Journal of Econometrics 156 (2010) 277–283

Contents lists available at ScienceDirect

Journal of Econometrics

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

Least squares model averaging by Mallows criterion☆

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

[a] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 1 October 2008
Received in revised form
14 September 2009
Accepted 12 October 2009
Available online 1 November 2009

JEL classification:
C51
C52

Keywords:
Asymptotic optimality
Continuous weights
Mallows criterion
Non-nested models

ABSTRACT

This paper is in response to a recent paper by Hansen (2007) who proposed an optimal model average estimator with weights selected by minimizing a Mallows criterion. The main contribution of Hansen's paper is a demonstration that the Mallows criterion is asymptotically equivalent to the squared error, so the model average estimator that minimizes the Mallows criterion also minimizes the squared error in large samples. We are concerned with two assumptions that accompany Hansen's approach. The first is the assumption that the approximating models are strictly nested in a way that depends on the ordering of regressors. Often there is no clear basis for the ordering and the approach does not permit non-nested models which are more realistic from a practical viewpoint. Second, for the optimality result to hold the model weights are required to lie within a special discrete set. In fact, Hansen noted both difficulties and called for extensions of the proof techniques. We provide an alternative proof which shows that the result on the optimality of the Mallows criterion in fact holds for continuous model weights and under a non-nested set-up that allows *any* linear combination of regressors in the approximating models that make up the model average estimator. These results provide a stronger theoretical basis for the use of the Mallows criterion in model averaging by strengthening existing findings.

1. Introduction

Model selection has always been an integral part of econometric modeling. The goal of model selection is to choose a single model that is regarded as the best among all candidate models considered in the initial stage of analysis. Once a model is chosen, all subsequent estimation and inference take place within the chosen model as if this model was given a priori. In reality, the properties of estimators and tests subsequent to model selection depend on the way the model has been selected in addition to the stochastic nature of the chosen model. Practitioners engaged in applied work, however, usually only take into the account the latter and report estimates obtained from the chosen model as if they were unconditional estimates even though they are actually conditional estimates. So, the end results may not be what they appear to be — what we believe to be a 95% confidence interval may actually be a 75% confidence interval, a hypothesis tested at the nominal 5% level may in fact have been tested at a much higher level, etc. Problems associated with inference after model selection have been investigated in Leeb and Pötscher (2003, 2005, 2008). In addition, as is known by many researchers, another major drawback of model selection is that it often results in estimators that are unstable, as a small perturbation of data can result in a very different model being selected (Yang, 2001).

An alternative to model selection is model averaging — rather than attaching to a single "winning" model, a model average estimator compromises across a set of competing models, and in doing so, incorporates model uncertainty into the conclusions about the unknown parameters. Another important motivation behind model averaging is that it provides a kind of insurance against selecting a very poor model and thus holds promise for improving the risk in regression estimation (Leung and Barron, 2006). Model averaging techniques from a Bayesian perspective have proliferated in the theoretical literature since the early 1970s, but these techniques were not in widespread use until recent advances in computing power facilitated their practical implementation. For an overview of Bayesian model averaging and a survey of the relevant literature, see Koop (2003, Chapter 11). Studies from a frequentist viewpoint have been fewer, but some important progress has been made in recent years. Buckland et al. (1997), for example, suggested combining models with weights based on the AIC or BIC scores of the competing models. Yang (2001) proposed an adaptive regression using a mixing method, and Yuan and Yang (2005) suggested a model screening step in conjunction with this method. Some results on the large sample

☆ All three authors contributed equally to this work and the order of authorship has nothing other than alphabetical significance.

* Corresponding author. Tel.: +852 27887146; fax: +852 27888560.
E-mail addresses: msawan@cityu.edu.hk (A.T.K. Wan), xinyu@amss.ac.cn (X. Zhang), ghzou@amss.ac.cn (G. Zou).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.10.030

behavior of likelihood based model average estimators were found in Hjort and Claeskens (2003). More recently, Leung and Barron (2006) proposed a mixture least squares estimator with weights that depend on the estimator's risk characteristics; they further derived a finite sample risk bound for the mixture estimator and showed that the performance of this estimator is generally comparable to the least squares estimator of the best individual model.[1]

In a recent article, Hansen (2007) proposed a least squares model average estimator with model weights selected by minimizing a criterion in the spirit of Mallows' C_p (Mallows, 1973). Hansen observed that the Mallows criterion is asymptotically equivalent to the squared error. Thus, the model average estimator that minimizes the Mallows criterion also minimizes the squared error in large samples. This is the main finding of Hansen's paper and is summarized in Theorem 1 of Hansen (2007). To prove the theorem, Hansen (2007) used a result of Li (1987), who demonstrated the asymptotic optimality of the Mallows criterion for model selection. In a sense Hansen's (2007) contribution may be viewed as an extension of Li (1987) from model selection to model averaging. Several recent contributions have applied the Mallows model averaging (MMA) approach advocated in Hansen (2007) to constructing forecasting combinations in stationary models estimated by least squares (Hansen, 2008), models with a structural break (Hansen, in press-a), and autoregressive models with a near unit root (Hansen, in press-b).

Hansen's (2007) approach and the subsequent extensions listed above mark a significant step toward the development of optimal weight choice in the frequentist model average estimator. Nonetheless, there are challenges inherent in Hansen's selection of model weights even if one confines scrutiny to a homoscedastic model as Hansen (2007) did. Our concern is two-fold. First, Hansen's model set-up follows that of Li (1987) by imposing the assumption that the regressors can be ordered in such a way that the model average estimator is a weighted sum of least squares estimators obtained from regression models that are strictly nested, with the mth regression model using the first k_m variables in the ordered set as regressors such that $0 < k_1 < k_2 < \cdots < k_M$, where the Mth model is the largest model included in the model average. In our view this nested set-up could be an issue, depending on the application; it is not just a technical condition but a key assumption about the way model averaging is performed. Hansen (2007, 2008)) also noted this problem. Second, Hansen (2007) studied the asymptotic properties of the model average estimator by constraining the model weights w_m to the special discrete set $H_n(N) = \left\{\sum_{m=1}^{M} w_m = 1,\ w_m \in \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}\right\}$ for some fixed integer N. Hansen (2008) also made a similar assumption. Hansen (2007) admitted that there is not much support for restricting the weights to this discrete set, but for technical reasons which he explained in his paper this restriction is necessary in order for the asymptotic results to go through. This has again called into question the generality of the approach. Hansen (2007) has explicitly called for extensions of proof techniques to allow for continuous weights.

