1. Filtering
2. Examples of filters
3. Maximum Likelihood Estimation
4. Kalman Filtering
5. Estimation of Kalman Filters
6. Next: Spectral Representation of time series
Final:

• Open book

• Study:
  – Lecture notes
  – Book
  – Homework
  – Additional Material
2 Filtering

- Filtering is about transforming a time series with certain “characteristics” or properties into another time series with different characteristics.
- The concept of filtering is very general and we will see it now in the time domain and later in the frequency domain.
- Filtering is best explained by examples.
- Define the lag operator \( L \) to be such that, for a time series \( X_t \),
  \[
  LX_t = X_{t-1}
  \]
  (sometimes you will see \( L \) denoted as \( B \), which people call the “backward” operator)
- The lag operator will be treated as a “number”, i.e. we can define all the usual math operations on it, such as addition, multiplication, inversion, etc.
- For example, we can write
  \[
  L^2 X_t = LLX_t = X_{t-2}
  \]
  \[
  \frac{X_t}{(1 - L)} = \sum_{j=0}^{\infty} X_{t-j} L^j
  \]
• Example: We can represent the familiar AR(1) process \( Y_t = \phi Y_{t-1} + \varepsilon_t \) as:

\[
Y_t = \phi LY_t + \varepsilon_t
\]

or

\[
Y_t - \phi LY_t = \varepsilon_t \\
(1 - \phi L) Y_t = \varepsilon_t
\]

• Example: We can represent an AR(2) process \( Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \) as:

\[
Y_t = \phi_1 LY_t + \phi_2 L^2 Y_t + \varepsilon_t
\]

or

\[
Y_t - \phi_1 LY_t - \phi_2 L^2 Y_t = \varepsilon_t \\
(1 - \phi_1 L - \phi_2 L^2) Y_t = \varepsilon_t
\]
• In the AR(1) case, suppose that $\phi = 0.95$. Then the process is persistent.

• But suppose we have an AR(2) process with $\phi_1 = 0.95$ and $\phi_2 = -0.2$.

• Q: Is the process persistent?

• A:

• Q: What exactly do we mean by a “persistent” process?

• A: We want to know:

$$\frac{\partial y_{t+k}}{\partial \epsilon_t}$$

• In other words, if there is a shock, or “news,” how long would the news impact the process and to what extent.

• In the AR(1) case,

$$y_{t+k} = \phi y_{t+k-1} + \epsilon_{t+k}$$

$$= \epsilon_{t+k} + \ldots + \phi^k \epsilon_t + \phi^{k+1} y_{t-k-1}$$

$$\frac{\partial y_{t+k}}{\partial \epsilon_t} = \phi^k$$

• $\frac{\partial y_{t+k}}{\partial \epsilon_t}$ (as a function of $k$) is known as an “impulse response” function.

• But what about in the AR(2) or AR(k) case?
• We can write an AR(2) process as a vector AR(1) process:
\[
\begin{bmatrix}
y_t \\
y_{t-1}
\end{bmatrix} = \begin{bmatrix}
\phi_1 & \phi_2 \\
1 & 0
\end{bmatrix} \begin{bmatrix}
y_{t-1} \\
y_{t-2}
\end{bmatrix} + \begin{bmatrix}
\varepsilon_t \\
0
\end{bmatrix}
\]
\[
Y_t = \Phi Y_{t-1} + \varepsilon_t
\]

• Then the dynamics of \(Y_t\) are driven by \(\Phi\), because
\[
Y_{t+k} = \varepsilon_{t+k} + \ldots + \Phi^k \varepsilon_t + \Phi^{k+1} Y_{t-1}
\]

• So, we have to characterize the behavior of \(\Phi^k\).

• But \(\Phi\) is a matrix and everything can get pretty messy.
• But there is an easy way out. Write:

\[ \Phi = T \Lambda T^{-1} \]

where \( T \) is the matrix of eigenvectors and \( \Lambda \) is a diagonal matrix with the eigenvalues along the diagonal.

