



FUNDAÇÃO
GETULIO VARGAS

EPGE

Escola de Pós-Graduação
em Economia

Ensaio Econômico

Escola de

Pós Graduação

em Economia

da Fundação

Getúlio Vargas

Nº 668

ISSN 0104-8910

***A Panel Data Approach to Economic
Forecasting: The Bias-Corrected Average
Forecast***

João Victor Issler, Luiz Renato Lima

Janeiro de 2008

**Os artigos publicados são de inteira responsabilidade de seus autores. As opiniões
neles emitidas não exprimem, necessariamente, o ponto de vista da Fundação
Getulio Vargas.**

A Panel Data Approach to Economic Forecasting: The Bias-Corrected Average Forecast*

João Victor Issler and Luiz Renato Lima[†]
Graduate School of Economics – EPGE
Getulio Vargas Foundation
email: jissler@fgv.br and luizr@fgv.br

First Draft: December, 2006.

This Version: February, 2008.

Abstract

In this paper, we propose a novel approach to econometric forecasting of stationary and ergodic time series within a panel-data framework. Our key element is to employ the (feasible) bias-corrected average forecast. Using panel-data sequential asymptotics we show that it is potentially superior to other techniques in several contexts. In particular, it is asymptotically equivalent to the conditional expectation, i.e., has an optimal limiting mean-squared error. We also develop a zero-mean test for the average bias and discuss the forecast-combination puzzle in small and large samples. Monte-Carlo simulations are conducted to evaluate the performance of the feasible bias-corrected average forecast in finite samples. An empirical exercise based upon data

*We are especially grateful for the comments and suggestions given by two anonymous referees, Marcelo Fernandes, Wagner Gaglianone, Antonio Galvão, Raffaella Giacomini, Clive Granger, Roger Koenker, Marcelo Medeiros, Marcelo Moreira, Zhijie Xiao, and Hal White. We also benefited from comments given by the participants of the conference “Econometrics in Rio.” We thank Wagner Gaglianone and Claudia Rodrigues for excellent research assistance. João Victor Issler thanks the hospitality of the Rady School of Management, and the Department of Economics of UCSD, where parts of this paper were written. Both authors thank the hospitality of University of Illinois, where the final revision was written. João Victor Issler thanks the support of CNPq-Brazil, CAPES, and FAPERJ. Luiz Renato Lima thanks the support of CAPES. The usual disclaimer applies.

[†]Corresponding Author.

from a well known survey is also presented. Overall, theoretical and empirical results show promise for the feasible bias-corrected average forecast.

Keywords: Forecast Combination, Forecast-Combination Puzzle, Common Features, Panel-Data, Bias-Corrected Average Forecast.

J.E.L. Codes: C32, C33, E21, E44, G12.

1 Introduction

Bates and Granger(1969) made the econometric profession aware of the benefits of forecast combination when a limited number of forecasts is considered. The widespread use of different combination techniques has lead to an interesting puzzle from the econometrics point of view – the well known *forecast combination puzzle*: if we consider a fixed number of forecasts ($N < \infty$), combining them using equal weights ($1/N$) fare better than using “optimal weights” constructed to outperform any other forecast combination in the mean-squared error (MSE) sense.

Regardless of how one combines forecasts, if the series being forecast is stationary and ergodic, and there is enough diversification among forecasts, we should expect that a weak law-of-large-numbers (WLLN) applies to well-behaved forecast combinations. This argument was considered in Palm and Zellner (1992), who asked the question “*to pool or not to pool*” forecasts? Recently, Timmermann (2006) used risk diversification – a principle so widespread in finance – to defend pooling of forecasts. Of course, to obtain this WLLN result, at least the number of forecasts has to diverge ($N \rightarrow \infty$), which entails the use of asymptotic panel-data techniques. In our view, one reason that pooling forecasts has not yet been given a full asymptotic treatment, with $N, T \rightarrow \infty$, is that forecasting is frequently thought to be a time-series experiment, not a panel-data experiment.

In this paper, we propose a novel approach to forecasting stationary and ergodic series within a panel-data framework. First, we use a two-way decomposition for the forecast error (Wallace and Hussein (1969)), where individual errors are the sum of a time-invariant forecast bias, an unforecastable aggregate zero-mean shock, and an idiosyncratic (or sub-group) zero-mean error term. Second, we show the equivalence between this two-way decomposition and a model where forecasts are a biased and error-ridden version

of the conditional expectation of the series – the optimal forecast in the MSE sense. Indeed, the latter is the *common feature* of all individual forecasts (Engle and Kozicki (1993)), while individual forecasts deviate from the optimal because of forecast misspecification. Third, letting $N, T \rightarrow \infty$, we use standard tools from panel-data asymptotic theory to show that the pooling of forecasts delivers optimal limiting forecasts in the MSE sense. In our key result, we prove that, in the limit, the feasible *bias-corrected average forecast* – equal weights in combining forecasts coupled with an estimated bias-correction term – is an optimal forecast identical to the conditional expectation.

The feasible bias-corrected average forecast is also parsimonious besides being optimal. The only parameter we need to estimate is the mean bias, for which we show consistency under the sequential asymptotic approach developed by Phillips and Moon (1999). Indeed, the only way we could increase parsimony in our framework is by doing without any bias correction. To test the usefulness of performing bias correction, we developed a zero-mean test for the average bias which draws upon the work of Conley (1999) on random fields.

As a by-product of the use of panel-data asymptotic methods, with $N, T \rightarrow \infty$, we advance the understanding of the forecast combination puzzle. The key issue is that simple averaging requires no estimation of weights, while optimal weights requires estimating N weights that grow unbounded in the asymptotic setup. We show that there is no puzzle under certain asymptotic paths for N and T , although that result does not hold generally. We are also able to discuss the puzzle in small samples, linking its presence to the curse of dimensionality which plagues so many estimators throughout econometrics¹.

Despite the scarcity of panel-data studies on the pooling of forecasts², there has been panel-data research on forecasting focusing on the pooling of information; see Stock and Watson (1999 and 2002a and b) and Forni et al. (2000, 2003). Pooling forecasts is related to forecast combination

¹We thank Roger Koenker for suggesting this asymptotic exercise to us, and an anonymous referee for casting the puzzle in terms of the curse of dimensionality.

²The notable exception is Palm and Zellner (1992), who discuss “to pool or not to pool” forecasts using a two-way decomposition. They make very limited use of the panel dimension of forecasts in their discussion. Davies and Lahiri (1995) use a three-way decomposition, but focus on forecast rationality instead of combination.

and operates a reduction on the space of forecasts. Pooling information operates a reduction on a set of highly correlated regressors. Forecasting can benefit from the use of both procedures, since, in principle, both yield asymptotically optimal forecasts in the MSE sense.

A potential limitation in the literature on pooling of information is that pooling is performed in a linear setup, and the statistical techniques employed were conceived as highly parametric – principal-component and factor analysis. That is a problem if the conditional expectation is not a linear function of the conditioning set or if the parametric restrictions used (if any) are too stringent to fit the information being pooled. In this case, pooling forecasts will be a superior choice, since the forecasts being pooled need not be the result of estimating a linear model under a highly restrictive parameterization. On the contrary, these models may be non-linear, non-parametric, and even unknown to the econometrician, as is the case of using a survey of forecasts. Moreover, the components of the two-way decomposition employed here are estimated using non-parametric techniques, dispensing with any distributional assumptions. This widens the application of the methods discussed in this paper.

The ideas in this paper are related to research done in two different fields. From econometrics, it is related to the common-features literature after Engle and Kozicki (1993). Indeed, we attempt to bridge the gap between a large literature on common features applied to macroeconomics, e.g., Vahid and Engle (1993, 1997), Issler and Vahid (2001, 2006) and Vahid and Issler (2002), and the econometrics literature on forecasting related to common factors, forecast combination, bias and intercept correction, perhaps best represented by the work of Bates and Granger (1969), Granger and Ramanathan (1984), Palm and Zellner (1992), Forni et al. (2000, 2003), Hendry and Clements (2002), Stock and Watson (2002a and b), Elliott and Timmermann (2003, 2004, 2005), Hendry and Mizon (2005), and, more recently, by the excellent surveys of Clements and Hendry (2006), Stock and Watson (2006), and Timmermann (2006) – all contained in Elliott, Granger and Timmermann (2006). From finance and econometrics, our approach is related to the work on factor analysis and risk diversification when the number of assets is large, to recent work on panel-data asymptotics, and to panel-data methods focusing on financial applications, perhaps best exemplified by the work of Chamberlain and Rothschild (1983), Connor and

Korajczyk (1986), Phillips and Moon (1999), Bai and Ng (2002), Bai (2005), and Pesaran (2005). Indeed, our approach borrows from finance the idea that we can only diversify idiosyncratic risk but not systematic risk. The latter is associated with the common element of all forecasts – the conditional expectation term – which is to what a specially designed forecast average converges to.

The rest of the paper is divided as follows. Section 2 presents our main results and the assumptions needed to derive them. Section 3 presents the results of a Monte-Carlo experiment. Section 4 presents an empirical analysis using the methods proposed here, confronting the performance of our bias-corrected average forecast with that of other types of forecast combination. Section 5 concludes.

2 Econometric Setup and Main Results

Suppose that we are interested in forecasting a weakly stationary and ergodic univariate process $\{y_t\}$ using a large number of forecasts that will be combined to yield an optimal forecast in the mean-squared error (MSE) sense. These forecasts could be the result of using several econometric models that need to be estimated prior to forecasting, or the result of using no formal econometric model at all, e.g., just the result of an opinion poll on the variable in question using a large number of individual responses. We can also imagine that some (or all) of these poll responses are generated using econometric models, but then the econometrician that observes these forecasts has no knowledge of them.

Regardless of whether forecasts are the result of a poll or of the estimation of an econometric model, we label forecasts of y_t , computed using conditioning sets lagged h periods, by $f_{i,t}^h$, $i = 1, 2, \dots, N$. Therefore, $f_{i,t}^h$ are h -step-ahead forecasts and N is either the number of models estimated to forecast y_t or the number of respondents of an opinion poll regarding y_t .

We consider 3 consecutive distinct time sub-periods, where time is indexed by $t = 1, 2, \dots, T_1, \dots, T_2, \dots, T$. The first sub-period E is labeled the “estimation sample,” where models are usually fitted to forecast y_t in the subsequent period, if that is the case. The number of observations in it is $E = T_1 = \kappa_1 \cdot T$, comprising $(t = 1, 2, \dots, T_1)$. For the other two, we follow the standard notation in West (1996). The sub-period R (for regression) is

labeled the post-model-estimation or “training sample”, where realizations of y_t are usually confronted with forecasts produced in the estimation sample, and weights and bias-correction terms are estimated, if need be. It has $R = T_2 - T_1 = \kappa_2 \cdot T$ observations in it, comprising $(t = T_1 + 1, \dots, T_2)$. The final sub-period is P (for prediction), where genuine out-of-sample forecast is entertained. It has $P = T - T_2 = \kappa_3 \cdot T$ observations in it, comprising $(t = T_2 + 1, \dots, T)$. Notice that $0 < \kappa_1, \kappa_2, \kappa_3 < 1$, $\kappa_1 + \kappa_2 + \kappa_3 = 1$, and that the number of observations in these three sub-periods keep a fixed proportion with T – respectively, κ_1 , κ_2 and κ_3 – being all $O(T)$. This is an important ingredient in our asymptotic results for $T \rightarrow \infty$.

