2. Background (Review of some Statistical Theory)

2.a. Basic properties of random sequences

2.b. Stationary Time Series, the Weak Law of Large Numbers and the Central Limit Theorem

2.c. Algebra of expectations and Martingale Difference Sequences

2.d. Synthetic theory of ML estimation

Luca Fanelli

University of Bologna

luca.fanelli@unibo.it
2a Basic properties of random sequences

Deterministic sequences

Let \( \{h_T, T = 1, 2, \ldots\} \equiv \{h_T\} \) be a sequence of real numbers.

If the sequence has a limit, \( h \), then this is denoted by

\[
\lim_{T \to \infty} h_T = h.
\]

This implies that for every \( \varepsilon > 0 \) there exists a positive, finite integer \( T_\varepsilon \) such that

\[
|h_T - h| < \varepsilon \quad \text{for} \quad T > T_\varepsilon.
\]

If \( h_T \) is a \( p \times 1 \) vector, \( \lim_{T \to \infty} h_T = h \) means that for every \( \varepsilon > 0 \) there exists a positive, finite integer \( T_\varepsilon \) such that

\[
\|h_T - h\|_2 < \varepsilon \quad \text{for} \quad T > T_\varepsilon.
\]
Recall section

Note that $\|v\|_2 = (v'v)^{1/2}$ is the Euclidean norm of the vector $v$.

This can be interpreted as a measure of the length of $v$ in the space $\mathbb{R}^p$, i.e. a measure of the distance of the vector $v$ from the vector $0_{p \times 1}$.

One can generalize this measure by defining the norm $\|v\|_A = (v'Av)^{1/2}$, where $A$ is a symmetric positive definite matrix; this norm measures the distance of $v$ from $0_{p \times 1}$ ‘weighted’ by the elements of the matrix $A$.

We will make use of quadratic forms of the type:

$$v'V^{-1}v$$

in which $v \sim N(0_{p \times 1}, V)$ is a $p \times 1$ vector distributed and $V$ is a $p \times p$ covariance matrix symmetric and positive definite. One has

$$v'V^{-1}v \sim \chi^2(p) \quad (\text{quadratic forms from Gaussian}).$$
Stochastic sequences

Henceforth $h_T$ will be considered a $p \times 1$ vector, except where stated otherwise.

Suppose now that each $h_T$ is a (continuous) random vector.

We consider sequence $\{h_T, T = 1, 2, \ldots\} \equiv \{h_T\}$.

We are interested in the concepts of convergence in probability and convergence in distribution.

The sequence of random vectors $\{h_T, T = 1, 2, \ldots\}$ converges in probability to the non-stochastic vector $h$ if for all $\epsilon > 0$:

$$\lim_{T \to \infty} P (\|h_T - h\|_2 < \epsilon) = 1;$$

we conventionally write $h_T \to_p h$. 
Slutsky’s Theorem

Let \( h_T \to_p h \), and let \( f(\cdot) \) be a vector of continuous functions, then \( f(h_T) \to_p f(h) \).

Other Important Theorems

Let \( h_T \to_p h \) and \( g_T \to_p g \). Then \( h_Tg_T' \to_p hg' \).

\[ f_1(h_T) \to_p f_1(h), \quad f_2(g_T) \to_p f_2(g), \quad f_1(\cdot) \text{ and } f_2(\cdot) \] continuous vector functions. \( f_1(h_T)f_2(g_T) \to_p f_1(h)f_2(g) \).
Recall.

Which is a $g \times 1$ vector function? If $v$ is a $p \times 1$ vector, then

$$f(v) := \begin{pmatrix} f_1(v) \\ \vdots \\ f_g(v) \end{pmatrix}$$

where each $f_i(v)$ is a function that takes as input the $p \times 1$ vector $v$, and gives in output the scalar $f_i(v), i = 1, \ldots, g$.

Therefore, we will use the concept of derivative:

$$\frac{\partial f(v)}{\partial v'} := \begin{pmatrix} \frac{\partial f_1(v)}{\partial v_1} \\ \vdots \\ \frac{\partial f_g(v)}{\partial v_1} \end{pmatrix} := \begin{bmatrix} \frac{\partial f_1(v)}{\partial v_1} & \cdots & \frac{\partial f_1(v)}{\partial v_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_g(v)}{\partial v_1} & \cdots & \frac{\partial f_g(v)}{\partial v_p} \end{bmatrix}$$

which is a $g \times p$ matrix known as Jacobian matrix.
The concept of convergence in probability leads us to the concept of consistency of an estimator.

