

Principal Component Analysis for Nonstationary Series

James D. Hamilton * Jin Xi †

University of California at San Diego

May 28, 2022

Revised: January 13, 2023

Abstract

This paper develops a procedure for uncovering the common cyclical factors that drive a mix of stationary and nonstationary variables. The method does not require knowing which variables are nonstationary or the nature of the nonstationarity. Applications to the term structure of interest rates and to the FRED-MD macroeconomic dataset demonstrate that the approach offers similar benefits to those of traditional principal component analysis with some added advantages.

Keywords: principal components, nonstationary, term structure of interest rates, economic activity indexes

*Email: jhamilton@ucsd.edu

†Email: x5jin@ucsd.edu

1 Introduction

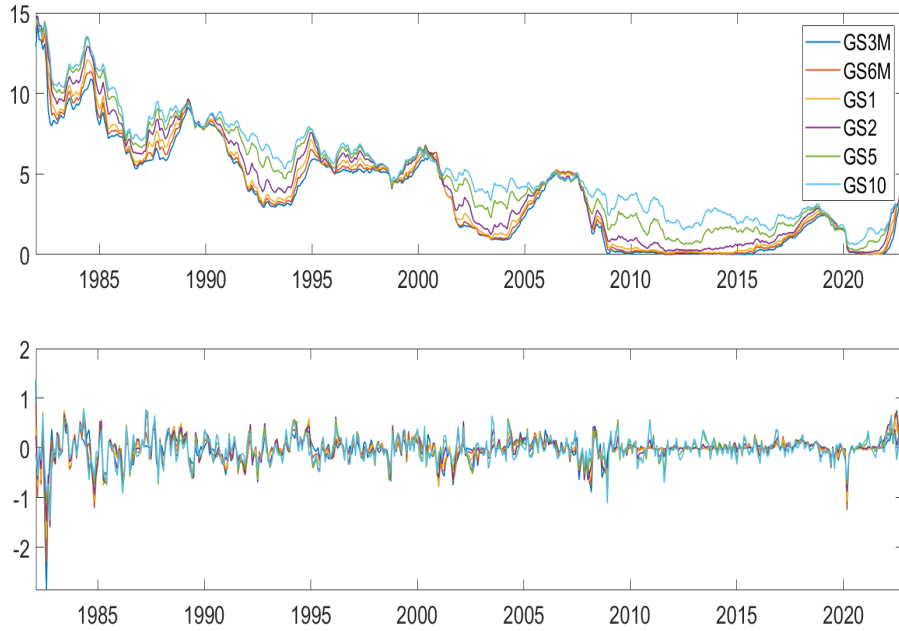
Principal component analysis (PCA) has become a key tool for building dynamic models of vector time series with a large cross-sectional dimension. The traditional approach subtracts the sample mean and divides each variable by its sample standard deviation and then finds linear combinations of the standardized variables that have maximum sample variance. These principal components are then used to build dynamic models for each individual series. For surveys of PCA and its usefulness in economics see [Bai and Ng \(2008\)](#) and [Stock and Watson \(2016\)](#).

One difficulty with PCA is that many of the time series encountered in economics and finance are nonstationary. For a nonstationary variable, the population mean is undefined and the sample standard deviation diverges to infinity as the number of time-series observations gets large. [Onatski and Wang \(2021\)](#) detailed some of the problems that can arise from trying to apply PCA to nonstationary data. The typical solution to this problem is for researchers to examine each series individually by hand to determine the transformation of that series that needs to be made before calculating principal components of the set of variables.

This approach has three shortcomings. First, while for some variables it may be fairly clear what transformation is necessary to achieve stationarity, for others it is far from obvious. For example, the top panel of [Figure 1](#) plots interest rates on U.S. Treasury securities with maturities ranging from three months to ten years from January 1982 to November 2022. There is a strong downward trend in all these interest rates over this sample. Should series like these be treated as stationary? If nonstationary, should we take their first differences or deviations from a time trend before performing PCA? Many finance applications take principal components of the yields without any transformation; see for example [Piazzesi \(2010\)](#) or [Joslin et al. \(2011\)](#). [McCracken and Ng \(2016\)](#) used either first differences or spreads between interest rates as the first step before including interest rates in PCA. [Crump and Gospodinov \(2022\)](#) recommended using either bond returns or the first differences of bond returns in place of yields themselves to reduce the persistence of these data. Many decisions like these, sometimes somewhat arbitrary, have to be made before applying PCA to large data sets.

The second concern is reproducibility. In order to communicate the methodology used in a particular study that makes use of PCA, the researcher needs to report the specific stationarity-

Figure 1: Yields on different maturities.



Notes to Figure 1. Top panel: yields on U.S. Treasury securities for maturities 3 months through 10 years, 1982:1 to 2022:11. Bottom panel: same series plotted as monthly changes.

inducing transformation that was used for each of the dozens or hundreds of variables studied. Another researcher who did not use the same transformations might obtain different results.

A third problem, and in our minds the most fundamental, is the appropriateness of the methodology itself. Suppose we somehow overcame the first problem and knew for certain the true nature of the trend in each individual series. Suppose for illustration we knew correctly that the first variable y_{1t} is a stationary $AR(1)$ process with autoregressive coefficient $\rho = 0.99$ while the second variable y_{2t} is a random walk. Then the currently prescribed procedure would instruct the researcher to use the first variable as is and the second variable in the form of first differences. But if we enter some variables in the form of levels as in the top panel of Figure 1 and others as first-differences as in the bottom panel, we would be mixing together data with very different properties. Would we expect that there is some linear combination of y_{1t} and Δy_{2t} that can summarize the common economic drivers behind the two variables? If differencing is the appropriate transformation for a random walk, it seems we should be using some similar transformation for

an $AR(1)$ process whose root is close but not quite equal to unity.

In this paper we propose an approach to PCA that solves all three of these problems. We interpret the researcher's goal to be to extract the common factors behind the cyclical components of each of the series studied. We follow [Hamilton \(2018\)](#) in defining the cyclical component of a variable to be the component that could not have been predicted on the basis of its own values two years earlier. Hamilton argued that this offers a good representation of what economists typically mean by the cyclical component of a time series and has the huge practical advantage that it can be consistently estimated using an OLS regression without having to know whether the series is stationary or the way in which it may be nonstationary. Our proposal is to filter each series using this regression and perform PCA on the regression residuals. Specifically, we estimate an OLS regression of y_{it} on $\{1, y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1}\}$, where $i = 1, \dots, N$ represents the index of the variable, h is the forecasting horizon ($h = 8$ for quarterly observations and $h = 24$ for monthly) and p is the number of lags used for the forecast ($p = 4$ for quarterly and $p = 12$ for monthly). We then calculate principal components of the residuals. This approach solves each of the three problems identified above. The procedure is fully automatic, requiring no subjective judgment calls by the researcher, and treats every variable in the same way. The transformation of y_{it} is a continuous function of the estimated autoregressive coefficients and implies no discontinuity in the way that persistent stationary series are treated as the largest autoregressive root tends toward unity.

[Bai and Ng \(2004\)](#) proposed that researchers should difference all the variables, whether stationary or not, and calculate principal components of the differenced data. Unit-root tests on the residuals can then be used to determine which individual series are unit-root processes as well as the nature of cointegration. [Bai \(2004\)](#) used this method to form an inference about the stochastic trends that are common across the variables. By contrast, the goal of our approach is to uncover the cyclical components that are common to both the stationary and nonstationary variables.

Section 2 presents our characterization of the cyclical component of an economic time series. Section 3 presents assumptions about the factor structure for the cyclical components and uses standard results to establish the consistency of PCA if the cyclical component was observed without error. Section 4 analyzes the case when the cyclical components are not known but have to be estimated using OLS regressions, and establishes consistency of the method in that setting. Sec-

tions 5 and 6 illustrate the promise of this approach in empirical analyses of interest rates and of the FRED-MD large macroeconomics data set.

2 The cyclical component of an economic time series

If the i th observed variable is a deterministic function of time plus a zero-mean stationary process, $y_{it} = \delta_i(t) + c_{it}$, we could describe the deterministic time trend as the limit of a forecast that would have been made in the arbitrarily distant past:

$$\delta_i(t) = \lim_{h \rightarrow \infty} \lim_{p \rightarrow \infty} E(y_{it} | y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1}).$$

By contrast, if the first difference of the i th variable is a zero-mean stationary process, [Beveridge and Nelson \(1981\)](#) suggested that we think of the trend as the forecast of the variable in the arbitrarily distant future. They decomposed $y_{it} = \delta_{it} + c_{it}$ where

$$\delta_{it} = \lim_{h \rightarrow \infty} \lim_{p \rightarrow \infty} E(y_{i,t+h} | y_{it}, y_{i,t-1}, \dots, y_{i,t-p+1}).$$

While these concepts of trend have some appeal, they have the significant practical drawback that both are based on the properties of forecasts at infinite horizons. They thus depend on conjectures of what happens at infinity, conjectures that are impossible to verify on the basis of a finite sample of observations. [Hamilton \(2018\)](#) suggested that we should instead define a trend in terms of finite-horizon forecasts whose properties we can observe. He proposed the decomposition

$$\begin{aligned} y_{it} &= P(y_{it} | 1, y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1}) + c_{it} \\ &= \alpha_{i0} + \alpha_{i1}y_{i,t-h} + \alpha_{i2}y_{i,t-h-1} + \dots + \alpha_{ip}y_{i,t-h-p+1} + c_{it} \end{aligned} \quad (1)$$

where $P(y|x)$ denotes the population linear projection of y on x . A forecast over a horizon h corresponding to two years is something we can get a reasonable idea about in a sample of typical size. Restricting the forecast to the class of population linear projections on a finite number of lags p is another feature that allows this characterization to be readily implemented. Hamilton argued that the primary reason we would go wrong in making a two-year-ahead forecast of most economic

time series is due to unforeseen cyclical changes. For example, the variable will be significantly below our forecast if the economy goes into a recession over the next two years, and significantly above our forecast if recovery from a downturn is more robust than expected. For this reason, he proposed referring to c_{it} as the cyclical component of the i th variable.

Hamilton showed that the cyclical component c_{it} defined by (1) is stationary for a broad class of nonstationary processes for y_{it} and can be consistently estimated with a simple OLS regression using levels of the variable. Specifically, if either: (i) y_{it} is stationary around a deterministic polynomial function of time of order $d_i \leq p$ satisfying

$$T^{-1/2} \sum_{s=1}^{[Tr]} (y_{it} - \delta_{i0} - \delta_{i1}t - \delta_{i2}t^2 - \dots - \delta_{i d_i}t^{d_i}) \implies \omega_i W_i(r)$$

where $[Tr]$ denotes the largest integer no greater than Tr , $W_i(r)$ denotes standard Brownian motion, and \implies denotes weak convergence of associated probability measures; or alternatively if (ii) d_i differences of y_{it} are stationary for some $d_i \leq p$ satisfying

$$T^{-1/2} \sum_{s=1}^{[Tr]} (\Delta^{d_i} y_{it} - \mu_i) \implies \omega_i W_i(r);$$

then [Hamilton \(2018\)](#) showed that the cyclical component c_{it} defined by the population linear projection (1) is stationary and that the estimated coefficients $\hat{\alpha}_i$ from an OLS regression give a consistent estimate of the population coefficients α_i .

It's instructive to consider some examples of what this means. If y_{it} is a stationary moving average process of order less than h , $\hat{\alpha}_{i0}$ converges to the population mean of y_{it} and $\hat{\alpha}_{ij} \xrightarrow{p} 0$ for $j = 1, \dots, p$. In this case, we would describe the cyclical component of y_{it} as simply the value of the variable minus its population mean. If y_{it} is an AR(1) process with autoregressive coefficient ϕ_i , then $\alpha_{i1} = \phi_i^h$. For example, if $\phi_i = 0.8$ for monthly data, $\alpha_{i1} = (0.8)^{24} = 0.005$, and again the cyclical component of y_{it} would essentially just be its deviation from the population mean. As the persistence of a stationary process increases, the regression removes the persistent component defined as the part of y_{it} that can be predicted two years in advance.

For an $I(1)$ process ($d_i = 1$), we have the accounting identity

$$y_{it} = y_{i,t-h} + \sum_{j=0}^{h-1} \Delta y_{i,t-j}. \quad (2)$$

Collect a constant term along with the $p - 1$ most recent changes in y_i as of date $t - h$ in the vector $q_{i,t-h} = (1, \Delta y_{i,t-h}, \Delta y_{i,t-h-1}, \dots, \Delta y_{i,t-h-p+2})'$. If Δy_{it} is $I(0)$, the population linear projection of $\Delta y_{i,t-j}$ on $q_{i,t-h}$ exists and is given by

$$P(\Delta y_{i,t-j} | q_{i,t-h}) = \pi'_{i,j-h} q_{i,t-h}$$

$$\pi_{i,j-h} = \left[E(q_{i,t-h} q'_{i,t-h}) \right]^{-1} E(q_{i,t-h} \Delta y_{i,t-j}).$$

This allows us to define the population linear projection of the $I(1)$ variable y_{it} on a constant and its p most recent levels as of date $t - h$ as

$$P(y_{it} | 1, y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1}) = y_{i,t-h} + \sum_{j=0}^{h-1} \pi'_{i,j-h} q_{i,t-h}$$

$$= \alpha_{i0} + \alpha_{i1} y_{i,t-h} + \alpha_{i2} y_{i,t-h-1} + \dots + \alpha_{ip} y_{i,t-h+1} \quad (3)$$

where for example $\alpha_{i1} = 1 + \sum_{j=0}^{h-1} \pi_{i,j-h,1}$ for $\pi_{i,j-h,1}$ the first element of the vector $\pi_{i,j-h}$. Thus in this case the value of the cyclical component in the decomposition (1) is $c_{it} = \sum_{j=0}^{h-1} (\Delta y_{i,t-j} - \pi'_{i,j-h} q_{i,t-h})$, which is stationary. An OLS levels regression chooses $\hat{\alpha}_{ij}$ to minimize the sample sum of \hat{c}_{it}^2 , and results in consistent estimates of the population parameters α_{ij} without needing to take differences of the original data.

For an $I(2)$ process, we could make use of the accounting identity

$$y_{it} = y_{i,t-h} + h \Delta y_{i,t-h} + \sum_{j=0}^{h-1} (j+1) \Delta^2 y_{i,t-j}. \quad (4)$$

The population linear projection of $\Delta^2 y_{i,t-j}$ on its $p - 2$ most recent values as of date $t - h$ again exists, and substituting these projection coefficients into (4) again gives a definition of the levels linear projection of the form of (3). OLS regression on levels then yields a residual that consistently estimates the stationary cyclical component c_{it} , which in this case is a function of the residuals from

population linear projections of $\Delta^2 y_{i,t-j}$ on $(1, \Delta^2 y_{i,t-h}, \Delta^2 y_{i,t-h-1}, \dots, \Delta^2 y_{i,t-h-p+3})'$. Again we did not need to know that the true $d_i = 2$ in order to estimate the regression (1). Hamilton (2018) showed that related results hold for an $I(d_i)$ process for any $d_i \leq p$ and also hold whenever y_{it} is stationary around a polynomial time trend of order less than or equal to p .

