Multi-Step Forecast Model Selection

Bruce E. Hansen

April 2010
Preliminary

Abstract

This paper examines model selection and combination in the context of multi-step linear forecasting. We start by investigating multi-step mean squared forecast error (MSFE). We derive the bias of the in-sample sum of squared residuals as an estimator of the MSFE. We find that the bias is not generically a scale of the number of parameters, in contrast to the one-step-ahead forecasting case. Instead, the bias depends on the long-run variance of the forecast model in analogy to the covariance matrix of multi-step forecast regressions, as found by Hansen and Hodrick (1980). In consequence, standard information criterion (Akaike, FPE, Mallows and leave-one-out cross-validation) are biased estimators of the MSFE in multi-step forecast models. These criteria are generally under-penalizing for over-parameterization and this discrepancy is increasing in the forecast horizon. In contrast, we show that the leave-h-out cross validation criterion is an approximately unbiased estimator of the MSFE and is thus a suitable criterion for model selection. Leave-h-out is also suitable for selection of model weights for forecast combination.

JEL Classification: C52, C53
Keywords: Mallows, AIC, information criterion, cross-validation, forecast combination, model selection

1 Introduction

Model selection has a long history in statistics and econometrics, and the methods are routinely applied for forecast selection. The most important theoretical contributions are those of Shibata (1980) and Ing and Wei (2005). Shibata (1980) studied an infinite-order autoregression with homoskedastic errors, and showed that models selected by the final prediction criterion (FPE) or the Akaike information criterion (AIC) are asymptotically efficient in the sense of asymptoticaly minimizing the mean-squared forecast error, when independent samples are used for estimation and for forecasting. Ing and Wei (2005) extended Shibata's analysis to the case where the same data is used for estimation and forecasting. These papers provide the foundation for the recommendation of the use of FPE, AIC or their asymptotic equivalents (including Mallows and leave-one-out cross-valdation) for forecast model selection.

In this paper we investigate the appropriateness of these information criterion in multi-step forecasting. We adopt multi-step mean squared forecast error (MSFE) as our measure of risk, and set our goal to develop an approximately unbiased estimator of the MSFE. Using conventional methods, we show that the MSFE is approximately the expected sample sum of squared residuals plus a penalty which is a function of the long-run covariance matrix.

In the case of one-step forecasting with homoskedastic errors, this penalty simplifies to twice the number of parameters multiplied by the error variance. This is the classic justification for why classic information criteria (AIC and its asymptotic equivalents) are approximately unbiased for the MSFE.

In the case of multi-step forecasting, however, the fact that the errors have overlapping dependence means that the correct penalty does not simplify to a scale of the number of parameters. This implies that the penalties used by classic information criteria are incorrect. The situation is identical to that faced in inference in forecasting regressions with overlapping error dependence, as pointed out by Hansen and Hodrick (1980). The overlapping dependence of multi-step forecast errors invalidates the information matrix equality. This affects information criteria as well as covariance matrices.

Our finding and proposed adjustment are reminiscent to the work of Takeuchi (1976), who investigated model selection in the context of likelihood estimation with possibly misspecified models. Takeuchi showed that the violation of the information matrix equality due to misspecification renders the Akaike information criterion inappropriate, and that the correct parameterization penality depends on the matrices appearing in the robust covariance matrix estimator. Unfortunately, Takeuchi's precient work has had little impact on empirical model selection practice.

We investigate the magnitude of this discrepancy in a simple model and show that the distortion depends on the degree of serial dependence, and in the extreme case of high dependence the correct penalty is h times the classic penalty, where h is the forecast horizon.

Once the correct penalty is understood, it possible to construct information criteria which are approximately unbiased for the MSFE. Our preferred criterion is the leave-h-out cross-validation criterion. We show that it is an approximately unbiased estimator of the MSFE. It works well in

practice, and is conceptually convenient as it does not require penalization or HAC estimation.

Interestingly, our results may not be in conflict with the classic optimality theory of Shibata (1980) and Ing and Wei (2005). As these authors investigated asymptotic optimality in an infinite-order autogression, in large samples the information criterion are comparing estimated AR(k) and AR($k+1$) models where k is tending to infinity. As the coefficient on the $k+1$'st autoregressive lag is small (and tends to zero as k tends to infinity), this is a context where the correct information penalty is classical and proportionate to the number of parameters. As the asymptotic optimality theory focuses on the selection of models with a large and increasing number of parameters, the distortion in the penalty discussed above may be irrelevant.

While many information criteria for model selection have been introduced, the most important are those of Akaike (1969, 1973), Mallows (1973), Takeuchi (1976), Schwarz (1978) and Rissanen (1986). The asymptotic optimality of the Mallows criterion in infinite-order homoskedastic linear regression models was demonstrated by Li (1987). The optimality of the Akaike criterion for optimal forecasting in infinite-order homoskedastic autoregressive models was shown by Shibata (1980), and extended by Banasali (1996), Lee and Karagrigoriou (2001), Ing (2003, 2004, 2007), and Ing and Wei (2003, 2005).

In addition to forecast selection we consider weight selection for forecast combination. 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). Stock and Watson (1999, 2004, 2005) have provided detailed empirical evidence demonstrating the gains in forecast accuracy through forecast combination. Hansen (2007) developed the Mallows criterion for weight selection in linear regression, and was shown to apply to one-step-ahead forecast combination by Hansen (2008). Hansen and Racine (2009) developed weight selection for model averaging using a leave-one-out criterion. In this paper we recommend the leave-h-out criterion for selection of weights for multi-step forecasting.