The purpose of this paper is to demonstrate that neither the nested model set-up nor the restriction of model weights to the set $H_n(N)$ is necessary for proving the MMA estimator's asymptotic optimality. It turns out that the asymptotic optimality of the MMA estimator, in the sense of achieving the smallest squared error, continues to hold under a general (non-nested) set-up that permits *any* linear combination of regressors in the models that make up the MMA estimator. In addition, our asymptotic theory does not restrict the allowable model weights to the discrete set $H_n(N)$. As we shall demonstrate, provided that a condition on the rate of convergence of the asymptotic risk is satisfied, the optimality of the Mallows criterion in fact holds also for the class of continuous model weights. Thus, the result obtained in this paper gives a stronger theoretical justification for the use of Hansen's MMA estimator in econometric applications. It is instructive to point out that Li's (1987) result on the asymptotic optimality of the Mallows criterion for model selection was in fact generalized by Shao (1997) for non-nested models. However, Shao's (1997) proof *cannot* be extended directly to the context of model averaging.

The remainder of this paper begins with a description of the model set-up in Section 2. Section 3 states the main results, with proofs of results contained in the Appendix.

2. Model set-up

Our description of the model set-up follows Hansen's (2007) notations. Wherever appropriate we will point out the differences in the two set-ups. Consider a random sample (y_i, x_i), $i = 1, \ldots, n$, where $x_i = (x_{i1}, x_{i2}, \ldots)$ is countably infinite and y_i is real-valued. Following Hansen (2007), we assume that the data-generating process is

$$y_i = \mu_i + e_i, \tag{1}$$

where $\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ij}$, $E(e_i|x_i) = 0$, and $E(e_i^2|x_i) = \sigma^2$. We consider a sequence of linear approximating models $m = 1, 2, \ldots, M$, where the mth model uses *any* k_m regressors belonging to x_i such that $k_m > 0$. So the mth approximating model is

$$y_i = \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} + e_i, \tag{2}$$

where $x_{i1(m)}, x_{i2(m)}, \ldots$ are variables in x_i that appear as regressors in the mth model, and the $\theta_{j(m)}$ are the corresponding coefficients. The approximation error is $b_{i(m)} = \mu_i - \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)}$. The total number of models M can be finite or infinite. The class of models covered here is substantially larger than that under the set-up of Li (1987) and Hansen (2007), where it is assumed that the regressors are ordered at the outset with the mth approximating model using the first k_m regressors in the ordered set such that $0 < k_1 < k_2 < \cdots < k_M$. That is, a model is always nested within the bigger models in the sequence. This could be an issue, depending on the application. Our set-up, on the other hand, is general, and in particular it allows the approximating models to be non-nested. It also avoids the problem of having to order the regressors at the outset, which, as Hansen (2007) commented, may be troubling in some circumstances.

In matrix notation, (1) and (2) may be written as $Y = \mu + e = X_{(m)}\Theta_{(m)} + b_{(m)} + e$ and $Y = X_{(m)}\Theta_{(m)} + e$, respectively, where $Y = (y_1, \ldots, y_n)'$, $\mu = (\mu_1, \ldots, \mu_n)'$, $X_{(m)}$ is an $n \times k_m$ matrix with ijth element $x_{ij(m)}$, $\Theta_{(m)} = (\theta_{1(m)}, \ldots, \theta_{k_m(m)})'$, $b_{(m)} = (b_{1(m)}, \ldots, b_{n(m)})'$ and $e = (e_1, \ldots, e_n)'$. Let $\hat{\Theta}_{(m)} = (X'_{(m)}X_{(m)})^{-} X'_{(m)}Y$ be the least squares estimator of $\Theta_{(m)}$ in the mth approximating model. The corresponding estimator of μ is $\hat{\mu}_{(m)} = X_{(m)}\hat{\Theta}_{(m)} = X_{(m)}(X'_{(m)}X_{(m)})^{-}X'_{(m)}Y \equiv P_{(m)}Y$, and $\hat{e}_{(m)} = Y - \hat{\mu}_{(m)}$ is the vector of residuals.

[1] Model average estimators are not without problems, however. For example, Kabaila (2002) demonstrated that post-model-selection estimators with asymptotically efficient properties can have rather inefficient small sample performance. As remarked on by one referee, Kabaila's (2002) findings are likely to carry over to model average estimators. Hence the results of this and other related papers on asymptotic loss efficiency of model average estimators may have limited small sample impact and care must be exercised in interpreting the results reported. Additionally, Pötscher (2006) showed that the difficulties known for the estimation of small sample distributions of post-model-selection estimators (Leeb and Pötscher, 2003) also occur for model average estimators.

Let $w = (w_1, \ldots, w_M)'$ be a weight vector in the unit simplex of R^M : $H_n = \left\{ w \in [0,1]^M : \sum_{m=1}^M w_m = 1 \right\}$. The model average estimator of μ is

$$\hat{\mu}(w) = \sum_{m=1}^{M} w_m P_{(m)} Y \equiv P(w)Y. \tag{3}$$

Let the average squared error loss and the corresponding risk be $L_n(w) = (\hat{\mu}(w) - \mu)'(\hat{\mu}(w) - \mu)$ and $R_n(w) = E\{L_n(w)|x_1, \ldots, x_n\}$, respectively. It can be shown that $R_n(w) = \|A(w)\mu\|^2 + \sigma^2 trP^2(w)$, where $A(w) = I - P(w)$. Hence we have

$$R_n(w) \geq \|A(w)\mu\|^2, \tag{4}$$

and

$$R_n(w) \geq \sigma^2 trP^2(w). \tag{5}$$

Note that under the extended set-up, Lemma 1(iii) of Hansen (2007) concerning the property of $P(w)$ continues to hold.

