NBER WORKING PAPER SERIES

ARE MORE DATA ALWAYS BETTER FOR FACTOR ANALYSIS?

Jean Boivin
Serena Ng

Working Paper 9829
http://www.nber.org/papers/w9829

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2003

The authors would like to thank National Science Foundation for financial support (SES-0136923 and SES-0214104). We thank the editor and two anonymous referees for useful comments and Pritha Mitra for useful research assistance. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research

Are More Data Always Better for Factor Analysis?
Jean Boivin and Serena Ng
NBER Working Paper No. 9829
July 2003
JEL No. E37, E47, C3, C53

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 the 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 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. Weighting the data by their properties when constructing the factors also lead to improved forecasts. Our simulation analysis is unique in that special attention is paid to cross-correlated idiosyncratic errors, and we also allow the factors to have stronger loadings on some groups of series than others. It thus allows us to better understand the properties of the principal components estimator in empirical applications.

Jean Boivin
Graduate School of Business
821 Uris Hall
Columbia University
New York, NY
and NBER
jb903@columbia.edu

Serena Ng
Department of Economics
University of Michigan
317 Lorch Hall
Ann Arbor, MI 48109-1022
serena.ng@umich.edu

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 the 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 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.

A factor model with mutually uncorrelated idiosyncratic errors is a 'strict factor model', and use of these models is not new. However, the new generation of 'large dimensional approximate' 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 is large in both the cross-section (N) and the time (T) dimensions. Allowing the errors to be correlated makes the framework suited for a wider range of economic applications. The large dimensional nature of the 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.

In applications, approximate factor models are typically estimated by the method of principal components, by which an eigenvalue decomposition of the sample covariance matrix (the static approach) or the spectral density matrix (the dynamic approach). The results thus far are encouraging is performed. 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 (2002b), 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 (2001a) 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 modeling 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 VARs 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. By different choices of the data, two researchers using the same estimator can end up with different 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. The asymptotic theory under which the method of principal components is based assumes that the cross-correlation in the errors is not too large, and that the variability of the common component is not too small. In practice, our data are typically drawn from a small number of broad categories (such as industrial production, prices, interest rates). Think of ordering the series within a category by the importance of its common component, and put together a dataset comprising of high ranked series from each category. Now expand this dataset by adding the lower ranked, or 'noisy' series. Two things will happen. The average size of the common component will fall as more series are added, and the possibility of correlated errors will increase as more series from the same category are included. When enough of the 'noisy' series are added, the average common component will be smaller, and/or the residual cross-correlation will eventually be larger than that warranted by theory, creating a situation where more data might not be desirable.

The objective of this paper is to provide an empirical assessment of the extent to which the properties of the data affect the factor estimates. To our knowledge, this paper is the first to focus on the finite sample properties of the principal component estimator in the presence of cross-section correlation in the idiosyncratic errors, which is a pervasive feature of the data. In the empirical application considered, the errors of 115 out of the 147 series have correlation coefficients larger than .5. 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 (2002b) to obtain standard (unweighted) factor estimates. We then consider procedures that weigh or drop some of 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

Suppose we are interested in 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

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

with forecast error variance $\widehat{\sigma}_p^2$, where $\widehat{\gamma}_j, j = 0, \dots, p$ are the least squares estimates. 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.

Now assume X_{it} admits a factor structure:

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

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, $\chi_{it} = \lambda_i^{0\prime} F_t$ is the common component of series i, and e_{it} is an idiosyncratic error with $E(e_i^2) = \sigma_i^2$. If we observe the factors F_t^0, we can consider the forecast:

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

whose forecast error variance is $\widehat{\sigma}^2_{\widehat{\varepsilon},0}$. 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 F_t^0 using data from N series. Then the feasible factor-augmented forecast, referred to by Stock and Watson (2002a) as a 'diffusion index' forecast, is

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

with forecast error $\widehat{\sigma}^2_{\widehat{\varepsilon}}$. Now the difference between the diffusion index forecast and the naive AR(p) forecast is $\widehat{\sigma}^2_{\widehat{\varepsilon}} - \widehat{\sigma}^2_p = \left[\widehat{\sigma}^2_{\widehat{\varepsilon}} - \widehat{\sigma}^2_{\widehat{\varepsilon},0}\right] + \left[\widehat{\sigma}^2_{\widehat{\varepsilon},0} - \widehat{\sigma}^2_p\right]$. 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 the generated regressors $\widehat{F}_{t,N}$. In finite samples, the desirability of a feasible diffusion index forecast depends crucially on the estimates of F_t^0. We follow the literature and consider the method of principal components.[1]

Let Σ_X and Σ_χ be the population variance of the data and of the unobserved common components associated with N observations, respectively. Let Ω be the covariance matrix of the idiosyncratic errors. These can be thought of as N-dimensional sub-matrices of the infinite dimensional population covariances. A factor model has population covariance structure $\Sigma_X = \Sigma_\chi + \Omega$. As F_t^0 is common to all variables, Σ_χ has r non-zero eigenvalues, and they increase with N. A fundamental feature of factor models is that the r largest eigenvalues of Σ_X also increase with N. This suggests that the space spanned by the factors can be estimated using an eigenvalue decomposition of the the sample estimate of Σ_X. Denote by $\widehat{\lambda}_i' = (\widehat{\lambda}_{i1} \ldots \widehat{\lambda}_{ir})$ the estimated loadings, and let $\widehat{F}_{t,N} = (\widehat{F}^1_{t,N} \ldots \widehat{F}^r_{t,N})'$ be the estimated factors. Let $v_j = (v_{1j}, \ldots v_{Nj})'$ be the eigenvector corresponding to the j^{th} largest eigenvalue of the $N \times N$ sample covariance 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 (2002a) and 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 correlation both serially and cross-sectionally. The

[1] Kapetanios and Marcellino (2002) compared the properties of the static and the dynamic principal components estimator. As both are aimed at large dimensional panels, the issues to be discussed are relevant to both. Our discussion follows the simpler static principal components estimator.

first condition can be relaxed.[2] 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 cross-section idiosyncratic variance.[3] Bai (2003) further showed that $\widehat{F}_{t,N}$, suitably scaled, is $\sqrt{N}$ consistent for F_t^0. The various asymptotic results lead to the natural presumption 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 ($\lambda_i^0 = \lambda \ \forall i$). 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 *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, F_t can be estimated 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_1^2, 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 $\sigma_1^2 < \sigma_2^2$. 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 $\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N^2} > \frac{\sigma_1^2}{N_1}$. Whether or not more data yield more efficient estimates depend very much on the properties of the additional series.

