FORECASTING USING CROSS-SECTION AVERAGE-AUGMENTED TIME SERIES REGRESSIONS *

Hande Karabiyik
Vrije Universiteit Amsterdam
and
Tinbergen Institute

Joakim Westerlund†
Lund University
and
Deakin University

December 16, 2019

Abstract

There is a large and growing literature concerned with forecasting time series variables using factor-augmented regression models. The workhorse of this literature is a two-step approach in which the factors are first estimated by applying the principal components method to a large panel of variables, and the forecast regression is estimated conditional on the first-step factor estimates. Another stream of research that has attracted much attention is concerned with the use of cross-section averages as common factor estimates in interactive effects panel regression models. The main justification for this second development is the simplicity and good performance of the cross-section averages when compared to estimated principal component factors. In view of this, it is quite surprising that no one has yet considered the use of cross-section averages for forecasting. Indeed, given the purpose to forecast the conditional mean, the use of the cross-sectional average to estimate the factors is only natural. The present paper can be seen as a reaction to this. The purpose is to investigate the asymptotic and small-sample properties of forecasts based on cross-section average-augmented regressions. In contrast to existing studies, the investigation is carried out while allowing the number of factors to be unknown.

JEL Classification: C12; C13; C33.

Keywords: Forecasting; Factor-augmented regressions; Cross-section average.

*Previous versions of this paper were presented at a seminar in Nottingham, at the 2017 RMSE in Rotterdam, at the 2019 NEM in Stockholm and at the 2019 IPDC in Vilnius. The authors would like to thank seminar and conference participants, and in particular Maurice Bun, Markus Eberhardt, David Harvey, Steve Leybourne, Johan Lyhagen, Ovidijus Stauskas and Lorenzo Trapani for many valuable comments and suggestions. Westerlund thanks the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Corresponding author: Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

1 Introduction

Consider the scalar variable y_t, observable for $t = 1, ..., T$ time periods. The data generating process (DGP) of this variable is the same as in the previous literature on forecasting using factor-augmented regressions (see, for example, Bai and Ng, 2006; Stock and Watson, 2002a, 2002b), and is given by

$$y_{t+h} = \alpha' F_t + \beta' W_t + \varepsilon_{t+h} = \delta' z_t + \varepsilon_{t+h}, \tag{1}$$

where $z_t = [F_t', W_t']'$, $\delta = [\alpha', \beta']'$, $h \geq 0$ is the forecast horizon, F_t is a $r \times 1$ vector of unobserved common factors, or "diffusion indices", W_t is a $n \times 1$ vector of known variables, including deterministic terms and lags of y_t, and ε_t is an error term. While we do not get to observe F_t, we assume that there is another variable available that is informative regarding F_t. Specifically, we assume the existence of an "external" $m \times 1$ panel data variable, $x_{i,t}$, which is observable not only across time but also across $i = 1, ..., N$ cross-sectional units. Again, similarly to the previous literature, the DGP of this variable is assumed to have the following common factor representation:

$$x_{i,t} = \lambda_i' F_t + e_{i,t}, \tag{2}$$

where λ_i is a $r \times m$ matrix of factor loadings and $e_{i,t}$ is a $m \times 1$ vector of errors that are "largely idiosyncratic".[1]

As is well known, if F_t and δ were known, and $E(\varepsilon_{t+h} | z_t, z_{t-1}, ...) = 0$, the mean square optimal forecast of y_t is given by

$$y_{T+h|T} = E(y_{T+h} | z_T, z_{T-1}, ...) = \delta' z_T. \tag{3}$$

Of course, F_t and δ are not known, and we therefore use $\widehat{y}_{T+h|T} = \widehat{\delta}' \widehat{z}_T$ in place of $y_{T+h|T}$. Here $\widehat{z}_t = [W_t', \widehat{F}_t']'$, where $\widehat{F}_t$ is an estimator of the space spanned by F_t, and $\widehat{\delta} = [\widehat{\beta}', \widehat{\alpha}']'$ is the least squares (LS) slope estimator in a regression of y_{t+h} onto $\widehat{z}_t$. An important question here is: How to construct $\widehat{F}_t$? The previous literature has focused almost exclusively on the case when $\widehat{F}_t$ is obtained using the principal components (PC) method (see Bai and Ng, 2006; Corradi and Swanson, 2014; Djogbenou et al., 2015, 2017; Stock and Watson, 2002a, 2002b, to mention a few).[2] The idea here is to chose $\widehat{F}_t$ so as to minimize the variance of the resulting idiosyncratic

[1] The panel variables in $x_{i,t}$ could be allowed to depend on W_t without affecting the results of the paper.

[2] Bai and Ng (2009) provide a review of some alternative computationally more demanding approaches, and give additional references.

errors, a problem that can be solved by performing an eigenvalue decomposition of the sample covariance matrix of $x_{i,t}$. The estimated PC factors can then be seen as weighted cross-section averages of $x_{i,t}$ with weights set equal to the eigenvectors of the sample covariance matrix of $x_{i,t}$.

The results obtained by using the PC method have been very encouraging (see, for example, Eickmeier and Ziegler, 2008). One reason for this is the "averaging out effect" that occurs when pooling from across the cross-sectional dimension of $x_{i,t}$, and that works by effectively smoothing any structural instabilities that might exist (see, for example, De Mol et al., 2008; Banerjee et al., 2008; Stock and Watson, 2009). A drawback is that PC is by construction sensitive to both the level and variation of the variance of $e_{i,t}$ (see, for example, Breitung and Tenhofen, 2011; Choi, 2012; Boivin and Ng, 2006). Hence, while the averaging makes it robust to certain features, PC is sensitive to the weights. This begs the question: Why not consider the equal weighted cross-section average (CA) of $x_{i,t}$ as an estimator of the space spanned by F_t? The fact that this question has not yet received an answer is particularly surprising given the good performance of the equal weighted average in combining forecasts from different models. In fact, the performance of the equal weighted forecast combination has been so good that it has given rise to what has become known as the "forecast combination puzzle" (see, for example, Stock and Watson, 2004; Timmermann, 2006). As a partial explanation, Smith and Wallis (2009) point out that, in analogy to the usual comparison of the LS and weighted LS estimators, relatively sophisticated combinations based on estimated weights suffer from an additional source of small-sample error that is not there when setting the weights equal, and that this might well account the difference in performance. Similar results have been documented in the literature on interactive effects panel data regression models, in which CA estimation of the common factors tends to lead to better performance than when said estimation is carried out using PC (see, for example, Chudik et al., 2011; Westerlund and Urbain, 2015).

The current paper is motivated by the discussion in the last paragraph. The purpose is to study the asymptotic and finite-sample properties of $\widehat{y}_{T+h|T}$ when F_t is estimated using the CAs of $x_{i,t}$. The use of these CAs in the current context not only simplifies considerably the implementation of the forecasting exercise, but is in fact quite natural in the sense that it uses the (sample) mean of $x_{i,t}$ to estimate the (conditional) mean of y_{T+h}. An important issue is, as it turns out, how many factors there are. The existing theory based on PC estimation assume that r is known (see, for example, Bai and Ng, 2006; Djogbenou et al., 2015, 2017; Gonçalves and Perron, 2014; Stock and Watson, 2002a), which is not realistic. One may of course argue, as is indeed commonly done, that r can be consistently estimated and therefore that the known r assumption is without loss of generality. As the bulk of the Monte Carlo evidence shows, however, r is a difficult object to estimate, which is reflected also in the empirical literature (see, for example, Breitung and Eickmeier, 2011; Breitung and Pigorsch, 2013; Moon and Weidner, 2015; Stock and Watson, 2005, 2009). Typically, one ends up with too many factors, and as a result many researches have chosen to work with a fixed number (see, for example, Cheng and Hansen, 2015; De Mol et al., 2008; Moon and Weidner, 2015; Stock and Watson, 2002a, 2002b, 2009). It is therefore important to consider the case when the number of estimated factors is larger than r, which closely related to the "oversampling problem" discussed by Boivin and Ng (2006). One novelty of the present paper is therefore to relax the otherwise so common known r assumption.

Clearly, $\widehat{y}_{T+h|T}$ depends not only on $\widehat{\delta}$ but also on $\widehat{F}_t$ through $\widehat{z}_t$. Thus, to study the behavior of $\widehat{y}_{T+h|T}$, we must examine the asymptotic properties of both $\widehat{\delta}$ and $\widehat{F}_t$. The number of factors is, as already pointed out, treated as unknown and estimated using the $m \geq r$ CAs of $x_{i,t}$.[3] Karabiyik et al. (2017) considered the pooled LS estimator of a factor-augmented panel data regression with r factors that are estimated using $m \geq r$ CAs. According to their results, while consistent, inference is impaired by the presence of a bias in the asymptotic distribution of the estimator. The exact form of the bias depends on whether $m = r$ or $m > r$; however, this is as far as the dependence on the unknown parameter r goes. Moreover, the bias is decreasing (increasing) in N (T), which means that if $T/N \to 0$ there is no dependence on m and r at all. The results reported in the present paper are quite different, which is partly expected given that the problem considered here is much more challenging than the one in Karabiyik et al. (2017). In particular, unlike in this other paper, here it is not enough to just control for the factors, but their predictive ability is in fact key in this paper. Moreover, while in Karabiyik et al. (2017) the model of interest is a panel data regression, here it is a predictive time series regression. We therefore want to do more with less (data); hence, the challenge.

According to the results, while $\widehat{\beta}$ is consistent and asymptotically normal, unless $m = r$, $\widehat{\alpha}$ is generally inconsistent (for the space spanned by α). Interestingly, in spite of this inconsistency, $\widehat{y}_{T+h|T}$ is still consistent for $y_{T+h|T}$ and asymptotically normal. Hence, the inconsistency of

[3]When $r < m$, the LS estimator will be inconsistent, since then there are unattended factors in (1) that may be correlated with W_t. We therefore restrict our attention to the case $m \geq r$.

$\widehat{\alpha}$ does not interfere with the consistency of $\widehat{y}_{T+h|T}$. It does, however, affect inference, as the asymptotic variance of $\widehat{y}_{T+h|T}$ is inestimable when $m > r$. This means that confidence intervals for $y_{T+h|T}$ will not have correct asymptotic coverage. However, the coverage error goes in the "right" direction in the sense that the asymptotic coverage is at least as large as the nominal coverage. Hence, while the asymptotic coverage is not correct, we can still control the type I error rate. These results illustrate quite clearly the importance of not requiring r to be known.

The balance of the paper is organized as follows. In Section 2, we present and discuss the assumptions, which are used in Section 3 to derive our asymptotic results. Section 4 presents the results of a small-scale Monte Carlo study. Section 5 is concerned with an empirical application aimed at forecasting eight US macroeconomic variables. Section 6 concludes. All proofs are given in the supplemental material.

2 Assumptions

The conditions under which we will be working are summarized in Assumptions A–D. Here and throughout this paper $\text{tr}\, A$, $\text{rk}\, A$, A^+ and $\|A\| = \sqrt{\text{tr}(A'A)}$ denote the trace, the rank, the generalized Moore–Penrose inverse, and the Frobenius (Euclidean) norm, respectively, of the matrix A. For any two matrices A and B, $\text{diag}(A, B)$ denotes the block-diagonal matrix that takes A (B) as the upper left (lower right) block. For any matrix A_i, we use $\overline{A} = N^{-1}\sum_{i=1}^{N} A_i$ to denote its CA. Moreover, $\to_d$ and $\to_p$ signify convergence in distribution and convergence in probability, respectively.

