WHY YOU SHOULD NEVER USE THE HODRICK-PRESCOTT FILTER

James D. Hamilton*

Abstract—Here's why. (a) The Hodrick-Prescott (HP) filter introduces spurious dynamic relations that have no basis in the underlying data-generating process. (b) Filtered values at the end of the sample are very different from those in the middle and are also characterized by spurious dynamics. (c) A statistical formalization of the problem typically produces values for the smoothing parameter vastly at odds with common practice. (d) There is a better alternative. A regression of the variable at date t on the four most recent values as of date $t-h$ achieves all the objectives sought by users of the HP filter with none of its drawbacks.

I. Introduction

OFTEN economic researchers have a theory that is specified in terms of a stationary environment and wish to relate the theory to observed nonstationary data without modeling the nonstationarity. Hodrick and Prescott (1981, 1997) proposed a very popular method for doing this, commonly interpreted as decomposing an observed variable into trend and cycle. Although drawbacks to their approach have been known for some time, the method continues to be widely adopted in academic research, policy studies, and analysis by private sector economists. For this reason, it seems useful to collect and expand on those earlier concerns here and note that there is a better way to solve this problem.

II. Characterizations of the Hodrick-Prescott Filter

Given T observations on a variable y_t, Hodrick and Prescott (1981, 1997) proposed interpreting the trend component g_t as a very smooth series that does not differ too much from the observed y_t.[1] It is calculated as

$$\min_{\{g_t\}_{t=-1}^{T}} \left\{ \sum_{t=1}^{T} (y_t - g_t)^2 + \lambda \sum_{t=1}^{T} [(g_t - g_{t-1}) - (g_{t-1} - g_{t-2})]^2 \right\}. \tag{1}$$

When the smoothness penalty $\lambda \to 0$, g_t would just be the series y_t itself, whereas when $\lambda \to \infty$ the procedure amounts to a regression on a linear time trend (i.e., produces a series whose second difference is exactly 0). The common practice is to use a value of $\lambda = 1{,}600$ for quarterly time series.

A closed-form expression for the resulting series can be written in vector notation by defining $\tilde{T} = T + 2$, $\underset{(T\times 1)}{y} = (y_T, y_{T-1}, \ldots, y_1)'$, $\underset{(\tilde{T}\times 1)}{g} = (g_T, g_{T-1}, \ldots, g_{-1})'$ and

$$\underset{(T\times\tilde{T})}{H} = \begin{bmatrix} \underset{(T\times T)}{I_T} & \underset{(T\times 2)}{0} \end{bmatrix}$$

$$\underset{(T\times\tilde{T})}{Q} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}.$$

The solution to equation (1) is then given by[2]

$$g^* = (H'H + \lambda Q'Q)^{-1} H'y = A^* y. \tag{2}$$

The inferred trend g_t^* for any date t is thus a linear function of the full set of observations on y for all dates.

As Hodrick and Prescott (1981) and King and Rebelo (1993) noted, the identical inference can alternatively be motivated from particular assumptions about the time-series behavior of the trend and cycle components. Suppose our goal was to choose a value for a $(T \times 1)$ vector a_t such that the estimate $\tilde{g}_t = a_t' y$ has minimum expected squared difference from the true trend:

$$\min_{a_t} E(g_t - a_t' y)^2. \tag{3}$$

The solution to this problem is the population analog to a sample regression coefficient and is a function of the variance of y and its covariance with g:

$$\tilde{g} = E(gy') \left[E(yy') \right]^{-1} y = \tilde{A} y. \tag{4}$$

As an example of a particular set of assumptions we might make about these covariances, let c_t denote the cyclical component and v_t the second difference of the trend component:

$$y_t = g_t + c_t, \tag{5}$$

$$g_t = 2g_{t-1} - g_{t-2} + v_t. \tag{6}$$

Suppose that we believed that v_t and c_t are uncorrelated white noise processes that are also uncorrelated with (g_0, g_{-1}), and let C_0 denote the (2×2) variance of (g_0, g_{-1}). These assumptions imply a particular value for $\tilde{A}$ in equation (4). As we let the variance of (g_0, g_{-1}) become arbitrarily large (represented as $C_0^{-1} \to 0$), then in every sample, the inference, equation (4), would be numerically identical to expression (2).

Received for publication January 3, 2017. Revision accepted for publication June 9, 2017. Editor: Yuriy Gorodnichenko.

* University of California, San Diego.

I thank Daniel Leff for outstanding research assistance on this project and Frank Diebold, Robert King, James Morley, and anonymous referees for helpful comments on an earlier draft of this paper.

[1] Phillips and Jin (2015) reviewed the rich prior history of generalizations of this approach.

[2] The appendix provides a derivation of equations (2) and (4). Cornea-Madeira (2017) provided further details on A^* and a convenient algorithm for calculating it.

The Review of Economics and Statistics, December 2018, 100(5): 831–843

doi:10.1162/rest_a_00706

Proposition 1. *For* $\lambda = \sigma_c^2/\sigma_v^2$ *and any fixed* T, *under conditions* (5) *and* (6) *with* c_t *and* v_t *white noise uncorrelated with each other and uncorrelated with* (g_0, g_{-1}), *the matrix* $\tilde{A}$ *in equation* (4) *converges to the matrix* A^* *in equation* (2) *as* $C_0^{-1} \to 0$.

The proposition establishes that if researcher 1 sought to identify a trend by solving the minimization problem, equation (1), while researcher 2 found the optimal linear estimate of a trend process that was assumed to be characterized by the particular assumption that v_t and c_t were both white noise, the two researchers would arrive at the numerically identical series for trend and cycle provided the ratio of σ_c^2 to σ_v^2 assumed by researcher 2 was identical to the value of λ used by researcher 1.

The Kalman smoother is an iterative algorithm for calculating the population linear projection, equation (4), for models where the variance and covariance can be characterized by some recursive structure.[3] In this case, equation (5) is the observation equation and equation (6) is the state equation. Thus, as Hodrick and Prescott noted, applying the Kalman smoother to the above state-space model starting from a very large initial variance for $(g_0, g_{-1})'$ offers a convenient algorithm for calculating the HP filter and is in fact a way that the HP filter is often calculated in practice. Nevertheless, this observation should also be a bit troubling for users of the HP filter, in that they never defend the claim that the particular structure assumed in proposition 1 is an accurate representation of the true data-generating process. Indeed, if a researcher did know for certain that these equations were the true data-generating process and, further, knew for certain the value of the population parameter $\lambda = \sigma_c^2/\sigma_v^2$, he would probably be unhappy with using equation (2) to separate cycle from trend! The reason is that if this state-space structure was the true DGP, the resulting estimate of the cyclical component $c_t = y_t - \tilde{g}_t$ would be white noise—it would be random and exhibit no discernible patterns. By contrast, users of the HP filter hope to see suggestive patterns in plots of the series that are supposed to be interpreted as the cyclical component of y_t.

Premultiplying equation (2) by $H'H + \lambda Q'Q$ gives a system of equations whose tth element is

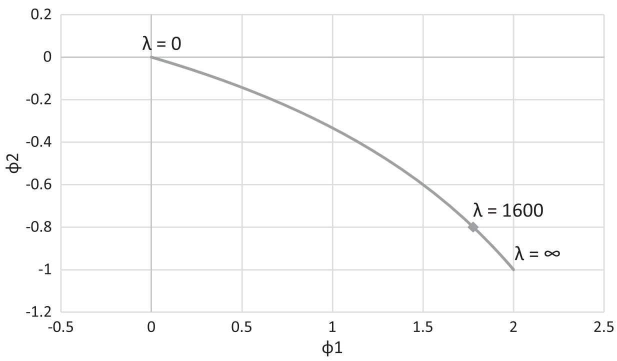
$$[1 + \lambda(1 - L^{-1})^2(1 - L)^2]g_t^* = y_t \quad \text{for } t = 1, 2, \ldots, T - 2 \tag{7}$$

for L the lag operator ($L^k x_t = x_{t-k}$, $L^{-k} x_t = x_{t+k}$). In other words, $F(L)g_t^* = y_t$ for

$$F(L) = 1 + \lambda(1 - L^{-1})^2(1 - L)^2. \tag{8}$$

The following proposition establishes some properties of this filter:[4]

FIGURE 1.—VALUES FOR ϕ_1 AND ϕ_2 IMPLIED BY DIFFERENT VALUES OF λ



Proposition 2. *For any* $\lambda : 0 < \lambda < \infty$, *the inverse of the operator* (8) *can be written as*

$$[F(L)]^{-1} = C\left[\frac{1 - (\phi_1^2/4)L}{1 - \phi_1 L - \phi_2 L^2} + \frac{1 - (\phi_1^2/4)L^{-1}}{1 - \phi_1 L^{-1} - \phi_2 L^{-2}} - 1\right], \tag{9}$$

where

$$\frac{1}{1 - \phi_1 z - \phi_2 z^2} = \sum_{j=0}^{\infty} R^j[\cos(mj) + \cot(m)\sin(mj)]z^j, \tag{10}$$

$$\frac{1}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} = \sum_{j=0}^{\infty} R^j[\cos(mj) + \cot(m)\sin(mj)]z^{-j},$$

$$\phi_1(1 - \phi_2) = -4\phi_2, \tag{11}$$

$$(1 - \phi_1 - \phi_2)^2 = -\phi_2/\lambda, \tag{12}$$

$$C = \frac{-\phi_2}{\lambda(1 - \phi_1^2 - \phi_2^2 + \phi_1^3/2)}, \tag{13}$$

$$R = \sqrt{-\phi_2}, \quad \cos(m) = \phi_1/(2R). \tag{14}$$

Roots of $(1 - \phi_1 z - \phi_2 z^2) = 0$ *are complex and outside the unit circle,* ϕ_1 *is a real number between* 0 *and* 2, ϕ_2 *a real number between* -1 *and* 0, *and* R *a real number between* 0 *and* 1.

Figure 1 plots the values of ϕ_1 and ϕ_2 generated by different values of λ. For $\lambda = 1{,}600$, $\phi_1 = 1.777$, and $\phi_2 = -0.7994$. These imply $R = 0.8941$, so that the absolute value of the weights decays with a half-life of about six quarters while $R^{60} = 0.0012$.[5]

[3] See, for example, Hamilton (1994, equation 13.6.3).

[4] Related results have been developed by Singleton (1988), King and Rebelo (1989, 1993), Cogley and Nason (1995), and McElroy (2008). Unlike these papers, here I provide simple direct expressions for the values of ϕ_1 and ϕ_2, and my analytical expressions of the HP filter entirely in terms of real parameters in equations (9) and (10) appear to be new.

[5] The other parameters for this case are $C = 0.056075$, $m = 0.111687$, and $\cot(m) = 8.9164$.

