Principal Component Analysis for Nonstationary Series

James D. Hamilton and Jin Xi
UCSD



Traditional principal component analysis

1) Figure out how to make each y_{it} stationary.

2) Calculate $\tilde{y}_{it} = (y_{it} - \hat{\mu}_i)/\hat{\sigma}_i$.

3) Find eigenvectors associated with biggest eigenvalues of $\hat{\mathbf{\Omega}} = T^{-1} \sum_{t=1}^{T} \tilde{\mathbf{y}}_t \tilde{\mathbf{y}}_t'$.

Problems with step 1:

a) Sometimes not clear what transformation to use.

b) Should we treat AR(1) with $\rho = 0.99$ completely differently from $\rho = 1$?

Consider trying to forecast y_{it} using a linear function of its m most recent values as of $t - h$:

$$y_{it} = \boldsymbol{\alpha}_i' \mathbf{z}_{i,t-h} + c_{it}$$

$$\mathbf{z}_{i,t-h} = (1, y_{i,t-h}, y_{i,t-h-1}, \ldots, y_{i,t-h-m+1})'$$

$$y_{it} = \boldsymbol{\alpha}_i' \mathbf{z}_{i,t-h} + c_{it}$$

Hamilton (REStat, 2018): for a large range of nonstationary processes

1) c_{it} is stationary.

2) $\boldsymbol{\alpha}_i$ can be consistently estimated by OLS.

Example 1: 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.

Example 2: 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.

Our proposal

1) Estimate by OLS $y_{it} = \boldsymbol{\alpha}_i' \mathbf{z}_{i,t-h} + c_{it}$

for $\mathbf{z}_{i,t-h} = (1, y_{i,t-h}, y_{i,t-h-1}, \ldots, y_{i,t-h-m+1})'$.

2) Calculate $\tilde{y}_{it} = (y_{it} - \hat{\boldsymbol{\alpha}}_i' \mathbf{z}_{i,t-h})/\hat{\sigma}_i$.

3) Find eigenvectors associated with biggest eigenvalues of $\hat{\boldsymbol{\Omega}} = T^{-1} \sum_{t=1}^{T} \tilde{\mathbf{y}}_t \tilde{\mathbf{y}}_t'$.

We suggest to use $h = 2$ years as definition of cyclical component of y_{it}.

Verifying that this works

Step 1: Assume that true cyclical components $c_{1t}, \ldots, c_{Nt}$ satisfy an approximate factor structure with true factors $f_{1t}, \ldots, f_{rt}$.

$$\underset{(N\times 1)}{\mathbf{C}_t} = \underset{(N\times r)}{\mathbf{\Lambda}} \underset{(r\times 1)}{\mathbf{F}_t} + \underset{(N\times 1)}{\mathbf{e}_t}$$

$$\lim_{N\to\infty} \sup_t \sum_{s=-\infty}^{\infty} |E[\mathbf{e}_t' \mathbf{e}_{t+s}/N]| < \infty$$

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

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

Verifying that this works

Step 2: Verify that principal components estimated from OLS residuals $\hat{c}_{1t}, \ldots, \hat{c}_{Nt}$ consistently estimate space spanned by true factors $f_{1t}, \ldots, f_{rt}$.

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

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

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

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

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

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

$$v_{it} = (\boldsymbol{\alpha}_{i0} - \hat{\boldsymbol{\alpha}}_i)' \mathbf{z}_{it}$$

single stationary regressor:

$$z_{it} \sim O_p(1) \qquad (\alpha_{i0} - \hat{\alpha}_i) \sim o_p(1)$$

single unit-root regressor:

$$T^{-1/2} z_{it} \sim O_p(1) \qquad T^{-1/2}(\alpha_{i0} - \hat{\alpha}_i) \sim o_p(1)$$

General case:

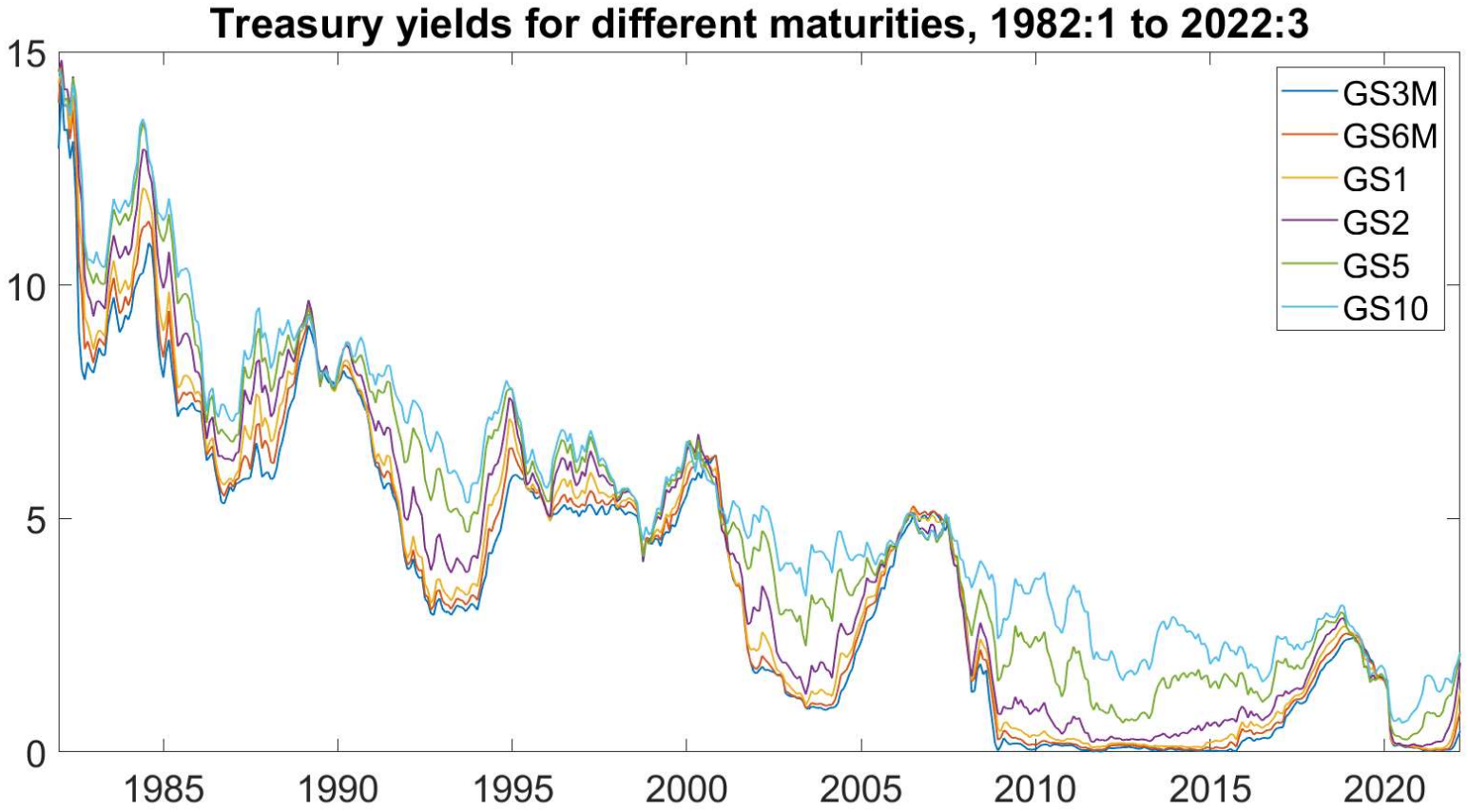
$$\sum_{t=1}^{T} v_{it}^2 = (\boldsymbol{\alpha}_{i0} - \hat{\boldsymbol{\alpha}}_i)' \sum_{t=1}^{T} \mathbf{z}_{it}\mathbf{z}_{it}'(\boldsymbol{\alpha}_{i0} - \hat{\boldsymbol{\alpha}}_i)$$

This is proportional to OLS Wald test of the (correct) null hypothesis that $\boldsymbol{\alpha}_{i0}$ is the true value.

