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 \(\lambda \rightarrow 0\), \(g_t\) would just be the series \(y_t\) itself, whereas when \(\lambda \rightarrow \infty\) 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 \(\tilde{T} = T + 2\), \(y = (y_T, \ldots, y_1)\) and \(g = (g_T, g_{T-1}, \ldots, g_1)\) and

\[
\begin{bmatrix}
I_T \\
T \times (T + 2)
\end{bmatrix} \begin{bmatrix}
0 \\
(T \times T) (T \times 2)
\end{bmatrix}
\]

The solution to equation (1) is then given by\(^2\)

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

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 \(\bar{g}_t = a_t'y\) 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 \(g\):

\[
\bar{g} = E(g'y) [E(y'y)]^{-1} y = \bar{A}'y. \tag{4}
\]

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 \((2 \times 2)\) variance of \((g_0, g_{-1})\). These assumptions imply a particular value for \(\lambda\) in equation (4). As we let the variance of \((g_0, g_{-1})\) become arbitrarily large (represented as \(C_0^{-1} \rightarrow 0\)), then in every sample, the inference, equation (4), would be numerically identical to expression (2).

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

* University of California, San Diego.

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

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

© 2018 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 \( \hat{A} \) in equation (4) converges to the matrix \( \hat{A}^\ast \) in equation (2) as \( C_{0}^{-1} \rightarrow 0 \).

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 that is supposed to be interpreted as the cyclical component of equations whose \( \lambda = \sigma_c^2/\sigma_v^2 \), 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 \( c_t = y_t - \hat{g}_t \) 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 \( H' + \lambda Q' \) gives a system of equations whose \( t \)th element is

\[
[1 + \lambda (1 - L^{-1})^2 (1 - L)]g^*_t = y_t
\]

for \( t = 1, 2, \ldots, T - 2 \) (7)

for \( L \) the lag operator \( (L^k x_t = x_{t-k}, L^{-k} x_t = x_{t+k}) \). In other words, \( F(L)g^*_t = \gamma_t \) for

\[
F(L) = 1 + \lambda (1 - L^{-1})^2 (1 - L)^2.
\]

The following proposition establishes some properties of this filter:

\[
\begin{align*}
[F(L)]^{-1} &= C \\
&= L \frac{1 - (\phi^2_1 \phi^2_2)}{1 - \phi_1 L - \phi_2 L^2} + L^{-1} \frac{1 - (\phi^2_1 \phi^2_2)}{1 - \phi_1 L^{-1} - \phi_2 L^{-2} - 1}.
\end{align*}
\]

where

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

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

\[
\phi_1 (1 - \phi_2) = -4 \phi_2,
\]

\[
(1 - \phi_1 - \phi_2)^2 = -\phi_2/\lambda,
\]

\[
R = \sqrt{\phi_2},
\]

\[
\cos(m) = \phi_1/(2R).
\]

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

\[
\begin{align*}
\lambda &= 0 \\
\lambda &= 1600 \\
\lambda &= \infty
\end{align*}
\]

Figure 1 plots the values of \( \phi_1 \) and \( \phi_2 \) generated by different values of \( \lambda \). 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 \( R^{60} = 0.0012 \).
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

\[
c_t = \frac{\lambda(1-L)^4}{F(L)} y_t = \frac{\lambda(1-L)^2}{F(L)} (1-L)^2 g_t = \frac{\lambda(1-L)^2}{F(L)} y_t - \epsilon_t.
\]

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 non-stationarity 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, Ateskon and Ohanian (2001) for inflation, or Balciilar 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 \( y_t = y_{t-1} + \epsilon_t \), where \( \epsilon_t \) is white noise and \((1-L)y_t = \epsilon_t\), 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

\[
c_t = \frac{\lambda(1-L)^3}{F(L)} \epsilon_{t+2}.
\]

For \( \lambda = 1,600 \) this is

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

with \( q_{0,t} = \epsilon_t - 3\epsilon_{t-1} + 3\epsilon_{t-2} - \epsilon_{t-3} \). \( q_{1,t} = \epsilon_t - 3.79\epsilon_{t-1} + 5.37\epsilon_{t-2} - 3.37\epsilon_{t-3} + 0.79\epsilon_{t-4} \), and \( q_{2,t} = -0.79\epsilon_{t+1} + 3.37\epsilon_t - 5.37\epsilon_{t-1} + 3.79\epsilon_{t-2} - \epsilon_{t-3} \). The underlying innovations \( \epsilon_t \) 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 \( \epsilon_{t-j} \)) and will in turn predict the future (as a result of dependence on future values of \( \epsilon_{t+j} \)). Since the coefficients that make up \([F(L)]^{-1}\) are determined solely by the value of \( \lambda \), 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. 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 clear 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.
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 $t$.\(^9\)

\(^9\) 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.
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 foreseeable 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 $\lambda$.