Expression (7) means that for t more than fifteen years from the start or end of a sample of quarterly data, the cyclical component $c_t = y_t - g_t^*$ is well approximated by

$$c_t = \lambda(1-L^{-1})^2(1-L)^2 g_t^* = \frac{\lambda(1-L^{-1})^2(1-L)^2}{F(L)} y_t = \frac{\lambda(1-L)^4}{F(L)} y_{t+2}. \tag{15}$$

As King and Rebelo (1993) noted, obtaining the cyclical component for these observations thus amounts to taking fourth differences of the original y_{t+2} and applying the operator $[F(L)]^{-1}$ to the result, so that the HP cycle might be expected to produce a stationary series as long as fourth-differences of the original series are stationary. However, De Jong and Sakarya (2016) noted there could still be significant nonstationarity coming from observations near the start or end of the sample, and Phillips and Jin (2015) concluded that for commonly encountered sample sizes, the HP filter may not successfully remove the trend even if the true series is only $I(1)$.

III. Drawbacks to the HP Filter

A. *Appropriateness for Typical Economic Time Series*

The presumption by users of the HP filter is that it offers a reasonable approach to detrending for a range of commonly encountered economic time series. The leading example of a time-series process for which we would want to be particularly convinced of the procedure's appropriateness would be a random walk. Simple economic theory suggests that variables such as stock prices (Fama, 1965), futures prices (Samuelson, 1965), long-term interest rates (Sargent, 1976; Pesando, 1979), oil prices (Hamilton, 2009), consumption spending (Hall, 1978), inflation, tax rates, and money supply growth rates (Mankiw, 1987) should all follow martingales or near martingales. To be sure, hundreds of studies have claimed to find evidence of statistically detectable departures from pure martingale behavior in all these series. Even so, there is indisputable evidence that a random walk is often extremely hard to beat in out-of-sample forecasting comparisons, as has been found, for example, by Meese and Rogoff (1983) and Cheung, Chinn, and Pascual (2005) for exchange rates, Flood and Rose (2010) for stock prices, Atkeson and Ohanian (2001) for inflation, or Balcilar et al. (2015) for GDP, among many others. Certainly if we are not comfortable with the consequences of applying the HP filter to a random walk, then we should not be using it as an all-purpose approach to economic time series.

For $y_t = y_{t-1} + \varepsilon_t$, where ε_t is white noise and $(1-L)y_t = \varepsilon_t$, Cogley and Nason (1995) noted that expression (15) means that when the HP filter is applied to a random walk, the cyclical component for observations near the middle of the sample will approximately be characterized by[6]

$$c_t = \frac{\lambda(1-L)^3}{F(L)} \varepsilon_{t+2}.$$

For $\lambda = 1{,}600$ this is

$$c_t = 89.72\left\{ -q_{0,t+2} + \sum_{j=0}^{\infty}(0.8941)^j[\cos(0.1117j) + 8.916\sin(0.1117j)](q_{1,t+2-j} + q_{2,t+2+j})\right\},$$

with $q_{0t} = \varepsilon_t - 3\varepsilon_{t-1} + 3\varepsilon_{t-2} - \varepsilon_{t-3}, q_{1t} = \varepsilon_t - 3.79\varepsilon_{t-1} + 5.37\varepsilon_{t-2} - 3.37\varepsilon_{t-3} + 0.79\varepsilon_{t-4}),$[7] and $q_{2t} = -0.79\varepsilon_{t+1} + 3.37\varepsilon_t - 5.37\varepsilon_{t-1} + 3.79\varepsilon_{t-2} - \varepsilon_{t-3}$. The underlying innovations ε_t are completely random and exhibit no patterns, whereas the series c_t is both highly predictable (as a result of the dependence on lags of ε_{t-j}) and will in turn predict the future (as a result of dependence on future values of ε_{t+j}). Since the coefficients that make up $[F(L)]^{-1}$ are determined solely by the value of λ, these patterns in the cyclical component are entirely a feature of having applied the HP filter to the data rather than reflecting any true dynamics of the data-generating process itself.

For example, consider the behavior of stock prices and real consumption spending.[8] The top panels of figure 2 show the autocorrelation functions for first-differences of these series, confirming that there is little ability to predict either from its own past values, as we might have expected from the literature cited at the start of this section. The lower panels show cross-correlations. Consumption has no predictive power for stocks, though stock prices may have a modest ability to anticipate changes in aggregate consumption.

Figure 3 shows the analogous results if we tried to remove the trend by HP filtering rather than first-differencing. The HP cyclical components of stock prices and consumption are both extremely predictable from their own lagged values as well as each other. The rich dynamics in these series are purely an artifact of the filter itself and tell us nothing about the underlying data-generating process. Filtering takes us from the very clean understanding of the true properties of these series that we can easily see in figure 2 to the artificial set of relations that appear in figure 3. The values plotted in figure 3 summarize the filter, not the data.

B. *Properties of the One-Sided HP Filter*

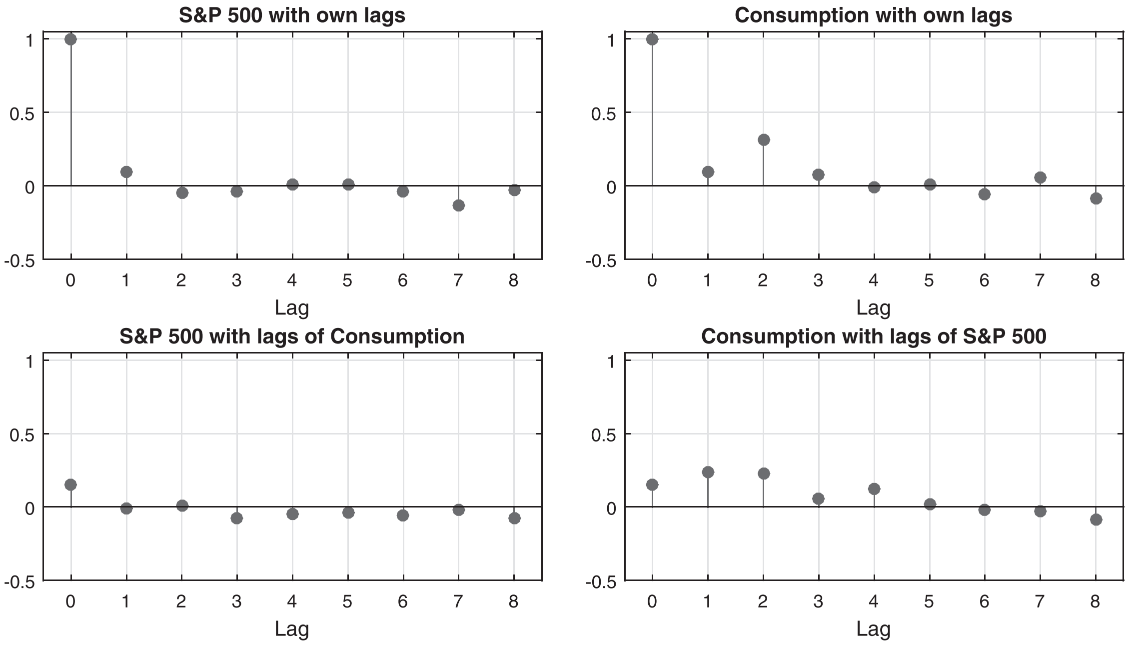
The HP trend and cycle have an artificial ability to "predict" the future because they are by construction a function of future realizations. One way we might try to get around this would be to restrict the minimization problem in equation (3), forcing a_t to load only on values $(y_t, y_{t-1}, \ldots, y_1)'$ that have

[6] Harvey and Jaeger (1993) also have a related discussion.

[7] The term q_{1t} is the expansion of $(1-L)^3[1-(\phi_1^2/4)L]\varepsilon_t$.

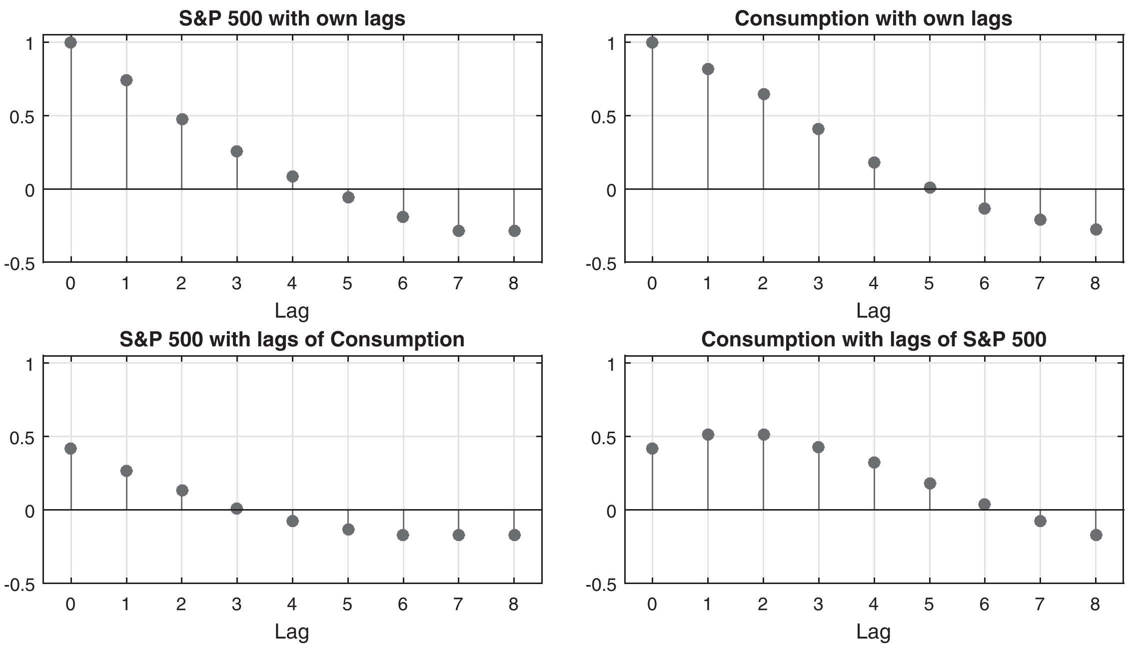
[8] Stock prices were measured as 100 times the natural log of the end-of-quarter value for the S&P 500 and consumption from 100 times the natural log of real personal consumption expenditures from the U.S. National Income and Product Accounts. All data for this figure are quarterly for the period 1950:1 to 2016:1.

FIGURE 2.—AUTOCORRELATIONS AND CROSS-CORRELATIONS FOR FIRST DIFFERENCE OF STOCK PRICES AND REAL CONSUMPTION SPENDING



(Upper left) Autocorrelations of log growth rate of end-of-quarter value for S&P 500. (Upper right) Autocorrelations of log growth rate of real consumption spending. (Lower panels) Cross-correlations.

