Today’s Agenda

1. MLE–Simple Introduction
   – GARCH estimation

2. Kalman Filtering

3. The Delta Method

4. Empirical Portfolio Choice

5. Wold Decomposition of Stationary Processes
Maximum Likelihood Estimation

(Preliminaries for GARCH/Stochastic Volatility & 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

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

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

\[
\Lambda(\theta) = \log f_{Y_{T-1} \ldots Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i}(\theta)
\]
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

\[ f_{Y_T,...,Y_1}(\theta) = f_{Y_T|Y_{T-1},...,Y_1}(\theta)f_{Y_{T-1}|Y_{T-2},...,Y_1}(\theta)\ldots f_{Y_1}(\theta) \]

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

\[
\Lambda (\theta) = \log f_Y + \sum_{t=2}^{T} \log f_{Y|Y_{t-1}}
\]

\[
= -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \left( \frac{\sigma^2}{1 - \phi^2} \right)
- \frac{\{y_1 - (c/(1 - \phi))\}^2}{2\sigma^2/(1 - \phi^2)}
- \frac{(T - 1)}{2} \log (2\pi) - \frac{(T - 1)}{2} \log (\sigma^2)
- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}
\]

- 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:

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

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

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

• 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:**

\[
T^{1/2} \left( \hat{\theta}^{MLE} - \theta \right) \sim ^a N(0, V)
\]

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

or

\[
V = \sum_{t=1}^{T} l\left( \hat{\theta}^{MLE}, y \right) l\left( \hat{\theta}^{MLE}, y \right)
\]

\[
l\left( \hat{\theta}^{MLE}, y \right) = \frac{\partial f}{\partial \theta} \left( \hat{\theta}^{MLE}, y \right)
\]

But we will not dwell on proving those properties.
Another Example: The log-likelihood of an AR(1)+ARCH(1) process

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

- where,
  \[ u_t = \sqrt{h_t} v_t \]

- ARCH(1) is:
  \[ h_t = \zeta + a u_{t-1}^2 \]
  where \( v_t \) is iid with mean 0, and \( E(v_t^2) = 1 \).

- GARCH(1,1): Suppose, we specify \( h_t \) as
  \[ h_t = \zeta + \delta h_{t-1} + a u_{t-1}^2 \]

- 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,
  \[ f_{Y_2|Y_1} \text{ is } N((c + \phi y_1, h_2) \]
  or for the ARCH(1)
  \[ f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi(\zeta + a u_1^2)}} \exp \left[ -\frac{(y_2 - c - \phi y_1)^2}{2(\zeta + a u_1^2)} \right] \]
Similarly,

\[ f_{Y_3|Y_2} \text{ is } N((c + \phi y_2), h_3) \]

or

\[ f_{Y_3|Y_2} = \frac{1}{\sqrt{2\pi} (\zeta + au_2^2)} \exp \left[ -\frac{(y_3 - c - \phi y_2)^2}{2 (\zeta + au_2^2)} \right] \]
Then, the conditional log likelihood can be written as

\[ \Lambda (\theta | y_1) = \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}} \]

\[ = - \frac{(T - 1)}{2} \log (2\pi) - \frac{(1)}{2} \sum_{t=2}^{T} \log \left( \zeta + au_{t-1}^2 \right) \]

\[ - \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2(\zeta + \alpha u_{t-1}^2)} \]

- The unknown parameters are collected in \( \theta = (c, \phi, \zeta, \alpha) \)

- 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.

\[ \max_{\theta} \Lambda (\theta) \]

- Example:mle_arch.m
Similarly for GARCH(1,1):

\[ \Lambda (\theta | y_1) = \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}} \]

\[ = -\frac{(T - 1)}{2} \log (2\pi) - \frac{1}{2} \sum_{t=2}^{T} \log (h_t) \]

\[ - \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2 h_t} \]

where

\[ h_t = \zeta + \delta h_{t-1} + \alpha u_{t-1}^2 \]

