Economic and Financial Aspects of Crude Oil Markets

Gabriele D'Amore

Department of Methods and Models for Economics, Territory and Finance

Sapienza University of Rome

A thesis submitted for the degree of

Doctor of Philosophy in International Economics and Finance

8-6-2017

I would like to dedicate this thesis to my uncle Antonio D'Amore without whom this thesis would not have been possible. He is a strong proponent of culture and intellectual curiosity and influenced me very much. For all his constant support and affection throughout this thesis process, I can never thank him enough. I also would like to dedicate this work to my beloved girlfriend Emanuela Bruno whose love, patience, constant encouragement and loyal support were crucial to keeping me going.

The Role of Market Speculation in Rising Oil Prices

Contents

European Energy Security: the Substitutability of European Crude Oil Imports from Russia

Contents

Predictability Information Criterion for Selecting Stochastic Pricing Models

Contents

The Role of Market Speculation in Rising Oil Prices

Gabriele D'Amore*

May 8, 2017

Abstract

The large oil price fluctuations occurred from 2003 to 2008 has raised many questions about their causes. Many empirical studies have attempted to understand how oil price fluctuations are driven by changes in both market fundamentals and speculative pressures. In this regard, some problems arise such as: the use of unreliable data like the global level of inventories or the specification of a vast number of arbitrary restrictions for the models. In this study I try to isolate, coherently with the view of Knittel and Pindyck(2016) [34] and inspired by Kilian's works, the speculative effect on the short-term spot price fluctuations determined by structural forward-looking behavioural shocks produced in the futures market. Exploiting a dataset used in Kilian and Murphy (2012)[31], CFTC data (period 1999M1-2008M6) and taking advantage of the standard theory of storage we will be able to verify, with a Blanchard-Quah structural approach, that the impact of these shocks is remarkable but not the prevalent one in magnitude. Instead, it would seem that speculative inventory holdings may have played a much more important role.

Keywords: Oil price, Financialization, Short Hedging, Open Interest, Standard Theory of Storage, SVAR.
JEL-Classification: C2 C32 Q40 Q43.

1 Introduction

The price of oil is showing strong fluctuations since 2003 when a sudden surge in oil prices, apparently inexplicable, has led the price of WTI at its historical peak in July 2008 at 145.31 dollars per barrel before falling in December of the same year at 30.28 dollars per barrel. The same type of behaviour is observable even later causing a drop of 75% in crude oil prices from June 2014 to January 2016.
The study of the hidden economic reasons determining these fluctuations is a major objective of the international policy makers. (See for example HM Government (2010) [21]; EU Commission (2008) [13]; Lieberman, 2008 [38]). More

*Sapienza University of Rome. Mail to: gabriele.damore@uniroma1.it. Corresponding author at: Department of Economics and Social Sciences, Piazzale AldoMoro, 5 - 00185 Rome(IT).

generally, the fluctuations in commodity prices, pose serious issues to the nations.
Usually, countries are affected by two types of opposing effects on the purchasing power: effects for the exporting countries, which suffer from the falling prices, causing the drop of the national gross domestic product; effects for the importing countries, which suffer from the generated inflation, to a greater extent based on the degree of poverty of the involved nation. However, the high complexity of the international price formation mechanism and the large number of actors involved, make the study of these fluctuations a difficult task.
Numerous studies have attempted to pursue this objective. Very often the literature offers econometric exploratory models, sometimes supported by well-defined economic theory. (see Alquist et al., (2011) [1]; Baumeister et al., (2010) [6]; Bencivenga et al., (2012) [7]; Juvenal and Petrella, (2014)[27]; Kilian, (2006) [29]; Kilian and Murphy (2012)[31](2014)[32]; Lippi and Nobili, (2012) [39]; D'Ecclesia et al.(2014)) [15].
These studies are focused on the impact on the price determined by some fundamental drivers.
An interesting line of research is based on the hypothesis that the observed unusual price dynamic from 2003 is due to the financialization of crude oil market. In fact, from 2003 the dynamic evolution of the price seems quite different compared to the past.
Curiously during the same period, money managers were found to be particularly active on the commodity exchange using crude oil as any other asset class (Turner et al., 2011 [48]). Over ten years, the Global Asset Under Managment (AUM) passed from 10 million to 400 million dollars, since 2001, with half of the contracts managed using short dates or nearby futures contracts representing around 50% in open interest.(D'Ecclesia et al.(2014)[15])
An important aspect studied in the literature is the measure of the effect of speculation on the spot price of crude oil.
Generally speaking, three factors need to be accurately studied: 1) the Supply, 2) the Fundamental Demand 3) The Speculative Demand. The first of these factors can be easily analysed, as there are available observable data, while for both the second and the third factor, there is no single, universally acceptable measure.
Notice that fundamental demand is intrinsically linked to the economic cycle, while the speculative demand, is generated by a plurality of actors with unknown targets which make it the most difficult factor to analyse. For instance, a speculative behaviour can be observed both in producers and refineries companies, that accumulate reserves when the selling price is lower than expected and vice versa, the remaining part is generated by professional speculators who trade futures contracts.
In my opinion, to understand the phenomenon, it is necessary to analyse the behaviour of different players, which operate according to different preferences. I believe that different long-term goals expressed by the various agents involved in the market lead to completely different short-term effects on price.
Following the classification provided by the Commodities Future Trading Commission (CFTC), we will distinguish two categories of traders: 1)"commercial" and 2) "non-commercial". Thanks to this classification we will study and isolate the impact of speculation on short-term dynamics of the crude oil price between 1999 and 2008, mainly inspired by Kilian's works and studies on speculative

activity conducted by Knittel and Pindyck(2016) [34]. I will present three major results: 1) differently to previous literature, we will be able to empirically measure the impact that diversifiable risk and upward expectations, of the non-commercials (speculators), have on the real spot price fluctuations; 2) we will show that the non-commercial upward expectations impact significantly on the short-term deviations in oil prices, furthermore; 3) by jointly studying the spot market and the futures market we will be able to implicitly distinguish the effect of forward-looking operations on the spot market.
In the next section, it has reviewed the literature on the impact of speculation on commodity prices. The literature often disagrees about the actual role of speculation in affecting oil prices.

2 Literary Review

In recent years about four explanations were provided by the literature to clarify the reasons for the sudden increase in the crude oil price of the first decade of the new millennium:
1) The first hypothesis concerns the finiteness of crude oil that causes a boundary limit to the increase of the production capacity ("peak oil hypothesis");
2) The second hypothesis is based on the growing demand from emerging countries such as China and India ("demand growth hypothesis");
3) The third hypothesis considers speculators as those responsible for an altered dynamic of prices due to the strengthening of a forward-looking demand ("speculation hypothesis");
4) the fourth hypothesis claims that the growth of the market liquidity determines the increase in the crude oil demand("excess liquidity hypothesis").
Singleton (2011) [44], and Hamilton and Wu (2011)[24] are in favor of the speculation hypothesis. Hamilton (2009) [23] provided an overview of possible causes about the changes in oil prices and concluded that speculation played a role in the increase in prices in the summer of 2008. Smith (2009) do not find evidence that the speculation has raised prices between 2004 and 2008. Kaufmann(2011)[28] is one of the few researchers favourable to the peak hypothesis. Kilian (2006) [29], Kilian and Murphy (2010))[31], Kilian and Hicks (2012)[30], Krugman (2008) [37] and Hamilton (2009)[22] [23], Dees et al.(2007,2008)[16] [17] find strong evidence that prices are determined by the growth of demand against a stable supply. Some of the most important theoretical and empirical works are B Fattouh, L Kilian, L Mahadeva (2012)[18] and Kilian (2006)[29], Kilian and Murphy (2012)[31], and Baumeister and Kilian (2012)[5], Alquist and Kilian (2010)[1], Hamilton (2009)[23], Smith (2009)[45], Knittel e Pindyck (2016)[34]. Many of these works generalise the crude oil market model proposed by Kilian (2006) [29] to examine the role of speculation and forward-looking behaviour with a focus on its role in spot and futures prices. Christopher R. Knittel and Robert S. Pindyck (2016) [34]identify two channels whereby speculators may influence the spot price of any storable commodity: 1) Speculation via the Futures Market. 2) Speculative Inventory Holdings.

3 Modelling the real price of oil

3.1 The structural approach in literature

The literature of both empirical and theoretical studies has often used oil prices to evaluate the variation of economic aggregates, considering such prices as exogenously given. However, it is still not totally clear what are the directions of dependency among economic aggregates and crude oil. Kilan (2006)[29] addressed this problem proposing a structural way to analyse crude oil prices inside of a system of endogenous variables. Through a simple but powerful model, he was able to study the linkage between oil price fluctuations and other fundamental variables by simply extracting the underlying structural shocks of the system.
He showed that the fluctuations of the global spot price of crude oil were due to the nature of underlying shocks and that forward-looking demand shock is the dominant factor. This finding has been considered sufficient to explain how the global economy was able to avoid recession during a period of high oil prices.
The fundamental assumptions of the Kilian's base model are: 1) zero short run price elasticity of supply of crude oil; 2) the introduction of a new measure of global real economic activity based on the global index of dry cargo single voyage freight rates maturate at a monthly frequency. This last monthly measure was introduced on the assumption that the world economic business cycle reverberates its fluctuations in this measure, being them historically positively correlated.(see, e.g., Isserlis 1938[26], Tinbergen, 1959[47], Stopford 1997[46] Klovand 2004[33]). The advantage of using this measure is it does not require exchange-rate based weights to be calculated because implicitly it already incorporates shifting country weights. Moreover, it automatically aggregates real economic activity in all countries (including, e.g., China, India). 3) Kilian interpreted any change in the spot price, not determined by structural shocks of supply and demand, as due to the expectations of the agents operating in the market (forward-looking view)(e.g.. New discoveries such as new off-shore oil fields, an expected global financial crisis or the uncertainty about future oil supply shortfalls, anticipation of a War,...).
These variations, not explained by fundamentals, are interpretable, in some sense, as the effect of the speculative behaviour of rational players.
This initial model has been continuously improved over time. The improvements mainly concern: 1) the introduction of new variables in the model; 2) the estimation of the VAR model imposing exclusion restrictions on the impact multiplier matrix (following Faust (1998)[19], Canova and De Nicolò (2002)[10], Uhlig (2005)[49]); 3) bounds on the magnitude of the short-run oil supply elasticities and the sign of the impact response (Kilian and Murphy (2012)[31]). They demonstrated that if only the sign restrictions are used, some elasticities become unreasonably large.
The increasing of crude oil market size and the contemporary abnormal behaviour of prices have led many researchers to investigate the connection between the spot price of oil and speculation. Speculation can involve both the purchase or sale of crude oil contracts for physical storage, which determines the accumulation of inventories and transactions of futures contracts. (See Büyükşahin and Harris (2011)[9]). Christopher R. Knittel and Robert S. Pindyck (2016) [34] have identified two channels whereby speculators can influ-

ence the spot price of any storable commodity: 1) speculation via the futures market; 2) speculative inventory holdings.
Since a considerable part of the speculative operations occurs on futures markets, it would be essential to clarify how futures prices can affect spot prices. Coherently with economic theory (see Giannone and Reichlin (2006)[20]), oil futures spread should not have predictive power.
Fattouh, Kilian and Mahadeva (2012)[18], argue that there is little evidence to support speculation in pushing the spot price of oil after 2003 and they support the thesis that spot and futures prices reflect common economic fundamentals.

3.2 The proposed approach

The approach I propose builds on the first family of structural models proposed by Kilian (2006) [29], for the study of the world market of crude oil. I expand the original system of endogenous variables:

- s_t^{det} the detrended world's crude oil production;
- g_t^{det} a detrended index of real economic activity representing the global business cycle;
- p_t^{det} the index of the detrended deflated crude oil prices calculated on the base of U.S. refiners' acquisition cost;

representing the spot market of crude oil, adding some key variables to the system. In order to build an empirical model that can extract the impact that the speculator's upward expectation have had on oil spot price fluctuations, we need the availability of: 1) data broken down by type of traders (hedgers or speculators); 2) one or more proxies of speculators upward expectations, being not possible to directly observe them.
To this end, I propose to include in the model the following additional variables observable in the futures market:

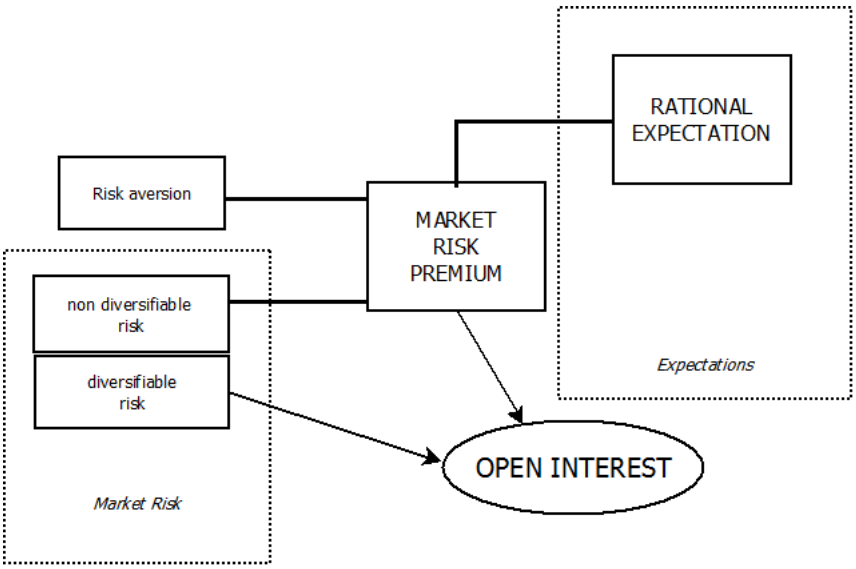
- $\triangle OI_t^{hedg}$ the percentage change in futures open interest for short positions held by commercial operators (hedgers).
- $\triangle OI_t^{spec}$ the percentage change in futures open interest for long positions held by non-commercial operators (speculators).

There are many reasons for this choice:

a. hedgers and speculators are two types of traders that usually are each other's counterpart (short hedging operations). This affects the entry or exit from the market of both and it induces endogeneity. Therefore, for the purpose of estimation of structural shocks, it is preferable to include the open interest of both traders in our system of *endogenous* variables.

b. We decide to draw disaggregated open interest data from futures market because they are naturally connected to the expectations of future spot prices. In fact, the futures market is widely used by operators to manage market risk on spot price or to speculate from their upward or downward expectations of the spot price.

Figure 1: Futures open interest for short positions held by commercial operators (hedgers)



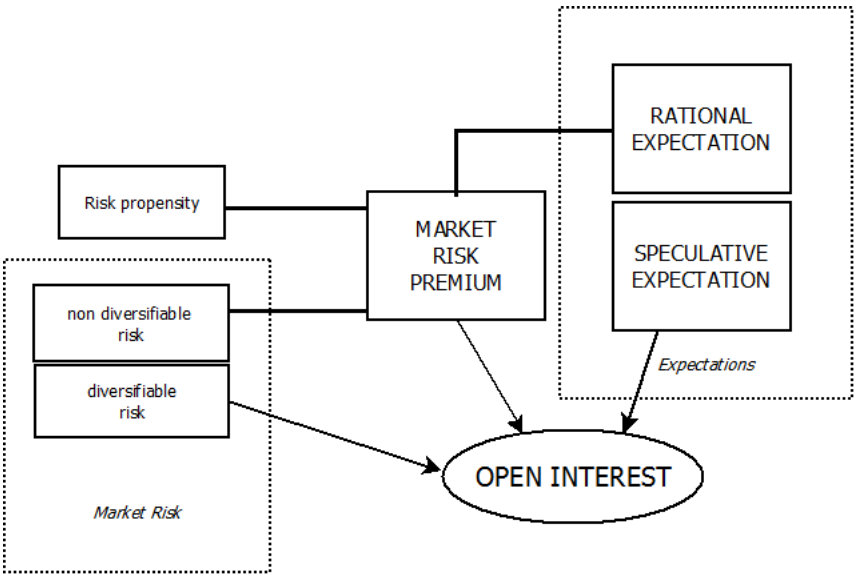
c. Disaggregated futures open interest, by type of transaction (long and short) and traders (hedgers and speculator), helps us to separately study the expectations characterising hedgers and speculators in the futures market. In fact, the open interest indicates the flow of money into futures market, for both long and short selling operations, or in simple terms, how the futures market is, in a particular moment, relevant to the trading activities of a multitude of operators:

   the *hedgers* sell futures to protect a portfolio of activities when they have a rational expectation of a decline in the spot price. We reasonably guess that the higher the market risk, (or the greater the risk aversion) and the higher the open interest on short operations of those operators. In this sense, the futures open interest of the hedger's short operations contains information on rational expectations, the hedgers' risk aversion and the market risk. (See figure 1)

   the *speculators* buy futures to speculate on price when there is an upward expectation. We reasonably expect that the higher the risk of the market, or the higher the speculators risk attitude and the higher the open interest on long operations. In this sense, the futures open interest of speculator's long operations contains information on speculators upward expectations, the speculators' attitude to risk (which can be analysed in terms of risk premium according to Keynesian theory of backwardation) and the market risk. (See figure 2)

the strategy I propose, to extrapolate the effect of speculators upward expecta-

Figure 2: Futures open interest for long positions held by non-commercial operators (speculators)



tion on the spot price, requires to disentangle the expectations in: 1) the market's rational expectation and 2) the speculative expectation of non-commercials (speculators) making use of the concept of:

- Market risk premium $\pi_t^{Mkt}{}_T$ with the *theory of normal backwardation*;

  Keynes (1930) theory of normal backwardation argues that *if producers of a commodity (the so-called commercial traders) want to hedge the market risk, they may want to do it by selling futures contracts. The counterpart (arbitrageurs) may be compensated for assuming that risk in the form of an appropriate futures price. This kind of compensation is called market risk premium.*

If we assume:

- the theory of normal backwardation supposing to be valid for the long run;
- no effect on spot price are due to speculative expectations in the long run;

the consequences are the following:
speculators have the incentive to take and hold the long position, in the futures market, to get the risk premium they expect, that does not necessarily coincide with the market risk premium. To put it another way, they speculate on the expected spot price. However in the long run market risk premium cannot be affected by speculators expectations, as it is assumed that in the long run spot price is not affected by speculative expectations.(For further details read section 4.1).

hedgers the higher will be the market risk premium the higher will be the incentive to cover the market risk taking and holding a short position and vice versa.
Notice that,

- in this context of long-run expectations the market risk premium is only affected by non-diversifiable risk.

In this regard, Hamilton (2014) [24] argues that: "*empirical support for this view has come from Carter, Rausser, and Schmitz (1983)[11], Chang (1985) [12], and De Roon, Nijman, and Veld (2000) [14], who interpreted the compensation as arising from the non-diversifiable component of commodity price risk.*"
Other differences that characterise the proposed approach compared to that of Kilian's are: 1)we impose restrictions on the long-term impacts matrix, instead of short-term impacts matrix. This will allow us to take advantage of the popular theory of storage; 2) According to Baumeister, Christiane, and Gert Peersman (2009) [6]we assume that the increase of hedging possibilities, provided by the growth of crude oil market size, is reducing the long-term price elasticity of supply.
Reasoned arguments in support of the methodology are provided below.

3.2.1 The Standard Theory of Storage

According to the standard theory of storage and the non-arbitrage principle, it can be shown that there is a long-run relationship between the price of crude oil and a number of variables, including the price of a futures contract (see (5)). This condition allows us to state that any speculation on the futures markets should immediately influence the spot price of crude oil in the long run.
Kilian (2006) [29] implicitly consider this relation for avoiding to include the price of futures contracts inside the model. However, it must be stressed that this connection is valid only *in the long term* being based on the non-arbitrage principle. Consequently, we couldn't be sure that the price of futures contracts likewise affect the short-term spot price.
From this observation, we believe that exploiting this relation in the long run framework is somehow fundamental to study the effect of speculation in the futures markets on the crude oil spot price.

3.2.2 Data

Oil prices are not the only instrument useful for assessing the speculation in the futures markets. A much cleaner signal of the speculative activity is provided by the CFTC, with data and periodic reports covering the positions of broad categories of traders, such as "commercial" and "non-commercial" traders.

3.2.3 The Elasticity Hypotheses

Kilian and Murphy(2014)[32] proposed to replace the so far adopted Kilian's hypothesis of zero short-run oil supply elasticity, considered "a good approximation but unlikely to be literally correct", with the imposition of a set of: 1)

restrictions on the sign of the contemporaneous impact matrix; 2) bonds on the oil supply elasticity.
We do not believe that this approach can be convincing for two reasons:
1)(the arbitrary nature of the constraints) the introduction of this methodology leads to a greater number of boundary constraints we cannot always test.
Kilian and Murphy's [31] demonstrated that sign restrictions alone, on the parameters of the model, are not sufficient to solve the drawback. Moreover, the imposition of bounds on the elasticity of supply does not guarantee the definition of a more accurate model. In fact, such bounds are arbitrarily estimated [1] by the authors, considering some specific historical episodes such as the outbreak of the Persian Gulf War on August 2, 1990.
2)(The effect of hedging activities) The growing phenomenon of financialization of the crude oil market is probably reducing the degree of price elasticity of both the demand and supply. The players (both consumers and producers) that are making use of these instruments for hedging are increasingly being insensitive to the spot prices. Baumeister, Christiane, and Gert Peersman (2009) [6] argued that "*opportunities for hedging could decrease the sensitivity of commercial dealers to oil price fluctuations in the spot market, contributing to less elastic oil supply and demand curves. The reduced price elasticities of supply and demand result in increased oil price volatility which further encourages the development of a market for derivatives*". According to Krichene (2005) [36], the massive estimated drop in long-run price elasticity of supply to 0.25 in 1973-2004 from 0.46 in 1918-73, shows a change from a competitive to a market-maker structure, which supports the hypothesis of Baumeister, Christiane and Gert Peersman. However, the dedicated literature often has provided countless contradictory estimates of the long-run price elasticity of both supply and demand depending on the time horizon and the reference market. An increasing number of works supports the idea of a very low long-run price elasticity of supply, for instance (as summarized by Naoyuki Yoshino, Farhad Taghizadeh-Hesary (2015)[50]) Krichene (2002)[35] computed the short run price elasticity from -0.08 to 0.08 and long-run price elasticity of 0,10-1.10 for the OPEC members from 1918-1999. Ramcharran (2002)[43] obtained negative and significant price elasticity for 7 of the 11 OPEC members. Askari and Krichene (2010)[2] got a short run price elasticity from -0.48 to 0.660 from 1970(Q1)-2008(Q4) and long-run price elasticity of -0,02-0.008. Naoyuki Yoshino, Farhad Taghizadeh-Hesary (2016)[50] computed a low long-run price elasticity 0.03.
Moreover, oil production is increasingly getting influenced by technological factors and oil disclosures.(e.g. fracking and shell oil revolution).
Furthermore, the price changes that occurred at the turn of 2014 and 2015, with a vertical fall in prices, have for several experts revealed that OPEC has had lost its power to influence global production of oil and thus to affect the prices in the long run.
All of these reasons let us suppose that a zero long-run price elasticity of supply as a good approximation for our extent.

[1] the authors call them empirically plausible bounds

4 The Spot Market

To describe the spot price formation process adequately, in the long run, we need first to explain the connection between supply and demand of crude oil in the long run.
We assume that spot price is affected by three structural shocks: 1) ϵ_t^s; 2) ϵ_t^d; 3) ϵ_t^{eco}.
According to D'Ecclesia et al.(2014) [15], Kilian (2006)[29], He et al. (2010)[25], and Dees et al. (2008)[17], among others, we employ on the demand side the following demand equation:

$$d_t = \alpha_{0\,t} + \alpha_1 p_t + \alpha_2 g_t + \alpha_3 \epsilon_t^d \tag{1}$$

where:

- d_t is the global demand of crude oil;
- p_t is the real oil price;
- g_t is a proxy for the global economic activity;
- ϵ_t^d is the structural shock of demand. It's the unexpected demand that cannot be explained by the economic activity (crude oil inventory holdings);

with $\alpha_1 < 0$, $\alpha_2 > 0$ and $\alpha_3 > 0$.
On the supply side, we propose the following supply equation that is supposed to be inelastic to price, for the aforementioned reasons:

$$s_t = \beta_{0\,t} + \beta_1 \epsilon_t^s$$

with $\beta_1 > 0$, where:

- s_t is the supply;
- ϵ_t^s is the structural shock of supply supposed to be, according to Kilian (2006)[29], mainly determined by technological factors and oil disclosures;

According to Kilian (2006)[29] we suppose the global economic activity g_t to depend on the crude oil provided by the market.

$$g_t = \gamma_{0\,t} + \gamma_1 s_t + \gamma_2 \epsilon_t^{eco}$$

where:

- ϵ_t^{eco} is the structural shock of the economy;

with $\gamma_1 > 0$ and $\gamma_2 > 0$.

In equilibrium, we derive the following reduced form for the long-term real oil price:

$$p_t^* = \frac{\beta_1}{\alpha_1}\epsilon_t^s - \frac{\alpha_3}{\alpha_1}\epsilon_t^d - \frac{\alpha_2}{\alpha_1}\gamma_0 - \frac{\alpha_2}{\alpha_1}\gamma_1\beta_1\epsilon_t^s - \frac{\alpha_2}{\alpha_1}\gamma_1\beta_0 - \frac{\alpha_2}{\alpha_1}\gamma_2\epsilon_t^{eco} - \frac{\alpha_0}{\alpha_1} + \frac{\beta_0}{\alpha_1}$$

detrending all the variables, we get the following system of equations

$$s_t^{det} = \beta_1\epsilon_t^s \tag{2}$$

$$g_t^{det} = \gamma_1\beta_1\epsilon_t^s + \gamma_2\epsilon_t^{eco} \tag{3}$$

$$p_t^{*\,det} = \beta_1\left(\frac{1}{\alpha_1} - \frac{\alpha_2}{\alpha_1}\right)\epsilon_t^s - \frac{\alpha_2}{\alpha_1}\gamma_2\epsilon_t^{eco} - \frac{\alpha_3}{\alpha_1}\epsilon_t^d \tag{4}$$

These equations, describing the real spot market, are three of the five equations that we are going to introduce in the empirical model. To derive the remaining two we need to study the forward-looking behaviours of the traders on the Futures market.

4.1 The Observability of the forward-looking behaviour

How can we study the effect of shifts in expectations on the spot price in the short run?

The economic theory argues that spot prices are affected by expectations. There are two main channels where market expectations manifest themselves: 1) the level of inventories; 2) the price of the futures markets. In theory, both phenomena are directly observable. For instance, the changes in demand, for above-ground crude oil inventories (see Kilian and Murphy (2014)[32], Alquist and Kilian (2010)[1], Christopher R. Knittel and Robert S. Pindyck (2016) [34]) should permit us to observe changes in the expectations of the demand and supply of crude oil. However, as explained in "Alquist and Kilian (2010)[1], unfortunately, "there are reasons to be skeptical of the reliability of global oil inventory data, especially in recent years". On the contrary, the futures market can provide much more reliable data in this sense.

Can we measure the expectations of speculators trading in the futures market?

The storage standard theory implies the existence of a long run relationship ensuring that a change in "expectations", about future conditions of supply and demand of crude oil, is automatically reflected in the spot market trough a change in futures prices or of the convenience yield. (see Alquist and Kilian (2010)[1]).

$$p_t^* = \frac{1}{1+r_T}\cdot[F_{t,T} + \psi_{t,T} - K_T] \tag{5}$$

where:

- p_t^* is the spot price at time t
- $F_{t,T}$ is the future price at time t with delivery time t+T
- K_T constant unit storage cost until time t+T

- $\psi_{t,T}$ is the marginal convenience yield or price of the storage
- r_T it's a constant discount rate

it would be interesting, in order to improve the adequacy of the model, to exploit somehow this theoretical result for studying the behaviour of the speculators in the short run when there is a change in the expectations of the long-run price. Christopher R. Knittel and Robert S. Pindyck (2016)[34] argue that there are essentially two prevailing channels for speculation: 1) speculation via the futures market and 2)speculation via inventory accumulation. Splitting the price formation process into two markets: the cash market[2] and the market for storage[3] they were able to show that speculative activity, regardless of the channel where it manifests itself, influences the change in spot prices and the level of inventories only in the short run. Any changes in long-term prices are only due to actual changes in demand and supply in the two markets. Therefore speculators can only swing the spot and futures price around the equilibrium price regardless the correct anticipation of the spot price change over time. They argue that such a *short-run* effect depends on the temporary shift of market expectation, determined by speculators', from the rational expectation.

$$\mathbb{E}_t\left(p^*_{t+T}\right) = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) + s_{t\,T} \tag{6}$$

- $s_{t\,T}$ shift in the expectation due to the speculators' wrong[4] expectations.
- $\mathbb{E}_t\left(\bar{p}^*_{t+T}\right)$ is the expected future spot price under rational expectations. It changes when fundamentals are expected to change.

While in the *long run* market does not expect any shift from the equilibrium spot price.

$$\mathbb{E}_t\left(p^*_{t+T}\right) = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) \tag{7}$$

In order to study the effect of the speculators expectations we propose to disentangle the expectations of two categories of traders: 1) the rational expectation of the commercials (hedgers) and 2) the speculative expectation of non-commercials (speculators), and estimate their short-run effects on the spot price by studying their positions on the market and their propensity to the risk.
To this end, we propose a definition of market risk premium in the period between t and t+T as the difference between the market expectation at time t of the spot price at time t+T and the future price at time t with delivery time t+T.

$$\pi^{Mkt}_{t\,T} = \mathbb{E}_t\left(p^*_{t+T}\right) - F_{t,T} \tag{8}$$

[2]The cash market is the market where purchases and sales for immediate delivery occur at the "spot price." The spot price does not equate production (including imports) and consumption (which might include exports) in the short-run. This misalignment of the spot price determines the change in inventories

[3]The market for storage is the market where the equilibrium level of inventory is determined.The price of storage is the cost for the privilege of holding a unit of inventory which is equal, in a competitive market like this, to the marginal value of the good or service (here simply called marginal convenience yield $\psi_{t,T}$).

[4]According to Knittel and Pindyck (2016) [34]we define wrong expectation as whatever expectation different from the rational expectation which is not followed by a change in fundamentals.

Notice that: under definitions 6 7 the market risk is affected by speculators expectations only in the short run.
Despite the fact that speculators expect a risk premium guessed on the base of their expectations

$$\pi_{t\,T}^{spec} = \mathbb{E}_t^{spec}\left(p_{t+T}^*\right) - F_{t,T}$$

where

- $\mathbb{E}_t^{spec}\left(p_{t+T}^*\right)$ is the speculator expectation.

in the long run, they can only get the market risk premium $\pi_{t\,T}^{Mkt}$.

5 The Futures Market

A Futures market is a place where participants buy and sell commodity/future contracts for delivering a commodity on a specified future date. Traditionally we define two categories of participants: hedgers and speculators. The *hedgers* are players that primarily enter into futures contracts to offset a risk exposure on the spot price, on the contrary, *speculators* are basically risk seekers and trade betting on the expectation of the future price of the futures contracts.
Both kinds of agents *"open a position"* in the futures market looking somehow at the risk and their risk propensity. However they *close positions* for different reasons: the hedgers, being interested to hedge against market risk on the spot price, decide depending on the duration and sustainability of this risk whereas the speculators decide to close a futures contract depending on their expectations and their propensity risk. We propose to analyse these decisions in an empirical model introducing detailed data referring to the open interest of hedgers and speculators.

5.1 Short Hedging operations in the period 2003-2008

If the speculators in the short term have had a role in the rise in prices, occurred between 2003 and 2008, it would mean that operations such as short hedging could have been the vehicle whereby speculators determined a distortion of the market price letting it grow out as described by Knittel and Pindyck (2016)[34].
The short hedge transactions are transactions carried out by hedgers who want to insure an inventory against the risk of falling spot prices. In this case, the hedger wants to sell a futures contract at a price $F_{t,T}$ on the market and keeps the position as long as a risk on the spot price will be counterbalanced. This opportunity is permitted as long as there is someone on the market that is willing to buy the futures contract at price $F_{t,T}$.
This kind of trader is usually a speculator having upward expectations.

$$\mathbb{E}_t^{spec}(p_{t+T}^*) > F_{t,T}$$

where $\mathbb{E}_t^{spec}\left(p_{t+T}^*\right)$ is the speculator's expectation at time t.
In this case, the speculator wants to buy a futures contract, at price $F_{t,T}$ on the market, and to keep the position as long as he is willing to bet on growing prices. The proposed empirical model, we are going to explain next section, let

us study how speculators' upward expectations [5] and the average risk propensity of the market affect oil price.

5.2 Open interest in the long run

The open interest is defined as the total number of long and short contracts that are held[6] by market participants at the end of each day. It is also defined as a measure of the flow of money into the futures market.
On the basis of what has been explained in the section 3.2, disaggregated futures open interest, by type of transaction (long and short) and traders (hedgers and speculator) is a source of useful information to characterise the futures market in an empirical model, as it is intrinsically connected to the concept of:

- risk propensity;
- market risk
- expectations;

which, as we already saw in section 3.2, influences the behaviour of hedgers and speculators in the futures market.
For this reason, making use of formula (20), according to Baker and Routledge(2011)[3][7] we approximate the open interest by a continuously differentiable function g.

We distinguish a different functional for each type of trader.
For the hedger, the open interest OI^{hedg} is defined as follows

$$OI^{hedg}_{t\,T} = g^{hedg}\left[\pi_{t\,T}, \Delta\sigma^2_{t\,T}\right] \tag{9}$$

for the speculator OI^{spec} is

$$OI^{spec}_{t\,T} = g^{spec}\left[\pi^{spec}_{t\,T}, \Delta\sigma^2_{t\,T}\right] \tag{10}$$

on the base of definition (6) and (20) we get the following expression (see proof 1 in the appendix):

$$OI^{spec}_{t\,T} = g^{spec}\left[\pi_{t\,T}, \Delta\sigma^2_{t\,T}, s_{t\,T}\right] \tag{11}$$

where:

- $\Delta\sigma^2_T$, is the residual risk (diversifiable risk);
- $\pi^{spec}_{t\,T}$, The risk premium for the speculator is represented by the premium[8] associated to his own expectation, where $\pi^{spec}_{t\,T} = \mathbb{E}_t\left(p^*_{t+T}\right) - F_{t,T}$. (See equation (6)).

[5] We are confident that this shock is inversely correlated respectively to the downward expectations shocks of speculators

[6] We define the held contracts as purchased futures contracts which are and not already closed into the market.

[7] Baker and Routledge (2011)[3] show that changes in open interest and risk premium on oil futures arise endogenously as a result of heterogeneous risk-aversion.

[8] The risk premium is the premium required for buying a futures contract.

- $\pi_{t\ T}$, is the risk premium associated to rational expectation, where $\pi_{t\ T} = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) - F_{t,T}$;
- $\mathbb{E}_t^{spec}\left(p^*_{t+T}\right)$ is the speculator expectation;
- $\mathbb{E}_t\left(\bar{p}^*_{t+T}\right)$ is the rational expectation of future price;
- $s_{t\ T} = \mathbb{E}_t^{spec}\left(p^*_{t+T}\right) - \mathbb{E}_t\left(\bar{p}^*_{t+T}\right)$ is the shift in the market expectation due to the speculator expectation. These expectations affect the demand in the futures market determining price fluctuation (Knittel and Pindyck (2016) [34]).

Notice that, the hedger does not use own expectations, as the speculator does, because the hedger does not bet on the future price but does only care to rationally offset the risk on the spot market.
Before we proceed to build the empirical model, we need to make some guesses we are going to check out after the calibration of the model. For example, guesses at how open interest theoretically reacts to a change in the risk aversion of the traders (see Baker and Routledge (2011)[3]). In order to do that, we proxy the risk aversion (or risk propensity) with the above mentioned risk premiums. We suppose that: the higher the risk aversion (propensity) of the hedgers, the greater (lower) the recourse to the futures market by the hedger and thus the greater(lower) the flow of money into the futures market (open interest) $\frac{\partial OI^{hedge}_{t\ T}}{\partial \pi_{t\ T}} \gtrless 0$. Viceversa, the higher the risk propensity of the speculators, the higher the recourse to the futures market and thus the higher the flow of money into the futures market (open interest) $\frac{\partial OI^{spec}_{t\ T}}{\partial \pi^{spec}_{t\ T}} > 0$. Notice that speculators are compensated for the assumption of the non-diversifiable risk [9]. The growing risk is instead an incentive for both agents to operate on the futures market although for different reasons. With the increase of residual risk, the hedger faces higher risk and therefore opens a futures contract $\frac{\partial OI^{hedg}_{t\ T}}{\partial \Delta\sigma^2} \geq 0$. At the same time risk lover speculators takes positions for the opposite reason $\frac{\partial OI^{spec}_{t\ T}}{\partial \Delta\sigma^2} \geq 0$. Moreover, the expected growth of the price entails the increase of the expected risk premium by the speculator and thus it acts as an incentive to enter or to stay in the futures market $\frac{\partial OI^{spec}_{t\ T}}{\partial \mathbb{E}^{spec}_t\left(p^*_{t+T}\right)} \geq 0$.

5.2.1 Open interest and risk premium in the short run

We assume that hedgers can be affected by speculators expectations, in the short run, since they believe that the expected risk premium is the market risk premium.

$$\frac{\partial OI^{hedge}_{t\ T}}{\partial p^*_t} = \frac{\partial OI^{hedge}_{t\ T}}{\partial \pi^{Mkt}_{t\ T}} \frac{\partial \pi^{Mkt}_{t\ T}}{\partial p^*_t} \tag{12}$$

$$\frac{\partial OI^{spec}_{t\ T}}{\partial p^*_t} = \frac{\partial OI^{spec}_{t\ T}}{\partial \pi^{spec}_{t\ T}} \frac{\partial \pi^{spec}_{t\ T}}{\partial p^*_t} \tag{13}$$

[9] Economic theory states that if traders use futures contracts to hedge against commodity price risk, the agent who takes the other side of the contracts (called arbitrageur) may receive compensation for accepting nondiversifiable risk in the form of positive expected returns from their positions (see Hamilton and Wu (2014))[24]

This consideration let us imagine a short run hedgers open interest affected by the shift s

$$OI_{t\,T}^{hedg} = g^{hedg}\left[\pi_{t\,T}^{Mkt}, \Delta\sigma_{t\,T}^{2}\right] \tag{14}$$

on the base of definition (20) of market risk premium and (6), we get the following expression :

$$OI_{t\,T}^{hedg} = g^{hedg}\left[\pi_{t\,T}, \Delta\sigma_{t\,T}^{2}, s_{t\,T}\right] \tag{15}$$

However, if market price converges to the rational value of the crude oil we will have a different impact of the spot price on the hedgers open interest.

5.2.2 Open interest and risk premium in the long run

In order to complete the empirical model, we study how open interest, for each agent, changes in the long run when spot price changes.

$$\frac{\partial OI_{t\,T}^{hedge}}{\partial p_t^*} = \frac{\partial OI_{t\,T}^{hedge}}{\partial \pi_{t\,T}} \frac{\partial \pi_{t\,T}}{\partial p_t^*} \tag{16}$$

$$\frac{\partial OI_{t\,T}^{spec}}{\partial p_t^*} = \frac{\partial OI_{t\,T}^{spec}}{\partial \pi_{t\,T}^{spec}} \frac{\partial \pi_{t\,T}^{spec}}{\partial p_t^*} \tag{17}$$

We assume that rational expectations are correct in the long run so that the hedger long run risk premium depends only on the long term price:

$$\pi_{t\,T} = -p_t^*(r_T) + \psi_{t,T} - K_T$$

$$\pi_{t\,T}^{spec} = \mathbb{E}_t^{spec}\left(p_{t+T}^*\right) - p_t^*(1 + r_T) + \psi_{t,T} - K_T$$

according to Pindick et. al (2016) [34], we impose that the marginal convenience yield depends on the spot price [10] $\psi_{t,T} = \psi_{t,T}(p_t^*)$ such that the first derivative of the marginal convenience yield in the level of price is positive $\psi'_{t,T}(p_t^*) > 0$ we get the following results:
The spot price impacts on risk premium as follows (see proof 3 in appendix):

$$\frac{\partial \pi_{t\,T}}{\partial p_t^*} = -(r_T) + \psi'_{t,T}(p_t^*) \tag{18}$$

$$\frac{\partial \pi_{t\,T}^{spec}}{\partial p_t^*} = -(1 + r_T) + \psi'_{t,T}(p_t^*) \tag{19}$$

by equation (5), we credibly suppose that $\frac{\partial \pi_{t\,T}^{spec}}{\partial p_t^*} \leq 0$ (see proof 2 in appendix)

[10] We assume that the producer is more inclined to spend more to store a higher evaluated good than a lower evaluated good (see for Pindick (1990)[42] (2016)[34])

6 Building the Empirical Model

Referring to the Kilian's model of 2006, it is proposed an empirical model of crude oil market using a SVAR with an identification scheme based on long run restrictions (Blanchard and Quah(1989)[8]); We define the following model

$$A_0 y_t = \alpha + \sum_{i=1}^{n} A_i y_{t-i} + \epsilon_t$$

where:

- ϵ_t is a vector containing the orthogonal structural innovations
- y_t consists of a vector containing in sequence: 1) the percentage change of the world's crude oil production, 2) a detrended index of real economic activity representing the global business cycle, 3) the index of the detrended deflated crude oil prices calculated on the base of U.S. refiners' acquisition cost 4) the percentage change in open interest for short positions held by commercial operators 5) the percentage change in open interest for long positions held by non-commercial operators.
- ϵ_t^s= The first shock can be determined by supply disruption associated with external events of production policy, by unexpected decisions of production policy or even by technological breakthroughs.
- ϵ_t^{eco}=The second shock regards to unexpected fluctuations in the global business cycle ('flow demand shock')
- ϵ_t^d=is the structural shock of the demand for crude oil. It is due to the change of the optimal level of the inventories on the Market for Storage (see Knittel and Pindyck (2016)[34]) where the equilibrium level of inventory is determined.

in order to study the remaining shocks, we must identify the factors influencing both the entry and exit decisions of individual agents about their positions in the futures market. Based on the theoretical study carried out in 5.2 we propose to add the following last two structural errors:

- ϵ_t^{Risk}= it is a function of the diversifiable risk (see figures 1 and 2) in the spot market for all maturities T of the traded futures contracts

  $$\epsilon_{t\,T}^{Risk} = f^{Risk}\left(\Delta\sigma_{t\,1}^2, \ldots, \Delta\sigma_{t\,n}^2\right)$$

  the function is increasing in $\Delta\sigma_{t\,T}^2$ for all T=1,...,n.
- ϵ_t^{Bull}= it's a function of the upward expectations for Non-Commerical traders (speculators) for all maturities T of the traded futures contracts.

  $$\epsilon_t^{Bull} = f^{Bull}\left[\mathbb{E}_t^{spec}\left(p_{t+1}^*\right) - \mathbb{E}_t\left(\bar{p}_{t+i}^*\right), \ldots, \mathbb{E}_t^{spec}\left(p_{t+n}^*\right) - \mathbb{E}_t\left(\bar{p}_{t+n}^*\right)\right]$$

  the function is increasing in the difference $\mathbb{E}_t^{spec}\left(p_{t+T}^*\right) - \mathbb{E}_t\left(\bar{p}_{t+T}^*\right)$ for all T=1,...,n.