Indeed, this intuition extends to factor analysis. Consider the special case when a researcher unintentionally included N_1 series twice. More precisely, there are $N = 2N_1$ series, but 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.

It is also useful to see why the properties of the errors are important from a different viewpoint.

[2] Bai and Ng (2001) showed uniform consistency 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.

[3] 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.

In a strict factor model, Ω is assumed to be a diagonal matrix. With N fixed, the maximum likelihood estimates of λ for arbitrary $\widetilde{\Omega}$ (a diagonal matrix) are the eigenvectors of $\widetilde{\Omega}^{-1/2}\widehat{\Sigma}_X\widetilde{\Omega}^{-1/2}$.[4] The principal components estimates, on the other hand, are the eigenvectors of $\widehat{\Sigma}_X$, which evidently correspond to the maximum likelihood estimates only when Ω is a scalar matrix. Accordingly, for a given N, the principal components estimator can be expected to be less precise the further are the errors from being homoskedastic and mutually uncorrelated.

3 Monte Carlo Simulations

In this section, we set up two monte carlo experiments to study how the data might affect the factor estimates. The series to be forecasted in both monte carlos are generated by

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

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, 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). Assumptions on e_{it} will be made precise below.

Our monte carlo experiments are designed to focus on the effects of heteroskedasticity and cross-correlated errors on the factor estimates. This is based on two considerations. First, previous simulation studies on the principal components estimator generally assume *iid* errors (see Kapetanios and Marcellino (2002) and Forni et al. (2000), for example). Stock and Watson (2002a) and Bai and Ng (2002) considered a case where an error is correlated with the one ordered before and after it. But as we will see, the cross-correlation found in the data is more substantial. Second, previous simulation studies tend to assume the error variances are constant across i. In the data, the variation in the errors is more substantial.

3.1 Model 1: Correlated and Noisy Errors

In this monte carlo experiment, the factors are assumed to be *iid*. The total number of series available is $N = N_1 + N_2 + N_3$. With $u_{it} \sim N(0,1)$, $i = 1, \ldots, N$ as the building block, we consider

[4] See Anderson (1984), p. 589.

three types of idiosyncratic errors:

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

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

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

The first N_1 series are what we call 'clean' series as their errors are mutually uncorrelated. The next N_2 series also have mutually uncorrelated errors, but differ from the N_1 series because $\sigma_2^2 > \sigma_1^2$. Each N_3 series is correlated with some C series that belong to the N_1 set. The $N_1 \times N_3$ matrix Ω_{13} has $C \times N_3$ non-zero elements and is the source of cross-correlation. More precisely, series $i \in [N_1 + N_2 + 1, N_1 + N_2 + N_3]$ is correlated with series $j \in [1, N_1]$ with coefficient ρ_{ij}. Since two series in the N_3 group can be correlated with the same series in the N_1 group, the errors of the N_3 type can also be mutually correlated. To isolate cases with cross-correlated errors from cases with large error variances, we let $\sigma_1 = \sigma_3 < \sigma_2$.[5] These assumptions on the idiosyncratic errors yield an Ω with the following property:

$$\Omega = \begin{bmatrix} \sigma_1^2 I_{N_1} & 0_{N_1 \times N_2} & \Omega_{13} \\ 0_{N_2 \times N_1} & \sigma_2^2 I_{N_2} & 0_{N_2 \times N_3} \\ \Omega_{13}' & 0_{N_3 \times N_2} & \Omega_{33} \end{bmatrix}.$$

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. We let $N_3 = n_3 N_1$. We vary n_3 and draw ρ_{ij} from a uniform distribution with a lower bound of .05 and an upper bound of .7. The σ_i^2 are chosen such that the factors explain κ_i of the variation in the data, given χ_{it}, with κ_3 set to .01.[6] We consider three N_1 values, five N_2 values, ten pairs of (N_3, C), nine sets of $(\kappa_1, \kappa_2, \kappa_y)$. Each of the 12,150 configurations is simulated M=1000 times.

Let x_{it} be the standardized data (with mean zero and unit variance) to which the method of principal components is used to extract k factors, where k can be different from r. This yields $\widehat{F}_t$, $\widehat{\chi}_{it}$, and $\widehat{e}_{it}$. Chamberlain and Rothschild (1983) showed that asset prices have an approximate factor structure if the largest eigenvalue (and hence all of the eigenvalues) of $\Omega = E(e_t e_t')$ is bounded. But the largest eigenvalue of Ω is bounded by $\max_i \sum_{j=1}^{N} |\tau_{ij}|$, where $\tau_{ij} = E(e_{it} e_{jt})$. Thus, under the assumptions of an approximate factor model, there should exist a P such that $\sum_{j=1}^{N} |\tau_{ij}| \leq P < \infty$ for all i and for all N. While P is useful for the development of theory, it does not provide a practical guide of how much cross-correlation is permitted in practice. Consider $\widehat{\tau}_i^* = \sum_{j=1}^{N} |\widehat{\tau}_{ij}|$,

[5] 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.

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

where $\widehat{\tau}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\widehat{e}_{it}\widehat{e}_{jt}$. We then use $\widehat{\tau}^{*} = \frac{1}{N}\max_i \widehat{\tau}_i^{*}$ as an indicator for P/N. This quantity should be small and decrease with N.

To assess the relative importance of the common component, we consider

$$R^2 = \frac{1}{N}\sum_{i=1}^{N} R_i^2, \quad \text{and} \quad R_q = R^2_{.9N} - R^2_{.1N}$$

where $R_i^2 = \frac{\sum_{t=1}^{T}\widehat{\chi}_{it}^2}{\sum_{t=1}^{T}x_{it}^2}$ indicates the relative importance of the common component in series i. The average of R_i^2 measures the average importance of the common component in the data as a whole. The cross-section dispersion of the common component is measured by R_q, the difference between the R_i^2 in the 90 and the 10 percentile.

Given $\widehat{F}_t$, β is estimated by OLS, and the diffusion index forecast $\widehat{y}_{\widehat{F},t+1|t}$ is obtained. We use three statistics to gauge the properties of the factor estimates and the forecasts:

$$\begin{aligned}
S_{F,F0} &= \frac{tr(F^{0\prime}\widehat{F}(\widehat{F}^{\prime}\widehat{F})^{-1}\widehat{F}^{\prime}F^{0})}{tr(F^{0\prime}F^{0})} \\
S_{y,y0} &= 1 - \frac{\sum_{t=1}^{T}(\widehat{y}_{F^0,t+1|t} - \widehat{y}_{\widehat{F},t+1|t})^2}{\sum_{t=1}^{T}\widehat{y}^2_{F^0,t+1|t}} \\
S_{\widehat{\beta},\beta} &= \frac{\sum_{t=1}^{T}(y_{\widehat{F},t+1|t} - \widehat{y}_{\widehat{F},t+1|t})^2}{\sum_{t=1}^{T}y^2_{\widehat{F},t+1|t}}.
\end{aligned}$$

Since we can only identify the space spanned by the factors, the second factor need not coincide with the second estimated factor. We therefore project each of the true factors on all estimated factors. A small $S_{F,F0}$ thus indicates a small discrepancy between the space spanned by the actual and the estimated factors. Similarly, 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 $S_{\widehat{\beta},\beta}$ statistic assesses the feasible diffusion index forecasts relative to the conditional mean. This latter evaluation is possible (but not in an empirical setting) because the model that generates the data is known in the simulations.