FIGURE 3.—AUTOCORRELATIONS AND CROSS-CORRELATIONS FOR HP CYCLICAL COMPONENT OF STOCK PRICES AND REAL CONSUMPTION SPENDING



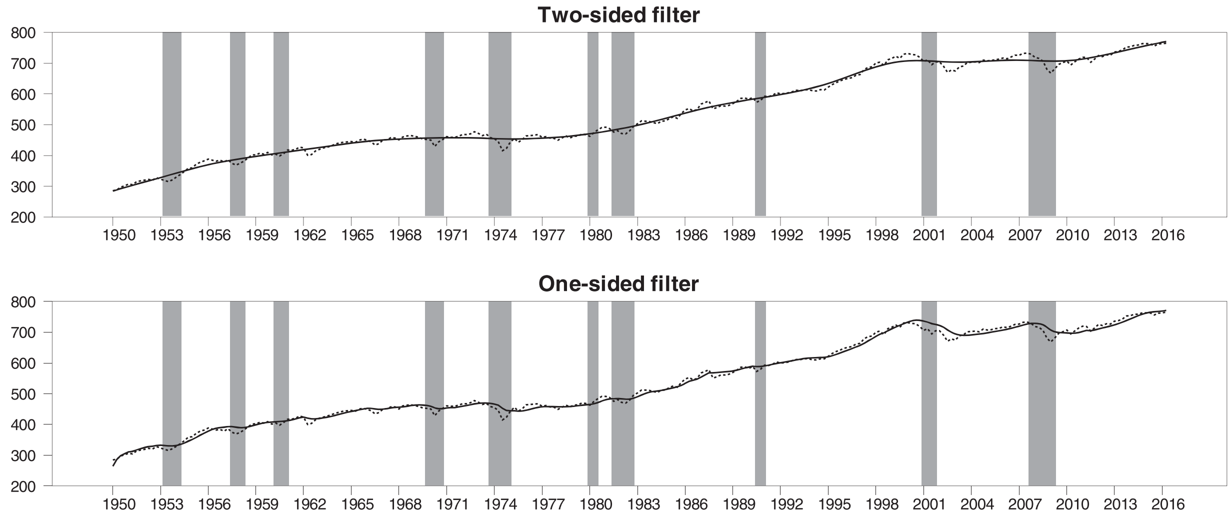
(Upper left) Autocorrelations of HP cycle for log of end-of-quarter value for S&P 500. (Upper right) Autocorrelations of HP cycle for log of real consumption spending. (Lower panels) Cross-correlations.

been observed as of date t rather than also using future values as was done in the HP filter, equation (4). The value of this one-sided projection for date t could be calculated by taking the end-of-sample HP-filtered series for a sample ending at t, repeated for each t.[9]

The top panel of figure 4 shows the result of applying the usual two-sided HP filter to stock prices. The trend is identified to have been essentially flat throughout the 2000s, with the prerecession booms and postrecession busts in stock prices viewed entirely as cyclical phenomena. The bottom panel shows the results of applying a one-sided HP filter to the

[9] An easier way to calculate the one-sided projection is with a single pass of the Kalman filter through the entire sample for the state-space model assumed in proposition 1 with C_0 large and $\sigma_c^2/\sigma_v^2 = 1{,}600$. The Kalman filter gives the one-sided projection, while the Kalman smoother gives the usual two-sided HP filter.

FIGURE 4.—COMPARISON OF ONE-SIDED AND TWO-SIDED HP FILTERS



Dotted line in both panels plots 100 times natural log of S&P 500 stock price index. The solid curve in the top panel plots the HP estimate of trend as inferred using the usual two-sided filter (calculated using the Kalman smoother for the state-space model in proposition 1), whereas the solid curve in the bottom panel plots trend from a one-sided HP filter (calculated using the Kalman filter for the same model). Shaded regions denote NBER recession dates.

same data. This would instead identify the trend component as rising during economic expansions and falling during recessions. The reason is that a real-time observer would not know in early 2009, for example, that stock prices were about to appreciate remarkably and, accordingly, would have judged much of the drop observed up to that date to be permanent. It is only with hindsight that we are tempted to interpret the 2008 stock market crash as a temporary phenomenon. Making use of unknowable future values in this way is in fact a fundamental reason that HP-filtered series exhibit the visual properties that they do, precisely because they impose patterns that are not a feature of the data-generating process and could not be recognized in real time. Some researchers might be attracted by the simple picture of the "long-run" component of stock prices summarized by the top panel of figure 4. But that picture is just something that their imagination has imposed on the data. And the end-of-sample value obtained from the procedure is actually quite different from what the researcher is seeing in the middle of the sample.

Moreover, although a one-sided filter would eliminate the problem of generating a series that is artificially able to predict the future, changes in both the one-sided trend and its implied cycle are readily forecastable from their own lagged values, and likewise by values of any other variables. Again, this is not a feature of the stock prices themselves, but instead an artifact of choosing to characterize the cycle and trend in this particular way.

C. *Data Coherent Values for* λ

A separate question is what value we should use for the smoothing parameter λ. Hodrick and Prescott (1997) motivated their choice of $\lambda = 1{,}600$ based on the prior belief that a large change in the cyclical component within a quarter would be around 5%, whereas a large change in the trend component would be around (1/8)%, suggesting a choice of $\lambda = \sigma_c^2/\sigma_v^2 = (5/(1/8))^2 = 1{,}600$. Ravn and Uhlig (2002) showed how to choose the smoothing parameter for data at other frequencies if indeed it would be correct to use 1,600 on quarterly data. These rules of thumb are almost universally followed.

It is worth noting that if the state-space representation in proposition 1 were indeed an accurate characterization of the trend that we were trying to infer, we would not need to make up a value for λ but could in fact estimate it from the data. If, for example, we assumed a normal distribution for the innovations $(v_t, c_t)'$, we could use the Kalman filter to evaluate the likelihood function for the observed sample $(y_1, \ldots, y_T)'$ and find the values for σ_v^2 and σ_c^2 that maximize the likelihood function.[10] This could alternatively be given a quasi-maximum likelihood interpretation as a generalized least squares minimization of the squared forecast errors weighted by reciprocals of their model-implied variance.

Table 1 reports MLEs of σ_v^2, σ_c^2, and λ for a number of commonly studied macroeconomic series. For every one of these, we would estimate a value for σ_c^2 whose magnitude is similar to, and in fact often smaller than, σ_v^2, and certainly not 1,600 times as large.[11] If we used a value of $\lambda = 1$ instead of $\lambda = 1{,}600$, the resulting series for g_t would differ little from the original data y_t itself; $\lambda = 1$ implies a value for R in expression (10) of 0.48, which decays with a half-life of less than one quarter.

Thus, not only is the HP filter very inappropriate if the true process is a random walk. As commonly applied with $\lambda = 1{,}600$, the HP filter is not even optimal for the only

[10] See, for example, Hamilton (1994, equations (13.4.1) and (13.4.2)). Note that although the inferred value for the trend g_t depends only on the ratio σ_c^2/σ_v^2, the parameters σ_c^2 and σ_v^2 are separately identifiable because σ_c^2 can be inferred from the average observed size of $(y_t - g_t)^2$.

[11] Nelson and Plosser (1982) have also made this observation.

TABLE 1.—MAXIMUM LIKELIHOOD ESTIMATES OF PARAMETERS OF STATE-SPACE FORMALIZATION OF THE HP FILTER FOR ASSORTED QUARTERLY MACROECONOMIC SERIES

	σ_c^2	σ_v^2	λ
GDP	0.115	0.468	0.245
Consumption	0.163	0.174	0.940
Investment	4.187	12.196	0.343
Exports	5.818	3.341	1.741
Imports	4.423	4.769	0.927
Government spending	0.221	1.160	0.191
Employment	0.006	0.250	0.023
Unemployment rate	0.014	0.092	0.152
GDP deflator	0.018	0.081	0.216
S&P 500	21.284	15.186	1.402
10-year Treasury yield	0.135	0.054	2.486
Fed funds rate	0.633	0.116	5.458
Real rate	0.875	0.091	9.596

example—namely, equations (5) and (6)—for which anyone has claimed that it might provide the ideal inference!

IV. A Better Alternative

Here I suggest an alternative concept of what we might mean by the cyclical component of a possibly nonstationary series: How different is the value at date $t+h$ from the value that we would have expected to see based on its behavior through date t?[12] This concept of the cyclical component has several attractive features. First, as den Haan (2000) noted, the forecast error is stationary for a wide class of nonstationary processes. Second, the primary reason that we would be wrong in predicting the value of most macro and financial variables at a horizon of $h = 8$ quarters ahead is cyclical factors such as whether a recession occurs over the next two years and the timing of recovery from any downturn.[13]

While it might seem that calculating this concept of the cyclical component requires us to know the nature of the nonstationarity and to have the correct model for forecasting the series, neither of these is the case. We can instead always rely on very simple forecasts within a restricted class: the population linear projection of y_{t+h} on a constant and the four most recent values of y as of date t. This object exists and can be consistently estimated for a wide range of nonstationary processes, as I now show.

A. *Forecasting When the True Process Is Unknown*

Suppose that the dth difference of y_t is stationary for some d. For example, $d = 2$ would mean that the growth rate is nonstationary but the change in the growth rate is stationary. Note that the dth difference is also stationary for any series with a deterministic time trend characterized by a dth-order polynomial in time. For any such process, we can write the value of y_{t+h} as a linear function of initial conditions at time t plus a stationary process. For example, when $d = 1$, letting $u_t = \Delta y_t$ we can write

$$y_{t+h} = y_t + w_t^{(h)}, \tag{16}$$

where the stationary component is given by $w_t^{(h)} = u_{t+1} + \cdots + u_{t+h}$. For $d = 2$ and $\Delta^2 y_t = u_t$,

$$y_{t+h} = y_t + h\Delta y_t + w_t^{(h)}, \tag{17}$$

where now $w_t^{(h)} = u_{t+h} + 2u_{t+h-1} + \cdots + hu_{t+1}$. This result holds for general d, as demonstrated in the following proposition:

Proposition 3. *If $(1-L)^d y_t$ is stationary for some $d \geq 1$, then for all finite $h \geq 1$,*

$$y_{t+h} = \kappa_h^{(1)} y_t + \kappa_h^{(2)} \Delta y_t + \cdots + \kappa_h^{(d)} \Delta^{d-1} y_t + w_t^{(h)},$$

with $\Delta^s = (1-L)^s$, $\kappa_\ell^{(1)} = 1$ for $\ell = 1, 2, \ldots$ and $\kappa_j^{(s)} = \sum_{\ell=1}^{j} \kappa_\ell^{(s-1)}$ for $s = 2, 3, \ldots, d$ and $w_t^{(h)}$ is a stationary process.

It further turns out that if $\Delta^d y_t \sim I(0)$ and we regress y_{t+h} on a constant and the d most recent values of y as of date t, the coefficients will be forced to be close to the values implied by the coefficients $\kappa_h^{(j)}$ in proposition 3. For example, if $\Delta^2 y_t$ is $I(0)$, then in a regression of y_{t+h} on $(y_t, y_{t-1}, 1)'$, the fitted values will tend to $y_t + h(y_t - y_{t-1}) + \mu_h$ for $\mu_h = E(w_t^{(h)})$ as the sample size gets large; that is, the coefficient on y_t will go to $1+h$, and the coefficient on y_{t-1} will go to $-h$. The implication is that the residuals from a regression of y_{t+h} on $(y_t, y_{t-1}, 1)'$ will be stationary whenever y itself is $I(2)$. The reason is that any other values for these coefficients would imply a nonstationary series for the residuals, whose sum of squares become arbitrarily large relative to those implied by the coefficients $1+h$ and $-h$ as the sample size grows large.

