Are More Data Always Better for Factor Analysis?

Jean Boivin* Serena Ng†

September 30, 2002

Abstract

Factors estimated from large macroeconomic panels are being used in an increasing number of applications. However, little is known about how the size and composition of the data affect the factor estimates. In this paper, we question whether it is possible to use more series to extract the factors and that yet the resulting factors are less useful for forecasting, and the answer is yes. Such a problem tends to arise when the idiosyncratic errors are cross-correlated. It can also arise if forecasting power is provided by a factor that is dominant in a small dataset but is a dominated factor in a larger dataset. In a real time forecasting exercise, we find that factors extracted from as few as 40 pre-screened series often yield satisfactory or even better results than using all 147 series. Our simulation analysis is unique in that special attention is paid to cross-correlated idiosyncratic errors, and we also allow the factors to have weak loadings on groups of series. It thus allows us to better understand the properties of the principal components estimator in empirical applications.

*Graduate School of Business, 719 Uris Hall, Columbia University, New York, NY. Email: jb903@columbia.edu

†Corresponding author: Department of Economics, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218. Email: Serena.Ng@jhu.edu

The authors would like to thank National Science Foundation for financial support. We thank Pritha Mitra for useful research assistance.

1 Introduction

Most macroeconomic empirical analysis are based on a handful of variables. For example, a typical VAR has around six and rarely more than ten variables. While abandoning information in a large number of series can be justified only under rather restrictive assumptions about the joint distribution of the data, use of large scale models remains an exception rather than a rule. In part, this is because of the computation burden involved with large models, and in part, this is because not every available series can be informative, so that including irrelevant information may also come at a cost. In recent years, a new strand of research has made it possible to use relevant information from a large number of variables while keeping the empirical framework small. These studies are based on the assumption that the data admit a factor structure and thus have a common-idiosyncratic decomposition. Factor analysis provides a formal way of defining what type of variation is relevant for the panel of data as a whole.

While use of factor models in empirical modelling is not new, the new generation of factor models differ from the classical ones in at least two important ways:- (i) the idiosyncratic errors can be weakly serially and cross-sectionally correlated, and (ii) the number of observations in 'large dimensional factor models' is large in both the cross-section (N) and the time (T) dimensions. Relaxing the assumption of 'strict factor models' to allow the errors to be correlated makes the framework suited for a wider range of economic applications. The large dimensional nature of the data panel makes it possible to exploit more data in the analysis. It also opens the horizon for consistent estimation of the factors, something that is not possible when the number of cross-section units is small.

Empirical work adopting a large dimensional approximate factor framework have predominantly been based upon the principal components estimator, by which the factor estimates are obtained from an eigenvalue decomposition of the the sample covariance matrix (the static approach) or the spectral density matrix (the dynamic approach). The results thus far are encouraging. Forni and Lippi (1997) and Forni and Reichlin (1998) found two factors formed from 450 disaggregated series to be helpful in understanding aggregate dynamics. Stock and Watson (2002), and Chan, Stock and Watson (1998) showed that the forecast errors of many macroeconomic variables can be reduced by extracting three factors from around 150 series. Forni, Hallin, Lippi and Reichlin (2001) obtained a similar result using 123 series to estimate two factors. Bernanke and Boivin (2002), Bernanke, Boivin and Eliasz (2002) used roughly the same data as Stock and Watson and found that information in the factors is relevant for the empirical modelling of monetary policy. Using as many as 479 series, Giannone, Reichlin and Sala (2002) also adopted a factor approach to

assess the conduct of monetary policy. Stock and Watson (2001) and Forni, Hallin, Lippi and Reichlin (2001b) used factors estimated from around 150 and 400 series respectively, to assess whether financial variables help forecast inflation and real activity. Forni and Reichlin (1998) used data from 138 regions in Europe to extract country and Europe-specific factors, while Cristadoro, Forni, Reichlin and Giovanni (2001) used 447 series to construct a four-factor core inflation index for the Euro area.

The studies cited above are evidently quite different from the small scale VAR analyses that have dominated the literature, as each study has used at least 100 cross-section units to estimate the factors. However, Watson (2000) also found that the marginal gain (in terms of forecast mean-squared error) from increasing N beyond 50 appears less substantial. Bai and Ng (2002) found that in simulations, the number of factors can be quite precisely estimated with N as small as 40 when the errors are *iid*. This suggests that N does not need to be extremely large for the principal components estimator to give reasonably precise estimates.

Could it be that increasing N beyond a certain point is not even desirable? This might appear to be an implausible outcome at first thought, as basic statistical principles suggest more data always improve statistical efficiency. However, whereas a typical panel is sampled to be representative of a cross section, with indicators provided by the data releasing agency to reflect the sampling design, the data used in macroeconomic type factor analysis is not subject to the same scrutiny. The factors are always defined with respect to a specific set of data, and 'correctness' of the dataset depends very much on the exercise on hand. Every rule used to select the data is in some sense ad-hoc. But by choice of data, two researchers using the same estimator can end up with more or less efficient factor estimates. The choice of data is thus not innocuous.

The basic intuition for why using more data to estimate the factors might not be desirable is as follows. In theory, there is an upper bound on the permissible amount of cross-correlation in the errors of an approximate factor model. In practice, because our data are drawn from a small number of broad categories (such as industrial production, prices, interest rates), the probability that the errors are correlated will likely increase with the number of series considered for analysis. Suppose we add data whose information content about the factors is very similar to those series already in the data set. While the new series add little to the variation in the common component of the data, it could have errors that are strongly cross-correlated with the included ones. When enough of such 'noisy' series are added, the residual cross-correlation could exceed the bound warranted by the approximate factor model. Unfortunately, this theoretical bound has no empirical counterpart and thus cannot be tested in practice. Nonetheless, it is important to understand to what extent the factor estimates are affected by this problem of cross correlation in the errors, and what can

be done to mitigate the problem. This provides the motivation for our analysis.

The objective of this paper is to provide a better understanding of the role of N in factor analysis, in a setup that mimics key features of typical macroeconomic applications. To our knowledge, this paper is the first to focus on the small sample properties of the principal component estimators in the presence of cross-section correlation in the idiosyncratic errors. This is especially important since this is a pervasive feature of the data and while the new generation of factor models allows some flexibility in that dimension, the limits to that flexibility are not well understood.[1]

The paper is organized as follows. Section 2 begins by using simple examples to show how and to what extent adding more data can have adverse effects on the factor estimates. We use monte carlo simulations in Section 3 to document the conditions under which adding more data can be undesirable. In Section 4, we use 147 series as in Stock and Watson (2002) to obtain standard (unweighted) factor estimates. We then consider procedures that weigh the data before extracting the principal components. We find that when used to forecast eight macroeconomic time series, the forecasts using the weighted estimates generally have smaller errors than the unweighted ones. In some sense, this result is encouraging, as it indicates that we have not fully exploited the potential of factor analysis. However, the results also point to a need to develop more efficient estimators as it is not simply N that determines estimation and forecast efficiency. The information that the data can convey about the factor structure is also important.

2 The Role of N in Theory

Denote by y_{t+1} the one-period ahead forecast of a series y_t. The model that generates y_t is not known. Given the history of y_t, a naive forecast can be obtained using an AR(p) model

$$\widehat{y}_{t+1|y_t,\dots y_1} = \widehat{\alpha}_0 + \sum_{j=1}^{p} \widehat{\alpha}_j y_{t-j+1} \tag{1}$$

with forecast error variance $\widehat{\sigma}_p^2$. Suppose we also observe N series, $X_t = (X_{1t}, \dots X_{Nt})'$, some of which are informative about y_{t+1}. If N is small (and smaller than T), we can consider the forecast

$$\widehat{y}_{t+1|y_t,\dots y_1,X_t} = \widehat{\eta}_0 + \sum_{i=1}^{N} \widehat{\eta}_{1i}' X_{it} + \sum_{j=1}^{p} \widehat{\gamma}_j y_{t-j+1}.$$

However, if N is large, such a forecast will not be efficient because sampling variability will increase with the number of regressors. When $N > T$, the forecast is not even feasible.

[1] As reported in Appendix I, the idiosyncratic error of 115 of the 147 series we considered have a correlation coefficient of more than .5 in absolute value with another series in the dataset.

Now assume X_{it} admits a factor structure:

$$X_{it} = \lambda_i^{0\prime} F_t^0 + e_{it} \equiv C_{it} + e_{it}, \quad i = 1, \dots N, t = 1, \dots T.$$

In the above, F_t^0 is a $r \times 1$ vector of factors common to all variables, λ_i^0 is the vector of factor loadings for series i, $C_{it} = \lambda_i^{0\prime} F_t$ is the common component of series i, and e_{it} is an idiosyncratic error. If we observe the factors F_t^0, we can consider the forecast:

$$\widehat{y}_{t+1|y_t,\dots y_1,F_t^0} = \widehat{\beta}_0 + \widehat{\beta}_1' F_t^0 + \sum_{j=1}^{p} \widehat{\gamma}_j y_{t-j+1} \tag{2}$$

whose forecast error variance is $\widehat{\sigma}^2_{\widehat{\epsilon},0}$. The appeal of (2) is that it allows information in a large number of observed data X_t to be summarized in a small number of variables, F_t^0. But F_t^0 is not observed. Let $\widehat{F}_{t,N}$ be a consistent estimate of the factors using information on N series up to time t. Then the feasible factor-augmented forecast, referred to by Stock and Watson (1998) as a 'diffusion index' forecast, is

$$\widehat{y}_{t+1|y_t,\dots y_1,\widehat{F}_{t,N}} = \widehat{\beta}_0 + \widehat{\beta}_1' \widehat{F}_{t,N} + \sum_{j=1}^{p} \widehat{\gamma}_j y_{t-j+1}, \tag{3}$$

with forecast error $\widehat{\sigma}^2_{\widehat{\epsilon}}$. Now the difference between the diffusion index forecast and the naive AR(p) forecast is $\widehat{\sigma}^2_{\widehat{\epsilon}} - \widehat{\sigma}^2_p = \left[\widehat{\sigma}^2_{\widehat{\epsilon}} - \widehat{\sigma}^2_{\widehat{\epsilon},0}\right] + \left[\widehat{\sigma}^2_{\widehat{\epsilon},0} - \widehat{\sigma}^2_p\right]$. If F_t^0 was observed, the first error would be irrelevant and a feasible diffusion forecast can do no worse than the AR(p) forecast. This follows from the fact that (2) nests (1) and the mean squared error from using the latter for forecasting cannot exceed the former. But the feasible forecast is based upon (3), which involves generated regressors $\widehat{F}_{t,N}$. In finite samples, the desirability of a feasible diffusion index forecast depends crucially on the errors in estimating F_t^0.