The Mallows criterion for model averaging is

$$C_n(w) = \left(Y - \hat{\mu}(w)\right)' \left(Y - \hat{\mu}(w)\right) + 2\sigma^2 trP(w). \tag{6}$$

Following Hansen (2007), let $\hat{w} = \arg\min_{w \in H_n}(C_n(w))$ be the weight vector based on the Mallows criterion, and the MMA estimator is (3) using $\hat{w}$ as the weight vector. Note that $EC_n(w) = EL_n(w) + n\sigma^2$, i.e., $C_n(w)$ is an unbiased estimator of the expected in-sample squared error plus a constant.

3. Optimality of the Mallows criterion under the non-nested set-up

This section presents the main result of this paper, which demonstrates the asymptotic optimality of the MMA estimator under the non-nested set-up described in Section 2. Specifically, the optimality result holds for all weights in the set H_n, which includes the discrete set $H_n(N)$ considered by Hansen (2007) as a special case. The following theorem, which we refer to as Theorem 1′, is a counterpart to Theorem 1 of Hansen (2007):

Theorem 1′. *As $n \to \infty$, if, for some fixed integer $1 \leq G < \infty$,*

$$E(e_i^{4G}|x_i) \leq \kappa < \infty, \tag{7}$$

and

$$M\xi_n^{-2G} \sum_{m=1}^{M} \left(R_n(w_m^0)\right)^G \to 0, \tag{8}$$

then

$$\frac{L_n(\hat{w})}{\inf_{w \in H_n} L_n(w)} \xrightarrow{p} 1, \tag{9}$$

where $\xi_n = \inf_{w \in H_n} R_n(w)$ and w_m^0 is an $M \times 1$ vector in which the mth element is one and the others are zeros.

Proof. See the Appendix. □

That is, the Mallows weight vector yields a squared error that is asymptotically identical to that of the infeasible optimal weight vector. This implies that the MMA estimator is asymptotically optimal in the class of model average estimators (3) with weight vector belonging to the set H_n which includes all weights.

Theorem 1′ is actually close to Hansen's (2007) Theorem 1, but it has been developed under the non-nested set-up and does not restrict the weight set H_n to any subset. Theorem 1′ depends on assumption (7), which places a bound on the conditional moments. This bound condition may be contrasted with the corresponding condition

$$E(|e_i|^{4(N+1)}|x_i) \leq \kappa < \infty, \tag{10}$$

required for Hansen's (2007) Theorem 1. But note that (10) depends on model weights since Hansen (2007) restricted the weight set to $H_n(N)$. Hansen remarked that the restriction of H_n to $H_n(N)$ can be made less binding by picking a large N, provided that the aforementioned conditional moment bound is satisfied. Clearly, the larger the value of N, the more stringent the bound condition (10). On the other hand, the bound condition (7) required for Theorem 1′ has the merit of being independent of model weights even though Theorem 1′ is based on a more general set-up.

The convergence condition given in (8), which is similar to a convergence condition in Li (1987), also deserves some attention. First, $\xi_n \to \infty$ is obviously a necessary condition for (8) to hold. Hansen's (2007) Theorem 1 also requires $\xi_n \to \infty$, and as Hansen (2007) remarked, this condition simply means there is no finite approximating model for which the bias is zero, which is conventional for non-parametric regression. Second, in addition to $\xi_n \to \infty$, condition (8) also requires $M \sum_{m=1}^M \left(R_n(w_m^0)\right)^G \to \infty$ at a rate slower than $\xi_n^{2G} \to \infty$ as $n \to \infty$. Thus, condition (8) that determines Theorem 1′ is stronger than the corresponding condition required for Theorem 1 of Hansen (2007). If we write $\eta_n = \max_{w \in H_n(0)} R_n(w)$, where $H_n(0)$ is the set comprising the vectors w_m^0, $m = 1, \ldots, M$, then condition (8) may be justified by noting that $M^2(\eta_n \xi_n^{-2})^G \to 0$ is a sufficient condition for (8). Given the rate of $\xi_n \to \infty$, the slower the rates of $M \to \infty$ and $\eta_n \to \infty$, the faster the rate of $M^2(\eta_n \xi_n^{-2})^G \to 0$. In practice, the rates of $M \to \infty$ and $\eta_n \to \infty$ can be reduced by removing the very poor models at the outset prior to model combining.

In what follows, we provide two explicit examples under which condition (8) holds. Proofs of results relating to these examples are given in the Appendix.

Example 1. Consider the data-generating process

$$Y(z_i) = \mu_i + \varepsilon_i = f(z_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $f(z_i) = \sum_{s=1}^{\infty} \theta_s \cos\left((s-1)z_i\right)/s$, $z_i = 2\pi(i-1)/n$ and the ε_i are i.i.d. $N(0, \sigma^2)$. This model was also studied by Shibata (1981) and Kabaila (2002). Suppose that $\theta_s = s^{-7/12}$, and that we smooth across estimators from models that comprise subsets of the first s_n ($s_n < S$) regressors, where S is the largest integer that is no greater than $n/2$. Altogether there are $M = 2^{s_n - 1}$ approximating models if we assume all models contain the intercept. We show in the Appendix that condition (8) holds provided that $s_n = O(\log n)$. It is also of interest to note that when $s_n = O(n^{1/3})$, Hansen's condition (i.e., $\xi_n \to \infty$ as $n \to \infty$) holds but condition (8) does not. This shows that condition (8) is indeed a stronger condition.

Example 2. This example considers a simple nested setting where the mth approximating model contains the first m regressors. This is also a natural example where condition (8) is satisfied[2] even though the nested setting is more restrictive than the general setting discussed in this paper, and we include this example here partly as a response to the referee's suggestion. Now, suppose that $R_n(w_m^0)$ is of order $nm^{-a} + m$, where $a > 0$. Assume that $M = O(n^v)$, where v is a constant and G_0 ($> 1/a$) is an integer that satisfies condition (7). We show in the Appendix that a sufficient condition for condition (8) to hold is $v < \min\{G_0/(1 + 2aG_0), 1/(1 + a)\}$. Applying this result to the regression function of Example 1 and assuming that the models are strictly nested, we observe that $R_n(w_m^0)$ has order $nm^{-13/6} + m$, and thus, condition (8) holds when $v < 3/13$.