If $\Delta^d y_t$ is stationary and we regress y_{t+h} on a constant and the p most recent values of y as of date t for any $p > d$, the regression will use d of the coefficients to make sure the residuals are stationary and the remaining $p+1-d$ coefficients will be determined by the parameters that characterize the population linear projection of the stationary variable $w_t^{(h)}$ on the stationary regressors $(\Delta^d y_t, \Delta^d y_{t-1}, \ldots, \Delta^d y_{t-p+d+1}, 1)'$. Proposition 4, which follows, provides a formal statement of these claims. In the proof of this proposition, I have followed Stock (1994) in defining a series u_t to be $I(0)$ if it has fixed mean μ and satisfies a functional central limit theorem.[14] This

[12] This idea is related to Beveridge and Nelson's (1981) definition of the trend component of y_t as $g_t = \lim_{h\to\infty} \lim_{p\to\infty} E(y_{t+h}|y_t, y_{t-1}, \ldots, y_{t-p+1})$ which limit exists, and can be calculated provided that $(1-L)y_t$ is a mean-zero stationary process. Beveridge and Nelson then (somewhat curiously) interpreted the cyclical component as $c_t = g_t - y_t$. By contrast, here we keep h and p fixed and take advantage of the fact that $g_t = E(y_{t+h}|y_t, y_{t-1}, \ldots, y_{t-p+1})$ exists for a broad range of nonstationary processes. We interpret the cyclical component at date $t+h$ as $c_{t+h} = y_{t+h} - g_t$.

[13] This same consideration suggests using $h = 24$ for monthly data and $h = 2$ for annual data.

[14] Stock (1994) demonstrated that an example of sufficient conditions that imply equation (18) is that $u_t = \mu + \sum_{j=0}^{\infty} \psi_j \eta_t$ where η_t is a martingale difference sequence with variance σ^2 and finite fourth moment, $\psi(1) \neq 0$, and

requires that the sample mean of u_t has a Normal distribution as the sample size T gets large, as does a sample mean that used only Tr observations for $0 < r \leq 1$. Formally,

$$T^{-1/2}\sum_{s=1}^{[Tr]}(u_t - \mu) \Rightarrow \omega W(r), \tag{18}$$

where $[Tr]$ denotes the largest integer less than or equal to Tr, $W(r)$ denotes standard Brownian motion, and "$\Rightarrow$" denotes weak convergence in probability measure. I will show that if the dth difference ($u_t = \Delta^d y_t$) satisfies equation (18) or if the deviation from a dth-order deterministic polynomial in time ($u_t = y_t - \delta_0 - \delta_1 t - \delta_2 t^2 - \cdots - \delta_d t^d$) satisfies equation (18), then we can remove the nonstationary component with the same simple regression.[15]

Proposition 4. *Suppose that either $u_t = \Delta^d y_t$ satisfies equation* (18) *or that $u_t = y_t - \sum_{j=0}^{d}\delta_j t^j$ with $\delta_d \neq 0$ satisfies equation* (18) *for some unknown d. Let $x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1}, 1)'$ for some $p \geq d$ and consider OLS estimation of $y_{t+h} = x_t'\beta + v_{t+h}$ for $t = 1, \ldots, T$ with estimated coefficient*

$$\hat{\beta} = \left(\sum_{j=1}^{T} x_t x_t'\right)^{-1}\left(\sum_{j=1}^{T} x_t y_{t+h}\right). \tag{19}$$

If $p = d$, the OLS residuals $y_{t+h} - x_t'\hat{\beta}$ converge to the variable $w_t^{(h)} - E(w_t^{(h)})$ in proposition 3. If $p > d$, the OLS residuals converge to the residuals from a population linear projection of $w_t^{(h)}$ on $(\Delta^d y_t, \Delta^d y_{t-1}, \ldots, \Delta^d y_{t-p+d+1}, 1)'$.

Proposition 4 establishes that if we estimate an OLS regression of y_{t+h} on a constant and the $p = 4$ most recent values of y as of date t,

$$y_{t+h} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + v_{t+h}, \tag{20}$$

the residuals,

$$\hat{v}_{t+h} = y_{t+h} - \hat{\beta}_0 - \hat{\beta}_1 y_t - \hat{\beta}_2 y_{t-1} - \hat{\beta}_3 y_{t-2} - \hat{\beta}_4 y_{t-3}, \tag{21}$$

offer a reasonable way to construct the transient component for a broad class of underlying processes. The series is stationary provided that fourth differences of y_t are stationary, a goal that HP intends but does not necessarily achieve. But whereas the approximation to the HP filter in equation (15) imposes all four unit roots, the sample regression would use only four differences if it is warranted by observed features of the data. The proposed procedure has a number of other advantages over HP. First, any finding that $\hat{v}_{t+h}$ predicts some other variable x_{t+h+j} represents a true ability of y to predict x rather than an artifact of the way we chose to detrend y, by virtue of the fact that $\hat{v}_{t+h}$ is a one-sided filter.[16] Second, unlike the HP cyclical series c_{t+h}, the value of $\hat{v}_{t+h}$ will by construction be difficult to predict from variables dated t and earlier.[17] If we find such predictability, it tells us something about the true data-generating process, for example, that x Granger-causes y. Third, the value of $\hat{v}_{t+h}$ is a model-free and essentially assumption-free summary of the data. Regardless of how the data may have been generated, as long as $(1-L)^d y_t$ is covariance stationary for some $d \leq 4$, there exists a population linear projection of y_{t+h} on $(y_t, y_{t-1}, y_{t-2}, y_{t-3}, 1)'$. That projection is a characteristic of the data-generating process that can be used to define what we mean by the cyclical component of the process and can be consistently estimated from the data. Given a dynamic stochastic general equilibrium or any other theoretical model that would imply an $I(d)$ process, we could calculate this population characteristic of the model and estimate it consistently from the data.

B. Properties in Some Common Settings

Random walk. Given the literature cited in section IIIA, it is instructive to examine the consequences if this procedure were applied to a random walk: $y_t = y_{t-1} + \varepsilon_t$. In this case, $d = 1$ and $w_t^{(h)} = \varepsilon_{t+h} + \varepsilon_{t+h-1} + \cdots + \varepsilon_{t+1}$. For large samples, the OLS estimates of equation (20) converge to $\beta_1 = 1$ and all other $\beta_j = 0$, and the resulting filtered series would simply be the difference

$$\tilde{v}_{t+h} = y_{t+h} - y_t, \tag{22}$$

that is, how much the series changes over an $h = 8$ quarter horizon or, equivalently, the sum of the observed changes over h periods. Note that for $h = 8$ the filter $1 - L^h$ wipes out any cycles with frequency of exactly one year, and thus is taking out both the long-run trend as well as any strictly seasonal components.[18] This also fits with the common understanding

$\sum_{j=0}^{\infty} j|\psi_j| < \infty$, in which case ω^2 in equation (18) is given by $\sigma^2[\psi(1)]^2$. Alternatively, Phillips (1987) derived equation (18) from primitive moment and mixing conditions on u_t.

[15] The reason to state these as two separate possibilities is that if the nonstationarity is purely deterministic, then the dth differences will not satisfy the functional central limit theorem. For example, if $y_t = \gamma_0 + \gamma_1 t + \varepsilon_t$ with ε_t white noise, then $\Delta y_t = \gamma_1 + \psi(L)\varepsilon_t$ for $\psi(L) = 1 - L$ and $\psi(1) = 0$. Of course when $u_t = \Delta^d y_t$ satisfies equation (18) with $\mu \neq 0$, the series y_t has both a dth-order stochastic as well as dth-order deterministic polynomial trends, so that case, along with pure stochastic trends ($\mu = 0$) and pure deterministic trends, are all allowed by proposition 4.

[16] While the OLS coefficients, equation (19), make use of future observations, this influence vanishes asymptotically, in contrast to HP's first-order dependence on future observations. Expression (22) below offers another alternative that allows zero inference of future observations for any sample size T.

[17] Note, however, that c_{t+h} by construction can be predicted by variables known at date $t + h - 1$. The value of c_{t+h} will be correlated with its own lagged values $c_{t+h-1}, c_{t+h-2}, \ldots, c_{t+1}$ but likely uncorrelated with $c_t, c_{t-1}, \ldots$

[18] As in Hamilton (1994), the filter $1 - L^8$ has power transfer function $(1 - e^{-8i\omega})(1 - e^{8i\omega}) = 2 - 2\cos(8\omega)$, which is 0 at $\omega = 0, \pi/4, \pi/2, 3\pi/4, \pi$ and thus eliminates not only cycles at the zero frequency but also cycles that repeat themselves every 8,4,8/3, or 2 quarters. See also Hamilton (1994, figures 6.5 and 6.6).

of what we would mean by the cyclical component. Because the simple filter, equation (22), does not require estimation of any parameters, it can also be used as a quick robustness check for concerns about the small-sample applicability of the asymptotic claims in proposition 4, as illustrated in the applications below.

Deterministic time trend. Another instructive example is a pure deterministic time trend of order $d = 1$: $y_t = \delta_0 + \delta_1 t + \varepsilon_t$ for ε_t white noise. In this case, $\Delta y_t = \delta_1 + \varepsilon_t - \varepsilon_{t-1}$ is stationary, and $w_t^{(h)} = \Delta y_{t+1} + \cdots + \Delta y_{t+h} = \delta_1 h + \varepsilon_{t+h} - \varepsilon_t$ is also stationary for any h. I show in the appendix that for this case, the limiting coefficients on $y_t, \ldots, y_{t-p+1}$ described by proposition 4 are each given by $1/p$ and the implied trend for y_{t+h} is

$$\delta_0 + \delta_1(t+h) + p^{-1}(\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_{t-p+1}). \tag{23}$$

Even for $p = 1$ this is not a bad estimate and for $p = 4$ should not differ much from the true trend $\delta_0 + \delta_1(t+h)$. Again regardless of the choice of p, the difference between y_{t+h} and equation (23) will be stationary.

Interpreting DSGE's. A third instructive example is when y_t is an element of a theoretical dynamic stochastic general equilibrium model that is stationary around some steady-state value μ. If the effects of shocks in the theoretical model die out after h periods, then the linear projection, equation (20), in the theoretical model is characterized by $\beta_0 = \mu$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. In other words, the component v_{t+h} is exactly the deviation from the steady state. If shocks have not completely died out after h periods, then part of what is being labeled trend by this method would include the components of shocks that persist longer than h periods. But for any value of h, the linear projection is a well-defined population characteristic of the theoretical stationary model, and there is an exactly analogous object one can calculate in the possibly nonstationary observed data. The method thus offers a way to make an apples-to-apples comparison of theory with data of the sort that users of the HP filter often desire but that the HP filter itself will always fail to deliver.