**Consistency of an estimator**

Let \( \{\hat{\theta}_T\} \) be a sequence of estimators (hence random vectors) of the unknown parameter \( \theta_0 \); then \( \hat{\theta}_T \) is said to be a **consistent estimator** of \( \theta_0 \) if \( \hat{\theta}_T \to p \theta_0 \).

Convergence in probability implies that the difference between \( \hat{\theta}_T \) and \( \theta_0 \) disappears with probability one as \( T \to \infty \). In the limit \( \hat{\theta}_T \) and \( \theta_0 \) are essentially identical.
Consider the random vector $h_T$ with distribution function (density) $F_T(c)$.

The sequence of random vectors $\{h_T\}$ with corresponding distribution functions $\{F_T(c)\}$ converges in distribution to the stochastic vector $h$ with distribution function $F(c)$, if and only if there exists $T_\epsilon$ for every $\epsilon > 0$, such that $\|F_T(c) - F(c)\|_2 < \epsilon$ for $T > T_\epsilon$ at all points of continuity $\{c\}$.

We conventionally write $h_T \to_D h$.

**Important result:** Imagine that

$$\hat{V} \to_p V, \quad g_T \to_D \mathcal{L}(0, V)$$

where $V$ is a non-random matrix. Then

$$g_T \to_D \mathcal{L}(0, \hat{V}).$$
The concept of convergence in distribution leads us to the concept of \textit{asymptotic distribution of an estimator}.

Under a set of regularity conditions, the typical asymptotic distribution of an estimator is multivariate Gaussian.

**Asymptotic normality of an estimator**

Let \( \{\hat{\theta}_T\} \) be a sequence of estimators (hence random vectors) of the unknown parameter \( \theta_0 \). The consistent estimator of \( \theta_0, \hat{\theta}_T \), is said to be \textit{asymptotically normal (or Gaussian)} if

\[
T^{1/2}(\hat{\theta}_T - \theta_0) \rightarrow_D N(0, V_\theta),
\]

where \( V_\theta \) is the asymptotic covariance matrix of the estimator.

Using the important result above, we can say that if we have a consistent estimator \( \hat{V}_\theta \) of \( V_\theta \), then

\[
T^{1/2}(\hat{\theta}_T - \theta_0) \rightarrow_D N(0, \hat{V}_\theta).
\]
2.b. Stationary time series, the Weak Law of Large Numbers and the Central Limit Theorem

In dealing with the estimators we are interested in this course, we will face certain sums of random variables.

The Weak Law of Larger Numbers (WLLN) and Central Limit Theorem (CLT) allows us to understand the limiting behaviour of these sums.

However, we need some restrictions on the nature of the random variables we consider.

Henceforth we will denote the the sequence of random vectors \( \{ h_t, t = 1, 2, ..., \infty \} = \{ h_t \} \) **stochastic process**.

A stochastic process involves, by definition, an infinite sequence of random vectors.
Given \( \{h_t\} \), consider the finite realization of \( \{h_t\} \):

\[
h_1, \ldots, h_T
\]

where \( T \) is a given sample length (some initial conditions may be fixed).

We call such realization **time series**.

In other words, for each stochastic process you can think about an associated time series.

Analogy: the stochastic process is the population, the time series is the sample counterpart.

It happens that the two terms are used interchangeably.

The stochastic processes and corresponding time series we have in mind are \( r_1, \ldots, r_T \), where \( r_t \) is the log-return of an asset.
Strictly stationary stochastic process

Let \( \{h_t\} \) be a stochastic process. Define the finite set \( \{t_1, t_2, \ldots, t_n\} \) for given integer \( n \). \( \{h_t\} \) is said to be a strictly stationary process if and only if the joint probability distribution function, \( F(\cdot) \), satisfies:

\[
F(h_{t_1}, h_{t_2}, \ldots, h_{t_n}) = F(h_{t_1+c}, h_{t_2+c}, \ldots, h_{t_n+c})
\]

for any integer constant \( c \).

Consequence of this definition:

for \( n:=1 \), \( F(h_{t_1}) = F(h_{t_1+c}) \) for any integer \( c \)

hence all moments of the process (if they exist) are constant over time because e.g.

\[
E(h_{t_1}) = E(h_{t_1+c})
\]

\[
Var(h_{t_1}) = Var(h_{t_1+c})
\]

\[
Cov(h_{t_1}, h_{t_2}) = Cov(h_{t_1+c}, h_{t_2+c}).
\]
Observe that a strictly stationary process does not imply independence.