Our proposal is therefore to estimate the same regression for every variable y_{it} , regardless of whether we think it is stationary and without making any conjecture about the nature of any nonstationarity. To allow for persistent seasonal components in y_{it} , we recommend choosing p to be the number of observations in a year. Our procedure estimates the following regression by OLS for every variable,

$$y_{it} = \alpha_{i0} + \alpha_{i1}y_{i,t-h} + \alpha_{i2}y_{i,t-h-1} + \dots + \alpha_{ip}y_{i,t-h-p+1} + c_{it}, \quad (5)$$

with $h = 8$ and $p = 4$ for quarterly data and $h = 24$ and $p = 12$ for monthly data.¹ We will describe the residual from the estimated regression \hat{c}_{it} as the estimated cyclical component of variable y_{it} and the residual from the population linear projection c_{it} as the true cyclical component. The value \hat{c}_{it} is a consistent estimate of c_{it} , and the true value c_{it} is stationary as long as any nonstationarity in y_{it} is characterized by either a polynomial time trend of order d_i or an $I(d_i)$ process with $d_i \leq p$. Our procedure is to perform PCA on the regression residuals $\{\hat{c}_{1t}, \dots, \hat{c}_{Nt}\}$.

One practical decision is whether a nonlinear transformation of the raw data is necessary for $\Delta^{d_i} y_{it}$ to be stationary for some d_i . If taking the change in the log is the correct way to produce a stationary series, then taking the change in the level would not produce a stationary series. We recommend using logs for variables like output or prices which are usually described in terms of growth rates. For such variables we use the log of the level of the variable, $y_{it} = \log Y_{it}$ as the variable in the regression (5). For variables like interest rates or the unemployment rate that are already quoted in percentage terms, we use the raw data $y_{it} = Y_{it}$ in the regression.

The true cyclical component c_{it} has mean zero and is stationary for a wide range of processes.

¹If the goal is simply to extract a stationary component, this could be accomplished using regression (5) with any finite h . If we set $h = 1$, we would be looking for a factor structure in one-period-ahead forecast errors, which might be interpreted as a factor structure characterizing common shocks to the variables. Although this would be feasible, our purpose in this paper is to calculate the common factors behind cyclical movements in the variables. If we increase h beyond the two-year horizon, we would be basing the definition of the stationary component on forecasts that are difficult to estimate reliably in samples of typical size. Our recommended forecasting horizon of two years is both feasible and offers a good characterization of the cyclical component of an economic variable.

However, the population values of α_{ij} are not known but must be estimated by regression. In Section 3 we characterize our assumptions about the factor structure that we hypothesize describes the true values of c_{it} , and use standard results to establish that these population factors could be consistently estimated if the true values of c_{it} were observed without error. Section 4 considers the case when we do not know the value of d_i for each series, do not know whether it is stationary or characterized by a deterministic time trend or an $I(d_i)$ process, and the c_{it} are not observed. In that section we analyze the consequences of performing PCA on the estimated OLS residuals \hat{c}_{it} .

3 Principal component analysis when the cyclical component is observed

In the previous section we defined the true cyclical component c_{it} to be the residual from a population linear projection of y_{it} on $(1, y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1})'$, and noted that c_{it} is stationary for a broad class of possible processes. In this section we provide sufficient conditions under which the true cyclical components for a collection of N different variables would have a factor structure that could be consistently estimated using PCA if we observed the true value of c_{it} for each variable. The set-up and results in this section closely follow [Stock and Watson \(2002\)](#).

3.1 Assumed factor structure of the true cyclical components

Collect the true cyclical components for the N different series at time t in an $(N \times 1)$ vector $C_t = (c_{1t}, \dots, c_{Nt})'$. We postulate that these are characterized by a factor structure of the form

$$C_t = \Lambda F_t + e_t . \tag{6}$$

$(N \times 1)$ $(N \times r)$ $(r \times 1)$ $(N \times 1)$

The number of latent factors r is much less than the number of variables N , but the r factors are assumed to account for most of the variance of C_t in a sense made formal below. Since the factors are unobserved, $C_t = \Lambda H^{-1} H F_t + e_t$ would imply the identical observable model as (6). Thus some normalizations are necessary in order to talk about consistently estimating the j th factor f_{jt} . In empirical estimation, practitioners typically resolve this ambiguity by estimating the j th column of Λ by the eigenvector associated with the j th largest eigenvalue of $T^{-1} \sum_{t=1}^T C_t C_t'$. Note that such

a procedure implies a normalization in which the columns of Λ are orthogonal to each other and the elements of F_t are uncorrelated with each other and ordered by the size of their variance. We follow [Stock and Watson \(2002\)](#) in how to characterize these conventions as the cross-section dimension N and time-series dimension T get large.²

Assumption 1 (factor structure).

- (i) $(\Lambda' \Lambda / N) \rightarrow I_r$.
- (ii) $E[F_t F_t'] = \Omega_{FF}$, where Ω_{FF} is a diagonal matrix with $\omega_{ii} > \omega_{jj} > 0$ for $i < j$.
- (iii) $|\lambda_{ij}| \leq \bar{\lambda} < \infty$.
- (iv) $T^{-1} \sum_t F_t F_t' \xrightarrow{p} \Omega_{FF}$.

In addition to implementing the property that eigenvectors of a symmetric matrix are orthogonal, Assumption 1(i) requires that each factor makes a nonnegligible contribution to the average variance of c_{it} across i . That is, if we were to imagine adding more variables (increasing N) with $\lambda_{ij} = 0$ for all i greater than some fixed N_0 , then Assumption 1(i) could not hold. Likewise 1(ii) and 1(iv) require that each factor continues to matter as the number of time-series observations T grows. These conditions are consistent with serial dependence of the factors, but rely on the fact that C_t is stationary.

Let γ denote an $(N \times 1)$ vector and $\Gamma = \{\gamma : \gamma' \gamma / N = 1\}$. Note that if γ were the j th column of Λ , the scalar $\gamma' \Lambda F_t / N$ would converge to f_{jt} and $(N^2 T)^{-1} \sum_{t=1}^T \gamma' \Lambda F_t F_t' \Lambda' \gamma \xrightarrow{p} \omega_{jj}$. The assumption that the idiosyncratic elements e_t do not have a factor structure requires that there is no value of γ for which the analogous operation applied to e_t would lead to anything other than zero: $\sup_{\gamma \in \Gamma} (N^2 T)^{-1} \sum_{t=1}^T \gamma' e_t e_t' \gamma \xrightarrow{p} 0$. [Stock and Watson \(2002\)](#) used the following assumptions to guarantee the absence of a factor structure in e_t .

Assumption 2 (moments of the errors).

- (i) $\lim_{N \rightarrow \infty} \sup_t \sum_{s=-\infty}^{\infty} |E[e_t' e_{t+s} / N]| < \infty$.
- (ii) $\lim_{N \rightarrow \infty} \sup_t N^{-1} \sum_{i=1}^N \sum_{j=1}^N |E[e_{it} e_{jt}]| < \infty$, where e_{it} denotes the i th element of e_t .

²See [Bai and Ng \(2013\)](#) and [Stock and Watson \(2016\)](#) for discussion of alternative normalizations.

$$(iii) \lim_{N \rightarrow \infty} \sup_{t,s} N^{-1} \sum_{i=1}^N \sum_{j=1}^N |cov[e_{is}e_{it}, e_{js}e_{jt}]| < \infty.$$

Some might be concerned that we have simply postulated that the true cyclical components are characterized by Assumptions 1 and 2. But something very similar is done in traditional applications that assume conditions like these characterize specified stationary transformations of the original data. Indeed, insofar as the cyclical components have a common primitive definition in terms of two-year-ahead forecast errors, we feel these assumptions are easier to defend in our application than in many others.

3.2 Consequences of applying PCA to the true cyclical components

Recall that the vector of true cyclical components C_t is stationary and has population mean zero. If C_t was observed directly, its estimated sample variance matrix would be $S = T^{-1} \sum_{t=1}^T C_t C_t'$ and a linear combination $\gamma' C_t$ for any $(N \times 1)$ vector γ would have sample variance $\gamma' S \gamma$. If C_t was observed, the first estimated principal component (denoted $\tilde{f}_{1t} = N^{-1} \tilde{\lambda}'_1 C_t$) would be defined as the linear combination that has maximum sample variance subject to a normalization condition such as $\gamma \in \Gamma = \{\gamma : \gamma' \gamma / N = 1\}$:

$$\tilde{\lambda}_1 = \arg \sup_{\gamma \in \Gamma} \tilde{R}(\gamma) \quad (7)$$

$$\tilde{R}(\gamma) = (N^2 T)^{-1} \gamma' \sum_{t=1}^T C_t C_t' \gamma. \quad (8)$$

Note we are normalizing $\tilde{\lambda}'_1 \tilde{\lambda}_1 / N = 1$ as we did asymptotically for the columns of Λ in Assumption 1(i). We also divide the sample variance of $\gamma' C_t$ by N^2 in anticipation of the result that $\tilde{R}(\tilde{\lambda}_1)$ will converge to a fixed constant as N and T grow. The solution to (7) is obtained by setting $\tilde{\lambda}_1$ proportional to the eigenvector of $S = T^{-1} \sum_{t=1}^T C_t C_t'$ associated with the largest eigenvalue. For example, if we calculated eigenvectors of this matrix using code that normalizes eigenvectors to have unit length and orders eigenvalues by decreasing size, $\tilde{\lambda}_1$ would be \sqrt{N} times the first eigenvector. The largest eigenvalue of S is equal to $T^{-1} \sum_{t=1}^T \tilde{f}_{1t}^2$, the sample variance of the first principal component. The j th principal component $N^{-1} \tilde{\lambda}'_j C_t$ is found by maximizing $\tilde{R}(\gamma)$ subject to the constraint that γ is orthogonal to $\tilde{\lambda}_1, \dots, \tilde{\lambda}_{j-1}$. The solution for $\tilde{\lambda}_j$ is proportional to the eigenvector of S associated with the j th largest eigenvalue.

Alternatively, if we observed the true factors F_t and loadings Λ , we could calculate the compo-

ment of the variance of $\gamma' C_t$ that is attributable to the r factors alone:

$$R^*(\gamma) = (N^2 T)^{-1} \gamma' \sum_{t=1}^T \Lambda F_t F_t' \Lambda' \gamma. \quad (9)$$

Stock and Watson (2002) showed that under Assumptions 1 and 2, the maximum value for (8) (which is given by the largest eigenvalue of S) and the supremum of (9) over all $\gamma \in \Gamma$ converge in probability to the same number ω_{11} , which is the population variance of the first factor, and that $\tilde{\lambda}'_1 C_t / N$ gives a consistent estimate of f_{1t} up to a sign. If we were to estimate $k > r$ principal components, the first r would consistently estimate f_{jt} up to a sign normalization and the last $k - r$ would asymptotically have zero variance. We restate their results in the following theorem.

Theorem 1. (Stock and Watson, 2002). *Suppose that Assumptions 1 and 2 hold. Let $R^*(\gamma)$ be the function in (9) and let $\tilde{f}_{1t}, \dots, \tilde{f}_{kt}$ denote the first k estimated principal components of C_t ($\tilde{f}_{jt} = \tilde{\lambda}'_j C_t / N$) with $k \geq r$. Let $\tilde{F}_t = (\tilde{f}_{1t}, \dots, \tilde{f}_{rt})'$ and $\tilde{\Lambda} = \begin{bmatrix} \tilde{\lambda}_1 & \dots & \tilde{\lambda}_r \end{bmatrix}$. Then as N and T go to infinity:*

(i) $\sup_{\gamma \in \Gamma} R^*(\gamma) \xrightarrow{p} \omega_{11};$

(ii) If $\lambda_1^* = \arg \sup_{\gamma \in \Gamma} R^*(\gamma)$ and $\lambda_j^* = \arg \sup_{\gamma \in \Gamma, \gamma' \lambda_1^* = \dots = \gamma' \lambda_{j-1}^* = 0} R^*(\gamma)$, then $R^*(\lambda_j^*) \xrightarrow{p} \omega_{jj}$ for $j = 1, \dots, r$;

(iii) $T^{-1} \sum_{t=1}^T \tilde{f}_{jt}^2 = \tilde{R}(\tilde{\lambda}_j) \xrightarrow{p} \omega_{jj}$ for $j = 1, \dots, r$;

(iv) $T^{-1} \sum_{t=1}^T \tilde{f}_{jt}^2 \xrightarrow{p} 0$ for $j = r + 1, \dots, k$;

(v) $\tilde{S} \tilde{\Lambda}' \Lambda / N \xrightarrow{p} I_r$ where \tilde{S} is a diagonal matrix whose row j column j element is $+1$ if $\tilde{\lambda}'_j \lambda_j > 0$ and -1 if $\tilde{\lambda}'_j \lambda_j < 0$.

(vi) $\tilde{S} \tilde{F}_t - F_t \xrightarrow{p} 0$.

4 Principal component analysis when the cyclical component must be estimated

In this section we assume that we do not observe the true cyclical component c_{it} of series i but have an estimate $\hat{c}_{it} = c_{it} + v_{it}$. Let $\hat{C}_t = (\hat{c}_{1t}, \dots, \hat{c}_{Nt})'$ and $V_t = (v_{1t}, \dots, v_{Nt})'$. We investigate the

properties of principal components calculated from the estimated cyclical components:

$$\hat{f}_{jt} = N^{-1} \hat{\lambda}'_j \hat{C}_t \quad (10)$$

$$\hat{\lambda}_j = \arg \sup_{\{\gamma \in \Gamma, \gamma' \hat{\lambda}_1 = \dots = \gamma' \hat{\lambda}_{j-1} = 0\}} \hat{R}(\gamma)$$

$$\hat{R}(\gamma) = (N^2 T)^{-1} \gamma' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \gamma.$$

We first note a high-level sufficient condition under which PCA applied to the estimated cyclical components \hat{C}_t gives consistent estimates of the true factors F_t . Let $v_{it} = \hat{c}_{it} - c_{it}$ denote the difference between the estimated and true cyclical component of series i at date t . The condition is that v_{it} converges in mean square to zero uniformly in i and t as T goes to infinity.

Assumption 3 (high-level conditions on v_{it}). For $\forall \delta > 0, \exists T_\delta : E(v_{it}^2) < \delta \forall T > T_\delta$ and $\forall i, t$.

The following result establishes that if the error in estimating the cyclical component satisfies Assumption 3, the results $(\hat{f}_{jt}, \hat{\lambda}_j, \hat{R}(\gamma))$ of applying PCA to the estimated cyclical components \hat{C}_t give consistent estimates of the magnitudes that characterize the true cyclical components C_t .

Theorem 2. Suppose that C_t and e_t in equation (6) satisfy Assumptions 1 and 2. Let $\hat{C}_t = C_t + V_t$ for $V_t = (v_{1t}, \dots, v_{Nt})'$ where v_{it} satisfy Assumption 3. Let $\hat{f}_{1t}, \dots, \hat{f}_{kt}$ denote the first k estimated principal components of \hat{C}_t ($\hat{f}_{jt} = \hat{\lambda}'_j \hat{C}_t / N$) with $k \geq r$ and let $\hat{F}_t = (\hat{f}_{1t}, \dots, \hat{f}_{kt})'$ and $\hat{\Lambda} = \begin{bmatrix} \hat{\lambda}_1 & \dots & \hat{\lambda}_r \end{bmatrix}$. Then as N and T go to infinity:

- (i) $T^{-1} \sum_{t=1}^T \hat{f}_{jt}^2 = \hat{R}(\hat{\lambda}_j) \xrightarrow{p} \omega_{jj}$ for $j = 1, \dots, r$;
- (ii) $T^{-1} \sum_{t=1}^T \hat{f}_{jt}^2 \xrightarrow{p} 0$ for $j = r + 1, \dots, k$;
- (iii) $\hat{S} \hat{\Lambda}' \Lambda / N \xrightarrow{p} I_r$ where \hat{S} is a diagonal matrix whose row j column j element is $+1$ if $\hat{\lambda}'_j \lambda_j > 0$ and -1 if $\hat{\lambda}'_j \lambda_j < 0$;
- (iv) $\hat{S} \hat{F}_t - F_t \xrightarrow{p} 0$.

