Today’s Agenda

• Announcements

1. – No more homework (after GARCH)
   – Papers Due: May 18, 2008 (Sunday, 11:59pm)
   – Final: 2-sided, 8.5 by 11 inch cheat-sheet allowed.

2. Forecasting

3. Finish GMM

4. Maximum Likelihood Estimation
   – ARCH estimation
   – GARCH estimation

5. Kalman Filtering
   – Estimation–MLE

6. Non-Linear Dependence
You suspect that executing a trade will have a significant effect of the price of a stock (price-impact). Therefore, before placing the trade, you want to investigate what is the effect of a trade on the price of the asset of interest. You run the regression

\[ p_{t+1} = \alpha + \beta V o_l_t + \varepsilon_{t+1} \]

- Is this regression well specified?
- What other regression can you run to investigate the price impact of your trade?
1 Forecasting

- Thus far, the focus has been on in-sample fit.
- Forecasting is an important component of empirical work. We want to find
  \[ r_{t+1|t} = E(r_{t+1|x_t}) \]
- This is usually done with linear regressions
  \[ r_{t+1} = \alpha + \beta x_t + \varepsilon_{t+1} \]
- We can estimate this regression with data \( \{r_t, x_t\}_{t=1}^T \).
  (We will have \( T - 1 \) observations to run the regression).
- We obtain \( \hat{\alpha}_T, \hat{\beta}_T \) and \( \{\hat{\varepsilon}_t\} \)
- The in-sample fit and model selection can be addressed as discussed above.
- Perhaps the best way to evaluate this model is to see how it performs out-of-sample!
- Suppose we want to form a forecast of period \( T + 1 \).
- We form a forecast:
  \[ \hat{r}^{F}_{T+1|T} = \hat{\alpha}_T + \hat{\beta}_T x_T \]
- Then, we can wait for the true relation \( r_{T+1} \) and form a mean square error (MSE), aka (MSFE):
  \[ MSE = E \left( \hat{r}^{F}_{T+1|T} - r_{T+1} \right)^2 \]
• We can form one period, two period, ..., $K$ period forecasts in the same fashion.

• People also look at the $RMSE = \sqrt{MSE}$.

• The forecasts can be formulated with or without re-estimating the parameters (my preference: without re-estimation). Tradeoffs.

• The forecasts that we have considered thus far, $r_{t+1} = \alpha + \beta x_t + \varepsilon_{t+1}$ are linear.

• We can also have non-linear forecasts (nothing tells us that the expectation function has to be linear!)
  
  $r_{t+1} = g(\beta; x_t) + \varepsilon_{t+1}$

  where $g(.)$ is a function of the parameters $\beta$ and the data $x_t$.

• Note: The linear forecasts can always be motivated as a first-order approximation of the non-linear forecasts.

• If we start playing with non-linear forecasts, we might over-fit the data: high in-sample $R^2$ without an out-of-sample improvement in MSE.

• Over-fitting is a huge problem in applied work.
• Useful benchmark:
\[
\frac{E \left( r_{T+1} - \hat{r}_{T+1|T} \right)^2}{E \left( r_{T+1} - \bar{r}_T \right)^2} = \frac{MSE(\text{model})}{MSE(\text{no model})} = \text{relative improvement}
\]

• Problem: It is difficult to derive a distribution for this measure. It is more of a heuristic benchmark

• We can say that if we use the model, our forecasts improve by ___ percents.

• Backtesting a trading strategy

• Q: Is backtesting truly an out-of-sample experiment?
1.1 Data-Snooping.

- Very often, the analyst runs a few regressions (few specifications of the same relation of interest) and reports only the most significant one.

- This procedure induces a data-snooping bias. To see that, consider a test at the $\alpha = 0.05$ level.

- This means that even if the null is true, in 100 random samples 5 tests would turn out to be significant by random chance.

- This is a serious problem. Given that we work with the same datasets over and over, we know what tends to work and what does not.

- Hence, we are biasing ourselves into formulating hypotheses that work, just because of the prior experience with the data.
  - Recall: With non-stationary data, this happens with probability 1.

- But we only have one realization from the true data generating process (DGP).

- There are no real solution to the data-snooping problem

- Possible remedy: Out of Sample forecasting.