A separate question is what value we should use for the smoothing parameter $\lambda$. 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 $\lambda = \sigma^2_t/\sigma^2_c = (5/(1/8))^2 = 1,600$. 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 $\lambda$, but could in fact estimate it from the data. If, for example, we assumed a normal distribution for the innovations $\nu_t$, $c_t$, we could use the Kalman filter to evaluate the likelihood function for the observed sample $(y_t, \ldots, y_T)$ and find the values for $\sigma^2_c$ and $\sigma^2_t$ that maximize the likelihood function. 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 $\sigma^2_c$, $\sigma^2_t$, and $\lambda$ for a number of commonly studied macroeconomic series. For every one of these, we would estimate a value for $\sigma^2_T$ whose magnitude is similar to, and in fact often smaller than, $\sigma^2_c$, and certainly not 1,600 times as large.

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 $\sigma^2_t/\sigma^2_c$, the parameters $\sigma^2_t$ and $\sigma^2_c$ are separately identifiable because $\sigma^2_t$ can be inferred from the average observed size of $(y_t - g_t)^2$.

11 Nelson and Plosser (1982) have also made this observation.
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 \( d \)th difference of \( y \) is stationary for some \( d \). For example, \( d = 2 \) would mean that the growth rate is not stationary but the change in the growth rate is stationary. Note that the \( d \)th difference is also stationary for any series with a deterministic time trend characterized by a \( d \)th-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

\[
y_{t+h} = y_t + w_t^{(h)},
\]

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

\[
y_{t+h} = y_t + h \Delta y_t + w_t^{(h)},
\]

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 \( \Delta^s = (1 - L)^s \), \( \kappa_h^{(1)} = 1 \) for \( s = 1, 2, \ldots \) and \( \kappa_h^{(s)} = \sum_{j=1}^{s-1} \kappa_h^{(s-j)} \) 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 \( \kappa_h \) 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 \( y_t \) 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 becomes 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 \((\Delta^s y_t, \Delta^s y_{t-1}, \ldots, \Delta^s y_{t-p+d+1}, 1)' \). 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 \( \mu \) and satisfies a functional central limit theorem.¹⁴

¹² This idea is related to Beveridge and Nelson’s (1981) definition of the trend component of \( y_t \) as \( g_t = \lim_{h \to \infty} \epsilon_{y_{t+h}}(y_{t+1}, \ldots, y_{t+h+1}) \) 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 \( c_t = y_t - g_t \). By contrast, here we keep \( h \) and \( p \) fixed and take advantage of the fact that \( g_t = E(y_{t+h}|y_{t-1}, \ldots, y_{t-p+1}) \) exists for a broad range of nonstationary processes. We interpret the cyclical component at date \( t + h \) as \( c_{t+h} = y_{t+h} - \hat{g}_t \).

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 \( u_t = \mu + \sum_{j=0}^\infty \psi_j \eta_j \) where \( \eta_j \) is a martingale difference sequence with variance \( \sigma^2 \) 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,
\[
T^{-1/2} \sum_{s=1}^{[Tr]} (u_t - \mu) \Rightarrow \omega W(r),
\]
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 \( d \)th difference \( u_t = \Delta^d y_t \) satisfies equation (18) or if the deviation from a \( d \)th-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 \( \delta_j \neq 0 \) satisfies equation (18) for some unknown \( d \). Let \( x_t = (y_t, y_{t+1}, \ldots, y_{t-p+1})' \) for some \( p \geq d \) and consider OLS estimation of \( y_{t+h} = x_t \hat{\beta} + v_{t+h} \) for \( t = 1, \ldots, T \) with estimated coefficient \( \hat{\beta} = \left( \sum_{j=1}^{T} x_j x_j' \right)^{-1} \left( \sum_{j=1}^{T} x_j y_{t+h} \right) \).

