



FINANCIAL ECONOMETRICS AND EMPIRICAL FINANCE - MODULE 2

General Exam – October 2018
Time Allowed: 2 hours and 20 minutes

Please answer all the questions by writing your answers in the spaces provided. There are two optional questions (7 and 8). No additional papers will be collected and therefore they will not be marked. You always need to carefully justify your answers and show your work. The exam is closed book, closed notes. No calculators are useful or permitted. You can withdraw until 10 minutes before the due time.

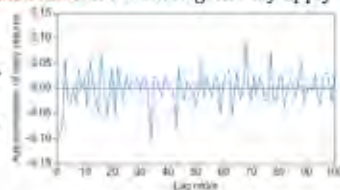
Question 1.A (10 points)

Describe in detail, also with reference to the examples that have been provided in the lectures, the six stylized facts that are typically displayed by asset returns. Discuss how you would proceed to test whether the unconditional distribution of asset returns is Gaussian. Make sure to also mention the main reasons why the returns may deviate from a Gaussian distribution.

Debriefing

Stylized facts on asset returns

- At daily or weekly frequencies, asset returns display **weak serial correlations** (in absolute value)
 - Returns **are not normal and display asymmetries and fat tails**
 - Asset returns display a few **stylized facts** that tend to generally apply and that are well-known
 - Refer to daily returns on the S&P 500 from January 1, 2001, through December 31, 2010
 - But these properties are much more general, see below
- ① Daily returns show weak autocorrelation:
- $$\text{Corr}(R_{t+1}, R_{t+1-\tau}) \approx 0, \quad \text{for } \tau = 1, 2, 3, \dots, 100$$
- Returns are almost impossible to predict from their own past
- ② The unconditional distribution of daily returns does not follow the normal distribution



Lecture 1: The Econometrics of Financial Returns – Prof. Guidolin

Stylized facts on asset returns

-
- The histogram is more peaked around zero than a normal distribution
 - Daily returns tend to have more small positive and fewer small negative returns than the normal distribution (fat tails)
 - The stock market exhibits occasional, very large drops but not equally large upmoves
 - Consequently, the distribution is **asymmetric** or **negatively skewed**

Lecture 1: The Econometrics of Financial Returns – Prof. Guidolin

Stylized facts on asset returns

- At high frequencies, the standard deviation of asset returns completely dominates the mean which is often not significant
- Squared and absolute returns have **strong serial correlations** and there is a leverage effect
- Correlations between asset returns are time-varying
- Std. dev. completely dominates the mean at short horizons
 - S&P 500: daily mean of 0.0056% and daily std. dev. of 1.3771%
- Variance, measured, for example, by squared returns, displays positive correlation with its own past
- Equity and equity indices display negative correlation between variance and returns, the leverage effect
- Correlation between assets appears to be time varying



Lecture 1: The Econometrics of Financial Returns - Prof. Guidolin

Model Specification Tests: Jarque-Bera Test

- Because the normal distribution is symmetric, the third central moment, denoted by μ_3 , should be zero; and the fourth central moment, μ_4 , should satisfy $\mu_4 = 3\sigma^4$
- A typical index of asymmetry based on the third moment (**skewness**), that we denote by \hat{S} , of the distribution of the residuals is

$$\hat{S} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{\epsilon}_t^3}{\hat{\sigma}_t^3}$$

- The most commonly employed index of tail thickness based on the fourth moment (**excess kurtosis**), denoted by \hat{K} , is

$$\hat{K} = \frac{1}{T} \sum_{t=1}^T \frac{\hat{\epsilon}_t^4}{3\hat{\sigma}_t^4} - 3$$

- If the residuals were normal, \hat{S} and \hat{K} would have a zero-mean asymptotic distribution, with variances $6/T$ and $24/T$, respectively
- The Jarque-Bera test concerns the composite null hypothesis:

$$H_0: \frac{\mu_3}{\sigma^3} = 0 \quad \text{and} \quad H_0: \frac{\mu_4}{\sigma^4} - 3 = 0$$

Lecture 2: Autoregressive Moving Average (ARMA) Models - Prof. Guidolin

24

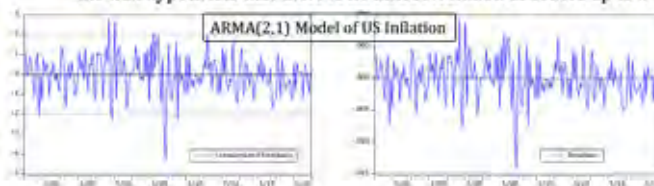
Model Specification Tests: Jarque-Bera Test

- Jarque and Bera prove that because the sample statistics

$$\lambda_1 = \frac{1}{6T} \sum_{t=1}^T \left(\frac{\hat{\epsilon}_t^3}{\hat{\sigma}_t^3} \right) \quad \lambda_2 = \frac{1}{24T} \sum_{t=1}^T \left(\frac{\hat{\epsilon}_t^4}{\hat{\sigma}_t^4} - 3 \right)$$

are $N(0,1)$ distributed, the null consists of a joint test that λ_1 and λ_2 are zero tested as $H_0: \lambda_1 + \lambda_2 = 0$, where $\lambda_1^2 + \lambda_2^2 \sim \chi_2^2$ as $T \rightarrow \infty$

- Compute **sample autocorrelations of residuals** and perform tests of hypotheses to assess whether there is any linear dependence
 - Same portmanteau tests based on the Q-statistic can be applied to test the null hypothesis that there is no autocorrelation at orders up to h



Question 1.B (4 points)

Ms. Granger, a junior analyst at Badcredit Bank, is trying to persuade her boss, Nic Dwarf, that $E[R_{t+1}] = 11\%$ is not incompatible with $E_t[R_{t+1}] = -11\%$. Do you agree with her claim? Support your answer by referring to the concept of conditional vs. unconditional distribution and moments.

Debriefing

Unconditional vs. Conditional objects

- Unconditional** moments and densities represent the long-run, average properties of times series of interest
- Conditional** moments and densities capture how our perceptions of RV dynamics changes over time as news arrive

- Our task will consist of building and estimating models for both the conditional variance and the conditional mean
 - E.g., $\mu_{t+1} = \theta_0 + \phi_1 R_t$ and $\sigma_{t+1}^2 = \lambda \sigma_t^2 + (1-\lambda) R_t^2$
- However, robust conditional mean relationships are not easy to find, and assuming a zero mean return may be a prudent choice

- One important notion in this course distinguishes between unconditional vs. conditional moments and/or densities
- An unconditional moment or density represents the long-run, average, "stable" properties of one or more random variables
 - Example 1:** $E[R_{t+1}] = 11\%$ means that on average, over all data, one expects that an asset gives a return of 11%

Lecture 1: The Econometrics of Financial Returns - Prof. Guidolin

13

Unconditional vs. Conditional objects

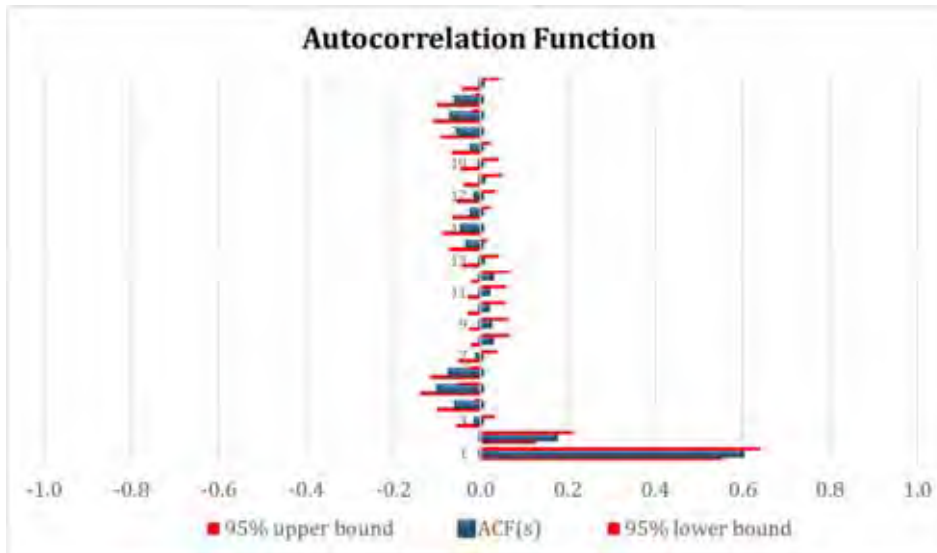
- Example 2:** $E[R_{t+1}] = 11\%$ is not inconsistent with $E_t[R_{t+1}] = -6\%$ if news are bad today, e.g., after a bank has defaulted on its obligations
- Example 3:** One good reason for the conditional mean to move over time is that $E_t[R_{t+1}] = \alpha + \beta R_t + \epsilon_{t+1}$, which is a predictive regression
 - Recall Homework 2 in Theory of Finance? Ok, that was a conditional mean model written in predictive form
- Example 4:** This applies also to variances, i.e., there is a difference between $\text{Var}[R_{t+1}] = \sigma^2$ and $\text{Var}_t[R_{t+1}] = \sigma_t^2$
- Example 5:** Therefore the **unconditional density** of a time series represents long-run average frequencies in one observed sample
- Example 6:** The **conditional density** describes the expected frequencies (probabilities) of the data based on currently available info
- When a series (or a vector of series) is **identically and independently (i.i.d. or IID) distributed over time**, then the conditional objects collapse into being unconditional ones
- Otherwise unconditional ones mix over conditional ones...

Lecture 1: The Econometrics of Financial Returns - Prof. Guidolin

14

Question 1.C (3 points)

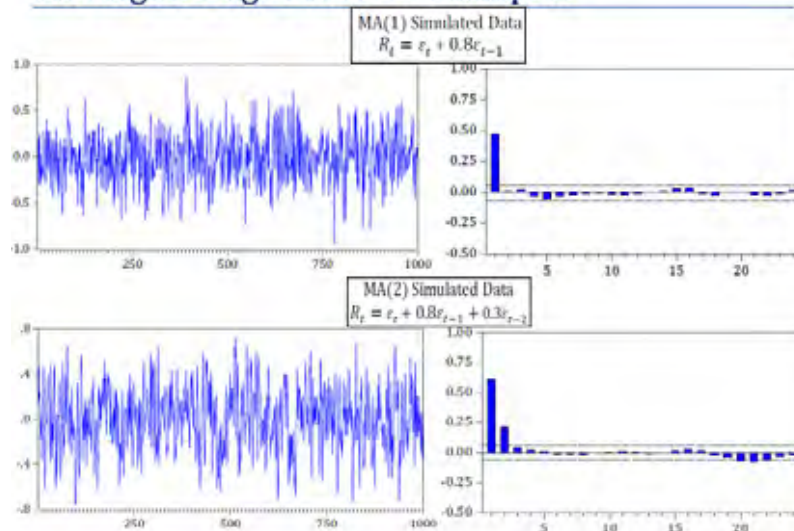
Mr. Dwarf has now assigned to Ms. Granger the task to analyze the features of the returns of an emerging market stock index. In particular, he is convinced that the returns follow the process $R_{t+1} = \mu + \sigma_{t+1}z_{t+1}$, with $z_{t+1} \sim IID D(0,1)$. As a first step, Ms. Granger has decided to estimate the correlogram of the series, which is reported below. On the basis of this evidence, do you believe that Mr. Dwarf's intuition about the process followed by the data is correct? Make sure to clearly justify your answer.



Debriefing.

The model assumed by Mr. Dwarf implies zero autocorrelation between the returns. However, this does not seem to be the case when we look at the correlogram. Indeed, the correlogram is compatible with an ARMA model, but not with a white noise process.

Moving Average Process : Examples



Question 2.A (10 points)

Consider the following VMA(∞) representation of a VAR(1) model

$$y_t = \mu + \sum_{i=1}^{\infty} \Theta_i u_{t-i} + u_t,$$

where $\Theta_i = A_1^i$ and A_1 is the matrix of the coefficients of the *reduced form* VAR(1). Can we interpret the coefficients Θ_i as impact multipliers of the true, *structural* innovations? If not, carefully explain why and discuss whether and under what conditions it is possible to retrieve the impact multipliers to structural innovations from the OLS estimates of a VAR in its reduced form. Finally, discuss which kind of information is entailed in a variance decomposition of forecast errors and specify whether some identification scheme must be imposed in order to retrieve such information.

Debriefing.

Identifying Structural from Reduced-Form VARs

- In a sense, **shocks to z , are more primitive, enjoy a higher rank, and move the system also through a contemporaneous impact on x** .
- The VAR(1) now acquires a triangular structure:

$$\begin{aligned}
 x_t &= \gamma_{10} - b_{12}z_t + \gamma_{11}x_{t-1} + \gamma_{12}z_{t-1} + \epsilon_t^x \\
 z_t &= \gamma_{20} + \gamma_{21}x_{t-1} + \gamma_{22}z_{t-1} + \epsilon_t^z
 \end{aligned}$$

$$y_t = \underbrace{B^{-1}\Gamma_0}_{a_0} + \underbrace{B^{-1}\Gamma_1}_{A_1}y_{t-1} + \underbrace{B^{-1}\epsilon_t}_{u_t} = a_0 + A_1y_{t-1} + u_t \quad \text{with } B \equiv \begin{bmatrix} 1 & b_{12} \\ 0 & 1 \end{bmatrix}$$
- This corresponds to imposing a Choleski decomposition on the covariance matrix of the residuals of the VAR in its reduced form
- Indeed, now we can re-write the relationship between the pure shocks (from the structural VAR) and the regression residuals as

$$\begin{bmatrix} u_t^x \\ u_t^z \end{bmatrix} = u_t \equiv B^{-1}\epsilon_t = \begin{bmatrix} 1 & b_{12} \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_t^x \\ \epsilon_t^z \end{bmatrix} = \begin{bmatrix} \epsilon_t^x - b_{12}\epsilon_t^z \\ \epsilon_t^z \end{bmatrix}$$

Lecture 4: Multivariate Time Series Analysis - Prof. Giordano 45

Identifying Structural from Reduced-Form VARs

- The reason is clear if we compare the number of parameters of the primitive system with the number recovered from the estimated VAR model

$$\begin{aligned}
 x_t &= \gamma_{10} - b_{12}z_t + \gamma_{11}x_{t-1} + \gamma_{12}z_{t-1} + \epsilon_t^x \\
 z_t &= \gamma_{20} - b_{21}x_t + \gamma_{21}x_{t-1} + \gamma_{22}z_{t-1} + \epsilon_t^z
 \end{aligned}$$
 8 mean parameters + 2
- $y_t = \underbrace{B^{-1}\Gamma_0}_{a_0} + \underbrace{B^{-1}\Gamma_1}_{A_1}y_{t-1} + \underbrace{B^{-1}\epsilon_t}_{u_t} = a_0 + A_1y_{t-1} + u_t$ 6 mean parameters + 3
 - 9 vs. 10: unless one is willing to restrict one of the parameters, it is not possible to identify the primitive system and the **structural VAR is under-identified**
- One way to identify the model is to use the type of recursive system proposed by Sims (1980): we speak of **triangularizations**
- In our example, it consists of imposing a restriction on the primitive system such as, for example, $b_{21} = 0$
- As a result, while z has a contemporaneous impact on x , the opposite is not true

Lecture 4: Multivariate Time Series Analysis - Prof. Giordano 44

Impulse Response Functions

- The formula can be used recursively to compute h -step-ahead predictions starting with $h = 1$:

$$E_t[y_{t+h} | \mathcal{I}_t] = a_0 + A_1 E_t[y_{t+h-1} | \mathcal{I}_t] + \dots + A_{h-1} E_t[y_{t+1} | \mathcal{I}_t]$$
- In essence, the same results that apply to AR models generalize
- VAR models are used in practice with the goal of understanding the **dynamic relationships** between the variables of interest

Definition (Impulse Response Function) In the context of a VAR model, an impulse response functions trace out the time path of the effects of an exogenous shock to one (or more) of the endogenous variables on some or all of the other variables in a VAR system.