- To construct \( h_t \), we have to filter the \( \{u_{t-1}\} \) series.
- For a given \( u_t \), \( h_0 \), and \( \zeta, \delta, \) and \( \alpha \), we construct \( h_t \).
- The \( h_t \) will allow us to evaluate the likelihood \( \Lambda (\theta | y_1) \).
- Optimize \( \Lambda (\theta | y_1) \) with respect to all the parameters, given the initial conditions.
- This recursive feature of the GARCH makes it harder to estimate with GMM.
2 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).
2.1 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
  \[ E(w_t v_t') = 0 \]

- Also \( E(v_t v_t') = Q, E(w_t w_t') = R \)

- 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): \[ Y_{t+1} - \mu = \phi_1 (Y_t - \mu) + \phi_2 (Y_{t-1} - \mu) + \varepsilon_{t+1} \]

• State equation:
\[
\begin{bmatrix}
Y_{t+1} - \mu \\
Y_t - \mu
\end{bmatrix} =
\begin{bmatrix}
\phi_1 & \phi_2 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
Y_t - \mu \\
Y_{t-1} - \mu
\end{bmatrix} +
\begin{bmatrix}
\varepsilon_{t+1} \\
0
\end{bmatrix}
\]

• 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:

$$z_{t+1|t} = E(z_{t+1} | y_t..., x_t... )$$

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 $P_{t|t-1} = E \left( \left( z_t - z_{t|t-1} \right) \left( z_t - z_{t|t-1} \right) \right)$, 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 $P_{1|0} = E\left(\left(z_{1|0} - z_1\right)\left(z_{1|0} - z_1\right)\right)$
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
   
   $$
   E \left( y_t - y_{t|t-1} \right) \left( y_t - y_{t|t-1} \right)' 
   = E \left[ H' (z_t - z_{t|t-1}) (z_t - z_{t|t-1})' H \right] + E [w_t w_t'] 
   = H' P_{t|t-1} H + R
   $$
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 $E(z_t|y_t, x_t, \ldots) = z_{t|t}$

\[
z_{t|t} = z_{t|t-1} + E \left( (z_t - z_{t|t-1}) (y_t - y_{t|t-1}) \right) \ast \left( E (y_t - y_{t|t-1}) (y_t - y_{t|t-1})' \right)^{-1} (y_t - y_{t|t-1})
\]

– 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 $z_{t+1|t}$
     \[
     E_t(z_{t+1}) = E \left( z_{t+1|y_t, x_t, \ldots} \right) \\
     = E \left( Fz_t + v_{t+1|y_t, x_t, \ldots} \right) \\
     = FE(z_t|y_t, x_t, \ldots) + 0 \\
     = Fz_t 
     \]
   - We can use the above equation to write
     \[
     E_t(z_{t+1}) = \{ z_{t|t-1} \\
     + P_{t|t-1} H \left( H'P_{t|t-1} H + R \right)^{-1} (y_t - A'x_t - H'z_{t|t-1}) \} \\
     = Fz_{t|t-1} \\
     + FP_{t|t-1} H \left( H'P_{t|t-1} H + R \right)^{-1} (y_t - A'x_t - H'z_{t|t-1}) 
     \]
   - We can also derive an equation for the error in forecast as a recursion
     \[
     P_{t+1|t} = F[P_t|t] \\
     - P_{t|t-1} H \left( H'P_{t|t-1} H + R \right)^{-1} H'P_{t|t-1}]F' \\
     + Q 
     \]
   - 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 $\varepsilon_t$
2.2 Estimation of Kalman Filters (MLE)

- 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

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

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

  \[
  f_{y_t|x_t,y_{t-1}} = (2\pi)^{-n/2} \left| H'P_{t|t-1}H + R \right|^{-1/2} \\
  \times \exp\left[ -\frac{1}{2} \left( y_t - A'x_t - H'z_{t|t-1} \right)' \left( H'P_{t|t-1}H + R \right)^{-1} \left( y_t - A'x_t - H'z_{t|t-1} \right) \right]
  \]
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.
2.3 Applications in Finance

- Anytime we have unobservable state variables
  - Filtering expected returns (Pastor and Stambaugh (JF, 2008))
  - Filtering variance (Brandt and Kang (JFE, 2007))

- Interpolation of data
  - Bernanke and Kuttner (JME?)