Whereas estimation is carried out by the method of maximum likelihood in a classical setting when N is small, large scale approximate factor models are estimated by either the static or the dynamic method of principal components.[2] As both estimators are aimed at large dimensional panels, the issues to be discussed are relevant to both. Our discussion follows the simpler static principal components estimator.

Let Σ_X and Σ_C be the $N \times N$ population covariance matrices of the data and of the unobserved common components, respectively. Let Ω be the covariance matrix of the idiosyncratic errors. These can be thought of as N-dimensional sub-matrices of the population covariances. A factor model has population covariance structure $\Sigma_X = \Sigma_C + \Omega$. As F_t^0 is common to all variables, the

[2]Kapetanios and Marcellino (2002) showed that the time domain approach seems to perform slightly better.

r largest eigenvalues of Σ_C increase with N. A fundamental feature of factor models is that the r largest eigenvalues of Σ_X will also diverge with N. This suggests that the space spanned by the factors can be estimated using an eigenvalue decomposition of the $\widehat{\Sigma}_X$, the sample estimate of Σ_X. Denote by $\widehat{\lambda}_i' = (\widehat{\lambda}_{i1} \dots \widehat{\lambda}_{ir})$ the estimated loadings, and let $\widehat{F}_{t,N} = (\widehat{F}^1_{t,N} \dots \widehat{F}^r_{t,N})'$ be the estimated factors. Let $v_j = (v_{1j}, \dots v_{Nj})'$ be the eigenvector corresponding to the j^{th} largest eigenvalue of the $N \times N$ sample covariance moment matrix, $\widehat{\Sigma}_X$. The j^{th} estimated factor is $\widehat{F}^j_{t,N} = \frac{1}{\sqrt{N}} \sum_{i=1}^N X_{it} v_{ij}$, and the corresponding loading is estimated as $\widehat{\lambda}_{ij} = \sqrt{N} v_{ij}$. Stock and Watson (1998), Bai and Ng (2002), showed that the factor space can be consistently estimated as $N, T \to \infty$ if (i) the errors are stationary, (ii) the factors have non-trivial loadings, and (iii) the idiosyncratic errors have weak some correlation both serially and mutually. The first condition can be relaxed.[3] The second condition is necessary for distinguishing the pervasive factors from the idiosyncratic noise. Under the third condition, Ω need not be a diagonal matrix for a given N. But, as N tends to infinity, the non-diagonal elements of Ω should go to zero, and the diagonal terms should approach the constant, average cross-section idiosyncratic variance.[4] Bai (2001) further showed that $\widehat{F}_{t,N}$, suitably scaled, is $\sqrt{N}$ consistent for F^0_t. The various asymptotic results lead to the natural presumption that that the factor estimates are more efficient the larger is N.

Intuition about the large sample properties of the factor estimates is best seen from a model with one factor ($r = 1$), and identical loadings ($\lambda^0_i = \lambda \ \forall i$). Given a panel of N cross sections, a decomposition of Σ_x (assumed known) would yield $\widehat{F}_{t,N} = F_t + \frac{1}{N} \sum_{i=1}^N e_{it}$, from which it follows that $\text{var}(\widehat{F}_{t,N}) = \text{var}(\frac{1}{N} \sum_{i=1}^N e_{it})$. If e_{it} is assumed to be iid, $\text{var}(\widehat{F}_{t,N}) = \frac{\sigma^2}{N}$ would decrease with N irrespective of the value of λ. The result is analogous to classical regression analysis when the loadings are observed. In that case, the least squares estimate of F_t can be obtained from a cross-section regression of the data at time t on the N loadings. If the errors are iid, then by the Gauss Markov Theorem, the estimator is efficient, and the variance of the estimates falls with N.

However, even in a regression setting, the relation between the variance of the least squares estimates and the sample size is not unambiguous when the iid assumption is relaxed. Consider estimation of the sample mean. Suppose N_1 series are drawn from a population with variance σ^2_1, from which we can compute a sample mean, $\bar{y}$. Suppose an additional N_2 series are drawn from a population with variance σ^2_2, where $\sigma^2_1 < \sigma^2_2$. With $N = N_1 + N_2$ series, we obtain a sample mean $\widetilde{y}$. It is easy to see that $\frac{\text{var}(\widetilde{y})}{\text{var}(\bar{y})} > 1$ if $\frac{N_1\sigma^2_1 + N_2\sigma^2_2}{N^2} > \frac{\sigma^2_1}{N_1}$. Whether or not more data yield more

[3] Bai and Ng (2001) showed uniform consistency even when the idiosyncratic errors are non-stationary. This means that even when the individual regressions are spurious, the common factors can be consistently estimated from a large dimensional panel.

[4] Connor and Korajzcyk (1986) provided the first results for this estimator using sequential asymptotics. The asymptotic properties of the dynamic estimator are analyzed in Forni, Hallin, Lippi and Reichlin (2000) under similar conditions.

efficient estimates depend very much on the properties of the additional series.

Indeed, this result extends to factor analysis. Consider a concrete example. Suppose we have N_1 series with $\sigma_i^2 = 1$, and N_2 series with $\sigma_i^2 = (1+\theta^2)$. Suppose also that all $N = N_1 + N_2$ series have the same loading. If the covariance matrix of the common component was observed, an eigenvalue decomposition of it would have precisely estimated that λ is identical for every i. Suppose instead of Σ_X, we decompose $\widehat{\Sigma_X}$. Its largest eigenvalue, is $\frac{1}{2}[N+1+\theta^2+N\vartheta]$, where $\vartheta \leq 1$ a function of N_1, N_2, and θ.[5] It is easy to show that the eigenvector corresponding to the largest eigenvalue of Σ_X will have two distinct values. Indeed, the loadings estimated for the first N_1 series will identical, as will the loadings estimated for the last N_2 series. But the estimated loadings will be the same across i only $\vartheta = 1$. When $\vartheta \neq 1$, the ratio of the two distinct values of the eigenvector is an indicator of the bias. If $N_1 = N_2 = 200$ and $\theta = 3$, this ratio is .98. If $N_1 = N_2$ is 500, the ratio is .99. When $(N_1, N_2) = (50,20)$, the ratio becomes .94 when $\theta = 2$, and drops to .87 when $\theta = 3$. Clearly, the ratio is close to unity when N_1 and/or N_2 are large. But for small yet realistic values of N_1 and N_2, the discrepancy can be non-trivial. Combining data with very different error variances can yield rather inaccurate factor estimates.

Similar results obtain when there is cross correlation in the errors. Suppose a researcher unintentionally included N_1 series twice, so that there are $N = 2N_1$ series, and thus N_1 pairs of the idiosyncratic errors are perfectly correlated. When the errors are *iid*, it can be shown that $\text{var}(\widehat{F}_{t,N}) = \frac{\sigma^2}{N_1}$ which depends on N_1, not the total number of series used, N. Nothing is gained by adding more data because the duplicated series increase the variation of the common component, but the variance of the now cross correlated errors is also larger by the same proportion. It is then not hard to construct cases when some series have errors so strongly correlated with others that adding them reduces rather than improves the efficiency of the factor estimates.

In both examples considered above, a practitioner with a large dataset can be misled into thinking that the estimates are more efficient than they really are, even with Σ_X assumed known. In practice this matrix has to be estimated. Sampling variability could make use of more data even more undesirable for the factor estimates. In the next section, the relation between the factor estimates, N, and the properties of the data, are explored using monte carlo simulations.

3 Monte Carlo Simulations

This section reports results from two monte carlo experiments designed to study how the choice of data might affect the factor estimates. The goal is to provide a better understanding of the

[5] To be precise, $\vartheta = \sqrt{\frac{(\theta^2+N_2-N_1-1)^2+4N_1N_2}{(N_1+N_2)^2}}$, which equals unity if $\theta = 1$.

properties of the estimator in practical situations.

The series to be forecasted in both monte carlos are generated by

$$y_{t+1} = \sum_{j=1}^{r} \beta_j F_{jt}^0 + \epsilon_{t+1} \equiv y_{F^0,t+1|t} + \epsilon_{t+1},$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, and σ_ϵ^2 is chosen such that the R^2 of the forecasting equation is κ_y. When β and F_t^0 are observed, we denote the forecast by $y_{F^0,t+1|t}$. The infeasible diffusion index forecast is $\widehat{y}_{F^0,t+1|t}$, which only requires estimation of β. The feasible diffusion index forecast is denoted $\widehat{y}_{\widehat{F},t+1|t}$, which requires estimation of both the factors and β. The X_{it} are simulated from a model with r factors:

$$X_{it} = \sum_{m=1}^{r} \lambda_{im} F_{mt} + e_{it}.$$

The factor loadings vary across i and are assumed to be N(1,1).

The data are always standardized to have mean zero and unit variance before applying the method of principal components. Let $x_{it} = X_{it} - \bar{X}_i$ be the demeaned data to which the method of principal components is used to extract k factors, where k can be different from r. For each sample drawn from a given data configuration, three statistics are considered:

$$\begin{aligned} S_{F,F0} &= \frac{tr(F^{0\prime}F^0\widehat{F}(\widehat{F}'\widehat{F})^{-1}\widehat{F}'F^0)}{tr(F^{0\prime}F^0)} \\ MSE_y &= \frac{\sum_{t=1}^{T}(y_{F^0,t+1|t} - \widehat{y}_{\widehat{F},t+1|t})^2}{\sum_{t=1}^{T} y_{F^0,t+1|t}} \\ S_{y,y0} &= 1 - \frac{\sum_{t=1}^{T}(\widehat{y}_{F^0,t+1|t} - \widehat{y}_{\widehat{F},t+1|t})^2}{\sum_{t=1}^{T} \widehat{y}_{F^0,t+1|t}^2}. \end{aligned}$$

Since we can only identify the space spanned by the factors, the second factor need not coincide with the second estimated factor. We project each of the true factors on all estimated factors and use the $S_{F,F0}$ statistic to summarize the discrepancy in the space spanned by the actual and the estimated factors. The MSE_y statistic assesses the feasible diffusion index forecasts relative to the conditional mean. This evaluation is possible (but not in an empirical setting) because the model that generates the forecast in the simulation is known. The $S_{y,y0}$ statistic measures how close are the feasible forecasts from the infeasible forecasts that assume the factors are observed. We report the average of these statistics over M=1000 replications.