- Let's use again a simple VAR(1) model:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} = a_0 + A_1 y_{t-1} + u_t$$
- We know that a stationary VAR has a MA(∞) representation:

$$y_t = \mu + \sum_{i=0}^{\infty} A_i u_{t-i} = \mu + \sum_{i=0}^{\infty} \Theta_i u_{t-i} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \theta_{1,0i} & \theta_{1,20i} \\ \theta_{2,0i} & \theta_{2,20i} \end{bmatrix} \begin{bmatrix} u_{1,t-i} \\ u_{2,t-i} \end{bmatrix} \approx \dots$$

Lecture 4: Multivariate Time Series Analysis - Prof. Giordano

Impulse Response Functions

- The two error processes, $\{u_{1,t}\}$ and $\{u_{2,t}\}$ can be represented in terms of the two sequences $\{\epsilon_{1,t}\}$ and $\{\epsilon_{2,t}\}$, i.e., the structural innovations:

$$\begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} = \frac{1}{1 - \lambda_1 \lambda_{21}} \begin{bmatrix} 1 & -\lambda_{12} \\ -\lambda_{21} & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$
- Therefore the model can be re-written as

$$y_t = \mu + \sum_{i=0}^{\infty} \Phi_i \epsilon_{t-i} \quad \Phi_i = \frac{A_i}{1 - \lambda_1 \lambda_{21}} \begin{bmatrix} 1 & -\lambda_{12} \\ -\lambda_{21} & 1 \end{bmatrix} + \frac{\Theta_i}{1 - \lambda_1 \lambda_{21}} \begin{bmatrix} 1 & -\lambda_{12} \\ -\lambda_{21} & 1 \end{bmatrix}$$
- The coefficients in Φ_i (**impact multipliers**) can be used to generate the effects of shocks to the innovations $\{\epsilon_{1,t}\}$ and $\{\epsilon_{2,t}\}$ on the time path of $\{y_{1,t}\}$ and $\{y_{2,t}\}$
- The cumulative effects of a one-unit shock (or impulse) to a structural shock on an endogenous variable after H periods can then be obtained by computing the sum $\sum_{i=0}^{H-1} \Phi_{t+i}$
- A VAR in its reduced form is under-identified and therefore we cannot compute the coefficients in Φ_i from the OLS estimates of the VAR in its standard form unless we impose restrictions

Lecture 4: Multivariate Time Series Analysis - Prof. Giordano

Impulse Response Functions

- One method to place these restrictions consists of the application of a Choleski decomposition: $y_t = \mu + \sum_{i=0}^{h-1} \Phi_i W W^{-1} u_{t-i}$ ($\Sigma = W \Sigma W'$, $\varepsilon_t = W^{-1} u_{t-1}$, and $\Phi_t = \Theta_t W$)
 - Because of the triangular structure of $W = B^{-1}$, a Choleski decomposition allows only the shock to the first variable to contemporaneously affect all the other variables in the system
 - A shock to the second variable will produce a contemporaneous effect on all the variables in the system, but the first one
 - This may of course be impacted in the subsequent period, through the transmission effects mediated by the autoregressive coefficients
 - A shock to the third variable will affect all the variables in the system, but the first two, and so on
- Therefore, a Choleski identification scheme forces a potentially important identification asymmetry on the system
- A different ordering of the variables in the system would have been possible, implying a reverse ordering of the shocks

Lecture 4: Multivariate Time Series Analysis- Prof. Guidolin

3.1 Impulse Response Functions

- VAR models can be used to understand the dynamic relationships between the variables of interest.

Impulse Response Function: In the context of a VAR model, an impulse response function traces out the time path of the effects of an exogenous shock to one (or more) of the endogenous variables on some or all of the other variables in a VAR system.

Impact multipliers: coefficients of the matrix Φ_0 .

Starting from the moving average representation of a VAR(1)

$$y_t = \alpha_0 + A_1 y_{t-1} + u_t = \mu + \sum_{i=0}^{\infty} A_i^j u_{t-i} = \mu + \sum_{i=0}^{\infty} \Phi_i u_t =$$

[7]

Variance Decompositions

- Understanding the properties of forecast errors from VARs is helpful in order to assess the interrelationships among variables
- Using the VMA representation of the errors, the h-step-ahead forecast error is $u_t(h) = y_{t+h} - E_t[y_{t+h}] = \sum_{i=0}^{h-1} \Phi_i \varepsilon_{t+i}$
 - See lecture notes for algebra of such representation
 - Because all white noise shocks the same variance, if we denote by $\sigma_{\varepsilon_t}^2(h)$ the h-step-ahead variance of the forecast of (say) y_t , we have: $\sigma_{\varepsilon_t}^2(h) = \sigma_{\varepsilon_t}^2 [\phi_{11}^2(0) + \phi_{11}^2(1) + \dots + \phi_{11}^2(h-1)] + \sigma_{\varepsilon_t}^2 [\phi_{21}^2(0) + \phi_{21}^2(1) + \dots + \phi_{21}^2(h-1)]$
 - Because all the coefficients in Φ_i are non-negative, the variance of the forecast error increases as the forecast horizon h increases
 - We decompose the h-step-ahead forecast error variance into the proportion due to each of the (structural) shocks $\sigma_{\varepsilon_t}^2 [\phi_{11}^2(0) + \phi_{11}^2(1) + \dots + \phi_{11}^2(h-1)]$ and $\sigma_{\varepsilon_t}^2 [\phi_{21}^2(0) + \phi_{21}^2(1) + \dots + \phi_{21}^2(h-1)]$
- Such proportions due to each shock is a **variance decomposition**

Lecture 4: Multivariate Time Series Analysis- Prof. Guidolin

Variance Decompositions

- Like in IRF analysis, variance decompositions of reduced-form VARs require identification (because otherwise we would be unable to go from the coefficients in θ to their counterparts in Φ)
 - Choleski decompositions are typically imposed
 - Forecast error variance decomposition and IRF analyses both entail similar information from the time series
 - Example on weekly US Treasury yields, 1990-2016 sample:

		Variance Decomposition of 10Y Yield					
		Period	S.E.	1M Yield	1Y Yield	5Y Yield	10Y Yield
Choleski ordering:	1	0.101	0.260	37.242	51.653	10.845	
	2	0.158	0.306	36.977	52.897	9.820	
	3	0.202	0.285	37.053	53.316	9.345	
	4	0.237	0.253	37.265	53.389	9.093	
	5	0.268	0.221	37.527	53.311	8.941	
	6	0.296	0.194	37.805	53.163	8.839	
	7	0.321	0.170	38.086	52.981	8.763	
	8	0.344	0.150	38.363	52.794	8.703	
	9	0.365	0.134	38.633	52.580	8.653	
	10	0.385	0.120	38.896	52.375	8.608	
	11	0.404	0.109	39.151	52.171	8.568	
	12	0.422	0.100	39.399	51.971	8.530	

Lecture 4: Multivariate Time Series Analysis- Prof. Guidolin

or, for example in the case of a VAR(1)

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{1,1(i)} & \phi_{1,2(i)} \\ \phi_{2,1(i)} & \phi_{2,2(i)} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-i} \\ \varepsilon_{2,t-i} \end{bmatrix}$$

where

$$\begin{bmatrix} \phi_{1,1} \\ \phi_{2,1} \end{bmatrix} = \frac{1}{1 - b_{1,2} b_{2,1}} \begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} \phi_{1,1} \\ \phi_{2,1} \end{bmatrix}$$

Then

$$\Phi_t = \frac{A_1^t}{1 - b_{1,2} b_{2,1}} \begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix} = \frac{\Theta_t}{1 - b_{1,2} b_{2,1}} \begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix}$$

and

$$y_t = \mu + \sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i}$$

For example, $\phi_{1,2(0)}$ is the instantaneous impact on $y_{1,t}$ of a one-unit change in $\varepsilon_{2,t}$.

Cumulative response of the variable j to a shock to the variable i :

$$\sum_{i=0}^H \phi_{j,i(i)}$$

For example, $\sum_{i=0}^H \phi_{1,2(i)}$ is the cumulative effects of a one-unit shock (or impulse) to $\varepsilon_{2,t}$ on the variable $y_{1,t}$ after H periods.

Long-run impact multipliers: impact multipliers when $H \rightarrow \infty$.

- The set of elements $\phi_{j,i(i)}$, with $i = 1, \dots, H$ is the impulse response function of the j th variable of the system, up to the period H .
- VAR in its reduced form is under-identified by construction and therefore $\phi_{j,i(i)}$ cannot be computed from the OLS estimates of the VAR in its standard form without imposing adequate restrictions.
- Choleski decompositions provide a minimal set of restrictions concerning the simultaneous relationships among variables that can be used to identify the structural model, but this method forces potentially important identification asymmetry on the system.
- IRFs are constructed using estimated coefficient, thus will contain sampling error. Therefore, it is advisable to construct confidence intervals around them to account for the uncertainty that derives from parameter estimation.

Question 2.B (3 points)

Max Earlgrey, a senior economist at BundBank Inc., is selecting the best VAR(p) model for a vector of time series that includes 1-month, 1-, 5-, and 10-year US Treasury bond rates. On the basis of a sample of 1,395 observations, he reports that a likelihood ratio test (LRT) of a VAR(1) vs. a VAR(2) gives a test statistic of 203.72. In addition, Max has determined that the LRT of a VAR(2) vs. a VAR(3) gives a test statistic of 38.56. Would Max be able to compute the (S)BIC for a VAR(2) using the information reported? In the affirmative case, please show how, otherwise clearly discuss why not and which additional information would he require.

Debriefing.

The formula for (S)BIC is the following

$$\ln|\tilde{\Sigma}_u(p)| + \frac{2}{T} \ln T(N^2p + N).$$

The number of observations (T) has been given to you in the text of the exercise and also N , the number of variables in the system is known (they are 4). However, you are not able to extract the value of $\ln|\tilde{\Sigma}_u(2)|$ from the information that you were given. Therefore, in order to be able to perform the computation, one of the following information is sufficient:

- $\ln|\tilde{\Sigma}_u(1)|$
- $\ln|\tilde{\Sigma}_u(2)|$ (obviously)
- $\ln|\tilde{\Sigma}_u(3)|$ (obviously)

Question 2.C (4 points)

A younger colleague of Dr. Earlgrey, Miss Granger, pointedly suggests that they shall conduct a full specification search, and produces the table below. Which is the model selected by each of the three information criteria? Do they all lead to the selection of the same model and, if not, is this plausible?

VAR Lag Order Selection Criteria
 Endogenous variables: ONEMONTH ONEYEAR FIVEYEARS TENYEARS
 Exogenous variables: C
 Date: 10/11/18 Time: 17:43
 Sample: 1/05/1990 12/30/2016
 Included observations: 1395

Lag	LogL	LR	AIC	SC	HQ
0	-4984.879	NA	7.152514	7.167541	7.158133
1	6162.080	22214.01	-8.805849	-8.730715	-8.777757
2	6264.603	203.7234	-8.929897	-8.794654	-8.879331
3	6284.063	38.55692	-8.934857	-8.739507	-8.861818
4	6299.304	30.11002	-8.933769	-8.678310	-8.838256
5	6354.346	108.4274	-8.989743	-8.674177	-8.871756
6	6375.059	40.68298	-8.996500	-8.620826	-8.856039
7	6390.868	30.96155	-8.996226	-8.560445	-8.833292
8	6406.572	30.66431	-8.995802	-8.499912	-8.810394
9	6419.565	25.29749	-8.991491	-8.435494	-8.783610
10	6443.217	45.91304	-9.002461	-8.386356	-8.772106
11	6460.887	34.20005	-9.004855	-8.328643	-8.752027
12	6473.561	24.45873	-9.000088	-8.263767	-8.724785
13	6490.912	33.38322	-9.002024	-8.205596	-8.704248
14	6504.644	26.34175	-8.998773	-8.142236	-8.678523

Debriefing.

The exercise requires you to find the model that **minimizes** each of the three information criteria. Namely, the AIC selects a VAR(11) model while both the SC and the HQ criterion select a more parsimonious VAR(2) model. It is perfectly plausible that the three criteria lead to the selection of different models, see slides below.