2 Model

Consider the h-step-ahead forecasting model

$$\begin{aligned} y_t &= \boldsymbol{x}_{t-h}'\boldsymbol{\beta} + e_t \\ E\left(\boldsymbol{x}_{t-h}e_t\right) &= 0 \\ \sigma^2 &= Ee_t^2 \end{aligned} \tag{1}$$

where $\boldsymbol{x}_{t-h}$ is $k \times 1$ and contains variables dated h periods before y_t. The variables (y_t, x_{t-h}) are observed for $t = 1, ..., n$, and the goal is to forecast y_{n+h} given x_n.

In general, the error e_t is a MA(h-1) process. For example, if y_t is an AR(1)

$$y_t = \alpha y_{t-1} + u_t \tag{2}$$

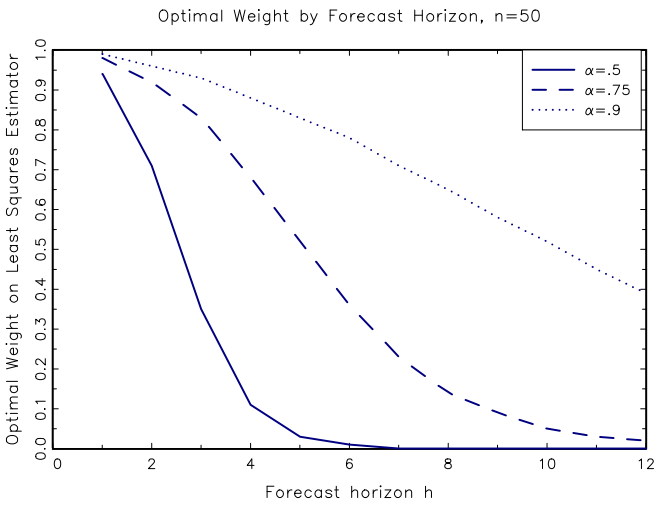

Figure 1: Optimal Weight by Forecast Horizon

with u_t iid and $Eu_t = 0$, then the optimal h-step-ahead forecast takes the form (1) with $x_{t-h} = y_{t-h}$, $\beta = \alpha^h$ and

$$e_t = u_t + \alpha u_{t-1} + \cdots + \alpha^{h-1} u_{t-h+1}$$

which is an MA(h-1) process

The forecast horizon affects the optimal choice of forecasting model. For example, suppose again that y_t is generated by (2). Let $\hat{\beta}_{LS}$ be the least-squares estimate[1] of β in (1) and consider the class of model average forecasts

$$\hat{y}_{n+h|n}(w) = w x_n \hat{\beta}_{LS} \tag{3}$$

where $w \in [0, 1]$. This is a weighted average of the unconstrained least-squares forecast $\hat{y}_{n+h|n} = x_n \hat{\beta}_{LS}$ and the constrained forecast $\tilde{y}_{n+h|n} = 0$. The mean-square forecast error of (3) is

$$MSFE(w) = E\left(y_{n+h} - \hat{y}_{n+h|n}(w)\right)^2.$$

The optimal weight w minimizes this expression, and varies with h, α and n. Using 100,000 simulation replications, the MSFE was calculated for $n = 50$. The optimal weight is displayed in Figure 1 as a function of of h for several values of α. We can see that the optimal weight w declines with forecast horizon, and for any horizon h the optimal weight w is increasing in the autoregressive parameter α. For the one-step-ahead forecast ($h = 1$), the optimal weight on the least-squares estimate is close to 1.0 for all values of α shown, but for the 12-step-ahead forecast, the optimal weight is close to zero for all but the largest values of α.

[1] Qualitatively similar results are obtained if we replace $\hat{\beta}_{LS}$ with $\hat{\alpha}^h$ where $\hat{\alpha}$ is the least-squares estimate of α from the AR(1) (2).

3 Forecast Selection

Using observations $t = 1, ..., n$, the forecasting equation (1) is estimated by least-squares, which we can write

$$\begin{aligned} y_t &= \boldsymbol{x}_{t-h}'\hat{\boldsymbol{\beta}} + \hat{e}_t \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n} \hat{e}_t^2 \end{aligned} \tag{4}$$

and is used to construct the out-of-sample forecast

$$\hat{y}_{n+h|n} = \boldsymbol{x}_n'\hat{\boldsymbol{\beta}}. \tag{5}$$

The MSFE of the forecast is

$$MSFE = E\left(y_{n+h} - \hat{y}_{n+h|n}\right)^2. \tag{6}$$

The goal is to select a forecasting model with low MSFE.

A common information criterion for model selection is the Akaike information criterion (AIC)

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{n}.$$

A similar criterion (but robust to heteroskedasticity) is leave–one-out cross-validation

$$CV_1 = \frac{1}{n}\sum_{t=1}^{n} \tilde{e}_{t,1}^2$$

where $\tilde{e}_{t,1}(m)$ is the residual obtained by least-squares estimation with the observation t omitted.

It turns out that these criteria are generally inappropriate for multi-step forecasting due to the moving average structure of the forecast error e_t. Instead, we recommend forecast selection based on the leave-h-out cross-validation criterion

$$CV_h = \frac{1}{n}\sum_{t=1}^{n} \tilde{e}_{t,h}^2 \tag{7}$$

where $\tilde{e}_{t,h}$ is the residual obtained by least-squares estimation with the $2h+1$ observations $\{t - h + 1, ..., t + h - 1\}$ omitted.

