1. Today’s Agenda

1. ARCH/GARCH
2. GMM (Intro by Example)
3. GMM (Formal Treatment)
4. A Realistic Example: Chan et al. (1992)
2 Autoregressive Conditional Heteroskedasticity (ARCH)

- Suppose we have the returns process \( \{ r_t \}_{t=1}^T \).
- First, we model the conditional mean, for example as
  \[
  r_t = \mu + \phi r_{t-1} + u_t
  \]
- We know that the unconditional first moments are
  \[
  E(r_t) = \frac{\mu}{1-\phi} \quad \text{and} \quad Var(r_t) = \frac{\sigma_u^2}{1-\phi^2}
  \]
- We also know that the conditional first moment
  \[
  E_{t-1}(r_t) = \mu + \phi r_t
  \]
  is time varying, even though the unconditional moment is not!
- Q: Can we also have the same situation for the second conditional moment, i.e. to have a time-varying conditional second moment, although the unconditional second moment is constant over time?
- A: Yes.
- Note: The unconditional second moment of \( u_t \) is \( \sigma_u^2 \).
• Motivation: Do we need to model the conditional second moment of returns and interest rates?
• A: Let’s see what the data shows.
• Intuitively, we want to model $u_t^2$ to follow an AR process just as we had $r_t$ follow an AR process.

• We can write such a process as

$$u_t^2 = \zeta + \alpha u_{t-1}^2 + w_t$$

where $w_t$ is a white noise process with $E(w_t^2) = \sigma_w^2$.

• Note that $E_{t-1}(u_t^2) = \zeta + \alpha u_{t-1}^2$, so that the conditional moment is time-varying although $E(u_t^2) = \sigma_u^2$.

• The above process is called an autoregressive conditional heteroskedasticity model of order 1, or ARCH(1).

• We can generalize to an ARCH(p) model as:

$$u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2 + w_t$$
• Note that a variance cannot be negative. We need to place certain restrictions on
\[ u_t^2 = \zeta + \alpha u_{t-1}^2 + w_t \]
in order to insure that \( u_t^2 \) is always positive.

• We need:
  - \( w_t \) to be bound from below by \(-\zeta\), where \( \zeta > 0 \).
  - \( \alpha \geq 0 \).
  - For covariance stationarity of \( u_t^2 \), we also need \( \alpha < 1 \) as in the other AR model.

• With all those conditions, we can see that
\[ \text{Var} (u_t) = E (u_t^2) = \sigma_u^2 = \frac{\zeta}{1 - \alpha} \]
• This is the ARCH(1) model.
• We have written it in its autoregressive form, similar to the AR processes.
• It is more convenient, but less intuitive, to present the ARCH(1) model as:

\[ u_t = \sqrt{h_t} v_t \]

where \( v_t \) is iid with mean 0, and \( E(v_t^2) = 1 \).

• Suppose that

\[ h_t = \zeta + a u_{t-1}^2 \]

then combining the above equations, we obtain:

\[ u_t^2 = h_t v_t^2 \]

• Now, since \( v_t \) is iid then

\[
E_{t-1}(u_t^2) = E_{t-1}(h_t^2 v_t^2) \\
= E_{t-1}(h_t^2) E_{t-1}(v_t^2) \\
= \zeta + a u_{t-1}^2
\]

as before.
• Reconciling the two definitions.
• From one side, we have $u_t^2 = h_t v_t^2$
• From another side, we have $u_t^2 = h_t + w_t$. Therefore
  \[ h_t v_t^2 = h_t + w_t \]
  or
  \[ w_t = h_t (v_t^2 - 1) \]
• From here, we can see that $E_{t-1} (w_t^2)$ is time varying, whereas $E (w_t^2) = \sigma_w^2$
• The ARCH process gives us conditional heteroskedasticity, but it turns out that $E_{t-1} \left( u^2_t \right)$ has turned out to be a very persistent process (in the data).

• We can capture such a process with an ARCH(p) process

$$u^2_t = \zeta + \alpha_1 u^2_{t-1} + \alpha_2 u^2_{t-2} + \ldots + \alpha_p u^2_{t-p} + w_t$$

where $p$ is very large.

• This solution is inefficient. There are too many parameters to estimate!

• Q: What to do?

• A: GARCH.

