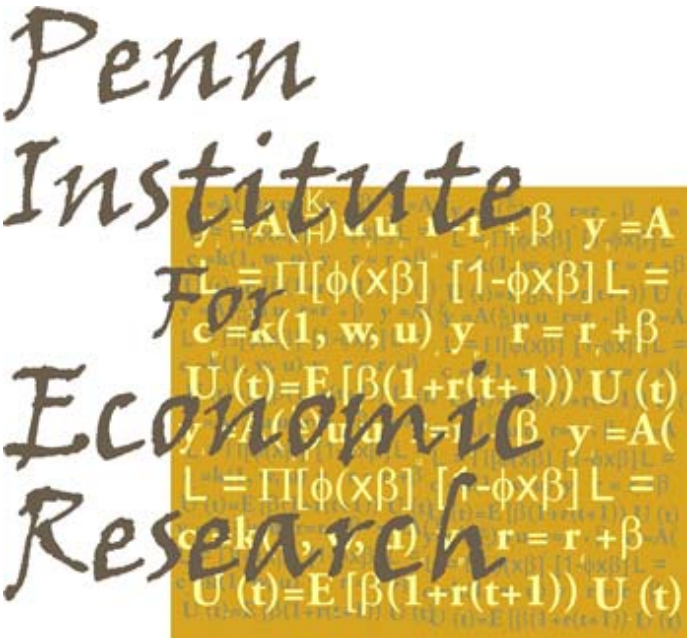

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 12-046

"Forecasting with Factor-Augmented Regression:
A Frequentist Model Averaging Approach"

by

Xu Cheng and Bruce E. Hansen

http://ssrn.com/abstract=2180921

Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach

Xu Cheng
University of Pennsylvania

Bruce E. Hansen*
University of Wisconsin

First Version: May, 2012
This Version: October, 2012

Abstract

This paper considers forecast combination with factor-augmented regression. In this framework, a large number of forecasting models are available, varying by the choice of factors and the number of lags. We investigate forecast combination using weights that minimize the Mallows and the leave-h-out cross validation criteria. The unobserved factor regressors are estimated by principle components of a large panel with N predictors over T periods. With these generated regressors, we show that the Mallows and leave-h-out cross validation criteria are approximately unbiased estimators of the one-step-ahead and multi-step-ahead mean squared forecast errors, respectively, provided that $N, T \to \infty$. In contrast to well-known results in the literature, the generated-regressor issue can be ignored for forecast combination, without restrictions on the relation between N and T.

Simulations show that the Mallows model averaging and leave-h-out cross-validation averaging methods yield lower mean squared forecast errors than alternative model selection and averaging methods such as AIC, BIC, cross validation, and Bayesian model averaging. We apply the proposed methods to the U.S. macroeconomic data set in Stock and Watson (2012) and find that they compare favorably to many popular shrinkage-type forecasting methods.

JEL Classification: C52, C53

Keywords: Cross-validation, factor models, forecast combination, generated regressors, Mallows

*The authors thank Frank Diebold, Frank Schorfhedie, and participants at the 2012 CIREQ Econometrics conference for helpful comments. Hansen thanks the National Science Foundation for research support.

1 Introduction

Factor-augmented regression has received much attention in high-dimensional problems where a large number of predictors are available over a long period. Assuming some unobserved latent factors generate the comovement of all predictors, one can forecast a particular series by the factors rather than by the original predictors, with the benefit of significant dimension reduction (Stock and Watson, 2002). In factor-augmented regression, the factors are determined and ordered by their importance in driving the covariability of many predictors, which may not be consistent with their forecast power for the particular series of interest, an issue discussed in Bai and Ng (2008, 2009). In consequence, model specification is necessary to determine which factors should be used in the forecast regression, in addition to specifying the number of lags of the dependent variable and the number of lags of the factors included. These decisions vary with the particular series of interest and the forecast horizon.

This paper proposes forecast combination based on frequentist model averaging criteria. The forecast combination is a weighted average of the predictions from a set of candidate models that vary by the choice of factors and the number of lags. The model averaging criteria are estimates of the mean square forecast errors (MSFE). Hence, the weights that minimize these model averaging criteria are expected to minimize the MSFE. Two different types of model averaging methods are considered: the Mallows model averaging (MMA; Hansen, 2007) and the leave-h-out cross-validation averaging (CVA$_h$; Hansen, 2010). For one-step-ahead forecasting, the CVA$_h$ method is equivalent to the jackknife model averaging (JMA) from Hansen and Racine (2012). The MMA and CVA$_h$ methods were designed for standard regression models with observed regressors. However, dynamic factor models involve unobserved factors and their estimation creates generated regressors. The effect of generated regressors on model selection and combination has not previously been investigated. This paper makes this extension and provides a theoretical justification for frequentist model averaging methods in the presence of estimated factors.

We show that even in the presence of estimated factors, the Mallows and leave-h-out cross-validation criteria are approximately unbiased estimators of the one-step-ahead and multi-step-ahead MSFE, respectively, provided that $N, T \to \infty$. In consequence, these frequentist model averaging criteria can be applied to factor-augmented forecast combination without modification. Thus for model selection and combination, the generated-regressor issue can be safely ignored. This is in contrast to inference on the coefficients, where Pagan (1984), Bai and Ng (2009), Ludvigson and Ng (2011), and Gonçalves and Perron(2011) have shown that the generated regressors affect the sampling distribution. It is worth emphasizing that our result is not based on asymptotic rates of convergence (such as assuming $T^{1/2}/N \to 0$ as in Bai and Ng (2006)); instead it holds because the

focus is on forecasting rather than parameter estimation. Indeed, in the context of a non-dynamic factor model (one without lagged dependent variables and no serial correlation) we show that the Mallows criterion is an unbiased estimate of the MSFE in finite samples, and retains the classic optimality developed in Li (1987), Andrews (1991) and Hansen (2007). In dynamic models our argument is asymptotic, but does not rely on differing rates of convergence.

Our simulations demonstrate the superior finite-sample performance of the MMA and CVA$_h$ forecasts in the sense of low MSFE. This is consistent with the optimality of MMA and JMA in the absence of temporal dependence and generated regressors (Hansen, 2007; Hansen and Racine, 2012). In addition, the advantage of CVA$_h$ is found most prominent in long-horizon forecast with serially correlated forecast errors.

We apply the proposed methods to the U.S. macroeconomic data set in Stock and Watson (2012) and find that they compare favorably to many popular shrinkage-type forecasting methods.

The frequentist model averaging approach adopted here extends the large literature on forecast combination, see Granger (1989), Clemen (1989), Diebold and Lopez (1996), Henry and Clements (2002), Timmermann (2006), and Stock and Watson (2006), for reviews. Stock and Watson (1999, 2004, 2012) provide detailed empirical evidence demonstrating the gains of forecast combination. The simplest forecast combination is to use equal weights. Compared to simple model averaging, MMA and CVA$_h$ are less sensitive to the choice of candidate models. Alternative frequentist forecast combination methods are proposed by Bates and Granger (1969), Granger and Ramanathan (1984), Timmermann (2006), Buckland, Burnham, and Augustin (2007), Burnham and Anderson (2002), and Hjort and Claeskens (2003). Hansen (2008) shows that MMA has lower MSFE in one-step-ahead forecasts than other methods.

Another popular model averaging approach is the Bayesian model averaging (BMA; Min and Zellner, 1993). The BMA has been widely used in econometric applications, including Sala-i-Martin, Doppelhofer, and Miller (2004), Brock and Durlauf (2001), Brock, Durlauf, and West (2003), Avramov (2002), Fernandez, Lay, and Steel (2001a,b), Garratt, Lee, Pesaran, and Shin (2003), and Wright (2008, 2009). Geweke and Amisano (2011) propose optimal density combination for forecast models. Compared to BMA, the frequentist model averaging approach here does not reply on priors and allows for misspecification through the balance of misspecification errors against overparameterization. Furthermore, our frequentist model averaging approach explicitly deals with generated-regressors, while BMA has no known adjustment.

As an alternative to the model averaging approach, forecasts can be based on one model picked by model selection. Numerous model selection criteria have been proposed, including the Akaike information criterion (AIC; Akaike, 1973), Mallows' C_p (Mallows, 1973), Bayesian information

criterion (BIC; Schwarz 1978), and cross-validation (Stone, 1974). Bai and Ng (2009) argue that these model selection criteria are unsatisfactory for factor-augmented regression because they rely on the specific ordering of the factors and the lags, where the natural order may not work well for the forecast of a particular series. This issue is alleviated in forecast combination by the flexibility of choosing candidate models. In addition, the above model selection procedures have not been investigated in the presence of generated regressors; ours is the first to make this extension.

