Journal of Econometrics 146 (2008) 342–350

Contents lists available at ScienceDirect

Journal of Econometrics

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

Least-squares forecast averaging

Bruce E. Hansen *

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

ARTICLE INFO

Article history:
Available online 7 September 2008

JEL classification:
C52
C53

Keywords:
Mallows
AIC
BIC
BMA
Forecast combination
Model selection

ABSTRACT

This paper proposes forecast combination based on the method of Mallows Model Averaging (MMA). The method selects forecast weights by minimizing a Mallows criterion. This criterion is an asymptotically unbiased estimate of both the in-sample mean-squared error (MSE) and the out-of-sample one-step-ahead mean-squared forecast error (MSFE). Furthermore, the MMA weights are asymptotically mean-square optimal in the absence of time-series dependence. We show how to compute MMA weights in forecasting settings, and investigate the performance of the method in simple but illustrative simulation environments. We find that the MMA forecasts have low MSFE and have much lower maximum regret than other feasible forecasting methods, including equal weighting, BIC selection, weighted BIC, AIC selection, weighted AIC, Bates–Granger combination, predictive least squares, and Granger–Ramanathan combination.

1. Introduction

Forecast combination has a long history in econometrics. While a broad consensus is that forecast combination improves forecast accuracy, there is no consensus concerning how to form the forecast weights. The most recent literature has focused on two particularly appealing methods—simple averaging and Bayesian averaging. The simple averaging method picks a set of models and then gives them all equal weight for all forecasts. The Bayesian averaging method computes forecast weights as a by-product of Bayesian model averaging (BMA).

This paper introduces a new method appropriate for linear models estimated by least squares. The method is to construct forecast combinations using the weights computed by Mallows Model Averaging (MMA), the weights which minimize the generalized Mallows criterion introduced in Hansen (2007). The Mallows criterion is an estimator of mean-squared error (MSE) and mean-squared forecast error (MSFE), and MMA weights are asymptotically optimal with respect to mean-square loss. We therefore expect MMA combination to produce point forecasts with low MSFE. In the context of linear regressions with iid data, Hansen (2007) introduced the idea of model averaging using the weights which minimize the Mallows criterion, demonstrated that the Mallows criterion is an unbiased estimate of the MSE, and showed that the Mallows averaging estimator is asymptotically optimal in the sense of achieving the best possible MSE. Our contribution in this paper is to propose using these weights for forecast combination, show that the Mallows criterion is asymptotically unbiased for the MSE and MSFE when the observations are a stationary time-series, and demonstrate the relative performance of the method in simulation environments.

As mentioned above, two powerful existing methods for forecast combination are simple averaging and Bayesian averaging. Both have been shown to be extremely versatile and successful in applications, yet neither is inherently satisfying. Simple averaging only makes sense if the class of models under consideration is reasonable. If a terrible model is included in the class of forecasting models, simple averaging will pay the penalty. This induces an inherent arbitrariness, and thus the method is incomplete unless augmented by a description of how the initial class of models is determined, which destroys the inherent simplicity of the method.

On the other hand, the fact that BMA relies on priors (over the class of models and over the parameters in the models) means that this method suffers from the arbitrariness which is inherent in prior specification. Furthermore, the BMA paradigm is inherently misspecified. It is developed under the assumption that the truth is one finite-dimensional parametric model out of a class of models under consideration. The goal is to find the "true" model out of this class. This paradigm and goal is inherently misspecified and misguided, as it is more appropriate to think of models as approximations, and that the "true" model is more complex than any of the models in our explicit class. When we fit models, we balance specification error (bias) against overparameterization (variance). The correct goal is to define the object of interest (such as forecast mean-squared error) and then evaluate methods based

* Tel.: +1 608 263 3880; fax: +1 608 263 3876.
E-mail address: bhansen@ssc.wisc.edu.
URL: http://www.ssc.wisc.edu/~bhansen.

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.08.022

on this criterion, without assuming that we necessarily have the correct model.

Mallows Model Averaging takes exactly the desired approach. The goal is to obtain the set of weights which minimizes the MSE over the set of feasible forecast combinations. The generalized Mallows criterion is an estimate of the MSE and MSFE, and the weights which minimize this criterion are asymptotically optimal in some settings.