- Time varying parameters
  - Time-varying Betas (Ghysels (JF, 1998))
3 Kalman Smoother

- For purely forecasting purposes, we need
  \[ z_{t|t-1} = E(z_t|I_{t-1}) \]
  where \( I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, y_1, x_{t-1}, \ldots x_1\} \) and the corresponding error \( P_{t|t-1} = E((z_t - z_{t|t-1})^2) \).

- But if we want to model a process (understand its properties), we might want to incorporate all the available information in \( I_T = \{y_T, y_{T-1}, \ldots, y_1, x_T, \ldots x_1\} \).

- In other words, we might want to estimate
  \[ z_{t|T} = E(z_t|I_T) \]
  There is definitely a look-ahead bias here, but that is the point. We want to include all available information in order to get a better glimpse into the properties of \( z_t \).

- Recall that from the KF, we have the sequences \( \{z_{t+1|t}\}, \{z_{t|t}\}, \{P_{t+1|t}\}, \{P_{t|t}\} \).
Suppose someone tells you the correct value of \( z_{t+1} \) at time \( t \). How can you improve upon the best forecast \( z_{t|t} \)? It turns out that we do the same updating as we did in step 3 of the KF:

\[
E(z_t|z_{t+1}, I_t) = z_{t|t} + E \left( (z_t - z_{t|t}) \left( z_{t+1} - z_{t+1|t} \right) \right) \left( E \left( z_{t+1} - z_{t+1|t} \right) \left( z_{t+1} - z_{t+1|t} \right)^\prime \right)^{-1} \left( z_{t+1} - z_{t|t} \right)
\]

1. – We can write this a bit more intuitively as:

\[
E(z_t|z_{t+1}, I_t) = z_{t|t} + J_t \left( z_{t+1} - z_{t+1|t} \right)
\]

where

\[
J_t = E \left( (z_t - z_{t|t}) \left( z_{t+1} - z_{t+1|t} \right) \right) \left( E \left( z_{t+1} - z_{t+1|t} \right) \left( z_{t+1} - z_{t+1|t} \right)^\prime \right)^{-1}
\]

\[
= P_{t|t} F P_{t+1|t}^{-1}
\]

– Because the process is Markovian, \( E(z_t|z_{t+1}, I_t) = E(z_t|z_{t+1}, I_T) \). We can’t do better than that!

Hence,

\[
E(z_t|z_{t+1}, I_T) = z_{t|t} + J_t \left( z_{t+1} - z_{t+1|t} \right)
\]

– Last step. We can show that

\[
E(z_t|I_T) = z_{t|T} = z_{t|t} + J_t \left( z_{t+1|T} - z_{t+1|t} \right)
\]
Hence, the KS algorithm is, after we obtain the KF \( \{z_{t+1|t}\}, \{z_{t|t}\}, \{P_{t+1|t}\}, \{P_{t|t}\} \)

2. Start at the end, \( z_{T|T} \).

3. Compute \( J_{T-1} = P_{T-1|T-1} F P_{T|T-1}^{-1} \)

4. Compute

\[
z_{T-1|T} = z_{T-1|T-1} + J_{T-1} \left( z_{T|T} - z_{T|T-1} \right)
\]

5. Use \( z_{T-1|T} \) to compute \( z_{T-2|T} \) and so on.

6. We can compute the associated MSE as

\[
P_{t|T} = P_{t|t} + J_t \left( P_{t+1|T} - P_{t+1|t} \right) J_t^	op
\]
4 Time-Varying Parameters

• An example of a time varying parameter model:
  \[ r_{t+1} = \alpha + \beta_t x_t + \epsilon_{t+1} \]
  \[ \beta_{t+1} = \gamma \beta_t + v_{t+1} \]

• Q: What equations are observations and what are the state equations?