Examples of strictly stationary processes:

1. \( \{\varepsilon_t\} \sim \text{iid} \)

2. \( \{\varepsilon_t\} \sim WNN(0_{p \times 1}, V) \) (Gaussian white noise).
To make the inference, it is sufficient to deal with a weaker conditions and in particular with covariance stationarity + ergodic processes.

Inference is the process by which we use data and observations to understand something about the underlying population.
Covariance (weak) stationary stochastic process

Let \( \{h_t\} \) be a stochastic process.

\( \{h_t\} \) is said to be a strictly stationary process if and only if the following conditions hold:

1. \( E(h_t) = h < \infty \);

2. \( Cov(h_t, h_{t-k}) \) is a matrix whose elements do not depend on \( t \).

Strict stationarity implies weak (covariance) stationarity. The converse is not true!
Examples of covariance stationary processes:

1. \( \{\varepsilon_t\} \sim \text{iid} \)

2. \( \{\varepsilon_t\} \sim WNN(0_{p\times1}, V) \) (Gaussian white noise).
Weak Law of Large Numbers (WLLN)

Let \( \{h_t\} \) be a strictly stationary stochastic process with \( E(h_t) := \mu \ (p \times 1) \). Then subject to certain regularity conditions

\[
\bar{h} := \frac{1}{T} \sum_{t=1}^{T} h_t \to_p \mu.
\]

Ergodic processes

Let \( \{h_t\} \) be a covariance stationary stochastic process with \( E(h_t) := \mu \ (p \times 1) \), \( \text{Var}(h_t) := V \) and \( \text{Cov}(h_t, h_{t-\ell}) = \Gamma(\ell) \), \( \ell = 1, 2, \ldots \). The process is ergodic for the mean if

\[
\bar{h} := \frac{1}{T} \sum_{t=1}^{T} h_t \to_p \mu.
\]
The process is ergodic for the variance if
\[ \frac{1}{T} \sum_{t=1}^{T} (h_t - \bar{h})(h_t - \bar{h})' \to_p V. \]

The process is ergodic for the covariances if
\[ \frac{1}{T} \sum_{t=1}^{T} (h_t - \bar{h})(h_{t-\ell} - \bar{h})' \to_p \Gamma(\ell), = 1, 2, \ldots \]
Central Limit Theorem (CLT)

Let \( \{h_t\} \) be a strictly stationary stochastic process with \( E(h_t) := \mu \ (p \times 1) \). Then subject to certain regularity conditions

\[
\frac{1}{T^{1/2}} \sum_{t=1}^{T} (h_t - \mu) \rightarrow_D N(0_{p \times 1}, \Omega)
\]

where the \( p \times p \) positive definite covariance matrix \( \Omega \) can be interpreted as

\[
\Omega := \lim_{T \to \infty} \text{Var} \left[ T^{-1/2} \sum_{t=1}^{T} (h_t - \mu) \right]
\]

and is also known as the long run covariance matrix.

Why is \( \Omega \) called long run covariance matrix? Because it exists also the short run covariance matrix:

\[
\text{Var}(h_t) := E \left[ (h_t - \mu)(h_t - \mu)' \right] := \Lambda.
\]
The difference between these two covariance matrices is given by the expression

\[
V \text{ar} \left[ T^{-1/2} \sum_{t=1}^{T} (h_t - \mu) \right] = \frac{1}{T} V \text{ar} \left[ \sum_{t=1}^{T} (h_t - \mu) \right] \\
= \frac{1}{T} \left\{ \sum_{t=1}^{T} V \text{ar} [(h_t - \mu)] + 2 \sum_{t=1}^{T} \sum_{s \neq t} T \text{ov}(h_t, h_s) \right\}.
\]

Thus for an independent stochastic process.

\[\Omega := \Lambda\]
Another important property

Let \( \{ M_t, t = 1, \ldots, T \} \) be a sequence of \( p \times p \) stochastic matrices (it means that each element of \( M_t \) is a random variable) such that \( M_T \rightarrow_p M \), where is a \( p \times p \) matrix of (non-random) constants, and let \( \{ v_t, t = 1, \ldots, T \} \) be a sequence of \( p \times 1 \) vectors such that \( v_T \rightarrow_D N(0, \Omega) \). Then

\[
M_T v_T \rightarrow_D N(0, M\Omega M').
\]
The WLLN and CLT hold also under weaker conditions which are:

- weak (covariance) stationarity

- ergodicity.

