{w^t w^m\left(\mu'\tilde{\mathbf{A}}_t'\tilde{\mathbf{A}}_m\mu - \mu'\mathbf{A}_t\mathbf{A}_m\mu\right)\right\}\right|/R_n(w) \\ &= \left|\sum_{t=1}^{M_n}\sum_{m=1}^{M_n}\left\{w^t w^m\left[\mu'(\mathbf{A}_t + \mathbf{Q}_t\mathbf{A}_t)'\right.\right.\right. \\ &\quad \left.\left.\left.\times\ (\mathbf{A}_m + \mathbf{Q}_m\mathbf{A}_m)\mu - \mu'\mathbf{A}_t\mathbf{A}_m\mu\right]\right\}\right|/R_n(w) \\ &= \left|\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\mathbf{A}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{A}_m\mu\right. \\ &\quad \left. + \ 2\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\mathbf{A}_t\mathbf{Q}_t\mathbf{A}_m\mu\right|/R_n(w), \end{aligned} \tag{A.38}$$

$$\begin{aligned} &\left|\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\mathbf{A}_t\mathbf{Q}_t\mathbf{A}_m\mu\right|/R_n(w) \\ &= \left|\mu'\sum_{m=1}^{M_n} w^m\mathbf{A}_m\mathbf{Q}_m\mathbf{A}(w)\mu\right|/R_n(w) \\ &\le \left(\mu'\sum_{m=1}^{M_n} w^m\mathbf{A}_m\mathbf{Q}_m \times \sum_{m=1}^{M_n} w^m\mathbf{Q}_m\mathbf{A}_m\mu\mu'\mathbf{A}(w)\mathbf{A}(w)\mu/R_n^2(w)\right)^{1/2} \\ &\le \left(\mu'\sum_{m=1}^{M_n} w^m\mathbf{A}_m\mathbf{Q}_m\sum_{m=1}^{M_n} w^m\mathbf{Q}_m\mathbf{A}_m\mu/R_n(w)\right)^{1/2} \\ &= \left(\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\mathbf{A}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{A}_m\mu/R_n(w)\right)^{1/2}, \end{aligned} \tag{A.39}$$

and

$$\begin{aligned} &\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\mathbf{A}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{A}_m\mu/R_n(w) \\ &\le 2\xi_n^{-1}\sum_{t=1}^{M_n}\sum_{m=1}^{M_n} w^t w^m\mu'\left(\mathbf{A}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{A}_m + \mathbf{A}_m\mathbf{Q}_m\mathbf{Q}_t\mathbf{A}_t\right)\mu \\ &\le \xi_n^{-1}\mu'\mu\,\delta\left(\mathbf{A}_t\mathbf{Q}_t\mathbf{Q}_m\mathbf{A}_m\right) \\ &\le \xi_n^{-1}\tilde{h}^2\mu'\mu. \end{aligned} \tag{A.40}$$

Now, by the Markov inequality, for any $\delta > 0$,

$$\begin{aligned} &\Pr\left\{\sup_{w\in\mathcal{H}_n}\tilde{h}e'e/R_n(w) \ge \delta\right\} \\ &\le \Pr\left\{e'e \ge \tilde{h}^{-1}\xi_n\delta\right\} \le E\left\{e'e\right\}\tilde{h}\xi_n^{-1}\delta^{-1} \\ &= \mathrm{tr}(\Omega)\tilde{h}\xi_n^{-1}\delta^{-1} \le \delta(\Omega)n\xi_n^{-1}\delta^{-1}\tilde{h}, \end{aligned} \tag{A.41}$$

$$\begin{aligned} &\Pr\left\{\sup_{w\in\mathcal{H}_n} e'\mathbf{P}(w)e/R_n(w) \ge \delta\right\} \\ &\le \Pr\left\{e'\mathbf{P}e \ge \xi_n\delta\right\} \le E\left\{e'\mathbf{P}e\right\}\xi_n^{-1}\delta^{-1} \\ &= \mathrm{tr}(\mathbf{P}\Omega)\xi_n^{-1}\delta^{-1} \le \delta(\Omega)r\xi_n^{-1}\delta^{-1}, \end{aligned} \tag{A.42}$$