• Then:

\[ \Phi^2 = \Phi \Phi = T \Lambda T^{-1} T \Lambda T^{-1} = T \Lambda^2 T^{-1} \]

• Then:

\[ \Phi^k = T \Lambda^k T^{-1} \]

• So the eigenvalues in \( \Lambda \) will govern the behavior of \( \Phi \). In particular, the highest eigenvalue of \( \Lambda \) will be the most important to consider.
• Recall that the eigenvalues of $\Phi$ (in the AR(2) case) are the solution to:

$$\begin{vmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 0$$

• Or:

$$\lambda^2 - \phi_1 \lambda - \phi_2 = 0$$
We can view \((1 - \phi_1 L - \phi_2 L^2)\) as a polynomial in the lag operator. Therefore, we can look for roots, or numbers \(\lambda_1\) and \(\lambda_2\) such that

\[
(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L) = 0
\]

\[
(1 - (\lambda_1 + \lambda_2)L + \lambda_1 \lambda_2 L^2) = 0
\]

Therefore, matching powers of \(L\), we can find the roots by solving the equations \((\lambda_1 + \lambda_2) = \phi_1\) and \(\lambda_1 \lambda_2 = -\phi_2\).
• Example: Suppose $\phi_1 = 0.6$ and $\phi_2 = -0.08$, so that $Y_t = 0.6LY_t - 0.08L^2Y_t + \varepsilon_t$. You can check that 

$$(1 - 0.6L + 0.08L^2) = (1 - 0.4L)(1 - 0.2L)$$

• But we want to be able to find $\lambda_1$ and $\lambda_2$ in general situations. How to do that?

• Think of a generic polynomial in $z$, (we use $z$ because the $L$ operator has a particular meaning, i.e. it is the lag operator), as above

$$(1 - \phi_1z - \phi_2z^2) = (1 - \lambda_1z)(1 - \lambda_2z)$$

• Now, we ask, what values of $z$ will set the right hand side to zero?

• The answer is: $z = 1/\lambda_1$ or $z = 1/\lambda_2$.

• But why is it important?

• Well, the same $z$ must also set the left hand side to zero.

• But we also know that the $z$ that sets

$$(1 - \phi_1z - \phi_2z^2) = 0$$

can be found by

$$z_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}$$

• In other words, $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$. 

• So, can we always find the $\lambda'$s from the $z'$s by inverting?

• Yes, except for the case when $\phi_1^2 + 4\phi_2 < 0$, because then we will have trouble evaluating $\sqrt{\phi_1^2 + 4\phi_2}$.

• In such a case $z_1$ and $z_2$ are complex numbers (and complex conjugates).

• We can write $z_1 = a + bi$, where $i = \sqrt{-1}$. In our example, you can check that $a = \phi_1/2$ and $b = \frac{1}{2}\sqrt{-\phi_1^2 - 4\phi_2}$

• We can write the complex number $z$ in the polar coordinate as

$$z_1 = R[\cos \theta + i \sin \theta]$$

where

$$R = \sqrt{a^2 + b^2}$$

$$\cos \theta = a/R$$

$$\sin \theta = b/R$$

• Now, we will use DeMoivre's theorem to write

$$z_1 = R[\cos \theta + i \sin \theta] = R.e^{i\theta}$$

• Similarly

$$z_2 = R[\cos \theta - i \sin \theta] = R.e^{-i\theta}$$
Therefore

\[ \lambda_1 = z_1^{-1} = R^{-1}e^{-i\theta} \]
\[ \lambda_2 = z_2^{-1} = R^{-1}e^{i\theta} \]

- The possibility of having complex roots is really cool. We can see that an AR(2) process with complex roots can generate a process that has a very pronounced cyclical component (sin and cos behavior).

- Q: Can we have a cyclical behavior in an AR(1) process?

- We will come back to those same ideas when we get into the frequency domain representation of a series.
• As a final remark, the lag operator is extremely useful to manipulate time series. For instance, we can find the mean and the variance of an AR(1) process simply as:

\[ Y_t = c + \phi L Y_t + \varepsilon_t \]

Hence, denoting \( E(Y_t) = \mu \), we obtain

\[
\begin{align*}
\mu &= c + \phi L \mu \\
\mu &= c + \phi \mu \\
\mu &= \frac{c}{1 - \phi}
\end{align*}
\]

• Similarly for the variance. Denoting the variance \( Var(Y_t) = \gamma \), we can write

\[
\begin{align*}
\gamma &= \phi^2 L^2 \gamma + \sigma^2 \\
\gamma - \phi^2 \gamma &= \sigma^2 \\
\gamma &= \frac{\sigma^2}{1 - \phi^2}
\end{align*}
\]
• Now that we are comfortable with the lag operator and polynomials in the lag operator, here is a filtering example:

• Suppose that $\varepsilon_t$ is an iid series with variance $Var(\varepsilon_t) = \sigma^2$.

