Principal Component Analysis for Nonstationary Series

James D. Hamilton, UCSD

Jin Xi, UCSD

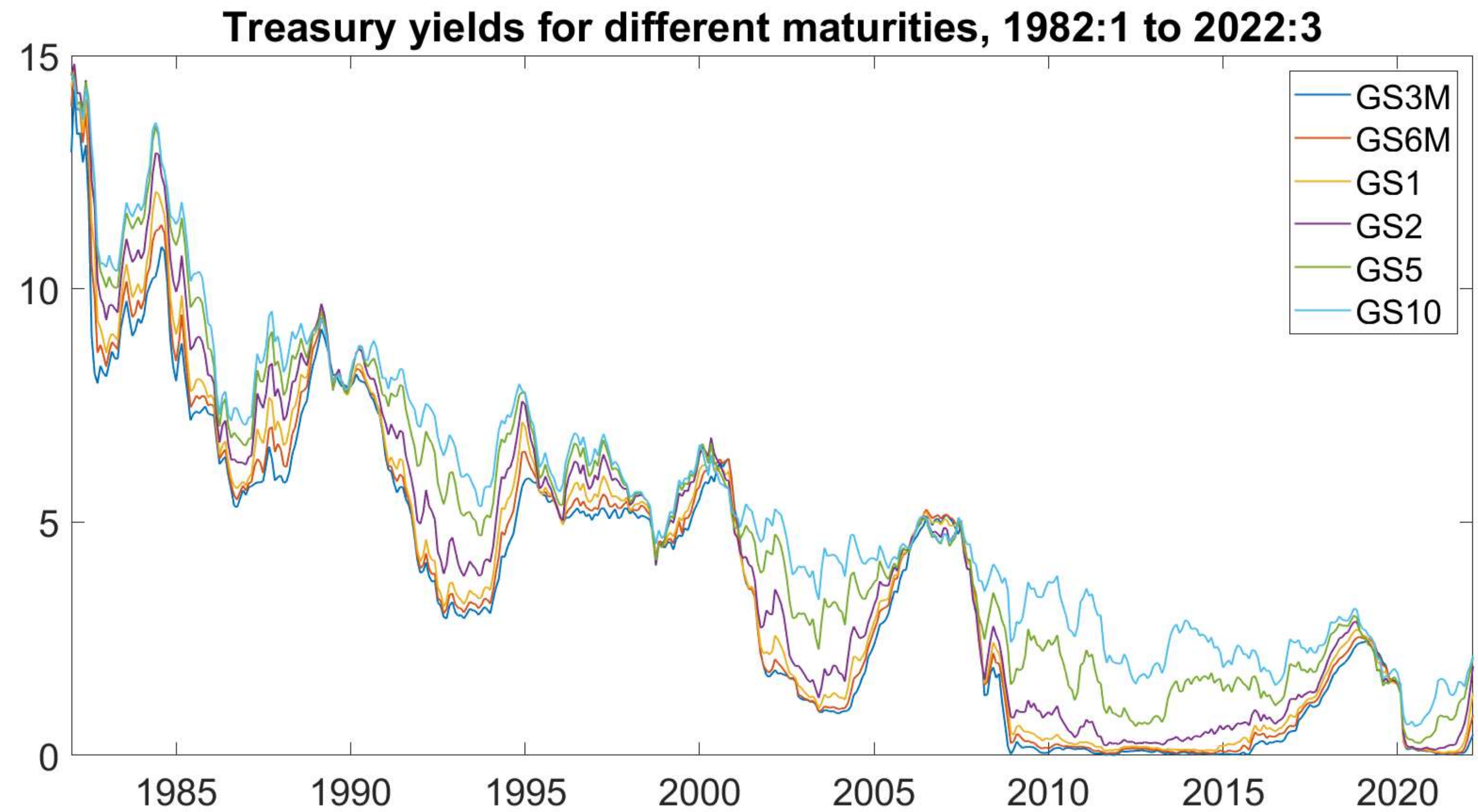
Approaches to large data sets

- Sparsity
  - assumption: most variables not useful
  - examples: LASSO, random forest
- Shrinkage
  - assumption: all variables used but each gets small weight
  - Principal components, ridge regression, Bayesian inference
- Problem: how use these methods when some variables may be nonstationary?

- Principal components: subtract sample mean from each variable and divide by standard deviation
- Calculate eigenvectors of correlation matrix associated with largest eigenvalues
- Use eigenvectors associated with largest eigenvalues to calculate linear combinations of variables

- Problem: if a variable is nonstationary, sample mean and standard deviation do not converge to any population parameter
- PCA when some variables are nonstationary can give very misleading results
  - Onatski and Wang, Econometrica 2021
- Usual approach: determine transformation necessary to make each individual variable stationary

Problem 1: necessary transformation can be unclear



- Many finance applications apply PCA to yields themselves
- McCracken and Ng (JBES 2016) use first-differences of yields or yield spreads
- Crump and Gospodinov (Econometrica 2022) use excess returns or first-differences of returns

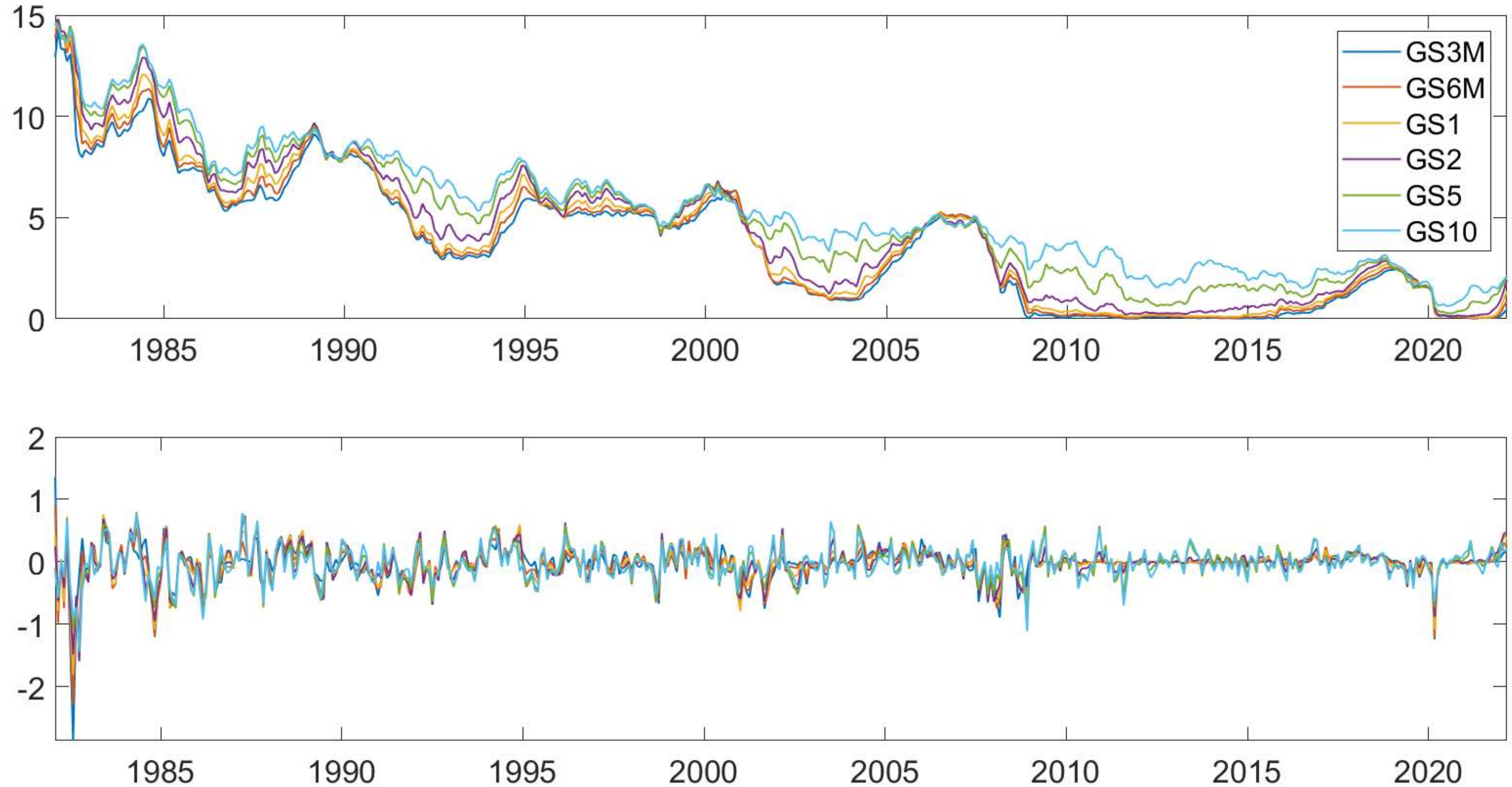
Problem 2: reproducibility

- Need to communicate decision used for every variable in the study
- Another researcher who did not use same transformations could get different answers

Problem 3: appropriateness of the method

- Suppose we knew for certain that variable 1 is random walk and variable 2 is AR(1) with coefficient 0.99
- Current approach would say use differences of variable 1 and levels of variable 2
- But these have very different properties

Levels and first-differences of yields



Hamilton (REStat, 2018)

- The error in predicting a variable 2 years from now as a linear function of recent values:
  - is a stationary population magnitude for a broad class of nonstationary processes such as ARIMA(p,d,q) or processes stationary around dth-order polynomial time trends
  - could be described as cyclical component of the series
  - can be consistently estimated by OLS regression without knowing d

Example: suppose Δy_{it} is stationary $(d = 1)$.

