Econometric Reviews
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lecr20

To Combine Forecasts or to Combine Information?

Huiyu Huang [a] & Tae-Hwy Lee [b]

[a] PanAgora Asset Management, Boston, Massachusetts, USA

[b] Department of Economics, University of California, Riverside, California, USA

Version of record first published: 15 Sep 2010.

To cite this article: Huiyu Huang & Tae-Hwy Lee (2010): To Combine Forecasts or to Combine Information?, Econometric Reviews, 29:5-6, 534-570

To link to this article: http://dx.doi.org/10.1080/07474938.2010.481553

Econometric Reviews, 29(5–6):534–570, 2010

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2010.481553

Taylor & Francis
Taylor & Francis Group

TO COMBINE FORECASTS OR TO COMBINE INFORMATION?

Huiyu Huang[1] and Tae-Hwy Lee[2]

[1]*PanAgora Asset Management, Boston, Massachusetts, USA*
[2]*Department of Economics, University of California, Riverside, California, USA*

□ *When the objective is to forecast a variable of interest but with many explanatory variables available, one could possibly improve the forecast by carefully integrating them. There are generally two directions one could proceed: combination of forecasts (CF) or combination of information (CI). CF combines forecasts generated from simple models each incorporating a part of the whole information set, while CI brings the entire information set into one super model to generate an ultimate forecast. Through linear regression analysis and simulation, we show the relative merits of each, particularly the circumstances where forecast by CF can be superior to forecast by CI, when CI model is correctly specified and when it is misspecified, and shed some light on the success of equally weighted CF. In our empirical application on prediction of monthly, quarterly, and annual equity premium, we compare the CF forecasts (with various weighting schemes) to CI forecasts (with principal component approach mitigating the problem of parameter proliferation). We find that CF with (close to) equal weights is generally the best and dominates all CI schemes, while also performing substantially better than the historical mean.*

Keywords Equally weighted combination of forecasts; Equity premium; Factor models; Forecast combination; Forecast combination puzzle; Information sets; Many predictors; Principal components; Shrinkage.

JEL Classification C3; C5; G0.

1. INTRODUCTION

When one wants to predict an economic variable using the information set of many explanatory variables that have been shown or conjectured to be relevant, one can either use a super model which combines all the available information sets or use the forecast combination methodology. It is commonly acknowledged in the literature that the forecast generated by all the information incorporated in one step (combination of information,

Address correspondence to Tae-Hwy Lee, Department of Economics, University of California, Riverside, CA 92521-0427, USA; E-mail: taelee@ucr.edu

or CI) is better than the combination of forecasts from individual models each incorporating partial information (combination of forecasts, or CF). For instance, Engle et al. (1984) have commented: "The best forecast is obtained by combining information sets, not forecasts from information sets. If both models are known, one should combine the information that goes into the models, not the forecasts that come out of the models." Granger (1989), Diebold (1989), Diebold and Pauly (1990), and Hendry and Clements (2004) have similar arguments. It seems that researchers in this field lean more towards favoring the CI scheme.

However, as Diebold and Pauly (1990) further point out, "... it must be recognized that in many forecasting situations, particularly in real time, pooling of information sets is either impossible or prohibitively costly." Likewise, when models underlying the forecasts remain partially or completely unknown (as is usually the case in practice, e.g., survey forecasts), one would never be informed about the entire information set. On the other hand, growing amount of literature have empirically demonstrated the superior performance of forecast combination. For recent work, see Stock and Watson (2004) and Giacomini and Komunjer (2005).[1]

The frequently asked questions in the existing literature are: "To combine or not to combine"[2] and "how to combine."[3] In this paper, we are interested in: "To combine forecasts or to combine information." This is an issue that has been addressed but not yet elaborated much (see Chong and Hendry, 1986; Diebold, 1989; Newbold and Harvey, 2001; Stock and Watson, 2004; Clements and Galvao, 2006 provide empirical comparisons). Indeed, quite often the combination of forecasts is used when the only things available are individual forecasts (for example the case of professional forecasters) while the underlying information and the model used for generating each individual forecast are unknown, thus the focus of "how to combine."

In this article we elaborate a different issue. Consider the situation that the predictor sets are available but the question is how to use these predictor sets. This forecasting situation is also prevalent in practice. For example, the empirical application we consider in Section 5, predicting excess stock market return using a couple of predictors with proven forecast ability for return in the finance literature. With predictor sets now available, combination of forecasts is no longer a method you end up with

[1]A similar issue is about forecast combination versus forecast encompassing, where the need to combine forecasts arises when one individual forecast fails to encompass the other. See Diebold (1989), Newbold and Harvey (2001), among others.

[2]See Palm and Zellner (1992) and Hibon and Evgeniou (2005).

[3]See, for example, Granger and Ramanathan (1984), Deutsch et al. (1994), Shen and Huang (2006), and Hansen (2008). Clemen (1989) and Timmermann (2006) provide excellent surveys on forecast combination and related issues.

due to lack of knowledge on the underlying information of individual forecasts, but one you can choose to get better out-of-sample forecasting performance than pooling all the predictors at once into a large model (CI). The common belief that CI is better than CF might be based on the in-sample analysis. On the contrary, from out-of-sample analysis, we often find CF performs better than CI. Many articles typically account for the out-of-sample success of CF over CI by pointing out various disadvantages CI may possibly possess. For example, (a) in many forecasting situations, particularly in real time, CI by pooling all information sets is either impossible or too expensive (Diebold, 1989; Diebold and Pauly, 1990; Timmermann, 2006); (b) in a data rich environment where there are many relevant input variables available, the super CI model may suffer from the well-known problem of curse of dimensionality (Timmermann, 2006); and (c) under the presence of complicated dynamics and nonlinearity, constructing a super model using CI may be likely misspecified (Hendry and Clements, 2004).

In this article, through a linear regression framework, for out-of-sample forecasting, under strict exogeneity of predictors, we show analytically that CI can be beaten by CF even when the CI model coincides with the data generation process (DGP) and when the CI model is misspecified. Intuitively, CF can be more successful than CI in out-of-sample forecasting largely due to: 1) the bias and variance trade-off between a small model (each individual forecasting model in CF is usually small) and a large model (CI model is usually large); and 2) in the stage of combining, CF combines individual forecasts that contain both information of the forecast target y and information of the predictors x, while CI combines information of the predictors only without taking into consideration their relationships with y. In this sense, CF may be viewed as a "supervised learning" mechanism (see, for example, Bai and Ng, 2008). We also shed some light on the (puzzling) success of the equally-weighted CF forecasts. Monte Carlo study is presented to illustrate the analytical results. Our analytical illustration provides some interpretation for simulation and empirical findings. Moreover, the analytical findings assist us to shed some light on the empirical success of equally weighted combination of forecasts which is deemed as a "puzzle" in forecast combination literature (Stock and Watson, 2004; Timmermann, 2006).

Finally, as an empirical application, we study the equity premium prediction for which we compare various schemes of CF and CI. Goyal and Welch (2008) explore the out-of-sample performance of many stock market valuation ratios, interest rates and consumption-based macroeconomic ratios toward predicting the equity premium. Here we bring the CF method into predicting equity premium and compare with CI. We implement CF with various weighting methods, including simple average, regression based approach (see Granger and Ramanathan, 1984),

and principal component forecast combination (see Stock and Watson, 2004). We find that CF with (close to) equal weights is generally the best and dominates all CI schemes, while also performing substantially better than the historical mean.

The article is organized as follows. Section 2 examines analytically the out-of-sample relative merits of CF in comparison with CI. Section 2 considers two cases, which set up the two experimental designs of Monte Carlo analysis in Section 4. In Section 3 we discuss the "forecast combination puzzle"—the empirical success of equally weighted combination of forecasts (which we call CF-Mean), and provide our attempts on understanding the puzzle in several ways. Furthermore, we discuss the weighting of CF in the shrinkage framework as in Diebold and Pauly (1990) and compare with CI. Section 5 presents an empirical application for equity premium prediction to compare the performance of various CF and CI schemes. In Section 5, we use the principal component (PC) models for CI and CF. Section 2 is about the theory of comparing CF and CI, but not about comparing CF-PC and CI-PC. Nevertheless, we include these two factor models (CF-PC, CI-PC) in the empirical section (Section 5) because of the following two reasons. First, the empirical section has a large number of predictors $N = 12$ and it would be unfair for CI as it could easily be contaminated by the large parameter estimation uncertainty. Hence, we consider a factor model for CI, namely CI-PC, for a fair comparison by mitigating the parameter estimation error. We include CI-PC in the empirical section, even if we do not include it in the analytical discussion and Monte Carlo experiment where $N = 2, 3$ is small. Second, more importantly, we note that CF-Mean is a single factor CF-PC model with the factor loading shrunken to a constant (Remark 3, Section 5.2). As we include CF-Mean, it may be natural to include the factor model without the shrinkage (CF-PC). Noting that CF-Mean is a shrinkage version of the CF-PC, we can also view the regression based CF (CF-RA) and its shrinkage version (denoted CF-RA(κ)) as a general shrinkage version of the CF-PC. Section 6 concludes.

2. OUT-OF-SAMPLE FORECAST: CF CAN BE BETTER THAN CI

Suppose we forecast a scalar variable y_{t+1} using the information set available up to time t, $\mathcal{I}_t = \{x_s\}_{s=0}^{t}$, where x_s is a $1 \times k$ vector of weakly stationary variables. Let $x_s = (x_{1s}\ x_{2s})$ be a non-empty partition. The CF forecasting scheme is based on two individual regression models

$$y_{t+1} = x_{1t}\beta_1 + \epsilon_{1,t+1}, \tag{1}$$

$$y_{t+1} = x_{2t}\beta_2 + \epsilon_{2,t+1}. \tag{2}$$

The CI takes a model[4]

$$y_{t+1} = x_{1t}\alpha_1 + x_{2t}\alpha_2 + e_{t+1}. \tag{3}$$

Forecast Models: Denote the one-step out-of-sample CI and CF forecasts as

$$\begin{aligned} \hat{y}^{\mathrm{CI}}_{T+1} &= x_T\hat{\alpha}_T = x_{1T}\hat{\alpha}_{1,T} + x_{2T}\hat{\alpha}_{2,T}, \\ \hat{y}^{\mathrm{CF}}_{T+1} &= w_1\hat{y}^{(1)}_{T+1} + w_2\hat{y}^{(2)}_{T+1} = w_1 x_{1T}\hat{\beta}_{1,T} + w_2 x_{2T}\hat{\beta}_{2,T}, \end{aligned} \tag{4}$$

where $\hat{y}^{(1)}_{T+1}$ and $\hat{y}^{(2)}_{T+1}$ are forecasts generated by forecasting models (1) and (2) respectively, and w_i $(i = 1, 2)$ denote the forecast combination weights. All parameters are estimated using strictly past information (up to time T) as indicated in subscript. Let $\hat{e}_{T+1} \equiv y_{T+1} - \hat{y}^{\mathrm{CI}}_{T+1}$ denote the forecast error by CI, $\hat{\epsilon}_{i,T+1} \equiv y_{T+1} - \hat{y}^{(i)}_{T+1}$ denote the forecast errors by the first $(i = 1)$ and the second $(i = 2)$ individual forecast, and $\hat{e}^{\mathrm{CF}}_{T+1} \equiv y_{T+1} - \hat{y}^{\mathrm{CF}}_{T+1}$ denote the forecast error by CF.

DGPs: We consider two cases for the DGP: (i) when the DGP is the same as the CI model (i.e., the CI model (3) is correctly specified for the DGP) and (ii) when the DGP has the additional variable set x_3 to generate y (i.e., the CI model (3) is misspecified for the DGP as it omits x_3).

We show that even in the first case when the CI model coincides with the DGP, CF can be better than CI in a finite sample (see Section 2.1 and Section 4 (Table 1) for the analysis and simulation results). When the CI model is not correctly specified for the DGP and suffers from the omitted variable problem, we show that CF can be better than CI even in a large sample $(T \to \infty)$. Section 2.2 and Section 4 (Table 2) provide analytical illustrations and simulation results, respectively.

2.1. When the CI Model is Correctly Specified

DGP1: Suppose that the DGP is the same as the CI model (3) which generates y from x_1 and x_2

$$y_{t+1} = x_{1,t}\theta_1 + x_{2,t}\theta_2 + \eta_{t+1}, \tag{5}$$

[4]Our CF and CI model set-ups (equations (1), (2) and (3)) are similar to Hendry and Clements (2004) (their equations (5) to (7)). However, they compare CF with the best individual forecast but here we compare CF with forecast by the CI model (the DGP in Hendry and Clements, 2004). Also note that Harvey and Newbold (2005) investigate gains from combining the forecasts from DGP and mis-specified models, and Clark and McCracken (2009) examine methods of combining forecasts from nested models, while in contrast, we consider combining forecasts from non-nested (mis-specified) individual forecasting models and compare with models incorporating all available information directly (CI, which may be correctly specified).

where $\eta_{t+1} \sim IID(0, \sigma_\eta^2)$ and $x_t = (x_{1,t} \;\; x_{2,t})$ with each $x_{i,t}$ being $1 \times k_i$ $(i = 1, 2)$ is strictly exogenous.[5] Let $\theta = (\theta_1' \;\; \theta_2')'$. To simplify the algebra in this section and the Monte Carlo simulation in Section 4, we assume the conditional mean of x_t is zero[6]

$$x_t' = \begin{pmatrix} x_{1,t}' \\ x_{2,t}' \end{pmatrix} \sim IN_k \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \right]. \tag{6}$$

Models (CI and CF): Consider predicting y_t one-step ahead using information $x_t = (x_{1,t} \;\; x_{2,t})$ up to time t. The forecasts by CI and CF are respectively, $\hat{y}_{T+1}^{\text{CI}} = x_{1,T}\hat{\alpha}_{1,T} + x_{2,T}\hat{\alpha}_{2,T}$ and $\hat{y}_{T+1}^{\text{CF}} = w_1\hat{y}_{T+1}^{(1)} + w_2\hat{y}_{T+1}^{(2)} = w_1 x_{1,T}\hat{\beta}_{1,T} + w_2 x_{2,T}\hat{\beta}_{2,T}$, with w_i $(i = 1, 2)$ denoting the forecast combination weights.

MSFE: Note that the unconditional MSFE by CI forecast is

$$\begin{aligned} MSFE^{\text{CI}} &= E\left\{E[\hat{e}_{T+1}^2 \mid \mathcal{I}_T]\right\} = E\left\{Var_T(y_{T+1}) + [E_T(\hat{e}_{T+1})]^2\right\} \\ &= E(\eta_{T+1}^2) + E[(\theta - \hat{\alpha}_T)' x_T' x_T (\theta - \hat{\alpha}_T)] \\ &= \sigma_\eta^2 + E[\eta' X (X'X)^{-1} x_T' x_T (X'X)^{-1} X'\eta] \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2 E\{tr[x_T' x_T (T^{-1}X'X)^{-1}]\}, \end{aligned} \tag{7}$$

where $Var_T(\cdot)$ and $E_T(\cdot)$ denote the conditional variance and the conditional expectation given information $\mathcal{I}_T$ up to time T. Note that, if $x_t \sim IN_k(0, \;\; \Omega)$, then $E\{tr[x_T' x_T (T^{-1}X'X)^{-1}]\} \simeq tr\{\Omega\Omega^{-1}\} = k$, the dimension of x_t. Therefore, $MSFE^{\text{CI}} \simeq \sigma_\eta^2 + T^{-1}\sigma_\eta^2 k = \sigma_\eta^2 + O\left(\frac{k}{T}\right)$.[7]

[5]We assume $E[\eta\eta'|X] = \sigma_\eta^2 I_T$ for both DGP1 and DGP2 that we use for the Monte Carlo analysis in Section 4. Note that a dynamic model augmented with dynamic terms (such as lagged dependent variables) may also be considered, for example, as used in Stock and Watson (SW: 2002a; 2002b; 2004; 2006; 2009) and Bai and Ng (BN: 2002; 2008; 2009):

$$y_{t+h} = c + \alpha' W_t + \beta' X_t + e_{t+h},$$

where X_t is a vector of predictors and W_t is a vector of predetermined variables such as lags of y_t. In our DGP1 and DGP2 in Sections 2 and 4, however, we consider a simple case when $\alpha = \mathbf{0}$ as in some articles of SW and BN, which can be thought of as a result of a residual regression (Frisch–Waugh–Lowell theorem) after regressing y on W and regressing X on W to project out W first. We attempt to consider the simplest possible designs for Monte Carlo experiment in Section 4, which match with the discussion in Section 2. In Section 2 we assume strict exogeneity that rules out lagged dependent variables, only to simplify the algebra for the two DGPs in Section 2. The same DGPs are used in Section 4. However, all the results in this article can be extended to a more complicated model, including lagged dependent variables and other predetermined variables, without the strict exogeneity assumption.