Assumption A. λ_i is a non-random matrix such that $\|\lambda_i\| < \infty$, $\overline{\lambda} \to \lambda$ as $N \to \infty$ and $\text{rk}\,\overline{\lambda} = \text{rk}\,\lambda = r \leq m$.

Assumption B.

1. $E(e_{i,t}) = 0_{m\times 1}$ and $E(\|e_{i,t}\|^8) < \infty$ for all i and t.

2. $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\Sigma_{e,ijts}\| < \infty$, $T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\Sigma_{e,ijts}\| < \infty$ and $\|\Sigma_{e,ijts}\| < \infty$ for all i, j, t and s, where $\Sigma_{e,ijts} = E(e_{i,t}e'_{j,s})$ with $\Sigma_{e,ijtt} = \Sigma_{e,ij}$ for all t.

3. $E(\|N^{-1}T^{-1/2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}(e_{i,t}e'_{i,t} - \Sigma_{e,ij})\|^2) < \infty$ and $E(\|N^{-1/2}\sum_{i=1}^{N}(e_{i,t}e'_{i,s} - \Sigma_{e,iits})\|^2)$ $< \infty$ for all t and s.

4. $N^{-1/2}\sum_{i=1}^{N} e_{i,t} \to_d N(0_{m\times 1}, \Sigma_e)$ as $N \to \infty$, where $\Sigma_e = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\Sigma_{e,ij}$ is positive definite.

Assumption C. z_t and $e_{i,t}$ are mutually independent groups. Dependence within each group is allowed.

Assumption D.

1. $E(\varepsilon_{t+h}|z_t, z_{t-1}, ...) = 0$ for $h > 0$.

2. ε_t is independent of $e_{i,s}$ for all i, t and s.

3. $E(\varepsilon_t^2) = \sigma_{\varepsilon,t}^2 \in (0,\infty)$, $T^{-1}\sum_{t=1}^{T-h}\sigma_{\varepsilon,t+h}^2 \to \sigma_\varepsilon^2 \in (0,\infty)$ as $T \to \infty$, $T^{-1}\sum_{t=1}^{T} z_t z_t'$ is positive definite almost surely (a.s.), $T^{-1}\sum_{t=1}^{T} z_t z_t' \to_p \Sigma_z$ as $T \to \infty$, where Σ_z is positive definite, and $E(\|z_t\|^4) < \infty$.

4. $T^{-1/2}\sum_{t=1}^{T-h}\varepsilon_{t+h}z_t \to_d N(0_{(r+n)\times 1}, \Sigma_{z\varepsilon})$ as $T \to \infty$, where the $(r+n)\times(r+n)$ matrix $\Sigma_{z\varepsilon} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T-h} E(\varepsilon_{t+h}^2 z_t z_t')$ is positive definite.

Assumptions A–D are almost the same as in Bai and Ng (2006), which is the main theoretical point of reference in the PC strand of the literature. One difference is Assumption A, which supposes that λ has rank $r \leq m$. This should be compared to Assumption B of Bai and Ng (2006), which requires that $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i'$ is positive definite. Hence, while in PC each factor has to have a nontrivial contribution to the variance of $x_{i,t}$, in the CA approach considered here it is the contribution to the mean that matters. Another difference when compared to the PC strand of the literature is that here the number of factors that can be permitted is bounded from above by m. Hence, unlike in PC, in CA m is typically larger than one. As we explain in detail in Section 5, however, in applications this extra data requirement is typically not an issue in the sense that the series contained in the panel data set can be divided into categories of variables, and where each category is interpreted quite naturally as a panel data variable. Hence, in this case, m is simply the number of categories. Assumption B allows $e_{i,t}$ to be weakly dependent across both time and cross-section units; however, it cannot be strongly dependent, as in the unit root case. Heteroskedasticity across the cross-section is permitted but not across time. This last restriction is only for ease of exposure, and can be easily relaxed, as we explain in Section 3. Assumption C allows for serially correlated factors and cross-sectionally correlated factor loadings. Assumption D does not rule out the case when y_t is included in z_t. It also allows F_t and W_t to be correlated both with each other and over time. It does, however, require that F_t and W_t are stationary.

3 Asymptotic results

We want to use $\widehat{y}_{T+h|T} = \widehat{\delta}'\widehat{z}_T$ to infer $y_{T+h|T} = \delta' z_T$. As already mentioned, the previous research has focused on the case in which r is known. One of the contributions of the current paper is to relax this assumption. As we will now show, not requiring $m = r$ has important consequences. We begin by noting how

$$\widehat{F}_t = \overline{x}_t = \overline{\lambda}' F_t + \overline{e}_t. \tag{4}$$

The fact that the $r \times m$ matrix $\overline{\lambda}$ is not necessarily square and invertible is important, because it means that the object to be estimated is not necessarily given by F_t. Intuitively, the estimated object must have the same dimension as $\widehat{F}_t$. Therefore, if $m > r$, $\widehat{F}_t$ cannot be estimating F_t. In order to account for this, it is useful to partition $\overline{\lambda}$ as $\overline{\lambda} = [\overline{\lambda}_r, \overline{\lambda}_{-r}]$, where $\overline{\lambda}_r$ is an $r \times r$ matrix of full rank, and $\overline{\lambda}_{-r}$ is $r \times (m-r)$. This partitioning is without loss of generality because $\overline{\lambda}$ has rank r under Assumption A. Let us correspondingly partition $\overline{e}_t = [\overline{e}'_{r,t}, \overline{e}'_{-r,t}]'$, where $\overline{e}_{r,t}$ and $\overline{e}_{-r,t}$ are $r \times 1$ and $(m-r) \times 1$, respectively. If $m = r$, then we define $\overline{\lambda} = \overline{\lambda}_r$ and $\overline{e}_t = \overline{e}_{r,t}$. Let us also define the $m \times m$ matrix $\overline{\Lambda}$, which is such if $m > r$, then

$$\overline{\Lambda} = \begin{bmatrix} \overline{\lambda}_r^{-1} & -\overline{\lambda}_r^{-1}\overline{\lambda}_{-r} \\ 0_{(m-r)\times r} & I_{m-r} \end{bmatrix} = [\overline{\Lambda}_r, \overline{\Lambda}_{-r}], \tag{5}$$

where $\overline{\Lambda}_r = [\overline{\lambda}_r^{-1\prime}, 0'_{(m-r)\times r}]'$ is $m \times r$ and $\overline{\Lambda}_{-r} = [-\overline{\lambda}'_{-r}\overline{\lambda}_r^{-1\prime}, I_{m-r}]'$ is $m \times (m-r)$. If $m = r$, we define $\overline{\Lambda} = \overline{\Lambda}_r = \overline{\lambda}_r^{-1} = \overline{\lambda}^{-1}$. It is useful to think of $\overline{\Lambda}$ as a type of "inverse" of $\overline{\lambda}$, which is such that $\overline{\lambda}\overline{\Lambda} = I_r$ if $m = r$ and $\overline{\lambda}\overline{\Lambda} = [I_r, 0_{r\times(m-r)}]$ if $m > r$.

In order to appreciate the significance of $\overline{\Lambda}$, it is illustrative to first consider the case when $m = r$. The fact that in this case $\overline{\Lambda} = \overline{\lambda}^{-1}$ means that (4) can be rewritten as

$$\overline{\Lambda}'\widehat{F}_t = \overline{\lambda}^{-1\prime}\widehat{F}_t = F_t + \overline{\lambda}^{-1\prime}\overline{e}_t. \tag{6}$$

By Assumption B, $\|\overline{e}_t\| = O_p(N^{-1/2})$ uniformly in t, which implies that $\|\overline{\lambda}^{-1\prime}\widehat{F}_t - F_t\| = O_p(N^{-1/2})$. Hence, while $\widehat{F}_t$ is not consistent for F_t, $\overline{\lambda}^{-1\prime}\widehat{F}_t$ is, which is enough for our purposes, because the rotation by $\overline{\lambda}$ is subsumed in the estimation of α. In order to appreciate this last point, note how $\widehat{\alpha}'\widehat{F}_t = \alpha' F_t + (\overline{\lambda}\widehat{\alpha} - \alpha)' F_t + \widehat{\alpha}'\overline{\lambda}'(\overline{\lambda}^{-1\prime}\widehat{F}_t - F_t)$. In Theorem 1 below we show that $\|\overline{\lambda}\widehat{\alpha} - \alpha\| = o_p(1)$. Hence, since $\|\overline{\lambda}^{-1\prime}\widehat{F}_t - F_t\| = o_p(1)$, we have that

$$\widehat{\alpha}'\widehat{F}_t = \alpha' F_t + o_p(1). \tag{7}$$

In other words, while we cannot estimate α and F_t separately, we can still consistently estimate their product, which is the quantity that matters for $y_{T+h|T}$. The main challenge when $m = r$ is to show that the estimation error in (7) is negligible also in the asymptotic distribution theory of $\widehat{y}_{T+h|T}$.

Let us now consider the case when $m > r$. Analogous to (6),

$$\overline{\Lambda}'\widehat{F}_t = \overline{\Lambda}'\overline{\lambda}' F_t + \overline{\Lambda}'\overline{e}_t = \begin{bmatrix} F_t \\ 0_{(m-r)\times 1} \end{bmatrix} + \begin{bmatrix} \overline{\Lambda}_r'\overline{e}_t \\ \overline{\Lambda}_{-r}'\overline{e}_t \end{bmatrix}. \tag{8}$$

Again, since $\|\overline{e}_t\| = O_p(N^{-1/2})$, we have that $\overline{\Lambda}'\widehat{F}_t$ is no longer consistent for F_t but for $F_t^0 = \overline{\Lambda}'\overline{\lambda}' F_t$, which under $m > r$ is equal to $F_t^0 = [F_t', 0_{(m-r)\times 1}']'$. This makes sense, because now $\widehat{F}_t$ is over-parameterized, and therefore the redundant factor estimates should be estimated to zero. The problem with this result is that F_t^0 is not just any parameter but is in fact an estimated regressor, which means that the associated second order moment matrix is asymptotically singular because of the zeros. This situation is similar to the one that occurs when fitting regressions involving regressors that are of different orders of integration (see, for example, Chang and Phillips, 1995), and the solution is the same. Specifically, in order to account for the limiting singularity, we introduce the $m \times m$ normalization matrix D_N, which is $D_N = I_m$ if $m = r$ and $D_N = \text{diag}(I_r, \sqrt{N} I_{m-r})$ if $m > r$. Since $D_N F_t^0 = F_t^0$, normalization by D_N does not affect the object of interest. Let us further introduce $\widehat{F}_t^0 = D_N \overline{\Lambda}'\widehat{F}_t$ and $\overline{e}_t^0 = D_N \overline{\Lambda}'\overline{e}_t$, which are such that $\widehat{F}_t^0 = \overline{\lambda}^{-1\prime}\widehat{F}_t$ and $\overline{e}_t^0 = \overline{\lambda}^{-1\prime}\overline{e}_t$ if $m = r$. If, on the other hand, $m > r$, then $\overline{e}_t^0 = [\overline{e}_{r,t}^{0\prime}, \overline{e}_{-r,t}^{0\prime}]' = [\overline{e}_t'\overline{\Lambda}_r, \sqrt{N}\overline{e}_t\overline{\Lambda}_{-r}]'$. Note how the last $m - r$ rows of $\overline{e}_t^0$ are scaled by $\sqrt{N}$. This is important because it prevents the last $m - r$ rows of $\widehat{F}_t^0$ from converging to zero, as is obvious from

$$\widehat{F}_t^0 = F_t^0 + \overline{e}_t^0 = \begin{bmatrix} F_t \\ \overline{e}_{-r,t}^0 \end{bmatrix} + o_p(1). \tag{9}$$

The normalization by D_N therefore resolves the asymptotic singularity issue, but at a cost of inducing a dependence on $\overline{e}_{-r,t}^0$, which is non-negligible. This dependence is the main challenge we face when studying the limiting behaviour of $\widehat{y}_{T+h|T}$ when $m > r$. The fact that $\overline{e}_{-r,t}^0$ is non-negligible suggests that over-specification of the number of factors will affect the asymptotic distribution theory.

