Confidence Intervals for Diffusion Index Forecasts with a Large Number of Predictors

Jushan Bai* Serena Ng †

October 2003

Preliminary, comments welcome

Abstract

This paper examines the situation in which we have a large number of series N, each with T observations, and each series has some predictive ability for the variable of interest, y. We estimate common factors from the panel of data and show that $\sqrt{T}$ consistent forecasts for the conditional mean of y can be obtained. The effect of "generated regressors" is shown to be asymptotically negligible when T/N goes to zero. Regardless of the magnitude of T/N, we show how confidence intervals can be constructed. The same formulas can be applied to stationary or nonstationary factors. In cases when the idiosyncratic errors of the panel are cross-correlated, we also develop a covariance matrix estimator that is robust to cross-section correlation and heteroskedasticity. We provide a consistency proof this CS-HAC estimator.

*Department of Economics, NYU, 269 Mercer St, New York, NY 10003 Email: Jushan.Bai@nyu.edu.

†Department of Economics, University of Michigan, Ann Arbor, MI 48109 Email: Serena.Ng@umich.edu
The authors acknowledge financial support from the NSF (grants SBR-9896329, SES-0137084, SES-0136923
We thank seminar participants at Yale University for useful comments

1 Introduction

The use of factors to achieve dimension reduction has been found to be empirically useful in analyzing macroeconomic time series, and adding factors to an otherwise standard regression or forecasting model is being used by an increasing number of researchers[1]. However, the theoretical properties of the method are not fully understood and important issues remain to be addressed. In the case of forecasting, how to construct confidence intervals remains unknown. This is a nontrivial problem as the forecasting equation involves "generated regressors.' In this paper, we derive the rate of convergence and the limiting distribution of the factor augmented forecasts, enabling the construction of confidence intervals.

The object of interest is the h-period ahead forecast of a series y_t. The information available includes a large number of predictors x_{it} $(i = 1, 2, ..., N; t = 1, 2, ..., T)$ and a smaller set of other variables W_t. For example, W_t might be lags of y_t. If N was small, we could formulate a forecasting model with all the x_{it} and W_t as predictors. But the forecasts will be less efficient the larger the number of predictors because more parameters have to be estimated. And when N exceeds T, judicious choice of variables is necessary even though not making use of all relevant data is known to entail efficiency loss.

We consider a single forecasting equation

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

where y_t is a scalar series, and W_t is a vector of observable series. The vector F_t is unobservable. Instead, we observe a panel of data x_{it} which contains information about F_t. We refer to

$$x_{it} = \lambda_i' F_t + e_{it} \tag{2}$$

the factor representation of the data, with F_t being a $r \times 1$ vector of common factors, λ_i is the corresponding vector of factor loadings, and e_{it} is an idiosyncratic error. Equations (1) and (2) above constitute what is referred to by Stock and Watson (2002a) as a 'diffusion index forecasting model' (DI). Its defining characteristic is that information about x_{it} is parsimoniously summarized in a low dimensional vector, F_t. In economic analysis, these represent comovements or common shocks to economic time series.

If F_t is observable, and assuming the mean of ε_t conditional on past information is zero,

[1] See, for example Cristadoro, Forni, Reichlin and Giovanni (2001), Forni, Hallin, Lippi and Reichlin (2001b), Bernanke and Boivin (2002)

the (mean-squared) optimal forecast of y_t is the conditional mean and is given by

$$y_{T+h|T} = E(y_{T+h}|z_T) = \alpha' F_T + \beta' w_T \equiv \delta' z_T,$$

where $z_t = (F_t', \ W_t')'$ for all t. But such a forecast is not feasible, because α, β, and F_t are all unobservable. The feasible forecast that replaces the unknown objects by their estimates is:

$$\widehat{y}_{T+h|T} = \widehat{\alpha}' \widetilde{F}_T + \widehat{\beta}' W_T = \widehat{\delta}' \widehat{z}_T,$$

where $\widehat{z}_t = (\widetilde{F}_t', \ W_t')'$ for all t. We use a 'tilde' for estimates of the factor model of x_{it}, while hatted variables are estimated from the forecasting equation. To be precise, $\widehat{\alpha}$ and $\widehat{\beta}$ are the least squares estimates obtained from a regression of y_{t+h} on $\widetilde{F}_t$ and W_t, $t = 1, \ldots T-h$. The factors, F_t, are estimated from x_{it} by the method of principal components using data up to period T and will be discussed further below.

It is clear that $\widehat{\alpha}$ and $\widehat{\beta}$ are functions of "generated regressors" $\widetilde{F}_1, \widetilde{F}_2, ..., \widetilde{F}_{T-h}$, and the forecast $\widehat{y}_{T+h|T}$ itself also depends on $\widetilde{F}_T$. Thus, to study the behavior of the forecasts, we must examine the statistical properties of the estimated parameters $(\widehat{\alpha}, \widehat{\beta})$ as well as those of the estimated factors. Stock and Watson (2002a) show that $(\widehat{\alpha}, \widehat{\beta})$ is consistent for (α, β) and $\widehat{y}_{T+h|T}$ is consistent for $y_{T+h|T}$. To construct confidence intervals, we must obtain the rate of convergence and the limiting distributions of these quantities.

We are specifically interested in the case of large dimensional panels. By a 'large panel', we mean that our theory will allow both N and T to tend to infinity. In practical terms, we assume that that are at least 50 time series observations for each panel unit, and there are 50 or more cross-section units. After discussing the appeal of the diffusion index forecasting methodology in Section 2, Section 3 presents the asymptotic theory and discusses how terms necessary for predictive inference can be constructed. The analysis is extended to non-stationary factors in Section 4. A by-product of the present exercise that is of interest in its own right is estimation of the error covariance matrix when heteroskedasticity and cross-section correlation are of unknown form. This is presented in Section 5. Section 6 presents simulation results to assess the adequacy of the asymptotic approximations in finite sample properties. Empirical applications are considered in Section 7. Proofs are given in the Appendix.

2 Why DI Can Reduce Forecast Error Variance

We first provide some intuition for the appeal of diffusion index forecasts. For ease of exposition, consider the one-step ahead forecast:

$$y_{t+1} = \alpha F_t + \varepsilon_{t+1}$$

where ε_t are iid $(0, \sigma_\varepsilon^2)$. Furthermore, assume that the scalar series F_t is an AR(1) process

$$F_t = \rho F_{t-1} + u_t$$

where u_t are iid $(0, \sigma_u^2)$ and u_s and ε_t are independent for all t and s. Suppose also for the moment that the model parameters are known.

If F_t is observable, the one-step ahead forecast of y_{t+1} at time t is given by αF_t so that the forecast error is ε_{t+1}, and the forecast error variance is σ_ε^2. If F_t is not observable, then y_t is an unobserved components model. The univariate time series forecast is based on the ARMA representation of y_t. In this case, y_t is an ARMA(1,1) process:

$$y_{t+1} = \rho y_t + z_{t+1} + \theta z_t$$

where z_t is a white noise process. Assuming the infinite past history of y_t $(...., y_{t-2}, y_{t-1}, y_t)$ is available, the one-step ahead forecast of y_{t+1} at time t is $\rho y_t + \theta z_t$. The forecast error is z_{t+1} and the forecast error variance is $\sigma_z^2 = E(z_{t+1}^2)$. It can be shown that $\sigma_z^2 > \sigma_\varepsilon^2$, so smaller forecasting error variance is obtained when F_t is observable. This is not surprising and conforms to the intuition that more information permits a better forecast.

The assumption that F_t is observable is of course not realistic. Nevertheless, if we observe a large number of indicators that have F_t as their common sources of variation, we can exploit this commonality to estimate the process F_t very well by the method of principal components (up to a transformation). This is the essence of the diffusion index forecasting. In the limit when N goes to infinity, the DI forecasts are the same as when F_t is observable. In this example, the reduction in forecast error is $\sigma_z^2 - \sigma_\varepsilon^2$, which is strictly positive. In cases with more complex dynamics and/or when W_t are present, knowledge of F_t can still be expected to yield better forecasts, because one can, in general, do no worse with more information.

In practice, the model parameters are also unknown. Parameter uncertainty contributes an $O(1/T)$ term to the forecast error variance. So if F_t was observable, but α is being estimated, it is well known that the variance of $y_{T+1} - \widehat{y}_{T+1|T}$ is $\sigma_\varepsilon^2 + O(T^{-1})$. As will be shown, estimating the factor process F_t will contribute another $O(1/N)$ term to the forecasting error variance. Our contribution is to show that when α and F_t both have to be estimated, the variance of $y_{T+1} - \widehat{y}_{T+1|T}$ is $\sigma_\varepsilon^2 + O(T^{-1}) + O(N^{-1})$. This is less than σ_z^2 when T and N are both large, so one can expect the diffusion index approach to yield better forecasts even when F_t is not observed. The importance of a large N must be stressed, however, because when N is fixed, consistent estimation of the factor process F_t is not possible even if λ_i are all observable. In the event when T and N are both large and are such that T/N goes to zero, we can further show that uncertainty in F_t is dominated by parameter uncertainty so that F_t can be treated as though it is observable. We now turn to the theory underlying these results.

3 Distribution Theory

In matrix notation, the factor model is $X = F\Lambda' + e$, where X is $T \times N$ data matrix, and $F = (F_1, ..., F_T)'$ and $\Lambda = (\lambda_1, ..., \lambda_N)'$, and e is $T \times N$ error matrix. The principal component estimates are denoted $\widetilde{F} = (\widetilde{F}_1, ..., \widetilde{F}_T)'$ and $\widetilde{\Lambda} = (\lambda_1, \ldots, \lambda_N)'$. These are the r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix $XX'/(TN)$ in decreasing order. For future reference, we also let $\widetilde{V}$ be a $r \times r$ diagonal matrix consisting of the r largest eigenvalues of $XX'/(TN)$. We need the following assumptions:

Assumption A: Common factors

1. $E\|F_t\|^4 \leq M$ and $\frac{1}{T}\sum_{t=1}^T F_t F_t' \xrightarrow{p} \Sigma_F$ for a $r \times r$ positive definite matrix Σ_F.

