Least Squares Model Averaging

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

January 2006
Revised: August 2006

Abstract

This paper considers the problem of selection of weights for averaging across least-squares estimates obtained from a set of models. Existing model average methods are based on exponential AIC and BIC weights. In distinction, this paper proposes selecting the weights by minimizing a Mallows' criterion, the latter an estimate of the average squared error from the model average fit. We show that our new Mallows' Model Average (MMA) estimator is asymptotically optimal in the sense of achieving the lowest possible squared error in a class of discrete model average estimators. In a simulation experiment we show that the MMA estimator compares favorably with those based on AIC and BIC weights. The proof of the main result is an application of Li (1987).

*Research supported by the National Science Foundation. I gratefully thank the Co-Editor (Whitney Newey), three referees, and Benedickt Potscher for helpful comments.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706

1 Introduction

This paper develops a new model averaging estimator for least squares regression. A model average estimator is a weighted average of estimates obtained from different models. The goal in model averaging is to reduce estimation variance while controlling omitted variable bias. We propose a Mallows' criterion for the selection of the model weights, an estimate of the squared error. The empirical weights are found by numerical minimization of this criterion. We show that this method of weight selection is asymptotically optimal in the sense that the fitted estimates asymptotically achieve the minimum squared error in a class of discrete model average estimators.

Model selection has a long history in statistics and econometrics, and different methods have been advocated based on distinct estimation criteria, including AIC (Akaike (1973)), Mallows' C_p (Mallows (1973)), BIC (Schwarz (1978)), delete-one cross-validation (Stone (1974)), generalized cross-validation (Craven and Wahba (1979)), and the focused information criterion (Claeskens and Hjort (2003)). For GMM and empirical likelihood estimation, analogous criterion have been proposed by Andrews and Lu (2001), Hong, Preston and Shum (2003), and Hall, Inoue, Jana and Shin (2005).

Model averaging is an alternative to model selection. There is a large Bayesian literature, and a growing frequentist literature. Seminal contributions to Bayesian model averaging (BMA) include Draper (1995) and Raftery, Madigan and Hoeting (1997), and for literature reviews see Hoeting, et. al. (1999) and Raftery and Zheng (2003). Some applications in econometrics include Doppelhofer, Miller and Sala-i-Martin (2000), Brock and Durlauf (2001), Avramov (2002), Fernandez, Ley and Steel (2001a,b), Garratt, Lee, Pesaran and Shin (2003), Brock, Durlauf and West (2003), and Wright (2003ab). In the frequentist literature, Buckland et. al. (1997) and Burnham and Anderson (2002) suggested exponential AIC weights. The risk properties of a similar class of estimators is examined in Leung and Barron (2004). Yang (2001) and Yuan and Yang (2005) proposed a mixing estimator. Hjort and Claeskens (2003) provided an asymptotic analysis of model average estimators in likelihood-based models.

Shrinkage and parameter penalization are other alternatives to model selection and averaging. Some recent contributions include the Lasso-type estimators of Knight and Fu (2000), the penalized likelihood estimators of Fan and Li (2001) and Fan and Peng (2004), and the empirical Bayes estimator of Knox, Stock and Watson (2004).

There is also a large literature discussing the effects of model selection on inference. Potscher (1991) shows that AIC selection results in distorted inference. Kabaila (1995) examined the impact on confidence regions. Buhlmann (1999) presents conditions under which

post-model-selection (PMS) estimators are adaptive. Leeb and Potscher (2003, 2005ab) examine the unconditional and conditional distribution of PMS estimators and argue that they cannot be uniformly estimated.

The approach we take in this paper is similar to that of selecting the number of terms in a series expansion. Andrews (1991a) and Newey (1997) studied the convergence rates for series estimators and give conditions for asymptotic normality, but do not give rules for selection. Shibata (1980, 1981, 1983) demonstrated the asymptotic optimality of AIC selection in the context of Gaussian regressions. Shibata's analysis was extended to non-Gaussian autoregressions by Lee and Karagrigoriou (2001). Li (1987) demonstrated the asymptotic optimality of model selection in homoskedastic linear regression using Mallows' criterion, cross-validation and generalized cross-validation. Andrews (1991b) extended Li's results to the case of heteroskedastic errors. An thorough review of the asymptotic properties of model selection criteria has been provided by Shao (1997). A critique of the optimality criterion used in these papers is made by Kabaila (2002).

We propose a model average estimator with weights selected by minimizing a Mallows' criterion. Our main contribution is a demonstration that the Mallows' criterion is asymptotically equivalent to the squared error, and thus our Mallows' Model Average (MMA) estimator asymptotically achieves the lowest possible squared error in the class of model average estimators. Our proof is an application of Theorem 2.1 of Li (1987).

There are two important limitations of our results. First, we restrict attention to regressions with conditionally homoskedastic errors. Andrews (1991) showed that model selection by Mallows' criterion is not optimal under heteroskedasticity. The optimality of MMA will similarly fail under heteroskedasticity. Second, our asymptotic theory restricts the model average weights to a discrete set due to the difficulty of establishing uniformity over a weight vector whose dimension is unbounded. Developing weight selection methods which allow for heteroskedasticity and extending the proof technique to allow for continuous weights are important topics for future research.

Section 2 discusses the estimation framework and model average estimators. Section 3 calculates the average squared error of the model average estimator. Section 4 introduces the Mallows' criterion for the model average estimator and its sampling properties. Section 5 presents simulation evidence in support of the new MMA estimator. Proofs of the results are presented in the Appendix. A Gauss program which calculates the MMA estimator is available on the author's webpage www.ssc.wisc.edu/~bhansen.