C. *Specification of p and h*

One might be tempted to use a richer model than equation (20) to forecast y_{t+h}, such as using a vector of variables, more than four lags, or even a nonlinear relation. However, such refinements are completely unnecessary for the goal of extracting a stationary component and have the significant drawback that the more parameters we try to estimate by regression, the more the small sample results are likely to differ from the asymptotic predictions. The simple univariate regression, equation (20), is estimating a population object that is well defined regardless of whether the variable is part of a large vector system with nonlinear dynamics. For this reason, just as the HP filter is always implemented as a univariate procedure, my recommendation is to follow that same strategy for the approach here.

A related issue is the choice of h. For any fixed h, there exists a sample size T for which the results of proposition 4 hold. However, a bigger sample size T will be needed the bigger is h. The information in a finite data set about very long-horizon forecasts is quite limited. If we are interested in business cycles, a two-year horizon should be the standard benchmark. It is also desirable with seasonal data to have both p and h be integer multiples of the number of observations in a year. Hence, for quarterly data, my recommendation is $p = 4$ and $h = 8$.

In other settings, the fundamental interest could be in shocks whose effects last substantially longer than two years but are nevertheless still transient. A leading example would be the recent interest in debt cycles prompted by data sets such as developed by Jordà, Schularick, and Taylor (2017). For such an application, I would use $h = 5$ years, with the regression-free implementation $(y_{t+5} - y_t)$ having particular appeal given the length of data sets available.

D. *Empirical Illustrations*

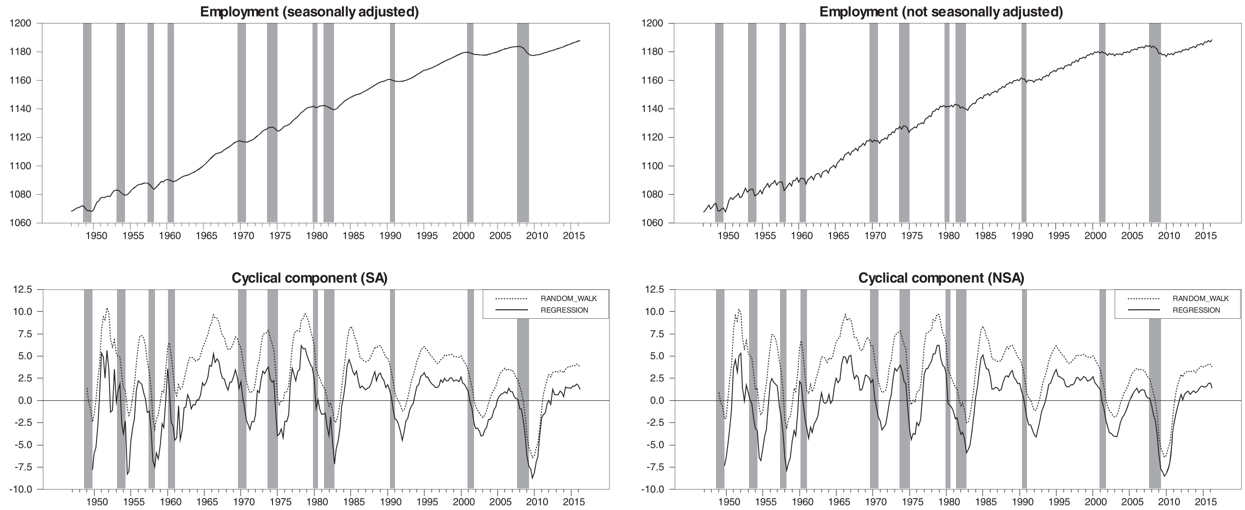
Figure 5 shows the results when this approach is applied to data on U.S. total employment. The raw seasonally adjusted data (y_t) are plotted in the upper left panel. The residuals from regression (20) estimated for these data are plotted in black in the lower-left panel, and the eight-lag difference (22) is in red. The latter two series behave very similarly in this case, as indeed I have found for most other applications. The primary difference is that the regression residual has sample mean 0 by construction (by virtue of the inclusion of a constant term in the regression), whereas the average value of equation (22) will be the average growth rate over a two-year period.

One interesting observation is that the cyclical component of employment starts to decline significantly before the NBER business cycle peak for essentially every recession. Note that this inference from figure 5 is summarizing a true feature of the data and is not an artifact of any forward-looking aspect of the filter.

The right panels of figure 5 show what happens when the same procedure is applied to seasonally unadjusted data. The raw data themselves exhibit a very striking seasonal pattern (see the top right panel). Notwithstanding, the cyclical factor inferred from seasonally unadjusted data (bottom right panel) is almost indistinguishable from that derived from seasonally adjusted data, confirming that this approach is robust to methods of seasonal adjustment.

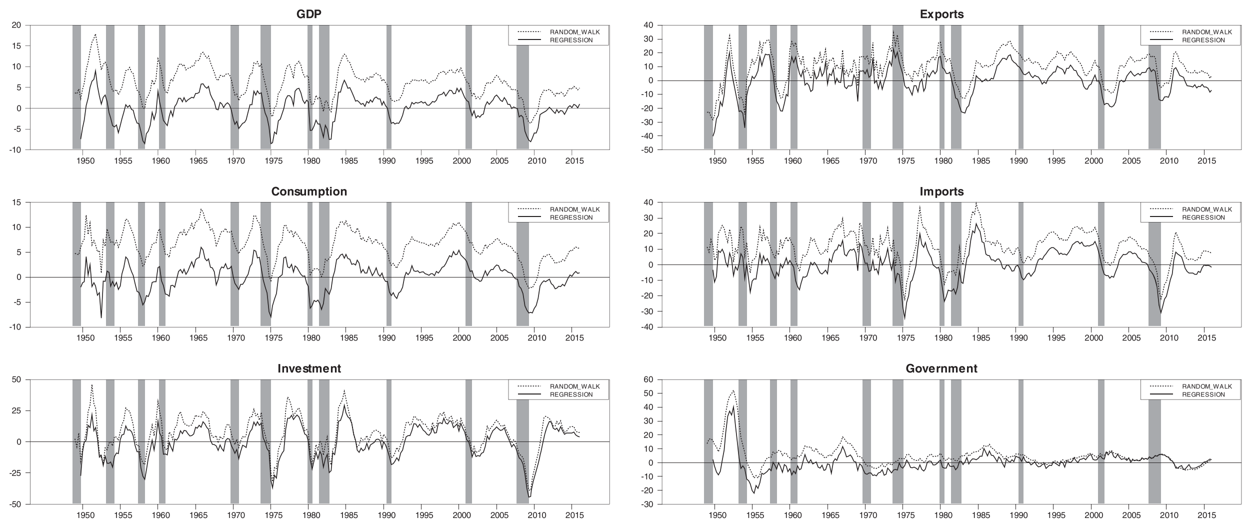
Figure 6 applies the method to the major components of the U.S. national income and product accounts. Investment spending is more cyclically volatile than GDP, while consumption spending is less so. Imports fall significantly during recessions, reflecting lower spending by U.S. residents on imported goods, and exports substantially less so, reflecting the fact that international downturns are often decoupled from

FIGURE 5.—REGRESSION AND EIGHT-QUARTER-CHANGE FILTERS APPLIED TO SEASONALLY ADJUSTED AND SEASONALLY UNADJUSTED EMPLOYMENT DATA



(Upper left) 100 times the log of end-of-quarter values for seasonally adjusted nonfarm payrolls. (Lower left) Solid line plots $y_t - \hat{\beta}_0 - \hat{\beta}_1 y_{t-8} - \hat{\beta}_2 y_{t-9} - \hat{\beta}_3 y_{t-10} - \hat{\beta}_4 y_{t-11}$ as a function of t, while dotted line plots $y_t - y_{t-8}$. Right panels show results when the identical procedure is applied instead to seasonally unadjusted data.

FIGURE 6.—RESULTS OF APPLYING REGRESSION (SOLID) AND EIGHT-QUARTER-CHANGE (DOTTED) FILTERS TO 100 TIMES THE LOG OF COMPONENTS OF U.S. NATIONAL INCOME AND PRODUCT ACCOUNTS



those in the United States. Detrended government spending is dominated by war-related expenditures: the Korean War in the early 1950s, the Vietnam War in the 1970s, and the Reagan military buildup in the 1980s.

Table 2 reports the standard deviation of the cyclical component of each of these and a number of other series, along with their correlation with the cyclical component of GDP. We find very little cyclical correlation between output and prices.[19] Both the nominal fed funds rate and the ex ante real fed funds rate (the latter based on the measure in Hamilton et al., 2016) are modestly procyclical, whereas the ten-year nominal interest rate is not.

V. Conclusion

The HP filter is intended to produce a stationary component from an $I(4)$ series, but in practice, it can fail to do so, and invariably imposes a great cost. It introduces spurious dynamic relations that are purely an artifact of the filter and have no basis in the true data-generating process, and there exists no plausible data-generating process for which common popular practice would provide an optimal decomposition into trend and cycle. There is an alternative approach that can isolate a stationary component from any $I(4)$ series,

[19] Den Haan (2000) found a positive correlation in application of a related methodology. I attribute the difference to differences in sample period.

TABLE 2.—STANDARD DEVIATION OF CYCLICAL COMPONENT AND CORRELATION WITH CYCLICAL COMPONENT OF GDP FOR ASSORTED MACROECONOMIC SERIES

	Regression Residuals		Random walk		
	SD	*GDP Correlation*	*SD*	*GDP Correlation*	Sample
GDP	3.38	1.00	3.69	1.00	1947:1–2016:1
Consumption	2.85	0.79	3.04	0.82	1947:1–2016:1
Investment	13.19	0.84	13.74	0.80	1947:1–2016:1
Exports	10.77	0.33	11.33	0.30	1947:1–2016:1
Imports	9.79	0.77	9.98	0.75	1947:1–2016:1
Government spending	7.13	0.31	8.60	0.38	1947:1–2016:1
Employment	3.09	0.85	3.32	0.85	1947:1–2016:2
Unemployment rate	1.44	−0.81	1.72	−0.79	1948:1–2016:2
GDP deflator	2.99	0.04	4.11	−0.13	1947:1–2016:1
S&P 500	21.80	0.41	22.08	0.38	1950:1–2016:2
10-year Treasury yield	1.46	−0.05	1.51	0.08	1953:2–2016:2
Fed funds rate	2.78	0.33	3.03	0.40	1954:3–2016:2
Real rate	2.25	0.39	2.60	0.42	1958:1–2014:3

Filtered series were based on the full sample available for that variable, while correlations were calculated using the subsample of overlapping values for the two indicators. Note that the regression residuals lose the first eleven observations and the random-walk calculations lose the first eight observations.

preserves the underlying dynamic relations, and consistently estimates well-defined population characteristics for a broad class of possible data-generating processes.

REFERENCES

Atkeson, Andrew, and Lee E. Ohanian, "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review* 25 (1) (2001), 2–11.

Balcilar, Mehmet, Rangan Gupta, Anandamayee Majumdar, and Stephen M. Miller, "Was the Recent Downturn in US Real GDP Predictable?" *Applied Economics* 47 (2015), 2985–3007.

Beveridge, Stephen, and Charles R. Nelson, "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 (1981), 151–174.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual, "Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive?" *Journal of International Money and Finance* 24 (2005), 1150–1175.