This paper complements the growing literature on forecasting with many regressors. In addition to those discussed above, many papers consider forecast in a data rich environment. Forni, Hallin, Lippi, and Reichlin (2002, 2005) consider the generalized dynamic factor model and frequency domain estimation. Bernanke, Boivin, and Eliasz (2005) propose forecast with factor-augmented vector autoregressive (FAVAR) model. A factor-augmented VARMA model is suggested by Dufour and Stevanovic (2010). The dynamic factor model is reviewed in Stock and Watson (2011). Bai and Ng (2008) form target predictors associated with the object of interest. Bai and Ng (2009) introduce the boosting approach. Stock and Watson (2012) describe a general shrinkage representation that covers special cases like pretest, BMA, empirical Bayes, and bagging (Inoue and Kilian, 2008). Pesaran, Pick and Timmermann (2011) also investigate multi-step forecasting with correlated errors and factor-augmentation, but in a multivariate framework. Kelly and Pruitt (2011) propose a three-pass-regression filter to handle many predictors. Tu and Lee (2012) consider forecast with supervised factor models. A comprehensive comparison among many competing methods is available in Kim and Swanson (2010). Ng (2011) provides an excellent review on variable selection and contains additional references.

The rest of the paper is organized as follows. Section 2 introduces the dynamic factor model and describes the estimators and combination forecasts. Section 3 provides a detailed description of forecast selection and combination procedures based on the Mallows and leave-h-out cross-validation criteria. Section 4 provides theoretical justification by showing the Mallows and leave-h-out cross-validation criteria are approximately unbiased estimators of the MSFE. Monte Carlo simulations and an empirical application to U.S. macroeconomic data are presented in Sections 5 and 6. Summary and discussions are provided in Section 7.

2 Model and Estimation

Suppose we have observations (y_t, X_{it}) for $t = 1, ..., T$ and $i = 1, ..., N$, and the goal is to forecast y_{T+h} using the factor-augmented regression model

$$y_{t+h} = \alpha_0 + \alpha(L)y_t + \beta(L)'F_t + \varepsilon_{t+h} \tag{2.1}$$

where $h \geq 1$ is the forecast horizon, $\sigma^2 = \mathbb{E}\varepsilon_t^2$, and $F_t \in \mathbb{R}^r$ are unobserved common factors satisfying

$$X_{it} = \lambda_i' F_t + e_{it}. \tag{2.2}$$

The vectors $\lambda_i \in \mathbb{R}^r$ are called the factor loadings, e_{it} is called an idiosyncratic error, and $\alpha(L)$ and $\beta(L)$ are lag polynomials of order p and q, respectively.[1] In matrix notation, (2.2) can be written as

$$X = F\Lambda' + e \tag{2.3}$$

where X is a $T \times N$, $F = (F_1, ..., F_T)'$ is $T \times r$, $\Lambda = (\lambda_1, ..., \lambda_N)'$ is $N \times r$, and e is a $T \times N$ error matrix. We assume that the number of factors r in (2.2) is known, though in practice r can be consistently selected by the information criteria in Bai and Ng (2002).[2]

Our contribution is to treat the structures of the lag polynomials $\alpha(L)$ and $\beta(L)$ in (2.1) as unknown, and to introduce methods to select the lag structures. Suppose that the forecaster is considering approximating models for (2.1) which include up to $p_{\max}$ lags of y_t and $q_{\max}$ lags of F_t. Thus the largest possible lag structure for (2.1) includes the regressors

$$z_t = (1, y_t, ..., y_{t-p_{\max}+1}, F_t', ..., F_{t-q_{\max}+1}')'. \tag{2.4}$$

Given this regressor set, write (2.1) as

$$y_{t+h} = z_t' b + \varepsilon_{t+h} \tag{2.5}$$

where b includes all coefficients from (2.1). Now suppose that the forecaster is considering M approximating models indexed by $m = 1, ..., M$, where each approximating model m specifies a subset $z_t(m)$ of the regressors z_t. The forecaster's m^{th} approximating model is then

$$y_{t+h} = z_t(m)' b(m) + \varepsilon_{t+h}(m), \tag{2.6}$$

or in matrix notation

$$y = Z(m)b(m) + \varepsilon(m). \tag{2.7}$$

We do not place any restrictions on the approximating models; in particular, the models

[1] We assume a sufficient number of observations are available in history for the estimation of (2.1) when the left hand side is y_1.

[2] The averaging methods proposed below also work in practice when r is unknown and the largest approximating model is chosen to include $r_{\max}$ number of factors, where $r_{\max} > r$. This is equivalent to employing irrelevant factor regressors in (2.1), which has insignificant effect on the optimal combination forecast.

may be nested or non-nested. However, the set of models should be selected judiciously so that the total number of models M is practically and computationally feasible. A simple choice is to take sequentially nested subsets of z_t. Another simple feasible choice is to set $z_t(m) = (1, y_t, y_{t-1}, ..., y_{t-m+1}, F_t^m, ..., F_{t-m+1}^m)$, where F_t^m denote the first m factors in F_t. Alternatively, a relatively simple choice is to set $z_t(m) = (1, y_t, y_{t-1}, ..., y_{t-p(m)+1}, F_t^m, ..., F_{t-q(m)+1}^m)$ where we separately vary $p(m)$ among $(1, 2, ..., P)$ and $q(m)$ among $(1, 2, ..., Q)$. The choice of lag structures is not critical to our treatment.

For estimation we replace the unobservable factors F by their principle component estimate $\widetilde{F} = (\widetilde{F}_1, ..., \widetilde{F}_T)' \in \mathbb{R}^{T \times r}$, which is the matrix of r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix XX'. Let $\widetilde{z}_t(m)$ denote $z_t(m)$ with the factors F_t replaced with their estimates $\widetilde{F}_t$, and set $\widetilde{Z}(m) = (\widetilde{z}_1(m), ..., \widetilde{z}_{T-h}(m))'$. The least squares estimate of $b(m)$ is then $\widehat{b}(m) = (\widetilde{Z}(m)'\widetilde{Z}(m))^{-1}\widetilde{Z}(m)'y$ with residual $\widehat{\varepsilon}_{t+h}(m) = y_{t+h} - \widetilde{z}_t(m)'\widehat{b}(m)$. The least squares estimate $\widehat{b}(m)$ is often called a "two-step" estimator as the regressor $\widetilde{z}_t(m)$ contains the estimate $\widetilde{F}_t$ also known as a "generated regressor".

The least squares forecast of y_{T+h} by the m^{th} approximating model is

$$\widehat{y}_{T+h|T}(m) = \widetilde{z}_T(m)'\widehat{b}(m). \tag{2.8}$$

Forecast combinations can be constructed by taking weighted averages of the forecasts $\widehat{y}_{T+h|T}(m)$. These take the form

$$\widehat{y}_{T+h|T}(w) = \sum_{m=1}^{M} w(m)\widehat{y}_{T+h|T}(m), \tag{2.9}$$

where $w(m)$, $m = 1, ..., M$, are forecast weights. Let $w = (w(1), ..., w(M))'$ denote the weight vector. We will require that the weights are non-negative and sum to one, e.g., $0 \leq w(m) \leq 1$ and $\sum_{m=1}^{M} w(m) = 1$, or equivalently that $w \in \mathcal{H}^M$, the unit simplex in $\mathbb{R}^M$. Forecast combination generalizes forecasting based on a single model as the latter obtains by setting $w(m) = 1$ for a single model m.

3 Forecast Selection and Combination

The problem of forecast selection is choosing the forecast $\widehat{y}_{T+h|T}(m)$ from the set $m = 1, ..., M$. The problem of forecast combination is selecting the weight vector w from $\mathcal{H}^M$. In this section we describe the Mallows and leave-h-out cross-validation criteria for forecast selection and combination.

Factor models are distinct from conventional forecasting models in that they involve generated regressors (the estimated factors). As shown by Pagan (1984), in general the presence of generated

regressors affects the asymptotic distribution of two-step parameter estimates such as $\widehat{b}(m)$. The details for dynamic factor models have been worked out by Bai and Ng (2006, 2009). Bai and Ng (2006) show that the generated regressor effect is asymptotically negligible if $T^{1/2}/N \to 0$, that is, if the cross-sectional dimension is sufficiently large so that the first-step estimation error is of a smaller stochastic order than the second-step estimation error. Bai and Ng (2009) refine this analysis, showing that the first stage estimation increases the asymptotic variance by a factor related to both T and N. Consequently, they propose to adjust the boosting stopping rule for MSE minimization. The lesson from this literature is that we should not neglect the effect of generated regressors when considering model selection.