[2] This setting is very similar to the series estimation setting considered by Newey (1997), and the order of $R_n(w_m^0)$ assumed here is precisely the same as that of the risk of the series estimator given in Theorem 1 of Newey (1997).

Now, let $\bar{k}_m = tr\left(P_{(m)}\right)$. Clearly, when $X_{(m)}$ has full column rank, $\bar{k}_m = k_m$. Further, let the M^*th approximating model be the model such that $\bar{k}_{M^*} = \max\{\bar{k}_1, \ldots, \bar{k}_M\}$, and $\hat{\sigma}^2_{M^*} = (n - \bar{k}_{M^*})^{-1}(Y - \hat{\mu}_{(M^*)})'(Y - \hat{\mu}_{(M^*)})$ be the least squares estimator of σ^2 based on the M^* th approximating model. The following theorem shows that provided that some mild conditions are satisfied, Theorem 1′ remains valid if σ^2 is unknown and replaced by $\hat{\sigma}^2_{M^*}$.

Theorem 2. *When σ^2 is unknown and replaced by $\hat{\sigma}^2_{M^*}$, Theorem 1′ remains valid if*

$$\mu'\mu/n = O(1), \tag{11}$$

and

$$\bar{k}^2_{M^*}/n \le \varphi < \infty, \tag{12}$$

as $n \to \infty$, where φ is an arbitrary constant.

Proof. See the Appendix. □

Condition (11), which concerns the average of μ_i^2, holds for Example 1 considered above and is quite common and reasonable. A similar assumption can be found in Shao (1997, p. 224). When all the $X_{(m)}$ are of full column rank, condition (12) places a constraint on the number of regressors in the largest approximating model. Shibata (1980) and Newey (1997) also made similar assumptions[3] in the contexts of their investigations. Additionally, Theorem 2 is also valid if $\hat{\sigma}^2_{M^*}$ is replaced by another least squares estimator of σ^2 obtained from another approximating model. However, using $\hat{\sigma}^2_{M^*}$ has an advantage because the convergence rates of (A.19) and (A.20) given in the Appendix depend in turn on the rate of $\mu'(I - P_{(M^*)})\mu/\xi_n^2 \to 0$. This latter rate is generally faster than the corresponding rate of $\mu'(I - P_{(m)})\mu/\xi_n^2 \to 0$ when $\hat{\sigma}^2_{M^*}$ is replaced by $\hat{\sigma}^2_m$ such that $m \ne M^*$. Also, as noted by one referee, provided that $\hat{\sigma}^2$ converges to σ^2 sufficiently quickly, Theorem 1′ is also valid if σ^2 is replaced by $\hat{\sigma}^2$. For instance, if $\bar{k}_{M^*}(\hat{\sigma}^2_{M^*} - \sigma^2) = O_p(1)$ as $n \to \infty$, then Theorem 1′ still holds even if σ^2 is replaced by $\hat{\sigma}^2_{M^*}$, and in this case, conditions (11) and (12) would not be required.

A final note concerns a special case of these results. If H_n is restricted to $H_n(0)$, then the problem reduces from model averaging to model selection. In this special case, condition (8) is replaced by the condition $\sum_{m=1}^M \left(R_n(w_m^0)\right)^{-G} \to 0$, and by using a different proof technique one can show that Theorem 1′ is still valid. Thus, under the non-nested set-up, the Mallows criterion is optimal as a model selection criterion. As mentioned before, this result was noted earlier by Shao (1997). Our proof for this result turns out to be the same as Shao's (1997) but it is entirely different from the proof for the general model averaging result discussed in Theorem 1′.

Acknowledgements

Wan's work was supported by a competitive earmarked research grant from the Hong Kong Research Grant Council (Grant No. CityU-102709). Zou's work was supported by the National Natural Science Foundation of China (Grant Nos. 70625004, 70221001 and 10721101). The authors thank the referees, Emmanuel Guerre, Bruce Hansen, Jan Magnus, Benedikt Pötscher and Yuhong Yang for their constructive comments and suggestions. In particular, they are indebted to the suggestion given by Bruce Hansen for improving a convergence condition in an earlier version of the paper.

[3] Assumption 3 of Shibata (1980) is considerably stronger than condition (12) used here. Theorem 8 of Newey (1997) also assumes a condition similar to condition (12).

Appendix

Proof of Theorem 1′. The proof of Theorem 1′ is an application of Whittle (1960) and Chebyshev's inequality. Note that

$$C_n(w) = L_n(w) + \|e\|^2 + 2\langle e, A(w)\mu\rangle + 2(\sigma^2 trP(w) - \langle e, P(w)e\rangle).$$

Theorem 1′ is valid if the following hold:

As $n \to \infty$,

$$\sup_{w\in H_n} |\langle e, A(w)\mu\rangle| / R_n(w) \xrightarrow{p} 0, \tag{A.1}$$

$$\sup_{w\in H_n} \left|\sigma^2 trP(w) - \langle e, P(w)e\rangle\right| / R_n(w) \xrightarrow{p} 0, \tag{A.2}$$

and

$$\sup_{w\in H_n} |L_n(w)/R_n(w) - 1| \xrightarrow{p} 0. \tag{A.3}$$

Using the triangle inequality, Bonferroni's inequality, Chebyshev's inequality, Theorem 2 of Whittle (1960) and Eq. (4), we observe, for any $\delta > 0$, that

$$\begin{aligned}
&P\left\{\sup_{w\in H_n} |\langle e, A(w)\mu\rangle| / R_n(w) > \delta\right\} \\
&\le P\left\{\sup_{w\in H_n} \sum_{m=1}^M w_m \left|e'\left(I - P_{(m)}\right)\mu\right| > \delta\xi_n\right\} \\
&= P\left\{\max_{1\le m\le M} \left|e'\left(I - P_{(m)}\right)\mu\right| > \delta\xi_n\right\} \\
&= P\left\{\left\{\left|\langle e, A(w_1^0)\mu\rangle\right| > \delta\xi_n\right\} \cup \left\{\left|\langle e, A(w_2^0)\mu\rangle\right| > \delta\xi_n\right\}\right. \\
&\quad \left. \cup \cdots \cup \left\{\left|\langle e, A(w_M^0)\mu\rangle\right| > \delta\xi_n\right\}\right\} \\
&\le \sum_{m=1}^M P\left\{\left|\langle e, A(w_m^0)\mu\rangle\right| > \delta\xi_n\right\} \\
&\le \sum_{m=1}^M E\left\{\frac{\langle e, A(w_m^0)\mu\rangle^{2G}}{\delta^{2G}\xi_n^{2G}}\right\} \le C_1\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^M \left\|A(w_m^0)\mu\right\|^{2G} \\
&\le C_1\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^M \left(R_n(w_m^0)\right)^G,
\end{aligned}$$

