Data Processing and Filters

Seth Leonard
seth@ottoquant.com

OttoQuant
www.ottoquant.com

May 20, 2020

The first stage of modeling time series data typically involves processing each input series individually as needed. Even if we will be using tools such as dynamic factor models, which accept mixed frequency data, noisy data, and data with missing observations, there are still series-specific issues which need addressing first. The two main concerns are stationary and seasonality. We may also wish to standardize the variance of the data. Non-stationarity due to growth (as opposed to regime switching, which we will deal with separately) can be dealt with via de-trending or differencing. De-trending often preserves some of the information contained in the levels of the data. However, it almost always requires observations both before and after the current date t. For this reason, it is typically unsuitable for long horizon forecasting. Differencing, on the other hand, allows us to convey the real time information in the data without any lag.

1 Frequency Domain in Brief

Most of our analysis will deal with realizations of observables Y_t relative to one or more observations of Y_τ at other points in time. That is, we will be dealing primarily with time domain techniques. Frequency domain provides an alternative approach to analyzing time series data as a weighted sum of period functions.

Following the notation in Hamilton (1994):

$$y_t = \mu + \int_0^{\pi} \alpha(\omega)\cos(\omega t)d\omega + \int_0^{\pi} \delta(\omega)\sin(\omega t)d\omega \tag{1}$$

1.1 The Population Spectrum

Our interest lies in analyzing the series y_t described as a periodic function such as that in (1). To this end, we can use Euler's formula to write (1) as

$$y_t = \frac{1}{2}\int_0^{\pi} e^{i\omega t}\big(\alpha(\omega) - i\delta(\omega)\big)d\omega + \frac{1}{2}\int_0^{\pi} e^{-i\omega t}\big(\alpha(\omega) + i\delta(\omega)\big)d\omega$$

By defining a new function $Z(\omega)$ we can combine the above integrals in one integral from $-\pi$ to π so that

$$y_t = \int_{-\pi}^{\pi} e^{i\omega t} Z(\omega)d\omega$$

where for $\omega >= 0$, $Z(\omega) = \alpha(\omega) - i\delta(\omega)$, and for $\omega < 0$, $Z(\omega) = \alpha(-\omega) + i\delta(-\omega)$. The complex conjugate of $Z(\omega)$ is then

$$\overline{Z(\omega)} = \alpha(\omega) + i\delta(\omega), \quad \omega >= 0$$

$$\overline{Z(\omega)} = \alpha(-\omega) - i\delta(-\omega), \quad \omega < 0$$

Using the fact that $E(x, x) = E(x, \bar{x})$, we can find the variance of y_t as

$$\begin{aligned} \gamma_k &= E(y_t y_{t-k}) \\ &= E(y_t \bar{y}_{t-k}) \\ &= \textstyle\int_{-\pi}^{\pi} e^{i\omega t} Z(\omega)d\omega \int_{-\pi}^{\pi} e^{-i\omega t}\overline{Z(\omega)}d\omega \\ &= \textstyle\int_{-\pi}^{\pi} e^{i\omega k} E\big(Z(\omega)\overline{Z(\omega)}\big)d\omega \end{aligned} \tag{2}$$

Defining $s(\omega) = E\big(Z(\omega)\overline{Z(\omega)}$ we then have

$$\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} s(\omega)d\omega$$

Using Euler's formula again, we can write $e^{i\omega k} = \cos(\omega k)$ so that

$$\gamma_k = 2\int_0^{\pi} \cos(\omega k)s(\omega)d\omega \tag{3}$$

Of course, we do not yet know what $s(\omega)$ is. One could invert (3). The solution is

$$s(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-i\omega k} \gamma_k \tag{4}$$

This is the population spectrum. To check this solution, we can plug it into (3) above. Thus we have

$$\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} \left(\frac{1}{2\pi} \sum_{j=-\infty}^{\infty} e^{-i\omega j} \gamma_j \right) d\omega$$

Isolating the terms that contain ω we have

$$\gamma_k = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j \int_{-\pi}^{\pi} e^{i\omega(k-j)} d\omega = \gamma_k$$

The last equality come from the fact that for $k = j$ the integral

$$\int_{-\pi}^{\pi} e^{i\omega(k-j)} d\omega = 1$$

and for $k \neq j$ the integral

$$\int_{-\pi}^{\pi} e^{i\omega(k-j)} d\omega = 0$$

Using Euler's formula, and noting that for a covariance-stationary process $\gamma_j = -\gamma_j$, we can simplify the result for $s(\omega)$ as follows:

$$\begin{aligned} s_y(\omega) &= \tfrac{1}{2\pi}\omega_0(\cos(0) - i\sin(0)) \\ &+ \tfrac{1}{2\pi}\Big(\textstyle\sum_{j=1}^{\infty} \gamma_j(\cos(\omega j) + \cos(-\omega j) - i\sin(\omega j) - i\sin(-\omega j)) \Big) \\ &= \tfrac{1}{2\pi}\Big(\gamma_0 + 2\textstyle\sum_{j=1}^{\infty} \gamma_j \cos(\omega j) \Big) \end{aligned}$$

As an example of working with equation (3), suppose only the first three autocovariances $\gamma_0, \gamma_1, \gamma_2$ are non-zero:

$$s_y(\omega) = \frac{1}{2\pi}\Big(\gamma_0 + 2\big(\gamma_1 \cos(\omega) + \gamma_2 \cos(2\omega)\big) \Big)$$

Letting $k = 0$ we have

$$\begin{aligned} \textstyle\int_{-\pi}^{\pi} s_y(\omega)\cos(0) d\omega &= \textstyle\int_{-\pi}^{\pi} \frac{1}{2\pi}\Big(\gamma_0 + 2\big(\gamma_1 \cos(\omega) + \gamma_2 \cos(2\omega)\big) \Big) d\omega \\ &= \tfrac{1}{2\pi}\gamma_0\omega \Big|_{-\pi}^{\pi} + 2\big(\gamma_1 \sin(\omega) + \gamma_2 \sin(2\omega)\big) \Big|_{-\pi}^{\pi} \\ &= \tfrac{2\pi}{2\pi}\gamma_0 \end{aligned}$$