The Mallows criterion for model selection was introduced by Mallows (1973), and is similar to the information criteria of Akaike (1973) and Shibata (1980). The asymptotic optimality of model selection by this class of criterion has been studied by Shibata (1980, 1981, 1983), Li (1987), Banasali (1996), Lee and Karagrigoriou (2001), Ing (2003, 2004, 2007) and Ing and Wei (2003, 2005). Akaike (1979) proposed using the exponentiated AIC as model weights, and this suggestion was picked up and expanded by Buckland et al. (1997) and Burnham and Anderson (2002). Hjort and Claeskens (2003) introduced a general class of frequentist model average estimators, including methods similar to Mallows model averaging.

The Bayesian information criterion was introduced by Schwarz (1978) as a method for model selection. There is a large literature on Bayesian Model Averaging; see the review by Hoeting et al. (1999). Some applications in econometrics include Sala-i-Martin et al. (2004), Brock and Durlauf (2001), Avramov (2002), Fernandez et al. (2001a,b), Garratt et al. (2003) and Brock et al. (2003).

The idea of forecast combination was introduced by Bates and Granger (1969), extended by Granger and Ramanathan (1984), and spawned a large literature. Some excellent reviews include Granger (1989), Clemen (1989), Diebold and Lopez (1996), Hendry and Clements (2002), Timmermann (2006) and Stock and Watson (2006). The idea of using Bayesian model averaging for forecast combination was pioneered by Min and Zellner (1993) and its usefulness recently demonstrated by Wright (2003a,b). Stock and Watson (1999, 2004, 2005) have provided detailed empirical evidence demonstrating the gains in forecast accuracy through forecast combination, and in particular have demonstrated the success of simple averaging (equal weights) along with Bayesian model averaging.

The plan of the paper is simple. Section 2 introduces the model and approximating linear forecasting models. Section 3 introduces forecast combination and reviews existing combination methods. Section 4 presents the Mallows criterion and the MMA forecast combination. Section 5 shows that the Mallows criterion is an asymptotically unbiased estimate of the in-sample mean-squared error. Section 6 shows that this MSE criterion is approximately equivalent to out-of-sample mean-squared forecast error. Section 7 reviews Hansen's (2007) demonstration of the asymptotic efficiency of MMA in independent samples. Section 8 presents the results of two simulation experiments.

2. Approximating models

Let $\{y_t, \mathbf{x}_t : t = 1, \ldots, n\}$ be a sample where y_t is real-valued and $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots)'$ is countably infinite. Consider the regression model

$$y_t = \mu_t + e_t \tag{1}$$

$$\mu_t = \sum_{j=1}^{\infty} a_j x_{jt} = \mathbf{a}'\mathbf{x}_t \tag{2}$$

$$E\left(e_t \mid \mathbf{x}_t\right) = 0 \tag{3}$$

$$E\left(e_t^2 \mid \mathbf{x}_t\right) = \sigma^2. \tag{4}$$

In Eq. (2), $\mathbf{a} = (a_1, a_2, \ldots)'$ denotes a countably infinite vector, and we assume that sum (2) converges in mean square.

The goal is to construct a point forecast f_{n+1} of y_{n+1} given $\mathbf{x}_{n+1}$. The optimal mean-square forecast is the conditional mean μ_{n+1}, and therefore empirical forecasts are estimates of μ_{n+1}.

As (2) possibly contains an infinite number of coefficients but the sample size is finite, a finite-dimensional approximating model is estimated in practice. Suppose we have a set of M approximating models where the m'th uses the first $k(m)$ regressors. (The leading case sets $k(m) = m$.) This imposes the strong assumption that the set of forecasting models is strictly nested. This is highly restrictive (as there are many cases where non-nested models are of interest) but is necessary for application of the methods discussed in this paper. Extending the methods to allow for non-nested models would be highly desirable.

Letting $\mathbf{x}_t(m) = (x_{1t}, x_{2t}, \ldots, x_{k(m)t})'$ and $\mathbf{a}(m) = (a_1, a_2, \ldots, a_{k(m)})'$ this model can be written as

$$y_t = \mathbf{a}(m)'\mathbf{x}_t(m) + e_t(m)$$

where

$$e_t(m) = e_t + \sum_{j=k(m)+1}^{\infty} a_j x_{jt},$$

or in matrix notation

$$\mathbf{y} = \mathbf{X}(m)\mathbf{a}(m) + \mathbf{e}(m).$$

The least-squares estimate of $\mathbf{a}(m)$ is

$$\hat{\mathbf{a}}(m) = \left(\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\mathbf{X}(m)'\mathbf{y},$$