The way we have defined it, in the asymptotic analysis $\widehat{F}_t^0$ is the relevant estimator to consider regardless of whether $m = r$ or $m > r$. We therefore want to replace $\widehat{F}_t$ with $\widehat{F}_t^0$, and in so doing it is useful to introduce $Q_N = \text{diag}(\overline{\Lambda} D_N, I_n)$, $\widehat{z}_t^0 = Q_N' \widehat{z}_t = [\widehat{F}_t^{0\prime}, W_t']'$ and $\delta^0 = [\alpha'(D_N \overline{\Lambda}' \overline{\lambda}')^+, \beta']'$. Since $\overline{\lambda}\overline{\Lambda} = [I_r, 0_{r\times(m-r)}]$, we have $\overline{\lambda}\overline{\Lambda} D_N = [I_r, 0_{r\times(m-r)}]$, which has full row rank r. This implies $(\overline{\lambda}\overline{\Lambda} D_N)^+ = [I_r, 0_{r\times(m-r)}]'$, and so $\delta^0 = [\alpha', 0_{1\times(m-r)}, \beta']'$. Making use of this notation, (9) and the fact that $(D_N \overline{\Lambda}' \overline{\lambda}')^+ D_N \overline{\Lambda}' \overline{\lambda}' = I_r$,

$$\begin{aligned} y_{T+h|T} &= \alpha' F_T + \beta' W_T \\ &= \alpha'(D_N \overline{\Lambda}' \overline{\lambda}')^+ D_N \overline{\Lambda}' \overline{\lambda}' F_T + \beta' W_T \\ &= \alpha'(D_N \overline{\Lambda}' \overline{\lambda}')^+ \widehat{F}_T^0 + \beta' W_T - \alpha'(D_N \overline{\Lambda}' \overline{\lambda}')^+ (\widehat{F}_T^0 - F_T^0) \\ &= \delta^{0\prime} \widehat{z}_T^0 - \alpha'(D_N \overline{\Lambda}' \overline{\lambda}')^+ (\widehat{F}_T^0 - F_T^0) \\ &= \delta^{0\prime} \widehat{z}_T^0 - \alpha' \overline{e}_{r,T}^0. \end{aligned} \tag{10}$$

Hence, since $\widehat{y}_{T+h|T} = \widehat{\delta}' \widehat{z}_T = \widehat{\delta}' Q_N^{-1\prime} \widehat{z}_T^0$, we can show that

$$\widehat{y}_{T+h|T} - y_{T+h|T} = T^{-1/2} \sqrt{T} (\widehat{\delta} - Q_N \delta^0)' Q_N^{-1\prime} \widehat{z}_T^0 + N^{-1/2} \alpha' \sqrt{N} \overline{e}_{r,T}^0. \tag{11}$$

The asymptotic distribution of $\widehat{y}_{T+h|T} - y_{T+h|T}$ therefore has two sources; $Q_N^{-1} \sqrt{T} (\widehat{\delta} - Q_N \delta^0)$ and $\sqrt{N} \overline{e}_{r,T}^0$, where the latter (former) is due to the estimation of F_t (δ). Lemma 1 is concerned with the first term. It is stated in terms of the $(n+m) \times 1$ vectors B and z_t^0, which are such that $B = [0_{r\times 1}', b', 0_{n\times 1}]'$ and $z_t^0 = [F_t', \overline{e}_{-r,t}^{0\prime}, W_t']'$ if $m > r$, and $B = 0_{(m+n)\times 1}$ and $z_t^0 = z_t$ if $m = r$. Here, $b = (\Lambda_{-r}' \Sigma_e \Lambda_{-r})^{-1} \Lambda_{-r}' \Sigma_e \Lambda_r \alpha$, where Λ_r and Λ_{-r} are from $\Lambda = [\Lambda_r, \Lambda_{-r}] = \lim_{N\to\infty} \overline{\Lambda}$.

Lemma 1. *Suppose that Assumptions A–D hold. Then, as N, $T \to \infty$ with $\sqrt{T}/N \to 0$,*

$$\sqrt{T} Q_N^{-1} (\widehat{\delta} - Q_N \delta^0) + \sqrt{T} N^{-1/2} B = \Sigma_{z^0}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T-h} z_t^0 \varepsilon_{t+h} + o_p(1),$$

where

$$\Sigma_{z^0} = \underset{N,T\to\infty}{\text{plim}} \frac{1}{T} \sum_{t=1}^{T-h} z_t^0 z_t^{0\prime}.$$

For $\widehat{z}_T^0$, we again make use of (9), giving

$$\widehat{z}_t^0 - z_t^0 = \begin{bmatrix} \overline{e}_{r,t}^0 \\ 0_{(n+m-r)\times 1} \end{bmatrix} = O_p(N^{-1/2}). \tag{12}$$

By using this and Lemma 1, (11) becomes

$$
\begin{aligned}
&\delta_{NT}(\widehat{y}_{T+h|T} - y_{T+h|T}) \\
&= \delta_{NT}T^{-1/2}[\sqrt{T}(\widehat{\delta} - Q_N\delta^0)'Q_N^{-1\prime} + \sqrt{T}N^{-1/2}B']\widehat{z}_T^0 + \delta_{NT}N^{-1/2}(\alpha'\sqrt{N}\bar{e}_{r,T}^0 - B'\widehat{z}_T^0) \\
&= \delta_{NT}T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}\varepsilon_{t+h}z_t^{0\prime}\Sigma_{z^0}^{-1}z_T^0 + \delta_{NT}N^{-1/2}(\alpha'\sqrt{N}\bar{e}_{r,T}^0 - B'z_T^0) + o_p(1),
\end{aligned} \tag{13}
$$

where $\delta_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$ accounts for the difference in normalization of the two terms on the right-hand side of (11). Consider the first of these terms. Assumptions B and D ensure that

$$
\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h} z_t^0\varepsilon_{t+h} \to_d N(0_{(m+n)\times 1}, \Sigma_{z^0\varepsilon}) \tag{14}
$$

as $T \to \infty$, where

$$
\Sigma_{z^0\varepsilon} = \lim_{N,T\to\infty}\frac{1}{T}\sum_{t=1}^{T-h} E(\varepsilon_{t+h}^2 z_t^0 z_t^{0\prime}).
$$

Hence, conditional on z_T^0, the first term on the right-hand side of (13) is asymptotically normal;

$$
\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}\varepsilon_{t+h}z_t^{0\prime}\Sigma_{z^0}^{-1}z_T^0 \to_d MN(0, \phi^0), \tag{15}
$$

where $MN(\cdot,\cdot)$ denotes a mixed normal distribution and

$$
\phi = \lim_{T\to\infty} z_T^{0\prime}\Sigma_{z^0}^{-1}\Sigma_{z^0\varepsilon}\Sigma_{z^0}^{-1}z_T^0.
$$

The conditioning on z_T^0 here is important, because while it is straightforward to show that $T^{-1/2}\sum_{t=1}^{T-h}\varepsilon_{t+h}z_t^{0\prime}\Sigma_{z^0}^{-1}z_T^0$ is asymptotically normal even unconditionally, the unconditional variance is given by $E(\phi) = \text{tr}\,[\Sigma_{z^0}^{-1}\Sigma_{z^0\varepsilon}\Sigma_{z^0}^{-1}\lim_{T\to\infty}E(z_T^0 z_T^{0\prime})]$, where $\lim_{T\to\infty}E(z_T^0 z_T^{0\prime})$ is inestimable under our assumptions.

Remark 1. Lemma 1 and (15) imply that

$$
\sqrt{T}Q_N^{-1}(\widehat{\delta} - Q_N\delta^0) + \sqrt{T}N^{-1/2}B \to_d MN(0, \phi) \tag{16}
$$

as $T \to \infty$. This result is worthy of some discussion. Suppose first that $m > r$. Let $\widehat{\alpha} = [\widehat{\alpha}_r', \widehat{\alpha}_{-r}']'$, where $\widehat{\alpha}_r$ and $\widehat{\alpha}_{-r}$ are $r\times 1$ and $(m-r)\times 1$, respectively. In this notation,

$$
Q_N^{-1}\sqrt{T}(\widehat{\delta} - Q_N\delta^0) + \sqrt{T}N^{-1/2}B = \sqrt{T}\begin{bmatrix} \overline{\lambda}\widehat{\alpha} - \alpha \\ N^{-1/2}(\widehat{\alpha}_{-r} + b) \\ \widehat{\beta} - \beta \end{bmatrix}, \tag{17}
$$

which according to (16) is asymptotically mixed normal. We can therefore show that $\overline{\lambda}\widehat{\alpha}$ and $\widehat{\beta}$ are $\sqrt{T}$-consistent for α and β, respectively, and asymptotically normal. The most striking observations are, however, related to $\widehat{\alpha}_{-r}$, the estimated coefficient of the redundant factor estimates. The first thing to note is that under the assumptions of Lemma 1, $\widehat{\alpha}_{-r}$ is not necessarily convergent, and that it is only when $\sqrt{T}N^{-1/2} \to \infty$ that $(\widehat{\alpha}_{-r} + b) \to_p 0_{(m-r)\times 1}$.[4] But even if $\widehat{\alpha}_{-r}$ is in fact convergent, the rate at which this happens is slower than the usual $\sqrt{T}$ rate. In fact, under the Lemma 1 requirement that $\sqrt{T}/N \to 0$, we have $(\sqrt{T}N^{-1/2})T^{-1/4} \to 0$, which means that the rate of convergence is slower than $T^{1/4}$. Hence, since $\widehat{\delta}$ is dominated by the component that converges slower, the overall convergence rate is given by $\sqrt{T}N^{-1/2}$. The limit of $\widehat{\alpha}_{-r}$ is also interesting. To the extent that $\widehat{\delta}$ can be viewed as an estimator of $Q_N\delta^0$, the true value of $\widehat{\alpha}_{-r}$ is given by the zero vector. The fact that $b = (\Lambda_{-r}'\Sigma_e\Lambda_{-r})^{-1}\Lambda_{-r}'\Sigma_e\Lambda_r\alpha \neq 0_{(m-r)\times 1}$ whenever $\alpha \neq 0_{r\times 1}$ means that $\widehat{\alpha}_{-r}$ is generally inconsistent.[5]

Let us now consider the case when $m = r$, in which $B = 0_{(m+n)\times 1}$, $z_t^0 = z_t$ and

$$Q_N^{-1}\sqrt{T}(\widehat{\delta} - Q_N\delta^0) = \sqrt{T}\begin{bmatrix} \widehat{\beta} - \beta \\ \overline{\lambda}\widehat{\alpha} - \alpha \end{bmatrix}. \tag{18}$$

Except for the rotation of $\widehat{\alpha}$, which is not the same as when using PC, when $m = r$ the result reported in (16) is the same as the one reported in Theorem 1 of Bai and Ng (2006). We also note that the requirement that $\sqrt{T}/N \to 0$ is the same as in this other paper. Hence, up to the rotation of $\widehat{\alpha}$, asymptotic distribution of $Q_N^{-1}\sqrt{T}(\widehat{\delta} - Q_N\delta^0)$ when $m = r$ is the same as for PC.