We now compare our time setup with that of West. He only considers two consecutive periods: R data points are used to estimate models and the subsequent P data points are used for prediction. His setup does not require estimating bias-correction terms or combination weights, so there is no need for an additional sub-period for estimating the models that generate the $f_{i,t}^h$ 's³. In the case of surveys, since we do not have to estimate models, our setup is equivalent to West's. Indeed, in his setup, $R, P \rightarrow \infty$ as $T \rightarrow \infty$, and $\lim_{T \rightarrow \infty} R/P = \pi \in [0, \infty]$. Here⁴:

$$\lim_{T \rightarrow \infty} \frac{R}{P} = \frac{\kappa_2}{\kappa_3} = \pi \in (0, \infty).$$

In our setup, we also let N go to infinity, which raises the question of whether this is plausible. On the one hand, if forecasts are the result of estimating econometric models, they will differ across i if they are based either upon different conditioning sets or upon different functional forms of the conditioning set (or both). Since there is an infinite number of functional forms that could be entertained for forecasting, this gives an infinite number of possible forecasts. On the other hand, if forecasts are the result of a survey, although the number of responses is bounded from above, for all practical purposes, if a large enough number of responses is obtained, then the behavior of forecast combinations will be very close to the limiting behavior when $N \rightarrow \infty$.

³Notice that the estimated models generate the $f_{i,t}^h$'s, but model estimation, bias-correction estimation and weight estimation cannot be performed all within the same sub-sample in an out-of-sample forecasting exercise.

⁴To include the endpoints of $\pi \in [0, \infty]$ we must give up either a training sample or a genuine out-of-sample period, asymptotically.

Recall that, if we are interested in forecasting y_t , stationary and ergodic, using information up to h periods prior to t , then, under a MSE risk function, the optimal forecast is the conditional expectation using information available up to $t - h$: $\mathbb{E}_{t-h}(y_t)$. Using this well-known optimality result, Hendry and Clements (2002) argue that the fact that the simple forecast average $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$ usually outperforms individual forecasts $f_{i,t}^h$ shows our inability to approximate $\mathbb{E}_{t-h}(y_t)$ reasonably well with individual models. However, since $\mathbb{E}_{t-h}(y_t)$ is optimal, this is exactly what these individual models should be doing.

With this motivation, we write the $f_{i,t}^h$'s as approximations to the optimal forecast as follows:

$$f_{i,t}^h = \mathbb{E}_{t-h}(y_t) + k_i + \varepsilon_{i,t}, \quad (1)$$

where k_i is the individual model time-invariant bias and $\varepsilon_{i,t}$ is the individual model error term in approximating $\mathbb{E}_{t-h}(y_t)$, where $\mathbb{E}(\varepsilon_{i,t}) = 0$ for all i and t . Here, the optimal forecast is a common feature of all individual forecasts and k_i and $\varepsilon_{i,t}$ arise because of forecast misspecification⁵. We can always decompose the series y_t into $\mathbb{E}_{t-h}(y_t)$ and an unforecastable component ζ_t , such that $\mathbb{E}_{t-h}(\zeta_t) = 0$ in:

$$y_t = \mathbb{E}_{t-h}(y_t) + \zeta_t. \quad (2)$$

Combining (1) and (2) yields,

$$\begin{aligned} f_{i,t}^h &= y_t - \zeta_t + k_i + \varepsilon_{i,t}, \text{ or,} \\ f_{i,t}^h &= y_t + k_i + \eta_t + \varepsilon_{i,t}, \text{ where, } \eta_t = -\zeta_t. \end{aligned} \quad (3)$$

Equation (3) is indeed the well known two-way decomposition, or error-component decomposition, of the forecast error $f_{i,t}^h - y_t$:

$$\begin{aligned} f_{i,t}^h &= y_t + \mu_{i,t} & i = 1, 2, \dots, N, & \quad t > T_1, \\ \mu_{i,t} &= k_i + \eta_t + \varepsilon_{i,t}. \end{aligned} \quad (4)$$

⁵If an individual forecast is the conditional expectation $\mathbb{E}_{t-h}(y_t)$, then $k_i = \varepsilon_{i,t} = 0$. Notice that this implies that its MSE is smaller than that of $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$, something that is rarely seen in practice when a large number of forecasts are considered.

It has been largely used in econometrics dating back to Wallace and Hussein (1969), Amemiya (1971), Fuller and Battese (1974) and Baltagi (1980). Palm and Zellner (1992) employ a two-way decomposition to discuss forecast combination in a Bayesian and a non-Bayesian setup⁶, and Davies and Lahiri (1995) employed a three-way decomposition to investigate forecast rationality within the “Survey of Professional Forecasts.”

By construction, our framework in (4) specifies explicit sources of forecast errors that are found in both y_t and $f_{i,t}^h$; see also the discussion in Palm and Zellner and Davies and Lahiri. The term k_i is the time-invariant forecast bias of model i or of respondent i . It captures the long-run effect of forecast-bias of model i , or, in the case of surveys, the time invariant bias introduced by respondent i . Its source is $f_{i,t}^h$. The term η_t arises because forecasters do not have future information on y between $t - h + 1$ and t . Hence, the source of η_t is y_t , and it is an additive aggregate zero-mean shock affecting all forecasts equally⁷. The term $\varepsilon_{i,t}$ captures all the remaining errors affecting forecasts, such as those of idiosyncratic nature and others that affect some but not all the forecasts (a group effect). Its source is $f_{i,t}^h$.

From equation (4), we conclude that k_i , $\varepsilon_{i,t}$ and η_t depend on the fixed horizon h . Here, however, to simplify notation, we do not make explicit this dependence on h . In our context, it makes sense to treat h as fixed and not as an additional dimension to i and t . In doing that, we follow West (1996) and the subsequent literature. As argued by Vahid and Issler (2002), forecasts are usually constructed for a few short horizons, since, as the horizon increases, the MSE in forecasting gets hopelessly large. Here, h will not vary as much as i and t , especially because $N, T \rightarrow \infty$ ⁸.

⁶Palm and Zellner show that the performance of non-Bayesian combinations obey the following MSE rank: (i) the unfeasible weighted forecast with known weights performs better or equal to the simple average forecast, and (ii) the simple average forecast may perform better than the feasible weighted forecast with estimated weights. Our main result is that the feasible bias-corrected average forecast is optimal under sequential asymptotics. We also propose an explanation to the forecast-combination puzzle based on the curse of dimensionality. Critical to these results is the use of large N, T asymptotic theory.

⁷Because it is a component of y_t , and the forecast error is defined as $f_{i,t}^h - y_t$, the forecast error arising from lack of future information should have a negative sign in (4); see (3). To eliminate this negative sign, we have defined η_t as the negative of this future-information component.

⁸Davies and Lahiri considered a three-way decomposition with h as an added dimension. The focus of their paper is forecast rationality. In their approach, η_t and $\varepsilon_{i,t}$ depend on

From the perspective of combining forecasts, the components k_i , $\varepsilon_{i,t}$ and η_t play very different roles. If we regard the problem of forecast combination as one aimed at diversifying risk, i.e., a finance approach, then, on the one hand, the risk associated with $\varepsilon_{i,t}$ can be diversified, while that associated with η_t cannot. On the other hand, in principle, diversifying the risk associated with k_i can only be achieved if a bias-correction term is introduced in the forecast combination, which reinforces its usefulness.

We now list our set of assumptions.

Assumption 1 We assume that k_i , $\varepsilon_{i,t}$ and η_t are independent of each other for all i and t .

Independence is an algebraically convenient assumption used throughout the literature on two-way decompositions; see Wallace and Hussein (1969) and Fuller and Battese (1974) for example. At the cost of unnecessary complexity, it could be relaxed to orthogonality, something we avoid here.

Assumption 2 k_i is an identically distributed random variable in the cross-sectional dimension, but not necessarily independent, i.e.,

$$k_i \sim \text{i.d.}(B, \sigma_k^2), \quad (5)$$

where B and σ_k^2 are respectively the mean and variance of k_i . In the time-series dimension, k_i has no variation, therefore, it is a fixed parameter.

The idea of dependence is consistent with the fact that forecasters learn from each other by meeting, discussing, debating, etc. Through their ongoing interactions, they maintain a current collective understanding of where their target variable is most likely heading to, and of its upside and downside risks. Given the assumption of identical distribution for k_i , B represents the market (or collective) bias. Since we focus on combining forecasts, a pure idiosyncratic bias does not matter but a collective bias does. In principle, we

h but k_i does not, the latter being critical to identify k_i within their framework. Since, in general, this restriction does not have to hold, our two-way decomposition is not nested into their three-way decomposition. Indeed, in our approach, k_i varies with h and it is still identified. We leave treatment of a varying horizon, within our framework, for future research.

could allow for heterogeneity in the distribution of k_i – means and variances to differ across i . However, that will be a problem in testing the hypothesis that forecast combinations are biased.

It is desirable to discuss the nature of the term k_i , which is related to the question of why we cannot focus solely on unbiased forecasts, for which $k_i = 0$. The role of k_i is to capture the long-run effect, in the time dimension, of the bias of econometric models of y_t , or of the bias of respondent i . We first discuss survey-based forecasts. In this case, a relevant question to ask is: why would forecasters introduce bias under an MSE risk function? Laster, Bennett and Geoum (1999), Patton and Timmermann (2006), and Batchelor (2007) list different arguments consistent with forecasters having a non-quadratic loss function. Following their discussion, we assume that all forecasters employ a combination of quadratic loss and a secondary loss function. Bias is simply a consequence of this secondary loss function and of the intensity in which the forecaster cares for it. The first example is that of a bank selling an investment fund. In this case, the bank’s forecast of the fund return may be upward-biased simply because it may use this forecast as a marketing strategy to attract new clients for that fund. Although the bank is penalized by deviating from $\mathbb{E}_{t-h}(y_t)$, it also cares for selling the shares of its fund. The second example introduces bias when there is a market for *pessimism* or *optimism* in forecasting. Forecasters want to be labeled as optimists or pessimists in a “branding” strategy to be experts on “worst-” or on “best-case scenarios,” respectively. Batchelor lists governments as examples of experts on the latter.

In the case of model-based forecasts, bias results from model misspecification. Here, it is important to distinguish between in-sample and out-of-sample model fitting. The fact that, in sample, a model approximates well the data-generating process (DGP) of y_t does not guarantee that it will in out-of-sample forecasting; see the discussion in Clements and Hendry (1996) and in Hendry and Clements (2002). Notice that bias correction is a form of intercept correction. Intuitively, if $k_i > 0$, model i will consistently overpredict the target variable y_t and it is reasonable to correct its forecasts downwards by the same amount as k_i . The equivalence between bias correction and intercept correction was discussed by Hendry and Clements (2002); we discuss this equivalence below. Alternatively, Palm and Zellner (1992) list the following reasons for bias in forecasts: carelessness; the use of a poor

or defective information set or incorrect model; and errors of measurement.

Assumption 3 The aggregate shock η_t is a stationary and ergodic MA process of order at most $h - 1$, with zero mean and variance $\sigma_\eta^2 < \infty$.