We consider the variations for all T because we will use aggregate open interest data for all maturities T.

6.1 Long run restrictions

We define the set of long run restrictions as follows:

$$y_t^* = B^{long}\epsilon_t$$

$$y_t^* = \begin{pmatrix} s_t^{det} \\ g_t^{det} \\ p_t^{*\,det} \\ \triangle\, OI_t^{hedg} \\ \triangle\, OI_t^{spec} \end{pmatrix} = \begin{bmatrix} b_{1,1} & 0 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & 0 & 0 & 0 \\ b_{3,1} & b_{3,2} & b_{3,3} & 0 & 0 \\ b_{4,1} & b_{4,2} & b_{4,3} & b_{4,4} & 0 \\ b_{5,1} & b_{5,2} & b_{5,3} & b_{5,4} & b_{5,5} \end{bmatrix} \begin{pmatrix} \epsilon_t^s \\ \epsilon_t^{eco} \\ \epsilon_t^d \\ \epsilon_t^{Risk} \\ \epsilon_t^{Bull} \end{pmatrix}$$

where:
y_t^* is the vector containing the long-term values.

B^{long} is the estimated identified long-run impact matrix.
s_t^{det} the percentage change of the world's crude oil production;
g_t^{det} is a detrended index of real economic activity representing the global business cycle;
$p_t^{*\,det}$ is the index of the detrended deflated crude oil prices calculated on the base of U.S. refiners' acquisition cost;
$\triangle\, OI_t^{hedg}$ is the change in futures open interest held by hedgers aggregated for all maturities T;
$\triangle\, OI_t^{spec}$ is the change in futures open interest held by speculators aggregated for all maturities T.

We chose to insert the variations $\triangle\, OI_t^{hedg}$ and $\triangle\, OI_t^{spec}$ instead of their levels in order to deal with an invertible VAR$(\cdot)$ model.

The first three long run restrictions are explained in section 4.

I impose that:

- $b_{1,5} = b_{2,5} = b_{3,5} = 0$
  we assume, as in Kilian (2006)[29] and Christopher R. Knittel and Robert S. Pindyck (2016)[34], that expectations of price changes have no real effect, in the long run. Therefore no impact is expected to come from the upward expectation shock ϵ_t^{Bull} as well ($b_{1,5} = b_{2,5} = b_{3,5} = 0$);

- $b_{3,4} = 0$
  Any change in diversifiable risk does not have effect on price $b_{3,4} = 0$ since market doesn't compensate unsystematic risk. This hypothesis is consistent with Büyükşahin and Harris (2011)[9] results, in which they find that price changes precede hedge funds and other non-commercial (speculator) positions;

- $b_{1,4} = b_{2,4} = 0$
  no real effects, in the long run, are also expected on the world's crude oil supply and real economic activity.($b_{1,4} = b_{2,4} = 0$);

- $b_{1,2} = b_{1,3} = b_{2,3} = 0$
  the other three restrictions ($b_{1,2} = b_{1,3} = b_{2,3} = 0$) are the simple consequence of the spot market definition made in section 4 where the detrended price $p_t^{*\,det}$, supply s_t^{det} and global economic activity g_t^{det} are determined by only three shocks that are: the structural shock of demand, ϵ_t^d; the structural shock of supply, ϵ_t^s; the structural shock of the economy, ϵ_t^{eco}. See equations: (2) (3) (4);

- $b_{4,5}$
  we suppose that, in the long term, market rational expectations coincide with the long-term spot price (see section 5.2.2). Consequently in order to measure the impact we need nothing but parameters $b_{4,1}; b_{4,2}; b_{4,3}$.

6.2 The meaning of the long term impact parameters

We argued in section 5.2 that open interests are affected by three factors: risk propensity, market risk and expectations (see figures 1 2), as they can explain the decisions, of individual agents, about their position in the futures market. We are going to consider this factors for explaining the impact to estimate due to the structural shocks:

- $b_{4,1}; b_{4,2}; b_{4,3}; b_{5,1}; b_{5,2}; b_{5,3}$

  we study the impact of the spot price on open interest through the market risk premium. (see section 5.2.1) with parameters $b_{4,1}; b_{4,2}; b_{4,3}; b_{5,1}; b_{5,2}; b_{5,3}$ we estimate the effect that the structural shocks affecting the fluctuations of the price in the long run (ϵ_t^s ϵ_t^{eco} ϵ_t^d) have had on the variation of the analyzed open interests $\triangle\, OI_t^{hedg}$ and $\triangle\, OI_t^{spec}$.

  $b_{4,1}; b_{4,2}; b_{4,3}$
  On the base of formula (16) (17) and the conclusions in section 5.2.2 we expect that, if the hedger is risk averse, as it is supposed to be $\frac{\partial OI_{t\,T}^{hedge}}{\partial \pi_{t\,T}} > 0$ and the convenience yield increases instantaneously in price less than the discount rate
  $$\psi'_{t,T}(p_t^*) \leq r_T$$
  for any T, the long term impacts $b_{4,1}; b_{4,2}; b_{4,3}$ are negative.

  $b_{5,1}; b_{5,2}; b_{5,3}$
  if speculators seek risk premium $\frac{\partial OI_{t\,T}^{hedge}}{\partial \pi_{t\,T}} > 0$ the long term impacts $b_{5,1}; b_{5,2}; b_{5,3}$ are negative since it has been demonstrated that $\frac{\partial \pi_{t\,T}^{spec}}{\partial p_t^*} \leq 0$ (see proof 2 in appendix) and $\frac{\partial OI_{t\,T}^{spec}}{\partial \pi_{t\,T}} > 0$ for every T (see section 5.2).

- $b_{4,4}; b_{5,4}$
  Parameters $b_{4,4}; b_{5,4}$ are the impacts of the residual risk on the change of open interests. We suppose that producers are unable to diversify their portfolio appropriately as the market does. Therefore they cover all the risks (diversifiable and undiversifiable) tendentiously through the futures markets. We expect both parameters are positives since open interest is

affected by all the market risk regardless of the fact that a part of this risk (the diversifiable risk) does not produce any effect on the price in the long run.

- $b_{5,5}$
  The last parameter gives the relevance of the shift in the expectations for the open interest of the speculator, in the long run, $\triangle\, OI_t^{spec}$

- $b_{1,1}; b_{2,1}; b_{2,2}; b_{3,1}; b_{3,2}; b_{3,3}$
  equations (2) (3) (4) have the parameters of spot price model provided in section 4.

  $b_{1,1} = \beta_1$
  is the instantaneous equilibrium impact of oil productions shocks on the percentage change of the world's crude oil production;

  $b_{2,1} = \gamma_1 \beta_1$
  is the instantaneous equilibrium impact of oil productions shocks on the detrended index or real economic activity representing the global business cycle g_t^{det}. Notice that, the impact of the the supply shocks β_1 determines an indirect multiplicative effect on it;

  $b_{2,2} = \gamma_2$
  is the instantaneous equilibrium impact that unexpected fluctuations in the global business cycle ϵ_t^{eco} have on the (proxy of the) global business cycle g_t^{det};

  $b_{3,1} = \beta_1 \left(\frac{1}{\alpha_1} - \frac{\alpha_2}{\alpha_1} \right)$
  is the instantaneous equilibrium impact of the oil productions shocks on the spot price. In this case the impact of the the supply shocks β_1 determines an indirect multiplicative effect on the quantity $\left(\frac{1}{\alpha_1} - \frac{\alpha_2}{\alpha_1} \right)$ which is positive in sign if and only if the demand for crude oil increases less than proportionally to the growth in global economic activity $\alpha_2 < 1$;

  $b_{3,2} = -\frac{\alpha_2}{\alpha_1} \gamma_2$
  is the instantaneous equilibrium impact that unexpected fluctuations in the global business cycle have on spot price. notice that it is proportional to the the instantaneous equilibrium impact γ_2 that unexpected fluctuations in the global business cycle ϵ_t^{eco} have on the (proxy of the) global business cycle g_t^{det};

  $b_{3,3} = -\frac{\alpha_3}{\alpha_1}$
  is the instantaneous equilibrium impact that unexpected fluctuations in the global demand for crude oil have on the spot price. This impact can be partially due to speculation via inventory holding.

6.3 Data

We use monthly data over the sample period 1999-M1: 2008-M6. In order to compare the results with Kilian's, I'm going to use the same dataset built for

Kilian and Murphy (2012) [11][31]. The remaining part of data (open interest of commercials and non-commercials agents) are obtained from U.S. Commodity Futures Trading Commission (CFTC)[12]
The *price of crude oil* is based on U.S. refiners' acquisition cost of imported crude oil, extrapolated backwards as in Barsky and Kilian (2002)[4] and deflated by the U.S. consumer price index without annualising the growth rate of oil production. Both real prices of oil and real economic activity index, representing the global business cycle, are expressed in log deviations from their trend and mean, respectively. The *global oil production* data are measured in millions of barrels of oil and have been expressed as cumulative percent changes. The *index of real activity* has been saved by cumulating medium rates of increase in dry cargo ocean shipping freight rates, deflating the nominal index by the U.S. CPI and linearly detrending. The data referring to open interest have been transformed, to make them stationary, computing the percentage change from month to month. CFTC does not provide a unique value for each month. Therefore, we chose to consider the numerical data provided closer to the end of the month. Data on the open interest refer to all[13] futures long positions by Non-Commercials traders and all short positions carried out by Commercials traders[14] where we consider Commercials traders as a proxy of hedgers, on the contrary, the Non-Commercials as a proxy of speculators.

6.4 Model settings and Estimates

According to Lutkepohl and Netsunajev (2014)[41] we consider a lower number of lags compared to Kilian's work (Kilian (2006)[29]). The lag order of the VAR-models is determined using four different criteria. The Akaike Info Criterion, Final Prediction Error and Hannan-Quinn Criterion suggest two as the optimal number of lags (searched up to 10 lags of levels) while the Schwarz Criterion found one to be the best choice. I decided to use a 2nd-lagged Vector Autoregressive Model VAR(2).
A first estimate doesn't provide many significant regressors. In order to improve the statistical significance of the estimated parameters, it was decided to impose a subset of restrictions on the parameters of VAR in reduced form. The adopted procedure is implemented in J-MulTi4(version 4.24 (Oct 15, 2009)) is called "System Testing Procedure" (see Lütkepohl et al. (2006)[40]) and it does consist in an algorithm which identifies non-statistically significant parameters to be canceled out depending on a threshold value of the t-ratio parameters [15]. The estimated long run impact matrix is presented in figure 3 while the short

[11] `http://onlinelibrary.wiley.com/store/10.1111/j.1542-4774.2012.01080.x/asset/supinfo/JEEA_1080_sm_data_files.zip?v=1&s=734cbd288fba558c9386bba48c24a0619b5784e3`

[12] `http://www.cftc.gov/oce/web/ReportData/futures_CrudeOil.html`

[13] We mean for all maturities see `http://www.cftc.gov/MarketReports/CommitmentsofTraders/ExplanatoryNotes/index.htm`

[14] CFTC defines commercial and non-traders as follows: *"All of a trader's reported futures positions in a commodity are classified as commercial if the trader uses futures contracts in that particular commodity for hedging as defined in CFTC Regulation 1.3, 17 CFR 1.3(z)... generally gets classified as a "commercial" trader by filing a statement with the Commission, on CFTC Form 40: Statement of Reporting Trader, that it is commercially "...engaged in business activities hedged by the use of the futures or option markets."* all the others traders are considered non-commercials.

[15] The algorithm has been used in succession increasing the threshold at each step.

run matrix is presented in figure 4. We will see some details in the next session. Generally speaking, all coefficients are asymptotically significant at 95% and show the expected signs, with the only exception of $b_{3,1}^{Long}$, $b_{4,1}^{Long}, b_{5,1}^{Long}, b_{4,2}^{Long}$ for the long run coefficients and $b_{3,1}^{Short}$, $b_{4,1}^{Short}, b_{5,1}^{Short}, b_{4,2}^{Short}, b_{5,2}^{Short}, b_{5,3}^{Short}$ for the contemporaneous impacts.
Figure 5 represents the impact of the price to a unit shock from each of the five endogenous variables of the model. They allow four conclusions:
1) the ϵ_t^s has a positive statistically meaningful impact on real price only in the long run;
2) ϵ_t^d produces a positive effect, always statistically significant;
3) ϵ_t^{eco} has a positive and highly significant impact with a peak after the 3-rd month. Notice that these shocks do not cause reverting behaviour and the impact on the price is persistent over time;
4) ϵ_t^{Risk} has a reverting behavior impact on spot price that vanishes after one year. Every time that the risk increase, there is a weak upward trend of variation followed by a stronger downward trend after about three months;
5) ϵ_t^{Bull} has an immediate strong, and statistically significant, impact on the spot price that mainly runs out after three months.

Figure 3: Estimated identified long run impact matrix

$$B^{long} = \begin{bmatrix} 0.7992 & 0 & 0 & 0 & 0 \\ (9.5572) & (0.0000) & (0.0000) & (0.0000) & (0.0000) \\ 60.1377 & 119.5432 & 0 & 0 & 0 \\ (2.0250) & (2.4450) & (0.0000) & (0.0000) & (0.0000) \\ 74.1014 & 188.2361 & 98.4971 & 0 & 0 \\ (1.4868) & (2.1638) & (3.1828) & (0.0000) & (0.0000) \\ -0.0053 & -0.0165 & -0.0469 & 0.0589 & 0 \\ (-0.2955) & (-0.5875) & (-1,9634) & (8.3474) & (0.0000) \\ -0.1304 & -0.2626 & -0.1399 & 0.0861 & 0.2081 \\ (-1.5992) & (-1,9623) & (-2.2975) & (3.9054) & (7.1013) \end{bmatrix}$$

Note: bootstraped t-values are reported in parentheses

7 Discussion

The study identifies the global business cycle shock as the one with the greatest impact on the price fluctuations, in the long term, $b_{3,2}^{Long} = 188.2361$, whereas the structural shock of the demand determines the largest impact in the short term $b_{3,3}^{short} = 6.6939$. The shocks of the expectations has the second most relevant impact but broadly lower $b_{3,5}^{short} = 1.6638$ with respect to $b_{3,3}^{short} = 6.6939$.
We notice that:

- ϵ_t^{Bull}
  the long run impact, associated to the shock ϵ_t^{Bull}, on the open interest of the non-commercials agent is positive($b_{5,5}^{Long} = 0.2081$) and likely attributed to a shift in the expectation of the speculator, as it doesn't affect rational traders(hedgers) open interest ($b_{4,5}^{Long} = 0$) and all the other real

Figure 4: Estimated contemporaneous impact matrix

$$B^{short} = \begin{bmatrix} 0.9318 & 0 & 0 & 0 & 0 \\ (13.1164) & (0.0000) & (0.0000) & (0.0000) & (0.0000) \\ 0.7428 & 4.0769 & 0 & 0 & 0 \\ (1,9807) & (10.9974) & (0.0000) & (0.0000) & (0.0000) \\ 0.2701 & 1.1568 & 6.6939 & 0.6882 & 1.6638 \\ (0.3453) & (1,9629) & (12.6386) & (2.4714) & (3.2721) \\ -0.0062 & -0.0002 & -0.0144 & 0.0726 & 0.0087 \\ (-0.8763) & (-0.0258) & (-1.9605) & (9.7115) & (2.4374) \\ -0.0506 & -0.0347 & -0.0215 & 0.1174 & 0.2838 \\ (-1.2076) & (-0.9909) & (-0.6492) & (4.1461) & (8.7609) \end{bmatrix}$$

Note: bootstraped t-values are reported in parentheses

variables coherently with the theory provided by Kilian (2006)[29] and Knittel and Pindyck (2016)[34]. The increase in the price expectations pushes traders, to take a long position or delay the exit from the acquired position in the futures market.
The positivity of the parameter $b_{4,5}^{Short} = 0.0087$ is probably due to the weak but statistically significant effect that the shifts in the speculator's expectations determine on the market risk perceived by the hedgers in the short run (see section 5.2.1). This means that in the short term the shock ϵ_t^{Bull} distorts the hedgers' rational expectations inducing them to take and hold a position on the market. It is estimated a contemporaneous and statistically significant positive impact ($b_{3,5}^{Short} = 1.6638$ see figure 5), on the short-term fluctuations in the spot price, comparable for magnitude to the impact of the global business cycle on the crude oil real price fluctuations $b_{3,2}^{Short} = 1.1568$.
Growing speculators expectations have played a relevant role in rising prices but quite lower compared to the global demand shocks (see $b_{3,3}^{Short} = 6.6939$ figure 5).

- ϵ_t^{Risk}.
The shock ϵ_t^{Risk} is likely to be associated with a rise of diversifiable risks as it has a positive impact on the change of the open interest for both types of agents ($b_{4,4}^{Short} = 0.0726; b_{5,4}^{Short} = 0.1174$ $b_{4,4}^{Long} = 0.0589; b_{5,4}^{Long} = 0.0861$) but no effect on the spot price, in the long run ($b_{3,4}^{Long} = 0$). Generally speaking, traders (both hedgers and speculators) are more interested in entering into futures contracts when the market risk increases, except in cases where traders are indifferent to risk. $b_{3,4}^{Short} = 0.6882$ suggests that the non-diversifiable market risk has a significant positive effect on price in the short term and the increase of diversifiable risks could have probably played a role in the last price hike from January 2007 to June 2008 (see figure 5).
The positive parameters $b_{4,4}^{Short} = 0.0726$ and $b_{4,4}^{Long} = 0.0589$ (see figure 3)

is likely due to the inability of hedgers to diversify their portfolio. Therefore, they have to take and hold positions, on the futures market, to cover all the market risks (diversifiable and undiversifiable) through the futures markets. In this way, both markets permit that diversifiable risk vanishes in the long run $b_{3,4}^{Long} = 0$.
The positive parameter $b_{5,4}^{Long}$ indicates that speculators bet even on diversifiable risk in the long term.

Watching at figure 5, we notice that the dynamic evolution of the impact of the two shocks ϵ_t^{Risk} and ϵ_t^{Bull} on spot price is mean reverting around zero and remarkably similar (coherently with Christopher R. Knittel and Robert S. Pindyck (2016)[34]). This shows how the market perceives the shift in the expectations of speculators as an additional source of diversifiable risk whose effects on the price disappear after less than a year.

- ϵ_t^{eco}
  In the long run, the unexpected fluctuations in the global business cycle have a profound impact on price $b_{3,2}^{Long} = 188.2361$, but it doesn't produce significant effects on the propensity of traders to take and hold a position in the future markets $b_{4,2}^{Long} = -0.0165$, $b_{5,2}^{Long} = -0.2626$. In the short run, the impact on price is is high but reduced $b_{3,2}^{Short} = 1.1568$ moreover no significant effects are observable on open interest $b_{4,2}^{Short} = -0.0002$, $b_{5,2}^{Short} = -0.0347$. This means that, on average, traders operate within a sufficiently short period, such that unexpected fluctuations in the global business cycle do not induce a change in open interest.

- ϵ_t^{d}
  The sudden shock in the global demand for crude oil is the most relevant factor for price fluctuations. In the long run, demand shock has an intensely positive impact $b_{3,3}^{Long} = 98.4971$ on price. In the short term, a remarkable positive impact is also observable on price $b_{3,3}^{Short} = 6.6939$, which is characterised by the highest magnitude with respect the others impacts. Notice that, the global demand shocks determine an impact on price which is more than four times stronger than that induced by market expectations shocks $b_{3,3}^{Short} = 1.6638$ in the short run. In the near term (the first 8 months), global demand shock ϵ_t^d has a stronger impact than the business cycle shock ϵ_t^{eco}. This let us imagine that speculative inventory holdings may have played a significant role in price fluctuations. The effects on open interest is negative both in the short and long run: $b_{3,4}^{Short} = -0.0144$ $b_{3,5}^{Short} = -0.0215$ $b_{3,4}^{Long} = -0.0469$ $b_{3,5}^{Long} = -0.1399$.

- ϵ_t^{s}
  The cumulative impact of the supply shocks is invariably positive from October 2002 June 2008. During the same period, a marked increment of the spot price has been observed. Unexpected fluctuations in supply could have had a positive effect on the growth of the price in the period. Notice that after 18 months a statistically significant positive effect on price become observable.

- Null parameters
  The null contemporaneous impact coefficients $b_{1,4}^{Short}; b_{1,5}^{Short}; b_{2,4}^{Short}; b_{2,5}^{Short};$ suggest that the expectations of speculators and the risk of non-diversifiable market do not produce real effects on the fluctuations of the global supply of crude oil and global business cycle;
  the null contemporaneous impact coefficients $b_{1,2}^{Short}$ and $b_{1,3}^{Short}$ indicate that unexpected fluctuations in the global business cycle and demand for crude oil don't determine an immediate effect on oil production;
  the null contemporaneous impact coefficients $b_{2,3}^{Short}$ indicates that unexpected fluctuations in the global demand for crude oil don't instantaneously affect the global business cycle;

  It is worth to notice that the effect that price fluctuations determines on the open interest, through the parameters $b_{4,1}^{Long}$,$b_{4,2}^{Long}$,$b_{4,3}^{Long}$,$b_{5,1}^{Long}$,$b_{5,2}^{Long}$,$b_{5,3}^{Long}$ $b_{4,1}^{Short}$,$b_{4,2}^{Short}$,$b_{4,3}^{Short}$,$b_{5,1}^{Short}$,$b_{5,2}^{Short}$,$b_{5,3}^{Short}$, is likely to be negative for both categories of traders in both the short and long term. in particular: the negative parameters $b_{4,1}^{Long}$,$b_{4,2}^{Long}$,$b_{4,3}^{Long}$ can be interpreted as due to a marginal change of the convenience yield, which on average is lower than the discount rate $\psi'_{t,T}(p_t^*) < (r_T)$ see equation (18). The negative parameters confirm the theoretical assumption $\frac{\partial \pi_{t\ T}^{spec}}{\partial p_t^*} \leq 0$ explained in appendix (see proof 2).

8 Conclusion

This study provides a structural interpretation of the price fluctuations in the global market for crude oil based on the analytic framework proposed by Christopher R. Knittel and Robert S. Pindyck (2016)[34] and Kilian (2006) [29]. I have proposed a SVAR empirical model with long-run restrictions, à la Blanchard-Quah, whereby we have been able to estimate the long and short-term effect, of structural shocks on the spot price fluctuations.
In order to explain the effect of the speculation activity on the spot price, I employed a dataset of monthly data built matching the dataset used in Kilian and Murphy (2012) [16][31] with a CFTC dataset of futures open interests disaggregated by type of trader (commercial and non commercial) and type of operation (long and short). Some theoretical considerations have been developed to derive and interpret structural shocks from the open interest.
A long-term association based on the standard theory of the storage has led us to explain the open interest as a function of the: 1) risk propensity; 2) market risk; 3) expectations of the market price. Thus, we have used these definitions to estimate the short and long run effect of structural shocks on the spot price. The empirical estimation of the model has shown some remarkable results: 1) a growing diversifiable risk results in a temporary increase of "short-term" spot price; 2) the up-warding changes in expectations, exerted by speculators on the futures market, produce a prompt significant short-run effect on the spot price; 3) short run spot price fluctuations react, with greater strength, to unexpected

[16] http://onlinelibrary.wiley.com/store/10.1111/j.1542-4774.2012.01080.x/asset/supinfo/JEEA_1080_sm_data_files.zip?v=1&s=734cbd288fba558c9386bba48c24a0619b5784e3

shocks of the global oil demand, which prove how important are the changes in inventory holdings. In particular, in the short run, I believe that the observed strong impact of the shocks of the global oil demand (inventory holdings) can likely be explained by the speculative demand. In fact, the short-run impact of the demand shock on the spot price strangely takes a long time (eight months in mean) to be overtaken by that of the business cycle shock, which is inexplicable if it's only due to strictly technical production needs.
In conclusion, in the short run, we observe that: 1) spot price fluctuations strongly reacts to a shift in expectations, determined by speculators operating in the futures market; 2) global crude oil demand shocks (unexpected changes in inventory holdings) influence the spot price to a greater extent. Speculative inventory holdings probably have a bigger effect on spot price than speculative operations in the futures market; 3) different types of risks (diversifiable and non-diversifiable) determine different effects on the spot price. The market does not immediately identify the nature of risk. It would seem that the global market needs at least one month to determine the nature of a diversifiable risk; 4) shocks in the global economic activity and global supply do not produce the strongest effect on the spot price immediately.

References

[1] Ron Alquist and Lutz Kilian. *What do we learn from the price of crude oil futures?* In: ***Journal of Applied Econometrics*** 25.4 (2010), pp. 539–573.

[2] Hossein Askari and Noureddine Krichene. *An oil demand and supply model incorporating monetary policy.* In: ***Energy*** 35.5 (2010), pp. 2013–2021.

[3] Steven D Baker and Bryan R Routledge. *The price of oil risk.* In: (2011).

[4] Robert B Barsky and Lutz Kilian. *Do we really know that oil caused the great stagflation? A monetary alternative.* In: ***NBER Macr. Annual 2001***. MIT Press, 2002, pp. 137–198.

[5] Christiane Baumeister and Lutz Kilian. *Real-time analysis of oil price risks using forecast scenarios.* In: (2011).

[6] Christiane Baumeister and Gert Peersman. *Sources of the volatility puzzle in the crude oil market.* In: ***Available at SSRN 1471388*** (2009).

[7] Cristina Bencivenga, Rita L D'Ecclesia, and Umberto Triulzi. *Oil prices and the financial crisis.* In: ***Review of Managerial Science*** 6.3 (2012), pp. 227–238.

[8] Olivier Jean Blanchard and Danny Quah. *The Dynamic Effects of Aggregate Demand and Supply Disturbances.* In: ***The American Economic Rev.*** 79.4 (1989), pp. 655–673.

[9] Bahattin Büyüksahin and Jeffrey H Harris. *Do speculators drive crude oil futures prices.* In: ***Energy Journal*** 32.2 (2011), pp. 167–202.

[10] Fabio Canova and Gianni De Nicolo. *Monetary disturbances matter for business fluctuations in the G-7.* In: ***Journal of Monetary Economics*** 49.6 (2002), pp. 1131–1159.

[11] Colin A Carter, Gordon C Rausser, and Andrew Schmitz. *Efficient asset portfolios and the theory of normal backwardation.* In: ***Journal of Political Economy*** 91.2 (1983), pp. 319–331.

[12] Eric C Chang. *Returns to speculators and the theory of normal backwardation.* In: ***The Journal of Finance*** 40.1 (1985), pp. 193–208.

[13] EU Commission. *Communication on "Food Prices in Europe".* In: 821 (2010).

[14] Frans A De Roon, Theo E Nijman, and Chris Veld. *Hedging pressure effects in futures markets.* In: ***The Journal of Finance*** 55.3 (2000), pp. 1437–1456.

[15] Rita L D'Ecclesia et al. *Understanding recent oil price dynamics: A novel empirical approach.* In: ***Energy Economics*** 46 (2014), S11–S17.

[16] Stephane Dees et al. *Modelling the world oil market: Assessment of a quarterly econometric model.* In: ***Energy Policy*** 35.1 (2007), pp. 178–191.

[17] Stéphane Dées et al. *Assessing the factors behind oil price changes.* In: (2008).

[18] Bassam Fattouh, Lutz Kilian, and Lavan Mahadeva. *The role of speculation in oil markets: What have we learned so far?* In: (2012).

[19] Jon Faust. *The robustness of identified VAR conclusions about money*. In: ***Carnegie-Rochester Conference Series on Public Policy***. Vol. 49. Elsevier. 1998, pp. 207–244.

[20] Domenico Giannone and Lucrezia Reichlin. *Does information help recovering structural shocks from past observations?* In: ***Journal of the European Economic Association*** 4.2-3 (2006), pp. 455–465.

[21] HM Government. *The 2007/08 Agricultural Price Spikes: Causes and Policy Implications*. In: (2010).

[22] James D Hamilton. *Understanding crude oil prices*. Tech. rep. National Bureau of Economic Research, 2008.

[23] James D Hamilton. *Causes and Consequences of the Oil Shock of 2007-08*. Tech. rep. National Bureau of Economic Research, 2009.

[24] James D Hamilton and Jing Cynthia Wu. *Risk premia in crude oil futures prices*. In: ***Journal of International Money and Finance*** 42 (2014), pp. 9–37.

[25] Yanan He, Shouyang Wang, and Kin Keung Lai. *Global economic activity and crude oil prices: A cointegration analysis*. In: ***Energy Economics*** 32.4 (2010), pp. 868–876.

[26] Leon Isserlis. *Tramp shipping cargoes, and freights*. In: ***Journal of the Royal Statistical Society*** 101.1 (1938), pp. 53–146.

[27] Luciana Juvenal and Ivan Petrella. *Speculation in the oil market*. In: ***Journal of Applied Econometrics*** 30.4 (2015), pp. 621–649.

[28] Robert K Kaufmann. *The role of market fundamentals and speculation in recent price changes for crude oil*. In: ***Energy Policy*** 39.1 (2011), pp. 105–115.

[29] Lutz Kilian. *Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market*. In: (2006).

[30] Lutz Kilian and Bruce Hicks. *Did unexpectedly strong economic growth cause the oil price shock of 2003–2008?* In: ***Journal of Forecasting*** 32.5 (2013), pp. 385–394.

[31] Lutz Kilian and Daniel P Murphy. *Why agnostic sign restrictions are not enough: understanding the dynamics of oil market VAR models*. In: ***Journal of the European Economic Association*** 10.5 (2012), pp. 1166–1188.

[32] Lutz Kilian and Daniel P Murphy. *The role of inventories and speculative trading in the global market for crude oil*. In: ***Journal of Applied Econometrics*** 29.3 (2014), pp. 454–478.

[33] Jan Tore Klovland. *Business cycles, commodity prices and shipping freight rates: Some evidence from the pre-WWI period*. In: (2002).

[34] Christopher R Knittel and Robert S Pindyck. *The simple economics of commodity price speculation*. In: ***American Economic Journal: Macroeconomics*** 8.2 (2016), pp. 85–110.

[35] Noureddine Krichene. *World crude oil and natural gas: a demand and supply model*. In: ***Energy economics*** 24.6 (2002), pp. 557–576.

[36] Noureddine Krichene. *A simultaneous equations model for world crude oil and natural gas markets*. In: (2005).

[37] Paul Krugman. *The oil nonbubble*. In: ***New York Times*** 12 (2008), p. 2008.

[38] J Leiberman. *Senators hear testimony on legislation to reduce food and energy prices*. In: ***News Release. Available online: http://lieberman. senate. gov/newsroom/release. cfm*** (2008).

[39] Francesco Lippi and Andrea Nobili. *Oil and the macroeconomy: a quantitative structural analysis*. In: ***Journal of the European Economic Association*** 10.5 (2012), pp. 1059–1083.

[40] Helmut Lütkepohl, M Kratzig, and Dmitri Boreiko. *VAR analysis in JMulTi*. In: ***JMul'Ti Documentation Downloads*** (2006).

[41] Helmut Lütkepohl and Aleksei Netsunajev. *Structural vector autoregressions with smooth transition in variances: The interaction between us monetary policy and the stock market*. In: (2014).

[42] Robert S Pindyck. *Inventories and the short-run dynamics of commodity prices*. Tech. rep. National Bureau of Economic Research, 1990.

[43] Harri Ramcharran. *Oil production responses to price changes: an empirical application of the competitive model to OPEC and non-OPEC countries*. In: ***Energy economics*** 24.2 (2002), pp. 97–106.

[44] Kenneth J Singleton. *Investor flows and the 2008 boom/bust in oil prices*. In: ***Management Science*** 60.2 (2013), pp. 300–318.

[45] James L Smith. *World oil: market or mayhem?* In: ***The journal of economic perspectives*** 23.3 (2009), pp. 145–164.

[46] Martin Stopford. *Maritime Economics. 2nd*. 1997.

[47] Jan Tinbergen. *Tonnage and freight*. In: ***Jan Tinbergen Selected Papers*** (1959), pp. 93–111.

[48] Adair Turner, Jon Farrimond, and Jonathan Hill. *The oil trading markets, 2003–10: analysis of market behaviour and possible policy responses*. In: ***Oxford Review of Economic Policy*** 27.1 (2011), pp. 33–67.

[49] Harald Uhlig. *What are the effects of monetary policy on output? Results from an agnostic identification procedure*. In: ***Journal of Monetary Economics*** 52.2 (2005), pp. 381–419.

[50] Naoyuki Yoshino and Farhad Taghizadeh-Hesary. *Monetary Policy and the Oil Market*. Springer, 2015.

APPENDIX

Proof 1

We want to demonstrate that speculator's risk premium can be decomposed into the sum of the market risk premium and a shift due to speculators' expectations. We first give some definitions:
we define market risk premium as a difference between the expected future spot price by the market $\mathbb{E}_t\left(p^*_{t+T}\right)$ and the futures price $F_{t,T}$

$$\pi^{Mkt}_{t\ T} = \mathbb{E}_t\left(p^*_{t+T}\right) - F_{t,T} \tag{20}$$

and the speculator risk premium as a difference between the expected future spot price by speculator $\mathbb{E}_t\left(p^*_{t+T}\right)$ and the futures price $F_{t,T}$.

$$\pi^{spec}_{t\ T} = \mathbb{E}^{spec}_t\left(p^*_{t+T}\right) - F_{t,T}$$

where $\mathbb{E}_t\left(p^*_{t+T}\right)$ is the market expectation and $\mathbb{E}^{spec}_t\left(p^*_{t+T}\right)$ is the speculator's expectation.

In the short term, the expected future spot price is the sum of two elements: the expected future spot price under rational expectations $\mathbb{E}_t\left(\bar{p}^*_{t+T}\right)$ and a shift $s_{t\ T}$ due to the speculator's expectation

$$\mathbb{E}_t\left(p^*_{t+T}\right) = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) + s_{t\ T}$$

Thus in the short run the market risk premium is given by the following expression:

$$\pi^{Mkt}_{t\ T} = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) - F_{t,T} + s_{t\ T}$$

if the shift in the market expectation is defined as follows

$$s_{t\ T} = \mathbb{E}^{spec}_t\left(p^*_{t+T}\right) - \mathbb{E}_t\left(\bar{p}^*_{t+T}\right)$$

market risk premium become

$$\pi^{Mkt}_{t\ T} = \mathbb{E}^{spec}_t\left(p^*_{t+T}\right) - F_{t,T} = \pi^{spec}_{t\ T}$$

this result is sufficient to demonstrate that formulas (10) and (11), in fact

$$\pi^{spec}_{t\ T} = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) - F_{t,T} + s_{t\ T}$$

moreover by definition
$\pi_{t\ T} = \mathbb{E}_t\left(\bar{p}^*_{t+T}\right) - F_{t,T}$
therefore

$$\pi^{spec}_{t\ T} = \pi_{t\ T} + s_{t\ T}$$

Proof 2

By equation (5), we quantify the increment of the spot price, from time t to time $t+1$, assuming that the discount rate r_T and the unit storage cost K_T to be constant:

$$p^*_{t+1} - p^*_t = \frac{1}{1+r_T} \cdot [F_{t+1,T} - F_{t,T} + \psi_{t+1,T} - \psi_{t,T}]$$

by transformation

$$1 + r_T - \frac{F_{t+1,T} - F_{t,T}}{p^*_{t+1} - p^*_t} = \frac{\psi_{t+1,T} - \psi_{t,T}}{p^*_{t+1} - p^*_t}$$

being futures prices positively correlated with spot prices we can write the following inequality

$$1 + r_T \geq \frac{\psi_{t+1,T} - \psi_{t,T}}{p^*_{t+1} - p^*_t}$$

Assuming that:
1) the time change is infinitesimal;
2) the producer is more inclined to spend more to store a higher evaluated good than a lower evaluated good , as in in Pindick (1990)[42] (2016)[34], meqaning that $\psi'_{t,T}(p^*_t) > 0$
we are allowed to write that

$$-(1 + r_T) + \psi'_{t,T}(p^*_t) \leq 0$$

Proof 3

$$\pi^{spec}_{t\,T} = \mathbb{E}_t \left(p^*_{t+T}\right) - p^*_t(1 + r_T) + \psi_{t,T} - K_T$$

by equation 6 we get:

$$= \mathbb{E}_t \left(\bar{p}^*_{t+T}\right) + s_{t\,T} - p^*_t(1 + r_T) + \psi_{t,T} - K_T$$

assuming $s_{t\,T} = \mathbb{E}^{spec}_t \left(p^*_{t+T}\right) - \mathbb{E}_t \left(\bar{p}^*_{t+T}\right)$ as done in section 5.2

$$= \mathbb{E}^{spec}_t \left(p^*_{t+T}\right) - p^*_t(1 + r_T) + \psi_{t,T} - K_T$$

Table 1: VAR in reduced form.

Reduced form:	$y_t = \alpha^* + A_1^* y_{t-1} + A_2^* y_{t-2} + u_t$
exogenous variables:	$y_t = \begin{bmatrix} s_t^{det} & g_t^{det} & p_t^{*\,det} & \triangle\, OI_t^{hedg} & \triangle\, OI_t^{spec} \end{bmatrix}$
deterministic variables:	α
endogenous lags:	2
exogenous lags:	0
sample range:	$[1999M1, 2008M8], T = 116$
estimation method:	EGLS
Restrictions on A_1^*	$A_1^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ a_{121} & a_{122} & 0 & 0 & 0 \\ 0 & a_{132} & a_{133} & 0 & a_{135} \\ 0 & a_{142} & a_{143} & 0 & a_{145} \\ 0 & 0 & a_{153} & 0 & a_{155} \end{bmatrix}$
Restrictions on A_2^*	$A_2^* = \begin{bmatrix} a_{211} & 0 & 0 & 0 & 0 \\ a_{221} & a_{222} & 0 & 0 & 0 \\ a_{231} & 0 & a_{233} & 0 & a_{235} \\ 0 & 0 & a_{243} & a_{244} & 0 \\ 0 & 0 & a_{253} & 0 & a_{255} \end{bmatrix}$
Restrictions on α^*	$\alpha^* = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix}$

Table 2: VAR in structural form.

Structural form:	$A_0 y_t = \alpha + A_1 y_{t-1} + A_2 y_{t-2} + \epsilon_t$ $y_t^* = B^{long}\epsilon_t \qquad u_t = B^{short}\epsilon_t$
exogenous variables:	$y_t = \begin{bmatrix} s_t^{det} & g_t^{det} & p_t^{*\,det} & \triangle\, OI_t^{hedg} & \triangle\, OI_t^{spec} \end{bmatrix}$
deterministic variables:	α
endogenous lags:	2
exogenous lags:	0
sample range:	$[1999M1, 2008M8], T = 116$
estimation method:	Blanchard-Quah
Restrictions on B^{long}	$B^{long} = \begin{bmatrix} b_{1,1} & 0 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & 0 & 0 & 0 \\ b_{3,1} & b_{3,2} & b_{3,3} & 0 & 0 \\ b_{4,1} & b_{4,2} & b_{4,3} & b_{4,4} & 0 \\ b_{5,1} & b_{5,2} & b_{5,3} & b_{5,4} & b_{5,5} \end{bmatrix}$

Table 3: Estimated lag 1 matrix coefficients.

Coefficients	value	t-values
$a_{1,2,1}^{Long}$	0,771	1.944
$a_{1,2,2}^{Long}$	1.323	15.901
$a_{1,3,2}^{Long}$	0.098	2.344
$a_{1,3,3}^{Long}$	1.257	14.004
$a_{1,3,5}^{Long}$	-5.404	-3.101
$a_{1,4,2}^{Long}$	0.000	1.238
$a_{1,4,3}^{Long}$	0.005	5.571
$a_{1,4,5}^{Long}$	0.042	-2.224
$a_{1,5,3}^{Long}$	0.027	6.793
$a_{1,5,5}^{Long}$	-0.186	-2.406

Table 4: Estimated lag 2 matrix coefficients.