• Note that this does not fit within the KF setup:
  \[ y_t = A' x_t + H' z_t + w_t \]
  \[ z_t = F z_{t-1} + v_t \]

• We need the generalization
  \[ y_t = A(x_t) + H(x_t)' z_t + w_t \]
  \[ z_{t+1} = F(x_t) z_t + v_{t+1} \]

• Note that \( F(x_t) \) and not \( F(x_{t+1}) \) in the state equation!
• Now, we have to assume that—we didn’t have to do it earlier!
\[
\begin{bmatrix}
w_t \\
v_{t+1}
\end{bmatrix} | x_t, I_{t-1} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q(x_t) & 0 \\ 0 & R(x_t) \end{bmatrix} \right)
\]
• Before, we had linearity in all variables. Now, we don’t.
• Given the conditional normal assumption, we can show that
\[
\begin{bmatrix}
z_t \\
y_t
\end{bmatrix} | x_t, I_{t-1} \sim N \left( \begin{bmatrix} z_{t|t-1} \\ A(x_t) + H(x_t)^t z_{t|t-1} \end{bmatrix}, V \right)
\]
\[
V = \begin{bmatrix}
P_{t|t-1} & P_{t|t-1}H(x_t) \\
H'(x_t)P_{t|t-1} & H'(x_t)P_{t|t-1}H(x_t) + R(x_t)
\end{bmatrix}
\]
where \( \{z_{t|t-1}\}, \{z_{t|t}\}, \{P_{t|t-1}\}, \{P_{t|t}\} \) are obtained from the KF procedure above.
• Notice that, conditional on \( x_t \), the time varying parameters are fixed.
• Estimation is easy (MLE), given the assumption.
• TVP Example:

\[ r_t = \beta_t \lambda_t + w_t \]

\[ \beta_{t+1} - \bar{\beta} = F (\beta_t - \bar{\beta}) + \nu_{t+1} \]

We have a CAPM with TV $\beta_s$ in mind.

• If we assume that

\[
\begin{bmatrix}
  w_t \\
  v_{t+1} \\
  x_t, I_{t-1}
\end{bmatrix}
\sim N \left( \begin{bmatrix}
  0 \\
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  \sigma^2 & 0 \\
  0 & Q
\end{bmatrix} \right)
\]

then we are within the KF framework.

• Substituting the state variable $z_t = (\beta_t - \bar{\beta})$ in the space equation, we can write

\[ r_t = \lambda_t \bar{\beta} + \lambda_t z_t + w_t \]

• We can plug in the MLE estimator directly.

• Note: We can allow an AR(p) dynamics in the state equation quite easily.

• Example: Ludvigson and NG (JFE, 2007)

\[
m_{t+1} = a'F_t + \beta'Z_t + \varepsilon_{t+1}
\]

\[
VOL_{t+1} = c'F_t + d'Z_t + u_{t+1}
\]

where \( VOL_{t+1} \) is the realized volatility in month \( t + 1 \). (Observable)

5 Brandt and Kang (JFE, 2004):

\[
r_{t+1} = \mu_t + \sigma_t u_{t+1}
\]

\[
\begin{bmatrix}
\ln \mu_t \\
\ln \sigma_t
\end{bmatrix} = d + A \begin{bmatrix}
\ln \mu_{t-1} \\
\ln \sigma_{t-1}
\end{bmatrix} + \varepsilon_t
\]

The Delta Method
• We estimate \( y = \theta x + \varepsilon \), and obtain \( \hat{\theta} \) but are interested in a function \( g(\theta) \), where \( g(\.) \) is some non-linear model.

• Example: we have a forecast of the volatility, \( \hat{\sigma}_t \) and want to test its economic significance.