Accounting identity:

$$y_{it} = y_{i,t-h} + \sum_{j=0}^{h-1} \Delta y_{i,t-j}$$

y_{it} can be written as linear function of $y_{i,t-h}$ plus something stationary.

Error predicting y_{it} from $y_{i,t-h}, y_{i,t-h-1}, \ldots, y_{i,t-h+p-1}$ is stationary.

OLS minimizes sample squared forecast errors and consistently estimates this population object.

Suppose $\Delta^2 y_{it}$ is stationary

$(d = 2)$.

Accounting identity:

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

y_{it} can be written as linear function of

$y_{i,t-h}, y_{i,t-h-1}$ plus something stationary.

y_{it} = observation on variable i in period t

$$y_{it} = \alpha_{i0} + \alpha_{i1} y_{i,t-h} + \alpha_{i2} y_{i,t-h-1} + \cdots$$
$$+ \alpha_{ip} y_{i,t-h-p+1} + c_{it}$$

c_{it} = population magnitude (exists for large class of possible data-generating processes for y_{it})

$\hat{c}_{it}$ = OLS residual

Proposal: estimate by OLS separately for each $i = 1, \ldots, N$

$$y_{it} = z_{it}'\alpha_i + c_{it}$$

$$z_{it}' = (1, y_{i,t-h}, y_{i,t-h-1}, \ldots, y_{i,t-h-p+1})'$$

Perform PCA on regression residuals $\hat{c}_{it}$.

In principle, would work for any finite h.

$h = 1$ would correspond to principal component of 1-month-ahead forecast errors which is not usual object of interest.

For h too large, c_{it} has lots of persistence and very large sample needed to estimate.

We recommend $h = 24$ and $p = 12$ for monthly data.

Suppose true cyclical components are characterized by an approximate factor structure as in Stock and Watson (JASA 2002):

$$\underset{(N\times 1)}{C_t} = \underset{(N\times r)}{\Lambda}\ \underset{(r\times 1)}{F_t} + \underset{(N\times 1)}{e_t}$$

$$\lim_{N\to\infty} \sup_t \sum_{s=-\infty}^{\infty} |E[e_t' e_{t+s}/N]| < \infty$$

$$\lim_{N\to\infty} \sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |E[e_{it} e_{jt}]| < \infty$$

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

$v_{it} = \hat{c}_{it} - c_{it}$

If $v_{it} \overset{m.s.}{\to} 0$ uniformly in i and t, then subject to normalization conditions,

$$\hat{f}_{jt} \overset{p}{\to} f_{jt} \ \forall j, t$$

$$T^{-1} \sum_{t=1}^{T} \hat{f}_{jt}^2 \overset{p}{\to} E(f_{jt}^2) \text{ for } j \leq r$$

$$T^{-1} \sum_{t=1}^{T} \hat{f}_{jt}^2 \overset{p}{\to} 0 \text{ for } j > r$$

Should we expect that $E(v_{it}^2) \to 0$?

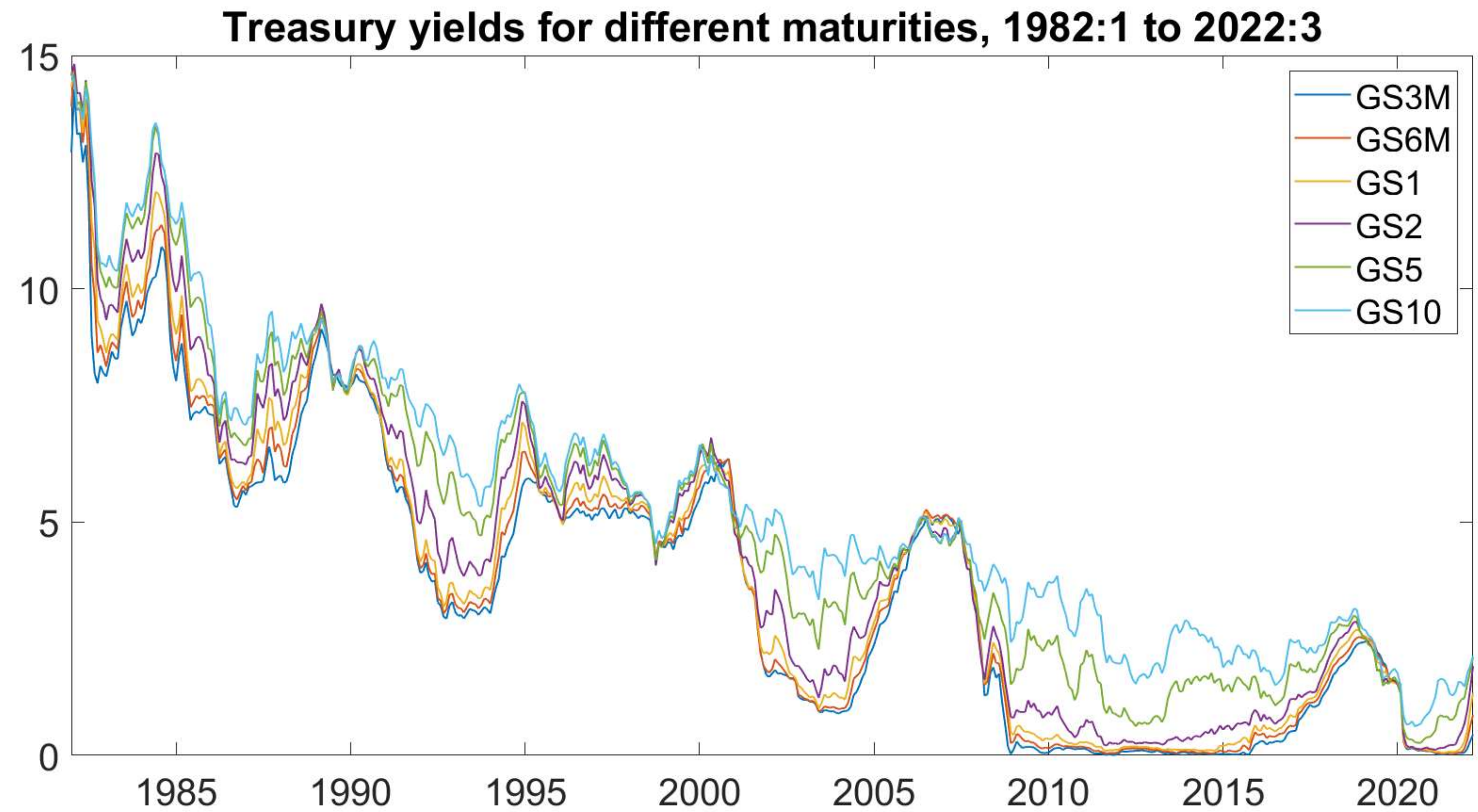
$$\sum_{t=1}^{T} v_{it}^2 = (\alpha_i - \hat{\alpha}_i)' \sum_{t=1}^{T} z_{it} z_{it}' (\alpha_i - \hat{\alpha}_i)$$

This is proportional to OLS Wald test of the (correct) null hypothesis that α_i is the true value.

$\sum_{t=1}^{T} v_{it}^{2}$ converges in distribution to some variable in a variety of stationary and nonstationary settings.

$$T^{-1}\sum_{t=1}^{T} v_{it}^{2} \xrightarrow{p} 0$$

Application 1: Describing the yield curve



Conventional PCA on levels:

$$\dot{y}_{it} = (y_{it} - \bar{y}_i)/\hat{\sigma}_i$$

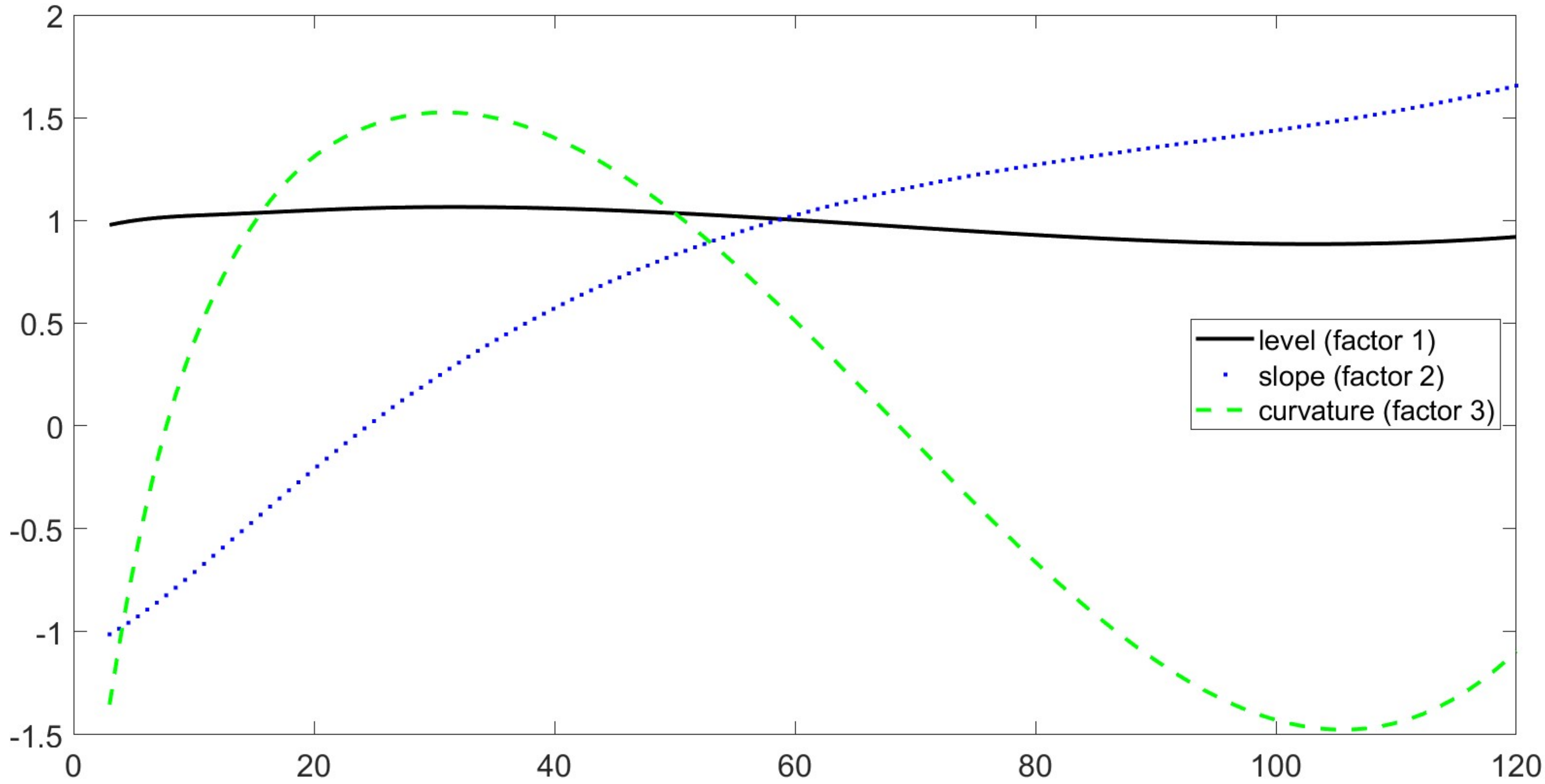
$$\underset{(N\times 1)}{\dot{y}_t} = \underset{(N\times r)}{\tilde{\Lambda}} \underset{(r\times 1)}{F_t} + \underset{(N\times 1)}{\tilde{e}_t}$$

$$\underset{(r\times 1)}{\tilde{F}_t} = \underset{(r\times N)}{\tilde{\Lambda}'} \underset{(N\times 1)}{\dot{y}_t}$$

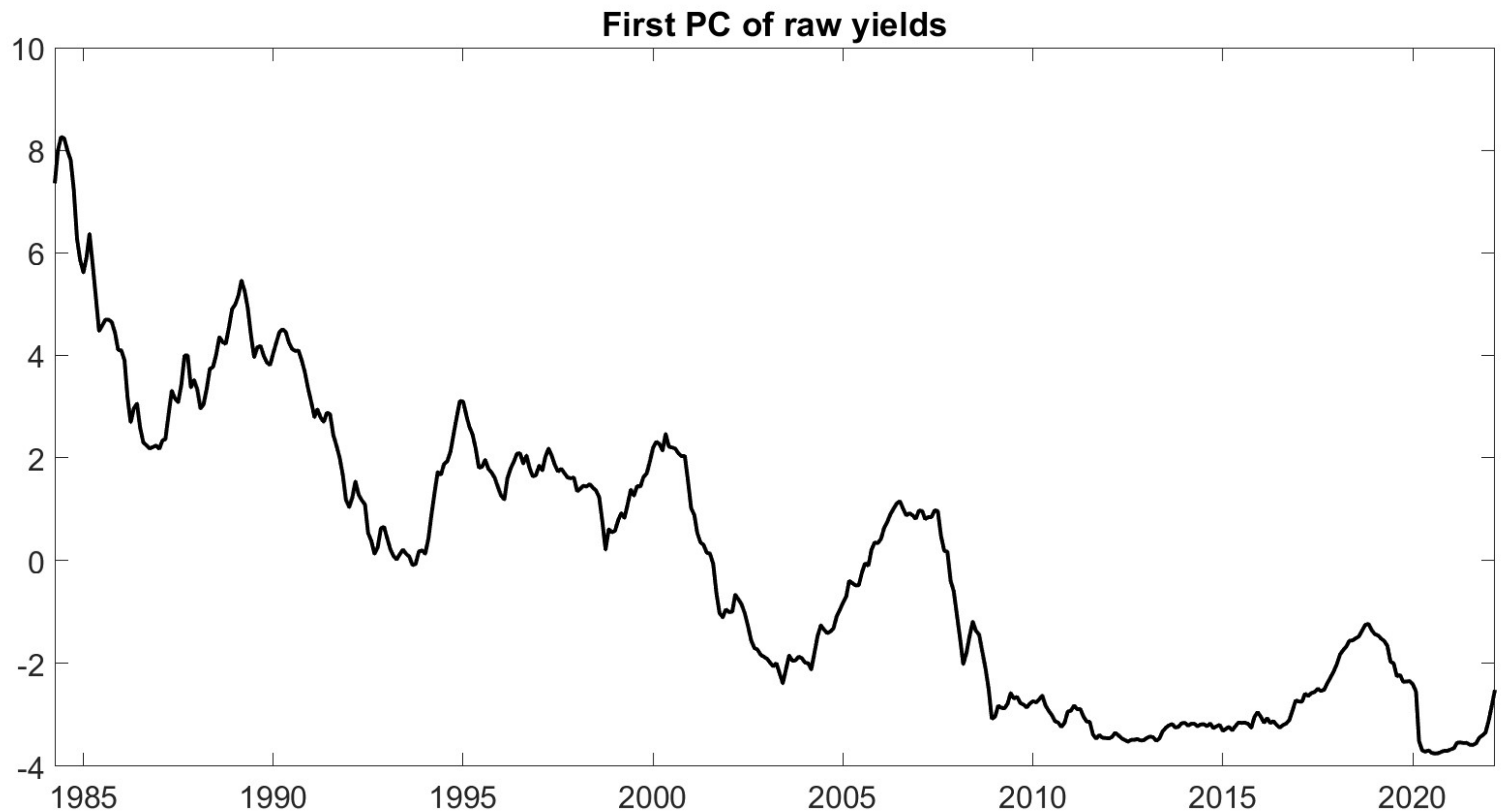
Let $\tilde{\lambda}_j =$ eigenvector of correlation matrix of raw yields associated with jth largest eigenvalue.

Consider plot of weights of $\tilde{\lambda}_j$ as a function of maturity of yield i.