Assumption B: Heterogeneous factor loadings

The loading λ_i is either deterministic such that $\|\lambda_i\| \leq M$ or it is stochastic such that $E\|\lambda_i\|^4 \leq M$. In either case, $\Lambda'\Lambda/N \xrightarrow{p} \Sigma_\Lambda$ as $N \to \infty$ for some $r \times r$ positive definite non-random matrix Σ_Λ.

Assumption C: Time and cross-section weak dependence and heteroskedasticity

1. $E(e_{it}) = 0$, $E|e_{it}|^8 \leq M$;

2. $E(e_{it}e_{js}) = \tau_{ij,ts}$, $|\tau_{ij,ts}| \leq \tau_{ij}$ for all (t,s) and $|\tau_{ij,ts}| \leq \gamma_{ts}$ for all (i,j) such that

$$\frac{1}{N}\sum_{i,j=1}^N \tau_{ij} \leq M, \ \frac{1}{T}\sum_{t,s=1}^T \gamma_{ts} \leq M, \text{ and } \frac{1}{NT}\sum_{i,j,t,s=1} |\tau_{ij,ts}| \leq M$$

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

Assumption D: $\{\lambda_i\}, \{F_t\}$, and $\{e_{it}\}$ are three groups of mutually independent stochastic variables.

Assumption E: Let $z_t = (F_t' \;\; W_t')')$. Then

1. $\frac{1}{T}\sum_{t=1}^{T} z_t z_t' \xrightarrow{p} \Sigma_{zz} = \begin{bmatrix} \Sigma_{FF} & \Sigma_{FW} \\ \Sigma_{WF} & \Sigma_{WW} \end{bmatrix} > 0$;

2. $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t \varepsilon_{t+h} \xrightarrow{d} N(0, \text{plim}\, \frac{1}{T}\sum_{t=1}^{T}(\varepsilon_{t+h}^2 z_t z_t'))$.

Assumptions A and B together imply r common factors. Assumption C allows for limited time series and cross section dependence in the idiosyncratic component. Heteroskedasticity in both the time and cross section dimensions is also allowed. Given Assumption C1, the remaining assumptions in C are easily satisfied if the e_{it} are independent for all i and t. The allowance for weak cross-section correlation in the idiosyncratic components leads to the *approximate factor structure* of Chamberlain and Rothschild (1983). It is more general than a *strict factor model* which assumes e_{it} is uncorrelated across i. Assumption D is standard in factor analysis. Assumption E ensures that the forecasting model is well specified and that the parameters of the model can be identified.

Theorem 1 *(Estimation) Suppose Assumptions A to E hold. Let $\widetilde{F}_t$ be the factor estimates obtained by the method of principal components, and let $\widehat{\alpha}$ and $\widehat{\beta}$ be the least squares estimates from a regression of y_{t+h} on $\widehat{z}_t = (\widetilde{F}_t' \; W_t')'$. Let $H = \widetilde{V}^{-1}(\widetilde{F}'F/T)(\Lambda'\Lambda/N)$. If $\sqrt{T}/N \to 0$,*

$$\sqrt{T}\left(\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} - \begin{bmatrix} H^{-1\prime}\alpha \\ \beta \end{bmatrix} \right) \xrightarrow{d} N\Big(0, D\Big),$$

where

$$D = \text{plim}\left(\frac{1}{T}\sum_{t=1}^{T-k} \widehat{z}_t \widehat{z}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T-k} \widehat{\varepsilon}_{t+h}^2 \widehat{z}_t \widehat{z}_t'\right) \left(\frac{1}{T}\sum_{t=1}^{T-k} \widehat{z}_t \widehat{z}_t'\right)^{-1}. \tag{3}$$

As is well known, the factor model is unidentified because $\alpha' L L^{-1} F_t = \alpha F_t$ for any invertible matrix L. Theorem 1 is a result pertaining to the difference between $\widehat{\alpha}$ and the space spanned by α. Consistency of the parameter estimates follows from the fact that the averaged squared deviations between $\widetilde{F}_t$ and HF_t vanish as N and T both tend to infinity, see Bai and Ng (2002). The consequence of having generated regressors in the forecasting equation has no effect on consistency of the parameter estimates. Letting $\widehat{\delta} = (\widehat{\alpha}' \; \widehat{\beta}')'$, and $\delta = (\alpha' H^{-1} \; \beta')'$, Theorem 1 can be compactly stated as

$$\sqrt{T}(\widehat{\delta} - \delta) \xrightarrow{d} N(0, Avar(\widehat{\delta})).$$

Stock and Watson (2002a) showed consistency of $\widehat{\delta}$ for δ. Here we establish the rate of convergence and the limiting distribution. Asymptotic normality of $\widehat{\delta}$ follows from that fact that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t\varepsilon_{t+h}$ obeys a central limit theorem. Because $\widetilde{F}_t$ is close to F_t, the same asymptotic result holds when z_t is replaced by $\widehat{z}_t$.

Since $Avar(\widehat{\delta})$ is the probability limit of (3), a consistent estimate for $Avar(\widehat{\delta})$ is one of the following:

$$\widehat{Avar(\widehat{\delta})} = \left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t'\right)^{-1}\left[\frac{1}{T}\sum_{t=1}^{T-h}\widehat{\varepsilon}_{t+h}^2\widehat{z}_t\widehat{z}_t'\right]\left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t'\right)^{-1} \tag{4a}$$

$$\widehat{Avar(\widehat{\delta})} = \widehat{\sigma}_\varepsilon^2\left[\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t'\right]^{-1}. \tag{4b}$$

Formula (4a) is the White-Eicker estimate of asymptotic variance and is robust to heteroskedasticity. However, if we assume homoskedasticity so that $E(\varepsilon_{t+h}^2|z_t) = \sigma_\varepsilon^2 \ \forall t$, a consistent estimate of $Avar(\widehat{\delta})$ can be obtained using (4b), where $\widehat{\sigma}_\varepsilon^2 = \frac{1}{T}\sum_{t=1}^{T-h}\widehat{\varepsilon}_{t+h}^2$. As stated, the asymptotic variance is valid when $z_t\varepsilon_{t+h}$ is serially uncorrelated. Extension of (4a) to allow for serial correlation in $z_t\varepsilon_{t+h}$ is straightforward. As shown in Andrews (1991), a heteroskedastic-autocorrelation consistent variance covariance (HAC) matrix that converges to the population covariance matrix can be constructed provided the bandwidth is chosen appropriately. It is noted, however, when ε_t is serially correlated, $y_{T+h|T}$ defined earlier will cease to be the conditional mean, given past information.

Consistency and asymptotic normality of the parameter estimates in the forecasting equation with $\widetilde{F}_t$ replacing the unobservable F_t have important implications for diffusion index forecasting.

Theorem 2 *(Conditional mean forecast) Let $\widehat{y}_{T+h|T} = \widehat{\delta}'\widehat{z}_T$ be the feasible h-step ahead forecast of y_{T+h}. Under the assumptions of Theorem 1,*

$$\frac{\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T})}{B_T}\xrightarrow{d}N(0,1)$$

where

$$B_T^2 = \widehat{z}_T'\,Avar(\widehat{\delta})\,\widehat{z}_T + (T/N)\widehat{\alpha}'\,Avar(\widetilde{F}_T)\,\widehat{\alpha}$$

The least-squares forecast of the conditional mean is root-T consistent and asymptotically normal, provided that T/N is bounded. More precisely, the convergence rate is $\min[\sqrt{N},\sqrt{T}]$ in view of the second term of B_T. The two terms in the forecast error variance follows from the fact that

$$\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T}) = \sqrt{T}(\widehat{\delta} - \delta)'\widehat{z}_T + (T/N)^{1/2}\alpha' H^{-1}\sqrt{N}(\widetilde{F}_T - HF_T).$$

The first component of the forecast error arises from having to estimate α and β, and the second term arises from having to estimate F_t. Now Bai (2003) showed that for each t,

$$\begin{aligned} \sqrt{N}(\widetilde{F}_t - HF_t) &\xrightarrow{d} N\Big(0, V^{-1}Q\Gamma_t Q'V^{-1}\Big) \\ &\equiv N\Big(0, Avar(\widetilde{F}_t)\Big), \end{aligned} \tag{5}$$

where $\widetilde{F}'F/T \xrightarrow{p} Q$, $\widetilde{V} \xrightarrow{p} V$, and $\Gamma_t = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N E(\lambda_i \lambda_j{}' e_{it}e_{jt})$. Thus, the cost of having to estimate F_t is a less efficient forecast. However, this error is negligible when $T/N \to 0$ because $\sqrt{N}(\widetilde{F}_t - HF_t)$ is $O_p(1)$. Intuitively, when N is large, factors F_t can be estimated more precisely so that estimation error can be ignored.