[6]Monte Carlo analysis in Section 4 shows that dynamics in the conditional mean do not affect our general conclusions in this section.

[7]This is also explained in Stock and Watson (2006) and Bai and Ng (2009) in different ways.

Similarly, the unconditional MSFE by CF forecast is

$$\begin{aligned} MSFE^{\mathrm{CF}} &= E\left\{E[(\hat{e}_{T+1}^{\mathrm{CF}})^2 \mid \mathcal{I}_T]\right\} = E\left\{Var_T(y_{T+1}) + [E_T(\hat{e}_{T+1}^{\mathrm{CF}})]^2\right\} \\ &= \sigma_\eta^2 + E\left\{\left[E_T(y_{T+1} - \hat{y}_{T+1}^{\mathrm{CF}})\right]^2\right\} \\ &= \sigma_\eta^2 + E\left\{\left[x_T\theta - \sum_{i=1}^{2} w_i x_{i,T}\hat{\beta}_{i,T}\right]^2\right\} \\ &= \sigma_\eta^2 + E\left\{\left[x_T\theta - \sum_{i=1}^{2} w_i x_{i,T}(X_i'X_i)^{-1}X_i'Y\right]^2\right\}. \end{aligned} \tag{8}$$

Comparison: Therefore, it follows that the CF forecast is better than the CI forecast under the MSFE loss if the following condition holds:

$$\begin{aligned} &T^{-1}\sigma_\eta^2 E\{tr[x_T'x_T(T^{-1}X'X)^{-1}]\} \\ &\quad > E\left\{\left[(x_{1,T}\theta_1 - w_1 x_{1,T}\hat{\beta}_{1,T}) + (x_{2,T}\theta_2 - w_2 x_{2,T}\hat{\beta}_{2,T})\right]^2\right\}. \end{aligned} \tag{9}$$

Note that $\hat{\alpha}_T \to \theta$, *a.s.* as $T \to \infty$ for the CI model. Note also that $\hat{\beta}_T \equiv (\hat{\beta}_{1,T}'\hat{\beta}_{2,T}')' \to \theta$, *a.s.* as $T \to \infty$ for the two individual forecasting models if the two sets of predictors x_1, x_2 are orthogonal, but $\hat{\beta}_T \nrightarrow \theta$ otherwise. Therefore, as $T \to \infty$, $MSFE^{\mathrm{CI}} \leq MSFE^{\mathrm{CF}}$ always follows.

For a finite T, however, even when the CI model (3) coincides with DGP1, the squared conditional bias by $\hat{y}_{T+1}^{\mathrm{CI}}$ can be greater than that by $\hat{y}_{T+1}^{\mathrm{CF}}$. This is mostly due to the parameter estimation error in $\hat{\alpha}_T$, which is often of a larger size $\left(O\left(\frac{k}{T}\right)\right)$ compared to the parameter estimation errors in individual models each use a smaller set of regressors $\left(O\left(\frac{k}{2T}\right)\right.$ if $k_1 = k_2 = k/2$),[8] thus leaving out room for CF forecast to beat CI forecast in terms of MSFE. In general, this can be understood through the bias and variance trade-off between large and small models.[9] To illustrate the finite sample potential gain of CF

[8]The number of parameters estimated in the CF method is actually larger than the number of parameters in the CI method if the combining weights w_i are to be estimated. In this section we focus on the case when the weights are given (not estimated). In Section 3 we will discuss the case when the weights are estimated, where we explain why the CF with equal weights can outperform the CF with estimated weights and also note the benefits of shrinking the estimated weights towards the equal weights.

[9]Harvey and Newbold (2005) have the similar finding: forecasts from the true (but *estimated*) DGP do not encompass forecasts from competing mis-specified models in general, particularly when T is small. By comparing the restricted and unrestricted models Clark and McCracken (2009) note also the finite sample forecast accuracy trade-off resulted from parameter estimation noise in their simulation and in empirical studies. Note that by contrasting CF with CI here we make a fair comparison in terms of information content – both CI and CF in our framework use same amount of information (x's) but in different ways (CI direct and CF indirect).

over CI more explicitly, we consider a simplified case where $k_1 = k_2 = 1$ (thus $k = 2$) and $\Omega_{2\times 2} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ (assuming $x_t \sim IN_k(0, \Omega)$). Let $w_1 = w$ and $w_2 = 1 - w$. It can be shown that equation (9)'s LHS $\simeq 2T^{-1}\sigma_\eta^2$, while its RHS $\simeq w^2[T^{-1}\sigma_\eta^2 + \theta_2^2(1-\rho^2)] + (1-w)^2[T^{-1}\sigma_\eta^2 + \theta_1^2(1-\rho^2)] + 2w(1-w)[T^{-1}\rho\sigma_\eta^2 - \theta_1\theta_2\rho(1-\rho^2)]$.[10] Rearranging terms, equation (9) can then be written into

$$T^{-1}\sigma_\eta^2 > \frac{(1-\rho^2)[w^2\theta_2^2 + (1-w)^2\theta_1^2 - 2w(1-w)\theta_1\theta_2\rho]}{1 + 2w(1-w)(1-\rho)}, \tag{10}$$

assuming $1 + 2w(1-w)(1-\rho) > 0$. The above condition is more likely to hold (so that CF outperforms CI in MSFE) when its LHS is large. This would happen when either T is small or σ_η^2 is big. Also note that with everything else held constant, the RHS of equation (10) is getting close to 0 when ρ is approaching to 1, therefore, when x_i's are highly collinear, CF will have more chance to beat CI. In our Monte Carlo analysis in Section 4, we will consider such a simple parameter setting, for which the above analytical conclusions will be confirmed by simulation findings.

2.2. When the CI Model is not Correctly Specified

Often in real time forecasting, DGP is unknown and the collection of predictors used to forecast the variable of interest is perhaps just a subset of all relevant ones. This situation frequently occurs when some of the relevant predictors are simply unobservable. For instance, in forecasting the output growth, total expenditures on R&D and brand building may be very relevant predictors but are usually unavailable. They may thus become omitted variables for predicting output growth. To account for these more practical situations, we now examine the case when the CI model is misspecified with some relevant variables omitted. In this case, we demonstrate that CF forecast can be superior to CI forecast even in a large sample. Intuitively, this is expected to happen likely because when the CI model is also misspecified, the bias-variance trade-off between large and small models becomes more evident, thus leading to possibly better chance for CF forecast (generated from a set of small models) to outperform CI forecast (generated from one large model).

DGP2: Suppose that the true DGP involves one more set of variables $x_{3,t}$ than DGP1

$$y_{t+1} = x_{1,t}\theta_1 + x_{2,t}\theta_2 + x_{3,t}\theta_3 + \eta_{t+1}, \tag{11}$$

[10]This can be seen from the derivations in the Appendix and let $k_1 = k_2 = 1$ and $\theta_3 = 0$.

where $\eta_{t+1} \sim IID(0, \sigma_\eta^2)$ and $x_t = (x_{1,t}\ x_{2,t}\ x_{3,t})$ with each $x_{i,t}$ being $1 \times k_i$ ($i = 1, 2, 3$) is strictly exogenous. To simplify the algebra again we assume the conditional mean of x_t is zero

$$x_t' = \begin{pmatrix} x_{1,t}' \\ x_{2,t}' \\ x_{3,t}' \end{pmatrix} \sim IN_k \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix} \right]. \tag{12}$$

Models (CI and CF): Suppose we use the same CI and CF models as in the previous subsection, forecasting y_{T+1} using the CI model (3) and the CF scheme given by (1) and (2) with the information set $\{(x_{1,s}\ x_{2,s})\}_{s=0}^{T}$. The CI model in (3) is misspecified by omitting $x_{3,t}$, the first individual model in (1) omits $x_{2,t}$ and $x_{3,t}$, and the second individual model in (2) omits $x_{1,t}$ and $x_{3,t}$. The forecasts by CI and CF are therefore respectively, $\hat{y}_{T+1}^{\mathrm{CI}} = x_{1,T}\hat{\alpha}_{1,T} + x_{2,T}\hat{\alpha}_{2,T}$ and $\hat{y}_{T+1}^{\mathrm{CF}} = w_1\hat{y}_{T+1}^{(1)} + w_2\hat{y}_{T+1}^{(2)} = w_1 x_{1,T}\hat{\beta}_{1,T} + w_2 x_{2,T}\hat{\beta}_{2,T}$, with w_i ($i = 1, 2$) denoting the forecast combination weights.

Let us consider the special case $w_1 + w_2 = 1$ and let $w \equiv w_1$ hereafter. The forecast error by CI is thus:

$$\hat{e}_{T+1} = y_{T+1} - \hat{y}_{T+1}^{\mathrm{CI}} = x_{1,T}(\theta_1 - \hat{\alpha}_{1,T}) + x_{2,T}(\theta_2 - \hat{\alpha}_{2,T}) + x_{3,T}\theta_3 + \eta_{T+1}. \tag{13}$$

The forecast errors by the first and the second individual forecast are, respectively:

$$\begin{aligned} \hat{\epsilon}_{1,T+1} &= y_{T+1} - \hat{y}_{T+1}^{(1)} = x_{1,T}(\theta_1 - \hat{\beta}_{1,T}) + x_{2,T}\theta_2 + x_{3,T}\theta_3 + \eta_{T+1}, \\ \hat{\epsilon}_{2,T+1} &= y_{T+1} - \hat{y}_{T+1}^{(2)} = x_{1,T}\theta_1 + x_{2,T}(\theta_2 - \hat{\beta}_{2,T}) + x_{3,T}\theta_3 + \eta_{T+1}. \end{aligned} \tag{14}$$

Hence the forecast error by CF is:

$$\hat{e}_{T+1}^{\mathrm{CF}} = y_{T+1} - \hat{y}_{T+1}^{\mathrm{CF}} = w\hat{\epsilon}_{1,T+1} + (1 - w)\hat{\epsilon}_{2,T+1}. \tag{15}$$

Let $z_t = (x_{1,t}\ x_{2,t})$, $Var(z_t) = \Omega_{zz}$, $Cov(z_t, x_{3,t}) = \Omega_{z3}$, $\xi_{3z,T} = x_{3,T} - z_T\Omega_{zz}^{-1}\Omega_{z3}$, $Var(\xi_{3z,T}) = \Omega_{\xi_{3z}} = \Omega_{33} - \Omega_{3z}\Omega_{zz}^{-1}\Omega_{z3}$, $\theta_{23} = (\theta_2'\ \theta_3')'$, $\theta_{13} = (\theta_1'\ \theta_3')'$, $\xi_{23.1,T} = (x_{2,T} - x_{1,T}\Omega_{11}^{-1}\Omega_{12} \quad x_{3,T} - x_{1,T}\Omega_{11}^{-1}\Omega_{13})$, $\xi_{13.2,T} = (x_{1,T} - x_{2,T}\Omega_{22}^{-1}\Omega_{21} \quad x_{3,T} - x_{2,T}\Omega_{22}^{-1}\Omega_{23})$, $Var(\xi_{23.1,T}) = \Omega_{\xi_{23.1}}$, and $Var(\xi_{13.2,T}) = \Omega_{\xi_{13.2}}$.

MSFE: See Appendix for derivation of MSFEs for the CI and CF models.

Comparison: We now compare CI with CF. Assume that the DGP consists of (11) and (12). From comparing MSFEs from (39) and (46) in Appendix, the CF forecast is better than the CI forecast in MSFE if the following condition holds:

$$\begin{aligned} \theta_3'\Omega_{\xi_{3z}}\theta_3 + g_T^{\mathrm{CI}} > {} & w^2\theta_{23}'\Omega_{\xi_{23.1}}\theta_{23} + (1 - w)^2\theta_{13}'\Omega_{\xi_{13.2}}\theta_{13} \\ & + 2w(1 - w)\theta_{23}'E[\xi_{23.1,T}'\xi_{13.2,T}]\theta_{13} + g_T^{\mathrm{CF}}, \end{aligned} \tag{16}$$

where $g_T^{\mathrm{CI}} = T^{-1}(k_1 + k_2)\sigma_\eta^2$ and $g_T^{\mathrm{CF}} = T^{-1}(w^2 k_1 + (1-w)^2 k_2)\sigma_\eta^2 + 2w(1-w)E[x_{1,T}(\hat{\beta}_{1,T} - E(\hat{\beta}_{1,T}))(\hat{\beta}_{2,T} - E(\hat{\beta}_{2,T}))'x'_{2,T}]$ are both $O(T^{-1})$.

The condition (16) under which CF is better than CI can be simplified when T goes to infinity. Note that it involves both small sample and large sample effect. If we ignore $O(T^{-1})$ terms or let $T \to \infty$, (16) becomes

$$\theta_3'\Omega_{\xi_{3z}}\theta_3 > w^2\theta_{23}'\Omega_{\xi_{23.1}}\theta_{23} + (1-w)^2\theta_{13}'\Omega_{\xi_{13.2}}\theta_{13} + 2w(1-w)\theta_{23}'E[\xi'_{23.1,T}\xi_{13.2,T}]\theta_{13}. \tag{17}$$

The variance of the disturbance term in the DGP model (11) no longer involves since it only appears in g_T^{CI} and g_T^{CF}, the two terms capturing small sample effect. While this large-sample condition may still look complicated, we note that all the terms in (17) are determined only by θ and Ω in DGP2.

Remark: We also note that there is a chance that the CI forecast is even worse than two individual forecasts. Note that

$$MSFE^{\mathrm{CI}} = \sigma_\eta^2 + T^{-1}(k_1 + k_2)\sigma_\eta^2 + \theta_3'\Omega_{\xi_{3z}}\theta_3,$$

and the MSFE's by individual forecasts $\hat{y}_{T+1}^{(1)}$ and $\hat{y}_{T+1}^{(2)}$ are, respectively

$$MSFE^{(1)} = \sigma_\eta^2 + T^{-1}k_1\sigma_\eta^2 + \theta_{23}'\Omega_{\xi_{23.1}}\theta_{23},$$
$$MSFE^{(2)} = \sigma_\eta^2 + T^{-1}k_2\sigma_\eta^2 + \theta_{13}'\Omega_{\xi_{13.2}}\theta_{13}.$$

Suppose $MSFE^{(1)} > MSFE^{(2)}$, i.e., the second individual forecast is better, then CI will be worse than the two individual forecasts if

$$T^{-1}k_2\sigma_\eta^2 + \theta_3'\Omega_{\xi_{3z}}\theta_3 > \theta_{23}'\Omega_{\xi_{23.1}}\theta_{23}. \tag{18}$$

This is more likely to happen if the sample size T is not large, and/or σ_η^2 is large. The Monte Carlo analysis in Section 4 also confirms this result (see Table 2).