3.1 Model 1: Correlated and Noisy Errors

Most of the studies on the finite sample properties of the principal components estimator assumed the idiosyncratic errors are *iid*, as in Kapetanios and Marcellino (2002) and Forni et al. (2000).

But an aspect of approximate factor models is that they allow for cross-correlation in the errors. Stock and Watson (1998) and Bai and Ng (2002) considered a case where an error is correlated with the one ordered before and after it, and find cross-correlated errors to have some but not strong, adverse effects on the forecasts and the number of factors being selected. But as we will see in the next section, the cross-correlation present in empirical applications is substantially stronger. To our knowledge, this is the first study in which the effect of cross-correlated errors on the factor estimates of an approximate factor model is the primary focus. It is also common practice to assume in the simulations that the common and idiosyncratic components are of equal importance, and that the error variances are constant across i. In practice, one is likely to encounter data with diverse properties.

In this monte carlo experiment, the factors are assumed to be *iid*. The total number of series available is $N = N_1 + N_2 + N_3$. The $N_1 \times N_3$ matrix Ω_{13} has $C \times N_3$ non-zero elements and is the source of cross-correlation. Unequal variance is captured by assuming $\sigma_1^2 < \sigma_2^2$. With $u_{it} \sim N(0,1)$ as the building block, we consider three types of idiosyncratic errors:

N1: $e_{it} = \sigma_1 u_{it}$,

N2: $e_{it} = \sigma_2 u_{it}$,

N3: $e_{it} = \sigma_3 \widetilde{e}_{it}, \widetilde{e}_{it} = u_{it} + \sum_{j=1}^{C_i} \rho_{ij} u_{jt}$.

The first N_1 series are what we call 'clean' series. They are mutually uncorrelated and whose variance is small relative to the common component. The next N_2 series differs from the first N_1 series only in that the factors have less explanatory power. Each of the next N_3 series is correlated with some C series that belong to the N_1 set. Residual cross correlation can arise when there is a factor that is not sufficiently pervasive. Thus, an alternative way to simulate data with the same error structure is to assume that there is a $r+1$ factor with loadings such that $\lambda_{i,r+1} \neq 0$ for C of the first N_1 series and for the last N_3 series.[6]

To isolate cases with cross-correlated errors from cases with large error variances, we let $\sigma_1 = \sigma_3 < \sigma_2$. To ensure that the presence of cross correlation does not reduce the importance of the factors, the $\widetilde{e}_{it}$ are standardized to have unit variance. The σ_i^2 are chosen such that the factors explain κ_i of the variation in the data, given λ.[7] These assumptions on the idiosyncratic errors are

[6] With this alternative method, interpretation of the number of factors and control over the size of the common component is more difficult.

[7] More precisely, $\kappa = \frac{\text{var}(C)}{\text{var}(C)+\sigma_e^2}$ which can be used to solve for σ_e^2, given var(C) implied by the other parameters.

meant to replicate an Ω with the following property:

$$\Omega = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0_{N_1 \times N_2} & \Omega_{13} \\ 0_{N_2 \times N_1} & \sigma_2^2 I_{N_2} & 0_{N_2 \times N_3} \\ \Omega_{13}' & 0_{N_3 \times N_2} & \sigma_3^2 I_{N_3} \end{bmatrix},$$

We consider data generated by up to three factors. For a given r (the true number of factors), we estimate $k = 1, \ldots 3$ factors. Thus, if $k < r$, the assumed number of factors is too small. Since the number of series correlated with the first data type can be no larger than N_1, we let $N_3 = n_3 N_1$ and vary n_3. We draw ρ_{ij} drawn from a uniform distribution with lower bound of .05 and upper bound of .7. We consider three N_1 values, five N_2 values, ten pairs of (N_3, C), nine sets of $(\kappa_1, \kappa_3, \kappa_y)$, giving a total of 12,150 configurations.

Some summary statistics of the simulations are given in the second panel of Table 1. Notably, the experimental design generates substantial variations in the factor estimates, with S_{F,F_0} ranging from .04 (almost unpredictable) to .99 (almost perfectly predictable). The mean-squared forecast errors ranges from .072 to .964, while $S_{y,y0}$ ranges from .071 to .99. We use the following statistics to measure the extent of correlation in the errors and the relative importance of the common component:

$$\tau = \max_i \frac{1}{N} \sum_{j=1}^{N} |\tau_{ij}| \quad R_2 = \frac{1}{N} \sum_{i=1}^{N} R_i^2 \qquad R_q = R_{.9N}^2 - R_{.1N}^2$$

where

$$\tau_{ij} = \frac{1}{T} \sum_{t=1}^{T} \widehat{e}_{it} \widehat{e}_{jt}, \qquad R_i^2 = \frac{\sum_{t=1}^{T} \widehat{c}_{it}^2}{\sum_{t=1}^{T} x_{it}^2}.$$

A factor model with mutually uncorrelated idiosyncratic errors is a 'strict factor model'. Chamberlain and Rothschild (1983) showed that asset prices have an approximate factor structure if the largest eigenvalue (and hence all of the eigenvalues) of the $N \times N$ covariance matrix $\Omega = E(e_t e_t')$ is bounded. But the largest eigenvalue of Ω is bounded by $\max_i \sum_{j=1}^{N} |\tau_{ij}|$. Thus, under the assumptions of an approximate factor model, there should exist a $P < \infty$ such that $\sum_{j=1}^{N} |\tau_{ij}| \leq P$ for all i and all N. While P provides a bound for the development of theory, it does not provide a practical guide as to how much cross-correlation is permissible. The τ statistic aims shed some light on how much cross correlation can be tolerated in empirical analysis. For a given i, R_i^2 indicates the relative importance of the common component. The average of R_i^2 measures the average size of the common component. The cross-section dispersion in the common component is measured by R_q, the difference between the R_i^2 in the 90 and the 10 percentile.

Because of the large number of configurations involved, we summarize the the variations in S_{F,F_0} and $S_{y,y0}$ and MSE_y using response surfaces. We begin with a general specification that

includes higher order terms and gradually drop the statistically insignificant ones. Table 1 reports the estimates, along with the robust standard errors. Recall that the larger is S_{F,F_0}, the more precise are the factor estimates. The larger is $S_{y,y0}$, the closer are the diffusion index forecasts to those generated by the (infeasible) forecasts based on observed factors. The larger is MSE_y, the larger is the forecast mean-squared error. Increasing N_1 by one increases S_{F,F_0} by less than one basis point, but increases S_{y,y_0} by more than one basis point. Not surprisingly, increasing N_1 reduces MSE_y, but at a declining rate.

Under-estimating the number of factors ($r > k$) reduces the precision of both the factor estimates and the forecasts, while overestimating ($r < k$) the number of factors has the opposite effect. This last result is surprising, as a priori, we had expected some efficiency loss from over-parameterization. One explanation for this could be that the over-parameterization considered (by just one factor) is very small compared to the sample size. The quantitative effect of under and over estimation are quite different for estimation and forecasting. A rule that is optimal from a factor estimation point of view may not be be optimal from a forecast point of view. As expected, the factor estimates are generally more precise the smaller the number of true factors (r) holding N fixed. An additional factor reduces S_{F,F_0} by about 7 basis points. The mean-squared forecast errors are higher the larger the number of true factors, as replacing the factors by the principal components estimates inevitably increases sampling variability.

Adding series with relatively large idiosyncratic variances has a first order effects on S_{F,F_0} and $S_{y,y0}$ that are positive, but second order effects that are negative. Thus, adding another series has efficiency gains when N_2 is small, but negative when N_2 is sufficiently large. To be precise, the effect of N_2 on S_{F,F_0} becomes negative when N_2 is around 22 (1.925/.088), all else equal. Coincidentally, the threshold of N_2 $S_{y,y0}$ is also around 22.

Whereas the coefficient on N_2 indicates the effect of adding noisy data to clean data, the coefficient on R_2 indicates if the factor estimates depend on the average size of the common component, or in other words, the strength of the factor structure. Clearly, the factors are more precisely estimated when the common component is important. However, the larger the dispersion in the importance of the common component, as indicated by R_q^2, the less precise are the estimates. Put differently, the estimates are more precise when then data are more homogeneous. One can get more efficient estimates from using a small set of homogenous data than adding to it a lot of data with large idiosyncratic errors.

As expected, S_{F,F_0} and $S_{y,y0}$ are both decreasing in N_3 and C, holding other parameters fixed. A summary statistic for the extent of cross correlation in the errors is τ, which itself depends on C, N_3, and the actual residual correlation. Increasing τ by .01 reduces S_{F,F_0} and $S_{y,y0}$ by .86 and

.90 basis points, respectively.[8]

The present monte-carlo exercise highlights two points. First, while increasing N_1 is desirable from both an estimation and forecasting standpoint, this is not always the case if we increase data of the N_2 and N_3 type. It can be misleading to use the total number of observations as a guide to efficiency. Second, the factor estimates and forecasts are clearly less efficient when the errors are cross correlated and/or have vastly unequal variances. This amounts to an Ω whose off-diagonal elements are non-zero, and which has substantial dispersion in the diagonal elements. But this is precisely the flexibility that an approximate factor model purports to offer.

To understand why the factor estimates are sensitive to the properties of the errors, it is useful to recall that in a classical setting with N fixed, the maximum likelihood estimates of λ for given Ω (a diagonal matrix) are the eigenvectors of $\Omega^{-1/2}\widehat{\Sigma}_X\Omega^{-1/2}$, see Anderson (1984), p. 589. The principal components estimates, on the other hand, are the eigenvectors of $\widehat{\Sigma}_X$, which evidently correspond to the maximum likelihood estimates when Ω is a scalar matrix. Accordingly, for a given N, the principal components estimator is not efficient when the errors are far from homoskedastic and mutually uncorrelated. The question in practice is whether Ω corresponding to the data being analyzed is sufficiently far from being diagonal as to violate the assumptions of the approximate factor model. Formal procedures that test if these assumptions (such as the bound on τ) are not available at the moment. In the next section, we will consider estimation of factors using different data configurations to gauge the extent of the problem in empirical applications and to see if improvements can be made.