4 Forecast Combination

Suppose that there are M forecasting models indexed by m, where the m'th model has $k(m)$ regressors, residuals $\hat{e}_t(m)$, variance estimate $\hat{\sigma}^2(m)$ and forecast $\hat{y}_{n+h|n}(m)$. We want to select a

set of weights $w(m)$ to make a forecast combination

$$\hat{y}_{n+h|n} = \sum_{m=1}^{M} w(m)\hat{y}_{n+h|n}(m).$$

To minimize the MSFE of one-step-ahead forecasts, Hansen (2008) proposed forecast model averaging (FMA). This selects the weights $w(m)$ to minimize the Mallows criterion

$$FMA(w) = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=1}^{M} w(m)\hat{e}_t(m)\right)^2 + 2\hat{\sigma}^2\sum_{m=1}^{M} w(m)k(m)$$

where $\hat{\sigma}^2$ is a preliminary estimate of σ^2. Hansen and Racine (2009) proposed Jackknife model averaging (JMA) which selects weights $w(m)$ to minimize the leave-one-out cross-validation criterion

$$CV_1MA(w) = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=1}^{M} w(m)\tilde{e}_{t,1}(m)\right)^2$$

where $\tilde{e}_{t,1}(m)$ is the residual obtained by least-squares estimation with observations t omitted. This is similar to FMA but is robust to heteroskedasticity. These criteria are appropriate for one-step-ahead forecast combination as they are approximately unbiased estimates of the MSFE.

In the case of multiple-step forecasting these criterion are not appropriate. Instead, we recommend selecting the weights $w(m)$ to minimize the leave-h-out cross-validation criterion

$$CV_hMA(w) = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=1}^{M} w(m)\tilde{e}_{t,h}(m)\right)^2.$$

5 Illustrations

To illustrate the difference, Figure 2 displays the MSFE of five estimators in the context of model (1) for $n = 50$ and $h = 4$: The unconstrained least-squares estimator, the selection estimators based on CV_1 and CV_h, and the combination estimates based on CV_1MA and CV_hMA. The data are generated by the equation (1) with $k = 8$, the first regressor an intercept and the remaining regressors normal AR(1)'s with coefficients 0.9, and setting $\beta = (\mu, 0, ..., 0)$. The regression error e_t is a normal MA(h-1) with equal coefficients and normalized to have unit variance. The selection and combination estimators are constructed from two base model estimators: (i) unconstrained least-squares estimation, and (ii) $\beta = (0, 0, ..., 0)$. The MSFE of the five estimators are displayed as a function of μ, and are normalized by the MSFE of the unconstrained least-squares estimator.

We can see a large difference in performance of the estimators. The estimator with the uniformly lowest MSFE (across μ) is the leave-h-out combination estimator CV_hMA, and the difference in MSFE is substantial.

Figures 3, 4, and 5 display the MSFE of the same estimators when the data are generated

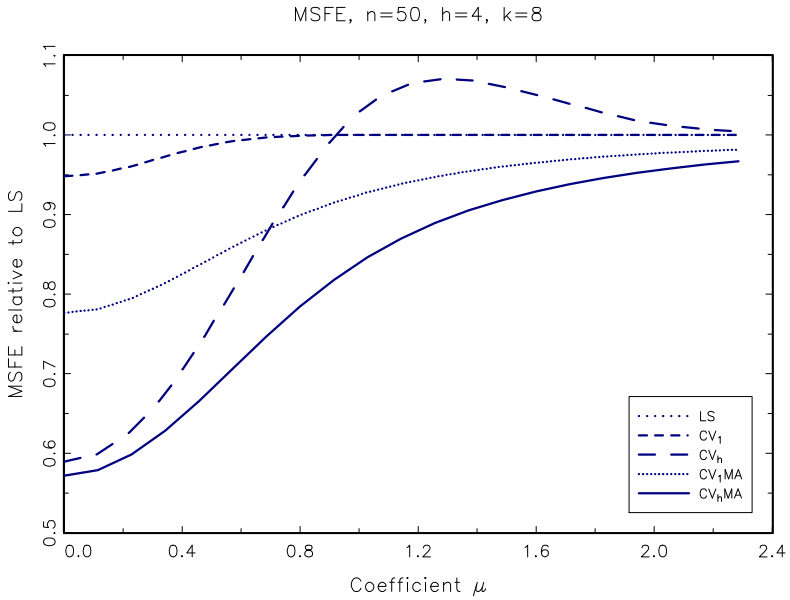

Figure 2: MSFE of 4-step-ahead forecast as a function of the coefficient μ

by an AR(1) (2) with coefficient α, for different forecast horizons h. The forecasting equation is a regression of y_t on an intercept and three lags of y_t : $y_{t-h}, y_{t-h-1}, ..., y_{t-h-3}$ ($k = 4$ regressors). The two base model estimators are: (i) unconstrained least-squares estimation, and (ii) $\beta = (0, 0, 0, 0)$. The sample size again is $n = 50$. Figure 3 displays the MSFE for $h = 4$, Figure 4 for $h = 8$ and Figure 5 for $h = 12$, and the MSFE is displayed as a function of the autoregressive coefficient α and is normalized by the MSFE of the unconstrained least-squares estimator. In nearly all cases the leave-h-out combination estimator CV_hMA has the lowest MSFE, with the only exception $h = 4$ for large α. For $h = 4$ the difference between the estimators is less pronounced, but for large h and α the difference between the MSFE of the h-step criteria estimators CV_hMA and CV_h and the 1-step estimators CV_1MA and CV_1 is quite large.