3. UNDERSTANDING THE FORECAST COMBINATION PUZZLE

In the empirical forecasting literature numerous papers have found that the equally-weighted forecast combination often outperforms the CF using estimated optimal forecasts. Stock and Watson (2004) refer this as a "forecast combination puzzle." Before we help illustrate analytical findings via Monte Carlo analysis in the next section, here we attempt to understand the puzzling empirical success of the CF with equal weights through some analysis. The Monte Carlo analysis in Section 4 confirms our explanation of the forecast combination puzzle. The Monte Carlo analysis

also provides some insights on the possibility that CF with equal weights can dominate CI even in a large sample.

While the weight w in CF has not yet been specified in the above analysis, we now consider CF with specific weights, in particular, the equal weights. Our aim of this section is to illustrate when and how CF with certain weights can beat CI in out-of-sample forecasting, and shed some light on the success of equally weighted CF.

Let $MSFE^{\mathrm{CI}} = E(\hat{e}_{T+1}^2) \equiv \gamma_{\mathrm{CI}}^2$, and $\gamma_i^2 \equiv E(\hat{\epsilon}_{i,T+1}^2)$ $(i = 1, 2)$ denote MSFE's by the two individual forecasts. Define $\gamma_{12} \equiv E(\hat{\epsilon}_{1,T+1}\hat{\epsilon}_{2,T+1})$. From equation (15), the MSFE of the CF forecast is

$$MSFE^{\mathrm{CF}} = w^2\gamma_1^2 + (1-w)^2\gamma_2^2 + 2w(1-w)\gamma_{12} \equiv \gamma_{\mathrm{CF}}^2(w). \tag{19}$$

CF-Mean: Consider the equally weighted CF, denoted "CF-Mean" $(w = \frac{1}{2})$:

$$\hat{y}_{T+1}^{\text{CF-Mean}} = \frac{1}{2}\hat{y}_{T+1}^{(1)} + \frac{1}{2}\hat{y}_{T+1}^{(2)}, \tag{20}$$

for which the MSFE is

$$MSFE^{\text{CF-Mean}} = E(y_{T+1} - \hat{y}_{T+1}^{\text{CF-Mean}})^2 = \frac{1}{4}(\gamma_1^2 + \gamma_2^2 + 2\gamma_{12}) \equiv \gamma_{\mathrm{CF}}^2\left(\frac{1}{2}\right). \tag{21}$$

CF-Optimal: Consider the "CF-Optimal" forecast with weight

$$w^* = \arg\min_w \gamma_{\mathrm{CF}}^2(w) = \frac{\gamma_2^2 - \gamma_{12}}{\gamma_1^2 + \gamma_2^2 - 2\gamma_{12}}, \tag{22}$$

obtained by solving $\partial\gamma_{\mathrm{CF}}^2(w)/\partial w = 0$ (as in Bates and Granger, 1969 but without assuming the individual forecasts are unbiased since we work directly on MSFE instead of error variances). Note that if we rearrange terms in (15), it becomes the Bates and Granger (1969) regression

$$\hat{\epsilon}_{2,T+1} = w(\hat{\epsilon}_{2,T+1} - \hat{\epsilon}_{1,T+1}) + \hat{e}_{T+1}^{\mathrm{CF}}, \tag{23}$$

from which estimate of w^* is obtained by the least squares.

We note that CF-Optimal always assigns a larger (smaller) weight to the better (worse) individual forecast, since the optimal weight w^* for the first individual forecast is less than $\frac{1}{2}$ if it is the worse one ($w^* = \frac{\gamma_2^2 - \gamma_{12}}{\gamma_1^2 + \gamma_2^2 - 2\gamma_{12}} < \frac{1}{2}$ if $\gamma_1^2 > \gamma_2^2$); and the weight is larger than $\frac{1}{2}$ when it is the better one ($w^* > \frac{1}{2}$ if $\gamma_1^2 < \gamma_2^2$). Also note that $w^* = \frac{1}{2}$ if $\gamma_1^2 = \gamma_2^2$. One practical problem is that w^* is unobservable.

We now explain how we may understand the puzzle in three ways, attributing the success of CF-Mean to (i) the finite sample estimation error of the forecast combining weights, (ii) the possible scenario when CF-Mean is indeed near optimal, and (iii) weak predictors. The Monte Carlo and empirical analysis in the subsequent sections confirm our arguments.

3.1. Understanding the Puzzle: When the Combining Weights are Estimated

This optimal weight w^* for CF needs to be estimated in practice, but it provides guidance for our analysis regarding the virtue of equal-weights in CF which empirically is often found to work better than many sophisticatedly estimated weights (Stock and Watson, 2004; Timmermann, 2006).

In practice, w^* may be estimated and the consistently estimated weight $\hat{w}$ may converge to w^* in large sample. When the in-sample estimation size T is large we use CF-Optimal (Bates and Granger, 1969; Granger and Ramanathan, 1984). However, when the noise is large and T is small, the estimated weight $\hat{w}$ may be in some distance away from w^*, and the gap between $\gamma^2_{\text{CF}}(\hat{w})$ and $\gamma^2_{\text{CF}}(w^*)$ may be wide enough such that it is possible to have the following ranking

$$\gamma^2_{\text{CF}}(w^*) < \gamma^2_{\text{CF}}\left(\frac{1}{2}\right) < \gamma^2_{\text{CF}}(\hat{w}). \tag{24}$$

Therefore, when the noise is large and T is small, we may be better off by using the CF-Mean instead of estimating the weights. Similarly, Smith and Wallis (2009) address the forecast combination puzzle by attributing to the effect of finite sample estimation error of the combining weights.

To explore more about weighting in CF, we further consider shrinkage estimators for w. In case when the above ranking of (24) holds, we can shrink the estimated weight $\hat{w}$ towards the equal weight $\frac{1}{2}$ to reduce the MSFE. We have discussed three alternative CF weights: (a) $w = \hat{w}$, (b) $w = \frac{1}{2}$, and (c) $w = w^*$. It is likely that w^* may be different from both $\hat{w}$ and $\frac{1}{2}$. The relative performance of CF with $\hat{w}$ and CF-Mean depends on which of $\hat{w}$ and $\frac{1}{2}$ is closer to w^*. Depending on the relative distance between $\hat{w}$ and w^*, between $\frac{1}{2}$ and w^*, and between $\hat{w}$ and $\frac{1}{2}$, the shrinkage of $\hat{w}$ towards $\frac{1}{2}$ may or may not work. The common practice of shrinking $\hat{w}$ towards $\frac{1}{2}$ may improve the combined forecasts as long as shrinking $\hat{w}$ towards $\frac{1}{2}$ is also to shrink $\hat{w}$ towards w^*.

As we will see from the simulation results in Section 4, shrinkage of $\hat{w}$ towards $\frac{1}{2}$ works quite well when the noise in the DGP is large and when the in-sample size T is small. When the noise is not large or T is large, CI is usually the best when it is correctly specified for the DGP. However, when CI is not correctly specified for the DGP, CI can be beaten by CF even in a large sample. The CF with $\hat{w}$, that is obtained

from the *R*egression *A*pproach for weights as suggested by Granger and Ramanathan (1984), denoted as CF-RA, and its shrinkage version towards the equal weights, denoted as CF-RA(κ) (the shrinkage parameter κ will be detailed in Section 4) generally works quite well. As Diebold and Pauly (1990) point out, CF-RA with no shrinkage (with $\kappa = 0$) and CF-Mean may be considered as two polar cases of the shrinkage. Of course, we note that more shrinkage to the equal weights is not necessarily better. However, if the weights are estimated when the noise is large and T is small, the sampling error (estimation error) may be very large to make the forecast error variance very large as well. The shrinkage toward CF-Mean is to reduce the variance at the cost of increasing the forecast bias. In general, the MSFE (the sum of the forecast error variance and squared bias) may be reduced by the shrinkage, which we will observe from the Monte Carlo results in Section 4.

3.2. Understanding the Puzzle: When CF-Mean is Close to CF-Optimal

However, we note that the above explanation for the success of CF-Mean attributing to the finite sample estimation error (as in Smith and Wallis, 2009 and as illustrated above) holds probably only when the unobservable optimal combination weight w^* is close to $\frac{1}{2}$ such that CF-Mean is approaching CF-Optimal hence dominating other sophisticated combinations where estimation errors often involve. It is unlikely that CF-Mean would outperform other CF with weights obtained by the regression equivalent of w^* when w^* is very close to 1 (or 0). Such values of w^* happen when the first (second) individual forecast is clearly better than or encompasses the second (first) individual forecast such that combination of the two has no gains. See Hendry and Clements (2004) for illustrations of situations where combination forecast gains over individual ones.

Therefore, in order to shed more light on the empirical success of simple average forecast combination, i.e., the CF-Mean, it is worth investigating under what kind of DGP structures and parameterization one could have $w^* \simeq \frac{1}{2}$ so that CF-Optimal $\simeq$ CF-Mean. We consider again DGP2 [equations (11) and (12)] discussed in Section 2.2 where the CI model is misspecified. The DGP1 in Section 2.1 where the CI model is correctly specified for the DGP is actually a special case of equation (11) when we let $\theta_3 \equiv \mathbf{0}$. First, we note again that $w^* = \frac{1}{2}$ if $\gamma_1^2 = \gamma_2^2$. Second, from the discussions in Section 2.2 we have

$$\begin{aligned} \gamma_1^2 &\equiv MSFE^{(1)} = \sigma_\eta^2 + T^{-1}k_1\sigma_\eta^2 + \left(\theta_2' \ \theta_3'\right)\Omega_{\xi_{23.1}}\begin{pmatrix}\theta_2\\ \theta_3\end{pmatrix}, \\ \gamma_2^2 &\equiv MSFE^{(2)} = \sigma_\eta^2 + T^{-1}k_2\sigma_\eta^2 + \left(\theta_1' \ \theta_3'\right)\Omega_{\xi_{13.2}}\begin{pmatrix}\theta_1\\ \theta_3\end{pmatrix}, \end{aligned} \tag{25}$$

where it is easy to show that

$$\Omega_{\xi_{23.1}} = \begin{pmatrix} \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} & \Omega_{23} - \Omega_{21}\Omega_{11}^{-1}\Omega_{13} \\ \Omega_{32} - \Omega_{31}\Omega_{11}^{-1}\Omega_{12} & \Omega_{33} - \Omega_{31}\Omega_{11}^{-1}\Omega_{13} \end{pmatrix},$$

and

$$\Omega_{\xi_{13.2}} = \begin{pmatrix} \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} & \Omega_{13} - \Omega_{12}\Omega_{22}^{-1}\Omega_{23} \\ \Omega_{31} - \Omega_{32}\Omega_{22}^{-1}\Omega_{21} & \Omega_{33} - \Omega_{32}\Omega_{22}^{-1}\Omega_{23} \end{pmatrix}.$$

Therefore, to make $\gamma_1^2 = \gamma_2^2$ (so that $w^* = \frac{1}{2}$) one sufficient set of conditions is $\theta_1 = \theta_2$ (implying $k_1 = k_2$) and $\Omega_{\xi_{23.1}} = \Omega_{\xi_{13.2}}$. The latter happens when $\Omega_{11} = \Omega_{22}$ and $\Omega_{13} = \Omega_{23}$. Intuitively, when the two individual information sets matter about the same in explaining the variable of interest, their variations (signal strengths) are also about the same, and they correlate with the omitted information set quite similarly, the resulting forecast performances of the two individual forecasts are thus about equal.

In our Monte Carlo study in Section 4, we consider the three designs of DGPs in Panels A, B, C of Table 2, such that the underlying optimal combination weight w^* is $\frac{1}{2}$. In these three designs of DGPs in Panels A, B, C of Table 2, we set $\theta_1 = \theta_2 = 0.3$ (with $k_1 = k_2 = 1$) and $\Omega_{11} = \Omega_{22} = 1$ and $\Omega_{13} = \Omega_{23} = 0.7$ or -0.7.

In addition, we also consider one exceptional case where we let $\theta_1 > \theta_2$ to make $\gamma_1^2 < \gamma_2^2$ so that $w^* > \frac{1}{2}$ to see how CF with different weights perform in comparison with CI. In the design of DGP in Panel D of Table 2, we set $\theta_1 = 3\theta_2 = 0.6$. Other cases such as $\Omega_{11} > \Omega_{22}$ give similar results (not reported). These four Monte Carlo cases will be detailed in Section 4, where we confirm our understanding of the forecast combination puzzle discussed in this section.

3.3. Understanding the Puzzle: When Predictors are Weak

Clark and McCracken (2009) argue that often in practical reality, the predictive contents of some variables of interest is quite low and hard to predict, especially for forecasting financial returns in the conditional mean. Likewise, the different individual information sets used to predict such variables in the (near) efficient financial markets are performing quite similarly (all equally bad, perhaps). When all or most of predictors are weak, a simple average combination of individual forecasts is often desirable since in such a situation CF-Mean may be quite close to CF-Optimal. We illustrate in Section 5 through an empirical study on forecasting the equity premium. Asset prices are hard to predict and oftentimes the predictors used to generate forecasts have quite limited predictive power, making them "weak predictors."

4. MONTE CARLO ANALYSIS

In this section we conduct Monte Carlo experiments in the context of Section 2 to illustrate under what specific situations CF can be better than CI in out-of-sample forecasting.

4.1. DGPs: Two Cases

We consider the same two cases that we considered in Section 2 – when the CI model is correctly specified for the DGP (corresponding to Section 2.1) and when it is not (corresponding to Section 2.2). We use the following two DGPs:

DGP1: with $x_t = (x_{1,t}\ x_{2,t})$, so that the CI model in (3) is correctly specified:

$$\begin{aligned} y_{t+1} &= x_{1,t}\theta_1 + x_{2,t}\theta_2 + \eta_{t+1},\ \eta_t \sim N(0, \sigma_\eta^2), \\ x_{i,t} &= \phi_i x_{i,t-1} + v_{i,t},\ v_t = (v_{1,t}\ v_{2,t}) \sim N(0, \Omega_{2\times 2}), \end{aligned} \tag{26}$$

DGP2: with $x_t = (x_{1,t}\ x_{2,t}\ x_{3,t})$, so that the CI model in (3) is not correctly specified:

$$\begin{aligned} y_{t+1} &= x_{1,t}\theta_1 + x_{2,t}\theta_2 + x_{3,t}\theta_3 + \eta_{t+1},\ \eta_t \sim N(0, \sigma_\eta^2), \\ x_{i,t} &= \phi_i x_{i,t-1} + v_{i,t},\ v_t = (v_{1,t}\ v_{2,t}\ v_{3,t}) \sim N(0, \Omega_{3\times 3}), \end{aligned} \tag{27}$$

where all $v_{i,t}$'s are independent of η_t. We consider different degrees of signal to noise with seven different values of $\sigma_\eta = 2^j$ $(j = -2, -1, 0, 1, 2, 3, 4)$.

The pseudo random samples for $t = 1, \ldots, R + P + 1$ are generated and R observations are used for the in-sample parameter estimation (with the fixed rolling window of size R) and the last P observations are used for pseudo real time out-of-sample forecast evaluation.[11] We experiment with $R = 100, 1000$, $P = 100$. The number of Monte Carlo replications is 1000 for $R = 100$ and 100 for $R = 1000$.

Different specifications for covariance matrix Ω and coefficient vector θ are used as discussed in Sections 2 and 3. We consider two sets of the different parameter values of Ω in Table 1, and four sets of different parameter values of Ω and θ in Table 2. In both Tables 1 and 2, all ϕ_i's are set at zero as the results are similar for different values of ϕ_i.

[11]The notation of R and P is adopted from West (1996). As we use a rolling forecasting scheme to estimate parameters using the R observations, the notation T that was used to denote the sample size for the in-sample estimation in Sections 2 and 3 is now R in Sections 4 and 5.