For the second term on the right-hand side of (13), we make use of the fact that

$$\begin{aligned} I_m - \Sigma_e\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\overline{\Lambda}_{-r}' &= [I_m - \Sigma_e\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\overline{\Lambda}_{-r}']\Sigma_e^{1/2}\Sigma_e^{-1/2} \\ &= \Sigma_e^{1/2}[I_m - \Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\overline{\Lambda}_{-r}'\Sigma_e^{1/2}]\Sigma_e^{-1/2} \\ &= \Sigma_e^{1/2}M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}}\Sigma_e^{-1/2}, \end{aligned} \tag{19}$$

where $M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} = I_m - \Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\overline{\Lambda}_{-r}'\Sigma_e^{1/2}$ and $\Sigma_e = \Sigma_e^{1/2}\Sigma_e^{1/2\prime}$ with $\Sigma_e^{1/2}$ being the lower triangular Choleski factor. As stated, $M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}}$ is only defined for $m > r$. If $m = r$,

[4] By using the results provided in the proof of Theorem 1, we can further show that $\sqrt{T}(\overline{\lambda}\widehat{\alpha} - \alpha)$ and $\sqrt{T}N^{-1/2}(\widehat{\alpha}_{-r} + b)$ are asymptotically independent.

[5] Interestingly enough, in spite of the problems of non-convergence and slow rate of convergence for $\widehat{\alpha}_{-r}$, the linear combination $\overline{\lambda}\widehat{\alpha} = \overline{\lambda}_r\widehat{\alpha}_r + \overline{\lambda}_{-r}\widehat{\alpha}_{-r}$ is $\sqrt{T}$-consistent for α. Of course, since $\overline{\lambda}$ is unknown, in practice this rate of convergence is not attainable. This finding is similar in spirit to the results reported by Chang and Phillips (1995), where the rate of convergence depends on whether or not the non-stationarity characteristics of the regressors are known.

we define $M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} = I_m$. By using this, the definitions of $\bar{e}^0_{r,t}$, $\bar{e}^0_{-r,t}$ and B, and letting $\overline{\Phi} = \Sigma_e^{-1/2\prime} M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} \Sigma_e^{1/2\prime}\overline{\Lambda}_r\alpha$,

$$\begin{aligned}\alpha'\sqrt{N}\bar{e}^0_{r,t} - B'z^0_t &= \alpha'\sqrt{N}\bar{e}^0_{r,t} - \alpha'\overline{\Lambda}_r'\Sigma_e\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\bar{e}^0_{-r,t} \\ &= \alpha'\overline{\Lambda}_r'[I_m - \Sigma_e\overline{\Lambda}_{-r}(\overline{\Lambda}_{-r}'\Sigma_e\overline{\Lambda}_{-r})^{-1}\overline{\Lambda}_{-r}']\sqrt{N}\bar{e}_t \\ &= \alpha'\overline{\Lambda}_r'\Sigma_e^{1/2} M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} \Sigma_e^{-1/2}\sqrt{N}\bar{e}_t \\ &= \overline{\Phi}'\sqrt{N}\bar{e}_t.\end{aligned}$$

The presence of $M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}}$ in $\overline{\Phi}$ means that $\overline{\Phi}'\sqrt{N}\bar{e}_t$ is asymptotically uncorrelated with $\bar{e}^{0\prime}_{-r,t}$, as is clear from

$$\begin{aligned}\lim_{N\to\infty} E(\overline{\Phi}'\sqrt{N}\bar{e}_t\bar{e}^{0\prime}_{-r,t}) &= \lim_{N\to\infty} \alpha'\overline{\Lambda}_r'\Sigma_e^{1/2} M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} \Sigma_e^{-1/2} N E(\bar{e}_t\bar{e}_t')\overline{\Lambda}_{-r} \\ &= \alpha'\overline{\Lambda}_r'\Sigma_e^{1/2} M_{\Sigma_e^{1/2\prime}\overline{\Lambda}_{-r}} \Sigma_e^{-1/2\prime}\overline{\Lambda}_{-r} = 0_{1\times(m-r)}. \end{aligned} \tag{20}$$

Hence, since $\overline{\Phi}'\sqrt{N}\bar{e}_t$ and $\bar{e}^0_{-r,t}$ are also (jointly) asymptotically normal by Assumption B, they must be asymptotically independent. Because z_t is independent of $e_{i,t}$ by the same assumption, it follows that z^0_t must be asymptotically independent of $\sqrt{N}\bar{e}_t$. The asymptotic distribution of $\overline{\Phi}'\sqrt{N}\bar{e}_t$ as $N \to \infty$ is given by

$$\overline{\Phi}'\sqrt{N}\bar{e}_t \to_d N(0_{m\times 1}, \Phi'\Sigma_e\Phi), \tag{21}$$

where $\Phi = \lim_{N\to\infty}\overline{\Phi} = \Sigma_e^{-1/2\prime} M_{\Sigma_e^{1/2\prime}\Lambda_{-r}} \Sigma_e^{1/2\prime}\Lambda_r\alpha$. The fact that z^0_t and $\sqrt{N}\bar{e}_t$ are asymptotically independent is important because it means that the unconditional asymptotic normality in (15) holds also conditionally on z^0_t. The normal in (15) is also independent of the one in (21) because while the former is determined by ε_{t+h}, the latter is determined by $e_{i,t}$, which are independent by Assumption D. Hence, by adding the results,

$$\begin{aligned}\delta_{NT}(\widehat{y}_{T+h|T} - y_{T+h|T}) &= \delta_{NT}T^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}\varepsilon_{t+h}z^{0\prime}_t\Sigma_{z^0}^{-1}z^0_T + \delta_{NT}N^{-1/2}\overline{\Phi}'\sqrt{N}\bar{e}_t + o_p(1) \\ &\to_d MN\left(0, \lim_{N,T\to\infty}(\delta^2_{NT}T^{-1}\phi + \delta^2_{NT}N^{-1}\Phi'\Sigma_e\Phi)\right)\end{aligned} \tag{22}$$

as N, $T \to \infty$. Theorem 1 is a direct consequence of this last result.

Theorem 1. *Under the conditions of Lemma 1,*

$$t(y_{T+h|T}) = \frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{T^{-1}\phi + N^{-1}\Phi'\Sigma_e\Phi}} \to_d N(0,1).$$

The fact that according to Theorem 1 the asymptotic distribution of $t(y_{T+h|T})$ is correctly centered at zero is noteworthy. Indeed, given how $Q_N^{-1}\sqrt{T}(\widehat{\delta} - Q_N\delta^0)$ enters (11) through $T^{-1/2}\sqrt{T}(\widehat{\delta} - Q_N\delta^0)'Q_N^{-1\prime}\widehat{z}_T^0$, one would expect $\sqrt{T}N^{-1/2}B$ to manifest itself as an asymptotic bias whenever $m > r$. However, this is not the case. The reason is that B is zero except for $m - r$ rows in the middle. Hence, since $\bar{e}_{-r,T}^0$ sits in the corresponding rows of z_T^0, the product $B'z_t^0$ is mean zero, even though B and z_T^0 are not. The asymptotic variance of $t(y_{T+h|T})$ comprises two terms; $T^{-1}\phi$, which emanates from the estimation of δ, and $N^{-1}\Phi'\Sigma_e\Phi$, which emanates from the estimation of F_t. Because the two variance terms vanish at different rates, the rate of consistency of $\widehat{y}_{T+h|T}$ is given by δ_{NT}, which is unexpected given the relatively slow convergence rate of $\widehat{\alpha}_{-r}$ discussed in Remark 1.

Theorem 1 holds regardless of whether $m > r$ or $m = r$. The only difference is that if $m = r$, then $z_t^0 = z_t$, which means that ϕ and Φ reduce to $\phi = \lim_{T\to\infty} z_T'\Sigma_z^{-1}\Sigma_{z\varepsilon}\Sigma_z^{-1}z_T$ and $\Phi = \lambda^{-1}\alpha$, respectively. The fact that the variance changes depending on whether $m = r$ or $m > r$ means that the known r assumption is not innocuous. This is true for CA but a similar result is expected to hold also for PC. The theoretical predictions reported in the PC strand of the literature based on knowing r should therefore be interpreted with caution, since, as pointed out in Section 1, in practice the number of factors is likely to be overestimated.

Remark 2. Theorem 1 does not impose any restrictions on N/T. However, the theorem simplifies if either $N/T \to 0$ or $T/N \to 0$. On the one hand, if $N/T \to 0$, then $\delta_{NT} = \sqrt{N}$ and so $t(y_{T+h|T})$ reduces to

$$t(y_{T+h|T}) = \frac{\sqrt{N}(\widehat{y}_{T+h|T} - y_{T+h|T})}{\sqrt{NT^{-1}\phi + \Phi'\Sigma_e\Phi}} = \frac{\sqrt{N}(\widehat{y}_{T+h|T} - y_{T+h|T})}{\sqrt{\Phi'\Sigma_e\Phi}} + o_p(1) \to_d N(0,1) \tag{23}$$

as N, $T \to \infty$ with $\sqrt{T}/N \to 0$ and $N/T \to 0$. If, on the other hand, $T/N \to 0$, then $\delta_{NT} = \sqrt{T}$ and so

$$t(y_{T+h|T}) = \frac{\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T})}{\sqrt{\phi}} + o_p(1) \to_d N(0,1) \tag{24}$$

as N, $T \to \infty$.

Remark 3. It is interesting to compare the above results with those reported by Bai and Ng (2006) for PC in the known r case. Let us therefore assume that $r = m = 1$. In this case, ϕ is asymptotically equivalent to the corresponding PC term given in Theorem 3 of

Bai and Ng (2006). As for $\Phi'\Sigma_e\Phi = \Sigma_e\alpha^2\lambda^{-2}$, the corresponding term in PC is given by $\Sigma_e\alpha^2(\lim_{N\to\infty} N^{-1}\sum_{i=1}^N \lambda_i^2)^{-1}$.[6] Hence, since $\overline{\lambda}^2 \leq N^{-1}\sum_{i=1}^N \lambda_i^2$ by the Cauchy–Schwarz inequality, we have that $\Sigma_e\alpha^2\lambda^{-2} \geq \Sigma_e\alpha^2(\lim_{N\to\infty} N^{-1}\sum_{i=1}^N \lambda_i^2)^{-1}$. The PC forecast is therefore more efficient than the CA forecast, which is partly expected because under homoskedasticity PC is asymptotically equivalent to maximum likelihood (see Bai, 2003, for a discussion). This is true in the special case considered here, and provided that T/N does not go to zero, so that $N^{-1}\Phi'\Sigma_e\Phi$ is in fact non-negligible. In general, nothing can be said about the relative efficiency of CA and PC. In Section 4, we therefore use Monte Carlo simulations to shed light on this issue.