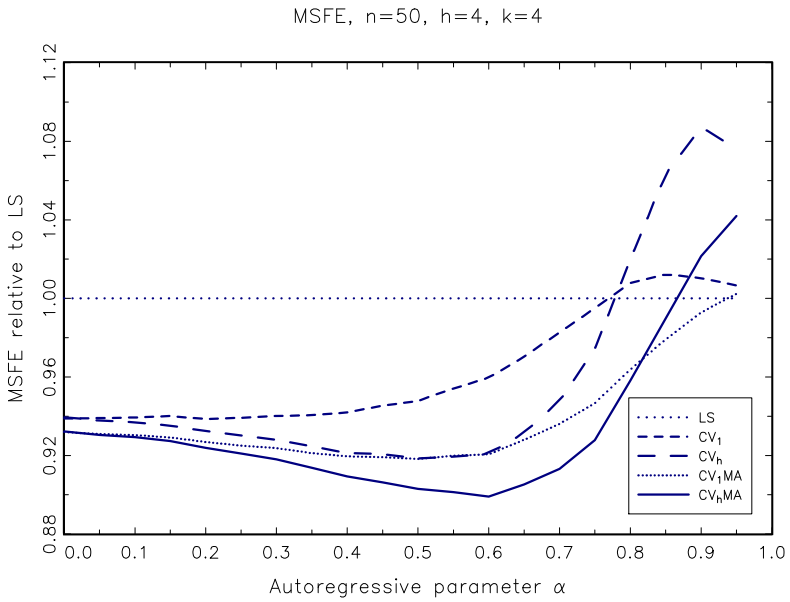

Figure 3: 4-step-ahead MSFE for AR(1) process

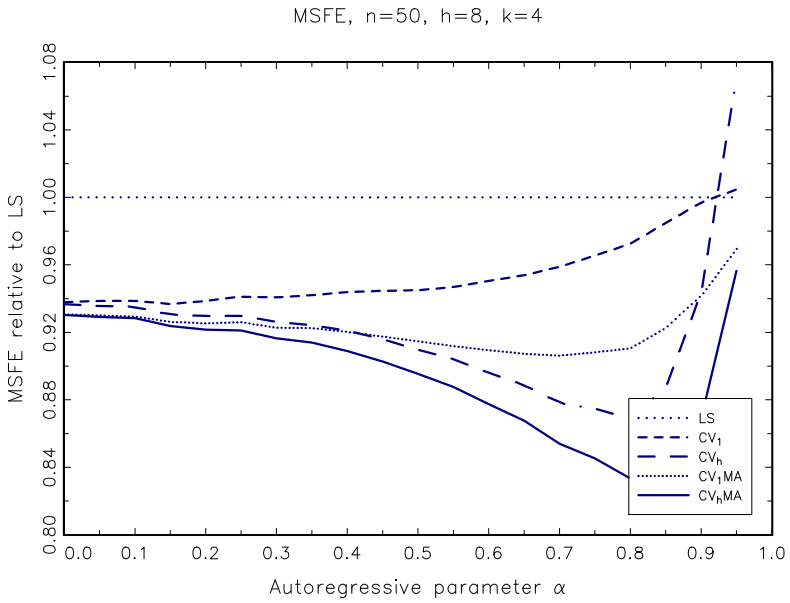

Figure 4: 8-step-ahead MSFE for AR(1) process

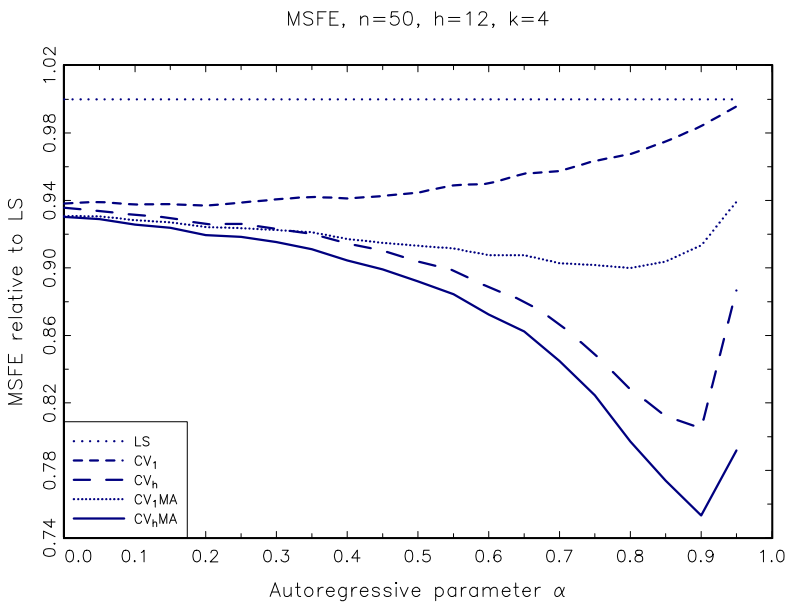

Figure 5: 12-step-ahead MSFE for AR(1) process

6 MSFE

We now develop a theory to justify the recommendations of the previous sections.

A common measure of forecast accuracy is the mean-square forecast error (MSFE) defined in (6). The basis for our theory is the following representation of the MSFE.