Under what conditions can we expect Assumption 3 to hold? For $\tilde{z}_{it} = (1, y_{i,t-h}, y_{i,t-h-1}, \dots, y_{i,t-h-p+1})'$, the true cyclical component c_{it} is the residual from a population linear projection of y_{it} on \tilde{z}_{it} and \hat{c}_{it} is the residual from the corresponding estimated regression:

$$c_{it} = y_{it} - \alpha'_{i0} \tilde{z}_{it}$$

$$\hat{c}_{it} = y_{it} - \hat{\alpha}'_i \tilde{z}_{it}$$

$$v_{it} = \hat{c}_{it} - c_{it} = (\alpha_{i0} - \hat{\alpha}_i)' \tilde{z}_{it}$$

$$(\alpha_{i0} - \hat{\alpha}_i) = - \left(\sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it} \right)^{-1} \left(\sum_{t=1}^T \tilde{z}_{it} c_{it} \right) \quad (11)$$

$$v_{it}^2 = (\alpha_{i0} - \hat{\alpha}_i)' \tilde{z}_{it} \tilde{z}'_{it} (\alpha_{i0} - \hat{\alpha}_i) \quad (12)$$

Consider first the case of a single stationary regressor ($p = 1, d_i = 0$) and $v_{it}^2 = (\alpha_{i0} - \hat{\alpha}_i)^2 y_{i,t-h}^2$. If y_{it} is stationary we expect that $\sqrt{T}(\alpha_{i0} - \hat{\alpha}_i) \xrightarrow{d} N(0, V_i)$ and $E(y_{i,t-h}^2) < \infty$, in which case v_{it}^2 should converge to zero as T gets large. Next consider the case of a single $I(1)$ regressor ($p = 1, d_i = 1$). Take for illustration the case of a random walk: $y_{it} = y_{i,t-1} + \varepsilon_{it} = \varepsilon_{it} + \varepsilon_{i,t-1} + \dots + \varepsilon_{i1}$ and $c_{it} = \varepsilon_{it} + \varepsilon_{i,t-1} + \dots + \varepsilon_{i,t-h+1}$. Then

$$\begin{aligned} \hat{\alpha}_i - \alpha_{i0} &= \left(\sum_{t=1}^T y_{i,t-h}^2 \right)^{-1} \left(\sum_{t=1}^T y_{i,t-h} c_{it} \right) \\ &= \left(\sum_{t=1}^T y_{i,t-h}^2 \right)^{-1} \left(\sum_{t=1}^T y_{i,t-h} (\varepsilon_{i,t-h+1} + \varepsilon_{i,t-h+2} + \dots + \varepsilon_{it}) \right) \end{aligned}$$

$$\begin{aligned} T(\hat{\alpha}_i - \alpha_{i0}) &= \left[T^{-2} \sum_{t=1}^T y_{i,t-h}^2 \right]^{-1} \left[T^{-1} \sum_{t=1}^T (y_{i,t-h} \varepsilon_{i,t-h+1}) + T^{-1} \sum_{t=1}^T (y_{i,t-h+1} - \varepsilon_{i,t-h+1}) \varepsilon_{i,t-h+2} \right. \\ &\quad \left. + \dots + T^{-1} \sum_{t=1}^T (y_{i,t-1} - \varepsilon_{i,t-h-1} - \varepsilon_{i,t+h-2} - \dots - \varepsilon_{i,t-h+1}) \varepsilon_{it} \right] \\ &\xrightarrow{d} \left[\sigma_i^2 \int_0^1 [W_i(r)]^2 dr \right]^{-1} \left[h \sigma_i^2 \int_0^1 W_i(r) dW_i(r) \right] \end{aligned}$$

which is h times the Dickey-Fuller distribution.³ Also for $t - h = [rT]$ for $[rT]$ the largest integer less than or equal to rT and $r \in (0, 1)$,

$$T^{-1} y_{i,[rT]}^2 \xrightarrow[T \rightarrow \infty]{d} \sigma_i^2 [W_i(r)]^2.$$

If we assume that $E(T^{-1} y_{i,t-h}^2)$ is bounded for all t , then $v_{it}^2 = T^{-1} [T(\alpha_{i0} - \hat{\alpha}_i)]^2 [T^{-1} y_{i,t-h}^2]$ should

³See Hamilton (1994, Proposition 17.3) for similar derivations.

again converge to zero for each t as $T \rightarrow \infty$.

Returning to the general case, we have from (12) and (11) that

$$\begin{aligned}\sum_{t=1}^T v_{it}^2 &= (\alpha_{i0} - \hat{\alpha}_i)' \left(\sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it} \right) (\alpha_{i0} - \hat{\alpha}_i) \\ &= \left(\sum_{t=1}^T c_{it} \tilde{z}'_{it} \right) \left(\sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it} \right)^{-1} \left(\sum_{t=1}^T \tilde{z}_{it} c_{it} \right).\end{aligned}\quad (13)$$

This will be recognized as the OLS Wald statistic for testing the true null hypothesis $H_0 : \alpha_i = \alpha_{i0}$ multiplied by $\hat{\sigma}_i^2 = (T - k)^{-1} \sum_{t=1}^T \hat{c}_{it}^2$, the average squared regression residual. To find the asymptotic distribution of $\sum_{t=1}^T v_{it}^2$ we can make use of the insight of Sims et al. (1990) that the residuals from a regression of y_{it} on \tilde{z}_{it} are numerically identical to the residuals from a regression of y_{it} on $z_{it} = R_i \tilde{z}_{it}$ for R_i any nonsingular matrix. For a particular form of nonstationarity, there is a particular value for R_i that makes the asymptotic properties of the residuals easiest to analyze. Note that we do not need to know the form of the nonstationarity since the observed residual $\hat{c}_{it} = y_{it} - \hat{\alpha}'_i \tilde{z}_{it}$ is numerically identical to the analyzed residual $y_{it} - \hat{\beta}'_i z_{it}$ and thus $v_{it} = \hat{c}_{it} - c_{it}$ is identical whichever way one chooses to describe the regression. For example, for $d_i = 2$ and $p = 4$, we would conveniently characterize the regressors as $z_{it} = (\Delta^2 y_{i,t-h}, \Delta^2 y_{i,t-h-1}, 1, \Delta y_{i,t-h}, y_{i,t-h})'$. The estimated residuals from a regression of y_{it} on z_{it} are numerically identical to the residuals from a regression on \tilde{z}_{it} . We do not need to know the value of d_i to implement the first regression, and therefore do not need to know the value of d_i to find the residuals from the second regression:

$$\begin{aligned}\sum_{t=1}^T v_{it}^2 &= \left(\sum_{t=1}^T c_{it} \tilde{z}'_{it} \right) \left(\sum_{t=1}^T \tilde{z}_{it} \tilde{z}'_{it} \right)^{-1} \left(\sum_{t=1}^T \tilde{z}_{it} c_{it} \right) \\ &= \left(\sum_{t=1}^T c_{it} \tilde{z}'_{it} (R_i')^{-1} \right) \left(\sum_{t=1}^T R_i^{-1} z_{it} z'_{it} R_i'^{-1} \right)^{-1} \left(\sum_{t=1}^T R_i^{-1} z_{it} c_{it} \right) \\ &= \left(\sum_{t=1}^T c_{it} z'_{it} \right) \left(\sum_{t=1}^T z_{it} z'_{it} \right)^{-1} \left(\sum_{t=1}^T z_{it} c_{it} \right).\end{aligned}\quad (14)$$

For any given unknown true value of d_i , there is also an unknown true value of a diagonal scaling matrix Y_{iT} that facilitates calculation of the asymptotic distribution:

$$\sum_{t=1}^T v_{it}^2 = \left(\sum_{t=1}^T c_{it} z'_{it} Y_{iT}^{-1} \right) \left(Y_{iT}^{-1} \sum_{t=1}^T z_{it} z'_{it} Y_{iT}^{-1} \right)^{-1} \left(Y_{iT}^{-1} \sum_{t=1}^T z_{it} c_{it} \right).\quad (15)$$

Again we do not need to know the value of Y_{iT} in order to know that (15) characterizes the fit-

ted residuals. The results in [Hamilton \(2018\)](#) establish that for a broad class of stationary and nonstationary processes,

$$\left(\sum_{t=1}^T c_{it} z'_{it} Y_{iT}^{-1} \right) \left(Y_{iT}^{-1} \sum_{t=1}^T z_{it} z'_{it} Y_{iT}^{-1} \right)^{-1} \left(Y_{iT}^{-1} \sum_{t=1}^T z_{it} c_{it} \right) \xrightarrow{d} q'_i Q_i q'_i.$$

Note the existence of limiting variables q_i and Q_i does not depend on any assumption that the c_{it} are serially uncorrelated. The result that $\sum_{t=1}^T v_{it}^2 \sim O_p(1)$ implies $T^{-1} \sum_{t=1}^T v_{it}^2 \sim o_p(1)$ when the regressors are stationary or nonstationary of any unknown order $d_i < p$. For example, if $d_i = 2$, $p = 4$ and $\Delta^2 y_{it}$ is a mean-zero $I(0)$ process, then

$$Y_{iT} = \begin{bmatrix} T^{1/2} & 0 & 0 & 0 & 0 \\ 0 & T^{1/2} & 0 & 0 & 0 \\ 0 & 0 & T^{1/2} & 0 & 0 \\ 0 & 0 & 0 & T & 0 \\ 0 & 0 & 0 & 0 & T^2 \end{bmatrix} \quad (16)$$

$$Q_i = \begin{bmatrix} \gamma_{i0} & \gamma_{i1} & 0 & 0 & 0 \\ \gamma_{i1} & \gamma_{i0} & 0 & 0 & 0 \\ 0 & 0 & 1 & \omega_i \int_0^1 W_i(r) dr & \omega_i \int W_i^{(2)}(r) dr \\ 0 & 0 & \omega_i \int_0^1 W_i(r) dr & \omega_i^2 \int_0^1 [W_i(r)]^2 dr & \omega_i^2 \int W_i(r) W_i^{(2)}(r) dr \\ 0 & 0 & \omega_i \int W_i^{(2)}(r) dr & \omega_i^2 \int W_i(r) W_i^{(2)}(r) dr & \omega_i^2 \int [W_i^{(2)}(r)]^2 dr \end{bmatrix} \quad (17)$$

for $\gamma_{ij} = E(\Delta^2 y_{it} \Delta^2 y_{i,t-j})$, $\omega_i^2 = \sum_{j=-\infty}^{\infty} \gamma_{ij}$, $W_i(r)$ standard Brownian motion, and $W_i^{(2)}(r) = \int_0^r W_i(s) ds$. Thus $T^{-1} \sum_{t=1}^T v_{it}^2$ should converge to zero for a broad class of processes.

Consider next the convergence of v_{it}^2 for each individual t . In the case of a single stationary regressor, $v_{it}^2 = (\alpha_{i0} - \hat{\alpha}_i)^2 y_{i,t-1}^2$ and $\sum_{t=1}^T v_{it}^2 = (\alpha_{i0} - \hat{\alpha}_i)^2 \sum_{t=1}^T y_{i,t-1}^2$ so

$$\frac{v_{it}^2}{T^{-1} \sum_{t=1}^T v_{it}^2} = \frac{y_{i,t-1}^2}{T^{-1} \sum_{t=1}^T y_{i,t-1}^2} \quad (18)$$

$$v_{it}^2 = (A_{it}/B_{iT}) T^{-1} \sum_{t=1}^T v_{it}^2 \quad (19)$$

for $A_{it} = y_{i,t-1}^2$ and $B_{iT} = T^{-1} \sum_{t=1}^T y_{i,t-1}^2$. As $T \rightarrow \infty$, $B_{iT} \xrightarrow{p} E(y_{i,t-1}^2)$ meaning that if $T^{-1} \sum_{t=1}^T v_{it}^2 \xrightarrow{p}$

0, then also $v_{it}^2 \xrightarrow{p} 0$. A sufficient condition to ensure that $E(v_{it}^2) < \delta$ is that $E(A_{it}/B_{iT})^2 < \kappa_4$ is uniformly bounded.

For the random-walk example, expression (18) again holds identically and we can again rewrite it in the form of (19) now defining $A_{it} = T^{-1}y_{i,t-1}^2$ and $B_{iT} = T^{-2}\sum_{t=1}^T y_{i,t-1}^2$. Then $T^{-1/2}y_{i,[Ts]} \xrightarrow{d} \sigma_i W_i(s)$ and

$$\frac{A_{it}}{B_{iT}} = \frac{T^{-1}y_{i,[Ts]}^2}{\int_0^1 T^{-1}y_{i,[Ts]}^2 ds} \xrightarrow{d} \frac{[W_i(s)]^2}{\int_0^1 [W_i(s)]^2 ds}.$$

This limiting distribution again has finite variance. We can ensure that Condition 3 holds for finite T and each t as before by assuming that $E(A_{it}/B_{iT})^2 < \kappa_4$.

For the general case,

$$\begin{aligned} v_{it}^2 &= (\alpha_{i0} - \hat{\alpha}_i)' \tilde{z}_{it} \tilde{z}_{it}' (\alpha_{i0} - \hat{\alpha}_i) \\ &= (\alpha_{i0} - \hat{\alpha}_i)' R_i^{-1} Y_{iT} Y_{iT}^{-1} R_i \tilde{z}_{it} \tilde{z}_{it}' R_i' Y_{iT}^{-1} Y_{iT} (R_i')^{-1} (\alpha_{i0} - \hat{\alpha}_i) \\ &= \bar{\alpha}_i' (Y_{iT}^{-1} z_{it} z_{it}' Y_{iT}^{-1}) \bar{\alpha}_i \\ &= T^{-1} \bar{\alpha}_i' A_{it} \bar{\alpha}_i \end{aligned}$$

for $\bar{\alpha}_i = Y_{iT} (R_i')^{-1} (\alpha_{i0} - \hat{\alpha}_i)$ and $A_{it} = T Y_{iT}^{-1} z_{it} z_{it}' Y_{iT}^{-1}$. Likewise $\sum_{t=1}^T v_{it}^2 = \bar{\alpha}_i' B_{iT} \bar{\alpha}_i$ for $B_{iT} = Y_{iT}^{-1} \left(\sum_{t=1}^T z_{it} z_{it}' \right) Y_{iT}^{-1}$. Thus

$$\frac{v_{it}^2}{T^{-1} \sum_{t=1}^T v_{it}^2} = \frac{\bar{\alpha}_i' A_{it} \bar{\alpha}_i}{\bar{\alpha}_i' B_{iT} \bar{\alpha}_i}. \quad (20)$$

In general, the ratio in (20) converges in distribution as $T \rightarrow \infty$ to a variable with finite variance, and condition 3 will hold if the ratio has finite variance for each t as well.