and

$$\begin{aligned} &\Pr\left\{\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)e\right|/\tilde{R}_n(w) > \delta\right\} \\ &\le \Pr\left\{\sup_{w\in\mathcal{H}_n}\sum_{m=1}^{M_n} w^m\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\ &= \Pr\left\{\max_{1\le m\le M_n}\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\ &= \Pr\left\{\left\{\left|\mu'\tilde{\mathbf{A}}'(w_1^o)e\right| > \delta\tilde{\xi}_n\right\}\cup\cdots\cup\left\{\left|\mu'\tilde{\mathbf{A}}'(w_{M_n}^o)e\right| > \delta\tilde{\xi}_n\right\}\right\} \\ &\le \sum_{m=1}^{M_n}\Pr\left\{\left|\mu'\tilde{\mathbf{A}}'(w_m^o)e\right| > \delta\tilde{\xi}_n\right\} \\ &\le \delta^{-2}\tilde{\xi}_n^{-2}\sum_{m=1}^{M_n} E\left(\mu'\tilde{\mathbf{A}}'(w_m^o)e\right)^2 \\ &= \delta^{-2}\tilde{\xi}_n^{-2}\sum_{m=1}^{M_n}\mathrm{tr}\left(\Omega\tilde{\mathbf{A}}(w_m^o)\mu\mu'\tilde{\mathbf{A}}'(w_m^o)\right) \end{aligned}$$

$$= \delta^{-2}\tilde{\xi}_n^{-2}\sum_{m=1}^{M_n}\mathrm{tr}\left(\mu'\tilde{\mathbf{A}}'(w_m^o)\Omega\tilde{\mathbf{A}}(w_m^o)\mu\right)$$
$$\le \delta^{-2}\tilde{\xi}_n^{-2}\mathcal{S}(\Omega)\sum_{m=1}^{M_n}\left\|\tilde{\mathbf{A}}(w_m^o)\mu\right\|^2$$
$$\le \delta^{-2}\tilde{\xi}_n^{-2}\mathcal{S}(\Omega)\sum_{m=1}^{M_n}\tilde{R}_n(w_m^o). \tag{A.43}$$

By (A.20), (A.21), (A.28), (A.35), (A.42), and conditions (12) and (22),

$$\sup_{w\in\mathcal{H}_n}|L_n(w)/R_n(w)-1| \xrightarrow{p} 0. \tag{A.44}$$

By (A.18)–(A.27), (A.38)–(A.40), and conditions (12), (22) and (23),

$$\sup_{w\in\mathcal{H}_n}\left|\tilde{R}_n(w)/R_n(w)-1\right| \xrightarrow{a.s.} 0. \tag{A.45}$$

By (A.43), (A.45), and conditions (12) and (21),

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)e\right|/\tilde{R}_n(w) \xrightarrow{p} 0. \tag{A.46}$$

By (A.18), (A.19), (A.37), (A.41), (A.42), (A.45), and conditions (12) and (22),

$$\sup_{w\in\mathcal{H}_n}\left|e'\tilde{\mathbf{P}}(w)e\right|/\tilde{R}_n(w) \xrightarrow{p} 0. \tag{A.47}$$

By (A.18)–(A.27), (A.29)–(A.36), (A.41), (A.42), (A.45) and conditions (12) and (22),

$$\sup_{w\in\mathcal{H}_n}\left|\tilde{L}_n(w)/\tilde{R}_n(w)-1\right| \xrightarrow{p} 0. \tag{A.48}$$