2 Model Average Estimator

Let $(y_i, x_i) : i = 1, ..., n$ be a random sample, where y_i is real-valued while $x_i = (x_{1i}, x_{2i}, ...)$ is countably infinite. The model is the homoskedastic linear regression

$$\begin{aligned} y_i &= \mu_i + e_i && (1) \\ \mu_i &= \sum_{j=1}^{\infty} \theta_j x_{ji} && (2) \\ E\left(e_i \mid x_i\right) &= 0 && (3) \\ E\left(e_i^2 \mid x_i\right) &= \sigma^2. && (4) \end{aligned}$$

We assume $E\mu_i^2 < \infty$ and (2) converges in mean-square. The linearity of (2) is not essential to the idea model averaging, but it greatly simplifies the algebraic calculations. As the elements of x_i may be terms in a series expansion, (2) includes nonparametric regression.

Consider a sequence of approximating models $m = 1, 2, ...$, where the $m'th$ model uses the first k_m elements of x_i, where $0 \leq k_1 < k_2 < \cdots$. The $m'th$ approximating model is

$$y_i = \sum_{j=1}^{k_m} \theta_j x_{ji} + b_{mi} + e_i \tag{5}$$

where the approximation error is $b_{mi} = \sum_{j=k_m+1}^{\infty} \theta_j x_{ji}$. In matrix notation, $Y = X_m \Theta_m + b_m + e$, where $Y = (y_1, ..., y_n)'$, X_m is the $n \times k_m$ matrix with ij'th element x_{ji}, $\Theta_m = (\theta_1, ..., \theta_{k_m})'$, $b_m = (b_{m1}, ..., b_{mn})'$, and $e = (e_1, ..., e_n)'$.

Lurking behind (5) is an explicit ordering of the regressors x_{ji}. In some cases (such as a series expansion) this may not be troubling, but in other cases a natural ordering of the regressors may not be obvious. In practice, it may be feasible to order the regressors by groups, and this may be a common application of model averaging.

Let $M = M_n \leq n$ be an integer for which $X_{k_M}' X_{k_M}$ is invertible. For all $m \leq M$, the least-squares estimate of Θ_m is $\hat{\Theta}_m = \left(X_m' X_m\right)^{-1} X_m' Y$. Let $W = (w_1, ..., w_M)'$ be a weight vector in the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H}_n = \left\{ W \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}. \tag{6}$$

A model average estimator of Θ_M is

$$\hat{\Theta} = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\Theta}_m \\ 0 \end{pmatrix}. \tag{7}$$

A model average estimator bears some resemblance to a shrinkage estimator. This can be seen most plainly when the regressors are orthogonal. In this case, the j'th element of the model average estimator $\hat{\Theta}$ is the j'th element of the unconstrained estimator $\hat{\Theta}_M$ multiplied by $\sum_{m=j}^{M} w_m$. Thus the coefficient estimates are shrunk towards zero, with the degree of shrinkage increasing with j. However, in the standard case where the regressors are not orthogonal such a simple representation is not possible.

In the $m'th$ approximating model (5), let $\mu_m = X_m\Theta_m$ so that $\mu = \mu_m + b_m$ where $\mu = (\mu_1, ..., \mu_n)'$. The estimate of μ in the $m'th$ approximating model is $\hat{\mu}_m = X_m\hat{\Theta}_m = P_mY$ where $P_m = X_m \left(X_m'X_m\right)^{-1} X_m'$. The model average estimate of μ is $\hat{\mu}(W) = X_M\hat{\Theta} = P\left(W\right)Y$ where $P\left(W\right) = \sum_{m=1}^{M} w_m P_m$ is the implied "hat" matrix.

As the matrix $P(W)$ plays an important role in the algebraic structure of the model average estimator, we discuss here some of its properties. Note that $P\left(W\right)$ is symmetric but generally not idempotent. Let $\lambda_{\max}(A)$ denote the largest eigenvalue of A and define

$$\Gamma_M = \begin{bmatrix} k_1 & k_1 & k_1 & \cdots & k_1 \\ k_1 & k_2 & k_2 & \cdots & k_2 \\ k_1 & k_2 & k_3 & \cdots & k_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_1 & k_2 & k_3 & \cdots & k_M \end{bmatrix}. \tag{8}$$

Lemma 1

1. $\operatorname{tr}\left(P\left(W\right)\right) = \sum_{m=1}^{M} w_m k_m \equiv k(W)$.
2. $\operatorname{tr}\left(P\left(W\right)P\left(W\right)\right) = \sum_{m=1}^{M}\sum_{l=1}^{M} w_m w_l \min\left(k_l, k_m\right) = W'\Gamma_M W$.
3. $\lambda_{\max}\left(P(W)\right) \leq 1$.

3 Squared Error

Define the average squared error $L_n\left(W\right) = \left(\hat{\mu}\left(W\right) - \mu\right)'\left(\hat{\mu}\left(W\right) - \mu\right)$ and conditional squared error $R_n(W) = E\left(L_n\left(W\right) \mid X\right)$ where $X = \left\{x_1, ..., x_n\right\}$.

Lemma 2

$$R_n(W) = W'\left(A_n + \sigma^2\Gamma_M\right)W, \tag{9}$$

where Γ_M *is defined in (8),*

$$A_n = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_M \\ a_2 & a_2 & a_3 & \cdots & a_M \\ a_3 & a_3 & a_3 & \cdots & a_M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_M & a_M & a_M & \cdots & a_M \end{bmatrix}, \tag{10}$$

and $a_m = b_m'\left(I - P_m\right)b_m$. *Furthermore* $a_l \geq a_m$ *for* $l \leq m$, *and* $A_n + \sigma^2\Gamma_M > 0$ *if* $a_1 > 0$.

Lemma 2 shows that the conditional squared error $R_n(W)$ is a quadratic function in the weight vector W; an ellipsoid in $\mathbb{R}^M$ centered at the zero vector. It is interesting to observe that the optimal weight vector W which minimizes $R_n(W)$ necessarily puts non-zero weight on at least two models, except in the special case that $a_1 = a_M$. To see this, suppose that $M = 2$, in which case $R_n(W)$ is uniquely minimized by $w_1 = \left(1 + \left(a_1 - a_2\right)/\sigma^2\left(k_2 - k_1\right)\right)^{-1}$ which is in $(0, 1)$ unless $a_1 = a_2$.