An estimate of $Avar(\widetilde{F}_t)$ (for any given t) can be obtained by first substituting $\widetilde{F}$ for F, and noting that $\widetilde{Q} = \widetilde{F}'\widetilde{F}/T$ is an r-dimensional identity matrix by construction ($\widetilde{Q}$ is an estimate for QH' whose limit is an identity). We can then consider the estimator

$$\widehat{Avar(\widetilde{F}_t)} = \widetilde{V}^{-1}\widetilde{\Gamma}_t\widetilde{V}^{-1},$$

where $\widetilde{\Gamma}_t$, a consistent estimate of $H'^{-1}\Gamma_t H^{-1}$ with $\Gamma_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N E(\lambda_i\lambda_j' e_{it}e_{jt})$, can be one of the following:

$$\widetilde{\Gamma}_t = \frac{1}{N}\sum_{i=1}^N \widetilde{e}_{it}^2 \widetilde{\lambda}_i\widetilde{\lambda}_i' \tag{6a}$$

$$\widetilde{\Gamma}_t = \widetilde{\sigma}_e^2 \frac{1}{N}\sum_{i=1}^N \widetilde{\lambda}_i\widetilde{\lambda}_i' \tag{6b}$$

$$\widetilde{\Gamma}_t = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \widetilde{\lambda}_i\widetilde{\lambda}_j' \frac{1}{T}\sum_{t=1}^T \widetilde{e}_{it}\widetilde{e}_{jt}. \tag{6c}$$

The various specifications of $\widetilde{\Gamma}_t$ accommodate flexible error structures in the factor model. Both (6a) and (6b) assume that e_{it} is cross-sectionally uncorrelated with e_{jt}. The estimator (6b) further assumes $E(e_{it}^2) = \sigma_e^2$ for all i and t. Under regularity conditions, $\widetilde{\sigma}_e^2 = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{e}_{it}^2 \xrightarrow{p} \sigma_e^2$. Although (6a) and (6b) both assume the idiosyncratic errors are cross-sectionally uncorrelated, it is not especially restrictive because much of the cross-correlation in the data is presumably captured by the common factors. At an empirical level, allowing for cross-correlation in the errors would entail estimation of $N(N-1)/2$ additional parameters. Because N is large by assumption, sampling variability could generate non-trivial efficiency loss. For small cross-section correlation in the errors, constraining them to be zero could sometimes be desirable. The estimators defined in (6a) and (6b) are useful even if residual cross-correlation are genuinely present.

When it is deemed inappropriate to assume zero cross-section correlation in the errors, the asymptotic variance of $\widetilde{F}_t$ can be estimated by (6c). Consistency of $\widetilde{V}^{-1}\widetilde{\Gamma}_t\widetilde{V}_t^{-1}$ for $V^{-1}Q\Gamma_t Q'V^{-1}$ is established in Bai (2003) when $\widetilde{\Gamma}_t$ is defined as in (6a) and (6b). However, the estimator $\widetilde{\Gamma}_t$ defined in (6c) is new. Consistency of $\widetilde{\Gamma}_t$ will be shown below under a more general set up and in a broader context. Suffice it to note for now that the estimator, which we will refer to as CS-HAC, is robust to cross-section correlation and heteroskedasticity in e_{it}, but requires $E(e_{it}e_{jt}) = \sigma_{ij}$ for all t. Loosely speaking, covariance stationarity of e_{it} implies that Γ_t does not depend on t so that residuals from other periods, not just t, can be used to estimate Γ_t. A law of large number can be then invoked. In particular, if $\zeta_t = \frac{1}{\sqrt{N}}\sum_{i=1}^N \lambda_i e_{it}$, then $\frac{1}{T}\sum_{t=1}^T \zeta_t^2 \xrightarrow{p} E(\zeta_t^2) = \Gamma$, validating consistent estimation of Γ in the presence of residual correlation cross-sectionally by (6c) .

As in standard regression models, the forecast error variance in the present setting decreases at rate T. But whereas for fixed T, the forecast error variance increases with the number of predictors through a loss in degrees of freedom, the efficiency of the diffusion index forecast improves with the number of predictors. This is because a large N enables more precise estimation of the common factors and thus results in more efficient forecasts. This property of the factor estimates is also in sharp contrast to that obtained in standard factor analysis that assumes a fixed N. With the sample size is fixed in one dimension, consistent estimation of the factor space is not possible however large T becomes.

Once appropriate estimators for $Avar(\widehat{\delta})$ and $Avar(\widetilde{F}_T)$ are chosen, the above results allow us to construct prediction intervals. This exercise is straightforward given asymptotic normality of the forecasts errors. For example, the 95% interval for the $y_{T+h|T}$ is

$$\left(\widehat{y}_{T+h|T} - \frac{1.96}{\sqrt{T}}\sqrt{\widehat{Avar}(\widehat{y}_{T+h|T})}, \quad \widehat{y}_{T+h|T} + \frac{1.96}{\sqrt{T}}\sqrt{\widehat{Avar}(\widehat{y}_{T+h|T})}\right),$$

where $\widehat{Avar}(\widehat{y}_{T+h|T})$ is equal to B_T^2, as defined in Theorem 2.

Although the conditional mean is a useful benchmark for the theoretical properties of forecasts, it is not observable. Thus, in practice, forecast comparisons are inevitably made in terms of y_{T+h}. Since $y_{T+h} = y_{T+h|T} + \varepsilon_{T+h}$, it follows that

$$\widehat{y}_{T+h|T} - y_{T+h} = (\widehat{y}_{T+h|T} - y_{T+h|T}) + \varepsilon_{T+h}.$$

In view of Theorem 2, the variance of $\widehat{y}_{T+h|T} - y_{T+h}$ is approximately equal to $\sigma_\varepsilon^2 + \frac{1}{T} Avar(\widehat{y}_{T+h|T})$. So if ε_t is normally distributed, then $\widehat{y}_{T+h|T} - y_{T+h}$ is also approximately normal and

$$\widehat{y}_{T+h|T} - y_{T+h} \sim N\Big(0, \sigma_\varepsilon^2 + \frac{1}{T} Avar(\widehat{y}_{T+h|T})\Big).$$

Clearly, the variance is dominated by the variance of ε_{T+h}. This result, which is standard in the forecasting literature, is preserved when the factors are estimated. It should, however, be stressed that the error arising from using $\widetilde{F}_t$ is asymptotically negligible only if Theorem 2 holds. It is thus essential that N and T are both large.

The above results will be useful in rather broader contexts, as having to conduct inference when the latent common factors are replaced by generated regressors is not uncommon. For example, Bernanke and Boivin (2002) considered factor-augmented vector autoregression (FVAR), which is nothing more than including the principal components estimates of the factors to an otherwise standard VAR. Because the factors are unobservable, impulse response functions and decomposition of variances should be adjusted by the error that arises from the unobservability of F_t. Our results enable such calculations. Observed variables are often used in place of the latent factors when testing various theories of asset returns. In Bai and Ng (2003), we develop tests to determine whether the observables are good proxies for the latent factors. That analysis, which amounts to assessing the in-sample forecasting ability of the latent factors, makes use of the results presented here, with h set to zero.

4 Nonstationary factors

The preceding analysis can be extended to nonstationary factors. Although nonstationary factors imply different rates of convergence for the estimated model parameters, we will show that for the purpose of constructing confidence intervals, the formula for stationary factors remains valid. This means that as a practical matter, it is not necessary to know the presence of I(1) factors when making predictive inference. The following analysis verifies these claims.

Assuming again that the forecasting equation is $y_{t+h} = \alpha' F_t + \beta' W_t + \varepsilon_{t+h}$, and the panel of data have a factor representation $x_{it} = \lambda_i' F_t + e_{it}$, but we now assume

$$F_t = F_{t-1} + u_t,$$

where u_t is a sequence of I(0) processes. To analyze this case of non-stationary factors, all previous assumptions are maintained, except for the following:

Assumption A': (1) $E\|u_t\|^{4+\delta} \leq M$ and $\frac{1}{T^2}\sum_{t=1}^{T} F_t F_t' \xrightarrow{d} \Sigma_F$, where Σ_F is positive definite (random) matrix with probability 1, and (2) ε_t is an iid sequence with zero mean and variance σ_ε^2, where ε_s is independent of $z_t = (F_t', W_t')'$ for all t and s.

Assumption A'(1) rules out cointegration among the components of F_t, although the results are applicable for this case. Cointegration among F_t is equivalent to the existence of both I(1) and I(0) factors, see Bai (2004). This case would require more complicated notation and will not be presented to simplify the exposition.

Assumption A'(2) imposes conditional homoskedasticity on ε_t. As a result, the following mixture normality is a reasonable assumption:

$$D_T^{-1}\sum_{t=1}^{T} z_t \varepsilon_{t+h} \xrightarrow{d} MN(0, \sigma_\varepsilon^2 \Omega) \tag{7}$$

where $MN(0, \sigma_\varepsilon^2\Omega)$ is shorthand notation for conditional normal distribution with covariance matrix $\sigma_\varepsilon^2\Omega$, conditional on Ω, where Ω is the limiting random matrix of $D_T^{-1}z'zD_T^{-1}$ where $D_T = TI_{r+p}$ if W_t is also I(1), and $D_T = (TI_r, \sqrt{T}I_p)$ if W_t is I(0). If some components of W_t are I(1), and others are I(0), D_T is adjusted accordingly. By definition, if $\xi \sim MN(0, \sigma_\varepsilon^2\Omega)$, then $\sigma_\varepsilon^{-1}\Omega^{-1/2}\xi \sim N(0, I)$.

Let $\widetilde{F}$ be a $T \times r$ matrix consisting of r eigenvectors (multiplied by T) of the matrix $XX'/(T^2N)$, corresponding to the first r largest eigenvalues (in deceasing order). Let $\widetilde{V}$ be the diagonal matrix consisting of these eigenvalues. Define $\widetilde{\Lambda} = X'\widetilde{F}/T^2$ and $H = \widetilde{V}^{-1}(\widetilde{F}'F/T^2)(\Lambda'\Lambda/N)$.

Theorem 3 *Suppose assumptions A', B-E and (7) hold. Let $\widehat{\alpha}$ and $\widehat{\beta}$ be the least squares estimators from a regression of y_{t+h} on $\widehat{z}_t = (\widetilde{F}_t'\ W_t')'$. If $\sqrt{T}/N \to 0$,*

$$(D_T^{-1}\widehat{z}'\widehat{z}D_T^{-1})^{1/2} D_T \left(\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} - \begin{bmatrix} H'^{-1}\alpha \\ \beta \end{bmatrix} \right) \xrightarrow{d} N\left(0, \sigma_\varepsilon^2 I\right) \tag{8}$$

where $\widehat{z} = (\widehat{z}_1, ..., \widehat{z}_{T-h})'$.

The theorem shows that $\widehat{\alpha}$ converges to $H'^{-1}\alpha$ at rate T and $\widehat{\beta}$ converges to β at rate $\sqrt{T}$ when W_t is I(0). These are the same rates as known F. Of course, for known F, we will directly estimate α instead of $H'^{-1}\alpha$. When the estimator is weighted by the random matrix $(D_T^{-1}\widehat{z}'\widehat{z}D_T^{-1})^{1/2}$, the limiting matrix is normal. The unweighted limiting distribution is mixture normal.