The stationary $p = 1$ example is a special case of this general formulation with $z_{it} = y_{i,t-h}$, $Y_{iT} = \sqrt{T}$, $A_{it} = y_{i,t-h}^2$ and $B_{iT} = T^{-1} \sum_{t=1}^T y_{i,t-h}^2$. For the random-walk example, $z_{it} = y_{i,t-h}$, $Y_{iT} = T$, $A_{it} = T^{-1} y_{i,t-h}^2$ and $B_{iT} = T^{-2} \sum_{t=1}^T y_{i,t-h}^2$. For the $p = 4, d_i = 2$ example, Y_{iT} is given by

(16) and $B_{iT} \xrightarrow{d} Q_i$ given in (17), while the lower right (3×3) block of A_{it} is characterized by

$$\begin{aligned}
& T \begin{bmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & T^{-2} \end{bmatrix} \begin{bmatrix} 1 \\ \Delta y_{i,t-h} \\ y_{i,t-h} \end{bmatrix} \begin{bmatrix} 1 & \Delta y_{i,t-h} & y_{i,t-h} \end{bmatrix} \begin{bmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & T^{-2} \end{bmatrix} \\
& = \begin{bmatrix} 1 & T^{-1/2} \Delta y_{i,t-h} & T^{-3/2} y_{i,t-h} \\ T^{-1/2} \Delta y_{i,t-h} & T^{-1} (\Delta y_{i,t-h})^2 & T^{-2} y_{i,t-h} \Delta y_{i,t-h} \\ T^{-3/2} y_{i,t-h} & T^{-2} y_{i,t-h} \Delta y_{i,t-h} & T^{-3} y_{i,t-h}^2 \end{bmatrix} \\
& \xrightarrow{d} \begin{bmatrix} 1 & \omega_i W_i(s) & \omega_i W_i^{(2)}(s) \\ \omega_i W_i(s) & \omega_i^2 [W_i(s)]^2 & \omega_i^2 W_i(s) W_i^{(2)}(s) \\ \omega_i W_i^{(2)}(s) & \omega_i^2 W_i(s) W_i^{(2)}(s) & \omega_i^2 [W_i^{(2)}(s)]^2 \end{bmatrix}
\end{aligned}$$

for $[sT] = t - h$ and $\omega_i^2 = \sum_{j=-\infty}^{\infty} E(\Delta^2 y_{it} \Delta^2 y_{i,t-j})$.

We now formally state sufficient conditions that guarantee that Assumption 3 holds.

Assumption 4 (sufficient conditions for Assumption 3).

- (i) The true cyclical component c_{it} is uniformly bounded, that is, there exists $\kappa_1 < \infty : \forall i \ c_{it}^2 < \kappa_1$.
- (ii) $\sum_{t=1}^T v_{it}^2 \xrightarrow{d} U_i = q_i' Q_i^{-1} q_i$ with $E(U_i) < \kappa_2 \ \forall i$.
- (iii) The convergence is uniform in i , that is, for all $\kappa_3, \varepsilon_3 > 0, \exists T_3(\kappa_3, \varepsilon_3) : \forall T > T_3(\kappa_3, \varepsilon_3)$ and $\forall i$,

$$\left| \text{Prob} \left(\sum_{t=1}^T v_{it}^2 > \kappa_3 \right) - \text{Prob}(U_i > \kappa_3) \right| < \varepsilon_3.$$

- (iv) Let $\lambda^{\max}(A_{it})$ denote the largest eigenvalue of $A_{it} = T Y_{iT}^{-1} z_{it} z_{it}' Y_{iT}^{-1}$ and $\lambda^{\min}(B_{iT})$ the smallest eigenvalue of $B_{iT} = Y_{iT}^{-1} \left(\sum_{t=1}^T z_{it} z_{it}' \right) Y_{iT}^{-1}$. There exists a $\kappa_4, T_4 < \infty$ such that $E \left[\lambda^{\max}(A_{it}) / \lambda^{\min}(B_{iT}) \right]^2 < \kappa_4$ for all $T > T_4$ and all i and t .

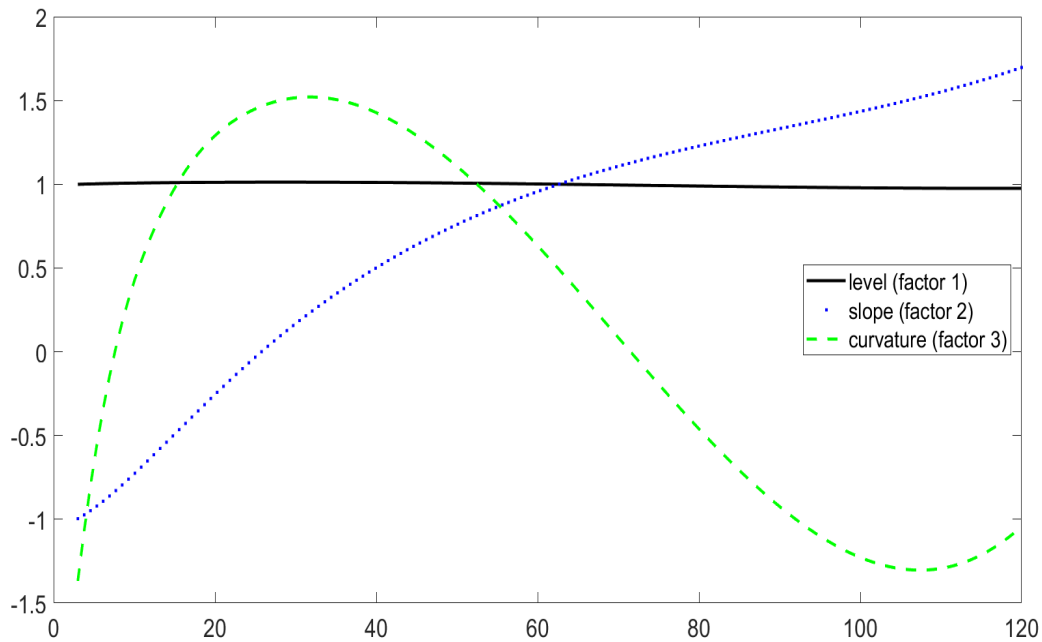
Presumably it is possible to replace condition 4(i) with restrictions on the tail behavior of c_{it} , though we have not attempted that here. Sufficient conditions for 4(ii) are analyzed in Hamilton (2018). The following result establishes that Assumption 4 can replace Assumption 3.

Theorem 3. Assumption 4 implies Assumption 3.

5 Summarizing the yield curve

Our first application uses the data on U.S. Treasury yields plotted in the top panel of Figure 1.⁴ In many finance applications the yields are assumed to be stationary and no transformation is performed before estimating the principal components. Nevertheless, there is a substantial downward trend in interest rates over this period.

Figure 2: Factor loadings for the original data on yields

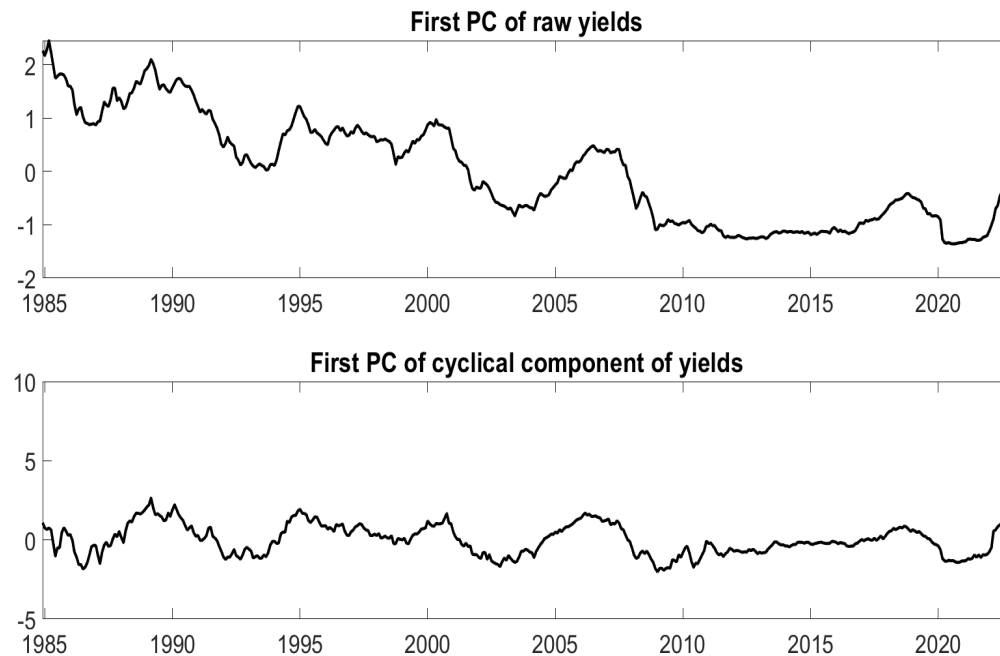


We start by reproducing the traditional finance application of PCA to the raw data on yields. The factor loadings associated with the first three principal components are plotted in Figure 2. The horizontal axis in Figure 2 shows the maturities in months, and the three curves display the estimates of the first three columns of Λ in equation (6) plotted as a function of maturity. The coefficient relating yields to the first factor (often referred to as the *level* factor) is roughly the same for all maturities; the level factor is essentially the average interest rate at time t . Loadings on the second factor (*slope*) are positive for long rates and negative for short rates, causing the second factor to be positive when the yield curve slopes up and negative when it slopes down. The third

⁴The data are constant maturity yields from the Federal Reserve Bank of St. Louis database <https://fred.stlouisfed.org/categories/115>.

factor (*curvature*) has a positive weight for 1- to 6-year bonds and negative weight for bonds with very short or very long maturity.

Figure 3: First Principal component of level and cyclical component of yields

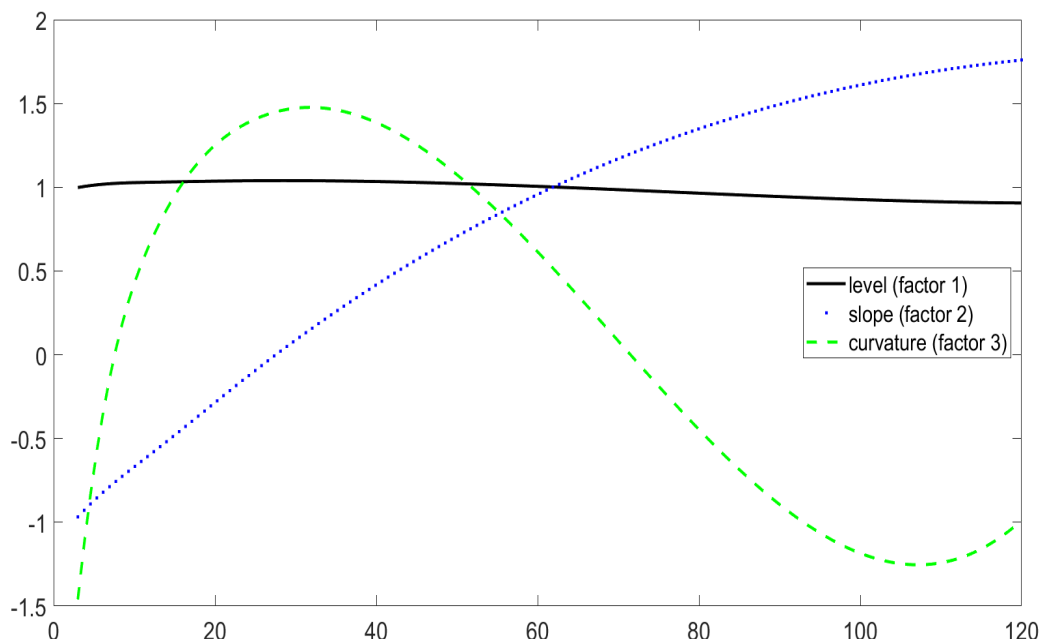


The top panel in Figure 3 plots the first principal component of the raw data. The estimate closely tracks the general movement of interest rates. But it is also clear that the dynamics come from both the downward trend and the cyclical movements.

We next regressed each yield on a constant and 12 of its lagged values 2 years earlier and performed PCA on the regression residuals. Figure 4 shows that the loadings on the three factors are quite similar to the ones in Figure 2, meaning that the factors are still very naturally described as the level, slope and curvature of the cyclical component of interest rates. One modest difference in the factor loadings is that the cyclical level factor loads a little more strongly on short-term interest rates. A more important difference is seen in the plot of the value of the first factor in the second panel Figure 3. By construction there is no trend in the bottom panel Figure 3, in contrast to the clear downward trend of the top panel.

For this application, all the variables share the same trend and the same cyclical tendencies, so PCA applied to the levels results in similar broad conclusions as PCA for the cyclical components.

Figure 4: Factor loadings for the cyclical component of yields



This will no longer be the case, however, if the analysis includes interest rates along with other variables that exhibit different trends. In such applications it can be very helpful to isolate the cyclical components of interest rates in order to relate these to the cyclical components of other variables. We consider such an example in our second application.

6 Characterizing large macroeconomic data sets

The use of large macroeconomic data sets was pioneered by [Stock and Watson \(1999\)](#), whose goal was to use the information of 168 different macroeconomic variables to produce better forecasts of inflation. They found that the first principal component of macroeconomic variables that are related to the level of real economic activity produced the best inflation forecasts over the period 1959:1 to 1997:9. Their findings led to the development of the Chicago Fed National Activity Index, which is the first principal component of a subset of 85 different measures of economic activity.⁵

⁵The variables used to calculate the Chicago Fed National Activity Index (CFNAI) fall into four broad groups: (1) production and income; (2) employment, unemployment, and hours; (3) personal consumption and housing; and (4) sales, orders, and inventories

McCracken and Ng (2016) reviewed subsequent uses of large macroeconomic sets and developed the FRED-MD database whose 2015:4 vintage covered 134 macroeconomic variables. These variables include monthly measures divided into 8 broad categories: (1) output and income; (2) labor market; (3) housing; (4) consumption, orders, and inventories; (5) money and credit; (6) interest and exchange rates; (7) prices; and (8) stock market. This data set offers benefits of continuity and continuous updating and is the basis for our analysis in this paper.

In previous uses of PCA on large macroeconomic data sets, each of the variables needed to be transformed using a detrending method that was selected individually for each series. For details of how this has been done see Federal Reserve Bank of Chicago (2021) for the CFNAI or the data appendix to McCracken and Ng (2016) for FRED-MD. Figure 5 illustrates these transformations for three important macroeconomic indicators. The first column plots the raw data, while the second column plots the data as transformed by McCracken and Ng (2016) in order to ensure stationarity, using the same data set as in their original paper. Everyone agrees that industrial production (row 1) is nonstationary, and all previous researchers have used first differences of the log of industrial production shown in panel (1,2). While there is little doubt that this is a good way to generate a stationary series for this variable, monthly growth rates of industrial production exhibit a lot of high-frequency fluctuations around the dominant cyclical patterns. For the unemployment rate (row 2), it is less clear whether the series should be regarded as stationary. McCracken and Ng (2016) used first differences of unemployment, which behave quite differently from the level. The purchasing managers composite index from the Institute of Supply Management (row 3) appears to be stationary, and McCracken and Ng (2016) entered this series directly into PCA without any transformation.

The top panel of Figure 6 plots the first principal component of the transformed series arrived at by McCracken and Ng (2016).⁶ This inherits some of the high-frequency fluctuations seen in the (1,2) and (2,2) panels of Figure 5. Indeed, McCracken and Ng (2016) regarded this series as too volatile to reliably identify business cycles and turning points, and instead plotted in their Figure 3 the accumulation of this series. The CFNAI (shown in panel 2 of Figure 6) is very similar to the first principal component of the FRED-MD macro data set.