Some summary statistics of the simulated data are given in the second panel of Table 1. Notably, the experimental design generates substantial variation 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.

Because of the large number of configurations involved, we summarize the results 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. 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 $S_{\widehat{\beta},\beta}$, 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. An additional factor reduces S_{F,F_0} by about 7 basis points. The mean-squared forecast error is also higher the larger the number of true factors, as having to estimate more factors inevitably increases sampling variability.

Adding series with relatively large idiosyncratic variances has 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 for $S_{y,y0}$ is also around 22.

Of special interest to us are two results. First, the common factors are more precisely estimated when the common component is important, as indicated by the coefficient on the variable labeled R^2 in Table 1. However, the larger the dispersion in the importance of the common component, as indicated by the effect on R_q, the less precise are the estimates. This suggests that adding data with large idiosyncratic errors or weak factor loadings need not be desirable. Second, S_{F,F_0} and $S_{y,y0}$ are both decreasing in N_3 and C, holding other parameters fixed. This suggests that the forecasts and the factor estimates are adversely affected by cross correlation in the errors. A summary statistic for the extent of cross correlation is $\widehat{\tau}^*$, since it depends on C, N_3, and ρ_{ij}. Evidently, increasing $\widehat{\tau}^*$ by .01 reduces S_{F,F_0} and $S_{y,y0}$ by .86 and .90 basis points, respectively.[7]

The present monte carlo exercise highlights the fact that 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. The factor estimates and forecasts are clearly less efficient when the errors are cross correlated and/or have vastly unequal idiosyncratic error variances.

3.2 Model 2: Oversampling

In empirical work, we almost always work with only a subset of the data available. To understand if it matters which N series are being used for analysis, we simulate data from a strict factor model in this subsection. We assume that there are two serially correlated factors driving the data, viz: $X_{it} = \lambda_{i1} F_{1t} + \lambda i 2 F_{2t} + e_{it}$, with

$$F_{mt} = .5F_{mt-1} + u_{mt}, \quad u_{mt} \sim N(0,1), \quad m = 1, 2..$$

[7] The second order effect is positive, but numerically small. Evaluated at $\widehat{\tau}^* = .1$, the second order effect is 1.81.

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

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

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

N_1: $X_{it} = .8F_{1t} + e_{it}, \sigma^2_i \sim N(0, 1 - .8^2)$;

N_2: $X_{it} = .6F_{2t} + e_{it}, \sigma^2_i \sim N(0, 1 - .6^2)$;

N_3: $X_{it} = .4F_{1t} + .1F_{2t} + e_{it}, \sigma^2 \sim N(0, 1 - .4^2 - .1^2)$;

N_4: $X_{it} = .1F_{1t} + .4F_{2t} + e_{it}, \sigma^2 \sim N(0, 1 - .1^2 - .4^2)$;

N_5: $X_{it} = e_{it}, \sigma^2_i \sim N(0, 1)$.

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 factor one is real and factor two 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 that in 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 with two or more factors estimated was there a noticeable improvement in S_{F,F_0}. With 100 series, S_{F,F_0} improves to only .955 in case 9. Thus as in the previous monte carlo, efficiency of the factor estimates is determined not simply by whether the sample size is 40 or 100, but also by the informativeness of the data about the factors.

One motivation for the present monte carlo is to highlight the fact that the factor space being estimated depends on the choice of data. In case 1 when the N_1 series was used, the first principal component estimates the space spanned by F_1. For this reason, extracting one factor given the N_1 dataset is adequate for forecasting y^A. Analogously, extracting one factor from the N_2 dataset is adequate for the purpose of forecasting y^B. However, if the first factor dominates the variation of the second, we will need to estimate two factors from $N_1 + N_2$ series to forecast y^B efficiently. Analogously, if we had data in which F_2 dominates F_1, such as case 4, 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 factors 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 $\sigma_\varepsilon^A = \sigma_\varepsilon^B$ 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 X will tend to be inferior to those that depend on the dominant factors in X.

As noted earlier, macroeconomic panels are 'put together' by the researcher, and as such, the factors are always sample dependent. As 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 to see if $\widehat{y}_{t+1|y_t,\ldots y_1,\widehat{F}_{t,N}}$ depends on N in real-time, 12 month ahead forecasting of a large number of economic time series with special attention to eight commonly studied economic indicators: 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 scaled 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 $k = 6$ factors.[8] We then use the BIC as suggested in Stock and Watson (2002b) to determine the number of factors used in the forecasting equation with ω set to 0.001.

The base case of X_{it} is a balanced panel of N=147 monthly series available from 1959:1 to 1998:12. Following Stock and Watson (2002b), 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 relative importance of the common component for each series are denoted $R^2_i(3)$ and $R^2_i(6)$, when three and six factors are being estimated respectively.

The forecasting exercise begins with 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 evaluate the k factor diffusion index forecasts relative to those of the AR(4) forecasts (i.e. with zero factors). The results, when k is chosen optimally by the BIC, are reported in the first row of columns 5 to 12 of Table 3. Incidentally, the BIC usually suggests two factors. An entry less than one indicates that the diffusion index forecast is superior to the naive AR(4) forecast. The results confirm the findings of Stock and Watson that the diffusion indices can be useful for real time forecasting, even though the factors have to be estimated.

[8] The main difference between our analysis and that of Stock and Watson is that we did not allow lags of the factors to enter the forecasting model.

But can more efficient factor augmented forecasts be obtained? A look at the properties of the data reported in Appendix A reveal several features. First, there are more series from some groups than others, so the problem of oversampling is conceivable. Second, many of the R_i^2s are very small. For example, three factors explain only .01 of the variation in IPUT (series 15). 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 is thus quite large. Our monte carlo results suggest that such dispersion can have adverse effects on the forecasts.