If \( p = d \), the OLS residuals \( y_{t+h} - \hat{\beta} x_t \) 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 \( t \),
\[
y_{t+h} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + v_{t+h},
\]
the residuals,
\[
v_{t+h} = y_{t+h} - \hat{\beta}_0 - \hat{\beta}_1 y_t - \hat{\beta}_2 y_{t-1} - \hat{\beta}_3 y_{t-2} - \hat{\beta}_4 y_{t-3},
\]
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,
\[
\sum_{i=0}^{\infty} \phi_i < \infty, \text{ in case } \sigma^2 \text{ in equation (18) is given by } \sigma^2 [ \phi(1) ]^2.
\]
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 \( d \)th differences will not satisfy the functional central limit theorem. For example, if \( y_t = y_0 + \gamma t + \varepsilon_t \) with \( \varepsilon_t \) white noise, then \( \Delta y_t = \gamma + \varepsilon(L) \gamma \) for \( \varepsilon(L) = 1 - L \) 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 \( d \)th-order stochastic as well as \( d \)th-order deterministic polynomial trends, so that case, along with pure stochastic trends (\( \mu = 0 \)) and pure deterministic trends, are all allowed by proposition 4.

---

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

\(^{16}\) 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 \( \sigma_{c_{i|h}} \) by construction can be predicted by variables known at date \( t + h - 1 \). The value of \( \sigma_{c_{i|h}} \) will be correlated with its own lagged values \( \sigma_{c_{i+h-1}}, \sigma_{c_{i+h-2}}, \ldots, \sigma_{i+h} \) but likely uncorrelated with \( \sigma_1, \sigma_2, \ldots, \sigma_h \).

\(^{18}\) As in Hamilton (1994), the filter \( 1 - L^h \) has power transfer function \((1 - e^{-2\pi h\omega})(1 - e^{-2\pi i \delta h \omega}) = 2 - 2 \cos(8\omega), \) which is \( 0 \) at \( \omega = \pi / 4, \pi / 2, 3\pi / 4, \pi \) 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 any compensation of any parameters, it can also be used as a quick robustness check for concerns about the small-sample applicability of the results of proposition 4, as illustrated in the applications below.

Deterministic time trend. Another instructive example is a pure deterministic trend of order \( d = 1 \): \( y_t = \delta_0 + \delta_1 t + \varepsilon_t \) for \( \varepsilon_t \) 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

\[
\delta_0 + \delta_1 (t + h) + p^{-1} (\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_{t-p+1}).
\]

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 \( \mu \). 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 \( y_{t+h} \) 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 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 \( y_{t+h} \) 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 imports.
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.
preserves the underlying dynamic relations, and consistently estimates well-defined population characteristics for a broad class of possible data-generating processes.

REFERENCES


APPENDIX

Derivations and Proofs

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

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

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

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

$$E(g_t - a_t'y)^2 = E(g_t - \tilde{\alpha}_t y + \tilde{\alpha}_t y - a_t'y)^2$$

$$= E((g_t - \tilde{\alpha}_t y)^2 + 2E((g_t - \tilde{\alpha}_t y)'(\tilde{\alpha}_t - a_t))$$

$$+ (\tilde{\alpha}_t - a_t)'E(yy') (\tilde{\alpha}_t - a_t).$$

The middle term equals 0 by the definition of $\tilde{\alpha}_t$.

$$E((g_t - \tilde{\alpha}_t y)'(\tilde{\alpha}_t - a_t)) = [E(g_t'y) - E(g_t'y)]E(yy') (\tilde{\alpha}_t - a_t).$$

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

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

$$E(v_t) = E(c_t) = 0,$$

$$E(v_t c_{t-j}) = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_z^2 \end{bmatrix} \quad \text{if} \quad j = 0,$$

$$E(v_t c_{t-j}) = \begin{bmatrix} \sigma_z^2 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{otherwise}.$$  \hfill (A1)

$$E(g_0) = E(g_{-1}) = 0,$$  \hfill (A2)

$$E \begin{bmatrix} g_0 \\ g_{-1} \end{bmatrix} = C_0,$$  \hfill (A3)

$$E \begin{bmatrix} v_t \\ c_t \end{bmatrix} g_{-1} = 0 \quad \text{for} \quad t = 1, \ldots, T.$$  \hfill (A4)

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

$$E(gg')H' \rightarrow \sigma_z^2 H'$$ \hfill (A6)

as $C_0^{-1} \rightarrow 0$. To do so, write equation (6) as $Qg = v$ for $v = (v_T, v_{T-1}, \ldots, v_1)$ and $Qg_0 = v_0$, for $v_0$, a $(2 \times 1)$ vector with mean 0 and variance $\sigma_z^2 I_2$. Also from equation (A5), $v_0$ is uncorrelated with $v$ and

$$Q_0 = \begin{bmatrix} P_0^{-1} \\ 0 \end{bmatrix} \begin{bmatrix} P_0 \\ \ast \end{bmatrix},$$

