Financial Econometrics

Sol Yates

March 4, 2024

Contents

Lecture 1: Introduction Lecture

Wed 31 Jan 11:19

1 Financial Time Series and their Characteristics

1.1 Asset Returns

[Lecture 1] [Reading]

Financial studies involve returns, instead of prices of assets.

Returns :

- Is a complete and scale free summary of the investment opportunity
- Are easier to handle than price series

p_t is the price of an asset at time index t. And assuming an asset pays no dividends.

Continuous Compounding

One period Simple Returns

Holding the asset for one period from date $t-1$ to date t would result in a simple gross return :

$$1 + R_t = \frac{P_t}{P_{t-1}} \text{ or } P_t = P_{t-1}(1 + R_t).$$

The corresponding one period simple net return or simple return is :

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Multi period Simple Returns
Holding the asset for k periods between dates t-k and t gives a k-period simple gross return :

$$\begin{aligned} 1 + R_t[k] = \frac{P_t}{P_{t-k}} &= \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \ldots \times \frac{P_{t-k+1}}{P_{t-k}}. \\ &= (1 + R_t)(1 + R_{t-1}) \ldots (1 + R_{t-k+1}). \\ &= \prod_{j=0}^{k-1} (1 + R_{t-j}). \end{aligned}$$

That is, the k-period simple gross return is just the product of the k one period simple gross returns involved. A compound return.

The actual time interval is important in discussing and comparing returns, if not given, it is implicitly assumed to be one year.

If an asset is gold for k years, then the annualized average return is defined as

$$\text{Annualized} R_t[k] = (\prod_{j=0}^{(k-1)} (1 + R_{t-j}))^{(\frac{1}{k})} - 1.$$

Which is a geometric mean of the k one period simple gross returns involved and can be computed by

$$= \exp(\frac{1}{k} \sum_{j=0}^{(k-1)} \ln(1 + r_{t-j})) - 1$$

Where it is easier to compute the arithmetic average than the geometric mean and the one-period returns tend so be small, one can use a first order Taylor expansion to approximate the annualized return and obtain

$$\approx \frac{1}{k} \sum_{j=0}^{(k-1)} R_{t-j}.$$

Continuous Compounding
Assume the interest rate of a bank deposit is 10% per annul, and the initial deposit is \$1

If the bank pays interest once a year, then the net value of the deposit becomes 1.1\$. If the bank pays interest semi-annually, the 6-month interest rate is 5% and the net value is $1(1 + \frac{0.1}{2})^2 = \$1.1025$ after the first year.

In general if the bank pays interest m times a year, then the interest rate for each payment is $10\%/m$ and the net value of the deposit becomes $1(1 + \frac{0.1}{m})^{(m)}$ one year later.

Continuously Compounded Returns

The natural logarithm of the simple gross return of an asset is called the continuously compounded return or log return :

$$R_t = \ln(1 + R_t) = \ln(P_t/P_{t-1}) = p_t - p_{t-1} \tag{1}$$

Where $p_t = \ln(P_T)$. Continuously compounded returns are advantageous since they are the sum of continuously compounded multi period return.

Portfolio Return

Simple net return of a portfolio consisting of N assets is a weighted average of the simple net returns of the assets involved, where the weight on each asset is the percentage of the portfolio's value invested in that asset. Where p is a portfolio that places weight w_i on asset i. Then the simple return of p at time t is

$$R_{p,t} = \sum_{i=1}^{N} w_i R_{it}.$$

Where R_{it} is the simple return of asset i.

The continuously compounded returns of a portfolio, do not have this property. Instead,

$$R_{p,t} sim \sum_{i=1}^{N} w_i r_{it}.$$

Where $r_{p,t}$ is the continuously compounded return of the portfolio at time t

Dividend Payment

If an asset pays periodically. Let D_t be the dividend payment of an asset between dates $t-1$ and P_t be the price of the asset at the end of period t. The dividend is this not included in P_t The simple net return and continuously compounded return at time t become

$$R_t = \frac{P_t + D_t}{P_{t-1}} - 1 \quad , \quad r_t = \ln(P_t + D_t) - \ln(P_{t-1}).$$

Excess Return

The difference between the asset's return and return on some reference asset, often taken to be rissoles such as short term US treasury bill. Simple excess return and log excess return of an asset are then defined as

$$Z_t = R_t - R_{0t} \quad , \quad z_t = r_t - r_{0t}.$$

Where R_{0t} and r_{0t} are the simple and log returns of the reference asset (resp)

Distributional Properties of Returns

Review of statistical distributions and their moments

Joint Distribution

$$F_{X,Y}(x, y : \theta) = P(X \leq x, Y \leq y : \theta).$$

Where $x \in R^{(p)}$, $y \in R^{(q)}$ and the inequality $\leq$ is a joint distribution function of X and Y with parameter θ. The behavior of X and Y is characterized by $F_{X,Y}(x, y : \theta$

If the joint probability density function $f_{x,y}(x, y : \theta)$ exists then

$$F_{X,Y}(x, y : \theta) = \int_{-\infty}^{x} \int_{-\infty}^{Y} f_{x,y}(w; z; \theta) dz dw.$$

Where X and Y are continuous random vectors

Marginal Distribution

Given by

$$F_X(X;\theta) = F_{X,Y}(x,\infty,\dots,\infty,\theta).$$

Thus, the marginal distribution of X is obtained by integrating out Y. A similar definition applies to the marginal distribution of Y If $k=1$ X is a scalar random variable and the distribution function becomes

$$F_X(x) = P(X \leq x;\theta).$$

Which is the CDF of X. The CDF of a random variable is nondecreasing and satisfies $F_X(-\infty)=0$and $f_X(\infty)=1$ For a given probability p, the smallest real number x_p such that $p \leq F_X(x_p)$ is called the 100 p th quantile of the random variable X

Conditional Distribution

The conditional distribution of X given $y \leq y$ is given by

$$F_{X|Y\leq y}(x;\theta) = \frac{P(X \leq X, Y \leq Y : \theta}{P(Y \leq Y : \theta)}.$$

Moments of a Random Variable

The l-th moment of a continuous random variable X is defined as

$$M_l' = E[X^l] = \int_{-\infty}^{\infty} x^l f(x)dx$$

Where E stands for expectation and $f(x)$ is the probability density function of x. The first moment is called the mean or expectation, measuring the central location of the distribution.

The l-th central moment of X is defined as

$$M_l = E[(X-\mu_x)^l] = \int_{-\infty}^{\infty} (x-\mu_x)^l f(x)dx$$

The second central moment, denoted σ_x^2 measures the variability of X and is called the variance of X. The positive square root σ_x of variance is the *standard deviation* of X.

The first two moments of a random variable uniquely determine a normal distribution.

The Third Central moment measures the symmetry of X with respect to its mean, whereas the *fourth central moment* measures the tail behaviour of X.

Skewness and *kurtosis* are normalised third and fourth central moments of X, are often used to summarise the extent of asymmetry and tail thickness

1.2 Descriptive Statistics

Let Y_t be a time-series of random variables with a history of realisations y_t with $t=1,\dots,T$

Mean

$$E[Y_t] = \mu \quad , \quad \hat{mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

Variance

$$V[Y_t] = E[(Y_t-\mu)^2] \quad , \quad \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{mu})^2$$

Skewness

$$S[Y_t] = E[\frac{(Y_t-\mu)^3}{\sigma^3}] \quad , \quad \hat{S} = \frac{1}{T}\sum_{t=1}^{T}[\frac{(Y_t-\mu)^3}{\sigma^3}][]$$

Kurtosis

$$K[Y_t] = E[\frac{(Y_t - \mu)^4}{\sigma^4}] \quad , \quad \hat{S} = \frac{1}{T}\sum_{t=1}^{T}[\frac{(Y_t - \mu)^4}{\sigma^4}][]$$

Jargue-Bera test, tests H_0 of normality of the series :

$$JB = \frac{T}{6}(\hat{S}^2 + \frac{(\hat{K} - 3)^2}{4})$$

Where k is the number of estimated parameters. This test statistic has a χ^2 distribution with 2 degrees of freedom (always). Tests 2 parameters jointly. Rejection when skewness is not 0 or kurtosis is not 3. Skewed or heavy tailed. Then use individual tests against 0 or 3 using WLLN and CLT. T test, standardizing appropriately.

Quantile-Quantile plots : plot theoretical quantiles against the empirical ones

Stylized Facts

- Return series do not follow a normal distribution
- The normal distribution does not explain the occurrence probability of extreme events
- Better assumptions are student-t or stable distributions
- On higher frequencies (intraday) the deviation from normality is more pronounced than on lower frequencies
- Aggregated return series, do however, tend to normality
- Return series posses fat tails
- Return series are leptokurtic or posses an overkurtosis (kurtosis $>$ 3)
- Large returns occur more often than expected
- Large returns are more often negative than positive which yields left skewed returns (skewness $<$ 0)
- Intraday returns are subject to typical trading session effects (seasonality, opening and closing issues)
- Returns are subject to volatility clustering, which is again more pronounced on higher frequencies
- Volatility is time varying
- Financial time series are correlated
- Correlations are also time varying

Standardized Return

$$(\frac{r_t - \mu}{\hat{\sigma}}).$$

Kurtosis is probably the most important, telling you about the number of extreme events. Say coca-cola vs tesla (kurtosis of 50). Can be seen as number of outliers around mean

Plotting histogram, kurtosis is heavy tails, extreme distribution lands exactly to the tails.