• We are going to filter, or transform the series $\varepsilon_t$ into another series, $X_t$.

• First, define the filter

$$F(L) = (1 - \theta L)$$

• Second, we obtain the $X_t$ series as

$$X_t = F(L) \varepsilon_t$$

$$= (1 - \theta L) \varepsilon_t$$

$$= \varepsilon_t - \theta \varepsilon_{t-1}$$

• The series $X_t$ is called a moving average of order 1, or MA(1), process.
• Note that the properties of $X_t$ are different from those of $\varepsilon_t$
  - $\text{Var}(\varepsilon_t) = \sigma^2$, $\text{Cov}(\varepsilon_t \varepsilon_{t-k}) = 0$
  - $\text{Var}(X_t) = \sigma \left(1 + \theta^2\right) > \text{Var}(\varepsilon_t)$
  - $\text{Cov}(X_tX_{t-1}) = \text{Cov}([\varepsilon_t - \theta \varepsilon_{t-1}] [\varepsilon_{t-1} - \theta \varepsilon_{t-2}]) = -\theta \sigma^2$
  - $\text{Cov}(X_tX_{t-j}) = 0$, j>1.
• So, the filtered series $X_t$ induces a slight (one period) serial correlation.
• Suppose we define another filter, $G(L) = F(L)^{-1} = 1/(1 - \theta L)$

• We will filter $\varepsilon_t$ to obtain a series $Y_t$, such that

$$Y_t = G(L)\varepsilon_t$$

• Note that $X_t$ and $Y_t$ are obtained from the same $\{\varepsilon_t\}$ process, but by applying different filters.

• We will investigate the properties of $Y_t$

$$Y_t = \frac{\varepsilon_t}{(1 - \theta L)}$$

$$Y_t(1 - \theta L) = \varepsilon_t$$

$$Y_t = \theta Y_{t-1} + \varepsilon_t$$

• $Y_t$ is our familiar AR(1) process. We know that

- $Var(Y_t) = \frac{\sigma^2}{1 - \theta^2}$

- $Cov(Y_tY_{t-k}) = \frac{\sigma^2\theta^k}{1 - \theta^2}$

• Therefore, the filter $G(L)$ induces a persistence that lasts longer than only one period.
• The $H(L) = (1 - L)$ filter.

• Recall that if we have a series that is non-stationary, my advice was to take first-difference, and the new series would be stationary. We will illustrate this advice with a (quite general) example.

• Suppose we have a series, $Y_t$, composed of a stationary and a non-stationary series:

\[
Y_t = Z_t + X_t \\
Z_t = Z_{t-1} + e_t \\
X_t = \phi X_{t-1} + u_t, \quad |\phi| < 1
\]

• Note that we can rewrite $Z_t (1 - L) = e_t$ and $X_t (1 - \phi L) = u_t$. 
• Therefore, we can write

\[ Y_t = Z_t + X_t \]
\[ Y_t = \frac{e_t}{1 - L} + \frac{u_t}{1 - \phi L} \]

\[
(1 - L) (1 - \phi L) Y_t = e_t (1 - \phi L) + u_t (1 - L) \\
(1 - L) (1 - \phi L) Y_t = \eta_t
\]

• But, from here we can immediately see that one of the roots of the polynomial in \( L \) is 1. Therefore, this process will be non-stationary.

• Q: Why don’t we like non-stationary processes?
• We will stationarize $Y_t$, by filtering it using $H(L)$.

• Let’s call the filtered series $\tilde{Y}_t = H(L)Y_t$

\[
\tilde{Y}_t = (1 - L) \left[ \frac{e_t}{1 - L} + \frac{u_t}{1 - \phi L} \right]
= e_t + \frac{u_t (1 - L)}{1 - \phi L}
= e_t + \frac{u_t}{1 - \phi L} - \frac{u_{t-1}}{1 - \phi L}
= stationary + stationary + stationary
= stationary

• Therefore, the filter $H(L)$ transformed the non-stationary series $Y_t$ into a stationary series $\tilde{Y}_t$. 
To summarize:
- We have defined the lag operator \( L \)
- Using the lag operator, we can define a filter function \( F(L) \)
- \( F(L) \) modifies the properties of a series \( Y_t \)
- It is useful to modify the properties of \( Y_t \) (think stationarity, seasonality, etc.)