3.2 Model 2: Oversampling

In empirical work, we choose to analyze a subset of data available. To understand if it matters which N series are selected, we simulate data from a strict factor model in this subsection. We assume that there are two serially correlated factors driving the data, viz: $X_{it} = \sum_{m=1}^{2} \lambda_{im}F_{mt} + e_{it}$,

$$F_{mt} = .5F_{mt-1} + u_{mt}, \quad u_{mt} \sim N(0,1).$$

Two series are to be forecasted and are generated as follows:

$$\begin{aligned} y^A_{t+1} &= \beta^A F_{1t} + \epsilon^A_{t+1} \\ y^B_{t+1} &= \beta^B F_{2t} + \epsilon^B_{t+1}, \end{aligned}$$

with $\sigma^A_\epsilon = \sigma^B_\epsilon$. There are five types of data in this monte carlo, with sample size $N_i, i = 1, \ldots 5$:

N1: $X_{it} = .8F_{1t} + e_{it}, \sigma^2_i \sim N(0, 1 - .8^2)$;

[8]The second order effect is positive, but numerically small. Evaluated at $\tau = .1$, the second order effect is 1.81.

N2: $X_{it} = .6F_{2t} + e_{it}, \sigma_i^2 \sim N(0, 1 - .6^2)$;

N3: $X_{it} = .4F_{1t} + .1F_{2t} + e_{it}, \sigma^2 \sim N(0, 1 - .4^2 - .1^2)$;

N4: $X_{it} = .1F_{1t} + .4F_{2t} + e_{it}, \sigma^2 \sim N(0, 1 - .1^2 - .4^2)$;

N5: $X_{it} = e_{it}, \sigma_i^2 \sim N(0, 1)$.

The simulated data have two features. First, some series are driven by one factor, some by two factors, and some do not obey a factor structure. Second, some series weigh factor 1 more heavily than factor 2 and vice versa. To fix ideas of the situation that the experiment attempts to mimic, suppose one factor is real and one is nominal. The N_1 series might be output and employment type series, the N_2 series might be prices, the N_3 series might be interest rate type series, and the N_4 series might be stock market type series. Variations in the N_5 series are purely idiosyncratic. The errors are mutually uncorrelated within and between groups. Cross correlation is not an issue in this experiment.

The simulation results are reported in Table 2. The main features of the previous monte carlo are also apparent here when the errors are not cross-correlated. First, under-estimating the number of factors has large efficiency loss, while over-estimating has little impact on the estimates or the forecasts. Second, the factor estimates are no less precise when the noisy data are dropped, even though the nominal sample size is smaller. Remarkably, when the number of assumed factors is at least as large as the underlying data, the space spanned by the factors can be quite precisely estimated by the method of principal components with as few as 40 series, provided the data are informative about the factors. With 40 series (case 3), S_{F,F_0} is .944. In none of the remaining cases that two or more factors were estimated was there a noticeable improvement in S_{F,F_0}. With 100 series (case 9), for example, S_{F,F_0} is .955; the factor estimates are only marginally better. Thus as in the previous monte carlo, efficiency of the factor estimates is determined not by whether the sample size is 40 or 100, but by what information the data has to bear about the factors.

One motivation for the present monte carlo is to highlight the fact that factor space being estimated depends on the choice of data. In case 1 when the N_1 series was used, the eigenvector estimates the space spanned by F_1. In case 2 when N_2 series was used, the eigenvector estimates the space spanned by F_2. For this reason, extracting one factor given the N_1 dataset is optimal for forecasting y^A, as is extracting one factor from the N_2 dataset for the purpose of forecasting y^B. However, a dataset formed with $N_1 + N_2$ series will have two factors. As the average variation of F_1 is larger than that of F_2 (in view of the loadings in this data), the estimator will first identify the space spanned by F_1, and then the space spanned by both F_1 and F_2. Thus, to forecast y^B, we

need to use two estimated factors to pick up the variation in F_2. Analogously, if we had data in which F_2 dominates F_1, such as case 4 with $N_2 + N_3$ series, forecasting y^A using one factor would have been disastrous. We refer to a situation in which the data are more informative about some factors than the others as 'oversampling'.

More generally, let m be the true number of factors in the forecasting equation. The foregoing results suggest that when the data are oversampled, the number of estimated factors that will efficiently forecast a series that depends on m factors will be larger than m, if the m factors are not the m most dominant factors in X. A criterion that determines the optimal number of factors in X can be a misleading indicator of the number of factor needed for forecasting a single series, y.

The problem of oversampling is helpful in understanding why in Table 2, y^A is always forecasted more precisely than y^B, even though both series have the same degree of predictability (since $\sigma_\epsilon^A = \sigma_\epsilon^B$ and $\beta^A = \beta^B$). This result arises because efficient forecasts of y^B requires inclusion of more estimated factors than y^A. But more estimated factors also induce more sampling variability into the forecasts. For this reason, forecasts of a series that depend on the less important factors in a given set of data will tend to be inferior to those that depend on the dominant factors.

As noted earlier, macroeconomic panels are 'put together' by the researcher, quite unlike a typical panel in which data are designed to be representative of a population of interest. The factors are always sample dependent. Correctness of the factors thus depend on the objective on hand. However, as also seen from the results, the forecast error for y^B using $N_2 + N_3$ series is larger than using N_2 series alone. Likewise, the forecast error for y^A from using $N_1 + N_3$ series is larger than using the N_1 series alone. This raises the possibility that if we think the series to be forecasted depends on F_1 and F_2, estimating F_1 from N_1 series and F_2 from N_2 series could outperform estimating F_1 and F_2 jointly from a larger dataset comprising of series with varying factor structures. This alternative will be explored in the next section.

4 The Role of N in Real Time Forecasting

The goal of this section is see to if $\widehat{y}_{t+1|y_t,\ldots y_1,\bar{F}_{t,N}}$ depends on N in real-time, 12 month ahead forecasting of 8 economic time series: industrial production (ip), real personal income less transfers (gmyxspq), real manufacturing trade and sales (msmtq), number of employees on nonagricultural payrolls (lpnag), the consumer price index (punew), the personal consumption expediture deflator (gmdc), the CPI less food and energy (puxx), and the producer price index for finished goods (pwfsa). The logarithms of the four real variables are assumed to be $I(1)$, while the logarithms of the four prices are assumed to be $I(2)$.

Let y_t generically denote one of the eight series after logarithmic transformation. Define the h

step ahead growth to be $y^h_{t+h} = 100[y_{t+h} - y_t]$ and the one period growth to be $z_t = 100 \cdot h[y_t - y_{t-1}]$. The diffusion index forecasts are obtained from the equation

$$\widehat{y}_{t+h|y_t,\ldots y_1,\widehat{F}_t} \equiv \widehat{y}_{t+h|t} = \widehat{\beta}_0 + \widehat{\beta}_1' \widehat{F}_{t,N} + \sum_{j=1}^{p} \widehat{\gamma}_j z_{t-j+1}, \tag{4}$$

where $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\gamma}$ are OLS estimates. The univariate forecasts are based on the model that excludes the factors. Since our primary interest is in the role of N, we fix p to 4 to compare univariate AR(4) forecasts with those augmented with up to $r = 6$ factors.[9]

The base case uses a balanced panel of N=147 series to estimate the factors. These are monthly series available from 1959:1 to 1998:12. Following Stock and Watson (2002), the data are standardized and transformed to achieve stationarity where necessary. The data can roughly be classified into 13 groups:- [1]: real output and income (series 1-19), [2]: employment and hours (series 20-44), [3]: retail and manufacturing trade (series 45-53), [4]: consumption (series 54-58), [5]: housing starts and sales (series 59-65), [6]: inventories (series 66-76), [7]: orders (series 77-92), [8]: stock prices (series 93-99), [9]: exchange rate (series 100-104), [10]: interest rates (105-120), [11]: money and credit (series 121-127), [12]: price indexes (series 128-145), [13]: misc (series 146-147). Details are given in Appendix A. The R^2_i for each series for three and six factors are also reported.

The forecasting exercise begins with estimation of the factors and of (4) using data from 1959:3-1970:1. A 12 period ahead forecast is formed by using values of the regressors at 1970:1 to give $y^h_{1970:1+h}$ for $h = 12$. The sample is updated by one period, the factors and the forecasting model are both re-estimated, and a 12 month forecast for 1971:2 is formed. The final forecast is made for 1998:12 in 1998:12-h. The rolling AR(4) forecasts are likewise constructed. We consider MSER(r,0), the mean squared error of a r factor diffusion index forecast relative to that of an AR(4) forecast with zero factors.

The base case results are reported in the first row of columns 5 to 12 of Table 3. A $r-$factor diffusion index forecast beats the AR(4) forecast if MSER(r,0)$<$ 1. For the output series, the forecast error from including one factor alone is sometimes larger than the simple AR(4) model, but adding more factors can lead to significant improvements. For example, for IP, MSER(1,0) is .973. With two factors, MSER(2,0) falls to .632, and relative error is further reduced to .589 with three factors. However, there is no further gain in increasing the number of factors beyond three to forecast the real series. The inflation forecast error tends to be lowest with $r = 1$. In the case of CPI inflation, MSER is .734 with one factor, .750 for two factors, and .806 with three factors. For gmdc, MSER(1,0) is .734. A three factor forecast would almost wipe out the gain from using

[9]The main difference between our analysis and that of Stock and Watson is that we did not use the BIC to select p, and we did not allow lags of the factors to enter the forecasting model.

the factors over the AR(4) model, as MSER(3,0) is .956. As there is no evidence to support use of more than three factors, results for more estimated factors are not reported.