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

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

so

$$E(gg') = \sigma_z^2 \begin{bmatrix} Q \\ Q_0 \end{bmatrix}^{-1} \begin{bmatrix} Q \\ Q_0 \end{bmatrix}^{-1}.$$ 

\[ (Q'Q)E(gg')H' + \sigma_z^2 H' = (Q_0'Q_0)E(gg')H', \]

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

Notice next from

$$\begin{bmatrix} y \\ T \times 1 \end{bmatrix} = H \begin{bmatrix} g \\ T \times 1 \end{bmatrix} + c$$

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

$$\tilde{A} = E(gy')E(yy')^{-1} = E(gg')H'[HE(gg')H' + \sigma_z^2 I_T]^{-1},$$

(A7)

Combining equations (2) and (A7),

$$(H'H + \lambda Q'Q)(A' - \tilde{A})HE(gg')H' + \sigma_z^2 I_T$$

$$= H'[HE(gg')H' + \sigma_z^2 I_T] - (H'H + \lambda Q'Q)E(gg')H',$$

(A8)

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

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

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

From the symmetry of $F(z)$ in $z$ and $z^{-1}$, we can without loss of generality normalize $\theta_1$ and $\theta_2$ to be inside the unit circle and write

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

Define $(1 - \phi_1 z - \phi_2 z^2) = (1 - \theta_1 z)(1 - \theta_2 z)$, namely, $\phi_1$ is the real number $\theta_1 + \theta_2$ and $\phi_2$ is the negative real number $-\theta_1 \theta_2$. Note also that the roots of $(1 - \phi_1 z - \phi_2 z^2) = 0$ are the complex conjugates $\theta_1^{-1}$ and $\theta_2^{-1}$, which are both outside the unit circle. This gives the bounds on $\phi_1$ and $\phi_2$ stated in proposition 2 as in Hamilton (1994). Then

$$F(z) = \frac{\lambda}{\phi_1}(1 - \phi_1 z - \phi_2 z^2)(1 - \phi_1 z^{-1} - \phi_2 z^{-2}).$$

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

$$F(1) = 1 = (1 - \phi_1 - \phi_2)(1 - \phi_1),$$

(A9)

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

$$F(-1) = 1 + 16 \lambda = (1 - \phi_1 - \phi_2)(1 - \phi_1).$$

(A10)

Taking the difference between these last two equations establishes $(4\phi_1 - 4\phi_2 \lambda)/(\phi_2 - \phi_1) = 16\lambda$, 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

$$\frac{1}{(1 - \phi_1 z - \phi_2 z^2)(1 - \phi_1 z^{-1} - \phi_2 z^{-2})}$$

$$= \frac{C_0 + C_1 z}{1 - \phi_1 z - \phi_2 z^2} + \frac{C_0 + C_1 z^{-1}}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} + B_0.$$  \hfill (A12)
Combining terms over a common denominator shows that equation (A12) will hold provided
\[
1 = (C_0 + C_1 z)(1 - \phi_1 z^{-1} - \phi_2 z^{-2}) \\
+ (C_0 + C_1 z)^{-1}(1 - \phi_3 z^{-1} - \phi_4 z^{-2}) \\
+ B_0(1 - \phi_1 z^{-1} - \phi_2 z^{-2}) \\
= [2C_0 - 2C_1 z + B_0(1 + \phi_1^2 + \phi_2^2)] \\
+ [C_0 - C_1 \phi_1 - C_1 \phi_2 - B_0 \phi_1 + B_0 \phi_2](z + z^{-1}) \\
- [C_0 \phi_2 + B_0 \phi_2(2z^2 + z^{-2})].
\]
The coefficient on \(z^2 + z^{-2}\) will be 0 provided \(B_0 = -C_0\) or provided
\[
1 = [2C_0 - 2C_1 \phi_1 - C_0 \phi_1^2 + C_0 \phi_2^2] + [C_1 - C_1 \phi_2 - C_0 \phi_1 \phi_2](z + z^{-1}).
\]
(A13)

The coefficient on \((z + z^{-1})\) will be 0 provided
\[
C_1 = \frac{C_0 \phi_1 \phi_2}{1 - \phi_4^2} = -C_0 \phi_1^2/4
\]
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
\[
1 = C_0(1 - \phi_1^2 - \phi_2^2)/2.
\]
Combining these results, we conclude that
\[
\frac{1}{(1 - \phi_1 z^{-1} - \phi_2 z^{-2})(1 - \phi_1 z^{-1} - \phi_2 z^{-2})} \\
= C_0 \frac{1 - (\phi_1^2/4)z^{-1}}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} \frac{1 - (\phi_2^2/4)z^{-1}}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} - 1.
\]
From equation (A9), we then obtain equation (9) with \(C = -C_0 \phi_4/\lambda\) as claimed in equation (13).