Theorem 1

$$MSFE = E\left(\hat{\sigma}^2\right) + \frac{2\sigma^2 B}{n} + O\left(n^{-3/2}\right) \tag{8}$$

where $\hat{\sigma}^2$ *is from (4) and*

$$\begin{aligned} B &= \sigma^{-2}\operatorname{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right) \\ \boldsymbol{Q} &= E\left(\boldsymbol{x}_{t-h}\boldsymbol{x}_{t-h}'\right) \\ \boldsymbol{\Omega} &= \sum_{j=-(h-1)}^{h-1} E\left(\boldsymbol{x}_{t-h-j}e_{t-j}\boldsymbol{x}_{t-h}'e_t\right). \end{aligned} \tag{9}$$

Therorem 1 shows that the sum of square errors is a biased estimate of the MSFE with the bias determined by the constant B. The constant B is a function of the matrix $\boldsymbol{\Omega}$ which appears in the covariance matrix for the parameter estimates $\hat{\boldsymbol{\beta}}$ as is standard for estimation with overlapping error dependence, as shown by Hansen and Hodrick (1980).

7 Conditionally Homoskedastic One-Step-Ahead Forecasting

Suppose that $h = 1$ and the forecast error is a conditionally homoskedastic martingale difference:

$$E\left(e_t \mid \Im_{t-1}\right) = 0$$
$$E\left(e_t^2 \mid \Im_{t-1}\right) = \sigma^2.$$

In this case

$$\boldsymbol{\Omega} = \boldsymbol{Q}\sigma^2 \tag{10}$$

so that the bias term (9) takes the simple form

$$B = k. \tag{11}$$

In this case Theorem 1 implies

$$MSFE = E\left(\hat{\sigma}^2\right) + \frac{2k\sigma^2}{n}.$$

This shows that in a homoskedastic MDS forecasting equation, the Mallows criterion

$$C = \hat{\sigma}^2 + \frac{2k\tilde{\sigma}^2}{n}$$

(where $\tilde{\sigma}^2$ is a preliminary consistent estimate of σ^2), Akaike's final prediction error (FPE) criterion

$$FPE = \hat{\sigma}^2\left(1 + \frac{2k}{n}\right)$$

and the exponential of the Akaike information criterion (AIC)

$$\exp\left(AIC\right) = \hat{\sigma}^2 \exp\left(\frac{2k}{n}\right) \simeq FPE$$

are all approximately unbiased estimators of MSFE.

These three information criteria essentially use the approximation $B \simeq k$ to construct the parameterization penalty, which is appropriate for one-step homoskedastic forecasting due to the information-matrix equality (10). However, (10) generally fails for multi-step forecasting as pointed out by Hansen and Hodrick (1980). When (10) fails, then AIC, FPE, and Mallows are biased information criteria.

7.1 Criterion Distortion in h-step Forecasting

We have shown that classic information criteria estimate the bias of the sum of squared errors using the approximation $B \simeq k$ which is valid for one-step homoskedastic forecasting but generally invalid for multi-step forecasting. Instead, the correct penalty is proportional to $B = \sigma^{-2}\operatorname{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right)$.

What is the degree of distortion due to the use of the traditional penality k rather than the correct penalty B? In this section we explore this question with a simple example.

Suppose that (1) holds with $\boldsymbol{x}_{t-h}$ and e_t mutually independent. In this case

$$B = k + 2\sum_{j=1}^{h-1} \operatorname{tr}\left(\left(E(\boldsymbol{x}_t, \boldsymbol{x}_t')\right)^{-1} E(\boldsymbol{x}_t, \boldsymbol{x}_{t-j})\right) \operatorname{corr}(e_t, e_{t-j}).$$

Notice that if both $\boldsymbol{x}_t$ and e_t are positively serially correlated, then $B > k$. In this sense, we see that generally the conventional approximation $B \simeq k$ is an underestimate of the correct penalty.

To be more specific, suppose that the elements of $\boldsymbol{x}_t$ are independent AR(1) processes with coefficient ρ, in which case

$$\operatorname{tr}\left(\left(E(\boldsymbol{x}_t, \boldsymbol{x}_t')\right)^{-1} E(\boldsymbol{x}_t, \boldsymbol{x}_{t-j})\right) = k\rho^j$$

and

$$B = k\left(1 + 2\sum_{j=1}^{h-1} \rho^j \operatorname{corr}(e_t, e_{t-j})\right).$$

Furthermore, suppose that e_t is a MA(h-1) with equal coefficients. Then

$$\operatorname{corr}(e_t, e_{t-j}) = 1 - \frac{j}{h}$$

and

$$B = k + 2k\sum_{j=1}^{h-1} \rho^j \left(1 - \frac{j}{h}\right).$$

This is increasing with ρ. The limit as $\rho \to 1$ is

$$\lim_{\rho \to 1} B = kh.$$

We have found that in this simple setting the correct penalty kh is proportional to both the number of parameters k and the forecast horizon h. The classic penalty k is off by a factor of h, and puts too small a penalty on the number of parameters. Consequently, classic information criteria will over-select for multi-step forecasting.

8 Leave-h-out Cross Validation

Our second major contribution is a demonstration that the leave-h-out CV criterion is an approximately unbiased estimate of the MSFE.