If \( \{r_t\} \) is such that the WLLN can be applied, then

\[
\frac{1}{T} \sum_{t=1}^{T} r_t \rightarrow p E(r_t),
\]

\[
\frac{1}{T} \sum_{t=1}^{T} (r_t - \hat{\mu}_r)^2 \rightarrow p E[(r_t - \mu_r)^2],
\]

\[
\hat{\sigma}^j_r := \frac{1}{T} \sum_{t=1}^{T} (r_t - \hat{\mu})^j \rightarrow p E[(r_t - \mu_r)^j], \quad j \geq 3,
\]

\[
\hat{S}(r_t) := \frac{\hat{\sigma}_r^3}{(\hat{\sigma}_r)^3} \rightarrow p S(r_t) := \frac{E[(X-\mu)^3]}{\sigma^3},
\]

\[
\hat{K}(r_t) := \frac{\hat{\sigma}_r^4}{(\hat{\sigma}_r)^2} \rightarrow p K(r_t) := \frac{E[(X-\mu)^4]}{\sigma^4},
\]

\[
\hat{\rho}(\tau) := Corr(r_t, r_{t-\tau}) \rightarrow p Cov(r_t, r_{t-\tau}), \tau = 1, 2, \ldots, \tau^{\text{max}}.
\]
2.c. Algebra of expectations

Let $w_t:=\begin{pmatrix} y_t \\ z_t \end{pmatrix}$ be bi-variate continuous random vector with density $f_w(w_t; \theta)$ (joint distribution).

Marginal distribution of $y_t$:

$$f_y(y_t; \theta_y) = \int_z f_w(w_t; \theta) dz_t$$

from which

$$E(y_t) = \int_y y_t \ f_y(y_t; \theta_y) dy_t$$

Marginal distribution of $z_t$:

$$f_z(z_t; \theta_z) = \int_y f_w(w_t; \theta) dy_t$$

from which

$$E(z_t) = \int_z z_t \ f_z(z_t; \theta_y) dz_t$$
Conditional distribution of $y_t$ given $z_t$:

$$f_{y|z}(y_t \mid z_t; \theta_{y|z}) := \frac{f_w(w_t; \theta)}{f_z(z_t; \theta_z)}$$

from which we derive the well known factorization:

$$f_w(w_t; \theta) = f_{y|z}(y_t \mid z_t; \theta_{y|z}) \times f_z(z_t; \theta_z).$$

and

$$E(y_t \mid z_t) = \int_y y_t \ f_{y|z}(y_t \mid z_t; \theta_{y|z}) \ dy_t.$$ 

Intuitively, $E(y_t \mid z_t)$ is as a function of $z_t$, i.e.

$$E(y_t \mid z_t) = g(z_t)$$

where, in general, the function $g(z_t)$ is nonlinear.

The only case in which $E(y_t \mid z_t) = g(z_t)$ is linear is when $w_t$ is Gaussian; the parameters in $B$ are functions of the parameters of the Gaussian distribution!
We can further generalize the concept of conditional expectations.

Let $v$ a $p \times 1$ stochastic vector and $\mathcal{F}$ a set containing random variables and all of their linear combinations (sigma-algebra or sigma-field).

Then

$$E(v \mid \mathcal{F}) := \left( \begin{array}{c} E(v_1 \mid \mathcal{F}) \\ \vdots \\ E(v_p \mid \mathcal{F}) \end{array} \right)$$

is a stochastic vector.

Intuitive interpretation: Take the point of view of an investigator that has to forecast the expected value of $v$ based on the available information contained in $\mathcal{F}$. It is as if we express the stochastic variability of the elements of $v$ as function of the stochastic elements in the set $\mathcal{F}$.

In some cases $\mathcal{F}=M$, where $M$ is a matrix whose elements are random variables:

$$E(v \mid M) := \left( \begin{array}{c} E(v_1 \mid M) \\ \vdots \\ E(v_p \mid M) \end{array} \right)$$

$p \times 1$
Properties

1. Linearity:
   \[ E(a'v_1 + b'v_2 \mid \mathcal{F}) = a'E(v_1 \mid \mathcal{F}) + b'E(v_2 \mid \mathcal{F}) \]
   where \( v_1 \) and \( v_2 \) stochastic vectors and \( a \) and \( b \) deterministic vectors of conformable dimensions;

2. Inclusion rule: if \( v \in \mathcal{F} \), then
   \[ E(v \mid \mathcal{F}) = vE(1 \mid \mathcal{F}) = v \]
   (If \( v \) is included in the set \( \mathcal{F} \) it can be shifted away from the expectations operator).

3. Law of iterated expectations 1:
   \[ E(v) = E_\mathcal{F} [E(v \mid \mathcal{F})] \]
   (This rule established a link between unconditional expectations and conditional expectations).