Since h is a bounded constant in our setup, η_t is the result of a cumulation of shocks to y_t that occurred between $t - h + 1$ and t . Being an $MA(\cdot)$ is a consequence of the Wold representation for y_t and of (2). If y_t is already an $MA(\cdot)$ process of order smaller than $h - 1$, then its order will be the same as that of η_t . Otherwise, the order is $h - 1$. In any case, it must be stressed that η_t is unpredictable, i.e., that $\mathbb{E}_{t-h}(\eta_t) = 0$. This a consequence of (2) and of the law of iterated expectations, simply showing that, from the perspective of the forecast horizon h , unless the forecaster has superior information, the aggregate shock η_t cannot be predicted.

Assumption 4: Let $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \dots, \varepsilon_{N,t})'$ be a $N \times 1$ vector stacking the errors $\varepsilon_{i,t}$ associated with all possible forecasts, where $\mathbb{E}(\varepsilon_{i,t}) = 0$ for all i and t . Then, the vector process $\{\varepsilon_t\}$ is assumed to be covariance-stationary and ergodic for the first and second moments, uniformly on N . Further, defining $\xi_{i,t} = \varepsilon_{i,t} - \mathbb{E}_{t-1}(\varepsilon_{i,t})$, the innovation of $\varepsilon_{i,t}$, we assume that

$$\lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E}(\xi_{i,t} \xi_{j,t})| = 0. \quad (6)$$

Because the forecasts are computed h -steps ahead, forecast errors $\varepsilon_{i,t}$ can be serially correlated. Assuming that $\varepsilon_{i,t}$ is weakly stationary is a way of controlling its time-series dependence. It does not rule out errors displaying conditional heteroskedasticity, since the latter can coexist with the assumption of weak stationarity; see Engle (1982).

Equation (6) limits the degree of cross-sectional dependence of the errors $\varepsilon_{i,t}$. It allows cross-correlation of the form present in a specific group of forecasts, although it requires that this cross-correlation will not prevent a weak law-of-large-numbers from holding. Following the forecasting literature with large N and T , e.g., Stock and Watson (2002b), and the financial econometric literature, e.g., Chamberlain and Rothschild (1983), the condition $\lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E}(\xi_{i,t} \xi_{j,t})| = 0$ controls the degree of cross-sectional dependence decay in forecast errors. It is noted by Bai (2005, p. 6),

that Chamberlain and Rothschild’s cross-sectional error dependence decay requires:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E}(\xi_{i,t} \xi_{j,t})| < \infty. \quad (7)$$

Notice that this is the same cross-sectional dependence decay used in Stock and Watson. Of course, (7) implies (6), but the converse is not true. Hence, Assumption 2 has a less restrictive condition than those commonly employed in the factor-model literature.

We state now basic results related to the classic question of “to pool or not to pool forecasts,” when only simple weights ($1/N$) are used; see, for example, Granger (1989) and Palm and Zellner (1992).

Proposition 1 *Under Assumptions 1-4, the mean-squared error in forecasting y_t , using the individual forecast $f_{i,t}^h$, is $\mathbb{E}(f_{i,t}^h - y_t)^2 = k_i^2 + \sigma_\eta^2 + \sigma_{\epsilon_i}^2$, where $\sigma_{\epsilon_i}^2$ is the variance of $\epsilon_{i,t}$, $i = 1, 2, \dots, N$.*

Proposition 2 *Under Assumptions 1-4, as $N \rightarrow \infty$, the mean-squared error in forecasting y_t , combining all possible individual forecasts $f_{i,t}^h$, is*

$$MSE_{average} = \mathbb{E} \left(\underset{N \rightarrow \infty}{plim} \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - y_t \right)^2 = B^2 + \sigma_\eta^2.$$

We can now compare the MSE of a generic individual forecast with that of an equally weighted ($1/N$) forecast combination by using the usual bias-variance standard decomposition of the mean squared error (MSE)

$$MSE = Bias^2 + VAR.$$

Proposition 1 shows that we can decompose individual MSE’s, MSE_i , as:

$$\begin{aligned} MSE_i &= k_i^2 + \sigma_\eta^2 + \sigma_{\epsilon_i}^2 \\ &= Bias_i^2 + VAR_i, \quad i = 1, 2, \dots, N \end{aligned}$$

where $Bias_i^2 = k_i^2$ and $VAR_i = \sigma_\eta^2 + \sigma_{\epsilon_i}^2$. Proposition 2 shows that averaging forecasts reduces variance, but not necessarily MSE,

$$\begin{aligned} MSE_{average} &= B^2 + \sigma_\eta^2 \\ &= Bias_{average}^2 + VAR_{average}, \end{aligned} \quad (8)$$

where $VAR_{average} = \sigma_\eta^2 < VAR_i = \sigma_\eta^2 + \sigma_{\epsilon_i}^2$, but comparing $Bias_{average}^2 = B^2$ with $Bias_i^2 = k_i^2$ requires knowledge of B and k_i . If $B = 0$, i.e., we are considering unbiased forecasts, on average, then $MSE_i = k_i^2 + \sigma_\eta^2 + \sigma_{\epsilon_i}^2$, while $MSE_{average} = \sigma_\eta^2$. Therefore, if the number of forecasts in the combination is large enough, combining forecasts with a zero collective bias will lead to a smaller MSE – as concluded in Granger (1989). However, if $B \neq 0$, we cannot conclude that the average forecast has MSE lower than that of individual forecasts, since B^2 may be larger or smaller than $k_i^2 + \sigma_{\epsilon_i}^2$.

One way to eliminate the term B^2 in (8) is to perform bias correction coupled with equal weights ($1/N$) in the forecast combination. The next set of results investigates the properties of the bias-corrected average forecast (BCAF).

Proposition 3 *If Assumptions 1-4 hold, then, the bias-corrected average forecast, given by $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i$, obeys $plim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i \right) = y_t + \eta_t$ and has a mean-squared error as follows:*

$MSE_{BCAF} = \mathbb{E} \left[plim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i \right) - y_t \right]^2 = \sigma_\eta^2$. Therefore, it is an optimal forecasting device in the MSE sense.

Proposition 3 shows that the bias-corrected average forecast is an optimal forecast in the MSE sense. Bias correction eliminates the term B^2 from the MSE expression, while the use of equal weights naturally eliminates the variance of idiosyncratic components and group effects. The only term left in the MSE is σ_η^2 , related to unforecastable news to the target variable after the forecast combination was computed – something we could not eliminate unless we had superior (future) information. From a finance perspective, all risks associated with terms that could be diversified were eliminated by using the bias-corrected average forecast. We were left only with the undiversifiable risk expressed in σ_η^2 . Therefore, the optimality result.

There are infinite ways of combining forecasts. We now present an alternative weighting scheme to equal weights $1/N$.

Corollary 4 *Consider the sequence of deterministic weights $\{\omega_i\}_{i=1}^N$, such that $|\omega_i| \neq 0$, $\omega_i = \mathbf{O}(N^{-1})$ uniformly, and $\sum_{i=1}^N \omega_i = 1$, even when $N \rightarrow \infty$.*

Then, under Assumptions 1-4,

$$\begin{aligned} \text{plim}_{N \rightarrow \infty} \left(\sum_{i=1}^N \omega_i f_{i,t}^h - \sum_{i=1}^N \omega_i k_i - y_t \right) &= \eta_t, \text{ and,} \\ \mathbb{E} \left[\text{plim}_{N \rightarrow \infty} \left(\sum_{i=1}^N \omega_i f_{i,t}^h - \sum_{i=1}^N \omega_i k_i \right) - y_t \right]^2 &= \sigma_\eta^2. \end{aligned}$$

and the same result of Proposition 3 follows when a generic $\{\omega_i\}_{i=1}^N$ is used instead of $1/N$.

This corollary to Proposition 3 shows that there is not a unique optimum in the MSE sense. Indeed, any other combination scheme consistent with a WLLN will be optimal as well. Of course, optimal population weights, constructed from the variance-covariance structure of models with stationary data, will obey the structure in Corollary 4. Hence, theoretically, optimal population weights cannot perform better than $1/N$ under bias correction. Therefore, there is no forecast-combination puzzle in the context of population weights.

Although the discussion using population weights is useful, the puzzle is associated with weights ω_i estimated using data. Then, the key question to be answered analytically is why forecasts based on estimated weights are not optimal. We now compare $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$ with a bias-corrected version of

$\sum_{i=1}^N \omega_i f_{i,t}^h$ that uses estimated weights. We follow the discussion in Hendry and Clements (2002) using N different forecasts instead of just 2. Weights ω_i can be estimated ($\hat{\omega}_i$) by running the following regression, minimizing MSE subject to $\sum_{i=1}^N \omega_i = 1$:

$$y = \delta \mathbf{i} + \omega_1 f_1 + \omega_2 f_2 + \dots + \omega_N f_N + v, \quad (9)$$

where y denotes the $R \times 1$ vector of observations of the target variable, f_1, f_2, \dots, f_N denotes, respectively, the $R \times 1$ vectors of observations of the N individual forecasts, and \mathbf{i} is a vector of ones. Estimation is done over the time interval $T_1 + 1, \dots, T_2$ (i.e., over the training sample).

There are two cases to be considered: the behavior of estimated weights in small samples and asymptotically, when $N, T \rightarrow \infty$. In large samples, consistent estimation of weights requires:

$$0 < \lim_{N, T \rightarrow \infty} \frac{N}{R} = \lim_{N, T \rightarrow \infty} \frac{N/T}{\frac{T_2 - T_1}{T}} = \frac{\lim_{N, T \rightarrow \infty} N/T}{\kappa_2} < 1, \quad (10)$$

which implies not only that $N \rightarrow \infty$ at a smaller rate than T , but that $\lim_{N, T \rightarrow \infty} N/T < \kappa_2$. Recall that $\kappa_2 = 1 - \kappa_1 - \kappa_3$, hence, $\lim_{N, T \rightarrow \infty} N/T \ll 1$. As long as this condition is achieved, weights are estimated consistently in (9) and we are back to Corollary 4: asymptotically, there is no forecast-combination puzzle. Notice that, because regression (9) includes an intercept, the forecast $\hat{f} = \hat{\delta}\mathbf{i} + \hat{\omega}_1 f_1 + \hat{\omega}_2 f_2 + \dots + \hat{\omega}_N f_N$ is unbiased with δ playing the role of a bias-correction term.

Because there is no puzzle asymptotically, the low accuracy of \hat{f} relative to that of $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$ must reflect a poor small-sample approximation of population weights ω_i s by the $\hat{\omega}_i$ s. In small samples, estimation of ω_i requires $N < R$. On the one hand, to get close to an optimal weighted forecast, we need a large N . On the other hand, the forecast $\hat{f} = \hat{\delta}\mathbf{i} + \hat{\omega}_1 f_1 + \hat{\omega}_2 f_2 + \dots + \hat{\omega}_N f_N$ is not immune to the “curse of dimensionality,” since, as N increases, we need to estimate an increasing number of weights. Increasing N contributes to raising the variance of estimated weights, which works against the consistency of the $\hat{\omega}_i$ s.

As shown in our simulation exercise below, when N approaches R from below, the variance of \hat{f} grows with N and is typically big enough as to yield an inferior forecast (MSE) relative to $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$, although the latter is biased. Thus, the weighted forecast cannot avoid the “curse of dimensionality” that plagues several estimates across econometrics. In this context, the curse of dimensionality in \hat{f} is an explanation to the forecast-combination puzzle. Of course, if we make N/R close to zero in finite samples, with N large, we allow $\hat{\omega}_i \xrightarrow{p} \omega_i$, since R will then be large. Moreover, the requirement of large N guarantees the optimality result in Corollary 4.