For $k = 1$ we will need to use the identities $\cos^2(x) = (\cos(2x) + 1)/2$ and $\cos(x)\cos(y) = \big(\cos(x-y) + \cos(x+y)\big)/2$. Then

$$\int_{-\pi}^{\pi} s_y(\omega)\cos(\omega)d\omega \;=\; \int_{-\pi}^{\pi} \tfrac{1}{2\pi}\Big(\gamma_0 + 2\big(\gamma_1\cos(\omega) + \gamma_2\cos(2\omega)\big)\Big)\cos(\omega))d\omega$$

The first term will be

$$\int_{-\pi}^{\pi} \frac{1}{2\pi}\gamma_0\cos(\omega) = \frac{1}{2\pi}\gamma_0\sin(\omega)\Big|_{-\pi}^{\pi} = 0$$

Dealing with the second term,

$$\begin{aligned}
&= \int_{-\pi}^{\pi} \tfrac{1}{\pi}\big(\gamma_1\cos^2(\omega) + \gamma_2\cos(2\omega)\cos(\omega)\big)d\omega \\
&= \int_{-\pi}^{\pi} \tfrac{1}{\pi}\big(\gamma_1(\cos(2\omega)+1) + \gamma_2\cos(\omega) + \gamma_2\cos(3\omega)\big)d\omega \\
&= \tfrac{1}{\pi}\Big(\gamma_1(\tfrac{1}{2}\sin(2\omega)+\omega)\Big|_{-\pi}^{\pi} + \gamma_2\sin(\omega)\Big|_{-\pi}^{\pi} + \gamma_2\tfrac{1}{3}\sin(3\omega)\Big|_{-\pi}^{\pi}\Big) \\
&= \gamma_1
\end{aligned}$$

Recalling that $s_y(\omega)$ is symmetric, the above is equivalent to

$$\gamma_k = 2\int_0^{\pi} s_y(\omega)\cos(\omega k)d\omega$$

1.2 Relationship to the Autocovariance Generating Function

Recall that for the MA(1) process

$$y_t = \mu + \varepsilon_t + b\varepsilon_{t-1} \tag{5}$$

the variance is $(1+b^2)\sigma^2$, first autocovariance was $b\sigma^2$ and all higher autocovariances were zero. For the MA(2) process

$$y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2}$$

we have, letting γ_i denote the i^{th} covariance

$$\begin{aligned}
\gamma_0 &= (1 + b_1^2 + b_2^2)\sigma^2 \\
\gamma_1 &= (b_1 + b_1 b_2)\sigma^2 \\
\gamma_2 &= b_2\sigma^2
\end{aligned}$$

More generally, for the MA(q) function

$$y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \ldots$$

the autocovariances are

$$\begin{aligned} \gamma_0 &= (1 + b_1^2 + b_2^2 + \ldots + b_q^2)\sigma^2 \\ \gamma_1 &= (b_1 + b_1 b_2 + \ldots + b_{q-1}b_q)\sigma^2 \\ &\vdots \\ \gamma_q &= b_q\sigma^2 \end{aligned}$$

From this result it will be convenient to define an autocovariance generating function (again following Hamilton (1994))

$$g_y(z) = \sum_{-\infty}^{\infty} \gamma_j z^j \tag{6}$$

For our MA(1) process this function is

$$\begin{aligned} g_y(z) &= b\sigma^2 z^{-1} + \big((1 + b^2)\sigma^2\big)z^0 + b\sigma^2 z^1 \\ &= \sigma^2(1 + bz)(1 + bz^{-1}) \end{aligned}$$

For the more general MA(q) case we can write the autocovariance generating function as

$$g_y(z) = \sigma^2(1 + b_1 z + b_2 z^2 + \ldots + b_q z^q)(1 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_q z^{-q})$$

This function is not particularly useful as it stands. However, by letting

$$z = \cos(\omega) - i\sin(\omega) = e^{-i\omega}$$

where $i = \sqrt{-1}$ and the second equality is Euler's formula, and defining

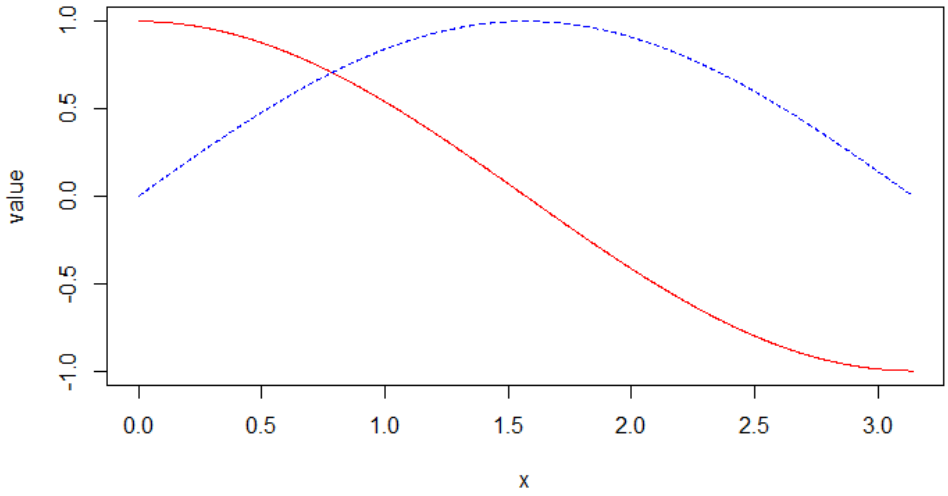
$$s_y(\omega) = \frac{1}{2\pi} g(e^{-i\omega}) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} \gamma_j e^{-i\omega j} \tag{7}$$

we have the population spectrum of y.

1.3 Analysis via the Spectrum

The above results are interesting, but not yet very useful. Recall the original goal of this exercise is to explain the process y_t as a periodic function. This raises a

few questions. First, why the form of equation (1)? That is, why a function of $\cos(\omega(t))$ and $\sin(\omega(t))$ over 0 to π? The reason we look at the function from 0 to π is that this describes one half of a full cycle, as illustrated below. Because cycles are symmetric, 0 to π is all we need; π to 2π will be identical, just with the opposite sign.



sin function, blue dash, and cos function, red solid

Moreover, we need both the sin and cos function as we would be unable to describe y_t with just one; for example the value of $\cos(\pi/2)$ is zero. We could, of course, have done the same over the interval $[-\pi/2, \pi/2]$, it simply convention to use $[0, \pi]$.