4 Mallows' Criterion

The Mallow's criterion for the model average estimator is

$$C_n(W) = \left(Y - X_M\hat{\Theta}\right)'\left(Y - X_M\hat{\Theta}\right) + 2\sigma^2 k(W) \tag{11}$$

where $k(W)$ defined in Lemma 1 is the effective number of parameters. Definition (11) depends on the unknown σ^2. We discuss the replacement of σ^2 with an estimate below.

The Mallows' criterion may be used to select the weight vector W. Define

$$\hat{W} = \underset{W \in \mathcal{H}_n}{\operatorname{argmin}}\, C_n(W) \tag{12}$$

the empirical Mallows' selected weight vector. As there is no closed-form solution to (12), the weight vector must be found numerically. For this calculation, it is convenient to write (11) in the following form. Let $\hat{e}_m$ be the $n \times 1$ residual vector from the m'th model, let $\bar{e} = (\hat{e}_1, ..., \hat{e}_M)$ be the $n \times M$ matrix collection of these residuals, and let $K = (k_1, ..., k_M)'$

be the $M \times 1$ vector of the number of parameters in the M models. Then (11) equals

$$C_n(W) = W'\bar{e}'\bar{e}W + 2\sigma^2 K'W \tag{13}$$

which is linear-quadratic in W. The solution (12) minimizes (13) subject to the nonnegativity and summation constraints (6). This is a classic quadratic programming problem, for which numerical algorithms are readily available. (For example, in the GAUSS programming language the procedure QPROG is appropriate.) The solution may be a unit vector or an interior value. If M is moderately large, a typical solution $\hat{W}$ can put zero weight on many of the individual models. The Mallows' Model Average (MMA) estimator is (7) using the weight vector $\hat{W}$.

We present two justifications for the Mallows' criterion. Our first is the classic observation that $C_n(W)$ is an unbiased estimate of the expected squared error plus a constant.

Lemma 3

$$EC_n(W) = EL_n(W) + n\sigma^2. \tag{14}$$

Our second justification is that if the weights are restricted to a discrete set, the empirical Mallows' weight vector asymptotically minimizes the squared error. Specifically, for some integer N let the weights w_m be restricted to the set $\{0, \frac{1}{N}, \frac{2}{N}, \cdots, 1\}$, and let $\mathcal{H}_n(N)$ be the subset of $\mathcal{H}_n$ restricted to this set of weights. Let

$$\hat{W}_N = \underset{W \in \mathcal{H}_n(N)}{\operatorname{argmin}} \; C_n(W)$$

be the Mallows' weight vector, the choice obtained by minimizing the Mallows' criterion over the discrete weight set $\mathcal{H}_n(N)$.

The following result is an application of Theorem 2.1 of Li (1987), who showed the asymptotic optimality of Mallows' criterion for model selection.

Theorem 1 *As $n \to \infty$, if*

$$\xi_n = \inf_{W \in \mathcal{H}_n} \; R_n(W) \to \infty \tag{15}$$

almost surely, and for some fixed integer $N < \infty$

$$E\left(|e_i|^{4(N+1)} \mid x_i\right) \leq \kappa < \infty \tag{16}$$

then

$$\frac{L_n\left(\hat{W}_N\right)}{\inf_{W \in \mathcal{H}_n(N)} L_n(W)} \to_p 1. \tag{17}$$

Note that the theorem places no restriction on M, the largest model included in the model average (other than the requirement that $X_{k_M}'X_{k_M}$ is invertible). Thus M may be fixed as $n \to \infty$ or $M = M_n$ may diverge to infinity.

Theorem 1 shows that the squared error obtained using the Mallows' weight vector $\hat{W}_N$ is asymptotically equivalent to the infeasible optimal weight vector. This means that the MMA estimator is asymptotically optimal in the class of model average estimators (7) where the weight vector W is restricted to the set $\mathcal{H}_n(N)$.

The restriction of $\mathcal{H}_n$ to $\mathcal{H}_n(N)$ can be made less binding by picking N large, which can be done so long as the conditional moment bound (16) holds. This restriction is imposed because the proof of (17) requires that $C_n(W)$ is asymptotically equivalent to $L_n(W)$ uniformly over W. The trouble is that the dimension of the set $\mathcal{H}_n$ is unbounded when $M_n \to \infty$ as $n \to \infty$, rendering conventional proof methods inapplicable.

Theorem 1 requires condition (15), which specifies that there is no finite approximating model m for which the bias is zero. This assumption is conventional for nonparametric regression. For example, if $\gamma_m \sim m^{-\alpha}$ then we have the explicit rate $\xi_n \sim n^{1/(1+2\alpha)}$. If (15) fails, then MMA will not satisfy the optimality (17).

In practice, σ^2 is unknown so (11) needs to be computed with a sample estimate. One choice is $\hat{\sigma}_K^2 = (n-K)^{-1}\left(Y - X_K\hat{\Theta}_K\right)'\left(Y - X_K\hat{\Theta}_K\right)$ where $k_K = K$ corresponds to a "large" approximating model. Other estimators for σ^2 have been proposed in the nonparametric regression literature. Lemma 3 continues to hold if $\hat{\sigma}_K^2$ is unbiased for σ^2, which holds if $b_K = 0$, so the $K'th$ approximating model has no bias. Theorem 1 holds as stated so long as $\hat{\sigma}_K^2$ is consistent for σ^2, which is valid as shown next.