Filtering is used everywhere
- Stereos
- Cell phones
- Search for extra-terrestrials
3 Examples of filters

- We have already given a few examples of familiar filters. Here we discuss some more:

- Intuitively, we know that when we take a moving average of a series, we effectively smooth the series, i.e. we iron out any non-linearities.

- Example:
1.

• In the above example, we have created a series 
  \( Y_t = Z_t + X_t \).

• \( Z_t = 0.9Z_{t-1} + e_t \) and \( X_t = 0.1X_{t-1} + u_t \)

• We think of \( Z_t \) as the strong “signal” and \( X_t \) is the almost-iid noise.

• In other words, \( Z_t \) is responsible for the big swings in \( Y_t \), whereas the jagged peaks are due to \( X_t \).
• We want to smooth out the noise and be left with the signal.

• We can do that, as in the figure, by applying a filter \( F(L) = \frac{1}{3}L + \frac{1}{3}L^0 + \frac{1}{3}L^{-1} \) to \( Y_t \).

• The filtered series is:
  \[
  \tilde{Y}_t = F(L)Y_t = \left( \frac{1}{3}L + \frac{1}{3}L^0 + \frac{1}{3}L^{-1} \right) Y_t = \frac{1}{3}Y_{t-1} + \frac{1}{3}Y_t + \frac{1}{3}Y_{t+1}
  \]

• This is nothing but a moving average of \( Y_t \) and its two adjacent values.

• In a similar fashion, we can smooth the series even more by taking a longer moving average.

• Example: \( F(L) = \frac{1}{7}L^3 + \frac{1}{7}L^2 + \frac{1}{7}L + \frac{1}{7}L^0 + \frac{1}{7}L^{-1} + \frac{1}{7}L^{-2} + \frac{1}{7}L^{-3} \)
2.

- The second filter, the MA(7) filter induces an even smoother behavior in the filtered series.

- So, by filtering the series, we have effectively removed the components of $Y_t$ that are moving fast, or that induce the mostly unpredictable behavior in $Y_t$.

- If you have used some of the financial websites (Yahoo!), you know that they offer the option of
displaying a price using a MA filter.

- But \( F(L) = \frac{1}{L^3}L^2 + \frac{1}{L}L + \frac{1}{L^3}L^0 + \frac{1}{L^2}L^{-1} + \frac{1}{L}L^{-2} + \frac{1}{L^3}L^{-3} \) is a special filter. It is:
  - Two sided
  - Symmetric
  - Equally-weighted

- The design of the filter depends on the application.
• The MA filter is the single most widely used filter in applied work (justifiably so, since it is simple and does the job).

• However, suppose we want to filter out a specific feature of the data.
  – For example, corporate taxes are due quarterly. Therefore, any series that is tightly related to taxes will have a quarterly-frequency component to it. In other words, every time taxes are due, the series will exhibit a certain pattern.
  – We might not want to have this quarterly pattern in taxes influence our results. In other words, we want to seasonally adjust (or filter) the series so that quarterly fluctuations are not part of the new series. We can do that with a quarterly filter.
  – There are prices, such as prices of commodities, prices of utilities that have a seasonal component. An entire industry of analysts is trying to forecast such fluctuations. For example, it is widely believed that the price of crude (and refined) oil has a 6 months cycle, peaking in the summer (travel season) and the winter (heating).
  – Although this is usually (historically) true, we should not forget that we are dealing with random
variables. For example, last year, the price of oil DECREASED during the summer months.
• So, the question is, how can we isolate the features of the data that happen every 3 or every 6 months?
• Can we decompose a series into periodic components, where each component has different periodicity?
• Yes, but for that we will need a great deal of tools (next time).
• Before we leave the time domain representation, we will introduce the Kalman Filter and its estimation.
4 Maximum Likelihood Estimation
(Preliminaries for Kalman Filtering)

• Suppose we have the series \( \{Y_1, Y_2, \ldots, Y_T\} \) with a joint density \( f_{Y_T \ldots Y_1}(\theta) \) that depends on some parameters \( \theta \) (such as means, variances, etc.)