The results derived in the previous section were for the population y_t. Suppose instead we simply had nine observations of y_t, and wanted to explain our observations in terms of evenly spaced points along $[0, \pi]$, which we will call ω_j. To ensure the ω_j's are evenly distributed, their values will be

$$\begin{aligned} \omega_1 &= \frac{2\pi}{T} \\ \omega_2 &= \frac{4\pi}{T} \\ \omega_3 &= \frac{6\pi}{T} \\ \omega_4 &= \frac{8\pi}{T} \end{aligned}$$

where $T = 9$. The discrete version of equation (1) is

$$y_t = \mu + \sum_{j=1}^{J} \left(\alpha_j \cos(\omega_j(t-1)) + \delta_j \sin(\omega_j(t-1))\right) + u_t \tag{8}$$

Since we have nine observations and nine parameters (μ, α_j, and δ_j for $j \in [1, 4]$), we can fit equation (8) perfectly. We can then write the sample variance as

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_t) = \frac{1}{2}\sum_{j=1}^{J}(\alpha_j^2 + \delta_j^2)$$

where the part of the variance due to cycles at frequency ω_j is

$$\frac{1}{2}(\alpha_j^2 + \delta_j^2)$$

From the previous results, this will be equivalent to

$$\frac{4\pi}{T}\hat{s}_y(\omega_j)$$

where $\hat{s(\omega)}$ is our estimate of $s(\omega)$, called the sample periodogram.

Of course in practice, we would not want to estimate the parameters of (8) this way, but use a much larger sample of y_t. This can be done via non-parametric or kernal estimates of the autocovariances and applying the results of section 1.2; see Hamilton (1994).

A closer look at equation (8) is helpful in exploring what the spectrum of a process tells us. For a large T, ω_1 will be close to zero. Thus the terms $\omega_1(t-1)$ will grow slowly as t becomes larger. Put differently, α_1 and δ_1 measure low frequency movements in y — $\omega_1(t-1)$ won't reach 2π, a full cycle, until the very end of the sample (obviously the term never quite reaches 2π). At the other end of the spectrum, $\omega_T(t-1)$ cycles through π nearly every period t (and 2π nearly every other period); terms on ω_T therefor describe high frequency movements of y_t.

As an example, we could calculate the spectrum of MA(2) process

$$y_t = \varepsilon + 2\varepsilon_{t-1} + \varepsilon_{t-2}$$

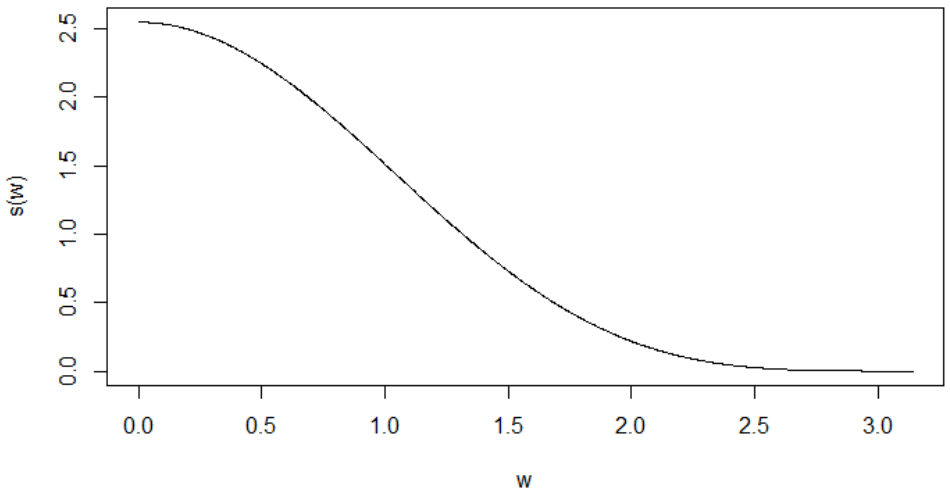
as follows. We have

$$\begin{aligned} \gamma_0 &= (1+4+1)\sigma^2 \\ \gamma_1 &= (2+2)\sigma^2 \\ \gamma_2 &= 1\sigma^2 \end{aligned}$$

thus

$$s(\omega) = \frac{1}{2\pi}\big(5 + 2(4 * \cos(\omega) + \cos(2\omega))\big)$$

The following figure plots this function for $\sigma = 1$:



Spectrum for a simple MA(2) process

Note that from our earlier results in (3), the area under the curve sums to the total (unconditional, or γ_0) variance of the process for y_t. Thus the variance of y_t is due primarily to cycles at low frequency, with the contribution from high frequencies approaching zero.

2 Univariate Filters and Trend Estimation

The discussion of spectral analysis above is a method for evaluating univariate data and univariate filters. If for example, we found a high proportion of variance in a series due to cycles at an annual frequency, we may wish to remove this seasonality before performing further analysis. Alternatively, we may wish to remove low frequency volatility in the data, such as a trend, to isolate seasonal or business cycle components. Keep in mind, however, that de-trending in this way to enforce stationarity typically depends on both past and future values, and is

thus inappropriate for forecasting (we will normally difference data instead). The following sections discuss a number of ways to accomplish these aims by filtering data. Discussion of the last method, Kalman filtering, is particularly detailed as the multivariate Kalman filter will become one of the main tools we use for analysis.

2.1 HP Filter

Suppose we can decompose our observations for a series y_t as

$$y_t = \tau_t + c_t + \varepsilon_t \tag{9}$$

where τ_t is a trend, c_t cyclical components of the data, and ε_t idiosyncratic components. The Hodrick-Prescott filter optimizes the loss function

$$\min_{\tau} \left\{ \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \Big((\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \Big)^2 \right\} \tag{10}$$

The first term penalizes deviations in τ from the observed series y_t. The second term penalizes volatility in τ, and λ determines the magnitude of this penalty on volatility. Thus, as λ increases the estimated series for τ becomes smoother. The filter has the advantage of a simple, straightforward interpretation. However, James D. Hamilton, author of Hamilton (1994) (one of the main references for these notes) advises against using the HP filter for several reasons, including a tendency to produce spurious correlations and the existence of better alternatives.