Factor loadings for first 3 PC of raw yields as a function of maturity in months



First PC of raw yields as a function of time

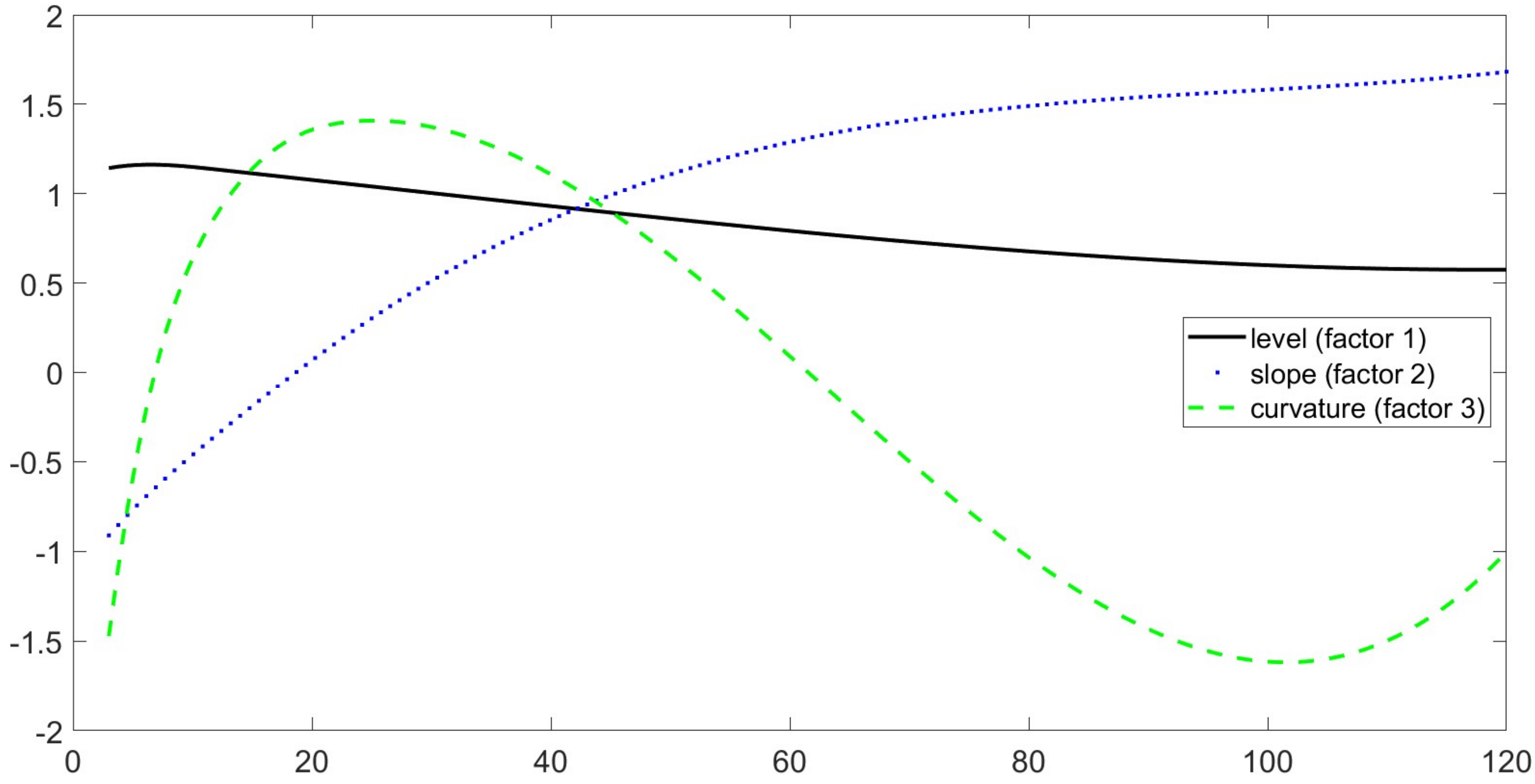


$\hat{c}_{it} =$ residual from OLS regression of y_{it} on $(1, y_{i,t-24}, y_{i,t-25}, \ldots, y_{i,t-35})$.

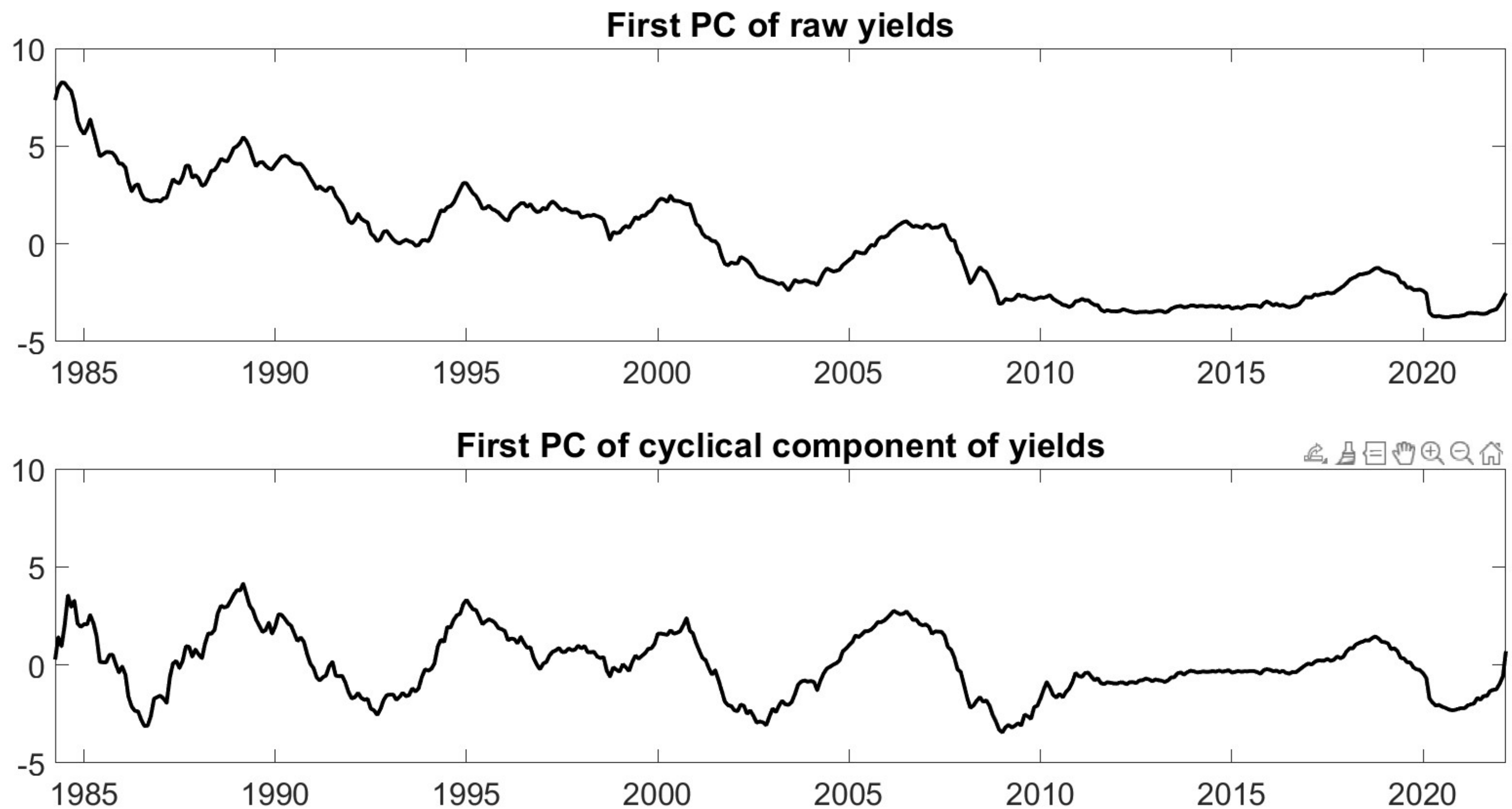
$\hat{\lambda}_j =$ eigenvector of correlation matrix of $\hat{c}_{it}$ associated with jth largest eigenvalue.

Now plot elements of $\hat{\lambda}_j$ as a function of maturity of yield i.

Factor loadings for first 3 PC of cyclical components of yields



First principal component of raw yields and cyclical component of yields



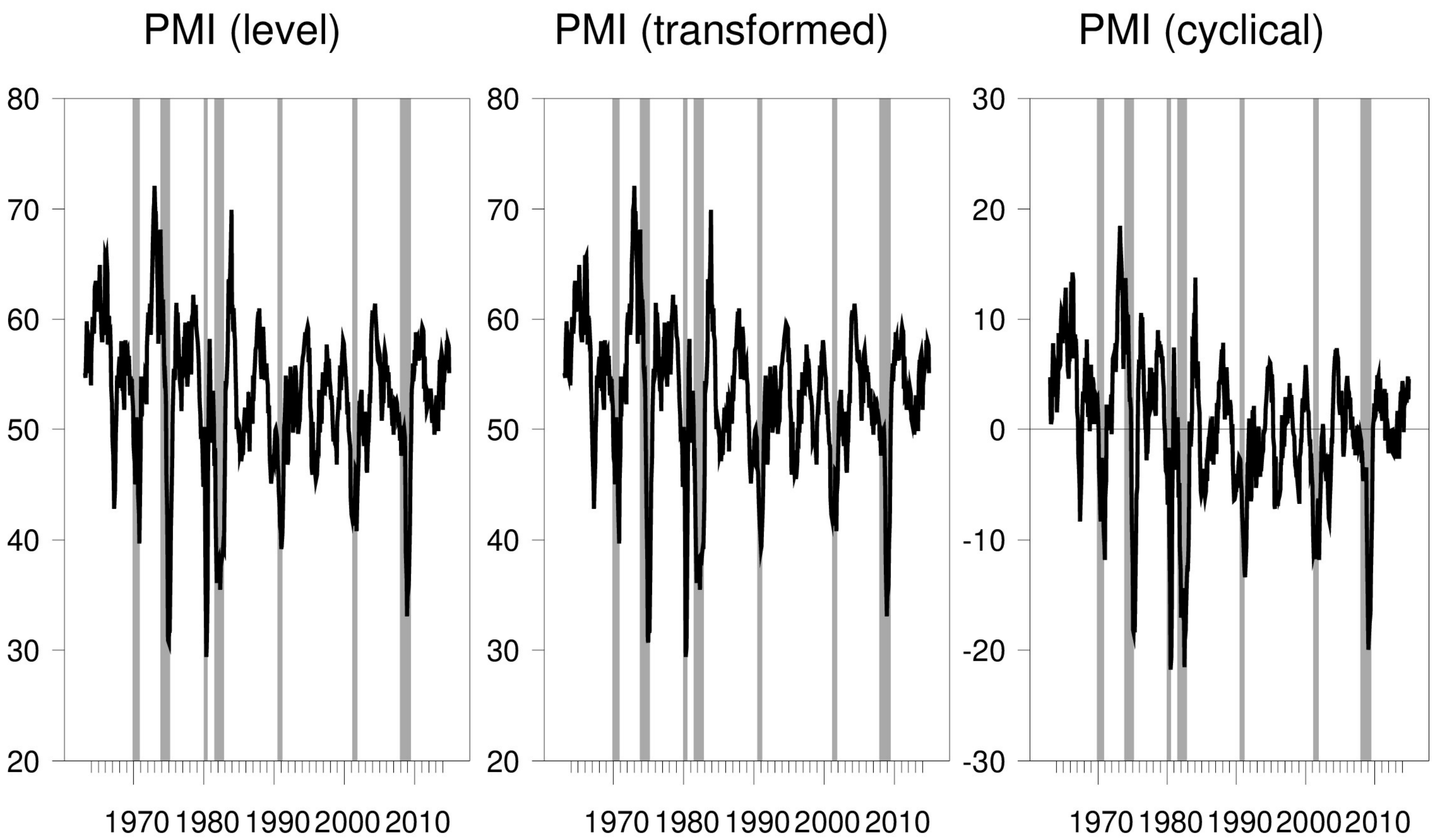
- For this application, PCA on levels works fine because all variables share the same trend component.
- Principal components capture both level and trend.
- If we mix U.S. nominal interest rates with other variables that have different trends, nonstationarity is bigger concern.