• We observe a realization of \( Y_t \).

• If we make some functional assumptions on \( f \), we can think of \( f \) as the probability of having observed this particular sample, given the parameters \( \theta \).

• The maximum likelihood estimate (MLE) of \( \theta \) is the value of the parameters \( \theta \) for which this sample is most likely to have been observed.

• In other words, \( \hat{\theta}^{MLE} \) is the value that maximizes \( f_{Y_T \ldots Y_1}(\theta) \).
Q: But, how do we know what \( f \)--the true density of the data--is?

A: We don’t.

Usually, we assume that \( f \) is normal, but this is strictly for simplicity. The fact that we have to make distributional assumptions limits the use of MLE in many financial applications.

Recall that if \( Y_t \) are independent over time, then

\[
f_{Y_T, \ldots, Y_1}(\theta) = f_{Y_T}(\theta_T) f_{Y_{T-1}}(\theta_{T-1}) \ldots f_{Y_1}(\theta_1) = \prod_{i=1}^{T} f_{Y_i}(\theta_i)
\]

Sometimes it is more convenient to take the log of the likelihood function, then

\[
\Lambda(\theta) = \log f_{Y_T, \ldots, Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i}(\theta)
\]
• However, in most time series applications, the independence assumption is untenable. Instead, we use a conditioning trick.

• Recall that
  \[ f_{Y_2Y_1} = f_{Y_2|Y_1} f_{Y_1} \]

• In a similar fashion, we can write
  \[ f_{Y_T,...,Y_1}(\theta) = f_{Y_T|Y_{T-1},...,Y_1}(\theta) f_{Y_{T-1}|Y_{T-2},...,Y_1}(\theta) \cdots f_{Y_1}(\theta) \]

• The log likelihood can be expressed as
  \[ \Lambda(\theta) = \log f_{Y_T,...,Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i|Y_{i-1},...,Y_1}(\theta_i) \]
Example: The log-likelihood of an AR(1) process

\[ Y_t = c + \phi Y_{t-1} + \varepsilon_t \]

Suppose that \( \varepsilon_t \) is iid \( N(0, \sigma^2) \)

Recall that \( E(Y_t) = \frac{c}{1-\phi} \) and \( Var(Y_t) = \frac{\sigma^2}{1-\phi^2} \)

Since \( Y_t \) is a linear function of the \( \varepsilon_t \)'s, then it is also Normal (sum of normals is a normal).

Therefore, the density (unconditional) of \( Y_t \) is Normal.

Result: If \( Y_1 \) and \( Y_2 \) are jointly Normal, then the marginals are also normal.

Therefore,

\[ f_{Y_2|Y_1} \text{ is } N \left( (c + \phi y_1), \sigma^2 \right) \]

or

\[ f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_2 - c - \phi y_1)^2}{2\sigma^2} \right] \]
Similarly,

\[ f_{Y_3|Y_2} \text{ is } N \left( (c + \phi y_2), \sigma^2 \right) \]

or

\[
f_{Y_3|Y_2} = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{(y_3 - c - \phi y_2)^2}{2\sigma^2} \right]
\]
Then, the log likelihood can be written as

\[
\Lambda(\theta) = \log f_Y + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}} \\
= -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \left(\frac{\sigma^2}{(1 - \phi^2)}\right) \\
- \frac{\{y_1 - (c/(1 - \phi))\}^2}{2\sigma^2/(1 - \phi^2)} \\
- \frac{(T - 1)}{2} \log (2\pi) - \frac{(T - 1)}{2} \log (\sigma^2) \\
- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}
\]

- The unknown parameters are collected in \( \theta = (c, \phi, \sigma) \)
- We can maximize \( \Lambda(\theta) \) with respect to all those parameters and find the estimates that maximize the probability of having observed such a sample.