2.2 Bandpass Filter

A bandpass filter is a filter that attempts to preserve variation in a series y_t for a given frequency, and remove variance from cycles at other frequencies. A low pass filter preserves variance from low frequency cycles (such as trends), which a high pass filter preserves variance from high frequency cycles.

Suppose we have a filter

$$h(L)y_t = (1 - L - L^2 - \ldots)y_t \tag{11}$$

where L is the lag operator. For example, we would write the first difference of y_t as

$$x_t = (1 - L)y_t$$

If y_t has the autocovariance generating function $g_y(z)$, then the autocovariance generating function for x_t will be

$$g_x(z) = h(z)h(z^{-1})g_y(z) \tag{12}$$

This is useful as we already know from section 1 that we can write the spectrum of a process as

$$s(\omega) = \frac{1}{2\pi} g(e^{-iw})$$

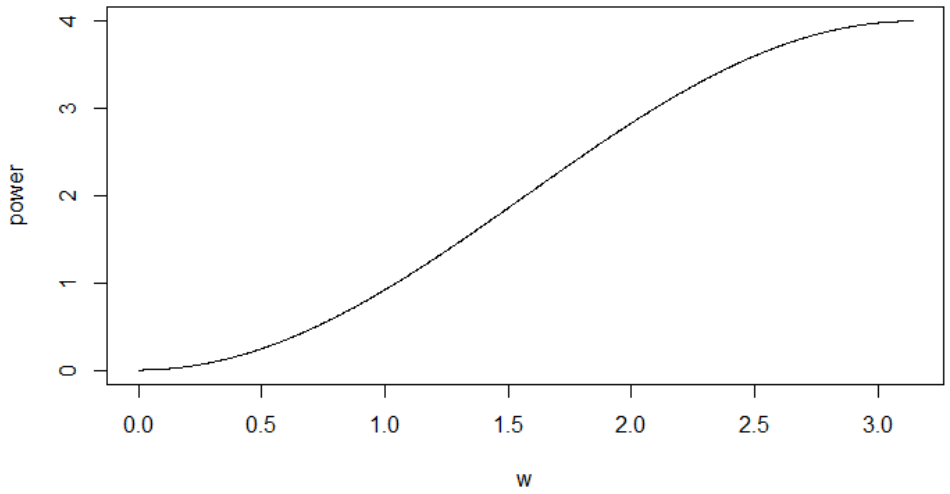
Thus the spectrum of the new process is

$$s_x(\omega) = h(e^{-iw})h(e^{iw})s_y(\omega)$$

For the example of differencing the data

$$\begin{aligned} h(e^{-i\omega}e^{i\omega}) &= (1 - e^{-i\omega})(1 - e^{i\omega}) \\ &= 1 - e^{-i\omega} - e^{i\omega} + 1 \\ &= 2 - 2\cos(\omega) \\ &= 2(1 - \cos(\omega)) \end{aligned}$$

The following figure plots this result over $[0, \pi]$.

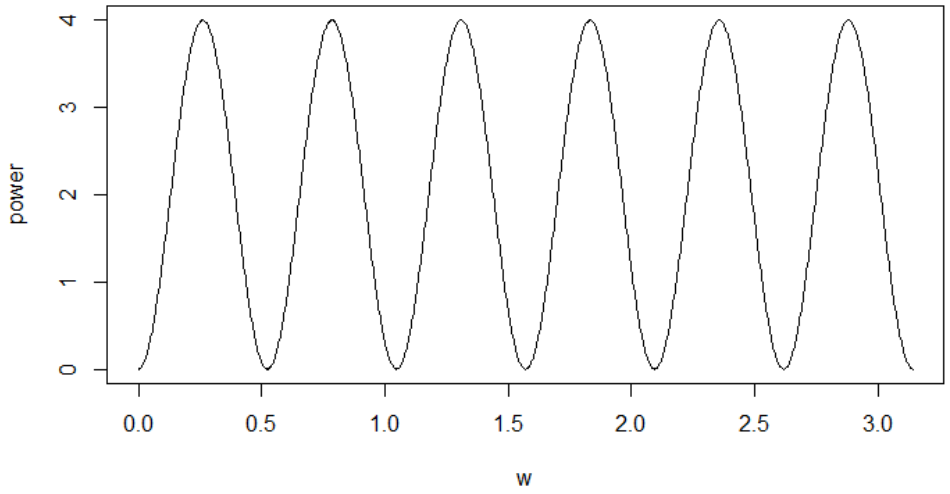


Transformation due to taking first differences

The is an example of a high-pass filter: low frequency cycles are removed and high frequency cycles are amplified. As a second example we can look at taking year on year differences. In this case we have (for monthly data) $h(L) = 1 - L^{12}$

$$\begin{aligned} h(e^{-i\omega}e^{i\omega}) &= (1 - e^{-12i\omega})(1 - e^{12i\omega}) \\ &= 1 - e^{-12i\omega} - e^{12i\omega} + 1 \\ &= 2 - 2\cos(12\omega) \\ &= 2(1 - \cos(12\omega)) \end{aligned}$$

The following figure plots this result over $[0, \pi]$.



Transformation due to taking year on year differences

Thus the year on year filter removes not only low frequencies and annual (12 month), but also frequencies at 6, 4, 2, 2.4, and 2 months!

2.3 L1-Norm Filters

The loss functions we typically use in econometrics minimize mean squared error. For example, we can derive the OLS estimator by minimizing

$$l = \sum_{i=1}^{N} (y_i - \beta x_i)^2$$

This loss function is l_2, 2 because the loss function is squared. There are times, however, when we may wish to minimize an absolute value, not a squared value. This represents l_1 loss. For example, lasso models are a popular means of shrinking parameter estimates to avoid problems of overparameterization (results can be similar to using a zero prior for Bayesian models, which we will come to later). Penalties on parameters can be l_2, that is

$$\lambda \sum_{i=1}^{k} \beta_i^2$$

which will shrink all parameter estimates towards zero, or l_1

$$\lambda \sum_{i=1}^{k} |\beta_i|$$

which will select parameters by setting some to zero and leaving others unchanged, or a mix of the two.