With $\Omega_{2\times 2} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ for DGP1, we have $\text{var}(\hat{\alpha}_R) = R^{-1}\sigma_\eta^2\Omega^{-1}$, and $\text{var}(\hat{\alpha}_{i,R}) = R^{-1}\sigma_\eta^2(1-\rho^2)^{-1}$ for $i = 1, 2$. In Table 1 we consider $\rho = corr(x_1, x_2) = 0$ and 0.8, measuring two different degrees of collinearity in the CI model.

4.2. CF with Estimated Weights and Its Shrinkage Toward CF-Mean

One of the CF methods we use is the Regression Approach (RA) for combining forecasts as suggested by Granger and Ramanathan (1984), denoted as CF-RA,

$$y_{t+1} = \text{intercept} + w_1\hat{y}_{t+1}^{(1)} + w_2\hat{y}_{t+1}^{(2)} + \text{error}, \qquad t = T_0, \ldots, R, \tag{28}$$

where the pseudo out-of-sample forecast is made for $t = T_0, \ldots, R$ with T_0 the time when the first pseudo out-of-sample forecast is generated (we choose it at the middle point of each rolling window). The three versions of the CF-RA methods are considered as in Granger and Ramanathan (1984), namely, (a) CF-RA1 for the unconstrained regression approach forecast combination, (b) CF-RA2 for the constrained regression approach forecast combination with zero intercept and the unit sum of the weights $w_1 + w_2 = 1$, and (c) CF-RA3 for the constrained regression approach forecast combination with zero intercept but without restricting the sum of the weights.

To illustrate more the parameter estimation effect on combination weights, we also consider CF with shrinkage weights based on CF-RA3. Let CF-RA3(κ) denote the shrinkage forecasts considered in Stock and Watson (2004, p. 412) with the shrinkage parameter κ controlling for the amount of shrinkage on CF-RA3 towards the equal weighting (CF-Mean). The shrinkage weight used is $w_{i,t} = \lambda\hat{w}_{i,t} + (1-\lambda)/N$ $(i = 1, 2)$ with $\lambda = \max\{0, 1 - \kappa N/(t - h - T_0 - N)\}$, $N = 2$ (the number of individual forecasts), and $h = 1$ (one step ahead forecast).[12] For simplicity we consider a spectrum of different values of κ, that are chosen such that CF-RA3(κ) for the largest chosen value of κ is closest to CF-Mean. We choose ten different values of κ with equal increment depending on the in-sample size R as presented in Tables 1 and 2.

4.3. Monte Carlo Results

Table 1 presents the Monte Carlo results for DGP1, for which we simulate two different cases with $\Omega_{2\times 2}$ being diagonal (Panel A, $\omega = 0$) and with $\Omega_{2\times 2}$ being non-diagonal (Panel B, $\omega = 0.8$).

[12]Stock and Watson (2009) show the various forecasting methods (such as Bayesian methods, Bagging, etc.) in the shrinkage representations.

Table 2 presents the Monte Carlo results for DGP2, for which the CI model is not correctly specified as it omits x_{3t}. We simulate four different cases with different values for $\Omega_{3\times3}$ and θ where unless specified otherwise we let $\theta_1 = \theta_2$, $\Omega_{11} = \Omega_{22}$, and $\Omega_{13} = \Omega_{23}$ to make optimal weight $w^* = \frac{1}{2}$.

TABLE 1 Monte Carlo simulation (when CI model is the DGP). This set of tables presents the performance of each forecasting schemes for predicting y_{t+1} out-of-sample where y_t is by DGP: $y_{t+1} = x_t\theta + \eta_{t+1}, \eta_t \sim N(0, \sigma_\eta^2)$; $x_{it} = \phi_i x_{it-1} + v_{it}, v_t \sim N(0, \Omega), i = 1, 2$. We report the out-of-sample MSFE of each forecasting scheme, where bolded term indicates smaller-than-CI case and the smallest number among them is highlighted

Panel A. No correlation: $\Omega = \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right); \phi_i = 0; \theta = \left(\begin{smallmatrix}0.5\\ 0.5\end{smallmatrix}\right)$
MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = l$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	4.9798	1.9708	1.2403	1.0493	1.0050	**0.9935**	**0.9900**
$\hat{y}^{(2)}$	4.9225	1.9879	1.2404	1.0507	1.0053	**0.9937**	**0.9903**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.1256	1.0930	1.0870	1.0810	1.0584	1.0353	1.0314
CF-RA2	3.0124	1.5074	1.1322	1.0357	1.0138	1.0073	1.0055
CF-RA3 ($\kappa = 0$)	1.1752	1.1004	1.0758	1.0563	1.0284	1.0095	1.0061
CF-RA3 ($\kappa = 1$)	1.1705	1.0965	1.0719	1.0526	1.0253	1.0071	1.0038
CF-RA3 ($\kappa = 3$)	1.1860	1.0938	1.0655	1.0458	1.0195	1.0026	**0.9995**
CF-RA3 ($\kappa = 5$)	1.2279	1.0984	1.0611	1.0396	1.0143	**0.9986**	**0.9956**
CF-RA3 ($\kappa = 7$)	1.2992	1.1101	1.0584	1.0341	1.0097	**0.9951**	**0.9923**
CF-RA3 ($\kappa = 9$)	1.3984	1.1288	1.0578	1.0293	1.0055	**0.9922**	**0.9895**
CF-RA3 ($\kappa = 11$)	1.5256	1.1545	1.0589	1.0251	1.0020	**0.9898**	**0.9872**
CF-RA3 ($\kappa = 13$)	1.6806	1.1875	1.0620	1.0216	**0.9990**	**0.9879**	**0.9854**
CF-RA3 ($\kappa = 15$)	1.8636	1.2272	1.0670	1.0188	**0.9965**	**0.9866**	**0.9842**
CF-RA3 ($\kappa = 17$)	2.0744	1.2743	1.0739	1.0166	**0.9946**	**0.9857**	**0.9834**
CF-RA3 ($\kappa = 19$)	2.3147	1.3288	1.0826	1.0152	**0.9933**	**0.9854**	**0.9832**
CF-Mean	2.9550	1.4763	1.1091	1.0142	**0.9922**	**0.9866**	**0.9845**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	5.0616	2.0509	1.2669	1.0573	1.0138	1.0051	1.0009
$\hat{y}^{(2)}$	4.8499	1.9921	1.2334	1.0665	1.0107	1.0021	**0.9998**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0111	1.0075	1.0076	1.0070	1.0072	1.0060	1.0071
CF-RA2	1.0111	1.0075	1.0076	1.0070	1.0072	1.0060	1.0071
CF-RA3 ($\kappa = 0$)	1.0174	1.0067	1.0079	1.0071	1.0059	1.0037	1.0019
CF-RA3 ($\kappa = 1$)	1.0174	1.0067	1.0079	1.0071	1.0059	1.0037	1.0019
CF-RA3 ($\kappa = 28$)	1.0174	1.0067	1.0079	1.0071	1.0059	1.0037	1.0019
CF-RA3 ($\kappa = 55$)	1.1185	1.0324	1.0141	1.0072	1.0040	1.0026	1.0008
CF-RA3 ($\kappa = 82$)	1.2370	1.0632	1.0211	1.0083	1.0034	1.0022	1.0004
CF-RA3 ($\kappa = 109$)	1.3997	1.1058	1.0309	1.0100	1.0030	1.0018	1.0001
CF-RA3 ($\kappa = 136$)	1.6051	1.1603	1.0432	1.0124	1.0028	1.0015	**0.9998**
CF-RA3 ($\kappa = 163$)	1.8578	1.2266	1.0582	1.0155	1.0028	1.0013	**0.9996**
CF-RA3 ($\kappa = 190$)	2.1532	1.3052	1.0759	1.0192	1.0030	1.0011	**0.9994**
CF-RA3 ($\kappa = 217$)	2.4929	1.3952	1.0962	1.0236	1.0034	1.0010	**0.9993**
CF-RA3 ($\kappa = 244$)	2.8784	1.4974	1.1192	1.0287	1.0041	1.0010	**0.9993**
CF-Mean	2.9463	1.5156	1.1233	1.0296	1.0042	1.0010	**0.9993**

(continued)

TABLE 1 Continued

Panel B. High correlation: $\Omega = \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right); \phi_i = 0; \theta = \left(\begin{smallmatrix} 0.5 \\ 0.5 \end{smallmatrix}\right)$
MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = 1$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	2.4295	1.3409	1.0808	1.0110	**0.9948**	**0.9911**	**0.9896**
$\hat{y}^{(2)}$	2.3969	1.3545	1.0788	1.0132	**0.9942**	**0.9913**	**0.9899**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0589	1.0568	1.0580	1.0571	1.0593	1.0406	1.0335
CF-RA2	1.1566	1.0455	1.0177	1.0108	1.0097	1.0085	1.0088
CF-RA3 ($\kappa = 0$)	1.0403	1.0366	1.0356	1.0343	1.0296	1.0125	1.0070
CF-RA3 ($\kappa = 1$)	1.0357	1.0331	1.0319	1.0307	1.0261	1.0099	1.0047
CF-RA3 ($\kappa = 3$)	1.0310	1.0265	1.0251	1.0240	1.0195	1.0050	1.0004
CF-RA3 ($\kappa = 5$)	1.0279	1.0214	1.0191	1.0179	1.0136	1.0007	**0.9967**
CF-RA3 ($\kappa = 7$)	1.0279	1.0171	1.0138	1.0125	1.0083	**0.9970**	**0.9934**
CF-RA3 ($\kappa = 9$)	1.0310	1.0140	1.0091	1.0076	1.0035	**0.9939**	**0.9908**
CF-RA3 ($\kappa = 11$)	1.0372	1.0121	1.0052	1.0033	**0.9994**	**0.9913**	**0.9886**
CF-RA3 ($\kappa = 13$)	1.0450	1.0109	1.0020	**0.9997**	**0.9959**	**0.9893**	**0.9870**
CF-RA3 ($\kappa = 15$)	1.0558	1.0113	**0.9996**	**0.9966**	**0.9931**	**0.9879**	**0.9860**
CF-RA3 ($\kappa = 17$)	1.0698	1.0125	**0.9980**	**0.9942**	**0.9908**	**0.9870**	**0.9855**
CF-RA3 ($\kappa = 19$)	1.0868	1.0148	**0.9970**	**0.9924**	**0.9891**	**0.9867**	**0.9855**
CF-Mean	1.1333	1.0245	**0.9974**	**0.9905**	**0.9876**	**0.9882**	**0.9875**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	2.4803	1.3861	1.0842	1.0204	1.0029	1.0016	1.0006
$\hat{y}^{(2)}$	2.3791	1.3458	1.0957	1.0206	1.0052	**0.9983**	**0.9980**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0063	1.0051	1.0052	1.0056	1.0038	1.0052	1.0053
CF-RA2	1.1327	1.0399	1.0109	1.0027	1.0002	1.0000	1.0000
CF-RA3 ($\kappa = 0$)	1.0047	1.0032	1.0038	1.0041	1.0031	1.0023	1.0005
CF-RA3 ($\kappa = 1$)	1.0047	1.0032	1.0038	1.0041	1.0031	1.0023	1.0005
CF-RA3 ($\kappa = 28$)	1.0063	1.0028	1.0031	1.0031	1.0023	1.0016	**0.9999**
CF-RA3 ($\kappa = 55$)	1.0095	1.0032	1.0028	1.0022	1.0015	1.0009	**0.9994**
CF-RA3 ($\kappa = 82$)	1.0174	1.0051	1.0027	1.0015	.0009	1.0003	**0.9991**
CF-RA3 ($\kappa = 109$)	1.0284	1.0075	1.0028	1.0009	1.0003	**0.9998**	**0.9988**
CF-RA3 ($\kappa = 136$)	1.0411	1.0111	1.0032	1.0005	**0.9998**	**0.9994**	**0.9986**
CF-RA3 ($\kappa = 163$)	1.0585	1.0158	1.0039	1.0003	**0.9994**	**0.9990**	**0.9985**
CF-RA3 ($\kappa = 190$)	1.0774	1.0213	1.0048	1.0002	**0.9991**	**0.9988**	**0.9985**
CF-RA3 ($\kappa = 217$)	1.1011	1.0276	1.0060	1.0002	**0.9989**	**0.9986**	**0.9986**
CF-RA3 ($\kappa = 244$)	1.1264	1.0351	1.0075	1.0004	**0.9988**	**0.9984**	**0.9988**
CF-Mean	1.1311	1.0367	1.0078	1.0005	**0.9988**	**0.9984**	**0.9988**

The four cases for Table 2 are presented in Panel A (where x_{1t} and x_{2t} are highly positively correlated with the omitted variable x_{3t}), in Panel B (where x_{1t} and x_{2t} are highly negatively correlated with the omitted variable x_{3t}), in Panel C (where everything is the same as in Panel B except with smaller θ_3), and in Panel D (where everything is the same as in Panel B except $\theta_1 = 3\theta_2$ to make $w^* \gg \frac{1}{2}$). See Section 3.2 for the discussion on this

parameterization in Panel D. In both Tables 1 and 2, all ρ_i's are set at zero as the results are similar for different values of ρ_i reflecting dynamics in x_{it} (and thus not reported for space).

TABLE 2 Monte Carlo simulation (when CI model is not the DGP). This set of tables presents the performance of each forecasting schemes for predicting y_{t+1} out-of-sample where y_t is by DGP: $y_{t+1} = x_t\theta + \eta_{t+1},\ \eta_t \sim N(0, \sigma_\eta^2);\ x_{it} = \phi_i x_{it-1} + v_{it},\ v_t \sim N(0, \Omega),\ i = 1, 2, 3$. Variable x_{3t} is omitted in each CF and CI schemes

Panel A. High positive correlations with the omitted variable: $\Omega = \begin{pmatrix} 1 & 0.6 & 0.7 \\ 0.6 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{pmatrix}; \phi_i = 0; \theta = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.6 \end{pmatrix}$

MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = 1$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	1.9823	1.5051	1.1665	1.0406	1.0011	**0.9932**	**0.9897**
$\hat{y}^{(2)}$	1.9761	1.5111	1.1656	1.0361	1.0024	**0.9919**	**0.9909**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0625	1.0620	1.0628	1.0615	1.0620	1.0434	1.0326
CF-RA2	1.2166	1.1157	1.0444	1.0179	1.0105	1.0077	1.0068
CF-RA3 ($\kappa = 0$)	1.0458	1.0428	1.0426	1.0396	1.0328	1.0154	1.0072
CF-RA3 ($\kappa = 1$)	1.0420	1.0391	1.0387	1.0359	1.0292	1.0126	1.0048
CF-RA3 ($\kappa = 3$)	1.0377	1.0333	1.0317	1.0290	1.0227	1.0075	1.0005
CF-RA3 ($\kappa = 5$)	1.0367	1.0296	1.0260	1.0229	1.0167	1.0029	**0.9967**
CF-RA3 ($\kappa = 7$)	1.0391	1.0279	1.0211	1.0174	1.0113	**0.9989**	**0.9934**
CF-RA3 ($\kappa = 9$)	1.0448	1.0281	1.0174	1.0126	1.0064	**0.9954**	**0.9907**
CF-RA3 ($\kappa = 11$)	1.0544	1.0306	1.0148	1.0084	1.0022	**0.9925**	**0.9884**
CF-RA3 ($\kappa = 13$)	1.0673	1.0351	1.0133	1.0050	**0.9985**	**0.9902**	**0.9867**
CF-RA3 ($\kappa = 15$)	1.0840	1.0418	1.0129	1.0022	**0.9953**	**0.9884**	**0.9855**
CF-RA3 ($\kappa = 17$)	1.1035	1.0503	1.0134	1.0000	**0.9928**	**0.9871**	**0.9849**
CF-RA3 ($\kappa = 19$)	1.1269	1.0610	1.0152	**0.9986**	**0.9908**	**0.9864**	**0.9848**
CF-Mean	1.1927	1.0928	1.0229	**0.9978**	**0.9885**	**0.9868**	**0.9864**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	2.0644	1.5331	1.1696	1.0452	1.0102	1.0024	**0.9998**
$\hat{y}^{(2)}$	2.0045	1.5254	1.1763	1.0523	1.0118	1.0035	**0.9986**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0045	1.0078	1.0043	1.0052	1.0050	1.0056	1.0016
CF-RA2	1.2092	1.1185	1.0305	1.0077	1.0027	1.0022	**0.9990**
CF-RA3 ($\kappa = 0$)	1.0025	1.0045	1.0042	1.0025	1.0030	1.0038	**0.9985**
CF-RA3 ($\kappa = 1$)	1.0025	1.0045	1.0041	1.0025	1.0030	1.0037	**0.9985**
CF-RA3 ($\kappa = 28$)	1.0040	1.0060	1.0028	1.0018	1.0022	1.0029	**0.9981**
CF-RA3 ($\kappa = 55$)	1.0106	1.0100	1.0025	1.0013	1.0016	1.0021	**0.9979**
CF-RA3 ($\kappa = 82$)	1.0226	1.0166	1.0031	1.0012	1.0011	1.0015	**0.9977**
CF-RA3 ($\kappa = 109$)	1.0392	1.0259	1.0047	1.0013	1.0007	1.0010	**0.9976**
CF-RA3 ($\kappa = 136$)	1.0608	1.0379	1.0072	1.0019	1.0005	1.0005	**0.9976**
CF-RA3 ($\kappa = 163$)	1.0880	1.0525	1.0107	1.0026	1.0004	1.0002	**0.9976**
CF-RA3 ($\kappa = 190$)	1.1197	1.0698	1.0151	1.0037	1.0005	**1.0000**	**0.9978**
CF-RA3 ($\kappa = 217$)	1.1564	1.0897	1.0205	1.0052	1.0006	**0.9999**	**0.9980**
CF-RA3 ($\kappa = 244$)	1.1981	1.1123	1.0267	1.0069	1.0010	**0.9999**	**0.9982**
CF-Mean	1.2056	1.1163	1.0279	1.0072	1.0010	**0.9999**	**0.9983**

(continued)

TABLE 2 Continued

Panel B. High negative correlations with the omitted variable: $\Omega = \begin{pmatrix} 1 & 0.6 & -0.7 \\ 0.6 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{pmatrix}; \phi_i = 0; \theta = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.6 \end{pmatrix}$

MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = 1$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	**0.9948**	**0.9915**	**0.9902**	**0.9906**	**0.9897**	**0.9900**	**0.9891**
$\hat{y}^{(2)}$	**0.9943**	**0.9920**	**0.9899**	**0.9900**	**0.9896**	**0.9891**	**0.9901**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0510	1.0396	1.0328	1.0293	1.0288	1.0268	1.0281
CF-RA2	1.0081	1.0077	1.0071	1.0072	1.0069	1.0073	1.0069
CF-RA3 ($\kappa = 0$)	1.0191	1.0142	1.0062	1.0045	1.0055	1.0040	1.0039
CF-RA3 ($\kappa = 1$)	1.0162	1.0114	1.0038	1.0023	1.0032	1.0018	1.0017
CF-RA3 ($\kappa = 3$)	1.0110	1.0065	**0.9996**	**0.9982**	**0.9989**	**0.9978**	**0.9977**
CF-RA3 ($\kappa = 5$)	1.0062	1.0020	**0.9957**	**0.9947**	**0.9952**	**0.9943**	**0.9942**
CF-RA3 ($\kappa = 7$)	1.0019	**0.9980**	**0.9925**	**0.9917**	**0.9921**	**0.9913**	**0.9912**
CF-RA3 ($\kappa = 9$)	**0.9981**	**0.9945**	**0.9899**	**0.9892**	**0.9895**	**0.9888**	**0.9887**
CF-RA3 ($\kappa = 11$)	**0.9948**	**0.9918**	**0.9877**	**0.9873**	**0.9874**	**0.9869**	**0.9868**
CF-RA3 ($\kappa = 13$)	**0.9924**	**0.9895**	**0.9860**	**0.9858**	**0.9858**	**0.9854**	**0.9853**
CF-RA3 ($\kappa = 15$)	**0.9900**	**0.9878**	**0.9849**	**0.9849**	**0.9848**	**0.9845**	**0.9844**
CF-RA3 ($\kappa = 17$)	**0.9885**	**0.9866**	**0.9843**	**0.9845**	**0.9843**	**0.9841**	**0.9840**
CF-RA3 ($\kappa = 19$)	**0.9876**	**0.9861**	**0.9843**	**0.9847**	**0.9843**	**0.9842**	**0.9841**
CF-Mean	**0.9876**	**0.9868**	**0.9861**	**0.9869**	**0.9863**	**0.9862**	**0.9863**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	1.0039	1.0012	1.0004	1.0002	**0.9996**	**0.9993**	**0.9990**
$\hat{y}^{(2)}$	1.0024	1.0015	**0.9982**	1.0007	**0.9992**	**0.9991**	**0.9987**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0048	1.0060	1.0056	1.0072	1.0028	1.0035	1.0029
CF-RA2	1.0019	1.0015	1.0012	1.0014	1.0012	1.0008	1.0002
CF-RA3 ($\kappa = 0$)	1.0039	1.0040	1.0032	1.0037	**0.9997**	1.0017	1.0011
CF-RA3 ($\kappa = 1$)	1.0039	1.0040	1.0031	1.0037	**0.9997**	1.0017	1.0011
CF-RA3 ($\kappa = 28$)	1.0029	1.0032	1.0021	1.0029	**0.9993**	1.0011	1.0005
CF-RA3 ($\kappa = 55$)	1.0019	1.0025	1.0013	1.0022	**0.9990**	1.0005	**1.0000**
CF-RA3 ($\kappa = 82$)	1.0014	1.0017	1.0005	1.0016	**0.9987**	1.0000	**0.9995**
CF-RA3 ($\kappa = 109$)	1.0010	1.0010	**0.9999**	1.0011	**0.9986**	**0.9996**	**0.9991**
CF-RA3 ($\kappa = 136$)	1.0005	1.0005	**0.9994**	1.0007	**0.9985**	**0.9993**	**0.9988**
CF-RA3 ($\kappa = 163$)	1.0000	1.0002	**0.9989**	1.0003	**0.9985**	**0.9991**	**0.9986**
CF-RA3 ($\kappa = 190$)	**0.9995**	**0.9998**	**0.9987**	1.0001	**0.9986**	**0.9989**	**0.9985**
CF-RA3 ($\kappa = 217$)	**0.9995**	**0.9998**	**0.9984**	**0.9999**	**0.9988**	**0.9989**	**0.9985**
CF-RA3 ($\kappa = 244$)	**0.9995**	**0.9995**	**0.9983**	**0.9999**	**0.9990**	**0.9989**	**0.9985**
CF-Mean	**0.9995**	**0.9995**	**0.9983**	**0.9998**	**0.9991**	**0.9989**	**0.9985**

(continued)

First, we observe that results presented in Tables 1 and 2 share some common features: MSFE increases with σ_η (the noise in the DGP), but as σ_η grows, CF-RA3(κ) and CF-Mean become better and better and can beat the CI model (whether correctly specified or not). For smaller R (=100), there are more chances for CF to outperform CI

TABLE 2 Continued

Panel C. High negative correlations with the omitted variable and relatively small θ_3:

$$\Omega = \begin{pmatrix} 1 & 0.6 & -0.7 \\ 0.6 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{pmatrix}; \phi_i = 0; \theta = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.2 \end{pmatrix}$$

MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = 1$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	1.3639	1.0973	1.0180	**0.9981**	**0.9911**	**0.9904**	**0.9891**
$\hat{y}^{(2)}$	1.3552	1.0988	1.0178	**0.9961**	**0.9914**	**0.9895**	**0.9902**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0606	1.0611	1.0643	1.0558	1.0395	1.0291	1.0289
CF-RA2	1.0879	1.0300	1.0139	1.0091	1.0075	1.0075	1.0068
CF-RA3 ($\kappa = 0$)	1.0421	1.0406	1.0396	1.0245	1.0129	1.0059	1.0043
CF-RA3 ($\kappa = 1$)	1.0384	1.0369	1.0358	1.0214	1.0102	1.0036	1.0020
CF-RA3 ($\kappa = 3$)	1.0322	1.0300	1.0286	1.0156	1.0053	**0.9994**	**0.9980**
CF-RA3 ($\kappa = 5$)	1.0272	1.0241	1.0222	1.0103	1.0010	**0.9957**	**0.9945**
CF-RA3 ($\kappa = 7$)	1.0248	1.0190	1.0162	1.0056	**0.9972**	**0.9925**	**0.9915**
CF-RA3 ($\kappa = 9$)	1.0248	1.0146	1.0110	1.0014	**0.9939**	**0.9899**	**0.9890**
CF-RA3 ($\kappa = 11$)	1.0248	1.0113	1.0064	**0.9979**	**0.9912**	**0.9878**	**0.9870**
CF-RA3 ($\kappa = 13$)	1.0272	1.0088	1.0025	**0.9949**	**0.9890**	**0.9862**	**0.9855**
CF-RA3 ($\kappa = 15$)	1.0309	1.0069	**0.9991**	**0.9925**	**0.9874**	**0.9851**	**0.9846**
CF-RA3 ($\kappa = 17$)	1.0359	1.0059	**0.9965**	**0.9907**	**0.9863**	**0.9845**	**0.9841**
CF-RA3 ($\kappa = 19$)	1.0433	1.0059	**0.9944**	**0.9894**	**0.9857**	**0.9845**	**0.9842**
CF-Mean	1.0656	1.0088	**0.9921**	**0.9886**	**0.9865**	**0.9863**	**0.9863**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	1.3648	1.1068	1.0346	1.0087	**0.9997**	**0.9986**	**0.9998**
$\hat{y}^{(2)}$	1.3585	1.1072	1.0230	1.0092	1.0008	1.0003	**0.9996**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0075	1.0063	1.0037	1.0072	1.0049	1.0025	1.0024
CF-RA2	1.0742	1.0240	1.0071	1.0045	1.0011	1.0006	1.0012
CF-RA3 ($\kappa = 0$)	1.0063	1.0041	1.0024	1.0061	1.0028	**0.9999**	1.0005
CF-RA3 ($\kappa = 1$)	1.0063	1.0041	1.0024	1.0060	1.0028	**0.9999**	1.0004
CF-RA3 ($\kappa = 28$)	1.0063	1.0037	1.0018	1.0051	1.0020	**0.9994**	**1.0000**
CF-RA3 ($\kappa = 55$)	1.0075	1.0037	1.0014	1.0043	1.0013	**0.9990**	**0.9996**
CF-RA3 ($\kappa = 82$)	1.0113	1.0044	1.0012	1.0037	1.0007	**0.9987**	**0.9992**
CF-RA3 ($\kappa = 109$)	1.0164	1.0059	1.0012	1.0031	1.0001	**0.9985**	**0.9990**
CF-RA3 ($\kappa = 136$)	1.0239	1.0078	1.0014	1.0027	**0.9997**	**0.9984**	**0.9989**
CF-RA3 ($\kappa = 163$)	1.0327	1.0103	1.0019	1.0024	**0.9993**	**0.9983**	**0.9988**
CF-RA3 ($\kappa = 190$)	1.0428	1.0133	1.0026	1.0022	**0.9990**	**0.9984**	**0.9989**
CF-RA3 ($\kappa = 217$)	1.0553	1.0166	1.0035	1.0022	**0.9987**	**0.9985**	**0.9990**
CF-RA3 ($\kappa = 244$)	1.0692	1.0211	1.0045	1.0023	**0.9986**	**0.9987**	**0.9992**
CF-Mean	1.0717	1.0218	1.0047	1.0023	**0.9985**	**0.9988**	**0.9993**

(continued)

given higher parameter estimation uncertainty in a small sample. Besides, the parameter estimation uncertainty makes the CF-RA2, which is argued to return asymptotically the optimal combination (Bates and Granger, 1969), performs undesirably. The best shrinkage value varies according

TABLE 2 Continued

Panel D. High negative correlations with the omitted variable and $\theta_1 = 3\theta_2$: $\Omega = \begin{pmatrix} 1 & 0.6 & -0.7 \\ 0.6 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{pmatrix}; \phi_i = 0; \theta = \begin{pmatrix} 0.6 \\ 0.2 \\ 0.6 \end{pmatrix}$

MSFE ratio

	$\sigma_\eta = 0.25$	$\sigma_\eta = 0.5$	$\sigma_\eta = 1$	$\sigma_\eta = 2$	$\sigma_\eta = 4$	$\sigma_\eta = 8$	$\sigma_\eta = 16$
			$R = 100,\ P = 100$				
$\hat{y}^{(1)}$	1.0014	**0.9960**	**0.9913**	**0.9906**	**0.9900**	**0.9900**	**0.9891**
$\hat{y}^{(1)}$	1.3459	1.1777	1.0529	1.0060	**0.9943**	**0.9900**	**0.9904**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0501	1.0515	1.0515	.0459	1.0353	1.0279	1.0290
CF-RA2	1.0210	1.0169	1.0140	1.0112	1.0090	1.0079	1.0070
CF-RA3 ($\kappa = 0$)	1.0243	1.0246	1.0232	1.0168	1.0103	1.0052	1.0043
CF-RA3 ($\kappa = 1$)	1.0210	1.0214	1.0200	1.0141	1.0078	1.0029	1.0021
CF-RA3 ($\kappa = 3$)	1.0167	1.0157	1.0142	1.0090	1.0032	**0.9988**	**0.9981**
CF-RA3 ($\kappa = 5$)	1.0143	1.0117	1.0093	1.0046	**0.9992**	**0.9952**	**0.9946**
CF-RA3 ($\kappa = 7$)	1.0143	1.0092	1.0053	1.0007	**0.9957**	**0.9921**	**0.9915**
CF-RA3 ($\kappa = 9$)	1.0162	1.0082	1.0021	**0.9975**	**0.9928**	**0.9896**	**0.9890**
CF-RA3 ($\kappa = 11$)	1.0205	1.0087	**0.9999**	**0.9948**	**0.9904**	**0.9875**	**0.9871**
CF-RA3 ($\kappa = 13$)	1.0272	1.0105	**0.9986**	**0.9928**	**0.9885**	**0.9860**	**0.9856**
CF-RA3 ($\kappa = 15$)	1.0363	1.0139	**0.9980**	**0.9914**	**0.9873**	**0.9850**	**0.9847**
CF-RA3 ($\kappa = 17$)	1.0472	1.0189	**0.9984**	**0.9905**	**0.9865**	**0.9845**	**0.9842**
CF-RA3 ($\kappa = 19$)	1.0606	1.0254	**0.9997**	**0.9903**	**0.9863**	**0.9845**	**0.9843**
CF-Mean	1.0983	1.0455	1.0058	**0.9919**	**0.9878**	**0.9865**	**0.9864**
			$R = 1000,\ P = 100$				
$\hat{y}^{(1)}$	1.0101	1.0060	**0.9992**	**0.9999**	**0.9988**	**0.9985**	**0.9995**
$\hat{y}^{(2)}$	1.3483	1.1861	1.0569	1.0214	1.0036	1.0010	**0.9996**
CI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CF-RA1	1.0034	1.0042	1.0034	1.0063	1.0047	1.0023	1.0027
CF-RA2	1.0097	1.0062	1.0026	1.0021	1.0007	1.0012	1.0010
CF-RA3 ($\kappa = 0$)	1.0019	1.0025	1.0024	1.0041	.0017	**0.9995**	1.0003
CF-RA3 ($\kappa = 1$)	1.0019	1.0025	1.0023	1.0040	1.0017	**0.9995**	1.0003
CF-RA3 ($\kappa = 28$)	1.0024	1.0027	1.0012	1.0031	1.0011	**0.9991**	**0.9998**
CF-RA3 ($\kappa = 55$)	1.0053	1.0045	1.0008	1.0025	1.0005	**0.9988**	**0.9994**
CF-RA3 ($\kappa = 82$)	1.0111	1.0075	1.0009	1.0022	1.0001	**0.9986**	**0.9991**
CF-RA3 ($\kappa = 109$)	1.0193	1.0122	1.0016	1.0021	**0.9998**	**0.9984**	**0.9988**
CF-RA3 ($\kappa = 136$)	1.0309	1.0182	1.0028	1.0024	**0.9996**	**0.9984**	**0.9987**
CF-RA3 ($\kappa = 163$)	1.0449	1.0254	1.0047	1.0029	**0.9996**	**0.9985**	**0.9987**
CF-RA3 ($\kappa = 190$)	1.0618	1.0344	1.0071	1.0037	**0.9996**	**0.9986**	**0.9988**
CF-RA3 ($\kappa = 217$)	1.0812	1.0449	1.0101	1.0048	**0.9997**	**0.9988**	**0.9989**
CF-RA3 ($\kappa = 244$)	1.1039	1.0566	1.0138	1.0061	**1.0000**	**0.9991**	**0.9992**
CF-Mean	1.1077	1.0586	1.0144	1.0064	1.0001	**0.9992**	**0.9992**

to different σ_η values, while generally a large amount of shrinkage (large κ) is found to be needed since the optimal combination strategy (except for Table 2 Panel D case) is about equal weighting. As mentioned in Section 3, shrinking too much to the equal weights is not necessarily good. The Monte Carlo evidence confirms this by noting that for a fixed

value of σ_η, CF-RA3(κ) with some values of κ is better than CF-Mean, and shrinking too much beyond that κ value sometimes make it deteriorate its performance.[13]