Using all these formulas in (A.16), we obtain the desired (OPT) condition. □

Proof of Theorem 3.1. This proof is an application of Lemma 4 and Theorem 2 of Ing and Wei (2003), and the proof of our Theorem 2.2. To begin, let us substitute $V_n(w)$, $\tilde{V}_n(w)$, ξ_n^*, $\tilde{\xi}_n^*$, $\sigma^2 I_n$ and "in probability", for $R_n(w)$, $\tilde{R}_n(w)$, ξ_n, $\tilde{\xi}_n$, Ω, and a.s. respectively used in Theorem 2.2 and its proof. Consider also the arguments used in the proof of Theorem 2.2, and note that condition (12) is implied by condition (C.4) and condition (C.1) implies that conditions (22)–(24) hold in probability. Now, taking into account all of the above, to prove Theorem 3.1, we need to verify that

$$\xi_n^{*-1}\tilde{h}e'e = o_p(1), \tag{A.49}$$
$$\xi_n^{*-1}e'\mathbf{P}e = o_p(1), \tag{A.50}$$

and

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{A}}'(w)e\right|/\tilde{V}_n(w) = o_p(1). \tag{A.51}$$

Eqs. (A.49)–(A.51) correspond to results implied by (A.41)–(A.43) respectively. From the decomposition in (A.4) and $\mu'e$ being unrelated to w, we can prove

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{P}}'(w)e\right|/\tilde{V}_n(w) = o_p(1), \tag{A.52}$$

instead of proving (A.51). Due to (A.45), for proving (A.52), we only need to prove

$$\xi_n^{*-1}\sup_{w\in\mathcal{H}_n}\left|\mu'\tilde{\mathbf{P}}'(w)e\right| = o_p(1). \tag{A.53}$$

In light of $\{e_1,\ldots,e_n\}$ being i.i.d. and $Ee_i^4<\infty$, we have

$$e'e = O_p(n). \tag{A.54}$$

By (A.18), (A.54), and (C.1), we obtain (A.49).

From Lemma 4 of Ing and Wei (2003) and conditions (C.3)–(C.4), we have

$$n^{-1}E(e'Y_L{Y_L}'e) = O(r_1). \tag{A.55}$$

So, by Markov's inequality,

$$n^{-1}r_1^{-1}e'Y_L{Y_L}'e = O_p(1). \tag{A.56}$$

In addition, by condition (C.2), we have

$$n^{-1}e'X^*X^{*\prime}e = O_p(1), \tag{A.57}$$

which, together with (A.56), leads to

$$n^{-1}r^{-1}e'\mathbf{X}\mathbf{X}'e = O_p(1). \tag{A.58}$$

Using Theorem 2 of Ing and Wei (2003) and conditions (C.3)–(C.4), we obtain

$$nE(\mathcal{S}((Y_L'Y_L)^{-1})) = O(1). \tag{A.59}$$

Combining with Markov's inequality, this yields

$$n\mathcal{S}((Y_L'Y_L)^{-1}) = O_p(1). \tag{A.60}$$

Let $\mathbf{J} = (Y_L'Y_L)^{-1}Y_L'X^*(X^{*\prime}\mathbf{M}X^*)^{-1/2}$. From Rao (1973), we know that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix}(Y_L'Y_L)^{-1}+\mathbf{J}\mathbf{J}' & -\mathbf{J}(X^{*\prime}\mathbf{M}X^*)^{-1/2}\\ -(X^{*\prime}\mathbf{M}X^*)^{-1/2}\mathbf{J}' & (X^{*\prime}\mathbf{M}X^*)^{-1}\end{pmatrix}$$
$$= \begin{pmatrix}(Y_L'Y_L)^{-1} & 0\\ 0 & 0\end{pmatrix} + 2\begin{pmatrix}\mathbf{J}\mathbf{J}' & 0\\ 0 & (X^{*\prime}\mathbf{M}X^*)^{-1}\end{pmatrix}$$
$$- \begin{pmatrix}\mathbf{J}\mathbf{J}' & \mathbf{J}(X^{*\prime}\mathbf{M}X^*)^{-1/2}\\ (X^{*\prime}\mathbf{M}X^*)^{-1/2}\mathbf{J}' & (X^{*\prime}\mathbf{M}X^*)^{-1}\end{pmatrix}. \tag{A.61}$$