• Statistical measure of fit: \( \text{MSE} = E \left\{ (\hat{\sigma}_t - \sigma_t)^2 \right\} \)

• Economic measure of fit: \( C(\hat{\sigma}_t, S_t, K, r, T) \) and compare it to \( C(\sigma_t, S_t, K, r, T) \), where \( C(\.) \) is the BS call-option formula.

• We want to know whether \( C(\hat{\sigma}_t, S_t, K, r, T) - C(\sigma_t, S_t, K, r, T) \) is economically and statistically different from zero.
• The Delta method

• If we have a consistent, asymptotically normal estimator

\[ \sqrt{T} \left( \hat{\theta} - \theta \right) \xrightarrow{d} N(0, V) \]

and \( g(.) \) is differentiable, then

\[ \sqrt{T} \left( g \left( \hat{\theta} \right) - g \left( \theta \right) \right) \rightarrow^{d} N(0, D'VD) \]

\[ D = \frac{\partial g}{\partial \theta} \mid \theta \]

• Sketch of the proof: From the Mean-Value Theorem, we can write

\[ g \left( \hat{\theta} \right) = g \left( \theta \right) + \frac{\partial g'}{\partial \theta} \mid \theta^M \left( \hat{\theta} - \theta \right) \]

where \( \theta^M \) lies between \( \hat{\theta} \) and \( \theta \). Since, \( \hat{\theta} \rightarrow^p \theta \), then

\( \theta^M \rightarrow^p \theta \) and \( \frac{\partial g'}{\partial \theta} \mid \theta^M \rightarrow^p \frac{\partial g'}{\partial \theta} \mid \theta \) (Continuous Mapping Theorem).
Then, we can write

\[ \sqrt{T} \left( g \left( \hat{\theta} \right) - g ( \theta ) \right) = \frac{\partial g'}{\partial \theta} |\theta^M \sqrt{T} \left( \hat{\theta} - \theta \right) \]

\[ - \frac{\partial g'}{\partial \theta} |\theta^M \rightarrow^p \frac{\partial g'}{\partial \theta} |\theta \]

\[ - \sqrt{T} \left( \hat{\theta} - \theta \right) \rightarrow^d N(0, V) \]

- **Slutsky Theorem:** \( \frac{\partial g'}{\partial \theta} |\theta^M \sqrt{T} \left( \hat{\theta} - \theta \right) \rightarrow^d \left[ \frac{\partial g'}{\partial \theta} |\theta \right] N(0, V) \)

- Or

\[ \sqrt{T} \left( g \left( \hat{\theta} \right) - g ( \theta ) \right) \rightarrow^d N(0, \left[ \frac{\partial g'}{\partial \theta} |\theta \right] V \left[ \frac{\partial g}{\partial \theta} |\theta \right]) \]
• Example: We run the regression (s.e. in parentheses)

\[ y_t = \alpha + \beta x_t + \varepsilon_t \]
\[ = 0.1 + 1.1 x_t + \varepsilon_t \]
\[ (0.04) \quad (0.3) \]

• A test of \( \beta = 1 \) yields, \( t = (1.1 - 1) / 0.3 = 0.33 \)

• We are interested in \( \ln(\hat{\beta}) \) and testing under the null of \( \ln(\beta) = 0 \). From the delta method, we know that

\[ \sqrt{T} \left( \ln(\hat{\beta}) - \ln(\beta) \right) \rightarrow^d N(0, D^2 V) \]

where \( D = \frac{1}{1.1} = 0.91, V = 0.3^2 = 0.09 \), or

\[ \sqrt{T} (0.095 - 0) \rightarrow^d N(0, 0.91^2 0.09) \]

and a test of \( \ln(\beta) = 0 \) is \( t = 0.095 / 0.2862 = 0.33 \).
6 Empirical Portfolio Choice–Mean-Variance Implementation

- The solution to the mean-variance problem:

\[
\min_x \text{var} \left( r_{p,t+1} \right) = x' \Sigma x \\
\text{s.t.} \quad E \left( r_p \right) = x' \mu = \bar{\mu}
\]

is

\[
x^* = \frac{\bar{\mu}}{\mu' \Sigma \mu} \times \Sigma^{-1} \mu \\
= \lambda \Sigma^{-1} \mu
\]

- Now, we have to rely on econometrics, to implement the solution.