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

$v_{it}^2 \overset{m.s.}{\rightarrow} 0$

Application 1. Term structure of interest rates



Conventional PCA on levels:

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

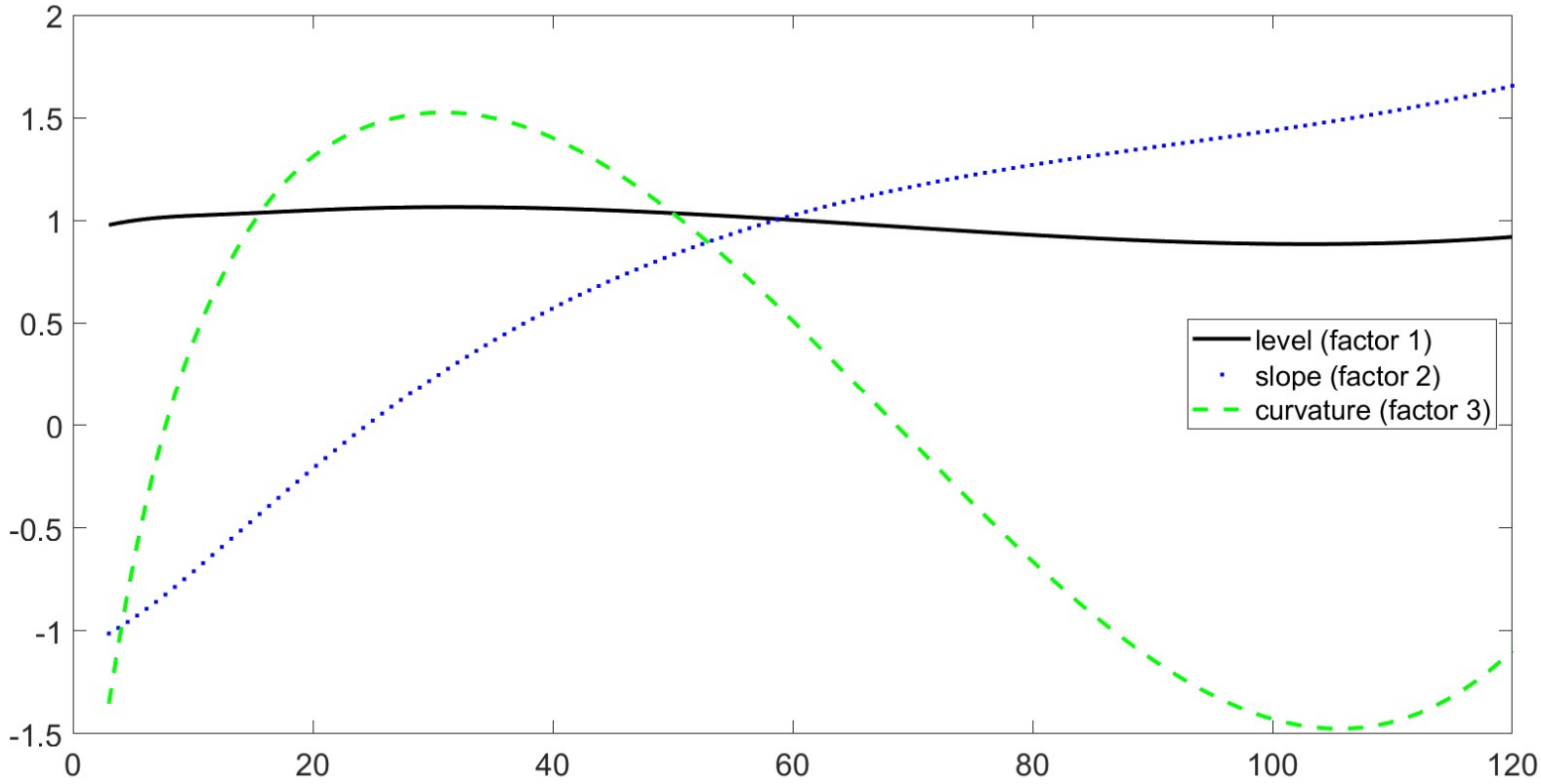
$$\underset{(N\times 1)}{\dot{\mathbf{y}}_t} = \underset{(N\times r)}{\tilde{\mathbf{\Lambda}}}\ \underset{(r\times 1)}{\mathbf{F}_t} + \underset{(N\times 1)}{\tilde{\mathbf{e}}_t}$$

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

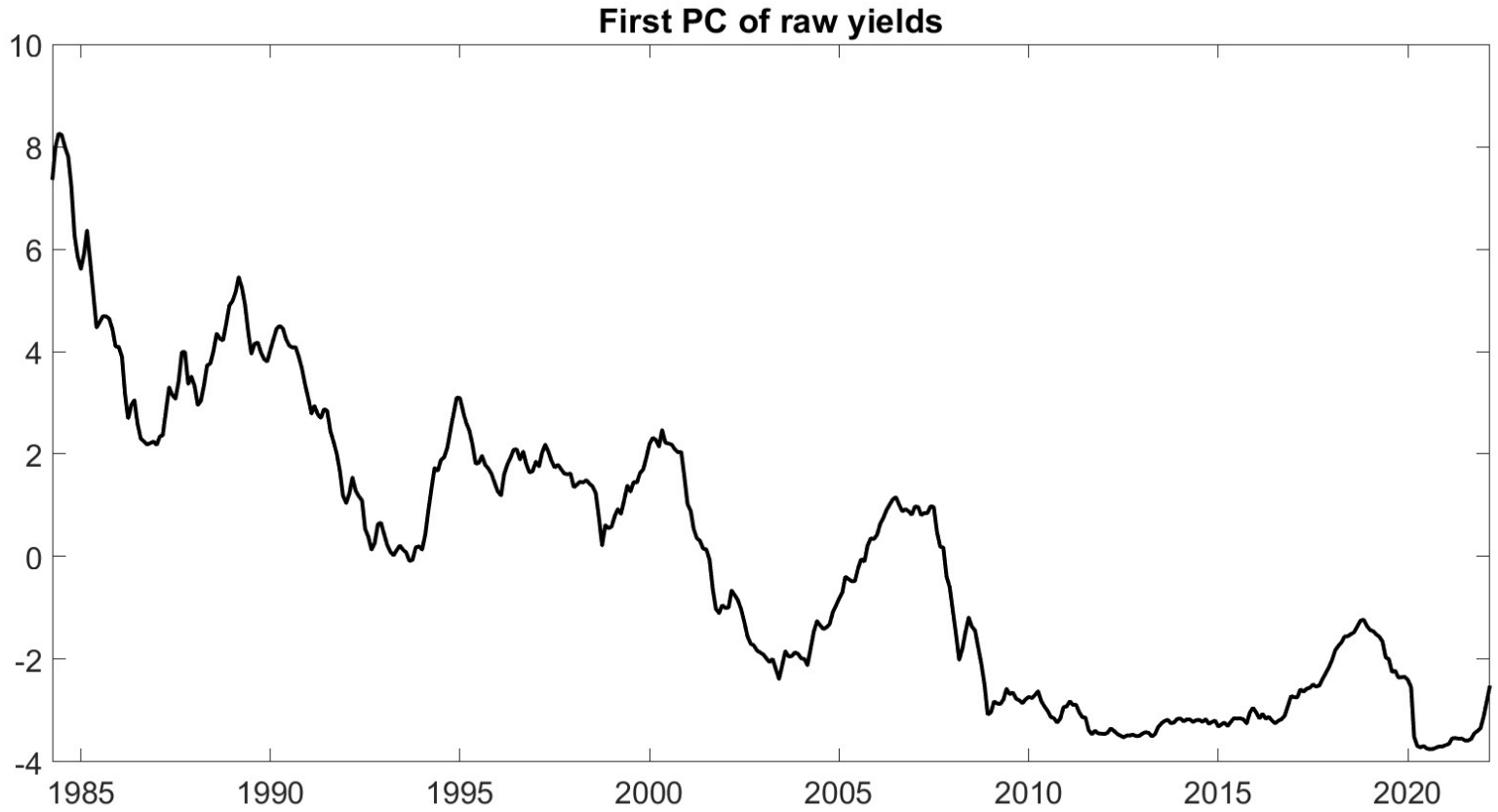
Let $\tilde{\lambda}_j =$ eigenvector of correlation matrix of raw yields associated with jth largest eigenvalue.