The results show that diffusion indexes can be useful for real time forecasting, even though the factors have to be estimated. However, more efficient forecasts are potentially available. A look at the data selected for analysis and estimated residuals reveal several features (see Appendix A). First, there are more series from some groups than others. Second, many of the R_i^2s are very small. For example, three factors explain only .01 of the variation in IPXMCA (series 16). Even with six factors, R_i^2 is improved to a mere .08, much smaller than series such as PMEMP that has an R_i^2 of .8. The dispersion in the importance of the common component, or in other words, the idiosyncratic error variance, is thus quite large. Third, there is substantial cross correlation in the idiosyncratic errors. To gauge the problem, we let τ_{ij} be the correlation coefficient between the $T \times 1$ vectors $\widehat{e}_i$ and $\widehat{e}_j$, the estimated residuals corresponding to the i^{th} and the j^{th} equation, respectively, from estimation of a six factor model over the entire sample, 71:1-97:12. We can identify a series, say, j_i^*, such that

$$j_i^{*1} = \max_j |\tau_{ij}| \qquad \tau_i^{*1} = |\tau_{ij_i^*}|.$$

That is, j_i^{*1} is the series whose idiosyncratic error is most correlated with series i, and τ_i^{*1} is the corresponding maximal correlation coefficient. For example, the IPCD and IPCN errors are both most correlated with IPC, with correlation coefficients of .66 and .69, respectively. The errors of FSPCOM and FSNCOM are even more correlated, with a τ_{ij} of .99 (see Appendix A). As the maximal correlation coefficient could be an outlier, we also report the second largest correlation coefficients (see the last two columns of Appendix A). Many of these coefficients remain quite high. In light of the properties of the data, the issues of oversampling, correlated errors, and noisy data could be relevant to the present forecasting exercise.

4.1 Weighted Principal Components

In classical regression analysis, generalized least squares is more efficient than ordinary least squares when the errors are non-spherical. This suggests that if we observe Ω, we can consider an efficient principal components estimator that weighs the data with Ω, by analogy to GLS. The problem is that we do not observe Ω. The analogous feasible GLS estimator would be to replace Ω by $\widehat{\Omega}$, the sample error covariance matrix from unweighted estimation of a k factor model. But $\widehat{\Omega}$ is a matrix of rank $N-k$ and thus not invertible. Thus, while minimizing $V(k) = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T e_{it}^2$ is suboptimal, minimizing $W^*(k) = \|\frac{1}{NT}\sum_{t=1}^T e_t'\widehat{\Omega}^{-1}e_t\|$ is infeasible. Another solution, considered in classical factor analysis, is to subtract $\widehat{\Omega}$, which is a diagonal matrix in classical analysis, from $\widehat{\Sigma}_X$. However, $\widehat{\Omega}$ is not diagonal in approximate factor models. The eigenvectors of $\widehat{\Sigma}_X - \widehat{\Omega}$ in fact have

the same span as $\widehat{\Sigma}_X$. There is, to our knowledge, no obvious way to exploit the entire $\widehat{\Omega}$ matrix to improve efficiency.

Although optimal weighting is not possible, it may still be possible to obtain a more efficient principal components estimator by doing some form of weighting. Consider the objective function

$$W(k) = \frac{1}{NT}\sum_{i=1}^{N} w_{iT} \sum_{t=1}^{T} e_{it}^2,$$

where w_{iT} is chosen to reflect the informativeness of series i. Notice that each series is weighted by a scalar, quite unlike feasible GLS whose weighting matrix typically has non-zero off diagonal elements. When N is large, this can be advantageous because it reduces sampling variability induced by any first step estimation.

As w_{iT} is meant to be data dependent, we rely on the residuals from a first step estimation to obtain the weights. If the number of factors is too small in the first step estimation, omitted factors would be treated as correlated errors which could lead to inaccurate weighting. Our first step estimation is based on a model with six factors. The results with eight factors are similar. For the sake of comparison, results using three factors in the first step are also reported.

We consider the following sets of weights:

Rule I: w_{iT} is the inverse of the i^{th} diagonal element of $\widehat{\Omega}_T$ estimated using data up to time T.

Rule 0: w_{iT} is the inverse of $\frac{1}{N}\sum_{j=1}^{N} |\widehat{\Omega}_T(i,j)|$. That is, the row sum of the covariance of the estimated residuals.

Rule 1: Let $j^* = \{j_i^*\}$ be the set of series whose error is most correlated with some other series. These series are all dropped. If $j_i^* = j_{i'}^*$, i.e. if series i and i' are most correlated with each other, series i' is dropped if $R_{i'}^2 < R_i^2$. As some series are in fact most correlated with the several series, of the 147 series, 73 are dropped, leaving us with 71 series.

Rule 2: From Rule 1, we also drop a series if its error is second most correlated with another series. This removes 38 series from the 71 series in the previous set, leaving us with 33 series.

Rule 1c The j^* set in Rule 1 is fixed during the rolling forecasts as it is based on full sample estimation. In real time, this set can be updated continuously. Rule 1c thus uses a set of X that need not be the same over time.

Rule 2c Follows Rule 2c but allows for continuous updating.

Rule I has the same objective function as the maximum likelihood estimator, which assumes Ω is a diagonal matrix. This procedure, also considered in Jones (2001) as a modification to the Connor and Korajzcyk (1998) estimator, might be desirable in a classical setting when the idiosyncratic errors are heteroskedastic but mutually uncorrelated, but not in approximate factor models when some residual cross-correlation is permitted. For this reason, we consider five other sets of weights

to account for cross-correlated errors. Rule 0 weights the data by the magnitude of residual cross-correlation. Like Rule I, all 147 series are used. Under Rules 1 and 2, w_{iT} is a binary variable. A series is either in or out based on the properties of the residuals over the entire sample. Rules 1c and 2c provide further flexibility by allowing w_{iT} to be updated as the sample size (in the time dimension) changes.

The above weighting schemes are aimed at accounting for the properties of Ω. However, our second monte carlo exercise also suggests that it may be more efficient to estimate F_2 from N_2 series rather than from $N_1 + N_2$ series when N_1 may not contain information about F_2. Inspection of $R_i^2(3)$ and $R^2(6)$ in the appendix reveals that many series (in particular, the real variables) are quite well explained by three factors, with small difference between $R_i^2(3)$ and $R_i^2(6)$. However, $R_i^2(6)$ for many series are substantially higher than $R_i^2(3)$ (in particular, the price variables). This suggests that factors that are important to some variables maybe unimportant in others. The 147 series are therefore reclassified into three categories. A 'real' (R) category consisting of 60 series from output, employment, retail and manufacturing trade, consumption, and miscellaneous. A 'nominal' (N) category consisting of 46 series relating to exchange rate, interest rates, money and credit, and price indexes. A 'volatile/leading indicator' (V) category consisting of 41 series of high volatility, including housing starts, inventories, orders, stock prices.[10] An appeal of this grouping is that it provides some economic interpretation to the factors. Obviously, this amounts to having three addition sets of w_{iT}. For example, the real factors are essentially extracted with w_{iT} such that it is one if a series is real and zero otherwise. After the data are classified, we then estimate three real factors using data exclusively from the real series, three from the nominal variables, and three from the volatile series. These $T \times 1$ vectors are denoted $\widehat{F}_j^k$, $j = 1, \ldots 3$, $k = R, N, V$. It remains to determine the order in which the factors are added to the forecasting equation. We consider 4 sets of orderings:

Rule A: F_1^R, F_2^R, F_3^R; Rule B: F_1^N, F_2^N, F_3^N;

Rule C: F_1^V, F_2^V, F_3^V; Rule D: F_1^R, F_1^V, F_1^N;

Under Rule A, only the real factors are used. Under Rule B, only the volatile factors are used. Under Rule C, only the real factors are used. Rule D eventually uses one factor from each category.

There is undoubtedly a certain ad-hocness in all these rules, but if cross-correlated errors are not prevalent in the data, or if data sampling is inconsequential, dropping the data should make the estimates less efficient, not more, since the sample size is smaller. The results using these weighting schemes are therefore revealing even if they are not optimal.

[10]The 'real' variables are thus from groups 1-4, plus 13, the 'nominal' variables are from groups 9-12, and the volatile group are variables from 5-8.

The results are reported in Table 3. Turning first to the output series, Rule C, the volatile factors, are apparently not useful at all in forecasting three of the four the real variables, nor is Rule A. The latter is perhaps to be expected since in the forecasting equation, the factors are augmented by four lags of each series, which themselves are real variables. However, Rule B is extremely effective in forecasting personal income, manufacturing trade, and employment. Although we have used only 46 series in the estimation, the one nominal factor so constructed generates more efficient forecasts than estimating three factors from all 147 series. Adding more nominal factors does not seem necessary.

The result that three of the four real series are better predicted by a factor extracted from predominantly nominal variables can be understood by noting that the first factor in the Stock and Watson data is a predominantly real factor. This means that the nominal factor is a dominated factor in the panel of 147 series. To pick up its variation, we need to include three estimated factors in the forecasting equation. As suggested by our second monte carlo, this is inferior to extracting the nominal factor from data in which this factor is dominant, which is the case with Rule B.

For forecasting the IP series, precise identification of the nominal factors (Rule B) is not useful at all, though adding the real and the volatile factors always beat the naive AR(4) forecast.[11] Downweighing series with a small common component or strong residual covariances are better than the base case. Notice that under Rule I, the error for a one factor forecast is .580, much smaller than the one factor base case forecast of .973, showing that even sub-optimal weighting can make an improvement. The best IP forecast is provided by Rule 2, which simply drops series with strongly correlated errors. With factors extracted from only 33 series, the forecasts errors are half of those of the AR(4) forecast, and ten percent smaller than if all 147 series are used. As well, using the same 33 series all the time seems to outperform continuously updating which series are to be dropped. Thus, for all four real series, we find that diffusion index forecasts can be made quite effective with factors extracted from a smaller number of series.

Turning now to the inflation series, using factors associated with Rule B are evidently uninformative. Again, this is not surprising, as the lagged prices will likely encompass much of the information in the nominal factors. Adding one real factor or one volatile factor both lead to smaller forecast errors than the AR(4) forecasts. Although none of the methods considered appear to perform noticeably better than the base case, the forecast errors of the base case which uses 147 series are closely matched by those with factors extracted from 33 series, ie Rule 2.

Two final observations can be made from Table 3. The first concerns how many factors are

[11] We also considered factors estimated from the combined set of real and nominal variables, real and volatile variables, and nominal and volatile variables. They are also worse than the base case.

really in the data and/or are required in forecasting exercises. Consider the three cases that extract the factors from all 147 series. Evidently, the one factor weighted forecasts (Rules I and 0) are quite similar to the three factor unweighted forecasts (base case). Over-parameterizing the number of factors appears to serve a similar purpose as downweighting the observations. But arguably, one factor is all that is required in the forecasting exercise.