\[
\max_{\theta} \Lambda(\theta)
\]
- Sometimes, we can even put constraints (such as \(|\phi| < 1|\))
- Q: Is it necessary to put the constraint \( \sigma^2 > 0 ? \)
• Note: If we forget the first observation, then we can write (setting $c = 0$) the FOC:

$$- \sum_{t=2}^{T} \frac{\partial}{\partial \phi} \frac{(y_t - \phi y_{t-1})^2}{2\sigma^2} = 0$$

$$\sum_{t=2}^{T} y_{t-1} (y_t - \phi y_{t-1}) = 0$$

\[ \hat{\phi} = \frac{\sum_{t=2}^{T} y_{t-1} y_t}{\sum_{t=2}^{T} y_{t-1}^2} \]

• RESULT: In the univariate linear regression case, OLS, GMM, MLE are equivalent!!!
• To summarize the maximum likelihood principle:
  (a) Make a distributional assumption about the data
  (b) Use the conditioning to write the joint likelihood function
  (c) For convenience, we work with the log-likelihood function
  (d) Maximize the likelihood function with respect to the parameters

• There are some subtle points.
  – We had to specify the unconditional distribution of the first observation
  – We had to make an assumption about the dependence in the series

• But sometimes, MLE is the only way to go.

• MLE is particularly appealing if we know the distribution of the series. Most other deficiencies can be circumvented.
• Now, you will ask: What are the properties of $\hat{\theta}^{MLE}$? More specifically, is it consistent? What is its distribution, where $\hat{\theta}^{MLE} = \arg \max \Lambda(\theta)$

• Yes, $\hat{\theta}^{MLE}$ is a consistent estimator of $\theta$.

• As you probably expect the asymptotic distribution of $\hat{\theta}^{MLE}$ is normal.

• Result:
  
  $$T^{1/2} \left( \hat{\theta}^{MLE} - \theta \right) \sim ^a N(0, V)$$

  $$V = \left[ -\frac{\partial^2 \Lambda(\theta)}{\partial \theta \partial \theta'} \bigg|_{\hat{\theta}^{MLE}} \right]^{-1}$$

  or

  $$V = \sum_{t=1}^{T} l(\hat{\theta}^{MLE}, y) l(\hat{\theta}^{MLE}, y)$$

  $$l(\hat{\theta}^{MLE}, y) = \frac{\partial f}{\partial \theta}(\hat{\theta}^{MLE}, y)$$

• But we will not dwell on proving those properties.
• Now we are ready to plunge into Kalman Filtering!!!!
• We will use MLE to estimate a complicated system of observable and unobservable components
• As I mentioned at the beginning of this course, the Kalman Filtering techniques used in finance are not exactly analogous to those used in engineering.
5 Kalman Filtering

- History: Kalman (1963) paper
- Problem: We have a missile that we want to guide to its proper target.
  - The trajectory of the missile IS observable from the control center.
  - Most other circumstances, such as weather conditions, possible interception methods, etc. are NOT observable, but can be forecastable.
  - We want to guide the missile to its proper destination.
- In finance the setup is very similar, but the problem is different.
- In the missile case, the parameters of the system are known. The interest is, given those parameters, to control the missile to its proper destination.
- In finance, we want to estimate the parameters of the system. We are usually not concerned with a control problem, because there are very few instruments we can use as controls (although there are counter-examples).
5.1 Setup (Hamilton CH 13)

\[ y_t = A'x_t + H'z_t + w_t \]
\[ z_t = Fz_{t-1} + v_t \]

where

- \( y_t \) is the observable variable (think “returns”)
  - The first equation, the \( y_t \) equation is called the “space” or the “observation” equation.
- \( z_t \) is the unobservable variable (think “volatility” or “state of the economy”)
  - The second equation, the \( z_t \) equation is called the “state” equation.
- \( x_t \) is a vector of exogenous (or predetermined) variables (we can set \( x_t = 0 \) for now).
- \( v_t \) and \( w_t \) are iid and assumed to be uncorrelated at all lags
  \[ E(w_tv_t') = 0 \]
- Also \( E(v_tv_t') = Q \), \( E(w_tw_t') = R \)
- The system of equations is known as a state-space representation.
- Any time series can be written in a state-space representation.
In standard engineering problems, it is assumed that we know the parameters \( A, H, F, Q, R \).

The problem is to give impulses \( x_t \) such that, given the states \( z_t \), the missile is guided as closely to target as possible.