the least-squares residuals are

$$\hat{\mathbf{e}}(m) = \mathbf{y} - \mathbf{X}(m)\hat{\mathbf{a}}(m),$$

and residual variance estimate

$$\hat{\sigma}^2(m) = \frac{1}{n}\hat{\mathbf{e}}(m)'\hat{\mathbf{e}}(m).$$

The least-squares forecast of y_{n+1} from this approximating model is then

$$\hat{f}_{n+1}(m) = \mathbf{x}_{n+1}(m)'\hat{\mathbf{a}}(m). \tag{5}$$

Some forecast combination methods require the calculation of recursive estimates. Stacking the first $t-1$ observations on y_t and $\mathbf{x}_t(m)'$ into the $(t-1)\times 1$ and $(t-1)\times k(m)$ matrices $\mathbf{y}_{t-1}$ and $\mathbf{X}_{t-1}(m)$, the recursive least-squares estimates of $\mathbf{a}(m)$ are

$$\hat{\mathbf{a}}_{t-1}(m) = \left(\mathbf{X}_{t-1}(m)'\mathbf{X}_{t-1}(m)\right)^{-1}\mathbf{X}_{t-1}(m)'\mathbf{y}_{t-1}.$$

The recursive least-squares forecasts of y_t from this approximating model are then

$$\hat{f}_t(m) = \mathbf{x}_t(m)'\hat{\mathbf{a}}_{t-1}(m), \tag{6}$$

the forecast errors are

$$\tilde{e}_t(m) = y_t - \hat{f}_t(m),$$

and an estimate of the forecast error variance is

$$\tilde{\sigma}^2(m) = \frac{1}{P+1}\sum_{t=n-P}^{n}\tilde{e}_t(m)^2 \tag{7}$$

for some integer P such that $\tilde{e}_{n-P}(m)$ is well defined.

3. Forecast combination

Section 2 defined a set of M approximating models and associated forecasts. We now consider combinations of these forecasts. Let $w(m)$ be a weight assigned to the m'th forecast. Define the vectors $\mathbf{w} = (w(1), \ldots, w(M))$ and $\hat{\mathbf{f}}_t = \left(\hat{f}_t(1), \hat{f}_t(2), \ldots, \hat{f}_t(M)\right)'$. The combination forecast of y_{n+1} is

$$\hat{\mathbf{f}}_{n+1}(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{f}}_{n+1} = \sum_{m=1}^{M} w(m)\hat{f}_{n+1}(m) = \mathbf{x}_{n+1}'\hat{\mathbf{a}}(\mathbf{w}) \tag{8}$$

where

$$\hat{\mathbf{a}}(\mathbf{w}) = \sum_{m=1}^{M} w(m)\hat{\mathbf{a}}(m) \tag{9}$$

is an averaging estimator of $\mathbf{a}$. In (8) and (9) we interpret $\hat{\mathbf{a}}(m)$ and $\hat{\mathbf{a}}(\mathbf{w})$ to be infinite-dimensional vectors where $\hat{\mathbf{a}}(m)$ has entries of 0 beyond the m'th element.

Several methods have been proposed for selection of the weight vector $\mathbf{w}$. A classic method introduced by Bates and Granger (1969) sets the weights to be inversely proportional to the estimated forecast error variances defined in (7):

$$w(m) = \frac{\tilde{\sigma}^2(m)^{-1}}{\sum_{j=1}^{M} \tilde{\sigma}^2(j)^{-1}}. \tag{10}$$

This can be viewed as a smoothed version of predictive least squares (Rissanen, 1986), which simply selects the model with smallest forecast variance $\tilde{\sigma}^2(m)$.

Granger and Ramanathan (1984) proposed selecting the weights by minimizing the sum of squared forecast errors from the combination forecast

$$Q(\mathbf{w}) = \sum_{t=n-P}^{n} \left(y_t - \hat{\mathbf{f}}_t'\mathbf{w}\right)^2. \tag{11}$$

The unrestricted minimizer (Granger–Ramanathan's Method A) is the least-squares coefficient

$$\hat{\mathbf{w}} = \left(\sum_{t=n-P}^{n} \hat{\mathbf{f}}_t\hat{\mathbf{f}}_t'\right)^{-1} \sum_{t=n-P}^{n} \hat{\mathbf{f}}_t y_t. \tag{12}$$

As described in Timmermann (2006, Sections 3.2 and 5.2) it may be prudent to minimize (11) subject to the convexity constraints $0 \le w(m) \le 1$ and additivity constraint $\sum_{m=1}^{M} w(m) = 1$. This is

$$\hat{\mathbf{w}} = \underset{0\le w(m)\le 1,\ \sum_{m=1}^{M} w(m)=1}{\operatorname{argmin}} Q(\mathbf{w}). \tag{13}$$

We call the forecast using (13) the constrained Granger–Ramanathan forecast.