Model Selection: Information Criteria

- The alternative is to use information criteria (often shortened to IC)
- They essentially **trade off the goodness of (in-sample) fit and the parsimony of the model** and provide a (cardinal, even if specific to an estimation sample) summary measure
 - We are interested in forecasting out-of-sample: using too many parameters we will end up fitting noise and not the dependence structure in the data, reducing the predictive power of the model (**overfitting**)
- Information criteria include in rather simple mathematical formulations two terms: one which is a function of the sum of squared residual (SSR), supplemented by a penalty for the loss of degrees of freedom from the number of parameters of the model
 - Adding a new variable (or a lag of a shock or of the series itself) will have two opposite effects on the information criteria: it will reduce the residual sum of squares but increase the value of the penalty term
- The **best performing (promising in out-of-sample terms) model will be the one that minimizes the information criteria**

Lecture 3: Autoregressive Moving Average (ARMA) Models - Prof. Giordano 11

Model Selection: Information Criteria

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}$$

$$HQIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \ln(\ln(T))$$

$$SBIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \ln(T)$$

$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^2$

- The SBIC is the one IC that imposes the strongest penalty ($\ln T$) for each additional parameter that is included in the model.
- The HQIC embodies a penalty that is somewhere in between the one typical of AIC and the SBIC

	Log-likelihood	AIC	SBIC	HQIC
AR(1)	877.209	-0.4622	-0.4313	-0.4616
AR(2)	923.9792	-0.3313	-0.474	-0.344
AR(3)	924.6248	-0.3263	-0.4564	-0.3699
AR(4)	924.6276	-0.5279	-0.4281	-0.4667
ARMA(1,1)	918.9171	-0.4929	-0.442	-0.4723
ARMA(2,1)	933.7956	-0.486	-0.4711	-0.5121
MA(1)	924.3973	-0.3354	-0.306	-0.303
MA(2)	924.6333	-0.5299	-0.4261	-0.4967
ARMA(1,1)	923.788	-0.5182	-0.4703	-0.4907
ARMA(2,1)	926.9096	-0.556	-0.4715	-0.522
ARMA(2,2)	926.9914	-0.552	-0.501	-0.5118
ARMA(3,2)	927.0902	-0.5417	-0.4226	-0.4938
ARMA(3,1)	927.0896	-0.5384	-0.4026	-0.4838

Information criteria for autoregressive models estimated by the CPY algorithm

Question 3.A (9 points)

What is a spurious regression? Carefully define its causes and potential consequences for the validity of results from standard econometric procedures. How would you go about detecting the spurious nature of a regression? Suppose that one regresses a I(2) time series on a I(1) time series, would that cause a spurious regression problem? What are the remedies to avoid spurious regressions?

Debriefing.

Pitfalls in De-Trending Applications

- Even when the trend-stationary component is absent, if the time series is $I(0)$ but it is incorrectly differenced d times, the resulting differentiated series will contain d unit roots in its MA components
- What if $y_t \sim I(d)$ but by mistake we differentiate it $d+r$ times?
 - ③a -- If $r > 0$, we are over-differencing the series, and as such ② applies, that is, the resulting over-differentiated series will contain r unit roots in its MA components and will therefore be not invertible
 - ③b -- If $r < 0$, we are not differencing the series enough and the resulting series will still contain $d-r$ and will remain nonstationary
- Why is it that we care so much for isolating and removing trends?
- It turns out that, at least **in general, using $I(d)$ series with $d > 0$ in standard regression analysis, in general exposes us to the peril of invalid inferences**
- We speak of **spurious regressions**
- Suppose that $y_t \sim I(1)$ and $x_t \sim I(1)$, e.g., stock prices and GDP

Lecture 5: Unit Roots, Cointegration and Error Correction Models - Prof. Giordano 12

2 The Spurious Regression Problem

Sum of Stationary and Non-Stationary Series: Consider N time series, $y_{1,t} \sim I(d_1), y_{2,t} \sim I(d_2), \dots, y_{N,t} \sim I(d_N)$. Then, unless special conditions occur, their weighted sum will be integrated with an order that is the maximum across all integration orders:

$$\sum_{i=1}^N w_i y_{i,t} \sim I(\max\{d_1, d_2, \dots, d_N\})$$

(Heuristic) proof in case of three series

Let $y_t \sim I(1)$, $x_t \sim I(1)$ and η_t be three series such that y_t and x_t are independent. The regression of y_t on x_t

$$y_t = a + bx_t + \eta_t$$

The Spurious Regression Problem

- You estimate a regression of y_t on x_t , $y_t = a + bx_t + \eta_t$, expecting the errors (say, η_t) to be white noise, as required by OLS, but instead:

$$\begin{aligned} \eta_t &= (y_0 + \mu_y t + \sum_{r=1}^t \varepsilon_r^y) - a - b(x_0 + \mu_x t + \sum_{r=1}^t \varepsilon_r^x) \\ &= (y_0 - a - bx_0) + (\mu_y - b\mu_x)t + \sum_{r=1}^t (\varepsilon_r^y - b\varepsilon_r^x) \\ &= (y_0 + \mu_y t + \sum_{r=1}^t \varepsilon_r^y) - a - b(x_0 + \mu_x t + \sum_{r=1}^t \varepsilon_r^x) = [(\mu_y - b\mu_x) + \varepsilon_t^y - b\varepsilon_t^x] + \\ &\quad + (y_0 - a - bx_0) + (\mu_y - b\mu_x)(t-1) + \sum_{r=1}^{t-1} (\varepsilon_r^y - b\varepsilon_r^x) \\ &= (\mu_y - b\mu_x) + \eta_{t-1} + (\varepsilon_t^y - b\varepsilon_t^x) \end{aligned}$$

- The very error terms of a regression are $I(1)$!
 - This occurs unless very special conditions occur, see below
- A spurious regression has the following features:
 - The residuals are $I(1)$ and as such any shock is a permanent change of the intercept of the regression, in no way news

Lecture 5: Unit Roots, Cointegration and Error Correction Models – Prof. Guidetti 13

The Spurious Regression Problem

- Standard OLS estimators are inconsistent and the associated inferential procedures are invalid and statistically meaningless
- The regression has a high R^2 and t-statistics that appear to be significant, but the results are void of any economic meaning
 - Do not fall in the spurious regression trap, do not just boast huge R-squares, in a finance they are more often symptoms of problems
 - This is not a small sample problem; in fact, these issues worsen as the sample size grows
 - These ideas generalize, at the cost of technical complexity when one would try and regress an $I(d)$ series on another $I(d)$ series
 - Or when we regress a deterministic trend on another trend
- The cure of the problem is to work with stationary first/ d -differenced series
 - E.g. we generate two independent sets of IID white noise variables and use them to simulate 1000 observations from two driftless RWs
 - The two RWs are expected to be unrelated

Lecture 5: Unit Roots, Cointegration and Error Correction Models – Prof. Guidetti 14

This follows from the result on limits of stationary and non-stationary random walks, in the case of a unit root $I(1)$ variable:

- y_t is the weighted sum (plus a constant) of the error $I(1)$ variables. Therefore $y_t \sim I(1)$, which is the highest integration order of the variables that are not constant.
- Starting from $y_t = a + bx_t + \eta_t$, while the regressand and regressor are both $I(1)$, it turns out that the regression errors must then also contain non-stationary random walks (even a unit root).
 - When x_t is a random walk with drift, the assumptions of the classical regression model (y_t and x_t are stationary, and the errors have a zero mean and a finite variance) are not respected.

Spurious regression – regression fallacy

- The residuals are $I(1)$ and as such any shock has a permanent effect on the residuals, being equivalent to a permanent change of the intercept of the model
- R -square is high and t-statistics appear to be significant, but the results are void of any economic meaning
- Standard OLS estimators are inconsistent and the associated inferential procedures are invalid and statistically meaningless

16

- Estimating spurious regressions and reporting and discussing their results is meaningless.
- Problems will also arise in regression analysis when the regressand and the regressors are integrated of different orders. Regression equations using such variables are meaningless.

Example: if in

$$\eta_t = (a_0 - a + bx_0) + (\mu_y - b\mu_x)t + \sum_{r=1}^t (\varepsilon_r^y - b\varepsilon_r^x)$$

only x_t is $I(1)$, the resulting regression errors would be $I(1)$

$$\eta_t^* = (\mu_y + \varepsilon_t^y) - a - bx_0 + \mu_x t + \sum_{r=1}^t \varepsilon_r^x = (\mu_y - a - bx_0) + \mu_x t + b \sum_{r=1}^t \varepsilon_r^x + \varepsilon_t^y$$

Question 3.B (3.5 points)

Deron Cleanington is quant analyst that monitors Italian rates. He knows for a fact that 3-month T-bills (BoT) are $I(1)$; however, he does not know much about 10-year note (BTp) rates. As a way to familiarize with the data, Deron regresses 10-year rates on 3-month rates, finding (p-values are in parenthesis):

$$r_t^{10Y} = 0.562 + 0.844r_t^{3m} + \hat{\eta}_t, \quad \begin{matrix} (0.000) & (0.000) \end{matrix}$$

Additional checks based on a Philipps-Perron unit root test reveal that the null hypothesis cannot be rejected both at a 5% and 1% test size. Deron concludes that *both* series contain a unit root but they are not cointegrated and that as such, being spurious, the estimated coefficients are invalid (biased and inconsistent). Do you agree with Deron's conclusions? Make sure to clearly justify your answer.

Debriefing.

Deron's may be correct but we have no evidence to back both his claims. In fact, the evidence provides is compatible with both:

$r_t^{10Y} \sim I(1)$, so that both r_t^{3m} and r_t^{10Y} contain a unit root but the regression (which is also a Engle-Granger's univariate cointegration test) indicates the absence of cointegration in the fact that the null of $I(1)$ residuals cannot be rejected (here you needed to recall that a Phillipps-Perron's test has a unit root null).

$r_t^{10Y} \sim I(0)$, so that the regression of r_t^{10Y} on r_t^{3m} is simply an “unbalanced” regression (another case of spurious regression) in which—by definition, because the sum of a $I(0)$ and a $I(1)$ series is $I(1)$ —the residuals are $I(1)$, as indeed established by the failure to reject by the PP test.

Yet, and in both cases, Deron is right when he claims that as result of the regression being either spurious or unbalanced, the estimated coefficients are invalid (biased and inconsistent).

Question 3.C (3.5 points)

Frank Tuvicci, a senior quant strategist at HappyHouse Hedge Fund, is having a heated discussion with a new junior colleague of his, John Marrone, about the nature of the time-series of US 1-month Treasury rates. Based on the evidence provided in Table 1, coming from a standard ADF test including both a constant and a trend, Frank has concluded that the series is $I(0)$; however, John claims that relying on a KPSS test (for which results are reported in Table 2) he has failed to reject the null hypothesis and therefore that the series must be $I(1)$. In general, is it possible that different tests may lead to different conclusions about the integration order of a series? In this specific case, based on the evidence displayed, do you think that both the claims of Frank and John were reasonable? Make sure to clearly justify your answer.

TABLE 1

Null Hypothesis: ONEMONTH has a unit root
 Exogenous: Constant, Linear Trend
 Lag Length: 5 (Automatic - based on SIC, maxlag=23)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-1.975860	0.6135
Test critical values: 1% level	-3.964599	
5% level	-3.413017	
10% level	-3.128509	

*MacKinnon (1996) one-sided p-values.

TABLE 2

Null Hypothesis: ONEMONTH is stationary
 Exogenous: Constant, Linear Trend
 Bandwidth: 30 (Newey-West automatic) using Bartlett kernel

	LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic	0.138773
Asymptotic critical values*:	
1% level	0.216000
5% level	0.146000
10% level	0.119000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)

Residual variance (no correction)	2.292096
HAC corrected variance (Bartlett kernel)	65.75177

Debriefing.

Although in general ADF-type and KPSS tests are sufficiently different to occasionally contradict each other, this is not the case. John is clearly making a mistake in interpreting the results of the KPSS test that he conducted. Indeed, he is failing to recognize that the KPSS is conducted under the null of (trend) stationarity. Therefore, John is rejecting the null of stationarity, which is perfectly compatible with the conclusion by Frank, who is rejecting the null of the presence of a unit root in the series.

Other Unit Root Tests

- In the case of real earnings, ADF test that includes an intercept gives an estimate of α of 0 with a t-ratio of 10.196 which leads to a failure to reject the null of a unit root.
- The presence of a time trend cannot be ruled out on theoretical grounds - an ADF test also including a linear time trend, gives an estimate of α of -0.002 which is -1.900 standard deviations away from 0 and that does not allow us to reject the null of a unit root.
- Phillips and Perron (1988)** propose a nonparametric method of controlling for serial correlation when testing for a unit root that is an alternative to the ADF test
 - Classical DF test + modify the t-ratio of α so that serial correlation in the residuals does not affect the asymptotic distribution of the test
 - See lecture notes for PP test statistic
 - Null hypothesis remains a unit root
- Kwiatkowski, Phillips, Schmidt, and Shin (1992)** have proposed a testing strategy **under the null of (trend-) stationarity**

Lecture 5: Unit Roots, Cointegration and Error Correction Models - Prof. Guidotti 20

Other Unit Root Tests

- KPSS statistic is based on residuals from a regression of the series on exogenous, deterministic factors: $y_{t+1} = \alpha' \delta + u_{t+1}$
- KPSS test is:
$$KPSS_T \equiv \frac{\sum_{t=1}^T (\sum_{s=1}^t \hat{u}_s)^2}{T^2 \sum_{t=1}^T \frac{1}{t^2} \sum_{s=1}^t \hat{u}_s \hat{u}_{t-s}}$$

Intercept and trend
- Re-examine whether S&P real stock prices, aggregate earnings, and aggregative dividends give evidence of a unit root, with PP tests:
 - $P_{t+1} = 0.110 + 0.002 P_t + \varepsilon_{t+1}$, $t_{\alpha}^{PP} = 0.905$ with p -value 0.996
 - $E_{t+1} = 0.017 - 0.001 E_t + \varepsilon_{t+1}$, $t_{\alpha}^{PP} = -1.535$ with p -value 0.516
 - $D_{t+1} = -0.0037 + 0.004 D_t + \varepsilon_{t+1}$, $t_{\alpha}^{PP} = 1.760$ with p -value 0.999
- All series contain a unit root and this should be taken into account
- KPSS tests lead to the same conclusion even though the null differs
- Although rejecting the null of a unit root does not imply "accepting" the alternative hypothesis of stationarity, **ADF-type and KPSS tests are sufficiently different to occasionally contradict each other**

Lecture 5: Unit Roots, Cointegration and Error Correction Models - Prof. Guidotti 21

Question 4.A (10 points)

What is a rolling variance forecast model? What type of ARMA model does it represent and what random variables does it concern? Make sure to discuss the main pros and cons of this model.