Third, there is substantial cross correlation in the idiosyncratic errors. To gauge the problem, we obtain $\widehat{\tau}_{ij}$, the correlation coefficient between the residuals for series i and j, obtained from estimation of a six (and three) factor model over the entire sample, 71:1-97:12. For each series i, we can identify

$$\widehat{\tau}_1^*(i) \quad = \quad \max_j |\widehat{\tau}_{ij}| = \widehat{\tau}_{ij_i^1}.$$

That is, j_i^1 is the series whose idiosyncratic error is most correlated with series i, and the correlation between series i and j_i^1 is $\widehat{\tau}_1^*(i)$. 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 have a correlation coefficient of .99 (see Appendix A). Evidently, the maximum residual cross correlation in the data is non-trivial. As the maximum correlation coefficient could be an outlier, we also report the second largest residual cross correlation for each series (see the last two and three columns of Appendix A). That is, we identify j_i^2. Then $\widehat{\tau}_2^*(i)$ is the second largest residual correlation for series i. Many of these correlation coefficients remain quite high.

As a final check, the quantity $\sum_j |\tau_{ij}|$ should be bounded under the assumptions of the approximate factor model. These are reported in the last column of Appendix A. This series has a mean of 14.62 and a standard deviation of 5.10. As we have 147 series in the analysis, the average cross-correlation is around .1. In many cases, $\sum_j |\widehat{\tau}_{ij}|$ is large even though $\widehat{\tau}_1^*(i)$ is not. See, for example, series LPCC and FSPUT. This suggests many of the τ_{ij} are non-zero. In particular, the idiosyncratic errors of the data in groups 1 (industrial production) and 7 (manufacturing series) exhibit especially strong correlation. These results suggest that 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 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 span the same space as those of $\widehat{\Sigma}_X$. There is, to our knowledge, no obvious way to exploit the entire $\widehat{\Omega}$ matrix to improve efficiency.

Although optimal weighting is not possible, some form of weighting may still be desirable. 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 the objective function weighs the variance of each series; the covariances are not weighted as in feasible GLS. When N is large, this can be advantageous because having to estimate N instead of $N(N+1)/2$ weights induces less sampling variability.

As w_{iT} is meant to be data dependent, we rely on a first step estimation of the factor model to get the estimated residuals. If the number of factors is too small in this step, the errors will be correlated by construction which could lead to inaccurate weighting. The base case of our first step extracts six factors from X, from which the residuals are used to construct $\widehat{\Omega}$ and the weights as defined below. For robustness check, we also consider using three factors in the first step. In the second step, a new set of factors is estimated using different weighted criteria. After the factors are re-estimated, the BIC is used to determine how many of them are used in forecasting. This step is repeated for each of the eight y series under investigation.

We consider the following sets of weights:

Rule SWa: w_{iT} is the inverse of the i^{th} diagonal element of $\widehat{\Omega}_T$, estimated using data up to time T.[9]

Rule SWb: w_{iT} is the inverse of $\frac{1}{N}\sum_{j=1}^{N}|\widehat{\Omega}_T(i,j)|$.

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

[9] One can also think of weighting each series by $\frac{\widehat{\omega}_i^{-1}}{\sum_{i=1}^{N}\widehat{\omega}_i^{-1}}$. This gives the same relative weight between series i and j as the one we considered, namely $\frac{\widehat{\omega}_j}{\widehat{\omega}_i}$.

Rule 1c The j^1 set in Rule 1 is based on $\widehat{\Omega}_T$, the full sample estimate of Ω. In real time, this estimate can be updated continuously. Rule 1c uses $\widehat{\Omega}_t$ to obtain w_{it}.

Rule 2c Follows Rule 2 but allows for continuous updating as in Rule 1c.

Rule SWa was also considered in Jones (2001). It is aimed at accounting for heteroskedasticity in the errors, but not cross-correlation. Rule SWb weights the data by the magnitude of residual cross-correlation. With Rules SWa and SWb, all 147 series are used to estimate the factors as the weights are never exactly zero. In contrast, w_{iT} is a binary variable under Rules 1 and 2. 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 the residuals. However, our second monte carlo exercise also suggests that it may be more efficient to estimate the factors from a small dataset than from one with additional series that does not contain information about the factor. Inspection of Appendix A reveals that while many series (in particular, the real variables) are quite well explained by the first three factors, many series (in particular, the price variables) are better explained by factors three through six. For example, the $R_i^2(6)$ for the price variables is generally much higher than $R_i^2(3)$. In light of the results in our second monte carlo experiment, the 147 series are reclassified into three categories. The 'real' (R) category consists of 60 series from output, employment, retail and manufacturing trade, consumption, and miscellaneous. The 'nominal' (N) category consists of 46 series relating to exchange rate, interest rates, money and credit, and price indexes. The 'volatile/leading indicator' (V) category consists of 41 series of high volatility, including housing starts, inventories, orders, and stock prices.[10] An appeal of this grouping is that it provides some economic interpretation to the factors. Obviously, this amounts to having three additional 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 nominal factors from the nominal variables, and three volatile factors 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:

$$\text{Rule A: } F_1^R, F_2^R, F_3^R \qquad \text{Rule B: } F_1^N, F_2^N, F_3^N$$
$$\text{Rule C: } F_1^V, F_2^V, F_3^V \qquad \text{Rule D: } F_1^R, F_1^N, F_1^V;$$

As is evident, Rule A uses only the real factors, Rule B uses only the nominal factors, while Rule C uses only the volatile factors. Rule D eventually has one factor from each category.

[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.

There is undoubtedly a certain ad-hocness in all these rules, but if cross-correlated errors are not prevalent in the data, or if oversampling of the data 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.

The results with weights obtained from a first step estimation of 6 and 3 factors are reported in Table 3 and 4, respectively. The real factors (Rule A) apparently have little predictive power for the real variables. This is perhaps to be expected since the forecasting equation already has four lags of each series, which are themselves real variables. However, the nominal factors, extracted from 46 series (Rule B), are extremely effective for forecasting all real series. It beats the forecasts from the factors extracted from all 147 series (Rule SW). This result can be explained by the fact that the first factor in the Stock and Watson data is understood to be a real factor. This means that the nominal factors are not the most important factors in the panel of 147 series. As suggested by our second monte carlo, extracting these factors from a large panel is inferior to extracting them from data in which the factor is dominant, which is the case with Rule B. This would be consistent with real series being over-sampled in the data set.

Turning now to the inflation series, using factors associated with Rule B are evidently uninformative, as lagged prices have likely encompassed 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 forecasts with 147 series are closely matched by those with factors extracted from 33 series, ie Rule 2.