The main question of this paper is: is there a forecast combination device that is asymptotically equivalent to the optimal weighted forecast, in the

MSE sense, but does not suffer from the curse of dimensionality? Our answer to this question is the BCAF in Proposition 3. It borrows only the positive properties of fixed-weight ($1/N$) and estimated-weight forecasts. From the former, it inherits the fact that its variance decreases with N . From the latter, it inherits bias correction.

Despite its optimal behavior, it is immediately seen that the BCAF $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i$ is unfeasible since the k_i 's are unknown. Therefore, below, we propose replacing k_i with a consistent estimator. The underlying idea behind the consistent estimator of k_i is that one observes the realizations of y_t and $f_{i,t}^h$, $i = 1 \dots N$, for the R training-sample observations. Hence, one can form a panel of forecasts:

$$\left(f_{i,t}^h - y_t \right) = k_i + \eta_t + \varepsilon_{i,t}, \quad i = 1, 2, \dots, N, \quad t = T_1 + 1, \dots, T_2, \quad (11)$$

where it becomes obvious that k_i represents the fixed effect of this panel.

We propose a non-parametric estimator of k_i . It does not depend on any distributional assumption on $k_i \sim \text{i.d.}(B, \sigma_k^2)$ and it does not depend on any knowledge of the models used to compute the forecasts $f_{i,t}^h$. This feature of our approach widens its application to situations where the “underlying models are not known, as in a survey of forecasts,” as discussed by Kang (1986).

Due to the nature of our problem, which consists of a large number of forecasts and a time-invariant bias term (k_i), we need to consider large N , large T asymptotic theory to devise a consistent estimator for k_i . Searching for a method that allows robust asymptotic results without imposing too many restrictions (on functional relations and the existence of higher moments), we consider the sequential asymptotic approach developed by Phillips and Moon (1999). There, one first fixes N and then allows T to pass to infinity using an intermediate limit. Phillips and Moon write sequential limits of this type as $(T, N \rightarrow \infty)_{\text{seq}}$. We now show how to estimate consistently k_i , B , η_t , and $\varepsilon_{i,t}$, in this context.

Proposition 5 *If Assumptions 1-4 hold, the following are consistent esti-*

mators of k_i , B , η_t , and $\varepsilon_{i,t}$, respectively:

$$\begin{aligned}\widehat{k}_i &= \frac{1}{R} \sum_{t=T_1+1}^{T_2} f_{i,t}^h - \frac{1}{R} \sum_{t=T_1+1}^{T_2} y_t, & \text{plim}_{T \rightarrow \infty} (\widehat{k}_i - k_i) &= 0, \\ \widehat{B} &= \frac{1}{N} \sum_{i=1}^N \widehat{k}_i, & \text{plim}_{(T,N \rightarrow \infty)_{\text{seq}}} (\widehat{B} - B) &= 0, \\ \widehat{\eta}_t &= \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} - y_t, & \text{plim}_{(T,N \rightarrow \infty)_{\text{seq}}} (\widehat{\eta}_t - \eta_t) &= 0, \\ \widehat{\varepsilon}_{i,t} &= f_{i,t}^h - y_t - \widehat{k}_i - \widehat{\eta}_t, & \text{plim}_{(T,N \rightarrow \infty)_{\text{seq}}} (\widehat{\varepsilon}_{i,t} - \varepsilon_{i,t}) &= 0.\end{aligned}$$

We now state our most important result.

Proposition 6 *If Assumptions 1-4 hold, the feasible bias-corrected average forecast $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B}$ obeys $\text{plim}_{(T,N \rightarrow \infty)_{\text{seq}}} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} \right) = y_t + \eta_t = \mathbb{E}_{t-h}(y_t)$ and has a mean-squared error as follows:*

$$\mathbb{E} \left[\text{plim}_{(T,N \rightarrow \infty)_{\text{seq}}} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} \right) - y_t \right]^2 = \sigma_\eta^2. \quad \text{Therefore it is an optimal forecasting device.}$$

Proposition 6 shows that the feasible BCAF is asymptotically equivalent to the optimal weighted forecast. Its advantage is that it employs equal weights. As $N \rightarrow \infty$, the number of estimated parameters is kept at unity: \widehat{B} . This is a very attractive feature of the BCAF compared to devices that combine forecasts using estimated weights. Our answer to the curse of dimensionality is parsimony, implied by estimating only one parameter – \widehat{B} . Here, we need not limit the asymptotic path of N, T as is the case of forecasts based on estimated weights.

We must stress that bias-correction can be viewed as a form of intercept correction as discussed in Palm and Zellner (1992) and Hendry and Mizon (2005), for example. From (9), we could retrieve \widehat{B} from an OLS regression of the form:

$$y = \delta \mathbf{i} + \omega_1 f_1 + \omega_2 f_2 + \dots + \omega_N f_N + v,$$

where the weights ω_i are constrained to be $\omega_i = 1/N$ for all i . There is only one parameter to be estimated, δ , and $\widehat{\delta} = -\widehat{B}$, where \widehat{B} is now cast in

terms of the previous literature⁹.

Finally, we propose a new test for the usefulness of bias correction ($H_0 : B = 0$) using the theory of random fields as in Conley (1999). It is potentially relevant when we view B as a *market* or a *consensus* bias, which is customary in the finance and macroeconomics literature, respectively. When $B = 0$, the feasible BCAAF becomes $\frac{1}{N} \sum_{i=1}^N f_{i,t}^h$.

Proposition 7 *Under the null hypothesis $H_0 : B = 0$, the test statistic:*

$$\hat{t} = \frac{\hat{B}}{\sqrt{\hat{V}}} \xrightarrow[(T, N \rightarrow \infty)_{\text{seq}}]{d} \mathcal{N}(0, 1),$$

where \hat{V} is a consistent estimator of the asymptotic variance of $\bar{B} = \frac{1}{N} \sum_{i=1}^N k_i$.

2.1 The BCAAF and Nested Models

It is important to discuss whether and when the techniques above are applicable to the situation where some (or all) of the models we combine are nested. The potential problem is that the innovations from nested models can exhibit high cross-sectional dependence, violating Assumption 4. In what follows, we introduce nested models into our framework in the following way. Consider a continuous set of models and split the total number of models N into M classes (or blocks), each of them containing m nested models, so that $N = mM$. In the index of forecasts, $i = 1, \dots, N$, we group nested models contiguously. Hence, models within each class are nested but models across classes are non-nested. We make the number of classes and the number of models within each class to be functions of N , respectively as follows: $M = N^{1-d}$ and $m = N^d$, where $0 \leq d \leq 1$. Notice that this setup considers all the relevant cases: (i) $d = 0$ corresponds to the case in which all models are non-nested; (ii) $d = 1$ corresponds to the case in which all models are nested and; (iii) the intermediate case $0 < d < 1$ gives rise to N^{1-d} blocks of nested models, all with size N^d .

⁹Recall that \hat{B} requires estimating only one \hat{k}_i separately using R observations. For forecasts based on estimated weights, joint estimation of $N + 1$ parameters is performed using these same R observations.

For each block of nested models, Assumption 4 may not hold because the innovations from that block can exhibit high cross-sectional dependence¹⁰. Regarding the interaction across blocks of nested models, it is natural to impose that the correlation structure of innovations across classes is such that Assumption 4 holds, since we should expect that the cross-sectional dependence of forecast errors across classes is weak. We formalize this by using the following extension to equation (6) in Assumption 4.

Assumption 5: Consider the covariance matrix of innovations, given by $(\mathbb{E}(\xi_{i,t}\xi_{j,t}))$, and partition it into blocks. There are M main-diagonal blocks, each with $m^2 = N^{2d}$ elements. These blocks contain the covariance structure of innovations for each class of nested models. There are also $M^2 - M$ off-diagonal blocks which represents the across-block covariance structure of innovations. Index the classes (blocks) of models by $r = 1, \dots, M$, and models within each class (block) by $s = 1, \dots, m$. For all t , and any r and s , we may re-index $\xi_{r,s,t}$ to $\xi_{(r-1)m+s,t} = \xi_{i,t}$, $i = 1, \dots, N$ ¹¹. Within each block r , we assume that:

$$0 \leq \lim_{m \rightarrow \infty} \frac{1}{m^2} \sum_{k=1}^m \sum_{s=1}^m |\mathbb{E}(\xi_{r,k,t}\xi_{r,s,t})| = \lim_{N \rightarrow \infty} \frac{1}{N^{2d}} \sum_{k=1}^{N^d} \sum_{s=1}^{N^d} |\mathbb{E}(\xi_{r,k,t}\xi_{r,s,t})| < \infty, \quad (12)$$

being zero when the smallest nested model is correctly specified. However, across any two blocks r and l , $r \neq l$, we assume that:

$$\lim_{m \rightarrow \infty} \frac{1}{m^2} \sum_{k=1}^m \sum_{s=1}^m |\mathbb{E}(\xi_{r,k,t}\xi_{l,s,t})| = \lim_{N \rightarrow \infty} \frac{1}{N^{2d}} \sum_{k=1}^{N^d} \sum_{s=1}^{N^d} |\mathbb{E}(\xi_{r,k,t}\xi_{l,s,t})| = 0. \quad (13)$$

We now discuss how to implement the BCAF in the presence of nested models. The matrix $(\mathbb{E}(\xi_{i,t}\xi_{j,t}))$ has $N^2 = m^2M^2$ elements, but only $m^2M = N^{2d}N^{1-d} = N^{1+d}$ represent covariances among nested models, in which condition (12) holds. For the remaining elements, (13) holds. Unless $d = 1$, the total number of elements representing covariances among nested

¹⁰As discussed in great detail below, Assumption 4 will hold for nested models if the smallest nested model is correctly specified.

¹¹For example, $\xi_{2,1,t}$ is the first innovation for the second block, and corresponds to $\xi_{m+1,t}$, the $(m+1)$ -th cross-sectional observation.

models (N^{1+d}) will grow at a rate smaller than N^2 – the total number of elements. Therefore, Assumption 4 will still hold in the presence of nested models when $0 < d < 1$, and our result on optimality of the BCAF will still follow. Notice that $0 < d < 1$ corresponds to the case in which nested and non-nested models are combined and the number of nested models grows with N .

We now consider two special cases: if $d = 0$, there are only non-nested models, since each class of models has only one element. Hence, Assumption 4 holds. If $d = 1$, we only have one class of models, and all models are nested. In general, Assumption 4 will not hold, and a law-of-large numbers will not apply to $\frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}$. We now discuss in detail the special case in which $d = 1$ and the smallest model is correctly specified. It is easy to verify that, for the smallest model, the following equation will hold with $k_i = \varepsilon_{i,t} = 0$ in population:

$$f_{i,t}^h = \mathbb{E}_{t-h}(y_t) + k_i + \varepsilon_{i,t}. \quad (14)$$

This happens because, under correct specification of the smallest model in the nesting scheme, its population forecast will be $\mathbb{E}_{t-h}(y_t)$, making $k_i = \varepsilon_{i,t} = 0$. This, in turn, makes $\xi_{i,t} = \varepsilon_{i,t} - \mathbb{E}_{t-1}(\varepsilon_{i,t}) = 0$ for the smallest model. When the smallest nesting model is correctly specified, all models that nest it will have identical population errors, i.e., will have $k_i = \varepsilon_{i,t} = 0$ as well. This happens because any model that nests the smallest one will have irrelevant estimated parameters that will converge to zero, in probability. Therefore, condition (12) will hold as an equality, and a law-of-large numbers will apply to $\frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}$ as a degenerate case, where there is no variation in population, among the elements being combined.