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

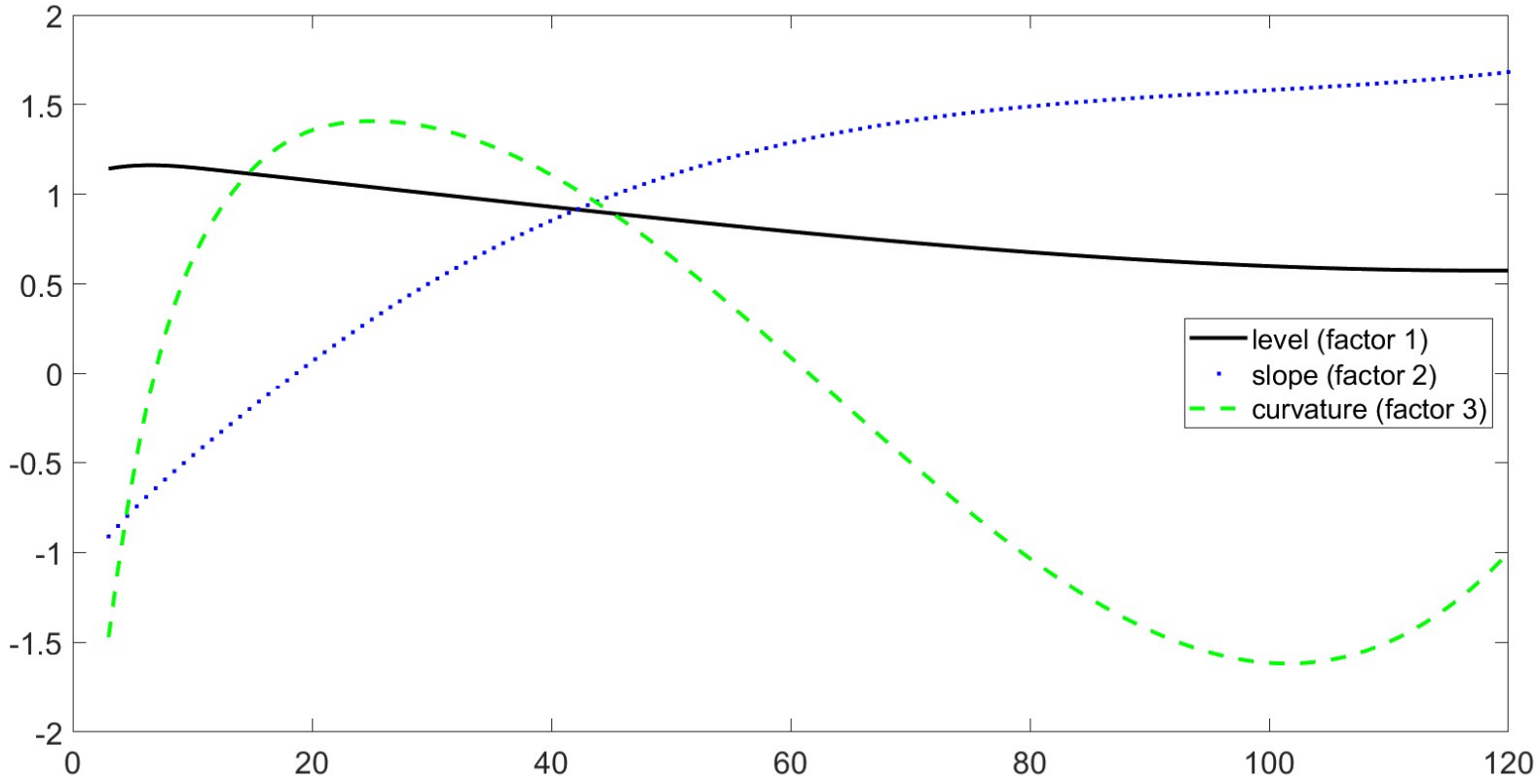


$\hat{c}_{it}$ = residual from OLS regression of y_{it} on $(1, y_{i,t-24}, y_{i,t-25}, \ldots, y_{i,t-35})$.

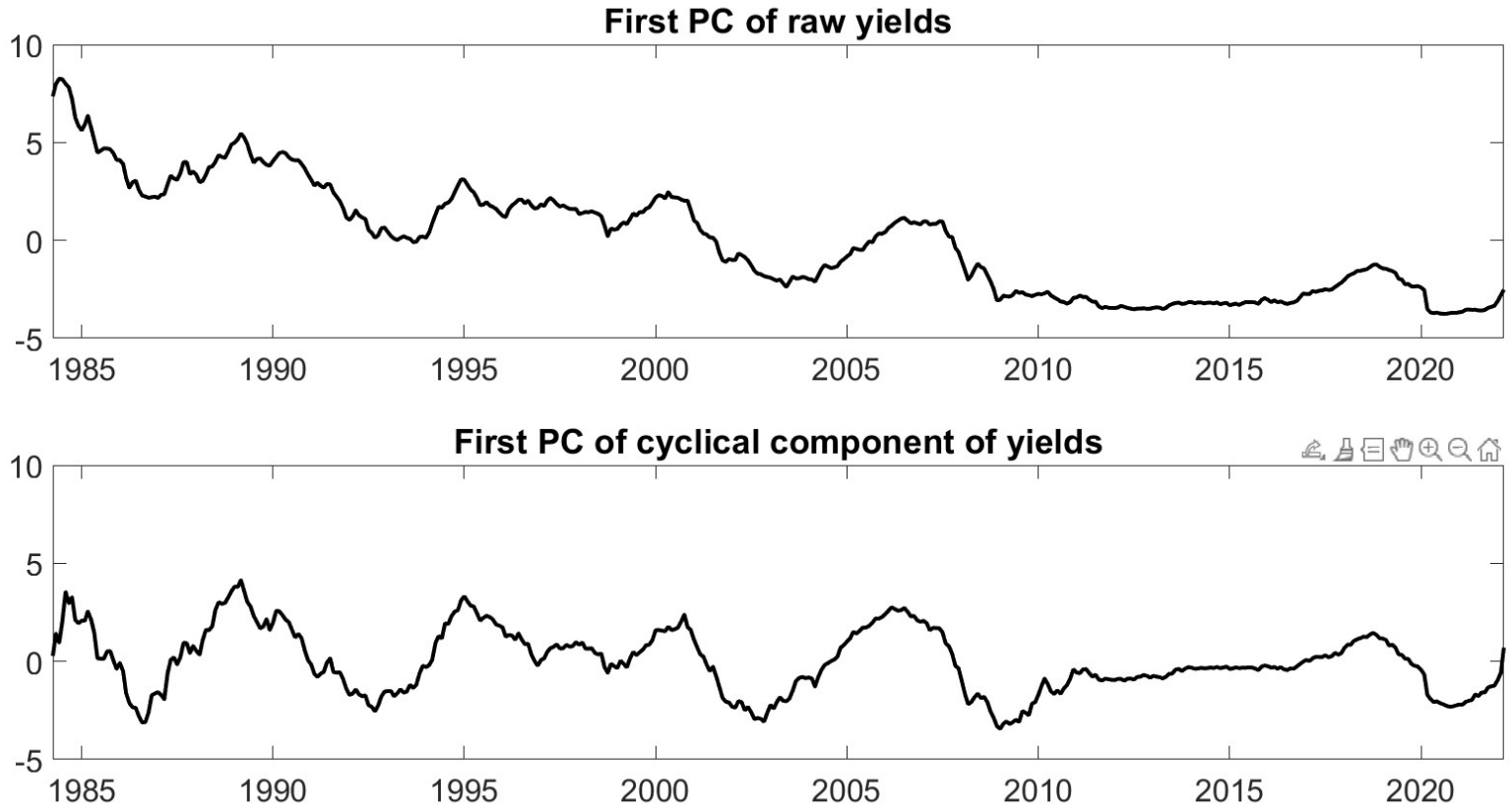
$\hat{\lambda}_j$ = eigenvector of correlation matrix of $\hat{c}_{it}$ associated with jth largest eigenvalue.