where C_1 is a constant. Then from condition (8), we obtain (A.1). Similarly,

$$\begin{aligned}
&P\{\sup_{w\in H_n} \left|\sigma^2 trP(w) - \langle e, P(w)e\rangle\right| / R_n(w) > \delta\} \\
&\le \sum_{m=1}^M P\left\{\left|\sigma^2 trP(w_m^0) - \langle e, P(w_m^0)e\rangle\right| > \delta\xi_n\right\} \\
&\le \sum_{m=1}^M E\left\{\frac{\left[\sigma^2 trP(w_m^0) - \langle e, P(w_m^0)e\rangle\right]^{2G}}{\delta^{2G}\xi_n^{2G}}\right\} \\
&\le C_2\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^M \left[trP^2(w_m^0)\right]^G \\
&\le C_2'\delta^{-2G}\xi_n^{-2G}\sum_{m=1}^M \left(R_n(w_m^0)\right)^G,
\end{aligned}$$

where C_2 and C_2' are constants. Thus, (A.2) is obtained from condition (8). In addition, it is readily seen that

$$\sup_{w\in H_n}|L_n(w)/R_n(w)-1|\xrightarrow{p}0$$
$$\Leftrightarrow \sup_{w\in H_n}\left|\frac{\|P(w)e\|^2-\sigma^2 trP^2(w)-2\langle A(w)\mu,P(w)e\rangle}{R_n(w)}\right|\xrightarrow{p}0.$$

To prove (A.3), it suffices to show, as $n\to\infty$, that

$$\sup_{w\in H_n}\left|\frac{\langle A(w)\mu,P(w)e\rangle}{R_n(w)}\right|\xrightarrow{p}0, \tag{A.4}$$

and

$$\sup_{w\in H_n}\left|\frac{\|P(w)e\|^2-\sigma^2 trP^2(w)}{R_n(w)}\right|\xrightarrow{p}0. \tag{A.5}$$

First, note that for any w and w^* belonging to H_n,

$$\left\|P(w^*)A(w)\mu\right\|^2\le\lambda^2_{\max}(P(w^*))\,\|A(w)\mu\|^2\le\|A(w)\mu\|^2,$$

where $\lambda_{\max}(P(w^*))$ is the largest eigenvalue of $P(w^*)$. We then obtain

$$\begin{aligned}
&P\left\{\sup_{w\in H_n}|\langle P(w)e,A(w)\mu\rangle|/R_n(w)>\delta\right\}\\
&\le P\left\{\sup_{w\in H_n}\left|e'\sum_{m=1}^M w_mP_{(m)}\sum_{m=1}^M w_m(I-P_{(m)})\mu\right|>\delta\xi_n\right\}\\
&\le P\left\{\sup_{w\in H_n}\sum_{t=1}^M\sum_{m=1}^M w_tw_m\left|e'P_{(t)}(I-P_{(m)})\mu\right|>\delta\xi_n\right\}\\
&\le P\left\{\max_{1\le t\le M}\max_{1\le m\le M}\left|e'P_{(t)}(I-P_{(m)})\mu\right|>\delta\xi_n\right\}\\
&= P\left\{\left\{\left|\left\langle P(w_1^0)e,A(w_1^0)\mu\right\rangle\right|>\delta\xi_n\right\}\right.\\
&\quad\cup\left\{\left|\left\langle P(w_1^0)e,A(w_2^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\quad\cup\cdots\cup\left\{\left|\left\langle P(w_1^0)e,A(w_M^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\quad\cup\left\{\left|\left\langle P(w_2^0)e,A(w_1^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\quad\cup\left\{\left|\left\langle P(w_2^0)e,A(w_2^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\quad\cup\cdots\cup\left\{\left|\left\langle P(w_2^0)e,A(w_M^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\quad\left.\cup\cdots\cup\left\{\left|\left\langle P(w_M^0)e,A(w_M^0)\mu\right\rangle\right|>\delta\xi_n\right\}\right\}\\
&\le\sum_{t=1}^M\sum_{m=1}^M P\left\{\left|\left\langle P(w_t^0)e,A(w_m^0)\mu\right\rangle\right|>\delta\xi_n\right\}\\
&\le\sum_{t=1}^M\sum_{m=1}^M E\left[\frac{\left\langle P(w_t^0)e,A(w_m^0)\mu\right\rangle^{2G}}{\delta^{2G}\xi_n^{2G}}\right]\\
&\le C_3\delta^{-2G}\xi_n^{-2G}\sum_{t=1}^M\sum_{m=1}^M\left\|P(w_t^0)A(w_m^0)\mu\right\|^{2G}\\
&\le C_3\delta^{-2G}\xi_n^{-2G}\sum_{t=1}^M\sum_{m=1}^M\left\|A(w_m^0)\mu\right\|^{2G}\\
&\le C_3\delta^{-2G}\xi_n^{-2G}M\sum_{m=1}^M\left(R_n(w_m^0)\right)^G,
\end{aligned}$$

where C_3 is a constant. Eq. (A.4) is thus obtained from condition (8). Likewise, recognizing that

$$tr\left[P^2(w^*)P^2(w)\right]\le\lambda^2_{\max}(P(w^*))trP^2(w)\le trP^2(w),$$