Theorem 4 *(Conditional mean forecast) Let $\widehat{y}_{T+h|T} = \widehat{\alpha}'\widetilde{F}_T + \widehat{\beta}'W_T$ be the feasible h-step ahead forecast of y_{T+h}. Under the assumptions of Theorem 3,*

$$\frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{C_T} \xrightarrow{d} N(0,1) \tag{9}$$

where

$$C_T^2 = \widehat{\sigma}_\varepsilon^2 \widehat{z}_T' (\widehat{z}'\widehat{z})^{-1}\widehat{z}_T + (1/N)\widehat{\alpha}' \widetilde{V}^{-1}\widetilde{\Gamma}_t\widetilde{V}^{-1}\widehat{\alpha}$$

The forecast error variance once again has two components. The first term of C_T^2 comes from the estimation of δ and is $O_p(T^{-1})$. The second term comes from the estimation of F_t and is $O_p(N^{-1})$. The overall convergence rate of $\widehat{y}_{T+h|T}$ to $y_{T+h|T}$ is $\min[\sqrt{N}, \sqrt{T}]$, as in the case of I(0) regressors.

It is known that the triple $(\widetilde{V}, \widetilde{F}, \widetilde{\Lambda})$ has to be normalized differently, depending on whether F_t is I(1) or I(0),[2]. One would then expect confidence intervals for stationary and non-stationary factors will be constructed differently. However, the expressions in Theorems 2 and Theorem 4 are in fact mathematically identical, implying that knowledge concerning the stationarity property of F_t is not essential for the construction of confidence intervals. This is because the number of times $\widetilde{F}_t$ appears in the numerator in (9) is the same as in denominator. The scalings thus offset. Indeed, observe that C_T^2 in Theorem 4 equals B_T^2/T in Theorem 2 when homoskedsaticity of ε_t is assumed. Correspondingly, the numerator of Theorem 2 is scaled, but that of Theorem 4 is not. Theorem 4 is stated under the assumption of conditional homoskedasticity, but it serves to highlight its main implication, namely, that the formulas for constructing forecasting confidence intervals derived for stationary common factors are also valid for nonstationary factors.

5 Covariance Matrix Estimator: the CS-HAC

In presenting the prediction intervals proposed above, we introduced the CS-HAC (cross-section correlation and heteroskedasticity robust) covariance matrix estimator. To motivate this estimator, consider the cross-section regression (with z_i observed):

$$y_i = z_i'\beta + \varepsilon_i \qquad i = 1, 2, ..., N,$$

where $\text{cov}(\varepsilon_i) = \Omega$. It is well known that the asymptotic variance of the least-squares estimator, $\widehat{\beta}$, is

$$Avar(\widehat{\beta}) = \lim_{N\to\infty} \left(\frac{z'z}{N}\right)^{-1} \left(\frac{z'\Omega z}{N}\right) \left(\frac{z'z}{N}\right)^{-1} \equiv \lim_{N\to\infty} S_{zz}^{-1}\Gamma S_{zz}^{-1}.$$

[2]Different scalings are used to derive proper rates of convergence and suitable limiting distributions.

For inference, the S_{zz} term poses no problem when N is large since $z_i, i = 1, \ldots N$ is observed. The middle term is more problematic because Ω is unobserved.

If e_i is uncorrelated with e_j, Ω is a diagonal matrix with diagonal elements $(\sigma_{11}, \sigma_{22}, ..., \sigma_{NN})$. Then the estimator

$$\widehat{\Gamma} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\varepsilon_i}^2 z_i z_i'$$

has the property that $\widehat{\Gamma} - \Gamma \xrightarrow{p} 0$, as shown in White (1980). Inference then follows textbook treatment.

What if Ω is not a diagonal matrix? It is attempting to replace σ_{ij} in the following expression

$$\Gamma = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} z_i z_j'$$

by $\widehat{\varepsilon_i}\widehat{\varepsilon_j}$ because $E\varepsilon_i\varepsilon_j = \sigma_{ij}$. Such an estimator is well known to be inconsistent, even if ε_i was observable. The reason is that

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \varepsilon_i \varepsilon_j z_i z_j' = \Big(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_i \varepsilon_i \Big) \Big(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_i \varepsilon_i \Big)'$$

converges to a random variable rather than a positive definite matrix.

The problem with estimating the error covariance matrix in cross-section regressions with heteroskedasticity and cross-section correlation of unknown form is akin to the problem of estimating the spectrum at the zero frequency in time series regressions. In a time series context, the natural estimator of summing T sample covariances also does not yield a consistent estimate of the population spectrum. The solution to this is well known. Under the assumption of weak dependence, σ_{ij} is negligible if $|i - j|$ is large. A consistent estimator, often referred to as the the HAC estimator, can be constructed with suitable choice of the truncation lag, see Newey and West (1987).

Using $|i-j|$ as a metric to judge when σ_{ij} is negligible is possible in time series regressions because the data are ordered in time. But unlike time series, a natural ordering of cross-section data arises as an exception rather than the rule. If the data can be ordered according to some economic distance, as in Conley (1999)), then consistent estimation of covariance is possible.

As a general matter, neither economic theory nor intuition can be expected to be of much help in obtaining a 'mixing condition' type ordering of the data. Since any permutation of the data is is an equally valid representation of information available, the different orderings also cannot be ranked. Instead of truncating terms 'far from' an observation, the common practice is to impose restrictions on the off-diagonal elements, or to parameterize Ω in terms of a finite number of parameters. Both approaches serve the purpose of reducing the number of unknowns in Ω from $O(N^2)$ to something more manageable.

We propose a third alternative. Instead of requiring economic theory to provide an ordering of the data, we require that repeated observations be available over time. The basic intuition is as follows. Under covariance stationarity, the time series observations allow us to consistently estimate the cross-section correlation matrix. Thus, although the cross-section regressions do not permit consistent estimation of the covariance matrix of interest, this is possible with estimation using large dimensional panels. We now formalize this intuition.

Theorem 5 *Consider the panel data regression model $y_{it} = z_{it}'\beta + \varepsilon_{it}$ for $i = 1, \dots N$ and $t = 1, \dots T$. Assuming ε_{it} are cross-sectionally weakly correlated as in Assumption C with $E(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij}$. In addition, ε_{it} is serially uncorrelated for every i. Then the least squares estimator of β is $\sqrt{NT}$ consistent and the asymptotic variance is the probability limit of $A_{NT}^{-1}\Gamma_{NT}A_{NT}^{-1}$, where $A_{NT} = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T z_{it}z_{it}'$, and*

$$\Gamma_{NT} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N \sigma_{ij}\frac{1}{T}\sum_{t=1}^T z_{it}z_{jt}'.$$

Consider the estimator

$$\widehat{\Gamma}_{NT} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N \Big(\frac{1}{T}\sum_{t=1}^T \widehat{\varepsilon}_{it}\widehat{\varepsilon}_{jt}\Big)\Big(\frac{1}{T}\sum_{t=1}^T z_{it}z_{jt}'\Big). \tag{10}$$

Then for arbitrary $N > 1$ and possibly $N >> T$, $\|\widehat{\Gamma}_{NT} - \Gamma_{NT}\| \xrightarrow{p} 0$ as $T \to \infty$.

The estimation Γ_t defined in (6c) of Section 3 is based on the estimator considered here. Assumption C is sufficient to guarantee $\sqrt{NT}$ consistency of β. It thus requires the cross-section correlation to be 'weak'. The notion of weak cross-section correlation, as defined in Chamberlain and Rothschild (1983) put bounds on the eigenvalues of Ω. Assumption C restates the condition in terms of the column sum of a matrix. By 'weak', we mean $\frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N |\sigma_{ij}| \le M$.

A similar problem was considered in Driscoll and Kraay (1998), though these authors did not assume weak cross section correlation. They are interested in estimating the error covariance matrix when the estimated parameters are $\sqrt{T}$ consistent. However, without imposing some discipline on the extent of the cross correlation, it is not clear that $\sqrt{T}$ consistency is enough to yield consistent estimation of the error covariance matrix. Perhaps for this reason, they provide results for Γ_{NT}/N when $N = N(T)$, which is arguably an unusual setup. We assume weak-cross section correlation, and can state precisely the conditions under which $||\widehat{\Gamma}_{NT} - \Gamma_{NT}||$ goes to zero.

The above result is stated for time varying regressors, and continue to hold, naturally, when the regressors do not vary over time. For example, if the regression is $y_{it} = z_i'\beta + \varepsilon_{it}$, then

$$\widehat{\Gamma}_{NT} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{jt}\right)z_i z_j'$$

it is easy to see that $\widehat{\Gamma}_{NT} - \Gamma_{NT} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{jt} - \sigma_{ij}\right)z_i z_j' \xrightarrow{p} 0$.

If we relax stationarity and no serial correlation to allow time series heteroskedasticity and series correlation, then Γ_{NT} is given by

$$\Gamma_{NT} = \frac{1}{TN}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E(z_{it}z_{js}'\varepsilon_{it}\varepsilon_{js})$$

Now consider a Newey-West type estimator. Define

$$\xi_{N,t} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}z_{it}\varepsilon_{it}, \quad \widehat{\xi}_{N,t} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}z_{it}\widehat{\varepsilon}_{it}.$$

$$\widehat{\Gamma}_{NT} = \widehat{D}_0 + \sum_{h=1}^{m}(1 - \frac{h}{m+1})(\widehat{D}_h + \widehat{D}_h')$$

where

$$\widehat{D}_h = \frac{1}{T}\sum_{t=h+1}^{T}\widehat{\xi}_{Nt}\widehat{\xi}_{N,t-h}', \quad h = 1, 2, ..., m$$

Theorem 6 *Assume that $m \to \infty$, as $T \to \infty$ with $m^3/T \to 0$. Then under Assumption C for ε_{it}, we have for arbitrary $N \geq 1$,*

$$||\widehat{\Gamma}_{NT} - \Gamma_{NT}|| \xrightarrow{d} 0.$$

as $T \to \infty$. The result holds for $N >> T$.