4. Law of iterated expectations 2: Suppose that \( \mathcal{G} \subseteq \mathcal{F} \); then
   \[ E [(v \mid \mathcal{F}) \mid \mathcal{G}] = E(v \mid \mathcal{G}). \]
Martingale Difference Sequences (MDS)

Given the stochastic sequences \( \{v_t\}_{t=0}^{\infty} \) and \( \{\mathcal{I}_t\}_{t=0}^{\infty} \), where:

- \( v_t \) is a \( p \times 1 \) vector,

- \( \mathcal{I}_t \) is a sigma field (information set) such that \( \sigma(v_t, v_{t-1}, \ldots, \mathcal{I}_t) = \mathcal{I}_t \);

- \( \mathcal{I}_t \subseteq \mathcal{I}_{t+1}, \ t = 1, 2, \ldots \)

we say that \( v_t \) is a Martingale Difference Sequence \( (v_t \sim \text{MDS}) \) with respect to \( \mathcal{I}_t \) if and only if

\[
E_t v_{t+1} := E(v_{t+1} \mid \mathcal{I}_t) = 0_{p \times 1}.
\]
Exercise 1.

Show that $v_t \sim \text{MDS}$ is an uncorrelated process.

We have to prove that

$$E(v_t v'_{t-\ell}) = 0_{p \times p} \quad \text{for } \ell=1,2,\ldots.$$ 

Observe that by the law of iterated expectations (type-1):

$$E(v_t v'_{t-\ell}) = E \left( E(v_t v'_{t-\ell} | I_{t-\ell}) \right) = E \left( E(v_t | I_{t-\ell}) v'_{t-\ell} \right)$$

$$= E \left( 0_{p \times 1} \times v'_{t-\ell} \right) = 0_{p \times 1}.$$ 

This exercise shows that a MDS corresponds to a sequence which is uncorrelated over time.

Moreover, the law of iterated expectations ensures that the unconditional mean of $v_t$ is zero: $E(v_t) = 0.$
Is a MDS a White Noise?

Depends on the assumption we make on the conditional variance of $v_t$:

- if $Var(v_{t+1} \mid \mathcal{I}_t) := \Sigma_v =$ const, then $Var(v_{t+1}) = \Sigma_v =$ const (why?) and $v_t$ is also a White Noise;

- if $Var(v_{t+1}) := \Sigma_{v,t}$, $v_t$ is uncorelated but is not a White Noise.

Is a MDS covariance stationary?

Exercise ....
**Conditional Variance**

$$\text{Var}(v \mid \mathcal{F}):= E\left( \left[ v - E(v \mid \mathcal{F}) \right]\left[ v - E(v \mid \mathcal{F}) \right]' \mid \mathcal{F} \right)$$

$$:= E\left( [vv'] \mid \mathcal{F} \right) - [E(v \mid \mathcal{F})]^2$$

is a $p \times p$ matrix.

By the Law of iterated expectations 1:

$$\text{Var}(v):= E_{\mathcal{F}} \left[ \text{Var}(v \mid \mathcal{F}) \right].$$

**Conditional Covariance**

$$\text{Cov}(v_1, v_2 \mid \mathcal{F}):= E\left( \left[ v_1 - E(v_1 \mid \mathcal{F}) \right]\left[ v_2 - E(v_2 \mid \mathcal{F}) \right]' \mid \mathcal{F} \right)$$

$$:= E\left( [v_1v_2] \mid \mathcal{F} \right)' - E(v_1 \mid \mathcal{F})E(v_2 \mid \mathcal{F})$$

is a $p_1 \times p_2$ matrix

$$\text{Cov}(v_1, v_2):= E_{\mathcal{F}} \left[ \text{Cov}(v_1, v_2 \mid \mathcal{F}) \right].$$
Crucial Decomposition

Given \( v \) and \( \mathcal{F} \) it always holds the decomposition:

\[
v_{\text{true value}} = E(v \mid \mathcal{F}) + \eta
\]

where

\[
\eta := v - E(v \mid \mathcal{F}).
\]

A typical econometric model is based on:

\[
y = E(y \mid \mathcal{F}) + u
\]

where

\[
Var(u \mid \mathcal{F}) := \left\{ \begin{array}{ll}
\text{constant} \\
g(x), x \in \mathcal{F}
\end{array} \right.
\]
We make intensive use of the following concepts:

1. Stochastic vector \( w_t := (y_t', z_t')' \);

2. Stochastic information set generated by \( z_t, w_{t-1}, \ldots, w_1 : \)

\[
\mathcal{F}_t := \{ z_t, w_{t-1}, \ldots, w_1 \}
\]

3. Specification of a model for the random vector

\[
E(y_{t+1} \mid \mathcal{F}_t)
\]

which is a function of the variables in \( \mathcal{F}_t \). Interpretation: we want to forecast the variables of interest at time \( t + 1 \), \( y_{t+1} \), conditional on the available information set.
4. Specification of a model for the random vector

\[ \text{Var}(y_{t+1} \mid \mathcal{F}_t) \]

which is a function of the variables in \( \mathcal{F}_t \). Interpretation: we want to forecast the variance of the variables of interest at time \( t + 1 \), conditional on the available information set.

Model for conditional volatility (variance)
In the previous set of slides we have stated that we are mainly interested in specifying a model and forecasting the quantity and the volatility of

\[ r_{T+1} := \log(P_{T+1}) - \log(P_T) \]

unknown at time \( T \) known at time \( T \)

Now, at time \( T \) the investor has a stochastic information set \( \mathcal{F}_T \), therefore we are mainly interested in the two following quantities:

\[ E(r_{T+1} \mid \mathcal{F}_T) \]

\[ Var(r_{T+1} \mid \mathcal{F}_T). \]

The three stylized facts seen in the Slides 1 suggest that if \( \mathcal{F}_T \) contains the past history of the returns, it is difficult to specify a model for \( E(r_{T+1} \mid \mathcal{F}_T) \) while it should be easier to specify a model for \( Var(r_{T+1} \mid \mathcal{F}_T). \)
Let us come back to our candidate model for asset returns

\[ r_t = \mu + \frac{u_t}{\sigma_t \epsilon_t} \]

where

(i) \( \mu := E(r_t) \);

(ii) \( \sigma_t \in \mathcal{F}_{t-1} \), where \( \mathcal{F}_{t-1} := \{ r_{t-1}, \ldots, r_1 \} \);

(iii) \( \{ \sigma_t \} \) is stationary with \( E(\sigma_t^4) < \infty \) and \( \text{Corr}(\sigma_t, \sigma_{t-\tau}) > 0, \tau = 1, 2, \ldots \)

(iv) \( \{ \epsilon_t \} \sim \text{iid} \mathcal{N}(0, 1) \)

(v) \( \{ \sigma_t \} \) and \( \{ \epsilon_t \} \) are stochastically independent.
The implications of the above specification are:

1. \( r_t \) is distributed as a Mixture of Gaussian that implies that \( K(r_t) > 3 \).

**Proof**: blackboard.

2. \( \text{Corr}(r_t, r_{t-\tau}) = 0 \), \( \tau = 1, 2, ... \)

**Proof**: blackboard.

3. \( \text{Corr}(r_t^2, r_{t-\tau}^2) > 0 \), \( \tau = 1, 2, ... \)

**Proof**: blackboard.
Further implications:

- $Var(u_t \mid F_{t-1}) = \sigma^2_t$

that means that $\sigma^2_t$ is the conditional variance of $u_t$ in the model

$$r_t = \mu + u_t, \quad t = 1, \ldots, T;$$

- $$r_t \mid \sigma^2_t \sim N(\mu, \sigma^2_t), \quad t = 1, \ldots, T;$$

that means that the unconditional distribution of $r_t$ is MixedGaussian

$$r_t \sim N(\,, \,)$$

and $\sigma^2_t$ serves as mixing variable.
2.d. Synthetic theory of ML estimation

We are concerned with maximum likelihood (ML) estimation for two reasons:

1. Key parametric estimation method (optimal properties under certain regularity conditions);

2. ARCH/GARCH processes are typically estimated by ML.
What is a parametric statistical model?

Statistical model = \{ stochastic distribution: sampling scheme: \}

⇒ we are able to write

\[ f(\text{DATA}; \theta) = L(\theta) \quad \text{likelihood function}. \]

The $b \times 1$ vector of unknown parameters $\theta$ belongs to the (open) space $\Theta$, and the ‘true’ value of $\theta$, $\theta_0$, is an interior point of $\Theta$. 
Suppose that $y_t$ is a $g \times 1$ vector which has density
$f(y_t; \theta)$