Choi, In, "Asymptotic Normality of the Least-Squares Estimates for Higher Order Autoregressive Integrated Processes with Some Applications," *Econometric Theory* 9 (1993), 263–282.

Cogley, Timothy, and James M. Nason, "Effects of the Hodrick-Prescott Filter on Trend and Difference Stationary Time Series: Implications for Business Cycle Research," *Journal of Economic Dynamics and Control* 19 (1995), 253–278.

Cornea-Madeira, Adriana, "The Explicit Formula for the Hodrick-Prescott Filter in Finite Sample," this REVIEW 99 (2017), 314–318.

De Jong, Robert M., and Neslihan Sakarya, "The Econometrics of the Hodrick-Prescott Filter," this REVIEW 98 (2016), 310–317.

Den Haan, Wouter J., "The Comovement between Output and Prices," *Journal of Monetary Economics* 46 (2000), 3–30.

Fama, Eugene F., "The Behavior of Stock-Market Prices," *Journal of Business* 38 (1965), 34–105.

Flood, Robert P., and Andrew K. Rose, "Forecasting International Financial Prices with Fundamentals: How Do Stocks and Exchange Rates Compare?" in Noel Gaston and Ahmed M. Khalid, eds., *Globalization and Economic Integration* (Northampton, MA: Elgar Publishing, 2010).

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy* 86 (1978), 971–987.

Hamilton, James D., *Time Series Analysis* (Princeton: Princeton University Press, 1994).

———- "Understanding Crude Oil Prices," *Energy Journal* 30 (2009), 179–206.

———- "Why You Should Never Use the Hodrick Prescott Filter," NBER working paper 23429 (2017).

Hamilton, James D., Ethan S. Harris, Jan Hatzius, and Kenneth D. West, "The Equilibrium Real Funds Rate: Past, Present and Future," *IMF Economic Review* 64 (2016), 660–707.

Harvey, Andrew C., and Albert Jaeger, "Detrending, Stylized Facts and the Business Cycle," *Journal of Applied Econometrics* 8 (1993), 231–247.

Hodrick, Robert J., and Edward C. Prescott, "Postwar U.S. Business Cycles: An Empirical Investigation," Northwestern University working paper (1981).

Hodrick, Robert J., and Edward C. Prescott, "Postwar U.S. Business Cycles: An Empirical Investigation," *Journal of Money, Credit and Banking* 29 (1997), 1–16.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. "Macrofinancial History and the New Business Cycle Facts" (pp. 213–263), in Martin Eichenbaum and Jonathan A. Parker, eds., *NBER Macroeconomics Annual 2016*, vol. 31 (Chicago: University of Chicago Press, 2017).

King, Robert, and Sergio T. Rebelo. "Low Frequency Filtering and Real Business Cycles," University of Rochester working paper (1989).

——— "Low Frequency Filtering and Real Business Cycles," *Journal of Economic Dynamics and Control* 17 (1993), 207–231.

Mankiw, N. Gregory,"The Optimal Collection of Seigniorage: Theory and Evidence," *Journal of Monetary Economics* 20 (1987), 327–341.

McElroy, Tucker, "Exact Formulas for the Hodrick-Prescott Filter," *Econometrics Journal* 11 (2008), 209–217.

Meese, Richard A., and Kenneth Rogoff, "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?" *Journal of International Economics* 14 (1983), 3–24.

Nelson, Charles R., and Charles R. Plosser, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics* 10 (1982), 139–162.

Park, Joon Y., and Peter C. B. Phillips, "Statistical Inference in Regressions with Integrated Processes: Part 2," *Econometric Theory* 5 (1989), 95–131.

Pesando, James E. "On the Random Walk Characteristics of Short- and Long-Term Interest Rates in an Efficient Market," *Journal of Money, Credit and Banking* 11 (1979), 457–466.

Phillips, Peter C. B., "Time Series Regression with a Unit Root," *Econometrica* 55 (1987), 277–301.

Phillips, Peter C. B., and Sainan Jin, "Business Cycles, Trend Elimination, and the HP Filter," Yale University working paper (2015).

Ravn, Morten O., and Harald Uhlig, "On Adjusting the Hodrick-Prescott Filter for the Frequency of Observations," this REVIEW 84 (2002), 371–380.

Samuelson, Paul, "Proof That Properly Anticipated Prices Fluctuate Randomly," *Industrial Management Review* 6 (1965), 41–49.

Sargent, Thomas J., "A Classical Macroeconometric Model for the United States," *Journal of Political Economy* 84 (1976), 207–237.

Sims, Christopher A., James H. Stock, and Mark W. Watson, "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica* 58 (1990), 113–144.

Singleton, Kenneth J., "Econometric Issues in the Analysis of Equilibrium Business Cycle Models," *Journal of Monetary Economics* 21 (1988), 361–386.

Stock, James H., "Unit Roots, Structural Breaks, and Trends" (pp. 2739–2841), in Robert F. Engle and Daniel L. McFadden, eds., *Handbook of Econometrics*, vol. 4 (Amsterdam: Elsevier, 1994).

APPENDIX

Derivations and Proofs

Derivation of Equation (2). The minimization problem (1) can be written as

$$\min_g \left\{(y-Hg)'(y-Hg)+\lambda(Qg)'(Qg)\right\}.$$

The derivative with respect to g is $-2H'(y-Hg)+2\lambda Q'Qg$ and setting this to 0 gives equation (2).

Derivation of Equation (4). Define $\tilde{a}_t = \left[E(yy')\right]^{-1}E(yg_t)$. For any a_t, we have

$$\begin{aligned}E(g_t-a_t'y)^2 &= E(g_t-\tilde{a}_t'y+\tilde{a}_t'y-a_t'y)^2\\ &= E(g_t-\tilde{a}_t'y)^2+2E[(g_t-\tilde{a}_t'y)y'](\tilde{a}_t-a_t)\\ &\quad+(\tilde{a}_t-a_t)'E(yy')(\tilde{a}_t-a_t).\end{aligned}$$

The middle term equals 0 by the definition of $\tilde{a}_t$:

$$E[(g_t-\tilde{a}_t'y)y'](\tilde{a}_t-a_t)=\{E(g_ty')-E(g_ty')[E(yy')]^{-1}E(yy')\}(\tilde{a}_t-a_t).$$

Hence $E(g_t-a_t'y)^2$ is minimized when $a_t=\tilde{a}_t$. Stacking $\tilde{a}_t'y$ into a $(\tilde{T}\times 1)$ vector gives equation (4).

Proof of Proposition 1. The assumptions can be written formally as

$$E(v_t)=E(c_t)=0, \tag{A1}$$

$$E\begin{bmatrix}v_t\\ c_t\end{bmatrix}\begin{bmatrix}v_{t-j} & c_{t-j}\end{bmatrix}=\begin{cases}\begin{bmatrix}\sigma_v^2 & 0\\ 0 & \sigma_c^2\end{bmatrix} & \text{if } j=0\\ 0 & \text{otherwise}\end{cases}, \tag{A2}$$

$$E(g_0)=E(g_{-1})=0, \tag{A3}$$

$$E\begin{bmatrix}g_0\\ g_{-1}\end{bmatrix}\begin{bmatrix}g_0 & g_{-1}\end{bmatrix}=C_0, \tag{A4}$$

$$E\begin{bmatrix}v_t\\ c_t\end{bmatrix}\begin{bmatrix}g_0 & g_{-1}\end{bmatrix}=0 \quad \text{for } t=1,\ldots T. \tag{A5}$$

I first establish that under equations (5) and (6) and (A1) to (A5),

$$(Q'Q)E(gg')H' \to \sigma_v^2 H' \tag{A6}$$

as $C_0^{-1}\to 0$. To do so, write equation (6) as $Qg=v$ for $v=(v_T,v_{T-1},\ldots,v_1)$ and $Q_0g=v_0$, for v_0, a (2×1) vector with mean 0 and variance $\sigma_v^2 I_2$. Also from equation (A5), v_0 is uncorrelated with v and

$$\underset{(2\times\tilde{T})}{Q_0}=\begin{bmatrix}\underset{(2\times T)}{0} & \underset{(2\times 2)}{P_0^{-1}}\end{bmatrix},$$

where P_0 is the Cholesky factor of C_0 $(P_0P_0'=C_0)$. Stacking these,

$$\begin{bmatrix}Q\\ Q_0\end{bmatrix}g=\begin{bmatrix}v\\ v_0\end{bmatrix}$$

so

$$E(gg')=\sigma_v^2\begin{bmatrix}Q\\ Q_0\end{bmatrix}^{-1}\begin{bmatrix}Q' & Q_0'\end{bmatrix}^{-1},$$

$$\begin{bmatrix}Q' & Q_0'\end{bmatrix}\begin{bmatrix}Q\\ Q_0\end{bmatrix}E(gg')=(Q'Q+Q_0'Q_0)E(gg')=\sigma_v^2 I_{\tilde{T}},$$

$$(Q'Q)E(gg')H'=\sigma_v^2H'-(Q_0'Q_0)E(gg')H',$$

which goes to σ_v^2H' as $P_0^{-1}\to 0$, as claimed in equation (A6).

Notice next from

$$\underset{(T\times 1)}{y} = \underset{(T\times\tilde{T})}{H}\underset{(\tilde{T}\times 1)}{g} + \underset{(T\times 1)}{c}$$

that $E(yy')=HE(gg')H'+\sigma_c^2I_T$ and $E(gy')=E(gg')H'+E(gc')=E(gg')H'$. Hence,

$$\begin{aligned}\tilde{A} &= E(gy')[E(yy')]^{-1}\\ &= E(gg')H'[HE(gg')H'+\sigma_c^2I_T]^{-1}.\end{aligned} \tag{A7}$$

Combining equations (2) and (A7),

$$\begin{aligned}&(H'H+\lambda Q'Q)(A^*-\tilde{A})[HE(gg')H'+\sigma_c^2I_T]\\ &= H'[HE(gg')H'+\sigma_c^2I_T]-(H'H+\lambda Q'Q)E(gg')H'\\ &= H'\sigma_c^2-(\sigma_c^2/\sigma_v^2)(Q'Q)E(gg')H',\end{aligned} \tag{A8}$$

which from equation (A6) goes to 0 as $C_0^{-1}\to 0$. Since the matrices premultiplying and postmultiplying the left side of equation (A8) are of full rank, this establishes that $A^*=\tilde{A}$ as claimed.