Coefficients	value	t-values
$a_{2,1,1}^{Long}$	-0.166	-1.796
$a_{2,2,1}^{Long}$	0.866	2.075
$a_{2,2,2}^{Long}$	-0.357	-4.328
$a_{2,3,1}^{Long}$	-1.637	-2.325
$a_{2,3,3}^{Long}$	-0.336	-3.817
$a_{2,3,5}^{Long}$	-2.589	-1.555
$a_{2,4,3}^{Long}$	-0.006	-6.163
$a_{2,4,4}^{Long}$	-0.172	-2.397
$a_{2,5,3}^{Long}$	-0.028	-7.260
$a_{2,5,5}^{Long}$	-0.178	-2.668

Table 5: Estimated lag 2 matrix coefficients.

Coefficients	value	t-values
α_1	0.090	1.033
α_2	0.581	1.408
α_3	1.108	1.407
α_4	-0.000	-0.028
α_5	0.059	1.926

Table 6: modulus of the eigenvalues of the reverse characteristic polynomial:

$\mid z \mid$							
1.0580	1.0899	2.6475	19.8168	1.8466	1.8466	2.4547	2.4547
2.4109	2.4109						

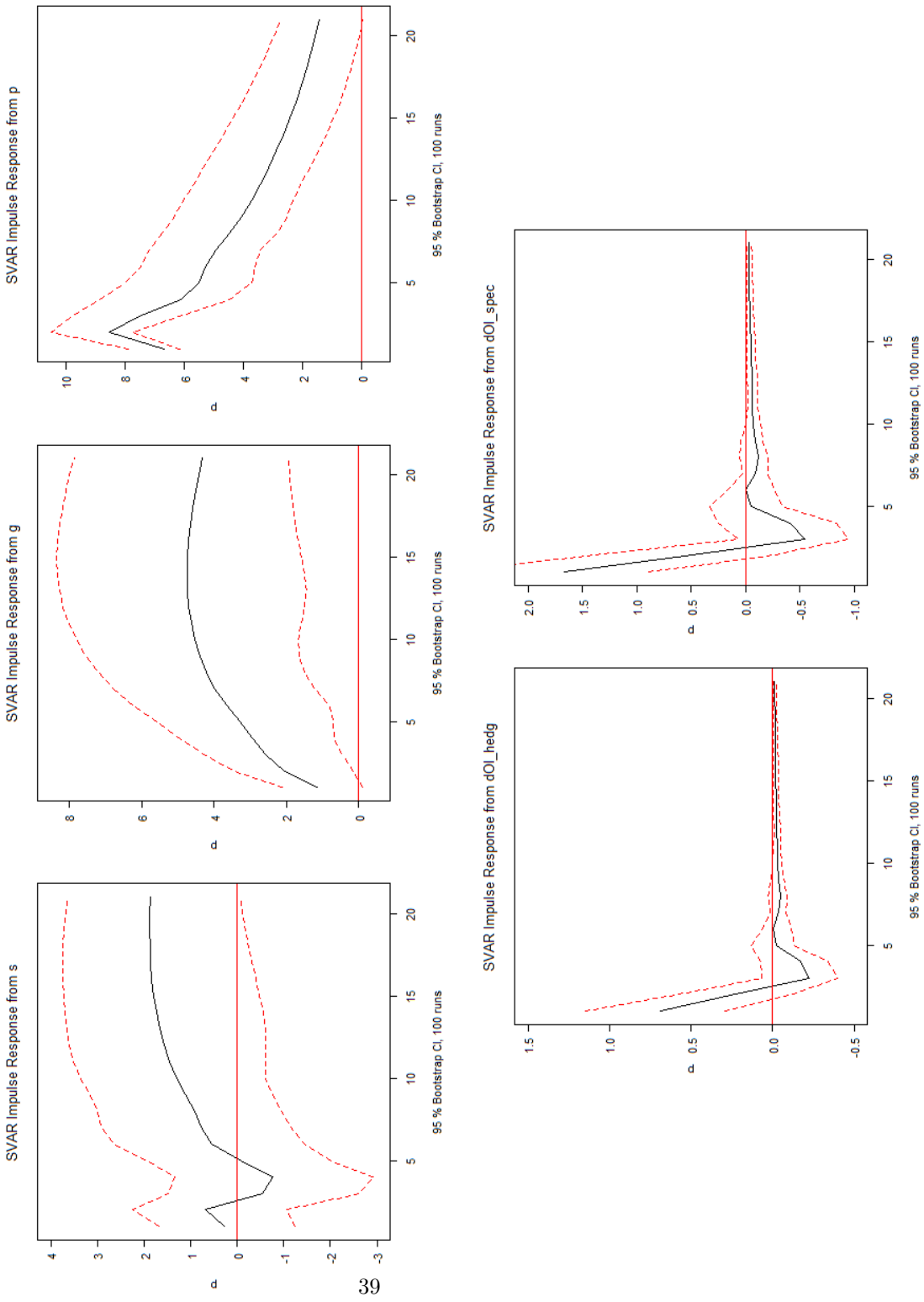
Table 7: Estimated identified long run impact coefficients.

Coefficients	Estimates	Standard Errors	t-values
$b_{1,1}^{Long}$	0,79920	0,0836	9,5572
$b_{2,1}^{Long}$	60,13770	29,6977	2,025
$b_{3,1}^{Long}$	74,10140	49,838	1,4868
$b_{4,1}^{Long}$	-0,00530	0,0178	-0,2955
$b_{5,1}^{Long}$	-0,13040	0,0816	-1,5992
$b_{2,2}^{Long}$	119,5432	48,8925	2,445
$b_{3,2}^{Long}$	188,2361	86,9931	2,1638
$b_{4,2}^{Long}$	-0,0165	0,0281	-0,5875
$b_{5,2}^{Long}$	-0,2626	0,1373	-1,9623
$b_{3,3}^{Long}$	98,4971	30,9462	3,1828
$b_{4,3}^{Long}$	-0,0469	0,025	-1,9634
$b_{5,3}^{Long}$	-0,1399	0,0609	-2,2975
$b_{4,4}^{Long}$	0,0589	0,0071	8,3474
$b_{5,4}^{Long}$	0,0861	0,022	3,9054
$b_{5,5}^{Long}$	0,20810	0,0293	7,1013

Table 8: Estimated identified short run impact coefficients.

Coefficients	Estimates	standard errors	t-values
$b_{1,1}^{Short}$	0,9318	0,071	13,1164
$b_{2,1}^{Short}$	0,7428	0,4172	1,9807
$b_{3,1}^{Short}$	0,2701	0,7822	0,3453
$b_{4,1}^{Short}$	-0,0062	0,007	-0,8763
$b_{5,1}^{Short}$	-0,0506	0,0419	-1,2076
$b_{2,2}^{Short}$	4,0769	0,3707	10,9974
$b_{3,2}^{Short}$	1,1568	0,6753	1,9629
$b_{4,2}^{Short}$	-0,0002	0,0066	-0,0258
$b_{5,2}^{Short}$	-0,0347	0,035	-0,9909
$b_{3,3}^{Short}$	6,6939	0,5296	12,6386
$b_{4,3}^{Short}$	-0,0144	0,0075	-1,9605
$b_{5,3}^{Short}$	-0,0215	0,0332	-0,6492
$b_{3,4}^{Short}$	0,6882	0,2785	2,4714
$b_{4,4}^{Short}$	0,0726	0,0075	9,7115
$b_{5,4}^{Short}$	0,1174	0,0283	4,1461
$b_{3,5}^{Short}$	1,6638	0,5085	3,2721
$b_{4,5}^{Short}$	0,0087	0,0036	2,4374
$b_{5,5}^{Short}$	0,2838	0,0324	8,7609

Figure 5: Impulse-response functions of price



The figure displays the impulse responses of the index of the detrended deflated crude oil prices to a one standard deviation change in the five structural shocks. The responses are significant at the 95% level.

Table 9: Legend of the figure 5:

Percentage change of the world's crude oil production	$s = s_t^{det}$
Detrended index of real economic activity representing the global business cycle;	$g = g_t^{det}$
Index of the detrended deflated crude oil prices calculated on the base of U.S. refiners' acquisition cost	$p = p_t^{*\,det}$
Change in futures open interest held by hedgers aggregated for all maturities T	$dOI_hedg =\triangle OI_t^{hedg}$
Change in futures open interest held by speculators aggregated for all maturities T	$dOI_spec =\triangle OI_t^{spec}$

Table 10: PORTMANTEAU TEST (H0:Rh=(r1,...,rh)=0)

Reference: Lütkepohl (1993), Introduction to Multiple Time Series Analysis, 2ed, p. 150.	
tested order:	16
test statistic:	344.8805
p-value:	0.9017
adjusted test statistic:	372.1226
p-value:	0.6039
degrees of freedom:	380.0000

Table 11: LM-TYPE TEST FOR AUTOCORRELATION with 16 lags

Reference: Doornik (1996), LM test and LMF test (with F-approximation)	
LM statistic:	437.9083
p-value:	0.0928
df:	400.0000

Table 12: ARCH-LM TEST with 16 lags

variable	teststat	p-Value(Chi^2)	skewness	kurtosis
u1	18.3582	0.3033	1.4054	0.1595
u2	7.9610	0.9500	0.5406	0.9173
u3	11.3066	0.7902	0.7967	0.6851
u4	17.3801	0.3614	1.3148	0.2081
u5	33.0255	0.0073	3.0819	0.0004

Table 13: MULTIVARIATE ARCH-LM TEST with 7 lags

VARCHLM test statistic:	1581.3156
p-value(chi^2):	0.4506
degrees of freedom:	1575.0000

European Energy Security: the Substitutability of European Crude Oil Imports from Russia

Gabriele D'Amore*

May 7, 2017

Abstract

The study is meant to be a contribution to the current debate on the diversification possibilities in EU for reducing the dependency on Russian crude oil and ensuring the energy security of the European Union (EU). We focus on the aggregate demand for crude oil in EU with the aim of investigating the degree of substitutability of crude oil imports from the Former Soviet Union countries (FSU)[1] and crude oil imports from four alternative regions (America, Middle East, Europe, Africa). Following Fuss (1977)[27] and Serletis (2010) [59] we employ an econometric model of intra-fuel substitution, using a nonlinear seemingly unrelated regression (SUR) estimator, to assess the aforementioned degree of substitutability in terms of Morishima elasticities of substitution. We use the most recent dataset, published by the European Commission, consisting of a collection of imported volumes and CIF[2] prices of crude oil by country of origin. The results indicate that the crude oil provided by former Soviet Union (FSU) countries is strongly substitutable with those imported from African and Middle Eastern countries whilst it is not substitutable with those imported from European and American countries.

Keywords: Flexible Functional Form, Translog Cost Function, Theoretical regularity, Crude Oil, TTIP, Russia, Morishima elasticities, Intra-Fuel Substitution, Energy Security
JEL-Classification: C2 D4.

1 Introduction

The securitization of the European Union energy supply is one of the most critical problems the EU member states are facing today. This issue has gained attention since the first major oil crisis of the 70's when the OPEC countries adopted a strategy of reducing supply as a warning to the West not to support Israel during the Arab-Israeli conflict. Energy crisis led Europe to search

*Sapienza University of Rome. Mail to: gabriele.damore@uniroma1.it. Corresponding author at: Department of Economics and Social Sciences, Piazzale AldoMoro, 5 - 00185 Rome(IT).

[1]The FSU countries are: Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan.

[2]CIF prices stand for cost, insurance and freight prices

for new sources of supply in order to try to limit the energy dependency on OPEC countries. Consequently, Russia and some other Eurasian countries were identified as alternative suppliers (see Helen (2010) [31]).
However according to authors like Morelli (2006) [50], Baran (2007) [6], Belkin (2008) [9], Woehrel (2010) [65], Badalov (2012) [5], this problem still remain crucial. The Russian solution to the European energy security problem, appears now to be anachronistic for three main reasons:

a. Recently the growing geopolitical instability of Eurasian countries has made EU concerned about Russia's partnership, see Mamlyuk (2015) [47]. According to European Commission, *"EU economy seems to be exposed to serious risks related to energy prices, including potential oil shocks risks"* (see Hedenus et al. (2010)[29]) so that the energy security has become a key objective of the European political agenda in the aftermath of the outbreak of the crisis in Ukraine.

b. At present "*EU energy security depends heavily on Russian supply, and Russian government would be lobbying to consolidate and increase the degree of dependence of Europe*". Almost 50% of its natural gas and 30% of its oil is imported from Russia and a consistent number of individual EU countries is geographically positioned so as to be naturally predisposed to assume a position of dependence on Russia becoming both economically and strategically vulnerable.
Many economists have expressed concern about this high degree of addiction, see Palonkorpi (2007) [53], Baran (2007) [6]. Other authors believe that a partnership with Russia is unavoidable as there is a high level of interdependence between the two economic areas such that this situation should remain unchanged for a long time, see Paillard (2010) [52]. Many countries have signed long-term contracts with Russia.

c. The increase of energy needs in time will likely enforce the degree of dependence of Europe on Russian supply. Total energy imports amount about 50%, and the European Commission expects this figure to rise to 65% by 2030. Last Eurostat statistic (preliminary data for 2012) of EU-28 Gross inland consumption (as % of total Mtoe) shows that crude oil and petroleum products consumption continue to dominate the energy mix.

For all the above reasons, Europe still appears to need a reliable security strategy that aims the diversification of the supply routes. This policy would reduce the dependence on imports overall and integrate European energy system to make it more resilient to external supply disruptions. However diversifying energy supply routes does not necessarily guarantee energy security since it strictly depends on the refinery capabilities to adjust production to different types of crude oil acquired at different costs.
A great unknown is about the possibility to substitute Russian crude oil with that coming from American continent and in particular with that produced in the United States which, by the way, over the last years, has become the leader in the extraction of oil with new techniques (fracking and shale oil). This unprecedented situation could potentially let the U.S. become a net exporter

nation of this new type of crude in the next future.
A question rise spontaneously, is American crude oil a possible alternative choice for Europe? In order to answer this question we need to investigate the benefits and the costs associated with this option and the other alternatives.
The central idea of the paper is to treat the quantities of imported crude oil as independent factors of the European production system and then to verify the degree of substitutability of crude oil imports from the Former Soviet Union (FSU) countries and crude oil imports from four alternative regions (America, Middle East, Europe, Africa) by calculating the Morishima elasticity of substitution (natural extension of the Hicksian elasticity , as we shall see) following the methodology developed by Fuss (1977)[27] and Serletis (2010) [59] and exploiting a nonlinear Seemingly Unrelated Regression (SUR) estimator.
The paper is composed of seven chapters:
chapter 1 is devoted mainly to listing the purposes of research; chapter 2 is devoted to the analysis of energy security in Europe with a particular focus on relations between Europe and the country's most important supplier of energy product, Russia. The chapter will also analyze, the degree of dependency on external sources and the critical factors determining the weaknesses in energy security in the European Union; chapter 3 is focused on the Russian oil supply, (production, transport routes, the Russian oil grades, taxation) and on its energy strategy up to 2030; chapter 4 is devoted to analyze the European crude oil demand and to list relevant factors affecting the demand itself (location of the petroleum extraction field, chemical properties, political and macroeconomic factors); chapter 5 and chapter 6 are devoted to explain the followed methodology and the related literature to support the choice of both the theoretical approach and the empirical analysis techniques; chapter 7 is dedicated to the empirical analysis and conclusions.

2 Energy Situations and Security: Energy Geopolitics in the EU and Russia

2.1 Energy security

There are several definitions of energy security in literature. The European Union does not distinguish between energy security and supply security, meaning that energy security is synonymous with stable and affordable supplies. Indeed, according to the 'Energy 2020' - a document of the European Commission - the main objective of the EU in terms of energy security is "*to ensure the uninterrupted physical availability of energy products and services on the market, at a price which is affordable for all consumers (private and industrial), while contributing to the EU's wider social and climate goals*" (European Electricity Grids Initiative and others (2010) [38]). Cristian von Hirschhausen defines it as a state where the risks related to high dependence on energy imports, political instability in producing and/or transit countries, as well as of other adverse contingencies, are mastered at reasonable economic costs (von Hirschhausen C., (2005) [60]).
For the International Energy Agency (IEA) the uninterrupted availability of energy sources at an affordable price is an inescapable characteristic for energy security. Apparently, the achievability of the energy security is affected by the

time horizon of its target. Therefore, long-term energy security requires timely investments to supply energy in line with economic developments and environmental needs. On the other hand, short-term energy security is possible when energy system can react promptly to sudden changes in the supply-demand balance (IEA, (2015)).
The World Energy Council (WEC) defines Energy Security as: "*the effective management of primary energy supply from domestic and external sources, the reliability of energy infrastructure, and the ability of the participating energy companies to meet current and future demand*" (Philip Lowe - World Energy Council (2015))[44]. In the light of that definition, WEC provides an index of energy sustainability which is called "Energy Trilemma Index" which ranks countries concerning their ability to provide sustainable energy policies on the base of 3 dimensions: Energy Security, Energy Equity and Environmental Sustainability. On the face of those definitions we can highlight some relevant dimensions for having energy security: 1) internal production: 2) logistics infrastructures 3) capability to exploit a differentiated bundle of energy resources 4) dealing with reliable suppliers.

2.2 Energy Policy of the European Union

The Treaty on the Functioning of the European Union (TFEU) directly states what is meant by energy security and gives a guideline for the energy policies of individual states establishing the principles by which these are to be formulated. In the article 194, it is established that "*Union policy on energy shall aim, in a spirit of solidarity between member states, to: (a) ensure the functioning of the energy market; (b) ensure security of energy supply in the Union; (c) promote energy efficiency and energy saving and the development of new and renewable forms of energy; and (d) promote the interconnection of energy networks*".
Moreover, the European Parliament and the Council are in charge of determining how to achieve these guiding principles, after consultation with the Economic and Social Committee and the Committee of the Regions, not affecting any "*member state's right to determine the conditions for exploiting its energy resources, its choice between different energy sources and the general structure of its energy supply*".
According to the European Commission [17], the European Union currently imports 53% of the energy it consumes (almost 90% of its crude oil, 66% of its natural gas and 42% of solid fuels). In 2013 the demand for external energy was worth approximately 400 € billion of which 300 € billion only for crude oil and oil products.
The energy dependence rate[3], was 53,4% in the EU28 in 2012 (table 1), nearly stable since 2008.

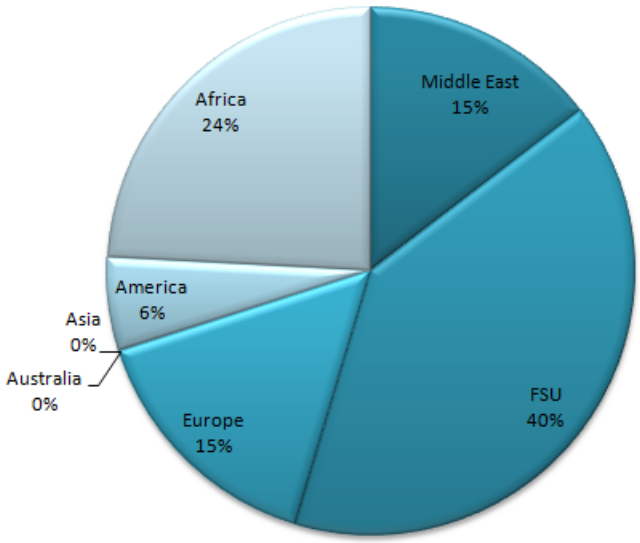
[3]The energy dependence rate is a percentage that stands for imports minus exports divided by gross consumption. Gross consumption is equal to the sum of the gross inland consumption and the fuel (oil) supplied to international marine bunkers. A positive dependency rate indicates a net importer of energy. A value lower than 100% occurs when net imports are lower than gross consumption. When the value is greater than 100%, energy products are supposed to be stored in inventories in the year of import.

Table 1: Dependency rates, 2013. - Eurostat Data

Area	Tot Dependency	Tot petroleum prod.	Natural Gas	Solid Fuel
Belgium	77,5	102	100,5	95,1
Bulgaria	37,8	103,7	93,2	16,4
Czech Rep.	27,9	96,3	100,2	-11,6
Denmark	12,3	-13,7	-23,1	90,7
Germany	62,7	96,1	87,2	44,5
Estonia	11,9	59,9	100	-0,1
Ireland	89	100,2	95,9	72,4
Greece	62,1	94,2	100	3,2
Spain	70,5	97,4	98,6	70,3
France	47,9	98,9	97,4	93,4
Croatia	52,3	77,1	31,8	110,1
Italy	76,9	90,7	88,1	96,2
Cyprus	96,4	101		100
Latvia	55,9	100,4	115,6	88,8
Lithuania	78,3	93,2	100	99,7
Luxembourg	96,9	100,3	99,6	100
Hungary	52,3	83,9	72,1	29,5
Malta	104,1	104,6		
Netherlands	26	94,7	-86,8	111,6
Austria	62,3	92,9	75,5	93,8
Poland	25,8	91,3	74,2	-10,4
Portugal	73,5	97,2	101,5	95,4
Romania	18,6	47	11,9	18,9
Slovenia	47,1	95,8	99,6	19,4
Slovakia	59,6	88,5	95,6	80,6
Finland	48,7	106,2	99,9	65,7
Sweden	31,6	101,5	99,1	82,4
UK	46,4	39,8	50,1	82
Iceland				
Norway	-470,2	-456,7	-1566,7	-87,4
Switzerland				
Montenegro	26,6	100		-1,2
Macedonia	47,9	93,7	100,1	9,7
Albania	25,1	25,6	0	99
Serbia	23,6	58,2	80,5	3,4
Turkey	73,3	92,5	97,8	54,7
EU-28	53,2	**87,4**	65,3	44,2

Differentiating the analysis by type of commodities 1 we can see that crude oil is the commodity for which Europe is more dependent as net importer. In fact, the dependency rate for all petroleum products is 86.4% for the EU28; almost all countries have a value higher than 80% apart from the Estonia, Croatia and Romania, Uk, Albania, Serbia with respectively 59.9%, 77.1%, 47%, 39,8%, 25.6%, 58,2%. The only net exporters of crude oil are Denmark with a value

Figure 1: Volume of Crude Oil Imports in the European Union (EU27) Period 1-12/2014 by origin. (EXTRA EU)



Source: EUROPEAN COMMISSION Directorate-General for Energy.
Note: the value is calculated using CIF prices

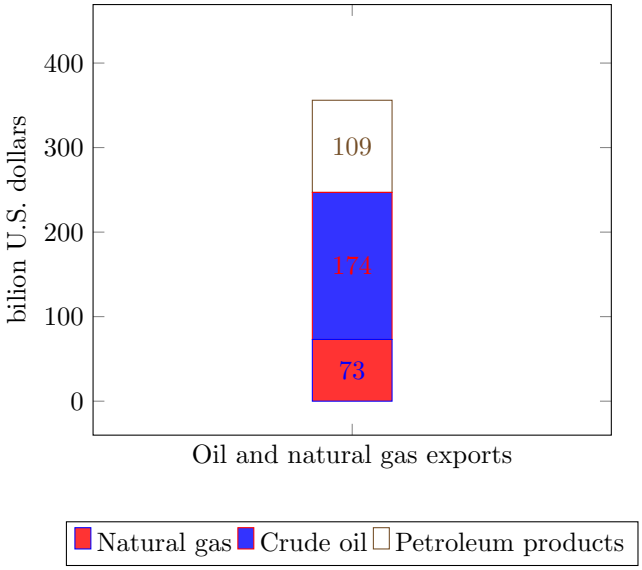
of -13.7% and Norway -456,7%. Over the last twenty-three years, this trend is consolidating. Europe owns only 0.6% of global oil and 2.0% of natural gas reserves available in the world. This situation makes Europe vulnerable to external energy suppliers, especially to those providing crude oil, which is the first fuel used in Europe.
Indisputably, Russia is the European Union's largest supplier. According to Eurostat (dataset:nrg123a), in 2014 the EU's (EU28) three biggest suppliers of crude oil (without NGL) are respectively Russia, Norway, and Nigeria with respectively 28,86%, 12,44% and 8,67%. According to the European Commission (Directorate-General for Energy)[4] if we also consider the Former Soviet Union nations (FSU) still under the economic influence of Russia, they supply to EU27 about 40% of the total value imported (see figure 1). In total, Europe imports 3.652.941 per 1000 bbl of crude oil. Former Soviet Union (FSU) nations supply the EU with around 1.462.333 per 1000 bbl of crude oil.
According to EU forecasts the overall dependence on energy imports is expected to grow if alternative energy sources would not be implemented in the nearest future.
In recent years the EU tried to make progress in the construction of an effective and sustainable energy supply plan. The greatest concern arises from the high level of dependence that some countries, particularly the east European ones, face on foreign suppliers such as Poland and Slovakia. Therefore, the overdepen-

[4]see eu-coi-from-extra-eu-2014-01-12.pdf that can be found at the following link http://ec.europa.eu/energy/sites/ener/files/documents/crude-oil-imports2014.zip

dence on Russia and the consequent monopolistic/oligopolistic pricing become the big problems to solve.
The European Union has established guidelines for that purpose: diversification of routes and suppliers, reduction of demand, transparency of pricing mechanisms and solidarity in the region. However, the low bargaining power poses huge problems for attaining those objectives. All efforts aimed at improving cooperation between the two economic areas would seem desirable for both. The high level of mutual dependency should induce both Russia and the EU to promote a "respectful relationship".
During the period the Cold War, EU and USSR adopted stable Import-Export relations. Today, both actors have made efforts to improve the "cooperation" in the field of energy that demonstrate the acknowledge of their mutual strategic significance. However, Eu and Russia are now pursuing conflicting objectives. Europe is intent on building a very ambitious plan of a "liberal" EU consumer market, although it plays the weaker role of the net "importing nation" of energy products, and Russia pursues aspirations to increasingly influence the European market with the intent to maximise strengthen "monopoly" position and therefore their profits.
Several attempts at dialogue have taken place over the last 15 years. One of them is the 6th EU-Russia Summit in Paris in October 2000 where the primary goal was " to provide reliability, security and predictability of energy relations on the free market in the long term and to increase confidence and transparency on both sides." However, after 15 years political reasons are prevailing on issues of a purely economic strategy. Notice that a large part of Russian federal budget depends on revenues from oil and natural gas activities. According to the Ministry of Finance, one-half of the Russia's federal budget revenue in 2013 came from gas and oil export customs duties and mineral extraction taxes. Moreover, energy exports play a central role in the development of Russian economy. According to EIA, oil and natural gas sales accounted for 68% of Russia's total export revenues in 2013 where 33% came from crude oil sales 21% from other petroleum products and 14% from natural gas 2. Notice that crude oil exports alone were greater in value than the value of all non-oil and natural gas exports. According to Kuboniwa et al.(2005) [43] the share of oil and gas sector in Russian GDP is likely to be underestimated by the official GDP statistics due to the prevalence of the transfer pricing.
The EU would likely prefer to maintain energy relations based on cooperation and interdependence with its largest energy supplier. At the same time Russia is adopting two strategies for preventing the achievement of European objectives: 1) entering into individual agreements with EU member states, with the aim of disrupting the Community plan of a single energy policy, by leveraging on the sovereignty of the various states in the matter of preparing the energy supply portfolio; 2) actively participating in projects with the intent to discourage Europe to adopt programs of diversification of energy suppliers. European position regarding Russia differs by countries[5] and Russia would seem to exploit these differences to its advantage and undermining the "European unity". Despite the cooperative efforts undertaken in recent years it is almost impossible to imagine a future based on the principle of transparency and the rules of the common market.

[5]For example, Finland is fully dependent on Russian gas; Spain receives none

Figure 2: Russia gross export sales, 2013



Source: U.S. Energy Information Administration, Russia Federal Customs Service
Note: Natural gas includes liquefied natural gas (LNG) sales.

2.3 The gas and oil consumption in the European Union

The energy security debate in Europe today focuses mainly on gas. At present, the old continent still bets on natural gas as the fuel of the future since "its green properties" and the availability of the technology allowing potentially its efficient use. This will probably induce the EU to grow natural gas consumption, in the light of the European Climate Change Programme (ECCP) which consider switching from coal to natural gas as an efficient choice to drive the EU greenhouse gas (GHG) emissions reduction. According to BP Energy Outlook 2035 [14] "*LNG net imports almost triple by 2035 and account for 30% of consumption in 2035*". Russian gas supply via pipeline is a primary source of supply, and, according to BP Energy Outlook 2035 [14], it is expected to grow by 15% and maintain a market share of around 31% by 2035. According to Eurogas, as cited in Bilgin [12], natural gas consumption will be critical for the economic future of the EU growing from 438 mtoe in 2005 to 625 mtoe in 2030. These expectations lead the European Commission to focus most of its attention on the gas market leaving the question of crude oil on the back burner since it can take advantage of an international and well-functioning market. Usually, crude oil is not part of the energy policy, **although it represents still more than one-third of the energy mix of the European Union**. Problems like the liquidity of the global oil market and the fundamental dependence of transport on oil do not create the same worries as for gas, and it is even not addressed in the European energy security policy. However, the question would probably deserve more attention. Although crude oil does not represent a strategic factor of

the future European energy portfolio, according to recent programs, the reduction of the consumption of such commodities will be relatively slow. According to BP Energy Outlook 2035 [14] in 2035 it is expected that oil and gas each will account for 29% of consumption followed by renewable which overtakes coal.

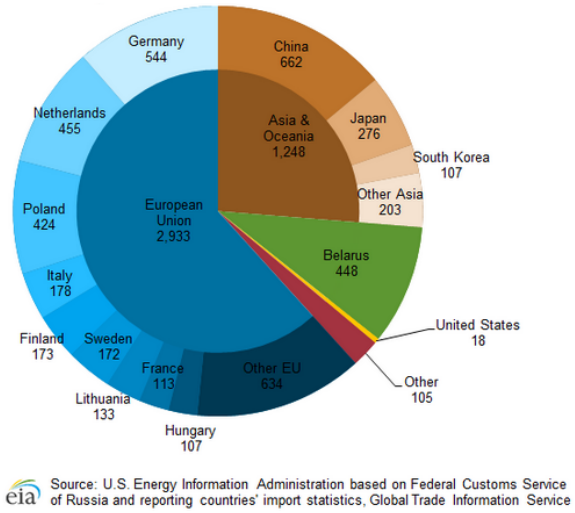
3 Russian Supply

3.1 Russian Federation as a World Energy Supplier

The huge Russian energy production capacity has been one of the primary levers that allowed Russia to increase its economic power and political stability. Russia has been and continues to be one of the most influential global energy suppliers in almost all energy sectors. According to the latest EIA's International Energy Statistics [24] Russia is the world's top crude oil (including lease condensate) producer with average liquids production of 10.551 million barrels per day (b/d) in 2016. It is also the third top world producer of petroleum and other liquids (after the United States and Saudi Arabia) with average liquids production of 11.240 million barrels per day (b/d) in 2016. Russia also produces significant amounts of primary coal and nuclear power making it be the sixth-largest major coal producer behind China, United States, India, Australia, and Indonesia with 393 million short tons in 2014 and the third-largest producer of nuclear power in the world in 2014. It is no coincidence that the top 4 products exported in 2014 by Russia are energy products: Petroleum (35%), Refined Petroleum (22%), Petroleum Gas (7%), Coal Briquettes (3.0%) (see The Atlas online[6]). Russia is not only a great producer but also a natural reserve of energy products for refining, owning in June 2016: 17,3 percent of world's total natural gas reserves, 17,6 percent of coal, and 6 percent of oil according to BP world statistics [15]. This enviable position has made this nation self-sufficient in fuels and power generation and a major exporter of energy products in the world. The latest energy plans announced by the Russian government are trying to push forward and reform the entire production industry, inducing the gradual replacement of exporting raw products with refined products to retain in the country all of the potential value added. The geographical position of Russia, in particular, the proximity to the Caspian Sea and Central Asian, allows for new business deals making Russia a key nation for energy production but also for the transit of the energetic products. Russia has been supplying about a third of Europe's oil and natural gas consumption, but, at the same time, it is also starting to export more to the energy-hungry East Asian markets. As if is not enough, Russia has not only vast proven reserves, but it is very likely that it also owns large deposits not yet counted in the regions to the East Siberia. According to the Oil and Gas Journal [22] (as cited in EIA 2016 [24]), Russia possess 80 billion barrels of proved oil reserves, which includes the brand new deposits in western Siberia province. Although most of Russia's oil production originates in West Siberia and the Urals-Volga regions, it is very likely that the production from East Siberia, Russia's Far East and the Russian Arctic will grow in the future.

[6]"The Atlas of Economic Complexity," Center for International Development at Harvard University, http://www.atlas.cid.harvard.edu

Figure 3: Crude oil and condensate exports by destination in 2014



Thousand barrels per day.
Source: EIA 2015

3.2 Russian Oil Export Supply

According to U.S. Energy Information Administration (EIA), Russian crude oil exports including lease condensate scored a relevant increase of 96.7% from 2648,36 thousand barrels per day in 1999 to 5211.14 thousand barrels per day in 2004. After this period a slight and persistent downward trend arose.
Revenues from oil exports are a primary source for both the Russian economy and government. They account for more than 68% of the value of total exports, based on information provided by Russia's Federal Customs Service. Mineral extraction taxes and export customs duties on oil and natural gas are responsible for about one-half of Russia's federal budget revenue in 2013 according to the Ministry of Finance. Additionally, crude oil exports alone were greater in value than the value of all non-oil and natural gas exports and produced value four times higher respect to natural gas. Europe has confirmed to be the primary destination of the exported flows with, in particular, 72% of the crude headed to Germany, Netherlands, Belarus, and Poland. However, the most recent data leave suspect of a sudden change of direction since Russian government is letting grow its ties in the east. Asia accounted for 26% of Russian crude exports in 2014, with China and Japan accounting for a growing share of total Russian exports. In May 2015 Russia become China's largest supplier of crude oil for the first time. Exports to China have more than doubled since 2010. According to Eastern Bloc Research other oil products are delivered with lower volumes in 2014, about 960,000 b/d of diesel and 1.6 million b/d of fuel oil, 100,000 b/d of gasoline and 60,000 b/d liquefied petroleum gas during the same year.

3.3 Relevant Factors Affecting Russian Oil Export Supply

3.3.1 Russian Oil Production

According to EIA, in 2014 Russia ranked third in the world for oil production, after Saudi Arabia and the United States, and first for crude oil supplies to Europe. In 2014 Russia produced an estimated 10.72 million b/d of petroleum and other liquids (of which almost 94% were crude oil including lease condensate)2. The production of crude oil of the modern Russian state is constantly growing since 1998 when production reached the level of 5.854 thousand barrels per day. During the period between 2000 and 2004, there was a relatively fast increase in productivity with the average rate of annual growth exceeded 7.5% per year and peaks of 9.8% in 2003 and 8.3% in 2004. However, from 2004, the speed of production growth has suffered a contraction with the annual growth slowed down to 2.7% in 2005, 2.2% in 2006 and 2% in 2007. From 2010 to 2014 the annual growth has reached the average level of 1.2%. It is likely that the deceleration was actually due to the hike in crude oil price that has discouraged large Russian companies to produce more given the wide profit margins. According to EIA, the annual average price of Russian crude oil rose gradually from $34.52 per barrel in 2004 to $94.77/barrel in 2008. In the background, this was also made possible thanks to the limited participation of foreign investors, due to government restrictions imposed to the cooperation in projects for developing national strategic oilfields. Usually, governments seek foreign investments, but Russian government wanted to protect the oil sector from outside influences especially during a period, characterised by high oil prices, when usually the big oil companies prefer to launch long-term programs.

3.3.2 Oil Transfer

Russian crude oil is roughly entirely transferred via pipelines. However, there are also alternative rail and sea routes available for reaching bordering countries or ports for exports. Transneft is the leader company in the sector with largest Russian pipeline network. Only small volumes of exports are shipped via rail and on vessels that load at independently-owned terminals. Russia uses three main channels of distribution of petroleum products headed to the West: Druzhba, which conveys oil to the Europe, Baltic Pipeline System 1 and Baltic Pipeline System 2, which have a strategic role of diversification of routes to reduce the degree of dependence on Druzhba pipeline route (Black Sea pipeline). All of these oil pipeline network exceptions of the Tengiz-Novorossiisk are under the control of the state owned Transneft company(IEA, 2010). Not very long time ago Russian government promoted new projects of construction of new oil pipelines intending to diversify the destination of petroleum products destined for export. The new routes pointing to the east (US, China and Japan) are: the Trans Sakhalin pipeline, Purpe-Samotlor Pipeline and the Eastern Siberia-Pacific Ocean (ESPO) Pipeline. EIA estimates that ESPO will have a capacity of 2.6 million b/d by 2020. The opening of this channel and the design and construction of new pipelines to the east is interpreted by many analysts as the choice of Russia to give priority, in the near future, to the supply of the Eastern countries rather than the western ones. New projects are pushing Russia to tie agreements with new Asian superpowers like China and India.

3.3.3 Russia's oil grades

Russia has several oil grades, Urals, Siberian Light, Sokol, Sakhalin (Ex Vityaz), REBCO and ESPO. The largest share of export contracts concerns the Urals blend oil supplied through the Baku-Novorossiysk pipeline and the Druzhba pipeline system. Urals blend consists of a mix of heavy sour crude from the Urals-Volga region and light sweet crude from West Siberia (Siberian Light crude). Markets consider this blend to be of inferior quality compared to the most famous Brent, for which it is underpriced with a spread. Siberian Light crude is a higher quality brand and thus more profitable when marketed on its own. However, the lack of adequate infrastructure makes full exploitation of this resource impossible at present and then it continues to be exported mainly by mixing it into the Urals blend. Other valuable oil grades are Sokol and Vityaz blend grade: the first is a light, sweet crude with an API gravity of 35.5° and 0.30% sulphur content according to ExxonMobil. According to Heinrich [30], Vityaz was a light (34.6° API), sweet (0.22% sulphur content), crude.
In 2014, Vityaz blend was replaced with a new grade of crude called Sakhalin blend which is loaded at the Prigorodnoye port, on the southern tip of Sakhalin Island. This oil grade is now delivered to Asian nations like: Japan, South Korea, Singapore and Indonesia. The Eastern Siberia-Pacific Ocean (ESPO) blend is a new mix of crude produced in several Siberian fields. The streaming began in December 2009 out of the Russian Eastern port of Kozmino, near Vladivostok. Since December 2009, this crude has been delivered both to Asia and the US. According to PLATTS [55], ESPO blend is a sweet, medium-light blend, with a gravity of 34.7°API and 0.535% sulphur content. The grade is exported to Asian countries like China through the recently constructed ESPO Pipeline or the Pacific coast port of Kozmino.

3.3.4 Taxation

Russia enforces two forms of peculiar taxation on crude oil: the minerals extraction tax (royalty) and the export tax. Crude oil boasts a special export tax regime if compared to other products. In fact, export tax on raw crude oil was being set up higher than other products to stimulate investment in refining capacity. Regarding minerals extraction tax, a form of facilities (or tax holidays) is being provided to encourage the extraction of the so-called difficult-to-develop resources. However, adjustments are frequently made by the government. For example, the last rebalancing between oil regimes, on January 1, 2015, when Russian government decided to raise the extraction tax and lower the export tax as a compensation for the increase in oil extraction. The taxation system remains one of the objectives of the government that imagines spurring development of difficult-to-develop resources despite the production at fields, set up mainly during the Soviet era, is falling down. As reported by Bloomberg in an article at the end of Agust 2015 [48]: "*Russia may lose about 100 million metric tons of output in 10 years at its key West Siberia fields, Energy Minister Alexander Novak said in June*". That's the reason why recently the Finance Ministry proposed reducing the reliance of taxes on duties tied to crude production with a new levy on earnings which should introduce a charge of 70 percent on profit from oil projects and would keep Russia's export tax. According to Ernst & Young The new tax system would start to be tested realistically not

earlier than 2017, and it would reduce the burden of the tax on the price of crude oil from 53% to 30% at the price of $45 per barrel.

3.4 Energy Strategy of Russia for the Period up to 2030

The Russian government in 2009 announced the launch of a long run strategic twenty-years energy program. The stated goal of the new strategy is to promote sustainable economic growth by exploiting the energy sector according to a principle of maximum efficiency. This program was launched to integrate and modify a previously carried out program of 2003. Three is the number of pillars on which the program rests: 1) the state is committed to conducting persistent activities in order to direct the development of key projects in the energy sector; 2) to create companies that simultaneously operate in the domestic market and the overseas market, representing Russia; 3) to support, through appropriate measures, companies that promote investment initiatives in areas where they can achieve objectives aligned with the state interests.
The Russian government has set precise production targets: 1) Crude oil production will rise close to the technical-economical upper limit and at the same time 2) The industry will contribute to state revenues and export income; 3) Russian government decided to promote and diversify the export destinations by the development of crude oil export pipelines and oil-shipping ports. Priority projects are the following crude oil pipelines: "Construction of the Burgas–Alexandroupolis crude oil pipeline," "the second-phase Baltic pipeline system" - Petroleum products pipelines: "Cebep (north)," "Iot (south)" - Crude-shipping ports: "Primorsk," "Ust-Luga," "Nakhodka".
Russia's total annual crude oil production is projected to increase up to 525 million tons by 2022 and up to 535 million tons by 2030. Major crude oil producing regions in Russia Western Siberia, Volga and Ural are expected to see the gradual and ongoing reduction in production. Planned productions in Eastern Siberia and the Far East are believed to can gradually replace the running out productions up to became the major oil production regions. However, to achieve the goal, massive investments in oil exploration and development projects will be needed, and Arctic offshore and shale resources are unlikely to be developed without the help of Western oil companies.
In the future, the Russian oil sector will assume most of the value added through the transportation of crude oil, and production and exports of high-value petrochemical and energy products thanks to a programmed technologically innovative revitalization of the oil sector. The Russian government plans to promote the gradual reduction in exports of crude oil, but at the same time, it will increase the export of petroleum products to hold back value added benefit of such productions to the economy. Internally the industry will stably, sustainably and efficiently satisfy Russia's domestic demand for crude oil and petroleum products.