6 Finite Sample Properties

We now use simulations to assess the finite sample properties of the procedures. A panel of data is generated as follows:

$$\begin{aligned} x_{it} &= \lambda_i' F_t + e_{it}, \quad i = 1, \dots N, t = 1, \dots, T \\ F_{jt} &= \rho_j F_{jt-1} + \sqrt{1-\rho_j^2} u_{jt} \quad j = 1, \dots, r \\ e_{it} &\sim (1+b^2) v_{it} + b v_{i+1,t} + b v_{i-1,t}. \\ \rho_j &= (.8)^j, \end{aligned}$$

where u_{jt} and v_{it} are mutually uncorrelated $N(0,1)$ random variables. We draw λ_i once from the standard normal distribution, and it does not change with N or T. In the simulations, we set $r = 2$ and assume that it is known. The series to be forecasted is

$$y_{t+h} = 1 + F_{1t} + F_{2t} + \varepsilon_{t+h}.$$

That is, $W_t = 1 \; \forall t$, α is the unit vector, and β equals 1. The simulation design follows Stock and Watson (2002a) closely. Configurations that include additional W_t series yield similar results and will not be presented.

Our main interest is in the coverage of the confidence intervals. Three types of confidence intervals will be presented:

Model (A): (6b) +(4b) ; Model (B): (6a) + (4a) ; Model (C): (6c) + (4a).

For each model, the coverage rates are reported for (i) the diffusion index forecast for the conditional mean, $\widehat{y}_{T+h|T}$; (ii) the infeasible forecast of the conditional mean $\widehat{y}^0_{T+h|T}$; (iii) the diffusion index forecast for y_{T+h}, and (iv) the infeasible forecast y^0_{T+h}. By infeasible forecast, we mean that F_t is used, and estimation of the factors is not necessary. A comparison of the feasible and infeasible forecasts gives an indication of the error arising from the estimation of F_t.

The results are presented in Tables 1, 2, and 3 respectively. The top panel are coverage rates when the forecasting model is correctly specified (in terms of the number of factors). Overall, the coverage rates are excellent. The probability that $y_{T+h|T}$ or y_{T+h} lies inside the estimated prediction intervals is always close to the nominal coverage rate of .95, even when N and T are only 50.

The idiosyncratic errors are cross-sectionally uncorrelated when $b = 0$, in which case all three estimators of $Avar(\widetilde{F}_t)$ are valid. Although (6c) should be less efficient, comparing the results in Table 1 and 2 with those in 3 reveal that estimating the cross-section correlation when none is present seems to have little effect on coverage. In the simulations, the errors are homoskedastic by design. The results using the heteroskedastic robust estimator in Tables 2 and 3 are also similar to those in Table 1 with homoskedasticity imposed.

When $b \neq 0$, use of (6c) is appropriate. Omitting cross-section correlation tends to weaken coverage marginally. This should not be taken as indication that cross-section correlation in the errors does not need to be dealt with. In situations when the cross-correlation is more prevalent, the effect will be amplified.

The bottom panel of Tables 1 to 3 consider situations when too few factors are used. In these cases, the coverage for the conditional mean is well below .95. This problem is not specific to diffusion index forecasting, however, as inference cannot be expected to be correct when the object of interest is misspecified. Nonetheless, the coverage for y_{T+h} remains accurate because the misspecification in the conditional mean leads to a correspondingly larger unconditional prediction error variance. Inference on y_{T+h} is not significantly affected by whether the error comes from the conditional mean, or from the residual component.

7 Empirical Application

Although diffusion index forecasts have been found to yield improvements over simple models, a major shortcoming is that only point forecasts are available. There exists no tools to assess uncertainty around the forecasts. With the distribution of the forecast errors presented in the previous section, it is now possible to compute prediction intervals.

To illustrate, we use as predictors the 150 series as in Stock and Watson (2002b).[3] We consider $h = 12$ period ahead forecast of the annual growth rate of industrial production, DIP, and inflation, DP. Thus, y_{t+12} is either $DIP = \log(IP_{t+12}) - \log(IP_t)$, or $DP = \log(PUNEW_{t+12}) - \log(PUNEW_t)$. For W_t, we include lags of the monthly first difference of the series, plus a constant. The forecasting exercise begins by estimating the factors using data on x_{it} from 1959:1 to 1969:1. We then obtain $\widehat{\alpha}$ and $\widehat{\beta}$ from a regression of y_t on $\widetilde{F}_{t-12}$ and W_{t-12}, for t=1959:1 to 1969:1. The forecast for T=1970:1 is computed as $\widehat{\alpha}'\widetilde{F}_{1969:1} + \widehat{\beta}'W_{1969:1}$. The sample is then extended by one month, the factors and all the parameters are re-estimated, and the forecast for $y_{1970:2}$ is formed. The procedure is repeated until the forecast for 1996:12 is made in 1995:12.

For the sake of comparison, we also consider the autoregressive forecast $\widehat{\beta}'W_{1969:1}$. We

[3]The data are taken from Mark Watson's web site http://www.princeton.edu/ mwatson.

first select the order of this autoregression using the BIC. The diffusion index model then augments this autoregression with the estimated factors. If the factors have no useful information, α should be zero, and the autoregressive forecast will be the optimal forecast.

Because the two series to be forecasted are one of the x_{it}s, the number of factors in y_t is the same as the number of common factors in the panel of data. This is determined using $\widehat{r} = \text{argmax}_{k=0,\ldots kmax} IC_P(k)$ where

$$ICP(k) = \log \widehat{\sigma}^2(k) + k \cdot g(N,T),$$

where $\widehat{\sigma}^2(k) = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{e}_{it}^2$. In Bai and Ng (2002), we showed that any penalty satisfying $g(N,T) \to 0$ and $\min[N,T]g(N,T) \to \infty$ is theoretically valid. Stock and Watson (2002b) used $g_1(N,T) = \frac{\log(\min[N,T])}{\min[T,T]}$. This penalty tends to favor a larger number of factors than $g_2(N,T) = (N+T)\frac{\log(NT)}{NT}$, an equally valid penalty except in the unusual case that $N = \exp(T)$. Obviously, the larger the number of factors, the less likely will the errors be cross-sectionally correlated. Thus, we consider two sets of confidence intervals. Configuration A uses $g_1(N,T)$ with $Avar(\widetilde{F}_t)$ specified by (6a). Configuration B uses $g_2(N,T)$ with $Avar(\widetilde{F}_t)$ specified by (6c). In both cases, (4a) is used for $Avar(\widehat{\delta})$. As it turns out, the results are quite similar, with results for configuration B slightly better. We will only report results for configuration B. It uses a smaller number of estimated factors, but correct for cross-section correlation in the idiosyncratic errors.

Industrial Production Figure 1a presents the autoregressive (AR) and the diffusion index forecasts for industrial production. Because DIP is only mildly serially correlated, the AR forecast (broken line) is roughly constant. The diffusion index forecast (dotted line) is more volatile, but tracks the actual data more closely. The average mean-squared error for the diffusion index and AR forecasts are 24.95 and 26.46, respectively. Figures 2a and 2b present the series to be forecasted, along with the 95% prediction interval as suggested by the diffusion index and the AR forecasts, respectively. The mean length of confidence intervals is 17.17 for the diffusion index model, and is 20.48 for the AR model. This agrees with the visual impression that the confidence intervals are narrower when the factors are used. The probability that the data lies within the prediction interval is .895 for the diffusion index forecast, and .935 for the AR forecast. The diffusion index model thus gives tighter forecasts, but has a few more misses.

Inflation The inflation forecasts are presented in Figures 3. As inflation displays stronger persistence, the AR forecast mirrors lagged inflation. The factors add information beyond what is in lagged inflation, reducing the MSE from 5.09 to 3.98. The data along with the 95% prediction interval are given in Figure 4. The prediction interval for the diffusion index forecasts are again tighter, with an average length of 5.19 compared to 7.41. However, the probability that the actual data point falls inside the prediction interval is only .746, compared with .848 for the AR forecast. This accords with Stock and Watson's observation that the diffusion index forecasting approach is less successful in forecasting inflation.

The fact that the data lie inside the AR prediction intervals with higher probability may appear to be a discouraging result at first glance. But such a result is a consequence that there is more uncertainty around the AR forecasts. Higher coverage arising from wider prediction intervals cannot be interpreted as evidence in favor of the AR forecasts, as with wide enough prediction intervals, the data must fall within the prescribed range.

The tools provided in this analysis provide a more complete picture of the ability of diffusion index forecasts. In the two applications considered, a notable aspect of diffusion index forecasts is the reduced adherence to lagged dynamics, even when the autoregressive structure is built in. Diffusion index forecasts are by no means 'black box forecasts'. Its attraction lies in its ability to incorporate information from a large number of predictors and not already in the lagged data in a parsimonious way.

8 Conclusion

The factor approach to forecasting is extremely useful in situations when a large number of indicator or predictor variables are present. The factors provides a significant reduction in the number of variables entering the forecasting equation while exploiting information in all available data. This latter aspect is important because it is by using information in all data available that permits consistent estimation of the factors. This paper contributes to the small but growing literature on factor forecasting by (i) showing that the conditional mean forecasts are $\sqrt{T}$ consistent, and (ii) presenting formulas for the forecast errors to permit predictive inference. As a by product, we suggest how the covariance matrix of cross-correlated errors can be consistently estimated. Our estimator is an extension of the heteroskedascity-consistent covariance matrix of White (1980) developed for the case of cross-sectionally uncorrelated errors.