The recent forecasting literature has devoted considerable attention to forecasts based on Bayesian Model Averaging (BMA). When the priors are diffuse the BMA weights approximately equal

$$w(m) = \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}(m)\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2}\mathrm{BIC}(j)\right)} \tag{14}$$

where

$$\mathrm{BIC}(m) = n\ln\left(\hat{\sigma}^2(m)\right) + k(m)\ln n$$

is the Bayesian Information Criterion (BIC) for model m. A related proposal (Buckland et al. (1997) and Burnham and Anderson (2002)) is smoothed AIC (SAIC) weights

$$w(m) = \frac{\exp\left(-\frac{1}{2}\mathrm{AIC}(m)\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2}\mathrm{AIC}(j)\right)} \tag{15}$$

where

$$\mathrm{AIC}(m) = n\ln\left(\hat{\sigma}^2(m)\right) + 2k(m)$$

is the Akaike Information Criterion (AIC).

The recent forecasting literature has also suggested the use of very simple combination methods, including the forecast mean which is equivalent to setting $w(m) = 1/M$, and the forecast median.

4. Mallows Model Averaging

In this paper we propose constructing forecast combinations using the weight vector selected by the Mallows Model Averaging (MMA) criterion of Hansen (2007), an extension of the classic Mallows criterion for model selection. The full-sample averaging estimator of the conditional mean μ_t is $\hat{\boldsymbol{\mu}}_t'\mathbf{w} = \hat{\mathbf{a}}(\mathbf{w})'\mathbf{x}_t$ where $\hat{\boldsymbol{\mu}}_t = \left(\hat{\mu}_t(1), \hat{\mu}_t(2), \ldots, \hat{\mu}_t(M)\right)'$ and $\hat{\mu}_t(m) = \mathbf{x}_t(m)'\hat{\mathbf{a}}(m)$. The MMA criterion is the penalized sum of squared residuals

$$\begin{aligned} C_n(\mathbf{w}) &= \sum_{t=1}^{n} \left(y_t - \hat{\boldsymbol{\mu}}_t'\mathbf{w}\right)^2 + 2\sum_{m=1}^{M} w(m)k(m)s^2 \\ &= \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\mathbf{w}'\mathbf{K}s^2 \end{aligned} \tag{16}$$

where $\hat{\mathbf{e}} = \left[\hat{\mathbf{e}}(1), \ldots, \hat{\mathbf{e}}(M)\right]$, $\mathbf{K} = (k(1), \ldots, k(M))'$ and

$$s^2 = \frac{1}{n - k(M)}\hat{\mathbf{e}}(M)'\hat{\mathbf{e}}(M)$$

is an estimate of σ^2 from the largest fitted model. The second equality in (16) uses the assumption that the weights sum to one, which will be imposed.

The MMA weight vector is the value of $\mathbf{w}$ which minimizes $C_n(\mathbf{w})$. Feasible values for $\mathbf{w}$ are weight vectors whose elements are non-negative and sum to one.[1] This set is the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H} = \left\{\mathbf{w} \in [0, 1]^M : \sum_{m=1}^{M} w^m = 1\right\}. \tag{17}$$

The definition of the Mallows weight vector is then

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{H}}{\operatorname{argmin}}\, C_n(\mathbf{w}). \tag{18}$$

The MMA forecasts are $\hat{\mu}_t'\hat{\mathbf{w}} = \hat{\mathbf{a}}(\hat{\mathbf{w}})'\mathbf{x}_t$. Due to the inequality constraints in (17), the solution to (18) is not analytically available for $M > 3$, and must be found numerically using quadratic programming, for which numerical solutions have been thoroughly studied and algorithms are widely available. Even when M is quite large the numerical solution to (18) is computationally fast using any standard algorithm.

[1] It is tempting to consider more weight vectors which do not satisfy these restrictions. However, if the Mallows criterion is used, we must restrict the weight vector to the unit simplex. If the non-negativity restriction is relaxed, the minimized values can be quite ill-behaved and have terrible empirical performance.

5. Mallows criterion and MSE

A traditional motivation for the Mallows criterion is that it is an (approximately) unbiased estimate of the in-sample mean-squared error. Hansen (2007) shows that this property holds for the criterion (16) for iid observations. In this section we develop an asymptotic analog for stationary dependent observations.