The leave-h-out estimator of β for observation t in regression (1) is

$$\tilde{\boldsymbol{\beta}}_{t,h} = \left(\sum_{|j-t|\geq h} \boldsymbol{x}_{j-h} \boldsymbol{x}'_{j-h} \right)^{-1} \left(\sum_{|j-t|\geq h} \boldsymbol{x}_{j-h} y_j \right)$$

where the summation is over all observations except the $2h+1$ observations $\{t-h+1, ..., t+h-1\}$ omitted. In other words, leaving out observations within $h-1$ periods of the time period t. The leave-h-out residual is

$$\tilde{e}_{t,h} = y_t - \tilde{\boldsymbol{\beta}}'_{t,h} \boldsymbol{x}_{t-h}$$

and the leave-h-out cross-validation criterion is

$$CV_h = \frac{1}{n} \sum_{t=1}^{n} \tilde{e}^2_{t,h}.$$

Theorem 2 $E\left(CV_h\right) = MSFE + o(1)$

Theorem 2 shows that leave-h-out CV is an appropriate h-step forecast selection criterion. This is in contrast to AIC, FPE, Mallows and leave-one-out CV, which are generally biased estimates of the MSFE for $h > 1$.

9 Computation

Let $\boldsymbol{X}$ be the matrix of stacked regressors $\boldsymbol{x}'_{t-h}$.

When $h = 1$, a well-known simplification for the leave-one-out residual is

$$\tilde{e}_{t,1} = \hat{e}_t \left(1 - \boldsymbol{x}'_{t-h} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_{t-h}\right)^{-1}.$$

For $h > 1$, a similar equation is not available. However, Racine (1997) showed that

$$\tilde{\boldsymbol{\beta}}_{t,h} = \left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'_{t,h} \left(\boldsymbol{I} - \boldsymbol{X}_{t,h} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'_{t,h} \right)^{-1} \boldsymbol{X}_{t,h} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \right] \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'_{t,h} \boldsymbol{y}_{t,h} \right)$$

where $\boldsymbol{X}_{t,h}$ and $\boldsymbol{y}_{t,h}$ are the blocks of $\boldsymbol{X}$ and $\boldsymbol{y}$ for the removed observations $\{t-h+1, ..., t, ..., t+h-1\}$. Now, let $\tilde{\boldsymbol{e}}_{t:h}$ be the $1+2h \times 1$ vector of residuals for the observations $\{t-h+1, ..., t, ..., t+h-1\}$ when the coefficient is estimated by $\tilde{\boldsymbol{\beta}}_{t,h}$, and let $\hat{\boldsymbol{e}}_{t:h}$ be the corresponding elements of $\hat{\boldsymbol{e}}$. Note that

$\tilde{e}_{t,h}$ is the middle element of the vector $\tilde{\boldsymbol{e}}_{t:h}$. Let $\boldsymbol{P}_{t,h} = \boldsymbol{X}_{t,h}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'_{t,h}$. We find

$$\begin{aligned}
\tilde{\boldsymbol{e}}_{t:h} &= \boldsymbol{y}_{t,h} - \boldsymbol{X}_{t,h}\tilde{\boldsymbol{\beta}}_{t,h} \\
&= \boldsymbol{y}_{t,h} - \boldsymbol{X}_{t,h}\hat{\boldsymbol{\beta}} + \boldsymbol{P}_{t,h}\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\boldsymbol{P}_{t,h}\boldsymbol{y}_{t,h} \\
&- \boldsymbol{P}_{t,h}\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\boldsymbol{X}_{t,h}\hat{\boldsymbol{\beta}} + \boldsymbol{P}_{t,h}\boldsymbol{y}_{t,h} \\
&= \left(\boldsymbol{I} + \boldsymbol{P}_{t,h}\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\boldsymbol{P}_{t,h} + \boldsymbol{P}_{t,h}\right)\hat{\boldsymbol{e}}_{t:h} \\
&- \left(\boldsymbol{P}_{t,h}\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1} - \boldsymbol{P}_{t,h} - \boldsymbol{P}_{t,h}\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\boldsymbol{P}_{t,h}\right)\boldsymbol{X}_{t,h}\hat{\boldsymbol{\beta}} \\
&= \left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\hat{\boldsymbol{e}}_{t:h}
\end{aligned}$$

using the matrix equalities

$$\begin{aligned}
\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})^{-1} &= (\boldsymbol{I} - \boldsymbol{P})^{-1} \\
\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})^{-1} - \boldsymbol{P} - \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})^{-1}\boldsymbol{P} &= 0.
\end{aligned}$$

This shows that a simple method to calculate $\tilde{e}_{t,h}$ is as the middle element of the vector

$$\tilde{\boldsymbol{e}}_{t:h} = \left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1}\hat{\boldsymbol{e}}_{t:h}. \tag{12}$$

Alternatively, the matrix equality

$$\begin{aligned}
\left(\boldsymbol{I} - \boldsymbol{P}_{t,h}\right)^{-1} &= \left(\boldsymbol{I} - \boldsymbol{X}_{t,h}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'_{t,h}\right)^{-1} \\
&= \boldsymbol{I} + \boldsymbol{X}_{t,h}\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'_{t,h}\boldsymbol{X}_{t,h}\right)^{-1}\boldsymbol{X}'_{t,h}
\end{aligned}$$

shows that

$$\tilde{\boldsymbol{e}}_{t:h} = \hat{\boldsymbol{e}}_{t:h} + \boldsymbol{X}_{t,h}\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'_{t,h}\boldsymbol{X}_{t,h}\right)^{-1}\boldsymbol{X}'_{t,h}\hat{\boldsymbol{e}}_{t:h}$$

and thus an alternative formula to calculate $\tilde{e}_{t,h}$ is

$$\tilde{e}_{t,h} = \hat{e}_t + \boldsymbol{x}'_{t-h}\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'_{t,h}\boldsymbol{X}_{t,h}\right)^{-1}\boldsymbol{X}'_{t,h}\hat{\boldsymbol{e}}_{t:h}. \tag{13}$$