3.4.1 Effect of recent sanctions and prices on Russian strategy

Recently, in order to attract sufficient capital to develop the twenty-years program successfully, the Russian government has approved some fiscal measures aimed at attracting foreign capital by offering a special tax rate to companies that invest in the Arctic offshore and low-permeability reservoirs, including

shale reservoirs. Many firms have already signalled their interest in the past, and several Western companies have entered a partnership with Russian companies, attracted especially by prospective of a large gain. According to EIA [23], Rosneft has signed agreements with ExxonMobil, Eni, Statoil, and China National Petroleum Company (CNPC) to explore the Arctic fields, LUKoil has signed agreements with Total to explore shale resources. Moreover, Shell, BP, and Statoil also signed agreements with Russian companies to explore shale resources. However, it is unlikely that these programs will be implemented or at least they will not be scaled down.
The reasons for this view reside in two unexpected events: 1) the sanctions imposed by Western nations on Russia in response to the actions and policies of the government of Russia concerning the Ukraine; 2) oil prices fell by more than half, from March 2014 to January 2015. The expected effect will be a reduced propensity to create new contracts to finance expensive new projects like deepwater, Arctic offshore and shale projects. Regarding the still open contracts, it does not seem that sanctions have discouraged enough Western companies to conclude the business and they continue to seem interested in the potential profit of such contracts.

4 Import Demand for Crude Oil in Europe

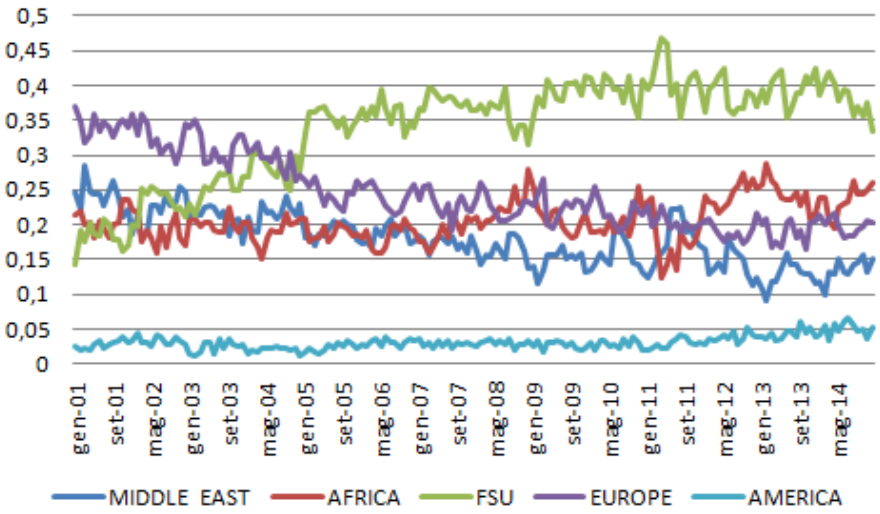
Over the years, Russia and other former Soviet Union countries (FSU[7]) have succeeded in acquiring a very significant weight in supplying energy commodities to Europe. Looking at the chart 4 you can see that the market for imports of crude oil is firmly under the control of those nations.
To understand what are the drivers of the European demand for crude oil on the foreign countries, we need to analyse what are the relevant factors that in general determines the demand in crude oil markets.

4.1 Spatial dimension of crude oil demand

Not all crude oil types are alike, factors such as the quality of crude oil, transportation costs, insurance costs, political and macroeconomic factors affect the relative costs of supply differently among different possible alternatives.
Probably some of the factors that more than others are determinant for the total refining costs of crude oil are the geographical position of the points of extraction and the relative position of the points of refining. The different qualities of crude oil available on the market, the transport costs and the related insurance costs and taxes depend mainly on the place of extraction and the local policy of the on-site distribution companies and governments.
For several decades oil was classified by the location of extraction. Obviously, physical distances are not a sufficient indicator to measure the real distances between the location of extraction and production. Real distances depend on the efficiency and effectiveness of the infrastructures located in the territories (transports and inventory capacity). There are four main modes of transportation of crude oil: pipeline, rail, truck, ship. Each of these has advantages and

[7]The FSU countries are: Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan.

Figure 4: Market shares of crude oil imports in Europe



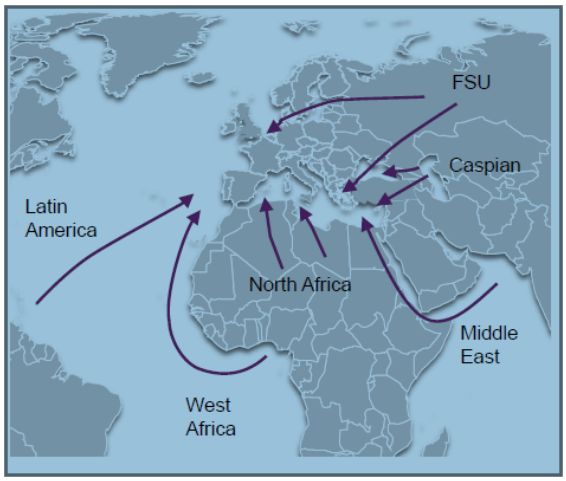
Source: based on Eurostat data

disadvantages affecting the actual distance between the producer and the extractor.
Pipelines are the most efficient and commonly used form of oil transportation. It permits to link altogether, with a small impact on greenhouse gas (GHG) emissions, the wellhead to gathering, processing facilities, refineries and tanker loading facilities. Another important factor is the national geopolitical strategy. It's worth remembering that the geographical distances do not matter when political reasons affect the terms of trade, as it happens whenever an embargo is imposed. The direction and the intensity of the crude oil import-export flow follow uneven relations of power, such as those connecting the EU to the rest of the world. Remember for instance the effect, on European oil import strategy, of the 1973 OPEC oil export embargo endorsed by many of the major Arab oil-producing states, in response to Western support of Israel during the Yom Kippur War. In that occasion, Europe intensified oil trade with the USSR and reduced the supplies from OPEC countries.

4.2 Quality dimension of demand

The oil industry classified crude oil into different types in order to measure the physical characteristics of the traded oil. For physical characteristics, we mean all the features that influence the production costs of every manufactured product derived from petroleum. This specification allows defining more precisely the intrinsic value of the oil to be processed. The lower are the costs of refining required for processing the oil the higher is its value. The market typically differentiates oil types by their densities (measured as API gravity) and their sulphur content. The density of oil is determined by the length of the hydrocarbons it contains. If the raw petroleum is of a high density the lower is the API

Figure 5: Map of the import possibilities



Source: Steve Cooper - Wood Mackenzie 2013

of crude oil (heavier crude oil). Sulphur is an undesirable characteristic of oil products. Types of crude oil containing high levels of sulphur are termed "sour", if they have relatively low levels of sulphur they are classified "sweet". Usually, oil blends are classified into four categories to refiners worldwide: light-sweet (30-40 API, $<$ 0.5 wt% S); light-sour (30-40 API, 0.5-1.5 wt% S); heavysour (15-30 API, 1.5-3.1 wt% S) and extra-heavy ($<$ 15 API and $>$ 3 wt% S). Refiners are generally willing to pay more for light, low sulphur crude oil. In that regard, Europe has potentially considerable flexibility over the selection of the quality since it is well-positioned to import crude from a huge number of different regions (see figure 5).
The high complexity of the refineries let Europe demand a mixed crude slate[8]. In fact, the industry of European refineries is not bounded by a limited upgrading capacity. Rather, a reduction of the overcapacity along with an increase in the demand for lighter crude oil supply, are likely to happen in the next future. The small net cash margins, due to the world's highest operating costs charged (see Lukach et al. (2015) [45]), are imposing the rationalisation of the production with the aim of the international competitiveness. Therefore, a reduction in crude oil supplied by Russia of any given quality would be potentially expected to result in the substitution by amounts of a better or equivalent quality from alternative attainable regions (see figure 5) ensuring falling operating costs.
According to Wood Mackenzie European crude demand will fall in the near-term and they predict crude slate will get sweeter and slightly heavier. In the long term, the decline of domestic and Russian crude should be compensated by increased Caspian, Middle East and America imports.

[8]Crude slate stands for the choice of crude oil used by refineries

4.3 Other dimensions of the demand

Several additional factors could also play a very relevant role. Some of them are: 1) European emphasis upon energy efficiency will likely influence future energy demand developments in the long run. For example, the European Union's (EU) biofuels target of 10% of energy content by 2020 in road transportation; 2) growing production in some regions could have an impact on the trade flows and storage in Europe; 3) political factors like wars can change the relative convenience between alternative supply areas; 4) macroeconomic factors, as economic growth, may result in a change in expected consumption that inevitably affects the business plans of refineries.

4.4 The effect of Russian Energy strategy on Eu crude oil demand

The expected reduction in exports of crude oil from Russia is not necessarily sufficient to compromise Europe's security of supply. The oil market could be big enough for substituting Russian oil with other sources. The European demand can absorb large volumes of crude qualitatively similar to Russia's Urals Blend. However, not all the potential suppliers can provide such type of crude oil. The European refinery companies were in big trouble in the last decade, and they are still continuing to be in the troubles today opting, most of the time, for reducing their exposure in the refining business. A cheap, high-quality crude oil supply would allow them to get higher value added. However, several factors, such as those mentioned above, like transport costs, taxes, etc. can heavily affect the demand. The search for an alternative to Russian oil, in the medium and long run, can not leave aside from the consideration of such factors. According to IEA a lower economic growth, a higher vehicle efficiency and the substitution of oil in transport with biofuels and natural gas will probably drop the demand for crude oil in Europe in a long-term decline. The surplus of capacity and weak refining margins of the European oil refining companies is one of the primary signals of the expected structural falling demand and a growing international competition of the Asia, Middle East and the USA. The long-term plans of the European refinery sector tend mainly to reduce the throughput volumes which will make even more complicated the search for an alternative commercial partner that should offset the falling Russian export that, according to the Energy Outlook of the Russian Academy of Sciences, will fall by almost half by 2040. However, the reduction in exports of some fuels to Europe is partially due to the willingness to anticipate the change in oil demand and the subsequent substitution effect between different petroleum products exported in Europe. For example, fuel oil is expected to be replaced by other fuels in the long-run. At present more than half of the cars in Europe are diesel-powered cars as a result of fiscal incentives. Therefore Russia is shifting towards exporting more diesel and jet fuel rather than fuel oil which is, at present, predominantly exported in Europe.

4.5 Petroleum Refinery Sector

From 2007, Europe is suffering from a severe crisis in the oil refining sector due to poor production margins that companies can generate, despite their great versatility. The awkward moment has often led several companies to sell their

plants or, alternatively, to convert them into terminals, thereby taking their capacity off the market forever. Two were the triggers, sliding demand amid a lingering economic recession and the resulting over-capacity. US, Middle East, India, and Russia can readily supply of gasoline, gas oil-diesel, jet kerosene, and all other hydrocarbon products at lower costs. Another important factor is the lower oil prices available on the international markets that are leading to sharp falls in European upstream investment. At the same time in Eastern Europe, Russian oil producers like Gazprom Neft announced plans to invest billions to upgrade refinery plants, with new advanced technologies, in order to produce better grades of products, increase the oil conversion rate, enhance energy efficiency. As a result, oil demand in Europe, already reached its lowest since at least 1995 and probably to the early 1990s, in a relatively short time, and is expected to fall further, pulled down by a bleak economic outlook.
Crude cost is the single most important determinant of the profitability of an oil company. With crude costs accounting for around 80% of the refinery expenditures, processing cheaper crude can have a very positive impact on refinery margins

5 Can Europe survive without Russian crude oil?

European demand reduction and simultaneous increase in Asian demand appear to be a sufficient evidence to persuade Russia to reverse course in export.
This turnabout involves for Europe an intrinsic geopolitical risk for three main reasons: 1) as a decline in production is inevitable and even though gas consumption is increasing at a very high rate, it is expected that oil products will presumably still remain the energy products with the highest demand for a long time; 2) the most profitable Asian trade, once intensified, could induce Russia to cut supplies to Europe at a rate higher than they would actually be able to bear; 3) geopolitical tensions between NATO countries and Russia suggest an embargo as a possible scenario which can be extensible to the former Soviet Union (FSU) nations which are still partially under the sphere of influence of Russia (consider that part of them belong to the Eurasian Economic Union). For these reasons it is necessary to check where the market, given the fragile situation of the European refining oil sector, may direct their demand in the absence of Russian supply.
Replacing 4800 thousand barrels per day of crude oil and lease imports from Russia in the short run could be a significant challenge, for European countries. To answer the original question we need to choose: a) a measure of the degree of substitution between crude oil supply of the former Soviet Union (FSU) countries and the others available on the international markets; b) a dataset; c) a theoretical approach for the estimate.

5.1 Methodology

5.1.1 The choice of the measure of substitutability

To find the most appropriate measure to be taken in this study, I will make a brief introduction of the principal measures of elasticity to allow a comparison.

Hicks was the first describing the concept of substitutability, often called the elasticity of substitution, in 1932 [32]. A number of other measures that generalizes the concept were provided years later (see Allen and Hicks (1934)[33], Allen (1938)[1], Uzawa (1962)[63], McFadden (1963)[49], Morishima (1967)[51], Blackorby and Russell (1989)[13]).
The original Hicks work, assumed two inputs, capital x_1 and labor x_2 and calculated the relative change in the input proportion x_1/x_2 due to the relative change in the marginal rate of technical substitution f_{x_2}/f_{x_1} while output Y was held constant.

$$\sigma = \frac{dLN\left(\frac{x_1}{x_2}\right)}{dLN\left(\frac{f_{x_2}}{f_{x_1}}\right)}$$

However this measure is suitable for the analysis of only two factors being the marginal rate of technical substitution f_{x_2}/f_{x_1} not determined uniquely otherwise.
In 1989 Blackorby and Russell [13]demonstrated, under the assumptions of perfect competition and profit maximization, the ratio f_{x_2}/f_{x_1} equals relative factor prices p_2/p_1 and exploited this finding for building a new elasticity measure they called Hicks' elasticity of substitution (HES)

$$HES = \frac{dLN\left(\frac{x_1}{x_2}\right)}{dLN\left(\frac{p_2}{p_1}\right)}$$

Others elasticity measures in a multifactor setting were built on the base of this definition.
One of the goals was to isolate the degree of elasticity between two goods among all those available on the market. The Hicks-Allen elasticity of substitution (HAES) is usually termed a measure of relative substitutability and it is based on the assumption that all inputs being flexible and cost minimised for fixed output

$$HAES_{ij} = \frac{\partial LN\left(\frac{x_i}{x_j}\right)}{\partial LN\left(\frac{p_j}{p_i}\right)}$$

Another popular measure of input substitution is the Allen partial elasticity of substitution (AES), introduced by Allen (1938). From Uzawa (1962)[63], the AES between two inputs, for a twice-differentiable cost function (TC) can be calculated by

$$AES_{ij} = \frac{\eta_{ij}}{S_j}$$

where $S_j = x_j \cdot p_j/C$ denotes the cost share of factor j while η_{ij} is a measure of absolute substitutability also called the cross-price elasticity

$$\eta_{ij} = \frac{\partial LN\left(x_i\right)}{\partial LN\left(p_j\right)}$$

this measure widely used in most of the empirical studies does not represent the degree of elasticity in the Hicksian sense as it does not provide any additional information respect to the cross elasticity, being it proportional to this value. Blackorby and Russell (1989) [13]show that the AES does not measure the ease of substitution. Moreover, the measure is not appropriate for input with a small cost share.
In 1989 the same two authors proposed an alternative measure, they called the Morishima Elasticity of Substitution (MES), that measures the change in relative factors for a change in the price level of one factor. This measure can be expressed in terms of measures cross-price elasticity and own-price-elasticity.

$$MES_{ij} = \frac{\partial LN\left(\frac{x_i}{x_j}\right)}{\partial LN\left(p_j\right)} = \frac{\partial LN\left(x_i\right)}{\partial LN\left(p_j\right)} - \frac{\partial LN\left(x_j\right)}{\partial LN\left(p_j\right)} = \eta_{ij} - \eta_{jj}$$

Alternatively, MES can be interpreted as the ratio of the relative change in the ratio of input i to input j to the relative change in the price of input j, for an infinitesimal change of that price.
In my study, I need a measure that let us get to know what is the expected substitution effect of the European aggregate demand, for crude oil imported from former Soviet Union (FSU) countries. We are solely interested in factors influencing the decisions of European refineries and oil producers of the former Soviet Union (FSU) area. Accordingly, we decide to perform the analysis letting one country price be constant to a variation of FSU crude oil price, to exclude the direct influence of this change in price in the other market. For that reason, a two-factor-one-price elasticity, where solely the price of factor j is flexible with all other prices being fixed, should be the more appropriate measure to employ in the analysis.

5.1.2 European production function

We will try to select a functional form of the production function observing several restrictions on the supply and demand functions, implied by economic theory, but letting it be sufficiently flexible to fit the data in the empirical analysis. Therefore elasticities of supply and demand will be not arbitrarily restricted by the only choice of the functional form.

5.1.3 The required data

We consider in the analysis a representative European refinery firm. We suppose that this firm takes the price as given and does not expect its output decisions to affect oil prices. If the market price changes, then the firm reassess its production decision. This model is built on the following five assumptions:
1. the European buyer acts as a price taker, meaning that the buyer decisions has no impact on the price charged for the crude oil. The buyer decides the amount to purchase that minimizes it's production costs taking the prices as given;
2. the market consists of many sellers. We will not assume anything about the sellers, meaning that the seller does not necessarily takes the price as given. Notice that this assumption does not violate the idea that that oil market prices are governed by an oligopoly, which usually is commonly supposed to be right;

3. we need to assume that the buyer does not care which seller provide the oil if all sellers charge the same price. Meaning that every crude oil sold by all sellers in the market is assumed to be homogeneous;
4. perfect information.
It is worth to be concerned about the assumption 3, among all the assumptions, since, as we already saw, that crude oil provided by a group of producers is usually qualitatively heterogeneous due to several factors like chemical properties or transports costs. However a way for ridding off all the heterogeneity, otherwise without our control, is to exploit the CIf prices[9] of the imported crude oil volumes. There are mainly two reasons: 1) in a deficit market, like the European one, the price of crude oil is primarily CIF, because a huge share of crude has to be transported; 2) the crude oil price has been set such that the CIF prices for different crude, from different parts of the world but of the same quality and quantity, equate when delivered to the same refinery. Consider the example of JP Favennec & R Baker (2001) [25] *"A Rotterdam refinery, able to buy Brent crude oil at $18/bbl FOB Sullom Voe (the Brent loading terminal in the Shetland Islands), will incur a freight cost of $0.40/bbl. For crude oil of the same quality, the refinery could pay a delivered cost at Rotterdam not exceeding $18 + 0.40/bbl, with no loss of profit. So, if an equivalent crude oil is available in West Africa and the freight rate from the loading port is $0.80/bbl, to be competitive the FOB price for the West African crude must be* $18.40 - 0.80$, *i.e. $17.60/bbl. This basic model applies everywhere, but it cannot be perfectly applied. There can be discounts as well as premiums against the delivered CIF price for a large number of reasons. Refiners must also pay port (harbour) dues which can vary enormously between countries and from port to port"*. Therefore in order to consider in the analysis most of the factors that affect trades, we prefer to employ CIF prices by country of origin.

6 Theoretical approach

6.1 Literature on interfuel substitution

According to Behar Stevens. (2009) [8] the oil price shocks in the 1970s spurred many studies on the elasticity of substitution between energy and other inputs in production. A big debate raised on the substitutability (or not) between energy and capital and on the difference between energy and different types of producing inputs. Over the years, the issue has attracted the attention of many researchers of the time whose works were based mainly on Diewert's (1971) [21] paper. Among the most relevant empirical energy demand analysis there are Berndt and Wood (1975) [11], Fuss (1977) [27], and Pindyck (1979) [54].
Those papers derive cost share (or input-output) equations applying Shephard's lemma and estimate the parameters with an econometric model, using relevant data and producing inferences about the demand for fuels. Berndt and Wood (1975) [11] were the first to estimate the elasticities of substitution with energy in a production function. The early studies by Hudson and Jorgenson (1974) [36],Berndt and Wood (1975)[11], Fuss (1977) [27] and Magnus (1979) [46]focused the attention on the degree of substitutability among production inputs

[9]CIF (cost, insurance and freight) price is the price for a crude oil cargo delivered to the discharge port

like capital, labour, materials and energy. They mainly found the same result of substitutability between energy and labour, but complementarity between energy and capital. Apostolakis (1990) [2] and Bacon (1992) [4] provided some surveys of the early studies of interfuel substitution elasticities in the OECD countries. Fuss(1977) [27] worked also on the substitutability among different energy inputs. He found evidence of substitutability for oil, gas, and coal, and no evidence between each of these energy inputs and electricity. Further studies obtained mixed results, see Uri (1979) [62], Considine (1989) [19], Hall (1986) [28]. However as Serletis (2015) [59] said, "*the major contributions in this area are quite outdated by now, since their data incorporate observations before the 1970s...*"
Research in this area focused on two different methodological directions to the investigation of interfuel substitution (energy elasticities) and the demand for energy. With the first direction one estimates long-run and short-run demand elasticities using respectively cointegration techniques and error-correction models, see Bentzen and Engsted (1993) [10]and Hunt and Manning (1989) [37].
Although the technique deals with econometric regularity issues has not a proper microeconomic foundation.
The second direction, developed by Diewert (1974) [21], requires both a differentiable form for the cost function, and the application of Shephard's (1953) [56] lemma to derive a demand system generation. Other relevant studies investigating interfuel substitution and the demand for energy are: Jones (1995) [39], Serletis and Shahmoradi (2008) [58], and Serletis et al. (2009), Serletis et.al.(2010) [59].
The majority of the works are based on locally flexible functional forms and, in particular, the translog, introduced by Christensen et al. (1975) [16] See, for example, Fuss (1977) [27] , Pindyck (1979) [54], Jones (1995) [39], and Urga and Walters (2003) [61], Serletis and Shahmoradi (2008) [58], Serletis et.al.(2010) [59].

6.2 Empirical specification

Following Fuss (1977)[27] and Serletis (2010) [59], it is assumed a neoclassical model of the aggregate production function of the EU industry with a KLEM production structure and a number m of energy inputs

$$y = f(K, L, E_1, E_2, ..., E_o, ..., E_m, M)$$

and the vector of factor prices p

$$p = (P_K, P_L, P_{E_1}, P_{E_2}, ..., P_{E_o}, ..., P_{E_m}, P_M)$$

where Y is EU gross output, E_k for $k = 1, ..., m$ are m energy inputs (o means crude oil input), L is labor input, M materials, K capital input.
We state the assumption of homothetic weak separability of the production function in the k energy sources[10] and of exogeneity of the factor prices p and the output level y (Shephard, 1953 [56])

[10] We impose the homothetic weak separability assumption because we suppose that the decisions concerning the quantity of imported crude oil are determined before those concerning the other quantities. Consequently we can introduce an underlying two-stage budgeting optimization process, see Appendix C. In order to use this optimization process we need to

$$y = f[K, L, E_1, E_2, ..., E_o(E_{o1}, E_{o2}, ..., E_{on}), ..., E_m, M]$$

where $\{E_{oi}\}_{i=1,...,n}$ is the set of crude oil import demands by i-th country of origin.
Under regularity conditions (see Appendix A and B) and using the described production function, Shephard (1953)[56] showed that, for the duality theory, the cost function corresponding to the homothetically weakly separable production function is weakly separable as well. Consequently the marginal rate of substitution between any two components E_{ok},E_{oj} with $k \neq j$ of E_o does not depend upon the value of all the other factors $K, L, E_1, E_2, ..., E_m, M$, see appendix C and D.
In order to study the elasticity of substitution among crude oil imports demands by country of origin $E_{o1}, E_{o2}, ..., E_{on}$, given the price vector $\pi = (P_{E_{o1}}, P_{E_{o2}}, ..., P_{E_{on}})$ and the input vector $z = (E_{o1}, E_{o2}, ..., E_{on})$ on the base of the assumption of homothetic weak separability, we can shrink[11] our attention just only on the aggregate cost function $C_{E_o}(\pi, y)$ (see appendix D) .

$$C_{E_o} = C(\pi, y) = min_{z_i \geq 0}\{\pi' z | f(x, t) \geq y\}$$

Under the hypothesis of a linear homogeneous production function in inputs it is possible to demonstrate that $C_{E_o}(\pi, y) = y \cdot C_{E_o}(\pi)$, where $C_{E_o}(\pi)$ is an unit cost function (see appendix F).
In the next chapter I impose a specific functional form to $C_{E_o}(\pi)$.

6.3 CES production function and the translog energy aggregate cost function

The next steps require the specification of the functional form of the unitary cost function for the aggregate imported crude oil C_{E_o}. The goal is to calculate constant substitution effects, meaning that a cost function with a constant elasticity of substitution and a (CES) production function must be employed. A formal specification of a CES production function with two inputs was developed by Arrow et al.(1961) [3]

$$y = \gamma \left(\delta x_1^{-\rho} + (1 - \delta) x_2^{-\rho}\right)^{-\frac{\nu}{\rho}}$$

where y is the output quantity, x_1 and x_2 are the input quantities and ρ, γ, δ and ν are parameters. To extend the CES class function to n input factors, we consider, is in the class of nested CES function. The functional specification of the n-input CES function is

$$y = \gamma \left(\sum_{i=1}^{n} \delta_i x_i\right)^{-\frac{\nu}{\rho}}$$

assume that firms decide the foreign supplies before considering quantities of the other energy and non-energy inputs. See (Pindyck (1979) [54] Mountain Hsiao (1989) [35]; Kemfert Welsch (2000)[40] ; Klepper Peterson (2006) [41]).

[11] The homothetic separability let us investigate the substitution possibilities among crude oil supplies without concerning at the substitution possibilities among those crude oil supplies and the other commodities

As the non-linearity of the CES function does not permit linearization analytically, it is frequently approximated by the so-called "Kmenta approximation" (see Kmenta (1967) [42]), which can be calculated by linear estimation techniques. Alternatively, it can be implemented a non-linear least-squares using various optimisation algorithms. Hoff (2004) [34]showed that a correctly specified extension to the n-input case requires non-linear parameters restrictions on a translog function. Hence, there is a quite limited benefit in using the Kmenta approximation in the n-input case.
In this paper, and in order to be consistent with the existing empirical literature, I'm going to employ a translog cost function to investigate energy demand issues and oil suppliers substitution as in Serletis et.al.(2010) [59].

$$LN(C_{E_o}) = LN(\alpha_0) + \sum_{i=1}^{n} \beta_i LN(P_{E_{oi}}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} LN(P_{E_{oi}}) LN(P_{E_{oj}})$$

where i, j are individual energy types, and $P_{E_{oi}}$ is the price of crude oil imported from country i.
Under regularity conditions (see appendix B), some parameter restrictions are imposed in order to satisfy the assumption of homogeneity of degree one in prices of the cost function.

$$\sum_{i=1}^{n} \beta_i = 1$$

$$\sum_{i=1}^{n} \gamma_{ij} = \sum_{i=1}^{n} \gamma_{ji} = 0, \forall j = 1, ..., n$$

6.4 The Cost Share Equations

The n cost share equations for the individual energy types $\{S_{E_{Oi}}\}_{i=1,...,n}$, can be obtained in terms of shares of the i-th supply cost $P_{E_{oi}} \cdot E_{oi}$ in the aggregated cost of the crude oil imports, C_{E_o}, by differentiating the log of the cost function with respect to the log of the price of input i and using Shephard's Lemma, on C_{E_o},

$$\frac{\partial LN(C_{E_o})}{\partial LN(P_{E_{oi}})} = \frac{P_{E_{oi}}}{C_{E_o}} E_{oi} = S_{E_{Oi}}, i = 1, \ldots, n$$

see appendix E.
We can now calculate the share equation by differentiating the translog cost function

$$S_{E_{Oi}} = \frac{\partial LN(C_{E_o})}{\partial LN(P_{E_{oi}})} = \beta_i + \sum_{j=1}^{n} \gamma_{ij} LN(P_{E_{oj}}), j = 1, \ldots, n$$

The resulting cost share equations 1 2 3 5 can then be used to investigate the demand for individual oil import supplier, $E_{o1}, ..., E_{on}$ and to estimate the structure of substitution among the different import oil suppliers.

The analysis has been developed considering data aggregated by regions as follows:[12]:

- Af: Africa
- F: Former Soviet Union Countries
- E: Europe
- Am: America
- Me: Middle East

$$S_{Af} = \beta_{Af} + \gamma_{AfAf}LN(P_{Af}) + \gamma_{AfF}LN(P_F) + \gamma_{AfE}LN(P_E) + \gamma_{AfAm}LN(P_{Am}) + \gamma_{AfMe}ln(P_{Me}) \quad (1)$$

$$S_F = \beta_F + \gamma_{FAf}LN(P_{Af}) + \gamma_{FF}LN(P_F) + \gamma_{FE}LN(P_E) + \gamma_{FAm}LN(P_{Am}) + \gamma_{FMe}LN(P_{Me}) \quad (2)$$

$$S_E = \beta_E + \gamma_{EAf}LN(P_{Af}) + \gamma_{EF}LN(P_F) + \gamma_{EE}LN(P_E) + \gamma_{EAm}LN(P_{Am}) + \gamma_{EMe}LN(P_{Me}) \quad (3)$$

$$S_{Am} = \beta_{Am} + \gamma_{AmAf}LN(P_{Af}) + \gamma_{AmF}LN(P_F) + \gamma_{AmE}LN(P_E) + \gamma_{AmAm}LN(P_{Am}) + \gamma_{AmMe}LN(P_{Me}) \quad (4)$$

$$S_{Me} = \beta_{Me} + \gamma_{MeAf}LN(P_{Af}) + \gamma_{MeF}LN(P_F) + \gamma_{MeE}LN(P_E) + \gamma_{MeAm}LN(P_{Am}) + \gamma_{MeMe}LN(P_{Me}) \quad (5)$$

7 Parameter estimation and data

Notice that we have 30 parameters, 6 in each of the five share equations.
The imposition, however, of the ten symmetry restrictions
$\gamma_{FAf} = \gamma_{AfF}, \gamma_{EAf} = \gamma_{AfE}, \gamma_{AmAf} = \gamma_{AfAm}, \gamma_{MeAf} = \gamma_{MeAm}, \gamma_{EF} = \gamma_{FE}, \gamma_{AmF} = \gamma_{FAm}, \gamma_{MeF} = \gamma_{FMe}, \gamma_{AmE} = \gamma_{EAm}, \gamma_{MeE} = \gamma_{EMe}, \gamma_{MeAm} = \gamma_{AmMe}$, due to the regularity conditions on $LN(C_{E_o})$, restricts to 20 the number of the unknown parameters, moreover the assumptions of homogeneity of degree one in the prices restrict the number of free parameters to 14.

$$\beta_{Af} + \beta_F + \beta_E + \beta_{Am} + \beta_{Me} = 1$$

$$\gamma_{AfAf} + \gamma_{AfF} + \gamma_{AfE} + \gamma_{AfAm} + \gamma_{AfMe} = 0$$

$$\gamma_{FAf} + \gamma_{FF} + \gamma_{FE} + \gamma_{FAm} + \gamma_{FMe} = 0$$

[12]Therefore S_{Af} is the share equation with respect to Africa, S_F is the share equation with respect to former Soviet Union (FSU) countries, S_E is the share equation with respect to Europa, S_{Me} is the share equation with respect to Middle East. Notice that the wording Europe accounts for cif prices and quantities of every EU country that imports crude oil from foreign nations belonging in turn to EU

$$\gamma_{EAf} + \gamma_{EF} + \gamma_{EE} + \gamma_{EAm} + \gamma_{EMe} = 0$$
$$\gamma_{AmAf} + \gamma_{AmF} + \gamma_{AmE} + \gamma_{AmAm} + \gamma_{AmMe} = 0$$
$$\gamma_{MeAf} + \gamma_{MeF} + \gamma_{MeE} + \gamma_{MeAm} + \gamma_{MeMe} = 0$$

We implement the share equation system empirically by adding a stochastic component to each share equation, and we assume that the error vector is multivariate normally distributed with zero mean and the error covariance matrix Ω. Hence the share equation system can be written as

$$S_{Af} = \beta_{Af} + \gamma_{AfAf} LN(P_{Af}) + \gamma_{AfF} LN(P_F) + \gamma_{AfE} LN(P_E) + \gamma_{AfAm} LN(P_{Am}) + \gamma_{AfMe} ln(P_{Me}) + \varepsilon_{Af} \quad (6)$$

$$S_F = \beta_F + \gamma_{FAf} LN(P_{Af}) + \gamma_{FF} LN(P_F) + \gamma_{FE} LN(P_E) + \gamma_{FAm} LN(P_{Am}) + \gamma_{FMe} LN(P_{Me}) + \varepsilon_F \quad (7)$$

$$S_E = \beta_E + \gamma_{EAf} LN(P_{Af}) + \gamma_{EF} LN(P_F) + \gamma_{EE} LN(P_E) + \gamma_{EAm} LN(P_{Am}) + \gamma_{EMe} LN(P_{Me}) + \varepsilon_E \quad (8)$$

$$S_{Am} = \beta_{Am} + \gamma_{AmAf} LN(P_{Af}) + \gamma_{AmF} LN(P_F) + \gamma_{AmE} LN(P_E) + \gamma_{AmAm} LN(P_{Am}) + \gamma_{AmMe} LN(P_{Me}) + \varepsilon_{Am} \quad (9)$$

$$S_{Me} = \beta_{Me} + \gamma_{MeAf} LN(P_{Af}) + \gamma_{MeF} LN(P_F) + \gamma_{MeE} LN(P_E) + \gamma_{MeAm} LN(P_{Am}) + \gamma_{MeMe} LN(P_{Me}) + \varepsilon_{Me} \quad (10)$$

Notice that the shares sum to unity determines the random disturbances corresponding to the five share equations sum to zero that implies a singular covariance matrix of errors. Barten (1969)[7] discovered that the full information maximum likelihood estimates of the parameters can be obtained by arbitrarily deleting one equation, since the estimates are invariant to the equation removed. Following Barten, we decide to get rid off the share equation S_{Me} from the analysis. Moreover using the assumptions of homogeneity of degree one in the prices, we get the following set of four equations

$$S_{Af} = \beta_{Af} + \gamma_{AfAf} LN(\frac{P_{Af}}{P_{Me}}) + \gamma_{FAf} LN(\frac{P_F}{P_{Me}}) + \gamma_{EAf} LN(\frac{P_E}{P_{Me}}) + \gamma_{AmAf} LN(\frac{P_{Am}}{P_{Me}}) + \varepsilon_{Af} \quad (11)$$

$$S_F = \beta_F + \gamma_{FAf} LN(\frac{P_{Af}}{P_{Me}}) + \gamma_{FF} LN(\frac{P_F}{P_{Me}}) + \gamma_{EF} LN(\frac{P_E}{P_{Me}}) + \gamma_{AmF} LN(\frac{P_{Am}}{P_{Me}}) + \varepsilon_F \quad (12)$$

$$S_E = \beta_E + \gamma_{EAf} LN(\frac{P_{Af}}{P_{Me}}) + \gamma_{EF} LN(\frac{P_F}{P_{Me}}) + \gamma_{EE} LN(\frac{P_E}{P_{Me}}) + \gamma_{AmE} LN(\frac{P_{Am}}{P_{Me}}) + \varepsilon_E \quad (13)$$

$$S_{Am} = \beta_{Am} + \gamma_{AmAf} LN(\frac{P_{Af}}{P_{Me}}) + \gamma_{AmF} LN(\frac{P_F}{P_{Me}}) + \gamma_{AmE} LN(\frac{P_E}{P_{Me}}) + \gamma_{AmAm} LN(\frac{P_{Am}}{P_{Me}}) + \varepsilon_{Am} \quad (14)$$

We employ the dataset of statistics on EU crude oil imports[13] provided by the European Commission. The sample consists of monthly data for five macro

[13] http://ec.europa.eu/energy/sites/ener/files/documents/crude-oil-imports2014.zip

region areas: Africa (Af), former Soviet Union (FSU) countries (F), Europe (E), America (Am) , Middle East (Me), over the period Gen 2001 to Dec 2014 for a total of 168 observations in all. Statistics include: volume(1000 bbl), CIF price($/bbl), % of total imports for every region of origin. The study exploits CIF prices[14]. CIF prices are indicated to discriminate the relative convenience among the crude oil imports between different crude oil imports (see the subsubsection 5.1.3). The parameters are estimated at the first stage by a linear SUR estimator.

7.1 Linear constrained SUR Model

The first model adopted is a constrained linear seemingly unrelated regressions model (see appendix J). Although the estimate is statistically significant (see the estimation results in appendix table 15), it can not be considered economically significant because it violates, in almost every single point of the sample, the curvature assumption of the regularity conditions. Unfortunately, quasi-concavity cannot be guaranteed by any parametric restriction since the property is data dependent for the translog functional form and therefore the common estimation procedure, based on the linear SUR model, does not provide estimates satisfying concavity assumption in the aftermath. In this case inferences based on such a cost function are not meaningful consequently the concavity of the estimated function turns out to be a serious issue.
However, although the estimate is statistically significant, it can not be considered economically significant as well, because it violates, in almost every single point of the sample, the curvature assumption of the regularity conditions. Unfortunately, quasi-concavity cannot be guaranteed by any parametric restriction since the property is data dependent for the translog functional form and therefore the common estimation procedure based on the linear SUR model does not provide estimates satisfying concavity assumption in the aftermath. In this case, inferences based on such a cost function are not meaningful, consequently the concavity of the estimated function turns out to be a serious issue.

7.2 Local Concavity and Non-Linear SUR Model

Curvature requires that the cost function is a concave function of prices or, equivalently, that the Hessian matrix of the cost function, H, is negative semi-definite (see appendix B). In the case of a translog functional form, quasi-concavity can be imposed locally at a reference point.
As shown by Diewert and Wales (1987), the Hessian matrix, H, is negative semidefinite if and only if the following matrix is negative semidefinite

$$\boldsymbol{G} = \boldsymbol{\Gamma} - \underline{\boldsymbol{s}} + \boldsymbol{s}'\boldsymbol{s}$$

where $\Gamma = [\gamma_{ij}]$, $s = (s_1, \cdots, s_n)$ is the share vector, and $\underline{s}$ is the $n \times n$ diagonal matrix which has the share vector s on the main diagonal.
At the reference point (where all prices are set to one), the G matrix can be simplified as follows:

[14]The CIF price is defined by the Commission of the European Communities [18] as the price which "include the fob price (the price actually invoiced at the port of loading), the cost of transport, insurance and certain charges linked to crude oil transfer operations"

$$G_{ij} = \gamma_{ij} + \beta_i\beta_j - \delta_{ij}\beta_i$$

with $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.
Ryan and Wales (2000) [57]and Feng and Serletis (2008) [26], demonstrated that local concavity can be imposed by setting at the reference point $\boldsymbol{G = H = -KK'}$ as follows:

$$\gamma_{ij} + \beta_i\beta_j - \delta_{ij}\beta_i = (\boldsymbol{-KK'})_{ij}, i, j = 1, \cdots, n$$

where $\boldsymbol{K}$ is a lower triangular matrix.