⁶We generated this figure using the exact data and code posted at <https://research.stlouisfed.org/econ/mccracken/fred-databases/>. Note that we have multiplied the series by -1 in order to give it the property that the factor declines in recessions, and that their Figure 3 plots the accumulations ($s_{1t} = \sum_{j=1}^t \tilde{f}_{1t}$) whereas our graph shows \tilde{f}_{1t} itself.

Figure 5: Level, transformed value, and cyclical component of industrial production, unemployment, and PMI Composite, 1962:3 to 2014:12

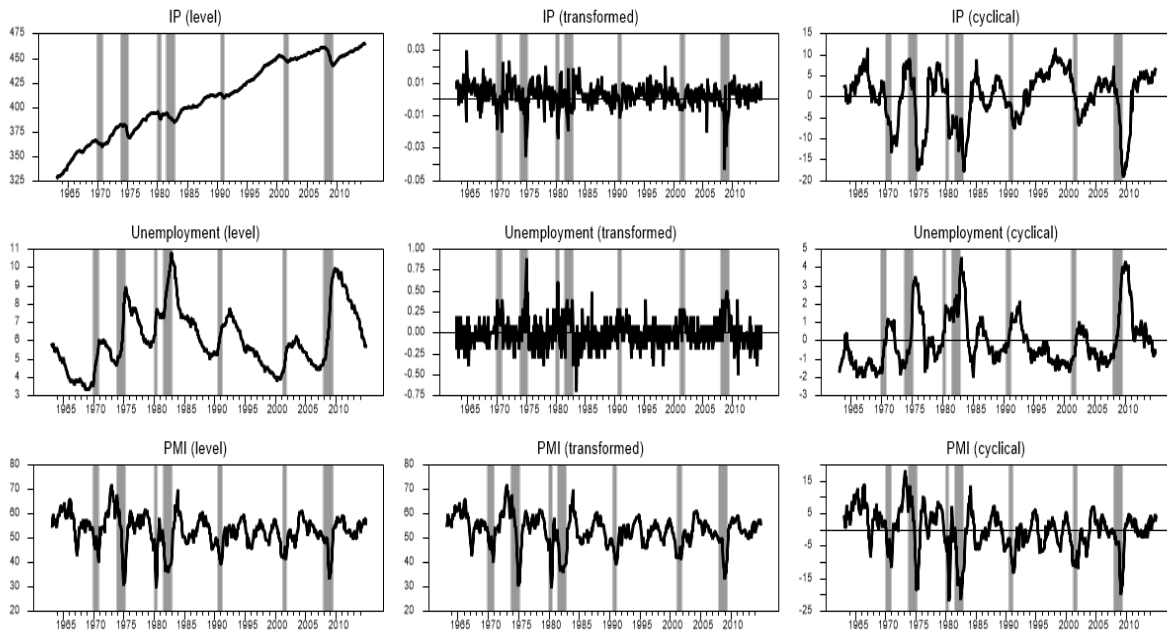
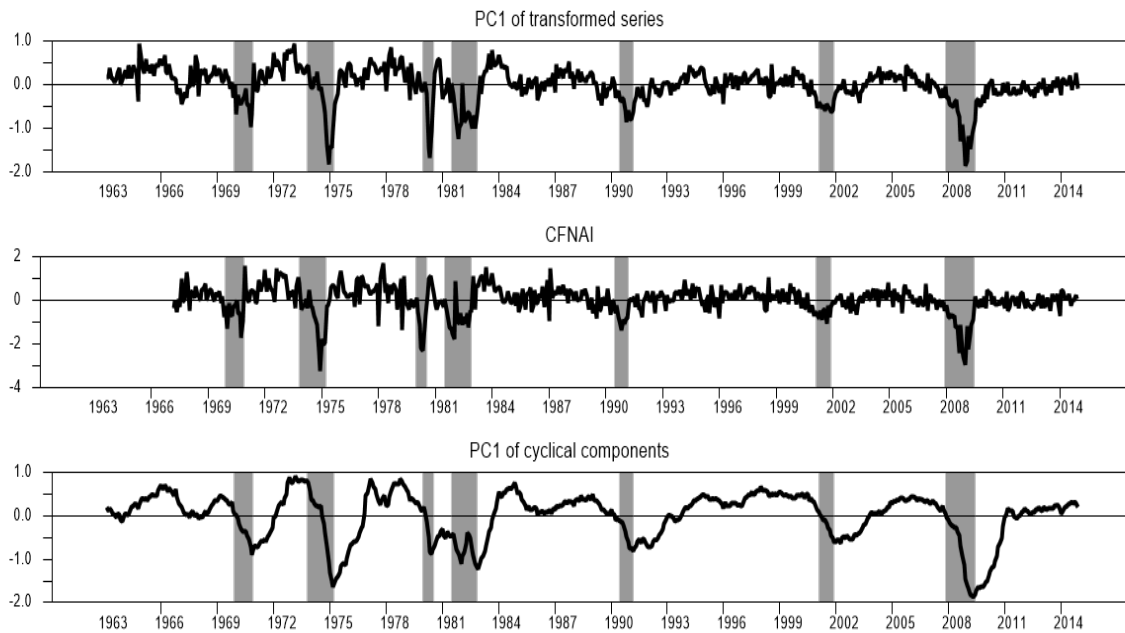


Figure 6: First PC of FRED-MD variables as transformed by McCracken and Ng (2016), the Chicago Fed National Activity Index, and first PC of cyclical components of FRED-MD variables, 1962:3 to 2014:12



The third column of Figure 5 plots the cyclical components of industrial production, unemployment, and PMI as estimated by the residuals of the OLS regression (5).⁷ PMI is almost impossible to predict two years in advance, and our cyclical component is almost identical to the original series. Thus both our method and the traditional approach use this variable essentially as is. There is some but not much predictability of the unemployment rate at the two-year horizon, so for this variable our transformation much more closely resembles the original series than it does the first-difference transformation. For industrial production, our approach takes out the broad trend while retaining the essential cyclical behavior observed in the raw data. The three variables in the third column, unlike those in the second column, all share a common characterization of what is happening over the business cycle. Consistent with a long tradition in business cycle research, when plotted this way PMI appears as a leading indicator, industrial production as a coincident indicator, and unemployment as a coincident or lagging indicator, with all three clearly following the same cycle.

The first principal component of the estimated cyclical components of the variables in the data set is plotted in the bottom panel of Figure 6. Unlike the CFNAI, this provides a very clean summary of historical business cycles. We would suggest that our series could be viewed as a stationary version of the series that McCracken and Ng were looking for when they accumulated the first principal component as calculated by their method.

6.1 Outliers

Previous users of large macro data sets devoted a lot of attention to outliers and implemented procedures to mitigate their influence. Prior to the COVID recession of 2020, the CFNAI discarded observations that were more than six times the interquartile range, as did [Stock and Watson \(1999\)](#) in some of their analysis. [McCracken and Ng \(2016\)](#) discarded observations that were more than ten times the interquartile range. This criterion identifies 79 different observations on 22 different variables as outliers in the 1960:3 to 2014:12 data set; for details see [Table B2](#).

⁷For those series that [McCracken and Ng \(2016\)](#) transformed using logs, first differences of logs, or second differences of logs (their transformations 4-6), we simply took the log of the variable before performing the regression. Thus for example the series plotted in the upper left panel of Figure 5 is 100 times the natural logarithm of the industrial production index. For those series that they used as is, as first differences, or second differences (their transformations 1-3), we simply used the variable as is. They employed a special transformation (7) for nonborrowed reserves. One would have expected to take logs of a variable like this, but the variable took on negative values in 2008. For this series their transformation was $y_{it} = \Delta(x_t/x_{t-1} - 1.0)$ and we used $y_{it} = x_t/x_{t-1}$.

This approach to identifying outliers requires one to know the form of the transformation that is needed to render each variable stationary. Can one accomplish the same task without knowing which variables are nonstationary or the form of the nonstationarity? If we knew the true cyclical component c_{it} we could easily identify outliers of c_{it} relative to its interquartile range. But outliers also exert undue influence on the estimated regression coefficients, tilting the coefficients so as to tend to make \hat{c}_{it} closer to zero than the true value c_{it} . We therefore calculated residuals using leave-one-out regressions.⁸ Outliers that exceed ten times the interquartile range for $h = 1$ -month-ahead forecasts are reported in the middle columns of Table B2. There is quite a bit of overlap between these outliers and those identified by conventional methods, though there are also some differences. For example, interest rate spreads in 1980 and some price measures in November 2008 register as outliers to the regression but not in the transformed series. Overall, the regression identifies 98 outliers in this data set compared to 79 identified by McCracken and Ng. If one were to use $h = 1$ -month-ahead forecasting regressions to extract a stationary component, then we would recommend removing or downweighting outliers as identified using leave-one-out interquartile ranges.

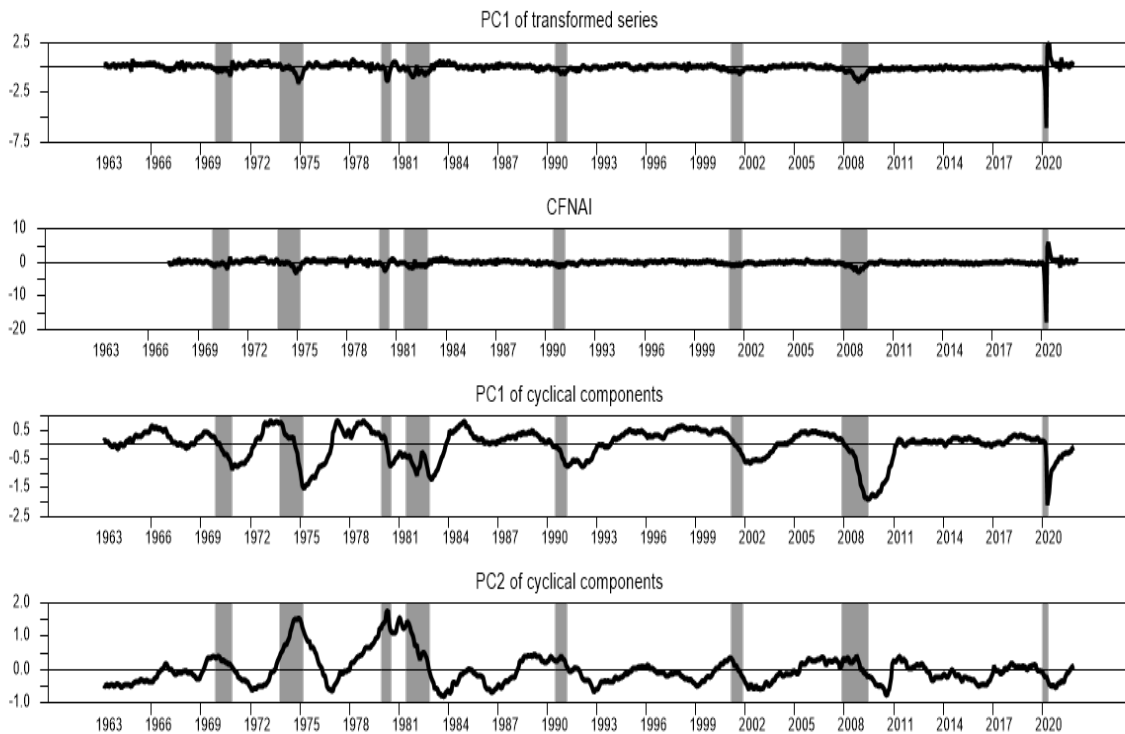
However, the errors associated with a two-year-ahead forecast are quite different from 1-month-ahead errors. For example, for a random walk, two-year-ahead errors are the accumulation of 24 different one-month-ahead forecast errors. From the Central Limit Theorem these sums exhibit much less kurtosis than individual one-month-ahead errors. We found outliers in the two-year-ahead regressions in only 2 of the 134 series, as reported in the last columns of Table B2. The behavior of total and nonborrowed reserves was certainly anomalous during the Federal Reserve’s response to the Great Recession, but nothing else in this sample is a clear outlier by this criterion. Our recommended procedure is to use 2-year-ahead regressions and make no corrections for outliers. The series that we have plotted in the bottom panel of Figure 6 is the unadjusted first principal component of the full set of OLS residuals \hat{c}_{it} .

Outliers are an even bigger issue when data for 2020 are included. For the 2022:4 vintage of FRED-MD, the McCracken-Ng procedure would identify 40 of the 127 variables as all being outliers in the single month of 2020:4. Despite dropping all of these 40 observations, the first

⁸That is, we calculated $\tilde{c}_{it} = y_{it} - \tilde{\alpha}_{i,t}\tilde{z}_{it}$ with $\tilde{\alpha}_{i,t} = \left(\sum_{s=1, s \neq t}^T \tilde{z}_{is}\tilde{z}_{is}'\right)^{-1} \left(\sum_{s=1, s \neq t}^T \tilde{z}_{is}y_{is}\right)$ estimated separately for each i and t and then divided $\tilde{c}_{i,t}$ by its observed interquartile range.

principal component calculated using their algorithm shows an enormous decline in this month. Indeed, in order to include the 2020 observations in the top panel of Figure 7, the scale must be so large that it makes all the previous cyclical fluctuations barely noticeable. This measure shows huge positive values in the subsequent two months that are also without precedent. The CFNAI modified its procedure for dealing with anomalous observations to handle these observations. But the CFNAI displays even more unprecedented negative and positive values in the spring of 2020, as seen in the second panel.

Figure 7: First PC of FRED-MD variables as transformed by [McCracken and Ng \(2016\)](#), the Chicago Fed National Activity Index, and first and second PC of cyclical components of FRED-MD variables, 1962:3 to 2022:3



By contrast, only two variables are identified as outliers for 2020:4 for purposes of our approach, these being new claims for unemployment insurance and the number unemployed for less than 5 weeks. The result of applying our procedure to the FRED-MD database up through 2022 is displayed in the third panel of Figure 7. Note that, unlike the top two panels, our series is plotted here on the same scale as in Figure 6. We would argue that our series correctly summarizes

cyclical movements in 2020 just as it did for earlier episodes. The magnitude of the peak-to-trough decline in the 2020 recession was comparable to other downturns, though it was distinguished by the remarkable speed with which it happened. The pace of recovery was also unprecedented, though the economy nevertheless remained cyclically weak for some time. Again we calculated the series that is displayed in the last two panels of Figure 7 without making any corrections or special treatment for outliers.

6.2 Missing observations

Another issue with large data sets comes from discontinued, newly added, or missing variables. McCracken and Ng (2016) adapted the Stock and Watson (2002) algorithm for unbalanced panels, though they found in their original data set that the results are essentially identical if one simply drops variables as needed to create a balanced panel. For our application, we have simply calculated principal components of \hat{c}_{it} on a balanced panel, though there is no obstacle to applying the Stock and Watson (2002) algorithm to an unbalanced panel of \hat{c}_{it} .⁹

6.3 Uses of macroeconomic cyclical factors

A key use of PCA is to summarize the statistical information in a large cross section of indicators; for illustrations see Bernanke et al. (2005), Bai and Ng (2008), Forni et al. (2009), Bai and Ng (2010), and Stock and Watson (2016). The movement in variable i that is captured by the j th factor alone is given by $\hat{\lambda}_{ij}\hat{f}_{jt}$. Since \hat{c}_{it} is normalized to have unit variance, the fraction of the variance of the stationary component of variable i that is explained by the j th macro factor is given by $\hat{\lambda}_{ij}^2 T^{-1} \sum_{t=1}^T \hat{f}_{jt}^2$.