$y_1, \ldots, y_T$ sample of length $T$

If the sampling scheme is such that all observations can be treated as independent over time (heroic assumption!) then

$$L(\theta) := f(DATA; \theta) := \prod_{t=1}^{T} f(y_t; \theta)$$

Otherwise

$$L(\theta) := f(DATA; \theta) := \prod_{t=1}^{T} f(y_t \mid F_{t-1}; \theta)$$

where

$$F_{t-1} := \{ y_{t-1}, \ldots, y_1 \} \quad \text{past history}$$

$$f(y_1 \mid F_0; \theta) := f(y_1; \theta).$$
It will be convenient to focus on the log-likelihood function:

$$\log L(\theta) := \sum_{t=1}^{T} \log f(y_t \mid \mathcal{F}_{t-1}; \theta).$$

The ML estimator of $\theta$ is obtained by solving the problem

$$\max_{\theta} \log L(\theta)$$

i.e.

$$\hat{\theta}_{ML} := \arg \max \log L(\theta).$$
As is known, in order to obtain $\hat{\theta}_{ML}$ it is necessary to solve the first-order conditions:

$$s_T(\theta) := \frac{\partial \log L(\theta)}{\partial \theta} := \left( \begin{array}{c} \frac{\partial \log L(\theta)}{\partial \theta_1} \\ \frac{\partial \log L(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial \log L(\theta)}{\partial \theta_b} \end{array} \right) = 0_{b \times 1}$$

that means that $\hat{\theta}_{ML}$ is such that $s_T(\hat{\theta}_{ML}) = 0_{b \times 1}$.

The $b \times 1$ vector $s_T(\theta)$ is known as the score (gradient) of the likelihood function.

It is not always possible to solve the first order conditions analytically; in many circumstances (e.g. non-linear restrictions) numerical optimization procedures are required.
Note that

\[ s_T(\theta) := \frac{\partial \log L(\theta)}{\partial \theta} := \frac{\partial \sum_{t=1}^{T} \log f(y_t | F_{t-1}; \theta)}{\partial \theta} \]

\[ = \sum_{t=1}^{T} \frac{\partial \log f(y_t | F_{t-1}; \theta)}{\partial \theta} = \sum_{t=1}^{T} s_t(\theta). \]

Each component \( s_t(\theta) \) of the score depends on the data, hence it is a random variable!
To be sure that $\hat{\theta}_{ML}$ is a maximum (and not a minimum), it is further necessary that the Hessian matrix

$$H_T(\theta) := \frac{\partial}{\partial \theta'} \left( \frac{\partial \log L(\theta)}{\partial \theta} \right) := \frac{\partial}{\partial \theta' s_T(\theta)} := \frac{\partial^2 \log L(\theta)}{\partial \theta' \partial \theta}$$

be (semi)negative definite at the point $\hat{\theta}_{ML}$, i.e. $H_T(\hat{\theta}_{ML}) \prec 0$. 
A crucial requirement for ML estimation is that

$$\log L(\theta_0) > \log L(\theta) \text{ for each } \theta \in \Theta \setminus \{\theta_0\}$$

condition that ensures the existence of a global maximum.

However, also situations of the time:

$$\log L(\theta_0) > \log L(\theta) \text{ for } \theta \in \mathcal{N}_{\theta_0} \subset \Theta$$

where $$\mathcal{N}_{\theta_0}$$ is a neighborhood of $$\theta_0$$ are potentially fine, local maximum.
Identification

Two points $\theta_1, \theta_2$ in $\Theta$ are said observationally equivalent if $\log L(\theta_1) = \log L(\theta_2)$; this means that the statistician is not able to establish whether we are doing inference on $\theta_1$ or $\theta_2$ on the basis of the data.

We say that a statistical model is globally identified iff there are no observational equivalent points in $\Theta$. It is a prerequisite for inference!

We say that a statistical model is locally identified iff there are no observational equivalent points in the neighborhood $\mathcal{N}_{\theta_0} \subset \Theta$.

Of course, a model may be locally identified but not globally identified.
We define \textbf{(Fisher) Information Matrix} the quantity:

\[ I_T(\theta) := E \left( -\frac{\partial^2 \log L(\theta)}{\partial \theta' \partial \theta} \right) = E \left( -H_T(\theta) \right) \]

and it can be shown that under a set of regularity conditions, including the hypothesis of correct specification of the statistical model, one has the equivalence:

\[ I_T(\theta) := E \left( -\frac{\partial^2 \log L(\theta)}{\partial \theta' \partial \theta} \right) = E \left( s_T(\theta) s_T(\theta) \right)' \]

\[ = E \left( \left[ \frac{\partial \log L(\theta)}{\partial \theta} \right] \left[ \frac{\partial \log L(\theta)}{\partial \theta} \right]' \right). \]