We can do something similar with the HP filter, using an l_1 loss instead of the standard l_2 loss. That is, equation (10) becomes

$$\min_{\tau} \left\{ \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \Big| (\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \Big| \right\} \tag{13}$$

The resulting trend will be peicewise linear. There are a few reasons this may be desirable. First, it may be convenient to have a locally linear trend. Second, shifts in a linear trend provide a nice corollary to the idea of regime shifts, which we will come to later.

2.4 Kalman Filtering

Kalman filtering and smoothing will become our workhorse method for later models, in particular for dynamic factor models (DFMs) and for variations of standard models, like VARs, written to incorporate missing, noisy, or mixed frequency data. For that reason we will cover the univariate filter in some detail here.

Kalman filtering, named for Kalman (1960), is a means of estimating the time series model

$$y_t = Hx_t + \varepsilon_t \tag{14}$$

$$x_t = Ax_{t-1} + e_t \tag{15}$$

where x_t is an unobserved state or states, y_t observed outcomes, and ε_t and e_t are normally distributed error terms with the covariance matrix $cov \begin{bmatrix} e_t \\ \varepsilon_t \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$. States, observed outcomes, and error terms can be either scalars or vectors. The Kalman filter works by first predicting an outcome $x_{t|t-1}$ where the subscripts indicate the prediction of x_t based on all observations from the initial period until period $t-1$ and then updating this prediction using the outcome from period t. Formally, we first predict

$$p(x_t|y_{1:t-1}) = \int p(x_t,\ x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \tag{16}$$

and then update this prediction based on the current observation y_t

$$p(x_t|y_t) \propto p(y_t|x_t)p(x_t|y_{1:t-1}) \tag{17}$$

In this way equation 17 is a Bayesian estimate of our unobserved states x_t using equation 16 as our prior. Before writing down what the prediction and updating equations will in fact be when our model follows 15 and 14, it's worth looking at a few features of the multivariate normal distribution.

2.4.1 Preliminaries

Suppose we know that the scalars x and y follow the multivariate normal distribution

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ m \end{bmatrix}, \Sigma \right) \tag{18}$$

where $\Sigma = \begin{bmatrix} P & C \\ C & S \end{bmatrix}$ and we're interested in determining the distribution $f(x|y)$. Using the definition of a conditional distribution we know that

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

or, since $f(y)$ is a normalizing constant,

$$f(x|y) \propto |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left(\begin{bmatrix} x-\mu \\ y-m \end{bmatrix}' \Sigma^{-1} \begin{bmatrix} x-\mu \\ y-m \end{bmatrix} \right) \right\}$$

The inverse of the covariance matrix is $\Sigma^{-1} = \frac{1}{PS-C^2}\begin{bmatrix} S & -C \\ -C & P \end{bmatrix}$ thus the exponential terms are

$$-\frac{1}{2(PS-C^2)}[(x-\mu)^2 S + (y-m)^2 P - 2(x-\mu)(y-m)C]$$

We can re-write this expression as

$$-\frac{1}{2(PS-C^2)}[(y-m)^2 P + (x-\mu-(y-m)CS^{-1})S(x-\mu-(y-m)CS^{-1}) - (y-m)^2 C^2 S^{-1}]$$

or, dumping the terms which don't contain our parameter of interest x into the normalizing constant,

$$-\frac{1}{2}[(x-(\mu+(y-m)CS^{-1}))\tilde{P}^{-1}(x-(\mu+(y-m)CS^{-1}))] \tag{19}$$

where $\tilde{P} = P - C^2 S^{-1}$. $f(x|y)$ is therefore normally distributed with mean $(\mu + (y-m)CS^{-1})$ and variance $\tilde{P} = P - C^2 S^{-1}$. This result generalizes to the case in which x and y are vectors with covariance matrix $\Sigma = \begin{bmatrix} P & C \\ C' & S \end{bmatrix}$ as

$$E(x|y) = \mu + CS^{-1}(y-m)$$

and

$$Var(x|y) = P - CS^{-1}C'$$

These results are essentially all we need to derive the Kalman filter.

2.4.2 The Kalman Filter

To use the results from section 2.4.1 requires two elements. The first is our model, equations (15) and (14). The second is the distribution for $\begin{bmatrix} x_{t|t-1} \\ y_t \end{bmatrix}$; this is not as obvious as it may seem since we never in fact observe x_t (or x_{t-1}). Therefore we need to define a new variable, call it $x_{t|t}$, which is our predicted value of x_t given observations $y_{1:t}$. Define the variance $var(x_{t|t}) = E_{t|t}(x_t - x_{t|t})(x_t - x_{t|t})'$ as $P_{t|t}$ and the variance $var(x_{t|t-1})$ (x_t given observations $1:t-1$) as $P_{t|t-1}$. Then the joint distribution for $x_{t|t-1}$ and y_t is

$$\begin{bmatrix} x_{t|t-1} \\ y_t \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} Ax_{t-1|t-1} \\ H(Ax_{t-1|t-1}) \end{bmatrix}, \Sigma_t\right)$$

where

$$\Sigma_t = \begin{bmatrix} P_{t|t-1} & C_t \\ C_t' & S_t \end{bmatrix}$$

From the results in section 2.4.1 we can immediately calculate the expected value of $x_{t|t-1}$ (which is unobserved) given y_t (which is observed), $E(x_{t|t-1}|y_t) = Ax_{t-1|t-1} + C_t S_t^{-1}(y_t - H(Ax_{t-1|t-1}))$, as well as $var(x_{t|t-1}|y_t) = P_{t|t} = P_{t|t-1} - C_t S_t^{-1} C_t'$. However, we still need to derive the values for Σ. From equation (15)

$$\begin{aligned} P_{t|t-1} &= var(x_{t|t-1}) \\ &= var\Big(Ax_{t-1|t-1} + e_t\Big) \\ &= AP_{t-1|t-1}A' + Q \end{aligned}$$

From equation (14) the variance of our predicted values for y_t will be

$$\begin{aligned} S_t &= var(y_t) \\ &= var\Big(Hx_{t|t-1} + \varepsilon_t\Big) \\ &= HP_{t|t-1}H' + R \end{aligned}$$