Table B1 reports the R^2 explained by the first macroeconomic cyclical factor for each of the 120 variables used in our analysis of the 2022 vintage dataset. The first factor alone accounts for about 3/4 of the variance of typical indicators of real output or income and about half of the variance

⁹A balanced panel was created from the 127 variables in the 2022:4 dataset by: using only data over 1960:1-2021:12; dropping the Michigan Survey of Consumer Sentiment (UMCSENT), trade-weighted exchange rate (TWEX-AFEGSMTH), and new orders for consumer goods (ACOGNO) and nondefense capital goods (ANDENO), which are the same four series dropped by McCracken and Ng to create a balanced panel from the 2015:4 dataset; dropping the VIX (VIXCLS), which was not included in the 2015:4 dataset and whose first value is July 1962; and dropping the financial commercial paper rate (CP3M) and the commercial paper-fed funds spread (COMPAPFF) which were not reported for April 2020. The particular variables used in our analysis of the 2022 vintage dataset are described in Table B1.

of typical indicators of labor-market conditions. The first macro factor is far less successful at describing financial indicators and is almost useless for capturing movements in nominal prices. It is interesting that when we add the second macro factor, the R^2 for the median price indicator rises to 64%. The first factor thus seems mainly to capture real economic conditions and the second characterizes nominal prices and interest rates.

The fourth panel in Figure 7 plots the second macro factor. This by construction is orthogonal to the first, and often continues to fall even as the recovery in real economic activity is beginning. This is consistent with the view that nominal variables may respond sluggishly to business-cycle developments.

We next explore the use of the cyclical macro factors in forecasting. [Stock and Watson \(1999\)](#) demonstrated that the first principal component of a large data set of real macroeconomic variables could be very helpful for forecasting inflation. Their finding gave rise to the Chicago Fed National Activity Index (CFNAI), a PCA-based indicator that is still widely used today. We compare the usefulness for forecasting of the CFNAI (denoted \hat{f}_t^{CF}), the first principal component calculated using the algorithm and data set of [McCracken and Ng \(2016\)](#) (denoted \hat{f}_t^{MN}), the first principal component of our cyclical components (denoted \hat{f}_t^{HX1}) calculated from the McCracken-Ng data set using equation (10), and the second principal component \hat{f}_t^{HX2} of the cyclical components. Our approach to comparing different forecasts is very similar to that used by [Stock and Watson \(1999\)](#) and [McCracken and Ng \(2016\)](#).

A particular model m uses a set of variables \mathbf{x}_t^m that are observed at date t to try to forecast the value of a variable of interest y_{t+h}^h that will not be observed until $t + h$:

$$y_{t+h}^h = \boldsymbol{\pi}^{m'} \mathbf{x}_t^m + u_{t+h}^{m,h}. \quad (21)$$

We estimated the value of $\boldsymbol{\pi}^m$ by OLS regression on the subsample $t = T_0, T_0 + 1, \dots, T_1 - h - 1$ and used these coefficients to forecast $y_{T_1}^h$. We then augmented the sample by one observation, estimating the regression over $t = T_0, T_0 + 1, \dots, T_1 - h$ and using those coefficients to forecast $y_{T_1+1}^h$. We repeated this for an evaluation period T_1 to T_2 and calculated the average squared forecast error over this evaluation period. We performed this analysis using three different evaluation periods. The first evaluation period is specified by $T_1 = 1970:1$ to $T_2 = 1996:12$, which was the

evaluation period in the original study by [Stock and Watson \(1999\)](#). The second evaluation period is $T_1 = 1997:1$ to $T_2 = 2014:12$, which corresponds to the new data used by [McCracken and Ng \(2016\)](#) that were not available to [Stock and Watson \(1999\)](#). The third evaluation period is $T_1 = 2015:1$ to $T_2 = 2021:12$, which is the new data available since publication of [McCracken and Ng \(2016\)](#). The models we considered were a pure autoregressive model,

$$\mathbf{x}_t^{AR} = (1, y_t^1, y_{t-1}^1, \dots, y_{t-5}^1)',$$

and models that add to the autoregressive model six lags of one of the principal components estimates. For example,

$$\mathbf{x}_t^{CF} = (\mathbf{x}_t^{AR}, \hat{f}_t^{CF}, \hat{f}_{t-1}^{CF}, \dots, \hat{f}_{t-5}^{CF})'.$$

This differs a little from the forecast evaluations performed by [Stock and Watson \(1999\)](#) and [McCracken and Ng \(2016\)](#) in that these authors used *BIC* to select different lag lengths for the autoregressive and principal components and for each subsample, whereas we set the lag length to six for every evaluation. Also, since \hat{f}_t^{CF} is only available beginning in 1967:3, we used $T_0 = 1967:9$ as the first date for estimation of all models. For the CFNAI we used the value of the index as it is currently reported for all historical dates.¹⁰ For the first two evaluation periods, we calculated \hat{f}_t^{MN} and \hat{f}_t^{HX} using the historical vintage of the FRED-MD data base available as of 2015:4.¹¹ For the third evaluation period, we re-estimated \hat{f}_t^{MN} and \hat{f}_t^{HX} using the data base available as of 2022:4.

For our first set of evaluations we set y_{t+h}^h to be the average inflation rate between month t and $t + h$, quoted at an annual rate,

$$y_{t+h}^h = (1200/h) \log(CPI_{t+h}/CPI_t),$$

where CPI_t denotes the level of the consumer price index in month t .¹² The column labeled *AR* in

¹⁰We downloaded \hat{f}_t^{CF} on September 12, 2022 from the FRED data base at <https://fred.stlouisfed.org/series/CFNAI>.

¹¹Our series for \hat{f}_t^{MN} for this subsample is thus almost (but not quite) identical to the series analyzed by [McCracken and Ng \(2016\)](#). We have been unable to identify the source of the small discrepancies.

¹²Here again our evaluation design differs slightly from that in [Stock and Watson \(1999\)](#) and [McCracken and Ng \(2016\)](#) in that those authors took the object of interest to be to forecast the *change* in the inflation rate as a function of lagged changes:

$$y_{t+h}^h = (1200/h) \log(CPI_{t+h}/CPI_t) - 1200 \log(CPI_t/CPI_{t-1}).$$

Table 1: Mean squared forecast errors for different models

sample	horizon	CPI					IP				
		AR	MN	CF	HX1	HX2	AR	MN	CF	HX1	HX2
1970-1996	h=1	7.91	0.99	1.00	1.03	0.90	76.73	0.94	0.97	0.96	1.02
	h=6	4.26	0.77	0.82	0.80	0.88	38.66	0.93	0.91	0.83	0.79
	h=12	5.32	0.62	0.70	0.74	1.33	27.19	1.06	1.01	1.21	0.87
1997-2014	h=1	12.26	1.04	1.03	1.02	1.09	58.90	0.83	0.85	0.98	1.00
	h=6	6.08	1.23	1.23	1.23	1.11	22.61	0.94	0.93	1.05	1.12
	h=12	4.21	1.22	1.22	1.28	1.17	20.11	1.01	0.96	1.06	1.11
2015-2022	h=1	6.66	1.54	1.93	1.40	1.05	727.45	0.95	1.72	1.04	0.98
	h=6	3.34	1.90	2.62	2.03	1.04	126.78	1.18	2.23	1.03	0.88
	h=12	2.73	1.66	2.48	1.69	1.01	55.40	1.18	2.32	0.87	0.83

Notes to Table 1. AR columns report simulated out-of-sample mean squared forecast error for purely autoregressive model evaluated over three different out-of-sample periods. MN columns report the MSE relative to the AR MSE when lags of the first principal component calculated using the procedures in [McCracken and Ng \(2016\)](#) are added to the autoregression, with a value less than one indicating the variable is useful for forecasting. CF columns report the relative MSE when lags of the Chicago Fed National Activity Index are added to the autoregression, HX1 when lags of the first principal component of the estimated cyclical components are added to the autoregression, and HX2 when lags of the second principal component of the estimated cyclical components are added to the autoregression.

Table 1 reports the simulated out-of-sample mean squared error of a purely autoregressive model for each of the three evaluation samples and for forecast horizons of $h = 1, 6, \text{ or } 12$ months.¹³ The top left panel reproduces the finding of [Stock and Watson \(1999\)](#) that an index like CFNAI significantly improves forecasts for longer horizons over the 1970-1996 period. The alternative measures \hat{f}_t^{MN} or \hat{f}_t^{HX1} offer similar improvements. The indexes offer little or no improvement for one-month-ahead forecasts over this period, but again are similar to each other. All three indexes are outperformed by simple autoregressive forecasts over either of the later two evaluation periods. The observation that inflation has become much harder to forecast in data since 1996 has been reported by a number of other researchers, including [Atkeson and Ohanian \(2001\)](#), [Fisher et al. \(2002\)](#), [Stock and Watson \(2007\)](#), and [Stock and Watson \(2008\)](#).¹⁴ Interestingly, the second

¹³[Stock and Watson \(1999\)](#) reported results for $h = 12$ months whereas [McCracken and Ng \(2016\)](#) reported results for $h = 1, 6, \text{ and } 12$.

¹⁴The result is also revealed in the numbers reported in Table 5 of [McCracken and Ng \(2016\)](#), though the authors did not comment on it.

cyclical factor \hat{f}_t^{HX2} does better than any of the other three indexes at forecasting inflation at the one-month horizon for the 1970-1996 sample, better than any of the other three at both the 6- and 12-month horizons for the 1997-2014 sample, and better at all three horizons for the 2015-2022 sample.

Table 1 also reports forecasts of industrial production, setting

$$y_{t+h}^h = (1200/h) \log(IP_{t+h}/IP_t),$$

for IP_t the level of the industrial production index in month t . The indexes \hat{f}_t^{MN} , \hat{f}_t^{CF} , and \hat{f}_t^{HX1} all help forecast industrial production over near horizons in the first two evaluation periods. The CFNAI does particularly poorly at forecasting either inflation or industrial production at any horizon for 2015-2022. Both \hat{f}_t^{HX1} and \hat{f}_t^{HX2} do significantly better than CFNAI in every case over this period.

We conclude that our approach offers similar benefits to conventional PCA when evaluated in terms of simulated out-of-sample forecasts, and does much better than measures like the CFNAI for recent data. We share the conclusion of the earlier literature that the usefulness of any principal-component-based measure for purposes of forecasting depends on the variable, evaluation period, and horizon of the forecast.

7 Conclusion

Calculating principal components of medium-horizon forecast errors is a viable approach to identifying the common cyclical factors that drive a large collection of potentially nonstationary economic indicators. This avoids the need to decide how to detrend each individual series and is much more promising than approaches such as the Chicago Fed National Activity Index for handling data that include the economic recession of 2020.

References

- Atkeson, A. and Ohanian, L. E. (2001). Are Phillips Curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1):2–11.
- Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1):137–183.
- Bai, J. and Ng, S. (2004). A PANIC attack on unit roots and cointegration. *Econometrica*, 72(4):1127–1177.
- Bai, J. and Ng, S. (2008). *Large Dimensional Factor Analysis*. Now Publishers Inc.
- Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606.
- Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.
- Bernanke, B. S., Boivin, J., and Eliasziw, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics*, 120(1):387–422.
- Beveridge, S. and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the ‘business cycle’. *Journal of Monetary Economics*, 7(2):151–174.
- Crump, R. K. and Gospodinov, N. (2022). On the factor structure of bond returns. *Econometrica*, 90(1):295–314.
- Federal Reserve Bank of Chicago (2021). Background on the Chicago Fed national activity index. <https://www.chicagofed.org/-/media/publications/cfnai/background/cfnai-background-pdf.pdf>.
- Fisher, J. D., Liu, C. T., and Zhou, R. (2002). When can we forecast inflation? *Economic Perspectives Federal Reserve Bank of Chicago*, 26(1):32–44.

- Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, 25(5):1319–1347.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.
- Hamilton, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics*, 100(5):831–843.
- Joslin, S., Singleton, K. J., and Zhu, H. (2011). A new perspective on Gaussian dynamic term structure models. *Review of Financial Studies*, 24(3):926–970.
- McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.
- Onatski, A. and Wang, C. (2021). Spurious factor analysis. *Econometrica*, 89(2):592–614.
- Piazzesi, M. (2010). Affine term structure models. In *Handbook of Financial Econometrics: Tools and Techniques*, pages 691–766. Elsevier.
- Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, pages 113–144.
- Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.
- Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.
- Stock, J. H. and Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33.
- Stock, J. H. and Watson, M. W. (2008). Phillips Curve inflation forecasts. *NBER Working Paper*, 14322.
- Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

A Proofs

The proof of Theorem 2 makes use of the following lemma.

Lemma 1. *Let S be any symmetric positive semidefinite ($N \times N$) matrix with diagonal elements $s_{ii}, i = 1, \dots, n$. Then*

$$\sup_{\gamma \in \Gamma} \frac{\gamma' S \gamma}{\gamma' \gamma} \leq \sum_{i=1}^N s_{ii}. \quad (\text{A-1})$$

Proof of Lemma 1.

Notice that the left side of (A-1) is equal to the largest eigenvalue of S . Since S is positive semidefinite, all eigenvalues are nonnegative so the largest eigenvalue is less than or equal to the sum of all the eigenvalues. But the sum of all the eigenvalues is equal to the trace of S , which is defined as the sum of its diagonal elements. Thus the left side of (A-1) must be less than or equal to the right side.

Proof of Theorem 2(i)-(ii),

Notice from $\hat{C}_t = C_t + V_t$ that

$$\begin{aligned} & (N^2 T)^{-1} \gamma' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \gamma \\ = & (N^2 T)^{-1} \gamma' \sum_{t=1}^T C_t C_t' \gamma + (N^2 T)^{-1} \gamma' \sum_{t=1}^T V_t V_t' \gamma + 2(N^2 T)^{-1} \gamma' \sum_{t=1}^T C_t V_t' \gamma. \end{aligned} \quad (\text{A-2})$$

We first show that the second and third terms on the right side of (A-2) converge in probability to 0 for all $\gamma \in \Gamma$. To show this for the second term, notice from Lemma 1 that

$$\begin{aligned} \sup_{\gamma \in \Gamma} (N^2 T)^{-1} \gamma' \sum_{t=1}^T V_t V_t' \gamma &= \sup_{\gamma \in \Gamma} (NT)^{-1} \frac{\gamma' \sum_{t=1}^T V_t V_t' \gamma}{\gamma' \gamma} \\ &\leq (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T v_{it}^2. \end{aligned} \quad (\text{A-3})$$

Since $E v_{it}^2 < \delta$ for all i and t , it follows that $E \left[(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T v_{it}^2 \right] < \delta$ and thus $\sup_{\gamma \in \Gamma} (N^2 T)^{-1} \gamma' \sum_{t=1}^T V_t V_t' \gamma \xrightarrow{p} 0$ by Markov's Inequality.