The \textbf{Asymptotic Information Matrix} is given by the quantity

\[ I_\infty(\theta) := \lim_{T \to \infty} \frac{1}{T} I_T(\theta) = \lim_{T \to \infty} \frac{1}{T} E(-H_T(\theta)). \]
A crucial result establishes that:

(a) a statistical model is globally identified iff

\[ \text{rank} \left[ I_\infty(\theta) \right] = b \quad \text{for} \quad \theta \in \Theta \]

(b) a statistical model is locally identified iff

\[ \text{rank} \left[ I_\infty(\theta) \right] = b \quad \text{for each} \quad \theta \in \mathcal{N}_{\theta_0} \subset \Theta. \]

Global identification \( \Rightarrow \) Local identification.
Fundamental result of ML estimation to know

Under standard conditions (including stationary ergodic processes), using some CLT which are valid for MDS:

\[ I_T(\theta_0)^{-1/2} s_T(\theta_0) \rightarrow_D N(0_{b \times 1}, I_b). \]

(What's the meaning of \( I_T(\theta_0)^{-1/2} \)?

Recall that \( I_T(\theta_0) \) is symmetric and positive definite.

Hence, by applying the spectral decomposition:

\[ I_T(\theta_0) := P G P' \]

where \( G \) is a diagonal matrix with the eigenvalues of \( I_T(\theta_0) \) on the main diagonal, i.e. \( G := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_b) \), where \( \lambda_i > 0, \ i = 1, \ldots, b \). Then

\[ I_T(\theta_0)^{-1/2} := P G^{-1/2} P', \quad G^{-1/2} := \text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_b^{-1/2}). \]
Properties of ML estimator

**Crucial Assumption:** the statistical model is correctly specified.

1. Let $\delta$ be a $1 \times a$ vector such that $\delta = g(\theta)$, where $g(\cdot)$ is a **continuous** vector function that maps the elements of the $g \times 1$ vector $\theta$ into $\delta$, $a \geq g$. Let $\hat{\theta}_{ML}$ be the ML estimator of $\theta$, then

$$\hat{\delta} = g(\hat{\theta}_{ML})$$

is the (indirect) ML estimator of $\delta$, i.e. $\hat{\delta} \equiv \hat{\delta}_{ML}$. Moreover (sandwitch rule),

$$\text{Var}(\hat{\delta}_{ML}) := \begin{bmatrix} \frac{\partial g(\hat{\theta}_{ML})}{\partial \theta'} \end{bmatrix}_{a \times g} \text{Var}(\hat{\theta}_{ML}) \begin{bmatrix} \frac{\partial g(\hat{\theta}_{ML})}{\partial \theta'} \end{bmatrix}'_{g \times a}$$

on condition that the **Jacobian matrix** $\frac{\partial g(\theta)}{\partial \theta'}$ be of full column rank $g$ evaluated at the ‘true’ point $\theta_0$. 
2. In general,

\[ E(\hat{\theta}_{ML}) \neq \theta_0 \]

namely the ML is not correct!

3. If the variables are generated by stationary and ergodic processes the ML estimator is consistent:

\[ \hat{\theta}_{ML} \rightarrow_p \theta_0. \]

4. If the variables are generated by stationary and ergodic processes the ML estimator is asymptotically Gaussian:

\[ T^{1/2} \left( \hat{\theta}_{ML} - \theta_0 \right) \rightarrow_D N(0_{b \times 1}, V_{\hat{\theta}}) \]

with asymptotic covariance matrix \( V_{\hat{\theta}} \). This property suggests that one can do standard inference in large samples!
5. If the variables are generated by stationary and ergodic processes the ML estimator estimator is asymptotically efficient:

\[ V_{\hat{\theta}} = [I_{\infty}(\theta)]^{-1} \]

in the sense that any other consistent asymptotic Gaussian estimator estimator \( \hat{\theta} \) has asymptotic covariance matrix \( V_{\hat{\theta}}^* \) such that \( V_{\hat{\theta}}^* \preceq [I_{\infty}(\theta)]^{-1} \).

The properties 3-4-5 are asymptotic properties and make the ML estimator ‘optimal’.

(Mild) Departures from Crucial Assumption: it may happen that the statistician specifies a distribution probability that is different from the one characterizing the DGP (e.g. the statistician specifies a Gaussian but the ‘true’ is a Student-t); in these circumstances the ML estimator is said to be a Quasi-ML (or Pseudo-ML) estimator and generally retains the property of consistency (3) and asymptotic normality (4) but, as expected, it is not longer asymptotic efficient (i.e. property 5 is lost).