Write model (1) in vector notation as $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$. The averaging estimator of $\boldsymbol{\mu}$ given the weight vector $\mathbf{w}$ is

$$\hat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{j=1}^{M} w(m)\mathbf{X}(m)\hat{\mathbf{a}}(m) = \mathbf{P}(\mathbf{w})\,\mathbf{y}$$

where

$$\mathbf{P}(\mathbf{w}) = \sum_{j=1}^{M} w(m)\mathbf{X}(m)\left(\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\mathbf{X}(m)'.$$

The averaging residual vector is $\hat{\mathbf{e}}(\mathbf{w}) = \mathbf{y} - \hat{\boldsymbol{\mu}}(\mathbf{w})$.

Define the in-sample mean-squared error

$$\begin{aligned} L_n(\mathbf{w}) &= E\left(\frac{1}{n}\sum_{t=1}^{n}\left(\mu_t - \hat{\mu}_t(\mathbf{w})\right)^2\right) \\ &= \frac{1}{n}E\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right). \end{aligned} \tag{19}$$

This is a summary measure of the fit of the averaging estimator.

Computing the sum of squared errors and expanding the square we find

$$\begin{aligned} \hat{\mathbf{e}}(\mathbf{w})'\hat{\mathbf{e}}(\mathbf{w}) &= \left(\mathbf{e} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\mathbf{e} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right) \\ &= \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right) + \mathbf{e}'\mathbf{e} \\ &\quad + 2\mathbf{e}'\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right) \\ &= \left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right) + \mathbf{e}'\mathbf{e} \\ &\quad + 2\mathbf{e}'\left(\mathbf{I} - \mathbf{P}(\mathbf{w})\right)\boldsymbol{\mu} - 2\mathbf{e}'\mathbf{P}(\mathbf{w})\,\mathbf{e}. \end{aligned}$$

Thus the expectation for the Mallows Criterion is

$$\begin{aligned} E\left(C_n(\mathbf{w})\right) &= E\left(\hat{\mathbf{e}}(\mathbf{w})'\hat{\mathbf{e}}(\mathbf{w}) + 2\sum_{m=1}^{M} w(m)k(m)s^2\right) \\ &\simeq n\left(L_n(\mathbf{w}) + \sigma^2\right) - 2\left(E\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\,\mathbf{e}\right) - \sigma^2\sum_{m=1}^{M} w(m)k(m)\right). \end{aligned}$$

The final term is asymptotically zero, as we now show.

Theorem 1. *If $(\mathbf{x}_t, e_t)$ is strictly stationary and ergodic, $E\mathbf{x}_t'\mathbf{x}_t < \infty$, and M and $\mathbf{w}$ are fixed as $n \to \infty$, then*

$$E\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\,\mathbf{e}\right) \longrightarrow \sigma^2\sum_{m=1}^{M} w(m)k(m).$$

Using Theorem 1, we see that $E\left(C_n(\mathbf{w})\right) \simeq n\left(L_n(\mathbf{w}) + \sigma^2\right)$. Thus the expectation of the Mallows criterion is asymptotically equivalent to the squared error (19). (The additive and multiplicative constants do not matter.) We see that this classic property of the Mallows criterion extends to model averaging with stationary dependent data.

6. MSE and MSFE

In this section we show that the mean-squared forecast error (MSFE) approximately equals $L_n(\mathbf{w})$ when the observations are strictly stationary, and thus the Mallows criterion can also be seen as an approximately unbiased estimate of the MSFE. The one-step-ahead out-of-sample forecast of y_{n+1} given $\mathbf{x}_{n+1}$ is

$$\hat{y}_{n+1} = \hat{\mathbf{a}}(\mathbf{w})'\mathbf{x}_{n+1}$$

so the one-step-ahead second-order MSFE is

$$\begin{aligned} R_n(\mathbf{w}) &= E\left(y_{n+1} - \hat{y}_{n+1}\right)^2 - \sigma^2 \\ &= E\left(e_{n+1} + \left(\mathbf{a} - \hat{\mathbf{a}}(\mathbf{w})\right)'\mathbf{x}_{n+1}\right)^2 - \sigma^2 \\ &= E\left(\left(\mathbf{a} - \hat{\mathbf{a}}(\mathbf{w})\right)'\mathbf{x}_{n+1}\right)^2 \\ &\simeq E\left(\left(\mathbf{a} - \hat{\mathbf{a}}(\mathbf{w})\right)'\mathbf{x}_t\right)^2 \\ &= E\left(\mu_t - \hat{\mu}_t(\mathbf{w})\right)^2 \\ &= L_n(\mathbf{w}). \end{aligned}$$

The approximation in the fourth line is valid for stationary observations due to the approximate independence of $\mathbf{x}_t$ and $\hat{\mathbf{a}}(\mathbf{w})$ in large samples (and similarly for $\mathbf{x}_{n+1}$ and $\hat{\mathbf{a}}(\mathbf{w})$).