And finally

$$\begin{aligned} C_t &= cov(x_{t|t-1}, y_t) \\ &= cov\Big(Ax_{t-1|t-1} + e_t, Hx_{t|t-1} + \varepsilon_t\Big) \\ &= cov\Big(Ax_{t-1|t-1} + e_t, H(Ax_{t-1|t-1} + e_t) + \varepsilon_t\Big) \\ &= P_{t|t-1}H' \end{aligned}$$

Thus equipped we can write the Kalman filter as follows. Our prediction for the mean and variance of x_t (without conditioning on y_t) is

$$\begin{aligned} x_{t|t-1} &= Ax_{t-1|t-1} \\ P_{t|t-1} &= AP_{t-1|t-1}A' + Q \end{aligned}$$

Our prediction for the mean and variance of y_t given observations $1 : t-1$ (before y_t is observed), denoted $y_{t|t-1}$, is

$$\begin{aligned} y_{t|t-1} &= Hx_{t|t-1} \\ S_t &= HP_{t|t-1}H' + R \end{aligned}$$

Our prediction for the covariance between x_t (again, without using y_t) and $y_{t|t-1}$ is

$$C_t = P_{t|t-1}H'$$

Note that the Kalman gain combines this covariance and the estimated variance of $y_{t|t-1}$ and is typically written as $K_t = C_t S_t^{-1}$. The above equations, called the prediction step, give us our prior. Our posterior estimates for the mean and variance of x_t given y_t (recall y_t is observed), called the updating step, are

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + C_t S_t^{-1}(y_{t|t} - y_{t|t-1}) \\ P_{t|t} &= P_{t|t-1} - C_t S_t^{-1} C_t' \end{aligned}$$

2.4.3 The Likelihood Function

We can write the likelihood of observing $\begin{Bmatrix} y_1 & y_2 & \dots & y_T \end{Bmatrix}$ as

$$f(y_{1:T}) = f(y_T|y_{1:T-1})f(y_{1:T-1}) = f(y_T|y_{1:T-1})f(y_{T-1}|y_{1:T-2})f(y_{1:T-2}) = \prod_{t=1}^{T} f(y_t|y_{1:t-1})$$

where $f(y_t|y_{1:t-1})$ is normally distributed with mean $y_{t|t-1}$ and variance S_t. Denoting the predictive error calculated by the Kalman filter in each period as $\nu_t = y_{t|t} - y_{t|t-1}$ we can thus write the likelihood of our observables as

$$\mathcal{L} = \prod_{t=1}^{T} (2\pi)^{-k/2} |S_t|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \nu_t' S_t^{-1} \nu_t \right\}$$

so that the log likelihood, which we typically use for any maximization problem, is

$$l = \kappa - \frac{1}{2} \sum_{t=1}^{T} \log(|S_t|) - \frac{1}{2} \sum_{t=1}^{T} \nu_t' S_t^{-1} \nu_t \tag{20}$$

where κ does not contain any parameters of interest and can thus be ignored in the maximization problem. The log likelihood is remarkably easy to calculate in practice as both ν_t and S_t are calculated in each period by the Kalman filter.

2.4.4 A Kalman Smoother

What the Kalman filter of the previous section delivers are estimates of $x_{t|t}$, that is, an estimate of x_t given observations from period 1 through t. However, if the states of the model are autocorrelated then presumably observations realized after period t also contain information about states in period t. The Kalman smoothers provide a means of employing this information so that our final estimate of states

becomes $x_{t|T}$, that is, an estimate of x_t given observations from period 1 through T. There are several approaches to Kalman smoothing; I'll outline the simple and popular Rauch-Tung-Striebel smoother. The process begins by running the Kalman filter and saving the values for $P_{t|t-1}$, $P_{t|t}$, $x_{t|t-1}$, and of course $x_{t|t}$. The key to the smoother is the fact that $f(x_t|x_{t+1}, y_{1:T}) = f(x_t|x_{t+1}, y_{1:t})$, which states that if we know x_{t+1} then further realizations of observables after period t do not add any additional information. We can summarize the relationship between $x_{t|t}$ and $x_{t+1|t}$ as

$$\begin{bmatrix} x_t|y_{1:t} \\ x_{t+1}|y_{1:t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_{t|t} \\ x_{t+1|t} \end{bmatrix}, \begin{bmatrix} P_{t|t} & P_{t|t}A' \\ A'P_{t|t} & P_{t+1|t} \end{bmatrix} \right) \tag{21}$$

where $P_{t+1|t} = AP_{t|t}A' + Q$. Using the same results for a joint normal distribution we used to derive the Kalman filter we then have[1]

$$\begin{array}{rcccl} E(x_t|x_{t+1}, y_{1:t}) & = & x_{t|T} & = & x_{t|t} + g_t(x_{t+1} - x_{t+1|t}) \\ Var(x_t|x_{t+1}, y_{1:t}) & = & P_{t|T} & = & P_{t|t} - g_t P_{t+1|t} g_t' \end{array}$$

where $g_t = P_{t|t}A'P_{t+1|t}^{-1}$. Since we do not in fact observe x_{t+1} we need to slightly modify the above equations. Using the law of iterated expectations for the first

$$E(x_t|y_{1:T}) = E(E(x_t|x_{t+1}, y_{1:t})|y_{1:T}) = x_{t|t} + g_t(x_{t+1|T} - x_{t+1|t}) \tag{22}$$

Using the law of iterated variance for the second

$$\begin{array}{rcl} Var(x_t|y_{1:T}) & = & E(Var(x_t|x_{t+1}, y_{1:t})|y_{1:T}) + Var(E(x_t|x_{t+1}, y_{1:t})|y_{1:T}) \\ & = & P_{t|t} - g_t P_{t+1|t} g_t' + g_t P_{t+1|T} g_t' \\ & = & P_{t|t} - g_t(P_{t+1|t} - P_{t+1|T}) g_t' \end{array} \tag{23}$$

Equations (22) and (23) form our smoother. We begin using our last filtered value for $\mu_{T|T}$ and $P_{T|T}$ and iterate backwards to the first period.