Application 2. Large macroeconomic data set

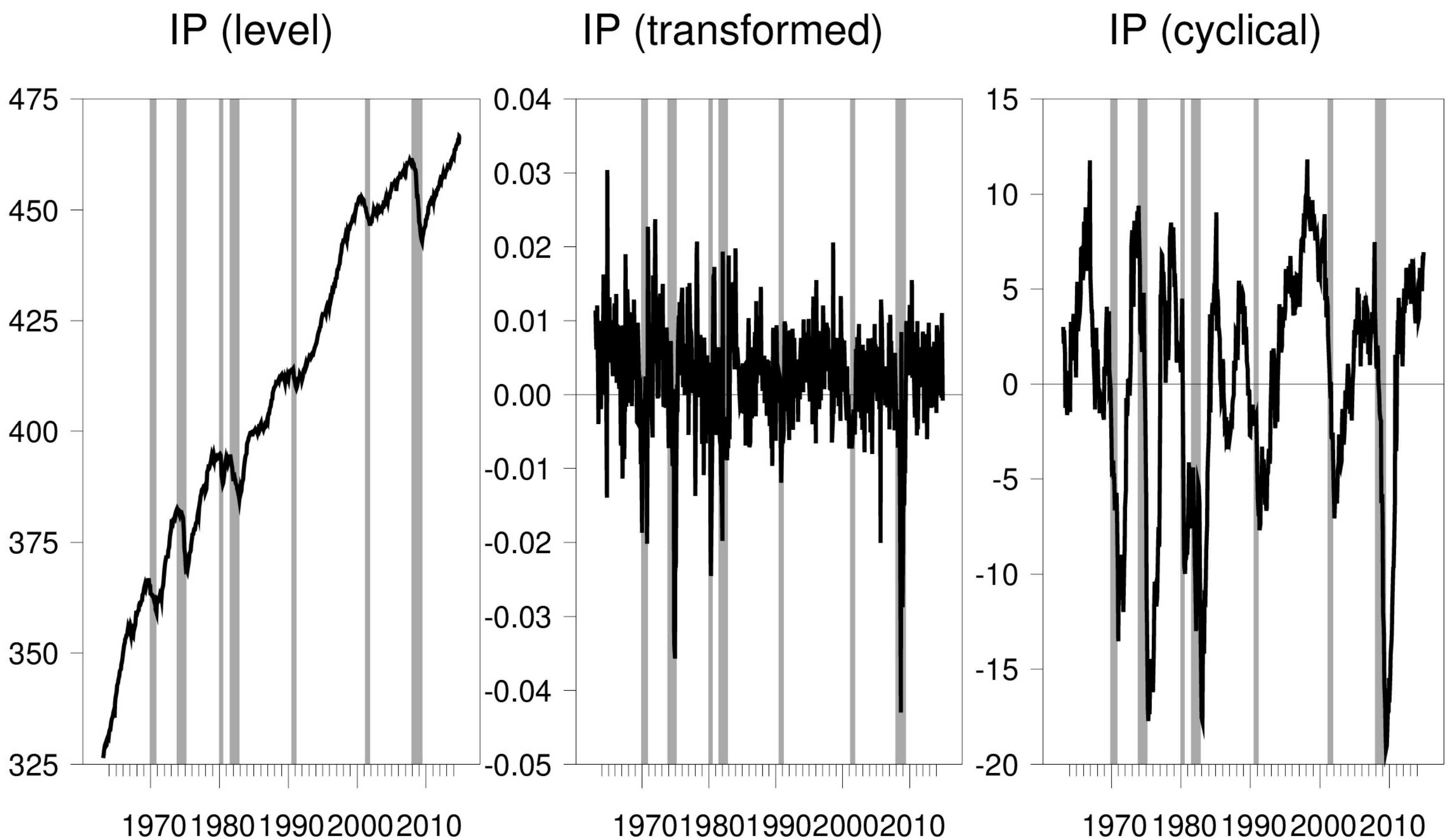
- Stock and Watson (JME 1999) found that first PC of a set of 85 different measures of real economic activity was best way to use big data set to predict inflation.
- This evolved into the Chicago Fed National Activity Index (CFNAI).

- McCracken and Ng (JBES 2016) developed FRED-MD data set
  - output and income; labor market; housing; consumption, orders, and inventories; money and credit; interest and exchange rates; prices; and stock market
  - 134 variables in 2015:4 vintage
  - continually updated
  - McCracken and Ng selected a transformation to make each variable stationary