Notice that the number of independent γ_{ij} equals the number of independent values K_{ij}.
Without considering symmetry, we know that since the elements of any row of the $\boldsymbol{\Gamma}$ matrix add up to zero and therefore only $(n-1) \times (n-1)$ elements of are linearly independent. Therefore, the $\boldsymbol{K}$ matrix must be of dimension $(n-1) \times (n-1)$ as well.
It can easily be shown that in our case with four fuels (n = 5), $i = Af; F; E; Am; Me$ the ij elements of $\boldsymbol{\Gamma}$ can be replaced by the k_{ij} elements of $\boldsymbol{K}$ as follows (see Feng and Serletis (2008)[26],)

$$\gamma_{AfAF} = -K^2_{AfAf} - \beta^2_{Af} + \beta_{Af}$$

$$\gamma_{FF} = -K^2_{FF} - K^2_{FAf} - \beta^2_{F} + \beta_{F}$$

$$\gamma_{EE} = -K^2_{EE} - K^2_{EF} - K^2_{EAf} - \beta^2_{E} + \beta_{E}$$

$$\gamma_{AmAm} = -K^2_{AmAm} - K^2_{AmE} - K^2_{AmF} - K^2_{AmAf} - \beta^2_{Am} + \beta_{Am}$$

$$\gamma_{AfF} = -K_{AfAf}K_{FAf} - \beta_{Af}\beta_{F}$$

$$\gamma_{FE} = -K_{FF}K_{EF} - K_{FAf}K_{EAf} - \beta_{F}\beta_{E}$$

$$\gamma_{EAm} = -K_{EE}K_{AmE} - K_{FF}K_{AmF} - K_{EAf}K_{AmAf} - \beta_{E}\beta_{Am}$$

$$\gamma_{AfE} = -K_{AfAf}K_{EAf} - \beta_{Af}\beta_{E}$$

$$\gamma_{FAm} = -K_{FF}K_{AmF} - K_{FAf}K_{AmAf} - \beta_{F}\beta_{Am}$$

$$\gamma_{AfAm} = -K_{AfAf}K_{AmAf} - \beta_{Af}\beta_{Am}$$

This system of parametric restrictions must be imposed at the reference point[15] to ensure that H is negative semidefinite. Clearly, the flexibility of the translog

[15]The employed reference point is the mean value for each i-th price considered in the analysis

specification is not destroyed, because the $n(n-1)/2$ elements of $\boldsymbol{K}$ just replace the $n(n-1)/2$ elements of $\boldsymbol{\Gamma}$ in the estimation. Notice that, by replacement, the coefficients γ_{ij} become nonlinear functions of K_{ij} ruling out any linear estimation technique as a possible empirical model to implement despite the fact that the original function is linear in parameters.
In order to estimates the new parameters we will employ a nonlinear SUR model (see appendix K). See the estimation results in appendix table 23.

7.3 Elasticity of Substitution

Allen Uzawa elasticity - σ_{ij}^a
The Allen Uzawa elasticity of substitution classifies a pair of inputs as direct substitutes (complements) if an increase in the price of $j-th$ commodity causes an increase (decrease) in quantity demanded of the other. It can be demonstrated that Allen Uzawa elasticity can be written as follows

$$\sigma_{ij}^a = \frac{\gamma_{ij} + S_{E_i} S_{E_j}}{S_{E_i} S_{E_j}} \quad \sigma_{ii}^a = \frac{\gamma_{ii} + S_{E_i}^2 - S_{E_i}}{S_{E_i}^2}$$

Morishima elasticity - σ_{ij}^m
Morishima elasticity of substitution rates a pair of inputs as direct substitutes (complements) if an increase in the price of j-th commodity causes the quantity of the other to increase (decrease) relatively to the quantity of the input whose price has changed.
In other words, σ_{ij}^m measures the impact of a change of producer's j price over the market share ratio between i and j when all other prices are kept constant, but all quantities adjust to their optimal levels.

$$\sigma_{ij}^m = S_{E_j}(\sigma_{ij}^a - \sigma_{ii}^a)$$

If $\sigma_{ij}^m > 0$ we say that inputs i and j are Morishima substitutes. If $\sigma_{ij}^m < 0$ we say that inputs i and j are Morishima complements.

Table 2: Allen Uzawa Elasticity

Allen Uzawa Elasticity	
σ^a_{AfAf}	-6,39783
σ^a_{AfF}	-1,45753
σ^a_{AfE}	0
σ^a_{AfAm}	0
σ^a_{AfMe}	9,997627
σ^a_{FAf}	-1,45753
σ^a_{FF}	-0,33205
σ^a_{FE}	0
σ^a_{FAm}	0
σ^a_{FMe}	2,277623
σ^a_{EAf}	0
σ^a_{EF}	0
σ^a_{EE}	0
σ^a_{EAm}	0
σ^a_{EMe}	-3,2E-16
σ^a_{AmAf}	0
σ^a_{AmF}	0
σ^a_{AmE}	0
σ^a_{AmAm}	0
σ^a_{AmMe}	6,86E-16
σ^a_{MeAf}	9,997627
σ^a_{MeF}	2,277623
σ^a_{MeE}	-3,2E-16
σ^a_{MeAm}	6,86E-16
σ^a_{MeMe}	-15,6229

Table 3: Morishima Elasticity

Morishima Elasticity	
σ^m_{AfF}	1,700889
σ^m_{AfE}	1,524486
σ^m_{AfAm}	0,225034
σ^m_{AfMe}	2,947425
σ^m_{FAf}	-0,22789
σ^m_{FE}	0,079121
σ^m_{FAm}	0,011679
σ^m_{FMe}	0,469143
σ^m_{EAf}	0
σ^m_{EF}	0
σ^m_{EAm}	0
σ^m_{EMe}	-5,8E-17
σ^m_{AmAf}	0
σ^m_{AmF}	0
σ^m_{AmE}	0
σ^m_{AmMe}	1,23E-16
σ^m_{MeAf}	5,187788
σ^m_{MeF}	6,162936
σ^m_{MeE}	3,722643
σ^m_{MeAm}	0,54951

7.4 Discussion

On the base of our analysis not all the alternatives to Former Soviet Union (FSU) oil are Morishima substitutes. There are only two reliable alternatives to the crude oil provided by FSU countries: Africa ($\sigma^m_{AfF} = 1,700889$), Middle East ($\sigma^m_{MeF} = 6,162936$) since these are the only sources, according to our estimation, to ensure energy replacement (a positive Morishima elasticity) as a result of an upward variation of the FSU oil prices.
We might notice that there is a huge distance between the two estimated elasticities, as a matter of fact, the substitutability of the FSU crude oil with Middle Eastern crude is more than three times greater than with African crude.
This situation could have important implications for the future European energy policy. For instance, the resolution of the EU embargo on oil exports from Iran will likely turn into a substitution effect of the imported crude oil from Russia to Iran in the long run. Russia has been one of the beneficiaries of the embargo having more than doubled exports into Iran's primary markets like Europe. This substitution effect is coherent with our estimate ($\sigma^m_{FMe} = 0,469143$). Probably because the benchmark export grade of FSU countries is similar to Iran's flagship blend. However, in the light of our estimated asymmetric elasticities, a turnaround flow from FSU countries to Middle Eastern countries could occur even stronger ($\sigma^m_{MeF} = 6,162936$). Our evidence also indicates that the relative demand for crude oil from FSU countries is inelastic to both Europe and America. Moreover, the new conflicts in Iraq would probably steer demand to African nations ($\sigma^m_{AfMe} = 2,947425$).

7.5 Conclusions

In order to study the imported crude oil demand and the degree of substitutability between FSU crude oil and several alternatives, I have proposed to use a euro area KLEM production function and a dataset of monthly data of crude oil imports (Volume (1000 bbl) and CIF prices ($/bbl)) provided by the European Commission. I have decided to sort the data into five groups by macro areas of origin (Africa, Former Soviet Union Countries (FSU), Europe, America and Middle East) and to employ CIF prices in order to get rid off, from the analysis, all the heterogeneity arising from a different distance, quality and taxation characterising the imported supply from the foreign countries.
Assuming that the European technology is: 1) aggregate homothetic weakly separable in quantities of imported crude oil; 2) linear homogeneous in inputs and assuming that 3) the aggregate industry cost function satisfies the regularity conditions, it has been demonstrated, on the base of the Composite Commodity Theorem, how to study the substitutability of European crude oil imports by Morishima elasticities.
With this type of elasticity, it has been possible to isolate the impact that the change of the oil price, in the Former Soviet Union (FSU) market, can determine on the EU relative demand of an alternative supply.
I have adopted two econometric estimation procedures. The first has required: 1) the choice of a flexible functional form of a unit cost function. The choice has fallen on a translog cost function; 2) a system of share equations[16]; 3) the

[16] A share equation explains the value which is imported from the i-th country on the total value imported

estimation of the parameters with a constrained linear SUR estimator.
Following the first procedure, it has been possible to obtain statistically significant estimates but not coherent with the assumption of the concavity of the cost function.
The second estimation procedure consisted in reparameterizing the original system of share equations so as to ensure the local concavity of the cost function at a reference point, that coincides to the sample mean price (see Diewert and Wales (1987)). This reconfiguration has made the problem not linear anymore and a nonlinear SUR model has been employed for the new parameter estimation.
The result has indicated that: the crude oil provided by former Soviet Union (FSU) countries is strongly substitutable with those imported from African and Middle Eastern countries while it is not substitutable with those imported from European and American countries.

References

[1] Roy George Douglas Allen, London School of Economics, and Political Science. *Mathematical analysis for economists*. Tech. rep. Macmillan London, 1938.

[2] Bobby E Apostolakis. *Energy—capital substitutability/complementarity: The dichotomy*. In: ***Energy Economics*** 12.1 (1990), pp. 48–58.

[3] Kenneth J Arrow et al. *Capital-labor substitution and economic efficiency*. In: ***The Review of Economics and Statistics*** (1961), pp. 225–250.

[4] Robert Bacon and Banco Mundial. *Measuring the possibilities of interfuel substitution*. 1031. Country Economics Department, the World Bank, 1992.

[5] Bakhtiyar Badalov. *The European Union and Russian Federation Energy Relations: Petrification or Revival?* In: (2012).

[6] Zeyno Baran. *EU energy security: time to end Russian leverage*. In: ***Washington*** 30.4 (2007), pp. 131–144.

[7] Anton P Barten. *Maximum likelihood estimation of a complete system of demand equations*. In: ***European economic review*** 1.1 (1969), pp. 7–73.

[8] Alberto Behar and Margaret Stevens. *The Allen/Uzawa elasticity of substitution under non-constant returns to scale*. In: (2009).

[9] Paul Belkin and Vince L Morelli. *The European Union's energy security challenges*. In: DTIC Document. 2007.

[10] Jan Bentzen and Tom Engsted. *Short-and long-run elasticities in energy demand: a cointegration approach*. In: ***Energy Economics*** 15.1 (1993), pp. 9–16.

[11] Ernst R Berndt and David O Wood. *Engineering and econometric interpretations of energy-capital complementarity*. In: ***The American Economic Review*** (1979), pp. 342–354.

[12] Mert Bilgin. *Geopolitics of European natural gas demand: Supplies from Russia, Caspian and the Middle East*. In: ***Energy Policy*** 37.11 (2009), pp. 4482–4492.

[13] Charles Blackorby and R Robert Russell. *Will the real elasticity of substitution please stand up?(A comparison of the Allen/Uzawa and Morishima elasticities)*. In: ***The American Economic Review*** (1989), pp. 882–888.

[14] BP. *BP Energy Outlook 2035 Country and regional insights - EU*. In: ().

[15] BP. *BP Statistical Review of World Energy June 2016*. In: ().

[16] Laurits R Christensen, Dale W Jorgenson, and Lawrence J Lau. *Transcendental logarithmic utility functions*. In: ***The American Economic Review*** (1975), pp. 367–383.

[17] European Commission. *Communication de la Commission au Parlement européen et au Conseil - Stratégie européenne pour la sécurité énergétique*. In: Document 52014DC0330 - ST 10409 2014 INIT (2-06-2014).

[18] COMMISSION OF THE EUROPEAN COMMUNITIES. *COMMISSION DECISION of 26 July 1999 implementing Council Decision 1999/280/EC regarding a Community procedure for information and consultation on crude oil supply costs and the consumer prices of petroleum products*. In: ***Official Journal of the European Communities*** (14-8-1999).

[19] Timothy J Considine. *Separability, functional form and regulatory policy in models of interfuel substitution*. In: ***Energy Economics*** 11.2 (1989), pp. 82–94.

[20] Michael Denny and Melvyn Fuss. *The use of approximation analysis to test for separability and the existence of consistent aggregates*. In: ***The American Economic Review*** (1977), pp. 404–418.

[21] W Erwin Diewert. *An application of the Shephard duality theorem: A generalized Leontief production function*. In: ***The Journal of Political Economy*** (1971), pp. 481–507.

[22] US EIA. *Worldwide look at reserves and production*. In: ***Oil Gas J*** (2015).

[23] US EIA. *Country Analysis Brief: Russia*. In: ***US Energy and Information Administration*** (2016).

[24] US EIA. *International energy statistics*. In: ***US Energy and Information Administration*** (2016).

[25] Jean-Pierre Favennec and Robin Baker. *Petroleum refining: refinery operation and management*. Editions Technip, 2001.

[26] Guohua Feng and Apostolos Serletis. *Productivity trends in US manufacturing: Evidence from the NQ and AIM cost functions*. In: ***Journal of Econometrics*** 142.1 (2008), pp. 281–311.

[27] Melvyn A Fuss. *The demand for energy in Canadian manufacturing: An example of the estimation of production structures with many inputs*. In: ***Journal of Econometrics*** 5.1 (1977), pp. 89–116.

[28] Vivian B Hall. *Major OECD country industrial sector interfuel substitution estimates, 1960–1979*. In: ***Energy Economics*** 8.2 (1986), pp. 74–89.

[29] Fredrik Hedenus, Christian Azar, and Daniel J.A. Johansson. *Energy security policies in EU-25—The expected cost of oil supply disruptions*. In: ***Energy Policy*** 38.3 (2010). Security, Prosperity and Community – Towards a Common European Energy Policy? Special Section with Regular Papers, pp. 1241 –1250. ISSN: 0301-4215. DOI: `http://dx.doi.org/10.1016/j.enpol.2009.01.030`. URL: `http://www.sciencedirect.com/science/article/pii/S0301421509000597`.

[30] Andreas Heinrich. *Export Pipelines from the CIS Region: Geopolitics, Securitization, and Political Decision-Making*. Vol. 10. Columbia University Press, 2014.

[31] Henry Helén. *The EU's energy security dilemma with Russia*. In: ***Polis J.*** 4 (2010), pp. 1–40.

[32] John R Hicks. *Marginal productivity and the principle of variation*. In: ***Economica*** 35 (1932), pp. 79–88.

[33] John R Hicks and Roy GD Allen. *A reconsideration of the theory of value. Part I.* In: ***Economica*** 1.1 (1934), pp. 52–76.

[34] Ayoe Hoff. *The linear Approximation of the CES Function with n Input Variables.* In: ***Marine Resource Economics*** (2004), pp. 295–306.

[35] Cheng Hsiao et al. *Modeling Ontario regional electricity system demand using a mixed fixed and random coefficients approach.* In: ***Regional Science and Urban Economics*** 19.4 (1989), pp. 565–587.

[36] Edward A Hudson and Dale W Jorgenson. *US energy policy and economic growth, 1975-2000.* In: ***The Bell Journal of Economics and Management Science*** (1974), pp. 461–514.

[37] Lester Hunt and Neil Manning. *ENERGY PRICE-and INCOME-ELASTICITIES OF DEMAND: SOME ESTIMATES FOR THE UK USING THE COINTEGRATION PROCEDURE.* In: ***Scottish Journal of Political Economy*** 36.2 (1989), pp. 183–193.

[38] European Electricity Grids Initiative et al. *Energy 2020: a strategy for competitive, sustainable and secure energy.* In: ***COM (2010)*** 639 (2010).

[39] Clifton T Jones. *A dynamic analysis of interfuel substitution in US industrial energy demand.* In: ***Journal of Business & Economic Statistics*** (1995), pp. 459–465.

[40] Claudia Kemfert and Heinz Welsch. *Energy-capital-labor substitution and the economic effects of CO 2 abatement: evidence for Germany.* In: ***Journal of Policy Modeling*** 22.6 (2000), pp. 641–660.

[41] Gernot Klepper and Sonja Peterson. *Marginal abatement cost curves in general equilibrium: The influence of world energy prices.* In: ***Resource and Energy Economics*** 28.1 (2006), pp. 1–23.

[42] Jan Kmenta. *On estimation of the CES production function.* In: ***International Economic Review*** 8.2 (1967), pp. 180–189.

[43] Masaaki Kuboniwa, Shinichiro Tabata, and Nataliya Ustinova. *How large is the oil and gas sector of Russia? A research report.* In: ***Eurasian Geography and Economics*** 46.1 (2005), pp. 68–76.

[44] Philip Lowe. *World Energy Council, 2015.* In: ().

[45] Ruslan Lukach et al. *EU Petroleum Refining Fitness Check: Impact of EU Legislation on Sectoral Economic Performance.* Tech. rep. Joint Research Centre (Seville site), 2015.

[46] Jan R Magnus. *Substitution between energy and non-energy inputs in the Netherlands 1950-1976.* In: ***International Economic Review*** (1979), pp. 465–484.

[47] Boris N Mamlyuk. *Ukraine Crisis, Cold War II, and International Law, The.* In: ***German LJ*** 16 (2015), p. 479.

[48] Elena Mazneva. *Russia Said to Move Closer to New Oil Tax to Spur Production.* In: ***Bloomberg*** (2015).

[49] Daniel McFadden. *Constant elasticity of substitution production functions.* In: ***The Review of Economic Studies*** (1963), pp. 73–83.

[50] Vince L Morelli. *The European Union's Energy Security Challenges.* In: DTIC Document. 2006.

[51] Michio Morishima. *A few suggestions on the theory of elasticity.* In: ***Keizai Hyoron (Economic Review)*** 16 (1967), pp. 144–150.

[52] Christophe-Alexandre Paillard. *Russia and Europe's mutual energy dependence.* In: ***Journal of international affairs*** 63.2 (2010), p. 65.

[53] Mikko Palonkorpi. *Energy security and the regional security complex theory.* In: ***Helsinki: Aleksanteri Institute/University of Helsinki*** (2007).

[54] Robert S Pindyck. *Interfuel substitution and the industrial demand for energy: an international comparison.* In: ***The Review of Economics and Statistics*** (1979), pp. 169–179.

[55] Platts. *Special Report: Russian crude oil exports to the Pacific Basin – an ESPO update February 2011.* In: ***The McGraw-Hill Companies*** (2011).

[56] Graham Pyatt. English. In: ***The Economic Journal*** 82.327 (1972), pp. 1059–1061. ISSN: 00130133. URL: `http://www.jstor.org/stable/2230285`.

[57] David L Ryan and Terence J Wales. *Imposing local concavity in the translog and generalized Leontief cost functions.* In: ***Economics Letters*** 67.3 (2000), pp. 253–260.

[58] Apostolos Serletis and Asghar Shahmoradi. *Semi-nonparametric estimates of interfuel substitution in US energy demand.* In: ***Energy Economics*** 30.5 (2008), pp. 2123–2133.

[59] Apostolos Serletis, Govinda R Timilsina, and Olexandr Vasetsky. *Interfuel substitution in the United States.* In: ***Energy Economics*** 32.3 (2010), pp. 737–745.

[60] Boriss Siliverstovs et al. *International market integration for natural gas? A cointegration analysis of prices in Europe, North America and Japan.* In: ***Energy Economics*** 27.4 (2005), pp. 603–615.

[61] Giovanni Urga and Chris Walters. *Dynamic translog and linear logit models: a factor demand analysis of interfuel substitution in US industrial energy demand.* In: ***Energy Economics*** 25.1 (2003), pp. 1–21.

[62] Noel D Uri. *Energy substitution in the UK, 1948–64.* In: ***Energy Economics*** 1.4 (1979), pp. 241–244.

[63] Hirofumi Uzawa. *Production functions with constant elasticities of substitution.* In: ***The Review of Economic Studies*** (1962), pp. 291–299.

[64] S. S. Wilks. *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.* In: ***The Annals of Mathematical Statistics*** 9.1 (1938), pp. 60–62. ISSN: 00034851. URL: `http://www.jstor.org/stable/2957648`.

[65] Steven Woehrel. *Russian energy policy toward neighboring countries.* In: DTIC Document. 2010.

APPENDIX 1

A
Shephard's Lemma

Let's denominate the production inputs in terms of a single $q = m + 3$ dimensional vector x of all the input of the production function.

$$x = (E_1, E_2, ..., E_o, ..., x_i, ..., E_m, L, M, K)' \geq 0$$

Let it be consistent with the EU aggregate technology production function

$$f(x) \geq y$$

We denote by vector of factor prices p

$$p = (P_{E_1}, P_{E_2}, ..., P_{E_o}, ..., p_i, ..., P_{E_m}, P_L, P_M, P_K)$$

the vector of input prices. Suppose all the price are given (price taker firms). Shephard duality theorem states that under certain condition (Regularity conditions) the cost function $C(p, y)$ can completely describe the technology (Duality theorem).

$$C(p, y) = min_{x_i \geq 0}\{p'x | f(x) \geq y\}$$

provided that cost function satisfies regularity conditions, we can obtain the cost minimizing input demand for the i-th factor by partially differentiating the cost function with respect to the i-th factor price

$$x_i(p, y) = \frac{\partial C}{\partial p_i}, i = 1, ..., q$$

B
Regularity Conditions

- **Positivity** $C(y, p)$ is a positive real-valued function

$$C(y, p) \geq 0, \forall p, y > 0 \text{ and } C(y, p) = 0$$

- **Monotonicity**

  Monotonicity requires positive marginal products of all inputs.

  $C(y, p)$ is a non-decreasing in p and y.

$$\frac{\partial C}{\partial p_i} \geq 0, i = 1, ..., q$$

  because $x_i \geq 0, i = 1, ..., q$ and

$$\frac{\partial C}{\partial y} \geq 0, i = 1, ..., q$$

- **Homogeneity** $C(y,p)$ is linear homogeneous in p. By Euler theorem we obtain the following expressions

$$\Sigma_{i=1}^{q} p_i \frac{\partial C}{\partial p_i} = C$$

$$\Sigma_{i=1}^{q} p_i \frac{\partial^2 C}{\partial p_i \partial p_j} = 0, \forall j$$

$$\Sigma_{i=1}^{q} p_i \frac{\partial^2 C}{\partial p_i \partial y} = \frac{\partial C}{\partial y}, \forall j$$

- **Simmetry** $C(y,p)$ is twice differentiable, Young theorem implies.

$$\frac{\partial^2 C}{\partial p_i \partial p_j} = \frac{\partial^2 C}{\partial p_j \partial p_i}, i,j = 1,...,q$$

or

$$\frac{\partial x_i}{\partial p_j} = \frac{\partial x_j}{\partial p_i}, i,j = 1,...,q$$

strictly speaking, the Hessian matrix must be symmetric

- **Concavity** $C(y,p)$ is concave in p if

$$\left[\frac{\partial^2 C}{\partial p_i \partial p_j}\right]_{\forall i,j} \text{ is a negative semidefinite matrix.}$$

C
Composite Commodity Theorem

It applies every time any single commodity, belonging to a group of commodities, has a price moving simultaneously and proportionally together with the others. This assumption seems to be consistent with commodities like different crude oil qualities or oil blends supplied by various nations and regions. We define the composite European imported crude oil commodity E_o to be the composite European crude oil imports expenditure differentiated by country of origin, $E_{o1}, E_{o2}, ..., E_{on}$. We assume their prices evolve in time proportionally together such that

$$P_{E_{oi}} = \Theta P^0_{E_{oi}}, \forall i = 1,...,n$$

where $P^0_{E_{oi}}$ is the initial price or just the price applied a time before $P_{E_{oi}}$.
Notice that next definitions of vector x and p differ from the previous provided in appendix A because this time we don't lump the crude oil supplies $E_{o1}, E_{o2}, ..., E_{on}$ into a composit commodity E_o. We denote by vector of inputs

$$x* = (E_1, E_2, ..., E_{o1}, E_{o2}, ..., E_{on}, ..., E_m, L, M, K)' \geq 0$$

We denote by vector of prices

$$p* = (P_{E_1}, P_{E_2}, ..., P_{E_{o1}}, P_{E_{o2}}, ..., P_{E_{on}}, ..., P_{E_m}, P_L, P_M, P_K)$$

The cost minimization problem given p and x can be formalized as:

$$Min_{x*}(p*)'(x*)$$

$$s.t. f(x*) = y$$

the correspondent Lagrangian is

$$\mathcal{L}(x*, \lambda; p*, y) = p'x * -\lambda(f(x*) - y)$$

applying envelop theorem with respect to Θ we obtain[17]

$$\frac{\partial C(p*, y)}{\partial \Theta} = \sum_{i=1}^{n} E_{oi} P^0_{E_{oi}}$$

In the light of the Shephard's lemma (see appendix A) we might interpret Θ as the price of the composite commodity

$$\Theta = P_{E_o}$$

while $\sum_{i=1}^{n} E_{oi} P^0_{E_{oi}}$ might be interpreted as the quantity Eo of the composite commodity.

$$Eo = \sum_{i=1}^{n} E_{oi} P^0_{E_{oi}}$$

So that we have

$$\frac{\partial C(p*, y)}{\partial P_{E_o}} = Eo$$

therefore we can equivalently consider, for the same cost minimization problem, next two vectors:
vector of inputs

$$x = (E_1, E_2, ..., E_o, ..., E_m, L, M, K)'$$

vector of prices

$$p = (P_{E_1}, P_{E_2}, ..., P_{E_o}, ..., P_{E_m}, P_L, P_M, P_K)$$

which are exactly the same vectors provided at appendix A.
The equivalent minimization problem is

$$Min_x p'x$$

$$s.t. f(x) = y$$

[17] any E_{oi} is a function of $p*$ and y

Given the price vector $\pi = (P_{E_{o1}}, P_{E_{o2}}, ..., P_{E_{on}})$ and the input vector $z = (E_{o1}, E_{o2}, ..., E_{on})$ The optimization problem may be solved in two stages (Denny and Fuss, 1975 [20]). In the second stage the economic agents optimize with respect to the fuel mix of E_o, therefore the resulting cost function is

$$C_{E_o} = C(\pi, y)$$

while in the first stage the optimisation is concerned with the capital K, labour L, materials M and Energy E_o. In this case, the resulting cost function is

$$C = C(p, y)$$

D MRS between E_{oi} and E_{oj}

From Appendix C

$$Eo = \sum_{i=1}^{n} E_{oi} P^0_{E_{oi}}$$

under the standard assumption of neoclassical economics goods and services are continuously divisible. Therefore, applying Dini's Theorem we can get the MRS between any two imported quantities of crude oil from different countries.

$$MRS = \frac{\partial E_{oi}}{\partial E_{oj}} = -\frac{\frac{\partial E_o}{\partial E_{oj}}}{\frac{\partial E_o}{\partial E_{oi}}} = -\frac{P^0_{E_{oj}}}{P^0_{E_{oi}}}$$

E Cost Share Equation

By definition

$$\frac{\partial LN(C)}{\partial LN(P_{E_{oi}})} = \frac{P_{E_{oi}}}{C} \frac{\partial C}{\partial P_{E_{oi}}}, i = 1, \ldots, n$$

by Shephard's Lemma A we get

$$\frac{\partial LN(C)}{\partial LN(P_{E_{oi}})} = \frac{P_{E_{oi}}}{C} E_{oi} = S_{E_{Oi}}, i = 1, \ldots, n$$

F Linear Homogeneous production function

We denote a vector of inputs x and a vector of prices p, a continuous and increasing in x aggregate technology production function $f(x, t)$ and a cost function $C(y, p)$.
If f is linearly homogeneous in x then $C(y, p) = y \cdot C(p)$.
Proof.

$$C(p, y) = min_x\{p'x : f(x) \geq y\}$$

$$= min_x\{p'x : f(x) = y\}$$

$$= min_x\{p'x : \frac{1}{y} \cdot f(x) = 1\}$$

$$= min_x\{p'x : f(\frac{x}{y}) = 1\}$$

$$= min_x\{y \cdot p'\frac{x}{y} : f(\frac{x}{y}) = 1\}$$

$$= y \cdot min_x\{p'\frac{x}{y} : f(\frac{x}{y}) = 1\}$$

$$= y \cdot min_q\{p'q : f(q) = 1\}$$

$$= y \cdot C(p)$$

similarly, it is possible to demonstrate that

$$C(\pi, y) = y \cdot C(\pi)$$

where π is the price vector of the imported crude oil $\pi = (P_{E_{o1}}, P_{E_{o2}}, ..., P_{E_{on}})$ and where the input vector is the quantity vector of the imported crude oil $z = (E_{o1}, E_{o2}, ..., E_{on})$

G Flexible Functional Form of a Cost Function

A flexible functional form f is an arbitrary function capable to approximate to the second order a given continuous differentiable function f^* at a given point $x^* \in \mathbb{R}^n$. Consequently any candidate function f must satisfy, as a matter of fact, a series of constraints:

a. $f(x^*) = f^*(x^*)$

b. $\bigtriangledown f(x^*) = \bigtriangledown f^*(x^*)$

c. $\bigtriangledown^2 f(x^*) = \bigtriangledown^2 f^*(x^*)$

In the paper, it has been employed a translog unit cost function which is a classical flexible functional form which satisfies the following conditions

a. $C(p^*) = C^*(p^*)$

b. $\bigtriangledown C(p^*) = \bigtriangledown C^*(p^*)$

c. $\bigtriangledown^2 C(p^*) = \bigtriangledown^2 C^*(p^*)$

Point a) determines that the flexible functional form of C inherits the regularity conditions at the point p^* consistently with the cost minimization problem. See appendix B.

Consider that:
Positivity has usually been checked before the estimation took place.
Monotonicity violations are frequent and empirically meaningful since this requirement is not automatically satisfied for most functional forms.

In the case of the translog function, monotonicity has been ensured when the following inequality holds

$$\frac{\partial C}{\partial p_i} = \frac{C}{p_i}\frac{\partial LN(C)}{\partial LN(p_i)} = \frac{C}{p_i}(\beta_i + \sum_{j=1}^{n}\gamma_{ij}LN(p_i)) \geq 0, i = 1, ..., q$$

therefore $C(p)$ is a non-decreasing in p.
Since both C and p_i are positive numbers, monotonicity depends on the sign of the term in parenthesis. Notice that if prices are all equal to one, what does really matter is the sign of $\{\beta_i\}$ for $i = 1, ..., q$.
Linear homogeneity (homogeneity of degree 1) in prices of cost function holds when: given an arbitrary constant value t the following expression hold

$$LN(t \cdot C_{E_o}(p)) = LN(C_{E_o}(t \cdot p))$$

We need to verify the conditions ensuring that the previous expression holds.

$$LN(C_{E_o}(t \cdot p)) = LN(\alpha_0) + \sum_{i=1}^{n}\beta_i LN(t \cdot P_i) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}LN(t \cdot p_i)LN(t \cdot p_j)$$

$$= LN(\alpha_0) + \textstyle\sum_{i=1}^{n}\beta_i LN(p_i) + LN(t)\sum_{i=1}^{n}\beta_i +$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}[LN(p_i)LN(p_j) + LN(p_i)LN(t) + LN(p_i)LN(t) + LN(t)^2]$$

Consequently, the following conditions

$$\sum_{i=1}^{n}\beta_i = 1$$

$$\sum_{i=1}^{n}\gamma_{ij} = \sum_{i=1}^{n}\gamma_{ji} = 0, \forall j = 1, ..., n$$

are sufficient for ensuring the homogeneity of degree 1 .

H Linear Constrained SUR Model

A constrained SUR model has been employed for estimating the gammas and betas of the system of equations 11 12 13 14 given a set of linear constraints.

$$S = Xb + \varepsilon \text{ such that } Rb = r$$

where $S = \begin{pmatrix} S_{Af} \\ S_F \\ \vdots \\ S_{Am} \end{pmatrix}, X = \begin{pmatrix} X_{Af} & 0 & \dots & 0 \\ 0 & X_F & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_{Am} \end{pmatrix}, b = \begin{pmatrix} b_{Af} \\ b_F \\ \vdots \\ b_{Am} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{Af} \\ \varepsilon_F \\ \vdots \\ \varepsilon_{Am} \end{pmatrix}$

With N=5 and K=168 we get that estimation of b

$$\hat{b}_{RFGLS} = \hat{b}_{FGLS} - \Big(X^{\mathsf{T}}(\hat{\Sigma}^{-1}\otimes I_N)X\Big)^{-1} R^{\mathsf{T}}\Big[R\Big(X^{\mathsf{T}}(\hat{\Sigma}^{-1}\otimes I_N)X\Big)^{-1}R^{\mathsf{T}}\Big]^{-1}\Big(R\hat{b}_{FGLS} - r\Big)$$

where

$$\hat{b}_{FGLS} = \Big(X^{\mathsf{T}}(\hat{\Sigma}^{-1}\otimes I_N)X\Big)^{-1} X^{\mathsf{T}}(\hat{\Sigma}^{-1}\otimes I_N)\, y$$

$$X_{Af} = X_F = X_E = X_{Am} = \begin{pmatrix} 1 & LN(\frac{P_{Af}}{P_{Me}})(t) & \dots & LN(\frac{P_{Am}}{P_{Me}})(t) \\ 1 & LN(\frac{P_{Af}}{P_{Me}})(t+1) & \dots & LN(\frac{P_{Am}}{P_{Me}})(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & LN(\frac{P_{Af}}{P_{Me}})(t+k) & \dots & LN(\frac{P_{Am}}{P_{Me}})(t+k) \end{pmatrix},$$

and

$$b_{Af} = \begin{pmatrix} \beta_{Af} \\ \gamma_{AfAf} \\ \vdots \\ \gamma_{AfAm} \end{pmatrix} \; b_F = \begin{pmatrix} \beta_F \\ \gamma_{FAf} \\ \vdots \\ \gamma_{FAm} \end{pmatrix} \; b_E = \begin{pmatrix} \beta_E \\ \gamma_{EAf} \\ \vdots \\ \gamma_{EAm} \end{pmatrix} \; b_{Am} = \begin{pmatrix} \beta_{Am} \\ \gamma_{AmAf} \\ \vdots \\ \gamma_{AmAm} \end{pmatrix}$$

I Likelihood ratio test of the nonlinear SUR model

Consider the likelihood $L = L(\theta \mid y)$ of a statistical model where y is the vector observations of n i.i.d observations drawn from a distribution with parameter θ belonging to a submanifold B_1 of $\mathbb{R}^d$ with dimension $\dim(B_1) = s$. Let $B_0 \subset B_1$ be a submanifold with dimension $\dim(B_0) = m$. We are interested in testing $H_0 \colon \{\theta \in B_0\}$ given the following definition of deviance d

$$d = 2(\ln(\text{ L}_{\text{alt model}}) - \ln(\text{ L}_{\text{null model}})) \sim \chi^2_{s-m}$$

Wilks' theorem [64] says that, under usual regularity assumptions, d is asymptotically χ^2 distributed with $s - m$ degrees of freedom when H_0 holds true. Notice that the log-likelihood function of the nonlinear SUR model is :

$$\ln(L) = -\frac{1}{2}\ln(|\mathbf{\Sigma}|\,) - \frac{1}{2}(\mathbf{y} - f(x;b))^{\mathrm{T}}(\Sigma^{-1}\otimes I)(\mathbf{y} - f(x;b)) - \frac{NT}{2}\ln(2\pi)$$

With unknown $\mathbf{\Sigma}$ matrix one can proceed with FGLS. Iterated FGLS provides maximum likelihood estimates.

$$\ln(L) = -\frac{1}{2}\ln(|\widehat{\mathbf{\Sigma}}|) - \frac{1}{2}(\mathbf{y} - f(x;b))^{\mathrm{T}}(\widehat{\Sigma}^{-1}\otimes I)(\mathbf{y} - f(x;b)) - \frac{NT}{2}\ln(2\pi)$$

where $\mathbf{y}$ is a vector containing all the dependent variables, $f(x;\beta)$ is a vector of functions of x, x is a vector of independent variables, b in a vector of coefficients, N is the number of equations, T is the number of observations, $\widehat{\mathbf{\Sigma}}$ is the expected variance-covariance matrix.

J Linear SUR model

Table 4: MODEL 1 Unrestricted

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,225997894	0,004167818	54,22451721	0
γ_{AfAf}	0,270871767	0,109667941	2,469926613	0,014544186
γ_{AfF}	0,204112209	0,146731842	1,391056005	0,16610433
γ_{AfE}	-0,621596382	0,153848064	-4,040326329	$8,21E-05$
γ_{AfAm}	0,028357764	0,052301819	0,5421946	0,588424924
β_F	0,396089768	0,009234807	42,89096458	0
γ_{FAf}	-0,345656323	0,24299582	-1,422478475	0,156798266
γ_{FF}	2,039369496	0,325119846	6,272669984	$3,07E-09$
γ_{FE}	-1,337385307	0,340887555	-3,923244742	0,000128534
γ_{FAm}	-0,121279981	0,115887316	-1,04653369	0,296864104
β_E	0,197591472	0,006269804	31,51477734	0
γ_{EAf}	-0,203487659	0,164977587	-1,233426085	0,219192523
γ_{EF}	-1,515719438	0,220734199	-6,866717742	$1,31E-10$
γ_{EE}	1,460349304	0,231439398	6,30985614	$2,53E-09$
γ_{EAm}	0,034682835	0,078679583	0,440811122	0,659933825
β_{Am}	0,040245835	0,001289202	31,2176302	0
γ_{AmAf}	0,078816687	0,033922824	2,323411737	0,021392173
γ_{AmF}	0,150203581	0,045387543	3,309356932	0,001151042
γ_{AmE}	-0,245451136	0,047588755	-5,15775494	$7,17E-07$
γ_{AmAm}	0,015939204	0,016178159	0,985229767	0,325971397

Constraints for imposing symmetry to the translog cost function

Table 5: MODEL 1 Restricted

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,224631	0,003886	57,79861	0
γ_{AfAf}	0,163638	0,099714	1,64108	0,101259
γ_{AfF}	0,033201	0,109594	0,302942	0,76203
γ_{AfE}	-0,41507	0,100991	-4,10994	4,46E-05
γ_{AfAm}	0,051956	0,029066	1,787538	0,074311
β_{F}	0,39467	0,007714	51,16593	0
γ_{FAf}	0,033201	0,109594	0,302942	0,76203
γ_{FF}	1,943267	0,256372	7,579877	1,18E-13
γ_{FE}	-1,34125	0,168309	-7,96897	7,11E-15
γ_{FAm}	0,097047	0,045174	2,148295	0,032055
β_{E}	0,198592	0,00548	36,23975	0
γ_{EAf}	-0,41507	0,100991	-4,10994	4,46E-05
γ_{EF}	-1,34125	0,168309	-7,96897	7,11E-15
γ_{EE}	1,322701	0,153997	8,589145	0
γ_{EAm}	-0,15954	0,037807	-4,21978	2,79E-05
β_{Am}	0,039573	0,001425	27,77348	0
γ_{AmAf}	0,051956	0,029066	1,787538	0,074311
γ_{AmF}	0,097047	0,045174	2,148295	0,032055
γ_{AmE}	-0,15954	0,037807	-4,21978	2,79E-05
γ_{AmAm}	0,039567	0,016445	2,40606	0,0164

Table 6: Likelihood ratio test on Restrictions of MODEL 1

LogLik	Df	Chisq	Pr($> Chisq$)
1651,797	NA	NA	NA
1659,33	6	15,06615	0,019748

we accept the restricted model with $\alpha = 0,01$

Imposition of null coefficients γ_{AfF} and γ_{FAf}

Table 7: MODEL 2 Restricted

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,224536287	0,003793103	59,19593	0
γ_{AfAf}	0,161092884	0,097877726	1,645858	0,100269711
γ_{AfF}	0	0	Inf	0
γ_{AfE}	-0,397789399	0,077785331	-5,113939	4,14E-07
γ_{AfAm}	0,055199569	0,026904786	2,051664	0,040597046
β_F	0,395354696	0,007461551	52,98559	0
γ_{FAf}	0	0	-inf	0
γ_{FF}	1,940981861	0,256081359	7,579552	1,18E-13
γ_{FE}	-1,32067965	0,155647604	-8,485063	0
γ_{FAm}	0,09650945	0,044834991	2,152548	0,031716639
β_E	0,19818872	0,005412239	36,61862	0
γ_{EAf}	-0,397789399	0,077785331	-5,113939	4,14E-07
γ_{EF}	-1,32067965	0,155647604	-8,485063	0
γ_{EE}	1,303569768	0,138314898	9,424652	0
γ_{FAm}	-0,162219172	0,03788176	-4,28225	2,13E-05
β_{Am}	0,039529034	0,001412721	27,98077	0
γ_{AmAf}	0,055199569	0,026904786	2,051664	0,040597046
γ_{AmF}	0,09650945	0,044834991	2,152548	0,031716639
γ_{AmE}	-0,162219172	0,03788176	-4,28225	2,13E-05
γ_{AmAm}	0,039292878	0,016520078	2,378492	0,017667602

Table 8: likelihood ratio test to compare model 2 and model 1 restricted

LogLik	Df	Chisq	Pr($> Chisq$)
1651,747	NA	NA	NA
1651,797	1	0,100434	0,75131

we accept model 2 with $\alpha = 0,01$

Table 9: MODEL 3 Restricted

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,224536	0,003793	59,19593	0
γ_{AfAf}	0,161093	0,097878	1,645858	0,10027
γ_{AfE}	-0,39779	0,077785	-5,11394	4,14E-07
γ_{AfAm}	0,0552	0,026905	2,051664	0,040597
β_{F}	0,395355	0,007462	52,98559	0
γ_{FF}	1,940982	0,256081	7,579552	1,18E-13
γ_{FE}	-1,32068	0,155648	-8,48506	0
γ_{FAm}	0,096509	0,044835	2,152548	0,031717
β_{E}	0,198189	0,005412	36,61862	0
γ_{EAf}	-0,39779	0,077785	-5,11394	4,14E-07
γ_{EF}	-1,32068	0,155648	-8,48506	0
γ_{EE}	1,30357	0,138315	9,424652	0
γ_{EAm}	-0,16222	0,037882	-4,28225	2,13E-05
β_{Am}	0,039529	0,001413	27,98077	0
γ_{AmAf}	0,0552	0,026905	2,051664	0,040597
γ_{AmF}	0,096509	0,044835	2,152548	0,031717
γ_{AmE}	-0,16222	0,037882	-4,28225	2,13E-05
γ_{AmAm}	0,039293	0,01652	2,378492	0,017668