Proof of Proposition 2. Let $\theta_1,\theta_2,\theta_3,\theta_4$ be the roots satisfying $F(\theta_i)=0$. As noted by King and Rebelo (1989), since $\lambda>0$, $F(z)$ in equation (8) is positive for all real z, meaning that θ_i comprise two pairs of complex conjugates. Since $F(z)=F(z^{-1})$, if θ_i is a root, then so is θ_i^{-1}. Thus, the values of θ_i are given by $Re^{im}, Re^{-im}, R^{-1}e^{im}$, and $R^{-1}e^{-im}$ for some fixed R and m; one pair is inside the unit circle and the other is outside. Noting that the coefficients on z^2 and z^{-2} in $F(z)$ are both λ, it follows that $F(z)$ can be written as

$$F(z)=\lambda(1-\theta_1z)(1-\theta_2z)(\theta_1^{-1}-z^{-1})(\theta_2^{-1}-z^{-1}).$$

From the symmetry of $F(z)$ in z and z^{-1}, we can without loss of generality normalize θ_1 and θ_2 to be inside the unit circle and write

$$F(z)=\frac{\lambda}{\theta_1\theta_2}(1-\theta_1z)(1-\theta_2z)(1-\theta_1z^{-1})(1-\theta_2z^{-1}).$$

Define $(1-\phi_1z-\phi_2z^2)=(1-\theta_1z)(1-\theta_2z)$, namely, ϕ_1 is the real number $\theta_1+\theta_2$ and ϕ_2 is the negative real number $-\theta_1\theta_2$. Note also that the roots of $(1-\phi_1z-\phi_2z^2)=0$ are the complex conjugates θ_1^{-1} and θ_2^{-1}, which are both outside the unit circle. This gives the bounds on ϕ_1 and ϕ_2 stated in proposition 2 as in Hamilton (1994). Then

$$F(z)=\frac{\lambda}{-\phi_2}(1-\phi_1z-\phi_2z^2)(1-\phi_1z^{-1}-\phi_2z^{-2}). \tag{A9}$$

Evaluating equations (8) and (A9) at $z=1$ gives

$$F(1)=1=(1-\phi_1-\phi_2)^2\lambda/(-\phi_2), \tag{A10}$$

as claimed in equation (12). Likewise, evaluating equations (8) and (A9) at $z=-1$ gives

$$F(-1)=1+16\lambda=(1+\phi_1-\phi_2)^2\lambda/(-\phi_2). \tag{A11}$$

Taking the difference between these last two equations establishes $(4\phi_1-4\phi_1\phi_2)\lambda/(-\phi_2)=16\lambda$ or $\phi_1(1-\phi_2)=-4\phi_2$, as claimed in equation (11). Note that since $\phi_2<0$ (required by complex roots), from equation (11), $\phi_1>0$.

I next establish that

$$\begin{aligned}&\frac{1}{(1-\phi_1z-\phi_2z^2)(1-\phi_1z^{-1}-\phi_2z^{-2})}\\ &=\frac{C_0+C_1z}{1-\phi_1z-\phi_2z^2}+\frac{C_0+C_1z^{-1}}{1-\phi_1z^{-1}-\phi_2z^{-2}}+B_0.\end{aligned} \tag{A12}$$

Combining terms over a common denominator shows that equation (A12) will hold provided

$$\begin{aligned}
1 &= (C_0 + C_1 z)(1 - \phi_1 z^{-1} - \phi_2 z^{-2}) \\
&\quad + (C_0 + C_1 z^{-1})(1 - \phi_1 z^1 - \phi_2 z^2) \\
&\quad + B_0(1 - \phi_1 z - \phi_2 z^2)(1 - \phi_1 z^{-1} - \phi_2 z^{-2}) \\
&= [2C_0 - 2C_1\phi_1 + B_0(1 + \phi_1^2 + \phi_2^2)] \\
&\quad + [C_1 - C_0\phi_1 - C_1\phi_2 - B_0\phi_1 + B_0\phi_1\phi_2](z + z^{-1}) \\
&\quad - [C_0\phi_2 + B_0\phi_2](z^2 + z^{-2}).
\end{aligned}$$

The coefficient on $(z^2 + z^{-2})$ will be 0 provided $B_0 = -C_0$ or provided

$$1 = [C_0 - 2C_1\phi_1 - C_0\phi_1^2 - C_0\phi_2^2] + [C_1 - C_1\phi_2 - C_0\phi_1\phi_2](z + z^{-1}). \tag{A13}$$

The coefficient on $(z + z^{-1})$ will be 0 provided

$$C_1 = \frac{C_0\phi_1\phi_2}{1 - \phi_2} = -C_0\phi_1^2/4 \tag{A14}$$

where the last equation made use of equation (11). Substituting equation (A14) into (A13), we see that equation (A12) will be true provided we set

$$1 = C_0(1 - \phi_1^2 - \phi_2^2 + \phi_1^3/2).$$

Combining these results, we conclude that

$$\begin{aligned}
&\frac{1}{(1 - \phi_1 z - \phi_2 z^2)(1 - \phi_1 z^{-1} - \phi_2 z^{-2})} \\
&\quad = C_0\left[\frac{1 - (\phi_1^2/4)z}{1 - \phi_1 z - \phi_2 z^2} + \frac{1 - (\phi_1^2/4)z^{-1}}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} - 1\right].
\end{aligned}$$

From equation (A9), we then obtain equation (9) with $C = -C_0\phi_2/\lambda$ as claimed in equation (13).

To derive equation (10), recall from Hamilton (1994) that

$$\frac{1}{1 - \phi_1 z - \phi_2 z^2} = \sum_{j=0}^{\infty} R^j[2\alpha\cos(mj) + 2\beta\sin(mj)]z^j. \tag{A15}$$

We know that the coefficient on z^j for $j = 0$ must be 1, requiring $[2\alpha\cos(0) + 2\beta\sin(0)] = 1$ or $\alpha = 1/2$. We likewise know that the coefficient on z^j for $j = 1$ is given by ϕ_1, so $R[\cos(m) + 2\beta\sin(m)] = \phi_1$, which, from equation (14), gives $R2\beta\sin(m) = \phi_1/2$ or $2\beta\sin(m) = \cos(m)$ so $2\beta = \cot(m)$. Substituting these values for α and β into equation (A15) gives equation (10).

Proof of Proposition 3. Recall the identity

$$y_{t+h} = y_t + \sum_{j=1}^{h} \Delta y_{t+j}, \tag{A16}$$

which immediately gives the result of proposition 3 for the case $d = 1$ as stated in equation (16). We likewise have the identity

$$\Delta y_{t+j} = \Delta y_t + \sum_{s=1}^{j} \Delta^2 y_{t+s}. \tag{A17}$$

Substituting equation (A17) into (A16) gives

$$y_{t+h} = y_t + \Delta y_t \sum_{j=1}^{h} 1 + w_t^{(h)},$$

for $w_t^{(h)} = \sum_{j=1}^{h}\sum_{s=1}^{j} \Delta^2 y_{t+s}$ as claimed in equation (17) for the case $d = 2$. We can proceed recursively using the identity $\Delta^k y_{t+s} = \Delta^k y_t + \sum_{r=1}^{s} \Delta^{k+1} y_{t+r}$ and substituting into the preceding expression. For any d, the resulting $w_t^{(h)}$ is a finite sum of stationary variables and therefore is itself stationary.

Proof of Proposition 4. Note that the fitted values and residuals implied by the coefficients in equation (19) are numerically identical to those if we were to do the (infeasible) regression $y_{t+h} = \tilde{x}_t'\alpha + v_{t+h}$ for

$$\tilde{x}_t = (\tilde{u}_t, \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-p+d+1}, 1, \Delta^{d-1}y_t, \Delta^{d-2}y_t, \ldots, \Delta y_t, y_t)'$$

with $\tilde{u}_t = \Delta^d y_t - \mu$. The latter regression is infeasible because we do not know the true values of μ and d. But because the fitted values are the same, once we find the properties of the second regression, we will also know the properties of the first. For example, for $d = 2$ and $p = 4$,

$$\tilde{x}_t = \begin{bmatrix} 1 & -2 & 1 & 0 & -\mu \\ 0 & 1 & -2 & 1 & \mu \\ 0 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ 1 \end{bmatrix} \equiv Hx_t \tag{A18}$$

and $\hat{\alpha} = \left(\sum Hx_t x_t' H'\right)^{-1}\left(\sum Hx_t y_{t+h}\right)$ so $\hat{\beta} = H'\hat{\alpha}$ for every sample. When $p = d$, we define the $(p+1) \times 1$ vector as $\tilde{x}_t = (1, \Delta^{d-1}y_t, \Delta^{d-2}y_t, \ldots, \Delta y_t, y_t)'$, that is, none of the $\tilde{u}_{t-j}$ variables appear in $\tilde{x}_t$ when $p = d$.

Define q to be the $(p+1) \times 1$ vector $q = (0, \ldots, 0, E(w_t^{(h)}), \kappa_h^{(d)}, \kappa_h^{(d-1)}, \ldots, \kappa_h^{(1)})'$, so that $\tilde{w}_t^{(h)} = w_t^{(h)} - E(w_t^{(h)}) = y_{t+h} - \tilde{x}_t'q$ and

$$\begin{aligned}
\hat{\alpha} &= \left(\sum \tilde{x}_t\tilde{x}_t'\right)^{-1} \sum \tilde{x}_t(\tilde{x}_t'q + \tilde{w}_t^{(h)}) \\
&= q + \left(\sum \tilde{x}_t\tilde{x}_t'\right)^{-1} \sum \tilde{x}_t\tilde{w}_t^{(h)}.
\end{aligned} \tag{A19}$$

We first consider the case when equation (18) holds for $\Delta^d y_t$ when there is no further drift and the initial value for all of the difference processes is 0—namely, the case when $\mu = 0$ and $\Delta^{d-j}y_t = \xi_t^{(j)}$ where $\xi_t^{(1)} = \sum_{j=1}^{t}\tilde{u}_j$ and $\xi_t^{(s)} = \sum_{j=1}^{t}\xi_j^{(s-1)}$ for $s = 2, 3, \ldots, d$. For this case, define

$$\Upsilon_T = \begin{bmatrix} T^{1/2}I_{p-d+1} & 0 & 0 & \cdots & 0 \\ 0 & T & 0 & \cdots & 0 \\ 0 & 0 & T^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & T^d \end{bmatrix}. \tag{A20}$$

Adapting the approach in Sims, Stock, and Watson (1990), we have from equation (A19) that

$$\begin{aligned}
T^{-1/2}\Upsilon_T(\hat{\alpha} - q) &= T^{-1/2}\Upsilon_T\left(\sum \tilde{x}_t\tilde{x}_t'\right)^{-1} \sum \tilde{x}_t\tilde{w}_t^{(h)} \\
&= T^{-1/2}\left[\Upsilon_T^{-1}\sum \tilde{x}_t\tilde{x}_t'\Upsilon_T^{-1}\right]^{-1}\Upsilon_T^{-1}\sum \tilde{x}_t\tilde{w}_t^{(h)} \\
&= \left[\Upsilon_T^{-1}\sum \tilde{x}_t\tilde{x}_t'\Upsilon_T^{-1}\right]^{-1}\left[T^{-1/2}\Upsilon_T^{-1}\sum \tilde{x}_t\tilde{w}_t^{(h)}\right].
\end{aligned} \tag{A21}$$