The Mallows (1973) criterion is a well-known unbiased estimate of the expected squared fit in the context of homoskedastic regression with independent observations. The criterion applies to any estimator whose fitted values are a linear function of the dependent variable y. In the context of model selection with estimated factors, the fitted regression vector is $\widetilde{Z}(m)\widehat{b}(m) = \widetilde{Z}(m)(\widetilde{Z}(m)'\widetilde{Z}(m))^{-1}\widetilde{Z}(m)'y$ and in the context of forecast combination the fitted regression vector is $\sum_{m=1}^{M} w(m)\widetilde{Z}(m)(\widetilde{Z}(m)'\widetilde{Z}(m))^{-1}\widetilde{Z}(m)'y$. In both cases the fitted values are a linear function of y if $\widetilde{Z}(m)$ is not a function of y, which occurs in any non-dynamic factor model (that is, model (2.1) without lagged dependent variables). This is because the generated regressors $\widetilde{Z}(m)$ are a function only of X. (Recall, $\widetilde{F}$ are the eigenvectors of XX' associated with the r largest eigenvalues.) Consequently, the Mallows criterion is directly applicable without modification to non-dynamic homoskedastic factor models, and Mallows selection and averaging retains the optimality properties described in Li (1987), Andrews (1991), and Hansen (2007). This is a simple yet exciting insight. It is also quite surprising given the failure of conventional inference in the presence of generated regressors. Our intuition is that while generated regressors inflate the variance of the parameter estimates, they symmetrically inflate the Mallows criterion, and thus the criterion remains informative.

Unfortunately this finite-sample argument does not apply directly to the dynamic model (2.1) with lagged dependent variables. Therefore in the next section we use asymptotic arguments to establish the validity of the Mallows criterion for the dynamic factor model. It follows that the unadjusted Mallows criterion is appropriate for forecast selection and combination for dynamic factor models.

We now describe the Mallows criterion for selection and combination. Let $k(m) = \dim(z_t(m))$ denote the number of regressors in the m^{th} model. The Mallows criterion for forecast selection is

$$C_T(m) = \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_t(m)^2 + \frac{2\widehat{\sigma}_T^2}{T}k(m), \tag{3.1}$$

where $\widehat{\sigma}_T^2$ is a preliminary estimate of σ^2. We suggest $\widehat{\sigma}_T^2 = (T-k(M))^{-1}\sum_{t=1}^T \widehat{\varepsilon}_t(M)^2$ using a large approximate model M so that $\widehat{\sigma}_T^2$ is approximately unbiased for σ^2. The Mallows selected model is $\widehat{m} = \operatorname{argmin}_{1\leq m\leq M} C_T(m)$ and the selected forecast is $\widehat{y}_{T+h|T}(\widehat{m})$. Numerically, this is accomplished by estimating each model m, calculating $C_T(m)$ for each model, and finding the model $\widehat{m}$ with the smallest value of the criterion.

For forecast combination, the Mallows criterion for weight selection is

$$C_T(w) = \frac{1}{T}\sum_{t=1}^T \left(\sum_{m=1}^M w(m)\widehat{\varepsilon}_t(m)\right)^2 + \frac{2\widehat{\sigma}_T^2}{T}\sum_{m=1}^M w(m)k(m). \tag{3.2}$$

The Mallows selected weight vector is obtained by finding the weight vector w which minimizes $C_T(w)$. We can write this as

$$\widehat{w} = \operatorname*{argmin}_{w\in\mathcal{H}^M} CV_T(w) \tag{3.3}$$

and the selected forecast is $\widehat{y}_{T+h|T}(\widehat{w})$. Following Hansen (2008) we call this the MMA forecast. Numerically, the solution (3.3) minimizes the quadratic function $C_T(w)$ subject to a set of equality and inequality constraints, and is easiest accomplished using a quadratic programming algorithm, which are designed for this situation. Quadratic programming routines are available in standard languages including Gauss, Matlab, and R.

The Mallows criterion is simple and convenient, but it is restrictive in that it requires the error ε_{t+h} to be conditionally homoskedastic and serially uncorrelated. The homoskedasticity restriction can be avoided by instead using leave-one-out cross validation as in Hansen and Racine (2012), which is a generally valid selection criterion under heteroskedasticity. The leave-one-out cross-validation criterion, however, still requires the error to be serially uncorrelated, yet when $h > 1$ the error ε_{t+h} is generally a moving average process and thus is serially correlated.

To incorporate serial correlation, Hansen (2010) has recommended using the leave-h-out cross-validation criterion which is the sum of squared leave-h-out prediction residuals.

To construct this criterion, define the leave-h-out prediction residual $\widetilde{\varepsilon}_{t+h,h}(m) = y_{t+h} - \widetilde{z}_t(m)'\widetilde{b}_{t,h}(m)$ where $\widetilde{b}_{t,h}(m)$ is the least squares coefficient from a regression of y_{t+h} on $\widetilde{z}_t(m)$ with the observations in periods $\{t-h+1, ..., t+h-1\}$ omitted. This leave-h-out residual uses the full-sample estimated factors $\widetilde{F}_t$. When $h = 1$ the prediction residual has the simple formula $\widetilde{\varepsilon}_{t+h,h}(m) = \widehat{\varepsilon}_{t+h}(m)(1-\widetilde{z}_t(m)'(\widetilde{Z}(m)'\widetilde{Z}(m))^{-1}\widetilde{z}_t(m))^{-1}$. For $h > 1$, Hansen (2010) has shown that

it can be computed via the formula

$$\widetilde{\varepsilon}_{t+h,h}(m) = \widehat{\varepsilon}_{t+h}(m) + \widetilde{z}_t'(m)\left(\sum_{|j-t|\geq h} \widetilde{z}_j(m)\widetilde{z}_j'(m)\right)^{-1}\left(\sum_{|j-t|<h} \widetilde{z}_j(m)\widehat{\varepsilon}_{j+h}(m)\right). \tag{3.4}$$

The cross-validation criterion for forecast selection is

$$CV_{h,T}(m) = \frac{1}{T}\sum_{t=1}^{T} \widetilde{\varepsilon}_{t,h}(m)^2. \tag{3.5}$$

The cross-validation selected model is $\widehat{m} = \operatorname{argmin}_{1\leq m\leq M} CV_{h,T}(m)$ and the selected forecast is $\widehat{y}_{T+h|T}(\widehat{m})$.

For forecast combination, the cross-validation criterion is

$$CV_{h,T}(w) = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{m=1}^{M} w(m)\widetilde{\varepsilon}_{t,h}(m)\right)^2. \tag{3.6}$$

The cross-validation selected weight vector minimizes $CV_{h,T}(w)$, that is,

$$\widehat{w} = \operatorname*{argmin}_{w\in\mathcal{H}^M} CV_{h,T}(w). \tag{3.7}$$

As for Mallows combination, (3.7) is conveniently solved via quadratic programming, as the criterion (3.6) is quadratic in w. The cross-validation selected combination forecast is $\widehat{y}_{T+h|T}(\widehat{w})$, and we call this the leave-h-out cross-validation averaging (CVA_h) forecast.

4 Asymptotic Theory

In this section, we provide theoretical justification for the Mallows criterion and the leave-h-out cross-validation criterion with estimated factors. In the first subsection we describe the technical assumptions, and in the second describe the connection between in-sample fit, mean-squared error, and mean-squared forecast error. In the third sub-section we show that the Mallows criterion is an approximately unbiased estimator of the MSFE in the case of one-step-ahead forecasts and conditional homoskedasticity. In the fourth we examine the leave-h-out cross-validation criterion, and show a similar result for multi-step forecasts allowing for conditional heteroskedasticity.

4.1 Assumptions

Let $\mathcal{F}_t = \sigma(y_t, X_t, y_{t-1}, X_{t-1}, ...)$ denote the information set at time t. Let C denote a generic constant. For a matrix A, $A > 0$ denotes A is positive definite.

Assumption R.

(i) $\mathbb{E}(\varepsilon_{t+h}|\mathcal{F}_t) = 0$.

(ii) $(z_t', \varepsilon_{t+h}, e_{1t}, ..., e_{Nt})$ is strictly stationary.

(iii) $\mathbb{E}||z_t||^4 \leq C$, $\mathbb{E}\varepsilon_t^4 \leq C$, and $\mathbb{E}(z_t z_t') > 0$.

(iv) $T^{-1/2}\sum_{t=1}^{T} z_t\varepsilon_{t+h} \to_d N(0, \Omega)$, where $\Omega = \sum_{|j|<h}\mathbb{E}(z_t z_{t-j}'\varepsilon_{t+h}\varepsilon_{t+h-j})$.