The relative numerical computation costs for (12) versus (13) roughly depends on the relative size of the matrices $\boldsymbol{P}_{t,h}$ and $\boldsymbol{X}'\boldsymbol{X}$. Thus roughly (12) is less costly if $2h+1 < k$, otherwise (13) is less costly. This is an important consideration as the leave-h-out criterion requires calculation of $\tilde{e}_{t,h}$ for all observations t

10 Proofs

The current proofs are sketches, and consequently I do not provide regularity conditions. Essentially, the regularity conditions will be strictly stationarity, mixing, and sufficiently finite moments.

Proof of Theorem 1: The MSFE is

$$\begin{aligned} MSFE &= E\left(y_{n+h} - \boldsymbol{x}_n'\hat{\boldsymbol{\beta}}\right)^2 \\ &= E\left(e_{n+h} - \boldsymbol{x}_n'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2 \\ &= \sigma^2 + E\left(\boldsymbol{x}_n'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2 \\ &\simeq \sigma^2 + E\left(\boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2. \end{aligned} \tag{14}$$

[Note: The approximation (14) needs more careful justification.]

Now the least-squares residual is

$$\hat{e}_t = e_t - \boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

so

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n}\hat{e}_t^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}e_t^2 - \frac{2}{n}\sum_{t=1}^{n}e_t\boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \frac{1}{n}\sum_{t=1}^{n}\left(\boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2. \end{aligned}$$

Its expected value is

$$E\left(\hat{\sigma}^2\right) = \sigma^2 - \frac{2}{n}E\left(\xi_n\right) + E\left(\boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2 \tag{15}$$

where

$$\begin{aligned} \xi_n &= \sum_{t=1}^{n}e_t\boldsymbol{x}_{t-h}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}e_t\boldsymbol{x}_{t-h}'\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_{t-h}\boldsymbol{x}_{t-h}'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}e_t\boldsymbol{x}_{t-h}'. \end{aligned}$$

Combining (14) and (15) we see that

$$MSFE \simeq E\left(\hat{\sigma}^2\right) + \frac{2}{n}E\left(\xi_n\right).$$

Now by the WLLN and CLT,

$$\xi_n \to_d Z'\boldsymbol{Q}^{-1}Z$$

where $Z \sim N\left(0, \boldsymbol{\Omega}\right)$. If ξ_n is uniformly integrable

$$E\left(\xi_n\right) \to E\left(Z'\boldsymbol{Q}^{-1}Z\right) = \mathrm{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right) = \sigma^2 B$$

and

$$E\left(\xi_n\right) = \operatorname{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right) + O\left(n^{-1/2}\right).$$

We have shown that

$$MSFE \simeq E\left(\hat{\sigma}^2\right) + \frac{2\sigma^2}{n}B + O\left(n^{-3/2}\right)$$

as claimed.

Proof of Theorem 2. In (13) we showed that

$$\tilde{e}_{t,h} = \hat{e}_t + \boldsymbol{x}_{t-h}'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_{t,h}'\boldsymbol{X}_{t,h}\right)^{-1}\boldsymbol{X}_{t,h}'\hat{\boldsymbol{e}}_{t:h}.$$

Thus

$$\begin{aligned}
CV_h &= \frac{1}{n}\sum_{t=1}^{n}\tilde{e}_{t,h}^2 \\
&\simeq \frac{1}{n}\sum_{t=1}^{n}\hat{e}_t^2 + \frac{2}{n}\sum_{t=1}^{n}\hat{e}_t\boldsymbol{x}_{t-h}'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_{t,h}'\boldsymbol{X}_{t,h}\right)^{-1}\boldsymbol{X}_{t,h}'\hat{\boldsymbol{e}}_{t:h} \\
&= \hat{\sigma}^2 + \frac{2}{n}\sum_{t=1}^{n}\hat{e}_t\boldsymbol{x}_{t-h}'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_{t,h}'\boldsymbol{X}_{t,h}\right)^{-1}\sum_{j=-(h-1)}^{h-1}\boldsymbol{x}_{t-h+j}\hat{e}_{t+j} \\
&= \hat{\sigma}^2 + \frac{2}{n}\operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_{t,h}'\boldsymbol{X}_{t,h}\right)^{-1}\sum_{j=-(h-1)}^{h-1}\hat{e}_t\boldsymbol{x}_{t-h}\boldsymbol{x}_{t-h+j}'\hat{e}_{t+j}\right) \\
&= \hat{\sigma}^2 + \frac{2}{n}\operatorname{tr}\left(\hat{\boldsymbol{Q}}_h^{-1}\hat{\boldsymbol{\Omega}}\right)
\end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{Q}}_h &= \frac{1}{n}\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_{t,h}'\boldsymbol{X}_{t,h}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_{t-h}\boldsymbol{x}_{t-h}' - \frac{1}{n}\sum_{j=-(h-1)}^{h-1}\boldsymbol{x}_{t-h-j}\boldsymbol{x}_{t-h-j}'
\end{aligned}$$

and

$$\hat{\boldsymbol{\Omega}} = \sum_{j=-(h-1)}^{h-1}\frac{1}{n}\sum_{t}\boldsymbol{x}_{t-h+j}\boldsymbol{x}_{t-h}'\hat{e}_{t+j}\hat{e}_t.$$