Debriefing.

Simple Models: Rolling Window Variance Forecast

- The most naive and yet surprisingly widespread models among practitioners are **simple rolling window models**:

$$\text{Var}_t[\varepsilon_{t+1}] = E_t[\varepsilon_{t+1}^2] \rightarrow \sigma_{t+1|t}^2(W) = \frac{1}{W} \sum_{s=1}^W \varepsilon_{t+1-s}^2 = \sum_{s=1}^W \left(\frac{1}{W}\right) \varepsilon_{t+1-s}^2$$
 - $\{\varepsilon_t\}$ consists of the empirical residuals of some conditional mean function model (a ARMA or a regression, say)
 - W is the rolling window length, the only parameter to be selected
- In short, this is a moving average model for squared residuals
- $W \ll T$ allows the model to capture time variation in conditional variance \Rightarrow predictive power that responds to market conditions
 - When $W = T$, the model gives the ML estimator of the variance
- This model has obvious limitations:
 - All past squared errors are given the same weight, $1/W$, irrespective of how old they are
 - Unclear how we should go about selecting the window length W as it represents the upper limit of a sum

Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidotti 9

2 Simple Univariate Parametric Models

2.1 Rolling Window Forecasts

Rolling window models of the conditional variance of $\{\varepsilon_t\}$ (time series process with zero conditional mean and $\sigma_{t+1|t} \equiv \text{Var}_t(\varepsilon_{t+1}) = E_t[\varepsilon_{t+1}^2]$ parametrized by the finite object $\theta \in \Theta \subseteq \mathbb{R}^k$. Then

$$\sigma_{t+1|t}^2(W) \equiv \frac{1}{W} \sum_{s=1}^W \varepsilon_{t+1-s}^2 = \sum_{s=1}^W \frac{1}{W} \varepsilon_{t+1-s}^2$$

where W = rolling window length is the only parameter in $\Theta \subseteq \mathbb{R}^k$

Simple Models: Rolling Window Variance Forecast

- Selection of W is left to subjective assessments, with the paradox that users with the same data, will deliver very different forecasts

Month	Return	MW Mean		Residual	Squared Residual	MW Variance Forecasts		
		W = 4	W = 4			W = 4	W = 3	W = 4
January	-11.55							
February	-4.35							
March	8.36							
April	2.05	-1.22	1.873	14.596				
May	-2.72	1.08	3.555	12.638				
June	9.15	0.17	-0.905	00.730				
July	3.10	-1.22	3.375	20.091				
August	0.06	-1.76	2.223	4.940	36.122			
September	-1.60	-1.53	-0.067	0.005	40.753	34.314		
October	1.39	1.10	0.288	0.083	38.167	31.008	20.439	
November	-8.32	-1.02	-3.303	10.907	11.270	20.041	25.441	
December	-11.24	1.60	-9.543	91.441	3.676	4.829	22.938	
					91.441	24.144	25.608	21.875
					Unconditional Avg.	23.609	22.138	21.830

- Especially when W is small, the forecasts generate "box shaped effects"

- When forecast spikes up, this may be due to either some small squared residual from $W + 1$ periods before been dropped or to a large time t squared residual
- The former event is hard to rationalize

Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidotti 10

- It is an equally weighted average over a sample of W recent and past observations of the squared residuals from some conditional mean model for the series of interest, say an asset return series.
- It represents a moving average model for squared residuals.
- The selection of $W \ll T$ allows the model to capture the time variation in conditional variance and employs it with some predictive power that responds to market conditions.

$$\lim_{W \rightarrow T} \sigma_{t+1|t}^2(W) = \lim_{W \rightarrow T} \frac{1}{W} \sum_{\tau=1}^W \epsilon_{t+1-\tau}^2 = \frac{1}{T} \sum_{\tau=1}^T \epsilon_{t+1-\tau}^2 = \hat{\sigma}_t^2$$

where $\hat{\sigma}_t^2$ = maximum likelihood sample variance estimator.

- Under a moving average model for returns, forecasts always exist since as new data arrive and all the sample information are used
- When all the data in the sample are used, the rolling window variance estimator becomes the sample variance.

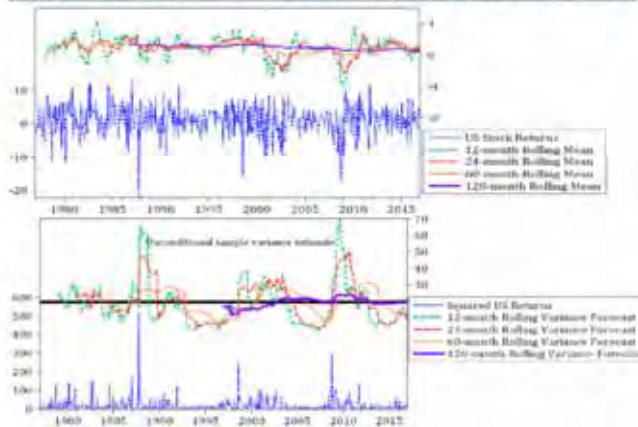
- The average of the rolling window variance forecasts is in general different from the sample variance estimator:

$$\frac{1}{T-W+1} \sum_{t=W}^T \sigma_{t+1|t}^2(W) = \frac{1}{(T-W+1)W} \sum_{t=W}^T \sum_{\tau=1}^W \epsilon_{t+1-\tau}^2 \neq \frac{1}{T} \sum_{\tau=1}^T \epsilon_{t+1-\tau}^2$$

- Limitations of computing variance forecasts using a rolling window model:

- All squared errors are given the same weight, $1/W$, irrespective of how old they are.
- It is unclear how W should be selected.
- Especially when W is small, the forecasts tend to generate frequent spikes, due to the fact that either some very small squared residual from $W - 1$ periods before has been dropped or at time t some large squared residual has been recorded and enters the calculation.

Simple Models: Rolling Window Variance Forecast



Question 4.B (3 points)

Mlado Vizov an analyst at Peeled & Head Ass. has just made a simple mathematical observation concerning a W -period rolling window variance estimator, $\sigma_{t+1|t}^2(W)$, namely that

$$\sigma_{t+1|t}^2(W) = \frac{1}{W} \sum_{\tau=1}^W \epsilon_{t+1-\tau}^2 = \epsilon_t^2 + \frac{1}{W-1} \sum_{\tau=1}^{W-1} \epsilon_{t-\tau}^2 = \epsilon_t^2 + \sigma_{t|t-1}^2(W-1).$$

Therefore, he claims that a rolling window variance estimator is just a special case of a RiskMetrics model, under the restriction that both the terms on the right-hand side are multiplied by a unit coefficient. Do you agree with his claim? Carefully explain your reasoning.

Debriefing.

As you know, a RiskMetrics model is simply written as

$$\sigma_{t+1|t}^2 = (1 - \lambda)\epsilon_t^2 + \lambda\sigma_{t|t-1}^2.$$

However, note that the variance process on the left- and right-hand sides of the RiskMetrics are the same: on the right we just have one lag of the process on the left. In the case pointed out by Mlado, we have instead that the process on the left, a W -observation rolling window variance estimator $\sigma_{t+1|t}^2(W)$ is structurally different from the process on the right, $\sigma_{t|t-1}^2(W-1)$, a $(W-1)$ -observation rolling window variance estimator, which just uses less data. Therefore, we can say that there is no restriction on RiskMetrics that can take us to a rolling window variance process and as a result Mlado is wrong.

Question 4.C (4 points)

Mr. Manly Beverly is due to give a presentation on the process followed by the conditional variance of the log-price returns on the 3-month futures on Wheat, traded on the Chicago Mercantile Exchange. The audience is composed of homogeneous type of customers: risk managers. Manly has obtained by ML methods consistent and asymptotically efficient estimates for the parameters of an EGARCH(1,1) process such that $\hat{\alpha} + \hat{\beta} = 1.1$. Moreover, also the point estimate of the parameter θ is positive. Therefore, he decides to advise risk managers to purchase options to hedge long-run risks originated by the price of wheat as quickly as possible, as an explosive variance process may justify a progressive increase in the price of long-term options, which may make hedging progressively more expensive, for a given size of risks. Is Manly's advise a sensible one in the light of the parameter estimates that he has obtained?

Debriefing.

Manly's concern is legitimate in the sense that an explosive conditional variance process would imply a long-run, ergodic variance that diverges to infinity and therefore that option positions set up to hedge increasingly distant maturity positions on wheat will become very expensive (here recall that the price of both long puts and calls is monotone increasing in their variance). However, his concern suffers from a flaw: the condition for stationarity of a EGARCH(1,1) is NOT $\hat{\alpha} + \hat{\beta} < 1$! Such conditions are generally more complex than in the GARCH case and—based on what we have said in the lecture—in the (1,1) case, they appear to only require that $\hat{\beta} < 1$, which may happily be satisfied in this case (i.e., we do not have enough information). Therefore, Manly's conclusion may be premature and ill-advised: he should study his econometrics better, before making claims in public.

Exponential GARCH Model

- Because $\sigma_{t+1|t}^2 = \exp(\ln \sigma_{t+1|t}^2)$ and $\exp(\cdot) > 0$, EGARCH always yields positive variance forecasts **without imposing restrictions**
 - $\{\sigma_{t+1|t}^2 + \theta(|z_{t-1}| - E|z_{t-1}|)\}$ is function of both the magnitude and the sign of past standardized residuals, and it allows the conditional variance process to respond asymmetrically to rises and falls in asset prices
 - It can be rewritten as: $const + [1 + ((I_{z_{t-1} > 0}) - (I_{z_{t-1} < 0}))^\theta] |z_{t-1}|$
 - Nelson's EGARCH has another advantage: in a GARCH, the parameter restrictions needed to ensure moment existence become increasingly stringent as the order of the moment grows
 - E.g., in case of ARCH(1), for an integer r , the $2r$ th moment exists if and only if $\alpha_1^r \prod_{k=1}^r (2k - 1) < 1$; for $r = 2$, existence of unconditional kurtosis requires $\alpha_1 < (1/3)^{1/2}$
 - In a EGARCH(p, q) case, if the error process η_t in the ARMA representation of the model has all moments and $\sum_{j=1}^q \beta_j^2 < 1$, then **all moments of an EGARCH process exist**
- How far better can EGARCH fare versus a standard GARCH model?
⇔ how important are asymmetries in conditional variance?

Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidolin 30

Question 5.A (10 points)

Define the statistical and economic nature of a EGARCH(p, q) model for conditional variance under a constraint of weak stationarity. How can you keep the forecasts of the conditional variance positive? What are in particular the reasons for the success of simple EGARCH(1,1) models in empirical applications? How would an EGARCH model capture the presence of (unconditional) skewness in typical time series of financial returns?

Debriefing.

Exponential GARCH Model

- Similarly to ARCH, GARCH captures thick-tailed returns and volatility clustering but it is not well suited to capture the "leverage effect" because $\sigma_{t+1|t}^2$ is only a function of ε_t^2 and not of their signs
- In the exponential GARCH (EGARCH) model of Nelson (1991), $\ln \sigma_{t+1|t}^2$ depends on both the size and the sign of lagged residuals and therefore **can capture asymmetries**

Definition (Exponential GARCH) In a EGARCH(p, q) model for the conditional log-variance, forecasts depends on a (non-negatively) weighted sum of past standardized errors (from some conditional mean function model, both in levels and in absolute values) and past log-variance forecasts:

$$\ln \sigma_{t+1|t}^2 = \omega + \sum_{i=1}^p \alpha_i [z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)] + \sum_{j=1}^q \beta_j \ln \sigma_{t-j+1|t-j}^2$$

The sequences (for fixed $i = 1, 2, \dots, p$) $\{z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)\}$ are zero-mean, IID random sequences in which, assuming $\theta < 0$):

If $z_{t-i} > 0$, $z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|) = \text{const} + (1 + \theta)z_{t-i}$, a linear function with slope $(1 + \theta) < 1$;

If $z_{t-i} < 0$, $z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|) = \text{const} + (1 - \theta)z_{t-i}$, a linear function with slope $(1 - \theta) > (1 + \theta)$.

Exponential GARCH model (EGARCH): In a EGARCH(p, q) model for the conditional log-variance, forecasts depends on a (non-negatively) weighted sum of past standardized errors (from some conditional mean function model, both in levels and in absolute values) and past log-variance forecasts:

$$\ln \sigma_{t+1|t}^2 = \omega + \sum_{i=1}^p \alpha_i [z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)] + \sum_{j=1}^q \beta_j \ln \sigma_{t-j+1|t-j}^2$$

where the sequences (for fixed $i = 1, 2, \dots, p$) $\{z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)\}$ are zero-mean, IID random sequences in which, assuming $\theta < 0$,

$$\text{If } z_{t-i} > 0 \text{ then } z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|) = \text{const} + (1 + \theta)z_{t-i}$$

that is a linear function with slope $(1 + \theta) < 1$.

$$\text{If } z_{t-i} < 0 \text{ then } z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|) = \text{const} + (1 - \theta)z_{t-i}$$

that is a linear function with slope $(1 - \theta) > (1 + \theta)$.

Or alternatively, EGARCH(1,1) can be specified as:

$$\ln \sigma_{t+1|t}^2 = \omega + \delta |z_t| + \theta z_t + (1 + \theta) \ln \sigma_{t|t-1}^2 = \omega + \delta \frac{\varepsilon_t}{\sigma_{t|t-1}} + \delta z_t + (1 + \theta) \ln \sigma_{t|t-1}^2$$

where δ captures the potential role of the asymmetries: when $\delta < 0$ then a negative residual increases the forecast of conditional variance more than a positive residual does.

It can be generalized to the case with p lags of the standardized residuals and q lags of past variance forecasts on the right-hand side.