Assumption R(i) implies that ε_{t+h} is conditionally unpredictable at time t, but when $h > 1$ it does not imply that ε_{t+h} is serially uncorrelated. This is consistent with the fact that the h-step-ahead forecast error ε_{t+h} typically is a moving average process of order $h-1$. Assumption R(ii) assumes the data is strictly stationary, which simplifies the asymptotic theory, and links the in-sample fit of the averaging estimator to its out-of-sample performance. (See Section 4.2 below for details.) Assumptions R(iii)-R(iv) are standard moment bounds and the central limit theorem, the latter satisfied under standard weak dependence conditions. The specific form of Ω in Assumption R(iv) follows from stationarity and Assumption R(i).

Assumption F.

(i) The factors satisfy $\mathbb{E}\|F_t\|^4 \leq C$ and $T^{-1}\sum_{t=1}^{T} F_t F_t' \to_p \Sigma_F > 0$.

(ii) The loading λ_i is either deterministic such that $\|\lambda_i\| \leq C$ or it is stochastic such that $\mathbb{E}\|\lambda_i\|^4 \leq C$. In either case, $N^{-1}\Lambda'\Lambda \to_p \Sigma_\lambda > 0$.

(iii) $\mathbb{E}(e_{it}) = 0$, $E|e_{it}|^8 \leq C$.

(iv) $\mathbb{E}(e_{it}e_{js}) = \sigma_{ij,ts}$, $|\sigma_{ij,ts}| \leq \overline{\sigma}_{ij}$ for all (t,s), and $|\sigma_{ij,ts}| \leq \tau_{ts}$ for all (i,j) such that $N^{-1}\sum_{i,j=1}^{N}\overline{\sigma}_{ij} \leq C$, $T^{-1}\sum_{t,s=1}^{T}\tau_{ts} \leq C$, and $(NT)^{-1}\sum_{i,j,t,s=1}|\sigma_{ij,ts}| \leq C$.

(v) For every (t,s), $\mathbb{E}|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq C$.

(vi) The variables $\{\lambda_i\}, \{F_i\}, \{e_{it}\}$ are three mutually independent groups. Dependence within each group is allowed.

(vii) For each t, $\mathbb{E}||(NT)^{-1/2}\sum_{s=1}^{T-h}\sum_{i=1}^{N}(F_s + \varepsilon_{s+h})(e_{it}e_{is} - E(e_{it}e_{is}))||^2 \leq C$.

(viii) For all (i,t), $\mathbb{E}||(NT)^{-1/2}\sum_{t=1}^{T-h}\sum_{i=1}^{N}\lambda_i e_{it}\varepsilon_{t+h}||^2 \leq M$, where $E(\lambda_i e_{it}\varepsilon_{t+h}) = 0$.

Assumption F is similar to Assumptions A-D in Bai and Ng (2006) and Assumptions 1-4 of Gonçalves and Perron (2011).[3] Assumptions F(i) and F(ii) ensure that there are r non-trivial

[3] Assumption F does not include Assumption C4 of Bai and Ng (2006) and Assumption 3(e) of Gonçalves and Perron (2011). The reason is that the objective of the present paper does not require invoking the asymptotic distribution of the estimated factors established in Bai (2003).

strong factors. This does not accommodate weak factors as in Onatski (2012). Assumptions F(iii)-F(v) allow for heteroskedasticity and weak dependence in both the time series and cross-sectional dimensions, an approximate factor structure as in Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986, 1993). Assumption F(vi) can be replaced by high-level moment conditions, such as Assumptions D and F2 of Bai (2003) and Assumptions 3(a), 3(c), and 3(d) of Gonçalves and Perron (2011). Assumption F(vii) and F(viii) impose weak dependence between the idiosyncratic errors and the regression error as well as bounded moments for the sum of some zero-mean random variables. They are analogous to Assumptions 3(b), 4(a), and 4(b) of Gonçalves and Perron (2011), who also provide sufficient conditions under mutual independence of $\{\lambda_i\}, \{e_{is}\}$ and $\{\varepsilon_{t+h}\}$. A condition similar to Assumption (vii) also is employed by Assumption F1 in Bai (2003).

4.2 MSE and MSFE

We first show that the MSFE is close to the expected in-sample squared error. To see this, write the conditional mean in (2.1) as μ_t so that the equation is $y_{t+h} = \mu_t + \varepsilon_{t+h}$ or as a $T \times 1$ vector as $y = \mu + \varepsilon$. Similarly for any forecast combination w, write $\widehat{\mu}_t(w) = \sum_{m=1}^{M} w(m)\widetilde{z}_t(m)'\widehat{b}(m)$ and in vector notation $y = \widehat{\mu}(w) + \widehat{\varepsilon}(w)$.

Now define the in-sample squared error

$$\begin{aligned} L_T(w) &= \frac{1}{T}\sum_{t+h=1}^{T}\left(\varepsilon_{t+h}^2 + \left(\mu_t - \widehat{\mu}_t(w)\right)^2\right) \\ &= \frac{1}{T}\varepsilon'\varepsilon + \frac{1}{T}\left(\mu - \widehat{\mu}(w)\right)'\left(\mu - \widehat{\mu}(w)\right). \end{aligned} \tag{4.1}$$

The first term is independent of the model weights. The second term measures the fit of the estimate $\widehat{\mu}(w)$ for the conditional mean μ. The expectation of the in-sample squared error is the in-sample mean-squared error:

$$\begin{aligned} MSE_T(w) &= \mathbb{E}L_T(w) \\ &= \mathbb{E}\left(\varepsilon_{t+h}^2 + \left(\mu_t - \widehat{\mu}_t(w)\right)^2\right). \end{aligned} \tag{4.2}$$

Now observe that the MSFE of the point forecast $\widehat{y}_{T+h|T}(w)$ is

$$\begin{aligned} MSFE_T(w) &= \mathbb{E}\left(y_{T+h} - \widehat{y}_{T+h|T}(w)\right)^2 \\ &= \mathbb{E}\left(\varepsilon_{T+h}^2 + \left(\mu_T - \widehat{\mu}_T(w)\right)^2\right) \\ &\simeq \mathbb{E}\left(\varepsilon_{t+h}^2 + \left(\mu_t - \widehat{\mu}_t(w)\right)^2\right) \\ &= MSE_T(w) \end{aligned} \tag{4.3}$$

the second equality holds since ε_{T+h} is uncorrelated with $\widehat{\mu}_T(w)$, and the approximation in the third line follows from stationarity of $(y_t, \widetilde{F}_t)$. This calculation shows that the MSFE is close to the MSE, which is the expected in-sample fit $L_T(w)$.

The Mallows and leave-h-out cross-validation criteria are designed as estimates of $L_T(w)$. The near equivalence with MSFE shows that these criteria are also estimates of MSFE and are thus appropriate forecast selection criteria.

The approximation rests on whether the distribution of $(y_t, \widetilde{F}_t)$ is approximately stationary. This holds since the principle component estimate $\widetilde{F}_t$ is a weighted average of $X_t = (X_{1t}, ..., X_{Nt})$, where the weight is an approximately orthogonal transformation of Λ, which holds under Assumption F as shown by Bai and Ng (2002) and Bai (2003). Combined with the stationarity and independence conditions in Assumptions R(ii) and F(vi), it follows that $(y_t, \widetilde{F}_t)$ is approximately stationary as claimed.

4.3 Mallows Criterion

In this section we restrict attention to the case of one-step forecasts ($h = 1$) and conditional homoskedasticity. Thus Assumption R(i) is strengthened to $\mathbb{E}(\varepsilon_{t+1}|\mathcal{F}_t) = 0$ and $\mathbb{E}(\varepsilon_{t+1}^2|\mathcal{F}_t) = \sigma^2$. Under these conditions we show that the Mallows criterion is an asymptotically unbiased estimate of the in-sample fit $L_T(w)$.