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



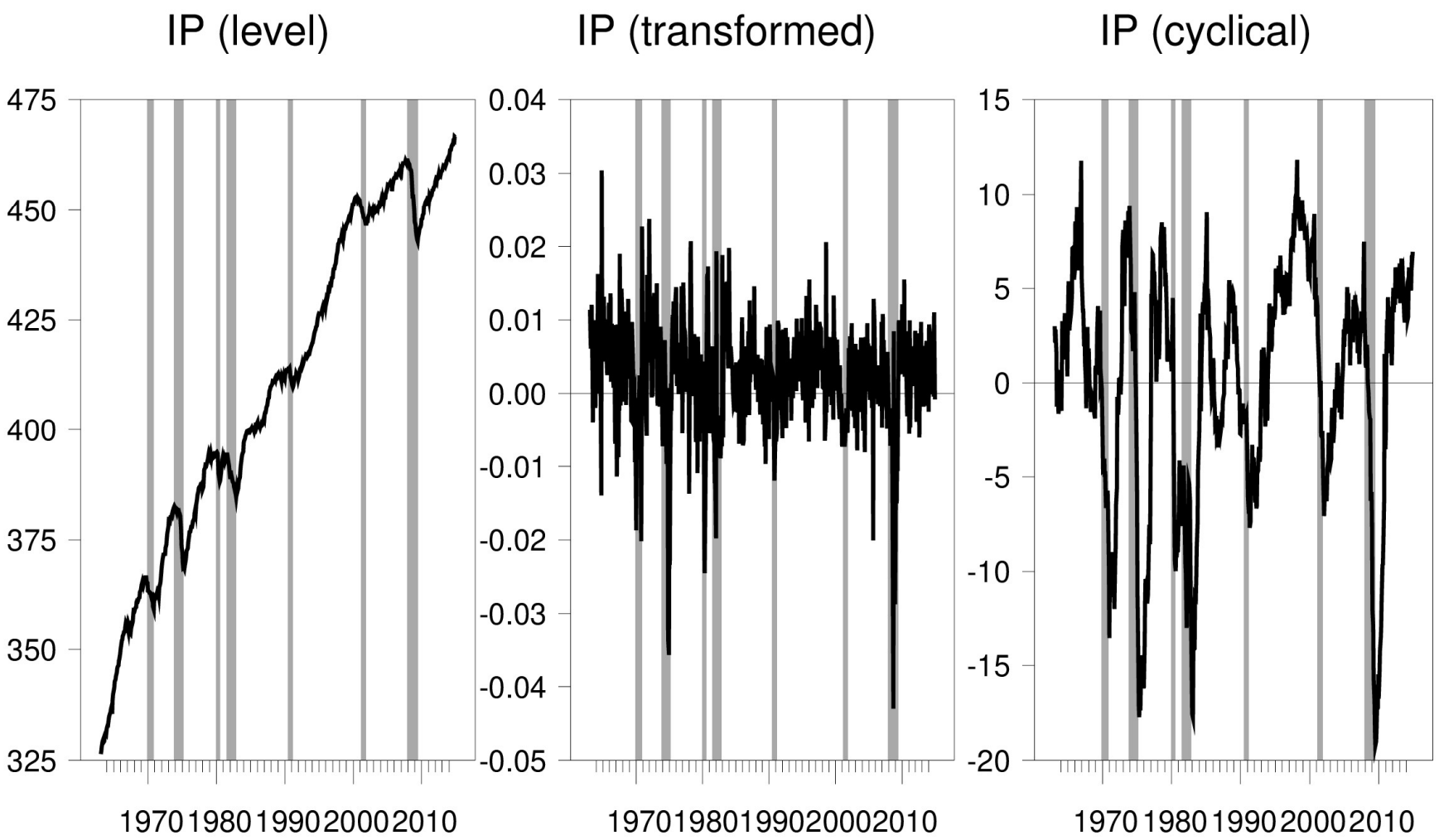
- For this application, PCA on levels works fine because all variables share the same trend component.
- Principal components capture both level and trend.
- If we mix U.S. nominal interest rates with other variables that have different trends, nonstationarity is bigger concern.

Application 2. Large macroeconomic data set

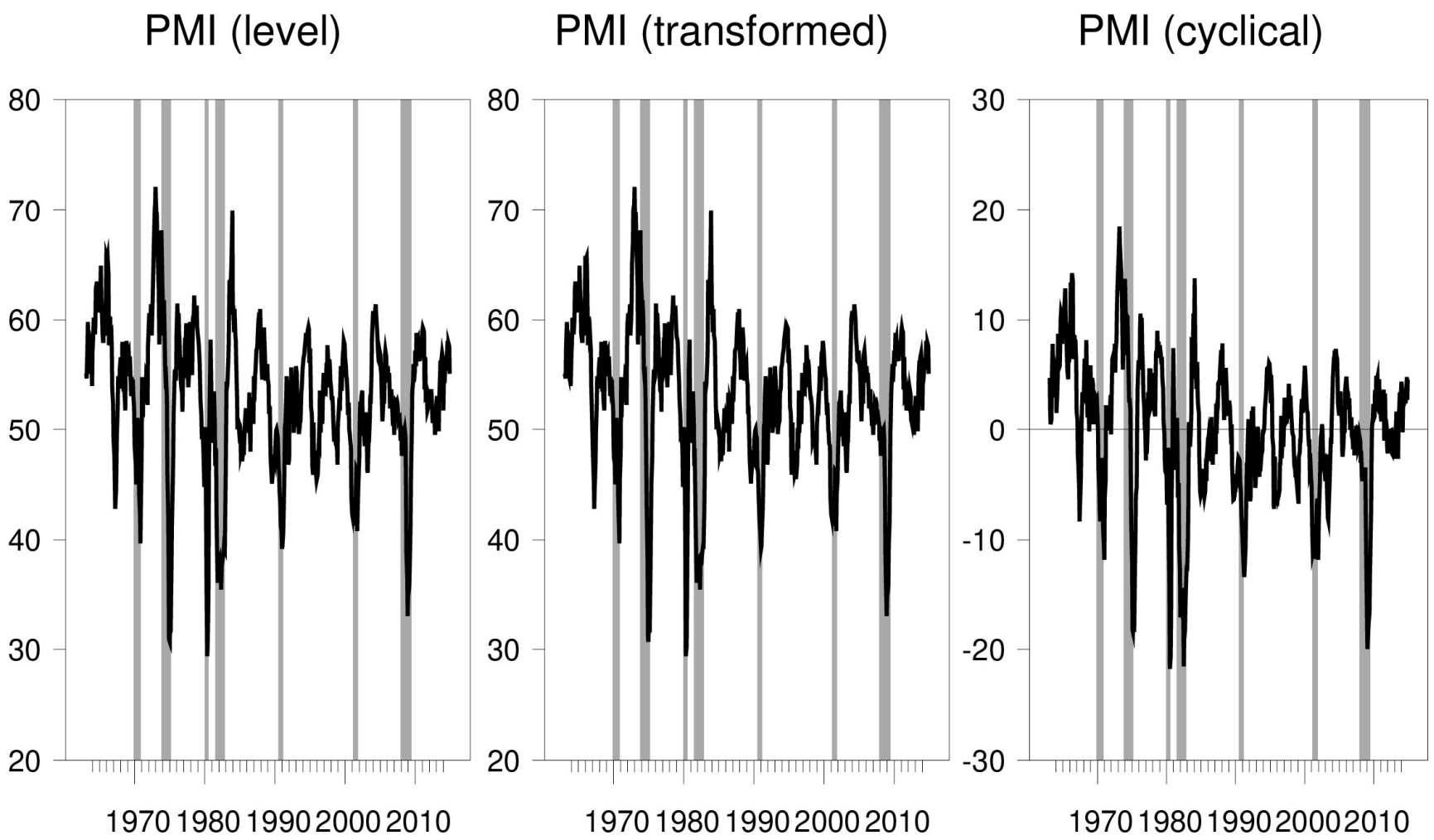
- Stock and Watson (JME 1999) found that first PC of a set of 85 different measures of real economic activity was best way to use big data set to predict inflation.
- This evolved into the Chicago Fed National Activity Index (CFNAI).