Consider first the last term in equation (A21):

$$T^{-1/2}\Upsilon_T^{-1}\sum \tilde{x}_t\tilde{w}_t^{(h)} = \begin{bmatrix} T^{-1}\sum \tilde{u}_t\tilde{w}_t^{(h)} \\ \vdots \\ T^{-1}\sum \tilde{u}_{t-p+d+1}\tilde{w}_t^{(h)} \\ T^{-1}\sum \tilde{w}_t^{(h)} \\ T^{-3/2}\sum \xi_t^{(1)}\tilde{w}_t^{(h)} \\ T^{-5/2}\sum \xi_t^{(2)}\tilde{w}_t^{(h)} \\ \vdots \\ T^{-d-1/2}\sum \xi_t^{(d)}\tilde{w}_t^{(h)} \end{bmatrix}. \tag{A22}$$

The first $p-d$ terms are just the sample means of stationary variables, which by the law of large numbers converge in probability to their expectation $E(\tilde{u}_{t-j}\tilde{w}_t^{(h)})$. Term $p-d+1$ likewise converges to $E(\tilde{w}_t^{(h)}) = 0$. Calculations analogous to those behind lemma 1(e) in Sims et al. (1990) show that the last d terms in equation (A22) also all converge in probability to 0.[20]

Turning next to the first term in equation (A21), the upper-left $(p-d) \times (p-d)$ block of $\Upsilon_T^{-1} \sum \tilde{x}_t \tilde{x}_t' \Upsilon_T^{-1}$ is characterized by

$$\begin{bmatrix} T^{-1}\sum \tilde{u}_t^2 & \cdots & T^{-1}\sum \tilde{u}_t\tilde{u}_{t-p+d+1} \\ \vdots & \cdots & \vdots \\ T^{-1}\sum \tilde{u}_{t-p+d+1}\tilde{u}_t & \cdots & T^{-1}\sum \tilde{u}_{t-p+d+1}^2 \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-d-1} \\ \vdots & \cdots & \vdots \\ \gamma_{p-d-1} & \cdots & \gamma_0 \end{bmatrix}$$

for $\gamma_j = E(\tilde{u}_t\tilde{u}_{t-j})$. From Sims et al. (1990) lemmas 1a and 1b, the lower-right $(d+1) \times (d+1)$ block satisfies

$$\begin{bmatrix} 1 & T^{-3/2}\sum \xi_t^{(1)} & T^{-5/2}\sum \xi_t^{(2)} & \cdots & T^{-d-1/2}\sum \xi_t^{(d)} \\ T^{-3/2}\sum \xi_t^{(1)} & T^{-2}\sum [\xi_t^{(1)}]^2 & T^{-3}\sum \xi_t^{(1)}\xi_t^{(2)} & \cdots & T^{-d-1}\sum \xi_t^{(1)}\xi_t^{(d)} \\ T^{-5/2}\sum \xi_t^{(2)} & T^{-3}\sum \xi_t^{(2)}\xi_t^{(1)} & T^{-4}\sum [\xi_t^{(2)}]^2 & \cdots & T^{-d-2}\sum \xi_t^{(2)}\xi_t^{(d)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ T^{-d-1/2}\sum \xi_t^{(d)} & T^{-d-1}\sum \xi_t^{(d)}\xi_t^{(1)} & T^{-d-2}\sum \xi_t^{(d)}\xi_t^{(2)} & \cdots & T^{-2d}\sum [\xi_t^{(d)}]^2 \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} 1 & \omega\int_0^1 W^{(1)}(r)dr & \omega\int_0^1 W^{(2)}(r)dr & \cdots & \omega\int_0^1 W^{(d)}(r)dr \\ \omega\int_0^1 W^{(1)}(r)dr & \omega^2\int_0^1 [W^{(1)}(r)]^2dr & \omega^2\int_0^1 W^{(1)}(r)W^{(2)}(r)dr & \cdots & \omega^2\int_0^1 W^{(1)}(r)W^{(d)}(r)dr \\ \omega\int_0^1 W^{(2)}(r)dr & \omega^2\int_0^1 W^{(2)}(r)W^{(1)}(r)dr & \omega^2\int_0^1 [W^{(2)}(r)]^2dr & \cdots & \omega^2\int_0^1 W^{(2)}(r)W^{(d)}(r)dr \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \omega\int_0^1 W^{(d)}(r)dr & \omega^2\int_0^1 W^{(d)}(r)W^{(1)}(r)dr & \omega^2\int_0^1 W^{(d)}(r)W^{(2)}(r)dr & \cdots & \omega^2\int_0^1 [W^{(d)}(r)]^2dr \end{bmatrix}$$

where $W^{(1)}(r)$ denotes standard Brownian motion and $W^{(j)}(r) = \int_0^r W^{(j-1)}(s)ds$. For the off-diagonal block of $\Upsilon_T^{-1} \sum \tilde{x}_t \tilde{x}_t' \Upsilon_T^{-1}$ we see using calculations analogous to Sims et al.'s lemma 1e that

$$\begin{bmatrix} T^{-1}\sum \tilde{u}_t & \cdots & T^{-1}\sum \tilde{u}_{t-p+d+1} \\ T^{-3/2}\sum \xi_t^{(1)}\tilde{u}_t & \cdots & T^{-3/2}\sum \xi_t^{(1)}\tilde{u}_{t-p+d+1} \\ T^{-5/2}\sum \xi_t^{(2)}\tilde{u}_t & \cdots & T^{-5/2}\sum \xi_t^{(2)}\tilde{u}_{t-p+d+1} \\ \vdots & \cdots & \vdots \\ T^{-d-1/2}\sum \xi_t^{(d)}\tilde{u}_t & \cdots & T^{-d-1/2}\sum \xi_t^{(d)}\tilde{u}_{t-p+d+1} \end{bmatrix} \xrightarrow{p} 0.$$

Bringing all these results together, it follows that

$$T^{-1/2}\Upsilon_T(\hat{\alpha} - q) \xrightarrow{p} \begin{bmatrix} g \\ 0 \end{bmatrix} \tag{A23}$$

$$g = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-d-1} \\ \vdots & \cdots & \vdots \\ \gamma_{p-d-1} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} E(\tilde{u}_t\tilde{w}_t^{(h)}) \\ \vdots \\ E(\tilde{u}_{t-p+d+1}\tilde{w}_t^{(h)}) \end{bmatrix}.$$

Note that g corresponds to the coefficients from a population linear projection of $\tilde{w}_t^{(h)}$ on $(\tilde{u}_t, \tilde{u}_{t-1}, \ldots, \tilde{u}_{t-p+d+1})'$. Thus, equation (A23) establishes that the first $p-d$ elements of $\hat{\alpha}$ converge to the stationary population projection coefficients g, the $p-d+1$ term to $E(w_t^{(h)})$, and the last d elements of $\hat{\alpha}$ converge to the $\kappa_h^{(j)}$ terms in q. Indeed, the latter estimates are superconsistent; they still converge to the terms in q even when multiplied by some positive power of T.

[20] That is, before multiplying by $T^{-1/2}$ the terms are all $O_p(1)$. For similar calculations, see lemma 1b in Choi (1993) and proposition 17.3e in Hamilton (1994).

Taking again the $p = 4$ and $d = 2$ example, equation (A18), the coefficients $\hat{\beta}$ from the actual regression of y_{t+h} on $(y_t, y_{t-1}, y_{t-2}, y_{t-3}, 1)'$ have plim

$$\hat{\beta} \xrightarrow{p} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ -2 & 1 & 0 & -1 & 0 \\ 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\mu & -\mu & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \mu h(h+1)/2 \\ h \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} g_1 + h + 1 \\ g_2 - 2g_1 - h \\ g_1 - 2g_2 \\ g_2 \\ \mu\{[h(h+1)/2] - g_1 - g_2\} \end{bmatrix}.$$

The above derivation assumed $\mu = 0$ so that there was no drift in $\Delta^d y_t$. If instead we had $\mu \neq 0$, then $\Delta^{d-1}y_t = \sum_{s=1}^t u_s = \sum_{s=1}^t \tilde{u}_s + t\mu = \xi_t^{(1)} + t\mu$, which is dominated for large t by the drift term $t\mu$ rather than the random walk term $\xi_t^{(1)}$, and $\Delta^{d-j}y_t = \xi_t^{(j)} + (1/j!)t^j\mu + o_p(t^j)$. In this case, we would simply replace Υ_T in the above derivations with

$$\tilde{\Upsilon}_T = \begin{bmatrix} T^{1/2}I_{p-d+1} & 0 & 0 & \cdots & 0 \\ 0 & T^{3/2} & 0 & \cdots & 0 \\ 0 & 0 & T^{5/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & T^{d+1/2} \end{bmatrix}. \tag{A24}$$

We would then arrive at the identical conclusion, equation (A23), this time using results a, c, and g from Sims et al.'s lemma 1 (1990).

Alternatively, adding a non-0 initial condition, for example, replacing $\xi_t^{(1)}$ with $\xi_t^{(1)} + \xi_0^{(1)}$ for $\xi_0^{(1)}$, any fixed constant produces a term that is still dominated asymptotically by $\xi_t^{(1)}$, and as in Park and Phillips (1989), the original convergence claims again all go through.

Finally, the derivations are very similar for the case of purely deterministic time trends, $y_t = \sum_{j=0}^d \delta_j t^j + u_t$. For this case, we have $\mu = E(\Delta^d y_t) = \delta_d$ and

$$\tilde{x}_t = \Bigg(\Delta^d u_t - \delta_d, \ldots, \Delta^d u_{t-p+d+1} - \delta_d, 1, \sum_{j=0}^{1} \delta_j^{(d-1)} t^j + \Delta^{d-1}u_t, \ldots, \sum_{j=0}^{d-1} \delta_j^{(1)} t^j + \Delta u_t, \sum_{j=0}^{d} \delta_j t^j + u_t\Bigg)',$$

where $\sum_{j=0}^{d-s} \delta_j^{(s)} t^j = \sum_{j=0}^{d-s+1} \delta_j^{(s-1)} t^j - \sum_{j=0}^{d-s+1} \delta_j^{(s-1)}(t-1)^j$ and $\delta_j^{(0)} = \delta_j$. Then for $\tilde{\Upsilon}_T$ as in equation (A24), we again have

$$T^{-1/2}\tilde{\Upsilon}_T^{-1} \sum \tilde{x}_t \tilde{w}_t^{(h)} \xrightarrow{p} (E(\Delta^d u_t - \delta_d)\tilde{w}_{t+h}, \ldots, E(\Delta^d u_{t-p+d+1} - \delta_d)\tilde{w}_{t+h}, 0, \ldots, 0)'.$$

The matrix $\tilde{\Upsilon}_T^{-1} \sum \tilde{x}_t \tilde{x}_t' \tilde{\Upsilon}_T^{-1}$ likewise has a block-diagonal plim, so for g, the coefficients of the population linear projection of $\tilde{w}_{t+h}$ on $(\Delta^d y_t - \delta_d, \ldots, \Delta^d y_{t-p+d+1} - \delta_d)'$, we have

$$\hat{\alpha} \xrightarrow{p} (g, E(w_t^{(h)}), \kappa_h^{(d)}, \ldots, \kappa_h^{(1)})'. \tag{A25}$$

Derivation of Equation (23). See Hamilton (2017).