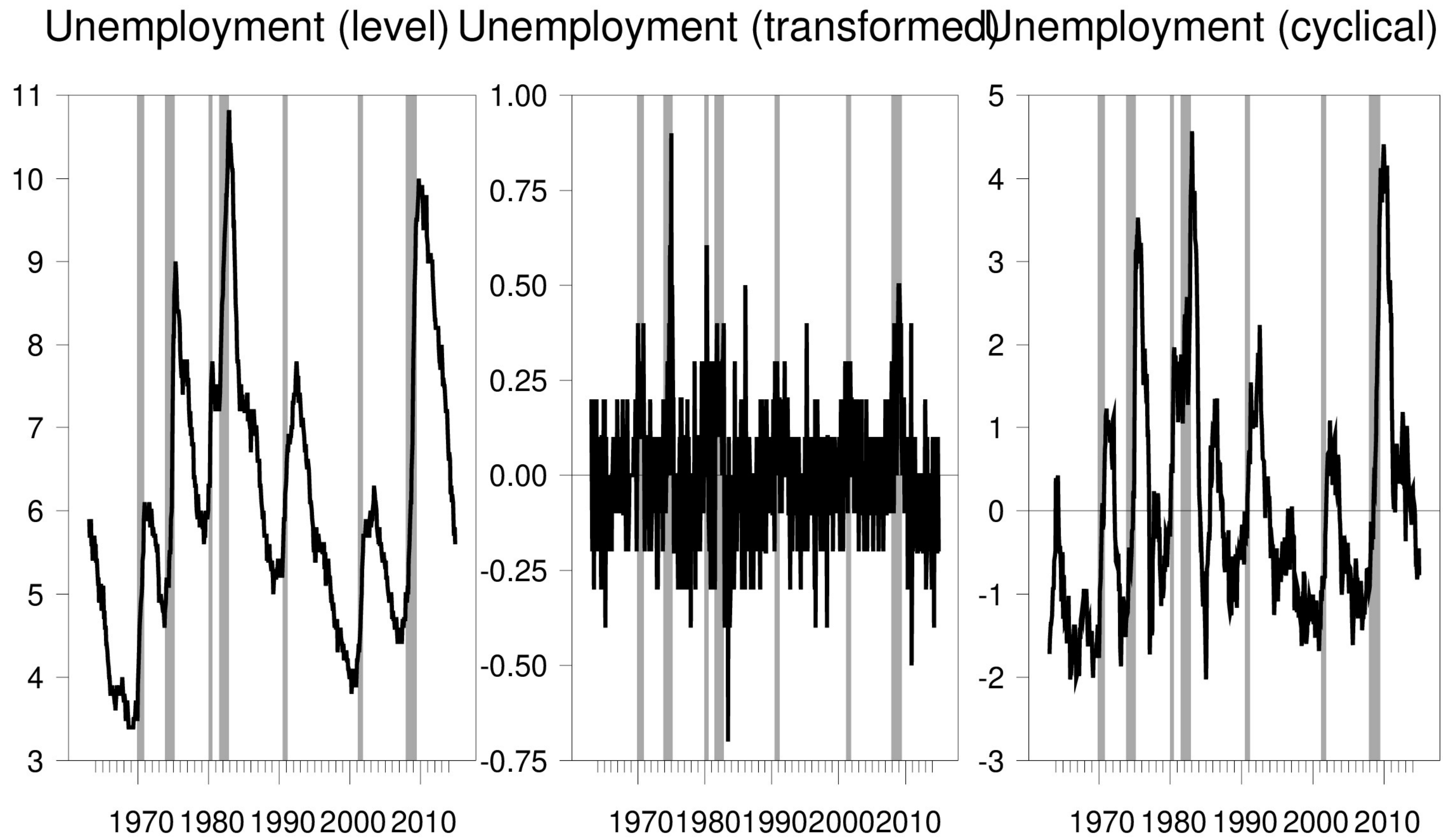
Plant managers index



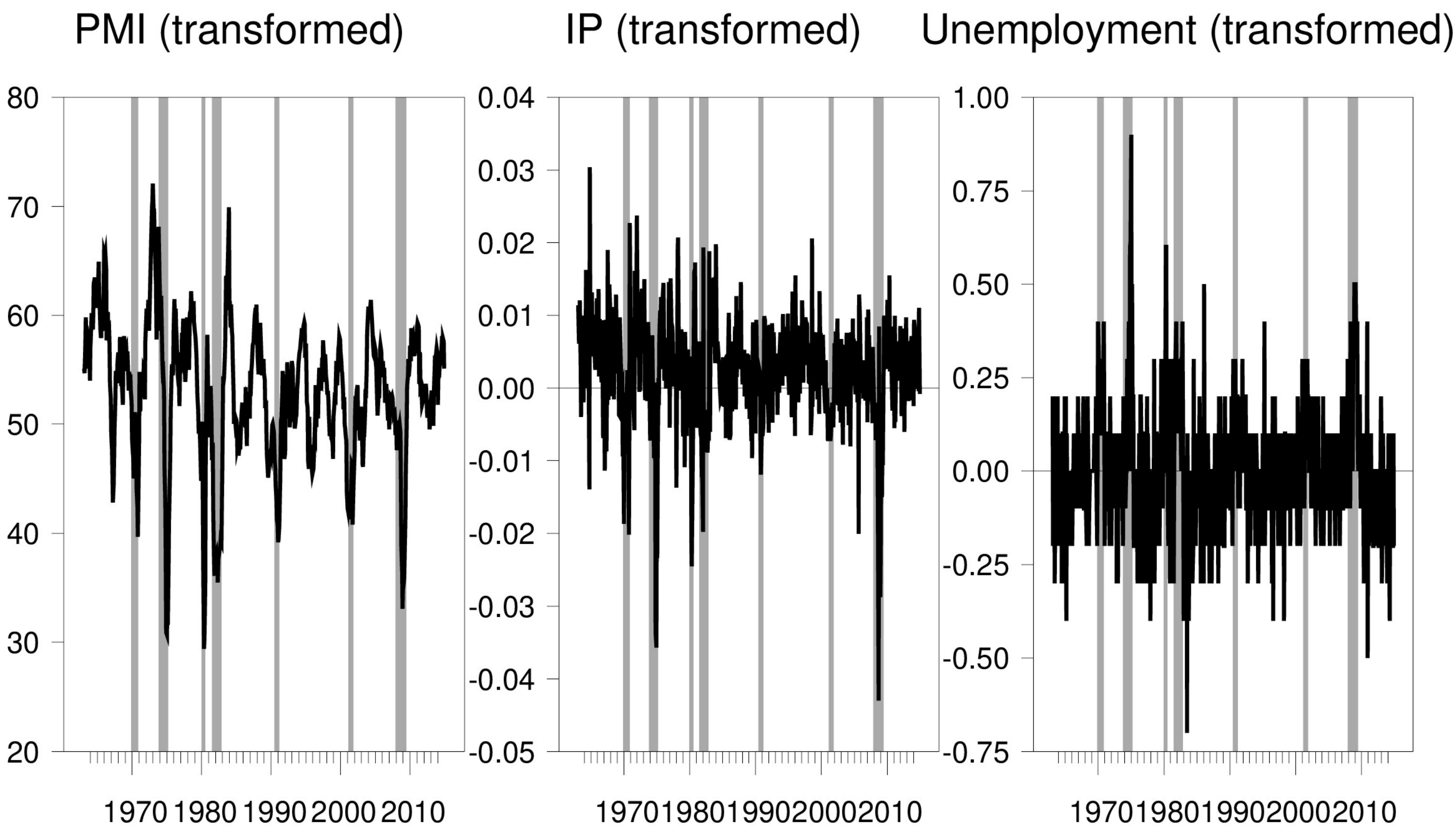
Log of industrial production index



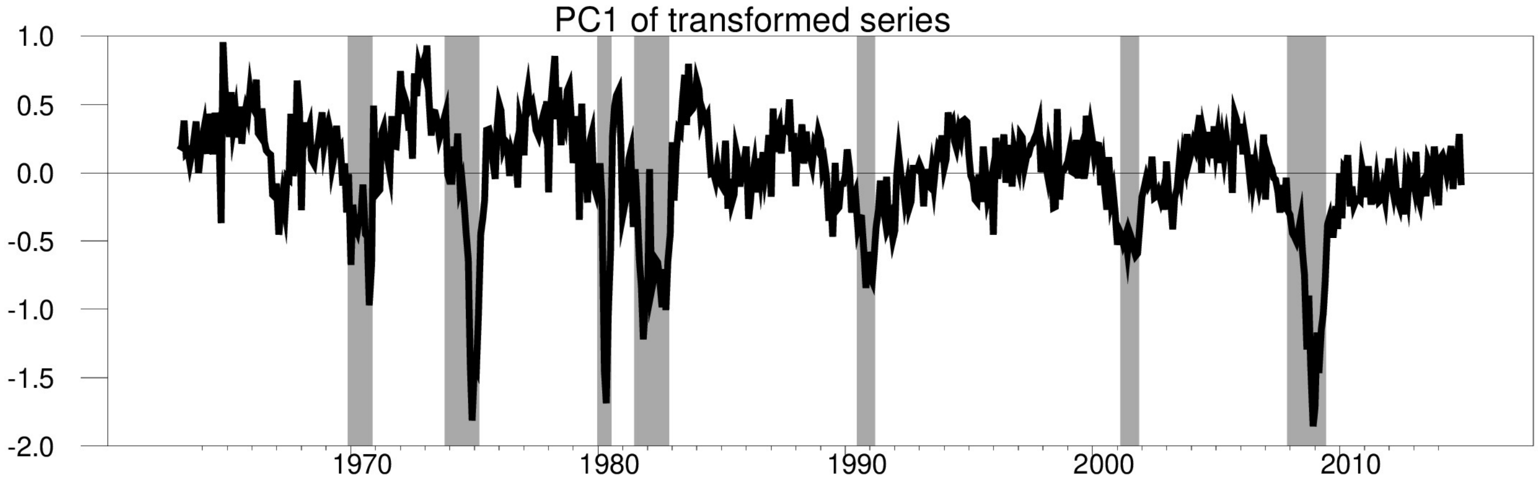
Unemployment rate

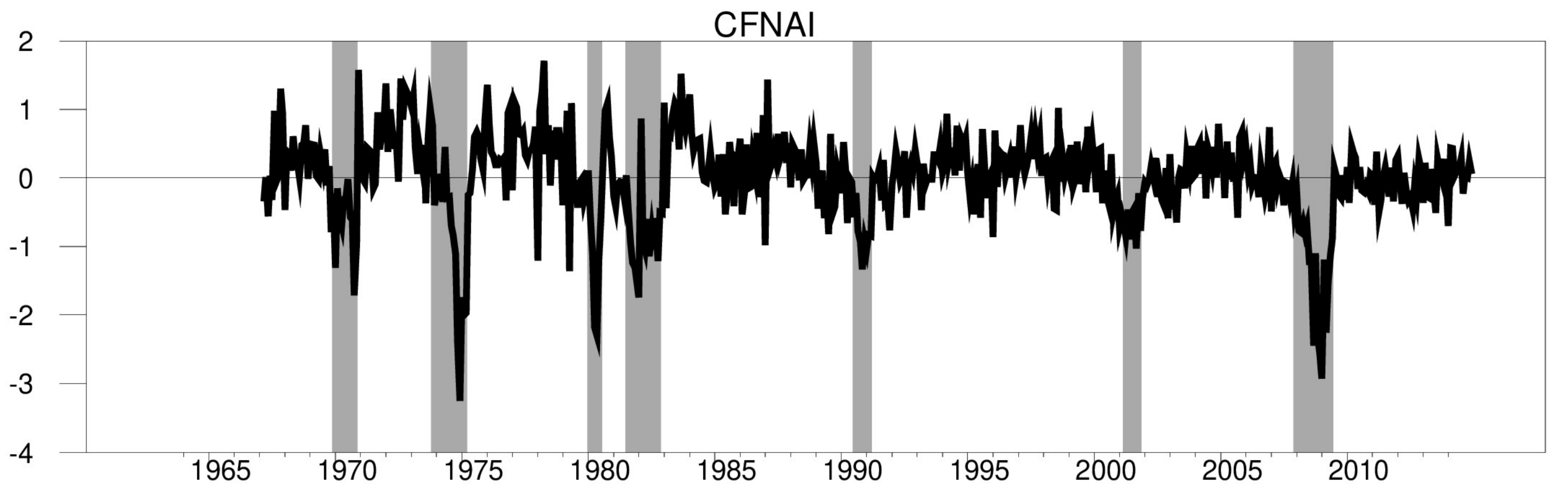


Series as transformed by McCracken and Ng

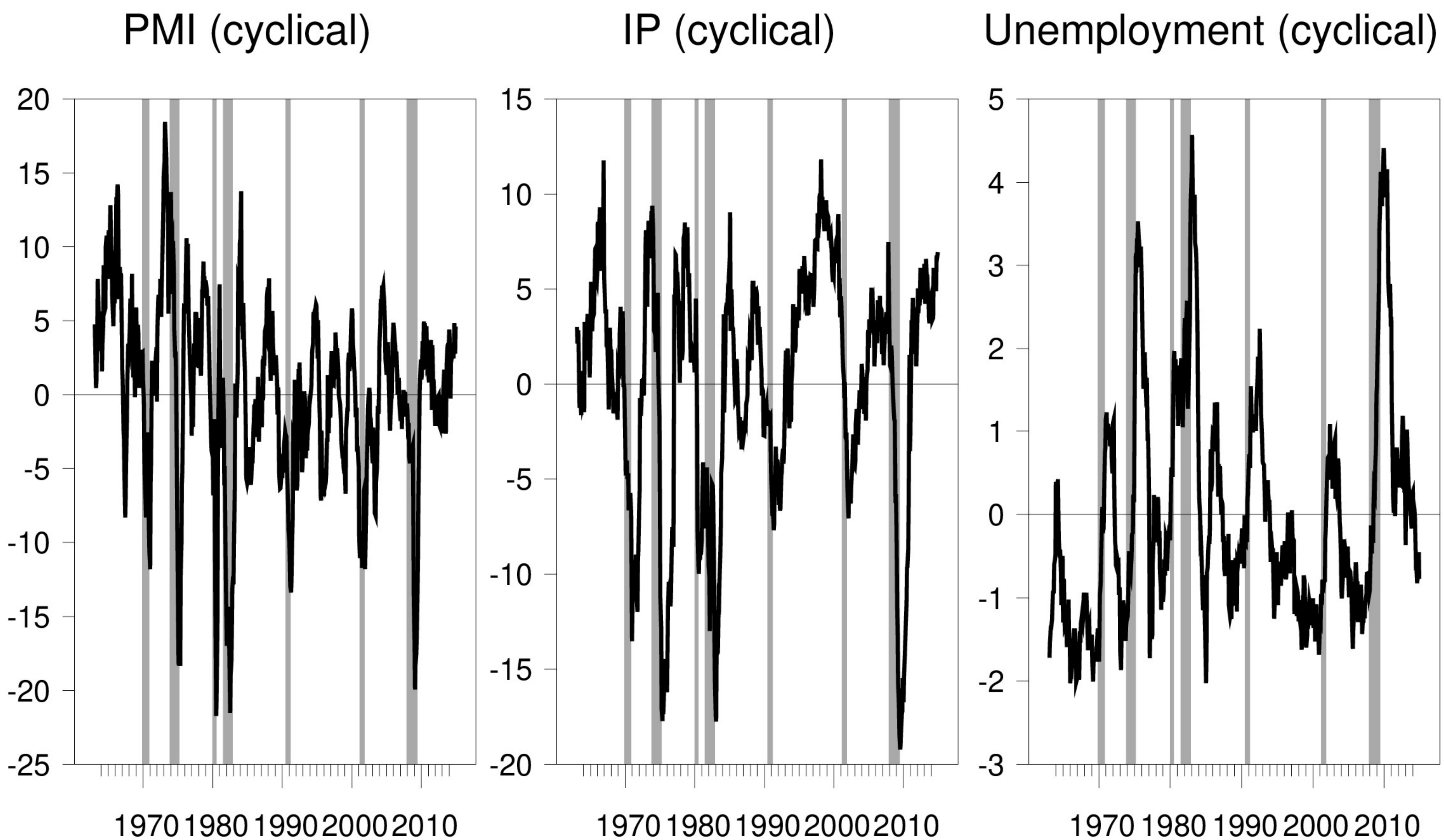


PC1 of transformed data and CFNAI

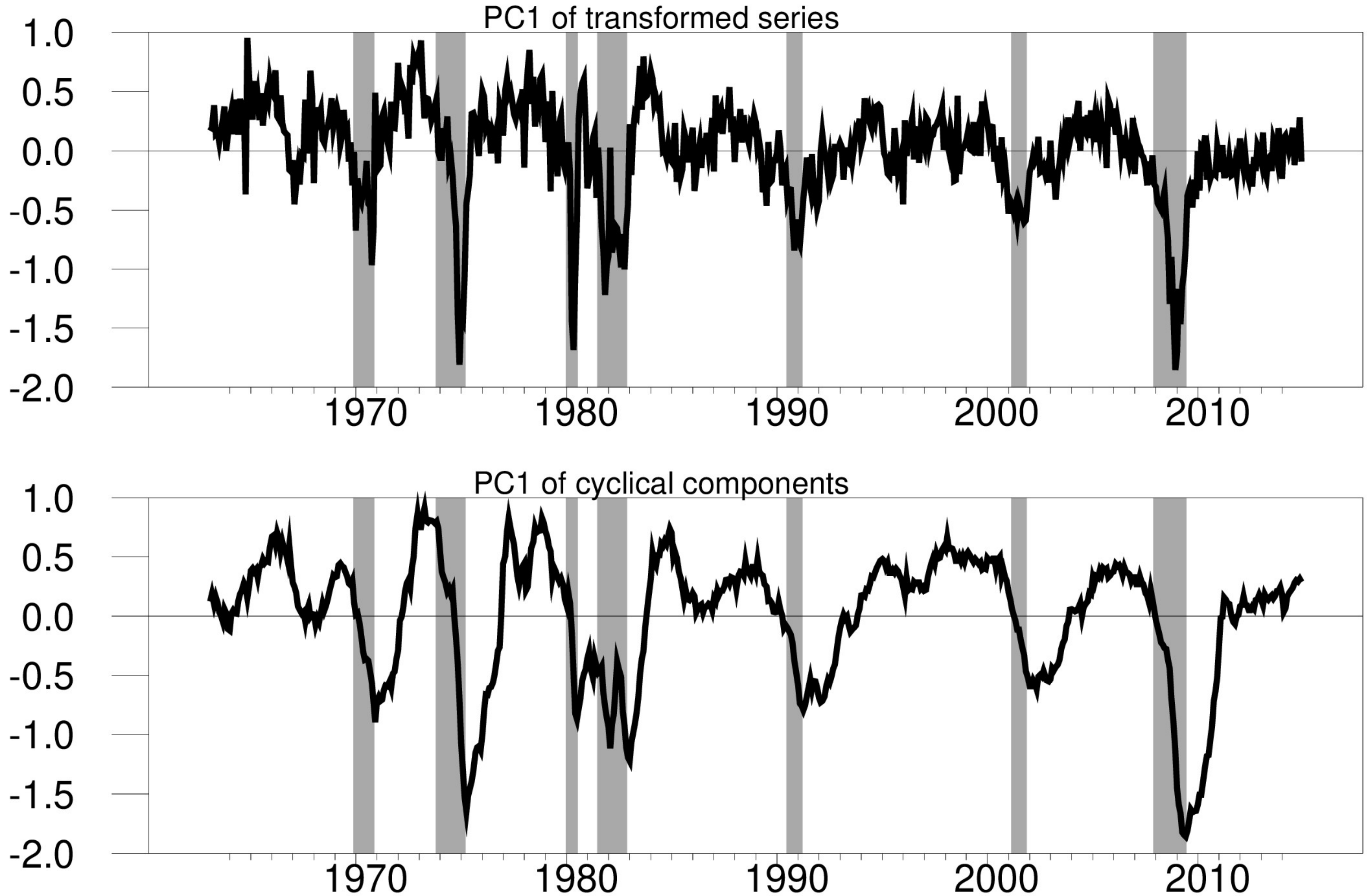




Cyclical components as identified by regressions



PC1 of transformed data and of cyclical components



Dealing with outliers

- Traditional approach to outliers:
  - Calculate interquartile range of transformed data
  - If observation exceeds *k* times the interquartile range, treat as missing
  - CFNAI historically used *k* = 6
  - McCracken-Ng used *k* =10 and found 79 outliers in 22 different variables in 1960-2014 data set

How identify outliers if don't know form of nonstationarity?

If we observed true c_{it}, could compare it with its interquartile range.

Can estimate $\hat{c}_{it}$, but outliers will unduly influence regression.

Consider regression that does not use y_{it} as dependent variable.
Use these coefficients to predict y_{it} and form "leave-one-out" residual $\tilde{c}_{it}$.
Compare $\tilde{c}_{it}$ with its interquartile range.
Leave-one-out regression with $h = 1$ identifies similar but not identical outliers as McCracken-Ng.
98 outliers in 31 different variables in 1960-2014 data set.