1.3 Distribution of Returns

The most general model for the log returns is its joint distribution function $F_r(r_{11},\ldots,r_N : r_{12},\ldots,r_{N2} : \ldots r_{IT}\ldots r_{NT} : Y;\theta)$

Where Y is a state vector consisting of variables that summarise the environment in which asset returns are determined and θ is a vector of parameters that uniquely determines the distribution funcion $F_r(\cdot)$, which governs the stochastic behaviour the returns r_{it} and $\mathbf{Y}$.

Often the state vector Y is treated as given and the main concern is the conditional distribution of $\{r)it\}$ given Y.

Some financial theories (CAPM) focus on the joint distirbution of N returns at a single tome index t. Whilst others look at the dynamic structure of individual asset returns

Since our main concern is the joint distribution of $\{r_{it}\}_{t=1}^{T}$ for asset i, it is useful to partition the joint distribution as :

$$\begin{aligned}F(r_{i_1},\ldots,r_{iT}:\theta) &= F(r_{i1})F(r_{i2}|r_{i1})\ldots F(r_{iT}|r_{iT-1},\ldots,r_{i1})\\ &= F(r_{i1})\prod_{t=2}^{T}F(r_{iT}|r_{it-1},\ldots,r_{i1})\end{aligned}$$

Where the parameter θ is omitted for brevity.

This partition the temporal dependencies of the log return r_{it}. With the main issue the specification of the conditional distribution $F(r_{it}|r_{i\ t-1})$ since different distributional specification lead to different theories in finance.

For instance the *random walk hypothesis* in which one version entails the conditional distribution $F(r_{it}|r_{i\quad t-1},\ldots r_{i1})$ is equal to the marginal distribution $F(r_{it})$ meaning returns are temporally independent and thus not predictable.

Normal Distribution

A traditional assumption is that the simple returns $\{R_{it}|t = 1,\ldots,T\}$ are **independently and identically distributed** as normal with fixed mean and variance.

However, this assumption encounters difficulties empirically,

- The lower bound of a simple return is -1, but the normal distribution may assume any value in the real line and hence has no lower bound
- If R_{it} is normally distributed then the multi period simple return $R_{it}[k]$ is not normally distributed because it is a product of one period returns
- The normality assumption is not supported by many empirical asset returns

Log normal Distribution

Another commonly used assumption is that he long returns r_t of an asset are independent and identically distributed (iid) as normal with mean μ and variance σ^2. The simple returns are then iid lognormal random variables with mean and variance given by

$$E[R_t] = \exp(\mu + \frac{\sigma^2}{2}) - 1$$

And

$$Var[R_t] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$$

Stable Distribution

The stable distribution are a natural generalisation of normal in that they are stable under addition, meeting the need of contionusly compounded returns r_t. Furthermore, stable disributinos are capabale of capturing excess kurtosis, shown by histroical stock returns

Hypothesis Test

Null $H_0 : s = 0$ vs $H_1 : S \neq 0$

$$\hat{t} + CLT \rightarrow^{(d)} N(0,1).$$

Tells you distribution under the null, then 95% of probability mass is between critical values, then outside of this, either suff evidence against the null or a type I error (5%) (at tails). Fundamentally, we cannot trust the null hypothesis.

Whatever we want to test, we put into the alternative. NO conclusion can be made if we fail to reject the null. If we collect evidence against the null then this is fundamentally different.

Lecture 2: Second Lecture - Review of Time Series

Fri 02 Feb 15:45

[Lecture 2] [Reading]

2 Time Series Basics

Stationarity

A time series $\{r_t\}$ is *strict stationary* if the joint distributing of $(r_{t_1}, \dots, r_{tk})$ is identical to that of $(r_{t_1+t}, \dots, r_{t_k+t})$ for all t where k is an arbitrary positive integer and $t_1, \dots, t_k$ is a collection of k positive integers.

That is, *strict stationarity* requires that the joint distributing is *time invariant* under a time shift, but of course this is hard to verify empirically and a very strong condition.

A weaker condition is that a time series $\{r_t\}$ is *weakly stationary* if both the mean of r_t and the covariance between r_t and r_{t-l} are time invariant.

That is,

$$E[r_t] = \mu \text{ a constant } Cov(r_t, r_{t-l}) = \dagger_l \text{ which only depends on l} \tag{2}$$

Where weak stationarity implies the time plot of the data would show that the T values fluctuate with *constant* variation around a fixed level. Enabling one to make inference conceding future observations.

But, implicitly we have assumed that the first 2 moments of r_t are finite

Where the covariance $\dagger_l = Cov(r_t, r_{t-l})$ is called the lag-$\updownarrow$ auto covariance of r_t. With 2 important properties :

1. $\dagger_0 = Var(r_t)$
2. $\dagger_{-l} = \dagger_l$

Correlation and autocorrelation functions

The correlation coefficient between 2 random variables X and Y is defined as

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}} \tag{3}$$

Where μ_x and μ_y are the mean of X and Y resp, assuming the variances exist also. This measures the *strength* of linear dependence between X and Y, and it can be shown that $-1 \leq \rho_{x,y} \leq 1$ and $\rho_{x,y} = \rho_{y,x}$. Where the two RV are uncorrelated if $\rho_{x,y} = 0$, which occurs iff X and Y are independent.

Autocorrelation Function (ACF) For a weakly stationary return series r_t when the linear dependence between r_t and its past values r_{t-i}, we can generalise the correlation concept to autocorrelation.

The correlation coefficient between r_t and r_{t-i} is called the lag-ℓ autocorrelation of r_t and is commonly denoted by ρ_e, which under the weak stationarity assumption is a function of ℓ only

We define

$$\rho_\ell = \frac{\text{Cov}(r_t, r_{t-\ell})}{\sqrt{\text{Var}(r_t)\text{Var}(r_{t-\ell})}} = \frac{\text{Cov}(r_t, r_{t-\ell})}{\text{Var}(r_t)} = \frac{y_\ell}{y_0}$$

Where $\text{Var}(r_t) = \text{Var}(r_{t-\ell})$ for a weakly stationary series

Stochastic Processes

- Chronologically ordered equidistant observations
- Generated by stochastic process
- Stochastic process - collection of RV (each Y_i is generated by different member of stochastic processes)
- **assumption** time series data has been generated by *stochastic process*

Definition 1. **stochastic process** is a family of random variables defined on a probability space

Definition 2. **time series** is a realisation of a stochastic process

Definition 3. **time series analysis** - only one history $Y_t(w)$, one state of the world $w \in \omega$ is available, but the goal is to derive the properties of $Y_t(\cdot)$ for a given t for different states of the world

Idea - how can we understand what is driving omegas? Different states of the world, since we observe y_t. So place some structure on y_t

Should be able to recognise :

- Non-stationary time series
- Autoregressive time series
- Kurtosis time series

Definition 4. Auto Covariance

Time series often show correlation between successive observations, this feature is called serial correlating or **autocorrelation**

Dependencies over time are described by auto covariance and autocorrelation functions

The j-th autocovariance of Y_t is given by

$$Cov[Y_t, y_{t-j}] = \gamma_{t,t-j} = E[Y_t - E[Y_t][Y_{t-j} - E[Y_{t-j}]$$

Correspondingly the variance of Y_t is defined as :

$$V[Y_T] = \gamma_{t,t} = E[Y_t - E[Y_t])^2]$$

Definition 5. Autocorrelation

The j-th autocorrelation of Y_t is given by :

$$\rho_{t,t-j} = \frac{Cov[Y_t, Y_{t-j}]}{V[Y_t]^{(\frac{1}{2})} V[Y_{t-j}^{(\frac{1}{2})}}$$

Definition 6. Covariance Stationary A time series $\{Y_t\}_{t=-\infty}^{(\infty)}$ is called covariance stationary, or weakly stationary, if :

$$\begin{aligned} E[Y_t] &= \mu_Y \\ V[Y_t] = \gamma_{t,t} &= \gamma_0 = \sigma_Y^2 < \infty \\ Cov[Y_t, Y_{t-j}] = \gamma_{t,t-j} &= \gamma_j < \infty \end{aligned}$$

For a covariance stationary process the j-th autocorrelation is given by :

White Noise

Definition 7. white noise A T is called this if it satisfies the following

$$E[Y_t] = 0 V[Y_t] = \sigma_Y^2 COv[Y_t, Y_s] = E[Y_t, Y_s] = 0$$

White noise is a weakly stationary process - all the ACFs are 0.