- Two step approach:
  - Solve the model
  - Estimate the parameters and plug them in!
PLUG-IN APPROACH:

We continue with the assumption that returns are i.i.d.

Then, we can estimate

\[ \hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} r_{t+1} \]

\[ \hat{\Sigma} = \frac{1}{T - N - 2} \sum_{t=1}^{T} (r_{t+1} - \hat{\mu}) (r_{t+1} - \hat{\mu})' \]

We plug in the estimates into the optimal solution

\[ \hat{x}^* = \frac{1}{\gamma} \hat{\Sigma}^{-1} \hat{\mu} \]

Under the normality assumption, this estimator is unbiased, or

\[ E(\hat{x}^*) = \frac{1}{\gamma} E(\hat{\Sigma}^{-1}) E(\hat{\mu}) \]

In the univariate case we can show by the delta method that

\[ Var(\hat{x}^*) = \frac{1}{\gamma^2} \left( \frac{\mu}{\sigma^2} \right)^2 \left( \frac{var(\hat{\mu})}{\mu^2} + \frac{var(\hat{\sigma}^2)}{\sigma^4} \right) \]
Example

Suppose we have 10 years of monthly data, or \( T = 120 \).

Suppose we have a stock with \( \mu = 0.06 \) and \( \sigma = 0.15 \).

Suppose that \( \gamma = 5 \).

Note that

\[
\hat{x}^* = \frac{1}{\gamma \sigma \mu} = \frac{1}{5 * 0.15^2} = 0.533
\]

Very close to the usual 60/40 advice by financial advisors!

With i.i.d. returns, the standard errors of the mean and variance are

\[
var(\mu) = \frac{\sigma}{\sqrt{T}} = \frac{0.15}{\sqrt{120}} = 0.014
\]

\[
var(\sigma^2) = \sqrt{2} \frac{\sigma^2}{\sqrt{T}} = \sqrt{2} \frac{0.15^2}{\sqrt{120}} = 0.003
\]

Plugging all these in the formula for \( \text{Var}(\hat{x}^*) \), we obtain

\[
\text{Var}(\hat{x}^*) = 0.14
\]

We can test hypotheses as with every other parameter of interest.
• Estimating $\Sigma$ is very problematic

• Many parameters to estimate
  – Suppose we have 500 assets in the portfolio. We have 125,250 unique elements to estimate.
  – In general, for N assets, we have $N(N + 1)/2$ unique elements to estimate!

• We need $\Sigma^{-1}$. Small estimation errors $\hat{\Sigma}$ results in very different $\hat{\Sigma}^{-1}$.

• Solution: Shrink the matrix
  $$\hat{\Sigma}^s = \delta S + (1 - \delta) \hat{\Sigma}$$

  where
  $$\delta \approx \frac{1}{T} \frac{A - B}{C}$$
  $$A = \sum_{i} \sum_{j} \text{asy var} \left( \sqrt{T} \hat{\sigma}_{i,j} \right)$$
  $$B = \sum_{i} \sum_{j} \text{asy cov} \left( \sqrt{T} \hat{\sigma}_{i,j}, \sqrt{T} s_{i,j} \right)$$
  $$C = \sum_{i} \sum_{j} \left( \hat{\sigma}_{i,j} - s_{i,j} \right)^2$$

  where $S$ is often taken to be $I$. For more discussions, see Ledoit and Wolf (2003)
• We can also shrink the weights directly
\[ x^s = \delta x_0 + (1 - \delta) x^* \]

• This approach is often used in applied work.