we obtain

$$\begin{aligned}
&P\left\{\sup_{w\in H_n}\left|\|P(w)e\|^2-\sigma^2trP^2(w)\right|/R_n(w)>\delta\right\}\\
&\le P\left\{\sup_{w\in H_n}\sum_{t=1}^M\sum_{m=1}^M w_tw_m\left|e'P_{(t)}P_{(m)}e-\sigma^2trP_{(t)}P_{(m)}\right|>\delta\xi_n\right\}\\
&\le P\{\max_{1\le t\le M}\max_{1\le m\le M}\left|e'P_{(t)}P_{(m)}e-\sigma^2trP_{(t)}P_{(m)}\right|>\delta\xi_n\}\\
&= P\left\{\left\{\left|\left\langle e,P(w_1^0)P(w_1^0)e\right\rangle-\sigma^2trP(w_1^0)P(w_1^0)\right|>\delta\xi_n\right\}\right.\\
&\quad\cup\left\{\left|\left\langle e,P(w_1^0)P(w_2^0)e\right\rangle-\sigma^2trP(w_1^0)P(w_2^0)\right|>\delta\xi_n\right\}\\
&\quad\cup\cdots\cup\left\{\left|\left\langle e,P(w_1^0)P(w_M^0)e\right\rangle-\sigma^2trP(w_1^0)P(w_M^0)\right|>\delta\xi_n\right\}\\
&\quad\cup\left\{\left|\left\langle e,P(w_2^0)P(w_1^0)e\right\rangle-\sigma^2trP(w_2^0)P(w_1^0)\right|>\delta\xi_n\right\}\\
&\quad\cup\left\{\left|\left\langle e,P(w_2^0)P(w_2^0)e\right\rangle-\sigma^2trP(w_2^0)P(w_2^0)\right|>\delta\xi_n\right\}\\
&\quad\cup\cdots\cup\left\{\left|\left\langle e,P(w_2^0)P(w_M^0)e\right\rangle-\sigma^2trP(w_2^0)P(w_M^0)\right|>\delta\xi_n\right\}\\
&\quad\left.\cup\cdots\cup\left\{\left|\left\langle e,P(w_M^0)P(w_M^0)e\right\rangle-\sigma^2trP(w_M^0)P(w_M^0)\right|>\delta\xi_n\right\}\right\}\\
&\le\sum_{t=1}^M\sum_{m=1}^M P\{\left|\left\langle e,P(w_t^0)P(w_m^0)e\right\rangle-\sigma^2trP(w_t^0)P(w_m^0)\right|>\delta\xi_n\}\\
&\le\sum_{t=1}^M\sum_{m=1}^M E\left[\frac{(\left\langle e,P(w_t^0)P(w_m^0)e\right\rangle-\sigma^2trP(w_t^0)P(w_m^0))^{2G}}{\delta^{2G}\xi_n^{2G}}\right]\\
&\le C_4\delta^{-2G}\xi_n^{-2G}\sum_{t=1}^M\sum_{m=1}^M(tr\{P^2(w_t^0)P^2(w_m^0)\})^G\\
&\le C_4\delta^{-2G}\xi_n^{-2G}\sum_{t=1}^M\sum_{m=1}^M(trP^2(w_m^0))^G\\
&\le C_4'\delta^{-2G}\xi_n^{-2G}M\sum_{m=1}^M\left(R_n(w_m^0)\right)^G,
\end{aligned}$$

where C_4 and C_4' are constants. Thus, by condition (8), (A.5) is also obtained. This completes the proof of Theorem 1′. □

Proof of results relating to Example 1. Here, we show that for the model given in Example 1, condition (8) is satisfied if $s_n=O(\log n)$. Let $\mu=(\mu_1,\mu_2,\ldots,\mu_n)'$, $\theta=(\theta_1,\theta_2,\ldots)'$, $X_n=(c_1/1,c_2/2,\ldots)$ and $c_s=(\cos((s-1)z_1),\ldots,\cos((s-1)z_n))'$. From Kabaila (2002, Appendix B), we have

$$X_n\theta=\begin{cases}c_1\Pi(1)+\sum_{s=2}^S c_s\left(\frac{\theta_s}{s}+\Xi(s)\right)+c_{S+1}\Pi(S+1),\\ \quad\text{for even } n,\\ c_1\Pi(1)+\sum_{s=2}^{S+1}c_s\left(\frac{\theta_s}{s}+\Xi(s)\right),\quad\text{for odd } n,\end{cases}$$

where $\Pi(s)=\sum_{i=0}^\infty\frac{\theta_{in+s}}{in+s}$ and $\Xi(s)=\sum_{i=1}^\infty\left(\frac{\theta_{in+s}}{in+s}+\frac{\theta_{in+2-s}}{in+2-s}\right)$. Let $X_n(j)$ be the matrix comprising the first j columns of X_n. Using results from Kabaila (2002, Appendix A), it is straightforward to show that

$$\mu'\mu=\begin{cases}n\Pi^2(1)+\frac{n}{2}\sum_{s=2}^S\left(\frac{\theta_s}{s}+\Xi(s)\right)^2+n\Pi^2(S+1),\\ \quad\text{for even } n,\\ n\Pi^2(1)+\frac{n}{2}\sum_{s=2}^{S+1}\left(\frac{\theta_s}{s}+\Xi(s)\right)^2,\ \text{for odd } n,\end{cases} \tag{A.6}$$

and

$$\begin{aligned}
&\theta'X_n'X_n(j)\left(X_n'(j)X_n(j)\right)^{-1}X_n'(j)X_n\theta=n\Pi^2(1)\\
&\quad+\frac{n}{2}\sum_{s=2}^j\left(\frac{\theta_s}{s}+\Xi(s)\right)^2
\end{aligned} \tag{A.7}$$

for $1 \le j < S$. It is readily seen that, for $s \in [2, S+1]$,

$$\begin{aligned}\Xi(s) &= \sum_{i=1}^{\infty}\left(\frac{1}{(in+s)^{19/12}}+\frac{1}{(in+2-s)^{19/12}}\right)\\ &< n^{-19/12}\sum_{i=1}^{\infty}\left(\frac{1}{i^{19/12}}+\frac{1}{(i/2)^{19/12}}\right)\\ &= o(1/n),\end{aligned} \tag{A.8}$$

and

$$\begin{aligned}\Pi(S+1) &= n^{-19/12}\frac{1}{(S/n+1/n)^{19/12}}\\ &\quad + n^{-19/12}\sum_{i=1}^{\infty}\frac{1}{(i+S/n+1/n)^{19/12}}\\ &< n^{-19/12}2^{19/12}+n^{-19/12}\sum_{i=1}^{\infty}\frac{1}{i^{19/12}} = o(1/n).\end{aligned} \tag{A.9}$$