Proof of Theorem 1

The forecasting model when F_t is observed can be written as:

$$\begin{aligned} y_{t+h} &= \alpha' F_t + \beta' W_t + \varepsilon_{t+h} \\ &= \alpha' H^{-1} \widetilde{F}_t + \beta' W_t + \varepsilon_{t+h} + \alpha' H^{-1}(HF_t - \widetilde{F}_t). \end{aligned}$$

This implies, for $Y = (y_h, y_{h+1}, ..., y_T)'$, $\varepsilon = (\varepsilon_h, ..., \varepsilon_T)'$, and $\widehat{z} = (\widehat{z}_1, ..., \widehat{z}_{T-h})'$,

$$Y = \widehat{z} \begin{bmatrix} H^{-1\prime}\alpha \\ \beta \end{bmatrix} + \varepsilon + (FH' - \widetilde{F})H^{-1\prime}\alpha.$$

Consider the regression $y_{t+h} = \alpha' \widetilde{F}_t + \beta' W_t + \varepsilon_{t+h}$. The least squares estimates are

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} = (\widehat{z}'\widehat{z})^{-1}\widehat{z}'Y,$$

and so

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} - \begin{bmatrix} H^{-1\prime}\alpha \\ \beta \end{bmatrix} = (\widehat{z}'\widehat{z})^{-1}\widehat{z}'\varepsilon + (\widehat{z}'\widehat{z})^{-1}\widehat{z}'(FH' - \widetilde{F})H^{-1}\alpha.$$

The second term is $o_p(T^{-1/2})$. This follows from $(\widehat{z}'\widehat{z}/T)^{-1} = O_p(1)$ and $T^{-1/2}\widehat{z}'(FH' - \widetilde{F}) = o_p(1)$ if $\sqrt{T}/N \to 0$, by Lemma 1 below. Consider the first term.

$$\frac{\widehat{z}'\varepsilon}{\sqrt{T}} = \begin{bmatrix} \frac{\widetilde{F}'\varepsilon}{\sqrt{T}} \\ \frac{W'\varepsilon}{\sqrt{T}} \end{bmatrix} = \begin{bmatrix} \frac{(\widetilde{F} - HF')\varepsilon}{\sqrt{T}} + \frac{HF'\varepsilon}{\sqrt{T}} \\ \frac{W'\varepsilon}{\sqrt{T}} \end{bmatrix}.$$

By Lemma 1 (part iii) below, $\frac{(\widetilde{F} - FH')'\varepsilon}{\sqrt{T}} \xrightarrow{p} 0$ if $\sqrt{T}/N \to 0$. Therefore,

$$\begin{aligned} \sqrt{T}\left(\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} - \begin{bmatrix} H^{-1\prime}\alpha \\ \beta \end{bmatrix}\right) &= \left(\frac{\widehat{z}'\widehat{z}}{T}\right)^{-1} \begin{bmatrix} \frac{HF'\varepsilon}{\sqrt{T}} \\ \frac{W'\varepsilon}{\sqrt{T}} \end{bmatrix} + o_p(1) \\ &= \left(\frac{\widehat{z}'\widehat{z}}{T}\right)^{-1} \begin{bmatrix} H & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} F'\varepsilon \\ W'\varepsilon \end{bmatrix} \frac{1}{\sqrt{T}} + o_p(1) \\ &= \left(\frac{\widehat{z}'\widehat{z}}{T}\right)^{-1} \begin{bmatrix} H & 0 \\ 0 & I \end{bmatrix} z'\varepsilon/\sqrt{T} + o_p(1). \end{aligned}$$

From $z'\varepsilon/\sqrt{T} \xrightarrow{d} N(0, \text{plim}\, \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t+h}^2 z_t z_t')$ by Assumption D2, the above is asymptotically normal. The asymptotic variance matrix is the probability limit of

$$\left(\frac{\widehat{z}'\widehat{z}}{T}\right)^{-1} \begin{bmatrix} H & 0 \\ 0 & I \end{bmatrix} \left(\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t+h}^2 z_t z_t'\right) \begin{bmatrix} H' & 0 \\ 0 & I \end{bmatrix} \left(\frac{\widehat{z}'\widehat{z}}{T}\right)^{-1}$$

From $HF_t = \widetilde{F}_t + o_p(1)$, the product of the middle three matrices is equal to $(\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t+h}^2 \widehat{z}_t \widehat{z}_t') + o_p(1)$, proving Theorem 1.

Lemma 1 *Under Assumptions A-E, (i)* $\frac{\widehat{z}'\widehat{z}}{T} = O_p(1)$*, (ii)* $\frac{\widehat{z}'(FH'-\widetilde{F})}{T} = O_p(\min^{-1}[N,T])$*, and (iii)* $\frac{(\widetilde{F}-FH)'\varepsilon}{T} = O_p(\min^{-1}[N,T])$.

Proof: to be added.

Proof of Theorem 2

Begin by rewriting

$$\begin{aligned}\widehat{y}_{T+h|T} - y_{T+h|T} &= \widehat{\alpha}'\widetilde{F}_T + \widehat{\beta}'W_T - \alpha' F_t - \beta' W_T \\ &= (\widehat{\alpha} - H^{-1\prime}\alpha)'\widetilde{F}_T + \alpha' H^{-1}(\widetilde{F}_T - HF_T) + (\widehat{\beta}-\beta)W_T.\end{aligned}$$

Multiplying by $\sqrt{T}$,

$$\begin{aligned}\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T}) &= \sqrt{T}\begin{bmatrix}\widehat{\alpha} - H^{-1\prime}\alpha \\ \widehat{\beta}-\beta\end{bmatrix}'\begin{bmatrix}\widetilde{F}_T \\ W_T\end{bmatrix} + \alpha' H^{-1}\frac{\sqrt{T}}{\sqrt{N}}\sqrt{N}(\widetilde{F}_T - HF_T) \\ &= \widehat{z}_T'\sqrt{T}(\widehat{\delta}-\delta) + (T/N)^{1/2}\alpha' H^{-1}\sqrt{N}(\widetilde{F}_T - HF_T)\end{aligned}$$

Thus, if T/N is bounded, $\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T}) = O_p(1)$ and is asymptotically normal because $\sqrt{T}(\widehat{\delta}-\delta)$ and $\sqrt{N}(\widehat{F}_T - HF_T)$ are asymptotically normal. Furthermore, the two distributions are asymptotically independent because the limiting distribution of $\sqrt{T}(\widehat{\delta}-\delta)$ is determined by $(\varepsilon_1, ..., \varepsilon_T)$ and the asymptotical distribution of $\sqrt{N}(\widetilde{F}_T - HF_T)$ is determined by cross-section disturbances at period T, e_{iT} for $i = 1, 2..., N$. Let $B_T^2 = \widehat{z}_T' Avar(\widehat{\delta})\widehat{z}_T + (T/N)\widehat{\alpha}' Avar(\widetilde{F}_T)\widehat{\alpha}$, then $\sqrt{T}(\widehat{y}_{T+h|T} - y_{T+h|T})/B_T \xrightarrow{d} N(0,1)$, proving Theorem 2.

Proof of Theorem 4 The expression C_T^2 is equal to B_T^2/T when (4b) is used in estimating $Avar(\widehat{\delta})$ of Theorem 2. This is because, unlike Theorem 2, the scaling factor $\sqrt{T}$ is not used in the numerator of (9). Therefore, Theorem 2 and Theorem 4 are identical in mathematical expressions. Nevertheless, the triple $(\widetilde{V}, \widetilde{F}, \widetilde{\Lambda})$ in Theorem 4 are estimated (or are scaled) differently, depending on whether F_t is I(1) or I(0).[4] It might appear that it is essential to know the stationarity property of F_t. It turns out that C_T^2 is invariant to different scalings. First consider the first term of C_T^2, which is $\widehat{z}_T'(\widehat{z}'\widehat{z})^{-1}\widehat{z}_T$. From $\widehat{z}_t = (\widetilde{F}_t', W_t')'$, it is clear that $\widetilde{F}_t$ appears twice in the numerator and twice in the denominator, thus immune to scaling. Next consider $\widehat{\alpha}'\widetilde{V}^{-1}\widetilde{\Gamma}_t\widetilde{V}^{-1}\widehat{\alpha}$. Given a data matrix X, let $(\widetilde{V}^s, \widetilde{F}^s, \widetilde{\Lambda}^s)$ be the estimated

[4]Different scalings are used to derive proper rates of convergence and suitable limiting distributions.

triple assuming F_t to be I(0), and let $(\widetilde{V}^n, \widetilde{F}^n, \widetilde{\Lambda}^n)$ be the corresponding triple assuming F_t to be I(1). Then $(\widetilde{V}^n, \widetilde{F}^n, \widetilde{\Lambda}^n) = (\widetilde{V}^s/T, \sqrt{T}\widetilde{F}^s, \widetilde{\Lambda}^s/\sqrt{T})$, by the definition of the estimation procedures. This implies that $\widehat{\alpha}^n = \widehat{\alpha}^s/\sqrt{T}$ (note $\widehat{\alpha}^n$ is the estimated regression coefficient when $\widetilde{F}^n$ is the regressor, and likewise for $\widehat{\alpha}^s$). Furthermore, the panel residuals $\widetilde{e}_{it}$ are invariant to scalings because $\widetilde{F}^n\widetilde{\Lambda}^{n\prime}$ is equal to $\widetilde{F}^s\widetilde{\Lambda}^{s\prime}$, it follows that $\widetilde{\Gamma}_t^n = \widetilde{\Gamma}_t^s/T$ in view of $\widetilde{\lambda}_i^n = \widetilde{\lambda}_i^s/\sqrt{T}$, see equations (6a)-(6c). From these relationships, it is easy to see that

$$\widehat{\alpha}^{n\prime}(\widetilde{V}^n)^{-1}\widetilde{\Gamma}_t^n(\widetilde{V}^n)^{-1}\widehat{\alpha}^n = \widehat{\alpha}^{s\prime}(\widetilde{V}^s)^{-1}\widetilde{\Gamma}_t^s(\widetilde{V}^s)^{-1}\widehat{\alpha}^s.$$

Thus, C_T^2 is the same whether F_t is assumed to be I(0) or I(1). The above argument is valid for F_t being I(2) or other processes. This result has the practical implication that forecasting confidence intervals derived for I(0) common factors are valid for nonstationary factors.