To see this, recalling the definitions of μ and $\widehat{\mu}(w)$ given in Section 4.2, we can see that $\widehat{\mu}(w) = \widetilde{P}(w)y = \widetilde{P}(w)\mu + \widetilde{P}(w)\varepsilon$, where $\widetilde{P}(w) = \sum_{m=1}^{M} w(m)\widetilde{P}(m)$ and $\widetilde{P}(m) = \widetilde{Z}(m)(\widetilde{Z}(m)'\widetilde{Z}(m))^{-1}\widetilde{Z}(m)'$. Thus the residual vector equals

$$\begin{aligned} \widehat{\varepsilon}(w) &= \varepsilon + \mu - \widehat{\mu}(w) \\ &= \varepsilon + \left(I - \widetilde{P}(w)\right)\mu - \widetilde{P}(w)\varepsilon. \end{aligned} \tag{4.4}$$

We calculate that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{m=1}^{M}w(m)\widehat{\varepsilon}_t(m)\right)^2 &= \frac{1}{T}\widehat{\varepsilon}(w)'\widehat{\varepsilon}(w)\\ &= L_T(w)+2\frac{1}{T}\left(\mu-\widehat{\mu}(w)\right)'\varepsilon\\ &= L_T(w)+2\frac{1}{T}\mu'\left(I-\widetilde{P}(w)\right)\varepsilon-2\frac{1}{T}\varepsilon'\widetilde{P}(w)\varepsilon. \qquad (4.5)\end{aligned}$$

It follows that

$$C_T(w)=L_T(w)+T^{-1/2}r_{1T}(w)+T^{-1}r_{2T}(w) \qquad (4.6)$$

where

$$\begin{aligned}r_{1T}(w) &= 2\frac{1}{\sqrt{T}}\mu'\left(I-\widetilde{P}(w)\right)\varepsilon\\ r_{2T}(w) &= -2\left(\varepsilon'\widetilde{P}(w)\varepsilon-\widehat{\sigma}_T^2\sum_{m=1}^{M}w(m)k(m)\right). \qquad (4.7)\end{aligned}$$

This shows that the Mallows criterion equals the in-sample fit $L_T(w)$ plus two remainder terms. We now show that $r_{1T}(w)$ and $r_{2T}(w)$ converge in distribution to zero mean random variables. This provides an asymptotic justification for treating $C_T(w)$ as an approximately unbiased estimate of $L_T(w)$. Consequently, selecting the weight vector (or model) to minimize $C_T(w)$ is a reasonable approximation to the minimization of $L_T(w)$, and hence the MSFE.

We first take $r_{2T}(w)$. First, note that if $\widehat{\sigma}_T^2$ is estimated using a large model which includes the true lags as a special case (or if the number of lags increases with sample size) then $\widehat{\sigma}_T^2\to_p\sigma^2$. Set $P(w)=\sum_{m=1}^{M}w(m)P(m)$ where $P(m)=Z(m)\left(Z(m)'Z(m)\right)^{-1}Z(m)'$. Under Assumption R, $\mathbb{E}(\varepsilon_{t+1}|\mathcal{F}_t)=0$ and $\mathbb{E}(\varepsilon_{t+1}^2|\mathcal{F}_t)=\sigma^2$, then $T^{-1/2}Z(m)'\varepsilon\to_d N(0,\sigma^2V(m))$ and $T^{-1}Z(m)'Z(m)\to_p V(m)$, where $V(m)=\mathbb{E}z_t(m)z_t'(m)$. It follows that $\varepsilon'P(m)\varepsilon\to_d\sigma^2\xi(m)$, where $\xi(m)\sim\chi^2_{k(m)}$. We deduce that

$$r_{2T}^0(w)=-2\left(\sum_{m=1}^{M}w(m)\left(\varepsilon'P(m)\varepsilon-\widehat{\sigma}_T^2k(m)\right)\right)\to_d-2\sum_{m=1}^{M}w(m)\sigma^2\left(\xi(m)-k(m)\right)=\zeta(w), \qquad (4.8)$$

where $\mathbb{E}\zeta(w)=0$.

We next show that $r_{2T}(w) - r_{2T}^0(w)$ is asymptotically negligible. To this end, write

$$\begin{aligned}
&\varepsilon' \widetilde{P}(m) \varepsilon \\
&= \left[T^{-1/2} Z_H(m)'\varepsilon + A_T\right]' \left[T^{-1} Z_H(m)' Z_H(m) + B_{1T} + B_{1T}' + B_{2T}\right]^{-1} \left[T^{-1/2} Z_H(m)'\varepsilon + A_T\right], \\
&A_T = T^{-1/2} \left(\widetilde{Z}(m) - Z_H(m)\right)' \varepsilon, \\
&B_{1T} = T^{-1} \left(\widetilde{Z}(m) - Z_H(m)\right)' Z_H(m), \\
&B_{2T} = T^{-1} \left(\widetilde{Z}(m) - Z_H(m)\right)' \left(\widetilde{Z}(m) - Z_H(m)\right),
\end{aligned} \tag{4.9}$$

and $Z_H(m) = Z(m)H(m)$ for some full-rank block-diagonal matrix $H(m)$ that transforms the factor column spaces in $Z(m)$.[4] Let $C_{NT} = \min[N, T]$. By Lemma A.1 of Bai and Ng (2006), $B_{1T} = O_p(C_{NT}^{-1})$ and $B_{2T} = O_p(C_{NT}^{-1})$ under Assumptions R and F, showing that the estimated factors approximately span the column spaces of the true factors in large sample. By Lemma A.1 of Gonçalves and Perron (2011), $A_T = O_p(C_{NT}^{-1})$, under Assumptions R and F.[5] It is worth pointing out that the normalization in A_T is $T^{-1/2}$, making it a stronger result than sample average. Because A_T, B_{1T}, and B_{2T} are all negligible as $N, T \to \infty$, we conclude that $r_{2T}(w) - r_{2T}^0(w) = o_p(1)$. We have shown that $r_{2T}(w) \to_d \zeta(w)$ when $N, T \to \infty$, as desired.

The arguments above are analogous to those in Bai and Ng (2006) on the effect of factor estimation on confidence intervals. However, the above results hold without imposing the strong $T^{1/2}/N \to 0$ condition used in Bai and Ng (2006).

We next take $r_{1T}(w)$. As in the above argument we can show that $r_{1T}(w) = r_{1T}^0(w) + o_p(1)$ where $r_{1T}^0(w) = \sum_{m=1}^M w(m) \frac{1}{\sqrt{T}} \mu' \left(I - P(m)\right) \varepsilon$. Notice that $\mu = Zb$ where $Z = (z_1, ..., z_T)'$ and b is the true coefficients in (2.5). Then under Assumption R,

$$\frac{1}{\sqrt{T}} \mu' \left(I - P(m)\right) \varepsilon = \frac{1}{\sqrt{T}} b' Z' \left(I - P(m)\right) \varepsilon \to_d S(m) \sim N(0, \sigma^2 Q(m)), \tag{4.10}$$

where $Q(m) = \operatorname{plim} b' Z' \left(I - P(m)\right) Zb$. Thus

$$r_{1T}(w) = r_{1T}^0(w) + o_p(1) \to_d S(w) = \sum_{m=1}^M w(m) S(m). \tag{4.11}$$

and $\mathbb{E}S(w) = 0$.

[4] The exact form of $H(m)$ is based on the transformation matrix H defined in Lemma A.1 of Bai and Ng (2006), with adjustments that each approximate model only invoves a subset of all factors and their lags. In addition, $H(m)$ is block-diagonal, where the upper-left block associated with the lags of y_t is an identity matrix. As such, $H(m)$ only rotates the columns of factors and their lags.

[5] Assumptions R and F imply all assumptions in Bai and Ng (2006) and Goncales and Perron (2011) used to obtain the desired results.

We have established the following result.

Theorem 1 *Suppose* $h = 1$, $\mathbb{E}(e_t^2|\mathcal{F}_{t-1}) = \sigma^2$, *and Assumptions* R *and* F *hold. For fixed* M *and* w, *and* $N, T \to \infty$,

$$C_T(w) = L_T(w) + T^{-1/2} r_{1T}(w) + T^{-1} r_{2T}(w),$$

where

$$r_{1T}(w) \to_d \zeta(w),$$

$$r_{2T}(w) \to_d S(w),$$

$\mathbb{E}\zeta(w) = 0$ *and* $\mathbb{E}S(w) = 0$.

Theorem 1 shows that for one-step homoskedastic forecasting, the Mallows criterion $C_T(w)$ is equal to the in-sample squared error $L_T(w)$ plus terms of smaller stochastic order with asymptotic zero means. Thus $C_T(w)$ is an asymptotically unbiased estimator of $\mathbb{E}L_T(w) \simeq MSFE_T(w)$. This holds for any weight vector w, and holds even though the regressors are estimated factors. This result is similar to the theory of Hansen (2008) for forecast combination without estimated factors.

While Theorem 1 establishes that the Mallows criterion is asymptotically unbiased for the MSFE, it does not establish that the selected weight vector is asymptotically efficient in the sense of Shibata (1980), Ing and Wei (2005), or Schorfheide (2005) for forecast selection, or Hansen (2007) in the case of model averaging. In particular, Ing and Wei (2005) show that in an infinite-order autoregressive (AR) model with i.i.d. innovations, the AR order selected by the Akaike or Mallows criterion is asymptotically optimal in the sense of minimizing the one-step-ahead MSFE among all candidate models. No similar result exists for forecast combination, and a rigorous demonstration of optimality is beyond the scope of this paper. Nevertheless, the asymptotic unbiasedness of the Mallows criterion shown in Theorem 1, the existing optimality results on Mallows model averaging, and the optimality theory of Ing and Wei (2005) together suggest that Mallows forecast combination in the presence of estimated factors is a reasonable weight selection method.