- 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

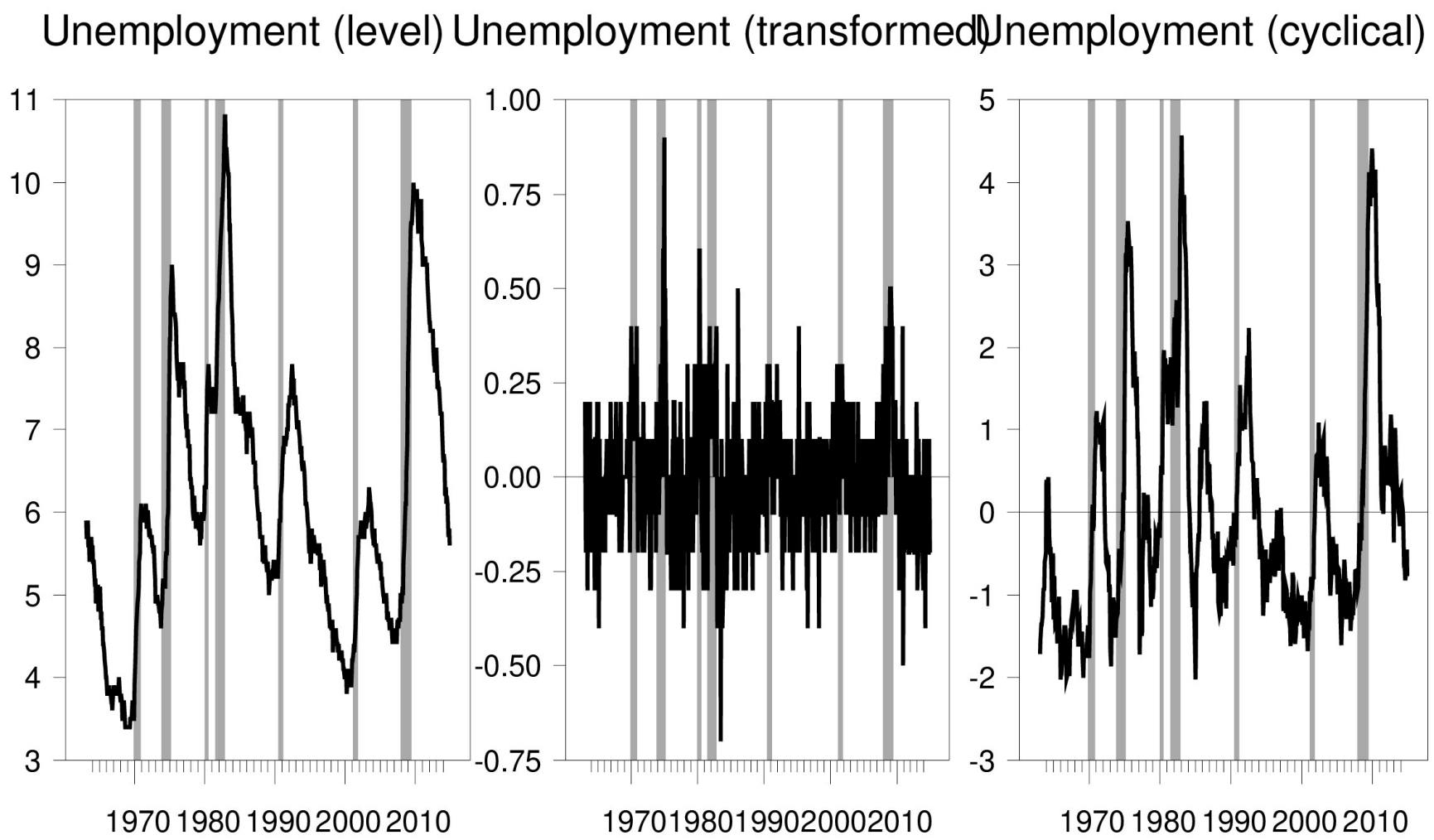
Log of industrial production index



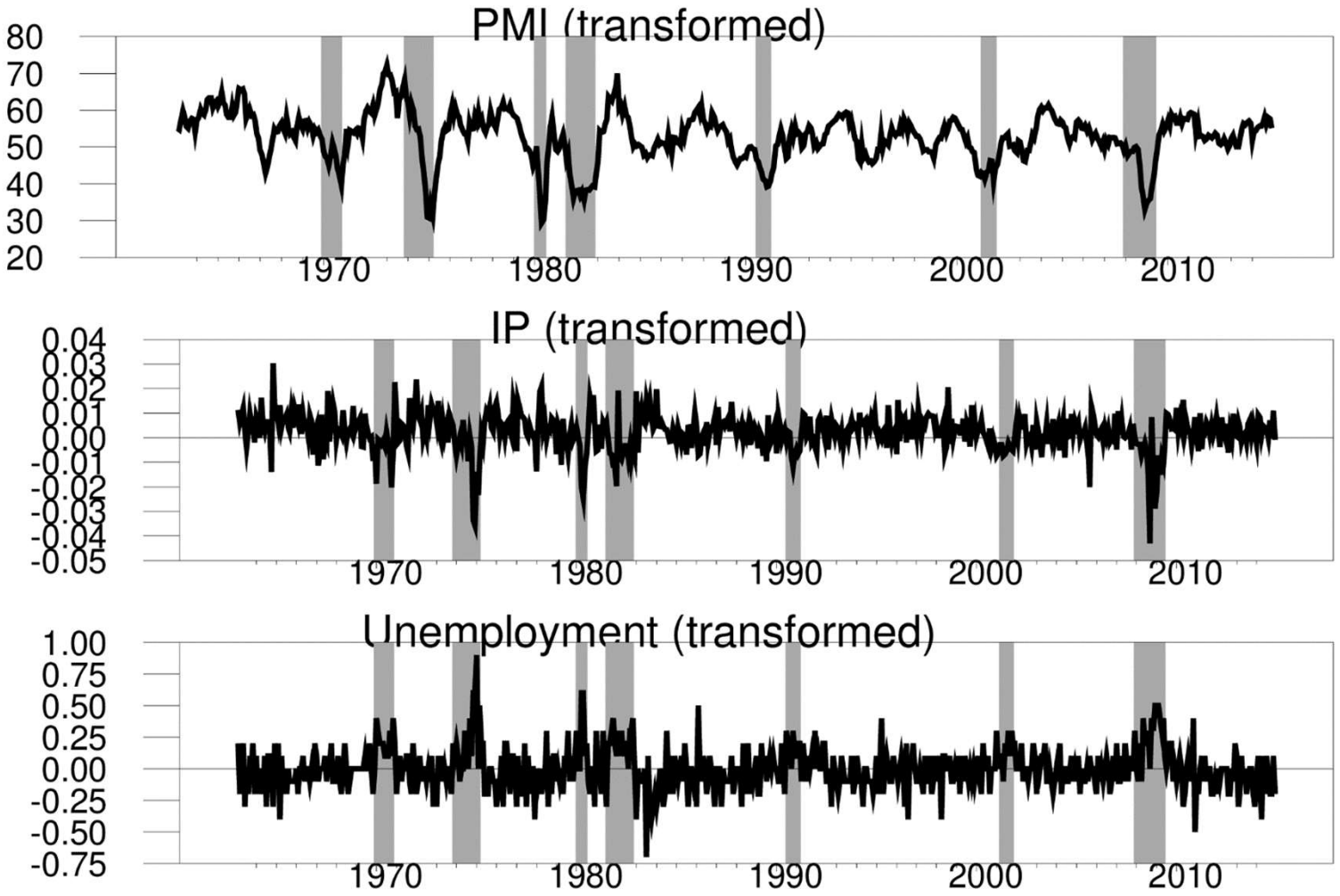