To derive equation (10), recall from Hamilton (1994) that
\[
\frac{1}{1 - \phi_1 z^{-1} - \phi_2 z^{-2}} = \sum_{j=1}^{\infty} R[j(2\alpha \cos(\mu j) + 2\beta \sin(\mu j)]z^j.
\]
(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 \(a = 1/2\). We likewise know that the coefficient of \(z^j\) for \(j = 1\) is given by \(\phi_1\) so \(R[\cos(0) + 2\beta \sin(0)] = \phi_1\), which, from equation (14), gives \(R[2\beta \sin(0)] = \phi_1/2 \phi_2 \phi_4 \sin(0) = \phi_1 \cos(0) = 2\beta = \cos(0)\). Substituting these values for \(a\) and \(b\) into equation (A15) gives equation (10).

**Proof of Proposition 3.** Recall the identity
\[
y_{i+h} = y_i + \sum_{j=1}^{h} \Delta y_{i+j},
\]
(A16)

which immediately gives the result of proposition 3 for the case \(d = 1\) as stated in equation (16). We likewise have the identity
\[
\Delta y_{i+j} = \Delta y_i + \sum_{j=1}^{h-1} \Delta^2 y_{i+j}.
\]
(A17)

Substituting equation (A17) into (A16) gives
\[
y_{i+h} = y_i + \Delta y_i \sum_{j=1}^{h} 1 + w_{ij}^{(h)},
\]
for \(w_{ij}^{(h)} = \sum_{j=1}^{h} \Delta^2 y_{i+j}\) as claimed in equation (17) for the case \(d = 2\). We can proceed recursively using the identity \(\Delta^d y_{i+h} = \Delta^d y_i + \sum_{j=1}^{h} \Delta^{d-1} y_{i+j}\) and substituting into the preceding expression. For any \(d\), the resulting \(w_{ij}^{(h)}\) is a finite sum of stationary variables and therefore 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 \(y_{i+h} = \tilde{\beta} \alpha + v_{i+h}\) for
\[
\tilde{\beta} = (\tilde{u}_i, \tilde{u}_{i-1}, \ldots, \tilde{u}_{i-p+d+1}, 1, \Delta^{d-1} y_{i}, \Delta^{d-2} y_{i}, \ldots, \Delta y_{i}, y_i)^T
\]
with \(\tilde{u}_i = \Delta^d y_i - \mu\). The latter regression is feasible because we do not know the true values of \(\mu\) and \(\Delta\). 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 \(d = 2, p = 4,\)
\[
\hat{\beta} = \begin{bmatrix} 1 & -2 & 0 & \mu \\ 0 & 1 & -2 & 1 & \mu \end{bmatrix} \begin{bmatrix} y_i \\ y_{i-1} \\ y_{i-2} \\ y_{i-3} \\ 1 \end{bmatrix} = H \tilde{\beta}
\]
(A18)

and \(\tilde{\beta} = (\sum H x_i H')^{-1} (\sum H x_i y_{i+h})\) so \(\hat{\beta} = H \tilde{\beta}\) for every sample. When \(p = d\), we define the \((p + 1) \times 1\) vector as \(\hat{\beta}_i = (1, \Delta^{d-1} y_{i}, \Delta^{d-2} y_{i}, \ldots, \Delta y_{i}, y_i)^T\), that is, none of the \(\tilde{u}_{i-j}\) variables appear in \(\hat{\beta}\) when \(p = d\).