Particularly, if r_t is normally distributed with mean 0 and variance $= \sigma^2$ the series is a Gaussian white noise

Definition 8. Autocorrelation Function Autocorrelation function of a covariance stationary process $\{Y_t\}_{t=-\infty}^{(\infty)}$ is the sequence of autocorrelations ρ_j for all $j = 0, 1, 2, \ldots$

Definition 9. Empirical Autocorrelation Function

The empirical (or sample) autocorrelation function of a time series Y_t is the sequence of sample autocorrelation coefficients $\hat{\rho}_j$ for all $j = 0, 1, 2, \ldots$:

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0} = \frac{\sum_{t=j+1}^{T} (Y_t - \overline{Y}(Y_{t-j} - \overline{Y})}{\sum_{t=1}^{T} (Y_t - \overline{Y}^2}$$

And

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} (Y_t - \overline{Y})(Y_{t-j} - \overline{Y}) \qquad \overline{Y} = \frac{1}{T} \sum_{t=1}^{T} Y_t$$

The graphical depictions of the empirical autocorrelation function is called an autocorrelogram

Definition 10. Partial Autocorrelation Function

Partial autocorrelation between Y_t and Y_{t-j} is the conditional correlation between Y_t and Y_{t-j} given (holding fixed) $Y_{t-1}, \dots, Y_{t-j+1}$

$$A_j = Cor[Y_t, Y_{t-j}|Y_{t-1}, \dots, Y_{t-j+1}]$$

Effects of in-between values are controlled for

Corresponding sample quantity $\hat{a}_j$ is called sample partial autocorrelation and is obtained as the OLS estimator of the coefficient a_j in model

$$Y_t = a_0 + a_1 Y_{t-1} + \dots + a_j Y_{t-j} + \mu_t$$

Definition 11. Sample Autocorrelation Function

If data generating process is a white noise process, then for large T:

$$\hat{\rho}_j \approx N(0, \frac{1}{T}), j = 1, 2, \dots$$

Means : $H_0 : \hat{\rho}_j = 0$ is rejected, if zero does not fall within the approximate 95% confidence interval

$$[r\hat{h}o_j - \frac{2}{\sqrt{T}}, r\hat{h}o_j + \frac{2}{\sqrt{T}}]$$

Equivalently, autocorrelations are not significant when $\hat{\rho}_j$ is within the approximate two standard error bound $\pm 2/\sqrt{T}$

Linear Time Series A time series r_t is said to be linear if it can be written as

$$r_t = \mu + \sum_{i=0}^{\infty} \psi_i \alpha_{t-i} \tag{4}$$

Where μ is the mean of r_t, $\psi_0 = 1$ and $\{\alpha_t\}$ is a sequence of iid RV with mean zero and well defined distributions (ii a white noise)

For this equation, the dynamic structure of r_t is governed by the coefficients ψ_i which are called the ψ weights of r_t

If r_t is weakly stationary, we can obtain its mean and variance easily by using the independence of $\{\alpha_t\}$ as

$$E[r_t] = \mu \qquad \text{Var}(r_t) = \sigma_\alpha^2 \int_{i=0}^{\infty} \psi_i^2 \tag{5}$$

Where σ_u^2 is the variance of a_t

Simple AR Models

If a monthly return of a value weighted index has a statistically significant lag-1 autocorrelation indicates that the lagged return r_{t-1} may be useful in predicting r_t, we can implement this in a model such as

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \tag{6}$$

Where $\{a_t\}$ is assumed to be a white noise series with mean zero and variance σ_a^2

This is analogous to the simple linear regression model in which r_t is the dependent variable and r_{t-1} is the explanatory variable.

This is actually an autoregressive (AR) model of order 1 (AR(1))

Note. Conditional on the past return, the current return is centred around $\phi_0 + \phi_1 r_{t-1}$ with CID σ_a AR(1) model implies that conditional on past return r_{t-1}, we have

$$E[r_t|r_{t-1}] = \phi_0 + \phi_1 r_{t-1} \qquad \text{Var}(r_t|r_{t-1}) = \text{Var}(a_t) = \sigma_a^2$$

This is a Markov property such that conditional on r_{t-1}, the return r_t is not correlated with r_{t-i} for $i > 1$

A straightforward generalisation of the AR(1) model is the AR(p) model

3 ARMA Processes

Definition 12. AR(p)-Process A time series is called an autoregressive process of order p if it satisfies a relationship of the type :

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \rho_p Y_{t-p} + \varepsilon_t$$

Where ε_t is a white noise error term

A(1) process : the simplest form of an A(p) process is obtained for $p = 1$ as

$$Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t$$

$$r_t = \phi_0 + \phi_1 r_{t-1} + \ldots + \phi_p r_{t-p} + a_t$$

Where p is a non-negative integer and $\{a_t\}$ is defined previously.

Essentially, this says that the past p variables $r_{t-i} i = 1, \ldots, p$jointly determined the conditional expectation of r_t given the past data.

AR(1) Properties

Considering an *AR(1) process:*

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \qquad -1 < \rho < 1$$

where u_t is a white noise process. This AR(1) process as the following properties

1. $\varepsilon_t = \sum_{i=0}^{\infty} \rho^i u_{t-i} = \quad \mathbf{MA}(\infty)$
2. $E[\varepsilon_t] = 0$
3. $V[\varepsilon_t] = \gamma_0 = \frac{\sigma_u^2}{1-\rho^2}$ with $V[u_t] = \sigma_u^2$

4.

$$\begin{aligned}\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) &= \gamma_1 = \rho_1 \frac{\sigma_u^2}{1-\rho^2}\\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-s}) &= \gamma_s = \rho_s \frac{\sigma_u^2}{1-\rho^2}\\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) &= \gamma_1 = \rho_1 \frac{\sigma_u^2}{1-\rho^2}\end{aligned}$$

Definition 13. MA(q)-Process A time series is called a **moving average process of order** q if it satisfies a relationship of the type

$$Y_t = \mu = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

Where ε_t is a white noise error term

MA(1) Process: the simplest form of an MA(q) process is obtained for $q = 1$ as

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Example. AR(1) Process

$$\sum_{i=0}^{\infty} \rho^{(i)} u_{t-i} =^{(wald)} MA(\infty)$$

3.1 ARMA

Often times the AR or MA models become cumbersome, so one may need a higher order model with many parameters to describe the dynamic structure of the data.

ARMA models are introduced essentially combining both AR and MA models, such that the number of parameters is kept small. Importantly, the GARCH model can be regarded as an ARMA model.

A time series r_t follows an ARMA(1, 1) model if it satisfies

$$r_t - \phi_1 r_{t-1} = \phi_0 + a_t - \theta_1 a_{t-1}$$

where $\{a_t\}$ is a white noise series. With the left and right giving the AR and MA part resp.

Lag operator let $\{Y_t\}_{t=-\infty}^{(\infty)}$ be a time series, then the lag operator $\mathcal{L}$ is defined by the relation

$$L^{(J)} \equiv Y_{t-j}$$

If $\{Y_t = c\}_{t=-\infty}^{(\infty)}$ where $c \in \mathbb{R}$, then $\mathcal{L}^{(j)} Y_t = L^{(j)} c = c$

ARMA(p,q) is a time series $\{Y_t\}_{t=-\infty}^{(\infty)}$ of the following form

$$\begin{aligned}\phi_p(L) Y_t &= c + \Theta(L) \varepsilon_t where\\ \phi_p(L) &= 1 - \phi_1 L - \phi_2 L^{(2)} - \ldots . - \phi_p L^{(p)}\\ \Theta(L) &= 1 + \theta_1 L + \theta_2 L^{(2)} + \ldots + \theta_q L^{(q)}\end{aligned}$$

With ε_t being a white noise and ϕ_p and Θ_q are called lag polynomials

Properties of ARMA(1, 1) Models

[2.6] These are generalisation of those of AR(1) models with some modifications to handle the MA(1) component.

Starting with the stationarity condition and taking expectation of the ARMA(1, 1) model :

$$E[r_t] - \phi_1 E[r_{t-1}] = \phi_0 + E[a_t] - \theta_1 E[a_{t-1}]$$

Because $E[\alpha_i] = 0$ for all I, the mean of r_t is

$$E[r_t] = \mu = \frac{\phi_0}{1 - \phi_0}$$

provided the series is weakly stationary.

Then, assuming for simplicity that $\phi_0 = 0$ we consider the autocovariance function of r_t

Multiplying the model by a_t and taking expectations we have

$$E[r_t a_t] = E[\alpha_t^2] - \theta_1 E[a_t a_{t-1}] = E[\alpha_t^2] = \sigma_a^2 \tag{7}$$

Then rewriting the model as

$$r_t = \phi_1 r_{t-1} + \alpha_t - \theta_1 a_{t-1}$$

and taking the variance of the prior equation, we have

$$\mathrm{Var}(r_t) = \phi_1^2 \mathrm{Var}(r_{t-1}) + \sigma_a^2 + \theta_1^2 \sigma_a^2 - 2\phi_1\theta_1 E[r_{t-1}a_{t-1}]$$

Where we make use of the fact that r_{t-1} and a_t are uncorrelated, then using 7 we obtain

$$\mathrm{Var}(r_t) - \phi_1^2 \mathrm{Var}(r_{t-1}) = \left(1 - 2\phi_1\theta_1 + \theta_1^2\right) \sigma_a^2$$

Therefore if the series is weakly stationary, then $\mathrm{Var}(r_t) = \mathrm{Var}(r_{t-1})$ and we have

$$\mathrm{Var}(r_t) = \frac{(1 - 2\phi_1\theta_1 + \sigma_1^2)\sigma_a^2}{1 - \phi_1^2}$$

3.2 ARMA estimation

ARMA(p, q) process:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 e_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

- Estimation via conditional Max likelihood
- **conditional** : derive the likelihood function under the assumption that the initial values of Y_t and ε_t are available
- **assume** : $\varepsilon_t \sim^{(iid)} N(0, \sigma^2)$
- ML parameter estimators are derided under the assumption of normality are quasi ML estimators
- Our goal is to estimate the vector $\theta = (c, \phi_1, \phi_2, \ldots, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$

ARMA estimation

Conditional log likelihood

Estimation is done under assumption that error term is normal.

LBJ test

Whether p is sufficiently long, if model specified correctly, then residuals shouldn't be correlated with each other.

Tells whether white noise property is plausible assumption

Critical values is from chi-squared dost, we test for absence of autocorrelation upto chosen lag order, leading into next weeks lecture of conditional heteroskedacity.