Consequently,

$$\mathcal{S}((\mathbf{X}'\mathbf{X})^{-1}) \le \mathcal{S}((Y_L'Y_L)^{-1}) + 2\max\left\{\mathcal{S}(\mathbf{J}\mathbf{J}'), \mathcal{S}((X^{*\prime}\mathbf{M}X^*)^{-1})\right\}$$
$$\le \mathcal{S}((Y_L'Y_L)^{-1}) + 2\max\left\{\mathcal{S}((Y_L'Y_L)^{-1})\mathcal{S}(n^{-1}X^{*\prime}X^*)\right.$$
$$\left.\times\,\mathcal{S}((n^{-1}X^{*\prime}\mathbf{M}X^*)^{-1}), n^{-1}\mathcal{S}((n^{-1}X^{*\prime}\mathbf{M}X^*)^{-1})\right\}. \tag{A.62}$$

From (A.60), (A.62) and condition (C.2), we have

$$n\mathcal{S}((\mathbf{X}'\mathbf{X})^{-1}) = O_p(1). \tag{A.63}$$

Now, by (A.58), (A.63), and (C.1), we obtain

$$\xi_n^{*-1}e'\mathbf{P}e = \xi_n^{*-1}e'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'e \le \xi_n^{*-1}\mathcal{S}((\mathbf{X}'\mathbf{X})^{-1})e'\mathbf{X}\mathbf{X}'e$$
$$= \xi_n^{*-1}rn\mathcal{S}((\mathbf{X}'\mathbf{X})^{-1})n^{-1}r^{-1}e'\mathbf{X}\mathbf{X}'e = o_p(1), \tag{A.64}$$

which is (A.50).

Also, from the argument of (A.64), we have

$$r^{-1}e'\mathbf{P}e = O_p(1). \tag{A.65}$$

To prove (A.53), we see that

$$\left|\mu'\tilde{\mathbf{P}}'(w)e\right| \le \left|\mu'\mathbf{P}(w)e\right| + \left|\mu'\mathbf{Q}(w)e\right| + \left|\mu'\mathbf{T}'(w)e\right|$$
$$= \left|\mu'\mathbf{P}\mathbf{P}(w)e\right| + \left|\mu'\mathbf{Q}(w)e\right| + \left|\mu'\mathbf{T}'(w)e\right|$$
$$\le \left(\mu'\mathbf{P}\mu e'\mathbf{P}(w)\mathbf{P}(w)e\right)^{1/2} + \left(\mu'\mu e'\mathbf{Q}(w)\mathbf{Q}(w)e\right)^{1/2}$$
$$+ \left(\mu'\mu e'\mathbf{T}(w)\mathbf{T}'(w)e\right)^{1/2}$$
$$\le \left(\mu'\mathbf{P}\mu e'\mathbf{P}(w)e\right)^{1/2} + \left(\mu'\mu\tilde{h}^2e'e\right)^{1/2} + \left(\mu'\mu\tilde{h}^2e'e\right)^{1/2}$$
$$\le \left(\mu'\mathbf{P}\mu e'\mathbf{P}e\right)^{1/2} + 2\left(\mu'\mu\tilde{h}^2e'e\right)^{1/2}$$

$$= \xi_n^* \left((r\xi_n^{*-2}\mu'\mathbf{P}\mu)(r^{-1}e'\mathbf{P}e) \right)^{1/2} + 2\xi_n^* \left((n^{-1}\mu'\mu)(nr^{-1}\tilde{h})(r\xi_n^{*-1})(\xi_n^{*-1}\tilde{h}e'e) \right)^{1/2}, \quad \text{(A.66)}$$

where the third "$\leq$" on the r.h.s. of the above is from (A.34)–(A.36). Combining (A.18), (A.49), (A.65), (A.66), and condition (C.1), we obtain (A.53). This completes the proof. □

References

Andrews, D.W.K., 1991. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. Journal of Econometrics 47, 359–377.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Bunea, F., Tsybakov, A.B., Wegkamp, M.H., 2007. Aggregation for Gaussian regression. Annals of Statistics 35, 1674–1697.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge, U.K.