From an empirical point of view, d can be regarded as a choice variable when implementing the BCAF. Choosing $d = 1$ (only one class of nested models) is an “excellent” choice when the class of model chosen is correctly specified (the smallest model is correctly specified). However, there is the (potentially high) risk of incorrect specification for the whole class, which will imply that the law-of-large numbers will not hold for $\frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}$ in implementing the BCAF. On the other hand, in choosing $d = 0$ (all models are non-nested), we completely eliminate nested models and the chance that the law-of-large numbers will not hold. In the intermediate case, $0 < d < 1$, we have nested models but can still apply a law-of-large numbers for $\frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}$ and implement the BCAF successfully. Here, keeping some nested models

poses no problem at all, since the mixture of models will still deliver the optimal forecast. From a practical point of view, the choice of $0 \leq d < 1$ seems to be superior. Here, we are back to the main theorem in finance about risk diversification: do not put all your eggs in the same basket, choosing a large enough number of diversified (classes of) models.

3 Monte-Carlo Study

3.1 Experiment design

We follow the setup presented in the theoretical part of this paper in which each forecast is the conditional expectation of the target variable plus an additive bias term and an idiosyncratic error. Our DGP is a simple stationary $AR(1)$ process:

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 y_{t-1} + \xi_t, \quad t = 1, \dots, T_1, \dots, T_2, \dots, T \\ \xi_t &\sim \text{i.i.d.} N(0, 1), \quad \alpha_0 = 0, \text{ and } \alpha_1 = 0.5, \end{aligned} \quad (15)$$

where ξ_t is an unpredictable aggregate zero-mean shock. We focus on one-step-ahead forecasts for simplicity. The conditional expectation of y_t is $\mathbb{E}_{t-1}(y_t) = \alpha_0 + \alpha_1 y_{t-1}$. Since ξ_t is unpredictable, the forecaster should be held accountable for $f_{i,t} - \mathbb{E}_{t-1}(y_t)$. These deviations have two terms: the individual specific biases (k_i) and the idiosyncratic or group error terms ($\varepsilon_{i,t}$). Because $\xi_t \sim \text{i.i.d.} N(0, 1)$, the optimal theoretical MSE is unity in this exercise.

The conditional expectation $\mathbb{E}_{t-1}(y_t) = \alpha_0 + \alpha_1 y_{t-1}$ is estimated using a sample of size 200, i.e., $E = T_1 = 200$, so that $\hat{\alpha}_0 \simeq \alpha_0$ and $\hat{\alpha}_1 \simeq \alpha_1$. In practice, however, forecasters may have economic incentives to make biased forecasts, and there may be other sources of misspecification arising from misspecification errors. Therefore, we generate forecasts as:

$$\begin{aligned} f_{i,t} &= \hat{\alpha}_0 + \hat{\alpha}_1 y_{t-1} + k_i + \varepsilon_{i,t}, \\ &= (\hat{\alpha}_0 + k_i) + \hat{\alpha}_1 y_{t-1} + \varepsilon_{i,t} \text{ for } t = T_1 + 1, \dots, T, \quad i = 1, \dots, N, \end{aligned} \quad (16)$$

where, $k_i = \beta k_{i-1} + u_i$, $u_i \sim \text{i.i.d.} \text{Uniform}(a, b)$, $0 < \beta < 1$, and $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \dots, \varepsilon_{N,t})'$, $N \times 1$, is drawn from a multivariate Normal distribution with size $R + P = T - T_1$, whose mean vector equals zero and covariance matrix equals Σ . We introduce heterogeneity and spatial dependence in the

distribution of $\varepsilon_{i,t}$. The diagonal elements of $\Sigma = (\sigma_{ij})$ obey: $1 < \sigma_{ii} < \sqrt{10}$, and off-diagonal elements obey: $\sigma_{ij} = 0.5$, if $|i - j| = 1$, $\sigma_{ij} = 0.25$, if $|i - j| = 2$, and $\sigma_{ij} = 0$, if $|i - j| > 2$. The exact values of the σ_{ii} 's are randomly determined through an once-and-for-all draw from a uniform random variable of size N , that is, $\sigma_{ii} \sim \text{i.i.d. Uniform}(1, \sqrt{10})$ ¹².

In equation (16), we built spatial dependence in the bias term k_i ¹³. The cross-sectional average of k_i is $\frac{a+b}{2(1-\beta)}$. We set the degree of spatial dependence in k_i by letting $\beta = 0.5$. For the support of u_i , we consider two cases: (i) $a = 0$ and $b = 0.5$ and; (ii) $a = -0.5$ and $b = 0.5$. This implies that the average bias is $B = 0.5$ in (i), whereas it is $B = 0$ in (ii). Finally, notice that the specification of $\varepsilon_{i,t}$ satisfies Assumption 4 in Section 2 as we let $N \rightarrow \infty$.

Equation (16) is used to generate different panels of forecasts in terms of the number of forecasters (N) and the size of the training-sample (R). We kept the number of out-of-sample observations equal to $P = 50$ in all cases. First, for $N = 10, 20, 40$, we let $R = 50$. For all these 3 cases, $\frac{N}{R}$ may not be small enough to guarantee a good approximation of optimal population weights by the $\hat{\omega}_i$ s. In order to approximate the asymptotic environment needed for optimality of the forecasts based on estimated weights, i.e., equation (10), we also considered the case in which $R = 500$, and 1,000, while keeping $N = 10$. For each N , we conducted 50,000 simulations in the experiment. In all cases, the total number of time observations used to fit models equals $E = T_1 = 200$.

3.2 Forecast approaches

In our simulations, we evaluate three forecasting methods: the feasible bias-corrected average forecast (BCAF), the forecast based on estimated weights, and the forecast based on fixed weights. For these methods, our results include aspects of the whole distribution of their respective biases and MSEs.

For the BCAF, we use the training-sample observations to estimate $\hat{k}_i =$

¹²The covariance matrix Σ does not change over simulations.

¹³The additive bias k_i is explicit in (16). It could be implicit if we had considered a structural break in the target variable as did Hendry and Clements (2002). There, an intercept change in y_t takes place right after the estimation of econometric models, biasing all forecasts. Hence, in their paper, intercept correction is equivalent to bias correction, which would be the case here too. However, a structural break would violate weak stationarity and that is why it is not attempted here.

$\frac{1}{R} \sum_{t=T_1+1}^{T_2} (y_t - f_{i,t})$ and $\widehat{B} = \frac{1}{N} \sum_{i=1}^N \widehat{k}_i$. Then, we compute the out-of-sample forecasts $\widehat{f}_t^{BCAF} = \frac{1}{N} \sum_{i=1}^N f_{i,t} - \widehat{B}$, $t = T_2 + 1, \dots, T$, and we employ the last P observations to compute $MSE_{BCAF} = \frac{1}{P} \sum_{t=T_2+1}^T (y_t - \widehat{f}_t^{BCAF})^2$.

For the forecast based on estimated weights (weighted average forecast), we use R observations of the training sample to estimate weights (ω_i) by OLS in:

$$y = \delta \mathbf{i} + \omega_1 f_1 + \omega_2 f_2 + \dots + \omega_N f_N + \varepsilon,$$

where the restriction $\sum_{i=1}^N \omega_i = 1$ is imposed in estimation. The weighted forecast is $\widehat{f}_t^{weighted} = \widehat{\delta} + \widehat{\omega}_1 f_{1,t} + \widehat{\omega}_2 f_{2,t} + \dots + \widehat{\omega}_N f_{N,t}$, and the intercept δ plays the role of bias correction. We employ the last P observations to compute $MSE_{weighted} = \frac{1}{P} \sum_{t=T_2+1}^T (y_t - \widehat{f}_t^{weighted})^2$.

For the forecast based on fixed weights (average forecast), there is no parameter to be estimated using training-sample observations. Out-of-sample forecasts are computed according to $f_t^{average} = \frac{1}{N} \sum_{i=1}^N f_{i,t}$, $t = T_2 + 1, \dots, T$, and its MSE is computed as $MSE_{average} = \frac{1}{P} \sum_{t=T_2+1}^T (y_t - f_t^{average})^2$.

Finally, for each approach, we also computed the out-of-sample mean biases. In small samples, the weighted forecast and the BCAF should have out-of-sample mean biases close to zero, whereas the mean bias of the average forecast should be close to $B = \frac{a+b}{2(1-\beta)}$.

3.3 Simulation Results

With the results of the 50,000 replications, we describe the empirical distributions of the bias and the MSE of all three forecasting methods. For each distribution we compute the following statistics: (i) kurtosis; (ii) skewness, (iii) mean and (iv) τ -th unconditional quantile, with $\tau = 0.01, 0.25, 0.50, 0.75$, and 0.99. In doing so, we seek to have a general description of all three forecasting approaches.

The main results are presented in Tables 1 through 4. In Table 1, where $B = 0.5$, the average bias across simulations of the BCAF and the weighted

forecast are practically zero. The mean bias of the simple average forecast is between 0.39 and 0.46, depending on N . In terms of MSE, the BCAF performs very well compared to the other two methods. The simple average forecast has a mean MSE at least 8.7% higher than that of the bias-corrected average forecast, reaching 17.8% higher when $N = 40$ ¹⁴. The forecast based on estimated weights has a mean MSE at least 22.7% higher, reaching 431.3% higher when $N = 40$. This last result is a consequence of the increase in variance as we increase N , with R fixed, and N/R close to unity. Notice that the average bias is virtually zero for $N = 10, 20, 40$. Since the MSE triples when N is increased from 10 to 40, all the increase in MSE is due to variance, revealing the curse-of-dimensionality working against the forecast based on estimated weights.

Table 2 presents the results when $B = 0$. In this case, the optimal forecast is the simple average, since there is no need to estimate a bias-correction term. In terms of MSE, comparing the simple-average forecast with the BCAF, we observe that they are almost identical – the mean MSE of the BCAF is about 1% higher than that of the average forecast, showing that not much is lost in terms of MSE when we perform an unnecessary bias correction. The behavior of the weighted average forecast is identical to that in Table 1.

Table 3 presents the result in which $B = 0.5$, $\frac{N}{R} \approx 0$, and $N = 10$, with $R = 500, 1, 000$. As expected, the $\hat{\omega}_i$ s are a very good approximation to optimal population weights. Despite that, we are still combining a relatively low number of forecasts: $N = 10$. Here, there is practically no difference in performance between the BCAF and the estimated-weight forecast. However, contrary to the results in Tables 1 and 2, equal-weight forecasts perform worse than estimated-weight forecasts. Indeed, the MSE of the simple average is at least 8.9% higher than that of the estimated-weight forecast, while the latter has an almost identical accuracy to the BCAF: no bias, and a variance (and MSE) that is more than 50% than that of the theoretical optimum $\sigma_\eta^2 = 1$.

Table 4 presents the result in which $B = 0.5$, $\frac{N}{R} \approx 0$, and $N = 40$, with $R = 2, 000, 4, 000$. Although the ratio $\frac{N}{R}$ was kept identical to that in Table

¹⁴When we compare the MSE of the fixed-weight forecast with that of the optimal forecast (which is unity in our simulation), we observe that the former converges towards $\sigma_\eta^2 + B^2$. Such a result is expected since \hat{B} converges in probability to B as $N \rightarrow \infty$.