Because the terms that appear in the variance of $t(y_{T+h|T})$ are unknown, as it stands, the asymptotic normal distribution theory reported in Theorem 1 is not very useful to us. We therefore look for consistent estimators of these terms. Analogous to Bai and Ng (2006), a natural candidate for an estimator of ϕ is given by $\widehat{\phi} = \widehat{z}_T'\widehat{\Sigma}_z^+\widehat{\Sigma}_{z\epsilon}\widehat{\Sigma}_z^+\widehat{z}_T$, where

$$\widehat{\Sigma}_{z\epsilon} = \frac{1}{T}\sum_{t=1}^{T-h}\widehat{\epsilon}_{t+h}^2\widehat{z}_t\widehat{z}_t', \tag{25}$$

$$\widehat{\Sigma}_z = \frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t', \tag{26}$$

with $\widehat{\epsilon}_{t+h} = y_{T+h} - \widehat{\delta}'\widehat{z}_t$. The estimator of $\Phi'\Sigma_e\Phi$ is given by $\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}$, where $\widehat{\Phi} = \widehat{\alpha}$ and the exact form of $\widehat{\Sigma}_e$ depends on whether or not $e_{i,t}$ is weakly correlated across the cross-section. Let $\widehat{e}_{i,t} = x_{i,t} - \widehat{\lambda}_i'\widehat{F}_t$, where $\widehat{\lambda}_i$ is the LS slope estimator in a time series regression of $x_{i,t}$ onto $\widehat{F}_t$. If $e_{i,t}$ is cross-section uncorrelated, we use

$$\widehat{\Sigma}_e = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{e}_{i,t}\widehat{e}_{i,t}', \tag{27}$$

whereas if $e_{i,t}$ is weakly cross-section correlated, we use

$$\widehat{\Sigma}_e = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\widehat{e}_{i,t}\widehat{e}_{j,t}', \tag{28}$$

where n is a cross-section truncation parameter satisfying $n/\delta_{NT}^2 \to 0$.[7] Regardless of which

[6] This follows from replacing λ by the corresponding PC rotation matrix H and then using Theorem 1 of Bai (2003).

[7] The truncation in $\widehat{\Sigma}_e$ under weak cross-section correlation does not require any additional assumptions other than Assumptions A–D. The reason is that the cross-section correlations can be consistently estimated using the time series variation as $T \to \infty$.

estimator that is being used, the feasible version of $t(y_{T+h|T})$ is given by

$$\widehat{t}(y_{T+h|T}) = \frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{T^{-1}\widehat{\phi} + N^{-1}\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}}}. \tag{29}$$

Remark 4. Assumption B rules out heteroskedasticity over time in $e_{i,t}$. An easy way to allow for general heteroskedasticity in both i and t when $e_{i,t}$ is cross-correlation free is to replace $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{e}_{i,t}\widehat{e}_{i,t}'$ with $N^{-1}\sum_{i=1}^{N}\widehat{e}_{i,T}\widehat{e}_{i,T}'$.

The next theorem is stated in terms of an $m \times 1$ vector Δ, which is such that $\Delta = 0_{m\times 1}$ if $m = r$ and

$$\Delta = \begin{bmatrix} -\lambda_r^{-1}\lambda_{-r}b \\ \sqrt{\tau}\sigma_\varepsilon(\Lambda_{-r}'\Sigma_e\Lambda_{-r})^{-1/2}Z_{m-r} \end{bmatrix}$$

if $m > r$, where $Z_{m-r} \sim N(0_{(m-r)\times 1}, I_{m-r})$ and τ is such that $N/T \to \tau$.

Theorem 2. *Suppose that Assumptions A–D hold. Then, as $N, T \to \infty$ with $\sqrt{T}/N \to 0$ and $N/T \to \tau < \infty$,*

$$\widehat{\phi} \to_p \phi,$$
$$\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi} \to_d (\Phi + \Delta)'\Sigma_e(\Phi + \Delta).$$

Theorem 2 shows that while $\widehat{\phi}$ is generally consistent for ϕ, $\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}$ is not. In fact, under $m > r$ the limit of $\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}$ is not even constant but random. The randomness of Δ disappears if $\tau = 0$. However, because of the presence of $\lambda_r^{-1}\lambda_{-r}b$ in Δ, $\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}$ is still not consistent for $\Phi'\Sigma_e\Phi$. The main exception is if $m = r$, in which case $\Delta = 0_{m\times 1}$, and so

$$\widehat{t}(y_{T+h|T}) = \frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{T^{-1}\phi + N^{-1}\Phi'\Sigma_e\Phi}} + o_p(1) \to_d N(0,1) \tag{30}$$

as $N, T \to \infty$ with $\sqrt{T}/N \to 0$. Hence, under these conditions, a $100(1-\gamma)\%$ confidence interval for $y_{T+h|T}$ can be easily constructed as

$$\begin{aligned} \mathrm{CI}_\gamma(y_{T+h|T}) = [&\widehat{y}_{T+h|T} - z_{\gamma/2}\cdot\sqrt{T^{-1}\widehat{\phi} + N^{-1}\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}}, \\ &\widehat{y}_{T+h|T} + z_{\gamma/2}\cdot\sqrt{T^{-1}\widehat{\phi} + N^{-1}\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}}], \end{aligned} \tag{31}$$

where $z_\gamma = \Psi^{-1}(1-\gamma)$ is the $(1-\gamma)$-th quantile of the standard normal cumulative distribution function, here denoted $\Psi(x)$. The asymptotic coverage of this confidence interval can be easily deduced from (30) and is given by

$$\lim_{N,T\to\infty} P[y_{T+h|T} \in \mathrm{CI}_\gamma(y_{T+h|T})] = \lim_{N,T\to\infty} P(|\widehat{t}(y_{T+h|T})| \le z_{\gamma/2}) = 1-\gamma. \tag{32}$$

This confidence interval is for the conditional mean of y_{T+h}. If we want a confidence interval for y_{T+h} itself, then we have to assume that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Under this restriction, a $100(1-\gamma)\%$ confidence interval for y_{T+h} is given by

$$\begin{aligned}\mathrm{CI}_\gamma(y_{T+h}) = [\, &\widehat{y}_{T+h|T} - z_{\gamma/2} \cdot \sqrt{\widehat{\sigma}_\varepsilon^2 + T^{-1}\widehat{\phi} + N^{-1}\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}}, \\ &\widehat{y}_{T+h|T} + z_{\gamma/2} \cdot \sqrt{\widehat{\sigma}_\varepsilon^2 + T^{-1}\widehat{\phi} + N^{-1}\widehat{\Phi}'\widehat{\Sigma}_e\widehat{\Phi}}\,],\end{aligned} \tag{33}$$

where $\widehat{\sigma}_\varepsilon^2 = T^{-1}\sum_{t=1}^{T-h}\widehat{\varepsilon}_{t+h}^2$, whose asymptotic coverage is again given by $1-\gamma$.

The problem with the above results is of course that they only apply when $m = r$, which is unlikely to be the case in practice. Interestingly, if we accept that test statistics and confidence intervals are conservative, asymptotically valid inference is possible also when $m > r$. This is formalized in Theorem 3, which holds for all $m \geq r$.

Theorem 3. *Suppose that Assumptions A–D hold. Then, as N, $T \to \infty$ with $T/N \to 0$,*

$$\lim_{N,T\to\infty} P(|\widehat{t}(y_{T+h|T})| > z_{\gamma/2}) \leq \gamma.$$

Theorem 3 implies that

$$\lim_{N,T\to\infty} P[y_{T+h|T} \in \mathrm{CI}_\gamma(y_{T+h|T})] \geq 1-\gamma. \tag{34}$$

This holds for all $m \geq r$. If, however, $m = r$, then the asymptotic coverage is exactly $1-\gamma$ and the $T/N \to 0$ requirement is no longer needed, as is evident from (30). The fact that the $T/N \to 0$ condition is only needed when $m > r$ reinforces the importance of not treating r as known in the asymptotic analysis. Hence, while Theorem 3 only requires that $m \geq r$, the results do depend on whether $m = r$ or $m > r$. In the next section, we use Monte Carlo simulations to investigate the effect of the conservativeness when $m > r$ in small samples. According to the results, the effect of using the conservative critical values is almost nonexistent. Finally, note that while stated in terms of $\mathrm{CI}_\gamma(y_{T+h|T})$, the result in (34) applies also to $\mathrm{CI}_\gamma(y_{T+h})$.

4 Monte Carlo simulations

In this section, we evaluate the small-sample properties of the CA-based forecasts. The results are compared to those obtained when using both the true factors and the PC-based method of Bai and Ng (2006). The DGP used for this purpose is similar to the DGP used in the Monte Carlo study of Bai and Ng (2006), and can be seen a restricted version of (1) and (2) that sets

$h = 4$, $W_1 = \cdots = W_T = 1$, $\alpha = 1_{m\times 1} = [1, ..., 1]'$, $\beta = 1$ and $\varepsilon_t \sim N(0, 1)$. Serial correlation in F_t is permitted through

$$F_t = \rho F_{t-1} + \sqrt{1-\rho^2} u_t, \tag{35}$$

where $\rho = 0.5$ and $u_t \sim N(0, 1)$. Bai and Ng (2006) assume that $m = r = 1$. According to our asymptotic results, however, it matters whether $m = r$ or $m > r$, and in this section we therefore set $m \in \{1, 2\}$. The DGP of $e_{i,t}$ is similar to the one considered by Bai and Ng (2006), and is given by

$$e_t = \Omega(u)^{1/2} v_t, \tag{36}$$

where $v_t = [v_{1,t}, ..., v_{N,t}]'$ is $N \times m$ with $v_{i,t} \sim N(0_{m\times 1}, \sigma^2_{v,i} I_m)$ being $m \times 1$. Hence, $\Omega(u)^{1/2}$ is the lower triangular Cholesky factor of the $N \times N$ Toeplitz matrix $\Omega(u)$, whose i-th diagonal element is u^i if $i \le 10$ and is zero otherwise. Hence, in this DGP, the cross-section correlation and cross-section heteroskedasticity of $e_{i,t}$ is controlled by u and $\sigma^2_{v,i}$, respectively. We consider a total of eight DGPs that differ only in how we parameterize r, m, u and $\sigma^2_{v,i}$ and λ_i. This is described in Table 1.

As already explained in Section 3, the appropriate choice of $\widehat{\Sigma}_e$ depends on whether $e_{i,t}$ is cross-section correlated, which is similar to PC. The results reported in this section are based on using the cross-section correlation robust variance estimator only when $e_{i,t}$ is in fact weak cross-section correlated, and in so doing we follow Bai and Ng (2006), and set $n = \lfloor \min\{\sqrt{N}, \sqrt{T}\} \rfloor$. Also, the PC results are based on taking the true number of factors, r, as known. We also simulate the forecast based on taking F_t as known.[8] We report the empirical coverage rate and mean squared forecast error (MSE) for both $\widehat{y}_{T+h}$ and $\widehat{y}_{T+h|T}$ when $\gamma = 0.05$. The results are based on 5,000 replications of samples of size $(N, T) \in \{30, 50, 100, 200\}$.

Results reported in Tables 2–9 can be summarized as follows:

- The coverage rates for $\widehat{y}_{T+h}$ are very close to the nominal 95% rate. This is true regardless of whether one uses CA or PC. For $\widehat{y}_{T+h|T}$, however, there is a marked difference in performance with the coverage of CA generally being much closer to the nominal rate, especially among the smaller values of N and T. Hence, in terms of coverage, it seems as that the simplicity of CA comes without cost. In fact, if anything, it seems as that

[8] When F_t is known, we simply ignore the terms induced by the estimation of F_t from the variance, as in Bai and Ng (2006).

the computationally most costly PC approach is also the one with poorest performance, which, as pointed out in Section 1, is largely in agreement with the findings of the previous forecast combination and panel literatures.