Table 1 (concluded)

			McKracken-Ng		Regression (h=1)		Regression (h =24)	
variable	id	description	no.	dates	no.	dates	no.	dates
AAAFFM	99	Aaa corporate fed funds spread	0		3	1980:5,1980:11, 1981:2	0	
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	
CUSR0000SAS	119	CPI services	0		1	1980:7	0	
DSERRG3M086SBEA	126	PCE consumption	1	2001:10	0		0	
MZMSL	131	MZM money stock	1	1983:1	1	1983:1	0	
DTCOLNVHFNM	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	

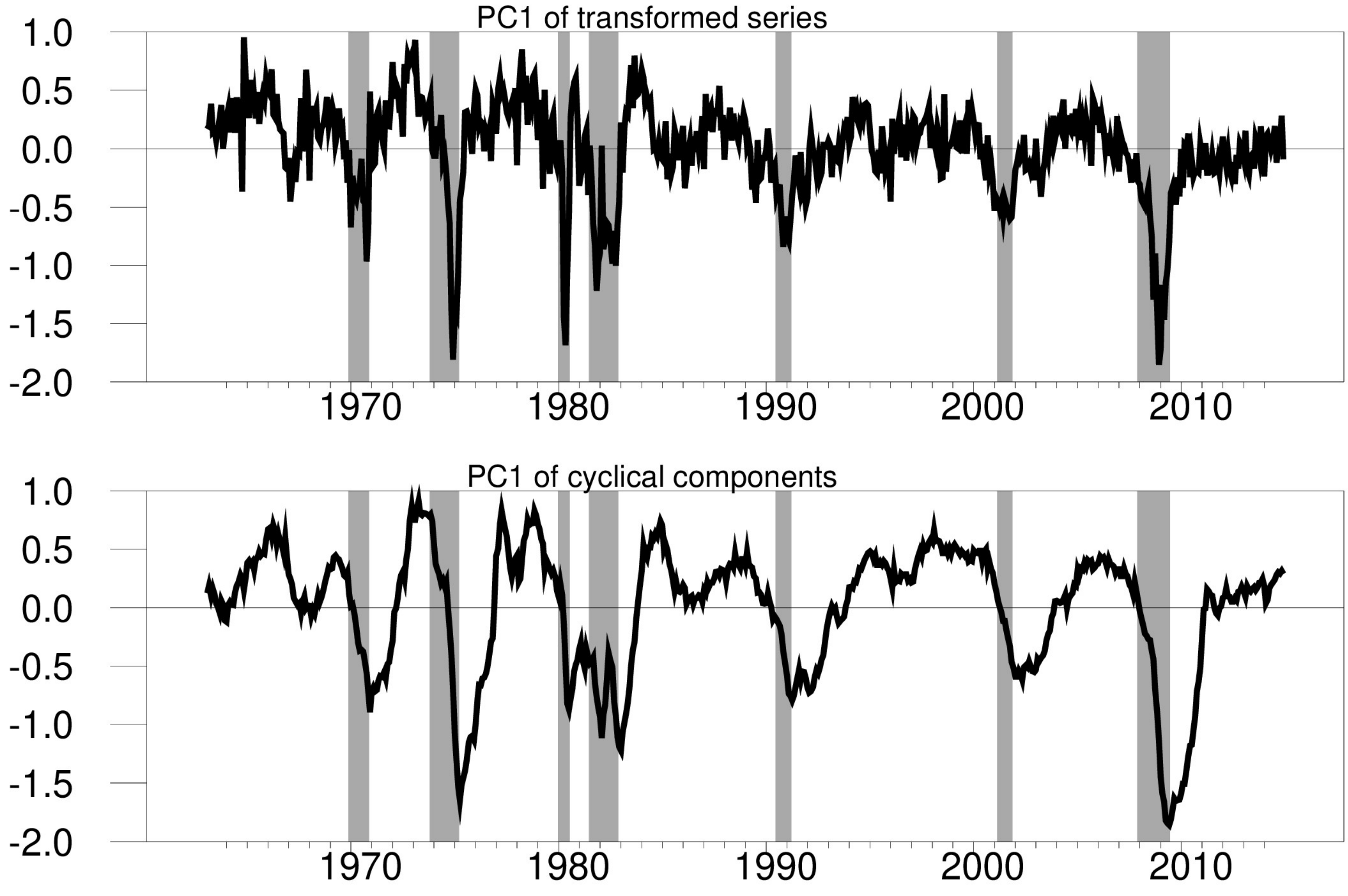
But regressions with $h = 24$ have far fewer outliers.

If y_{it} is random walk, then c_{it} is sum of 24 individual innovations.

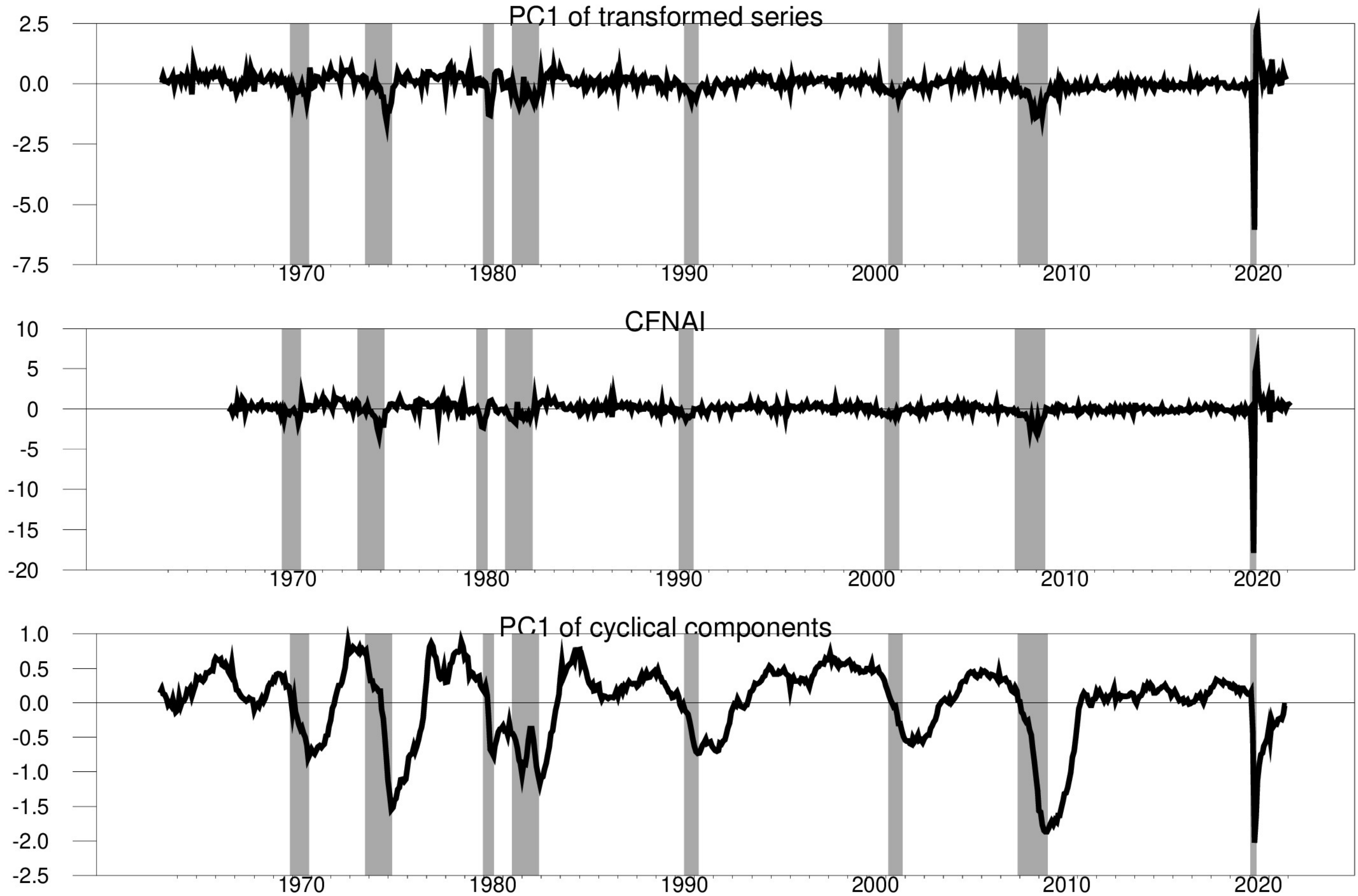
By CLT, c_{it} has a distribution much closer to Normal distribution.

In 1960-2014, outliers detected in only two variables (nonborrowed and total reserves) essentially all in the Great Recession.

Our recommended procedure makes no corrections for outliers



- When dataset is expanded to include recent data, McCracken-Ng identifies 40 outliers in 2020:4 observations alone
- CFNAI modified their treatment of outliers to accommodate COVID observations
- Even so, the index value in 2020:4 for both McCracken-Ng and CFNAI is a huge outlier; must plot on new scale


PC1 of transformed series
2.5
0.0
-2.5
-5.0
-7.5
1970
1980
1990
2000
2010
2020
CFNAI
10
5
0
-5
-10
-15
-20
1970
1980
1990
2000
2010
2020
PC1 of cyclical components
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
1970
1980
1990
2000
2010
2020

- Cyclical components using h = 24 show outliers for only two variables in 2020:4
  - Initial claims for unemployment insurance
  - Number unemployed for 5 weeks or less
- We construct PC1 just as before with no changes and no outlier corrections
- PC1 of cyclical components is plotted on same scale before and after 2020