ARCH-ML test

Tests for conditional heteroskedacity in regression residuals

Pick ARMA based on this, if modelled successfully then null of LBJ test shouldn't be rejected and there shouldn't be any conditional heteroskedacity

Lecture 3: ARCH Models

Mon 12 Feb 09:01

[Lecture 3] [Reading]

Review

Week 1

Leptokurtic Property - How to measure a lot of outliers? Kurtosis. The kurtosis of our distribution is larger than 3 (4th moment of distribution). Since $K[r_t] > 3$ where $e \sim \mathcal{N}(\mu, \sigma^2)$

Left-Skewness - more negative returns than positive ones. $S[r_t] < 0$

Volatility clustering - periods of high volatility are followed by periods of high volatility. The volatile periods on the markets (across S of return distribution) they *cluster.* Market volatility is persistent.

Shape of daily returns - Compared to say a normal bell curve, is this a good distribution? Weekly more normal then daily, monthly more normal than weekly. Thus *aggregate returns tend to normality*

- Should know these by heart
- And be able to apply them and tell graphically

Week 2

Time series analysis

ARMA Models ($ARMA(1, 1)$ $y_t = c + \rho_1 + y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$

Stationarity ADF test

Model selection ACF /DACF

Information criteria - Bayesian information criteria helps to choose whether ARMA(1, 2) or MA(1) is better for data.

At the end we estimate by quasi-likelihood since $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, which is key to today's material.

It is important to realise this assumption is quite strong, the shortcut for this type of estimator is quasi-likelihood.

$\Theta = (c, \rho_1, \theta_1, \sigma^2)$, in empirical terms the maximum likelihood estimators minimise the negative log-likelihood, we can only find the minimum using gradient descent, hence minimising the negative.

$\hat{\theta}_{QML} = \underset{\Theta}{argmin}$

Autoregressive process order 1

Log likelihood - maximises function to find combination of parameters of model such that our ε_t's are normal

For Financial Econometrics, once plot ACF and PCF, once looking at squared residuals, we have seen a lot of significant lags in the squared residuals. LBQ test and ARCH-LM test whether there is remaining autocorrelation within the squares residuals.

These tests tell us that $\hat{\sigma^2}$ tell us there is autocorrelation across time within the residuals, only problem of model misspecification comes from squared residuals, variance of error term.

3.3 Conditional Heteroskedacity

In any ARMA model there is some expectation

$$Y_t = E[y_t|F_{t-1}] + \varepsilon_t$$

$c + \rho y_{t-1} + \theta\varepsilon_{t-1}$. F is filtration, past information and ε_t is new information/shock today.

White noise (DSA) :

$$\begin{aligned}\varepsilon_t &\sim WN \\ E[\varepsilon_t] &= 0 \\ V[\varepsilon_t] &= \sigma^2 \\ Cov(\varepsilon_t, \varepsilon_t) &= 0\end{aligned}$$

What is the difference between conditional and unconditional moments?

Conditional : $V[\varepsilon_t]$ and Unconditional $V[\varepsilon_t|F_{t-1}]$

$$\begin{aligned}Y_t &= c + \rho y_{t-1} + \varepsilon_t \\ E[y_t] &= E[c + \rho y_{t-1} + \varepsilon_t] \\ &= c + \rho E[y_{t-1}] \\ E[y_t] &= E[y_{t-1}] \\ E[y_t] &= \frac{c}{1-\rho}\end{aligned}$$

That is,

$$\frac{c}{1-\rho} \quad vs \quad c + \phi y_{t-1} \qquad (*)$$

$$\begin{aligned}&E[y_t|F_{t-1}] \\ &E[c + \rho y_{t-1} + \varepsilon_t|F_{t-1}] \\ C + \rho E[y_{t-1}|F_{t-1}] &+ E[\varepsilon_t|F_{t-1}] \\ &C + \rho y_{t-1} + 0\end{aligned}$$

White Noise

- $E[\varepsilon_t] = 0$
- $V[\varepsilon_t] = \sigma^2$
- $cov[\varepsilon_t, \varepsilon_t] = 0$

The unconditional moment in (*) is more important.

White noise assumption, assumes both conditional and unconditional are constant over time, that is

$$V[\varepsilon_t] = V[\varepsilon_t, |F_{t-1}] = \sigma^2$$

$V[\varepsilon_t] = \sigma^2$ but $V[\varepsilon_t, F_{t-1}]$ is time varying (conditional second moment).

We start with $\varepsilon_t = \mathcal{L}_t \cdot \sigma_t$ where $\mathcal{L}_t \sim \mathcal{N}(0,1)$ and ARCH (1) : $\sigma_t^2 = w + \alpha\varepsilon_{t-1}$

As we have just done with AR1, now look at conditional and unconditional second moment of ARCH(1).

$V[\varepsilon_t]$ and

$$\begin{aligned} E[\varepsilon_t] = E[\mathcal{L}_t\sigma_t] = \\ E[\mathcal{L}_t]E[\sigma_t] \\ 0 \cdot E[\sigma_t] = 0 \\ V[\varepsilon_t] = E[\varepsilon_t]^2 E[\mathcal{L}_t^2 \cdot \sigma_t^2] = \\ E[\mathcal{L}_t^2] \cdot E[\sigma_t^2] = E[\sigma_t^2] \end{aligned}$$

$V[\varepsilon_t|F_{t-1}]$ (not right yet)

$$\begin{aligned} E[\varepsilon_t|F_{t-1}] = \\ E[\mathcal{L}_t]E[\sigma_t] \\ 0 \cdot E[\sigma_t] = 0 \\ V[\varepsilon_t] = E[\varepsilon_t]^2 E[\mathcal{L}_t^2 \cdot \sigma_t^2] = \\ E[\mathcal{L}_t^2] \cdot E[\sigma_t^2] = E[\sigma_t^2] \end{aligned}$$

General Settings

So far we have focused on the estimation of the conditional mean function $E[Y_t|F_{t-1}]$:

$$Y_t = E[Y_t|F_{t-1}] + \varepsilon_t$$

Where ε_t is a weak white noise, that is, ε_t is serially uncorrelated : $Cov[\varepsilon_t, \varepsilon_{t-j}] = 0 \;\; \forall j \neq 1$

ARCH(1) Processes

A process σ_t^2 is called an ARCH(1) process if

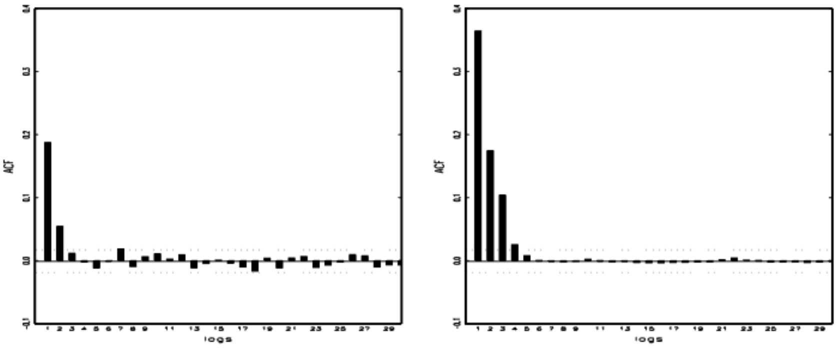
$$\sigma_t^2 + w + \alpha\varepsilon_{t-1}^2$$

With w>0 and $\alpha \geq 0$

Properties of Arch(1)

- Arch (!) Conditional variance σ_t^2 is strictly positive if $w > 0$ and $a \geq 0$
- Opposite to the historical volatility estimator, the arch 1 volatility is a weighted average of past information that gives more weight to the recent information than to the distant one
- Arch 1 process can be written as an A(1) process in ε_t^2
- Consequently ε_t^2 is stationary if $|\alpha| < 1$

- Given that both process ε_t and ε_t^2 and $E[\varepsilon_t] = 0$ then the unconditional variance of ε_t, $E[\varepsilon_t]$ is given by
$$\sigma_\varepsilon^2 = V[\varepsilon_t] = E[\varepsilon_t^2] = \frac{w}{1-\alpha}$$
- ARCH(1) captures the clustering effect : when volatility is high, it more probably stays high
- The kurtosis is always large t



utocorrelation functions of squared time series with ARCH(1) conditional variance with $\alpha = 0.2$ (left panel) and $\alpha = 0.7$ (right panel)

Figure 1

Conditional variance moment, we observe a high persistence in daily log returns in order to cauterises this lag persistence, this lag has to be large too. But the estimation of this A(50) model becomes very cumbersome, likelihoods optimise numerically, once you start imposing Stationarity conditions this it rot ensure generating something with a stationary second moments, these are some solutions to polynomial equations so we run into large p issues.

In tutorial we look at arch's in simulation study

Lecture 4: GARCH

Mon 19 Feb 09:00

[FE-L4] [3.5, 3.6, 3.8,3.9]

Recap

ARMA

1. $E[\varepsilon_t] = 0$
2. $V\varepsilon_t = \sigma^2$
3. $\text{Cov}(\varepsilon_t, \varepsilon_s) = 0$ that is no serial correlation

Tutorial 2 : S&P 500 Daily log returns → ARMA(p,q) → BIC then use residual diagnostics

$$@_t = y_t - \hat{E}[y_t|F_{t-1}] \to MA(\mathcal{L})$$

Week 3
NP / Rob Engel 2003

$$\begin{aligned} \varepsilon_t &= \sigma_t \mathcal{L}_t \\ \mathcal{L}_t &\sim \mathcal{N}(0,1) \\ \sigma_t^2 &= w + \alpha \varepsilon_{t-1}^2 < - \end{aligned}$$

1. $a \geq 0$ and $\omega > 0$ - to ensure positivity of conditional variance
2. $|\alpha| < 1$ Stationarity of conditional variance

ARCH(1)

$$\begin{cases} \sigma = w + \alpha \varepsilon_{t-1} < - \\ \varepsilon_t + \mathcal{L}_t \sigma_t \\ \quad \text{rewrite } \ \sigma_t^2 = w + \alpha \varepsilon_{t-1}^2 + \varepsilon_t^2 - \varepsilon_t^2 \\ AR(1) \ in \ \varepsilon_t^2 \rightarrow \varepsilon_t^2 = w + \alpha \varepsilon_{t-1}^2 + \left(\varepsilon_t^2 - \sigma_t^2 \right) \end{cases} \qquad Video \begin{cases} E[V_t] = 0 \\ V[v_t] < \infty v_t = \sigma^2 \\ Cov(v_t, v_{t-s}) = 0 \end{cases}$$

Pros

- Volatility clustering (video)
- Rise persistence at the cost of ARCH (p)
- Leptokurtic property $\alpha^2 \in (0, \frac{1}{3})$

Cons

- Leverage effect : $E[\mathcal{L}_t^3] = 0$
- Long memory (ACF)

What can we do with our Garch models to capture all remaining things in ACF?