Theorem 2 *If $K \to \infty$ and $K/n \to 0$ as $n \to \infty$ then $\hat{\sigma}_K^2 \to_p \sigma^2$ as $n \to \infty$.*

5 Finite Sample Investigation

We now investigate the finite sample MSE of the our model average estimator in a simple simulation experiment. The setting is the infinite-order regression $y_i = \sum_{j=1}^{\infty}\theta_j x_{ji} + e_i$. We set $x_{1i} = 1$ to be the intercept, and the remaining x_{ji} are iid $N(0,1)$. The error e_i is $N(0,1)$ and independent of x_i. (Other experiments, not reported, showed that the results are not sensitive to alternative distributions for the regressors and regression error.) The parameter are determined by the rule $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$. The population $R^2 = c^2/(1+c^2)$ is controlled by the parameter c.

The sample size is varied between $n = 50$, 150, 400, and 1000. The parameter α is varied between .5, 1.0 and 1.5. The larger α implies that the coefficients θ_j decline more quickly

with j. The number of models M is determined by the rule $M = 3n^{1/3}$ (so $M = 11$, 16, 22, and 30 for the four sample sizes). The coefficient c was selected to control the population R^2 to vary on a grid between 0.1 and 0.9.

We consider five estimators: (1) AIC model selection (AIC); (2) Mallows' model selection (Mallows) (3) smoothed AIC (S-AIC); (4) smoothed BIC (S-BIC) and (5) Mallows' Model Averaging (MMA). The AIC criterion for model Θ_m is $AIC_m = n \ln \hat{\sigma}_m^2 + 2m$. The AIC model selection estimator is $\hat{\Theta}_{\hat{m}}$ where $\hat{m}$ minimizes AIC_m. S-AIC was introduced by Buckland et. al. (1997) and embraced by Burnham and Anderson (2002) and Hjort and Claeskens (2003). It is the least-squares model average estimator (7) with the weights $w_m = \exp\left(-\frac{1}{2}AIC_m\right) / \sum_{j=1}^{M} \exp\left(-\frac{1}{2}AIC_m\right)$. S-BIC is a simplified form of Bayesian model averaging. It is the least-squares model average estimator (7) with the weights $w_m = \exp\left(-\frac{1}{2}BIC_m\right) / \sum_{j=1}^{M} \exp\left(-\frac{1}{2}BIC_m\right)$ where $BIC_m = n \ln \hat{\sigma}_m^2 + \ln(n)m$.

To evaluate the estimators we compute the risk (expected squared error). We do this by computing averages across 100,000 simulation draws. For each parameterization, we normalize the risk by dividing by the risk of the infeasible optimal least-squares estimator (the risk of the best-fitting model m).

The risk calculations are displayed in Figures 1 to 3 for $\alpha = .5$, 1.0 and 1.5, respectively. In each figure, the four panels display sample sizes. In each panel, risk (expected squared error) is displayed on the y-axis, and the population R^2 on the x-axis. The two dotted lines correspond to AIC and Mallows' selection. The dashed, dash-dotted, and solid lines correspond to S-AIC, S-BIC, and MMA, respectively.

In each panel, the AIC and Mallows' selection methods have quite similar risk. The smoothed AIC estimator achieves a lower risk than AIC model selection, which is consistent with the findings of the earlier literature. The S-AIC and MMA estimators are nearly equivalent for the case $\alpha = 1.5$ and large n, otherwise MMA achieves a lower risk than S-AIC. In many cases, its normalized risk is less than one, meaning that it is lower than that of infeasible optimal model selection.

It is also instructive to contrast the performance of the MMA and S-BIC estimators. The MMA estimator achieves lower risk in most cases, but S-BIC has lower risk when n and R^2 are small, and its relative performance improves when α is large. In particular, S-BIC has much lower risk when $\alpha = 1.5$ and $n = 50$. Their relative performance depends strongly on sample size, with the S-BIC estimator showing increasing relative risk, and the MMA showing decreasing relative risk, as n increases. In many cases, however, the risk of the S-BIC estimator is quite poor relative to the other methods.

6 Appendix

Proof of Lemma 1. Parts 1 and 2 follow from the facts that $\operatorname{tr}(P_m) = k_m$, $\operatorname{tr}(P_m P_l) = \operatorname{tr}\left(P_{\min(k_l,k_m)}\right) = \min(k_l, k_m)$, and simple algebra. Part 3 uses the fact that P_m is idempotent so that

$$\lambda_{\max}(P(W)) = \max_{\eta} \frac{\eta' P(W) \eta}{\eta'\eta} \leq \sum_{m=1}^{M} w_m \max_{\eta} \frac{\eta' P_m \eta}{\eta'\eta} = 1.$$

Proof of Lemma 2. Note that $\mu - \hat{\mu}(W) = (I - P(W))\mu - P(W)e$ and thus

$$L_n(W) = \mu'(I - P(W))(I - P(W))\mu - 2e'P(W)B_nW + e'P(W)P(W)e. \tag{18}$$

Lemma 1 and assumption (4) imply that

$$E\left(e'P(W)P(W)e \mid X\right) = \sigma^2 \operatorname{tr}(P(W)P(W)) = \sigma^2 W'\Gamma_M W.$$

Taking conditional expectations of (18) we obtain

$$E(L_n(W) \mid X) = \mu'(I - P(W))(I - P(W))\mu + W'\sigma^2\Gamma_M W.$$

Define $b_m^* = (I - P_m)\mu = (I - P_m)b_m$ and $B_n = [b_1^*, ..., b_M^*]$. Then

$$(I - P(W))\mu = \sum_{m=1}^{M} w_m b_m^* = B_n W. \tag{19}$$

Note that for $l \leq m$, $P_l P_m = P_l$ and $(I - P_m)b_l = (I - P_m)b_m$. Then

$$b_l^{*\prime} b_m^* = b_l'(I - P_l)(I - P_m)b_m = b_l'(I - P_m)b_m = b_m'(I - P_m)b_m = a_m$$

and thus $B_n'B_n = A_n$. It follows that $\mu'(I - P(W))(I - P(W))\mu = W'B_n'B_nW = W'A_nW$ and we obtain (9). Furthermore, for $l \leq m$ note that

$$b_m'(I - P_m)b_m = b_l'(I - P_m)b_l = b_l'(I - P_l)b_l - b_l'P_m(I - P_l)P_m b_l \leq b_l'(I - P_l)b_l$$

and thus $a_m \geq a_l$ as claimed.