Combined with Theorem 1, we deduce that the Mallows criterion is an approximately unbiased estimate of the MSFE $R_n(\mathbf{w})$.

7. Asymptotic efficiency of MMA

The previous sections showed that the Mallows criterion is an asymptotically unbiased estimator of the MSE $L_n(\mathbf{w})$ and MSFE $R_n(\mathbf{w})$ for strictly stationary observations. This suggests that the weights which minimize the Mallows criterion may be asymptotically optimal with respect to this criterion. We have not been able to establish this result for dependent data, but building on the work of Li (1987) for efficient model selection procedures, Hansen (2007) established this asymptotic optimality property when the observations are independent. We report this result here for completeness.

For some finite integer N let the weights $w(m)$ be restricted to the set $\{0, \frac{1}{N-1}, \frac{2}{N-1}, \ldots, 1\}$, let $\mathcal{H}^*$ be the subset of $\mathcal{H}$ restricted to this set of weights, and let $\hat{\mathbf{w}}^*$ be the MMA weights selected when restricted to $\mathcal{H}^*$:

$$\hat{\mathbf{w}}^* = \underset{\mathbf{w}\in\mathcal{H}^*}{\operatorname{argmin}}\, C_n(\mathbf{w}).$$

Theorem 2. *[(Hansen, 2007)] If $(y_t, \mathbf{x}_t)$ are iid, satisfy (1)–(4),*

$$\sup_t E\left(e_t^{4N} \mid \mathbf{x}_t\right) < \infty, \tag{A.1}$$

$$\inf_{\mathbf{w}\in\mathcal{H}} E\left(\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\mathbf{w})\right) \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right) \to \infty, \tag{A.2}$$

and $k(M)/n \to 0$, then the MMA weights $\hat{\mathbf{w}}^*$ are optimal in the sense that

$$\frac{L_n\left(\hat{\mathbf{w}}^*\right)}{\inf_{\mathbf{w}\in\mathcal{H}^*} L_n(\mathbf{w})} \to_p 1$$

as $n \to \infty$.

Theorem 2 shows that MMA is asymptotically efficient in the sense that its in-sample MSE is equivalent to the infeasible optimal averaging estimator. The assumptions are fairly minimal. Assumption (A.1) is a strong moment bound but may be a technical artifact of the proof method. The restriction of $\mathcal{H}$ to $\mathcal{H}^*$ also appears to be a technical artifact of the proof method.[2] However, Assumption (A.2) is important. It states that there is no finite m for which $\mu_t(m)$ equals μ_t—equivalently, that all finite-dimensional models are approximations. This is an important condition for it is well known that when the true model is finite dimensional, then this class of model selection procedures is asymptotically inefficient.

Theorem 2 excludes dependent data and is therefore not directly relevant for time-series forecasting. However, there is a closely related theory of optimal model selection for one-step-ahead forecasting in the context of autoregressive models, and we expect that the optimality property should carry over to forecast combination. The most relevant contribution is by Ing and Wei (2005) who consider stationary infinite-order autoregressions with iid innovations. Let $\hat{y}_{n+1}(m)$ be the one-step-ahead forecast of y_{n+1} using the least-squares estimate of an autoregressive (AR) model of order m and let $R_n(m)$ denote the second-order mean-squared forecast error from these estimates:

$$R_n(m) = E\left(y_{n+1} - \hat{y}_{n+1}(m)\right)^2 - \sigma^2.$$

Let $\hat{m}$ be the AR order selected by the Akaike or Mallows criterion. Ing and Wei (2005) show that $\hat{m}$ is asymptotically optimal in the sense that

$$\frac{R_n\left(\hat{m}\right)}{\inf_{1\le m\le M} R_n(m)} \to_p 1. \tag{20}$$

Model selection is a restricted version of model averaging. Given the similarity with Theorem 2 we conjecture that the optimality result (20) will extend to the case of general weight vectors.

However, the technical challenges for establishing this result are formidable. It is necessary to extend Lemmas 1, 2 and 3 of Ing and Wei (2005) to the case of averaging estimators. (Ideally the results would resemble Theorem 2 of Whittle (1960).) Such an extension is beyond the scope of the present paper and is left for future research.