Danilov, D., Magnus, J.R., 2004. On the harm that ignoring pretesting can cause. Journal of Econometrics 122, 27–46.

Goldenshluger, A., 2009. A universal procedure for aggregating estimators. Annals of Statistics 37, 542–568.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, B.E., 2008. Least squares forecast averaging. Journal of Econometrics 146, 342–350.

Hansen, B.E., Racine, J., 2012. Jackknife model averaging. Journal of Econometrics 167, 38–46.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899.

Hjort, N.L., Claeskens, G., 2006. Focussed information criteria and model averaging for Cox's hazard regression model. Journal of the American Statistical Association 101, 1449–1464.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Ing, C.K., Wei, C.Z., 2003. On same-realization prediction in an infinite order autoregressive process. Journal of Multivariate Analysis 85, 130–155.

Kapetanios, G., Labhard, V., Price, S., 2008a. Forecasting using Bayesian and information-theoretic model averaging. Journal of Business and Economic Statistics 26, 33–41.

Kapetanios, G., Labhard, V., Price, S., 2008b. Forecast combination and the bank of England's suite of statistical forecasting models. Economic Modelling 25, 772–792.

Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first stage prediction averaging. Econometrica 78, 697–718.

Leung, G., Barron, A.R., 2006. Information theory and mixing least-squares regressions. IEEE Transactions on Information Theory 52, 3396–3410.

Li, K.C., 1987. Asymptotic optimality for C_L, cross-validation and generalized cross-validations: discrete index set. Annals of Statistics 15, 958–975.

Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. On optimal weight choice in a frequentist model average estimator. Journal of the American Statistical Association 106, 1053–1066.

Liu, C.A., 2011. A plug-in averaging estimator for regressions with heteroskedastic errors. Working paper, Department of Economics, University of Wisconsin, Madison.

Newey, W., 1997. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics 79, 147–168.

Pesaran, M., Schleicher, C., Zaffaroni, P., 2009. Model averaging in risk management with an application to futures markets. Journal of Empirical Finance 16, 280–305.

Patton, A., Politis, D.N., White, H., 2009. Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. Econometric Reviews 28, 372–375.

Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econometric Reviews 23, 53–70.

Racine, J., 1997. Feasible cross-validatory model selection for general stationary processes. Journal of Applied Econometrics 12, 169–179.

Rao, C.R., 1973. Linear Statistical Inference and its Applications, second ed.. Wiley, New York, U.S.A.

Schomaker, M., Wan, A.T.K., Heumann, C., 2010. Frequentist model averaging with missing observations. Computational Statistics and Data Analysis 54, 3336–3347.

Shao, J., 1997. An asymptotic theory for linear model selection (with discussion). Statistica Sinica 7, 221–264.

Shibata, R., 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics 8, 147–164.

Shibata, R., 1981. An optimal selection of regression variables. Biometrika 68, 45–54.

Wan, A.T.K., Zhang, X., 2009. On the use of model averaging in tourism research. Annals of Tourism Research 36, 525–532.

Wan, A.T.K., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. Journal of Econometrics 156, 277–283.

Wang, H., Zhang, X., Zou, G., 2009. Frequentist model averaging estimation: a review. Journal of Systems Science and Complexity 22, 732–748.

Whittle, P., 1960. Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and its Applications 5, 302–305.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–586.

Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? Journal of the American Statistical Association 100, 1202–1214.

Zhang, P., 1992. Inference after variable selection in linear regression models. Biometrika 79, 741–746.

Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Annals of Statistics 39, 174–200.

Zhang, X., Wan, A.T.K., Zhou, S.Z., 2012. Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. Journal of Business and Economic Statistics 30, 132–142.