We now show that $A_n + \sigma^2\Gamma_M > 0$, which holds if for all $\alpha \neq 0$, $\alpha'\left(A_n + \sigma^2\Gamma_M\right)\alpha > 0$. If $\alpha = \iota_1$ is the first unit vector then $\alpha'\left(A_n + \sigma^2\Gamma\right)\alpha = a_1^2 + \sigma^2 k_1^2 > 0$. Otherwise if $\alpha \neq \iota_1$ note

that $\alpha' A_n \alpha = \alpha' B_n' B_n \alpha \geq 0$ and by the definition of Γ and some algebraic manipulations,

$$\alpha' \Gamma_M \alpha = k_1 \left(\sum_{m=1}^{M} \alpha_m \right)^2 + (k_2 - k_1) \left(\sum_{m=2}^{M} \alpha_m \right)^2 + \cdots + (k_M - k_{M-1}) \alpha_M^2 > 0$$

Thus $\alpha' \left(A_n + \sigma^2 \Gamma_M \right) \alpha > 0$ as required.

Proof of Lemma 3. By straightforward algebra

$$C_n(W) - L_n(W) = e'e + 2e' \left(I - P(W) \right) \mu - 2 \left(e' P(W) e - \sigma^2 k(W) \right). \tag{20}$$

Lemma 1 and assumption (4) imply that

$$E \left(e' P(W) e \mid X \right) = \sigma^2 \operatorname{tr} \left(P(W) \right) = \sigma^2 k(W). \tag{21}$$

Taking expectations of (20), equation (14) follows directly .

Proof of Theorem 1. Theorem 2.1 of Li (1987) established (17) for a broad class of linear estimators. It is sufficient to verify that his equations (A.1), (A.2) and (A.3) hold almost surely, conditional on X. Indeed, (A.1) is implied by part 3 of Lemma 1, and (A.2) holds by (16). It remains to show (A.3), which in our notation is

$$\sum_{W \in \mathcal{H}_n(N)} R_n(W)^{-(N+1)} \to 0 \tag{22}$$

almost surely as $n \to \infty$.

For integers $1 \leq j_1 \leq j_2 \leq \cdots \leq j_N$ let $W_{j_1, j_2, \ldots, j_N}$ be the weight vector which sets $w_{j_l} = 1/N$ for $l = 1, ..., N$, and the remainder zero. We can write

$$\mathcal{H}_n(N) = \{ W_{j_1, j_2, \ldots, j_N} : 1 \leq j_1 \leq j_2 \leq \cdots \leq j_N \leq M \}.$$

The restriction of the weights to the form $1/N$ is without loss of generality since the weak ordering of the integers j_k allows ties. We then have

$$\sum_{W \in \mathcal{H}_n(N)} R_n(W)^{-(N+1)} \leq \sum_{j_N=1}^{\infty} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} R_n(W_{j_1, j_2, \ldots, j_N})^{-(N+1)}. \tag{23}$$

Now break the sum into two groups based on whether $k_{j_N} < \xi_n$ or $k_{j_N} \geq \xi_n$. For the first group (which has less than ξ_n^N elements), use the bound $R_n(W) \geq \xi_n$ from (15) and for the

second group use the simple bound

$$R_n(W_{j_1,j_2,\ldots,j_N}) \geq \sigma^2 W'_{j_1,j_2,\ldots,j_N} \Gamma_M W_{j_1,j_2,\ldots,j_N} \geq \frac{\sigma^2}{N^2} k_{j_N} \geq \frac{\sigma^2}{N^2} j_N$$

where the first inequality is implied by (9) and the second uses the definitions of Γ_M and $W_{j_1,j_2,\ldots,j_N}$.

Using these bounds,

$$\begin{aligned}
\sum_{j_N=1}^{\infty} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} R_n(W_{j_1,j_2,\ldots,j_N})^{-(N+1)} &\leq \xi_n^{-1} + \sum_{j_N=\xi_n}^{\infty} \sum_{j_{N-1}=1}^{j_N} \cdots \sum_{j_1=1}^{j_2} \left(\frac{\sigma^2}{N^2} j_N\right)^{-(N+1)} \\
&\leq \xi_n^{-1} + \left(\frac{\sigma^2}{N^2}\right)^{-(N+1)} \sum_{j_N=\xi_n}^{\infty} j_N^{-2} \\
&\leq \xi_n^{-1} + \left(\frac{\sigma^2}{N^2}\right)^{-(N+1)} \xi_n^{-1} \\
&\to 0
\end{aligned}$$

almost surely as $n \to \infty$. Together with (23), this establishes (22) as desired.

Proof of Theorem 2. Since $\hat{e}_K = Y - X_K \hat{\Theta}_K = (I - P_K)\, e + (I - P_K)\, b_K$, we see that

$$\hat{\sigma}_K^2 = \frac{1}{n-K} e' (I - P_K)\, e + \frac{1}{n-K} b'_K (I - P_K)\, b_K + 2 \frac{1}{n-K} e' (I - P_K)\, b_K. \tag{24}$$

We examine the terms on the right side of (24). First, since $Ee' (I - P_K)\, e = \sigma^2 (n - K)$, by Theorem 2 of Whittle (1960)

$$E \left| e' (I - P_K)\, e - \sigma^2 (n-K) \right|^2 \leq C_2 \kappa^{1/(N+\delta)} \operatorname{tr} \left((I - P_K)\, (I - P_K) \right) = C_2 \kappa^{1/(N+\delta)} (n - K).$$

Thus for any $\delta > 0$, by Markov's inequality

$$\begin{aligned}
P\left(\left| \frac{1}{n-K} e' (I - P_K)\, e - \sigma^2 \right| > \delta \right) &\leq \frac{E \left| e' (I - P_K)\, e - \sigma^2 (n-K) \right|^2}{\delta^2 (n-K)^2} \\
&\leq \frac{C_2 \kappa^{1/(N+\delta)}}{\delta^2 (n-K)} \to 0
\end{aligned}$$

so $(n-K)^{-1} e' (I - P_K)\, e \to_p \sigma^2$. Second,

$$\frac{1}{n-K} E \left(b'_K (I - P_K)\, b_K \right) \leq \frac{n}{n-K} E b_{Ki}^2 \to 0$$

since $K \to \infty$ as $n \to \infty$ and the square integrability of μ_i implies $Eb_{Ki}^2 \to 0$ as $K \to \infty$. This implies $(n-K)^{-1} b_K' (I - P_K) b_K \to_p 0$. Similarly, the third term on the right side of (24) is $o_p(1)$ and we conclude that $\hat{\sigma}_K^2 \to_p \sigma^2$.