In finance, we want to estimate the unknown parameters \( A, H, F, Q, R \) in order to understand where the system is going, given the states \( z_t \). There is little attempt at guiding the system. In fact, we usually assume that \( x_t = 1 \) and \( A = E(Y_t) \), or even that \( x_t = 0 \).
• Note: Any time series can be written as a state space.

• Example: AR(2): \( Y_{t+1} - \mu = \phi_1 (Y_t - \mu) + \phi_2 (Y_{t-1} - \mu) + \epsilon_{t+1} \)

• State equation:

\[
\begin{bmatrix}
Y_{t+1} - \mu \\
Y_t - \mu
\end{bmatrix} =
\begin{bmatrix}
\phi_1 & \phi_2 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
Y_t - \mu \\
Y_{t-1} - \mu
\end{bmatrix} +
\begin{bmatrix}
\epsilon_{t+1} \\
0
\end{bmatrix}
\]

• Observation equation:

\( y_t = \mu + \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} Y_{t+1} - \mu \\ Y_t - \mu \end{bmatrix} \)

• There are other state-space representations of \( Y_t \). Can you write down another one?
• As a first step, we will assume that $A, H, F, Q, R$ are known.

• Our goal would be to find a best linear forecast of the state (unobserved) vector $z_t$. Such a forecast is needed in control problems (to take decisions) and in finance (state of the economy, forecasts of unobserved volatility).

• The forecasts will be denoted by:

$$z_{t+1|t} = E (z_{t+1|t}|y_{t...}, x_{t...})$$

and we assume that we are only taking linear projections of $z_{t+1}$ on $y_{t...}, x_{t.....}$ Nonlinear Kalman Filters exist but the results are a bit more complicated.

• The Kalman Filter calculates the forecasts $z_{t+1|t}$ recursively, starting with $z_{1|0}$, then $z_{2|1}, ...$ until $z_{T|T-1}$.

• Since $z_{t|t-1}$ is a forecast, we can ask how good of a forecast it is?

• Therefore, we define $P_{t|t-1} = E \left( (z_t - z_{t|t-1}) (z_t - z_{t|t-1}) \right)$, which is the forecasting error from the recursive forecast $z_{t|t-1}$. 
The Kalman Filter can be broken down into 5 steps

1. Initialization of the recursion. We need $z_{1|0}$. Usually, we take $z_{1|0}$ to be the unconditional mean, or $z_{1|0} = E(z_1)$.

(Q: how can we estimate $E(z_1)$? )

The associated error with this forecast is $P_{1|0} = E \left( (z_{1|0} - z_1) (z_{1|0} - z_1) \right)$
2. Forecasting $y_t$ (intermediate step)
   The ultimate goal is to calculate $z_{t|t-1}$, but we do that recursively. We will first need to forecast the value of $y_t$, based on available information:
   $$E(y_t|x_t, z_t) = A'x_t + H'z_t$$
   From the law of iterated expectations,
   $$E_{t-1}(E_t(y_t)) = E_{t-1}(y_t) = A'x_t + H'z_{t|t-1}$$
   The error from this forecast is
   $$y_t - y_{t|t-1} = H'(z_t - z_{t|t-1}) + w_t$$
   with MSE
   $$E\left((y_t - y_{t|t-1})(y_t - y_{t|t-1})'\right)$$
   $$= E\left[H'(z_t - z_{t|t-1})(z_t - z_{t|t-1})'H\right] + E[w_tw_t']$$
   $$= H'P_{t|t-1}H + R$$
3. Updating Step ($z_{t|t}$)

- Once we observe $y_t$, we can update our forecast of $z_t$, denoting it by $z_{t|t}$, before making the new forecast, $z_{t+1|t}$.

- We do this by calculating $E(z_{t|t} \mid y_t, x_t, ...) = z_{t|t}$

$$z_{t|t} = z_{t|t-1} + E \left( (z_t - z_{t|t-1}) (y_t - y_{t|t-1}) \right) \ast$$

$$\left( E \left( y_t - y_{t|t-1} \right) (y_t - y_{t|t-1})' \right)^{-1} \left( y_t - y_{t|t-1} \right)$$

- We can write this a bit more intuitively as:

$$z_{t|t} = z_{t|t-1} + \beta (y_t - y_{t|t-1})$$

where $\beta$ is the OLS coefficient from regressing $(z_t - z_{t|t-1})$ on $(y_t - y_{t|t-1})$.