2.4.5 The Steady State Filter

Note that in the above Kalman filter neither $P_{t|t}$, S_t, nor C_t depend on the realization of y_t or the expected values of x_t (they do, however, depend on the number

[1]The result for $P_{t|T}$ comes from the fact that

$$\begin{array}{rcl} Var(x_t|x_{t+1}, y_{1:t}) & = & P_{t|t} - P_{t|t}A'P_{t+1|t}^{-1}AP_{t|t} \\ & = & P_{t|t} - P_{t|t}A'P_{t+1|t}^{-1}P_{t+1|t}P_{t+1|t}^{-1}AP_{t|t} \\ & = & P_{t|t} - g_t P_{t+1|t} g_t' \end{array}$$

of series observed each period). Thus, if our series is covariance stationary, the number of observations remains the same each period, and if we happen to know the long run value of C_t, call it C, and S_t, call it S, we could simplify our Kalman filter as

$$\begin{aligned} \mu_{t|t-1} &= A\mu_{t-1|t-1} \\ m_{t|t-1} &= H\mu_{t|t-1} \\ \mu_{t|t} &= \mu_{t|t-1} + CS^{-1}(y_{t|t} - m_{t|t-1}) \end{aligned} \tag{24}$$

This is the steady state Kalman filter. To obtain these steady state values, we can simply run the system of difference equations that determine the relevant variances and covariances until convergence. This system is

$$\begin{aligned} P_{t|t-1} &= AP_{t-1|t-1}A' + Q \\ S_t &= HP_{t|t-1}H' + R \\ C_t &= P_{t|t-1}H' \\ P_{t|t} &= P_{t|t-1} - C_tS_t^{-1}C_t' \end{aligned} \tag{25}$$

2.4.6 Initial Values

Notice that in section 2.4.2 we said that since we don't know x_t, we can predict it with $x_{t-1|t-1}$, and similarly we use $P_{t-1|t-1}$ to calculate the variance of this prediction. This kicking-the-can-down-the-road approach runs into a problem at our first observation; what do we use for $x_{0|0}$ and $P_{0|0}$. There are numerous approaches to dealing with initial values. The simplest is to follow that outlined by Hamilton (1994) and use the long run mean and its associated variance. The results in the section have assumed our matrix of observables is demeaned, so that the long run value for factors $x_{0|0}$ will be zero. The variance is slightly more involved. Begining with an observation of x_p at some time in the past, the variances $P_{p+i|p}$ will be

$$\begin{aligned} P_{p+1|p} &= Q \\ P_{p+2|p} &= AQA' + Q \\ P_{p+3|p} &= AAQA'A' + AQA' + Q \\ &\vdots \end{aligned}$$

Defining $M_i = P_{p+i|p}$ for notational convenience,

$$\begin{aligned} M_1 &= Q \\ M_2 &= AM_1A' + Q \\ M_3 &= AM_2A' + Q \\ &\vdots \\ M_\infty &= AM_\infty A' + Q \end{aligned}$$

The solution to $M_\infty = AM_\infty A' + Q$ is

$$vec(M_\infty) = \mathbf{A}^{-1} vec(Q) \tag{26}$$

where

$$\mathbf{A} = (I - A \otimes A)$$

and I is an identity matrix of $k \times k$ where k is the number of factors in x_t, and $\otimes$ is the Kronecker product. M_∞ is thus the variance of $x_{0|0}$ that we can use to initialize our filter.

2.4.7 An Example

Suppose we have a model described by

$$x_t = \begin{bmatrix} 1 & -.5 \\ .1 & .7 \end{bmatrix} x_{t-1} + e_t$$

$$y_t = \begin{bmatrix} .5 & 1 \\ -1 & 2 \\ 1 & -1 \\ 1 & -.5 \end{bmatrix} x_t + \varepsilon_t$$

where x_t is a 2×1 vector of unobserved factors, y_t is a 4×1 vector of observed data, $e_t \sim N(0, I_2)$, and $\epsilon_t \sim N(0, I_4)$ and we would like to construct the unobserved factors x_t from the observed data.

In this case we already know the parameters A, Q, H, and R. Figure 1 plots the results for 200 observations.

2.4.8 Missing Observations

Using the example from the previous section, suppose the second series in y_t is not observed in a given period t. We can simply re-write our model for the data we do observe by dropping the rows of H and rows and columns of R corresponding to the missing data. The dimensions of the unobserved factors x_t remain the same, thus this does not present any problems in updating our factor predictions from one period to the next. Explicitly, if the second series of x_t were missing in period

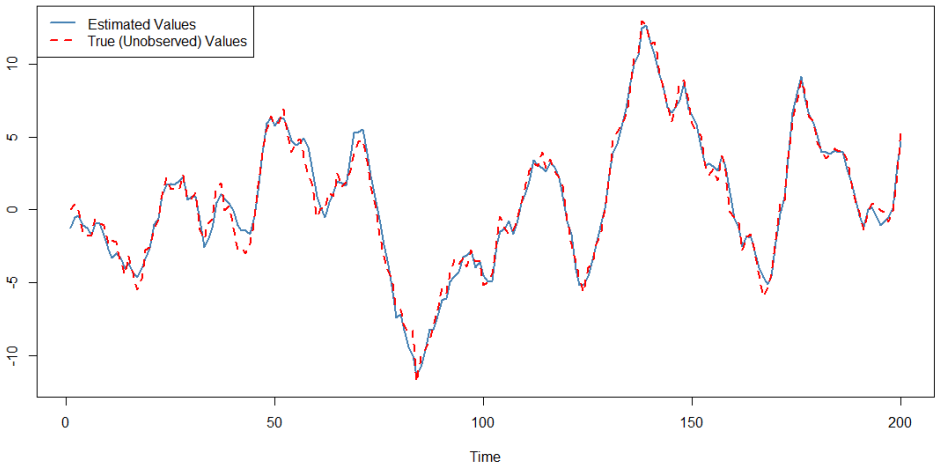


Figure 1: True series $x_{1,t}$ versus the series estimated from observations y_t

t then our transition equation would become

$$y_t = \underbrace{\begin{bmatrix} .5 & 1 \\ 1 & -1 \\ 1 & -.5 \end{bmatrix}}_{H_t} x_t + \varepsilon_t$$

and R becomes a 3×3 identity matrix.