References

[1] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[2] Andrews, D. W. K. (1991a): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59, 307-345.

[3] Andrews, D. W. K. (1991b): "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, 359-377.

[4] Andrews, D. W. K. and B. Lu (2001): "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models," *Journal of Econometrics*, 101, 123-164.

[5] Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Finance,* 64, 423-458.

[6] Brock, W. and S. Durlauf (2001): "Growth empirics and reality," *World Bank Economic Review*, 15, 229-272.

[7] Brock, W., S, Durlauf, and K. D. West (2003): "Policy analysis in uncertain economic environments," *Brookings Papers on Economic Activity*, 1, 235-322.

[8] Buckland, S.T., K.P. Burnhamn and N.H. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603-618.

[9] Buhlmann, Peter (1999): "Efficient and adaptive post-model-selection estimators," *Journal of Statistical Planning and Inference*, 79, 1-9.

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

[11] Claeskens, G. and N. L. Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900-916.

[12] Craven, P. and G. Wahba (1979): "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, 31, 377-403.

[13] Doppelhofer, G., R. Miller and X. Sala-i-Martin (2004): "Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach," *American Economic Review*, 94, 813-835..

[14] Draper, D. (1995): "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society, Series B*, 57, 45-70.

[15] Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348-1360.

[16] Fan, J. and H. Peng (2004): "Nonconcave penalized likelihood with a diverging number of parameters," *Annals of Statistics*, 32, 928-961.

[17] Fernandez, C. E. Ley and M.F.J. Steel (2001a): "Benchmark priors for Bayesian model averaging," *Journal of Econometrics,* 100, 381-427.

[18] Fernandez, C. E. Ley and M.F.J. Steel (2001b): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563-576.

[19] Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin (2003): "Forecasting uncertainties in macroeconomic modelling: An application to the UK economy," *Journal of the American Statistical Association*, 98, 829-838.

[20] Hall, A. R., A. Inoue, K. Jana, and C. Shin (2005): "Information in Generalized Method of Moments Estimation and Entropy Based Moment Selection," *Journal of Econometrics*, forthcoming.

[21] Hjort, N. L. and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899.

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

[23] Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19, 923-943.

[24] Kabaila, P. (1995): "The effect of model selection on confidence regions and prediction regions," *Econometric Theory*, 11, 537-549.

[25] Kabaila, P. (2002): "On variable selection in linear regression," *Econometric Theory*, 18, 913-925.

[26] Knight, K. and W. Fu (2000): "Asymptotics for Lasso-type Estimators," *Annals of Statistics*, 28, 1356-1378.

[27] Knox, T., J. H. Stock and M. W. Watson (2004): "Empirical Bayes Regression with Many Regressors," working paper.

[28] Lee, S. and A. Karagrigoriou (2001): "An asymptotically optimal selection of the order of a linear process," *Sankhya*, 63, Series A, 93-106.

[29] Leeb, H. and B. M. Pötscher (2003): "The finite-sample distribution of post-model-selection estimators and uniform versus non-uniform approximations," *Econometric Theory*, 19, 100-142.

[30] Leeb, H. and B. M. Pötscher (2005a): "Model selection and inference: Facts and Fiction," *Econometric Theory*, 21, 21-59.

[31] Leeb, H. and B. M. Pötscher (2005b): "Can one estimate the conditional distribution of post-model-selection estimators?" *Annals of Statistics*, forthcoming.

[32] Leung, G. and A. R. Barron (2004): "Information theory and mixing least-squares regressions," working paper.

[33] Li, Ker-Chau (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set," *Annals of Statistics*, 15, 958-975.

[34] Mallows, C.L. (1973): "Some comments on C_p," *Technometrics*, 15, 661-675.

[35] Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147-168.

[36] Potscher, B. M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[37] Raftery, A. E., D. Madigan, J. A. Hoeting (1997): "Bayesian model averaging for regression models," *Journal of the American Statistical Association*, 92, 179-191.

[38] Raftery, A. E. and Y. Zheng (2003) "Long-run performance of Bayesian model averaging," working paper, University of Washington.

[39] Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

[40] Shao, J. (1997): "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, 221-264.

[41] 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.

[42] Shibata, R. (1981): "An optimal selection of regression variables," *Biometrika*, 68, 45-54.

[43] Shibata, R. (1983): "Asymptotic mean efficiency of a selection of regression variables," *Annals of the Institute of Statistical Mathematics*, 35, 415-423.

[44] Stone, M. (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, Series B*, 36, 276-278.

[45] Whittle, P. (1960): "Bounds for the moments of linear and quadratic forms in independent variables," *Theory of Probability and Its Applications*, 5, 302-305.

[46] Wright, J. H. (2003a): "Bayesian model averaging and exchange rate forecasting," *Federal Reserve Board International Finance Discussion Papers*, 779.

[47] Wright, J. H. (2003b): "Forecasting US Inflation by Bayesian Model Averaging," *Federal Reserve Board International Finance Discussion Papers*, 780.

[48] Yang, Y. (2001): "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574-586.

[49] Yuan, Z. and Y. Yang (2005): "Combining linear regression models: When and How?" *Journal of the American Statistical Association*, 100, 1202-1214.