• The GARCH, Generalized ARCH allows us to capture the persistence of conditional volatility in a parsimonious way.
• GARCH: Recall that we could write:
\[ u_t = \sqrt{h_t} v_t \]
where \( h_t = \zeta + au_{t-1}^2 \) for an ARCH process.

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

• The direct link between \( h_t \) and \( h_{t-1} \) is exactly what is needed to capture the dependence between \( \sigma_t^2 \) and \( \sigma_{t-1}^2 \).

• A process with \( h_t = \zeta + \delta h_{t-1} + au_{t-1}^2 \) is called a GARCH(1,1).

• The unconditional second moment of \( u_t \) is:
\[ E(u_t^2) = \sigma^2_u = \frac{\zeta}{1 - (\alpha + \delta)} \]

• To insure stationarity, we need to restriction:
\[ \delta + \alpha < 1. \]
• Of course, we can generalize to a GARCH(p,q) as:
\[ u_t^2 = \zeta + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \ldots + \delta_p h_{t-p} + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2 \]

• Similarly, the parameters of the conditional and the unconditional second moment are related by:
\[
E \left( u_t^2 \right) = \frac{\zeta}{1 - (\alpha_1 + \ldots + a_q + \delta_1 + \ldots + \delta_p)}
\]

• Also, \( \alpha_i \geq 0 \) and \( \delta_j \geq 0 \)
• For most practical purposes a GARCH(1,1) is GREAT.
  – There is a trade-off. You introduce more parameters to capture the accurate dynamics, but there are more parameters to estimate.
  – Those parameters have restrictions. The estimation is tricky.
  – Bottom line, for 95% of the applications, GARCH(1,1) does a great job.

• GARCH is successful, because it can capture the persistence in $E_{t-1}(u_t^2)$, which is the most significant feature that needs to be captured.
Another useful model to estimate is the IGARCH model, or integrated GARCH

The IGARCH(1,1) is a GARCH(1,1) where

$$\delta + \alpha = 1$$

If this condition is satisfied, it can be shown that the conditional variance of $$u_t$$ is infinite.

The processes $$u_t$$ and $$u_t^2$$ are not covariance stationary.

However, the process $$u_t$$ is stationary (i.e. its conditional density does not depend on $$t$$).

The IGARCH is important because it captures the important case of a strong dependence that leads to non-stationarity.

Q: Do variances have to be stationary?
The GARCH literature has gone crazy chasing after the perfect conditional heteroskedasticity model. Some of the models we have are:

- ARCH in Means
- Exponential GARCH
- Nonlinear GARCH
- Asymmetric GARCH
- Fractionally Integrated GARCH (FIGARCH)
- ABS.ASYM.M.FIGARCH ????
• The interest in forecasting conditional volatility using past volatility has been greater in the applied field more so than in academia.

• The GARCH models are unsatisfactory, from an economic perspective, because
  – Explaining vol with past vol tells us nothing about the underlying economic factors that cause the volatility to move.
  – If a structural break occurs, a recursive model will fail miserably...a structural model might not.
  – In general, the GARCH models are difficult to generalize to a multivariate setting.
Q: Why not use (Schwert, French, and Stambaugh (1989))

\[ \hat{\sigma}_t^2 = \left( \frac{1}{n} \sum_{i=1}^{n} r_{i,t}^2 \right) \times 22 \]

as a “non-parametric” method, where \( r_{i,t} \) is the \( i \)-th daily return in month \( t \).

ARCH and GARCH methods are called “parametric” methods.

Q: Which one is better?

A: It depends on the application at hand. For forecasting, it is difficult to beat a GARCH(1,1).
Future Research:

- The conditional volatility literature is huge.
- There are very few papers on conditional correlation.
- But, understanding how and why correlations vary with time is crucial in finance.
  - Think of a portfolio choice problem (correlations matter a lot)
  - Think of a CAPM model. The correlation between the market return and the return of an asset is assumed constant (and, hence the risk premium is constant), but it need not be.
- There are no good models that try to capture the correlation as function of macroeconomic factors (except SVV (2001))—good for a research paper.
• This is great, but how do we estimate a GARCH model.
• OLS clearly does not work.
• We need a method that would allow us to do non-linear estimation
• We also need a general method that we can apply to any nonlinear problem, with minimal assumptions.
• We do not want to have assumptions on the distribution of the residuals.
• Therefore, we need GMM (Generalized Method of Moments)
3 Simple Introduction to GMM

• Recall that any variable $x_t$ has a distribution $F_x(x)$. If $x$ has moments $E(x^j), j = 1, \ldots$ then those moments can be used to retrieve $F_x(x)$.