- The coverage rates reported for the case when $m = r = 1$ are very similar to those reported for $m = 2 > r = 1$, suggesting the use of conservative critical values in the latter case has little or no effect on the coverage of CA. The fact that CA continues to perform relatively well also when $m > r$ is particularly noteworthy given that PC treats r as known, which means that, unlike CA, PC do not overestimate the number of factors.

- As expected, the MSE of $\widehat{y}_{T+h}$ is generally smallest when the forecast is based on the true factors, although the difference when compared to CA and PC is typically very small. Similarly, while the MSE results for CA and PC are not equal, the differences are small. In particular, while PC tend to perform best when $r = m = 1$ (as expected given the discussion following Corollary 1), in the more realistic case when $m > r$, it is the other way around. The variance of $\widehat{y}_{T+h}$ when $\varepsilon_t \sim N(0,1)$ is the variance of $\widehat{y}_{T+h|T}$ plus one (see Section 3). Consistent with this we see that the MSE of $\widehat{y}_{T+h|T}$ is roughly that of $\widehat{y}_{T+h}$ less one. Except for this difference, however, the results reported for $\widehat{y}_{T+h|T}$ are qualitatively the same as those reported for $\widehat{y}_{T+h}$.

All-in-all, we find that the relatively simple and user-friendly CA-based forecasting approach tends to perform at least as good as the main competitor based on PC.

5 Empirical illustration

In this section, we revisit the Stock and Watson (2005) data set, which has been heavily used in the PC literature.[9] An incomplete list of users of this data set include Bai and Ng (2008), Breitung and Pigorsch (2013), Breitung and Eickmeier (2011), and Hallin and Liŝka (2007). The data set has also been revised and extended in several directions (see, for example, Stock and Watson, 2009). The data comprise of 132 monthly macroeconomic series for the US and stretch the period 1960:1 to 2003:12. Many of the series are non-stationary, which, as pointed out in Section 2, is not permitted under our assumptions. The data are therefore transformed by taking logs, first or second differences when necessary, as in Stock and Watson (2005). The

[9]The data are available at http://www.princeton.edu/∼mwatson/wp.html.

variables to be forecasted are the same as in Stock and Watson (2002b). The first four are measures of real economic activity that are used to construct the Index of Coincident Economic Indicators maintained by the Conference Board. They are total industrial production (IPS10), real personal income less transfers (A0M051), real manufacturing and trade sales (A0M057), and the number of employees on non-farm payrolls (PAYEMS). The remaining four series are price indices. They are consumer price index (PUNEW), personal consumption expenditure implicit price deflator (GMDC), consumer price index less food and energy (CPILFESL), and consumer price index for finished goods (PWFSA).[10]

In most studies of the Stock and Watson (2005) data set the number of common factors is a key issue. Different approaches have been used; however, most studies estimate between six and 10 factors, which, as Stock and Watson (2005) note in their Section 4.3 called "Why So Many Factors?", seems excessive given the estimates reported in the bulk of the empirical factor model literature. Stock and Watson (2005) themselves apply PC and two of the information criteria of Bai and Ng (2002) with which they estimate seven factors, although they mention that the criteria are almost flat for between six and 10 factors. In other words, there is substantial uncertainty over the appropriate number of factors to use, which means that it is important to use forecasting approaches that are robust in this regard. The CA-based approach considered here only requires that the number of CAs is not smaller than the number of factors, and is in this sense more suitable than existing PC-based forecasting approaches, which all rely on correct specification of the number of factors.

The robustness with respect to the number of factors is one reason for preferring CA, as opposed to PC. Another reason is that it enables easy interpretation of the estimated factors. One of the issues with estimated PC factors is that they are difficult to interpret, because each factor estimate is influenced to some degree by all the series in the panel data set, and the orthogonalization in PC means that no one of them will correspond exactly to a precise economic concept. A common approach is to try to label the factors according to their relationship with the underlying series (see, for example, Ludvigson and Ng, 2009). Stock and Watson (2005) look at the marginal contribution of each factor estimate to each of the series in the data set, which are organized into 14 categories (as in Stock and Watson, 2002b), namely, (1) real output and income, (2) employment and hours, (3) real retail, manufacturing and trade sales, (4)

[10]Please see Table A.1 of Stock and Watson (2005) for a complete definition of all the 132 series by the mnemonics given here in parentheses.

consumption, (5) housing starts and sales, (6) real inventories, (7) orders, (8) stock prices, (9) exchange rates, (10) interest rates and spreads, (11) money and credit quantity aggregates, (12) price indexes, (13) average hourly earnings, and (14) miscellaneous. According to the results, while the first and most important factor loads mainly on series in the real output and income, and employment and hours categories, the second loads mainly on interest rates, consumption and stock prices. The second factor also loads on inflation, as does the third factor. The fourth factor loads on interest rates, the fifth factor loads mainly on employment and hours, and the sixth and seventh factors load mainly on exchange rates.

Of course, while suggestive, any labeling of the PC factors is naturally imperfect. An advantage of working with CAs is that they lend themselves to easy interpretation. In the current application, it is natural to average within categories and to interpret the resulting factor estimates accordingly. This will ensure not only that the estimated factors are economically meaningful, but also that there are enough CAs to capture the underlying factors. An issue arises, as the number of series within each category is not the same, that is, N is not the same across categories. This has two implications. One implication is that while the definition of $\widehat{F}_t$ is trivially extendable to the case with different number of cross-section units for each average, the definition of $\widehat{\Sigma}_e$ in the cross-section uncorrelated case is not. In this section, we therefore use the definition in cross-section correlated case, which is not only easily extendable (by just summing across the estimated idiosyncratic errors within each category), but also very general. The other implication is that because the number of series within each category varies greatly, from 27 (employment and hours) to just one (consumption and miscellaneous), we cannot always rely on consistency of the estimated factors. The way we solve this issue is to simply treat the averages for the smallest categories as known factors. The predictive content of each of the 14 factor candidates is evaluated by grid-searching over all possible combinations, and using the Schwarz Bayesian information criterion (BIC) to select the one to use (as in Stock and Watson, 2002a, 2002b).

Following the bulk of the previous literature (see, for example, Bai and Ng, 2008, and Stock and Watson, 2002a, 2002b), we use an expanding estimation window using all available data at the point of the forecast to estimate unknown parameters and factors. Another reason to prefer CA over PC is the extreme ease with which this type of rolling window out-of-sample forecasting can be carried out. Note in particular how, unlike with PC, the CAs do not have to be reestimated for each forecast. The first forecast is based on using the first 100 observations covering the period 1960:1–1968:4 for the estimation. In interest of comparison, the CA-based forecast is compared to that obtained when using PC. Stock and Watson (2002b) consider three PC-based forecasting models (called "diffusion index forecasts") that differ only in the lag structure of the fitted forecasting model. Interestingly, in most cases more elaborate lag structures offer little or no improvement over the simplest "DI" specification with only two factors and a constant. In this section, we consider a slightly more general model with $\widehat{z}_t = [\widehat{F}_t', 1, y_t]'$, where the dimension of $\widehat{F}_t$ is again selected by the BIC. As in the previous literature, we consider three forecasting horizons; $h = 6$, $h = 12$ and $h = 24$.

Table 10 reports the results on the MSE of the CA- and PC-based forecasts (times 100), computed relative to the MSE of a simple univariate autoregressive (AR) model that sets $\widehat{z}_t = [1, y_t]'$. Before we come to these results, however, we discuss the selection of the CAs. An important finding in Stock and Watson (2002a, 2002b) is that while when the selection is done at the estimated panel factor model for $x_{i,t}$ (as in Stock and Watson, 2005), as already mentioned, the required number of factors tends to be quite large, when the selection is done at the estimated time series forecasting model for y_{t+h} the required number of factors is much smaller, between one and three. Consistent with this finding, the number of CAs for each of the eight forecasted variables ranges between one and five with an average of 2.25. The CAs with the highest selection rates are real output and income, employment and hours, and price indices, which is broadly in agreement with the results of Stock and Watson (2005).

Looking now at Table 10, we see that the factors-based forecasts always outperform the AR benchmark. Hence, since the only the difference between these two sets of forecasts is the factors, we can infer that their inclusion leads to improved forecasting accuracy. We also see that the magnitude of the relative MSEs is roughly in agreement with the results reported by Stock and Watson (2002a, 2002b). More importantly, when we compare the relative MSE of the two factor-based forecasts we see that CA is almost uniformly better than PC. The only exceptions are for GMDC, CPILFESL and PWFSA when $h = 6$, in which case PC performs best, although the difference in performance is only marginal. The same cannot be said for the other variables and horizons. On the contrary, here the difference in performance can be quite substantial. To take an extreme example, consider PUNEW when $h = 24$. While the relative MSE of PC is 88.5, the relative MSE of CA is much smaller, 65.5, which represents an improvement of 26%. The average gain in performance obtained by using CA rather than PC is 5% when $h = 6$, 9% when $h = 12$ and 14% when $h = 24$. The average gain therefore increases with the forecasting horizon. A possible explanation for this is that CA is relatively less affected by the uncertainty that comes from increasing h. Hence, consistent with the Monte Carlo results reported in Section 5, we find that there is little or no cost to using the relatively simple CA-based forecasting approach. In fact, if anything, it is the other way around.

6 Conclusion

The existing forecasting literature for factor-augmented regressions is based almost exclusively on PC. The present paper is the first to consider CA as an alternative to PC. The main theoretical contribution is to show that the forecasts obtained based on CAs of $m \geq r$ panel data variables are consistent and asymptotically normal as N, $T \to \infty$ with $\sqrt{T}/N \to 0$. A problem arises in the empirically relevant case when $m > r$. In particular, the use of too many CAs causes an inconsistency in the estimator of the asymptotic variance of the conditional mean, which means that the coverage of the resulting confidence intervals is incorrect. The coverage is, however, shown to be upwards biased, which means that the confidence intervals will be conservative. This last result is important, because the previous PC-based literature assumes that r is known, which in CA is tantamount to requiring that $m = r$. The aforementioned problem caused by over-specification of the number of factors has therefore been completely overlooked.

The Monte Carlo and empirical results reveal that CA tends to perform at least as well as PC. CA is also computationally very attractive and enables easy interpretation of the estimated factors. It should therefore be a valuable addition to the already existing menu of forecasting tools.

References

Bai, J., and S. Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* **70**, 191–221.

Bai, J., and S. Ng (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica* **74**, 1133–1150.

Bai, J., and S. Ng (2008). Forecasting Economic Time Series using Targeted Predictors. *Journal of Econometrics* **146**, 304–317.

Bai, J., and S. Ng (2009). Boosting Diffusion Indices. *Journal of Applied Econometrics* **24**, 607–629.

Banerjee, A., M. Marcellino and I. Masten (2008). Forecasting Macroeconomic Variables Using Diffusion Indexes in Short Samples with Structural Change. In Rapach, D., and M. Wohar (Eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*. Emerald Publishing. Bingley, UK.

Boivin, J., and S. Ng (2006). Are More Data Always Better for Factor Analysis. *Journal of Econometrics* **132**, 169–194.

Breitung, J., and S. Eickmeier (2011). Testing for Structural Breaks in Dynamic Factor Models. *Journal of Econometrics* **163**, 71–84.

Breitung, J., and J. Tenhofen (2011). GLS Estimation of Dynamic Factor Models. *Journal of the American Statistical Association* **106**, 1150–1166.

Breitung, J., and U. Pigorsch (2013). A Canonical Correlation Approach for Selecting the Number of Dynamic Factors. *Oxford Bulletin of Economics and Statistics* **75**, 23–36.

Chang, Y., and P. C. B. Phillips (1995). Time Series Regression with Mixtures of Integrated Regressors. *Econometric Theory* **12**, 1033–1094.