Overall, Rules 2 and SWb produce results that are comparable, and often times better, than SW for all eight series considered. As these results are specific to a selected set of series, one might wonder whether our findings are general. To address this issue, we repeat the forecasting exercise for each of the 147 series in X and assess the performance of each rule. The results from averaging over all the real and all the nominal series are reported in Tables 3 and 4. Rule B continues to do well for the real series, though SWb is not far behind, suggesting that reducing the extent of cross-correlation in the errors can improve the factor forecasts of the real variables. As for the nominal variables, the weighting schemes did not improve the SW forecasts, but they did no worse either, suggesting again that using a smaller number of series to construct the factors could have been adequate. We also use two graphs to summarize the 147 forecasts. Figure 1 shows the percentage of series for which a given rule is the best one. While SW is best in about 10 percent of the cases, Rules SWa, SWb and B are better more often. As none of them systematically perform better than all the others, it is not clear from Figure 1 alone which rule one should prefer. Therefore in Figure

2, we present the percentage of series for which a given rule beats SW. In about 65 percent of the series, SWb outperforms SW.

While we make no claim that these rules are optimal, two observations are useful to highlight. First, the fact that Rules 2 and B used fewer than 50 series 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 possibility of 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.

Second, instead of dropping series with highly correlated errors, we can also downweigh their influence on the objective function. This is perhaps a more appealing way of dealing with the intrinsic properties of the data from a statistical perspective because no information is wasted. This is what rules SWa and SWb attempt to accomplish, and as we can see, with encouraging outcomes. We can expect more formal weighting schemes than the crude ones used here to be able to further enhance the properties of principal components estimator.

5 Conclusion

A feature stressed in recent applications of factor models is the use of data from 'large' panels. 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. But in simulations and the empirical examples considered, the factors extracted from as few as 40 series seem to do no worse, and in many cases, better than the ones extracted from 147 series.

In applications, the number of series available for analysis can be quite large. Suppose we have 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, plus a small number of aggregated ones. Should we use all series? What are the consequences 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 suggest that sample size alone does not determine the properties of the estimates. The quality of the data must be taken into account. There is room to further exploit the diffusion index forecasting technology by efficiently incorporating information about the properties of the data in the construction of the factors.

Appendix I: Data

Series	name	tcode	group	$R_i^2(3)$	$R_i^2(6)$	j_1	j_2	$\widehat{\tau}_1^*$	$\widehat{\tau}_2^*$	$\sum_i \lvert\widehat{\tau}_{ij}\rvert$
1	IP	5	1	0.70	0.73	11	9	0.89	0.81	25.53
2	IPP	5	1	0.62	0.68	3	4	0.93	0.78	22.74
3	IPF	5	1	0.55	0.61	2	4	0.93	0.81	20.69
4	IPC	5	1	0.41	0.48	3	2	0.81	0.78	20.52
5	IPCD	5	1	0.37	0.45	4	3	0.66	0.57	17.34
6	IPCN	5	1	0.14	0.16	4	13	0.69	0.66	14.77
7	IPE	5	1	0.44	0.48	3	2	0.55	0.47	11.33
8	IPI	5	1	0.42	0.45	11	1	0.41	0.41	14.24
9	IPM	5	1	0.51	0.53	1	11	0.81	0.67	17.47
10	IPMND	5	1	0.36	0.37	13	36	0.61	0.27	11.89
11	IPMFG	5	1	0.71	0.75	1	12	0.89	0.85	25.83
12	IPD	5	1	0.62	0.67	11	1	0.85	0.78	22.57
13	IPN	5	1	0.41	0.43	6	10	0.66	0.61	17.42
14	IPMIN	5	1	0.05	0.05	9	1	0.47	0.34	8.48
15	IPUT	5	1	0.01	0.08	57	45	0.45	0.24	10.44
16	IPXMCA	1	1	0.70	0.77	43	42	0.63	0.63	11.69
17	PMI	1	1	0.75	0.77	77	18	0.87	0.85	18.43
18	PMP	1	1	0.68	0.71	17	77	0.85	0.82	15.91
19	GMYXPQ	5	1	0.38	0.39	140	144	0.27	0.25	11.97
20	LHEL	5	2	0.32	0.36	21	30	0.69	0.18	9.66
21	LHELX	5	2	0.45	0.49	20	138	0.69	0.21	11.14
22	LHEM	5	2	0.24	0.27	23	33	0.90	0.27	9.25
23	LHNAG	5	2	0.28	0.30	22	9	0.90	0.20	9.07
24	LHUR	1	2	0.49	0.72	43	42	0.72	0.67	14.57
25	LHU680	1	2	0.41	0.63	28	29	0.78	0.60	16.86
26	LHU5	1	2	0.38	0.82	27	67	0.72	0.32	14.32
27	LHU14	1	2	0.53	0.89	26	29	0.72	0.62	19.69
28	LHU15	1	2	0.50	0.83	29	25	0.88	0.78	19.59
29	LHU26	1	2	0.56	0.87	28	27	0.88	0.62	18.71
30	LPNAG	5	2	0.72	0.76	31	37	0.93	0.71	16.52
31	LP	5	2	0.70	0.77	30	32	0.93	0.67	16.28
32	LPGD	5	2	0.72	0.77	34	31	0.74	0.67	16.42
33	LPCC	5	2	0.22	0.29	32	31	0.42	0.34	13.24
34	LPEM	5	2	0.69	0.72	35	32	0.92	0.74	16.34
35	LPED	5	2	0.62	0.65	34	32	0.92	0.67	16.52
36	LPEN	5	2	0.44	0.46	34	32	0.41	0.35	11.45
37	LPSP	5	2	0.39	0.41	30	31	0.71	0.57	12.74
38	LPTX	5	2	0.35	0.39	37	31	0.55	0.39	12.11
39	LPFR	5	2	0.17	0.20	62	78	0.31	0.31	10.76
40	LPS	5	2	0.20	0.21	37	30	0.45	0.33	9.79
41	LPGOV	5	2	0.11	0.21	37	30	0.47	0.31	7.30
42	LPHRM	1	2	0.20	0.21	43	24	0.93	0.67	12.87
43	LPMOSA	1	2	0.10	0.16	42	24	0.93	0.72	13.16
44	PMEMP	1	2	0.80	0.80	17	18	0.73	0.64	15.11
45	MSMTQ	5	3	0.62	0.77	72	46	0.80	0.64	20.54
46	MSMQ	5	3	0.58	0.64	73	47	0.91	0.87	16.96
47	MSDQ	5	3	0.53	0.59	46	73	0.87	0.78	15.68
48	MSNQ	5	3	0.23	0.26	86	46	0.66	0.45	11.89
49	WTQ	5	3	0.18	0.32	51	74	0.85	0.82	14.06
50	WTDQ	5	3	0.30	0.34	49	74	0.58	0.49	11.75
51	WTNQ	5	3	0.04	0.18	49	74	0.85	0.70	11.02
52	RTQ	5	3	0.20	0.33	75	54	0.79	0.61	15.50
53	RTNQ	5	3	0.07	0.16	56	52	0.81	0.59	13.15
54	GMCQ	5	4	0.18	0.36	55	56	0.75	0.61	16.81
55	GMCDQ	5	4	0.13	0.22	58	54	0.86	0.75	13.41
56	GMCNQ	5	4	0.07	0.19	53	54	0.81	0.61	12.53
57	GMCSQ	5	4	0.06	0.14	15	54	0.45	0.26	7.48
58	GMCANQ	5	4	0.10	0.19	55	54	0.86	0.57	11.06
59	HSFR	4	5	0.42	0.61	64	62	0.87	0.80	21.58
60	HSNE	4	5	0.36	0.42	59	43	0.55	0.49	13.92
61	HSMW	4	5	0.44	0.47	59	63	0.54	0.37	14.81
62	HSSOU	4	5	0.22	0.59	59	64	0.80	0.76	18.85
63	HSWST	4	5	0.22	0.41	59	64	0.77	0.72	15.93
64	HSBR	4	5	0.33	0.57	59	62	0.87	0.76	20.62
65	HMOB	4	5	0.08	0.34	62	64	0.51	0.41	13.63
66	IVMTQ	5	6	0.43	0.44	67	68	0.59	0.55	16.14
67	IVMFGQ	5	6	0.40	0.43	68	66	0.89	0.59	13.12
68	IVMFDQ	5	6	0.37	0.39	67	66	0.89	0.55	12.15
69	IVMFNQ	5	6	0.12	0.17	67	28	0.48	0.24	8.06
70	IVWRQ	5	6	0.12	0.12	66	74	0.47	0.41	7.93
71	IVRRQ	5	6	0.09	0.11	75	66	0.61	0.51	10.69
72	IVSRQ	2	6	0.67	0.79	45	85	0.80	0.59	21.99
73	IVSRMQ	2	6	0.64	0.67	46	47	0.91	0.78	17.49
74	IVSRWQ	2	6	0.19	0.31	49	51	0.82	0.70	13.27
75	IVSRRQ	2	6	0.12	0.24	52	71	0.79	0.61	16.10
76	PMNV	1	6	0.61	0.62	17	78	0.52	0.39	11.88