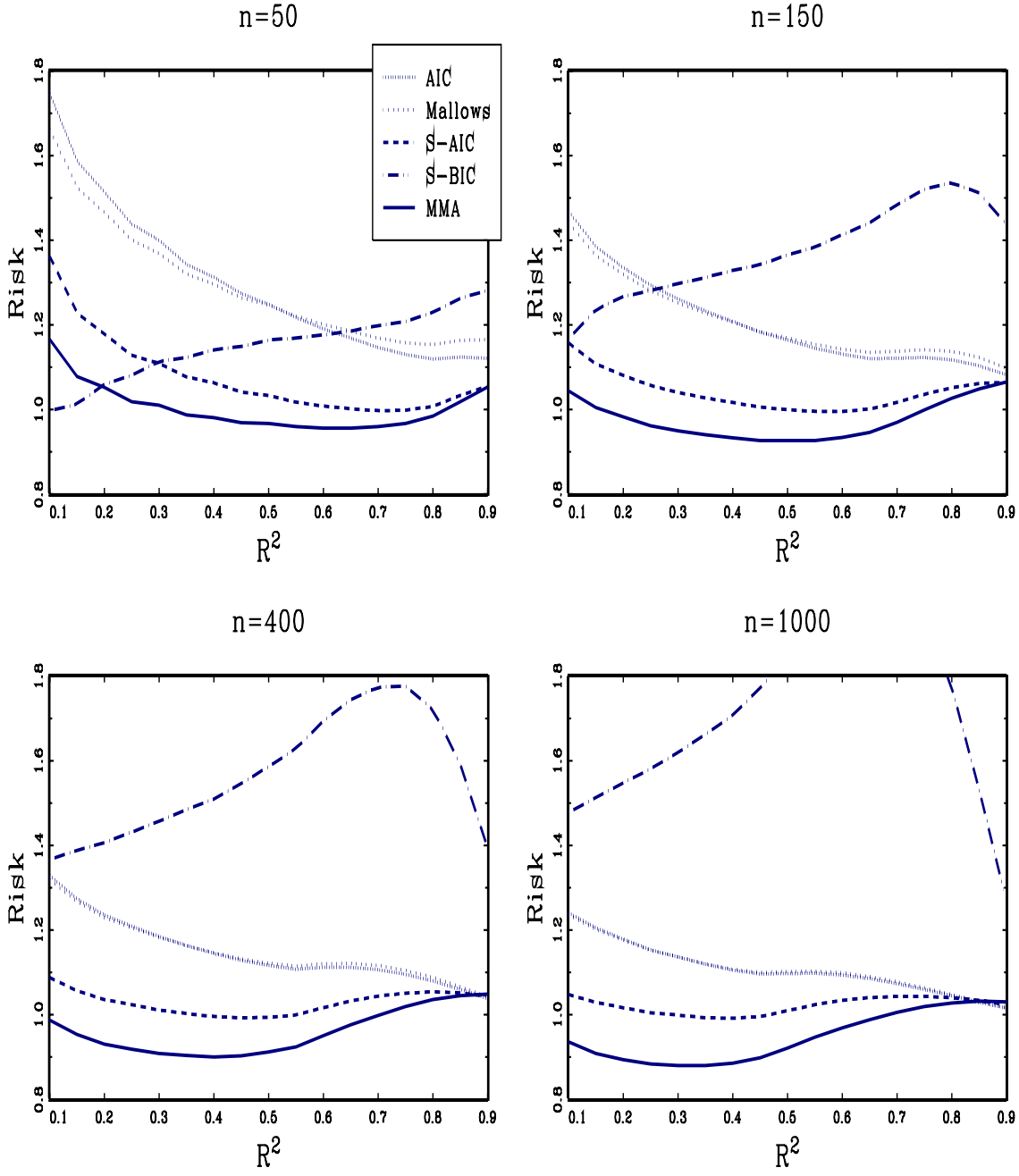

Figure 1: $\alpha = .5$

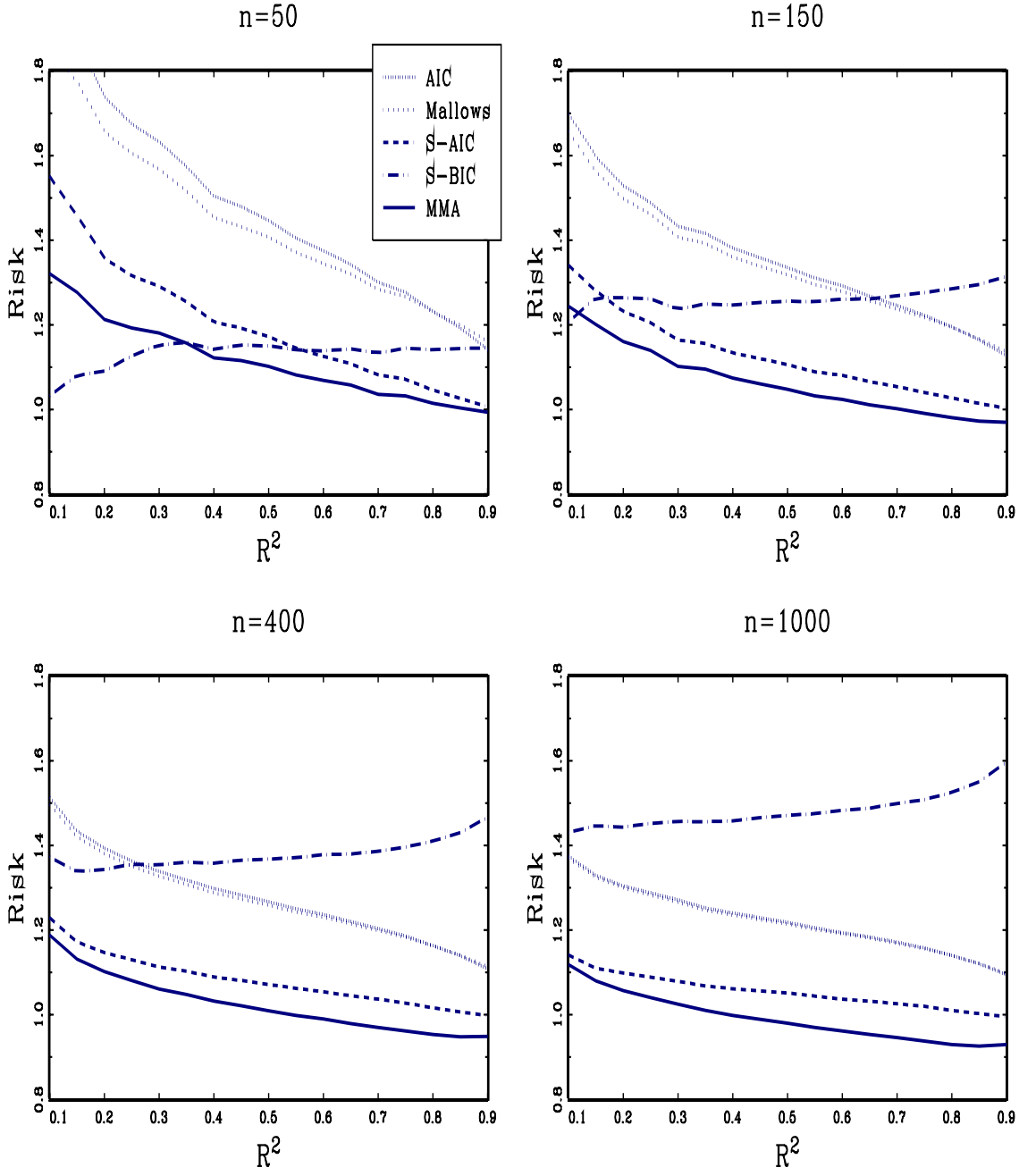