Second, we notice that results in Tables 1 and 2 differ in several ways. In Table 1 (when the CI model is correctly specified for the DGP), for smaller R and when the correlation between $x_{1,t}$ and $x_{2,t}$ is high, CF with shrinkage weights can beat CI even when disturbance in DGP (σ_η) is relatively small. When R gets larger, however, the advantage of CF vanishes. These Monte Carlo results are consistent with the analysis in Section 2.1, where we show CF may beat CI only in a finite sample. In contrast, by comparing the four panels in Table 2 (when the CI model is not correctly specified for the DGP), we find that when $x_{1,t}$ and $x_{2,t}$ are highly negatively correlated with the omitted variable $x_{3,t}$ and θ_3 is relatively large (Panel B), the advantage of CF (for even small values of σ_η) does not vanish as R gets larger. Moreover, we observe that even the individual forecasts can outperform CI in a large sample for large σ_η under this situation. The negative correlation of $x_{1,t}$ and $x_{2,t}$ with the omitted variable $x_{3,t}$, and the large value of θ_3 play an important role for CF to outperform CI in a large sample, which is conformable with the analysis in Section 2.2. In addition, Panel D of Table 2 shows that when x_1 contributes clearly more than x_2 in explaining the variable of interest y, the first individual forecast dominates the second one (making the optimal combination weight w^* close to 1 hence CF-Mean is clearly not working) when the noise in the DGP is not large. However, when the noise in the DGP is overwhelmingly large (signal to noise ratio is very low) such that the two individual forecasts are similarly bad, a close to equal weight is still desirable.

5. EMPIRICAL STUDY: EQUITY PREMIUM PREDICTION

In this section we study the relative performance of CI versus CF in predicting equity premium out-of-sample with many predictors including various financial ratios and interest rates. For a practical forecasting issue like this, we conjecture that CF scheme should be relatively more advantageous than CI scheme. Possible reasons are, first, it is very unlikely

[13]Our sample size used in the Monte Carlo experiments are $R = 100, 1000$, that are quite large for the small models with only 2 regressors (Table 1) and with 3 regressors (Table 2) in the CI model, making CI work quite comparably with CF. Even $R = 100$ may not be small enough to see the drastic difference (generally only about 2% improvement). The MSFE improvement by CF over CI would be more likely for a small sample size. Hence, we conducted the simulation with $R = 20$, for which CF models improve MSFE upon the performance of CI much more drastically by $8\% \sim 9\%$. For space, the results with $R = 20$ are not presented but included in the supplementary appendix that is made available on our webpage. In fact, the effect of the estimation sample size R can also be seen from the empirical results, e.g., comparing the three Panels in Table 3, with different size of R. When $R = 42$ is small (Panel C), CF models drastically improve MSFE over CI models.

that the CI model (no matter how many predictors are used) will coincide with the DGP for equity premium given the complicated nature of financial markets and therefore it is likely misspecified. Second, we deem that the conditions under which CF is better than CI as we illustrated in Sections 2.2 and 3.2 may easily be satisfied in this empirical application.

We obtained the monthly, quarterly and annual data over the period of 1927 to 2003 from the homepage of Amit Goyal (http://www.bus.emory.edu/AGoyal/). Our data construction replicates what Goyal and Welch (2008) did. The equity premium, y, is calculated by the S&P 500 market return (difference in the logarithms of index values in two consecutive periods) minus the risk free rate in that period. Our explanatory variable set, x, contains 12 individual variables: dividend price ratio, dividend yield, earnings price ratio, dividend payout ratio, book-to-market ratio, T-bill rate, long term yield, long term return, term spread, default yield spread, default return spread and lag of inflation, as used in Goyal and Welch (2008). Goyal and Welch (2008) explore the out-of-sample performance of these variables toward predicting the equity premium and find that not a single one would have helped a real-world investor outpredict the then-prevailing historical mean of the equity premium while pooling all by simple OLS regression performs even worse, and then conclude that "the equity premium has not been predictable." This supports our "weak predictors" argument discussed in Section 3.3 for explaining the success of CF-Mean.

Campbell and Thompson (2008) argue that once sensible restrictions are imposed on the signs of coefficients and return forecasts, forecasting variables with significant forecasting power in-sample generally have a better out-of-sample performance than a forecast based on the historical mean. Lewellen (2004) studies in particular the predictive power of financial ratios on forecasting aggregate stock returns through predictive regressions. He finds evidence of predictability by certain ratios over certain sample periods. In our empirical study, we bring the CF methodology into predicting equity premium and compare with CI since the analysis in Section 2 demonstrates that CF method indeed has its merits in out-of-sample forecasting practice. In addition, we investigate this issue of predictability by comparing various CF and CI schemes with the historical mean benchmark over different data frequencies, sample splits and forecast horizons.

5.1. CI Schemes

Two sets of CI schemes are considered. The first is the OLS using directly x_t (with dimension $N = 12$) as the regressor set while parameter estimate is obtained using strictly past data. The forecast is constructed as $\hat{y}_{T+h} = (1 \;\; x_T')\hat{\alpha}_T$. Let us call this forecasting scheme: CI-Unrestricted,

namely the kitchen-sink model. The second set of CI schemes aims at the problem associated with high dimension. It is quite possible to achieve a remarkable improvement on prediction by reducing dimensionality if one applies a factor model by extracting the Principal Components (PC) (Stock and Watson 2002a,b, 2004). The procedure is as follows:

$$x_t = \Lambda F_t + v_t, \tag{29}$$

$$y_{t+h} = (1 \; F_t')\gamma + u_{t+1}, \tag{30}$$

where Λ is $N \times r$ and F_t is $r \times 1$. In equation (29), by applying the classical principal component methodology, the latent common factors $F = (F_1 \; F_2 \; \cdots \; F_T)'$ is solved by:

$$\widehat{F} = X\widehat{\Lambda}/N \tag{31}$$

where N is the size of x_t, $X = (x_1 \; x_2 \; \cdots \; x_T)'$, and factor loading $\widehat{\Lambda}$ is set to $\sqrt{N}$ times the eigenvectors corresponding to the r largest eigenvalues of $X'X$ (see, for example, Bai and Ng, 2002). Once $\hat{\gamma}_T$ is obtained from (30) by regression of y_t on $(1 \; \widehat{F}_{t-1}')$ $(t = 1, 2, \ldots, T)$, the forecast is constructed as $\hat{y}_{T+h}^{\text{CI-PC}} = (1 \; \widehat{F}_T')\hat{\gamma}_T$ (let us denote this forecasting scheme as CI-PC).

If the true number of factors r is unknown, it can be estimated by minimizing some information criteria. Bai and Ng (2002) focus on estimation of the factor representation given by equation (29) and the asymptotic inference for r when N and T go to infinity. Equation (30), however, is more relevant for forecasting and thus it is our main interest. Moreover, we note that the N in our empirical study is only 12. We use AIC and BIC for which estimated number of factors k is selected by

$$\min_{1 \le k \le k_{\max}} IC_k = \ln(SSR(k)/T) + g(T)k, \tag{32}$$

where $k_{\max}$ is the hypothesized upper limit chosen by the user (we choose $k_{\max} = 12$), $SSR(k)$ is the sum of squared residuals from the forecasting model (30) using k estimated factors, and the penalty function $g(T) = 2/T$ for AIC and $g(T) = \ln T/T$ for BIC.[14] Additionally, we consider fixing k *a priori* at a small value like $1, 2, 3$.

[14]In model selection, it is well known that BIC is consistent in selecting the true model, and AIC is minimax-rate optimal for estimating the regression function. Yang (2005) shows that for any model selection criterion to be consistent, it must behave suboptimally for estimating the regression function in terms of minimax rate of convergence. Bayesian model averaging cannot be minimax-rate optimal for regression estimation. This explains that the model selected for in-sample fit and estimation would be different than the model selected for out-of-sample forecasting.

5.2. CF Schemes

We consider five sets of CF schemes where individual forecasts are generated by using each element x_{it} in x_t: $\hat{y}_{T+h}^{(i)} = (1 \quad x_{iT})\hat{\beta}_{i,T}$ $(i = 1, 2, \ldots, N)$. The first CF scheme, CF-Mean, is computed as $\hat{y}_{T+1}^{\text{CF-Mean}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{T+1}^{(i)}$. Second, CF-Median is to compute the median of the set of individual forecasts, which may be more robust in the presence of outlier forecasts. These two simple weighting CF schemes require no estimation in weight parameters.[15]

To explore more information in the data, thirdly, we estimate the combination weights w_i by regression approach (Granger and Ramanathan, 1984):

$$y_{t+h} = w_0 + \sum_{i=1}^{N} w_i \hat{y}_{t+h}^{(i)} + e_{t+1}, \tag{33}$$

and form predictor CF-RA, $\hat{y}_{T+h}^{\text{CF-RA}} = \hat{w}_0 + \sum_{i=1}^{N} \hat{w}_i \hat{y}_{T+h}^{(i)}$. Similarly as in Section 4 (Monte Carlo analysis), we experiment the three different versions of CF-RA. Fourth, we shrink CF-RA3 towards equally weighted CF by choosing increasing values of shrinkage parameter κ.

Finally, we extract the principal components from the set of individual forecasts and form predictor that may be called as CF-PC (combination of forecasts using the weighted principal components): see Chan et al. (1999).[16] Let $\hat{\mathbf{y}}_{t+h} := (\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \ldots, \hat{y}_{t+h}^{(N)})'$. Now, consider a factor model of $\hat{\mathbf{y}}_{t+h}$ (in the same way that a factor model of x_t in equation (29) for CI-PC

[15]Starting from Granger and Ramanathan (1984), based on earlier works such as Bates and Granger (1969) and Newbold and Granger (1974), various feasible optimal combination weights have been suggested, which are static, dynamic, time-varying, or Bayesian: see Diebold and Lopez (1996). Chan et al. (1999) and Stock and Watson (2004) utilize the principal component approach to exploit the factor structure of a panel of forecasts to improve upon Granger and Ramanathan (1984) combination regressions. They show this principal component forecast combination is more successful when there are large number of individual forecasts to be combined. The procedure is to first extract a small set of principal components from a (large) set of forecasts and then estimate the (static) combination weights for the principal components. Deutsch et al. (1994) extend Granger and Ramanathan (1984) by allowing dynamics in the weights which are derived from switching regression models or from smooth transition regression models. Li and Tkacz (2004) introduce a flexible non-parametric technique for selecting weights in a forecast combination regression. Empirically, Stock and Watson (2004) consider various CF weighting schemes and find the superiority of simple weighting schemes over sophisticated ones (such as time-varying parameter combining regressions) for output growth prediction in a seven-country economic data set.

[16]Also see Stock and Watson (2004), where it is called Principal Component Forecast Combination. In Aguiar-Conraria (2003), a similar method is proposed: Principal Components Combination (PCC), where the Principal Components Regression (PCR) is combined with the Forecast Combination approach by using each explanatory variable to obtain a forecast for the dependent variable, and then combining the several forecasts using the PCR method. This idea, as noted in the article, follows the spirit of Partial Least Squares in the Chemometrics literature thus is distinguished from what proposed in Chan et al. (1999).

is considered):[17]

$$\hat{\mathbf{y}}_{t+h} = \Lambda F_{t+h} + v_{t+h}. \tag{34}$$

In equation (34), by applying the classical principal component methodology, the latent common factors $F = (F_{1+h} \; F_{2+h} \; \cdots \; F_{T+h})'$ is solved by:

$$\widehat{F} = \widehat{Y}\widehat{\Lambda}/N \tag{35}$$

where $\widehat{Y} = (\hat{\mathbf{y}}_{1+h} \; \hat{\mathbf{y}}_{2+h} \; \cdots \; \hat{\mathbf{y}}_{T+h})'$, and factor loading $\widehat{\Lambda}$ is set to $\sqrt{N}$ times the eigenvectors corresponding to the r largest eigenvalues of $\widehat{Y}'\widehat{Y}$. If the true number of factors r is unknown, it can be estimated by minimizing some information criteria such as those of Bai and Ng (2002), AIC, or BIC, to get the estimated number of factors k. Let $\widehat{F}_{t+h} := (\widehat{F}^{(1)}_{t+h}, \ldots, \widehat{F}^{(k)}_{t+h})'$ denote the first k principal components of $\hat{\mathbf{y}}_{t+h} = (\hat{y}^{(1)}_{t+h}, \ldots, \hat{y}^{(N)}_{t+h})'$ for $t = T_0, \ldots, T$.[18] Then the forecasting equation is

$$\begin{aligned} y_{t+h} &= (1 \; \widehat{F}'_{t+h})\gamma + u_{t+h} \\ &= \gamma_0 + \sum_{i=1}^{k} \gamma_i \widehat{F}^{(i)}_{t+h} + u_{t+h}. \end{aligned}$$

Once $\hat{\gamma}_T$ is obtained by regression of y_{t+h} on $(1 \; \widehat{F}'_{t+h})$ $(t = T_0, \ldots, T)$, the CF-PC forecast is then constructed as $\hat{y}^{\text{CF-PC}}_{T+h} = (1 \; \widehat{F}'_T)\hat{\gamma}_T = \hat{\gamma}_{0T} + \sum_{i=1}^{k} \hat{\gamma}_{iT}\widehat{F}^{(i)}_{t+h}$ (let us denote this forecasting scheme as CF-PC).

Remark 1. Chan et al. (1999) choose $k = 1$ since the factor analytic structure for the set of individual forecasts they adopt permits one single factor—the conditional mean of the variable to be forecast. Our specifications for individual forecasts in CF, however, differ from those in Chan et al. (1999) in that individual forecasting models considered here use different and non-overlapping information sets, not a common total information set (which makes individual forecasts differ solely from specification error and estimation error) as assumed in Chan et al. (1999). Therefore, we consider $k = 1, 2, 3$. In addition to that, k is also chosen by the information criteria AIC or BIC, as discussed in Section 5.1.

[17]We use the same notation F, v, u, γ, Λ in this sub-section for CF-PC as in the previous subsection on CI-PC, only to make it easy to read. These are different in the two models and should be understood in the context of each model.