Notes: tcode is the transformation code, taken from Stock and Watson (2002a). 1=no transformation, 2=first difference, 4=logarithm, 5=first difference of logarithms, 6=second difference. The data can be classified into 13 groups, 1=real output and income, (2)=employment and hours, (3)=retail and manfacturing trade, (4)=consumption,

Appendix I: continued

Series	name	tcode	group	$R_i^2(3)$	$R_i^2(6)$	j_1	j_2	$\widehat{\tau}_1^*$	$\widehat{\tau}_2^*$	$\sum_i \lvert\widehat{\tau}_{ij}\rvert$
77	PMNO	1	7	0.66	0.71	17	18	0.87	0.82	16.53
78	PMDEL	1	7	0.52	0.57	17	76	0.63	0.39	14.84
79	MOCMQ	5	7	0.55	0.58	47	46	0.51	0.42	13.21
80	MDOQ	5	7	0.50	0.61	84	85	0.99	0.91	24.79
81	MSONDQ	5	7	0.12	0.21	91	92	0.83	0.83	17.44
82	MO	5	7	0.63	0.73	84	80	0.90	0.88	26.15
83	MOWU	5	7	0.44	0.55	85	84	0.98	0.90	24.69
84	MDO	5	7	0.50	0.61	80	85	0.99	0.92	24.99
85	MDUWU	5	7	0.39	0.50	83	84	0.98	0.92	24.38
86	MNO	5	7	0.36	0.39	48	87	0.66	0.52	14.52
87	MNOU	5	7	0.18	0.21	86	90	0.52	0.42	9.66
88	MU	5	7	0.41	0.47	89	85	0.99	0.43	17.21
89	MDU	5	7	0.39	0.46	88	85	0.99	0.44	16.84
90	MNU	5	7	0.22	0.25	87	134	0.42	0.21	10.36
91	MPCON	5	7	0.09	0.17	92	81	1.00	0.83	16.70
92	MPCONQ	5	7	0.08	0.16	91	81	1.00	0.83	16.64
93	FSNCOM	5	8	0.28	0.64	94	95	0.98	0.96	22.10
94	FSPCOM	5	8	0.28	0.64	95	93	0.99	0.98	22.23
95	FSPIN	5	8	0.27	0.62	94	93	0.99	0.96	21.45
96	FSPCAP	5	8	0.22	0.54	95	94	0.87	0.84	18.44
97	FSPUT	5	8	0.24	0.45	93	94	0.33	0.32	14.01
98	FSDXP	2	8	0.30	0.64	95	94	0.77	0.77	18.72
99	FSPXE	2	8	0.18	0.39	94	95	0.58	0.57	16.42
100	EXRUS	5	9	0.10	0.20	101	103	0.84	0.83	13.71
101	EXRGER	5	9	0.08	0.15	102	100	0.87	0.84	12.87
102	EXRSW	5	9	0.08	0.15	101	100	0.87	0.81	12.95
103	EXRJAN	5	9	0.06	0.15	100	102	0.83	0.57	10.87
104	EXRCAN	5	9	0.03	0.10	19	100	0.19	0.17	7.57
105	FYFF	2	10	0.23	0.26	106	94	0.30	0.28	11.08
106	FYGT5	2	10	0.26	0.45	107	108	0.94	0.77	17.68
107	FYGT10	2	10	0.23	0.43	106	108	0.94	0.83	17.65
108	FYAAAC	2	10	0.28	0.49	107	109	0.83	0.82	16.92
109	FYBAAC	2	10	0.32	0.50	108	107	0.82	0.69	15.40
110	FYFHA	2	10	0.22	0.33	107	106	0.60	0.60	14.32
111	FM1	6	11	0.05	0.10	112	113	0.62	0.45	7.69
112	FM2	6	11	0.04	0.08	113	111	0.64	0.62	9.15
113	FM3	6	11	0.02	0.04	112	111	0.64	0.45	7.73
114	FM2DQ	5	11	0.24	0.29	62	16	0.36	0.35	11.99
115	FMFBA	6	11	0.03	0.03	116	117	0.67	0.32	7.71
116	FMRRA	6	11	0.03	0.03	115	117	0.67	0.49	8.71
117	FMRNBC	6	11	0.05	0.08	116	115	0.49	0.32	7.91
118	PMCP	1	12	0.47	0.50	147	141	0.55	0.38	11.81
119	PWFSA	6	12	0.03	0.28	120	129	0.88	0.33	9.12
120	PWFCSA	6	12	0.03	0.30	119	129	0.88	0.30	9.23
121	PSM99Q	6	12	0.02	0.03	97	107	0.20	0.20	6.99
122	PUNEW	6	12	0.08	0.68	131	83	0.33	0.26	12.84
123	PU83	6	12	0.03	0.09	124	120	0.26	0.20	5.24
124	PU84	6	12	0.04	0.31	123	129	0.26	0.25	8.67
125	PU85	6	12	0.00	0.01	128	117	0.19	0.12	4.96
126	PUC	6	12	0.08	0.69	134	130	0.36	0.26	13.52
127	PUCD	6	12	0.01	0.04	133	134	0.28	0.24	7.32
128	PUS	6	12	0.02	0.05	129	134	0.32	0.32	8.10
129	PUXF	6	12	0.04	0.33	119	128	0.33	0.32	9.71
130	PUXHS	6	12	0.09	0.66	126	82	0.26	0.21	10.95
131	PUXM	6	12	0.08	0.62	122	128	0.33	0.22	12.62
132	GMDC	6	12	0.09	0.67	135	134	0.66	0.36	12.45
133	GMDCD	6	12	0.00	0.04	127	132	0.28	0.23	5.41
134	GMDCN	6	12	0.10	0.71	132	126	0.36	0.36	13.96
135	GMDCS	6	12	0.01	0.10	132	126	0.66	0.25	7.30
136	LEHCC	6	13	0.02	0.02	137	33	0.26	0.24	8.46
137	LEHM	6	13	0.01	0.02	35	136	0.31	0.26	8.81
138	SFYCP90	1	10	0.23	0.75	140	141	0.40	0.32	12.95
139	SFYGM3	1	10	0.51	0.83	140	141	0.83	0.46	16.57
140	SFYGM6	1	10	0.53	0.86	139	141	0.83	0.77	22.09
141	SFYGT1	1	10	0.52	0.80	140	142	0.77	0.62	19.82
142	SFYGT5	1	10	0.67	0.86	143	145	0.96	0.83	25.60
143	SFYGT10	1	10	0.69	0.86	142	144	0.96	0.91	25.90
144	SFYAAAC	1	10	0.70	0.83	145	143	0.92	0.91	24.65
145	SFYBAAC	1	10	0.74	0.85	144	143	0.92	0.88	23.35
146	SFYFHA	1	10	0.70	0.86	143	144	0.86	0.83	23.29
147	HHSNTN	1	10	0.38	0.58	118	25	0.55	0.40	10.65