Purchasing managers index



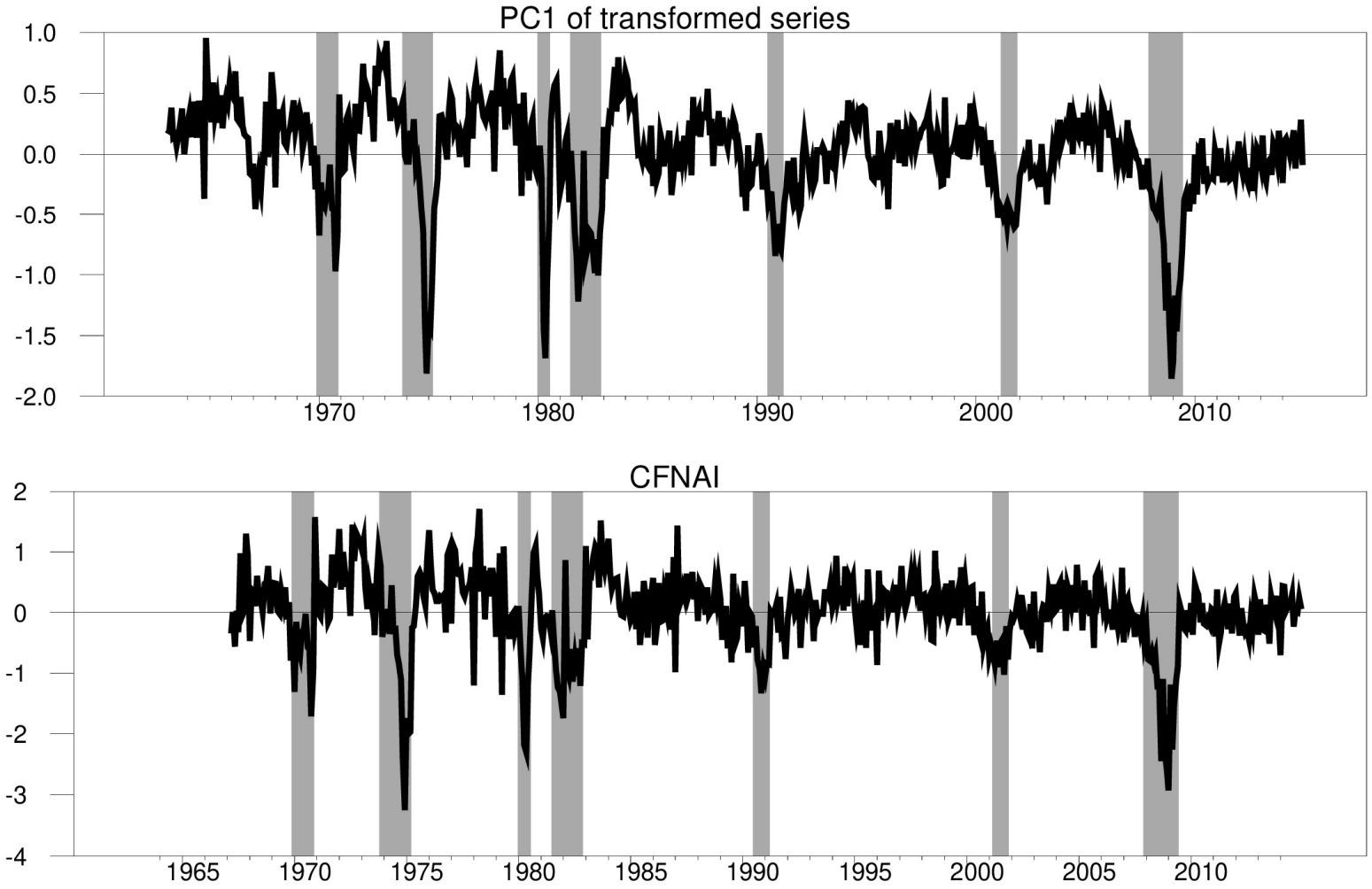
Unemployment rate



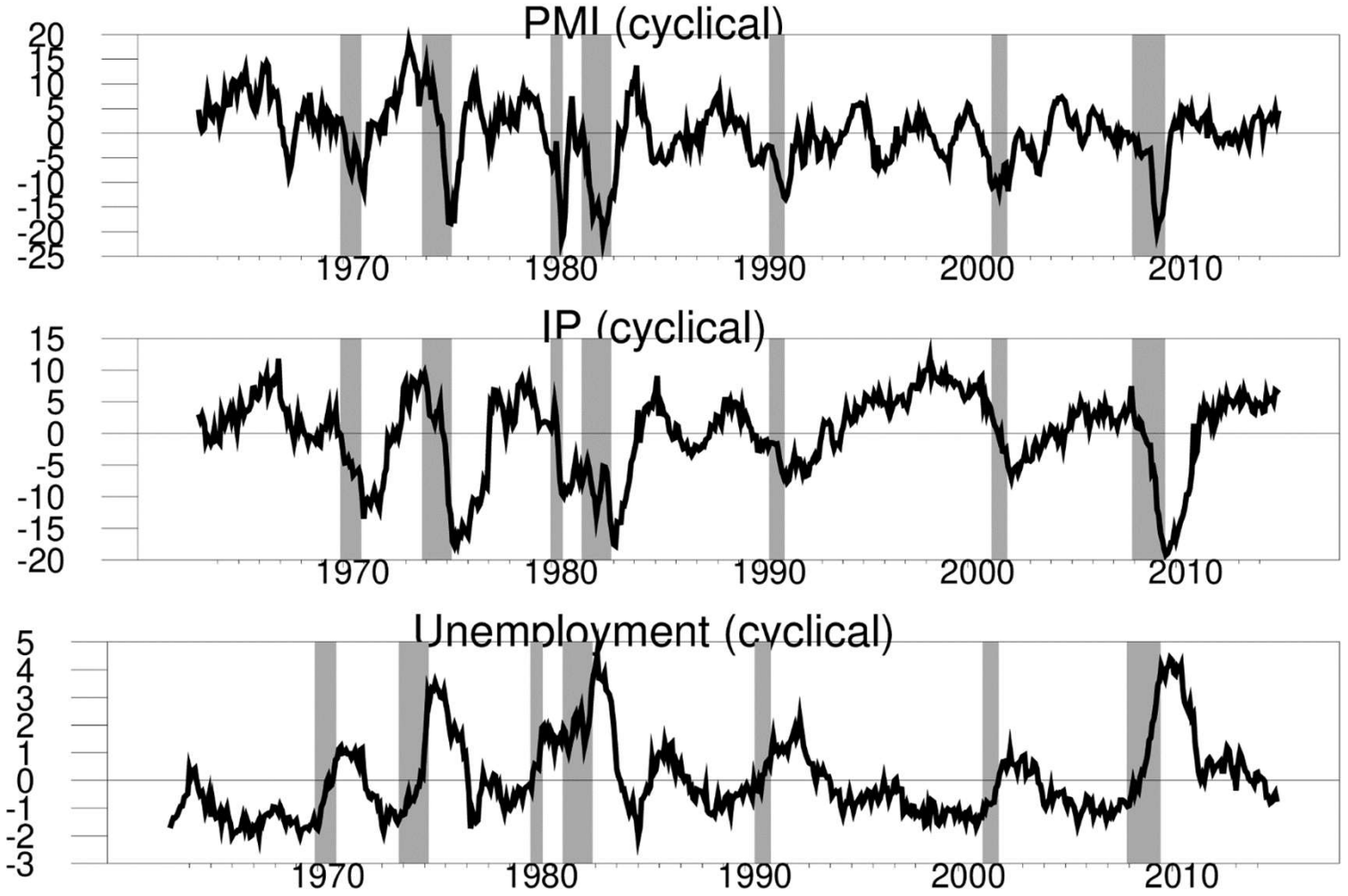
Series as transformed by McCracken and Ng