3.4 GARCH

A process σ_t^2 is called an GARCH(1, 1) process if

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

With $\omega > 0$, $\alpha \geq 0$ and $\beta \geq 0$

Properties

- ε_t^2 is stationary if $\alpha + \beta < 1$
- Both processes ε_t and ε_t^2 are stationary and $E[\varepsilon_t] = 0$ then the unconditional variance of ε_t $V[\varepsilon_t]$ which is equal to the unconditional mean of ε_t^2
- No leverage effects as in the ARCH
-

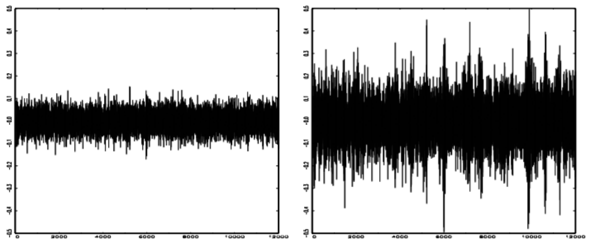


Figure 2: Simulated GARCH Models

Left $\alpha = 0.01$ and $\beta = 0.8$. Right $\alpha = 0.08$ and $\beta = 0.9$ If allow close to 1 then can generate longer persistence, usually the memory of the daily log returns is us more persistent. Most have very low memory, thus people came up with GARCH(p, q)

M1 GARCH(1, 1)

- It takes into account / able to model more persistent conditional volatility processes
- Mitigating the tradeoff between generating a leptokurtic distribution of ε_t and the persistence iof the ACF if ε_t^2 as compared to ARCH(1)

M2 ARCH(1)

GARCH captures over kurtosis, even if we could like sum of $\alpha + \beta$ to 1, we still have an opportunity to generate a over kurtosis (>3)

We can also show GARCH reveals larger excess kurtosis than the arch model, we can compare which is larger than the other, $\frac{6\alpha^2}{1-2a^2-(\alpha+\beta)^2)}$

Can show A(1) is equal to $MA(\infty)$, same applies for GARCH for $ARCH(\infty)$

$\alpha + \beta$ providers the necessary information on the degree of volatility clustering

GARCH(p, q)

Just extension of GARCH(1, 1), key notation is polynomial for lag operator, lags shift an observation 1 period ahead (power 2 = 2 period ahead). But except for notation, nothing fundamental changes.

To lie outside of the root circle, in practice to estimate such a model, ensure positivity constraints, then also have to ensure process modelling is stationary - the constraints on stationary on highly non linear. This very quickly becomes a complicated non linear constraint, thus a numerical issue driven by Stationarity constraint (non linear) imposed by IRMA (p, q), but if allow for more p and q lags, then model is able to generate over kurtosis then the persistence of the series, the properties become better but at the cost of optimising over something with highly non-linear constraint.

Further Types of GARCH models

ARCH providers an exponential decay, have to know GARCHS for risk modelling.

Integrated GARCH(1, 1)

- Specific to high frequency time series
- Describes a very large persistence in the conditional variance
- Is strictly stationary

- Propose α and β sum upto 1, GARCH STRUCTURE there to ensure non stationary process
- Risk metrics assumes that daily log returns follows process with infinite variance, that is we are not dealing with well defined statistical processes in real life, as seen by lack of first 2 moments

RiskMetrics^{TM} A special case of the IGARCH(1, 1) process

- From estimating the
- Gives forecast
- λ calibrates on loads of different stocks in the 90s
- Fix the β with λ

Exponential GARCH Aimed at capturing asymmetric shocks, now modelling h_{t-1} log transformation of σ_t^2, assuming it follows GARCH looking process, and modify the ARCH part

- Modelling logs of variance because we want to get rid of parameter constraints, if modelling logs can be positive, negative, get rid of these issues by modelling logs
-

Threshold GARCH TGARCH (1, 1) with indicator function, if shock was negative, bit easier to look at, if γ is positive, then ...

Tgarch, E garch if model left skewed
Tgarch(1,1) GJR-Garch
Usual garch(1, 1) : $\sigma_t^2 = w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
Tgarch(1, 1) : $\sigma_t^2 = w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
News impact curve : $NIC(\varepsilon_t|\sigma_{t-1}^2 = \sigma_{t-2}^2, \dots, = \sigma_t^2)$
GARCH(1, 1) : $w + \beta\sigma_t^2 + \alpha\varepsilon_{t-1}^2$ TGARCH(1, 1) $=$

$$\begin{cases} W + \beta\sigma_t^2 + \alpha\varepsilon_{t-1}^2 \varepsilon_{t-1} < 0 \\ W + \beta\sigma_t^2 + \alpha + \delta\varepsilon_{t-1}^2, \varepsilon_{t-1} < 0 \end{cases}$$

NIC : Egarch(1,1)

$$H_t = \ln(\sigma_t^2) = w + \alpha\mathcal{L}_{t-1} + \gamma(|z_{t-1}| - \sqrt{\frac{2}{a}}) \exp(h_t) = \sigma_t^2 = \exp^w \cdot \exp^{\alpha z_{t-1}} \cdot \exp^{\gamma(|z_{t-1}| - \sqrt{\frac{2}{a}})} \sigma_t^2 = \exp^w \cdot \sigma^2$$

$$\varepsilon_t > 0$$
$$\varepsilon_t < 0$$

If shock positive then $\exp^{\alpha+\gamma} \cdot \varepsilon_t/\sigma_t$

NIC: once you write down NIC, then it becomes more evident what model parameters give you which response, EGarch $\alpha < 0$, z_t between 0 and 1

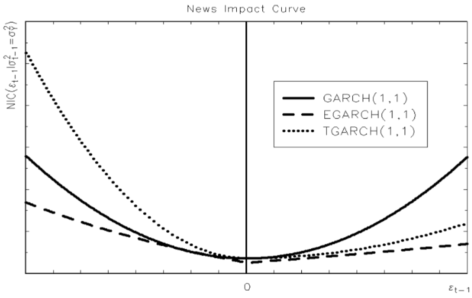


Figure 3: News Impact Curve

Model quality based on one picture, isn't exactly true, in order to plot NIC. Plug in σ, γ, β (ω) , plot is based on one set of parameters, can easily be reversed.

So has something to do with data rather than overall quality of model,

Recap ARCH, GARCH, IGARCH, EGARCH, TGARCH. Financial econometrics model conditional second moment, but what about first moment?

- Conditional mean? (1st moment), why are we interested in the second moment?
- We are risk averse etc, but
- In week 2 we have talked about how to model, ARMA - expected value of y_t then T2 we estimated conditional mean models, but the returns are on average 0, there is very slight autoregressive coefficients, but overall there is **no time series structure** in the conditional mean :

$$E[r_t|F_{t-1}] = 0$$

- WE have compared the ACF for daily log returns r_t, but in the actual return series, the history of returns is completely uninformative of the future
- In autocorrelation function few squared return we see a lot going on, and it doesn't die out, squared return is a proxy of conditional variance

Why do we model conditional second moment?

There is no time series structure to first moment, but there is in conditional second moment. Then we think how can we model our conditional variance of return process?

Nobel prize given for ARMA framework where ε_t can be white noise process. Then, even GARCH is not enough.

Then RiskMetrics comes and assumes infinite variance of daily returns, albeit a popular way of thinking. How much does turbulence persevere in market, how long after do we have to be conservative in our risk approaches

EGARCH, TGARCH. TGARCH more intuitive, EGARCH model the log variances and so can relax the positivity constraints, we don't care whether shocks are negative. Essentially a philosophical introduction to risk-modelling

Lecture 5: Model Estimation and Forecasting

Mon 26 Feb 08:58

[FE-L5] [Reading]

3.5 Recap

Week 1 : Leverage effects (skewness + testing whether neg) , volatility clustering (time series) , long memory (ACF of squared returns series) , leptokurtic property (sample skewness testing against 3). Properties (plots/ test)

Week 2 : limitations of ARMA modelling, which assume innovations are white noise - nothing about conditional heteroskedacity). Unconditional - variance of innovations is constant over time, but evidence empirically that conditional 2nd moment seems to be time variant.

Week 3 : Rob engles ARCH ARCH(1) model $\sigma_t^2 = f(\{_{t-1})$. Pro - volatility clustering, con - leverage of $\{_t$, but long memory for very large p, kurtosis $\alpha^2 \in (0, \frac{1}{s})$

Week 4 : GARCH(1,1) - pro - volatility clustering and long memory and overkurtisis, con - leverage
TGARCH, EGARCH → leverage. M (IGARCH) .

Maximum Likelihood

Quasi Maximum Likelihood

Maximum likelihood - have data $x_1, \dots, x_t$ then **assume** this data follows *some* distribution.