- (Leamer (1978), Lo and MacKinlay (1990)).
1.2 Practical Forecasting Considerations

We often have questions on how to best implement such forecasts. For instance:

- How much data to use (1-year, 5-years, 10-years)?
  - At what frequency?
  - At what horizon do we want to forecast?
  - What if the relation is not stable, i.e. if $\beta$ changes over time (e.g., CAPM)
1.3 Direct vs Indirect Forecasts

- In a direct forecast, we use low-frequency data to directly estimate the relation at the horizon of interest.

- In an iterated forecast, we use high-frequency data to estimate the model. The forecasts from the high-frequency data are iterated forward to the desired horizon.

- There is no general rule which approach is preferred.

- Ex: GARCH

  – Result: Iterated forecasts perform better.
1.4 Parameter Change

- Suppose that from $t = 1, \ldots, T_1$, we have the DGP:
  \[ y_t = \beta X_t + \varepsilon_t \]

- and from $T_1 + 1, \ldots, T$, we have a sudden change in parameters and the DGP is:
  \[ y_t = \gamma X_t + u_t \]

- First, is this plausible?

- Yes? Lucas (1972) Critique. People are not atoms. They might change their behavior.

- How do we deal with this problem:
  \[
  \begin{bmatrix}
  y_t \\
  y_t
  \end{bmatrix} = \begin{bmatrix}
  X_t & 0 \\
  0 & X_t
  \end{bmatrix} \begin{bmatrix}
  \beta \\
  \gamma
  \end{bmatrix} + \begin{bmatrix}
  \varepsilon_t \\
  u_t
  \end{bmatrix}
  \]

- This is nothing but two separate regressions. So we can run them separately. This is the unrestricted model with $\hat{\varepsilon}_t^u = \hat{\varepsilon}_t + \hat{u}_t$.

- The null hypothesis is that there is no change between the two samples, or $\gamma = \beta$. 
• The restricted model is: 
\[
\begin{bmatrix}
y_t \\
y_t
\end{bmatrix} = \begin{bmatrix}
X_t \\
X_t
\end{bmatrix} \begin{bmatrix}
\beta \\
\beta
\end{bmatrix} + \begin{bmatrix}
e_t \\
e_t
\end{bmatrix}
\]
which is nothing but estimating the regression over the entire sample, \( t = 1, \ldots, T \).

• The restricted SS is: 
\( \hat{\varepsilon}_t^r = \hat{e}_t \)

• Therefore, we can use the F-test to test the restriction. 
\[
F = \frac{\left( \sum (\hat{\varepsilon}_t^r)^2 - \sum (\hat{\varepsilon}_t^u)^2 \right) / k}{\sum (\hat{\varepsilon}_t^u)^2 / (T - 2k)}
\]
where \( k \) is the number of parameters in \( \beta \) (in our case, \( \beta = 1 \)), and \( T \) is the number of observations.

• There are many ways of specifying breaks in the parameters.
But this is the simplest way of modelling time-variation in the model.

Suppose that from \( t = 1, \ldots, T_1 \), we have the DGP:
\[
y_t = \beta X_t + \varepsilon_t
\]

and from \( T_1 + 1, \ldots, T \), we have a sudden change in the DGP as:
\[
y_t = \gamma X_t + \eta Z_t + \delta Z_t^{1-\alpha} + u_t
\]

We have a change not only in the parameters (from \( \beta \) to \( \gamma \)), but also in the functional form.

The problem is even worse. Often one would detect an instability in the parameters, but it would be attributed only to a parameter change (previous example).

People have a tendency to cling on to a particular model.

But the model changes (Think APT with factors coming in and out and in different forms!)

Is there hope of dealing with such uncertainty?

We have to always be aware of the possibility that our model might change!
● Another complication: In the previous examples, we assumed that we know the location of the break.

● In reality, we don’t!

● We can deal with this problem in the following way (Andrews (1998)):
  – Run a sequence of F tests at each point
  – Pick the maximum of the F-tests, \( \max_t F \)
  – This test has a supF distribution (not \( F \) because we have to take into account the uncertainty in the location of the break)
  – We can also deal with multiple breaks
    * Choose the second largest break
    * Choose the third largest break
    * Etc.
  – We have to start at some point, say 10% of the sample size on each side of the sample
  – The test has little power if the break happens to occur at the beginning of the sample.
2 GMM–Formal Treatment, Tests, Use and Misuse

- We start the estimation from an “orthogonality” condition:
  \[ E (h (w_t; \theta_0)) = 0 \]
  where
  - \( h (w_t; \theta) \) is a \( r \) dimensional vector of moment conditions, which depends on the data on some unknown parameters to be estimated.
  - The parameters are collected in vector \( \theta \) of dimension \( a \), where \( a \leq r \). The true value of \( \theta \) is denoted by \( \theta_0 \).
  - Note that \( h (., .) \) is a random variable.