- The bigger is the relationship between the two forecasting errors, the bigger the correction must be.
- It can be shown that 
\[ z_{t|t} = z_{t|t-1} + P_{t|t-1} H \left( H' P_{t|t-1} H + R \right)^{-1} (y_t - A' x_t - H' z_{t|t-1}) \]
- This updated forecast uses the old forecast \( z_{t|t-1} \), and the just observed values of \( y_t \) and \( x_t \).
4. Forecast $z_{t+1|t}$.

- Once we have an update of the old forecast, we can produce a new forecast, the forecast of $z_{t+1|t}$

$$E_t(z_{t+1}) = E(z_{t+1}|y_t, x_t, \ldots)$$

$$= E(Fz_t + v_{t+1}|y_t, x_t, \ldots)$$

$$= FE(z_t|y_t, x_t, \ldots) + 0$$

$$= Fz_t$$

- We can use the above equation to write

$$E_t(z_{t+1}) = F\{z_t|t-1$$

$$+ P_{t|t-1} H (H'H_{t|t-1} H + R)^{-1} (y_t - A'x_t - H'z_{t|t-1}) \}$$

$$= Fz_{t|t-1}$$

$$+ FP_{t|t-1} H (H'H_{t|t-1} H + R)^{-1} (y_t - A'x_t - H'z_{t|t-1})$$

- We can also derive an equation for the error in forecast as a recursion

$$P_{t+1|t} = F[P_{t|t}$$

$$- P_{t|t-1} H (H'H_{t|t-1} H + R)^{-1} H'P_{t|t-1}]F'$$

$$+ Q$$

5. Go to step 2, until we reach $T$. Then, we are done.
• Summary: The Kalman Filter produces
  – The optimal forecasts of $z_{t+1|t}$ and $y_{t+1|t}$ (optimal within the class of linear forecasts)
  – We need some initialization assumptions
  – We need to know the parameters of the system, i.e. $A, H, F, Q, R$.
• Now, we need to find a way to estimate the parameters $A, H, F, Q, R$.
• By far, the most popular method is MLE.
• Aside: Simulations Methods–getting away from the restrictive assumptions of $\varepsilon_t$
6 Estimation of Kalman Filters

• Suppose that $z_1$, and the shocks $(w_t, v_t)$ are jointly normally distributed.

• Under such an assumption, we can make the very strong claim that the forecasts $z_{t+1|t}$ and $y_{t+1|t}$ are optimal among any functions of $x_t$, $y_{t-1}$.... In other words, if we have normal errors, we cannot produce better forecasts using the past data than the Kalman forecasts!!

• If the errors are normal, then all variables in the linear system have a normal distribution.

• More specifically, the distribution of $y_t$ conditional on $x_t$, and $y_{t-1}$,... is normal, or

$y_t|x_t, y_{t-1}... \sim N \left( A'x_t + H'z_{t|t-1}, \left( H'P_{t|t-1}H + R \right) \right)$

• Therefore, we can specify the likelihood function of $y_t|x_t, y_{t-1}$ as we did above.

$$f_{y_t|x_t,y_{t-1}} = (2\pi)^{-n/2} \left| H'P_{t|t-1}H + R \right|^{-1/2}$$

$$\times \exp \left[ -\frac{1}{2} \left( y_t - A'x_t - H'z_{t|t-1} \right)' \left( H'P_{t|t-1}H + R \right)^{-1} \times \left( y_t - A'x_t - H'z_{t|t-1} \right) \right]$$
• The problem is to maximize

\[
\max_{A,H,F,Q,R} \sum_{t=1}^{T} \log f_{y_t|x_t,y_{t-1}}
\]

• Words of wisdom:
  - This maximization problem can easily get unmanageable to estimate, even using modern computers. The problem is that searching for global max is very tricky.
    * A possible solution is to make as many restrictions as possible and then to relax them one by one.
    * A second solution is to write a model that gives theoretical restrictions.
  - Recall that there are more than 1 state space representations of an AR process. This implies that some of the parameters in the state-space system are not identified. In other words, more than one value of the parameters (different combinations) can give rise to the same likelihood function.
    * Then, which likelihood do we choose?
    * Have to make restrictions so that we have an exactly identified problem.