Table 10: likelihood ratio test to compare model 3 and model 1 restricted

LogLik	Df	Chisq	Pr($> Chisq$)
1651,747	NA	NA	NA
1651,797	1	0,100434	0,75131

we accept model 3 with $\alpha = 0,01$

Table 11: likelihood ratio test to compare model 3 and model 2

LogLik	Df	Chisq	Pr($> Chisq$)
1651,747	NA	NA	NA
1651,747	0	1,82E-12	0

we reject model 3 with $\alpha = 0,01$

Table 12: MODEL 4 Restricted

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,226531	0,003618	62,61352	0
γ_{AfE}	-0,29321	0,045483	-6,44663	2,21E-10
γ_{AfAm}	0,042151	0,02556	1,649087	0,099606
β_{F}	0,394884	0,007451	53,00003	0
γ_{FF}	1,931928	0,255865	7,550576	1,45E-13
γ_{FE}	-1,30827	0,155254	-8,42663	2,22E-16
γ_{FAm}	0,097881	0,044627	2,19332	0,028632
β_{E}	0,197415	0,005385	36,66241	0
γ_{EAf}	-0,29321	0,045483	-6,44663	2,21E-10
γ_{EF}	-1,30827	0,155254	-8,42663	2,22E-16
γ_{EE}	1,220422	0,129062	9,456099	0
γ_{EAm}	-0,15589	0,03749	-4,15812	3,63E-05
β_{Am}	0,039729	0,0014	28,36893	0
γ_{AmAf}	0,042151	0,02556	1,649087	0,099606
γ_{AmF}	0,097881	0,044627	2,19332	0,028632
γ_{AmE}	-0,15589	0,03749	-4,15812	3,63E-05
γ_{AmAm}	0,037456	0,016362	2,289281	0,022378

Table 13: likelihood ratio test to compare model 4 and model 1 restricted

LogLik	Df	Chisq	Pr($> Chisq$)
1650,407	NA	NA	NA
1651,797	2	2,779529	0,249134

we accept model 4 with $\alpha = 0,01$

Table 14: likelihood ratio test to compare model 4 and model 2

LogLik	Df	Chisq	Pr($> Chisq$)
1650,407	NA	NA	NA
1651,747	1	2,679095	0,101673

we accept model 4 with $\alpha = 0,01$

Table 15: MODEL 5 Restricted

	Estimate	Std. Error	t value	Pr($>\|t\|$)
β_{Af}	0,22402	0,003286	68,18389	0
γ_{AfE}	-0,31153	0,044223	-7,0446	4,68E-12
β_F	0,395801	0,007426	53,29994	0
γ_{FF}	1,964287	0,255229	7,696175	5,11E-14
γ_{FE}	-1,34431	0,153687	-8,7471	0
γ_{FAm}	0,08964	0,043506	2,060394	0,039752
β_E	0,198198	0,005357	36,9973	0
γ_{EAf}	-0,31153	0,044223	-7,0446	4,68E-12
γ_{EF}	-1,34431	0,153687	-8,7471	0
γ_{EE}	1,281488	0,12318	10,40339	0
γ_{EAm}	-0,12341	0,030571	-4,03684	6,05E-05
β_{Am}	0,039963	0,001366	29,26154	0
γ_{AmF}	0,08964	0,043506	2,060394	0,039752
γ_{AmE}	-0,12341	0,030571	-4,03684	6,05E-05
γ_{AmAm}	0,032717	0,015837	2,065898	0,039227

Table 16: likelihood ratio test to compare model 5 and model 1 restricted

LogLik	Df	Chisq	Pr($> Chisq$)
1648,558	NA	NA	NA
1651,797	3	6,478833	0,090501

we accept model 5 with $\alpha = 0,01$

Table 17: likelihood ratio test to compare model 5 and model 4

LogLik	Df	Chisq	Pr($> Chisq$)
1648,557647	NA	NA	NA
1650,407299	1	3,699304	0,054435

we accept model 5 with $\alpha = 0,01$

K Non-Linear SUR model

Table 18: MODEL 1

	Estimate	Std. Error	t value	Pr(> |t|)
β_{Af}	0,202362	0,002244	90,18675	0
β_F	0,344352	0,005628	61,1806	0
β_E	0,238215	0,00399	59,7054	0
β_{Am}	0,035116	0,000807	43,50582	0
K_{AfAf}	-0,52239	0,103131	-5,06527	5,32E-07
K_{FAf}	-0,20935	0,222152	-0,94236	0,346358
K_{EAf}	0,0315	0,202048	0,155904	0,876157
K_{AmAf}	0,000209	0,057198	0,003653	0,997087
K_{FF}	7,31E-08	2033196	3,59E-14	1
K_{EF}	1,39E-07	6104902	2,28E-14	1
K_{AmF}	-7,76E-08	2064520	-3,76E-14	1
K_{EE}	3,65E-08	24889491	1,47E-15	1
K_{AmE}	2,09E-07	1,42E+08	1,47E-15	1
K_{AmAm}	-3,22E-07	92937424	-3,46E-15	1

Table 19: MODEL 2

	Estimate	Std. Error	t value	Pr($>\|t\|$)
β_{Af}	0,202601	0,002191	92,45253	0
β_F	0,344123	0,005416	63,53614	0
β_E	0,238396	0,003852	61,88933	0
β_{Am}	0,035139	0,000789	44,54157	0
K_{AfAf}	0,528708	0,079838	6,622293	7,40E-11
K_{FAf}	0,164849	0,121118	1,361062	0,173964
K_{EAf}	0,007115	0,109612	0,064912	0,948264
K_{AmAf}	0,00152	0,03249	0,046769	0,962711

Table 20: likelihood ratio test to compare model 2 and model 1

LogLik	Df	Chisq	Pr($> Chisq$)
1569	NA	NA	NA
1576	6	15.00288	0.02023429

we accept model 2 with $\alpha = 0,01$

Table 21: MODEL 3

	Estimate	Std. Error	t value	Pr($>\|t\|$)
β_{Af}	0,202703	0,002187	92,69634	0
β_F	0,34428	0,005378	64,01153	0
β_E	0,238335	0,003785	62,97139	0
β_{Am}	0,035074	0,000794	44,18828	0
K_{AfAf}	0,53279	0,057404	9,281487	0
K_{FAf}	0,188784	0,090573	2,084318	0,037519
K_{AmAf}	-0,01012	0,031636	-0,31997	0,749095

Table 22: likelihood ratio test to compare model 3 and model 2

LogLik	Df	Chisq	Pr($>Chisq$)
1576	NA	NA	NA
1578	1	2.969443	0.08485108

we accept model 3 with $\alpha = 0,01$

Table 23: MODEL 4

	Estimate	Std. Error	t value	Pr($>\|t\|$)
β_{Af}	0,202486	0,002192	92,38942	0
β_F	0,344288	0,005376	64,04375	0
β_E	0,238282	0,003784	62,97101	0
β_{Am}	0,035173	0,000777	45,2618	0
K_{AfAf}	0,512166	0,054159	9,456666	0
K_{FAf}	0,198392	0,087297	2,272616	0,023373003

Table 24: likelihood ratio test to compare model 4 and model 3

LogLik	Df	Chisq	Pr($> Chisq$)
1578	NA	NA	NA
1579	1	.946628	0.1629507

we accept model 4 with $\alpha = 0,01$

R Code

```
dat <- read.delim2("dat.txt")


View(dat)

# Define 1st equation
dat$S.af <- I(dat$P.af * dat$Q.af) / I(dat$P.af * dat$Q.af+
dat$P.f * dat$Q.f+ dat$P.e * dat$Q.e+ dat$P.am* dat$Q.am+
dat$P.me* dat$Q.me)
# Define 2nd equation
dat$S.f <- I(dat$P.f * dat$Q.f) / I(dat$P.af * dat$Q.af+
dat$P.f * dat$Q.f+ dat$P.e * dat$Q.e+ dat$P.am* dat$Q.am+
dat$P.me* dat$Q.me)
# Define 3rd equation
dat$S.e <- I(dat$P.e * dat$Q.e) / I(dat$P.af * dat$Q.af+
dat$P.f * dat$Q.f+ dat$P.e * dat$Q.e+ dat$P.am* dat$Q.am+
dat$P.me* dat$Q.me)
# Define 4th equation
dat$S.am <- I(dat$P.am * dat$Q.am) / I(dat$P.af * dat$Q.af+
dat$P.f * dat$Q.f+ dat$P.e * dat$Q.e+ dat$P.am* dat$Q.am+
dat$P.me* dat$Q.me)
# Define 5th equation
dat$S.me <- I(dat$P.me * dat$Q.me) / I(dat$P.af * dat$Q.af+
dat$P.f * dat$Q.f+ dat$P.e * dat$Q.e+ dat$P.am* dat$Q.am+
dat$P.me* dat$Q.me)

# add logarithms of the price ratios
dat$lP.af <- I(log(dat$P.af /dat$P.me))
dat$lP.f <- I(log(dat$P.f /dat$P.me))
dat$lP.e <- I(log(dat$P.e /dat$P.me))
dat$lP.am<- I(log(dat$P.am /dat$P.me) )
#------------------------------------------------------------------------
# MODEL 1 UNRESTRICTED

# simultaneous equation - MODEL 1

S1 <- S.af ~ 1+lP.af + lP.f + lP.e+ lP.am
S2 <- S.f ~ 1+lP.af + lP.f + lP.e+ lP.am
S3 <- S.e ~ 1+lP.af + lP.f + lP.e+ lP.am
S4<- S.am ~ 1+lP.af + lP.f + lP.e+ lP.am

# unrestricted estimation

library("lmtest", lib.loc="~/R/win-library/3.2")
library("systemfit", lib.loc="~/R/win-library/3.2")
list(Africa = S1, FSU = S2, Europe = S2, America = S2)
fitsur1un <- systemfit(list(Africa = S1, FSU = S2, Europe = S3,
America = S4), "SUR" ,data=dat)
summary(fitsur1un)
```

```
#-------------------------------------------------------------------------
# MODEL 1 RESTRICTED


# constraints for imposing symmetry to the translog cost function

restrictM <- matrix(c(0,0,1,0,0,  0,-1,0,0,0,  0,0,0,0,0,  0,0,0,0,0,
                      0,0,0,-1,0,  0,0,0,0,0,  0,1,0,0,0,  0,0,0,0,0,
                      0,0,0,0,-1,  0,0,0,0,0,  0,0,0,0,0,  0,1,0,0,0,
                      0,0,0,0,0,  0,0,0,-1,0,   0,0,1,0,0,  0,0,0,0,0,
                      0,0,0,0,0,  0,0,0,0,-1,   0,0,0,0,0,  0,0,1,0,0,
                      0,0,0,0,0,  0,0,0,0,0,   0,0,0,0,-1,  0,0,0,1,0),
                    byrow=TRUE, nrow=6, ncol=20)


# Theil's F test for testing all constraints

test.constr.1<-linearHypothesis( fitsur1un, restrictM, test = c( "FT",
"F", "Chisq" ))
# if $ p-value>\epsilon$, $ H_{0}$ the restriction is accepted.
#Otherwise, it is rejected. We are confident at 1% about the restriction
#hypothesis


# restricted estimation - imposition of the simmetry assumption to the
# translog cost function

fitsur1re<- systemfit(list(Africa = S1, FSU = S2, Europe = S3,
America = S4), method="SUR" ,data=dat, restrict.matrix=
restrictM)
summary(fitsur1re)


#  likelihood Ratio test 1 - accept with alpha=1% p-value=0.01974838
install.packages("lmtest")
library("lmtest", lib.loc="~/R/win-library/3.2")
lrTest1 <- lrtest( fitsur1re, fitsur1un )
print( lrTest1 )

#-------------------------------------------------------------------
# MODEL 2
# selection model - constraints 2 imposition of null coefficients
# Africa_lP.f and FSU_lP.af

restrictM2 <- matrix(c(0,0,0,0,0,  0,1,0,0,0,  0,0,0,0,0,  0,0,0,0,0,
                       0,0,0,-1,0,  0,0,0,0,0,  0,1,0,0,0,  0,0,0,0,0,
                       0,0,0,0,-1,  0,0,0,0,0,  0,0,0,0,0,  0,1,0,0,0,
                       0,0,0,0,0,  0,0,0,-1,0,   0,0,1,0,0,  0,0,0,0,0,
                       0,0,0,0,0,  0,0,0,0,-1,   0,0,0,0,0,  0,0,1,0,0,
                       0,0,0,0,0,  0,0,0,0,0,   0,0,0,0,-1,  0,0,0,1,0,
                       0,0,1,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0),
                     byrow=TRUE, nrow=7, ncol=20)
```

```
# Theil's F test for testing all constraints - 2

test.constr.2<-linearHypothesis( fitsur1un, restrictM2, test = c( "FT",
 "F", "Chisq" ))


# restricted estimation 2

fitsur2<- systemfit(list(Africa = S1, FSU = S2, Europe = S3, America =
S4), method="SUR" ,data=dat, restrict.matrix= restrictM2)
summary(fitsur3)
#  likelihood Ratio test 2 - accept with alpha=1% p-value=0.0339223

lrTest2 <- lrtest( fitsur2, fitsur1un )
print( lrTest2 )

#  likelihood Ratio test 3 - accept with alpha=1% and 5% p-value=0.751309


lrTest3 <- lrtest( fitsur2, fitsur1re )
print( lrTest3 )


#------------------------------------------------------------------------
# MODEL 3
# equazioni simulatenee da stimare -modello 3
S12 <- S.af ~ 1+lP.af + lP.e+ lP.am
S22 <- S.f ~ 1 + lP.f + lP.e+ lP.am
S32 <- S.e ~ 1+lP.af + lP.f + lP.e+ lP.am
S42<- S.am ~ 1+lP.af + lP.f + lP.e+ lP.am

# selection model - constraints 3

restrictM3 <- matrix(c(
  0,0,-1,0,  0,0,0,0,  0,1,0,0,0,  0,0,0,0,0,
  0,0,0,-1,  0,0,0,0,  0,0,0,0,0,  0,1,0,0,0,
  0,0,0,0,  0,0,-1,0,   0,0,1,0,0,  0,0,0,0,0,
  0,0,0,0,  0,0,0,-1,   0,0,0,0,0,  0,0,1,0,0,
  0,0,0,0,  0,0,0,0,   0,0,0,0,-1,  0,0,0,1,0),
  byrow=TRUE, nrow=5, ncol=18)
# Theil's F test for testing all constraints - 2

test.constr.3<-linearHypothesis( fitsur1un, restrictM3, test = c( "FT",
"F", "Chisq" ))

# restricted estimation 3

fitsur3<- systemfit(list(Africa = S12, FSU = S22, Europe = S32,
America = S42),, method="SUR" ,data=dat, restrict.matrix= restrictM3)

summary(fitsur3)

#  likelihood Ratio test 4 - accept with alpha=1% and 5% 0.7513099
```

```
lrTest4 <- lrtest( fitsur3, fitsur1re )
print( lrTest4 )

#  likelihood Ratio test 5 - accept with alpha=1% and 5% 0.7513099


lrTest5 <- lrtest( fitsur3, fitsur1re )
print( lrTest5 )


#---------------------------------------------------------------------
# MODEL 4
# equazioni simulatenee da stimare -modello 4
S14 <- S.af ~ 1 + lP.e+ lP.am
S24 <- S.f ~ 1 + lP.f + lP.e+ lP.am
S34 <- S.e ~ 1+lP.af + lP.f + lP.e+ lP.am
S44<- S.am ~ 1+lP.af + lP.f + lP.e+ lP.am

# selection model - constraints 4

restrictM4 <- matrix(c(
  0,1,0,  0,0,0,0,  0,-1,0,0,0,  0,0,0,0,0,
  0,0,1,  0,0,0,0,  0,0,0,0,0,  0,-1,0,0,0,
  0,0,0,  0,0,-1,0,   0,0,1,0,0,  0,0,0,0,0,
  0,0,0,  0,0,0,-1,   0,0,0,0,0,  0,0,1,0,0,
  0,0,0,  0,0,0,0,   0,0,0,0,-1,  0,0,0,1,0),
  byrow=TRUE, nrow=5, ncol=17)

# restricted estimation 4

fitsur4<- systemfit(list(Africa = S14, FSU = S24, Europe = S34,
 America = S44), method="SUR" ,data=dat, restrict.matrix= restrictM4)

summary(fitsur4)

#  likelihood Ratio test 6 - accept with alpha=1% and 5% p-value= 0.10167



lrTest6 <- lrtest( fitsur4, fitsur3 )
print( lrTest6 )

#  likelihood Ratio test 7 - accept with alpha=1% and 5% p-value=0.249134


lrTest7 <- lrtest( fitsur4, fitsur1re )
print( lrTest7 )

#---------------------------------------------------------------------
# MODEL 5
# equazioni simulatenee da stimare -modello 5
S15 <- S.af ~ 1 + lP.e
S25 <- S.f ~ 1 + lP.f + lP.e+ lP.am
S35 <- S.e ~ 1+lP.af + lP.f + lP.e+ lP.am
```

```
S45<- S.am ~ 1 + lP.f + lP.e+ lP.am

# selection model - constraints 5

restrictM5 <- matrix(c(
  0,1,  0,0,0,0,  0,-1,0,0,0,  0,0,0,0,
  0,0,  0,0,-1,0,   0,0,1,0,0,  0,0,0,0,
  0,0,  0,0,0,-1,   0,0,0,0,0,  0,1,0,0,
  0,0,  0,0,0,0,   0,0,0,0,-1,  0,0,1,0),
  byrow=TRUE, nrow=4, ncol=15)

# restricted estimation 5

fitsur5<- systemfit(list(Africa = S15, FSU = S25, Europe = S35,
America = S45),, method="SUR" ,data=dat, restrict.matrix= restrictM5)

summary(fitsur5)

#  likelihood Ratio test 8 - accept with alpha=1% and 5% p-value= 0.05443


lrTest8 <- lrtest( fitsur5, fitsur4 )
print( lrTest8 )

#  likelihood Ratio test 9 - accept with alpha=1% p-value= 0.01043849


lrTest9 <- lrtest( fitsur5, fitsur1re )
print( lrTest9 )

# conservo i valori numerici trasformandoli in dati
print(summary(fitsur5)$coefficients)

coeff1un<-summary(fitsur1un)$coefficients
coeff1re<-summary(fitsur1re)$coefficients
coeff2<-summary(fitsur2)$coefficients
coeff3<-summary(fitsur3)$coefficients
coeff4<-summary(fitsur4)$coefficients
coeff5<-summary(fitsur5)$coefficients

#-----------------------------------------------------------------

# adding the implicit coefficients calculated for Middle east to the
#already calculated coefficients (see the paper)
#for MODEL 1
new.coeff1re<-c(coeff1re[,1],
Middle.East_Intercept=1-coeff1re[1,1]-coeff1re[6,1]-coeff1re[11,1]-
coeff1re[16,1],

Middle.East_LP.af=-coeff1re[2,1]- coeff1re[3,1]- coeff1re[4,1]-
coeff1re[5,1],

Middle.East_LP.f=-coeff1re[7,1]- coeff1re[8,1]- coeff1re[9,1]-
coeff1re[10,1],
```

```
Middle.East_LP.e=-coeff1re[12,1]- coeff1re[13,1]- coeff1re[14,1]-
coeff1re[15,1],

Middle.East_LP.am=-coeff1re[17,1]- coeff1re[18,1]- coeff1re[19,1]-
coeff1re[20,1],

Middle.East_LP.me=-(-coeff1re[2,1]- coeff1re[3,1]- coeff1re[4,1]-
coeff1re[5,1])-( -coeff1re[7,1]- coeff1re[8,1]-
coeff1re[9,1]-coeff1re[10,1])-( -coeff1re[12,1]-
coeff1re[13,1]- coeff1re[14,1]-coeff1re[15,1])-( -
coeff1re[17,1]- coeff1re[18,1]- coeff1re[19,1]-
coeff1re[20,1]) )

new.coeff1re<-cbind(new.coeff1re)


#for MODEL 5
new.coeff5<-c(coeff5[,1]
  ,Intercept.me5=1-coeff5[1,1]-coeff5[3,1]-coeff5[7,1]-coeff5[12,1],

  Middle.East_LP.af5=- coeff5[2,1],

  Middle.East_LP.f5=-coeff5[4,1]- coeff5[5,1]-coeff5[6,1],

  Middle.East_LP.e5=-coeff5[8,1]- coeff5[9,1]- coeff5[10,1]-coeff5[11,1],

  Middle.East_LP.am5=- coeff5[13,1]- coeff5[14,1]-coeff5[15,1],

  Middle.East_LP.me5=-(- coeff5[2,1])-( -coeff5[4,1]- coeff5[5,1]-
coeff5[6,1])-( -coeff5[8,1]- coeff5[9,1]- coeff5[10,1]-
coeff5[11,1])-( - coeff5[13,1]- coeff5[14,1]-
coeff5[15,1]))

new.coeff5<-cbind(new.coeff5)

#-----------------------------------------------------------------
# cheking curvature of the cost function
library("micEcon", lib.loc="~/R/win-library/3.2")

new.coeff.list<-list(new.coeff5)


co5<- c(
  const=0,
  Af=new.coeff5[1,1],
  F=new.coeff5[3,1],
  E=new.coeff5[7,1],
  Am=new.coeff5[12,1],
  Me=new.coeff5[16,1],
  AfAf=0,
  Aff=0,
  Afe=new.coeff5[2,1],
  AfAm=0,
```

```
  Afme=new.coeff5[17,1],
  ff=new.coeff5[4,1],
  fe=new.coeff5[5,1],
  fam=new.coeff5[6,1],
  fme=new.coeff5[18,1],
  ee=new.coeff5[10,1],
  eam=new.coeff5[11,1],
  eme=new.coeff5[19,1],
  amam=new.coeff5[15,1],
  amme=new.coeff5[20,1],
  meme=new.coeff5[21,1]
)


co5<-cbind(co5)
library("numDeriv", lib.loc="~/R/win-library/3.2")
F5=function(x) c(
exp(co5[1]+co5[2]*log(x[1])+co5[3]*log(x[2])+co5[4]*log(x[3])+
co5[5]*log(x[4])+co5[6]*log(x[5])+
    co5[7]* log(x[1])*log(x[1])*0.5+
    co5[8]* log(x[1])*log(x[2])+
    co5[9]* log(x[1])*log(x[3])+
    co5[10]* log(x[1])*log(x[4])+
    co5[11]* log(x[1])*log(x[5])+
    co5[12]* log(x[2])*log(x[2])*0.5+
    co5[13]* log(x[2])*log(x[3])+
    co5[14]* log(x[2])*log(x[4])+
    co5[15]* log(x[2])*log(x[5])+
    co5[16]* log(x[3])*log(x[3])*0.5+
    co5[17]* log(x[3])*log(x[4])+
    co5[18]* log(x[3])*log(x[5])+
    co5[19]* log(x[4])*log(x[4])*0.5+
    co5[20]* log(x[4])*log(x[5])+
    co5[21]* log(x[5])*log(x[5])*0.5
))


library("numDeriv", lib.loc="~/R/win-library/3.2")
F6=function(x) c(
  co5[1]+co5[2]*log(x[1])+co5[3]*log(x[2])+
  co5[4]*log(x[3])+co5[5]*log(x[4])+co5[6]*log(x[5])+
        co5[7]* log(x[1])*log(x[1])*0.5+
        co5[8]* log(x[1])*log(x[2])+
        co5[9]* log(x[1])*log(x[3])+
        co5[10]* log(x[1])*log(x[4])+
        co5[11]* log(x[1])*log(x[5])+
        co5[12]* log(x[2])*log(x[2])*0.5+
        co5[13]* log(x[2])*log(x[3])+
        co5[14]* log(x[2])*log(x[4])+
        co5[15]* log(x[2])*log(x[5])+
        co5[16]* log(x[3])*log(x[3])*0.5+
        co5[17]* log(x[3])*log(x[4])+
        co5[18]* log(x[3])*log(x[5])+
        co5[19]* log(x[4])*log(x[4])*0.5+
```

```
       co5[20]* log(x[4])*log(x[5])+
       co5[21]* log(x[5])*log(x[5])*0.5
  )




hess5<-hessian(F5,c(dat[1,2], dat[1,3], dat[1,4], dat[1,5], dat[1,6]))
jacob5<-jacobian(F5,c(dat[1,2], dat[1,3], dat[1,4], dat[1,5], dat[1,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_1<-quasiconcavity( hess35 )
print(qcv5_1)



hess5<-hessian(F5,c(dat[2,2], dat[2,3], dat[2,4], dat[2,5], dat[2,6]))
jacob5<-jacobian(F5,c(dat[2,2], dat[2,3], dat[2,4], dat[2,5], dat[2,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_2<-quasiconcavity( hess35 )
print(qcv5_2)

hess5<-hessian(F5,c(dat[3,2], dat[3,3], dat[3,4], dat[3,5], dat[3,6]))
jacob5<-jacobian(F5,c(dat[3,2], dat[3,3], dat[3,4], dat[3,5], dat[3,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_3<-quasiconcavity( hess35 )
print(qcv5_3)

hess5<-hessian(F5,c(dat[4,2], dat[4,3], dat[4,4], dat[4,5], dat[4,6]))
jacob5<-jacobian(F5,c(dat[4,2], dat[4,3], dat[4,4], dat[4,5], dat[4,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_4<-quasiconcavity( hess35 )
print(qcv5_4)

hess5<-hessian(F5,c(dat[5,2], dat[5,3], dat[5,4], dat[5,5], dat[5,6]))
jacob5<-jacobian(F5,c(dat[5,2], dat[5,3], dat[5,4], dat[5,5], dat[5,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_5<-quasiconcavity( hess35 )
print(qcv5_5)
```

```
hess5<-hessian(F5,c(dat[6,2], dat[6,3], dat[6,4], dat[6,5], dat[6,6]))
jacob5<-jacobian(F5,c(dat[6,2], dat[6,3], dat[6,4], dat[6,5], dat[6,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_6<-quasiconcavity( hess35 )
print(qcv5_6)

hess5<-hessian(F5,c(dat[7,2], dat[7,3], dat[7,4], dat[7,5], dat[7,6]))
jacob5<-jacobian(F5,c(dat[7,2], dat[7,3], dat[7,4], dat[7,5], dat[7,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_7<-quasiconcavity( hess35 )
print(qcv5_7)

hess5<-hessian(F5,c(dat[7,2], dat[7,3], dat[7,4], dat[7,5], dat[7,6]))
jacob5<-jacobian(F5,c(dat[7,2], dat[7,3], dat[7,4], dat[7,5], dat[7,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_7<-quasiconcavity( hess35 )
print(qcv5_7)


hess5<-hessian(F5,c(dat[8,2], dat[8,3], dat[8,4], dat[8,5], dat[8,6]))
jacob5<-jacobian(F5,c(dat[8,2], dat[8,3], dat[8,4], dat[8,5], dat[8,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_8<-quasiconcavity( hess35 )
print(qcv5_8)

hess5<-hessian(F5,c(dat[9,2], dat[9,3], dat[9,4], dat[9,5], dat[9,6]))
jacob5<-jacobian(F5,c(dat[9,2], dat[9,3], dat[9,4], dat[9,5], dat[9,6]))
hess25 <- rbind(jacob5,hess5)
jacob25<-c(0,jacob5)
hess35<-cbind(jacob25,hess25)
library("miscTools", lib.loc="~/R/win-library/3.2")
qcv5_8<-quasiconcavity( hess35 )
print(qcv5_9)


#-----------------------------------------------------------------
# Compute the nearest negative Hessian matrix of the translog cost
# function LN(C) considering that if a monotone transformation
# of a function is concave then the original function is quasiconcave

library("Matrix", lib.loc="C:/Program␣Files/R/R-3.2.1/library")
```

```
hess52<-hessian(F6,c(dat[168,2], dat[168,3], dat[168,4], dat[168,5],
dat[168,6]))

# in order ti do that we use the PD function on the negative value of the
#hessian matrix obtained with the last linear constrained SUR model
neghess5<- nearPD(-hess5)
print(neghess5)
neghess5<-cbind(neghess5)
startpoint<-cbind(-neghess5[[1]]@x[])
View(startpoint)

#-----------------------------------------------------------------
#-----------------------------------------------------------------
# Normalize every single price with respect to its mean price

dat$P.af2<- dat$P.af./mean(dat$P.af.)
dat$P.f2<- dat$P.f./mean(dat$P.f.)
dat$P.e2<- dat$P.e./mean(dat$P.e.)
dat$P.am2<- dat$P.am./mean(dat$P.am.)
dat$P.me2<- dat$P.me./mean(dat$P.me.)


dat$S.af2 <- I(dat$P.af2 * dat$Q.af) / I(dat$P.af2 * dat$Q.af+
dat$P.f2 * dat$Q.f+ dat$P.e2 * dat$Q.e+ dat$P.am2* dat$Q.am+
dat$P.me2* dat$Q.me)
dat$S.f2 <- I(dat$P.f2 * dat$Q.f) / I(dat$P.af2 * dat$Q.af+
dat$P.f2 * dat$Q.f+ dat$P.e2 * dat$Q.e+ dat$P.am2* dat$Q.am+
dat$P.me2* dat$Q.me)
dat$S.e2 <- I(dat$P.e2 * dat$Q.e) / I(dat$P.af2 * dat$Q.af+
dat$P.f2 * dat$Q.f+ dat$P.e2 * dat$Q.e+ dat$P.am2* dat$Q.am+
dat$P.me2* dat$Q.me)
dat$S.am2 <- I(dat$P.am2 * dat$Q.am) / I(dat$P.af2 * dat$Q.af+
dat$P.f2 * dat$Q.f+ dat$P.e2 * dat$Q.e+ dat$P.am2* dat$Q.am+
dat$P.me2* dat$Q.me)
dat$S.me2 <- I(dat$P.me2 * dat$Q.me) / I(dat$P.af2 * dat$Q.af+
dat$P.f2 * dat$Q.f+ dat$P.e2 * dat$Q.e+ dat$P.am2* dat$Q.am+
dat$P.me2* dat$Q.me)
# add logarithms
dat$lP.af2 <- I(log(dat$P.af2 /dat$P.me2))
dat$lP.f2 <- I(log(dat$P.f2 /dat$P.me2))
dat$lP.e2 <- I(log(dat$P.e2 /dat$P.me2))
dat$lP.am2<- I(log(dat$P.am2 /dat$P.me2) )



# co[7]=-k11^2-b1^2+b1
# co[12]=-k22^2-k21^2-b2^2+b2
# co[16]=-k33^2-k32^2-k31^2-b3^2+b3
# co[19]=-k44^2-k43^2-k42^2-k41^2-b4^2+b4
# co[8]=-(k11)*(k21)-b1*b2
# co[13]=-(k22)*(k32)-(k21^2)*(k31)-b2*b3
# co[17]=-(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-b3*b4
# co[9]=-(k11)*(k31)-b1*b3
# co[14]=-(k22)*(k42)-(k21^2)*(k41)-b2*b4
```

```
# co[10]=-(k11)*(k41)-b1*b4
# co[1]=b0
# co[2]=b1
# co[3]=b2
# co[4]=b3
# co[5]=b4
# co[6]=1-b1-b2-b3-b4
# co[11]=-(-k11^2-b1^2+b1)-(-(k11)*(k21)-b1*b2)-(-
#(k11)*(k31)-b1*b3)-(-(k11)*(k41)-b1*b4)
# co[15]=-(-(k11)*(k21)-b1*b2)-(-k22^2-k21^2-b2^2+b2)-(-(k22)*(k32)-
#(k21^2)*(k31)-b2*b3)-(-(k22)*(k42)-(k21^2)*(k41)-b2*b4)
# co[18]=-(-(k11)*(k31)-b1*b3)-(-(k22)*(k32)-(k21^2)*(k31)-b2*b3)-
#( -k33^2-k32^2-k31^2-b3^2+b3)-(-(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-
#b3*b4)
# co[20]=-(-(k11)*(k41)-b1*b4)-(-(k22)*(k42)-(k21^2)*(k41)-b2*b4)-
#( -(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-b3*b4)-( -k44^2-k43^2-k42^2-
#k41^2-b4^2+b4)
# co[21]=-(-(-k11^2-b1^2+b1)-(-(k11)*(k21)-b1*b2)-(-(k11)*(k31)-b1*b3)-
#(-(k11)*(k41)-b1*b4))-( -(-(k11)*(k21)-b1*b2)-(-k22^2-k21^2-b2^2+b2)-(-
#(k22)*(k32)-(k21^2)*(k31)-b2*b3)-(-(k22)*(k42)-(k21^2)*(k41)-b2*b4))-( -
#(-(k11)*(k31)-b1*b3)-(-(k22)*(k32)-(k21^2)*(k31)-b2*b3)-
#( -k33^2-k32^2-k31^2-b3^2+b3)-(-(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-
#b3*b4))-( -(-(k11)*(k41)-b1*b4)-(-(k22)*(k42)-(k21^2)*(k41)-b2*b4)-
#( -(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-
#b3*b4)-( -k44^2-k43^2-k42^2-k41^2-b4^2+b4))

# MODEL 1 - our new system to fit

library("systemfit", lib.loc="~/R/win-library/3.2")
S52 <- S.af2 ~ b1+(-k11^2-b1^2+b1)*lP.af2 + (-(k11)*(k21)-b1*b2)*lP.f2 +
(-(k11)*(k31)-b1*b3)*lP.e2+ (-(k11)*(k41)-b1*b4)*lP.am2

S62<- S.f2 ~ b2+(-(k11)*(k21)-b1*b2)*lP.af2 +
(-k22^2-k21^2-b2^2+b2)*lP.f2 + (-(k22)*(k32)-
(k21^2)*(k31)-b2*b3)*lP.e2+(-(k22)*(k42)-
(k21^2)*(k41)-b2*b4)*lP.am2

S72 <- S.e2 ~ b3+(-(k11)*(k31)-b1*b3)*lP.af2 + (-(k22)*(k32)-
(k21^2)*(k31)-b2*b3)*lP.f2 + (-k33^2-k32^2-k31^2-b3^2+b3)*lP.e2+
(-(k33)*(k43)-(k32^2)*(k42)-(k31^2)*(k41)-b3*b4)*lP.am2

S82<- S.am2 ~ b4+(-(k11)*(k41)-b1*b4)*lP.af2 + (-(k22)*(k42)-
(k21^2)*(k41)-b2*b4)*lP.f2 + (-(k33)*(k43)-(k32^2)*(k42)-
(k31^2)*(k41)-b3*b4)*lP.e2+ (-k44^2-k43^2-k42^2-k41^2-b4^2+b4)*lP.am2

# our starting point is based on the nearest negative definite Hessian
#matrix previously calculated (see the paper)

start.values <- c(
  b1=runif(1, -1, 1),
  b2=runif(1, -1, 1),
  b3=runif(1, -1, 1),
  b4=runif(1, -1, 1),
  k11=0.69930544430958,
```

```
  k21=runif(1, -1, 1),
  k31=runif(1, -1, 1),
  k41=runif(1, -1, 1),
  k22=runif(1, -1, 1),
  k32=runif(1, -1, 1),
  k42=runif(1, -1, 1),
  k33=0.000000036516126261654200000000,
  k43=runif(1, -1, 1),
  k44=runif(1, -1, 1)
)


model <- list(S52,S62,S72,S82)

# Compute the nonlinear SUR
Nlfitsur<- nlsystemfit( "SUR", model, start.values, data=dat)
Nlfitsur$b
Nlfitsur$p
Nlfitsur$r2
Nlfitsur$adjr2
sum(Nlfitsur$p[])

# MODEL 2 - our new system to fit

S52 <- S.af2 ~ b1+(-k11^2-b1^2+b1)*lP.af2 + (-(k11)*(k21)-b1*b2)*lP.f2 +
(-(k11)*(k31)-b1*b3)*lP.e2+ (-(k11)*(k41)-b1*b4)*lP.am2

S62<- S.f2 ~ b2+(-(k11)*(k21)-b1*b2)*lP.af2 + (-0^2-k21^2-b2^2+b2)*lP.f2
+(-(0)*(0)-(k21^2)*(k31)-b2*b3)*lP.e2+ (-(0)*(0)-(k21^2)*(k41)-
b2*b4)*lP.am2

S72 <- S.e2 ~ b3+(-(k11)*(k31)-b1*b3)*lP.af2 + (-(0)*(0)-(k21^2)*(k31)-
b2*b3)*lP.f2 + (-0^2-0^2-k31^2-b3^2+b3)*lP.e2+ (-(0)*(0)-(0^2)*(0)
-(k31^2)*(k41)-b3*b4)*lP.am2

S82<- S.am2 ~ b4+(-(k11)*(k41)-b1*b4)*lP.af2 + (-(0)*(0)-(k21^2)*(k41)-
b2*b4)*lP.f2 + (-(0)*(0)-(0^2)*(0)-(k31^2)*(k41)-b3*b4)*lP.e2+
(-0^2-0^2-0^2-k41^2-b4^2+b4)*lP.am2

start.values <- c(
  b1=runif(1, -1, 1),
  b2=runif(1, -1, 1),
  b3=runif(1, -1, 1),
  b4=runif(1, -1, 1),
  k11=runif(1, -1, 1),
  k21=runif(1, -1, 1),
  k31=runif(1, -1, 1),
  k41=runif(1, -1, 1)
)


model <- list(S52,S62,S72,S82)

# Compute the nonlinear SUR
```

```
Nlfitsur2<- nlsystemfit( "SUR", model, start.values, data=dat)
Nlfitsur2$b
Nlfitsur2$p

Nlfitsur2$r2*
Nlfitsur2$adjr2
sum(Nlfitsur2$p)


# Test the goodness of fit of the model 2

# compute likelihood of the model 1

n<-Nlfitsur$n/Nlfitsur$g
p<-Nlfitsur$g
# Matrix inverse
x<-Nlfitsur$rcovest
kk=svd(x)
zz=kk$v%*%diag(1/kk$d)%*%t(kk$u)
# logaritmo del determinante della matrice di varianza covarianza del
# modello SUR nonlineare
logdet<-determinant((kronecker(x,Diagonal(n, 1))))$modulus[1]
# Loglikelihood function of the residuals
r<-c(Nlfitsur$resid[,1],Nlfitsur$resid[,2],Nlfitsur$resid[,3],
Nlfitsur$resid[,4])
jj<-kronecker(zz,Diagonal(n, 1))
L=-((n*p)/2)*logdet-0.5*t(r)%*%jj%*%(r)-1/2*log(2*pi)

# likelihood model 2
n<-Nlfitsur2$n/Nlfitsur2$g
p<-Nlfitsur2$g
# Matrix inverse
x<-Nlfitsur2$rcovest
kk=svd(x)
zz=kk$v%*%diag(1/kk$d)%*%t(kk$u)
# logarithm of the determinant of the variance covariance matrix ofthe
#nonlinear SUR model
logdet<-determinant((kronecker(x,Diagonal(n, 1))))$modulus[1]
# Loglikelihood function of the residuals
r<-c(Nlfitsur2$resid[,1],Nlfitsur2$resid[,2],Nlfitsur2$resid[,3],
Nlfitsur2$resid[,4])
jj<-kronecker(zz,Diagonal(n, 1))
L2=-(1/2)*logdet-0.5*t(r)%*%jj%*%(r)-(n*p)/2*log(2*pi)

kl<-2*(L2-L)
# Find the 95th percentile of the Chi-Squared distribution with 7 degrees
# of freedom.
qchisq(.95, df=14-8)        # 7 degrees
#of freedom
# is the area or probability in the upper tail of the chi-square
# distribution at the 95th percentile of the Chi-Squared distribution
# with 7 degrees of freedom.

1-pchisq(kl[1,1], df=14-8) # P-value of the loglikelihood ratio test -
```

```
# p-value=1 we accept model 2



# MODEL 3 - our new system to fit

S52 <- S.af2 ~ b1+(-k11^2-b1^2+b1)*lP.af2 + (-(k11)*(k21)-b1*b2)*lP.f2 +
(-(k11)*(0)-b1*b3)*lP.e2+ (-(k11)*(k41)-b1*b4)*lP.am2

S62<- S.f2 ~ b2+(-(k11)*(k21)-b1*b2)*lP.af2 + (-0^2-k21^2-b2^2+b2)*lP.f2
+ (-(0)*(0)-(k21^2)*(0)-b2*b3)*lP.e2+ (-(0)*(0)-(k21^2)*(k41)-
b2*b4)*lP.am2

S72 <- S.e2 ~ b3+(-(k11)*(0)-b1*b3)*lP.af2 + (-(0)*(0)-(k21^2)*(0)-
b2*b3)*lP.f2 +(-0^2-0^2-0^2-b3^2+b3)*lP.e2+ (-(0)*(0)-(0^2)*(0)-
(0^2)*(k41)-b3*b4)*lP.am2

S82<- S.am2 ~ b4+(-(k11)*(k41)-b1*b4)*lP.af2 + (-(0)*(0)-(k21^2)*(k41)-
b2*b4)*lP.f2 + (-(0)*(0)-(0^2)*(0)-(0^2)*(k41)-b3*b4)*lP.e2+
(-0^2-0^2-0^2-k41^2-b4^2+b4)*lP.am2

start.values <- c(
  b1=runif(1, -1, 1),
  b2=runif(1, -1, 1),
  b3=runif(1, -1, 1),
  b4=runif(1, -1, 1),
  k11=runif(1, -1, 1),
  k21=runif(1, -1, 1),
  k41=runif(1, -1, 1)
)



model <- list(S52,S62,S72,S82)

# Compute the nonlinear SUR
Nlfitsur3<- nlsystemfit( "SUR", model, start.values, data=dat)
Nlfitsur3$b
Nlfitsur3$p
Nlfitsur3$r2
Nlfitsur3$adjr2
sum(Nlfitsur3$p[])


# Test the goodness of fit of the model 3 with respect to model 2

# likelihood model 3
n<-Nlfitsur3$n/Nlfitsur3$g
p<-Nlfitsur3$g
# Matrix inverse
x<-Nlfitsur3$rcovest
kk=svd(x)
zz=kk$v%*%diag(1/kk$d)%*%t(kk$u)
# logarithm of the determinant of the variance covariance
```

```
# matrix of the nonlinear SUR model
logdet<-determinant((kronecker(x,Diagonal(n, 1))))$modulus[1]
# Loglikelihood function of the residuals
r<-c(Nlfitsur3$resid[,1],Nlfitsur3$resid[,2],Nlfitsur3$resid[,3],
Nlfitsur3$resid[,4])
jj<-kronecker(zz,Diagonal(n, 1))
L3=-(1/2)*logdet-0.5*t(r)%*%jj%*%(r)-(n*p)/2*log(2*pi)

kl<-2*(L3-L2)
# Find the 95th percentile of the Chi-Squared distribution with 7 degrees
#of freedom.
qchisq(.95, df=8-7)        # 7 degrees of freedom
# is the area or probability in the upper tail of the chi-square
# distribution at the 95th percentile of the Chi-Squared distribution
# with 7 degrees of freedom.

1-pchisq(kl[1,1], df=8-7) # P-value of the loglikelihood ratio test
# p-value=0.08485123 we accept model 3



# MODEL 4 - our new system to fit
S52 <- S.af2 ~ b1+(-k11^2-b1^2+b1)*lP.af2 + (-(k11)*(k21)-b1*b2)*lP.f2 +
(-(k11)*(0)-b1*b3)*lP.e2+(-(k11)*(0)-b1*b4)*lP.am2

S62<- S.f2 ~ b2+(-(k11)*(k21)-b1*b2)*lP.af2 + (-0^2-k21^2-b2^2+b2)*lP.f2
+ (-(0)*(0)-(k21^2)*(0)-b2*b3)*lP.e2+ (-(0)*(0)-(k21^2)*(0)-
b2*b4)*lP.am2

S72 <- S.e2 ~ b3+(-(k11)*(0)-b1*b3)*lP.af2 + (-(0)*(0)-(k21^2)*(0)-
b2*b3)*lP.f2 + (-0^2-0^2-0^2-b3^2+b3)*lP.e2+ (-(0)*(0)-(0^2)*(0)-
(0^2)*(0)-b3*b4)*lP.am2

S82<- S.am2 ~ b4+(-(k11)*(0)-b1*b4)*lP.af2 + (-(0)*(0)-(k21^2)*(0)-
b2*b4)*lP.f2 +(-(0)*(0)-(0^2)*(0)-(0^2)*(0)-b3*b4)*lP.e2+ (-0^2-0^2-0^2-
0^2-b4^2+b4)*lP.am2

start.values <- c(
  b1=runif(1, -1, 1),
  b2=runif(1, -1, 1),
  b3=runif(1, -1, 1),
  b4=runif(1, -1, 1),
  k11=runif(1, -1, 1),
  k21=runif(1, -1, 1)
)


model <- list(S52,S62,S72,S82)

# Compute the nonlinear SUR
Nlfitsur4<- nlsystemfit( "SUR", model, start.values, data=dat)
Nlfitsur4$b
Nlfitsur4$p
Nlfitsur4$r2
```

```
Nlfitsur4$adjr2
sum(Nlfitsur4$p[])

# Test the goodness of fit of the model 4 with respect to model 3

# likelihood model 4
n<-Nlfitsur4$n/Nlfitsur4$g
p<-Nlfitsur4$g
# Matrix inverse
x<-Nlfitsur4$rcovest
kk=svd(x)
zz=kk$v%*%diag(1/kk$d)%*%t(kk$u)
# logarithm of the determinant of the variance covariance matrix of
#the nonlinear SUR model
logdet<-determinant((kronecker(x,Diagonal(n, 1))))$modulus[1]
# Loglikelihood function of the residuals
r<-c(Nlfitsur4$resid[,1],Nlfitsur4$resid[,2],Nlfitsur4$resid[,3],
Nlfitsur4$resid[,4])
jj<-kronecker(zz,Diagonal(n, 1))
L4=-(1/2)*logdet-0.5*t(r)%*%jj%*%(r)-(n*p)/2*log(2*pi)

kl<-2*(L4-L3)
# Find the 95th percentile of the Chi-Squared distribution with 7 degrees
#of freedom.
qchisq(.95, df=7-6)        # 7 degrees of freedom
# is the area or probability in the upper tail of the chi-square
# distribution at the 95th percentile of the Chi-Squared distribution
# with 7 degrees of freedom.

1-pchisq(kl[1,1], df=7-6) # P-value of the loglikelihood ratio test
# p-value=0.162945 we accept model 4


print( Nlfitsur )
print( Nlfitsur2 )
print( Nlfitsur3 )
print( Nlfitsur4 )
```