For the last term in (A-2),

$$\left| (N^2 T)^{-1} \gamma' \sum_{t=1}^T C_t V_t' \gamma \right| \leq \left[(N^2 T)^{-1} \gamma' \sum_{t=1}^T C_t C_t' \gamma \right]^{1/2} \left[(N^2 T)^{-1} \gamma' \sum_{t=1}^T V_t V_t' \gamma \right]^{1/2}.$$

The first term converges in probability to a number no larger than $\omega_{11}^{1/2}$ from Theorem 1, and the second converges in probability to 0 for all γ from (A-3). Thus for all $\gamma \in \Gamma$, $(N^2T)^{-1}\gamma' \sum_{t=1}^T C_t V_t' \gamma \xrightarrow{p} 0$. We thus conclude from (A-2) that

$$(N^2T)^{-1}\gamma' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \gamma - (N^2T)^{-1}\gamma' \sum_{t=1}^T C_t C_t' \gamma \xrightarrow{p} 0 \quad (\text{A-4})$$

for all $\gamma \in \Gamma$. Since

$$T^{-1} \sum_{t=1}^T \hat{f}_{1t}^2 - T^{-1} \sum_{t=1}^T \tilde{f}_{1t}^2 = \sup_{\gamma \in \Gamma} (N^2T)^{-1} \gamma' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \gamma - \sup_{\gamma \in \Gamma} (N^2T)^{-1} \gamma' \sum_{t=1}^T C_t C_t' \gamma,$$

it follows that this difference converges in probability to zero, establishing result (i) of Theorem 2 for $j = 1$. Analogous calculations establish results (i) and (ii) for $j = 2, \dots, k$.

Proof of Theorem 2(iii).

Notice that $(N^2T)^{-1} \hat{\Lambda}' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \hat{\Lambda}$ is a diagonal matrix for all N and T by the definition of $\hat{\Lambda}$ with diagonal elements converging in probability to ω_{jj} by result (i):

$$(N^2T)^{-1} \hat{\Lambda}' \sum_{t=1}^T \hat{C}_t \hat{C}_t' \hat{\Lambda} \xrightarrow{p} \Omega_{FF}. \quad (\text{A-5})$$

Equation (A-4) then establishes that $(N^2T)^{-1} \hat{\Lambda}' \sum_{t=1}^T C_t C_t' \hat{\Lambda} \xrightarrow{p} \Omega_{FF}$. We also know from results (R2)-(R6) in Stock and Watson (2002) that

$$(N^2T)^{-1} \gamma' \sum_{t=1}^T C_t C_t' \gamma - (N^2T)^{-1} \gamma' \sum_{t=1}^T \Lambda' F_t F_t' \Lambda \gamma \xrightarrow{p} 0$$

for all $\gamma \in \Gamma$, meaning

$$\begin{aligned} (N^2T)^{-1} \hat{\Lambda}' \sum_{t=1}^T C_t C_t' \hat{\Lambda} &\xrightarrow{p} (N^2T)^{-1} \hat{\Lambda}' \sum_{t=1}^T \Lambda F_t F_t' \Lambda \hat{\Lambda} \\ &= (N^{-1} \hat{\Lambda}' \Lambda) \left(T^{-1} \sum_{t=1}^T F_t F_t' \right) (N^{-1} \Lambda' \hat{\Lambda}) \\ &\xrightarrow{p} H \Omega_{FF} H' \end{aligned} \quad (\text{A-6})$$

for $H = \text{plim} (N^{-1} \hat{\Lambda}' \Lambda)$. Combining results (A-4)-(A-6),

$$\Omega_{FF} = H \Omega_{FF} H'. \quad (\text{A-7})$$

Let \hat{h}'_j denote the j th row of $\hat{\Lambda}' \Lambda / N$,

$$\hat{h}'_j = \hat{\lambda}'_j \frac{\Lambda}{(N \times r)} / N,$$

for $\hat{\lambda}'_j$ the j th row of $\hat{\Lambda}'$. Then

$$\hat{h}'_j \hat{h}_j = \frac{\hat{\lambda}'_j}{\sqrt{N}} \frac{\Lambda \Lambda'}{N} \frac{\hat{\lambda}_j}{\sqrt{N}}.$$

This is less than or equal to the largest eigenvalue of $\Lambda \Lambda' / N$, which converges to 1. Letting $h'_j = (h_{j1}, h_{j2}, \dots, h_{jr})'$ denote the j th row of H , we thus have

$$\hat{h}'_j \hat{h}_j \xrightarrow{p} h_{j1}^2 + h_{j2}^2 + \dots + h_{jr}^2 \leq 1.$$

The (1,1) element of (A-7) states

$$h'_1 \Omega_{FF} h_1 = h_{11}^2 \omega_{11} + h_{12}^2 \omega_{22} + \dots + h_{1r}^2 \omega_{rr} = \omega_{11}.$$

Since $\omega_{11} > \omega_{22} > \dots > \omega_{rr} > 0$, this requires $h_{11}^2 = 1$ and $h_{12} = \dots = h_{1r} = 0$. Thus the (1,1) element of $\hat{\Lambda}' \Lambda / N$ converges in probability to ± 1 and other elements of the first row converge to zero.

The (2,2) element of (A-7) states

$$h_{21}^2 \omega_{11} + h_{22}^2 \omega_{22} + \dots + h_{2r}^2 \omega_{rr} = \omega_{22} \quad (\text{A-8})$$

where $h_{21} = \text{plim} \hat{\lambda}_2 \lambda_1 / N$. Regress λ_1 on $\hat{\lambda}_1$ with residual q_1 :

$$\lambda_1 = \hat{k}_1 \hat{\lambda}_1 + q_1 \quad (\text{A-9})$$

$$\hat{k}_1 = (\hat{\lambda}'_1 \hat{\lambda}_1 / N)^{-1} (\hat{\lambda}'_1 \lambda_1 / N)$$

$$q_1' \hat{\lambda}_1 = 0$$

$$\lambda_1' \lambda_1 / N = \hat{k}_1^2 (\hat{\lambda}_1' \hat{\lambda}_1 / N) + q_1' q_1 / N.$$

We saw above that $\hat{k}_1^2 \xrightarrow{p} 1$, which along with $\lambda_1' \lambda_1 / N \rightarrow 1$ and $\hat{\lambda}_1' \hat{\lambda}_1 / N = 1$ establishes $q_1' q_1 / N \xrightarrow{p} 0$. Premultiply (A-9) by $\hat{\lambda}_2' / N$:

$$\hat{\lambda}_2' \lambda_1 / N = \hat{k}_1 \hat{\lambda}_2' \hat{\lambda}_1 / N + \hat{\lambda}_2' q_1 / N = \hat{\lambda}_2' q_1 / N.$$

But from Cauchy-Schwarz

$$(\hat{\lambda}_2' q_1 / N)^2 \leq (\hat{\lambda}_2' \hat{\lambda}_2 / N) (q_1' q_1 / N) \xrightarrow{p} 0.$$

Thus $\hat{\lambda}_2' \lambda_1 / N \xrightarrow{p} h_{21} = 0$ and (A-8) becomes

$$h_{22}^2 \omega_{22} + h_{23}^2 \omega_{33} + \cdots + h_{2r}^2 \omega_{rr} = \omega_{22}.$$

Since $\omega_{22} > \omega_{33} > \cdots > \omega_{rr}$ and $h_{22}^2 + h_{23}^2 + \cdots + h_{2r}^2 \leq 1$, this requires $h_{22}^2 = 1$ and all other elements of the second row of H to be zero, establishing the second row of the claim in Theorem 2(iii). Proceeding iteratively through rows 3, 4, ..., r establishes the rest of the result in (iii).

Proof of Theorem 2(iv).

Write

$$\begin{aligned} \hat{S}\hat{F}_t - F_t &= N^{-1} \hat{S}\hat{\Lambda}'\hat{C}_t - F_t \\ &= N^{-1} \hat{S}\hat{\Lambda}'(\Lambda F_t + e_t + V_t) - F_t \\ &= (N^{-1} \hat{S}\hat{\Lambda}'\Lambda - I_r)F_t + N^{-1} \hat{S}\hat{\Lambda}'e_t + N^{-1} \hat{S}\hat{\Lambda}'V_t. \end{aligned} \tag{A-10}$$

The task is to show that all three terms in (A-10) have plim 0. That $(N^{-1} \hat{S}\hat{\Lambda}'\Lambda - I_r)F_t \xrightarrow{p} 0$ follows immediately from result (iii). For the second term,

$$N^{-1} \hat{S}\hat{\Lambda}'e_t = N^{-1} (\hat{S}\hat{\Lambda}' - \Lambda')e_t + N^{-1} \hat{S}\Lambda'e_t. \tag{A-11}$$

Consider the square of the j th element of the first term in (A-11):

$$\left[\frac{(\hat{s}_j \hat{\lambda}'_j - \lambda'_j) e_t}{N} \right]^2 \leq \left[\frac{(\hat{s}_j \hat{\lambda}'_j - \lambda'_j)(\hat{s}_j \hat{\lambda}_j - \lambda_j)}{N} \right] \left[\frac{e'_t e_t}{N} \right]. \quad (\text{A-12})$$

The first term in (A-12) is

$$\frac{(\hat{s}_j \hat{\lambda}'_j - \lambda'_j)(\hat{s}_j \hat{\lambda}_j - \lambda_j)}{N} = \frac{\hat{s}_j^2 \hat{\lambda}'_j \hat{\lambda}_j}{N} - \frac{\lambda'_j \hat{s}_j \hat{\lambda}_j}{N} - \frac{\hat{s}_j \hat{\lambda}'_j \lambda_j}{N} + \frac{\lambda'_j \lambda_j}{N},$$

which converges in probability to zero by Theorem 2(iii). The second term in (A-12) is $O_p(1)$, by result (R1) in Stock and Watson (2002), meaning the plim of (A-12) is zero. The second term in (A-11) also converges in probability to zero as in Stock and Watson (2002) Result (R15). Hence $N^{-1} \hat{S} \hat{\Lambda}' e_t \xrightarrow{p} 0$.

For the third term in (A-10), $N^{-1} \hat{S} \hat{\Lambda}' V_t$, note that the j th element is $N^{-1} \hat{s}_j \hat{\lambda}'_j V_t$ whose square is

$$N^{-2} \hat{\lambda}'_j V_t V_t' \hat{\lambda}_j \leq N^{-1} \sum_{i=1}^N v_{it}^2 \xrightarrow{p} 0 \quad (\text{A-13})$$

with the inequality following from Lemma 1 and the convergence in probability from Assumption 3 and Markov's Inequality.

Proof of Theorem 3.

We first demonstrate that Assumptions 4(i)-(iii) imply $E \left(T^{-1} \sum_{t=1}^T v_{it}^2 \right) < \delta$ for all $T > T_\delta$. Note from (14) $U_{iT} = \sum_{t=1}^T v_{it}^2$ can be written as the sum of squares of the fitted values from a regression of c_{it} on z_{it} , which by construction it must be smaller than the sum of squares of c_{it} itself:

$$U_{iT} < \sum_{t=1}^T c_{it}^2 < T \kappa_1.$$

We therefore know

$$\begin{aligned} E[T^{-1} U_{iT}] &= E[T^{-1} U_{iT} \mathbf{1}\{T^{-1} U_{iT} \leq \delta/2\}] + E[T^{-1} U_{iT} \mathbf{1}\{\delta/2 \leq T^{-1} U_{iT} \leq \kappa_1\}] \\ &< (\delta/2) + \kappa_1 \text{Prob}(T^{-1} U_{iT} \geq \delta/2). \end{aligned}$$

The condition $E[T^{-1} \sum_{t=1}^T v_{it}^2] < \delta$ will then follow if we can show that for any $\delta > 0 \exists T_\delta$ such

that $\text{Prob}(T^{-1}U_{iT} \geq \delta/2) \leq \delta/(2\kappa_1)$ whenever $T > T_\delta$. Given any $\delta > 0$, let $\kappa_3 = 4\kappa_1\kappa_2/\delta$ and $\varepsilon_3 = \delta/(4\kappa_1)$. Then from Assumption 4(iii), $\exists T_3(\kappa_3, \varepsilon_3)$ (call this $T_3(\delta)$) such that $\forall T \geq T_3(\delta)$,

$$\text{Prob}(U_{iT} \geq \kappa_3) < \text{Prob}(U_i \geq \kappa_3) + \varepsilon_3 = \text{Prob}(U_i \geq \kappa_3) + \delta/(4\kappa_1).$$

From Markov's Inequality and the definition of κ_3 this means

$$\begin{aligned} \text{Prob}\left(U_{iT} \geq \frac{4\kappa_1\kappa_2}{\delta}\right) &< \frac{E(U_i)}{\kappa_3} + \frac{\delta}{4\kappa_1} \\ &\leq \frac{\kappa_2}{(4\kappa_1\kappa_2/\delta)} + \frac{\delta}{4\kappa_1} \\ &= \frac{\delta}{2\kappa_1} \end{aligned} \tag{A-14}$$

with the second inequality coming from Assumption 4(ii). Let $T_\delta = \max\{T_3(\delta), 8\kappa_1\kappa_2/\delta^2\}$. For all $T \geq T_\delta$ we know from this definition of T_δ that

$$\begin{aligned} \text{Prob}(T^{-1}U_{iT} > \delta/2) &\leq \text{Prob}(T_\delta^{-1}U_{iT} > \delta/2) \\ &= \text{Prob}(U_{iT} > \delta T_\delta/2) \\ &\leq \text{Prob}(U_{iT} > 4\kappa_1\kappa_2/\delta). \end{aligned} \tag{A-15}$$

Putting (A-14) together with (A-15) establishes that for all $T > T_\delta$, $\text{Prob}(T^{-1}U_{iT} \geq \delta/2) \leq \delta/(2\kappa_1)$ which was to be shown.

We next show that $E\left(T^{-1}\sum_{t=1}^T v_{it}^2\right) < \delta$ along with Assumption 4(iv) imply that Assumption 3 holds. Notice that

$$v_{it}^2 = T^{-1}\bar{\alpha}'_i A_{it}\bar{\alpha}_i \leq T^{-1}(\bar{\alpha}'_i\bar{\alpha}_i)\lambda^{\max}(A_{it})$$

$$\sum_{t=1}^T v_{it}^2 = \bar{\alpha}'_i B_{iT}\bar{\alpha}_i \geq (\bar{\alpha}'_i\bar{\alpha}_i)\lambda^{\min}(B_{iT})$$

$$\frac{v_{it}^2}{T^{-1}\sum_{t=1}^T v_{it}^2} \leq \frac{\lambda^{\max}(A_{it})}{\lambda^{\min}(B_{iT})}$$

$$\begin{aligned}
E(v_{it}^2) &\leq E \left[\frac{\lambda^{\max}(A_{it})}{\lambda^{\min}(B_{iT})} T^{-1} \sum_{t=1}^T v_{it}^2 \right] \\
&\leq \left(E \left[\frac{\lambda^{\max}(A_{it})}{\lambda^{\min}(B_{iT})} \right]^2 \right)^{1/2} \left(E \left[T^{-1} \sum_{t=1}^T v_{it}^2 \right]^2 \right)^{1/2}
\end{aligned} \tag{A-16}$$

by Cauchy-Schwarz. The first term in (A-16) is less than $\sqrt{\kappa_4}$ by Assumption 4(iv). For the second term, $T^{-1} \sum_{t=1}^T v_{it}^2 < \kappa_1$ so $E \left[T^{-1} \sum_{t=1}^T v_{it}^2 \right]^2 < \kappa_1 E \left[T^{-1} \sum_{t=1}^T v_{it}^2 \right]$. For any $\delta > 0$, let $\delta_4 = \delta / (\kappa_1 \sqrt{\kappa_4})$. There exists a T_4 such that $E(T^{-1} \sum_{t=1}^T v_{it}^2) < \delta_4$ for $T > T_4$ and all i , meaning $E(v_{it}^2) < \delta$. Since this is uniform in i , it follows that $E \left[N^{-1} \sum_{i=1}^N v_{it}^2 \right] < \delta$, which was to be shown.