4.4 Multi-Step Forecast with Leave-h-out cross-validation Averaging

When $h > 1$ or the errors are possibly conditionally heteroskedastic the Mallows criterion applies an incorrect parameterization penalty. Instead, following Hansen (2010) we recommend the leave-h-out cross-validation criterion for forecast selection and combination. In this section we provide a theoretical foundation for this criterion in the presence of estimated factors.

First, as is shown in the proof of Theorem 2 of Hansen (2010), the cross-validation criterion is approximately equal to a penalized sum-of-squared errors. To see this, use the computation formula

(3.4) to write

$$\sum_{m=1}^{M} w(m)\widetilde{\varepsilon}_{t,h}(m) = \widehat{\varepsilon}_{t+h}(w) + \sum_{m=1}^{M} w(m)\widetilde{z}_t'(m)\left(\sum_{|j-t|\geq h}\widetilde{z}_j(m)\widetilde{z}_j'(m)\right)^{-1}\left(\sum_{|j-t|<h}\widetilde{z}_j(m)\widehat{\varepsilon}_{j+h}(m)\right) \tag{4.12}$$

where $\widehat{\varepsilon}_{t+h}(w) = \sum_{m=1}^{M} w(m)\widehat{\varepsilon}_{t+h}(m)$. Applied to definition (3.6) we find

$$\begin{aligned}
& CV_{h,T}(w) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\varepsilon}_{t+h}(w) + \sum_{m=1}^{M} w(m)\widetilde{z}_t'(m)\left(\sum_{|j-t|\geq h}\widetilde{z}_j(m)\widetilde{z}_j'(m)\right)^{-1}\left(\sum_{|j-t|<h}\widetilde{z}_j(m)\widehat{\varepsilon}_{j+h}(m)\right)\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{t+h}(w)^2 + \\
&\quad \frac{2}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{t+h}(w)\sum_{m=1}^{M} w(m)\widetilde{z}_t'(m)\left(\sum_{|j-t|\geq h}\widetilde{z}_j(m)\widetilde{z}_j'(m)\right)^{-1}\left(\sum_{|j-t|<h}\widetilde{z}_j(m)\widehat{\varepsilon}_{j+h}(m)\right) + T^{-2}r_{3T} \\
&= \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{t+h}(w)^2 + \frac{2}{T}\sum_{m=1}^{M} w(m)\,\mathrm{tr}\left(\widehat{V}(m)^{-1}\widehat{\Omega}(w,m)\right) + T^{-2}r_{3T}
\end{aligned} \tag{4.13}$$

where

$$\begin{aligned}
\widehat{V}(m) &= \frac{1}{T}\sum_{|j-t|\geq h}\widetilde{z}_j(m)\widetilde{z}_j'(m), \\
\widehat{\Omega}(w,m) &= \sum_{|j|<h}\frac{1}{T}\sum_{t=1}^{T}\widetilde{z}_{t-j}(m)\widetilde{z}_t'(m)\widehat{\varepsilon}_{t+h}(w)\widehat{\varepsilon}_{t+h-j}(m),
\end{aligned} \tag{4.14}$$

and

$$\begin{aligned}
r_{3T} &= \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{m=1}^{M} w(m)\widetilde{z}_t'(m)\left(\frac{1}{T}\sum_{|j-t|\geq h}\widetilde{z}_j(m)\widetilde{z}_j'(m)\right)^{-1}\left(\sum_{|j-t|<h}\widetilde{z}_j(m)\widehat{\varepsilon}_{j+h}(m)\right)\right)^2 \\
&= O_p(1).
\end{aligned} \tag{4.15}$$

Combined with expansion (4.5), we find

$$\begin{aligned}
CV_{h,T}(w) &= L_T(w) + T^{-1/2}r_{1T}(w) + T^{-1}r_{2T}^*(w) + T^{-2}r_{3T}, \text{ where} \\
r_{2T}^*(w) &= -2\left(\varepsilon'\widetilde{P}(w)\varepsilon - \sum_{m=1}^{M} w(m)\,\mathrm{tr}\left(\widehat{V}(m)^{-1}\widehat{\Omega}(w,m)\right)\right).
\end{aligned} \tag{4.16}$$

Under Assumption R,

$$\begin{aligned} T^{-1/2}Z(m)'\varepsilon &\to_d G(m) \sim N(0,\Omega(m)), \text{ where} \\ \Omega(m) &= \sum_{|j|<h} \mathbb{E}(z_t(m)z_{t-j}'(m)\varepsilon_{t+h}\varepsilon_{t+h-j}). \end{aligned} \tag{4.17}$$

Combined with the arguments presented in the previous section, we deduce that

$$\begin{aligned} \varepsilon'\widetilde{P}(m)\varepsilon &\to_d G(m)'V(m)^{-1}G(m), \\ \widehat{V}(m) &\to_p V(m) = \mathbb{E}z_t(m)z_t(m)', \\ \widehat{\Omega}(w,m) &\to_p \Omega(m). \end{aligned} \tag{4.18}$$

It follows that

$$r_{2T}^*(w) \to_d S^*(w) = -2\sum_{m=1}^{M} w(m)\left(G(m)'V(m)^{-1}G(m) - \operatorname{tr}\left(V(m)^{-1}\Omega(m)\right)\right). \tag{4.19}$$

Since $\mathbb{E}G(m)G(m)' = \Omega(m)$ it is not hard to calculate that $\mathbb{E}S^*(w) = 0$. We have established the following result.

Theorem 2 *Suppose Assumptions* R *and* F *hold. For any $h \geq 1$, fixed M and w, and $N, T \to \infty$,*

$$CV_{h,T}(w) = L_T(w) + T^{-1/2}r_{1T}(w) + T^{-1}r_{2T}^*(w) + T^{-2}r_{3T},$$

where

$$r_{1T}(w) \to_d \zeta(w),$$

$$r_{2T}^*(w) \to_d S^*(w),$$

$\mathbb{E}\zeta(w) = 0$ *and* $\mathbb{E}S^*(w) = 0$, *and* $r_{3T} = O_p(1)$.

Theorem 2 is similar in form to Theorem 1. It shows that the leave-h-out cross-validation criterion is equal to the in-sample squared error $L_T(w)$ plus terms of smaller stochastic order with asymptotic zero means. Thus $CV_{h,T}(w)$ is an asymptotically unbiased estimator of $\mathbb{E}L_T(w) \simeq MSFE_T(w)$. This holds for any weight vector w, even though the regressors are estimated factors, for any forecast horizon h, and allows for conditional heteroskedasticity. Theorem 2 extends Theorem 2 of Hansen (2010) to forecasting with factor-augmentation.

The conventional Mallows criterion imposes an incorrect penalty because $\Omega(m) \neq \sigma^2 V(m)$, as in Hansen and Hodrick (1980). This inequality arises when the error ε_{t+h} is serially correlated (which

occurs when $h > 1$) or conditionally heteroskedastic. This insight suggests that the performance of the Mallows criteria will deteriorate when the serial dependence of the forecast error is strong and the forecast horizon is long, and this is conformed by our simulations. A potential solution is to use an alternative penalty (e.g., a robust Mallows criterion). We recommend the leave-h-out cross-validation criterion as it makes this adjustment automatically, works well in finite samples, and is conceptually straightforward to generalize to more complicated settings.

5 Finite Sample Investigation

In this section, we investigate the finite-sample MSFE of the MMA and CVA_h methods. The data generating process is analogous to that considered in Bai and Ng (2009), but we focus on linear models and add moving average dynamics to the multi-step forecast error. Let F_{jt} denote the j^{th} component of F_t. For $j = 1, ..., r$, $i = 1, ..., N$, and $t = 1, ..., T$, the approximate factor model is

$$\begin{aligned} X_{it} &= \lambda_i F_t + \sqrt{r} e_{it}, \\ F_{jt} &= \alpha_j F_{jt-1} + u_{jt}, \\ e_{it} &= \rho_i e_{it-1} + \epsilon_{it}, \end{aligned} \tag{5.1}$$

where $r = 4$, $\lambda_i \sim N(0, rI_r)$, $\alpha_j \sim U[0.2, 0.8]$, $\rho_i \sim U[0.3, 0.8]$, $(u_{jt}, \epsilon_{it}) \sim N(0, I_2)$, i.i.d. over t, for all j and i. The values of α_j and ρ_i are drawn once and held fixed over simulation repetitions. The regression equation for forecast is

$$\begin{aligned} y_{t+h} &= \beta_1 F_{2t} + \beta_2 F_{4t} + \beta_3 F_{2t-1} + \beta_4 F_{4t-1} + \beta_5 F_{2t-2} + \beta_6 F_{4t-2} + \varepsilon_{t+h}, \\ \varepsilon_{t+h} &= \sum_{j=1}^{h-1} \pi^j v_{t+h-j}, \end{aligned} \tag{5.2}$$

where $v_t \sim N(0, 1)$, i.i.d. over t, and $\{v_t\}$ is independent of $\{u_{js}\}$ and $\{\epsilon_{is}\}$ for any t and s. As such, only two factors and their lags are relevant for forecasting. The parameters are $\beta = (\beta_1, ..., \beta_6) = c[0.5, 0.5, 0.2, 0.2, 0.1, 0.1]$, where c is a scaling parameter ranging from 0.2 to 1.2 for $h = 1$. For multi-step forecasting, the moving average parameter π ranges from 0.1 to 0.9 and the scale parameter c is held at 1. The sample size is $N, T = 100$ and 50,000 simulation repetitions are conducted.