- The “Method of Moments” principle states that we can estimate parameters by working with sample moments instead of population moments (Why?).

- Therefore, instead of working with
  \[ E (h (w_t; \theta_0)) = 0 \]
  which we cannot evaluate (Why?), we work with its sample analogue:
  \[ g (w_t; \theta) = \frac{1}{T} \sum_{t=1}^{T} h (w_t; \theta) \]
Example: OLS

\[ y_t = \beta x_t + \varepsilon_t \]
\[ E(x_t \varepsilon_t) = 0 \]
\[ E(x_t^2 \varepsilon) = 0 \]
\[ E(x_t^3 \varepsilon_t) = 0 \]

Here, we have three moment conditions \((r = 3)\), and one parameter to estimate \((a = 1)\).

You can think of

\[ h(w_t; \theta) = \begin{bmatrix} x_t (y_t - \beta x_t) \\ x_t^2 (y_t - \beta x_t) \\ x_t^3 (y_t - \beta x_t) \end{bmatrix} \]

\(w_t = (y_t, x_t)\) and \(\theta = \beta\).

We will work with the sample analogues

\[ g(w_t; \theta) = \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix} x_t (y_t - \beta x_t) \\ x_t^2 (y_t - \beta x_t) \\ x_t^3 (y_t - \beta x_t) \end{bmatrix} \]

Note, that from the e... theorem, we have

\[ g(w_t; \theta) \rightarrow^p E(h(w_t; \theta)) \]
Since there might be more moment conditions than parameters to estimate, we will work with the quadratic

\[ Q = g(w_1; \theta)' W_T g(w_1; \theta) \]

where \( W_T \) is a positive definite matrix that depends on the data.

The above quadratic can be minimized with respect to \( \theta \) using analytic or numerical methods (depending on the complexity of \( h \)).

It would be “logical” to put more weight on moments whose variance is smaller. Therefore, we want the matrix \( W_T \) to be inversely related to \( Var (h(w_1; \theta)) \), or \( W_T = Var (h(w_1; \theta))^{-1} \).

Before we pose the problem, we note that the weighing matrix \( Var (h(w_1; \theta))^{-1} \) does not take into account the dependence in the data. Therefore, we will work with

\[ \Gamma_j = E (h(w_t; \theta) h(w_{t-j}; \theta)), \quad j = 0, 1, 2, \ldots, \infty \]

\[ S = \sum_{j=0}^{\infty} \Gamma_j \]

The matrix \( S \) takes into account the dependence in the data.
– Long-run variance
– $2\pi$ spectrum at frequency zero.
• It turns out that we can prove (CLT with serially dependent data)
  \[ \sqrt{T} \left( g \left( w_t; \theta_0 \right) \right) \sim^a N(0, S) \]

• Note that if \( \Gamma_j = 0, j \geq 1 \) (serially independent data), then
  \[ S = \text{Var} \left( h \left( w_t; \theta \right) \right) = E \left( h \left( w_t; \theta \right) h \left( w_t; \theta \right) \right) . \]

• Finally we will let \( W_T = S_T^{-1} \)

• Therefore, the problem is:
  \[ Q = g \left( w_t; \theta \right)' S_T^{-1} g \left( w_t; \theta \right) \]

• The FOC is:
  \[ \left\{ \frac{\partial g}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g \left( w_t; \hat{\theta} \right) = 0 \]

• So, what are the properties of \( \hat{\theta} \)?
• Denote

\[ D'_{x=r} = \begin{bmatrix} \frac{\partial g_1(w_i; \hat{\theta})}{\partial \theta'} \\ \frac{\partial g_r(w_i; \hat{\theta})}{\partial \theta'} \end{bmatrix} \]

• We will show that

\[ \sqrt{T} \left( \hat{\theta} - \theta_0 \right) \sim^a N \left( 0, \left( DS^{-1}D' \right)^{-1} \right) \]
• The “proof” follows a few very simple steps
  – Use the Mean-Value theorem, to write
    \[ g(w_t; \hat{\theta}) = g(w_t; \theta_0) + D'_T \left( \hat{\theta} - \theta_0 \right) \]
  