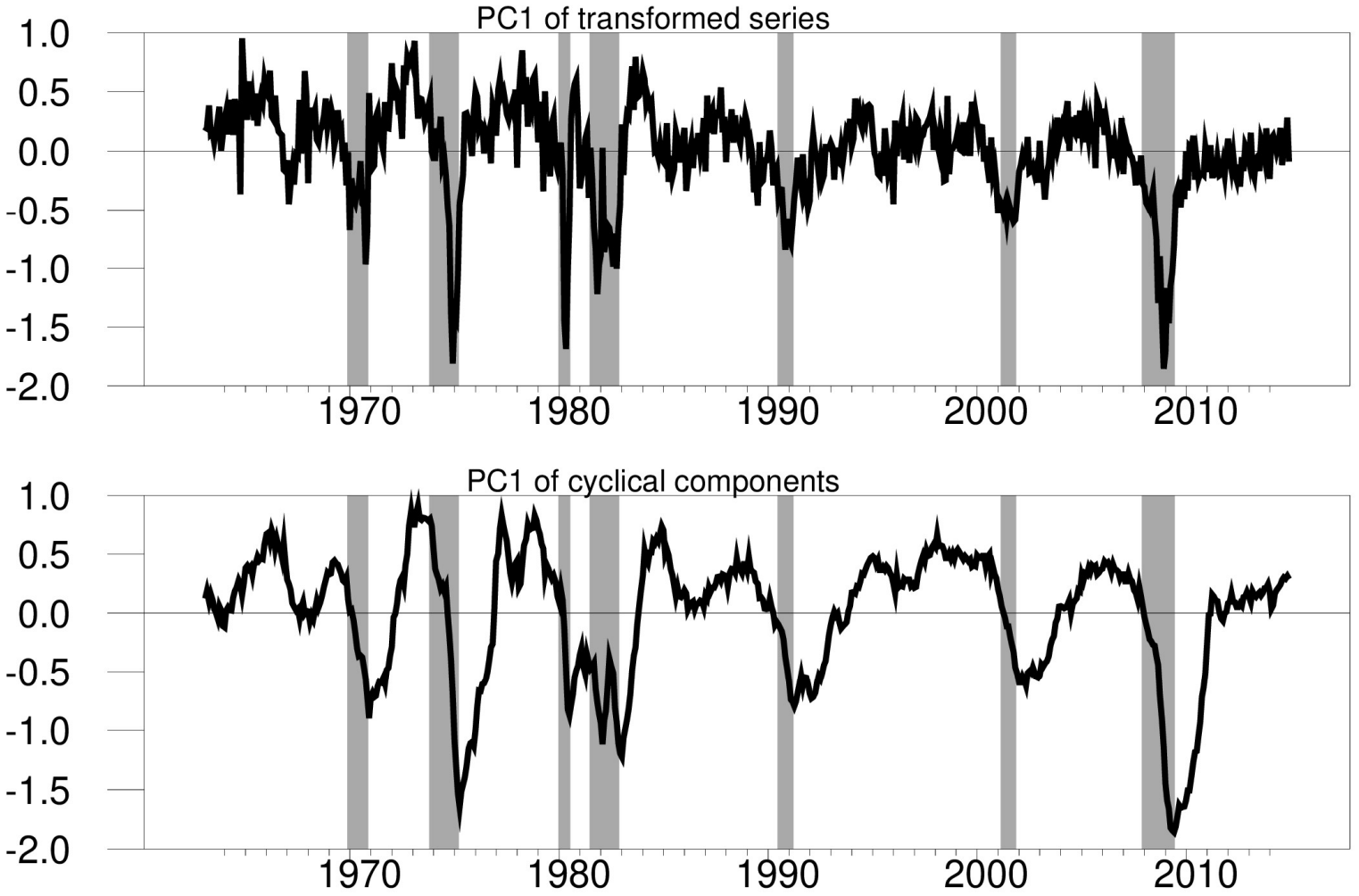
PC1 of transformed data and CFNAI



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

Our approach: if y_{it} is a random walk,

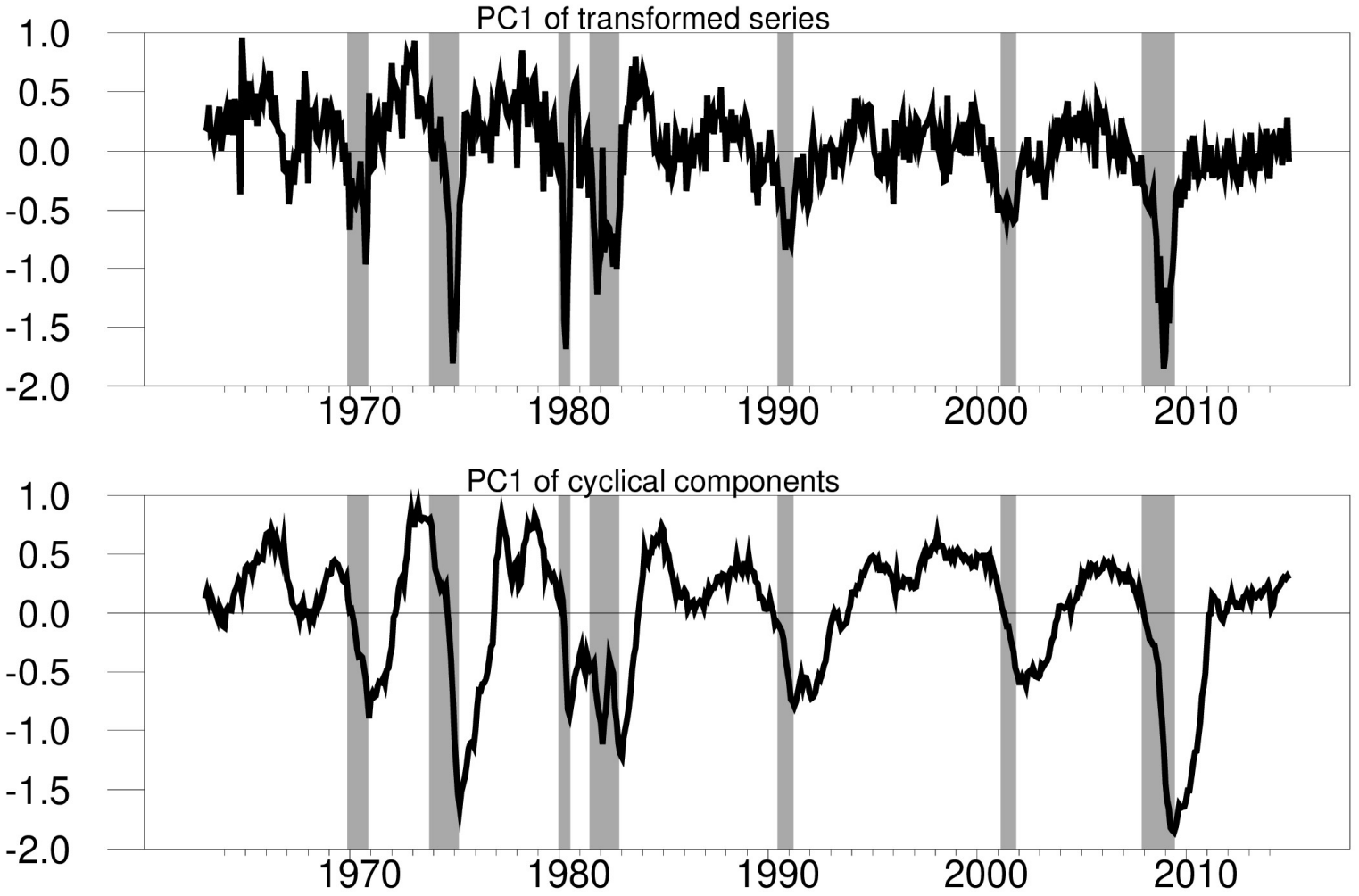
then $c_{it} = \sum_{j=0}^{h} \varepsilon_{i,t-h}$.