Now, denote $P^{(s_n)} = X_n(s_n)\left(X_n'(s_n)X_n(s_n)\right)^{-1}X_n'(s_n)$. Then from (A.6)–(A.9), we have

$$\begin{aligned}&\mu'(I-P^{(s_n)})\mu\\ &= \begin{cases}\dfrac{n}{2}\displaystyle\sum_{s=s_n+1}^{S}\left(\frac{\theta_s}{s}+\Xi(s)\right)^2+n\Pi^2(S+1), & \text{for even } n,\\ \dfrac{n}{2}\displaystyle\sum_{s=s_n+1}^{S+1}\left(\frac{\theta_s}{s}+\Xi(s)\right)^2, & \text{for odd } n,\end{cases}\\ &= \begin{cases}\dfrac{n}{2}\displaystyle\sum_{s=s_n+1}^{S}\left(\frac{1}{s^{19/12}}+o\left(\frac{1}{n}\right)\right)^2+o\left(\frac{1}{n}\right), & \text{for even } n,\\ \dfrac{n}{2}\displaystyle\sum_{s=s_n+1}^{S+1}\left(\frac{1}{s^{19/12}}+o\left(\frac{1}{n}\right)\right)^2, & \text{for odd } n,\end{cases}\\ &= \frac{n}{2}\sum_{s=s_n+1}^{S}\frac{1}{s^{19/6}}+o(1) = O(ns_n^{-13/6})+o(1).\end{aligned} \tag{A.10}$$

Recognizing that the vectors in $X(s_n)$ are orthogonal (Kabaila, 2002, Appendix A), we have, for any (m, t) pair of approximating models, that $P^{(s_n)} - P_{(m)} - P_{(t)} - P_{(m)}P_{(t)}$ is symmetric and idempotent. Thus,

$$\begin{aligned}&\mu'(I-P_{(m)})(I-P_{(t)})\mu - \mu'(I-P^{(s_n)})\mu\\ &\quad = \mu'\left(P^{(s_n)}-P_{(m)}-P_{(t)}+P_{(m)}P_{(t)}\right)\mu \ge 0,\end{aligned}$$

which, when combined with (A.10), leads to

$$\begin{aligned}\xi_n &\ge \inf_{w\in H_n}\|A(w)\mu\|^2\\ &= \inf_{w\in H_n}\sum_{m=1}^{M}\sum_{t=1}^{M}w_m w_t\mu'(I-P_{(m)})(I-P_{(t)})\mu\\ &\ge \mu'(I-P^{(s_n)})\mu = O(ns_n^{-13/6})+o(1).\end{aligned} \tag{A.11}$$

Let the first approximating model include no regressor other than the intercept. Similar to the derivation of (A.10), and recognizing that $\mu'(I-P_{(m)})\mu \le \mu'(I-P_{(1)})\mu$ for $m > 1$, we have

$$\begin{aligned}\eta_n \le \mu'(I-P_{(1)})\mu+\sigma^2 s_n &= \frac{n}{2}\sum_{s=2}^{S}\frac{1}{s^{19/6}}+o(1)+\sigma^2 s_n\\ &= O(n)+O(s_n).\end{aligned} \tag{A.12}$$

Combining (A.11) and (A.12), it is readily seen that

$$M\xi_n^{-2G}\sum_{m=1}^{M}\left(R_n(w_m^0)\right)^G \le \frac{M^2\left(\mu'(I-P_{(1)})\mu+\sigma^2 s_n\right)^G}{\left(\mu'(I-P^{(s_n)})\mu\right)^{2G}}. \tag{A.13}$$

Now, when $s_n = O(\log n)$, the right-hand side of (A.13) tends to zero for a sufficiently large G. Since ε_i is normally distributed, condition (7) is satisfied for any value of G. Therefore, condition (8) holds when $s_n = O(\log n)$. Note also that, if we let $s_n = O(n^{1/3})$, then for any fixed G,

$$\begin{aligned}M\xi_n^{-2G}\sum_{m=1}^{M}\left(R_n(w_m^0)\right)^G &\ge \frac{M\left(R_n(w_1^0)\right)^G}{\left(R_n(w_1^0)\right)^{2G}}\\ &= \frac{2^{s_n-1}}{(\mu'(I-P_{(1)})\mu+\sigma^2)^G}\to\infty,\end{aligned}$$

which implies that condition (8) is not satisfied. However, the same assumption would imply, from (A.11), that $\xi_n \to \infty$. In other words, if $s_n = O(n^{1/3})$, Hansen's condition holds but our condition (8) does not.

Proof of results relating to Example 2. Here, we show that, under the model setting described in Example 2, a sufficient condition for condition (8) to hold is $v < \min\{G_0/(1+2aG_0), 1/(1+a)\}$. Note that as $R_n(w_m^0)$ is of order $nm^{-a}+m$, we have $\mu'(I-P_{(m)})\mu = O(nm^{-a}+m)$, which, when combined with $P_{(m)}P_{(t)} = P_{(t)}$, $\mu'(I-P_{(m)})\mu \le \mu'(I-P_{(t)})\mu$ for $m > t$, and the condition $v < 1/(1+a)$, yields

$$\begin{aligned}\xi_n &\ge \inf_{w\in H_n}\|A(w)\mu\|^2\\ &= \inf_{w\in H_n}\sum_{m=1}^{M}\sum_{t=1}^{M}w_m w_t\mu'(I-P_{(m)})(I-P_{(t)})\mu\\ &\ge \mu'(I-P_{(M)})\mu = O(n^{1-av}).\end{aligned} \tag{A.14}$$

Also, from the assumptions that $R_n(w_m^0)$ is of order $nm^{-a}+m$ and $G_0 > 1/a$, together with the condition $v < 1/(1+a)$, we have

$$M\sum_{m=1}^{M}\left(R_n(w_m^0)\right)^{G_0} = O(n^{v+G_0})+O(n^{2v+vG_0}) = O(n^{v+G_0}). \tag{A.15}$$

Combining (A.14) and (A.15), it is readily seen that, if $v < \min\{G_0/(1+2aG_0), 1/(1+a)\}$, then condition (8) holds with $G = G_0$.