The second observation is that Rule 2 which uses just 33 series tends to produce forecasts that are comparable (if not better) in each of the eight cases than the base case that uses 147 series. It should again be stressed that we make no claim that the rule is optimal. But the result underscores our main point that use of more data to extract the factors does not necessarily yield better results. Reducing the sample size can sometimes help sharpen the factor structure and enables more efficient estimation. As well, use of more series in the estimation increases the probability that the data have correlated errors. Both observations serve as a reminder that the principal components estimator has many desirable properties if certain regularity conditions are satisfied. The selection of data is not innocuous because it determines how close are these conditions from being violated in practice.

5 Conclusion

A feature stressed in recent applications of factor models is the use of data from 'large' panels data. Because the theory is developed for large N and T, there is a natural tendency for researchers to use as much data as are available. In this paper, we show that more data do not necessarily yield more efficient estimates and forecasts. The information conveyed by the data also determines efficiency. As there is no formal guide to data selection, two researchers, both using the principal components estimator, can end up with rather different estimates. In simulations and the empirical examples considered, the factors extracted from as few as 40 series seem no less, and in many cases more, efficient than the ones extracted from 147 series.

The present study is motivated by issues that a practitioner will likely encounter in practice. In general applications, the number of series potentially usable for analysis can be quite large. Suppose we have already included the all-item CPI in the analysis. Would it be useful to also include CPI ex-food and energy? Suppose we have a large number of disaggregated series, along with a small number of aggregated ones. Should we use all series? What is the cost of oversampling data from particular groups? Is there a trade-off between the quantity and quality of the data? There is at the moment no guide to what data should be used in factor analysis. Our results nonetheless suggests that sample size alone does not determine the properties of the estimates. The quality of the data must be taken into account. The result that the factors can be quite precisely estimated with just 40 series also suggests considering estimators that are computationally more demanding

but that could be more efficient than the principal components estimator.

Appendix I: Data

				Rule								
Series	name	tcode	group	1	2	3	$R_i^2(3)$	$R_i^2(6)$	j_1^*	j_2^*	$\max_j \lvert\tau_{ij}^1\rvert$	$\max_j \lvert\tau_{ij}^2\rvert$
1	IP	5	1	1	0	0	0.70	0.73	11	9	0.89	0.81
2	IPP	5	1	1	1	1	0.62	0.68	3	4	0.93	0.78
3	IPF	5	1	1	0	0	0.55	0.61	2	4	0.93	0.81
4	IPC	5	1	1	0	0	0.41	0.48	3	2	0.81	0.78
5	IPCD	5	1	1	1	1	0.37	0.45	4	3	0.66	0.57
6	IPCN	5	1	1	0	0	0.14	0.16	4	13	0.69	0.66
7	IPE	5	1	1	0	0	0.44	0.48	3	2	0.55	0.47
8	IPI	5	1	1	0	0	0.42	0.45	11	1	0.41	0.41
9	IPM	5	1	1	0	0	0.51	0.53	1	11	0.81	0.67
10	IPMND	5	1	1	1	0	0.36	0.37	13	36	0.61	0.27
11	IPMFG	5	1	1	1	0	0.71	0.75	1	12	0.89	0.85
12	IPD	5	1	1	0	0	0.62	0.67	11	1	0.85	0.78
13	IPN	5	1	1	0	0	0.41	0.43	6	10	0.66	0.61
14	IPMIN	5	1	1	1	1	0.05	0.05	9	1	0.47	0.34
15	IPUT	5	1	1	0	0	0.01	0.08	57	45	0.45	0.24
16	IPXMCA	1	1	1	1	0	0.70	0.77	43	42	0.63	0.63
17	PMI	1	1	1	1	1	0.75	0.77	77	18	0.87	0.85
18	PMP	1	1	1	0	0	0.68	0.71	17	77	0.85	0.82
19	GMYXPQ	5	1	1	0	0	0.38	0.39	140	144	0.27	0.25
20	LHEL	5	2	1	0	0	0.32	0.36	21	30	0.69	0.18
21	LHELX	5	2	1	1	1	0.45	0.49	20	138	0.69	0.21
22	LHEM	5	2	1	0	0	0.24	0.27	23	33	0.90	0.27
23	LHNAG	5	2	1	1	1	0.28	0.30	22	9	0.90	0.20
24	LHUR	1	2	1	1	1	0.49	0.72	43	42	0.72	0.67
25	LHU680	1	2	1	0	0	0.41	0.63	28	29	0.78	0.60
26	LHU5	1	2	1	0	0	0.38	0.82	27	67	0.72	0.32
27	LHU14	1	2	1	1	0	0.53	0.89	26	29	0.72	0.62
28	LHU15	1	2	1	0	0	0.50	0.83	29	25	0.88	0.78
29	LHU26	1	2	1	1	1	0.56	0.87	28	27	0.88	0.62
30	LPNAG	5	2	1	1	0	0.72	0.76	31	37	0.93	0.71
31	LP	5	2	1	0	0	0.70	0.77	30	32	0.93	0.67
32	LPGD	5	2	1	0	0	0.72	0.77	34	31	0.74	0.67
33	LPCC	5	2	1	1	0	0.22	0.29	32	31	0.42	0.34
34	LPEM	5	2	1	1	0	0.69	0.72	35	32	0.92	0.74
35	LPED	5	2	1	0	0	0.62	0.65	34	32	0.92	0.67
36	LPEN	5	2	1	0	0	0.44	0.46	34	32	0.41	0.35
37	LPSP	5	2	1	0	0	0.39	0.41	30	31	0.71	0.57
38	LPTX	5	2	1	1	0	0.35	0.39	37	31	0.55	0.39
39	LPFR	5	2	1	1	0	0.17	0.20	62	78	0.31	0.31
40	LPS	5	2	1	1	1	0.20	0.21	37	30	0.45	0.33
41	LPGOV	5	2	1	1	1	0.11	0.21	37	30	0.47	0.31
42	LPHRM	1	2	1	1	0	0.20	0.21	43	24	0.93	0.67
43	LPMOSA	1	2	1	0	0	0.10	0.16	42	24	0.93	0.72
44	PMEMP	1	2	1	1	1	0.80	0.80	17	18	0.73	0.64
45	MSMTQ	5	3	1	0	0	0.62	0.77	72	46	0.80	0.64
46	MSMQ	5	3	1	0	0	0.58	0.64	73	47	0.91	0.87
47	MSDQ	5	3	1	0	0	0.53	0.59	46	73	0.87	0.78
48	MSNQ	5	3	1	0	0	0.23	0.26	86	46	0.66	0.45
49	WTQ	5	3	1	1	0	0.18	0.32	51	74	0.85	0.82
50	WTDQ	5	3	1	1	1	0.30	0.34	49	74	0.58	0.49
51	WTNQ	5	3	1	0	0	0.04	0.18	49	74	0.85	0.70
52	RTQ	5	3	1	1	0	0.20	0.33	75	54	0.79	0.61
53	RTNQ	5	3	1	0	0	0.07	0.16	56	52	0.81	0.59
54	GMCQ	5	4	1	1	1	0.18	0.36	55	56	0.75	0.61
55	GMCDQ	5	4	1	1	0	0.13	0.22	58	54	0.86	0.75
56	GMCNQ	5	4	1	1	0	0.07	0.19	53	54	0.81	0.61
57	GMCSQ	5	4	1	1	0	0.06	0.14	15	54	0.45	0.26
58	GMCANQ	5	4	1	0	0	0.10	0.19	55	54	0.86	0.57
59	HSFR	4	5	1	1	1	0.42	0.61	64	62	0.87	0.80
60	HSNE	4	5	1	0	0	0.36	0.42	59	43	0.55	0.49
61	HSMW	4	5	1	1	1	0.44	0.47	59	63	0.54	0.37
62	HSSOU	4	5	1	0	0	0.22	0.59	59	64	0.80	0.76
63	HSWST	4	5	1	0	0	0.22	0.41	59	64	0.77	0.72
64	HSBR	4	5	1	0	0	0.33	0.57	59	62	0.87	0.76
65	HMOB	4	5	1	1	0	0.08	0.34	62	64	0.51	0.41
66	IVMTQ	5	6	1	0	0	0.43	0.44	67	68	0.59	0.55
67	IVMFGQ	5	6	1	1	0	0.40	0.43	68	66	0.89	0.59
68	IVMFDQ	5	6	1	0	0	0.37	0.39	67	66	0.89	0.55
69	IVMFNQ	5	6	1	0	0	0.12	0.17	67	28	0.48	0.24
70	IVWRQ	5	6	1	1	0	0.12	0.12	66	74	0.47	0.41
71	IVRRQ	5	6	1	0	0	0.09	0.11	75	66	0.61	0.51
72	IVSRQ	2	6	1	1	1	0.67	0.79	45	85	0.80	0.59
73	IVSRMQ	2	6	1	1	1	0.64	0.67	46	47	0.91	0.78
74	IVSRWQ	2	6	1	1	1	0.19	0.31	49	51	0.82	0.70
75	IVSRRQ	2	6	1	0	0	0.12	0.24	52	71	0.79	0.61
76	PMNV	1	6	1	0	0	0.61	0.62	17	78	0.52	0.39