If all the observations are missing for a period t, which will be the case when making out of sample forecasts (as opposed to nowcasts) for example, then we simply ignore the updating step of our filter. That is, our prediction for next period factors is

$$x_{t|t} = Ax_{t-1|t-1}$$

and the variance of this prediction is

$$P_{t|t} = AP_{t-1|t-1}A' + Q$$

3 ARMA Estimation

An Autoregressive Moving Average, or ARMA model, combines lags of observed variables with lags of shocks. Note that one often sees references to ARIMA mod-

els; the I stands for integrated, as in non-stationary. Non-stationarity is typically dealt with prior to estimation (by differencing or detrending), so we will leave the I out. A simple example is an ARMA(1,1) model, that is, one AR component and one MA component

$$y_t = ay_{t-1} + \varepsilon_t + b\varepsilon_{t-1} \tag{27}$$

3.1 MSE Minimization

Estimating models with only AR components is much simpler than those with MA components due to the fact that the errors ε_t are unobserved. While one can estimate AR models (and VAR models) by OLS, that is not true of MA models. One option is to minimize the loss function

$$l = (y_t - \hat{y}_t)^2 \tag{28}$$

where $\hat{y}_t$ is our estimate of y_t. We can then minimize this function over parameters and pre-sample values of shocks. For example, we can minimize the loss function (28) for (27) over a, b, and ε_{-1}. Alternatively, we can treat pre-sample values of shocks, ε_{-1} in this example, as zero.

3.2 Maximum Likelihood Estimation

Assuming normally distributed error terms, we can write the log likeihood for our ARMA model as

$$\mathcal{L}(\theta) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t}{2\sigma^2}$$

The log likelihood requires estimating the additional parameter σ^2, the variance of the shocks ε_t. As before, we can also include pre-sample values of shocks in the parameter vector, or, as suggested by Box and Jenkins (1970), we can set these pre-sample shocks to zero.

Alternatively, we can cast our ARMA(p,q) model in state space form and use the tools developed in section 2.4.2. There is more than one way to cast our ARMA(1,1) model in state space, but one possibility is to use the observation

equation

$$y_t = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ \varepsilon_t \end{bmatrix}$$

and the transition equation

$$\begin{bmatrix} x_t \\ \varepsilon_t \end{bmatrix} = \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \varepsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_t \end{bmatrix}$$

The (log) likelihood function for this model is given by (20), which we can maximize for the parameters a, b, and the variance of shocks σ^2. This formulation has the added advantage that it automatically deals with missing values in y_t.

3.3 Seasonal ARMA Models

Instead of looking at the impact of past values in periods $t-1$ through $t-p$ or past shocks in periods $t-1$ through $t-q$ on our observed series y_t, we may be interested in what happened to our series one year ago, for example. This allows us to capture seasonal variations in our data. We'll follow convention and denote such a model an SARMA(p,q)(P,Q) model where the upper case P and Q denote seasonal lags. For example, suppose we had monthly data and wanted to fit an SARMA(1,0)(1,1) model. We can write this model as

$$y_t = ay_{t-1} + cy_{t-12} + \varepsilon_t + h\varepsilon_{t-12} \tag{29}$$

The SARMA(1,0)(1,2) model will be

$$y_t = ay_{t-1} + cy_{t-12} + \varepsilon_t + h_1\varepsilon_{t-12} + h_2\varepsilon_{t-24}$$

and so on. To remove the impact of seasonality in our observe series y_t, we can remove the estimated components for seasonal lags. Thus for 29, our seasonally adjusted series will be

$$y_t^{SA} = y_t - cy_{t-12} - h\hat{\varepsilon}_{t-12}$$

This may seem like a complicated way to get at seasonality. Why not, for example, use dummy variables, with a dummy for each month? There are a few advantages to seasonal ARMA estimates. First, we will typically have fewer parameters to estimate; dummies on months will only really be viable with a large number of years. Second, seasonal ARMA estimation allows seasonal effects to fluctuate over time; dummies assume the seasonal volatility will be the same every period. Thus if the seasonal impact of, for example, summer holidays is diminishing over time, our seasonal ARMA will capture some of that change.

Note that estimation of seasonal ARMA models by Kalman filtering may not be practical due to the fact that casting the model in state space requires continues lags, i.e. $t-1$ through $t-12$, for example. Mean squared error estimation allows us to only include seasonal lags of interest, and not everything inbetween.

4 Seasonal Adjustment

In practice, seasonal adjustment is typically more involved than the simple process outlines in section 3.3. One big challenge is the fact that ARMA models require stationarity, while much of the raw data we work with is non-stationary. While differencing data will be our normal approach to this issue for forecasting, detrending tends to be more effective for seasonal adjustment. Moreover, extreme values (outliers) may skew estimates of adjustments. For these reasons, seasonal adjustment is typically an iterative process in which we will:

1. estimate a trend of the (log) data and remove the trend to get a stationary series;
2. estimate a seasonal ARMA model on the stationary series and save seasonal factors;
3. remove or reduce any outliers — extreme values of error terms in our SARMA model;
4. remove seasonal factors from (log) level data;
5. repeat steps 1-4 until we are satisfied with the resulting seasonally adjusted data.

For monthly or lower frequency data the U.S. Census Bureau's X-13ARIMA-SEATS software is freely available and included in some statistical software such as IRIS for Matlab or the seasonal package in R. For higher frequency data, we will need to implement our own version of this process.

5 Further Reading

Hamilton (1994) is the main reference for most of the material in these notes,

particularly sections 1 and 2.2. For Kalman filtering, smoothing, and state space methods Durbin and Koopman (2012) is the definitive reference. Ljungqvest and Sargent (2012) Chapter 2 also covers the material in these notes, though the orientation is more towards theoretical applications.

References

Box, G. and Jenkins, G., 1970. *Time series analysis: forecasting and control*. Holden-Day.

Durbin, J. and Koopman, S. J., 2012. *Time Series Analysis by State Space Methods*. Oxford University Press.

Hamilton, J., 1994. *Time Series Analysis*. Princeton University Press.

Kalman, R., 01 1960. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82:35–45.

Ljungqvest, L. and Sargent, T. J., 2012. *Recursive Macroeconomic Theory*. MIT Press.