Which is function of the parameters, say $x_t \sim N(\mu, \sigma^2)$ and $\Theta = (\mu, \sigma^2)$

Then have PDF of data $f(x_t, \mu, \sigma^2) = -\frac{1}{\sqrt{2\pi\sigma^2}\exp(-\frac{(x-\mu)^2}{2\sigma})}$. If assume normal dist,then each and every value of x_t you know probability this data came from this distributing, then voter the entire sample you can take the likelihood function

$$\mathcal{L}|_{\mu,\sigma^2} = \prod_{t=1}^{T} f(x_t \mu, \sigma^2)$$
$$= f(x_1|\mu, \sigma^2) \cdot f(x_2|\mu, \sigma^2) \dots$$

So take log likelihood that is a function of data for given value of parameters μ, σ^2

$$\log \mathcal{L}(xq, \dots, x_t|\mu, \sigma^2) = \ln(\prod_{t=1}^{T} f(x_t), |\mu, \sigma^2)$$

In any time series we work with quasi likelihood, in classical ML you must be able to evaluate likelihood function at each and every point. At an autoregressive process of order 1 (AA(1)).

Have $\varepsilon_t \sim N(0, \sigma^2)$ so $\varepsilon_t = y_t - c - \phi y_{t-1}$ which us $N(0, 1)$

Then we have

Why quasi-likelihood?

Likelihood for first population : $f(e_1|c, \phi, \sigma^2)$, we assume y_0 is ...

Now likelihood function becomes function of data and parameters, but *also initial values* depending on how many autoregressive lags are there. → it is not really a likelihood. The conditioning makes it a quasi-likelihood

3.6 Estimation, Model choice and forecasting

Use knowledge of Max likelihood to ascertain which model fits the data best

Assume

$$R_t = c + \varepsilon_t$$
$$E_t = \mathcal{L}_t \sigma_t \sigma_t^2 = w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

Estimate with Max likelihood.

$$\varepsilon_t = r_t - cE_t|_{\mathcal{F}_{t-1}} \sim N(0, \sigma_t^2)$$

$E[c_t|\mathcal{F}_{t-1}]$ and $V[\varepsilon_t|\mathcal{F}_{t-1}]$

Where $\sigma_t = f(\mathcal{F}_{t-1})$ and $\varepsilon_t|_{F_{t-1}} = \mathcal{L}|F_{t-1}\sigma_t|\mathcal{F}_{\sqcup-\infty}$

Where $f(\varepsilon|F_{t-1}, \theta) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp(-\frac{e_t^2}{s\sigma_t^2})$

And

$$\theta = (c, w, \alpha, \rho)$$

$$= \frac{1}{\sqrt{2\pi(w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2)}} \cdot \exp(-\frac{(v_t - c)^2}{2(w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_t^2 - 1)})$$

$$\sigma_0^2 = \frac{w}{1 - \alpha - \beta} = \frac{1}{T} 2(r_t - \hat{\mu})^2$$

Normally distributed innovations. From likelihood theory, the best is the one with the largest likelihood.

Estimation of GARCH Models

Model : $Y_t = X_t'\gamma + \varepsilon_t$

The conditional variance of ε_t follows a GARCH(p, q) model

- M = max(p, q) Numbers of initial observations $t = -m + 1, -m + 2, \ldots, 0$

Conditional maximum likelihood

Normal Z_t

Student t Z_t

Assume $z_t \sim T(v)$ (std student t dist) , then :

$$\begin{aligned} E[Z_t] &= 0 \\ V[Z_t] &= \frac{v}{v - 2} \\ \text{Density Function} \quad \frac{\Gamma[(\nu + 1)/2]}{(\pi\nu)^{1/2}\Gamma[\nu/2]} &\left[1 + \frac{z_t^2}{\nu}\right]^{-(\nu+1)/2} \end{aligned}$$

Often estimated using standardised student t distribution, which is symmetric so expected value is 0, and

In PS1, there was ex on student t distribution with different degrees of freedom - the larger the dof, the closer to normal RV, smaller the hevier the tails (more outliers). 1 dof - Cauchy distribution

3.7 Model Choice and Diagnostics

Verify if there are ARCH effects in

- The origunal series of intrest Y_t
- The residuals froma mean regression $\hat{\varepsilon}_t$ The residuals standardised by the estimated GARCHS $\hat{z}_t = \frac{\hat{\varepsilon}_t}{\sqrt{\hat{\sigma_t^2}}}$

Test for ARCH effects

ARCH-M test

Auxiliary regression on the series of interest $\overline{x}_t$ (original series, residuals, standardised residuals):

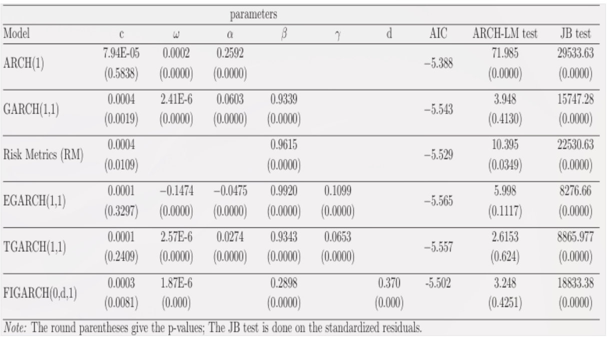
$$\overline{x}_t^2 = \psi + \alpha_1 \overline{x}_{t-1}^2 + \alpha_2 \overline{x}_{t-2}^2 + \ldots + \alpha_m \overline{x}_{t-m}^2 + \varepsilon_t$$

With $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$ and $H_A : H_0$ is not true

Standardised Residual Diagnostics

Assuming you already estimate a GARCH model for series

Verify if there are still ARCH effects left in the series (if the estimated GARCH model is correctly specified) by performing standardised residual diagnostic tests on the residuals standardised by the estimated GARCH conditional volatility ($\hat{z}_t = \frac{\varepsilon_t}{\sqrt{\hat{\sigma^2}_t}}$)



	parameters								
Model	c	ω	α	β	γ	d	AIC	ARCH-LM test	JB test
ARCH(1)	7.94E-05 (0.5838)	0.0002 (0.0000)	0.2592 (0.0000)				−5.388	71.985 (0.0000)	29533.63 (0.0000)
GARCH(1,1)	0.0004 (0.0019)	2.41E-6 (0.0000)	0.0603 (0.0000)	0.9339 (0.0000)			−5.543	3.948 (0.4130)	15747.28 (0.0000)
Risk Metrics (RM)	0.0004 (0.0109)			0.9615 (0.0000)			−5.529	10.395 (0.0349)	22530.63 (0.0000)
EGARCH(1,1)	0.0001 (0.3297)	−0.1474 (0.0000)	−0.0475 (0.0000)	0.9920 (0.0000)	0.1099 (0.0000)		−5.565	5.998 (0.1117)	8276.66 (0.0000)
TGARCH(1,1)	0.0001 (0.2409)	2.57E-6 (0.0000)	0.0274 (0.0000)	0.9343 (0.0000)	0.0653 (0.0000)		−5.557	2.6153 (0.624)	8865.977 (0.0000)
FIGARCH(0,d,1)	0.0003 (0.0081)	1.87E-6 (0.000)		0.2898 (0.0000)		0.370 (0.000)	-5.502	3.248 (0.4251)	18833.38 (0.0000)

Note: The round parentheses give the p-values; The JB test is done on the standardized residuals.

Figure 4: Estimation of different GARCH Models

Arch(1) is capturing overkurtosis, since it is able to generate outliers (α is sig diff from 0). But intercept is not sig different from 0.

Arch-LM test and JB test are tested on . . . , both tests are redirected, there is remaining heteroskedacity, α relatively mild.

GARCH - passing arch lm test, decay in ACF is very slow, α, β close to 1, very persistent, but able to measure conditional heteroskedacity

RM - re estimated on data, p val for ARCH lm is 0.04, depends on confidence interval determines rejection. But none are looking like norm RV

E(T) GARCH - neagtive shocks (response to future volaltity) γ positive. Egarch model log variances,

EGARCH - α - if shock negative then log of variance should be multiplied with negative variance (asymmetric response, how much is shock differnt from abs value of expected shock)

α and γ? Negative and positive for egarch - at 5% sig level, all garchs seem to model sufficiently long memory using model parameters, out of these (ignoring fact dont past JB test of normality)

When we talked about ARMA we talked about AIC, BIC allowing us to compare different models estimated using ML, but different models have different parameters, so to control for this have different penalty functions (k denotes parameters).

Even asymmetric GARCH are unable to account for negative (β), we see in the data. Thus we require advanced financial econometrics

Garch loved since it is easy to forecast risk with them, central banks require risk forecasting on daily basis - using GARCH(1,1) is very easy for this.

Exercise 1. TGARCH(1,1) Estimated $\widehat{\sigma_t^2} = \hat{w} + \hat{\alpha}\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \dots$

$$E[\sigma_{t-1}^2|\mathcal{F}_t] = \hat{w} + \alpha r_t^2 + \hat{\beta}\widehat{\sigma_t^2} + \dots$$
$$E[\sigma_{t-1}^2|\mathcal{F}_t] = w + aE[\varepsilon_t^2|\mathcal{F}_t] + \beta E[\sigma_{t+1}^2|\mathcal{F}_t] + \dots$$

Expected value

$$W + \alpha E[\sigma_{t+1}^2|\mathcal{F}_t] + \beta E[\sigma_t^2|\mathcal{F}_t] + \gamma E[\pi(z_t)]$$

Forecasting with Risk Metrics

Let σ_t^2 follow a riskmetrics model:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)Y_{t-1}^2$$

Where $\lambda = 0.94$

3.8 Variance Forecast Evaluation

σ^2 is not observed, it may be replaced by proxies such as

- $\sigma_{t+h}^2 = r_{t+h}^2$ (squared daily returns)
- $\sigma_{t+h}^2 = RV_{t+h}$ daily realised variance

Or alternatively, we evaluate the variance forecasts within economic applications :

- Value at risk, expected shortcuts,
- Asset pricing etc

Good forecasting performance does not translate to good in sample fit (tradeoff?)