3, we have now increased both N and R proportionally. As in Table 3, the $\hat{\omega}_i$ s are a very good approximation to optimal population weights, since R is large. However, contrary to the results there, now we are also combining an increasing number of forecasts: 40 vs. 10. As a result, not only BCAF and the estimated-weight forecast outperform the fixed-weight forecast, but they also are closer to the theoretical optimum $\sigma_\eta^2 = 1$ (about 15% worse here as opposed to more than 50% in Table 3).

One key insight from the results in Tables 1-4 is that no combination device outperforms the BCAF by a wide margin, and, even when the BCAF is not constructed to be optimal – Table 2 – its performance is practically identical to that of the optimal forecast.

4 Empirical Application

4.1 The Central Bank of Brazil’s “Focus Forecast Survey”

The “Focus Forecast Survey,” collected by the Central Bank of Brazil, is a unique panel database of forecasts. It contains forecast information on almost 120 institutions, including commercial banks, asset-management firms, and non-financial institutions, which are followed throughout time with a reasonable turnover. Forecasts have been collected since 1998, on a monthly frequency, and a fixed horizon, which potentially can serve to approximate a large N, T environment for techniques designed to deal with unbalanced panels – which is not the case studied here. Besides the large size of N and T , the Focus Survey also has the following desirable features: the anonymity of forecasters is preserved, although the names of the top-five forecasters for a given economic variable are released by the Central Bank of Brazil; forecasts are collected at different frequencies (monthly, semi-annually, annually), as well as at different forecast horizons (e.g., short-run forecasts are obtained for h from 1 to 12 months); there is a large array of macroeconomic time series included in the survey.

To save space, we focus our analysis on the behavior of forecasts of the monthly inflation rate in Brazil (π_t), in percentage points, as measured by the official Consumer Price Index (CPI), computed by FIBGE. In order to obtain the largest possible balanced panel ($N \times T$), we used $N = 18$ and a time-series sample period covering 2002:11 through 2006:3 ($T = 41$).

Of course, in the case of a survey panel, there is no estimation sample. We chose the first $R = 26$ time observations to compute \widehat{B} – the average bias – leaving $P = 18$ time-series observations for out-of-sample forecast evaluation. The forecast horizon chosen was $h = 6$, this being an important horizon to determine future monetary policy within the Brazilian Inflation-Targeting program.

The results of our empirical exercise are presented in Tables 5 and 6. First, we note that all the 18 individual forecasts perform worse than combinations, which is consistent with the discussion in Hendry and Clements (2002). The results in Table 5 show that the average bias is positive for the 6-month horizon, 0.06187, and marginally significant, with a p-value of 0.063. This is a sizable bias – approximately 0.75 percentage points in a yearly basis, for an average inflation rate of 5.27% a year. In Table 6, out-of-sample forecast comparisons between the simple average and the bias-corrected average forecast show that the former has a MSE 18.2% bigger than that of the latter. We also computed the MSE of the weighted forecast. Since we have $N = 18$ and $R = 26$, $N/R = 0.69$. Hence, the weighted average cannot avoid the curse of dimensionality, yielding an MSE 390.2% bigger than that of the BCAF.

It is important to stress that, although the feasible BCAF was conceived for a large N, T environment, the empirical results here show an encouraging performance even in a small N, T context. Also, the forecasting gains from bias correction are non-trivial.

5 Conclusions and Extensions

In this paper, we propose a novel approach to the econometric forecast of stationary and ergodic series y_t within a panel-data framework, where the number of forecasts and the number of time periods increase without bound. The basis of our method is a two-way decomposition of forecast errors. As shown here, this is equivalent to forecasters trying to approximate the optimal forecast under quadratic loss – the conditional expectation $\mathbb{E}_{t-h}(y_t)$, which is modelled as the *common feature* of all individual forecasts. Standard tools from panel-data asymptotic theory are used to devise an optimal forecasting combination that delivers $\mathbb{E}_{t-h}(y_t)$. This optimal combination uses equal weights and an estimated bias-correction term. The use of equal

weights avoids estimating forecast weights, which contributes to the reduction of forecast variance, although potentially at the cost of an increase in bias. The use of an estimated bias-correction term eliminates any possible detrimental effect arising from equal weighting. We label this optimal forecast as the (feasible) *bias-corrected average forecast*.

We show that the feasible BCAAF delivers the optimality result even under the presence of nested models and we fully characterize it by using a novel framework. As a by-product of the use of panel-data asymptotic methods, with $N, T \rightarrow \infty$, we advance the understanding of the forecast combination puzzle by showing that the low accuracy of the forecasts based on estimated weights, relative to those based on fixed weights, reflects a poor small-sample approximation of optimal population weights ω_i s by estimated weights. In small samples, estimation of ω_i requires $N < R$. On the one hand, to get close to an optimal weighted forecast, we need a large N . On the other hand, the forecast based on estimated weights is not immune to the “curse of dimensionality,” since, as N increases, we need to estimate an increasing number of weights. In our simulations, we show that the curse of dimensionality works against forecasts based on estimated weights, increasing their MSE as N approaches R .

Finally, we show that there is no forecast-combination puzzle under certain asymptotic paths for N and T , but not in all cases. Indeed, if $N \rightarrow \infty$ at a rate strictly smaller than T , then ω_i is consistently estimated, the weighted forecast with bias correction (intercept) is optimal, and there is no puzzle. Our simulations approximate this asymptotic environment by considering various cases in which $N/T \approx 0$. As expected, the forecast based on estimated weights outperforms the simple fixed-weight combination and has the same performance as the BCAAF. Since the case in which $N/T \approx 0$ is rarely observed in practice, we should not expect forecasts based on estimated-weight combinations to be accurate. On the other hand, the feasible BCAAF is asymptotically equivalent to the optimal-weight forecast but has a superior performance in small samples.

References

- [1] Amemiya, T. (1971), “The estimation of the variances in the variance-components model”, *International Economic Review*, vol. 12, pp. 1-13.

- [2] Baltagi, Badi H., 1980, "On Seemingly Unrelated Regressions with Error Components," *Econometrica*, Vol. 48(6), pp. 1547-1551.
- [3] Bai, J., (2005), "Panel Data Models with Interactive Fixed Effects," Working Paper: New York University.
- [4] Bai, J., and S. Ng, (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.
- [5] Bates, J.M. and Granger, C.W.J., 1969, "The Combination of Forecasts," *Operations Research Quarterly*, vol. 20, pp. 309-325.
- [6] Batchelor, R., 2007, "Bias in macroeconomic forecasts," *International Journal of Forecasting*, vol. 23, pp. 189-203.
- [7] Chamberlain, Gary, and Rothschild, Michael, (1983). "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, vol. 51(5), pp. 1281-1304.
- [8] Clark, T.E. and McCracken, M.W., 2007, "Combining Forecasts for Nested Models," Working Paper: Kansas City FED, forthcoming in the *Journal of Econometrics*.
- [9] Clements, M.P. and Hendry, D.F. (1996), "Intercept Corrections and Structural Change", *Journal of Applied Econometrics* , vol. 11, pp. 475-494.
- [10] Clements, M.P. and D.F. Hendry, 2006, Forecasting with Breaks in Data Processes, in C.W.J. Granger, G. Elliott and A. Timmermann (eds.) *Handbook of Economic Forecasting*, pp. 605-657, Amsterdam, North-Holland.
- [11] Conley, T.G., 1999, "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, Vol. 92 Issue 1, pp. 1-45.
- [12] Connor, G., and R. Korajczyk (1986), "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, 15, 373-394.
- [13] Davies, A. and Lahiri, K., 1995, "A new framework for analyzing survey forecasts using three-dimensional panel data," *Journal of Econometrics*, vol. 68(1), pp. 205-227

- [14] Elliott, G., C.W.J. Granger, and A. Timmermann, 2006, Editors, *Handbook of Economic Forecasting*, Amsterdam: North-Holland.
- [15] Elliott, G. and A. Timmermann (2005), "Optimal forecast combination weights under regime switching", *International Economic Review*, 46(4), 1081-1102.
- [16] Elliott, G. and A. Timmermann (2004), "Optimal forecast combinations under general loss functions and forecast error distributions", *Journal of Econometrics* 122:47-79.
- [17] Engle, R.F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, pp. 987-1006.
- [18] Engle, R. F., Issler, J. V., 1995, "Estimating common sectoral cycles," *Journal of Monetary Economics*, vol. 35, 83–113.
- [19] Engle, R.F. and Kozicki, S. (1993). "Testing for Common Features", *Journal of Business and Economic Statistics*, 11(4): 369-80.
- [20] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), "The Generalized Dynamic Factor Model: Identification and Estimation", *Review of Economics and Statistics*, 2000, vol. 82, issue 4, pp. 540-554.
- [21] Forni M., Hallin M., Lippi M. and Reichlin L., 2003 "The Generalized Dynamic Factor Model one-sided estimation and forecasting," *Journal of the American Statistical Association*, forthcoming.
- [22] Fuller, Wayne A. and George E. Battese, 1974, "Estimation of linear models with crossed-error structure," *Journal of Econometrics*, Vol. 2(1), pp. 67-78.
- [23] Granger, C.W.J., 1989, "Combining Forecasts-Twenty Years Later," *Journal of Forecasting*, vol. 8, 3, pp. 167-173.
- [24] Granger, C.W.J., and R. Ramanathan (1984), "Improved methods of combining forecasting", *Journal of Forecasting* 3:197–204.
- [25] Hendry, D.F. and M.P. Clements (2002), "Pooling of forecasts", *Econometrics Journal*, 5:1-26.

- [26] Hendry, D.F. and Mizon, G.E. (2005): “Forecasting in the Presence of Structural Breaks and Policy Regime Shifts”, in “*Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*,” D.W.K. Andrews and J.H. Stock (eds.), Cambridge University Press.
- [27] Issler, J. V., Vahid, F., 2001, “Common cycles and the importance of transitory shocks to macroeconomic aggregates,” *Journal of Monetary Economics*, vol. 47, 449–475.
- [28] Issler, J. V., Vahid, F., 2006, “The missing link: Using the NBER recession indicator to construct coincident and leading indices of economic activity,” *Annals Issue of the Journal of Econometrics on Common Features*, vol. 132(1), pp. 281-303.
- [29] Kang, H. (1986), “Unstable Weights in the Combination of Forecasts,” *Management Science* 32, 683-95.
- [30] Laster, David, Paul Bennett and In Sun Geoum, 1999, “Rational Bias In Macroeconomic Forecasts,” *The Quarterly Journal of Economics*, vol. 114, issue 1, pp. 293-318
- [31] Palm, Franz C. and Arnold Zellner, 1992, “To combine or not to combine? issues of combining forecasts,” *Journal of Forecasting*, Volume 11, Issue 8 , pp. 687-701.
- [32] Patton, Andrew J. and Allan Timmermann, 2006, “Testing Forecast Optimality under Unknown Loss,” forthcoming in the *Journal of the American Statistical Association*.
- [33] Pesaran, M.H., (2005), “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure.” Working Paper: Cambridge University, forthcoming in *Econometrica*.
- [34] Phillips, P.C.B. and H.R. Moon, 1999, “Linear Regression Limit Theory for Nonstationary Panel Data,” *Econometrica*, vol. 67 (5), pp. 1057–1111.
- [35] Stock, J. and Watson, M., “Forecasting Inflation”, *Journal of Monetary Economics*, 1999, Vol. 44, no. 2.