(5)-housing starts and sales, (6)= inventories, (7)=orders, (8)=stock prices, (9)=exchange rate, (10)=interest rates, (11)=money and credit, (12)=prices, (13)=misc. $R_i^2(3)$ and $R_i^2(6)$ are the fraction of variation in series i explained by three and six factors, respectively. j_1 and j_2 are the series whose errors are most correlated and second most correlated with series i. The corresponding correlation coefficients are $\widehat{\tau}_1^*$ and $\widehat{\tau}_2^*$, respectively.

Table 1: Response Surface for Monte Carlo 1

	Estimates			Summary Statistics of Simulated Data			
	100 × S_{F,F_0}	100 × $S_{y,y0}$	100 × $S_{\widehat{\beta},\beta}$	mean	s.d.	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)				
$\widehat{\tau}^*$	-86.512	-90.361	134.357	.140	.058	.068	.452
	(9.30)	(7.22)	(13.15)				
$[\widehat{\tau}^*]^2$	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

We considered a total of 12150 configurations of Model 1, each simulated 1000 times. The parameters that we vary between configurations are N_1, N_2, N_3, C, r, k, κ_1, κ_2 and κ_y. We use R^2, R_q^2 and τ^* to summarize the properties of the data. Summary statistics for the simulated data are given in the second panel of Table 1. For each configuration, forecast performance is measured using three statistics: $S_{F,F0}$, $S_{y,y0}$, and MSE. We use a response surface to summarize the sensitivity of forecast performance to the properties of the factor model. The coefficients are reported in columns 2, 3, and 4.

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}	$S^A_{\widehat{\beta},\beta}$	$MSE^B_{\widehat{\beta},\beta}$
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 Relative to AR(4), 6 Factors: 71:1-97:12

rule	N	ip	gmyxspq	msmtq	lpnag	Real	punew	gmdc	puxx	pwfsa	Nominal
SW	147	0.632	0.906	0.580	0.919	0.759	0.734	0.832	0.843	0.825	0.808
1	71	0.756	1.006	0.694	0.986	0.860	0.745	0.833	0.827	0.837	0.810
2	33	0.615	0.835	0.576	0.867	0.723	0.740	0.843	0.844	0.818	0.811
1c	71	0.661	0.819	0.620	0.872	0.743	0.771	0.874	0.880	0.840	0.841
2c	33	0.715	0.801	0.640	0.853	0.752	0.794	0.869	0.876	0.864	0.851
SWa	147	0.625	0.813	0.569	0.618	0.656	0.831	0.912	0.934	0.908	0.896
SWb	147	0.594	0.921	0.564	0.797	0.719	0.743	0.848	0.845	0.810	0.811
A	60	1.009	1.008	1.012	0.983	1.003	0.789	0.863	0.853	0.868	0.843
B	46	0.585	0.753	0.543	0.645	0.632	1.024	1.075	1.100	0.983	1.045
C	41	1.019	1.055	1.035	0.963	1.018	0.800	0.893	0.841	0.882	0.854
D	60	0.612	0.750	0.530	0.813	0.676	0.789	0.863	0.890	0.868	0.852
AR(4)		0.050	0.027	0.046	0.017	0.035	0.021	0.016	0.019	0.035	0.023

The number of factors is selected using the information criterion proposed by Stock and Watson (2002b), with $\omega = 0.001$. SW is the base case using 147 series. Let τ_{ij} be cross-correlation between the residuals of series i and j from full sample estimation of a six factor model. Define $j_i^{*1} = \max_j |\tau_{ij}|$ to the series most correlated with series i. Rule 1 removes all series in $\{j_i^*\}$. For each i, we find the series with the second largest $|\tau_{ij}|$. Rule 2 additionally removes those series. Rules 1c and 2c are based on rolling estimation of the correlation matrix, so the series that are dropped can change as we roll the sample. Rules A, B, and C extract three real, three nominal, and three volatile factors factors from subsets of the data. Rule D uses one real, one nominal, and one volatile factor in the forecasting exercise.