As a special case, we consider the regression function given in Example 1. Similar to the derivation of (A.10), we have

$$R_n(w_m^0) = \frac{n}{2}\sum_{s=m+1}^{S}\frac{1}{s^{19/6}}+\sigma^2 m+o(1) = O(nm^{-13/6}+m),$$

which satisfies the assumption in the current example with $a = 13/6$. A sufficient condition for (8) to hold is $v < 3/13$ as the noise component is normal.

Proof of Theorem 2. If σ^2 is estimated by $\hat\sigma^2_{M^*}$, the Mallows criterion may be written as $\hat C_n(w) \equiv (Y-\hat\mu(w))'(Y-\hat\mu(w)) + 2\hat\sigma^2_{M^*}trP(w) = C_n(w)+2trP(w)(\hat\sigma^2_{M^*}-\sigma^2)$. Hence, from the result of Theorem 1′, it suffices to prove that, as $n\to\infty$,

$$\sup_{w\in H_n} trP(w)\left|\hat\sigma^2_{M^*}-\sigma^2\right|/R_n(w) \xrightarrow{p} 0. \tag{A.16}$$

Obviously,

$$\begin{aligned}&\sup_{w\in H_n}\frac{trP(w)}{R_n(w)}\left|\hat\sigma^2_{M^*}-\sigma^2\right| \le \frac{\bar k_{M^*}}{\xi_n}\left|\hat\sigma^2_{M^*}-\sigma^2\right|\\ &= \frac{\bar k_{M^*}}{\xi_n}\left|\frac{Y'(I-P_{(M^*)})Y}{n-\bar k_{M^*}}-\sigma^2\right|\\ &\le \frac{\bar k_{M^*}}{n-\bar k_{M^*}}\frac{\mu'(I-P_{(M^*)})\mu}{\xi_n}+\frac{2\bar k_{M^*}\left|\mu'(I-P_{(M^*)})e\right|}{\xi_n(n-\bar k_{M^*})}\\ &\quad +\frac{\bar k_{M^*}\left|e'(I-P_{(M^*)})e-\sigma^2(n-\bar k_{M^*})\right|}{\xi_n(n-\bar k_{M^*})}.\end{aligned} \tag{A.17}$$

Now, from (8), we have

$$\frac{\mu'(I - P_{(M^*)})\mu}{\xi_n^2} \to 0 \tag{A.18}$$

as $n \to \infty$. Using (A.18) and conditions (11) and (12), we obtain, as $n \to \infty$,

$$\begin{aligned}&\frac{\bar{k}_{M^*}}{n - \bar{k}_{M^*}} \frac{\mu'(I - P_{(M^*)})\mu}{\xi_n} \\ &\leq \left(\frac{\bar{k}_{M^*}^2}{n - \bar{k}_{M^*}} \frac{\mu'(I - P_{(M^*)})\mu}{\xi_n^2} \frac{\mu'\mu}{n - \bar{k}_{M^*}} \right)^{1/2} \to 0. \end{aligned} \tag{A.19}$$

Using Theorem 2 of Whittle (1960), Chebyshev's inequality, (7), (12) and (A.18), we observe that, for any $\delta > 0$, as $n \to \infty$,

$$\begin{aligned}&P\left\{ \frac{2\bar{k}_{M^*} \left|\mu'(I - P_{(M^*)})e\right|}{\xi_n(n - \bar{k}_{M^*})} > \delta \right\} \\ &\leq E\left(\mu'(I - P_{(M^*)})e\right)^2 \frac{4\bar{k}_{M^*}^2}{\delta^2 \xi_n^2 (n - \bar{k}_{M^*})^2} \\ &\leq \frac{C_5 4\bar{k}_{M^*}^2 \mu'(I - P_{(M^*)})\mu}{\delta^2 \xi_n^2 (n - \bar{k}_{M^*})^2} \to 0, \end{aligned} \tag{A.20}$$

and

$$\begin{aligned}&P\left\{ \frac{\bar{k}_{M^*} \left|e'(I - P_{(M^*)})e - \sigma^2(n - \bar{k}_{M^*})\right|}{\xi_n(n - \bar{k}_{M^*})} > \delta \right\} \\ &\leq E\left(e'(I - P_{(M^*)})e - \sigma^2(n - \bar{k}_{M^*})\right)^2 \frac{\bar{k}_{M^*}^2}{\delta^2 \xi_n^2 (n - \bar{k}_{M^*})^2} \\ &\leq \frac{C_6 \bar{k}_{M^*}^2 (n - \bar{k}_{M^*})}{\delta^2 \xi_n^2 (n - \bar{k}_{M^*})^2} \to 0, \end{aligned} \tag{A.21}$$

where C_5 and C_6 are constants. Combining (A.17) and (A.19)–(A.21), it is straightforward to see that (A.16) holds. This completes the proof of Theorem 2. □

References

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

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., 2009a Averaging estimators for a regression with a possible structural break. Econometric Theory (in press-a).

Hansen, B.E., 2009b. Averaging estimators for autoregressions with a near unit root. Journal of Econometrics (in press-b).

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

Kabaila, P., 2002. On variable selection in linear regression. Econometric Theory 18, 913–925.

Koop, G., 2003. Bayesian Econometrics. John Wiley and Sons, Chichester, England.

Leeb, H., Pötscher, B.M., 2003. The finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. Econometric Theory 19, 100–142.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21, 21–59.

Leeb, H., Pötscher, B.M., 2008. Can one estimate the unconditional distribution of post-model-selection estimators. Econometric Theory 24, 338–376.

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_p, C_L, cross validation and generalized cross-validations: Discrete index set. Annals of Statistics 15, 958–975.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

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

Pötscher, B.M., 2006. The distribution of model averaging estimators and an impossibility result regarding its estimation. IMS Lecture Note-Monograph Series Time Series and Related Topics 52, 113–129.

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 liner process. Annals of Statistics 8, 147–164.

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

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.