8. Finite sample investigation

We now investigate the finite sample MSFE of the our model average estimator in two simulation designs. The first is the regression model

$$y_t = \theta_0 + \sum_{k=1}^{K} \theta_k x_{kt} + e_t.$$

The error $e_t \sim N(0, \sigma^2)$ is independent of the regressors x_{kt}, which are independent AR(1) processes $x_{kt} = \rho x_{kt-1} + u_{kt}$ where $\rho = 0.5$ and $u_{kt} \sim N(0, 0.75)$. We normalize $\sigma^2 = 1$ and set the regression coefficients by the rule

$$\theta_k = c\gamma_k$$
$$\gamma_k = \frac{k^\alpha \beta^k}{\sum_{j=1}^{K} j^{2\alpha}\beta^{2j}}.$$

By varying α and β a variety of patterns for θ_k as a function of k can be determined. Eight representative patterns were selected, and the coefficients γ_k are displayed in the left panels of Figs. 1 and 2. In the first six panels the coefficients γ_k are monotonically decreasing in k, as would be expected if the regressors were correctly ordered. The final two panels display cases of non-monotonic coefficients.

The parameter c was set by the rule $c = \sqrt{R^2/(1-R^2)}$ and R^2 varied on a grid from 0.1 to 0.9, as this is the population R^2 of this regression. We set the sample size to $n = 200$ and set $K = 12$. Forecasts of y_{n+1} given x_{n+1} are based on the linear regressions:

$$y_t = \hat{\theta}_0 + \sum_{k=1}^{m} \hat{\theta}_k x_{kt} + \hat{e}_t(m)$$

for $m = 0, \ldots, 12$, estimated by OLS.[3]

The second design is the moving average model

$$y_t = \sum_{k=0}^{\infty} \theta_k e_{t-k}$$

where e_t is iid $N(0, \sigma^2)$ and $n = 200$. We normalize $\sigma^2 = 1$ and set the coefficients using the rule

$$\theta_k = (1+k)^\alpha \beta^k.$$

By varying α and β a variety of moving average patterns can be generated. We varied α among {0, 0.25, 0.50 and 1.0} and varied β on a grid from .6 to .9. The moving average coefficient patterns are displayed in the four left panels of Fig. 3.

Forecasts of y_{n+1} are based on AR(m) models:

$$y_t = \hat{\mu} + \hat{\alpha}_1 y_{t-1} + \cdots + \hat{\alpha}_m y_{t-m} + \hat{e}_t(m)$$

for $m = 0, \ldots, 12$, estimated by OLS.

In both experiments a wide set of forecast combination methods were compared, including AIC selection, smoothed AIC (SAIC) (15), BIC selection, weighted BIC (WBIC) (14), PLS selection, median forecast, mean forecast, Bates–Granger combination (10), Granger–Ramanathan combination (12), Constrained Granger–Ramanathan combination (13), and MMA combination (18).

We compare the forecasting methods based on out-of-sample second-order mean-square forecast error (MSFE)

$$\text{MSFE} = \frac{n}{\sigma^2}\left(E\left(y_{n+1} - \hat{y}_{n+1}\right)^2 - \sigma^2\right).$$

The error variance σ^2 is subtracted because it is the leading term in the MSFE, and is common across forecast methods. The scaling (n/σ^2) is used to render the results scale-free. We compute MSFE by computing averages across 20,000 simulation draws.

Comparing the forecast methods across the parameter settings, we find that eight of the methods are dominated (either uniformly or nearly so) by one of the three other methods. Specifically, BIC selection is dominated by WBIC, AIC by SAIC and SAIC by MMA. The mean and median forecasts are dominated by the Bates–Granger combination. Granger–Ramanathan is strongly dominated by constrained Granger–Ramanathan and PLS, which are in turn dominated by MMA. To keep our graphs uncluttered, only the three undominated methods (WBIC, Bates–Granger, and MMA) are displayed.

For each experiment, MSFE is displayed in the right panels of Figs. 1–3. (The left panels display the coefficient pattern.) Figs. 1 and 2 display results for the first simulation design (the regression model) and Fig. 3 displays results for the second design (the

[2] It is highly desirable, but technically quite challenging, to relax this restriction. It is important to emphasize that this restriction is only relevant for the optimality theory, not for empirical application of the method.

[3] To assess robustness to misspecification, the simulation was also done with the regression models restricted to $m = 0, \ldots, 6$. The results were qualitatively unchanged.