Table 1: Coverage Rates and MSE:

$$\widehat{Avar(\widehat{\delta})} = \widehat{\sigma}_{\varepsilon}^2 \left[\frac{1}{T} \sum_{t=1}^{T} \widehat{z}_t \widehat{z}_t' \right]^{-1},$$

$$\widetilde{\Gamma}_t = \widetilde{\sigma}_e^2 \frac{1}{N} \sum_{i=1}^{N} \widetilde{\lambda}_i \widetilde{\lambda}_i \quad \forall\, t.$$

				Coverage Probability				MSE			
N	T	b	k	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$
50	50	0.00	2	0.94	0.93	0.93	0.92	0.15	0.09	1.17	1.15
100	50	0.00	2	0.94	0.92	0.94	0.94	0.12	0.09	1.09	1.07
200	50	0.00	2	0.95	0.92	0.93	0.93	0.09	0.08	1.16	1.16
50	100	0.00	2	0.95	0.92	0.94	0.94	0.10	0.04	1.17	1.09
50	200	0.00	2	0.96	0.94	0.96	0.95	0.07	0.02	1.07	1.03
200	100	0.00	2	0.96	0.94	0.95	0.94	0.05	0.04	1.07	1.07
100	200	0.00	2	0.96	0.94	0.95	0.94	0.04	0.02	1.04	1.02
200	200	0.00	2	0.95	0.94	0.95	0.95	0.03	0.02	1.03	1.03
100	400	0.00	2	0.97	0.95	0.96	0.96	0.03	0.01	0.95	0.91
50	50	0.50	2	0.91	0.93	0.94	0.92	0.23	0.09	1.22	1.15
100	50	0.50	2	0.93	0.92	0.94	0.94	0.16	0.09	1.12	1.07
200	50	0.50	2	0.94	0.92	0.93	0.93	0.10	0.08	1.16	1.16
50	100	0.50	2	0.93	0.92	0.94	0.94	0.15	0.04	1.24	1.09
50	200	0.50	2	0.94	0.94	0.96	0.95	0.13	0.02	1.14	1.03
200	100	0.50	2	0.96	0.94	0.95	0.94	0.06	0.04	1.08	1.07
100	200	0.50	2	0.95	0.94	0.95	0.94	0.07	0.02	1.09	1.02
200	200	0.50	2	0.96	0.94	0.96	0.95	0.04	0.02	1.03	1.03
100	400	0.50	2	0.97	0.95	0.96	0.96	0.06	0.01	0.99	0.91
50	50	0.00	1	0.55	0.93	0.90	0.92	1.13	0.09	2.22	1.15
100	50	0.00	1	0.52	0.92	0.92	0.94	0.99	0.09	1.97	1.07
200	50	0.00	1	0.53	0.92	0.93	0.93	0.90	0.08	2.01	1.16
50	100	0.00	1	0.52	0.92	0.94	0.94	1.05	0.04	2.14	1.09
50	200	0.00	1	0.50	0.94	0.94	0.95	0.93	0.02	2.01	1.03
200	100	0.00	1	0.44	0.94	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.00	1	0.43	0.94	0.94	0.94	0.94	0.02	2.03	1.02
200	200	0.00	1	0.38	0.94	0.95	0.95	0.90	0.02	1.80	1.03
100	400	0.00	1	0.40	0.95	0.96	0.96	0.86	0.01	1.78	0.91
50	50	0.50	1	0.57	0.93	0.91	0.92	1.15	0.09	2.24	1.15
100	50	0.50	1	0.54	0.92	0.92	0.94	1.00	0.09	1.99	1.07
200	50	0.50	1	0.53	0.92	0.93	0.93	0.91	0.08	2.02	1.16
50	100	0.50	1	0.55	0.92	0.93	0.94	1.09	0.04	2.21	1.09
50	200	0.50	1	0.53	0.94	0.94	0.95	0.96	0.02	2.05	1.03
200	100	0.50	1	0.47	0.94	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.50	1	0.45	0.94	0.93	0.94	0.96	0.02	2.07	1.02
200	200	0.50	1	0.39	0.94	0.96	0.95	0.91	0.02	1.81	1.03
100	400	0.50	1	0.43	0.95	0.96	0.96	0.86	0.01	1.79	0.91

Table 2: Coverage Rates and MSE:
$$\widehat{Avar(\widehat{\delta})} = \left(\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t'\right)^{-1}\left[\frac{1}{T}\sum_{t=1}^{T-h}\widehat{\varepsilon}_{t+h}^2\widehat{z}_t\widehat{z}_t'\right]\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t'\right)^{-1},$$
$$\widetilde{\Gamma}_t = \widetilde{\sigma}_e^2\frac{1}{N}\sum_{i=1}^{N}\widetilde{\lambda}_i\widetilde{\lambda}_i \quad \forall\, t.$$

				Coverage Probability				MSE			
N	T	b	k	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$
50	50	0.00	2	0.92	0.85	0.93	0.92	0.15	0.09	1.17	1.15
100	50	0.00	2	0.92	0.85	0.94	0.94	0.12	0.09	1.09	1.07
200	50	0.00	2	0.94	0.86	0.93	0.92	0.09	0.08	1.16	1.16
50	100	0.00	2	0.93	0.89	0.94	0.94	0.10	0.04	1.17	1.09
50	200	0.00	2	0.93	0.91	0.96	0.95	0.07	0.02	1.07	1.03
200	100	0.00	2	0.95	0.90	0.95	0.94	0.05	0.04	1.07	1.07
100	200	0.00	2	0.94	0.92	0.95	0.94	0.04	0.02	1.04	1.02
200	200	0.00	2	0.94	0.92	0.95	0.95	0.03	0.02	1.03	1.03
100	400	0.00	2	0.95	0.94	0.96	0.96	0.03	0.01	0.95	0.91
50	50	0.50	2	0.88	0.85	0.94	0.92	0.23	0.09	1.22	1.15
100	50	0.50	2	0.91	0.85	0.94	0.94	0.16	0.09	1.12	1.07
200	50	0.50	2	0.93	0.86	0.93	0.92	0.10	0.08	1.16	1.16
50	100	0.50	2	0.92	0.89	0.94	0.94	0.15	0.04	1.24	1.09
50	200	0.50	2	0.92	0.91	0.96	0.95	0.13	0.02	1.14	1.03
200	100	0.50	2	0.95	0.90	0.95	0.94	0.06	0.04	1.08	1.07
100	200	0.50	2	0.92	0.92	0.94	0.94	0.07	0.02	1.09	1.02
200	200	0.50	2	0.94	0.92	0.96	0.95	0.04	0.02	1.03	1.03
100	400	0.50	2	0.94	0.94	0.96	0.96	0.06	0.01	0.99	0.91
50	50	0.00	1	0.51	0.85	0.90	0.92	1.13	0.09	2.22	1.15
100	50	0.00	1	0.50	0.85	0.92	0.94	0.99	0.09	1.97	1.07
200	50	0.00	1	0.51	0.86	0.93	0.92	0.90	0.08	2.01	1.16
50	100	0.00	1	0.48	0.89	0.94	0.94	1.05	0.04	2.14	1.09
50	200	0.00	1	0.46	0.91	0.94	0.95	0.93	0.02	2.01	1.03
200	100	0.00	1	0.42	0.90	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.00	1	0.40	0.92	0.94	0.94	0.94	0.02	2.03	1.02
200	200	0.00	1	0.34	0.92	0.95	0.95	0.90	0.02	1.80	1.03
100	400	0.00	1	0.34	0.94	0.96	0.96	0.86	0.01	1.78	0.91
50	50	0.50	1	0.52	0.85	0.90	0.92	1.15	0.09	2.24	1.15
100	50	0.50	1	0.52	0.85	0.92	0.94	1.00	0.09	1.99	1.07
200	50	0.50	1	0.50	0.86	0.93	0.92	0.91	0.08	2.02	1.16
50	100	0.50	1	0.51	0.89	0.94	0.94	1.09	0.04	2.21	1.09
50	200	0.50	1	0.49	0.91	0.94	0.95	0.96	0.02	2.05	1.03
200	100	0.50	1	0.45	0.90	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.50	1	0.42	0.92	0.93	0.94	0.96	0.02	2.07	1.02
200	200	0.50	1	0.36	0.92	0.96	0.95	0.91	0.02	1.81	1.03
100	400	0.50	1	0.39	0.94	0.96	0.96	0.86	0.01	1.79	0.91

Table 3: Coverage Rates and MSE, $h = 4$:

$$\widehat{Avar(\widehat{\delta})} = \left(\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t'\right)^{-1}\left[\frac{1}{T}\sum_{t=1}^{T-h}\widehat{\varepsilon}_{t+h}^2\widehat{z}_t\widehat{z}_t'\right]\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t'\right)^{-1},$$

$$\widetilde{\Gamma}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\widetilde{\lambda}_i\widetilde{\lambda}_j'\frac{1}{T}\sum_{t=1}^{T}\widetilde{e}_{it}\widetilde{e}_{jt} \quad \forall\, t.$$