Exponential GARCH Model

- Because $\sigma_{t+1|t}^2 = \exp(\ln \sigma_{t+1|t}^2)$ and $\exp(\cdot) > 0$, EGARCH always yields positive variance forecasts **without imposing restrictions**
 - $\{z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)\}$ is function of both the magnitude and the sign of past standardized residuals, and it allows the conditional variance process to respond asymmetrically to rises and falls in asset prices
 - It can be rewritten as: $\text{const} + [1 + (\mathbb{I}_{z_{t-i} > 0} - \mathbb{I}_{z_{t-i} < 0})\theta]z_{t-i}$
 - Nelson's EGARCH has another advantage: in a GARCH, the parameter restrictions needed to ensure moment existence become increasingly stringent as the order of the moment grows
 - E.g., in case of ARCH(1), for an integer r , the r th moment exists if and only if $\alpha_1^r [1 + (2r - 1)\alpha_1] < 1$; for $r = 2$, existence of unconditional kurtosis requires $\alpha_1 < (1/3)^{1/2}$
 - In a EGARCH(p, q) case, if the error process η_t in the ARMA representation of the model has all moments and $\sum_{i=1}^p \alpha_i^2 < 1$, then all **moments of an EGARCH process exist**
- How far better can EGARCH fare versus a standard GARCH model?
 - how important are asymmetries in conditional variance?

Lecture 6: Univariate Volatility Modelling: ARCH and GARCH - Prof. Juhász

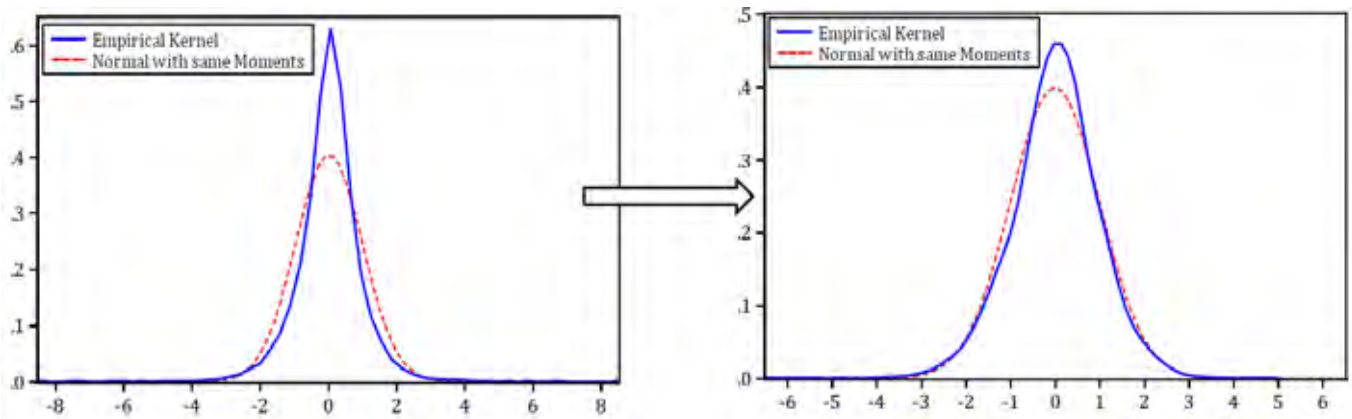
- EGARCH is a model that directly express forecasts not of future conditional variance, but of future conditional log-variance, where $\sigma_{t+1|t}^2$ depends on both the size and the sign of lagged residuals.
- $\{z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)\}$ is function of both the magnitude and the sign of past standardized residuals and it allows the conditional variance process to respond asymmetrically to rises and falls in asset prices compared to their mean.
- $\{z_{t-i} + \theta(|z_{t-i}| - E|z_{t-i}|)\}$ can be rewritten as

$$\text{const} + [1 + (\mathbb{I}_{z_{t-i} > 0} - \mathbb{I}_{z_{t-i} < 0})\theta]z_{t-i}$$
 - EGARCH captures the leverage effect and, more generally, the existence of asymmetries in conditional variance.
- No restrictions on the parameters are necessary to ensure non-negativity of the conditional variances and it is possible to find cases in which either negative or positive past shocks end up decreasing the forecast of variance instead of increasing it (asymmetric effect).
- In a EGARCH(p, q) case, if the error process $\{\eta_{t+i}\}$ in the ARMA representation of the model has all moments and $\sum_{i=1}^p \alpha_i^2 < 1$, then all moments for the EGARCH process will exist.

As for the reference to skewness, as discussed in the lectures, when a conditional variance model implies asymmetric, leverage effects, this means that large negative returns imply an increase in conditional variance that exceeds the increase induced by large positive returns; therefore negative returns may induce even larger negative returns (because variance is high) and this will end up inflating the left tail of the unconditional distribution vs. the right tail, which will translate in the presence of unconditional skewness.

Question 5.B (3 points)

Ms. Martina Kalvin is analyzing the series of US daily excess stock returns for a long 1963-2016 sample. On the left, you can see the kernel density estimator of the standardized residuals from a homoskedastic ARMA(2,1) model with Gaussian shocks; as a benchmark, the kernel density estimator is compared to a N(0,1). On the right, you can see the kernel density estimator of the standardized residuals from a ARMA(2,1)/ EGARCH(1,1) model with Gaussian shocks; as a benchmark, the kernel density estimator is compared to a N(0,1). Martina does *not* specify whether she has estimated her model by MLE or QMLE.



Ms. Kelvin claims that moving from left to the right, the validity of the model records a considerable improvement: do you agree and why? However, Martina reckons that the ARMA(2,1)/EGARCH(1,1) model with Gaussian shocks should be rejected: why would she claim that? Assuming you agree with her, what do you think may cause the difficulties that characterized the Gaussian ARMA(2,1)/EGARCH(1,1) model? Carefully motivate your replies.

Debriefing.

Probably it is trivial, but when moving from the left to the right panel, the improvements in the quality of the fit provided by the model to the data are evident: on the left, the blue empirical kernel density strongly departs from the Gaussian $N(0,1)$ benchmark under which the conditional mean model was estimated; on the right, the empirical kernel density of the standardized residuals approaches, even though it remains visibly different to the $N(0,1)$ benchmark under which the ARMA(2,1)/EGARCH(1,1) model has been estimated. However, whether or not the model can be rejected depends entirely on whether the model had been estimated by either MLE or QMLE:

— if the model were estimated by MLE, then a rejection would be justified, because also in the right panel there is a significant deviation of the empirical kernel from the assumed $N(0,1)$ distribution;

— if the model were estimated by QMLE, then a rejection would *not* be justified, because the significant deviation of the empirical kernel from the assumed $N(0,1)$ distribution in the right panel is not only admissible, but even expected, given that the Gaussian distribution for the shocks has been just assumed as an approximation.

Question 5.C (4 points)

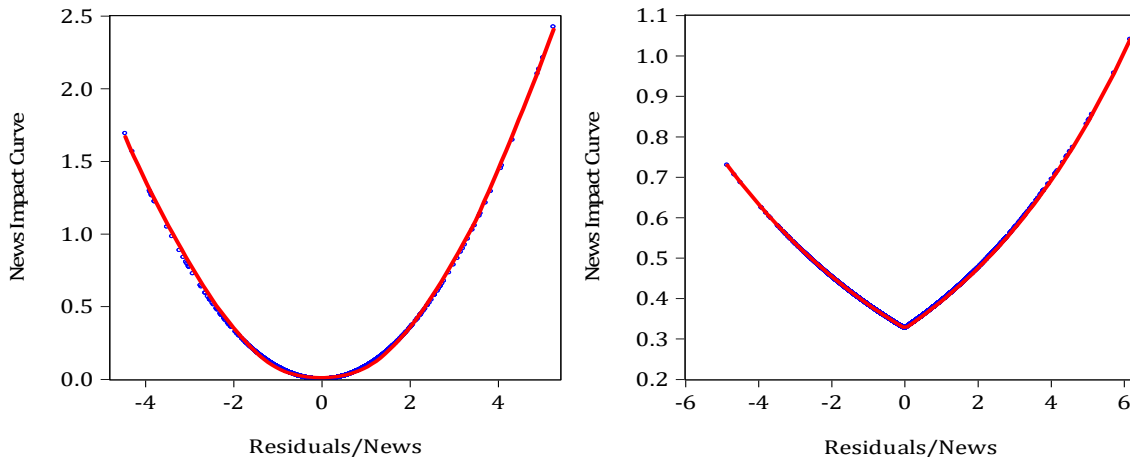
Eventually, Martina has estimated a ARMA(2,1)/EGARCH(1,1) model with GED errors that turns out to be as follows (standard errors are in parentheses):

$$R_{t+1} = \underset{(0.043)}{0.074} + \underset{(0.017)}{0.042}R_t + \underset{(0.019)}{0.053}R_t - \underset{(0.013)}{0.033}\epsilon_t + \underset{(0.008)}{0.022}\epsilon_t + \epsilon_{t+1}$$

$$\ln\sigma_{t+1|t}^2 = \underset{(0.008)}{-0.157} + \underset{(0.006)}{0.177} \left| \frac{\epsilon_t}{\sigma_{t|t-1}^2} \right| + \underset{(0.004)}{0.011} \frac{\epsilon_t}{\sigma_{t|t-1}^2}$$

$$+ \underset{(0.120)}{0.990} \ln\sigma_{t|t-1}^2 \quad \epsilon_t \text{ IID } GED(0, \sigma_{t+1|t}^2; 1.530) \quad \underset{(0.352)}{}$$

However, Martina has forgotten to label the following two pictures concerning the news impact curve (NIC) derived from the estimated ARMA(2,1)/EGARCH(1,1) model and she no longer remembers whether it is the left or the right plots that represents the NIC of the model above.



Can you help her finding the appropriate NIC that refers to the model above? Carefully explain what has guided your selection/answer and why the remaining plot is not plausibly derived from a ARMA(2,1)/EGARCH(1,1) model. Martina is also confused as to whether her estimates are either ML or QML. Can you help her? Make sure to justify your answer.

Debriefing.

The ARMA(2,1)/EGARCH(1,1) NIC is the rightmost one. We can detect that from the existence of a kink induced by the appearance of the absolute value of the standardized error on the right-hand side of the conditional variance model. As you will recall from your basic math courses, the kink point occurs in correspondence to the change of sign of the absolute value, when the function fails to be differentiable. The leftmost plot is everywhere differentiable and as such it cannot represent the NIC from a EGARCH model. Finally, because the model has been estimated assuming GED and not Gaussian errors, clearly the estimation methods must be full-information maximum likelihood (MLE) and cannot be QML, that would incorrectly assume a pseudo normal density for the errors.

Question 6.A (9 points)

Describe the theoretical justifications as well as the practical implementations of tests of the forecasting validity of a conditional heteroscedasticity model based on the linear regression

$$\epsilon_{t+1}^2 = a + b\hat{\sigma}_{t+1|t}^2 + e_{t+1},$$

where e_{t+1} is a white noise shock and $\hat{\sigma}_{t+1|t}^2$ are the one-step ahead conditional variance forecasts derived from a given model. How would you estimate this linear model? Under what circumstances the null that the model yields unbiased and efficient forecasts will be rejected? Discuss whether you would also use the regression R-square to assess the validity of the variance model. Make sure to clearly justify your answers.

Debriefing.

Are ARCH Models Enough?

- What does it mean that a CH models yield "good" forecasts? A requirement is that on average the realized squared residuals must equal the variance forecasts that a model offers:

$$\sigma_{t+1|t}^2 = E[\epsilon_{t+1}^2] = \sigma_{t+1}^2 - \epsilon_{t+1|t} = \sigma_{t+1|t}^2 + \epsilon_{t+1|t}$$
White noise
- Empirically, it implies that **two simple restrictions must be satisfied in the regression**

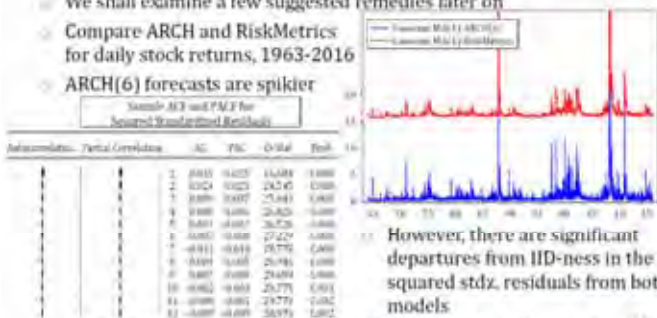
$$\sigma_{t+1|t}^2 = \mu + b\sigma_{t+1|t}^2 + \epsilon_{t+1|t}$$
- $a = 0$ and $b = 1$, jointly (when this occurs, $\sigma_{t+1|t}^2$ offers an unbiased predictor of squared residuals, used as a proxy of realized variance)
- The regression R^2 must be "large"
- However, this test of predictive performance may be fallacious: the process $\{\epsilon_t^2\}$ invariably provides a poor proxy for the process followed by the true but unobserved time-varying variance, $\{\sigma_t^2\}$
- This follows from

$$\begin{aligned} \text{Var}[\epsilon_{t+1}^2] &= E[(\epsilon_{t+1}^2 - \sigma_{t+1|t}^2)^2] = E[(\sigma_{t+1|t}^2 - \sigma_{t+1}^2 + \epsilon_{t+1|t})^2] \\ &= \sigma_{t+1|t}^4 E[(\epsilon_{t+1|t}^2 - 1)^2] = \sigma_{t+1|t}^4 E[\epsilon_{t+1|t}^4 + \epsilon_{t+1|t}^2 - 2\epsilon_{t+1|t}^2] \\ &= \sigma_{t+1|t}^4 (1 + \text{kurt}(\epsilon_{t+1|t}) - 2) = \sigma_{t+1|t}^4 (\text{kurt}(\epsilon_{t+1|t}) - 1) \end{aligned}$$

Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidetti 19

Are ARCH Models Enough?