Tutorial 1. 5 Last week simulated GARCH, this week estimating GARCH and forecasting based on the estimates. In PS4 we have simulated $y_t = c + \psi y_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t$ Chose some parameters, then simulated based upon those parameters, and once we had innovations, we simulated for values of ARMA parameters, simulated the ARMA recursions This week have the daily log returns of SNP500, estimate ARMA and GARCH parameters which are coming from data, why cant we just plug these parameters in and use them in the simulation? Taking our $\widetilde{\sigma^2}$ and simulate returns, why cant we do this and why instead do we forecast where $\widehat{\sigma_t^2} = \hat{w} + \gamma r_{t-1}^2 + \hat{\beta}\sigma_{t-1}^2$ Simulated series which resemble data properties is defined as $E[\sigma_{t+1}^2|\mathcal{F}_t]$

Lecture 6: Kalman Filter

Mon 04 Mar 09:04

[Lecture]

$$E[\sigma_{t+1}^2|\mathcal{F}_t] = w + \alpha\varepsilon_t^2 + \beta\sigma_t^2$$

If we think about the classical ARMA-GARCH framework, we have

1. returns with some conditional mean $+\varepsilon_t$ where

$$\begin{aligned} r_t &= E[r_t|\mathcal{F}_{t-1}] + \varepsilon_t \\ \varepsilon_t &= \mathcal{L}_t \cdot \sigma_t(\mathcal{F}_{t-1}) \\ \mathcal{L}_t &\sim \mathcal{N}(0,1) \end{aligned}$$

If we would like to assume that returns is driven by

$$\begin{aligned} r_t &= \varepsilon_t = \mathcal{L}_t\sigma_t \\ \sigma_t^2 &= f(\mathcal{F}_{t-1}) + q_t \\ \varepsilon_t|\mathcal{F}_{t-1} &\sim \mathcal{N}(0,1) \\ f(\varepsilon_t) &= \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma_t^2}\right) \end{aligned}$$

Multivariate normal distribution

$$\begin{bmatrix} r \\ y \end{bmatrix} \approx N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{pmatrix}\right)$$

We need the $x|y$ distribution to derive the Kalman filter, the transformation can get us certain properties

Expected value (these are population parameters, fixed values, numbers measuring fixed variance)

$$\begin{aligned} E[z] &= E[x] - \Sigma_{xy}\Sigma_{yy}^{-1}(E[y] - \mu_y) \\ &= E[Z'Z] \quad \text{if scalar, then} \quad E[z^2] \\ &= E\left[\left(x - \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\left(x - \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)\right)\right] \\ &\quad \text{algebra} \ldots \\ \text{cov}(x,y) &= E[xy] - E[x]E[y] \\ &\to u_x u_y - E[xy'] = -\text{cov}(x,y) = -\Sigma_{xy} \\ &\quad \text{and so the whole term} \\ &= E[X'X] - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}' \\ \text{cov}(z,z) &= E[ZZ'] - E[Z]E[Z]' \\ &= E[xx'] \end{aligned}$$

We have 2 RV with joint distribution, we want to understand the conditional distribution

$$X|y \approx N\left(\ldots\right)$$

Then if we take the transformation z

$$z = x - \Sigma_{xy}$$

1. $E[z] = E[x]$
2. $V[z] = \Sigma_{xx} - \Sigma_{yy}^{-1}\Sigma_{xy}$

3. $\text{Cov}(y, z) = 0$

$x = z + \Sigma_{xy} + \Sigma_{yy}^{-1}(y - \mu y)$
$E[x|y] = E[z|y] + \Sigma_{xy}\Sigma_{yy}^{-1}(E[y|y] - \mu_y)$
First term $= E[z]$, second term is same third is y so

$$E[x|y] = \mu_x = \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$

$V[x|y]$

$$x|y \approx N(\mu_x)$$

Local Trend Model

In order to understand the SV model, consider a simple local trend model first

$$y_t = \mu_t + e_t, \qquad e_t \sim N\left(0, \sigma_e^2\right), \tag{8}$$

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim N\left(0, \sigma_\eta^2\right), \tag{9}$$

Where $\{e_t\}$ and $\{\eta_t\}$ are 2 independent Gaussian white noise series and $t = 1, \ldots, T$

The initial value μ_1 is either *given or follows a known distribution* and is independent of $\{e_t\}$ and $\{\eta_t\}$ for t>0.

Here μ_t is is a pure *random walk* with initial value μ_1 and y_t is an observed version of μ_t with added noise e_t μ_t is referred to as the *trend* of the series which is not directly observable, and y_t is the observed data with observational noise e_t.

The models above can be used to analyse the *realised* volatility of an asset price, where μ_t represents the underlying log volatility of the asset price and y_t is the log of realised volatility.

The model is a special linear *gaussian state space model*, with the variable μ_t called the *state* of the system at time t (not directly observed).

The y model provides the link between the data y_t and the state μ_t and is called the *observation equation* with measurement error e_t

The next μ_{t+1} governs the time evolution of the state variable and is the state equation with innovate η_t.

If $\sigma_e = 0$ then $y_t = \mu_t$ and there is no measurement error, which is an ARMA (0, 1, 0) model

If $\sigma_e > 0$ then there exist measurement error and y_t is an ARMA(0, 1, 1) model satisfying

$$(1 - B)y_t = (1 - \theta B)a_t \tag{10}$$

where $\{a_t\}$ is a gaussian white noise with mean zero and variance σ_a^2

Then, the values of θ and σ_{eta} are determined by σ_e and σ_η

From the initial model we have

$$\begin{aligned} (1 - B)\mu_{t+1} = \eta_t \quad \text{or} \quad \mu_{t+1} = \frac{1}{1 - B}\eta_t \\ \text{then we can rewrite} \quad y_t = \mu_t + e_t = y_t = \frac{1}{1 - B}\eta_{t-1} + e_t \end{aligned} \tag{11}$$

And multiplying by B we have

$$(1 - B)y_t = \eta_{t-1} + e_t - e_{t-1}$$

Then letting $(1 - B)y_t = w_t$ we have $w_t = \eta_{t-1} + e_t - e_{t-1}$ And under the model assumptions it is easy to see that w_t is gaussian, $\text{Var}(w_t) = 2\sigma_e^2 + \sigma_\eta^2$ and $\text{Cov}(w_t, w_{t-1}) = -\sigma_e^2$ and $\text{Cov}(w_t, w_{t-j}) = 0$ for j>1

Then consequently w_t follows an MA(1) model and can be written as $w_t = (1 - \theta B)a_t$

And by equating the variance and lag-1 autocovariance of $w_t = (1 - \theta B)a_t = \eta_{t-1} + e_t - e_{t-1}$ and we have

$$(1 + \theta^2)\sigma_a^2 = 2\sigma_e^2 + \sigma_\eta^2$$

$$\theta\sigma_a^2 = \sigma_e^2$$

Then for a given σ_e^2 and σ_η^2 we consider the ratio of these to form a quadratic function of θ, having 2 solutions which we select the one that satisfies $|\theta|<1$.

Idea is that if you have state equation with large variance, you wont be able to recover much.

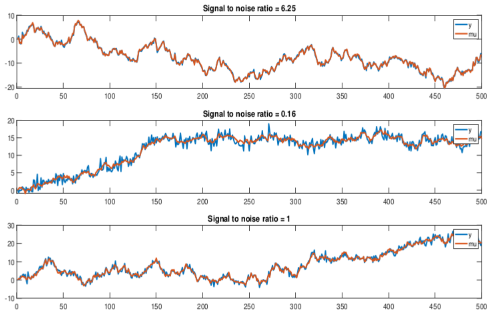


Figure 5: State Space Model

If signal to noise ratio = 0.16, observe blue try to recover red. It is not very informative, if ratio is 6 then signal is very informative

Kalman Filter

The aim of the analysis is to infer properties of the state μ_t alone from the data and the model. Let $F_t = \{y_1, \ldots, y_t\}$ be the information available at time t (inclusive) and assume that the model is known, including all parameters.

Three estimates of interest

1. Filtering : recover μ_t (remove measurement error)
2. Smoothing : estimate μ_t given all available information up to time T
3. Prediction : forecast μ_{t+k}

Analogy - filtering is figuring out the word you are reading based on knowledge accumulated from the beginning of the note, predicting is to guess the next word and smoothing is to decipher a particular word once you have read through the note.