  – Pre-multiply both sides by \( \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} \) to get
    \[ \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g(w_t; \hat{\theta}) \]
  
  = \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g(w_t; \theta_0) + \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} D'_T \left( \hat{\theta} - \theta_0 \right) \]
  
  – But, we know by definition that
    \[ \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g(w_t; \hat{\theta}) = 0 \]
Or
\[
\left\{ \frac{\partial g}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\} \quad \frac{S_T^{-1} g(w_t; \theta_0)}{S_T^{-1} D_T' \left( \hat{\theta} - \theta_0 \right)} = - \left\{ \frac{\partial g}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\} \quad \frac{S_T^{-1} D_T' \left( \hat{\theta} - \theta_0 \right)}{S_T^{-1} D_T'}
\]

- Rearranging, we get
\[
\left( \hat{\theta} - \theta_0 \right) = D_T^{-1} g(w_t; \theta_0)
\]

- Then,
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = D_T^{-1} \sqrt{T} g(w_t; \theta_0)
\]

- But, recall that
\[
\sqrt{T} \left( g(w_t; \theta_0) \right) \sim^a N(0, S')
\]

- Hence,
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) \sim^a N(0, D_T^{-1} S D_T^{-1'})
\]
\[
\sim^a N \left( 0, \left( D_T S^{-1} D_T' \right)^{-1} \right)
\]
Final Result: The GMM estimates are asymptotically normally distributed with a variance-covariance matrix equal to \( (D_T S^{-1} D_T')^{-1} / T \).

This is a huge result. All we needed was a set of moment conditions, nothing else!

We also need the data \( w_t \) to be stationary.

Many “standard” problems can be written as GMM.

The real power of GMM is that one framework can handle a lot of interesting problems.
It should be immediately obvious that the number of orthogonality conditions and the conditioning information matter

- In practice, the “type” of conditioning information will have a great impact on the estimates $\hat{\theta}$. Think instrumental variables.
- The question is: Which moments to choose?
- This is quite discomforting. If slight variations in our problem yields widely different estimates of $\theta$, what can we conclude?
Also: Estimating the matrix $S$ makes a huge difference. Recall that
\[
\Gamma_j = E(h(w_t; \theta) h(w_{t-j}; \theta))
\]
\[
S = \sum_{j=0}^{\infty} \Gamma_j
\]

Using sample analogues to obtain $\hat{\Gamma}_j$ and $\hat{S}$ is not the right way. Newey and West (1987) have proposed a “corrected” way, which is:
\[
\hat{S} = \hat{\Gamma}_0 + \sum_{v=1}^{q} \left(1 - \frac{v}{q+1}\right) \left(\hat{\Gamma}_j + \hat{\Gamma}'_j\right)
\]

Even the Newey-West method does not yield good results when the dimension of the system is large.

Moreover, the truncation point, $q$, introduces another source of error.
People have shown that small perturbations in $\hat{S}$ results in big differences in the estimates $\hat{\theta}$. In other words, suppose we use a matrix

$$\hat{S} + P$$

where $P$ has small values on its diagonals (perturbing the variances only). This results in widely different estimates. So, small differences in estimating $\hat{S}$ matter a lot.

The mechanics of why this is so reside in taking inverses...

Since we only need the optimal weighing matrix $S$ for efficiency (smallest variance), is it possible to find a matrix that, although not yielding efficient estimates, yields robust estimates?

In practice: The best (most robust) results are obtained with $I$, the identity matrix.

Empirical rule of thumb: Try the identity matrix first. Then, try the optimal weighing matrix, $\hat{S}$. If the results are widely different, stick with $I$. 
In his 1982 seminal paper, L.P. Hansen argued that the multiplicity of the moments, or the over-identifications, are an advantage, rather than a disadvantage.

- Even though we cannot have \( g(\hat{\theta}, w_t) = 0 \), it must be the case that at, and close to, the true value \( \theta_0 \), \( g() \) will be close to zero.