- When either $\sigma_{t+1|t}^2$ (hence, σ_{t+1}^2) or the kurtosis of the stdz. residuals are high, $\text{Var}[\epsilon_{t+1}^2]$ will be large, and using squared residuals to proxy instantaneous variances exposes a researcher to a lot of noise
- This choice is almost guaranteed to yield low regression R^2
- We shall examine a few suggested remedies later on
- Compare ARCH and RiskMetrics for daily stock returns, 1963-2016
- ARCH(6) forecasts are spikier



Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidetti 20

Generalized ARCH Models

- Although R^2 s are not irrelevant, positive significant estimates of intercepts \Rightarrow predicted variance is too low vs. realized variance
- The two slope coefficients significantly less than 1 \Rightarrow realized variance moves over time less vs. what is predicted

Question 6.B (4.5 points)

Mikki Paranoich, an independent researcher, has estimated two models to predict the one-day ahead variance of US aggregate excess stock returns. For two models, call them A and B, Mikki has obtained the following results from a regression of squared residuals (from a MA(1) model that has been pre-specified using Box-Jenkins analysis) on variance predictions, $\hat{\sigma}_{A,t+1|t}^2$ and $\hat{\sigma}_{B,t+1|t}^2$ (estimated standard errors are reported in parantheses):

$$\begin{aligned} \epsilon_{t+1}^2 &= \underset{(0.281)}{0.434} + \underset{(0.377)}{0.549} \hat{\sigma}_{A,t+1|t}^2 + \hat{\epsilon}_{t+1}^A & R^2 &= 0.029, \\ \epsilon_{t+1}^2 &= \underset{(0.038)}{-0.106} + \underset{(0.023)}{1.146} \hat{\sigma}_{B,t+1|t}^2 + \hat{\epsilon}_{t+1}^B & R^2 &= 0.159. \end{aligned}$$

Moreover, in the case of model A, a test of the joint null of $a = 0$ and $b = 1$ using an F-test leads to a rejection. Scatter plots of the squared residuals vs. variance predictions with a regression

- Because the CH models are time series models to be used in forecasting, a good CH model should be able to "adequately" predict future variance.
 - \Rightarrow Minimum requirement: the realized squared residuals must equal the variance forecasts.

$$\sigma_{t+1|t}^2 = E[\epsilon_{t+1}^2] = \epsilon_{t+1}^2 - \epsilon_{t+1|t} \Rightarrow \epsilon_{t+1}^2 = \sigma_{t+1|t}^2 + \epsilon_{t+1|t}$$

where $\epsilon_{t+1|t}$ are zero-mean, white noise forecast errors.

$$\Rightarrow \epsilon_{t+1}^2 = a + b\sigma_{t+1|t}^2 + \epsilon_{t+1|t} \text{ where}$$

- (a) $a = 0$ and $b = 1$ jointly $\Rightarrow \sigma_{t+1|t}^2$ offers an unbiased predictor of squared residuals
- (b) forecast errors are small.

- Main problem with $\epsilon_{t+1}^2 = a + b\sigma_{t+1|t}^2 + \epsilon_{t+1|t}$ (ϵ_{t+1}^2) provides a very poor proxy for the process followed by the true but unobserved time-varying variance, $\{\sigma_{t+1}^2\}$. In fact

$$\begin{aligned} \text{Var}[\epsilon_{t+1}^2] &= E[(\epsilon_{t+1}^2 - \sigma_{t+1|t}^2)^2] = E[(\sigma_{t+1|t}^2 - \sigma_{t+1}^2 + \epsilon_{t+1|t})^2] = \sigma_{t+1|t}^4 E[(\epsilon_{t+1|t}^2 - 1)^2] \\ &= \sigma_{t+1|t}^4 (E[\epsilon_{t+1|t}^4] + \epsilon_{t+1|t}^2 - 2\epsilon_{t+1|t}^2) = \sigma_{t+1|t}^4 (1 + \text{kurt}(\epsilon_{t+1|t}) - 2) = \sigma_{t+1|t}^4 (\text{kurt}(\epsilon_{t+1|t}) - 1) \end{aligned}$$

thus when either $\sigma_{t+1|t}^2$ or the kurtosis of the the standardized residuals are high, then $\text{Var}[\epsilon_{t+1}^2]$ will be large.

Equivalently, taking the coefficient of variation, $E[\epsilon_{t+1}^2] / \sqrt{\text{Var}[\epsilon_{t+1}^2]}$ as a measure of the variability of an estimator, then

$$\frac{E[\epsilon_{t+1}^2]}{\sqrt{\text{Var}[\epsilon_{t+1}^2]}} = \frac{\sigma_{t+1|t}^2}{\sqrt{\sigma_{t+1|t}^4 (\text{kurt}(\epsilon_{t+1|t}) - 1)}} = \frac{1}{\sqrt{\text{kurt}(\epsilon_{t+1|t}) - 1}}$$

that declines as $\text{kurt}(\epsilon_{t+1|t})$ increases.

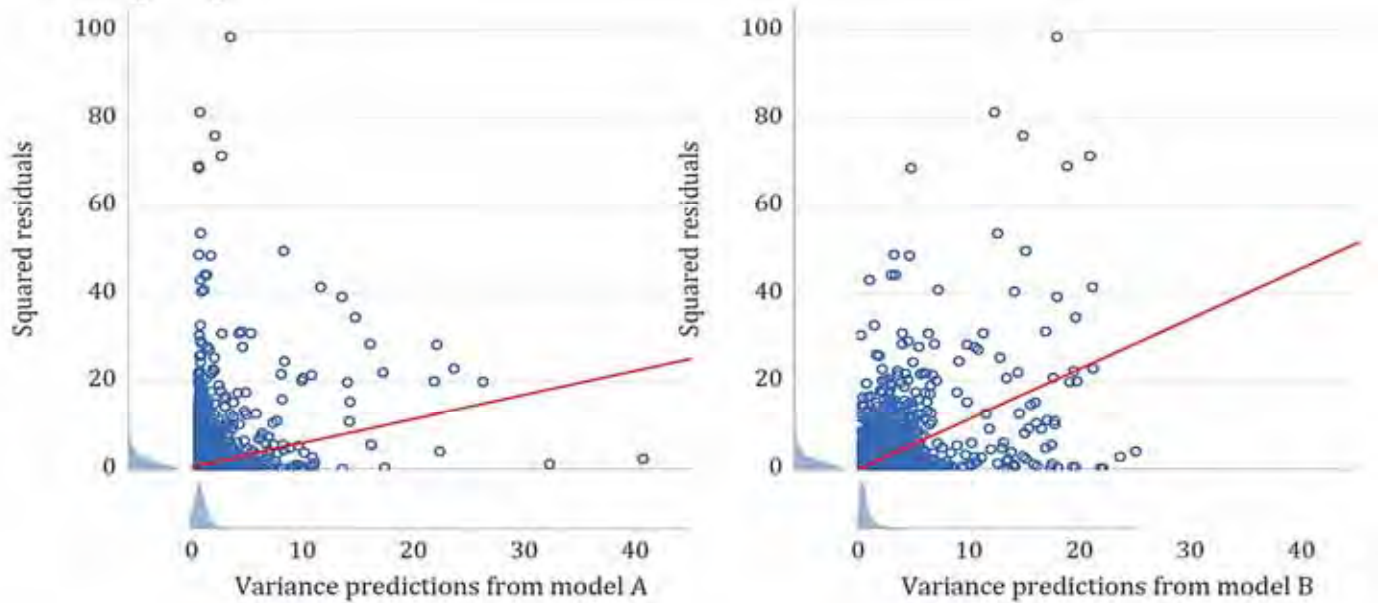
Are ARCH Models Enough?

- All forecasting power of past squared residuals is well captured
- It seems that past US equities losses lead to subsequent higher variance, the leverage effect
- Predictive accuracy regressions give (std errors in parentheses):
 - RiskMetrics: $\epsilon_{t+1}^2 = 0.142 + 0.854(\sigma_{t+1|t}^{\text{RM}})^2 + \epsilon_{t+1}$ $R^2 = 13.6\%$
 - ARCH(6): $\epsilon_{t+1}^2 = 0.197 + 0.791(\sigma_{t+1|t}^{\text{ARCH}})^2 + \epsilon_{t+1}$ $R^2 = 13.8\%$
- Crucial to report standard errors and not p-values because the simple null hypothesis of $b = 1$ requires that we calculate the t ratios:

$$\begin{aligned} \text{RiskMetrics: } t_{b=1}^{\text{RM}} &= \frac{0.854 - 1}{0.018} = \frac{0.146}{0.018} = -8.111 \\ \text{ARCH(6): } t_{b=1}^{\text{ARCH}} &= \frac{0.791 - 1}{0.017} = \frac{0.209}{0.017} = -12.29 \end{aligned}$$
- The null of $a = 0$ may be rejected with p-values close to 0.000
- Given individual rejections, pointless to apply F-tests of joint hypothesis

Lecture 6: Univariate Volatility Modelling, ARCH and GARCH - Prof. Guidetti 21

line superimposed are as follows:



Which of the two models, if any, can be considered to be a valid prediction tool? Make sure to clearly justify your answers.

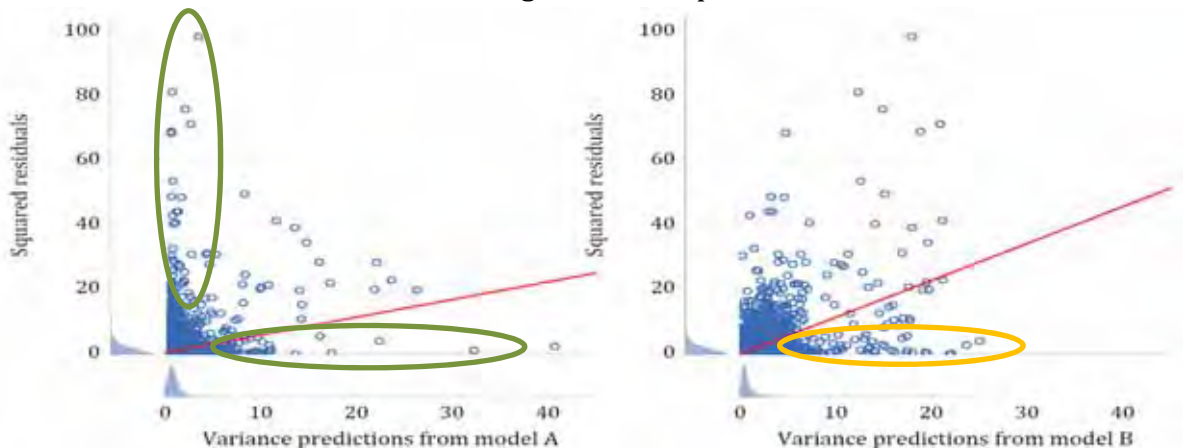
Debriefing.

In the case of model A, we have:

$$t_{a=0}^A = \frac{0.434}{0.281} = 1.545 < 2 \Rightarrow \text{fail to reject null of } a = 0$$

$$t_{b=1}^A = \frac{0.549 - 1}{0.377} = -1.196 \Rightarrow |1.196| < 2 \Rightarrow \text{fail to reject null of } b = 1$$

Therefore, in this case the model fails to be rejected. Yet, despite the parametric structure of the model is not rejected, the R-square in this case is largely disappointing. The left panel of the picture shows an interesting phenomenon: in a non-negligible fraction of the sample, recorded variance is large and exceeds 20 (careful, this is not a percentage!) but the model predicts a variance of almost zero, which is a reason for concern; in a few cases, we also record the opposite pattern: the recorded squared error is small and a below 1-2, but model A returns predictions that exceed 5 or even 10, see the green circles in the copy of the figures below. These regularities contribute to a rather small regression R-square.



In the case of model B, we have:

$$t_{a=0}^B = \frac{-0.106}{0.038} - 2.762 \Rightarrow |2.762| > 2 \Rightarrow \text{reject null of } a = 0$$

$$t_{b=1}^B = \frac{1.146 - 1}{0.027} = 6.448 \gg 2 \Rightarrow \text{reject null of } b = 1$$

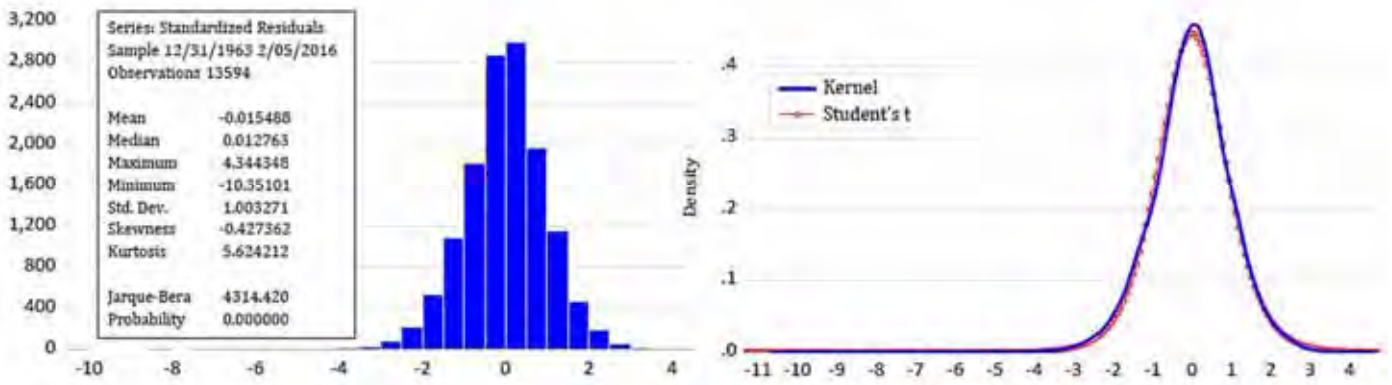
Clearly, in this case to test the joint null of $a = 0$ and $b = 1$ using an F-test will lead to a rejection. However, the R-square of this regression is not as low and disappointing as the one we have gotten for model A, and corresponds to almost the maximum one may hope to get with this type of data. This is qualitatively confirmed by the rightmost plot of the figure, in which high squared residuals are always matched by non-zero a substantial variance predictions: when variance will be high, the model will forecast that. However, remains visible and actually gets even stronger (see orange circle) the second type of bias: in a considerable fraction of the sample, the recorded squared error is small and a below 1-2, but model B returns predictions that exceed 5 or even 10, which implies that a fraction of the time, mode B predicts a high variance that fails to materialize in the data.