• Caution: Some variables do not have moments (Cauchy distribution case).

• Suppose we have random variables $x_t, y_t, z_t$.
  - A population moment of those variables is $E[g(x_t, y_t, z_t)]$
  - A sample moment of those variables is $\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t)$
  - By the ergodicity theorem (or the LLN in cross section, we know that) $\frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t) \rightarrow^{p} E[g(x_t, y_t, z_t)]$
    under some mild conditions on the function $g(.)$. 

• In other words, the distance between the sample and the population moment goes to zero in probability as $T \to \infty$:

$$\left\{ \frac{1}{T} \sum_{t=1}^{T} g(x_t, y_t, z_t) - E[g(x_t, y_t, z_t)] \right\} \to^p 0$$

• Can we use this “insight” to estimate parameters. Suppose that the function $g$ depends not only on the data but also on the unknown parameters, $\theta$.

• We want to choose the parameter $\theta$ in order to minimize the distance between the data and the population moment.

• In a simpler example, let’s concentrate on a univariate case. Then $g(x|\theta) = \mu$, the population mean. In other words, $\theta = \mu$.

• The problem becomes (trivially):

$$\left\{ \frac{1}{T} \sum_{t=1}^{T} x_t - \mu \right\} \to^p 0$$
Here is a more interesting example: OLS as GMM

The model is:
\[ y_t = x_t\beta + \varepsilon_t \]

The FOC in the OLS case could be written as:
\[ E(x_t\varepsilon_t) = 0 \]

This is a moment condition that also depends on parameters. To see that, write
\[ E(x_t(y_t - x_t\beta)) = 0 \]
\[ E(x_t y_t) = \beta E(x_t^2) \]
\[ \beta = \frac{E(x_t y_t)}{E(x_t^2)} \]

Therefore, approximating the population means by their sample analogues, we get
\[ \frac{1}{T} \sum_{t=1}^{T} (x_t(y_t - x_t\beta)) = 0 \]
\[ \frac{1}{T} \sum_{t=1}^{T} x_t y_t = \beta \frac{1}{T} \sum_{t=1}^{T} x_t^2 \]
\[ \hat{\beta} = \frac{\frac{1}{T} \sum_{t=1}^{T} x_t y_t}{\frac{1}{T} \sum_{t=1}^{T} x_t^2} \]
• But we can also write another moment condition:
\[ E(x_t^2 \varepsilon_t) = 0 \]
• Then, as above
\[ E(x_t^2 (y_t - x_t \beta)) = 0 \]
\[ \beta = \frac{E(x_t^2 y_t)}{E(x_t^3)} \]
• Therefore, using sample moments to approximate population moments, we get
\[ \hat{\beta}_2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t \]
\[ \frac{1}{T} \sum_{t=1}^{T} x_t^3 \]
• We can also use
\[ E(g(x_t) \varepsilon_t) = 0 \]
for some function \( g(.) \). Note: You should also be able to show that \( E(x_t \varepsilon_t) = 0 \) implies \( E(g(x_t) \varepsilon_t) = 0 \). Then, for a known function \( g(.) \)
\[ \hat{\beta}_g = \frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t \]
\[ \frac{1}{T} \sum_{t=1}^{T} g(x_t) x_t \]
• Oupss! Problem. We have one parameter, $\beta$, but three possible estimators

$$\hat{\beta} = \frac{1}{T} \sum_{t=1}^{T} x_t y_t \rightarrow^p \beta$$

$$\hat{\beta}_2 = \frac{1}{T} \sum_{t=1}^{T} x_t^2 y_t \rightarrow^p \beta$$

$$\hat{\beta}_g = \frac{1}{T} \sum_{t=1}^{T} g(x_t) y_t \rightarrow^p \beta$$

• Which one do we choose?

• Result: Under some very restrictive assumptions (i.e. exogeneity of $x_t$, homoskedasticity, uncorrelated $\varepsilon_t$, etc), the OLS is the best linear unbiased estimator (BLUE).

• In other words, in has the smallest variance among all linear unbiased estimators.

• However, who knows if those assumptions are satisfied. In all likelihood, they are not.