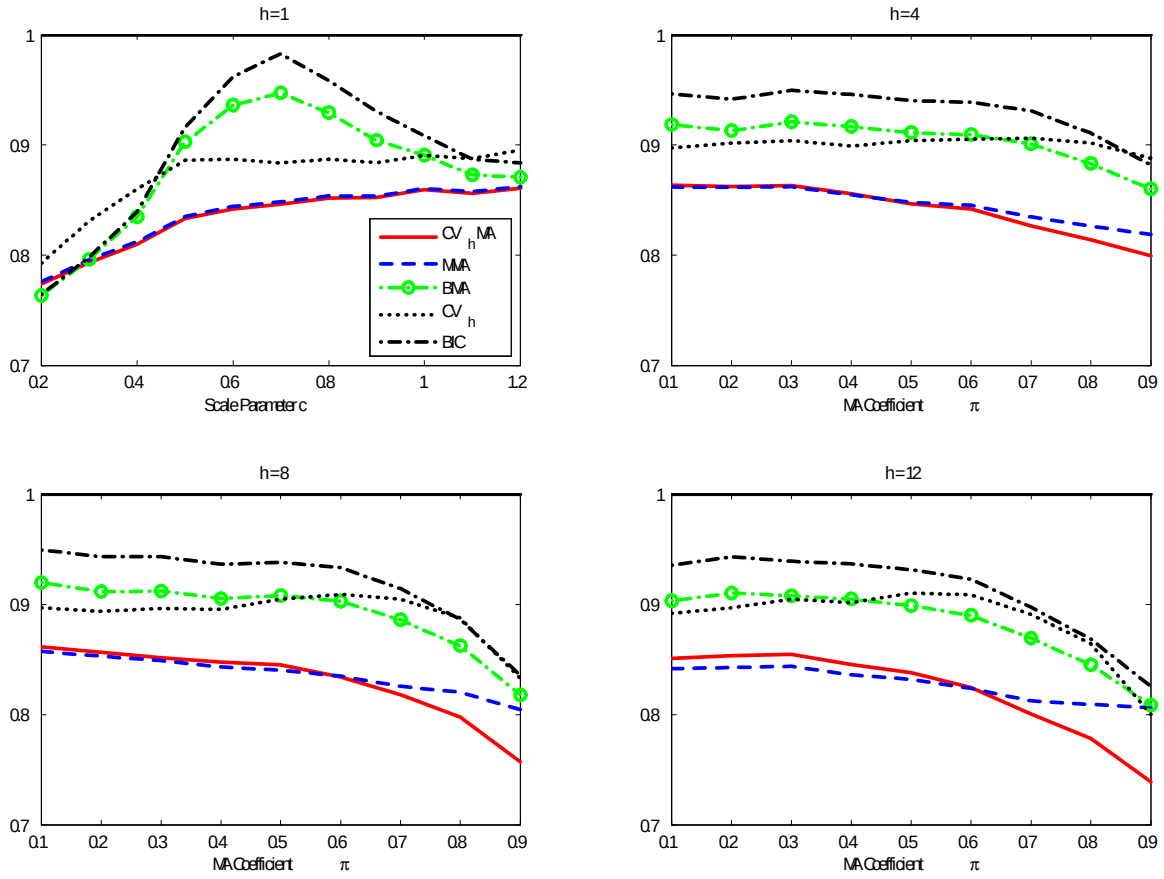

Figure 1. Relative MSFE to LS for $h = 1, 4, 8$, and 12. CV$_h$MA is leave-h-out cross validation model averaging. MMA is Mallows model averaging. BMA is Bayesian model averaging. CV$_h$ is model selection with leave-h-out cross validation. BIC is model selection with Bayesian information criterion.

The set of candidate regressors for model averaging and model selection is

$$\mathcal{Z}_t = (1, y_t, ..., y_{t-p_{\max}}, \widetilde{F}_t', ..., \widetilde{F}_{t-p_{\max}}'), \tag{5.3}$$

where $p_{\max} = 5$. The number of factors in $\widetilde{F}_t$ is selected by IC_{p2} in Bai and Ng (2002). Model averaging are performed over sequentially nested models. We also considered alternative ways to arrange the factors and their lags in $\mathcal{Z}_t$ and simulation results show the same pattern. Model selection methods also are computed over sequentially nested models.

We compare the MSFE of various model averaging and model selection methods. The model averaging methods include leave-h-out cross-validation averaging (CVA$_h$), jackknife model averaging (JMA), Mallows model averaging (MMA), and Bayesian model averaging (BMA).[6] The model

[6]The Bayesian model averaging is computed with weight $w(m) = \exp(-BIC(m)/2)/\sum_{i=1}^{M}\exp(-BIC(i)/2)$, where $BIC(m)$ is the BIC for the m^{th} model.

selection methods include leave-h-out cross validation, jackknife cross validation, Mallows model selection, AIC, and BIC. Selected procedures are reported in Figure 1, with the omitted ones dominated by some of the reported procedures. The relative MSFE in Figure 1 is normalized by the MSFE for the least-squares forecast with all regressors in $\mathcal{Z}_t$. Thus a value smaller than 1 implies superior performance relative to unconstrained least-squares.

Figure 1 shows that CVA_h has the best overall performance, followed by MMA. For the one-step-ahead forecast, CVA_h and MMA are comparable. They dominate all other methods except when the scale parameter c is around 0.2, an extreme situation with very low signal-to-noise ratio in the forecast equation. For the multi-step forecasts, the advantage of CVA_h is prominent when the forecast horizon is long and the serial dependence in the forecast error is strong. For example, when $h = 8$ and $\pi = 0.8$, the relative MSFE for CVA_h is 80%, around 10% smaller than that for model selection by BIC or cross validation, 7% smaller than that for BMA, and 3% smaller than that for MMA. Simulation results demonstrate the same pattern when we experiment with different specifications of the regression coefficients and the true number of factors and lags.

6 Empirical Application

In this section, we apply the MMA, JMA, and CVA_h to forecast U.S. macroeconomic series and compare them to various shrinkage-type methods discussed in Stock and Watson (2012). We adopt the approach in Stock and Watson (2012) that places nonzero weights on principle components beyond the first few. Thus, results here complement those in Stock and Watson (2012) by adding frequentist forecast combination methods to the list covered by their shrinkage representation, such as pretest methods, Bayesian model averaging, empirical Bayes, and bagging.

The data set, taken from Stock and Watson (2012), consists of 143 U.S. macroeconomic time series with quarterly observations from the second quarter of 1960 to the last quarter of 2008. The series are transformed by taking logarithm and/or differencing as described in Table B.1 of Stock and Watson (2012). The principle component estimates of the factors are computed from the 109 lower-level disaggregate series and all 143 series are used as the dependent variables to be forecast.

Following Stock and Watson (2012), the MSFE is computed in two ways: a rolling pseudo out-of-sample forecast method (Table 1) and a cross-validation method (Table 2). The length of the rolling window is 100-h. The rolling results pertain to the post-1984 "Great Moderation" period due to the need for a large startup sample.

We report relative root mean squared error (RMSE) relative to the dynamic factor model with 5 factors (DFM-5). Stock and Watson (2012) show that DFM-5 improves upon AR(4) model in

Table 1. Relative RMSE to DFM5, Rolling Forecast, 1985-2008

	$h=1$			$h=2$			$h=4$		
percentile	0.250	0.500	0.750	0.250	0.500	0.750	0.250	0.500	0.750
CVA_h	0.983	1.003	1.016	0.962	0.992	1.014	0.964	0.985	1.012
JMA	0.983	1.003	1.016	0.962	0.996	1.013	0.972	0.994	1.020
MMA	0.992	1.009	1.031	0.974	1.004	1.025	0.975	1.007	1.034
BMA	0.993	1.014	1.053	0.976	1.009	1.038	0.979	1.014	1.047

Table 2. Relative RMSE to DFM5, Cross Validation, Subsample 1985-2008

	$h=1$			$h=2$			$h=4$		
percentile	0.250	0.500	0.750	0.250	0.500	0.750	0.250	0.500	0.750
CVA_h	0.974	0.992	1.007	0.956	0.981	0.996	0.923	0.958	0.981
JMA	0.974	0.992	1.007	0.958	0.980	0.998	0.924	0.961	0.985
MMA	0.982	0.998	1.014	0.960	0.986	1.008	0.928	0.966	0.995
BMA	0.965	0.991	1.013	0.953	0.983	1.006	0.924	0.964	0.999

more than 75% of series and the shrinkage methods offer little or no improvements over DFM-5 on average. Hence, DFM-5 serves as a good benchmark for the comparison.