Predictability Information Criterion for Selecting Stochastic Pricing Models

Gabriele D'Amore*

May 8, 2017

Abstract

Pricing models of derivative instruments usually fail to provide reliable results when risks rise and financial crises occur. More advanced stochastic pricing models try to improve the fitting results adding risk factors and/or parameters to the models, incurring the risk of overfitted results. Drawing on these observations, it is proposed a generalisation of the Akaike Information Criterion (AIC) suitable to evaluate forecasting power of alternative stochastic pricing models[1] for any fixed arbitrary forecasting time-horizon. The Predictability Information Criterion (PIC) differs from the classical criteria for evaluating statistical models as it assumes that the random variable to study can (or cannot) be partially predictable, which makes it particularly suitable for studying stochastic pricing models coherently with the semimartingale definition of the price process. On the basis of this assumption the criterion measures and compares the uncertainty of the predictions of two different alternative models when prices are (or are not) predictable. We conclude with a focus on the crude oil market by comparing GBM and OU stochastic processes that are commonly used for modeling West Texas Intermediate (WTI) oil spot price returns in derivative pricing models.

Keywords: Model Selection, Information Theory, Predictability, GBM, OU, Crude Oil.
JEL-Classification: C19, G120, D81

1 Introduction

Recently, market competition and technological advancement have helped the flourishing of many sophisticated models, facilitating the creation of new, increasingly complex, derivatives and probably the financialization of old markets such as those of commodities. However, although those new products have allowed an expansion of the financial markets, increasing the liquidity, they have also introduced more complexity making them difficult to model and price.

*Sapienza University of Rome. Mail to: gabriele.damore@uniroma1.it. Corresponding author at: Department of Economics and Social Sciences, Piazzale AldoMoro, 5 - 00185 Rome(IT).

[1] In the sense that each one provides a prediction for the same economic variable

Nowadays, far more sophisticated models than the traditional ones have become necessary. This has led quantitative finance plays a leading role in pricing, risk management and hedging practices since the past two decades. However, despite the intellectual effort of quants, it is suspected that inadequate pricing models are frequently the cause of most of the losses on derivatives. For example, as reported in Cont(2006) [26]the application of inadequate mathematical models in risk management practices has been the cause of massive losses by financial institutions on several occasions. In this respect, Cont mentioned the $ 83 million loss incurred by the Bank of Tokyo / Mitsubishi, due to the overvaluation of a portfolio of swaps and options, and the £ 50 million loss suffered by NatWest Capital Markets in London because of a mispriced portfolio of German and U.K. interest rate options and swaptions.
The literature on stochastic modelling constantly deals with this issue improving the quality of the stochastic models regarding the underlying asset(s), but this seems not to be sufficient.
There are several limitations of quantitative pricing model studies:
1) "Model fitting": quantitative finance and economic theory have considerable difficulties in adequately describing and predicting the dynamic and the shape of the future risks underlying the price of assets and derivatives, resulting in the inability to determine their present value correctly. Paul Wilmott (2009) points out two relevant issues: a) the extreme complexity of the model prevents to understand whether the model is adequate to describe reality; b) the attempts to improve the quality of the stochastic models usually increases the degree of model complexity, but it doesn't necessarily mean it's actually worth it. To seek the extreme precision in the model calibration is a mere illusion when models are highly sensitive to changes in the parameters (high complexity) or when the analysed phenomena are highly unstable because of human behaviour.
2) "The ongoing evolution of the economic theory": economic theory concerning the pricing of assets is still imperfect and evolving. Some questions still have high relevance:
a) *is the asset price predictable?* Cochran (2009)[25] wrote: "in the early 1970s... stock returns were considered close to unpredictable and prices close to random walks". All are dramatically different nowadays: long-term stock returns are considered predictable (long run, business cycle correlation) and prices move on news of discount rate changes (see Lucas (1978)[80]).
b) *How do asset prices and returns behave over time?* Market noise plays a relevant role, Timmeramn et.al (2004)[106] said: "prices and values need not be closely related" since "Investors' information can be so 'noisy' (see Black) at times that prices are far removed from fundamentals". Intrinsic value is, however, a non-observable variable (see Timmerman et al. (2004)[106]) .
Föllmer and Schweizer (1993)[45] proved under rational expectations that several stochastic diffusion processes can describe the equilibrium price dynamic depending on various types of agents' behaviour on the market. For instance, information traders or fundamentalists believe that the actual stock price is attracted to its fundamental value. In this case, they proved that logarithmic price process induced by information traders behaves like an Ornstein-Uhlenbeck process around a time dependent level. On the contrary, a model of noise trading suggests that the logarithmic price process should behave like a random walk. For these reasons, unfortunately, model selection remain a no easy task. In the light of the above-listed limitations, an ideal model selection criterion for pricing

models, for each time horizon, should:

a. evaluate what is the model providing less "uncertain" predictions;

b. ensure that wrong predictions, due to market noise, do not affect the measure of uncertainty employed for the comparison, because we're aware of their unpredictable nature;

c. evaluate the stability of this kind of uncertainty, whatever is the intrinsic value of the asset, which is "unobservable" (for instance when the intrinsic value of a portfolio is sensibly lower than the market price or vice-versa);

d. penalise highly complex models.

Following these points, I propose a statistical tool I called Predictability Information Criterion (PIC) for choosing a statistical model among alternative specified parametric models. For each proposed model, this approach estimates the uncertainty of the predictions in terms of Kullback–Leibler divergence, between market log returns and modelled log returns, and selects the model with the lowest divergence.
We formalise the market log-price from time 0 to arbitrary time s as a one-dimensional real-valued ergodic semimartingale process. According to this definition, we define $\widetilde{y}_{re}(t)$ to be the corresponding log return process. Fixed the forecasting time horizon s, the criterion requires:
a) a random sample $\left\{\widetilde{y}^{i}_{re}(s)\right\}_{i=1,\dots,n}$ of i.i.d. random variables, drawn from the same density distribution $f(y_{re}(s))$, or a linear transformation of weakly dependent random variables (see proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]);
b) a theoretical solution $\widetilde{y}^{M}_{th}(s)$ for each M-th proposed stochastic process whose randomness is provided by a family of parametric probability density functions $\{f_t(y^{M}_{th}(s),\theta);\theta \in \Theta^{M}\}$. The predictability information criterion (PIC) is estimated by replacing, for each M-th model, the unknown parameter vectors θ with the asymptotically correct estimator $\hat{\theta}$ (a likelihood estimator or a quasi-maximum likelihood estimator) and replacing $f(y_{re}(s))$ with a kernel density distribution.
The method allows extrapolating the divergence "excluding" any effect generated by wrong predictions due to market noise, which is not predictable by definition. The PIC favours less complex models penalising those with a higher number of parameters. The PIC can be interpreted as a generalisation of the Akaike Information Criterion (AIC).
At the end of the study, it is shown an example where it is compared, at several time horizons, a Geometric Brownian Motion (GBM) to an Ornstein-Uhlenbeck process (OU) using daily West Texas Intermediate (WTI) crude oil spot prices, collected by EIA from 02/01/1986 to 21/03/2016.

2 Review on pricing theory and predictability

In the financial economics literature, there is a broad consensus that the movements of asset prices can be described by stochastic models created on some filtered probability space. Typically, it is believed that such a probability measure can be identified through a mix of statistical-econometric methodology

and some economic assumptions. The imposed economic assumptions generally pertain to the concept of some degree of market efficiency. This framework of analysis has determined a long debate on the predictability of assets returns. Cochran(2009) [25] reminds us that: "much of asset pricing theory stems from one simple concept: price equals expected discounted payoff". Therefore the predictability of an asset would require the ability to predict what payoff will be paid to the asset owner in the future. One of the first studies that has implicitly dealt with the problem is due to Samuelson(1965)[91] with his seminal paper 'Proof that properly anticipated prices fluctuate randomly'. The study came to a conclusion that, in competitive markets, if the spot stock's price is supposed to be equal to the expected discounted value of its futures random dividends, the expected percentage price movements cannot accurately be anticipated or predicted. Samuelson was thus able to formalise the idea that, in equilibrium, the competitive prices must move following a random walk if agents share the same expectations. Price sequence may also exhibit statistical dependencies with past data but past variations cannot be exploited for deducing the direction and the intensity of tomorrow's price change. Fama (1963 [40]; 1965[41]; 1965[42], 1970[81]) was the first author who explicitly used the concept of efficient markets hypothesis to describe the dynamics of prices. In summary, the hypothesis of efficient markets (EMH) believes that rational agents immediately use all the available information. Therefore any change in the equilibrium price can not be anticipated because of the randomness of the updating information available on the market that instantly changes the expectations about the future price. This intuitive idea has formalised as the hypothesis that the equilibrium price X_t in an efficient market should be a martingale. The efficient market hypothesis (EMH) is usually supposed to hold in financial markets, including commodity derivatives markets. Under EMH commodity prices reflect nothing but information on fundamentals and agents' intertemporal preferences. In 1990 Sims[98]wrote that:"Price changes for a durable good with small storage costs must, in a frictionless competitive market, be in some sense unpredictable", however the empirical analyses on predictability, conducted in support of this hypothesis, are voluminous but currently considered controversial. Several empirical studies have confirmed that asset returns are predictable, at least partially, having a significant predictable component inside.(see, for example, Keim and Stambaugh (1986) [65], Campbell and Shiller (1988)[22], Fama and French (1988)[43], Hodrick (1992)[55], Stambaugh (1999)[101], Goyal and Welch (2003)[49], Valkanov (2003)[107], Lewellen (2004)[77], and Boudoukh, Richardson, and Whitelaw (2006)[17]).
However, such a component is hardly detected out of the sample (Bossaerts and Hillion (1999)[16]). Also, many theoretical studies pointed out that the theoretical hypothesis of efficient markets appears to be highly restrictive. Lucas, Jr. (1978)[80] and Stephen F. Leroy (1973) [75], among others, argue that asset prices in equilibrium models do not, in general, follow martingale processes. For instance, Lucas proposed an equilibrium price model where, given a set of securities, agents "solve a dynamic optimization problem whereby consumption" is maximally smoothed. As Lucas wrote in his article: "Within [the framework of his model], it is clear that the presence of a diminishing marginal rate of substitution [...] is inconsistent with the [martingale] property."
Mainly two theoretical explanations are provided in literature: 1) the predictability is the consequence of achieving equilibrium prices (and Bossaerts

Green (1989) [15]; Berk and Green (2004)[13]); 2) in behavioral finance the predictability is determined by the set of incorrect individual expectations, called cognitive biases, which have an impact on market prices (Bondt and Thaler (1985)[14]). In the context of the first explanation, a more general and flexible model to describe the dynamic of the future price was introduced in the literature.
In general, the mathematical structure of utility maximisation problems and the risk aversion of the market's participants essentially implies that an optimal trading strategy only exists if the discounted price process X_t is a semimartingale process[2]. This class differs with respect to martingales as they admit a predictable component inside.
From the theoretical point of view, this model is ideally suited to explain the pricing process on the base of a more flexible version of market efficiency, namely no-arbitrage hypothesis. A pricing model, based on this assumption, means to rule out any possibility to get a "positive expected gain, over the risk free return, without any downside risk" (see Föllmer, Hans, et al.(2013) [44]). A reach letterature studied the consequence of no-arbitrage assumption (Harrison and Kreps(1979)[50], Kreps (1981) [69], Harrison and Pliska (1981)[51] Cox and Huang (1991)[31] Delbaen and W. Schachermayer(1994) [33], Ansel and Stricker (1991)[8], Constantinos Kardaras and Eckhard Platen(2011) [64]). Particularly important is the paper of Delbaen and W. Schachermayer (1994)[33] where they demonstrated that there is an hide connection between the economic notion of no arbitrage (precisely the "No Free Lunch with Vanishing Risk" (NFLVR) condition) and the mathematical concept of semimartingale for describing the price dynamic. In this respect, relevant studies of semartingales in continuous-time processes are by Harrison and Kreps (1979)[50], Kreps (1981), Harrison and Pliska (1981, 1983)[51][52], Duffie and Huang (1985, 1986)[38][39], Duffie (1986, 1988)[37][95], Huang (1985, 1987)[57][58], Pliska (1986)[89], and Cox and Huang (1989, 1991)[30][31].)

3 Model Selection and Akaike Information Criteria

Model selection means to choose a model inside a group of candidate models $\mathcal{M}$ to describe a certain amount y having available a sample of data. In order to take a decision, it is necessary to define some criteria for evaluating what characteristics should the best model have among the ones to compare. The accuracy of the model to fit the data is quite relevant but not enough because this principle alone does not necessarily ensure the quality of the model. Sometimes models appear to be excellent in describing the data just because affected by overfitting problem, meaning that the model basically fits the data found in the sample, rather than representing the underlying pattern of the population. Thus, a principle of parsimony is frequently added to the selection process,

[2] A real-valued process X defined on a filtered probability space is called a semimartingale if it can be decomposed as a sum of a martingale M_t (unpredictable part) and a càdlàg adapted process of locally bounded variation A_t (predictable part)

$$X_t = M_t + A_t$$

meaning that, it is preferred, among a group of candidate model $\mathcal{M}$, the one having the lowest degree of complexity, for instance a lower number of parameters (Occam's razor).
To make a selection, a series of statistical tools, named information criteria, have been proposed from 70's. Frequently these information criteria are split into two types:
1) the first type of information criteria (e.g., AIC, AIC_c and TIC) selects the model on the basis of a distance called Kullback-Leibler information and does not assume the existence of a true model within the set of models to be compared;
2)the second type of information criteria assumes the existence of a "true" model of the data generator within the set of models to be compared.
Akaike was the first to formulate an information criterion based on Kullback Leibler divergence. The objective of this measure is to calculate the distance between each proposed model and the data generator and select the model having the shortest divergence. This means that the chosen model does not necessarily coincide with the one having generated the observed data. The particular feature of this measure is that it can make such a calculation despite the data generating process is not directly observable.
Akaike showed that the empirical Kullback Leibler divergence, that can be expressed in terms of the empirical log-likelihood function at its maximum point, is biased (see Kenneth P. Burnham, David R. Anderson (2002)[21]) and propose an approximated correction factor. This correction acts like an Occam's razor letting the criterion favour the models that are likely to make good predictions with the expected lowest overfitting problem.
Some of the most relevant papers treating AIC are:
Akaike (1973)[2][1] where it is proved that minimising the expected Kullback-Leibler divergence between the true density and estimated density of the model is approximately equivalent to minimising AIC;
Akaike (1974)[3] in this paper the AIC is proposed to be applied for model selection in particular for time series;
Akaike (1981)[4] where it is studied the Bayesian approach to model selection and describes the Akaike approach in Bayesian terms. He pointed out some differences, and in particular, Akaike finds a link with his method when priors and likelihood are expectations of the predictive distribution.
Other comparisons between AIC and Bayesian selection method are in Akaike (1983)[5] Akaike (1985)[6] Stone (1979)[103] Leamer (1979)[74] Atkinson (1981)[9] Chow (1981)[24] Shibata (1981)[97] Nishii (1984)[84] J Kuha(2004)[70].
A large number of the AIC generalisations have been proposed over the years. Some of the most famous are: The Corrected Akaike's Information Criterion (AIC_c) (Sugiura (1978)[104]; and Sakamoto et. al. (1986)[90]) , Takeuchi's information criterion (ICT) (Takeuchi 1976 [105]).
The goal of AIC_c is to improve the performance of the estimator proposed by Akaike when the number of parameters is too high compared to the sample size. Hurvich and Tsai(1989) [59] suggested a version suitable for small sample size.
The Takeuchi's information criterion (ICT) is a modified appropriate version of AIC when the candidate model cannot be considered a good approximation of the real data generating process.
A multitude of alternative methods, based on different theoretical assumptions, were proposed like: BIC(Schwarz 1978 [94], Hoeting et al. 1999[56]) (Spiegelhal-

ter, Best, Carlin, and Van der Linde 2002 [100], Van der Linde, 2005[78]), EIC (Ishiguro, Sakamoto, and Kitgawa 1997[61]), FIC (Wei 1992[108]), GIC (Nishii 1984 [83]), NIC (Murata, Yoshizawa, and Amari 1991 [7]) and TIC (Takeuchi 1976[105]) and many others.
Among these above criteria, it is worth mentioning in more detail one of them: the Bayesian Information Criterion (BIC) that was introduced by Schwarz (1978)[94] as an asymptotic approximation to a transformation of the Bayesian posterior probability of a proposed model. The assumptions made for BIC consist of an equal prior probability for each model and, a converging selection procedure to the right model, if this belongs to the set of candidates.

4 Preliminary Notions

In order to investigate the issue, we are going to provide in the following sections a:

1) definition of uncertain economic quantity;
2) definition of uncertain prediction of the economic quantity;
3) definition of prediction of the economic quantity;
4) theory that let us measure the uncertainty of the prediction;
5) theory that let us measure the relative uncertainty about the predictions of two alternative models considering their different complexity.

4.1 Definitions

I propose a definition of uncertain economic quantity, uncertain prediction of the economic quantity and prediction of an economic quantity, respectively as follows:

Uncertain Economic Quantity $\widetilde{y}_{re}$

Let's define an uncertain economic quantity (i.e. the future rate of return of an investment) $\widetilde{y}_{re}$ as a continuous real valued random variable with a non-parametric density function $f_{re}(y_{re})$. An observation of this variable is called y_{re}.

Uncertain Prediction of the Economic Quantity $\widetilde{y}_{th}$

Let's define the uncertain prediction of an economic quantity (theoretical model) $\widetilde{y}_{th}$ (i.e. the solution of a stochastic differential equation at a given future time) as a continuous real valued random variable with a parametric density function $f_{y_{th}}(y_{th}|\theta)$, and a d-dimensional vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$.

Prediction of the Economic Quantity y_{th}

The prediction of the economic quantity is a real valued deterministic variable we define as a chosen outcome among the possibles provided by the uncertain prediction of the economic quantity
We define the prediction to be certain if it has been known that, the theoretical outcome y_{th}, is equal to the observed outcome y_{re}

$$y_{th} = y_{re}$$

before the knowledge of the outcome y_{re} is available .

Consequently we can define the uncertainty about the prediction as the uncertainty about the equality between values y_{th} and y_{re}.
In order to measure this uncertainty, we will refer to the information theory.

5 Measures of Information and Uncertainty

How can we measure information and uncertainty? Both concepts are related to the concept of doubt. Information arises when doubt disappears and vice versa (Kuhlthau 1993 [71] and Leroy, Singell 1987 [76]).
Therefore, given a random variable,

the greater is the number of possible outcomes, the more we are uncertain about the outcome a priori.

Intuitively, if it is given a prior knowledge over the source of the outcome, it must change the degree of uncertainty over the possible outcomes.
Shannon (1948) [96] treated the probability distribution as a kind of knowledge about the source of the outcome. He was able, just using some axioms, to quantify the expected uncertainty about observing an outcome, or equivalently, the amount of uncertainty represented by a probability distribution (Jaynes (1957) [63]). In the context of communication theory, it amounts to the minimal number of bits that should be transmitted to specify the outcome.
In order to exploite information theory in the analysis we need further definitions based on Brissaud (2005) page 70. [19], Hartley(1928) [53]and Shannon (1948) [96].
In order to explain next definitions let's consider the following hypotheses: suppose a discrete random variable has "N" possible outcomes; let's call "true outcome" one of them; suppose a decision maker has to decide what is the true outcome among the possible, knowing the N possible outcomes but unknowing what is the true outcome among them.
Then we can define:

Information I

The information I is defined to be a prior knowledge of the set of possible outcomes. It coincides to the knowledge that the decision-maker has got to properly identify the right outcome correctly. Given a probability space $(\Gamma, \mathscr{A}, \mathbb{P})$ consider $A \in \mathscr{A}$ to be an outcome which can happen randomly with probability $\mathbb{P}(A) \equiv p$, therefore the information due to the outcome A is defined as:

$$I = log(p) \tag{1}$$

Notice that, the information is always a non-positive number and it is increasing in p. It can be interpreted as the lack of uncertainty (Hartley(1928) [53])

Uncertainty U

We define the uncertainty[3] as the Hartley measure (1928) [53] of the outcome A. It is a measure, related to the sigma-algebra $\mathscr{A}$, of the number of possible alternatives when the chosen outcome is A.

$$U = -log(p) \tag{2}$$

Shannon Entropy H

In the literature (see Kolmogorov 1965 [68] Klir 2005 [67], Casquilho 2014 [23]) it is widely recognized that, starting from several well-justified axiomatic characterizations, Shannon entropy (Shannon (1948) [96]) is the only meaningful functional for measuring the expected uncertainty (or expected lack of information) in probability theory.
Let $\widetilde{x}$ be a *discrete* random variable with probability p.
The entropy is defined as the expected uncertainty over all the possible outcomes x.

$$H(\widetilde{x}) = -\sum_{x} p(x)log(p(x)) \tag{3}$$

(see Kenneth P. Burnham, David R. Anderson (2002)[21])
Shannon entropy quantifies the average uncertainty about the true outcome.

Differential Entropy H_{diff}

The differential entropy of a real valued *continuous* variable $\widetilde{x}$ is defined (see Cover, Thomas M (1991) [29], Kenneth P. Burnham, David R. Anderson (2002)[21])) as follows:

$$H(\widetilde{x})_{diff} = -\int_{-\infty}^{\infty} f(x) \cdot log(f(x))dx$$

where $f(x)$ is the probability density function of $\widetilde{x}$.
Notice that:

1) the differential entropy is not a measure of the average amount of information contained in a continuous r.v $\widetilde{x}$;
2) a continuous random variable contains an infinite amount of information,

[3]The Hartley measure is $H(A) = log(|A|) = log(\frac{1}{\frac{1}{|A|}}) = log(1/p) = -log(p)$ see Klir(2005) pag 27 [67].

because the entropy of the continuous variable $\widetilde{x}$ is infinite[4] or approximately equal to

$$H(\widetilde{x}) \approx -lim_{\Delta \to o} log(\Delta) - \int_{-\infty}^{\infty} f(x) \cdot log(f(x)) dx =$$
$$= -lim_{\Delta \to o} log(\Delta) - H_{diff}$$

which diverges to infinity almost always.
However

*the difference between two differential entropies can be used as an indicator for comparing the uncertainty of two continuous r.v. quantized to the same precision.*The mutual information is a typical example.

Mutual information MI

Given two random variables: $\widetilde{x}$, $\widetilde{y}$, the mutual information index $MI(\widetilde{x}, \widetilde{y})$ is defined as the reduction of expected uncertainty in $\widetilde{x}$ by knowing $\widetilde{y}$

$$MI(\widetilde{x}, \widetilde{y}) = H(\widetilde{x}) - H(\widetilde{x}|\widetilde{y})$$

(see Cover, Thomas M (1991) [29] and Kenneth P. Burnham, David R. Anderson (2002)[21])).

6 Degree of Relative Predictability

6.1 Residual Entropy related to model M

Given a economic model called **model M** *the residual entropy associated with the model M is defined as the mutual information between the uncertain economic variable* $\widetilde{y}_{re}$ *and the uncertain prediction provided by the model,* $\widetilde{y}_{th}^{M}$. *Therefore, this index can be expressed as the measure of the reduction of the expected uncertainty about the future value of the economic variable* $\widetilde{y}_{re}$ *once has became available the uncertain prediction* $\widetilde{y}_{th}^{M}$ *provided by the model M.*

$$H_{res}^{M}(\widetilde{y}_{re}, \widetilde{y}_{th}^{M}) = H(\widetilde{y}_{re}) - H(\widetilde{y}_{re}|\widetilde{y}_{th}^{M}) \tag{4}$$

The reduction, furthermore, is explainable in terms of Kullback-Leibler divergence as well (see Kenneth P. Burnham, David R. Anderson (2002)[21])

$$H_{res}^{M}(\widetilde{y}_{re}, \widetilde{y}_{th}^{M}) = \mathbb{E}\left[log\left(\frac{p(\widetilde{y}_{re}, \widetilde{y}_{th}^{M})}{p(\widetilde{y}_{re}) \cdot p(\widetilde{y}_{th}^{M})}\right)\right]$$

where:

- $p(\widetilde{y}_{re}, \widetilde{y}_{th}^{M})$ is the joint probability distribution function of the continuous random variables $\widetilde{y}_{re}$and $\widetilde{y}_{th}^{M}$;
- $p(\widetilde{y}_{re})$ is the probability distribution function of the continuous random variable $\widetilde{y}_{re}$;

[4]It is equal to zero if $H_{diff} \to -\infty$

- $p(\widetilde{y}_{th})$ is the probability distribution function of the continuous random variable $\widetilde{y}_{th}$.

Let's suppose that two alternative [5] stochastic models are available: **model A** and **model B**

We want to propose a measure that let us compare the degree of predictability of the economic variable $\widetilde{y}_{re}$ provided by the two models.

For this purpose we intend to proceed as follows: 1) we need to measure, for each model, the reduction of uncertainty about the future outcome $\widetilde{y}_{re}$, once the prediction of the model is provided; 2) we need to find a way to compare the two measurements. We will obviously prefer the model which guarantees the highest reduction.
Therefore, just using the definition of residual entropy provided in the last paragraph, we are now allowed to propose a *measure of the degree of relative predictability as the difference of residual entropies calculated over alternative models A and B* [6]

$$H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th}) = H^A_{res}(\widetilde{y}_{re}, \widetilde{y}^A_{th}) - H^B_{res}(\widetilde{y}_{re}, \widetilde{y}^B_{th}) \tag{5}$$

if:

$H_{A,B} > 0 \Rightarrow$ Model A is expected to be more uncertain with respect to model B

$H_{A,B} < 0 \Rightarrow$ Model B is expected to be more uncertain with respect to model A

7 Developing a measure for the Degree of Relative Predictability

We need to give a specific measure of the degree of relative predictability $H_{A,B}$ for the arbitrary models A and B, given our a priori knowledge (see (12)). For now, we assume that all the density functions are completely known. We will relax this assumption in chapter 8.
Under some conditions, explained in the following subsection, it is possible to write the formula of the degree of relative predictability as a function of an assigned $\bar{y}_{re}$ given the values of the pointwise predictions for both models $\bar{y}^A_{th}$, $\bar{y}^B_{th}$, the parametric densities $f_{y^A_{th}}, f_{y^B_{th}}$, with their vectors of parameter θ^A and θ^B, and the density $f_{y_{re}}$

$$H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th}) = H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})(\bar{y}_{re}|\bar{y}^A_{th}, \bar{y}^B_{th}, \theta^A, \theta^B, f_{y_{re}}) \tag{6}$$

later it will be given the explicit formula (16).

[5] In this sentence each "alternative" is meant to be any model built to provide a prediction for the same economic variable $\widetilde{y}_{re}$

[6] Each model provides a specific theoretical model solution y^A_{th} and y^B_{th}

The aim of this formula is to show the dependence of the relative predictability, with respect to all the possible values that the random variable $\bar{y}_{re} \in \bar{Y}_{re} \subseteq \mathbb{R}$ can assume with its probability of occurrence.
First of all, let's see some needed conditions for the calculation.

7.1 Indeterminacy, uncertainty and wrong predictions

Let's define $\{\widetilde{y}_{re}(t) : t \in T\}$ and $\{\widetilde{y}^M_{th}(t) : t \in T\}$ to be two real valued stochastic processes defining respectively an economic quantity and it's stochastic model (eg. the log return based on the price of a commodity). Let's define *0* to be the present time and s a future time such that the economic variable $\widetilde{y}_{re}$ and the theoretical model solution $\widetilde{y}^M_{th}$, as defined before, are two random variables belonging to those stochastic processes observable at the future time s, $\widetilde{y}_{re}(s) = \widetilde{y}_{re}$ and $\widetilde{y}^M_{th}(s) = \widetilde{y}^M_{th}$.
Let's suppose that these two stochastic processes, $\widetilde{y}_{re}(t)$ and $\widetilde{y}^M_{th}(t)$, are semimartingales, meaning that, according to Doob's decomposition theorem, they both can be respectively decomposed in a sum of two components: a càdlàg adapted process of locally bounded variation and a local martingale.
Therefore, in the context of our analysis, $\widetilde{y}_{re}$ and $\widetilde{y}^M_{th}$ can be interpreted as follows:
1) the nondeterministic part of $\widetilde{y}_{re}$ describes the inherent unpredictable part that characterized the future economic quantity to predict. The deterministic part $\bar{y}_{re}$ is the only predictable part of the economic variable, however *it cannot be directly observed and it cannot be separated from the unpredictable part* $\widetilde{e}_{re}$;
2) instead $\widetilde{y}^M_{th}$ is affected by a random prediction error, whose range and frequency are due to the ability of the economist to predict the economic quantity well enough, whereas the deterministic part is the pointwise prediction provided by the model M

$$\widetilde{y}_{re} = \underbrace{\bar{y}_{re}}_{predictable} + \underbrace{\widetilde{e}_{re}}_{unpredictable} \tag{7}$$

$$\widetilde{y}^M_{th} = \underbrace{\bar{y}^M_{th}}_{prediction} + \underbrace{\widetilde{e}^M_{th}}_{prediction\ error} \tag{8}$$

where $\widetilde{e}^M_{th}$ refers to risk in modeling correctly the true variable and $\bar{y}^M_{th}$ is the pointwise prediction provided by the theoretical model.
Moreover on the base of our definitions, we assume that $\widetilde{e}_{re}$ and $\widetilde{e}^M_{th}$ are independent variables.
Thus, on the base of these assumptions, the relative predictability of prediction models must be linked both to the indeterminacy of the economic variable and to the uncertainty of the theoretical model.
The theoretical model may differ from the economic variable for four nonexclusive reasons:

1st type) $\bar{y}^M_{th} \neq \bar{y}_{re}$

2nd type) $e^M_{th} \neq e_{re}$

3th type) $\bar{y}^M_{th} \neq e_{re}$

4th type) $e_{th}^{M} \neq \bar{y}_{re}$

where e_{re} and e_{th}^{M} are realizations respectively of $\widetilde{e}_{re}$ and $\widetilde{e}_{th}^{M}$.
However, *not all the possible causes of wrong predictions should reasonably affect the degree of uncertainty of predictions.*
The errors of the 3rd and 4th type are desirable and unavoidable.
The error of the 1st type is a sign of the inability of the theoretical model to estimate the predictable component (expectation) $\bar{y}_{re}$ correctly. The reason for this defeat is due to modelling assumptions (economic or statistical) which are unsuitable to explain the effective expectation (wrong functional form, incorrect explanatory variables introduced, lack of explanatory variables,...).
The error of the 2nd type is exclusively due to the independence of the unpredictable part of the economic variable and the risk in modelling the true variable correctly.
Given the definition of residual entropy, we need that

a) both unpredictable components e_{re} and e_{th}^{M} must affect the degree of predictability

however,

b) the 2nd type error must not affect the uncertainty quantification because meaningless.

Let's explain point b) with an example: suppose that the economic variable and the theoretical model share the same expectation and the same error probability density function.

$$\bar{y}_{th}^{M} = \bar{y}_{re}$$

$$f_{e_{th}^{M}} = f_{e_{re}}$$

We should consider this situation, independently of the value assumed by the error components, as the best possible result in a modelling sense. Under that perspective, we can define the two models as identical in spite of the solutions are different[7]. It doesn't make any sense to consider the two models as different in this case, because the prediction error is only due to "unpredictable" components, while we are supposed to be only interested in the prediction of the predictable components.

In order to satisfy contemporaneously point a) and point b) we have to calculate the entropy of the prediction conditioning the probabilities of $\widetilde{y}_{re}$ and $\widetilde{y}_{th}^{M}$, such that

$$e_{th}^{M} = e_{re} \tag{9}$$

Basically, it is proposed to study the degree of relative predictability by calculating for each model M the residual entropy of a prevision H_{res}^{M} imposing the condition (9).

[7]Note that we have set, the uncertain prediction of the economic quantity (solution of an economic model) as always given by a sum of a deterministic forecast and a random error component which must be imagined as coming from a data generating number which is produced randomly from a given distribution.

7.2 Conditions for Calculating the Residual Entropy of a Prediction

In order to compute the residual entropy of a prediction we need to make some assumptions about the knowledge available at the evaluation moment regarding the probability distribution functions of the variables $\widetilde{y}_{re}$ and $\widetilde{y}_{th}^{M}$:
1) for each model M we impose a parametric distribution of $\widetilde{y}_{th}^{M}$ with a vector of parameters θ. We justify this choice by the fact that theoretical models very often impose a parametric distribution to the model; 2)we impose a non-parametric distribution to $\widetilde{y}_{re}$; 3) we suppose either $\bar{y}_{re}$ or $\bar{y}_{th}^{M}$ to be known, so that we don't assign any uncertainty to those variables; 4) by definitions of $\widetilde{e}_{re}$ and $\widetilde{e}_{th}^{M}$ we consider $\widetilde{y}_{re}$ and $\widetilde{y}_{th}^{M}$ to be independent continuous random variables. Moreover, for the sake of simplicity we suppose they are equivalent (they share the same support).
All of the previous assumptions are going to be summarised, from here on out, by the information set Ψ^{M}

$$\Psi^{M} = \{\bar{y}_{re}, \bar{y}_{th}^{M}, \theta, f_{y_{re}}\} \tag{10}$$

However, as we already seen, to properly isolate the uncertainty of the predictions, we need to impose the condition (9) as well. For this purpose, it is defined a new random variable $\widetilde{z}$ and a new information set Ω^{M}

$$\widetilde{z} = \widetilde{e}_{th^{M}} - \widetilde{e}_{re} \tag{11}$$

the null outcome z of the random variable $\widetilde{z}$ is added to the previous information set determining Ω^{M}

$$\Omega^{M} = \{z = 0, \bar{y}_{re}, \bar{y}_{th}^{M}, \theta, f_{y_{re}}\} \tag{12}$$

where $\Theta^{M} \subseteq \mathbb{R}^{d_M}$ is the d_M-dimensional parameter space of the density function of the unpredictable part of the theoretical model M (that is $f_{e_{th}^{M}}$) and $\theta \in \Theta^{M}$ is a given point in that space. The presence of $f_{y_{re}}$ inside the information sets means that we are going to consider the density distribution of y_{re} to be a non-parametric (for now) "known" density function. We estimate the vector of parameters θ and the density y_{re} in the next sections.

7.3 Residual Entropy of a Prediction $H_{res}^{M}(\widetilde{y}_{re}, \widetilde{y}_{th}^{M})$

Now, we are allowed to calculate the residual entropy, as defined in 6.1, for a generic model M (it means it will be valid for any model included the generic couple of models A and B) conditioning both densities to the information set Ω^{M}

$$H_{res}^{M}(\widetilde{y}_{re}, \widetilde{y}_{th}^{M}) = \mathbb{E}_{\widetilde{y}_{re}, \widetilde{y}_{th}^{M}|\Omega^{M}}\left[log\left(\frac{p_{\widetilde{y}_{re}, \widetilde{y}_{th}^{M}|\Omega^{M}}(\widetilde{y}_{re}, \widetilde{y}_{th}^{M}|\Omega^{M})}{p_{\widetilde{y}_{re}|\Omega^{M}}(\widetilde{y}_{re}|\Omega^{M}) p_{\widetilde{y}_{th}^{M}|\Omega^{M}}(\widetilde{y}_{th}^{M}|\Omega^{M})}\right)\right] \tag{13}$$

Where:

- $p_{\widetilde{y}_{re},\widetilde{y}_{th}^M|\Omega^M}$ is the joint probability distribution of $\widetilde{y}_{re}$ and $\widetilde{y}_{th}^M$ given the information set Ω^M;
- $p_{\widetilde{y}_{re}|\Omega^M}$ is the probability distribution of $\widetilde{y}_{re}$ given the information set Ω^M;
- $p_{\widetilde{y}_{th}^M|\Omega^M}$ is the probability distribution of $\widetilde{y}_{th}^M$ given the information set Ω^M.