• Q: Can we stack all the moments in a vector as

$$E(g(\beta|x)) = E\left[\begin{array}{c} x_t \varepsilon \\ x_t^2 \varepsilon \\ g(x_t) \varepsilon_t \end{array}\right] = 0$$

and choose the value of $\beta$ that satisfies the three sample moments?
• A: Off course, not! Three equations, potentially nonlinear, with only one unknown....Who knows how many solutions there are, if any.

• But, we can construct a quadratic function, as:

$$E \left( g \left( x \mid \beta \right)' W g \left( x \mid \beta \right) \right) = 0$$

for some symmetric positive definite matrix $W$.

• Now, we have the information into the three equations, weighted by the elements of the matrix $W$.

• Problem: What matrix $W$ to choose?

• A: Any symmetric positive matrix will give us consistent estimates (i.e. $\hat{\beta}_W \rightarrow^p \beta$), but we are concerned with efficiency, or smallest possible standard errors around $\hat{\beta}_W$. 
• ENDOGENEITY: Instrumental Variables (IV) and GMM.

• By construction, we had $E(\varepsilon|x) = 0$ implied that $E(\varepsilon x) = 0$. In other words, the residuals and the explanatory variables are uncorrelated.

• However, in structural models, it is often the case that we want to run regressions when this requirement is not satisfied. For example:

$$FirmValue_t = \alpha + \beta Debt_t + \varepsilon_t$$

• But it is not reasonable to assume that Debt is an exogenous variable. For example, new (relatively low Firm Value) firms do not have access to debt. Indeed, we might try to run the opposite regression:

$$Debt_t = \delta + \zeta FirmValue_t + \upsilon_t$$

• So, here

$$E(Debt_t\varepsilon_t) = E((\delta + \zeta FirmValue_t + \upsilon_t)\varepsilon_t) \neq 0$$

• Q: If $E(Debt_t\varepsilon_t) \neq 0$, can we still have $\hat{\beta} \rightarrow^p \beta$?

• Breaking the $E(Debt_t\varepsilon_t) = 0$ condition is the cardinal sin in empirical work!!!

• Q: What to do?

• Note: We can argue that most equations in finance suffer from this endogeneity problem.
• Well, we can look for a variable, $Z_t$ which is:
  - Correlated with $Debt_t$ (it proxies for $Debt_t$)
  - Is uncorrelated with $\varepsilon_t$.

• If such a variable exists, then we can write the moment condition:
  \[ E(Z_t\varepsilon_t) = 0 \]

• Or, if we look at the sample moments, then
  \[
  \frac{1}{T} \sum_{t=1}^{T} Z_t\varepsilon_t = \frac{1}{T} \sum_{t=1}^{T} Z_t(y_t - \beta x_t) \\
  \frac{1}{T} \sum_{t=1}^{T} Z_t y_t = \beta \frac{1}{T} \sum_{t=1}^{T} Z_t x_t \\
  \]

\[
\hat{\beta}_{IV} = \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t y_t}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t} = \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t (\beta x_t + \varepsilon_t)}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t} \\
= \beta + \frac{\frac{1}{T} \sum_{t=1}^{T} Z_t \varepsilon_t}{\frac{1}{T} \sum_{t=1}^{T} Z_t x_t}
\]

• Then we can show that $\hat{\beta}_{IV} \rightarrow_p \beta$, whereas $\hat{\beta}_{ols}$ does not.

• This estimator was motivated from GMM.
• All this is great, but how do we choose the instrument $Z_t$?
• This is usually the big question.
• Usually, $Z_t = Debt_{t-k}$, because
\[
E(Z_t \varepsilon_t) = E(Debt_{t-k} \varepsilon_t) = E((\delta + \zeta Firm Value_{t-k} + \nu_{t-k}) \varepsilon_t)
\]
ARCH/GARCH Estimation using GMM:

- Recall the model:
  \[ y_t = x_t \beta + u_t \]
  \[ u_t^2 = h_t + w_t = \zeta + a u_{t-1}^2 + w_t \]

- The moment conditions are:
  \[ E(u_t x_t) = E((y_t - x_t \beta) x_t) = 0 \quad (1) \]
  \[ E\left(u_t^2 - \frac{\zeta}{1 - \alpha}\right) = 0 \]
  \[ E(w_t z_t) = E\left((u_t^2 - \zeta + a u_{t-1}^2) z_t\right) = 0 \]

- Note: 3 moments and three parameters (\(\beta, \zeta, \alpha\)).