- [36] Stock, J. and Watson, M., “Macroeconomic Forecasting Using Diffusion Indexes”, *Journal of Business and Economic Statistics*, April 2002a, Vol. 20 No. 2, 147-162.
- [37] Stock, J. and Watson, M., “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association*, 2002b.
- [38] Stock, J. and Watson, M., 2006, “Forecasting with Many Predictors,” In: Elliott, G., C.W.J. Granger, and A. Timmermann, 2006, Editors, *Handbook of Economic Forecasting*, Amsterdam: North-Holland, Chapter 10, pp. 515-554.
- [39] Timmermann, A., 2006, “Forecast Combinations,” in Elliott, G., C.W.J. Granger, and A. Timmermann, 2006, Editors, *Handbook of Economic Forecasting*, Amsterdam: North-Holland, Chapter 4, pp. 135-196.
- [40] Vahid, F. and Engle, R. F., 1993, “Common trends and common cycles,” *Journal of Applied Econometrics*, vol. 8, 341–360.
- [41] Vahid, F., Engle, R. F., 1997, “Codependent cycles,” *Journal of Econometrics*, vol. 80, 199–221.
- [42] Vahid, F., Issler, J. V., 2002, “The importance of common cyclical features in VAR analysis: A Monte Carlo study,” *Journal of Econometrics*, 109, 341–363.
- [43] Wallace, H. D., Hussain, A., 1969, “The use of error components model in combining cross-section and time-series data,” *Econometrica*, 37, 55–72.
- [44] West, K., 1996, “Asymptotic Inference about Predictive Ability,” *Econometrica*, 64, (5), pp. 1067-84.

A Appendix

A.1 Proofs of Propositions in Section 2

Proof of Proposition 1: Start with:

$$f_{i,t}^h - y_t = k_i + \eta_t + \varepsilon_{i,t}.$$

Then, the MSE of individual forecasts is:

$$\begin{aligned}
MSE_i &= \mathbb{E} \left(f_{i,t}^h - y_t \right)^2 = \mathbb{E} (k_i + \eta_t + \varepsilon_{i,t})^2 \\
&= \mathbb{E} (k_i^2) + \mathbb{E} (\eta_t^2) + \mathbb{E} (\varepsilon_{i,t}^2) \\
&= k_i^2 + \sigma_\eta^2 + \sigma_{\varepsilon_i}^2,
\end{aligned} \tag{17}$$

where $\sigma_{\varepsilon_i}^2$ is the variance of $\varepsilon_{i,t}$. Assumption 1 is used in the second line of (17). We also use the fact that k_i is a constant in the time-series dimension in the last line of (17). ■

Proof of Proposition 2: Start with the cross-sectional average of (4):

$$\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - y_t = \frac{1}{N} \sum_{i=1}^N k_i + \eta_t + \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}.$$

Computing the probability limit of the right-hand side above gives,

$$\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N k_i + \eta_t + \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}. \tag{18}$$

We will compute the probability limits in (18) separately. The first one is a straightforward application of the law of large numbers:

$$\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N k_i = B.$$

The second will turn out to be zero. Our strategy is to show that, in the limit, the variance of $\frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}$ is zero, a sufficient condition for a weak law-of-large-numbers to hold for $\{\varepsilon_{i,t}\}_{i=1}^N$.

Because $\varepsilon_{i,t}$ is weakly stationary and mean-zero, for every i , there exists a scalar Wold representation of the form:

$$\varepsilon_{i,t} = \sum_{j=0}^{\infty} b_{i,j} \xi_{i,t-j} \tag{19}$$

where, for all i , $b_{i,0} = 1$, $\sum_{j=0}^{\infty} b_{i,j}^2 < \infty$, and $\xi_{i,t}$ is white noise.

In computing the variance of $\frac{1}{N} \sum_{i=1}^N \sum_{j=0}^{\infty} b_{i,j} \xi_{i,t-j}$ we use the fact that there is no cross correlation between $\xi_{i,t}$ and $\xi_{i,t-k}$, $k = 1, 2, \dots$. Therefore,

we need only to consider the sum of the variances of terms of the form $\frac{1}{N} \sum_{i=1}^N b_{i,k} \xi_{i,t-k}$. These variances are given by:

$$\text{VAR} \left(\frac{1}{N} \sum_{i=1}^N b_{i,k} \xi_{i,t-k} \right) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N b_{i,k} b_{j,k} \mathbb{E} (\xi_{i,t} \xi_{j,t}), \quad (20)$$

due to weak stationarity of ε_t . We now examine the limit of the generic term in (20) with detail:

$$\begin{aligned} \text{VAR} \left(\frac{1}{N} \sum_{i=1}^N b_{i,k} \xi_{i,t-k} \right) &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N b_{i,k} b_{j,k} \mathbb{E} (\xi_{i,t} \xi_{j,t}) \leq \\ \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |b_{i,k} b_{j,k} \mathbb{E} (\xi_{i,t} \xi_{j,t})| &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |b_{i,k} b_{j,k}| |\mathbb{E} (\xi_{i,t} \xi_{j,t})| \leq \end{aligned} \quad (21)$$

$$\left(\max_{i,j} |b_{i,k} b_{j,k}| \right) \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E} (\xi_{i,t} \xi_{j,t})|. \quad (22)$$

Hence:

$$\begin{aligned} \lim_{N \rightarrow \infty} \text{VAR} \left(\frac{1}{N} \sum_{i=1}^N b_{i,k} \xi_{i,t-k} \right) &\leq \lim_{N \rightarrow \infty} \left(\max_{i,j} |b_{i,k} b_{j,k}| \right) \times \\ \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E} (\xi_{i,t} \xi_{j,t})| &= 0, \end{aligned}$$

since the sequence $\{b_{i,j}\}_{j=0}^{\infty}$ is square-summable, yielding $\lim_{N \rightarrow \infty} \left(\max_{i,j} |b_{i,k} b_{j,k}| \right) < \infty$, and Assumption 4 imposes $\lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N |\mathbb{E} (\xi_{i,t} \xi_{j,t})| = 0$.

Thus, all variances are zero in the limit, as well as their sum, which gives:

$$\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t} = 0.$$

Therefore,

$$\begin{aligned} \mathbb{E} \left(\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - y_t \right)^2 &= \mathbb{E} (B + \eta_t)^2 \\ &= B^2 + \sigma_{\eta}^2. \end{aligned} \quad (23)$$

■

Proof of Proposition 3: From the proof of Proposition 2, we have:

$$\begin{aligned} \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - y_t - \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N k_i &= \eta_t + \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t} \\ &= \eta_t, \end{aligned}$$

leading to:

$$\mathbb{E} \left[\text{plim}_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i \right) - y_t \right]^2 = \sigma_\eta^2.$$

■

Proof of Proposition 5: Although y_t, η_t and $\varepsilon_{i,t}$ are ergodic for the mean, $f_{i,t}^h$ is non ergodic because of k_i . Recall that, $T_1, T_2, R \rightarrow \infty$, as $T \rightarrow \infty$. Then, as $T \rightarrow \infty$,

$$\begin{aligned} \frac{1}{R} \sum_{t=T_1+1}^{T_2} f_{i,t}^h &= \frac{1}{R} \sum_{t=T_1+1}^{T_2} y_t + \frac{1}{R} \sum_{t=T_1+1}^{T_2} \varepsilon_{i,t} + \frac{1}{R} \sum_{t=T_1+1}^{T_2} \eta_t + k_i \\ &\xrightarrow{p} \mathbb{E}(y_t) + k_i + \mathbb{E}(\varepsilon_{i,t}) + \mathbb{E}(\eta_t) \\ &= \mathbb{E}(y_t) + k_i \end{aligned}$$

Given that we observe $f_{i,t}^h$ and y_t , we propose the following consistent estimator for k_i , as $T \rightarrow \infty$:

$$\begin{aligned} \widehat{k}_i &= \frac{1}{R} \sum_{t=T_1+1}^{T_2} f_{i,t}^h - \frac{1}{R} \sum_{t=T_1+1}^{T_2} y_t, \quad i = 1, \dots, N \\ &= \frac{1}{R} \sum_{t=T_1+1}^{T_2} (y_t + k_i + \eta_t + \varepsilon_{i,t}) - \frac{1}{R} \sum_{t=T_1+1}^{T_2} y_t \\ &= k_i + \frac{1}{R} \sum_{t=T_1+1}^{T_2} \varepsilon_{i,t} + \frac{1}{R} \sum_{t=T_1+1}^{T_2} \eta_t \\ &\text{or,} \\ \widehat{k}_i - k_i &= \frac{1}{R} \sum_{t=T_1+1}^{T_2} \varepsilon_{i,t} + \frac{1}{R} \sum_{t=T_1+1}^{T_2} \eta_t. \end{aligned}$$

Using this last result, we can now propose a consistent estimator for B :

$$\widehat{B} = \frac{1}{N} \sum_{i=1}^N \widehat{k}_i = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{R} \sum_{t=T_1+1}^{T_2} f_{i,t}^h - \frac{1}{R} \sum_{t=T_1+1}^{T_2} y_t \right].$$

First let $T \rightarrow \infty$,

$$\begin{aligned} \widehat{k}_i &\xrightarrow{p} k_i, \text{ and,} \\ \frac{1}{N} \sum_{i=1}^N \widehat{k}_i &\xrightarrow{p} \frac{1}{N} \sum_{i=1}^N k_i. \end{aligned}$$

Now, as $N \rightarrow \infty$, after $T \rightarrow \infty$,

$$\frac{1}{N} \sum_{i=1}^N k_i \xrightarrow{p} B,$$

Hence, as $(T, N \rightarrow \infty)_{\text{seq}}$,

$$\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} (\widehat{B} - B) = 0.$$

We can now propose a consistent estimator for η_t :

$$\widehat{\eta}_t = \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} - y_t = \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N \widehat{k}_i - y_t.$$

We let $T \rightarrow \infty$ to obtain:

$$\begin{aligned} \text{plim}_{T \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N \widehat{k}_i - y_t \right) &= \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i - y_t \\ &= \eta_t + \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}. \end{aligned}$$

Letting now $N \rightarrow \infty$, we obtain $\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t} = 0$ and:

$$\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} (\widehat{\eta}_t - \eta_t) = 0.$$

Finally,

$$\widehat{\varepsilon}_{i,t} = f_{i,t}^h - y_t - \widehat{k}_i - \widehat{\eta}_t, \text{ and } f_{i,t}^h - y_t = k_i + \eta_t + \varepsilon_{i,t}.$$

Hence :

$$\widehat{\varepsilon}_{i,t} - \varepsilon_{i,t} = (k_i - \widehat{k}_i) + (\eta_t - \widehat{\eta}_t).$$

Using the previous results that $\text{plim}_{T \rightarrow \infty} (\widehat{k}_i - k_i) = 0$ and $\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} (\widehat{\eta}_t - \eta_t) = 0$, we obtain:

$$\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} (\widehat{\varepsilon}_{i,t} - \varepsilon_{i,t}) = 0.$$

■

Proof of Proposition 6: We let $T \rightarrow \infty$ first to obtain:

$$\begin{aligned} \text{plim}_{T \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} \right) &= \text{plim}_{T \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N \widehat{k}_i \right) \\ &= \frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \frac{1}{N} \sum_{i=1}^N k_i = y_t + \eta_t + \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t}. \end{aligned}$$