Appendix I: continued

				Rule								
Series	name	tcode	group	1	2	3	$R_i^2(3)$	$R_i^2(6)$	j_1^*	j_2^*	$\max_j \lvert\tau_{ij}^1\rvert$	$\max_j \lvert\tau_{ij}^2\rvert$
77	PMNO	1	7	1	0	0	0.66	0.71	17	18	0.87	0.82
78	PMDEL	1	7	1	0	0	0.52	0.57	17	76	0.63	0.39
79	MOCMQ	5	7	1	1	1	0.55	0.58	47	46	0.51	0.42
80	MDOQ	5	7	1	0	0	0.50	0.61	84	85	0.99	0.91
81	MSONDQ	5	7	1	1	1	0.12	0.21	91	92	0.83	0.83
82	MO	5	7	1	0	1	0.63	0.73	84	80	0.90	0.88
83	MOWU	5	7	1	0	1	0.44	0.55	85	84	0.98	0.90
84	MDO	5	7	1	1	1	0.50	0.61	80	85	0.99	0.92
85	MDUWU	5	7	1	0	0	0.39	0.50	83	84	0.98	0.92
86	MNO	5	7	1	1	1	0.36	0.39	48	87	0.66	0.52
87	MNOU	5	7	1	0	0	0.18	0.21	86	90	0.52	0.42
88	MU	5	7	1	1	1	0.41	0.47	89	85	0.99	0.43
89	MDU	5	7	1	0	0	0.39	0.46	88	85	0.99	0.44
90	MNU	5	7	1	0	1	0.22	0.25	87	134	0.42	0.21
91	MPCON	5	7	1	0	1	0.09	0.17	92	81	1.00	0.83
92	MPCONQ	5	7	1	0	0	0.08	0.16	91	81	1.00	0.83
93	FSNCOM	5	8	1	0	0	0.28	0.64	94	95	0.98	0.96
94	FSPCOM	5	8	1	0	1	0.28	0.64	95	93	0.99	0.98
95	FSPIN	5	8	1	0	0	0.27	0.62	94	93	0.99	0.96
96	FSPCAP	5	8	1	0	0	0.22	0.54	95	94	0.87	0.84
97	FSPUT	5	8	1	0	0	0.24	0.45	93	94	0.33	0.32
98	FSDXP	2	8	1	1	1	0.30	0.64	95	94	0.77	0.77
99	FSPXE	2	8	1	0	0	0.18	0.39	94	95	0.58	0.57
100	EXRUS	5	9	1	0	0	0.10	0.20	101	103	0.84	0.83
101	EXRGER	5	9	1	1	1	0.08	0.15	102	100	0.87	0.84
102	EXRSW	5	9	1	0	0	0.08	0.15	101	100	0.87	0.81
103	EXRJAN	5	9	1	0	1	0.06	0.15	100	102	0.83	0.57
104	EXRCAN	5	9	1	1	1	0.03	0.10	19	100	0.19	0.17
105	FYFF	2	10	1	0	0	0.23	0.26	106	94	0.30	0.28
106	FYGT5	2	10	1	0	1	0.26	0.45	107	108	0.94	0.77
107	FYGT10	2	10	1	0	0	0.23	0.43	106	108	0.94	0.83
108	FYAAAC	2	10	1	0	0	0.28	0.49	107	109	0.83	0.82
109	FYBAAC	2	10	1	0	1	0.32	0.50	108	107	0.82	0.69
110	FYFHA	2	10	1	0	0	0.22	0.33	107	106	0.60	0.60
111	FM1	6	11	1	1	1	0.05	0.10	112	113	0.62	0.45
112	FM2	6	11	1	0	1	0.04	0.08	113	111	0.64	0.62
113	FM3	6	11	1	0	0	0.02	0.04	112	111	0.64	0.45
114	FM2DQ	5	11	1	1	1	0.24	0.29	62	16	0.36	0.35
115	FMFBA	6	11	1	0	1	0.03	0.03	116	117	0.67	0.32
116	FMRRA	6	11	1	0	0	0.03	0.03	115	117	0.67	0.49
117	FMRNBC	6	11	1	1	1	0.05	0.08	116	115	0.49	0.32
118	PMCP	1	12	1	1	1	0.47	0.50	147	141	0.55	0.38
119	PWFSA	6	12	1	0	0	0.03	0.28	120	129	0.88	0.33
120	PWFCSA	6	12	1	0	1	0.03	0.30	119	129	0.88	0.30
121	PSM99Q	6	12	1	1	1	0.02	0.03	97	107	0.20	0.20
122	PUNEW	6	12	1	0	0	0.08	0.68	131	83	0.33	0.26
123	PU83	6	12	1	0	0	0.03	0.09	124	120	0.26	0.20
124	PU84	6	12	1	1	1	0.04	0.31	123	129	0.26	0.25
125	PU85	6	12	1	0	1	0.00	0.01	128	117	0.19	0.12
126	PUC	6	12	1	0	0	0.08	0.69	134	130	0.36	0.26
127	PUCD	6	12	1	1	1	0.01	0.04	133	134	0.28	0.24
128	PUS	6	12	1	0	0	0.02	0.05	129	134	0.32	0.32
129	PUXF	6	12	1	0	0	0.04	0.33	119	128	0.33	0.32
130	PUXHS	6	12	1	0	1	0.09	0.66	126	82	0.26	0.21
131	PUXM	6	12	1	1	1	0.08	0.62	122	128	0.33	0.22
132	GMDC	6	12	1	0	1	0.09	0.67	135	134	0.66	0.36
133	GMDCD	6	12	1	0	0	0.00	0.04	127	132	0.28	0.23
134	GMDCN	6	12	1	0	0	0.10	0.71	132	126	0.36	0.36
135	GMDCS	6	12	1	0	0	0.01	0.10	132	126	0.66	0.25
136	LEHCC	6	13	1	0	1	0.02	0.02	137	33	0.26	0.24
137	LEHM	6	13	1	0	0	0.01	0.02	35	136	0.31	0.26
138	SFYCP90	1	10	1	0	0	0.23	0.75	140	141	0.40	0.32
139	SFYGM3	1	10	1	0	0	0.51	0.83	140	141	0.83	0.46
140	SFYGM6	1	10	1	1	1	0.53	0.86	139	141	0.83	0.77
141	SFYGT1	1	10	1	0	0	0.52	0.80	140	142	0.77	0.62
142	SFYGT5	1	10	1	0	0	0.67	0.86	143	145	0.96	0.83
143	SFYGT10	1	10	1	0	1	0.69	0.86	142	144	0.96	0.91
144	SFYAAAC	1	10	1	0	0	0.70	0.83	145	143	0.92	0.91
145	SFYBAAC	1	10	1	0	1	0.74	0.85	144	143	0.92	0.88
146	SFYFHA	1	10	1	0	1	0.70	0.86	143	144	0.86	0.83
147	HHSNTN	1	10	1	0	0	0.38	0.58	118	25	0.55	0.40

Table 1: Response Surface for Monte Carlo 1

	Response Surface			Summary Statistics			
	$100 \times S_{F,F_0}$	$100 \times S_{y,y0}$	$100 \times MSE_y$	mean	s.e.	min	max
N_1	0.874	1.172	-1.214	40	16.33	20	60
	(18.38)	(20.05)	(25.33)				
N_1^2	-0.006	-0.007	0.007				
	(9.16)	(10.08)	(12.08)				
$(r-k)>0$	-21.873	-2.195	1.888	.4444	.684	0	2
	(100.92)	(7.69)	(8.37)				
$(k-r)>0$	14.388	6.882	-8.244	.4444	.684	0	2
	(77.40)	(32.08)	(44.16)				
r	-7.058	2.176	2.635	2	.816	1	3
	(28.80)	(8.35)	(12.36)				
N_2	1.925	1.983	-1.862	5	5.773	0	15
	(21.70)	(18.27)	(21.09)				
N_2^2	-0.088	-0.089	0.083				
	(13.70)	(11.15)	(12.72)				
R_2	145.383	171.275	-187.590	.350	.143	.071	.721
	(32.37)	(29.45)	(37.57)				
R_2^2	-49.465	-105.62	94.508				
	(8.86)	(14.74)	(15.07)				
R_q	-12.137	18.773	-32.184	.462	.187	.089	.850
	(3.46)	(4.49)	(8.93)				
R_q^2	-22.023	-62.641	68.285				
	(5.66)	(13.43)	(16.69)				
N_3	-0.483	-0.663	0.689	15.2	16.237	1	60
	(19.01)	(20.31)	(24.94)				
N_3^2	0.007	0.008	-0.009				
	(15.60)	(15.52)	(19.46)				
C	-0.333	-0.655	0.649	24	22.301	0	75
	(16.77)	(25.83)	(30.06)				
C^2	0.001	0.004	-0.003				
	(3.79)	(10.72)	(11.84)				
τ^*	-86.512	-90.361	134.357	.140	.058	.068	.452
	(9.30)	(7.22)	(13.15)				
τ^{*2}	180.718	358.357	-309.556				
	(8.73)	(12.25)	(13.04)				
cons	34.884	14.159					
	(27.23)	(8.62)					
R^2	.8449	.6232	.7184				
S_{F,F_0}				.603	.292	.04	.991
S_{y,y_0}				.740	.23	.071	.99
mse_y				.458	.222	.072	.964

Table 2: Monte Carlo 2 with $r = 2$ factors:

	N_1	N_2	N_3	N_4	N_5	N	k	S_{F,F_0}	MSE_{yA}	MSE_{yB}
1	20	0	0	0	0	20	1	0.488	0.173	1.084
	20	0	0	0	0	20	2	0.491	0.188	1.090
	20	0	0	0	0	20	3	0.493	0.203	1.096
2	0	20	0	0	0	20	1	0.464	0.991	0.284
	0	20	0	0	0	20	2	0.467	0.995	0.299
	0	20	0	0	0	20	3	0.470	1.005	0.305
3	20	20	0	0	0	40	1	0.499	0.222	1.070
	20	20	0	0	0	40	2	0.944	0.185	0.300
	20	20	0	0	0	40	3	0.944	0.194	0.311
4	0	20	20	0	0	40	1	0.471	0.957	0.356
	0	20	20	0	0	40	2	0.870	0.414	0.299
	0	20	20	0	0	40	3	0.871	0.419	0.307
5	20	0	20	0	0	40	1	0.487	0.182	1.079
	20	0	20	0	0	40	2	0.506	0.192	1.063
	20	0	20	0	0	40	3	0.519	0.202	1.053
6	0	20	0	40	0	60	1	0.477	0.976	0.273
	0	20	0	40	0	60	2	0.492	0.976	0.283
	0	20	0	40	0	60	3	0.504	0.976	0.291
7	20	20	20	0	0	60	1	0.494	0.221	1.066
	20	20	20	0	0	60	2	0.943	0.184	0.296
	20	20	20	0	0	60	3	0.944	0.196	0.304
8	20	20	0	40	0	80	1	0.501	0.825	0.579
	20	20	0	40	0	80	2	0.955	0.187	0.253
	20	20	0	40	0	80	3	0.955	0.197	0.262
9	20	20	20	40	0	100	1	0.502	0.594	0.838
	20	20	20	40	0	100	2	0.955	0.186	0.251
	20	20	20	40	0	100	3	0.955	0.196	0.262
10	20	20	0	0	40	80	1	0.499	0.223	1.070
	20	20	0	0	40	80	2	0.942	0.187	0.304
	20	20	0	0	40	80	3	0.942	0.200	0.313
11	20	0	20	0	40	80	1	0.487	0.183	1.079
	20	0	20	0	40	80	2	0.492	0.198	1.079
	20	0	20	0	40	80	3	0.497	0.210	1.085
12	20	20	20	0	40	100	1	0.494	0.221	1.066
	20	20	20	0	40	100	2	0.942	0.185	0.300
	20	20	20	0	40	100	3	0.942	0.201	0.309
13	20	20	0	40	40	120	1	0.500	0.825	0.579
	20	20	0	40	40	120	2	0.954	0.188	0.254
	20	20	0	40	40	120	3	0.954	0.202	0.262
14	20	20	20	40	40	140	1	0.502	0.594	0.838
	20	20	20	40	40	140	2	0.954	0.186	0.252
	20	20	20	40	40	140	3	0.954	0.200	0.261