[18]In computing the out-of-sample equity premium forecasts by rolling window scheme with window size R, we set $T = R$ and choose T_0, the time when the first pseudo out-of-sample forecast is generated, at the middle point of the rolling window.

Remark 2. The biggest difference between CF-PC and CI-PC lies in the set of variables we use to extract the principal components (PC). In CI-PC, PC's are computed from x's directly, without accounting for their relationship with the forecast target variable y. This problem with CI-PC leads Bai and Ng (2008) to consider first selecting a subset of predictors ("targeted predictors") of x's that are informative in forecasting y, then using the subset to extract factors. In contrast, since CF-PC is one type of CF where we combine forecasts not the information sets directly, PC's in CF-PC are computed from the set of individual forecasts $\hat{y}$'s that contain both information on x's and on all past values of y. This actually provides us further intuitions on why CF may be more successful than CI, along the line of "supervised learning."

If $k = N$, there is no difference, i.e., CI-PC and CF-PC are the same. When $k < N$, the principal components of the forecasts from CF and the principal components of predictors in CI will differ from each other, because the linear combinations maximizing covariances of forecasts (for which the supervision operates for the relationship between y and x) and the linear combinations maximizing the covariances of predictors (for which there is no supervision) will be different.

Remark 3. CF-PC is the weighted combined forecasts. To see this, write the $N \times k$ matrix of estimated loadings of the k factors as

$$\widehat{\Lambda} = \begin{pmatrix} \hat{\lambda}_{11} & \cdots & \hat{\lambda}_{1k} \\ \vdots & & \vdots \\ \hat{\lambda}_{N1} & \cdots & \hat{\lambda}_{Nk} \end{pmatrix}. \tag{36}$$

Then the first k estimated CF-PC factors are

$$\widehat{F} = (\widehat{F}_{1+h}\ \widehat{F}_{2+h}\ \cdots\ \widehat{F}_{T+h})' = \widehat{Y}\widehat{\Lambda}/N,$$

or its tth column is

$$\begin{aligned} \widehat{F}_{t+h} &= \widehat{\Lambda}'\hat{\mathbf{y}}_{t+h}/N \\ &= \left(\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \hat{\lambda}_{i1}\hat{y}_{t+h}^{(i)} \quad \cdots \quad \tfrac{1}{N}\sum_{i=1}^{N} \hat{\lambda}_{ik}\hat{y}_{t+h}^{(i)}\right)' \\ &= (\widehat{F}_{t+h}^{(1)}, \ldots, \widehat{F}_{t+h}^{(k)})'. \end{aligned}$$

Note that by construction, CF-PC factors $(\widehat{F}_{t+h}^{(1)}, \ldots, \widehat{F}_{t+h}^{(k)})$ are the weighted combined forecasts, with the weights given by columns of $\widehat{\Lambda}$.

In particular, with $\hat{\lambda}_{i1} = 1$ (for all $i = 1, \ldots, N$), we note that $\widehat{F}_{t+h}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\lambda}_{i1}\hat{y}_{t+h}^{(i)} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{t+h}^{(i)}$ is the CF-Mean. CF-Mean is a particular form

of CF-PC with $k = 1$, obtained by shrinking $\hat{\lambda}_{i1}$ to a fixed constant. Therefore, CF-Mean is a single factor CF-PC model with the factor loading shrunken to a constant. Noting that CF-Mean is a shrinkage version of the CF-PC, we can also view the regression based CF (CF-RA) and its shrinkage version CF-RA(κ) as a general shrinkage version of the CF-PC.

It is important to recall that the consistent estimation of $\widehat{\Lambda}$ requires a large N (Bai, 2003; Stock and Watson, 2002b). When N is not large enough, CF-Mean (without estimating $\hat{\lambda}_{i1}$) can dominate CF-PC models (with estimating $\hat{\lambda}_{ij}$, $j = 1, \ldots, k$), as we will see the empirical results below with $N = 12$.[19]

5.3. Empirical Results

Table 3 presents the out-of-sample performance of each forecasting scheme for equity premium prediction across different forecast horizons h, different frequencies (monthly, quarterly, and annual in Panels A, B, and C) and different in-sample/out-of-sample splits R and P. Data range from 1927 to 2003 in monthly, quarterly and annual frequencies. All models are estimated using OLS over rolling windows of size R. To compare each model with the benchmark Historical Mean (HM) we report its MSFE ratio with respect to HM.[20]

First, similarities are found among Panels A, B, and C. While not reported for space, although there are a few cases some individual forecasts return relatively small MSFE ratio, the performance of individual forecasts is fairly unstable and each similarly bad. In contrast, we clearly observe the genuinely stable and superior performance of CF-Mean and CF with shrinkage weights (particularly with a large amount of shrinkage imposed so that the weights are close to equal weights), compared to almost all CI schemes across different frequencies, especially for shorter forecast horizons and for the forecast periods with earlier starting date. CF-Median also appears to perform quite well. This confirms the discussion in Section 3 (particularly Subsection 3.3) where we shed light on the reasons for the success of simple average combination of forecasts, and is fairly consistent with this understanding of the puzzle in the presence of "weak predictors" for the equity premium prediction.

[19]If N is very large and $\widehat{\Lambda} \xrightarrow{p} \Lambda$, CF-PC may work better than CF-Mean. Otherwise, the parameter estimation of the factor loading can contaminate CF-PC and make it worse than CF-Mean, which is in line with our understanding of the forecast combination puzzle discussed in Section 3.1. The consistent estimation of $\widehat{\Lambda} \xrightarrow{p} \Lambda$ amounts to the consistent estimation of the forecast combination weight w in Section 3.1 and Smith and Wallis (2009).

[20]The MSFE ratios are computed with respect to the CI benchmark for Monte Carlo analysis (Section 4, Tables 1, 2), while they are computed with respect to the historical-mean benchmark for empirical analysis as in Campbell and Thompson (2008). We present the tables only with the MSFE ratios. However, we make the MSFE values available from our webpage.

Second, MSFE ratios of the good models that outperform HM are smaller in Panel B (quarterly prediction) and Panel C (annual prediction) than in Panels A1 and A2 (monthly predictions). This indicates that with these good models we can beat HM more easily for quarterly and annual series than for monthly series.

Third, CF-PC with a fixed number of factors (1 or 2) frequently outperforms HM as well, and by contrast, the CI schemes rarely beat HM by a considerable margin. Generally BIC performs better than AIC by selecting a smaller k (the estimated number of factors) but worse than using a small fixed k (=1, 2, 3).

Fourth, within each panel, we find that generally it is hard to improve upon HM for more recent out-of-sample periods (forecasts beginning in 1980) and for longer forecast horizons, since the MSFE ratios tend to be larger under these situations. It seems that the equity premium becomes less predictable in recent years than older years.

Fifth, we note that the in-sample size R is smaller for the forecast period starting from the earlier year. In accordance with the conditions

TABLE 3 Equity premium prediction

Panel A1. Monthly prediction, forecasts begin 1969m1 ($R = 504$ and $p = 420$)

	MSFE ratio							
	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Historical mean	1.0000		1.0000		1.0000		1.0000	
CF-Mean	**0.9820**		**0.9860**		**0.9890**		**0.9891**	
CF-Median	**0.9887**		**0.9915**		**0.9913**		**0.9904**	
CF-RA1	1.0585		1.0660		1.0325		1.1548	
CF-RA2	1.0975		1.0847		1.0538		1.1225	
CF-RA3 ($\kappa = 0$)	1.0795		1.0581		1.0310		1.1240	
CF-RA3 ($\kappa = 1$)	1.0670		1.0487		1.0250		1.1116	
CF-RA3 ($\kappa = 3$)	1.0443		1.0317		1.0141		1.0889	
CF-RA3 ($\kappa = 5$)	1.0248		1.0172		1.0049		1.0684	
CF-RA3 ($\kappa = 7$)	1.0086		1.0052		**0.9974**		1.0503	
CF-RA3 ($\kappa = 9$)	**0.9956**		**0.9956**		**0.9916**		1.0346	
CF-RA3 ($\kappa = 11$)	**0.9859**		**0.9884**		**0.9875**		1.0213	
CF-RA3 ($\kappa = 13$)	**0.9794**		**0.9837**		**0.9851**		1.0103	
CF-RA3 ($\kappa = 15$)	**0.9762**	Mean/SD	**0.9815**	Mean/SD	**0.9844**	Mean/SD	1.0017	Mean/SD
CF-PC (AIC)	1.0429	9.13/3.26	1.0697	8.62/3.45	1.0363	4.74/4.23	1.0158	1.90/2.45
CF-PC (BIC)	**0.9828**	1.30/1.06	**0.9962**	1.14/0.49	1.0029	1.18/0.42	**0.9993**	1.06/0.24
CF-PC ($k = 1$)	**0.9858**		**0.9903**		**0.9989**		1.0049	
CF-PC ($k = 2$)	**0.9801**		**0.9953**		1.0000		**0.9995**	
CF-PC ($k = 3$)	**0.9912**		1.0076		1.0090		1.0065	
CI-Unrestricted	1.0103		1.0661		1.0400		1.0712	
CI-PC (AIC)	1.0142	8.70/2.18	1.0537	7.47/2.49	1.0655	6.22/2.82	1.0147	2.35/0.84
CI-PC (BIC)	1.0523	3.29/1.85	1.0655	2.48/1.39	1.0478	1.92/0.99	1.0071	1.38/0.63
CI-PC ($k = 1$)	**0.9998**		1.0009		**0.9996**		**0.9934**	
CI-PC ($k = 2$)	1.0060		1.0151		1.0134		**0.9944**	
CI-PC ($k = 3$)	1.0673		1.0805		1.0612		1.0115	

(continued)

TABLE 3 Continued

Panel A2. Monthly prediction, forecasts begin 1980m1 ($R = 636$ and $P = 288$)

	MSFE ratio							
	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Historical mean	1.0000		1.0000		1.0000		1.0000	
CF-Mean	**0.9938**		**0.9980**		**0.9981**		**0.9995**	
CF-Median	**0.9993**		1.0023		**0.9986**		1.0026	
CF-RA1	1.0606		1.0361		1.0873		1.0649	
CF-RA2	1.0590		1.0637		1.0811		1.0946	
CF-RA3 ($\kappa = 0$)	1.0821		1.0605		1.1108		1.0690	
CF-RA3 ($\kappa = 1$)	1.0741		1.0547		1.1008		1.0642	
CF-RA3 ($\kappa = 4$)	1.0523		1.0389		1.0734		1.0509	
CF-RA3 ($\kappa = 7$)	1.0338		1.0256		1.0501		1.0391	
CF-RA3 ($\kappa = 10$)	1.0187		1.0147		1.0310		1.0288	
CF-RA3 ($\kappa = 13$)	1.0069		1.0063		1.0161		1.0200	
CF-RA3 ($\kappa = 16$)	**0.9985**		1.0005		1.0053		1.0128	
CF-RA3 ($\kappa = 19$)	**0.9935**		**0.9970**		**0.9986**		1.0071	
CF-RA3 ($\kappa = 22$)	**0.9917**	Mean/SD	**0.9961**	Mean/SD	**0.9961**	Mean/SD	1.0029	Mean/SD
CF-PC (AIC)	1.0741	10.33/3.27	1.0251	8.74/3.98	1.0815	9.33/3.95	1.0198	4.26/4.55
CF-PC (BIC)	**0.9937**	1.30/0.77	1.0063	1.02/0.14	1.0104	1.02/0.13	1.0161	1/0
CF-PC ($k = 1$)	**0.9896**		1.0038		1.0089		1.0161	
CF-PC ($k = 2$)	**0.9918**		1.0091		1.0154		1.0148	
CF-PC ($k = 3$)	**0.9960**		1.0086		1.0150		1.0200	
CI-Unrestricted	1.0592		1.1344		1.0525		1.0495	
CI-PC (AIC)	1.0522	8.63/1.87	1.1274	7.68/2.12	1.0607	6.95/2.53	1.0197	2.68/1.14
CI-PC (BIC)	1.0639	3.02/1.72	1.0578	2.35/1.31	1.0199	1.64/1.08	1.0376	1.56/0.72
CI-PC ($k = 1$)	1.0131		1.0150		1.0200		1.0194	
CI-PC ($k = 2$)	1.0175		1.0251		1.0274		1.0315	
CI-PC ($k = 3$)	1.0617		1.0623		1.0575		1.0376	

(continued)

under which CF can be superior to CI as discussed in Section 2, the smaller in-sample size may *partly* account for the success of CF-Mean over the forecast period starting from the earlier year in line of the argument about parameter estimation uncertainty.

In summary, Table 3 shows that CF-Mean, or CF-RA3 using estimated weights shrunken towards equal weights, are simple but powerful methods to predict the equity premium out-of-sample in comparison with the CI schemes, and to beat the HM benchmark. This may be due to the estimation uncertainty of the factor loadings as discussed in Remark 3 of Section 5.2. When N is not large enough, CF-Mean (without estimating $\hat{\lambda}_{i1}$) can dominate CF-PC models (with estimating $\hat{\lambda}_{ij}$, $j = 1, \ldots, k$) due to the similar reason discussed in Section 3.1.

6. CONCLUSIONS

In this article, we show the relative merits of combination of forecasts (CF) compared to combination of information (CI). In the literature, it is

TABLE 3 Continued

Panel B. Quarterly prediction

	MSFE ratio							
	Forecasts begin 1969q1 ($R = 168$ and $P = 140$)				Forecasts begin 1980q1 ($R = 212$ and $P = 96$)			
	$h = 1$		$h = 4$		$h = 1$		$h = 4$	
Historical mean	1.0000		1.0000		1.0000		1.0000	
CF-Mean	**0.9589**		**0.9768**		**0.9899**		1.0071	
CF-Median	**0.9689**		**0.9831**		**0.9992**		1.0172	
CF-RA1	1.2436		1.7457		1.3127		1.2568	
CF-RA2	1.3942		1.6537		1.3120		1.3482	
CF-RA3 ($\kappa = 0$)	1.2981		1.6728		1.4901		1.2819	
CF-RA3 ($\kappa = 0.25$)	1.2660		1.6185		1.4554		1.2656	
CF-RA3 ($\kappa = 0.5$)	1.2354		1.5665		1.4219		1.2499	
CF-RA3 ($\kappa = 1$)	1.1791	Mean/SD	1.4690	Mean/SD	1.3586	Mean/SD	1.2198	Mean/SD
CF-PC (AIC)	1.3136	7.08/4.40	1.3484	3.31/3.98	1.2224	8.69/4.05	1.0959	4.17/4.87
CF-PC (BIC)	1.0512	1.27/0.66	1.0451	1.06/0.23	1.0136	1.25/0.78	1.0499	1.01/0.10
CF-PC ($k = 1$)	1.0036		1.0286		**0.9987**		1.0501	
CF-PC ($k = 2$)	**0.9993**		1.0287		1.0176		1.0306	
CF-PC ($k = 3$)	1.0214		1.0464		1.0216		1.0467	
CI-Unrestricted	1.0835		1.2182		1.3046		1.2026	
CI-PC (AIC)	1.1488	7.66/2.21	1.1104	2.56/1.35	1.2942	8.73/2.10	1.0708	2.97/1.84
CI-PC (BIC)	1.2094	2.36/0.95	1.0409	1.35/0.78	1.1799	2.67/1.60	1.2350	2.01/1.49
CI-PC ($k = 1$)	**0.9991**		**0.9932**		1.0414		1.0543	
CI-PC ($k = 2$)	1.0207		1.0091		1.0846		1.1257	
CI-PC ($k = 3$)	1.2214		1.0875		1.2112		1.1467	

(continued)

commonly believed that CI is optimal. This belief is valid for in-sample fit but when it comes to out-of-sample forecasting, CI is no longer undefeated. In Section 2, through stylized linear forecasting regressions we illustrate analytically the circumstances when the forecast by CF can be superior to the forecast by CI, when CI model is correctly specified and when it is misspecified. We also shed some light on how CF with (close to) equal weights may work by noting that, apart from the parameter estimation uncertainty argument (Smith and Wallis, 2009), in practical situations the information sets we selected that are used to predict the variable of interest are often with about equally low predictive content therefore a simple average combination is often close to optimal (discussed in Section 3). Our Monte Carlo analysis in Section 4 provides some insights on the possibility that CF with shrinkage or CF with equal weights can dominate CI even in a large sample.