Question 6.C (2.5 points)

In the case of model B, Mikky proceeds then to look for ways to improve the model and its predictive performance. He obtains the following evidence:

Correlogram of Standardized Residuals Squared

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.012	0.012	2.0284	0.154
		2	0.013	0.013	4.4512	0.108
		3	0.003	0.003	4.5777	0.205
		4	0.006	0.006	5.0391	0.283
		5	0.005	0.005	5.3715	0.372
		6	-0.009	-0.009	6.4624	0.373
		7	-0.005	-0.005	6.7768	0.452
		8	0.007	0.007	7.3683	0.497
		9	0.008	0.008	8.1469	0.519
		10	0.003	0.003	8.2723	0.602
		11	0.001	0.001	8.3001	0.686
		12	-0.009	-0.009	9.3877	0.670
		13	-0.003	-0.003	9.5413	0.731
		14	0.006	0.006	9.9991	0.762
		15	-0.007	-0.007	10.644	0.777
		16	-0.006	-0.005	11.066	0.805
		17	-0.006	-0.005	11.488	0.830
		18	-0.006	-0.006	11.927	0.851
		19	0.003	0.003	12.058	0.883
		20	-0.012	-0.012	14.030	0.829



Heteroskedasticity Test: ARCH

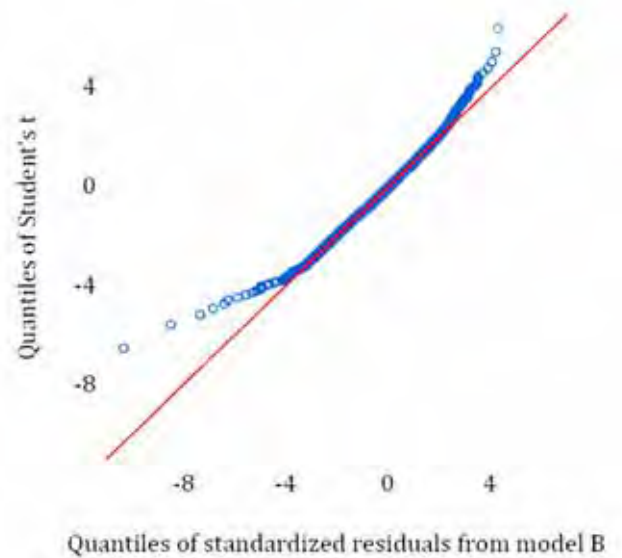
F-statistic	0.708403	Prob. F(15,13563)	0.7786
Obs*R-squared	10.63025	Prob. Chi-Square(15)	0.7783

Test Equation:

Dependent Variable: WGT_RESID^2
 Method: Least Squares
 Date: 06/10/18 Time: 12:28
 Sample (adjusted): 1/21/1964 2/05/2016
 Included observations: 13579 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.977614	0.037462	26.09627	0.0000
WGT_RESID^2(-1)	0.011953	0.008586	1.392130	0.1639
WGT_RESID^2(-2)	0.013269	0.008587	1.545238	0.1223
WGT_RESID^2(-3)	0.002626	0.008588	0.305755	0.7598
WGT_RESID^2(-4)	0.005673	0.008588	0.660620	0.5089
WGT_RESID^2(-5)	0.004777	0.008588	0.556286	0.5780
WGT_RESID^2(-6)	-0.009433	0.008588	-1.098376	0.2721
WGT_RESID^2(-7)	-0.004899	0.008588	-0.570402	0.5684
WGT_RESID^2(-8)	0.006865	0.008588	0.799391	0.4241
WGT_RESID^2(-9)	0.007395	0.008588	0.861054	0.3892
WGT_RESID^2(-10)	0.002875	0.008588	0.334799	0.7378
WGT_RESID^2(-11)	0.001457	0.008588	0.169673	0.8653
WGT_RESID^2(-12)	-0.009217	0.008588	-1.073269	0.2832
WGT_RESID^2(-13)	-0.003398	0.008588	-0.395663	0.6924
WGT_RESID^2(-14)	0.006185	0.008587	0.720219	0.4714
WGT_RESID^2(-15)	-0.006734	0.008587	-0.784227	0.4329

R-squared	0.000783	Mean dependent var	1.007221
Adjusted R-squared	-0.000322	S.D. dependent var	2.171752



Keep in mind that model B has been estimated assuming that the standardized shocks are drawn from a t-Student distribution, which justifies the selection of benchmark in the kernel density plot (third plot going clockwise). What is your advice to Mikki as to ways to improve the predictive power of model B? Make sure to clearly justify your answer.

Debriefing.

In fact, it all looks rather good apart from one piece of evidence: the kernel density comparison and especially the quantile-quantile plot reveal that the t-student inflates the tails of the predicted density *excessively* given the tail thickness expressed by the data (also because the estimated number of degrees of freedom, less than 8, appears to be really small). On the contrary there is no evidence that any residual ARCH structure is left in the data or of asymmetries that are not captured (see the kernel plot), even though further tests for asymmetries using the LM principles or news impact curves might be explored. Finally, note

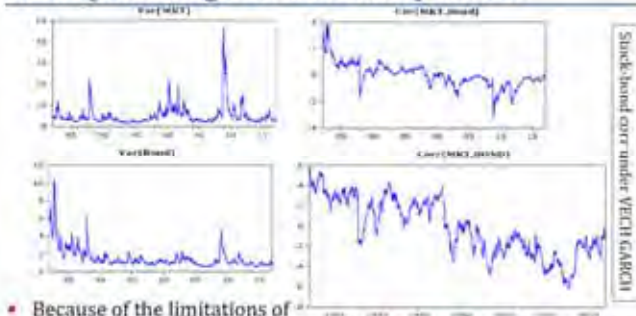
that one piece of evidence is rather redundant and unhelpful—the histogram provides the background to test for normality, but there is no presumption here that the data may come from a Gaussian distribution.

OPTIONAL Question 7 (4 points)

Describe the structure of a BEKK (Baba-Engle-Kraft-Kroner) multivariate GARCH (p, q) model. Make sure to illustrate its key advantages and disadvantages. How many parameters would you need to estimate in the BEKK(1,1) case?

Debriefing.

BEKK (Baba-Engle-Kraft-Kroner) GARCH



Because of the limitations of VECH, during the 1990s, one multivariate GARCH model surged to popularity, Engle and Kroner's (1995) **BEKK** (p, q):

Non-negative NaN matrices $\Sigma_{t+1|t} = CC' + \sum_{i=1}^p A_i (\epsilon_{t+1,i} \epsilon'_{t+1,i}) A_i' + \sum_{j=1}^q B_j \Sigma_{t+1-j} B_j'$

Lecture 6: Multivariate GARCH and Conditional Correlation Models - Prof. Guidetti 34

2. Estimate constant correlations using a simple sample estimator based on the standardized returns $\hat{u}_{t+1} = \hat{R}_{t+1} / \hat{\sigma}_{t+1}$ derived from the first step using GARCH-type models. Here, we exploit the fact that the conditional covariance of the \hat{u}_{t+1} variables equals the conditional correlation of two returns.

$$\hat{\rho}_{ij} = \frac{1}{T} \sum_{t=1}^T \hat{u}_{t+1,i} \hat{u}_{t+1,j}$$

Such constant correlations are then inserted inside $\hat{\Gamma}$ to estimate the constant correlation matrix.

6.3 BEKK GARCH

Given the picture provided above and the fact that DCC is a model popularized around the turn of the millennium, one may ask what was the state of multivariate GARCH modelling in practice before DCC became as popular as it is today. Apart from the uncomfortable case of CCC models that assume constant correlations over time, during the 1990s one of the most popular multivariate GARCH models had been Engle and Kroner's (1995) BEKK GARCH(p, q)²².

$$\Sigma_{t+1} = CC' + \sum_{i=1}^p A_i (\mathbf{R}_{t+1-i} \mathbf{R}'_{t+1-i}) A_i' + \sum_{j=1}^q B_j \Sigma_{t+1-j} B_j'$$

where the matrices $\{A_i\}_{i=1}^p$ and $\{B_j\}_{j=1}^q$ are non-negative and symmetric. This spatial product-sandwich form that is used to write the BEKK ensures the PSD property without imposing further restrictions, which represent the key reason for the success of BEKK models. In fact, this full matrix BEKK is easier to estimate than vech-GARCH models, even though it remains rather complex to handle. In practice, the popular form of BEKK that many empirical analysts have come to appreciate is a simpler (1,1) diagonal BEKK that restricts the matrices **A** and **B** to be diagonal matrices. BEKK models possess three attractive properties:

- 1. A BEKK is a truncated, low-dimensional application of a theorem by which all non-negative, symmetric $N \times N$ matrices (say, **M**) can be decomposed (for instance) as

$$\mathbf{M} = \begin{bmatrix} m_{11} & \dots & m_{1N} \\ \vdots & \ddots & \vdots \\ m_{N1} & \dots & m_{NN} \end{bmatrix} = \sum_{i=1}^N \begin{bmatrix} m_{i1} & \dots & m_{iN} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}$$

the appropriately selected vectors $m_{i\cdot}$. In a sense, mathematically it is no surprise that BEKK models often offer a good fit to the dynamics of variance.

- 2. As already mentioned, it easily ensures PSDness of the covariance matrix.
- 3. BEKK is invariant to linear combinations: e.g., if \mathbf{R}_{t+1} follows a BEKK GARCH(p, q), then any portfolio formed from the N securities or assets in \mathbf{R}_{t+1} will also follow a BEKK.

²²It is now 200 months since BEKK means "Baba-Engle-Kraft-Kroner" and the average weekly computer time spent at the four econometrics labs involved to do this development.

However, the number of parameters in BEKK remains rather large

$$0.5N(N+1) + 0.5pN(N+1) + 0.5qN(N+1) = 0.5N(N+1)(1+p+q) = O(N^2)$$

Often, this has still made DCC models preferable in practice. However, the number of parameters in BEKK is substantially inferior to those appearing in a full VEC specification. This happens because the parameters governing the dynamics of the covariance equation in BEKK models are the products of the corresponding parameters of the two corresponding variance equations in the same model.

The second and third properties of BEKK models can only be appreciated contrasting the features of BEKK under linear aggregation with the properties of alternative multivariate GARCH models, for instance even a simple diagonal vech ARCH. Not all multivariate GARCH models are invariant with respect to linear transformations.²² For instance, for the case of two asset return series ($N = 2$), consider a simple diagonal multivariate ARCH(1) model obtained from a simplification of the diagonal GARCH(p, q) introduced early on:

$$\text{vech}(\Omega_t) = (\Omega_0 - \mathbf{A}) + \text{vech} \left(\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_i \mathbf{R}_i' \right) + \mathbf{A} + \text{vech}(\mathbf{R}_{t-1} \mathbf{R}_{t-1}') \quad (10)$$

where the helpful variance-targeting restriction has already been imposed and \mathbf{A} is a diagonal matrix. Because we have set $N = 2$, Ω_t will be a 2×2 matrix, \mathbf{A} is a 2×2 diagonal matrix, \mathbf{R}_i is a 2×1 vector of asset returns, $\text{vech}(\Omega_t)$ is a 2×1 vector of unique elements from Ω_t , $\text{vech}(\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_i \mathbf{R}_i')$ is a 2×1 vector of unique elements from the sum of cross-product matrices $\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_i \mathbf{R}_i'$, $\text{vech}(\mathbf{R}_{t-1} \mathbf{R}_{t-1}')$ is a 2×1 vector of unique elements from the lagged cross-product matrix $\mathbf{R}_{t-1} \mathbf{R}_{t-1}'$. The number of coefficients to be estimated is of course $\frac{1}{2} n^{11}$, n^{22} , and n^{33} in the representation

$$\begin{bmatrix} \sigma_{1,1,t} \\ \sigma_{1,2,t} \\ \sigma_{2,2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} a^{11} & 0 & 0 \\ 0 & a^{22} & 0 \\ 0 & 0 & a^{33} \end{bmatrix} \begin{bmatrix} \mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \\ \mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} \\ \mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \end{bmatrix} + \begin{bmatrix} a^{11} & 0 & 0 \\ 0 & a^{22} & 0 \\ 0 & 0 & a^{33} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{t-1} \\ \mathbf{R}_{t-1} \mathbf{R}_{t-1} \\ \mathbf{R}_{t-1}^2 \end{bmatrix} = \begin{bmatrix} (1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + a^{11} \mathbf{R}_{t-1}^2 \\ (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} + a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \\ (1-a^{33})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + a^{33} \mathbf{R}_{t-1}^2 \end{bmatrix}$$

As for the conditions that guarantee that $\sigma_{1,1,t} > 0$ and $\sigma_{2,2,t} > 0$ at all times, i.e., that ensure PSDness of the model, check

$$\begin{aligned} (1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + a^{11} \mathbf{R}_{t-1}^2 &> 0 \text{ if and only if } a^{11} < (0, 1) \\ (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + a^{22} \mathbf{R}_{t-1}^2 &> 0 \text{ if and only if } a^{22} < (0, 1) \end{aligned}$$

²²In literature of a model, we mean that it stays in the same class if a linear transformation is applied to $\mathbf{R}_{t-1} = \mathbf{T} \mathbf{R}_{t-1}$, where \mathbf{T} is a square matrix of constants and \mathbf{R}_{t-1} corresponds to asset returns (with/without including the original assets). It seems reasonable that a model should be invariant, otherwise the question arises which base assets should be included.

and variances simultaneously while satisfying the positivity requirement for the variances and keeping Ω_t semi-positive definite at all times. Equivalently, if one wants to impose that the diagonal vech ARCH(1) model delivers a filtered covariance matrix Ω_t that is semi-positive definite at all times, the diagonal model itself must be turned into a constant-covariance multivariate ARCH model, so you understand that $a^{22} = 0$ implies $\sigma_{1,2,t} = \mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} = \sigma_{1,2,t}$ so that

$$\rho_{1,2,t} = \frac{\sigma_{1,2,t}}{\sqrt{\sigma_{1,1,t} \sigma_{2,2,t}}} = \frac{\sigma_{1,2,t}}{\sqrt{a^{22} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \sum_{i=1}^p \mathbf{R}_{i,t}^2}}$$

and dynamics in conditional correlations will exclusively come from dynamics in volatilities.²³