Even if ε_{it} is very non-Gaussian, by CLT

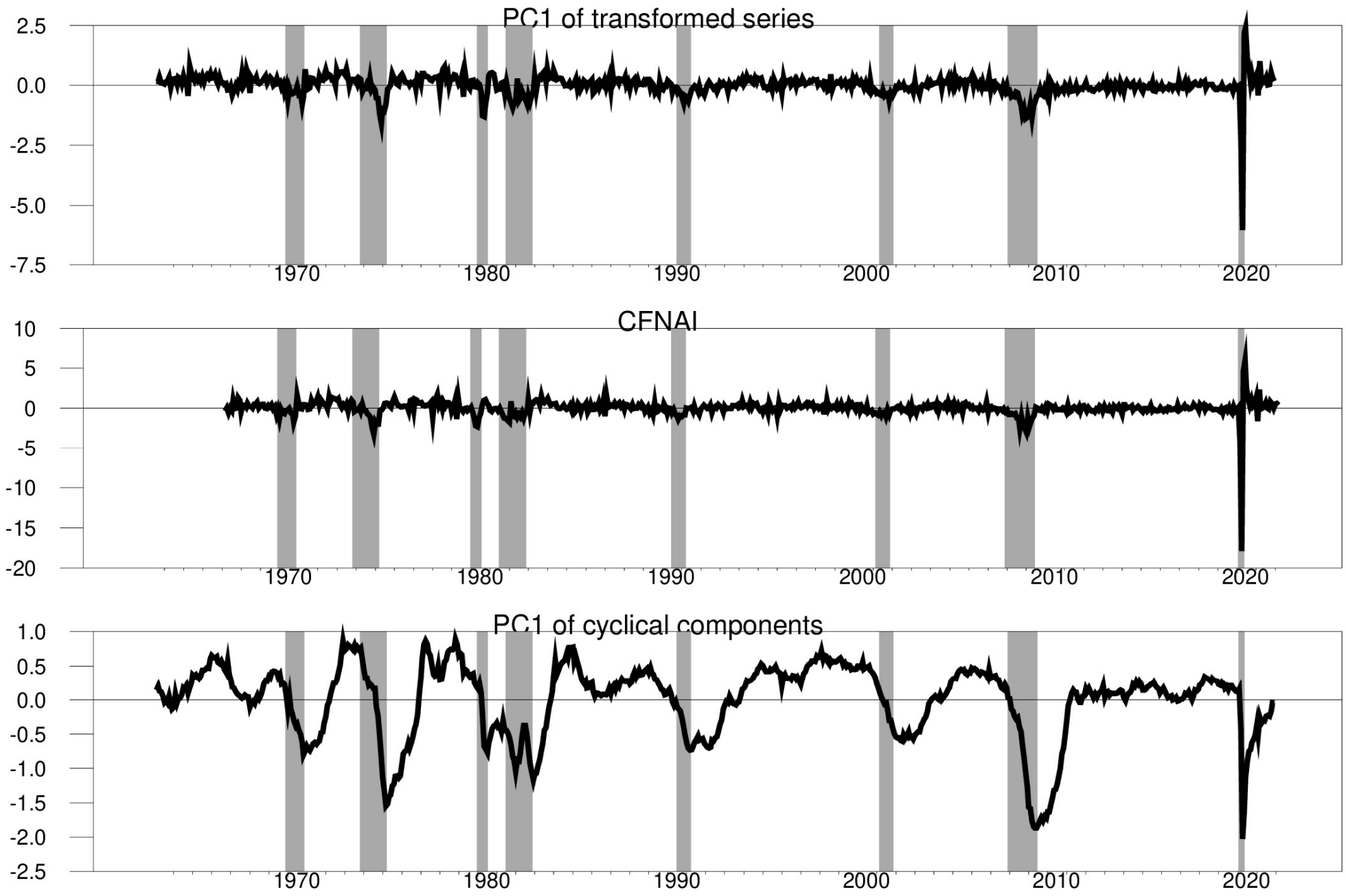
c_{it} is not far from Gaussian.

We find outliers in $\hat{c}_{it}$ for only two variables in this sample (total and nonborrowed reserves during Great Recession).
We recommend not doing anything special about outliers; just use data as is.

Our recommended procedure makes no corrections for outliers



- When dataset is expanded to include recent data, McCracken-Ng algorithm 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

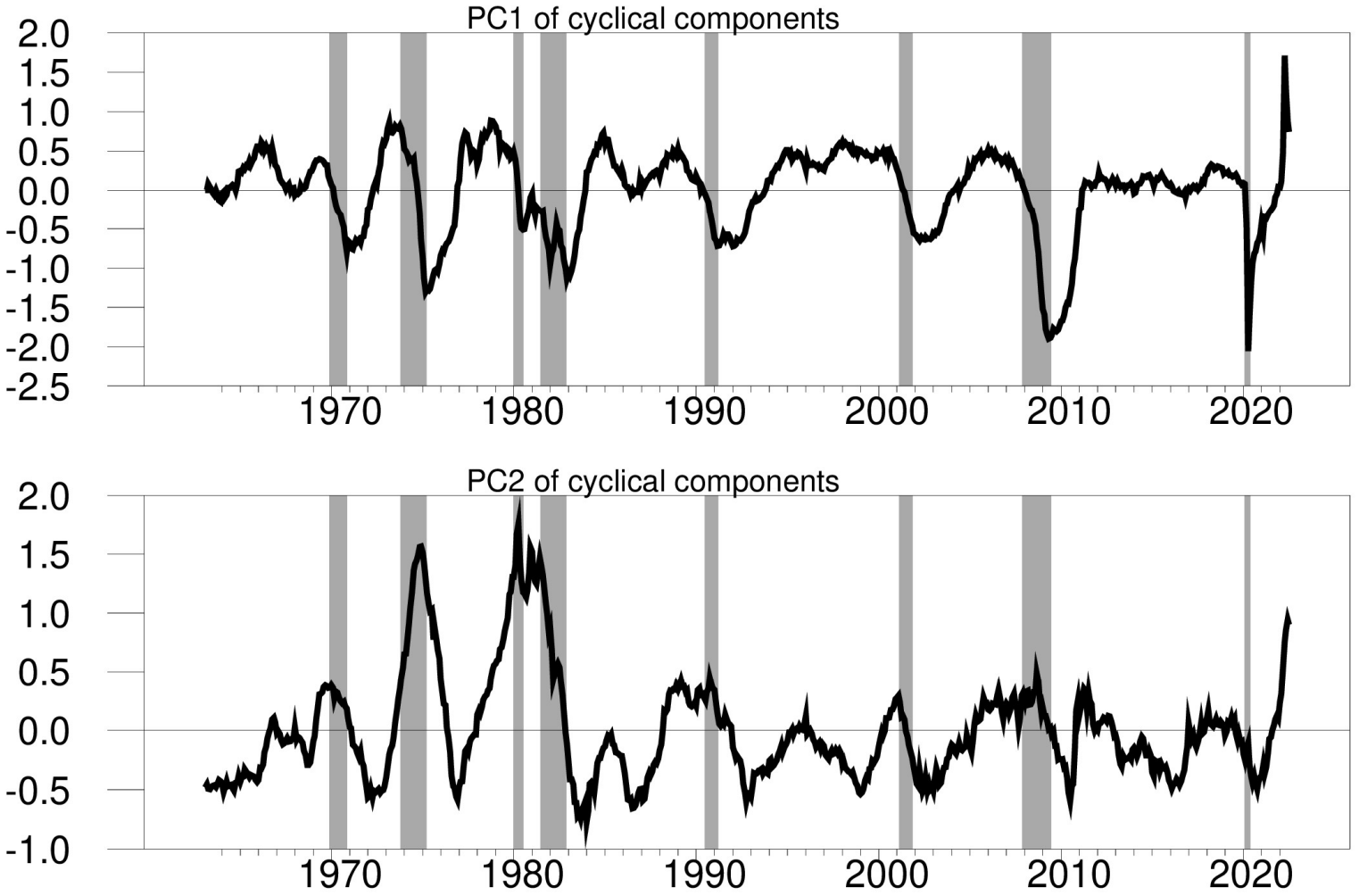

PC1 of transformed series
2.5
0.0
-2.5
-5.0
-7.5
1970
1980
1990
2000
2010
2020
CFNAI
10
5
0
-5
-10
-15
-20
1970
1980
1990
2000
2010
2020
PC1 of cyclical components
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
1970
1980
1990
2000
2010
2020

- 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

Median R^2 for variables in group explained by first and second cyclical PC, 1963-2022

	HX1	HX1&2
Group 1. Output and income	0.72	0.75
Group 2. Labor market	0.55	0.60
Group 3. Housing	0.19	0.37
Group 4. Consumption, orders, and inventories	0.36	0.70
Group 5. Money and credit	0.08	0.19
Group 6. Interest and exchange rates	0.05	0.52
Group 7. Prices	0.00	0.64
Group 8. Stock market	0.19	0.37

First and second cyclical PC, 1963-2022



Evaluate forecasts similarly to Stock and Watson (JME, 1999) and McCracken and Ng (JBES, 2016)

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

$$y_{t+h}^h = \boldsymbol{\pi}^{CF\prime}\mathbf{x}_t^{CF} + u_{t+h}^{m,h}$$

$$\mathbf{x}_t^{CF} = (1, y_t^1, y_{t-1}^1, \ldots, y_{t-5}^1, \hat{f}_t^{CF}, \hat{f}_{t-1}^{CF}, \ldots, \hat{f}_{t-5}^{CF})'$$

Estimate through $t = T_1$, forecast $y_{T_1+1+h}^h$.

Expanding windows, different evaluation periods.

Table 1: Mean squared forecast errors for different models

		CPI					IP				
sample	horizon	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.