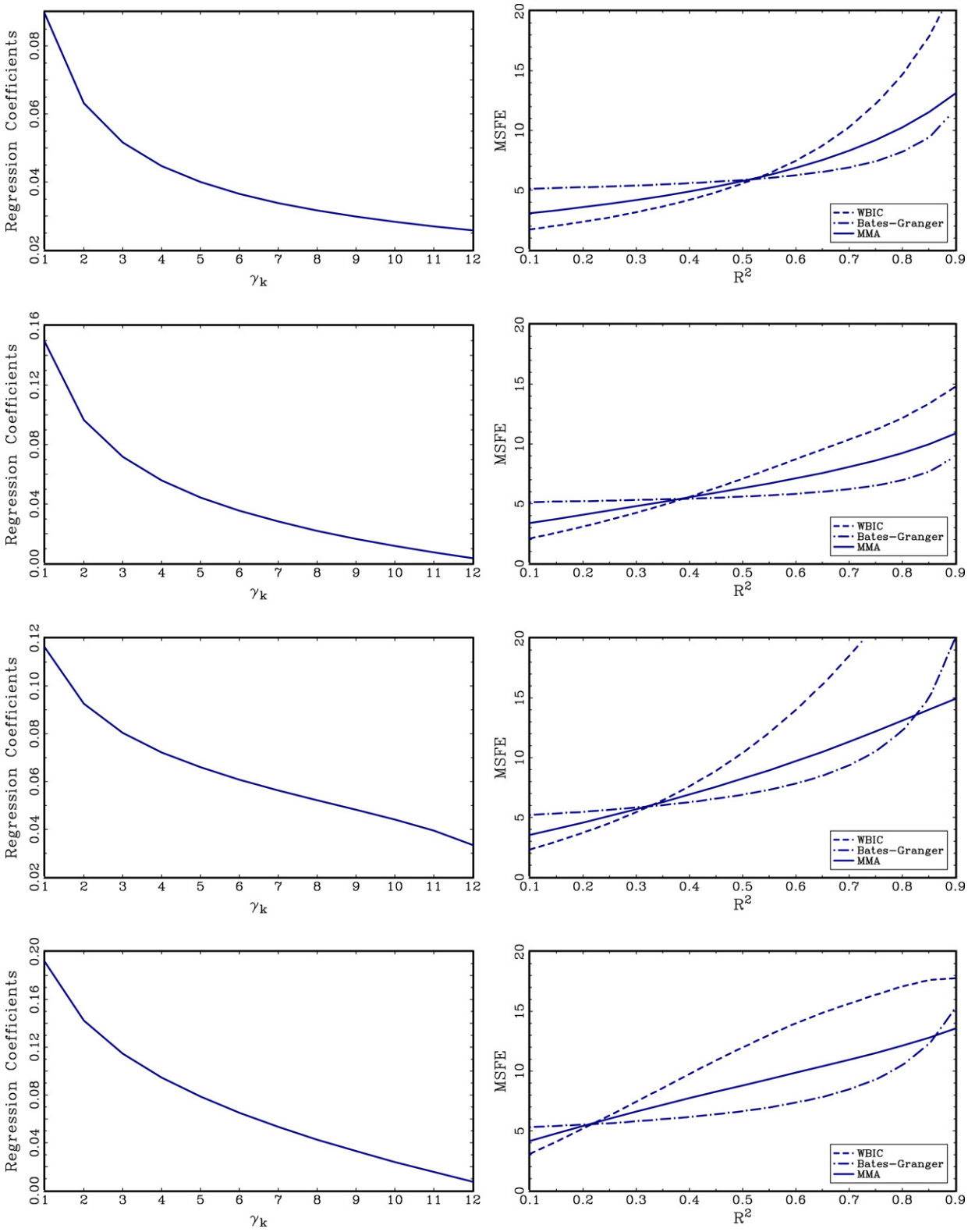

Fig. 1. Regression model, cases 1–4.

moving average model). In Figs. 1 and 2 the coefficient pattern γ_k is displayed in the left panel, and the right panel displays MSFE as R^2 is varied from 0.1 to 0.9. In Fig. 3 the coefficient α is held fixed at one of the values $\{0, 0.25, 0.50$ and $1.0\}$, and the right panel displays MSFE as β is varied from 0.6 to 0.9.

In all plots, the MSFE of WBIC forecasts is displayed with the dashed line, the MSFE of Bates–Granger forecasts with the dash-dotted line, and the MSFE of MMA forecasts with the solid line. Examining the twelve panels in the three figures, there is no one method which uniformly dominates the others.