B Data appendix.

Table B1: R^2 for each variable explained by first and second principal components, 1962:3 to 2022:1

Group 1. Output and income				
Index	FRED	Description	PC1	PC1&2
1	RPI	Real Personal Income	0.47	0.51
2	W875RX1	Real personal income ex transfer receipts	0.74	0.76
6	INDPRO	IP Index	0.85	0.85
7	IPFPNSS	IP: Final Products and Nonindustrial Supplies	0.87	0.87
8	IPFINAL	IP: Final Products (Market Group)	0.82	0.82
9	IPCONGD	IP: Consumer Goods	0.67	0.83
10	IPDCONGD	IP: Durable Consumer Goods	0.65	0.82
11	IPNCONGD	IP: Nondurable Consumer Goods	0.48	0.54
12	IPBUSEQ	IP: Business Equipment	0.67	0.73
13	IPMAT	IP: Materials	0.74	0.74
14	IPDMAT	IP: Durable Materials	0.75	0.76
15	IPNMAT	IP: Nondurable Materials	0.70	0.70
16	IPMANSICS	IP: Manufacturing (SIC)	0.86	0.87
17	IPB51222S	IP: Residential Utilities	0.02	0.03
18	IPFUELS	IP: Fuels	0.04	0.06
19	CUMFNS	Capacity Utilization: Manufacturing	0.74	0.74
		Median	0.72	0.75

Table B1 (continued)

Group 2. Labor market				
Index	FRED	Description	PC1	PC1&2
20	HWI	Help-Wanted Index for United States	0.52	0.52
21	HWIURATIO	Ratio of Help Wanted/No. Unemployed	0.54	0.54
22	CLF16OV	Civilian Labor Force	0.17	0.25
23	CE16OV	Civilian Employment	0.71	0.73
24	UNRATE	Civilian Unemployment Rate	0.73	0.73
25	UEMPMEAN	Average Duration of Unemployment (Weeks)	0.30	0.32
26	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	0.43	0.47
27	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	0.68	0.68
28	UEMP15OV	Civilians Unemployed - 15 Weeks and Over	0.67	0.68
29	UEMP15T26	Civilians Unemployed for 15-26 Weeks	0.67	0.67
30	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	0.58	0.60
31	CLAIMSx	Initial Claims	0.46	0.47
32	PAYEMS	All Employees: Total nonfarm	0.79	0.81
33	USGOOD	All Employees: Goods-Producing Industries	0.80	0.85
34	CES1021000001	All Employees: Mining and Logging: Mining	0.00	0.42
35	USCONS	All Employees: Construction	0.70	0.70
36	MANEMP	All Employees: Manufacturing	0.71	0.73
37	DMANEMP	All Employees: Durable goods	0.73	0.77
38	NDMANEMP	All Employees: Nondurable goods	0.51	0.51
39	SRVPRD	All Employees: Service-Providing Industries	0.65	0.67
40	USTPU	All Employees: Trade, Transportation and Utilities	0.75	0.78
41	USWTRADE	All Employees: Wholesale Trade	0.62	0.76
42	USTRADE	All Employees: Retail Trade	0.66	0.66
43	USFIRE	All Employees: Financial Activities	0.35	0.35
44	USGOVT	All Employees: Government	0.06	0.06
45	CES0600000007	Avg Weekly Hours : Goods-Producing	0.45	0.50
46	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	0.55	0.63
47	AWHMAN	Avg Weekly Hours : Manufacturing	0.46	0.54
115	CES0600000008	Avg Hourly Earnings : Goods-Producing	0.00	0.36
116	CES2000000008	Avg Hourly Earnings : Construction	0.04	0.29
117	CES3000000008	Avg Hourly Earnings : Manufacturing	0.01	0.28
		Median	0.55	0.60

Group 3. Housing				
Index	FRED	Description	PC1	PC1&2
48	HOUST	Housing Starts: Total New Privately Owned	0.22	0.37
49	HOUSTNE	Housing Starts, Northeast	0.25	0.40
50	HOUSTMW	Housing Starts, Midwest	0.21	0.42
51	HOUSTS	Housing Starts, South	0.17	0.29
52	HOUSTW	Housing Starts, West	0.16	0.26
53	PERMIT	New Private Housing Permits (SAAR)	0.18	0.36
54	PERMITNE	New Private Housing Permits, Northeast (SAAR)	0.23	0.44
55	PERMITMW	New Private Housing Permits, Midwest (SAAR)	0.20	0.48
56	PERMITS	New Private Housing Permits, South (SAAR)	0.14	0.31
57	PERMITW	New Private Housing Permits, West (SAAR)	0.14	0.24
		Median	0.19	0.37

Table B1 (continued)

Group 4. Consumption, orders, and inventories

Index	FRED	Description	PC1	PC1&2
3	DPCERA3M086SBEA	Real personal consumption expenditures	0.61	0.77
4	CMRMTSPL	Real Manu. and Trade Industries Sales	0.84	0.87
5	RETAIL	Retail and Food Services Sales	0.36	0.43
58	AMDMNO	New Orders for Durable Goods	0.64	0.70
59	AMDMUO	Unfilled Orders for Durable Goods	0.15	0.41
60	BUSINV	Total Business Inventories	0.10	0.70
61	ISRATIO	Total Business: Inventories to Sales Ratio	0.25	0.26
		Median	0.36	0.70

Group 5. Money and credit

Index	FRED	Description	PC1	PC1&2
62	M1SL	M1 Money Stock	0.04	0.05
63	M2SL	M2 Money Stock	0.02	0.06
64	M2REAL	Real M2 Money Stock	0.00	0.47
65	BOGMBASE	Monetary Base	0.19	0.19
66	TOTRESNS	Total Reserves of Depository Institutions	0.27	0.29
67	NONBORRES	Reserves Of Depository Institutions	0.00	0.00
68	BUSLOANS	Commercial and Industrial Loans	0.11	0.25
69	REALLN	Real Estate Loans at All Commercial Banks	0.21	0.22
70	NONREVSL	Total Nonrevolving Credit	0.29	0.29
71	CONSPI	Adjusted Monetary Base	0.08	0.15
118	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	0.02	0.02
119	DTCTHFNM	Total Consumer Loans and Leases Outstanding	0.16	0.19
120	INVEST	Securities in Bank Credit at All Commercial Banks	0.03	0.08
		Median	0.08	0.19

Group 6. Interest and exchange rates

Index	FRED	Description	PC1	PC1&2
76	FEDFUNDS	Effective Federal Funds Rate	0.20	0.75
77	TB3MS	3-Month Treasury Bill:	0.22	0.74
78	TB6MS	6-Month Treasury Bill:	0.23	0.74
79	GS1	1-Year Treasury Rate	0.20	0.72
80	GS5	5-Year Treasury Rate	0.05	0.60
81	GS10	10-Year Treasury Rate	0.01	0.55
82	AAA	Moody's Seasoned Aaa Corporate Bond Yield	0.00	0.56
83	BAA	Moody's Seasoned Baa Corporate Bond Yield	0.03	0.57
84	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	0.02	0.42
85	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	0.01	0.39
86	T1YFFM	1-Year Treasury C Minus FEDFUNDS	0.00	0.24
87	T5YFFM	5-Year Treasury C Minus FEDFUNDS	0.06	0.37
88	T10YFFM	10-Year Treasury C Minus FEDFUNDS	0.11	0.43
89	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	0.22	0.53
90	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	0.30	0.52
91	EXSZUS	Switzerland / U.S. Foreign Exchange Rate	0.00	0.02
92	EXJPUS	Japan / U.S. Foreign Exchange Rate	0.01	0.01
93	EXUSUK	U.S. / U.K. Foreign Exchange Rate	0.05	0.12
94	EXCAUS	Canada / U.S. Foreign Exchange Rate	0.00	0.05
		Median	0.05	0.52

Table B1 (concluded)

Group 7. Prices				
Index	FRED	Description	PC1	PC1&2
95	WPSFD49207	PPI: Finished Goods	0.00	0.74
96	WPSFD49502	PPI: Finished Consumer Goods	0.00	0.71
97	WPSID61	PPI: Intermediate Materials	0.00	0.65
98	WPSID62	PPI: Crude Materials	0.02	0.40
99	OILPRICE _x	Crude Oil, spliced WTI and Cushing	0.00	0.49
100	PPICMM	PPI: Metals and metal products:	0.07	0.38
101	CPIAUCSL	CPI: All items	0.00	0.81
102	CPIAPPSL	CPI: Apparel	0.00	0.32
103	CPITRNSL	CPI: Transportation	0.00	0.52
104	CPIMEDSL	CPI: Medical	0.21	0.42
105	CUSR0000SAC	CPI : Commodities	0.00	0.73
106	CUSR0000SAD	CPI : Durables	0.07	0.35
107	CUSR0000SAS	CPI : Services	0.02	0.63
108	CPIULFSL	CPI : All Items Less Food	0.01	0.75
109	CUSR0000SA0L2	CPI : All items less shelter	0.00	0.78
110	CUSR0000SA0L5	CPI : All items less medical care	0.00	0.81
111	PCEPI	PCE: Chain Price Index	0.00	0.76
112	DDURRG3M086SBEA	PCE: Durable Goods Price Index	0.05	0.35
113	DNDGRG3M086SBEA	PCE: Nondurable Goods Price Index	0.00	0.77
114	DSERRG3M086SBEA	PCE: Services Price Index	0.00	0.63
		Median	0.00	0.64
Group 8. Stock market				
Index	FRED	Description	PC1	PC1&2
72	S&P 500	S&P's Common Stock Price Index: Composite	0.29	0.33
73	S&P: indust	S&P's Common Stock Price Index: Industrials	0.25	0.28
74	S&P div yield	S&P's Composite Common Stock: Dividend Yield	0.13	0.41
75	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio	0.03	0.46
		Median	0.19	0.37
		Overall median	0.20	0.51

Notes to Table B1. Index refers to the index number of the variable in our database. FRED refers to variable name in the FRED database. PC1 is the fraction of the variance of the cyclical component of that variable that is explained by the first principal component. PC1&2 is the fraction of the variance of the cyclical component of that variable that is explained by the first and second principal components combined.

Table B2: Outliers associated with McCracken-Ng algorithm, 1-month-ahead regressions, and 24-month-ahead regressions

variable	id	description	McCracken-Ng		Regression (h=1)		Regression (h =24)	
			no.	dates	no.	dates	no.	dates
RPI	1	real personal income	1	2013:1	1	2013:1	0	
W875RX1	2	RPI less transfers	1	2013:1	1	2013:1	0	
IPDMAT	14	durables industrial production	1	1959:12	0		0	
CES1021-000001	35	employment: mining	6	1971:10,1971:12, 1977:12,1978:4, 1981:4,1981:6	6	1971:10,1971:12 1977:12,1978:4, 1981:4,1981:6	0	
MANEMP	37	employment: manufacturing	0		2	1970:10,1970:12	0	
DMANEMP	38	employment: durables	0		4	1964:10,1964:11, 1970:10,1970:12	0	
USGOVT	45	employment: government	0		2	1960:3,2010:5	0	
BUSINV	68	total business inventories	1	1982:1	1	1982:1	0	
M1SL	70	M1 money stock	2	2001:10,2009:1	0		0	
AMBSL	73	St. Louis monetary base	9	2008:9,2008:10, 2008:12-2009:3, 2010:2,2010:3, 2011:2	8	2008:9-2008:11, 2009:2,2009:3, 2009:10,2010:3, 2011:2	0	
TOTRESNS	74	total reserves	3	2001:10,2008:9, 2008:11	4	2001:9,2001:10, 2008:9,2008:10,	22	2008:11-2010:8
NON-BORRES	75	nonborrowed reserves	14	2001:10, 2008:1-2008:5, 2008:9, 2008:11-2009:5	18	2001:9, 2007:12-2009:4	22	2001:9, 2007:12-2008:4, 2008:9-2009:4, 2010:2,2010:3, 2010:5-2010:8, 2010:10,2010:12
REALLN	77	real estate loans	3	2006:10,2006:11, 2008:11	2	2006:10,2008:10	0	
NONREVSL	78	total non-revolving credit	4	1977:1,1977:2, 2010:12,2011:1	2	1977:1,2010:12	0	

Table B2 (continued)

variable	id	description	McCracken-Ng		Regression (h=1)		Regression (h =24)	
			no.	dates	no.	dates	no.	dates
CONSPI	79	nonrevolving consumer credit	0		3	1977:1,2010:12, 2013:1	0	
FEDFUNDS	84	fed funds rate	8	1979:10,1980:3, 1980:5,1980:11, 1980:12,1981:2, 1981:5,1982:8	6	1979:10,1980:3, 1980:5,1980:10, 1980:11,1981:5	0	
CP3M	85	3-month commer- cial paper rate	7	1980:3,1980:5, 1980:11,1980:12, 1981:5,1981:11, 1982:8	4	1980:3,1980:5, 1980:10,1981:5	0	
TB3MS	86	3-month Tbill rate	5	1980:3,1980:5, 1981:5,1981:11, 1982:8	4	1980:4,1980:5, 1981:5,1981:6	0	
TB6MS	87	6-month Tbill rate	2	1980:5,1981:11	2	1980:4,1980:5	0	
GS1	88	1-year Treasury rate	2	1980:5,1981:11	0		0	
COMPAPFF	93	com paper fed funds spread	0		2	1981:1,1981:6	0	
TB3SMFFM	94	3-month Tbill ff spread	0		3	1980:4,1981:2, 1981:6	0	
TB6SMFFM	95	6-month Tbill ff spread	0		4	1980:4,1980:5, 1980:12,1981:2	0	
T1YFFM	96	1-year Tbill ff spread	0		4	1980:4,1980:5, 1980:12,1981:2	0	
T5YFFM	97	5-year Treas ff spread	0		2	1980:5,1981:2	0	
T10YFFM	98	10-year Treas fed funds spread	0		1	1980:5	0	
AAAFFM	99	Aaa cor- porate fed funds spread	0		3	1980:5,1980:11, 1981:2	0	

Table B2 (concluded)

variable	id	description	McCracken-Ng		Regression (h=1)		Regression (h =24)	
			no.	dates	no.	dates	no.	dates
BAAFFM	100	Baa corporate fed funds spread	0		2	1980:5,1980:11	0	
PPIITM	108	PPI intermediate materials	0		1	2008:11	0	
PPICRM	109	PPI crude materials	1	2001:2	0		0	
OILPRICE	110	crude oil price	2	1974:1,1974:2	1	1974:1	0	
CPITRNSL	115	CPI transportation	0		1	2008:11	0	
CUS-R0000SAS	119	CPI services	0		1	1980:7	0	
DSERRG3-M086SBEA	126	PCE consumption	1	2001:10	0		0	
MZMSL	131	MZM money stock	1	1983:1	1	1983:1	0	
DTCOLN-VHFNM	132	motor vehicle loans	3	1977:12,2010:3,2010:4	1	2010:3	0	
DTCTHFNM	133	consumer loans	2	2010:12,2011:1	2	2010:12,2011:1	0	
total			79		98		44	