In accordance with the analytical findings, our empirical application on the equity premium prediction confirms the advantage of CF in real time forecasting. We compare CF with various weighting methods,

TABLE 3 Continued

Panel C. Annual prediction				
MSFE ratio				
	Forecasts begin 1969 ($R = 42$ and $p = 35$)		Forecasts begin 1980 ($R = 53$ and $p = 24$)	
	$h = 1$		$h = 1$	
Historical mean	1.0000		1.0000	
CF-Mean	**0.9096**		**0.9828**	
CF-Median	**0.9390**		1.0188	
CF-RA1	5.1820		6.4651	
CF-RA2	4.0819		3.2646	
CF-RA3 ($\kappa = 0$)	4.3141		5.0635	
CF-RA3 ($\kappa = 0.25$)	2.2625		3.3712	
CF-RA3 ($\kappa = 0.5$)	1.1408		2.1293	
CF-RA3 ($\kappa = 1$)	**0.9096**	Mean/SD	**0.9965**	Mean/SD
CF-PC (AIC)	4.6260	10.14/2.59	5.8805	10.08/3.39
CF-PC (BIC)	3.6133	5.29/4.62	2.2426	4.46/4.70
CF-PC ($k = 1$)	1.0034		1.1012	
CF-PC ($k = 2$)	**0.9376**		1.1211	
CF-PC ($k = 3$)	1.0507		1.3079	
CI-Unrestricted	1.9013		1.9979	
CI-PC (AIC)	1.9067	5.34/3.33	1.9196	6.33/3.16
CI-PC (BIC)	1.5243	3.03/1.87	1.5385	1.88/1.33
CI-PC ($k = 1$)	1.0340		1.2502	
CI-PC ($k = 2$)	1.0596		1.3183	
CI-PC ($k = 3$)	1.3754		1.3814	

Note: Data range 1927 m1 to 2003 m12; "$k_{\max}$," the maximum hypothesized number of factors, is set at 12; "h" is the forecast horizon; We report MSFE ratio which is the MSFE of each method over that of the Historical Mean model; "k" is the number of factors included in the principal component approaches; "Mean/SD" is the mean and standard deviation of the estimated number of factors over the out-of-sample. The case when Historical Mean benchmark is outperformed is indicated **bold**, and the smallest number among them is highlighted.

including simple average, regression based approach with principal component method (CF-PC), to CI models with principal component approach (CI-PC). We find that CF with (close to) equal weights dominates about all CI schemes, and also performs substantially better than the historical mean benchmark model. These empirical results highlight the merits of CF that we analyzed in Sections 2 and 3, and they are also consistent with much of literature about CF, for instance, the empirical findings by Stock and Watson (2004) where CF with various weighting schemes (including CF-PC) is found favorable when compared to CI-PC.

APPENDIX: DERIVATION OF MSFEs FOR SECTION 2.2

Define $\theta_{12} \equiv (\theta_1' \ \theta_2')'$ and $\delta_{\hat{\alpha}} \equiv \hat{\alpha}_T - E(\hat{\alpha}_T)$. Note that

$$\begin{aligned} E(\hat{\alpha}_T) &= E\left[\left(\sum z_t' z_t\right)^{-1} \sum z_t' y_{t+1}\right] \\ &= \theta_{12} + E\left[\left(\sum z_t' z_t\right)^{-1} \sum z_t' x_{3t}\right]\theta_3 \\ &= \theta_{12} + \Omega_{zz}^{-1}\Omega_{z3}\theta_3, \end{aligned} \tag{37}$$

and $Var(\hat{\alpha}_T) = T^{-1}\sigma_\eta^2\Omega_{zz}^{-1}$, so $\delta_{\hat{\alpha}} = \hat{\alpha}_T - \theta_{12} - \Omega_{zz}^{-1}\Omega_{z3}\theta_3$. Thus, the conditional bias by the CI forecast is

$$\begin{aligned} E(\hat{e}_{T+1} \mid \mathcal{I}_T) &= x_{1T}(\theta_1 - \hat{\alpha}_{1,T}) + x_{2,T}(\theta_2 - \hat{\alpha}_{2,T}) + x_{3,T}\theta_3 \\ &= z_T(\theta_{12} - \hat{\alpha}_T) + x_{3,T}\theta_3 = z_T(-\Omega_{zz}^{-1}\Omega_{z3}\theta_3 - \delta_{\hat{\alpha}}) + x_{3,T}\theta_3 \\ &= -z_T\delta_{\hat{\alpha}} + \xi_{3z,T}\theta_3, \end{aligned} \tag{38}$$

where $\mathcal{I}_T$ denotes the total information up to time T. It follows that

$$\begin{aligned} MSFE^{\mathrm{CI}} &= E[Var_T(y_{T+1})] + E[(E(\hat{e}_{T+1} \mid \mathcal{I}_T))^2] \\ &= \sigma_\eta^2 + E[(-z_T\delta_{\hat{\alpha}} + \xi_{3z,T}\theta_3)(-z_T\delta_{\hat{\alpha}} + \xi_{3z,T}\theta_3)'] \\ &= \sigma_\eta^2 + E[z_T Var(\hat{\alpha}_T) z_T'] + \theta_3' E[\xi_{3z,T}'\xi_{3z,T}]\theta_3 \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2 E[z_T\Omega_{zz}^{-1}z_T'] + \theta_3'\Omega_{\xi_{3z}}\theta_3 \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2 tr\{\Omega_{zz}^{-1}E[z_T'z_T]\} + \theta_3'\Omega_{\xi_{3z}}\theta_3 \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2(k_1 + k_2) + \theta_3'\Omega_{\xi_{3z}}\theta_3. \end{aligned} \tag{39}$$

Similarly, for the two individual forecasts, define $\delta_{\hat{\beta}_i} \equiv \hat{\beta}_{i,T} - E(\hat{\beta}_{i,T})$ $(i = 1, 2)$. Given that

$$\begin{aligned} E(\hat{\beta}_{1,T}) &= E\left[\left(\sum x_{1,t}' x_{1,t}\right)^{-1} \sum x_{1,t}' y_{t+1}\right] \\ &= \theta_1 + E\left[\left(\sum x_{1,t}' x_{1,t}\right)^{-1} \sum x_{1,t}'(x_{2,t}\theta_2 + x_{3,t}\theta_3)\right] \\ &= \theta_1 + \Omega_{11}^{-1}(\Omega_{12}\theta_2 + \Omega_{13}\theta_3), \end{aligned} \tag{40}$$

and

$$E(\hat{\beta}_{2,T}) = \theta_2 + \Omega_{22}^{-1}(\Omega_{21}\theta_1 + \Omega_{23}\theta_3), \tag{41}$$

the conditional biases by individual forecasts are:

$$\begin{aligned} E(\hat{\epsilon}_{1,T+1} \mid \mathcal{J}_T) &= x_{1,T}(\theta_1 - \hat{\beta}_{1,T}) + x_{2,T}\theta_2 + x_{3,T}\theta_3 = -x_{1,T}\delta_{\hat{\beta}_1} + \xi_{23.1,T}\theta_{23}, \\ E(\hat{\epsilon}_{2,T+1} \mid \mathcal{J}_T) &= x_{1,T}\theta_1 + x_{2,T}(\theta_2 - \hat{\beta}_{2,T}) + x_{3,T}\theta_3 = -x_{2,T}\delta_{\hat{\beta}_2} + \xi_{13.2,T}\theta_{13}. \end{aligned} \tag{42}$$

Hence, similar to the derivation for $MSFE^{\mathrm{CI}}$, it is easy to show that

$$\begin{aligned} MSFE^{(1)} &= \sigma_\eta^2 + E[(-x_{1,T}\delta_{\hat{\beta}_1} + \xi_{23.1,T}\theta_{23})(-x_{1,T}\delta_{\hat{\beta}_1} + \xi_{23.1,T}\theta_{23})'] \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2 E[x_{1,T}\Omega_{11}^{-1}x_{1,T}'] + \theta_{23}'\Omega_{\xi_{23.1}}\theta_{23} \\ &= \sigma_\eta^2 + T^{-1}\sigma_\eta^2 k_1 + \theta_{23}'\Omega_{\xi_{23.1}}\theta_{23}, \end{aligned} \tag{43}$$

and

$$MSFE^{(2)} = \sigma_\eta^2 + T^{-1}\sigma_\eta^2 k_2 + \theta_{13}'\Omega_{\xi_{13.2}}\theta_{13}, \tag{44}$$

by noting that $Var(\hat{\beta}_{i,T}) = T^{-1}\sigma_\eta^2\Omega_{ii}^{-1}$ $(i = 1, 2)$.

Using equation (15), the conditional bias by the CF forecast is

$$E(\hat{e}_{T+1}^{\mathrm{CF}} \mid \mathcal{J}_T) = wE(\hat{\epsilon}_{1,T+1} \mid \mathcal{J}_T) + (1-w)E(\hat{\epsilon}_{2,T+1} \mid \mathcal{J}_T). \tag{45}$$

It follows that

$$\begin{aligned} MSFE^{\mathrm{CF}} &= \sigma_\eta^2 + E[(E(\hat{e}_{T+1}^{\mathrm{CF}} \mid \mathcal{J}_T))^2] \\ &= \sigma_\eta^2 + E[w^2(E(\hat{\epsilon}_{1,T+1} \mid \mathcal{J}_T))^2 + (1-w)^2(E(\hat{\epsilon}_{2,T+1} \mid \mathcal{J}_T))^2 \\ &\quad + 2w(1-w)E(\hat{\epsilon}_{1,T+1} \mid \mathcal{J}_T)E(\hat{\epsilon}_{2,T+1} \mid \mathcal{J}_T)] \\ &= \sigma_\eta^2 + w^2[T^{-1}\sigma_\eta^2 k_1 + \theta_{23}'\Omega_{\xi_{23.1}}\theta_{23}] + (1-w)^2[T^{-1}\sigma_\eta^2 k_2 + \theta_{13}'\Omega_{\xi_{13.2}}\theta_{13}] \\ &\quad + 2w(1-w)E[x_{1T}\delta_{\hat{\beta}_1}\delta_{\hat{\beta}_2}'x_{2T}' + \theta_{23}'\xi_{23.1,T}'\xi_{13.2,T}\theta_{13}] \\ &= \sigma_\eta^2 + g_T^{\mathrm{CF}} + w^2\theta_{23}'\Omega_{\xi_{23.1}}\theta_{23} + (1-w)^2\theta_{13}'\Omega_{\xi_{13.2}}\theta_{13} \\ &\quad + 2w(1-w)\theta_{23}'E[\xi_{23.1,T}'\xi_{13.2,T}]\theta_{13}, \end{aligned} \tag{46}$$

where $g_T^{\mathrm{CF}} = T^{-1}(w^2k_1 + (1-w)^2k_2)\sigma_\eta^2 + 2w(1-w)E[x_{1T}\delta_{\hat{\beta}_1}\delta_{\hat{\beta}_2}'x_{2T}']$.

ACKNOWLEDGMENTS

We are grateful to two anonymous referees for their constructive suggestions and extensive comments. We would also like to thank Tim Bollerslev, Gloria González-Rivera, Bruce Hansen, Lutz Kilian, Michael McCracken, Barbara Rossi, George Tauchen, Aman Ullah, as well as the participants of the Applied Macro Workshop at Duke University,

Forecasting Session at North American ES Summer 2006 Meetings, Seminar at UC Riverside, and MEG 2006 Meeting, for helpful discussions and comments. All errors are our own.

REFERENCES

Aguiar-Conraria, L. (2003). *Forecasting in Data-Rich Environments*. Portugal: Cornell University and Minho University.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71(1):135–171.

Bai, J., Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* 70:191–221.

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* 24:607–629.

Bates, J. M., Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly* 20:451–468.

Campbell, J. Y., Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–1531.

Chan, Y. L., Stock, J. H., Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review* 1:91–121.

Chong, Y. Y., Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economics Studies* LIII:671–690.

Clark, T. E., McCracken, M. W. (2009). Combining forecasts from nested models. *Oxford Bulletin of Economics and Statistics* 71:303–329.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* 5:559–583.

Clements, M. P., Galvao, A. B. (2006). Combining predictors and combining information in modelling: forecasting US recession probabilities and output growth. In: Costas Milas, Philip Rothman, Dick van Dijk, eds. *Nonlinear Times Series Analysis of Business Cycles*, Chapter 2. Emeral Group Publishing, pp. 55–73.

Coulson, N. E., Robins, R. P. (1993). Forecast combination in a dynamic setting. *Journal of Forecasting* 12:63–67.

Deutsch, M., Granger, C. W. J., Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting* 10:47–57.

Diebold, F. X. (1989). Forecast combination and encompassing: reconciling two divergent literatures. *International Journal of Forecasting* 5:589–592.

Diebold, F. X., Lopez, J. A. (1996). Forecast evaluation and combination. NBER Working Paper, No. 192.

Diebold, F. X., Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting* 6:503–508.

Engle, R. F., Granger, C. W. J., Kraft, D. F. (1984). Combining competing forecasts of inflation using a bivariate ARCH model. *Journal of Economic Dynamics and Control* 8:151–165.

Giacomini, R., Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics* 23:416–431.

Goyal, A., Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21(4):1455–1508.

Granger, C. W. J. (1989). Invited review: combining forecasts – twenty years later. *Journal of Forecasting* 8:167–173.

Granger, C. W. J., Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting* 3:197–204.

Hansen, B. E. (2008). Least squares forecast averaging. *Journal of Econometrics* 146:342–350.

Harvey, D. I., Newbold, P. (2005). Forecast encompassing and parameter estimation. *Oxford Bulletin of Economics and Statistics* 67, Supplement 0305–9049.

Hendry, D. F., Clements, M. P. (2004). Pooling of forecasts. *Econometrics Journal* 7:1–31.

Hibon, M., Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting* 21:15–24.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics* 74:209–235.

Li, F., Tkacz, G. (2004). Combining forecasts with nonparametric kernel regressions. *Studies in Nonlinear Dynamics and Econometrics* 8(4), Article 2.

Newbold, P., Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society* 137:131–165.

Newbold, P., Harvey, D. I. (2001). Forecast combination and encompassing. In: Clements, M. P., Hendry, D. F., eds. *A Companion to Economic Forecasting*. Oxford, UK: Blackwell Publishers.

Palm, F. C., Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting* 11:687–701.

Shen, X., Huang, H.-C. (2006). Optimal model assessment, selection, and combination. *Journal of the American Statistical Association* 101:554–568.

Smith, J., Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics* 71:331–355.

Stock, J. H., Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics* 44:293–335.

Stock, J. H., Watson, M. W. (2002a). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20:147–162.

Stock, J. H., Watson, M. W. (2002b). Forecasting using principal components from a large number of predictors. *Journal of the 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*, Chapter 10. Amsterdam: North Holland.

Stock, J. H., Watson, M. W. (2009). Generalized Shrinkage Methods for Forecasting Using Many Predictors. Harvard University and Princeton University.

Timmermann, A. (2006). Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmermann, A., eds. *Handbook of Economic Forecasting*. Amsterdam: North Holland. Forthcoming.

West, K. (1996). Asymptotic inference about prediction ability. *Econometrica* 64:1067–1084.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92(4):937–950.