Define \(q\) to be the \((p + 1) \times 1\) vector \(q = (0, 0, 0, E(w_i^{(h)}), k_{b,0}, k_{b,0}^{d-1}, \ldots, k_{b,h})\), so that \(\tilde{u}_i^{(h)} = \hat{w}_i^{(h)} - E(w_i^{(h)}) = y_{i+h} - \tilde{y}_i q\) and
\[
\hat{\beta} = \sum \tilde{y}_i q^{(h)} \sum \tilde{y}_i q^{(h)} = q + \sum \tilde{y}_i q^{(h)}
\]
(A19)

We first consider the case when equation (18) holds for \(\Delta^d y_i\) 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 \(\Delta^{d-1} y_i = \xi_i^{(h)}\) where \(\xi_i^{(h)} = \sum_{j=1}^{h} \tilde{u}_j \) and \(\xi_i^{(h)} = \sum_{j=1}^{h} \tilde{u}_j^{(h)}\) for \(s = 2, 3, \ldots, d\). For this case, define
\[
Y_T = \begin{bmatrix} T^{1/2} (\tilde{u}_{i-p+d+1} & 0 & 0 & \cdots & 0) \\ 0 & T & 0 & \cdots & 0 \\ 0 & 0 & T & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & T^d \end{bmatrix}
\]
(A20)

Adapting the approach in Sims, Stock, and Watson (1990), we have from equation (A19) that
\[
T^{-1/2} Y_T (\hat{\beta} - q) = T^{-1/2} Y_T (\sum \tilde{y}_i q^{(h)} \sum \tilde{y}_i q^{(h)} \sum \tilde{y}_i q^{(h)} \sum \tilde{y}_i q^{(h)})
\]
(A21)

Consider first the last term in equation (A21):
\[
T^{-1/2} Y_T (\sum \tilde{y}_i q^{(h)} \sum \tilde{y}_i q^{(h)})
\]
(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 \(E(\tilde{u}_{i-p+d+1})\). Term \(p-d+1\) likewise converges to \(E(\tilde{u}_{i-p+d+1}^{(h)}) = 0\). 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 \( \mathbf{Y}_d^{-1} \sum \hat{\mathbf{x}} \mathbf{Y}_d^{-1} \) is characterized by

\[
\begin{bmatrix}
T^{-1} \sum \hat{a}_t^\omega & \cdots & T^{-1} \sum \hat{a}_{t-p+d+1} \\
\vdots & \ddots & \vdots \\
T^{-1} \sum \hat{a}_{t-p+d+1} \hat{a}_t & \cdots & T^{-1} \sum \hat{a}_{t-p+d+1} \hat{a}_{t-p+d+1}
\end{bmatrix}
\]

Taking again the \( p = 4 \) and \( d = 2 \) example, equation (A18), the coefficients \( \hat{\beta} \) from the actual regression of \( y_{t+1,0} \) on \( (y_{t+4,1}, y_{t+3,1}, y_{t+2,1}, y_{t+1,1}) \) have plim

\[
\hat{\beta} \xrightarrow{p} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ -2 & 1 & 0 & -1 & 0 \\ 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\mu & -\mu & 1 & 0 & 0 \end{bmatrix}
\]

\[
\begin{bmatrix} g_1 \\ s_2 \\ \mu h (h + 1)/2 \\ h \\ 1 \end{bmatrix}
\]

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

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 \( \Delta^d y_t = \sum_{j=1}^d \hat{a}_t + \mu t = \hat{\xi}^d_t + \mu t \), which is dominated for large \( t \) by the drift term \( \mu t \) rather than the random walk term \( \hat{\xi}^d_t \), and \( \Delta^{d-1} y_t = \hat{\xi}^{d-1}_t + (1/\mu) t \mu + O_p(t) \). In this case, we would simply replace \( \mathbf{Y}_d \) in the above derivations with

\[
\tilde{\mathbf{Y}}_d = \begin{bmatrix}
T^{-1/2} I_{d-1} & 0 & \cdots & 0 \\
0 & T^{-1/2} I_{d-2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & T^{-1/2} I_{d-1} \\
0 & 0 & \cdots & 0
\end{bmatrix}
\]

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 \( \hat{\xi}^d_t \) with \( \hat{\xi}^d_t + \hat{\xi}^{d+1}_t \), any fixed constant produces a term that is still dominated asymptotically by \( \hat{\xi}^d_t \), 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, \( y_t = \sum_{j=0}^d \beta_j t + \epsilon_t \). For this case, we have \( \mu = E(\Delta^d y_t) = \delta_d \) and

\[
\hat{\beta} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ -2 & 1 & 0 & -1 & 0 \\ 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\mu & -\mu & 1 & 0 & 0 \end{bmatrix}
\]

\[
\begin{bmatrix} g_1 \\ s_2 \\ \mu h (h + 1)/2 \\ h \\ 1 \end{bmatrix}
\]

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


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