- Note that, since
  \[
  \sqrt{T}(g(w_t; \theta_0)) \sim^a N(0, S)
  \]
  then, it must be the case that
  \[
  Tg(w_t; \theta_0)'S^{-1}g(w_t; \theta_0) \sim^a \chi^2(r)
  \]

- It might be conjectured that the same would hold true for \( \hat{\theta} \), or that
  \[
  Tg(w_t; \hat{\theta})'S^{-1}g(w_t; \hat{\theta}) \sim^a \chi^2(r)
  \]

- However, this intuition is false, because it is not necessarily the same linear combination of \( g(w_t; \hat{\theta}) \) that would be close to zero. Instead, it can be shown that
  \[
  Tg(w_t; \hat{\theta})'S^{-1}g(w_t; \hat{\theta}) \sim^a \chi^2(r - a)
  \]

- Note: This statistic is trivial to estimate. Plug \( \hat{\theta} \) into \( g(.) \), etc.
The test

\[ J = T g \left( w_t; \hat{\theta} \right)' S^{-1} g \left( w_t; \hat{\theta} \right) \]

called the rank test, the test for over-identifying restrictions, Hansen’s J test, etc. has been used extensively in finance.

Indeed, people have relied exclusively on this test to judge the fit of their models.

Problems:
- As discussed above, the over-identifying restrictions are subject to the “which moments” critique.
- The J test also depends crucially on \( S \), which cannot be estimated accurately.

Not surprisingly, the J test rejects a lot of models. But, people are now aware of its deficiencies.
• The GMM framework is rich enough that we can think of many other ways of testing the hypotheses of interest. As a starting point, we can break the orthogonality restrictions into those that identify and those that over-identify the parameters

\[
E \left( \begin{array}{c} h_1 (w_t; \theta_0) \\ h_2 (w_t; \theta_0) \end{array} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

• Some people have suggested to see how the estimates would change as we add more restrictions to \( h_2 \) (say, starting from no over-identifying restrictions, and adding progressively).

• This set-up has also yielded insights into the stability properties of the moments and (or versus) the estimates.
- Example (Hansen, EMA 1982; Hansen and Singleton, EMA 1982):

\[ p_t = E_t \left( m(z_t | \theta) x_{t+1} \right) \]

- \( m(z_t | \theta) \) is the stochastic discount factor that depends on some state variables and parameters.
- \( x_{t+1} \) is the payoff from the asset.

- Example from last lecture:

\[ p_t = E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} x_{t+1} \right] \]

- We don’t need to log-linearize anything with GMM.
- Non-linearities are not a problem.
- Robustness is a real issue.
- But this is not what people use. Why?
• We can re-write the pricing equation as:

\[ 1 = E_t \left( m(z_t | \theta) \frac{x_{t+1}}{p_t} \right) \]

\[ E_t (m(z_t | \theta) R_{t+1}) - 1 = 0 \]

where \( R_{t+1} = x_{t+1}/p_t \). In the case of stocks, think \( x_{t+1} = p_{t+1} + d_{t+1} \).

• In this case, returns are stationary.

• If \( z_t \) in the stochastic discount factor is also stationary, then the ergoticity theorem conditions will be satisfied (under some other mild conditions).

• Example from last lecture:

\[ E_t \left[ \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \left(1 + R_{t+1}\right) \right] - 1 = 0 \]

• We have the data \( \{c_t, R_t\} \).

• Have to estimate one parameter, \( \gamma \) (assuming \( \beta \) is 0.995).
3 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

$$f_{Y_{T-1}Y_1}(\theta) = f_{Y_T}(\theta_T)f_{Y_{T-1}}(\theta_{T-1})...f_{Y_1}(\theta_1) = \Pi_{i=1}^{T} f_{Y_i}(\theta_i)$$

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

$$\Lambda(\theta) = \log f_{Y_{T-1}Y_1}(\theta) = \sum_{i=1}^{T} \log f_{Y_i}(\theta)$$
• 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

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

• The log likelihood can be expressed as

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

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

\[ = -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \left( \frac{\sigma^2}{(1 - \phi^2)} \right) \]

\[ - \frac{1}{2 \sigma^2 / (1 - \phi^2)} \left\{ y_1 - \left( \frac{c}{1 - \phi} \right) \right\}^2 \]

\[ - \frac{(T - 1)}{2} \log (2\pi) - \frac{(T - 1)}{2} \log (\sigma^2) \]

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

- 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 $\hat{\theta}^{\text{MLE}}$? More specifically, is it consistent? What is its distribution, where

$$\hat{\theta}^{\text{MLE}} = \arg \max \Lambda(\theta)$$

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

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

Result:

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

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

or

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

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

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

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

- 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)
  \[
f_{Y_2|Y_1} = \frac{1}{\sqrt{2\pi(\zeta + au_1^2)}} \exp \left[ -\frac{(y_2 - c - \phi y_1)^2}{2(\zeta + au_1^2)} \right]\]
Similarly, 

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

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

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

- 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):

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

where

\[ h_t = \zeta + \delta h_{t-1} + \alpha u_{t-1}^2 \]
4 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).
4.1 Setup (Hamilton CH 13)

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

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
  \[ E(w_t v'_t) = 0 \]
  \[ E(v_t v'_t) = Q, \ E(w_t w'_t) = R \]

- 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 $P_{t|t-1} = E\left((z_t - z_{t|t-1}) (z_t - z_{t|t-1})\right)$, 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 $P_{1|0} = E((z_{1|0} - z_1)(z_{1|0} - z_1))$
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

$$E(y_t - y_{t|t-1})' (y_t - y_{t|t-1})' = E[H'(z_t - z_{t|t-1})' (z_t - z_{t|t-1})' H] + E[w_tw_t']$$

$$= H'P_{t|t-1}H + R$$
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 $E(z_{t|t}|y_t, x_t, \ldots) = z_{t|t}$
   
   \[
   z_{t|t} = z_{t|t-1} + E \left( (z_t - z_{t|t-1}) (y_t - y_{t|t-1}) \right) \times \left( E (y_t - y_{t|t-1}) (y_t - y_{t|t-1})' \right)^{-1} (y_t - y_{t|t-1})
   \]
   – 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 $z_{t+1|t}$
     
     $E_t(z_{t+1}) = E(z_{t+1}|y_t, x_t, ...)$
     
     $= E(Fz_t + v_{t+1}|y_t, x_t, ...)$
     
     $= FE(z_t|y_t, x_t, ...) + 0$
     
     $= Fz_t|t$

   - We can use the above equation to write
     
     $E_t(z_{t+1}) = F\{z_{t|t-1}$
     
     $\quad + P_{t|t-1}H (H'P_{t|t-1}H + R)^{-1} (y_t - A'x_t - H'z_{t|t-1})$
     
     $= Fz_{t|t-1}$
     
     $\quad + FP_{t|t-1}H (H'P_{t|t-1}H + R)^{-1} (y_t - A'x_t - H'z_{t|t-1})$

   - We can also derive an equation for the error in forecast as a recursion
     
     $P_{t+1|t} = F[P_{t|t}$
     
     $\quad - P_{t|t-1}H (H'P_{t|t-1}H + R)^{-1} H'P_{t|t-1}]F' + Q$

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

- Anytime we have unobservable state variables (Brandt et al. (2005))
  \[ r_{t+1} = a + b\sigma_t + e_{t+1} \]
  \[ \sigma_t = \mu + \phi\sigma_{t-1} + w_t \]

- Interpolation of data (Bernanke (1989))

- Time varying parameters (Stock and Watson (1996)):
  \[ r_{t+1} = \alpha + \beta_t x_t + e_{t+1} \]
  \[ \beta_t = \eta + \phi\beta_{t-1} + v_t \]
5 Non-linear Dependence

- In finance, the characterization of the variation of returns and the dependence between returns is crucial.
- Thus far, the variation has been captured by the variance.
- The dependence has been captured by the covariance.
- The covariance is a good measure only with linear models.
- Non-linear dependence needs to be addressed.
- Suppose we have two assets $R$ and $F$ with returns $R_S$ and $R_F$ whose distribution is $F_{R_S}$ and $F_{R_F}$.
- The “complete” behavior is defined by their distributions.
• We might want to know how the entire distribution of the two variables. In other words, we want to know the joint distribution, $F_{R_S,R_F}$.

• Q: Why?

• A: Because there might be nonlinearity in the relation between the two variables.

• Find a general way of relating the two distributions, $F_{R_S}$ and $F_{R_F}$

• For instance, how would the tails of the distributions change during extreme events? More? Less?
  – With credit derivatives, suppose that the credit rating of the underlying (Bond) changes.

• Or, can we model the dependence non-linearly?

• Would we see skewness, kurtosis, etc.?