				Coverage Probability				MSE			
N	T	b‘	k	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$	$\widehat{y}_{T+h\|T}$	$\widehat{y}^0_{T+h\|T}$	$\widehat{y}_{T+h}$	$\widehat{y}^0_{T+h}$
50	50	0.00	2	0.92	0.85	0.93	0.92	0.15	0.09	1.17	1.15
100	50	0.00	2	0.91	0.85	0.94	0.94	0.12	0.09	1.09	1.07
200	50	0.00	2	0.94	0.86	0.93	0.92	0.09	0.08	1.16	1.16
50	100	0.00	2	0.92	0.89	0.94	0.94	0.10	0.04	1.17	1.09
50	200	0.00	2	0.94	0.91	0.95	0.95	0.07	0.02	1.07	1.03
200	100	0.00	2	0.95	0.90	0.95	0.94	0.05	0.04	1.07	1.07
100	200	0.00	2	0.94	0.92	0.95	0.94	0.04	0.02	1.04	1.02
200	200	0.00	2	0.94	0.92	0.95	0.95	0.03	0.02	1.03	1.03
100	400	0.00	2	0.95	0.94	0.96	0.96	0.03	0.01	0.95	0.91
50	50	0.50	2	0.96	0.85	0.95	0.92	0.23	0.09	1.22	1.15
100	50	0.50	2	0.96	0.85	0.95	0.94	0.16	0.09	1.12	1.07
200	50	0.50	2	0.96	0.86	0.94	0.92	0.10	0.08	1.16	1.16
50	100	0.50	2	0.98	0.89	0.95	0.94	0.15	0.04	1.24	1.09
50	200	0.50	2	0.99	0.91	0.96	0.95	0.13	0.02	1.14	1.03
200	100	0.50	2	0.99	0.90	0.95	0.94	0.06	0.04	1.08	1.07
100	200	0.50	2	0.99	0.92	0.95	0.94	0.07	0.02	1.09	1.02
200	200	0.50	2	0.99	0.92	0.96	0.95	0.04	0.02	1.03	1.03
100	400	0.50	2	1.00	0.94	0.96	0.96	0.06	0.01	0.99	0.91
50	50	0.00	1	0.51	0.85	0.90	0.92	1.13	0.09	2.22	1.15
100	50	0.00	1	0.50	0.85	0.92	0.94	0.99	0.09	1.97	1.07
200	50	0.00	1	0.51	0.86	0.93	0.92	0.90	0.08	2.01	1.16
50	100	0.00	1	0.48	0.89	0.94	0.94	1.05	0.04	2.14	1.09
50	200	0.00	1	0.46	0.91	0.94	0.95	0.93	0.02	2.01	1.03
200	100	0.00	1	0.42	0.90	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.00	1	0.40	0.92	0.94	0.94	0.94	0.02	2.03	1.02
200	200	0.00	1	0.35	0.92	0.95	0.95	0.90	0.02	1.80	1.03
100	400	0.00	1	0.35	0.94	0.96	0.96	0.86	0.01	1.78	0.91
50	50	0.50	1	0.61	0.85	0.91	0.92	1.15	0.09	2.24	1.15
100	50	0.50	1	0.57	0.85	0.93	0.94	1.00	0.09	1.99	1.07
200	50	0.50	1	0.56	0.86	0.93	0.92	0.91	0.08	2.02	1.16
50	100	0.50	1	0.60	0.89	0.94	0.94	1.09	0.04	2.21	1.09
50	200	0.50	1	0.59	0.91	0.95	0.95	0.96	0.02	2.05	1.03
200	100	0.50	1	0.50	0.90	0.94	0.94	1.01	0.04	2.11	1.07
100	200	0.50	1	0.50	0.92	0.94	0.94	0.96	0.02	2.07	1.02
200	200	0.50	1	0.43	0.92	0.96	0.95	0.91	0.02	1.81	1.03
100	400	0.50	1	0.49	0.94	0.96	0.96	0.86	0.01	1.79	0.91

References

Andrews, D. W. K. (1991), Heteroskedastic and Autocorrelation Consistent Matrix Estimation, *Econometrica* **59**, 817–854.

Bai, J. and Ng, S. (2002), Determining the Number of Factors in Approximate Factor Models, *Econometrica* **70:1**, 191–221.

Bai, J. and Ng, S. (2003), Evaluating Latent and Observed Factors in Macroeconomics and Finance, manuscript in preparation.

Bai, J. S. (2003), Inference on Factor Models of Large Dimensions, *Econometrica* **71:1**, 135–172.

Bai, J. S. (2004), Estimating Cross-Section Common Stochastic Trends in Non-Stationary Panel Data, *Journal of Econometrics.*

Bernanke, B. and Boivin, J. (2002), Monetary Policy in a Data Rich Environment, *Journal of Monetary Economics.*

Chamberlain, G. and Rothschild, M. (1983), Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets, *Econometrica* **51**, 1305–1324.

Conley, T. (1999), GMM Estimation with Cross-Section Dependence, *Journal of Econometrics* **92**, 1–45.

Cristadoro, R., Forni, M., Reichlin, L. and Giovanni, V. (2001), A Core Inflation Index for the Euro Area, manuscript, www.dynfactor.org.

Driscoll, H. and Kraay, A. (1998), Consistent Covariance Matrix Estimation with Spatially-Dependent Panel Data, *Review of Economics and Statistics* **80:4**, 549–560.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2001b), Do Financial Variables Help in Forecasting Inflation and Real Activity in the Euro Area, manuscript, www.dynfactor.org.

Newey, W. and West, K. (1987), A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* **55**, 703–708.

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

Stock, J. H. and Watson, M. W. (2002b), Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business and Economic Statistics* **20:2**, 147–162.

White, H. (1980), A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica* **48**, 817–38.

Figure 1: Conditional Mean Forecast: DIP

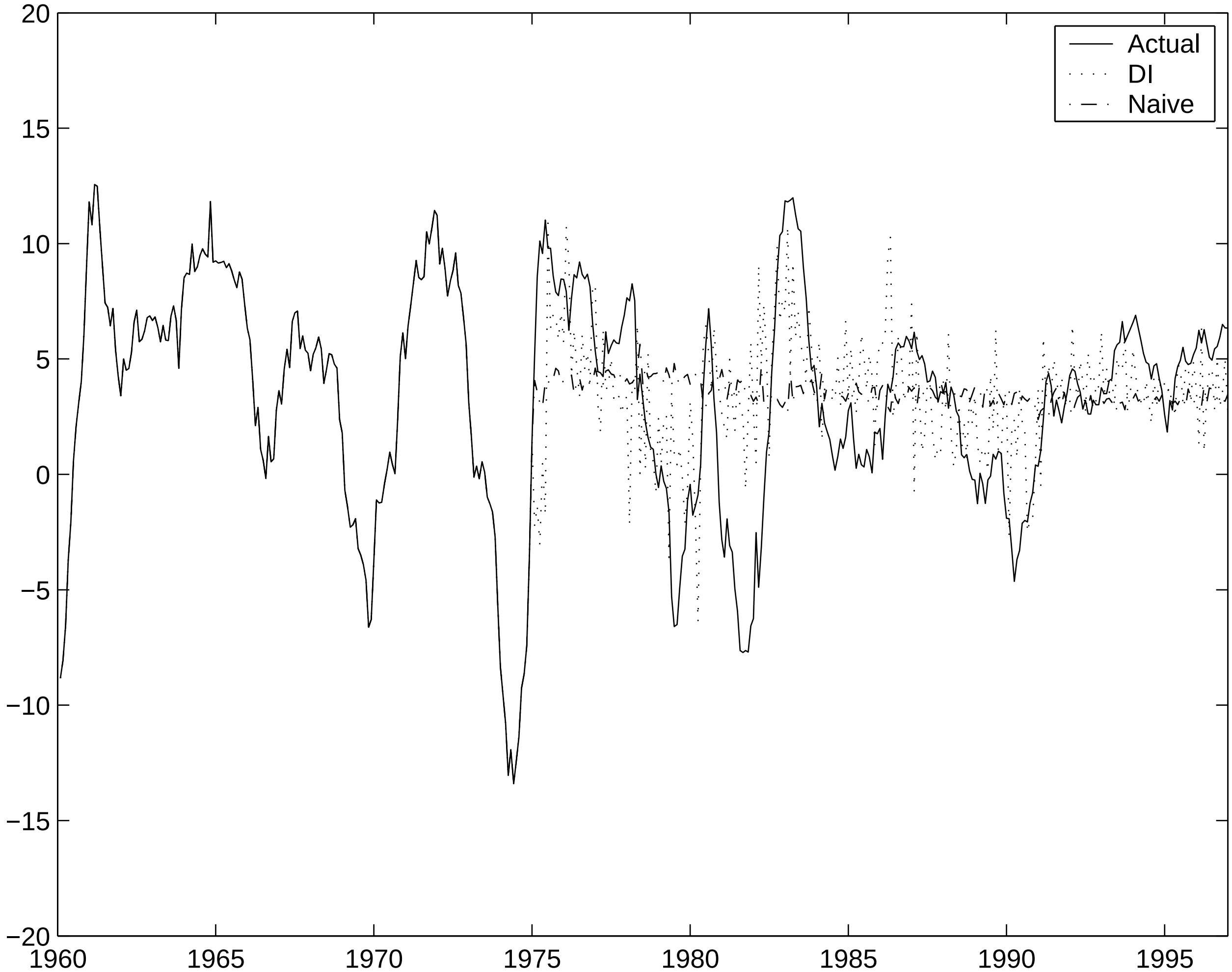

Figure 2a: Diffusion Index Forecast: DIP

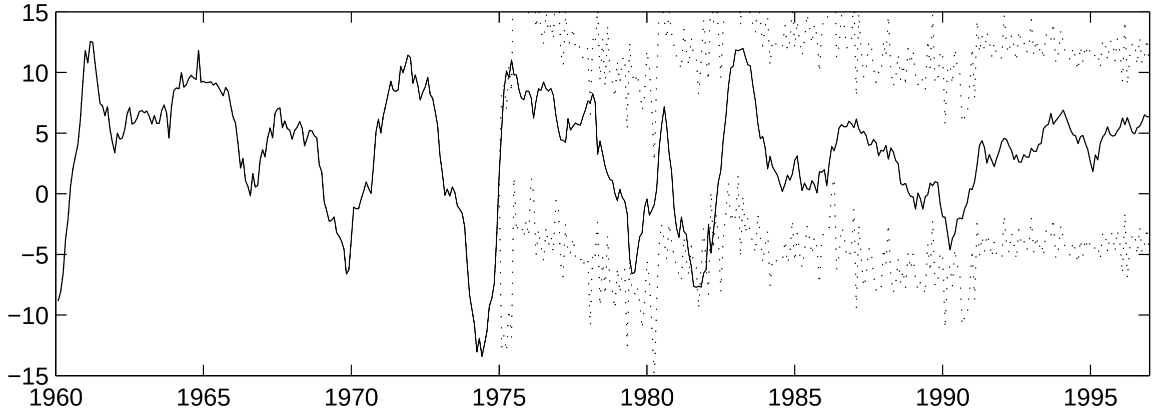

Figure 2b: Naive Forecast: DIP