Letting now $N \rightarrow \infty$ we obtain $\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \varepsilon_{i,t} = 0$ and:

$$\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} \right) = y_t + \eta_t = \mathbb{E}_{t-h}(y_t),$$

from (2) and (3), which is the optimal forecast. The MSE of the feasible bias-corrected average forecast is:

$$\mathbb{E} \left[\text{plim}_{(T, N \rightarrow \infty)_{\text{seq}}} \left(\frac{1}{N} \sum_{i=1}^N f_{i,t}^h - \widehat{B} \right) - y_t \right]^2 = \sigma_\eta^2.$$

showing that we are back to the result in Proposition 3. ■

Proof of Proposition 7: Under $H_0 : B = 0$, we have shown in Proposition 5 that \widehat{B} is a $(T, N \rightarrow \infty)_{\text{seq}}$ consistent estimator for B . To compute the consistent estimator of the asymptotic variance of \overline{B} we follow Conley(1999), who matches spatial dependence to a metric of economic distance. Denote by $\text{MSE}_i(\cdot)$ and $\text{MSE}_j(\cdot)$ the MSE in forecasting of forecasts i and j respectively. For any two generic forecasts i and j , we use $\text{MSE}_i(\cdot) - \text{MSE}_j(\cdot)$ as a measure of distance between these two forecasts. For N forecasts, we can choose one of them to be the benchmark, say, the first one, computing $\text{MSE}_i(\cdot) - \text{MSE}_1(\cdot)$ for $i = 2, 3, \dots, N$. With this measure of spatial dependence at hand, we can construct a two-dimensional estimator of the asymptotic variance of \overline{B} and \widehat{B} following Conley(1999, Sections 3 and 4). We label \overline{V} and \widehat{V} the estimates of the asymptotic variances of \overline{B} and of \widehat{B} , respectively.

Once we have estimated the asymptotic covariance of \overline{B} , we can test the null hypothesis $H_0 : B = 0$, by using the following t-ratio statistic:

$$t = \frac{\overline{B}}{\sqrt{\overline{V}}}.$$

By the central limit theorem, $t \xrightarrow[N \rightarrow \infty]{d} N(0, 1)$ under $H_0 : B = 0$. Now consider $\hat{t} = \frac{\hat{B}}{\sqrt{\hat{V}}}$, where \hat{V} is computed using $\hat{k} = (\hat{k}_1, \hat{k}_2, \dots, \hat{k}_N)'$ in place of $k = (k_1, k_2, \dots, k_N)'$. We have proved that $\hat{k}_i \xrightarrow{p} k_i$ as $T \rightarrow \infty$, then the test statistics t and \hat{t} are asymptotically equivalent and therefore

$$\hat{t} = \frac{\hat{B}}{\sqrt{\hat{V}}} \xrightarrow[(T, N \rightarrow \infty)_{\text{seq}}]{d} \mathcal{N}(0, 1).$$

■

A.2 Tables and Figures

Table 1: Monte-Carlo Results

$R = 50$ $a = 0$; $b = 0.5$

	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 10$						
skewness	-0.009	0.003	0.004	0.404	0.438	0.632
kurtosis	3.043	3.037	3.038	3.221	3.298	3.823
mean	0.000	0.391	-0.001	1.561	1.697	1.916
τ -th= 0.01 quantile	-0.590	-0.091	-0.643	0.911	0.989	1.080
τ -th= 0.25 quantile	-0.169	0.252	-0.186	1.337	1.448	1.608
τ -th= 0.50 quantile	-0.001	0.391	-0.001	1.540	1.672	1.872
τ -th= 0.75 quantile	0.167	0.530	0.184	1.763	1.918	2.179
τ -th= 0.99 quantile	0.578	0.879	0.641	2.394	2.642	3.138
$N = 20$						
skewness	0.010	-0.013	0.011	0.442	0.444	0.961
kurtosis	3.115	3.084	3.138	3.321	3.321	5.011
mean	0.000	0.440	-0.002	1.286	1.466	2.128
τ -th= 0.01 quantile	-0.532	-0.001	-0.690	0.754	0.853	1.117
τ -th= 0.25 quantile	-0.153	0.316	-0.195	1.098	1.247	1.723
τ -th= 0.50 quantile	-0.001	0.440	-0.004	1.266	1.444	2.053
τ -th= 0.75 quantile	0.151	0.565	0.192	1.452	1.659	2.448
τ -th= 0.99 quantile	0.535	0.876	0.687	1.987	2.275	3.851
$N = 40$						
skewness	-0.015	-0.006	-0.016	0.438	0.448	2.852
kurtosis	3.147	3.090	3.600	3.324	3.338	22.203
mean	0.000	0.465	0.000	1.147	1.351	6.094
τ -th= 0.01 quantile	-0.515	0.050	-1.209	0.673	0.786	2.165
τ -th= 0.25 quantile	-0.145	0.346	-0.315	0.980	1.150	4.021
τ -th= 0.50 quantile	0.000	0.465	0.002	1.130	1.331	5.337
τ -th= 0.75 quantile	0.141	0.583	0.312	1.295	1.529	7.243
τ -th= 0.99 quantile	0.509	0.876	1.209	1.771	2.100	17.669

Table 2: Monte-Carlo Results

$R = 50, a = -0.5; b = 0.5$

	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 10$						
skewness	-0.009	0.005	0.004	0.404	0.395	0.632
kurtosis	3.043	3.016	3.038	3.221	3.228	3.823
mean	0.000	0.000	-0.001	1.561	1.547	1.916
τ -th= 0.01 quantile	-0.590	-0.511	-0.643	0.911	0.905	1.080
τ -th= 0.25 quantile	-0.169	-0.149	-0.186	1.337	1.326	1.608
τ -th= 0.50 quantile	-0.001	0.000	-0.001	1.540	1.526	1.872
τ -th= 0.75 quantile	0.167	0.147	0.184	1.763	1.745	2.179
τ -th= 0.99 quantile	0.578	0.516	0.641	2.394	2.369	3.138
$N = 20$						
skewness	0.010	-0.015	0.011	0.442	0.414	0.961
kurtosis	3.115	3.071	3.138	3.321	3.283	5.011
mean	0.000	0.000	-0.002	1.286	1.272	2.128
τ -th= 0.01 quantile	-0.532	-0.462	-0.690	0.754	0.746	1.117
τ -th= 0.25 quantile	-0.153	-0.130	-0.195	1.098	1.089	1.723
τ -th= 0.50 quantile	-0.001	0.000	-0.004	1.266	1.254	2.053
τ -th= 0.75 quantile	0.151	0.130	0.192	1.452	1.435	2.448
τ -th= 0.99 quantile	0.535	0.456	0.687	1.987	1.951	3.851
$N = 40$						
skewness	-0.015	-0.005	-0.016	0.438	0.414	2.852
kurtosis	3.147	3.090	3.600	3.324	3.266	22.203
mean	-0.002	0.000	0.000	1.147	1.133	6.094
τ -th= 0.01 quantile	-0.515	-0.426	-1.209	0.673	0.667	2.165
τ -th= 0.25 quantile	-0.145	-0.123	-0.315	0.980	0.971	4.021
τ -th= 0.50 quantile	-0.002	0.000	0.002	1.130	1.116	5.337
τ -th= 0.75 quantile	0.141	0.121	0.312	1.295	1.278	7.243
τ -th= 0.99 quantile	0.509	0.424	1.209	1.771	1.733	17.669

Table 3: Monte-Carlo Results

$N = 10, R = 500, 1,000, a = 0; b = 0.5$

	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 10, R = 500$						
skewness	0.001	-0.015	-0.002	0.408	0.452	0.408
kurtosis	3.058	3.065	3.066	3.308	3.390	3.314
mean	0.000	0.391	0.000	1.532	1.697	1.559
τ -th= 0.01 quantile	-0.434	-0.091	-0.437	0.893	0.982	0.913
τ -th= 0.25 quantile	-0.124	0.254	-0.126	1.312	1.448	1.334
τ -th= 0.50 quantile	0.000	0.391	0.000	1.512	1.672	1.539
τ -th= 0.75 quantile	0.123	0.529	0.124	1.729	1.918	1.760
τ -th= 0.99 quantile	0.433	0.870	0.436	2.359	2.645	2.394
	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 10, R = 1,000$						
skewness	0.009	0.001	0.011	0.414	0.453	0.413
kurtosis	3.064	3.051	3.061	3.300	3.374	3.295
mean	0.000	0.392	0.000	1.535	1.695	1.541
τ -th= 0.01 quantile	-0.424	-0.088	-0.424	0.902	0.990	0.909
τ -th= 0.25 quantile	-0.122	0.253	-0.122	1.310	1.447	1.321
τ -th= 0.50 quantile	0.000	0.392	0.000	1.507	1.670	1.520
τ -th= 0.75 quantile	0.122	0.531	0.122	1.724	1.916	1.739
τ -th= 0.99 quantile	0.428	0.876	0.430	2.343	2.635	2.360

Table 4: Monte-Carlo Results

$N = 40, R = 2,000, 4,000, a = 0; b = 0.5$

	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 40, R = 2,000$						
skewness	-0.024	0.006	-0.022	0.407	0.452	0.408
kurtosis	3.151	3.139	3.143	3.252	3.330	3.252
mean	0.000	0.466	0.000	1.127	1.355	1.149
τ -th= 0.01 quantile	-0.370	0.049	-0.374	0.668	0.786	0.680
τ -th= 0.25 quantile	-0.103	0.348	-0.103	0.965	1.153	0.983
τ -th= 0.50 quantile	0.000	0.464	0.000	1.121	1.334	1.133
τ -th= 0.75 quantile	0.103	0.582	0.105	1.271	1.533	1.296
τ -th= 0.99 quantile	0.363	0.879	0.366	1.728	2.109	1.760
	Bias Distribution			MSE Distribution		
	BCAF	Average	Weighted	BCAF	Average	Weighted
$N = 40, R = 4,000$						
skewness	0.003	0.006	0.002	0.409	0.458	0.415
kurtosis	3.132	3.106	3.125	3.255	3.325	3.256
mean	0.000	0.465	0.000	1.127	1.355	1.137
τ -th= 0.01 quantile	-0.361	0.053	-0.364	0.666	0.790	0.673
τ -th= 0.25 quantile	-0.103	0.348	-0.103	0.966	1.153	0.975
τ -th= 0.50 quantile	0.000	0.464	0.000	1.111	1.333	1.122
τ -th= 0.75 quantile	0.102	0.581	0.103	1.271	1.533	1.283
τ -th= 0.99 quantile	0.363	0.878	0.364	1.725	2.112	1.739

Table 5: The Brazilian Central Bank Focus Survey
Computing Average Bias and Testing the No-Bias Hypothesis

Horizon (h)	Avg. Bias \hat{B}	$H_0 : B = 0$ p-value
6	0.06187	0.063

Notes: (1) $N = 18$, $R = 26$, $P = 15$, and $h = 6$ months ahead.

Table 6: The Brazilian Central Bank Focus Survey
Comparing the MSE of the Simple Average Forecast with that of
the Bias-Corrected Average Forecast and the Weighted Average
Forecast

Forecast Horizon (h)	(a) MSE BCAF	(b) MSE Average Forecast	(c) MSE Weighted Avg. Forecast	(b)/(a)	(c)/(a)
6	0.0683	0.0808	0.2665	1.182	3.902

Notes: (1) $N = 18$, $R = 23$ and $P = 18$, and $h = 6$ months ahead.