Applying the central limit theorem we find that

$$\operatorname{tr}\left(\hat{\boldsymbol{Q}}_h^{-1}\hat{\boldsymbol{\Omega}}\right) = \operatorname{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right) + O_p(n^{-1/2})$$

and thus

$$\begin{aligned} CV_h &= \hat{\sigma}^2 + \frac{2}{n}\operatorname{tr}\left(\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\right) + O_p(n^{-3/2}) \\ &= \hat{\sigma}^2 + \frac{2\sigma^2 B}{n} + O_p(n^{-3/2}). \end{aligned}$$

Combined with Theorem 1 and assuming that $\operatorname{tr}\left(\hat{\boldsymbol{Q}}_h^{-1}\hat{\boldsymbol{\Omega}}\right)$ is uniformly integrable, it follows that

$$\begin{aligned} E\left(CV_h\right) &= E\left(\hat{\sigma}^2\right) + \frac{2\sigma^2 B}{n} + O(n^{-3/2}) \\ &= MSFE + O(n^{-3/2}) \end{aligned}$$

as stated.

References

[1] Akaike, Hirotugu, 1969, Fitting autoregressive models for prediction, Annals of the Institute of Statistical Mathematics 21, 243-247.

[2] Akaike, Hirotugu, 1973, Information theory and an extension of the maximum likelihood principle, in: B. Petroc and F. Csake, (Eds.), Second International Symposium on Information Theory.

[3] Banasali, R.J., 1996, Asymptotically efficient autoregressive model selection for multistep prediction. Annals of the Institute of Statistical Mathematics 48, 577-602.

[4] Bates, J.M. and C.M.W. Granger, 1969, The combination of forecasts. Operations Research Quarterly 20, 451-468.

[5] Brock, William and Stephen Durlauf, 2001, Growth empirics and reality. World Bank Economic Review 15, 229-272.

[6] Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559-581.

[7] Diebold, F. X. and J. A. Lopez, 1996, Forecast evaluation and combination, in: Maddala and Rao, (Eds.), Handbook of Statistics, Elsevier, Amsterdam.

[8] Granger, C.W.J., 1989, Combining Forecasts – Twenty Years Later. Journal of Forecasting 8, 167-173.

[9] Granger, C.W.J. and R. Ramanathan, 1984, Improved methods of combining forecast accuracy. Journal of Forecasting 19, 197-204.

[10] Hansen, Bruce E., 2007, Least Squares Model Averaging. Econometrica 75, 1175-1189.

[11] Hansen, Bruce E., 2008, Least Squares Forecast Averaging. Journal of Econometrics 146, 342-350

[12] Hansen, Bruce E. and Jeffrey Racine, 2009, Jackknife Model Averaging, working paper.

[13] Hansen, Lars Peter and R. J. Hodrick, 1980, "Forward Exchange-Rates As Optimal Predictors of Future Spot Rates - An Econometric-Analysis," Journal of Political Economy 88, 829-853.

[14] Hendry, D.F. and M. P. Clements, 2002, Pooling of forecasts. Econometrics Journal 5, 1-26.

[15] Ing, Ching-Kang, 2003, Multistep prediction in autoregressive processes. Econometric Theory 19, 254-279.

[16] Ing, Ching-Kang, 2004, Selecting optimal multistep predictors for autoregressive processes of unknown order. Annals of Statistics 32, 693-722.

[17] Ing, Ching-Kang, 2007, Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. Annals of Statistics 35, 1238-1277.

[18] Ing, Ching-Kang and Ching-Zong Wei, 2003, On same-realization prediction in an infinite-order autoregressive process. Journal of Multivariate Analysis 85, 130-155.

[19] Ing, Ching-Kang and Ching-Zong Wei, 2005, Order selection for same-realization predictions in autoregressive processes. Annals of Statistics 33, 2423-2474.

[20] Lee, Sangyeol and Karagrigoriou, Alex, 2001, An asymptotically optimal selection of the order of a linear process, Sankhya 63 A, 93-106.

[21] 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.

[22] Mallows, C.L., 1973, Some comments on C_p. Technometrics 15, 661-675.

[23] Rissanen, J., 1986, Order estimation by accumulated prediction errors. Journal of Applied Probability 23A, 55-61.

[24] Schwarz, G., 1978, Estimating the dimension of a model. Annals of Statistics 6, 461-464.

[25] Shibata, Ritaei, 1980, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics 8, 147-164.

[26] Stock, J.H. and M. W. Watson, 1999, A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in: R. Engle and H. White, (Eds.), Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger. Oxford University Press.

[27] Stock, J.H. and M. W. Watson, 2004, Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405-430.

[28] Stock, J.H. and M. W. Watson, 2005, An empirical comparison of methods for forecasting using many predictors. working paper, NBER.

[29] Stock, J.H. and M. W. Watson, 2006, Forecasting with many predictors, in: G. Elliott, C.W.J Granger and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Vol 1. Elsevier, Amsterdam, 515-554.

[30] Takeuchi, K., 1976, Distribution of informational statistics and a criterion of model fitting, Suri-Kagaku, 153, 12-18.

[31] Timmermann, Allan, 2006, Forecast Combinations, in: G. Elliott, C.W.J Granger and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Vol 1. Elsevier, Amsterdam, 135-196.