- Replace those conditions by their sample analogues:
  \[ \frac{1}{T} \sum_{t=1}^{T} \left((y_t - x_t \beta) x_t\right) \]
  \[ \left(u_t^2 - \frac{\zeta}{1 - \alpha}\right) \]
  \[ \left((u_t^2 - \zeta + a u_{t-1}^2) z_t\right) = 0 \]

- Solve for (\(\hat{\beta}, \hat{\zeta}, \hat{\alpha}\)).
• GMM is a very powerful way of looking at an estimation problem.
• All we need is a moment condition that holds.
• The problem does not have to be linear.
• No distributional assumptions are needed.
• We can use GMM to estimate
  – The non-linearized version of the Consumption CAPM.
  – Nonlinear process, such as ARCH, GARCH, etc.
  – Interesting interest rate models (Chan et al (1992)).
4 GMM–Formal Treatment, Tests, Use and Missuse

• 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

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

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 \( \text{Var}(h(w_t; \theta)) \), or \( W_T = \text{Var}(h(w_t; \theta))^{-1} \).

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

\[
\Gamma_j = E(h(w_t; \theta) h(w_{t-j}; \theta))
\]

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

The matrix \( S \) takes into account the dependence in the data.
• It turns out that we can prove (CLT with serially dependent data)
  \[ \sqrt{T} (g (w_t; \theta_0)) \sim^a N(0, S) \]

• Note that if \( \Gamma_j = 0, j \geq 1 \) (serially independent data), then
  \( S = Var (h (w_t; \theta)) = E (h (w_t; \theta) h (w_t; \theta)) \).

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

• Therefore, the problem is:
  \( Q = g (w_t; \theta)' S_{-1}^T g (w_t; \theta) \)

• The FOC is:
  \[ \left\{ \frac{\partial g}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right\}' S_{-1}^T g (w_t; \hat{\theta}) = 0 \]

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

\[
D_T^{'} = \begin{bmatrix}
\frac{\partial g_1 (w_t; \hat{\theta})}{\partial \theta'}
\frac{\partial g_r (w_t; \hat{\theta})}{\partial \theta'}
\end{bmatrix}
\]

• 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
    \[ g(w_t; \hat{\theta}) = g(w_t; \theta_0) + D'_T (\hat{\theta} - \theta_0) \]
  – 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 (\hat{\theta} - \theta_0)
    – But, we know by definition that
      \[
      \left\{ \frac{\partial g}{\partial \theta'} \right|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g(w_t; \hat{\theta}) = 0
      \]
\[
\left\{ \frac{\partial g}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\}' S_T^{-1} g (w_t; \theta_0) = - \left\{ \frac{\partial g}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\}' S_T^{-1} D_T' \left( \hat{\theta} - \theta_0 \right)
\]

- 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

\[
\sqrt{T} (g (w_t; \theta_0)) \sim^a N(0, S)
\]

- 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 \( \left( D_T S^{-1} D_T' \right)^{-1} / T \).

• 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

- 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
\[
\Gamma_j = \mathbb{E} \left( h \left( w_t; \theta \right) h \left( w_{t-j}; \theta \right) \right)
\]
\[
S = \sum_{j=0}^{\infty} \Gamma_j
\]

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

- It might be conjectured that the same would hold true for \( \hat{\theta} \), or that
  \[
  T g(w_t; \hat{\theta})' S^{-1} g(w_t; \hat{\theta}) \sim^a \chi^2(r)
  \]

- 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
  \[
  T g(w_t; \hat{\theta})' S^{-1} g(w_t; \hat{\theta}) \sim^a \chi^2(r - a)
  \]

- 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.
Realistic Example: Chan, Karolyi, Longstaff, and Sanders (1992, JF):

- **Motivation**: Estimate a “very” general process for the short rate.

- **Q**: Why?

- **A**: Because, many models in finance have to specify a short rate. Usually they specify the short rate process that is most convenient, i.e. the short rate process that allows them to derive nice, simple, closed-form results.

- But one has to realize that the results from those models are conditional upon having the right short-rate process.
The proposed process for the short rate is:
\[ dr = (\alpha + \beta r) dt + \sigma r^\gamma dZ \]

- The short rate is written in continuous time to relate it to most finance models.
- The estimation is done in discrete time.
- The discretization is done by taking that \( dt \) is one month.
- No other adjustments are made.