Cheng, X., and B. E. Hansen (2015). Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach. *Journal of Econometrics* **86**, 280–293.

Choi, I. (2012). Efficient Estimation of Factor Models. *Econometric Theory* **28**, 274–308.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and Strong Cross Section Dependence and Estimation of Large Panels. *Econometric Journal* **14**, C45–C90.

Corradi, V., and N. R. Swanson (2014). Testing for Structural Stability of Factor Augmented Forecasting Models. *Journal of Econometrics* **182**, 100–118.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? *Journal of Econometrics* **146**, 318–328.

Djogbenou, A., S. Gonçalves, and B. Perron (2015). Bootstrap Inference in Regressions with Estimated Factors and Serial Correlation. *Journal of Time Series Analysis* **36**, 481–502.

Djogbenou, A., S. Gonçalves, and B. Perron (2017). Bootstrap Prediction Intervals for Factor Models. *Journal of Business & Economics Statistics* **35**, 53–69.

Eickmeier, S., and C. Ziegler (2008). How Successful are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach. *Journal of Forecasting* **27**, 237–265.

Gonçalves, S., and B. Perron (2014). Bootstrapping Factor-Augmented Regression Models. *Journal of Econometrics* **82**, 156–173.

Hallin, M., and R. Liška (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association* **102**, 603–617.

Karabiyik, H., S. Reese and J. Westerlund (2017). On the Role of the Rank Condition in CCE Estimation of Factor-Augmented Panel Regressions. *Journal of Econometrics* **197**, 60–64.

Ludvigson, S., and S. Ng (2009). Macro Factors in Bond Risk Premia. *Review of Financial Studies* **22**, 5027–5067.

Moon, H. R., and M. Weidner (2015). Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects. *Econometrica* **83**, 1543–1579.

Park, J., and P. C. B. Phillips (2000). Nonstationary Binary Choice Models. *Econometrica* **68**, 1249–1280.

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica* **74**, 967–1012.

Pesaran, M. H., T. Schuermann, and L. V. Smith (2009). Forecasting Economic and Financial Variables with Global VARs. *International Journal of Forecasting* **25**, 642–675.

Smith, J., and K. F. Wallis (2009). A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics* **71**, 331–355.

Stock, J. H., and M. W. Watson (2002a). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* **97**, 1167–1179.

Stock, J. H., and M. W. Watson (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics* **20**, 147–162.

Stock, J. H., and M. W. Watson (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* **23**, 405–430.

Stock, J. H., and M. W. Watson (2005). Implications of Dynamic Factor Models for VAR Analysis. NBER Working Paper No. 11467.

Stock, J. H., and M. W. Watson (2009). Forecasting in Dynamic Factor Models Subject to Structural Instability. In Shephard, N., and J. Castle (Eds.), *The Methodology and Practice of Econometrics: Festschrift in Honor of David F. Hendry*, 1–57. Oxford University Press.

Timmermann, A. (2006). Forecast Combinations. In Elliot, G., C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, 135–196. Amsterdam. Elsevier.

Westerlund, J., and J.-P. Urbain (2015). Cross-Sectional Averages Versus Principal Components. *Journal of Econometrics* **85**, 372–377.

Table 1: Monte Carlo DGPs.

Label	r	m	u	$\sigma^2_{v,i}$	λ_i
DGP1	1	1	0	1	$U[0,1]$
DGP2	1	1	0.5	1	$U[0,1]$
DGP3	1	1	0	$U[0.5,1.5]$	$U[0,1]$
DGP4	1	1	0.5	$U[0.5,1.5]$	$U[0,1]$
DGP5	1	2	0	1	$[U[0,1],U[0,0.5]]$
DGP6	1	2	0.5	1	$[U[0,1],U[0,0.5]]$
DGP7	1	2	0	$U[0.5,1.5]$	$[U[0,1],U[0,0.5]]$
DGP8	1	2	0.5	$U[0.5,1.5]$	$[U[0,1],U[0,0.5]]$

Notes: r and m refer to the number of factors in F_t and the number of panel data variables in $x_{i,t}$, respectively. u measures the extent of cross-section correlation in the "idiosyncratic" errors driving $x_{i,t}$, and $\sigma^2_{v,i}$ is the variance of those errors. λ_i is the $r \times m$ factor loading matrix.

Table 2: Monte Carlo results for DGP1 with $m = r = 1$, and cross-section uncorrelated and homoskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.97	0.97	0.95	0.91	0.89	0.95	0.90	0.96	0.13	0.15	0.07
50	30	0.96	0.96	0.96	0.97	0.94	0.94	0.95	0.92	0.95	0.11	0.13	0.07
100	30	0.97	0.97	0.96	0.91	0.88	0.90	0.96	0.94	0.96	0.09	0.11	0.07
200	30	0.96	0.96	0.96	0.95	0.92	0.94	0.95	0.94	0.95	0.08	0.11	0.07
30	50	0.96	0.96	0.96	1.00	0.96	0.94	0.94	0.85	0.96	0.11	0.11	0.04
50	50	0.96	0.96	0.96	1.01	0.98	0.97	0.95	0.88	0.95	0.08	0.09	0.04
100	50	0.96	0.96	0.96	0.93	0.92	0.92	0.96	0.92	0.96	0.06	0.08	0.04
200	50	0.97	0.97	0.97	0.94	0.92	0.93	0.95	0.93	0.95	0.05	0.07	0.04
30	100	0.96	0.95	0.95	1.06	1.03	0.98	0.93	0.77	0.95	0.09	0.08	0.02
50	100	0.95	0.95	0.95	1.05	1.03	1.01	0.94	0.82	0.95	0.06	0.06	0.02
100	100	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.88	0.95	0.04	0.05	0.02
200	100	0.96	0.96	0.96	0.93	0.92	0.92	0.94	0.90	0.95	0.03	0.04	0.02
30	200	0.96	0.95	0.96	1.03	1.00	0.95	0.93	0.65	0.94	0.08	0.07	0.01
50	200	0.96	0.96	0.96	1.01	0.99	0.97	0.95	0.75	0.95	0.05	0.04	0.01
100	200	0.95	0.95	0.95	1.00	0.99	0.98	0.95	0.82	0.96	0.03	0.03	0.01
200	200	0.95	0.95	0.95	1.01	1.01	1.01	0.94	0.88	0.95	0.02	0.02	0.01

Notes: "CA", "PC" and "F" refer to the results based on CAs, estimated PC factors based on knowing the true number of factors, r, and the true factors, respectively. "CR" and "MSE" denote the coverage rate and the empirical mean squared forecast error, respectively. See Table 1 for a detailed description of the DGP.

Table 3: Monte Carlo results for DGP2 with $m = r = 1$, and cross-section correlated and homoskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.97	0.97	0.98	0.92	0.89	0.92	0.94	0.96	0.16	0.16	0.07
50	30	0.96	0.97	0.96	0.99	0.94	0.94	0.94	0.94	0.95	0.13	0.14	0.07
100	30	0.97	0.97	0.96	0.92	0.88	0.90	0.94	0.96	0.96	0.10	0.11	0.07
200	30	0.96	0.96	0.96	0.95	0.92	0.94	0.95	0.95	0.95	0.08	0.11	0.07
30	50	0.96	0.96	0.96	1.02	0.97	0.94	0.91	0.91	0.96	0.13	0.12	0.04
50	50	0.96	0.96	0.96	1.02	0.99	0.97	0.93	0.93	0.95	0.10	0.10	0.04
100	50	0.96	0.96	0.96	0.94	0.92	0.92	0.94	0.95	0.96	0.07	0.08	0.04
200	50	0.97	0.97	0.97	0.94	0.91	0.93	0.94	0.96	0.95	0.06	0.07	0.04
30	100	0.96	0.96	0.95	1.08	1.04	0.98	0.89	0.88	0.95	0.11	0.09	0.02
50	100	0.95	0.95	0.95	1.07	1.04	1.01	0.91	0.91	0.95	0.08	0.07	0.02
100	100	0.96	0.96	0.96	0.97	0.95	0.95	0.93	0.94	0.95	0.05	0.05	0.02
200	100	0.96	0.96	0.96	0.94	0.92	0.92	0.93	0.95	0.95	0.04	0.04	0.02
30	200	0.96	0.96	0.96	1.05	1.01	0.95	0.89	0.84	0.94	0.10	0.08	0.01
50	200	0.96	0.96	0.96	1.03	1.00	0.97	0.92	0.90	0.95	0.06	0.05	0.01
100	200	0.96	0.95	0.95	1.01	0.99	0.98	0.93	0.93	0.96	0.04	0.03	0.01
200	200	0.95	0.95	0.95	1.02	1.01	1.01	0.93	0.94	0.95	0.03	0.03	0.01

Notes: See the notes to Table 2.

Table 4: Monte Carlo results for DGP3 with $m = r = 1$, and cross-section uncorrelated and heteroskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.97	0.97	1.00	0.95	0.89	0.93	0.85	0.96	0.19	0.20	0.07
50	30	0.97	0.96	0.96	1.01	0.97	0.94	0.95	0.88	0.95	0.15	0.16	0.07
100	30	0.97	0.96	0.96	0.93	0.90	0.90	0.95	0.92	0.96	0.11	0.13	0.07
200	30	0.96	0.96	0.96	0.96	0.93	0.94	0.95	0.93	0.95	0.09	0.12	0.07
30	50	0.96	0.96	0.96	1.05	1.01	0.94	0.92	0.78	0.96	0.17	0.16	0.04
50	50	0.96	0.96	0.96	1.04	1.01	0.97	0.93	0.83	0.95	0.12	0.12	0.04
100	50	0.96	0.96	0.96	0.95	0.93	0.92	0.95	0.89	0.96	0.08	0.09	0.04
200	50	0.97	0.97	0.97	0.94	0.92	0.93	0.94	0.92	0.95	0.06	0.08	0.04
30	100	0.96	0.95	0.95	1.12	1.08	0.98	0.91	0.67	0.95	0.15	0.13	0.02
50	100	0.96	0.95	0.95	1.09	1.06	1.01	0.93	0.74	0.95	0.10	0.09	0.02
100	100	0.96	0.96	0.96	0.98	0.96	0.95	0.94	0.81	0.95	0.06	0.06	0.02
200	100	0.96	0.96	0.96	0.94	0.93	0.92	0.94	0.87	0.95	0.04	0.05	0.02
30	200	0.96	0.95	0.96	1.09	1.05	0.95	0.91	0.55	0.94	0.14	0.12	0.01
50	200	0.96	0.96	0.96	1.04	1.02	0.97	0.93	0.64	0.95	0.08	0.07	0.01
100	200	0.96	0.95	0.95	1.02	1.00	0.98	0.94	0.73	0.96	0.05	0.04	0.01
200	200	0.96	0.95	0.95	1.02	1.01	1.01	0.94	0.81	0.95	0.03	0.03	0.01

Notes: See the notes to Table 2.