Tables 1-2 can be viewed as extensions of Table 2 and Table S-2A in Stock and Watson (2012), with three frequentist model averaging methods added to existing results.[7] The same forecast horizons, $h = 1, 2, 4$, are considered. Entries in the Tables are percentiles of distributions of RMSEs over the 143 variables being forecast. A value smaller than 1 at the median implies that the method considered is superior to DFM-5 for more than half of all series.

Table 1 shows that for $h = 4$ with rolling method, CVA_h improves upon DFM-5 by at least 1.5% for half of all series and by at least 3.6% for one-fourth of all series. In contrast, Table 2 of Stock and Watson (2012) shows that all shrinkage methods considered are inferior to DFM-5 for more than half of all series. JMA (equivalently, CVA_1) is only slightly inferior to CVA_h and MMA is comparable to other shrinkage methods. The same trend holds for $h = 2$, although the difference is not as significant as that for $h = 4$. When $h = 1$, all averaging and shrinkage methods are comparable to DFM-5.

Table 2 shows that for $h = 4$, CVA_h improves upon DFM-5 by at least 4.2% for half of all series and by at least 1.9% for three-fourth of all series, where MSFE is computed by cross-validation methods. In this case, other shrinkage methods also offer improvements upon DFM-5 for some series, but no method does so for as many as three-fourth of all series, according to Table S-2A in Stock and Watson (2012). A category analysis as in Stock and Watson (2012) shows that these

[7] The results on BMA is taken from Stock and Watson (2012). Comaprable results on AR(4), OLS, pretest, bagging, and Logit methods are also available in Stock and Watson (2012) and its supplement.

frequentist forecast combination methods also tend to do well when some shrinkage methods show improvements and there remain hard-to-forecast series.

7 Conclusion

This paper proposes frequentist model averaging approach for forecast combination with the factor-augmented regression, where the unobserved factors are estimated by the principle components of a large panel of predictors. The Mallows model averaging (MMA) and the leave-h-out cross-validation averaging (CVA_h) criteria are shown to be approximately unbiased estimators of the MSFE in one-step and multi-step forecasts, respectively, provided $N, T \to \infty$ in the panel data. Thus, the generated regressor issue is negligible, without any requirement on the relative size of N and T. Monte Carlo simulations and empirical application support the theoretical result that these frequentist model averaging criteria are designed to mirror the MSFE such that the weight vector selected approximately minimizes the MSFE.

The forecast combination methods proposed in this paper can be extended and adapted to a broader class of applications. One extension is to generalize the single variable forecast to the multivariate forecast in the factor-augmented vector autoregressive (FAVAR) model by Bernanke, Boivin, and Eliasz (2005). Second, nonlinear factor-augmented regression should be considered, as discussed in Bai and Ng (2009). Finally, interval forecast based on model averaging is an important but challenging topic (Leeb and Pötscher, 2003, 2008). These topics are investigated in future research.

References

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

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

Avramov, D., 2002. Stock return predictability and model uncertainty. Journal of Finance 64, 423-458.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135-171.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191-221.

Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74, 1133-1150.

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146, 304-317.

Bai, J., Ng, S., 2009. Boosting diffusion indices. Journal of Applied Econometrics 4, 607-629.

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

Bernanke B., Boivin, J., Eliasz, P. S., 2005. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) Approach. The Quarterly Journal of Economics 120, 387-422.

Brock, W., Durlauf, S., 2001. Growth empirics and reality. World Bank Economic Review 15, 229-272.

Brock, W., Durlauf, S., West, K. D., 2003. Policy analysis in uncertain economic environments. Brookings Papers on Economic Activity 1, 235-322.

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

Burnham, K. P., Anderson, D. R., 2002. Model selection and multimodel inference: A Practical Information-Theoretic Approach, Second ed. Springer, New York.

Chamberlain, G, Rothschild, M, 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51, 1281-1304.

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

Connor, G., Korajczyk, R. A., 1986. Performance measurement with the arbitrage pricing theory : A new framework for analysis. Journal of Financial Economics 15, 373-394.

Connor, G., Korajczyk, R. A., 1993. A test for the number of factors in an approximate factor model. Journal of Finance 48, 1263-1291.

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

Dufour, J.-M., Stevanovic, D., 2010. Factor-augmented VARMA models: identification, estimation, forecasting and impulse responses. University of Montreal, Working paper.

Fernandez, C., Ley, E., Steel, M., 2001a. Benchmark priors for Bayesian model averaging. Journal of Econometrics 100, 381-427.

Fernandez, C., Ley, E., Steel, M., 2001b. Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16, 563-576.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic factor model: Identification and estimation. Review of Economics and Statistics 82, 540-554.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association 100, 830-840.

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

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164, 130-141.

Granger, C.W.J., 1989. Combining forecasts twenty years later. Journal of Forecasting 8, 167-173.

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

Gonçalves S., Perron, B., 2011. Bootstrapping factor-augmented regression models. University of Montreal, Working paper.

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

Hansen, B. E., 2008. Least squares forecasting averaging. Journal of Econometrics 146, 342-350.

Hansen, B. E., 2010. Multi-step forecast model selection. University of Wisconsin, Working Paper.

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

Hansen, L. P., and Hodrick, R. J., 1980. Forward exchange-rates as optimal predictors of future spot rates - An econometrics analysis. Journal of Political Economy 88, 829-853.

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

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

Ing, C. -K., Wei, C. -Z., 2005. Order selection for same-realization predictions in autoregressive processes. Annals of Statistics 33, 2423-2474.

Inoue, A., Kilian, L., 2008. How useful is bagging in forecasting economics time series? A case study of U.S. CPI inflation. Journal of the American Statistical Association 103, 511-522.

Kelly, B., Pruitt, S., The three-pass regression filter: A new approach to forecasting using many predictors. Fama-Miller Working Paper; Chicago Booth Research Paper No. 11-19.

Kim, H., Swanson, N., 2010. Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. Rutgers University, Working paper.

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

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

Li, K.-C. 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set. Annals of Statistics 15, 958-975.

Ludvigson, S., Ng, S., 2011. A factor analysis of bond risk premia. In: Ullah, A. and Giles D. (Eds). Handbook of Empirical Economics and Finance, Chapman and Hall, 313-372.

Mallows, C.L., 1973. Some comments on Cp. Technometrics 15, 661-675.

Min, C.-K., Zellner, A., 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. Journal of Econometrics 56, 89-118.

Ng, S., 2011. Variable selection in predictive regressions. Columbia University, Working Paper.

Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics 168, 244-258.

Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. International Economic Review 25, 221-247.

Pesaran, M. H., Pick, A., Timmermann, A., 2011. Variable selection, estimation and inference for multi-period forecasting problems. Journal of Econometrics 164, 173-187.

Schorfheide, F., 2005. VAR Forecasting under misspecification. Journal of Econometrics 128, 99-136.

Sala-i-Martin, Xavier, Doppelhofer, Gernot, Miller, Ronald I., 2004. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. American Economic Review 94, 813-835.

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.

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

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

Stock, J. H., Watson, M. W., 2002. Forecasting using principle components from a large number of predictors. Journal of American Statistical Association 97, 1167-1179.

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

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

Stock, J. H., Watson, M. W., 2011. Dynamic factor models. In: Clements, M. P. and Hendry, D. F. (Eds). Oxford Handbook of Forecasting, Oxford: Oxford University Press.

Stock, J. H., Watson, M. W., 2012. Generalized shrinkage methods for forecasting using many predictors. Journal of Business and Economic Statistics, forthcoming.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B 36, 111-147.

Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 1. Elsevier, Amsterdam, 135-196.

Tu, Y., Lee, T.-H., 2012. Forecasting using supervised factor models. Working paper, University of California, Riverside.

Wright, J. H., 2008. Bayesian model averaging and exchange rate forecasting. Journal of Econometrics 146, 329-341.

Wright, J. H., 2009. Forecasting US Inflation by Bayesian Model Averaging. Journal of Forecasting 28, 131-144.