1. Merton
2. Vasicek
3. CIR SR
4. Dothan
5. GBM
6. Brennan-Schwartz
7. CIR VR
8. CEV

\[
\begin{align*}
\text{Merton:} & \quad dr = \alpha dt + \sigma dZ \\
\text{Vasicek:} & \quad dr = (\alpha + \beta r) dt + \sigma dZ \\
\text{CIR SR:} & \quad dr = (\alpha + \beta r) dt + \sigma r^{1/2} dZ \\
\text{Dothan:} & \quad dr = \sigma rdZ \\
\text{GBM:} & \quad dr = \beta rd t + \sigma rdZ \\
\text{Brennan-Schwartz:} & \quad dr = (\alpha + \beta r) dt + \sigma rdZ \\
\text{CIR VR:} & \quad dr = \sigma r^{3/2} dZ \\
\text{CEV:} & \quad dr = \beta rd t + \sigma r^\gamma dZ \\
\end{align*}
\]

### Parameter Restrictions Imposed by Alternative Models of Short-Term Interest Rate

Alternative models of the short-term riskless rate of interest \( r \) can be nested with appropriate parameter restrictions within the unrestricted model
\[ dr = (\alpha + \beta r) dt + \sigma r^\gamma dZ \]

<table>
<thead>
<tr>
<th>Model</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( \sigma^2 )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merton</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Vasicek</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CIR SR</td>
<td>0( \gamma )</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dothan</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GBM</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Brennan-Schwartz</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0( \gamma )</td>
</tr>
<tr>
<td>CIR VR</td>
<td>0</td>
<td>0</td>
<td>0( \gamma )</td>
<td>1</td>
</tr>
<tr>
<td>CEV</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Aside:

• The above problems assume a particular form of the variance process. They are called “stochastic volatility” models (when $\gamma \neq 0$).

• Note the difference in modelling. Recall the ARCH/GARCH literature.

• In ARCH/GARCH, the focus is on the second conditional moment.

• In the “stochastic volatility” models, the entire process is modelled and the variance depends on the level of the process.

• In practice and in implementation, there is little difference.
Implementation:


- The short rate is:

  \[ \Delta r_{t+1} = \alpha + \beta r_t + \varepsilon_{t+1} \]

  or

  \[ r_{t+1} = \alpha + (1 + \beta) r_t + \varepsilon_{t+1} \]

  \[ E(\varepsilon_{t+1}^2) = \sigma^2 r_t^{2\gamma} \]

- Note that while the continuous time processes assumed that \( dZ_t \) is normally distributed, this is purely for tractability.

- In econometrics, we don’t need to make sure an assumption. The shock \( \varepsilon_{t+1} \) can have any distribution.

- The parameters to be estimated are: \( \alpha, \beta, \gamma, \) and \( \sigma^2 \).
GMM

- Specify the moment conditions (We need at least 4 moment conditions. Why?)

\[
E(h(w_t; \theta)) = 0
\]

\[
E \left( \begin{array}{c}
\varepsilon_{t+1} \\
\varepsilon_{t+1} r_t \\
\varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma} \\
(\varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma}) r_t
\end{array} \right) = 0
\]

where \( w_t = (r_{t-1}, r_t) \), \( \theta = (\alpha, \beta, \gamma, \sigma^2) \).

- The sample moment is:

\[
g(w_t; \theta) = \frac{1}{T} \sum_{t=1}^{T} \left[ \begin{array}{c}
\varepsilon_{t+1} \\
\varepsilon_{t+1} r_t \\
\varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma} \\
(\varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma}) r_t
\end{array} \right]
\]

- We are done! For the rest, follow the algorithm.

- The same algorithm applies for all problems that can be written as a moment condition.
We could also proceed in the following way:

- Regress $r_t$ on $r_{t-1}$ using OLS. Get estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\epsilon}_t$.
- Estimate $\sigma^2$ and $\gamma$ using non-linear least squares:
  \[
  \min_{\sigma^2, \gamma} \sum_{t=1}^{T} \left( \hat{\epsilon}_{t+1}^2 - \sigma^2 r_t^2 \gamma \right)^2
  \]
  - Would we obtain the same answer?
  - What method is preferable in theory?
  - What method is preferable in practice?
• A good project: Redo the Chan et al. (1992) paper with data up to 2001 and with monthly and daily data. Compare the results.