Table 4: Forecast Errors Relative to AR(4), 3 Factors: 71:1-97:12

rule	N	ip	gmyxspq	msmtq	lpnag	Real	punew	gmdc	puxx	pwfsa	Nominal
SW	147	0.632	0.906	0.580	0.919	0.759	0.734	0.832	0.843	0.825	0.808
1	71	0.702	0.940	0.626	0.947	0.804	0.721	0.816	0.806	0.819	0.790
2	33	0.918	1.084	0.798	0.948	0.937	0.743	0.859	0.822	0.852	0.819
1c	71	0.701	0.890	0.612	0.901	0.776	0.752	0.834	0.860	0.845	0.822
2c	33	0.650	0.897	0.585	0.877	0.752	0.770	0.851	0.903	0.851	0.844
SWa	147	0.624	0.836	0.590	0.778	0.707	0.808	0.913	0.867	0.861	0.862
SWb	147	0.628	0.833	0.594	0.916	0.743	0.741	0.837	0.830	0.833	0.810
A	60	1.009	1.008	1.012	0.983	1.003	0.789	0.863	0.853	0.868	0.843
B	46	0.585	0.753	0.543	0.645	0.632	1.024	1.075	1.100	0.983	1.045
C	41	1.019	1.055	1.035	0.963	1.018	0.800	0.893	0.841	0.882	0.854
D	60	0.612	0.750	0.530	0.813	0.676	0.789	0.863	0.890	0.868	0.852
AR(4)		0.050	0.027	0.046	0.017	0.035	0.021	0.016	0.019	0.035	0.023

The number of factors is selected using the information criterion proposed by Stock and Watson (2002b), with $\omega = 0.001$. SW is the base case using 147 series. Let τ_{ij} be cross-correlation between the residuals of series i and j from full sample estimation of a three factor model. Define $j_i^{*1} = \max_j |\tau_{ij}|$ to the series most correlated with series i. Rule 1 removes all series in $\{j_i^*\}$. Rule 1 removes all series in $\{j_i^*\}$. For each i, we find the series with the second largest $|\tau_{ij}|$. Rule 2 additionally removes those series. Rules 1c and 2c are based on rolling estimation of the correlation matrix, so the series that are dropped can change as we roll the sample. Rules A, B, and C extract three real, three nominal, and three volatile factors factors from subsets of the data. Rule D uses one real, one nominal, and one volatile factor in the forecasting exercise.

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, *Econometrica* **70:1**, 191–221.

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

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.

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. (2001a), 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. (2001), Forecasting Output and Inflation: the Role of Asset Prices, *Journal of Economic Literature* **47:1**, 1–48.

Stock, J. H. and Watson, M. W. (2002a), Diffusion Indexes, *Journal of the American Statistical Association* **97:460**, 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.

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

Figure 1:
Fraction of series (over all 147) a rule outperforms all the others. Selection based on 6 factors

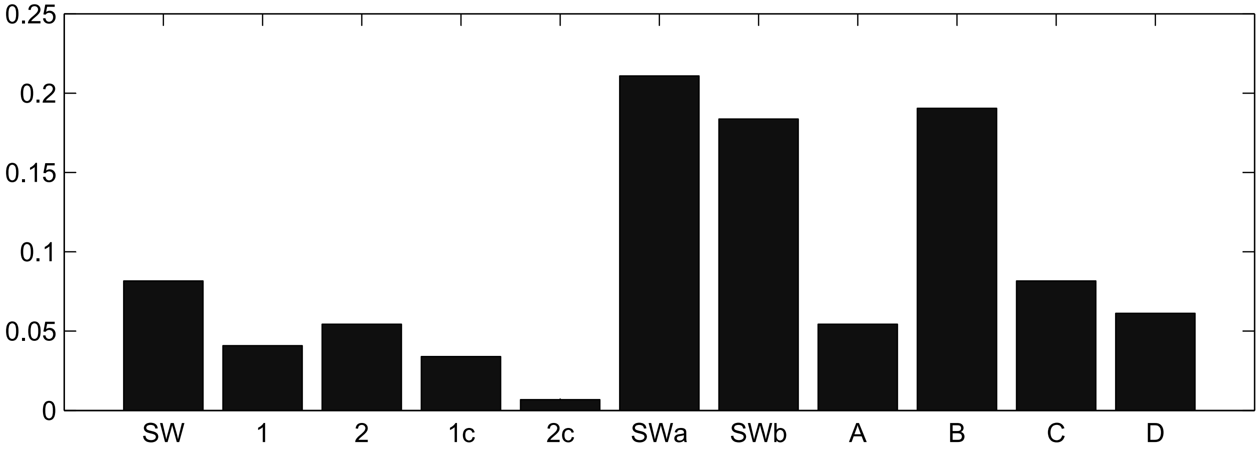
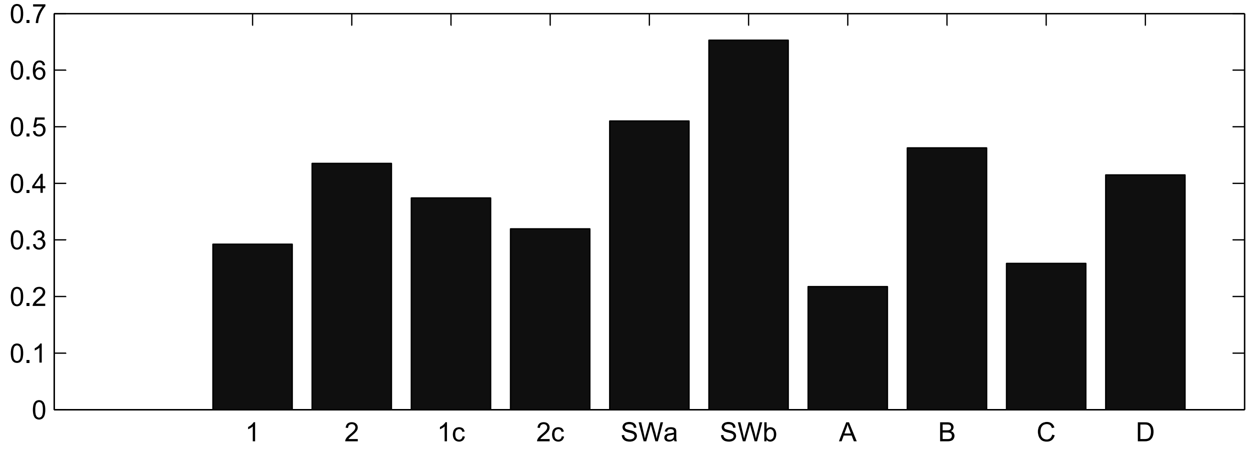


Figure 2:
Fraction of series (over all 147) a given rule outperforms SW. Selection based on 6 factors.


1c
2c
SWa
SWb
A
B
C
D