• There is a lot of interest in the profession in finding adequate (alternative) measures of dependence that go beyond simple correlations.
There are two (three) main ways to go:

- Copulas
- Tail dependence
- Quantile Regressions

These are different approaches at modeling the non-linear dependence between the distributions $F_{RS}$ and $F_{RF}$.
5.1 Copulas

The problem is as follows. Find a general way of relating the two distributions, $F_{RS}$ and $F_{RF}$ in order to find the joint distribution $F_{RS,RF}$.

- In other words, we want to do the following
  \[ F_{RS,RF} = C(F_{RS}, F_{RF}) \]
  where $C(., .)$ is some appropriately chosen function.

- The goal is to choose the right $C(., .)$. 
• Note: We cannot choose any $C(.,.)$. There are some restrictions that are imposed by the fact that we are dealing with distributions.

• Simple example; $C(x, y) = x \cdot y$
  – Q: When is this copula appropriate?

• Another example: $C(x, y) = \min(x, y)$

• Another example: $C(x, y) = N_\rho \left( N^{-1}(x), N^{-1}(y) \right)$ where $\rho$ is the correlation.
● Two widely used copulas:
  – Gumbel’s bivariate exponential:
    \[ C_\theta(x, y) = x + y - 1 + (1 - x)(1 - y) e^{-\theta \ln(1 - x)(1 - y)} \]
    where \( \theta \) measures the non-linear dependence.
    For \( \theta = 0 \), we have
    \[ C_{\theta=0}(x, y) = x \cdot y \]
  – AMH Copulas (Ali-Mikhail-Haq):
    \[ C_\theta(x, y) = \frac{xy}{1 - \theta (1 - x)(1 - y)} \]
    where \( \theta \) measures the non-linear dependence.
    For \( \theta = 0 \), we have
    \[ C_{\theta=0}(x, y) = x \cdot y \]
● In general, if you work with copulas:
  – You want to find a flexible form for $C(.,.)$
  – You want to be able to estimate all the dependence parameters.
5.2 Tail dependence

The main question is as follows. What is the probability that we will have an extreme (tail) realization of $R_F$ given that we observe an extreme tail realization of $R_S$?

- Note: We are not looking at typical co-movements (covariance) but at tail co-movements?

- A natural definition of tail dependence is:
  - On the positive side:
    \[
    \lambda_+ = \Pr (R_S > \text{extreme positive value} \mid R_F > \text{extreme positive value}) \\
    = \lim_{u \to 1^-} \Pr (R_S > F_{R_s}^{-1} (u) \mid R_F > F_{R_F}^{-1} (u))
    \]
  - On the negative side:
    \[
    \lambda_- = \Pr (R_S < \text{extreme negative value} \mid R_F > \text{extreme negative value}) \\
    = \lim_{u \to 0} \Pr (R_S < F_{R_s}^{-1} (u) \mid R_F < F_{R_F}^{-1} (u))
    \]
• I.e., what is the chance of observing a very positive (negative) return in the underlying asset, given that we have observed a very positive (negative) return in the hedging asset.

• Instead of average dependence, we are looking at tail dependence.

• Intuitively, copulas and tail dependence are related.

• There are also mathematical connections between the two:

\[
\lambda_+ = \lim_{u \to 1} \frac{1-2u+C(u,u)}{1-u},
\]

\[
\lambda_- = \lim_{u \to 0} \frac{C'(u, u)}{u}
\]

• For instance, for the Gumbel’s copula

\[
\lambda_+ = 1 - 2^\theta
\]

• Note: \( \lambda_+ = 0 \) means no tail dependence (when \( \theta = 0 \)). But the distributions might be otherwise dependent.
5.3 Quantile regression:

- Model the conditional quantiles of the distribution
- Instead of
  \[ R_F = \alpha + \beta R_S + \varepsilon \]
  where we model the conditional mean, we want to model the conditional quantiles.
- We do that by specifying:
  \[ Q_{R_F}(\tau) = \alpha(\tau) + \beta(\tau) R_S + \varepsilon \]
  for \( \tau \) between 0 and 1 where \( Q_{R_F}(\tau) \) is the \( \tau \) quantile of \( R_F \).
- Recall the definition of a quantile: (inverse of distribution).
- This is called a quantile regression, because we are modeling the quantiles of the distribution.
  - For \( \tau = 0.5 \), we are estimating how the median of the distribution of \( R_F \) depends on \( R_S \).
  - Similarly, for \( \tau = 0.05 \) or \( \tau = 0.99 \), we are estimating the extreme quantiles of the distribution.
- Quantile regressions are easy to estimate (similar to OLS).