Table 5: Monte Carlo results for DGP4 with $m = r = 1$, and cross-section correlated and heteroskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.97	0.97	1.05	0.96	0.89	0.91	0.91	0.96	0.23	0.21	0.07
50	30	0.97	0.97	0.96	1.04	0.98	0.94	0.92	0.93	0.95	0.18	0.17	0.07
100	30	0.97	0.97	0.96	0.94	0.90	0.90	0.94	0.95	0.96	0.12	0.13	0.07
200	30	0.96	0.96	0.96	0.97	0.93	0.94	0.94	0.95	0.95	0.10	0.12	0.07
30	50	0.96	0.96	0.96	1.10	1.03	0.94	0.89	0.89	0.96	0.21	0.17	0.04
50	50	0.96	0.96	0.96	1.07	1.02	0.97	0.92	0.91	0.95	0.15	0.13	0.04
100	50	0.97	0.96	0.96	0.97	0.93	0.92	0.93	0.94	0.96	0.10	0.10	0.04
200	50	0.96	0.97	0.97	0.95	0.92	0.93	0.94	0.95	0.95	0.07	0.08	0.04
30	100	0.96	0.96	0.95	1.16	1.09	0.98	0.88	0.85	0.95	0.19	0.14	0.02
50	100	0.96	0.95	0.95	1.12	1.07	1.01	0.90	0.89	0.95	0.13	0.10	0.02
100	100	0.96	0.96	0.96	1.00	0.97	0.95	0.92	0.92	0.95	0.08	0.07	0.02
200	100	0.96	0.96	0.96	0.95	0.93	0.92	0.93	0.94	0.95	0.05	0.05	0.02
30	200	0.96	0.96	0.96	1.13	1.07	0.95	0.87	0.81	0.94	0.18	0.13	0.01
50	200	0.96	0.96	0.96	1.08	1.03	0.97	0.91	0.89	0.95	0.11	0.08	0.01
100	200	0.96	0.95	0.95	1.04	1.01	0.98	0.92	0.92	0.96	0.07	0.05	0.01
200	200	0.96	0.95	0.95	1.03	1.02	1.01	0.93	0.93	0.95	0.04	0.03	0.01

Notes: See the notes to Table 2.

Table 6: Monte Carlo results for DGP5 with $m = 2 > r = 1$, and cross-section uncorrelated and homoskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.96	0.96	0.96	0.95	0.97	0.96	0.98	0.90	0.95	0.13	0.15	0.07
50	30	0.97	0.97	0.96	0.89	0.90	0.90	0.98	0.92	0.96	0.12	0.13	0.07
100	30	0.97	0.96	0.96	0.92	0.92	0.94	0.98	0.94	0.95	0.11	0.12	0.07
200	30	0.97	0.96	0.96	0.89	0.89	0.91	0.98	0.94	0.95	0.11	0.11	0.07
30	50	0.97	0.96	0.96	0.95	0.96	0.93	0.97	0.86	0.96	0.09	0.11	0.04
50	50	0.97	0.96	0.96	0.91	0.93	0.92	0.98	0.89	0.96	0.08	0.09	0.04
100	50	0.97	0.97	0.97	0.92	0.91	0.93	0.98	0.91	0.95	0.07	0.08	0.04
200	50	0.96	0.96	0.96	0.94	0.93	0.96	0.98	0.93	0.95	0.06	0.07	0.04
30	100	0.95	0.95	0.95	1.02	1.03	0.99	0.96	0.77	0.95	0.06	0.08	0.02
50	100	0.96	0.96	0.96	0.96	0.97	0.95	0.97	0.82	0.95	0.05	0.06	0.02
100	100	0.96	0.96	0.96	0.92	0.94	0.92	0.97	0.88	0.95	0.04	0.05	0.02
200	100	0.96	0.95	0.95	0.99	0.99	1.00	0.98	0.91	0.95	0.04	0.04	0.02
30	200	0.96	0.95	0.96	0.98	1.01	0.96	0.95	0.66	0.95	0.05	0.07	0.01
50	200	0.96	0.95	0.95	0.99	1.00	0.98	0.97	0.73	0.96	0.03	0.04	0.01
100	200	0.95	0.95	0.95	1.01	1.01	1.01	0.97	0.82	0.95	0.03	0.03	0.01
200	200	0.96	0.95	0.95	0.98	0.99	0.98	0.97	0.88	0.95	0.02	0.02	0.01

Notes: See the notes to Table 2.

Table 7: Monte Carlo results for DGP6 with $m = 2 > r = 1$, and cross-section correlated and homoskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.96	0.97	0.96	0.96	0.98	0.96	0.96	0.93	0.95	0.15	0.15	0.07
50	30	0.97	0.97	0.96	0.90	0.90	0.90	0.97	0.95	0.96	0.13	0.14	0.07
100	30	0.96	0.96	0.96	0.92	0.92	0.94	0.98	0.95	0.95	0.12	0.12	0.07
200	30	0.97	0.97	0.96	0.89	0.89	0.91	0.98	0.96	0.95	0.11	0.11	0.07
30	50	0.97	0.96	0.96	0.96	0.97	0.93	0.95	0.92	0.96	0.10	0.11	0.04
50	50	0.97	0.96	0.96	0.92	0.95	0.92	0.96	0.94	0.96	0.09	0.09	0.04
100	50	0.97	0.97	0.97	0.92	0.91	0.93	0.97	0.95	0.95	0.08	0.08	0.04
200	50	0.96	0.96	0.96	0.94	0.94	0.96	0.97	0.95	0.95	0.07	0.07	0.04
30	100	0.95	0.95	0.95	1.03	1.04	0.99	0.93	0.88	0.95	0.08	0.09	0.02
50	100	0.96	0.96	0.96	0.97	0.98	0.95	0.95	0.91	0.95	0.06	0.07	0.02
100	100	0.96	0.96	0.96	0.93	0.94	0.92	0.96	0.93	0.95	0.05	0.05	0.02
200	100	0.96	0.95	0.95	0.99	0.99	1.00	0.97	0.95	0.95	0.04	0.04	0.02
30	200	0.96	0.96	0.96	0.99	1.01	0.96	0.92	0.85	0.95	0.06	0.08	0.01
50	200	0.96	0.95	0.95	1.00	1.01	0.98	0.94	0.90	0.96	0.04	0.05	0.01
100	200	0.96	0.95	0.95	1.01	1.02	1.01	0.96	0.93	0.95	0.03	0.03	0.01
200	200	0.95	0.95	0.95	0.98	0.99	0.98	0.96	0.95	0.95	0.02	0.02	0.01

Notes: See the notes to Table 2.

Table 8: Monte Carlo results for DGP7 with $m = 2 > r = 1$, and cross-section uncorrelated and heteroskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.96	0.96	0.98	1.02	0.96	0.97	0.85	0.95	0.16	0.19	0.07
50	30	0.97	0.97	0.96	0.91	0.92	0.90	0.98	0.89	0.96	0.14	0.16	0.07
100	30	0.97	0.96	0.96	0.93	0.93	0.94	0.98	0.92	0.95	0.12	0.13	0.07
200	30	0.97	0.96	0.96	0.89	0.89	0.91	0.98	0.93	0.95	0.11	0.12	0.07
30	50	0.97	0.96	0.96	0.98	1.01	0.93	0.96	0.79	0.96	0.12	0.16	0.04
50	50	0.97	0.96	0.96	0.93	0.96	0.92	0.97	0.83	0.96	0.10	0.12	0.04
100	50	0.97	0.97	0.97	0.92	0.93	0.93	0.97	0.88	0.95	0.08	0.10	0.04
200	50	0.96	0.96	0.96	0.94	0.94	0.96	0.98	0.92	0.95	0.07	0.08	0.04
30	100	0.95	0.95	0.95	1.05	1.08	0.99	0.94	0.68	0.95	0.09	0.13	0.02
50	100	0.96	0.96	0.96	0.97	0.99	0.95	0.95	0.74	0.95	0.07	0.09	0.02
100	100	0.96	0.96	0.96	0.93	0.95	0.92	0.96	0.82	0.95	0.05	0.06	0.02
200	100	0.96	0.95	0.95	0.99	1.00	1.00	0.97	0.87	0.95	0.04	0.05	0.02
30	200	0.96	0.96	0.96	1.01	1.05	0.96	0.93	0.55	0.95	0.08	0.11	0.01
50	200	0.96	0.95	0.95	1.01	1.03	0.98	0.95	0.62	0.96	0.05	0.07	0.01
100	200	0.95	0.95	0.95	1.02	1.03	1.01	0.96	0.72	0.95	0.03	0.04	0.01
200	200	0.96	0.95	0.95	0.99	0.99	0.98	0.96	0.81	0.95	0.03	0.03	0.01

Notes: See the notes to Table 2.

Table 9: Monte Carlo results for DGP8 with $m = 2 > r = 1$, and cross-section correlated and heteroskedastic errors.

		$\widehat{y}_{T+h}$						$\widehat{y}_{T+h\|T}$					
		CR			MSE			CR			MSE		
N	T	CA	PC	F	CA	PC	F	CA	PC	F	CA	PC	F
30	30	0.97	0.97	0.96	1.00	1.03	0.96	0.95	0.91	0.95	0.18	0.21	0.07
50	30	0.97	0.97	0.96	0.92	0.93	0.90	0.97	0.94	0.96	0.15	0.16	0.07
100	30	0.96	0.96	0.96	0.93	0.94	0.94	0.97	0.95	0.95	0.13	0.13	0.07
200	30	0.97	0.97	0.96	0.90	0.89	0.91	0.97	0.95	0.95	0.11	0.12	0.07
30	50	0.97	0.96	0.96	1.01	1.03	0.93	0.94	0.89	0.96	0.15	0.17	0.04
50	50	0.97	0.96	0.96	0.95	0.97	0.92	0.95	0.92	0.96	0.12	0.13	0.04
100	50	0.97	0.97	0.97	0.93	0.93	0.93	0.97	0.93	0.95	0.09	0.10	0.04
200	50	0.96	0.96	0.96	0.94	0.94	0.96	0.97	0.95	0.95	0.07	0.08	0.04
30	100	0.96	0.95	0.95	1.08	1.10	0.99	0.92	0.85	0.95	0.12	0.14	0.02
50	100	0.97	0.96	0.96	0.99	1.01	0.95	0.94	0.89	0.95	0.09	0.10	0.02
100	100	0.96	0.96	0.96	0.94	0.96	0.92	0.95	0.92	0.95	0.06	0.07	0.02
200	100	0.96	0.96	0.95	1.00	1.01	1.00	0.96	0.94	0.95	0.05	0.05	0.02
30	200	0.96	0.96	0.96	1.03	1.07	0.96	0.90	0.82	0.95	0.11	0.13	0.01
50	200	0.96	0.96	0.95	1.02	1.04	0.98	0.93	0.87	0.96	0.07	0.09	0.01
100	200	0.96	0.95	0.95	1.03	1.03	1.01	0.95	0.91	0.95	0.04	0.05	0.01
200	200	0.96	0.95	0.95	0.99	1.00	0.98	0.96	0.93	0.95	0.03	0.03	0.01

Notes: See the notes to Table 2.

Table 10: Relative MSE $\times$ 100.

	$h = 6$		$h = 12$		$h = 24$	
Variable	CA	PC	CA	PC	CA	PC
IPS10	70.65	79.52	54.69	62.23	41.87	49.39
A0M051	70.04	76.59	60.21	62.09	60.34	66.55
A0M057	74.80	84.70	58.30	63.72	39.86	43.93
PAYEMS	75.99	83.78	52.13	58.35	37.42	39.03
PUNEW	67.35	68.96	66.44	74.83	65.50	88.48
GMDC	66.30	65.93	69.03	71.70	71.35	86.03
CPILFESL	71.98	68.79	73.25	82.87	76.81	99.12
PWFSA	66.94	66.44	62.49	68.73	64.50	69.82

Notes: The results reported in the table are the relative mean squared forecast error (MSE) when compared to the AR model. "CA" and "PC" refer to the results based on CA and PC, respectively, and h is the forecast horizon.