Figure 2: $\alpha = 1.0$

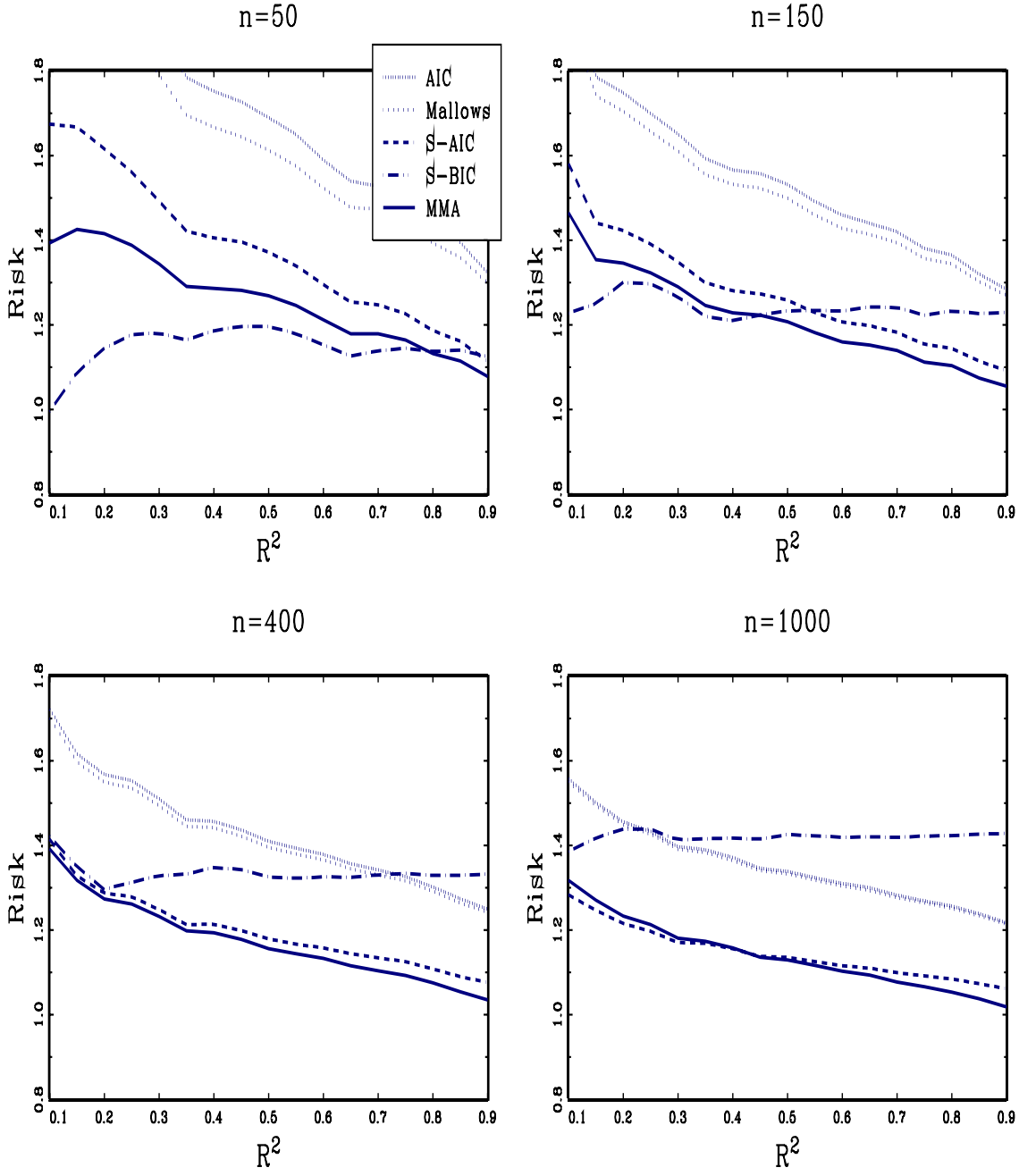

Figure 3: $\alpha = 1.5$