Properties of Multivariate Normal Distribution Considering a multivariate normal distribution

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{pmatrix}\right)$$

Kalman filter is tool which characterises the conditional distribution of μ_t given the data. Given the date we observe what is the distribution of μ_t

Notation

Let $\mu_{t|j} = E[\mu_t|F_j]$ and $\Sigma_{t|j} = \text{Var}(\mu_t|F_j)$ be the conditional mean and variance of μ_t given F_j. $y_{t|j}$ is the conditional mean of y_t given F_j

And $v_t = y_t - y_{t|t-1}$ and $V_t = \text{Var}(v_t|F_{t-1})$ be the 1 step ahead forecast error and its variance of y_t given F_{t-1}

The forecast error v_t is independent of F_{t-1} so that the conditional variance is the same as the unconditional variance, that is $\text{Var}(v_t|F_{t-1}) = \text{Var}(v_t)$

Then

$$Y_{t|t-1} = E[y_t|F_{t-1}] = E[\mu_t + e_t|F_{t-1}] = E[\mu_t|F_{t-1}] = \mu_{t|t-1}$$

And consequently,

$$v_t = y_t - y_{t|t-1} = y_t - \mu_{t|t-1}$$

and

$$V_t = \text{Var}(y_t - \mu_{t|t-1}|F_{t-1}) = \text{Var}(\mu_t + e_t - \mu_{t|t-1}|F_{t-1})$$

$$= \text{Var}((\mu_t - \mu_{t|t-1}|F_{t-1})) + \text{Var}(e_t|F_{t-1}) = \Sigma_{t|t-1} + \sigma_e^2$$

And then it is easy to see that

$$E[v_t] = E[E[y_t - y_{t|t-1}]|F_{t-1}] = E[y_{t|t-1} - y_{t|t-1}] = 0$$

$$\text{Cov}(v_t, y_j) = E[v_t, y_j] = E[E[v_t y_j|F_{t-1}]] = E[y_j|E[v_t|F_{t-1}]] = 0, \qquad j < t$$

Then as expected the 1 step ahead forecast error is uncorrelated with $y_j$for j<t. And furthermore for the linear model in eq. (8) and eq. (9) $\mu_{t|t} = E[\mu_t|F_t] = E[\mu_t|F_{t-1}, v_t]$ and $\Sigma_{tZt} = \text{Var}(\mu_t|F_t) = \text{Var}(\mu_t|F_{t-1}, v_t)$

That is, the information set f_t can be written as $F_t = \{F_{t-1}, y_t\}$

Theorem 1 :

Properties of MV normal distribution useful to the Kalman filter under normality Suppose that x, y and z are 3 RV such that their joint distribution is MV normal, additionally assume that the diagonal block covariance Σ_{ww} is non singular for $w = x, y, z$ and $\Sigma_{yx} = 0$, then

1. $E[x|y] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$
2. $\text{Var}(x|y) = \Sigma_{xx} - \Sigma_{xx}\Sigma_{yy}^{-1}\Sigma_{yx}$
3. $E[x|y, z] = E[x|y] + \Sigma_{xz}\Sigma_{zz}^{-1}(z - \mu_z)$
4. $\text{Var}(x|y, z) = \text{Var}(x|y) - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_i$

Then the conditional distribution of x given y is

$$x|y \sim\sim \mathcal{N}(\mu_x + \Sigma_x y \Sigma_y y^{-1}(y - \mu_y), \, \Sigma-)$$

$$\begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} |\mathcal{F}_{t-1}$$

Goal is the conditional distribution $\mu_t|F_t$ based on new data y_t and the conditional distribution $\mu_t|F_{t-1}$

$$\left(\begin{array}{c} \mu_t \\ \nu_t \end{array} \right) \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_{t|t-1} \\ 0 \end{array} \right), \left(\begin{array}{cc} \Sigma_{t|t-1} & \Sigma_{t|t-1} \\ \Sigma_{t|t-1} & V_t \end{array} \right) \right)$$

Prediction

Initialise idea on conditional mean and variance of unobserved latent state variable (signals), take data minus initial value/expectation.

Look at forecast error variance

How does it work?

First remove measurement error then estimate μ_t given all available information, then forecast. Distribution of μ_t given information set $\mathcal{F}_t$ today. In order to today recover the value of μ_t need to update conditional expectation so that take into account signal to noise ratio. How much new noise contributes to the conditional variance expectation.

Filter latent process based on information $t-1$ then we update forecast once new information has arrived. Kalman gain measures how much information does the new shock at time t add to uncertainty (?). Dent take information as given y_t has noise itself e so we only update conditional expectations proportionally to the signal to noise ratio.

Recover, then smoothing re estimating μ_t (trying to mitigate effect of starting values), then based on this forecast latent process. All based on one property of MVR norm.

After we know this we can write it down given this formula

Major idea of KF is to write down some expectations of latent process, then update these according to Kalman gain which measures model uncertainty plus new variance originating from noisy data. Which are inherently small in financial data. New data is not very informative (nothing in autocorrelation structure), so strongly depends on starting values. These values are not eaten up by new data as they may in physics.

[**11**]

Kalman Filter

The goal of the Kalman filter is to update knowledge of the state variable recursively when a new data point becomes available. That is, knowing the conditional distribution of μ_t given F_{t-1} and the new data y_t, we would like to obtain the conditional distribution of μ_t given F_t where as before $F_j = \{y_1 \ldots, y_j\}$ since $F_t = \{F_{t-1}, v_t\}$ giving y_t and F_{t-1} is equivalent to giving v_t and F_{t-1}.

To derive the KF, it suffices to consider the joint conditional distribution of $(\mu_t, v_t)'$ given F_{t-1} before applying the above theorem

From the definition [**11.1.2**]

4 Tutorial 6

[**PS6**]

Is the signal to noise ratio the same no matter the number of simulations?

SNR is defined as $\frac{\sigma_e^2}{\sigma_\eta^2}$, the variances determine this, here they do not depend on anything, it is always defined by the variance of the error term in state equation and in the observation model.

B) mu and filter s

Our forecast error is defined as in [**slide 11**]

To understand the code, write the recursions using this slide

```
for (t in 1:T){
 predict_mu[t] = filter_mu[t]
 predict_S[t]= filter_S[t]
 v[t] = y[t]-predict_mu[t]
 V[t] = predict_S[t]+s_e^2
 K[t] = predict_S[t]/V[t]
```

```
  filter_mu[t+1] = predict_mu[t] + K[t]*v[t]
  filter_S[t+1] = predict_S[t]*(1-K[t])+s_eta^2
  print(V[t])
}
```

Old conditional expectation $\mu_{t|t+1} = \mu_{t|t+1} + K_t \cdot \nu_t$

In the for loop we take our conditional predictions and feed them 1 step ahead in the next iteration, so we have filtered out in iteration t becomes a prediction in $t+1$ so $\nu_{t+1} = y_{t+1} i \nu_{t|t+1}$

Kalman filter updating - the filtered at t becomes a prediction for t+1

Also possible to start at 2 and do filtering at t-1

What are the filter initialisations?

Can also draw from MV norm, this is local linear trend model with strong SNR, starting values matter less here, though this is the usual way to initialise.

Filter gets updated proportional to Kalman gain.

$$\begin{aligned}
\nu = y_1 - u_{1|0} = y_1 - \ predictnu[1] \\
V_1 = predictS[1] + \sigma_e^2 \\
K_1 = \ldots \\
filter[1] = predict.mu[1] + V[1] \cdot K[1] \\
\text{then for t = 2}
\end{aligned}$$

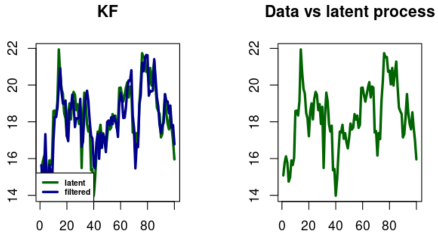


Figure 6

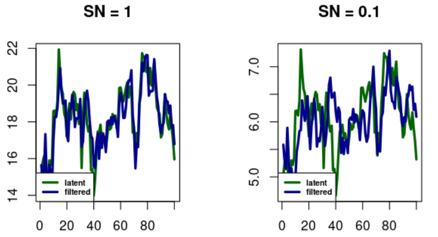


Figure 7

CI

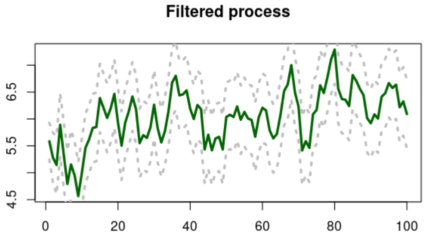


Figure 8

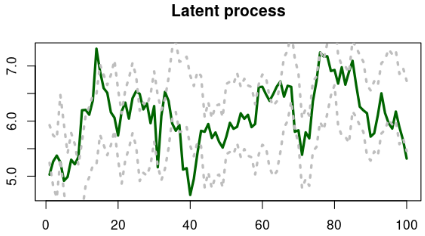


Figure 9

Fundamental property / assumption on μ_t, we assumed distribution given is **normal** and thus once we obtain filtration we can use this to write down CI. Conditional variance is given by filtered Σ.

Hypothetical Exam Question Why would CI be much wider where $\sigma_e^2 = 0.9$ and $\sigma_q^2 = 0.3$ than $\sigma_e^2 = 0.9$ and $\sigma_q^2 = 0.9$

The second has a SNR of 1, so higher ratio means the closer the observation and state equations are. Thinking about how CI are calculated, $\pm 2 \cdot \Sigma_{t|t}$ or $\Sigma_{t|t} = \Sigma_{t|t} + \sigma_q^2$

[**recursive slide**] - what about Kalman gain, need to think about how σ_η^2 influences variance, what does increase in σ_e^2, this is in denominator of Kalman gain :

This is all a recursive process

Why do we opt to calculate negative log likelihood?

```
 kf_loglik = function(y,s_e,s_eta){
fit = kf_recursions(y,s_e,s_eta)
# compute the negative log likelihood
l = 0.5*log(2*pi)+0.5*(log(fit$V))+ 0.5*((fit$v^2)/fit$V)
ll =sum(l)
return(ll)
}
```

Estimates are RV, they can lie anywhere so have to do negative.