Let's now examine the issue concerning the fact that while BEKK is "closed" under linear aggregation, a simpler diagonal vech-GARCH model is not. Consider a portfolio of the two assets, with weights w and $(1-w)$. We show that in spite of the fact that \mathbf{R}_{t-1} is characterized by a diagonal bivariate ARCH(1), the portfolio return $R_t^p = wR_{1,t} + (1-w)R_{2,t}$ has a variance process $\sigma_{p,p,t} \equiv \text{Var}_{t-1}[R_t^p]$ that fails to display the typical "diagonal form", i.e., $(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + a^{22} \mathbf{R}_{t-1}^2$. Note first that

$$\begin{aligned} \sigma_{p,p,t} &\equiv \text{Var}_{t-1}[R_t^p] = \text{Var}_{t-1}[wR_{1,t} + (1-w)R_{2,t}] \\ &= w^2 \sigma_{1,1,t} + (1-w)^2 \sigma_{2,2,t} + 2w(1-w)\sigma_{1,2,t} \\ &= w^2 (1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + w^2 a^{11} \mathbf{R}_{t-1}^2 + (1-w)^2 (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \\ &\quad + (1-w)^2 a^{22} \mathbf{R}_{t-1}^2 + 2w(1-w)(1-a^{12})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} + 2w(1-w)a^{12} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \end{aligned}$$

which cannot be written in diagonal form, $(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p (wR_{1,t} + (1-w)R_{2,t})^2 + a^{22}(wR_{1,t} + (1-w)R_{2,t})^2$ because for no definition of a^{22} it is possible to show that

$$\begin{aligned} w^2 (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + (1-w)^2 (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + 2w(1-w)(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} &= \\ = w^2 (1-w)^2 \mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + (1-w)^2 (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 + 2w(1-w)(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} \end{aligned}$$

²³It may not be obvious, notice that the formulae given above do in itself sufficient to derive that $a^{11} \in (0, 1)$ and $a^{22} \in (0, 1)$ and that you do not need to deal with the lower bound of the filtered covariance coefficient beyond a^{11}, a^{22} , and a^{33} . Just for completion, let us also consider the case of imposing that $\rho_{1,2,t} \in (-1, 1)$. The lower bound means that

$$-a^{12} + a^{12} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq \sqrt{(a^{11} + a^{11} \mathbf{R}_{t-1}^2)(a^{22} + a^{22} \mathbf{R}_{t-1}^2)}$$

$$\begin{aligned} (a^{12})^2 - (a^{12} \mathbf{R}_{t-1} \mathbf{R}_{t-1})^2 &\leq (a^{11})^2 + a^{11} \mathbf{R}_{t-1}^2 + (a^{22})^2 + a^{22} \mathbf{R}_{t-1}^2 \\ (a^{12})^2 - (a^{12})^2 \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 &\leq 2a^{11} a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq a^{11} a^{22} + a^{11} a^{22} \mathbf{R}_{t-1}^2 + a^{11} a^{22} \mathbf{R}_{t-1}^2 - a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 \end{aligned}$$

which is equivalent to

$$(a^{12})^2 - (a^{11})^2 \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + (a^{22})^2 - (a^{22})^2 \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 \geq 2a^{11} a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} - 2a^{11} a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq 0$$

which is the same condition used above.

BEKK (Baba-Engle-Kraft-Kroner) GARCH

The special "sandwich" structure of the coefficient matrices guarantees that $\Sigma_{t+1|t}$ is (semi-)PD without imposing other restrictions

The popular BEKK that many empiricists have come to appreciate is a simpler (1,1) diagonal BEKK that restricts the matrices \mathbf{A} and \mathbf{B}

BEKK models possess three attractive properties:

- When symmetry of \mathbf{A} and \mathbf{B} is imposed, a BEKK is a truncated, low-dimensional application of a theorem by which all nonnegative, symmetric $N \times N$ matrices (say, \mathbf{M}) can be decomposed as:

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \sum_{i=1}^N \begin{bmatrix} \mathbf{m}_{i,1} \mathbf{m}_{i,1} & \mathbf{m}_{i,1} \mathbf{m}_{i,2} \\ \mathbf{m}_{i,2} \mathbf{m}_{i,1} & \mathbf{m}_{i,2} \mathbf{m}_{i,2} \end{bmatrix}$$

for appropriately selected vectors $\mathbf{m}_{i,j}$.

BEKK ensures (5)PD-ness of $\Sigma_{t+1|t}$, because by construction, the sandwich form and outer vector products have this property

BEKK is invariant to linear combinations, i.e., if \mathbf{R}_{t+1} follows a BEKK GARCH(p, q), then any ptf. of the N assets in \mathbf{R}_{t+1} will also follow a BEKK. See lecture notes for examples and counterexamples under Vech ARCH

However, the number of parameters in BEKK remains rather large
Lecture 5: Multivariate GARCH and Conditional Correlation Models - Prof. Guidolin 15

At this point the filtered (predicted) correlation coefficient has expression

$$\rho_{1,2,t} = \frac{a^{22} + a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1}}{\sqrt{a^{11} + a^{11} \mathbf{R}_{t-1}^2} \sqrt{a^{22} + a^{22} \mathbf{R}_{t-1}^2}}$$

and, as it is obvious, $\rho_{1,2,t}$ should belong to $[-1, 1]$ $\forall t \geq 1$. Here we have chosen the notation $a^{11} \equiv (1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2$, $a^{22} \equiv (1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2$, and $a^{33} \equiv (1-a^{33})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2$. Focusing on the upper bound of the interval this means that

$$(a^{22} + a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1})^2 \leq (a^{11} + a^{11} \mathbf{R}_{t-1}^2)(a^{22} + a^{22} \mathbf{R}_{t-1}^2)$$

or

$$(a^{22})^2 + (a^{22})^2 \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + 2a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq a^{11} a^{22} + a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + a^{11} a^{22} \mathbf{R}_{t-1}^2 + a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2$$

which is equivalent to

$$(a^{11} a^{22} - (a^{22})^2) \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + (a^{11} a^{22} - (a^{22})^2) + a^{11} a^{22} \mathbf{R}_{t-1}^2 + a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 - 2a^{11} a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq 0$$

which cannot hold for a continuous distribution for the asset returns (even constraining $(a^{11} a^{22} - (a^{22})^2) \geq 0$ and $(a^{11} a^{22} - (a^{22})^2) \geq 0$),²⁴

$$a^{22} a^{11} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 - a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 - 2a^{11} a^{22} \mathbf{R}_{t-1} \mathbf{R}_{t-1} \geq 0$$

in general does not hold for $a^{22} = 0$. However, notice that if one sets $a^{22} = 0$, then the previous inequality simplifies to

$$a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + \left\{ \left[(1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \right] \left[(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \right] - \left[\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} \right]^2 \right\} + a^{22} a^{11} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 + a^{11} a^{22} \mathbf{R}_{t-1}^2 \mathbf{R}_{t-1}^2 \geq 0$$

which has a chance to hold if a^{11} and a^{22} are such that

$$\left[(1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \right] \left[(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2 \right] \geq \left[\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t} \right]^2$$

which also means that

$$\rho_{1,2,t} = \frac{\sigma_{1,2,t}}{\sigma_{1,1,t} \sigma_{2,2,t}} = \frac{\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t} \mathbf{R}_{i,t}}{\sqrt{(1-a^{11})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2} \sqrt{(1-a^{22})\mathbf{T}^{-1} \sum_{i=1}^p \mathbf{R}_{i,t}^2}} \geq 1$$

the unconditional correlation implied by the data and the diagonal bivariate ARCH(1) process is well-behaved. Therefore, if $a^{11} \in (0, 1)$ and $a^{22} \in (0, 1)$, then $a^{22} = 0$ (and some other restriction on a^{11} and a^{22}) must be imposed. This means that it is impossible to model the dynamics of volatilities

²⁴"At all times" has really meant "for all possible realizations of the volatilities bivariate vector \mathbf{R}_t , which can assume $(-1, \infty) \times (-\infty, 1)$ ", which alludes to the fact that even under limited responsibility, it seems most assets play at least a little role in the return.

and especially that

$$\omega^2 a^{11} R_{1,t+1}^2 + (1-\omega)^2 a^{22} R_{2,t+1}^2 + 2\omega(1-\omega)a^{21} R_{1,t+1}R_{2,t+1} = \omega^2 a^{11} R_{1,t}^2 + (1-\omega)^2 a^{22} R_{2,t}^2 + 2\omega(1-\omega)a^{21} R_{1,t}R_{2,t}$$

This means that the Diagonal multivariate ARCH model fails to be invariant to linear combinations: if you start with N assets that follow a Diagonal multivariate ARCH model, the resulting portfolio of assets will fail to follow a similar Diagonal model, which is of course problematic if not confounding. As you should be reading in the paper by Bauwens et al. (2006), the problem of (12) that causes it to fail the invariance property is very simple to visualize, while in

$$\text{vec}(\Omega_t) = (I_N - A) \text{vec} \left(I^{-1} \sum_{s=1}^T R_s R_s' \right) + A \text{vec} (R_{t-1} R_{t-1}')$$

A is diagonal, R_t^2 can be written as $(\omega(1-\omega)R_t = \omega'R_t$ and $\text{Var}_{t-1}(R_t^2) = \omega' \Omega_t \omega$ implies the need to use a vector of coefficients $\omega'A$ which is no longer a diagonal matrix (of course, it is not even a matrix).

It is also easy to see what you need to do in order for the invariance property to obtain: if you set $a^{11} = a^{22} = a^{21}$, then when $a^{21} = a^{11}$

$$\begin{aligned} & \omega^2(1-a^{21})I^{-1} \sum_{s=1}^T R_{1s}^2 + (1-\omega)^2(1-a^{21})I^{-1} \sum_{s=1}^T R_{2s}^2 + 2\omega(1-\omega)(1-a^{21})I^{-1} \sum_{s=1}^T R_{1s}R_{2s} \\ &= \omega^2(1-a^{11})I^{-1} \sum_{s=1}^T R_{1s}^2 + (1-\omega)^2(1-a^{22})I^{-1} \sum_{s=1}^T R_{2s}^2 + 2\omega(1-\omega)(1-a^{21})I^{-1} \sum_{s=1}^T R_{1s}R_{2s} \\ &= \omega^2 a^{11} R_{1,t+1}^2 + (1-\omega)^2 a^{22} R_{2,t+1}^2 + 2\omega(1-\omega)a^{21} R_{1,t+1}R_{2,t+1} \\ &= \omega^2 a^{21} R_{1,t+1}^2 + (1-\omega)^2 a^{21} R_{2,t+1}^2 + 2\omega(1-\omega)a^{21} R_{1,t+1}R_{2,t+1} \end{aligned}$$

will trivially hold. But this means that the only way for a Diagonal multivariate ARCH to possess the invariance property is for it to actually be a Scalar multivariate ARCH, in which the same ARCH coefficient applies to all conditional equations.

Question 8 (4 points)

One analyst working on your desk has been given the task to identify and estimate alternative regimes in the dynamic relationship among monthly US excess equity returns, Japanese excess equity returns, and the rate of change in the implied volatility of SPX options. The analyst has reported the following estimation output (p-values are in parentheses).

$$\begin{bmatrix} x_{t+1}^{US} \\ x_{t+1}^{Japan} \\ rVXO_{t+1} \end{bmatrix} = \begin{cases} \begin{bmatrix} -2.088 & -6.157 & -6.785 \end{bmatrix}' & \text{if } S_{t+1} = 1 \\ \begin{bmatrix} 0.958 & 0.296 & -3.780 \end{bmatrix}' & \text{if } S_{t+1} = 2 \\ \begin{bmatrix} 2.256 & 4.227 & 20.226 \end{bmatrix}' & \text{if } S_{t+1} = 3 \end{cases} + \begin{cases} \begin{bmatrix} 0.237 & 0.033 & -14.914 \\ 0.018 & 0.028 & -1.891 \\ 0.002 & 0.002 & -0.050 \end{bmatrix} & \text{if } S_{t+1} = 1 \\ \begin{bmatrix} 0.167 & 0.011 & -12.254 \\ 0.204 & 0.027 & -7.586 \\ 0.004 & 0.002 & -0.390 \end{bmatrix} & \text{if } S_{t+1} = 2 \\ \begin{bmatrix} -0.135 & 0.011 & -8.715 \\ 0.235 & 0.214 & -6.385 \end{bmatrix} & \text{if } S_{t+1} = 3 \end{cases} \cdot \begin{bmatrix} x_t^{US} \\ x_t^{Japan} \\ rVXO_t \end{bmatrix} \\ + \begin{cases} \begin{bmatrix} 3.104 & -0.154 & -0.151 \\ -2.269 & 4.746 & -0.321 \\ -5.118 & -16.64 & 10.92 \end{bmatrix} & \text{if } S_{t+1} = 1 \\ \begin{bmatrix} 2.332 & 0.312 & 0.097 \\ 3.067 & 4.215 & -0.009 \\ 2.914 & -0.489 & 12.88 \end{bmatrix} & \text{if } S_{t+1} = 2 \\ \begin{bmatrix} 3.661 & 0.118 & -0.335 \\ 2.924 & 6.768 & -0.468 \\ -36.45 & -94.14 & 29.72 \end{bmatrix} & \text{if } S_{t+1} = 3 \end{cases} \cdot \begin{bmatrix} z_{t+1}^{US} \\ z_{t+1}^{Japan} \\ rVXO_{t+1} \end{bmatrix} \quad \hat{P} = \begin{bmatrix} 0.397 & 0.055 & 0.548 \\ 0.000 & 0.841 & 0.159 \\ 0.392 & 0.340 & 0.268 \end{bmatrix}$$

How many regimes did he specify for this model? Are the regimes persistent? What is the duration of each of the regimes? Does the past rate of growth of implied volatility forecast

positive or negative excess US returns, and when? After taking such vector autoregressive structure into account, is the correlation between shocks to US excess equity returns and the rate of growth of implied volatility positive or negative, and when? Please justify your answers with reference to the estimation outputs provided.

Debriefing.

The analyst has specified three Markov regimes (states), as one can see from the fact that the estimated transition matrix is a 3x3 one. Only one regime, the second, is persistent, in the sense that $\Pr(S_{t+1} = 2|S_t = 2) = 0.814 > 0.5$; the other two regimes are non persistent, in the sense that $\Pr(S_{t+1} = j|S_t = j) < 0.5$ for $j = 1$ and 3. The durations of the three regimes are $(1 - 0.397)^{-1} = 1.658$, $(1 - 0.841)^{-1} = 6.289$, and $(1 - 0.268)^{-1} = 1.366$ months. Based on the p-values reported, the past rate of growth of implied volatility forecasts negative excess US returns (coefficients of -14.91 and -12.25, with zero p-values) in regimes 1 and 2, while the predictive power is more doubtful in regime 3 (coefficient of -8.72, but with a p-value exceeding 0.10). Net of the VAR effects, the correlation between shocks to US excess equity returns and the rate of growth of implied volatility is negative and rather sizable in regimes 1 and 3 (correlations are -0.15 and -0.34), and positive but smaller (0.10) in regime 2.