Table 3: Forecast Errors of Diffusion Index Models Relative to AR(4): 71:1-97:12

rule	N	r	ip	gmyxspq	msmtq	lpnag	punew	gmdc	puxx	pwfsa
SW	147	1	0.973	1.002	0.981	0.914	0.734	0.832	0.801	0.825
1	71	1	0.950	0.990	0.961	0.986	0.745	0.833	0.798	0.837
2	33	1	0.935	0.933	0.945	0.884	0.740	0.843	0.828	0.818
1C	71	1	0.852	0.935	0.896	0.889	0.771	0.871	0.873	0.840
2C	33	1	0.867	0.875	0.880	0.852	0.794	0.869	0.844	0.864
I	147	1	0.580	0.659	0.616	0.627	0.867	0.978	0.969	0.896
0	147	1	0.729	0.899	0.780	0.825	0.743	0.853	0.817	0.810
A	60	1	0.915	1.008	1.001	0.983	0.789	0.863	0.826	0.868
B	46	1	1.041	0.753	0.543	0.645	1.024	1.075	1.099	0.983
C	41	1	0.867	1.052	1.035	0.963	0.800	0.893	0.830	0.882
D	60	1	0.915	1.008	1.001	0.983	0.789	0.863	0.826	0.868
SW	147	2	0.632	0.821	0.580	0.724	0.750	0.864	0.778	0.840
1	71	2	0.748	0.901	0.697	0.782	0.752	0.844	0.773	0.849
2	33	2	0.720	0.886	0.675	0.751	0.754	0.867	0.789	0.837
1C	71	1	0.648	0.802	0.610	0.719	0.783	0.894	0.835	0.848
2C	33	1	0.614	0.743	0.598	0.688	0.818	0.919	0.867	0.889
I	147	2	0.642	0.812	0.572	0.680	0.738	0.855	0.801	0.815
0	147	2	0.616	0.813	0.553	0.701	0.725	0.846	0.769	0.814
A	60	2	0.978	1.094	1.045	1.037	0.792	0.838	0.814	0.892
B	46	2	1.050	0.744	0.529	0.627	1.012	1.075	1.084	0.987
C	41	2	0.820	1.055	1.018	0.976	0.786	0.882	0.797	0.866
D	106	2	0.948	0.750	0.530	0.648	0.821	0.933	0.911	0.866
SW	147	3	0.589	0.790	0.604	0.723	0.806	0.956	0.890	0.883
1	71	3	0.609	0.762	0.589	0.726	0.765	0.890	0.859	0.864
2	33	3	0.519	0.670	0.568	0.675	0.776	0.910	0.877	0.878
1C	71	1	0.644	0.812	0.618	0.766	0.787	0.889	0.897	0.834
2C	33	1	0.579	0.741	0.535	0.658	0.805	0.901	0.870	0.867
I	147	3	0.650	0.841	0.601	0.713	0.804	0.935	0.879	0.881
0	147	3	0.573	0.772	0.564	0.706	0.813	0.961	0.899	0.873
A	60	3	0.983	1.087	1.038	1.026	0.785	0.831	0.818	0.883
B	46	3	1.048	0.719	0.517	0.630	1.048	1.114	1.113	1.003
C	41	3	0.763	1.118	1.068	0.962	0.720	0.845	0.771	0.823
D	147	3	0.865	0.789	0.541	0.639	0.789	0.940	0.881	0.856
AR(4)		.035	0.050	0.027	0.046	0.017	0.021	0.016	0.019	0.035

$B(\tau_{ij} = \frac{1}{N}\sum_i \sum_j |\tau_{ij}|$. Selection based on 6 factors

Table 4: Forecast Errors of Diffusion Index Models Relative to AR(4): 71:1-97:12

rule	N	r	ip	gmyxspq	msmtq	lpnag	punew	gmdc	puxx	pwfsa
SW	147	1	0.973	1.002	0.981	0.914	0.734	0.832	0.801	0.825
1	71	1	0.977	1.023	0.987	0.939	0.721	0.816	0.781	0.819
2	33	1	1.012	1.030	1.030	0.948	0.757	0.859	0.801	0.852
1C	71	1	0.931	0.959	0.949	0.922	0.752	0.834	0.816	0.845
2C	33	1	0.912	0.925	0.937	0.902	0.770	0.851	0.825	0.851
I	147	1	0.645	0.765	0.708	0.792	0.808	0.943	0.877	0.861
0	147	1	0.906	1.007	0.936	0.957	0.741	0.837	0.778	0.833
A	60	1	0.915	1.008	1.001	0.983	0.789	0.863	0.826	0.868
B	46	1	1.041	0.753	0.543	0.645	1.024	1.075	1.099	0.983
C	41	1	0.867	1.052	1.035	0.963	0.800	0.893	0.830	0.882
D	60	1	0.915	1.008	1.001	0.983	0.789	0.863	0.826	0.868
SW	147	2	0.632	0.821	0.580	0.724	0.750	0.864	0.778	0.840
1	71	2	0.702	0.877	0.626	0.760	0.734	0.842	0.763	0.831
2	33	2	0.920	1.041	0.879	1.035	0.726	0.828	0.707	0.861
1C	71	1	0.696	0.880	0.612	0.731	0.768	0.888	0.825	0.851
2C	33	1	0.662	0.817	0.579	0.785	0.800	0.896	0.855	0.870
I	147	2	0.643	0.824	0.574	0.708	0.757	0.865	0.821	0.836
0	147	2	0.615	0.815	0.572	0.720	0.749	0.862	0.788	0.840
A	60	2	0.978	1.094	1.045	1.037	0.792	0.838	0.814	0.892
B	46	2	1.050	0.744	0.529	0.627	1.012	1.075	1.084	0.987
C	41	2	0.820	1.055	1.018	0.976	0.786	0.882	0.797	0.866
D	106	2	0.948	0.750	0.530	0.648	0.821	0.933	0.911	0.866
SW	147	3	0.589	0.790	0.604	0.723	0.806	0.956	0.890	0.883
1	71	3	0.621	0.793	0.574	0.686	0.704	0.823	0.782	0.815
2	33	3	0.770	0.921	0.680	0.819	0.655	0.780	0.724	0.790
1C	71	1	0.617	0.817	0.612	0.707	0.779	0.936	0.896	0.850
2C	33	1	0.631	0.823	0.578	0.696	0.760	0.917	0.842	0.853
I	147	3	0.626	0.842	0.624	0.817	0.833	0.980	0.937	0.895
0	147	3	0.582	0.783	0.594	0.747	0.807	0.962	0.911	0.882
A	60	3	0.983	1.087	1.038	1.026	0.785	0.831	0.818	0.883
B	46	3	1.048	0.719	0.517	0.630	1.048	1.114	1.113	1.003
C	41	3	0.763	1.118	1.068	0.962	0.720	0.845	0.771	0.823
D	147	3	0.865	0.789	0.541	0.639	0.789	0.940	0.881	0.856
AR(4)		.035	0.050	0.027	0.046	0.017	0.021	0.016	0.019	0.035

$B(\tau_{ij} = \frac{1}{N}\sum_i \sum_j |\tau_{ij}|$. Selection based on 3 factors

References

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Bai, J. and Ng, S. (2001), A PANIC Attack on Unit Roots and Cointegration, mimeo, Boston College.

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

Bai, J. S. (2001), Inference on Factor Models of Large Dimensions, mimeo, Boston College.

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

Bernanke, B., Boivin, J. and Eliasz, P. (2002), Factor Augmented Vector Autoregressions (FVARs) and the Analysis of Monetary Policy, mimeo, Columbia University.

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

Chan, Y., Stock, J. and Watson, M. W. (1998), A Dynamic Factor Model Framework for Forecast Combinations, mimeo, Princeton University.

Connor, G. and Korajzcyk, R. (1986), Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis, *Journal of Financial Economics* **15**, 373–394.

Connor, G. and Korajzcyk, R. (1998), Risk and Return in an Equilibrium APT Application of a New Test Methodology, *Journal of Financial Economics* **21**, 225–289.

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

Forni, M. and Lippi, M. (1997), *Aggregation and the Microfoundations of Dynamic Macroeconomics*, Oxford University Press, Oxford, U.K.

Forni, M. and Reichlin, L. (1998), Let's Get Real: a Factor-Analytic Approach to Disaggregated Business Cycle Dynamics, *Review of Economic Studies* **65**, 453–473.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), The Generalized Dynamic Factor Model: Identification and Estimation, *Review of Economics and Statistics* **82:4**, 540–554.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2001), Cooincident and Leading Indicators for the Euro Area, *Economic Journal* **111**, C82–85.

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.

Giannone, D., Reichlin, L. and Sala, L. (2002), Tracking Greenspan: Systematic and Unsystematic Monetary Policy Revisited, manuscript, www.dynfactor.org.

Jones, C. (2001), Extracting Factors from Heteroskedsatic Asset Returns, *Journal of Financial Economics* **62:2**, 293–325.

Kapetanios, G. and Marcellino, M. (2002), A Comparison of Estimation Methods for Dynamic Factor Models of Large Dimensions, draft, Bocconi University.

Stock, J. H. and Watson, M. W. (1998), Diffusion Indexes, NBER Working Paper 6702.

Stock, J. H. and Watson, M. W. (2001), Forecasting Output and Inflation: the Role of Asset Prices, NBER Working Paper 8180.

Stock, J. H. and Watson, M. W. (2002), Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business and Economic Statistics*.

Watson, M. W. (2000), Macroeconomic Forecasting Using Many Predictors, mimeo, Princeton University.