• Problem with shrinkage: Ad-hoc. No economic justification for it or for \( \delta \).

• Bayesian framework

• Economic constraints [Jagannathan and Ma (JF, 2003)]
Another solution: Factor models for stock $i$

$$r_{i,t} = \alpha_i + \beta_i f_m + \varepsilon_{i,t}$$

We can take variances to show that

$$\sum_r = \sigma_m^2 \beta \beta' + \sum_r$$

where $\beta$ is a vector of the betas and $\sum_r$ is a diagonal matrix with diagonal elements the variances of $\varepsilon_{i,t}$.

Now, the problem is reduced significantly!

What about time variation in $\mu$ and $\Sigma$!
7 Wold Decomposition: Stationary Processes

- Q: Isn’t the AR(1) (or ARMA(p,q)) model restrictive?
- No, because of the Wold decomposition result

Wold’s (1938) Theorem: Any zero-mean covariance stationary process \( Y_t \) can be represented in the form

\[
Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t
\]

where \( \psi_0 = 1 \) and \( \sum_{j=0}^{\infty} \psi_j^2 < \infty \) (square summable). The term \( \varepsilon_t \) is white noise and represents the linear projection error of \( Y_t \) on lagged \( Y_t \)'s

\[
\varepsilon_t = Y_t - E(Y_t|Y_{t-1}, Y_{t-2}, \ldots).
\]

The value \( \kappa_t \) is uncorrelated with \( \varepsilon_{t-j} \) for any \( j \) and is a purely deterministic term.

- Can we estimate all \( \psi_j \) in the Wold’s decomposition?
- The stationary ($|\phi| < 1$) AR(1) model can be written as

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

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

\[ Y_t = (1 - \phi L)^{-1} \varepsilon_t \]

\[ = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \]

or $\psi_j = \phi^j$. This is the restriction for the AR(1) model.

- The stationary ARMA(1,1) model can be written as

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

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

\[ Y_t = \frac{\varepsilon_t}{(1 - \phi L)} + \frac{\theta \varepsilon_{t-1}}{(1 - \phi L)} \]

\[ = \varepsilon_t + \sum_{j=1}^{\infty} \phi^{j-1} (\phi + \theta) \varepsilon_{t-j} \]

or $\psi_j = \phi^{j-1} (\phi + \theta)$.

- And so on.
Another interesting process: Fractionally differencing

\[ Y_t = (1 - L)^{-d} \varepsilon_t \]

where \( d \) is a number between 0 and 0.5.

It can be shown (Granger and Joyeux (1980), and Josking (1981)) that

\[ Y_t = \sum_{j=0}^{\infty} \eta_j \varepsilon_{t-j} \]

\[ \eta_j = \frac{1}{j!} (d + j - 1) (d + j - 2) (d + j - 3) \ldots (d + 1) \]

\[ \eta_j \approx (j + 1)^{d-1}, \text{ for large } j \]

Plot of \( \eta_j \) for \( d = 0.25 \) and \( \phi^j \) for \( \phi = 0.5 \) and \( \phi = 0.95 \)
There is a similar representation in the spectral domain

Spectral Representation Theorem [e.g., Cramer and Leadbetter (1967)]: Any covariance stationary process $Y_t$ with absolutely summable autocovariances can be represented as

$$Y_t = \mu + \int_0^\pi [\alpha(\omega) \cos(\omega t) + \delta(\omega) \sin(\omega t)] d\omega$$

where $\alpha(.)$ and $\delta(.)$ are zero-mean random variables for any fixed frequency $\omega \in [0, \pi]$. Also, for any frequencies $0 < \omega_1 < \omega_2 < \omega_3 < \omega_4 < \pi$, $\int_{\omega_1}^{\omega_2} \alpha(\omega) d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} \alpha(\omega) d\omega$ and the variable $\int_{\omega_1}^{\omega_2} \delta(\omega) d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} \delta(\omega) d\omega$.

Different (but equivalent) way of looking at a time-series.