After some manipulation (see appendix "*calculation of* $H_{res}(\widetilde{y}_{re},\widetilde{y}_{th})$"13) we finally get

$$H_{res}^M(\widetilde{y}_{re},\widetilde{y}_{th}^M) = -\,\mathrm{E}_{y_{th}^M|\Omega^M}\left[log\left(f_{y_{th}^M|\Omega^M}(\widetilde{y}_{th}^M|\Omega^M)\right)\right] - lim_{\Delta\to 0} log(\Delta) \tag{14}$$

where:

$$f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M) = \frac{f_{e_{re}}(y_{th}^M - \bar{y}_{th}^M|\Psi^M)\cdot f_{e_{th}^M}(y_{th}^M - \bar{y}_{th}^M|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^M}(e_{th}^M|\Psi^M)\cdot f_{e_{re}}(e_{th}^M|\Psi^M)de_{th}^M}$$

and

- $f_{e_{re}}(\cdot|\Psi^M) \equiv f_{e_{re}}(\cdot)$ is the conditional probability density function of the risk related to the economic variable;
- $f_{e_{th}}(\cdot|\Psi^M)$ is the conditional probability density function of the error component of the model.

Therefore, $H_{res}^M(\widetilde{y}_{re},\widetilde{y}_{th}^M)$ is just reduced to the entropy of a new random variable $y_{th}^M|\Omega^M$ with a semiparametric density function $f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M)$ defined in appendix (see formula (37)).
It has been also demonstrated (see appendix "*Consequences on the calculation of* $H_{res}(\widetilde{y}_{re},\widetilde{y}_{th})$"13) under the information set Ω, that the theoretical model $\widetilde{y}_{th}$ behaves like a function of the economic variable $\widetilde{y}_{re}$, meaning that we are allowed to simplify the previous formula (14) as follows:

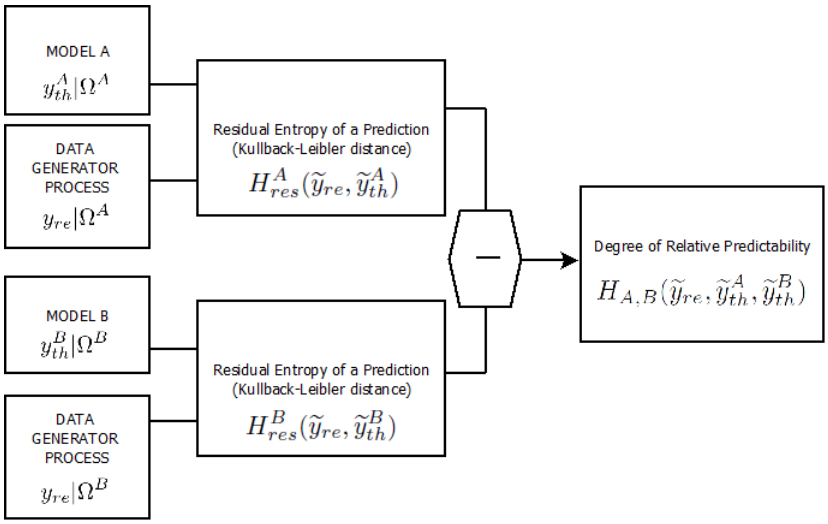
$$H_{res}^M(\widetilde{y}_{re},\widetilde{y}_{th}^M) = -\,\mathbb{E}_{y_{re}}\left[log\left(f_{y_{re}|\Omega^M}(\widetilde{y}_{re}|\Omega^M)\right)\right] - lim_{\Delta\to 0} log(\Delta) \tag{15}$$

where

$$f_{y_{re}|\Omega^M}(y_{re}|\Omega^M) = \frac{f_{y_{re}}(y_{re})\cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{y_{re}}(y_{re})\cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)dy_{re}}$$

and

- $f_{y_{re}}(\cdot)$ is the true density of the economic variable;
- $f_{e_{th}}(\cdot|\Psi^M)$ is the conditional probability density function of the error component of the model.



7.4 Degree of Relative Predictability $H_{A,B}(\widetilde{y}_{re}, \widetilde{y}_{th}^{A}, \widetilde{y}_{th}^{B})$

Recovering formula (5) we have defined the degree of relative predictability as the difference between two residual entropies $H_{res}^{A}(\widetilde{y}_{re}, \widetilde{y}_{th}^{A})$ and $H_{res}^{B}(\widetilde{y}_{re}, \widetilde{y}_{th}^{B})$. Therefore, applying formula (15) to formula (5), we finally get

$$H_{A,B}(\widetilde{y}_{re}, \widetilde{y}_{th}^{A}, \widetilde{y}_{th}^{B}) = \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) log \left(\frac{f_{y_{re}|\Omega^{B}}(y_{re}|\Omega^{B})}{f_{y_{re}|\Omega^{A}}(y_{re}|\Omega^{A})} \right) dy_{re} \tag{16}$$

where

$$\Omega^{A} = \{z = 0, \bar{y}_{re}, \bar{y}_{th}^{A}, \theta^{A}, f_{y_{re}}\}$$

$$\Omega^{B} = \{z = 0, \bar{y}_{re}, \bar{y}_{th}^{B}, \theta^{B}, f_{y_{re}}\}$$

8 Building the Estimator of $H_{A,B}$

Standard statistical practice ignores to model uncertainty. Data analysts typically select a model from a class of models and then proceed as if the chosen model had generated the data. This approach ignores the uncertainty in model selection leading to overconfident inferences and decisions that are riskier than expected. In a certain sense, $H_{A,B}$ can overcome that problem because: 1) it selects the model that comes closest to the true data generating process; 2) it compares the ability of the models to make predictions.
$H_{A,B}$ is defined in (6) as a function of $\bar{y}_{re}$ given $\bar{y}_{th}^{A}, \bar{y}_{th}^{B}, f_{e_{th}^{A}}, f_{e_{th}^{B}}, f_{y_{re}}$, however, despite the analysis can be carried out ex-post the resolution of uncertainty, it requires to estimate some quantities already included in the information set.
By definition (8) both economic models A and B provide a prevision so that $\bar{y}_{th}^{A}$, $\bar{y}_{th}^{B}$ are *known* while probably all the densities cannot be.
We assume, from now on, that the density $f_{e_{th}^{A}}$ is defined on the parameter space Θ_A, $f_{e_{th}^{B}}$ is defined on the parameter space Θ_B while the density $f_{y_{re}}$ is defined on a non-parametric space $\mathscr{F}$.

We justify this choice by the fact that the density of the theoretical error component, of an economic model M, is very often parametric while $f_{y_{re}}$ can be only guessed on the base of an observation sample, meaning that we'll prefer it to be freely detected without specifying in advance a class of parametric models.
We notice, by the definition (6), that the quantity $H_{A,B}$ depends on θ^A and θ^B. Consequently, it is possible that some parameters can increase the relative uncertainty of predictions more than others. However, if θ^A and θ^B can be regarded as random variables, this quantity may be averaged over $\Theta \equiv \Theta^A \times \Theta^B$ according to a probability density function f_{θ^A,θ^B} (see Lindley (1956)[79]).
I define the degree of relative predictability as a function ϕ of the vectors of parameters θ^A, θ^B and the density $f_{y_{re}}$.

$$H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th}) = \phi_{A,B}(\theta^A, \theta^B, f_{y_{re}})$$

In order to build an estimator for $\phi_{A,B}$ we consider the set of models

$$\mathcal{P} = \left\{ \mathbb{P}_{\theta^A \times \theta^B \times f_{y_{re}}} \;\middle|\; \theta^A \in \Theta^A, \theta^B \in \Theta^B, f_{y_{re}} \in \mathscr{F} \right\}$$

[8]

and we define $\phi_{A,B}$ to be a characteristic of a particular member $\mathbb{P}_{\theta^A \times \theta^B \times f_{y_{re}}}$ that can be written as a mapping from the space $\Theta^A \times \Theta^B \times \mathscr{F}$ to $\mathbb{R}$

$$\phi_{A,B} : \Theta^A \times \Theta^B \times \mathscr{F} \mapsto \mathbb{R}$$

where:

- $\Theta^A \subseteq \mathbb{R}^{d_A}$ is the d_A-dimensional parameter space of $f_{e^A_{th}}$
- $\Theta^B \subseteq \mathbb{R}^{d_B}$ is the $d_{.B}$-dimensional parameter space of $f_{e^B_{th}}$
- $\mathscr{F}$ is an infinite-dimensional space

Therefore we are going to treat the vectors of parameters $\hat{\theta}^A$ and $\hat{\theta}^B$ as random variables, and according to Lindley (1956)[79] we define the expected degree of relative predictability as follows:

$$\mathbb{E}_\Theta[H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})] = \mathbb{E}_\Theta[\phi_{A,B}(\hat{\theta}^A, \hat{\theta}^B, f_{y_{re}})]$$

$\mathbb{E}_\Theta[\phi_{A,B}()]$ results to be again a characteristic that can be written as a mapping from the non-parametric space $\mathscr{F}$ to $\mathbb{R}$.

$$\mathbb{E}_\Theta[\phi_{A,B}()] : \mathscr{F} \mapsto \mathbb{R}$$

Given a random sample $\left\{\widetilde{y}^i_{re}\right\}_{i=1,\dots,n}$, a possible choice for constructing the estimator is the plug-in estimator of $\mathbb{E}_\Theta[\phi_{A,B}]$

$$\mathbb{E}_{\Theta_n}[H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})] = \mathbb{E}_{\Theta_n}[\phi_{A,B}(\hat{\theta}^A_n, \hat{\theta}^B_n, \hat{f}^n_{y_{re}})] \tag{17}$$

where

- $\mathbb{E}_{\Theta_n}$ is the expectation operator over the estimators $\hat{\theta}^A_n$ and $\hat{\theta}^B_n$;
- $\hat{\theta}^A_n$ and $\hat{\theta}^B_n$ are the vector parameters estimator respectively for model A and B;
- $\hat{f}^n_{y_{re}}$ is any possible consistent estimator of the unconditional density function $f_{y_{re}}$ plugged into the mapping $\phi_{A,B}$.

[8] for simplicity we avoid to write the other known parameters $\bar{y}^A_{th}$, $\bar{y}^B_{th}$

8.1 A Proposed Estimator and its Asymptotic Properties

Given two vectors of estimated parameters for the two models: θ_o^A and θ_o^B, I'm going to propose an estimator

$$\mathbb{E}_{\Theta_n}\left[\phi_{A,B}(\hat{\theta}_n^A, \hat{\theta}_n^B, \hat{f}_{y_{re}}^n)\right]$$

such that the **weak consistency** of the estimator is verified as follows:

$$\text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_n}\left[\phi_{A,B}(\hat{\theta}_n^A, \hat{\theta}_n^B, \hat{f}_{y_{re}}^n)\right] = \mathbb{E}_{\Theta}\left[\phi_{A,B}(\theta_0^A, \theta_0^B, f_{y_{re}})\right] \tag{18}$$

where: the point θ_0^A is in the interior of Θ^A; $\theta_0^{\mathbf{B}}$ is in the interior of Θ^B and $f_{y_{re}}$ is the true unconditional density function of y_{re} for now on.
For the sake of simplicity, we need to change our notation to easily show the consistency of the estimator to suggest. For the generic model M:

- $L(\hat{\theta}_n)$ **the expected value of the log likelihood function with parameters estimator.**

  We denote by $L(\hat{\theta}_n)$ the expected value of the log likelihood function with a parameter estimator $\hat{\theta}_n$(see appendix (42) and (41)),

  $$L(\hat{\theta}_n) = \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) log\left(f_{y_{re}|\Omega}(y_{re}|\Omega^M)\right) dy_{re}$$

  with

  $$\Omega^M = \{z = 0, \bar{y}_{re}, \bar{y}_{th}^M, \hat{\theta}_n, f_{y_{re}}\}$$

- $L(\theta_0)$ **the expected value of the log likelihood function with parameters estimate.**

  We denote by $L(\hat{\theta_0})$ the expected value of the log likelihood function with the parameter estimate θ_0(see appendix (42) and (41)),

  $$L(\theta_0) = \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) log\left(f_{y_{re}|\Omega}(y_{re}|\Omega^M)\right) dy_{re}$$

  with

  $$\Omega^M = \{z = 0, \bar{y}_{re}, \bar{y}_{th}^M, \theta_0, f_{y_{re}}\}$$

- $L_n(\hat{\theta}_n)$ **a sample expectation value of the log likelihood function with parameters estimator.**

  One of the drawbacks to solve is to make inference over a probability density function $f_{y_{re}|\Omega^M}(y_{re}|\Omega^M)$ which is artificially constructed, through the use of two parametric probability densities and a non-parametric probability density. Consequently, no data generated from this probability are directly available and

this condition makes it difficult to define a consistent sample expectation of the likelihood function.

Let $\{\widetilde{y}^i_{re}\}_{i=1,\dots,n}$ and $\{\widetilde{y}^j_{re}\}_{j=1,\dots,n}$ be two random samples of i.i.d. random variables drown from the same density distribution $f_{y_{re}}(y_{re})$ (or weakly dependent random variables. See proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]), therefore I propose to employ the following sample expectation formula of the likelihood function:

$$L_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{\frac{1}{2nh}\sum_{j=1}^{n}\mathbf{1}\Big(|\frac{\widetilde{y}^j_{re}-\widetilde{y}^i_{re}}{h}|\leq 1\Big) f_{e_{th}}(\widetilde{y}^i_{re}-\bar{y}_{re}|\Psi^M)}{\frac{1}{n}\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}^i_{re}-\bar{y}_{re}|\Psi^M)}\right) \tag{19}$$

with

$$\Psi^M = \{\bar{y}_{re}, \bar{y}^M_{th}, \hat{\theta}_n, \hat{f}^n_{y_{re}}\}$$

where the plug-in estimator, of the density function of the economic variable, is an uniform kernel density function $\hat{f}^n_{y_{re}} = \frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}\Big(|\frac{\widetilde{y}^i_{re}-y_{re}}{h}|\leq 1\Big)$ (see appendix (54) for further details).
Notice that, the so constructed formula has three sources of uncertainty arising from the presence of two random samples and of the vector of random parameters.

- $L_n(\theta_0)$ **a sample expectation value of the log likelihood function with parameters estimate.**

  The next log likelihood function differs from the previous only because of the presence of the estimate θ_0 in substitution of the estimator $\hat{\theta}_n$,

$$L_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{\frac{1}{2nh}\sum_{j=1}^{n}\mathbf{1}\Big(|\frac{\widetilde{y}^j_{re}-\widetilde{y}^i_{re}}{h}|\leq 1\Big) f_{e_{th}}(\widetilde{y}^i_{re}-\bar{y}_{re}|\Psi^M)}{\frac{1}{n}\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}^i_{re}-\bar{y}_{re}|\Psi^M)}\right) \tag{20}$$

  with

$$\Psi^M = \{\bar{y}_{re}, \bar{y}^M_{th}, \theta_0, \hat{f}^n_{y_{re}}\}$$

moreover denoting

$$\mathbb{E}_{\Theta_n}[\phi_{A,B}(\hat{\theta}^A_n, \hat{\theta}^B_n, \hat{f}^n_{y_{re}})] = \mathbb{E}_{\Theta_n}[L_n(\hat{\theta}^B_n) - L_n(\hat{\theta}^A_n)]$$

and

$$\mathbb{E}_{\Theta}\left[\phi_{A,B}(\theta^A_0, \theta^B_0, f_{y_{re}})\right] = L(\theta^B_0) - L(\theta^A_0)$$

we can simplify the weak consistency condition (18) with the new notation.

$$\text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n^B) - L_n(\hat{\theta}_n^A)\right] - \left(L(\theta_0^B) - L(\theta_0^A)\right) = 0$$

$$\text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n^B)\right] - L(\theta_0^B) - \text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n^A)\right] - L(\theta_0^A) = 0$$

In order to ensure that the plug-in estimator (17) is at least weakly consistent it is sufficient that, for every model M, the estimator $\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n)\right]$ is at least weakly convergent to $L(\theta_0)$.

$$\text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n)\right] = L(\theta_0) \tag{21}$$

This convergence is ensured by generalized Slutsky theorem (see Demidenko (2013)[34]), given the estimator $L_n(\hat{\theta}_n)$ as defined in (19), and the asymptotic density for the vector of parameters as defined in (41).

8.2 Methodology

Formula 21 and 17 show that the expected degree of relative predictability converges in probability to a simple difference as follows:

$$\text{plim}_{n\to\infty}\,\mathbb{E}_{\Theta_\text{n}}[H_{A,B}(\widetilde{y}_{re}, \widetilde{y}_{th}^A, \widetilde{y}_{th}^B)] = L(\theta_0^B) - L(\theta_0^A)$$

this result can be potentially exploited for building a new information criterion however, in order to calculate $\mathbb{E}_{\Theta_\text{n}}[H_{A,B}(\widetilde{y}_{re}, \widetilde{y}_{th}^A, \widetilde{y}_{th}^B)]$, we should solve an integral and this is usually time-consuming. For this reason, in chapter 8.3, I replace the proposed estimator with an approximation, that converges to the same value $L(\theta_0^B) - L(\theta_0^A)$, that I'm going to call Predictability Information Criterion (PIC).
In order to derive the new estimator, we will proceed by steps:

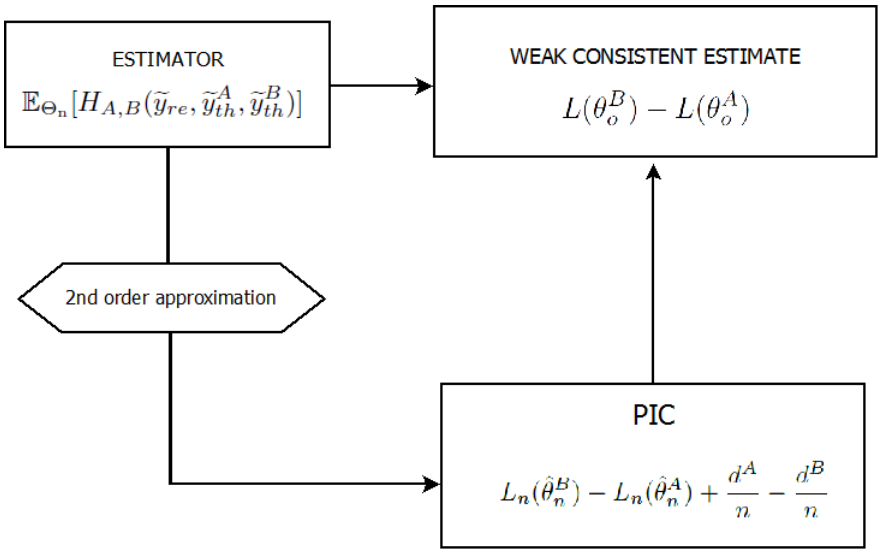
a. we compute a 2nd order functional expansion respectively of the above-mentioned $L_n(\hat{\theta}_n)$ and $L(\hat{\theta}_n)$;

b. we compute the expectation $\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right]$;

c. studying the asymptotic behavior of $\mathbb{E}_{\Theta_\text{n}}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right]$, we get an alternative estimator (PIC) converging in probability to the same difference $L(\theta_0^B) - L(\theta_0^A)$.

8.3 The Predictability Information Criterion (PIC) with Maximum Likelihood Estimators (MLE)

We compute the second order expansions of $L_n(\hat{\theta}_n)$ and $L(\hat{\theta}_n)$ centered around θ_o. We then study the asymptotic properties of the two expansions.



$L_n(\hat{\theta}_n) = L_n(\theta_o) - \left(\hat{\theta}_n - \theta_o\right)' D_\theta \{L_n(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' D^2_\theta \{L_n(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + O_p(n^{-2})$
$L(\hat{\theta}_n) = L(\theta_o) - \left(\hat{\theta}_n - \theta_o\right)' D_\theta \{L(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' D^2_\theta \{L(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + O_p(n^{-3/2})$
applying formula (59), (47) and (46)(see the appendix for more details) we get the following results:

$$L_n(\hat{\theta}_n) = L_n(\theta_o) - \left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' D^2_\theta \{L_n(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + o_p(1) \tag{22}$$

$$L(\hat{\theta}_n) = L(\theta_o) - \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right) + O_p(n^{-3/2})$$

We bring together the formulas and we ensemble a difference
$L_n(\hat{\theta}_n) - L(\hat{\theta}_n) = [L_n(\theta_o) - L(\theta_o)] - \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' D^2_\theta \{L_n(\theta_o)\} \left(\hat{\theta}_n - \theta_o\right) + o_p(1)$
using (61) we simplify as follows

$$L_n(\hat{\theta}_n) - L(\hat{\theta}_n) = L_n(\theta_o) - L(\theta_o) - \left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right) + o_p(1)$$

Now we calculate the expectation in $\hat{\theta}_n$ (see appendix (41) for further details)

$$\mathbb{E}_{\Theta_n}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right] = L_n(\theta_o) - L(\theta_o) - \mathbb{E}_{\Theta_n}\left[\left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right)\right] + o_p(1)$$

$$= L_n(\theta_o) - L(\theta_o) - \frac{1}{n}\left[Trace\left(I(\theta_o) \cdot \Sigma\right) + 0\right] + o_p(1)$$

If the estimator $\hat{\theta}_n$ achieves the Cramér–Rao lower bound (CRLB) at θ_o then Σ is equal to the inverse of the Fisher's Information Matrix $\Sigma = [I(\theta_o)]^{-1}$, where $I(\theta_o)$ is the Fisher's matrix calculated in θ_o.
The expression can be simplified as follows:

$$= L_n(\theta_o) - L(\theta_o) - \frac{d}{n} + o_p(1)$$

where d is the number of parameters of the model.

If $\hat{\theta}_n$ does converge in probability to θ_o, as it is supposed to do, looking at equation (22) we can easily verify, by Slutsky theorem, that:

$$L_n(\hat{\theta}_n) - L_n(\theta_o) = o_p(1)$$

consequently, we get the following expression:

$$\mathbb{E}_{\Theta_n}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right] = L_n(\hat{\theta}_n) - L(\theta_o) - \frac{d}{n} + o_p(1) \tag{23}$$

by generalized Slutsky theorem (see Demidenko (2013)[34]) and delta method it is possible to demonstrate the following convergence:

$$\text{plim}_{n\to+\infty}\, \mathbb{E}_{\Theta_n}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right] = 0 \tag{24}$$

using both results 23 24, we can get the following asymptotic result:

$$\text{plim}_{n\to+\infty}\left[L_n(\hat{\theta}_n) - \frac{d}{n}\right] = L(\theta_o) \tag{25}$$

on the base of this result we get a simplified alternative estimator of $\mathbb{E}_{\Theta_n}[H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})]$ that we call Predictability Information Criterion (PIC) with maximum likelihood estimtor (MLE).

$$PIC^{MLE}_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th}) = L_n(\hat{\theta}^B_n) - L_n(\hat{\theta}^A_n) + \frac{d^A}{n} - \frac{d^B}{n} \tag{26}$$

In fact, with the asymptotic result (25)), it's easy to see that both $PIC^{MLE}_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})$ and $H_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th})]$ converge in probability to the same limit $L(\theta^B_0) - L(\theta^A_0)$.

An explicit formulation of PIC is:

$$\boxed{PIC^{MLE}_{A,B}(\widetilde{y}_{re}, \widetilde{y}^A_{th}, \widetilde{y}^B_{th}) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{f_{e_{th}}(\widetilde{y}^i_{re} - \bar{y}_{re}|\Psi^B)\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}^i_{re} - \bar{y}_{re}|\Psi^A)}{f_{e_{th}}(\widetilde{y}^i_{re} - \bar{y}_{re}|\Psi^A)\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}^i_{re} - \bar{y}_{re}|\Psi^B)}\right) + \frac{d^A}{n} - \frac{d^B}{n}} \tag{27}$$

Where:

- d^A is the number of parameters of the model A.
- d^B is the number of parameters of the model B.
- n is the number of data in the sample
- $\Psi^A = \{\bar{y}_{re}, \bar{y}^A_{th}, \hat{\theta}^A_n, \hat{f}^n_{y_{re}}\}$
- $\Psi^B = \{\bar{y}_{re}, \bar{y}^B_{th}, \hat{\theta}^B_n, \hat{f}^n_{y_{re}}\}$

According to 6.1, if:

$PIC^{MLE}_{A,B} > 0 \Rightarrow$ the prediction of model A is expected to be more uncertain with respect to model B

$PIC^{MLE}_{A,B} < 0 \Rightarrow$ the prediction of model B is expected to be more uncertain with respect to model A

9 The Predictability Information Criterion (PIC) with Quasi-Maximum Likelihood Estimators (QMLE)

Give a real-valued diffusion process endowed with the following univariate stochastic differential equation:

$$dX_t = b(X_t, \theta)dt + \sigma(X_t, \theta)\, dW_t \tag{28}$$

with the vector of parameters θ, it is possible that the transition probability density does not have a closed form expression. Consequently, the maximum likelihood estimator of the vector of parameters θ cannot always be achievable. In these cases, quasi-maximum likelihood procedures try to fix the issue estimating the vector of parameters by the maximization of an approximated likelihood function.
The Quasi-Maximum Likelihood Estimators (QMLE) are built such that they are consistent and asymptotically normal. However such estimators are less efficient of the maximum likelihood estimators. In order to compute the PIC, implementing the quasi-maximum likelihood estimators, I'm going to adopt a procedure that has been already used by Yoshida (1992) [110], Genon-Catalot and Jacod (1993)[47], and Kessler (1997) [66] to estimate stochastic differential equations. In particular, I consider the version implemented in Iacus (2016) [60],Brouste et al. (2014) [20], Iacus (2011) [102] (pag 207), De Gregorio (2012) [32], which is obtained by discretization of the continuous time stochastic differential equation (28)(Euler-Maruyama scheme) assuming that the increments are conditionally independent Gaussian random variables.
The method splits the vector of parameters θ into two parts, since some of them converge at different rate [9]. $\theta = (\alpha, \beta)$ in particular the only q parameters appearing in σ can be estimated efficiently, where: $\alpha = (\alpha_1, \ldots, \alpha_p)' \in \Theta^p \subset \mathbb{R}^p$ and $\beta = (\beta_1, \ldots, \beta_q)' \in \Theta^q \subset \mathbb{R}^q$.
Given the random sample $\left\{X_t^i\right\}_{i=1,\ldots,n+1}$ observed only at n +1 equidistant discrete times t_i , such that, if the process X_t is ergodic, it is possible to demonstrate that the proposed Quasi Maximum Likelihood Estimator (QMLE) is a consistent estimator of θ_0 and asymptotically Gaussian with rate of convergence given by $\varphi(n)^{-1/2}$

$$\varphi(n)^{-1/2}\left(\hat{\theta}_n - \theta_o\right) \xrightarrow{d} N\left[0,\ I(\theta_o)^{-1}\right] \tag{29}$$

where:

- $\varphi(n) = \begin{pmatrix} \frac{1}{n \cdot h_n} I_p & 0 \\ 0 & \frac{1}{n} I_q \end{pmatrix}$
- $I(\theta_o)$ is the Fisher's matrix calculated in θ_o

[9] To be consistent with the definition of the parameter spaces provided in chapter8, I consider the dimension of the vector of parameters, for the generic model M, to be equal to the sum of the dimensions of α and β, meaning that $d^M = p^M + q^M$, where: d^M is the dimension of the parameters in model M; p^M is the dimension of the vector of parameters called α in model M, q^M is the dimension of the vector of parameters called β in model M. Coherently respect to what done so far, I do not indicate the apex M anymore, considering it implicit.

- I_p is the identity matrix of order p
- I_q is the identity matrix of order q
- h_n time interval between each consecutive pairs of data points observed in the sample such that: $h_n = t_i - t_{i-1} < \infty$ for $1 \leq i \leq n$. This value decreases when the number of observations, n, increases.

Proceeding in a similar manner to chapter 8.3, we can derive an equivalent representation of the predictability information criterion in the presence of the QMLE.

$$\mathbb{E}_{\Theta_n}\left[L_n(\hat{\theta}_n) - L(\hat{\theta}_n)\right] = L_n(\theta_o) - L(\theta_o) - \left[Trace\left(I(\theta_o)\cdot\varphi(n)\cdot I(\theta_o)^{-1}\right) + 0\right] + o_p(1)$$

$$= L_n(\theta_o) - L(\theta_o) - \left[Trace\left(\varphi(n)\cdot I_d\right)\right] + o_p(1)$$

$$= L_n(\theta_o) - L(\theta_o) - \frac{p}{n\cdot h_n} - \frac{q}{n} + o_p(1) \tag{30}$$

We call Predictability Information Criterion (PIC) with Quasi-Maximum Likelihood Estimators (QMLE) the following expression

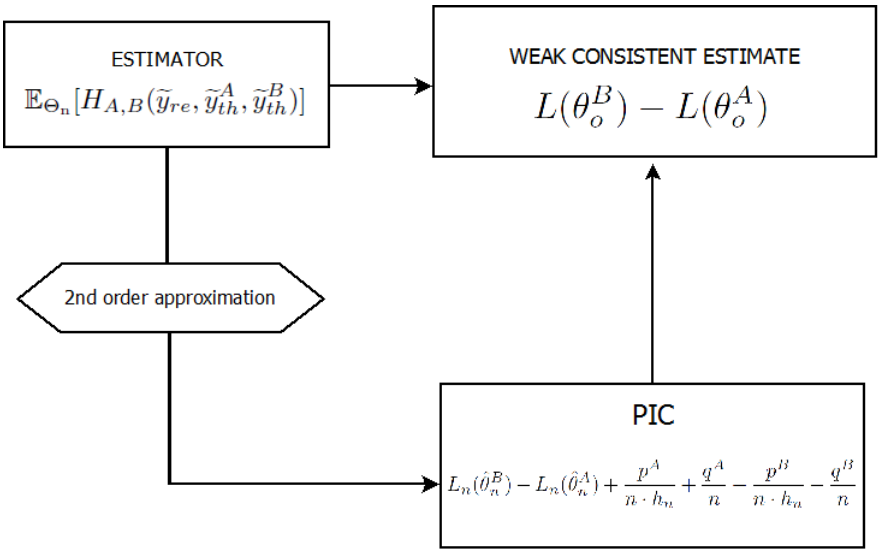
$$PIC_{A,B}^{QMLE}(\widetilde{y}_{re}, \widetilde{y}_{th}^{A}, \widetilde{y}_{th}^{B}) = L_n(\hat{\theta}_n^B) - L_n(\hat{\theta}_n^A) + \frac{p^A}{n\cdot h_n} + \frac{q^A}{n} - \frac{p^B}{n\cdot h_n} - \frac{q^B}{n} \tag{31}$$

An explicit formulation of PIC with Quasi-Maximum Likelihood Estimators (QMLE) is:

$$\boxed{PIC_{A,B}^{QMLE}(\widetilde{y}_{re}, \widetilde{y}_{th}^{A}, \widetilde{y}_{th}^{B}) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{f_{e_{th}}(\widetilde{y}_{re}^{i} - \bar{y}_{re}|\Psi^B)\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}_{re}^{i} - \bar{y}_{re}|\Psi^A)}{f_{e_{th}}(\widetilde{y}_{re}^{i} - \bar{y}_{re}|\Psi^A)\sum_{i=1}^{n} f_{e_{th}}(\widetilde{y}_{re}^{i} - \bar{y}_{re}|\Psi^B)}\right) + \frac{p^A}{n\cdot h_n} + \frac{q^A}{n} - \frac{p^B}{n\cdot h_n} - \frac{q^B}{n}} \tag{32}$$

where:

- p^A is the number of parameters inside $\sigma(X_t, \theta)$ of the model A.
- q^A is the number of remaining parameters of the model A.
- p^B is the number of parameters inside $\sigma(X_t, \theta)$ of the model B.
- q^B is the number of remaining parameters of the model B.
- $\Psi^A = \{\bar{y}_{re}, \bar{y}_{th}^A, \hat{\theta}_n^A, \hat{f}_{y_{re}}^n\}$
- $\Psi^B = \{\bar{y}_{re}, \bar{y}_{th}^B, \hat{\theta}_n^B, \hat{f}_{y_{re}}^n\}$
- n is the number of data in the sample
- $\text{plim}_{n\to+\infty} n\cdot h_n = \infty$
- $\text{plim}_{n\to+\infty} n\cdot h_n^2 = 0$



- $\text{plim}_{n \to +\infty} h_n = 0$

according to 6.1, if:

$PIC^{QMLE}_{A,B} > 0 \Rightarrow$ the prediction of model A is expected to be more uncertain with respect to model B.
$PIC^{QMLE}_{A,B} < 0 \Rightarrow$ the prediction of model B is expected to be more uncertain with respect to model A.

10 Oil Price Modeling

Price modeling of commodities is a very complex and difficult task since it needs to consider contemporaneously both the high complexity of the price discovery in the global market (in recent years commodities become a commercial asset of money managers worldwide) (Geman, 2005[46]) and macroeconomic factors they depend on (there are often many empirical studies providing conflicting results about the dependency structure among variables).
Modeling these peculiar prices has become a major objective of those who intend to use, or create, complex derivatives, now become widely used in the industrial production, for both speculative and investment purposes. Recently, the economic literature has started to address the problem of modelling crude oil prices again.
Much of the early literature considers the price of crude oil as affected by a single risk factor (single factor model): Paddock et al., 1988[85]; Brennan and Schwartz, 1985[18]; McDonald and Siegel, 1985[82]). They usually define the dynamics of prices as due to a Geometric Brownian Motion (GBM).
The choice comes from the analogy that characterises the movement of these prices with those of shares on the capital markets. Choosing a Geometric Brownian Motion means assuming that oil prices are expected to grow exponentially,

at a constant rate over time, and their variance to increase in proportion to time. However, many empirical studies suggest that its implementation involves drawbacks both in the descriptive power, and because it might induce mispricing of derivative securities.
GBM seems to counter the economic theory of the commodities, according to which the distance of the spot price from the theoretical long term price (intrinsic value) should affect the volumes of productions. For example, when the oil price is above the theoretical level, there should be an incentive to increase the production which reduces prices, up to the theoretical level, in the long run. When the price is lower than the theoretical level, there should be an incentive to reduce production and to wait for the rise of oil prices in the long run. In fact, a reduced oil supply on the market generally determines a rise in prices until it hooks the theoretical, in the long-run. So, although there is a temporary permissible misalignment, due to the stochastic perturbations, it is expected a well defined dynamics in the long run.
As remembered by Geman (2005)[46], two major commodity indexes in the 90's introduced a mean-reverting drift, in the stochastic differential equation, for driving oil price dynamics in their simulations. Moreover, many empirical studies supported this hypothesis: Schwartz (1997) [93] Pindyck (1999, 2001)[87][88]; Laughton and Jacoby 1993, 1995[72][73]; Cortazar and Schwartz 1994[27]; Dixit and Pindyck 1994[36]; Smith and McCardle 1999[99]; Dias 2004[35]; Begg and Smit 2007[12]; Willigers and Bratvold 2009[109]).
Note that, employing a GBM to model prices, that actually follow a mean-reverting dynamic, means overestimating the degree of risk related to future prices (See B. Jafarizadeh and Bratvold 2012 [62]).
That's why it is usually preferred to employ a mean reverting process which allows a better description of the oil market price dynamics (Pindyck, 2001 [88]).
Over the years, an increasing number of increasingly complex models has become the subject of investigation for economists. Discovering new stochastic processes, best suited to fit the data observed in the markets, is one of the primary objectives of banks and financial institutions that are seeking to build sophisticated financial instruments. The most successful models are those having two or three stochastic factors (Gibson and Schwartz (1990)[48], Cortazar and Schwartz (1994)[27], Schwartz (1997)[93], Pilipovic (1998) [86], Baker et al. (1998)[10], Hilliard and Reis (1998)[54], Schwartz and Smith (2000)[92], Cortazar and Schwartz (2003) [28]) or a mixture o stochastic processes such as those with jumps.
The need to introduce additional risk factors is mainly due to the stochastic nature of the long-term price. In general, more advanced models often make use of a greater number of parameters to be estimated. However, this increase can effectively improve the fitting properties of the model even without actually describing the right process. This results in misleading indications about the performance of the model to predict the future data. Therefore, in general, *the increase of the degree of adherence of the model to the data is not necessarily a positive feature*, from a statistical point of view. This big hidden drawback of such models is called the overfitting problem[10]. .

[10]There is overfitting when the model tends to adapt to the sample of data instead of the model that generates such a data.

10.1 Geometric Brownian Motion (GBM)

The Geometric Brownian Motion (GBM) is a continuous time stochastic process frequently used in finance to model the dynamic of asset prices X_t. The stochastic differential equation describing the process is constituted by the sum of a deterministic and a stochastic part:

$$dX_t = \mu X_t \, dt + \sigma X_t \, dW_t$$

where W_t is a Wiener process, μ is the percentage drift and σ is the instantaneous standard deviation. Some peculiar characteristics of the process are: 1) the variance is linearly increasing in time to a constant value σ^2; 2) the expected returns of oil price is constant and independent of the value of the process $\mathbb{E}(dX/X) = \alpha dt$. The Geometric Brownian Motion (GBM) has mathematical properties that make it very easy to implement. For example, it admits a closed form analytical solution

$$\ln\left(\frac{X_t}{X_0}\right) = \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t$$

with a normal probability density function.

$$\ln\left(\frac{X_t}{X_0}\right) \sim N\left[(\mu - \sigma^2/2)t, \sigma^2 t\right]$$

Some authors, such as Dixit (1992), believe that the GBM is a process that also guarantees further advantages, from a model-fitting point of view and when there is irreversibility of the risky investment plans. Whereas for Pindyck (1999) [87] the GBM doesn't provide meaningful results.

10.2 Ornstein–Uhlenbeck process (OU)

The Ornstein–Uhlenbeck process is a continuous-time stochastic process used to model random variables that tend to oscillate around a trend value (mean reversion).
This process could be considered as an extension, to a continuous variable, of a discrete autoregressive process of order 1 AR(1). The stochastic differential equation is

$$dS_t = (\mu - \kappa S_t) \, dt + \sigma \, dW_t$$

$\kappa > 0$, μ and $\sigma > 0$.
This process is not immediately applicable to asset prices because it does not exclude negative price values. Therefore Dixit and Pindyck (1994) [36] have proposed to include the change in prices rather than price levels.
Defining the asset prices again as X_t, we are allowed to interpret $S_t = \ln(X_t)$ for modelling price differences. It admits the following closed form solution:

$$\ln\left(\frac{X_t}{X_0}\right) = \frac{\mu}{\kappa} + \left(\ln(X_0) - \frac{\mu}{\kappa}\right) e^{-\kappa t} - \ln(X_0) + \sigma e^{-kt} \int_0^t e^{ks} dW_s$$

we can calculate the transition density function

$$\ln\left(\frac{X_t}{X_0}\right) \sim N\left[\frac{\mu}{\kappa} + \left(\ln(X_0) - \frac{\mu}{\kappa}\right) e^{-\kappa t} - \ln(X_0), \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa t}\right)\right]$$

11 Comparison of GBM and OU for Crude Oil Market - Empirical evidence

We are going to provide a practical example of the application of the Predictability Information Criterion (PIC). The example consists of determining which stochastic process, between GBM and OU, produces less uncertain forecasts for daily Cushing, OK WTI Spot Price FOB (Dollars per Barrel), according to the proposed criteria PIC, for each possible scenario $\bar{y}_{re}$ and forecasting horizon s. Historical price data, provided by EIA, start from 02/01/1986 and stop at 21/03/2016.
We define $\{\widetilde{y}_{re}(t) : t \in T\}$ to be a semimartingale process, consisting in the log-return of the WTI spot price from present time 0 to time t.

$$\widetilde{y}_{re}(t) = \ln\left(\frac{X_t}{X_0}\right)$$

Fixing a time horizon s and assuming that the price process is observed on an equally-spaced time line t = -n,...,-2,-1,0,1, 2,...,s,... (with $s < n$). We define the sample of the log returns with time horizon s as a set

$$\left\{\ln\left(\frac{X_{s(i+1)}}{X_{i\cdot s}}\right)\right\}_{i=-1,\dots,-n}$$

of i.i.d., (or weakly dependent. See proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]), random variables with density function $f_{y_{re}(s)}$.
The analysis employs a realisation of this sample for both calculating PIC and the estimates of the parameters of the two models.
We also define $\{\widetilde{y}_{th}^{GBM}(t) : t \in T\}$ and $\{\widetilde{y}_{th}^{OU}(t) : t \in T\}$ to be two parametric processes describing the behavior of the log-return of the WTI Spot Price from time 0 to time t. The first one assumes the crude oil spot price dynamically changes according to a geometric Brownian motion (GBM), the second one according to an Ornstein–Uhlenbeck process (OU).
We proceed with the estimation of the two stochastic models parameters making use of "yuima" library available in R. This library provides the parameters of the model based on the Quasi-Maximum Likelihood Estimator (QMLE) (Iacus (2016) [60],Brouste et al. (2014) [20], Iacus (2011) [102] pag 207).

Once we computed the parameters and fixed a time horizon s, we can calculate the PIC value of the two processes, using the logarithmic rates of returns, for each given value of $\bar{y}_{re}$ (unknown predictable part of the log returns) as follows:
for the Geometric Brownian Motion:

$$\widetilde{y}_{th}^{GBM}(s) = \ln\left(\frac{X_s}{X_0}\right) = \underbrace{\left(\mu - \frac{\sigma^2}{2}\right)s}_{prediction} + \underbrace{\sigma W_s}_{prediction\ error}$$

for the Ornstein–Uhlenbeck process:

$$\widetilde{y}_{th}^{OU}(s) = \ln\left(\frac{X_s}{X_0}\right) = \underbrace{\frac{\mu}{\kappa} + \left(\ln(X_0) - \frac{\mu}{\kappa}\right)e^{-\kappa s} - \ln(X_0)}_{prediction} + \underbrace{\sigma e^{-ks}\int_0^s e^{kj}dW_j}_{prediction\ error}$$

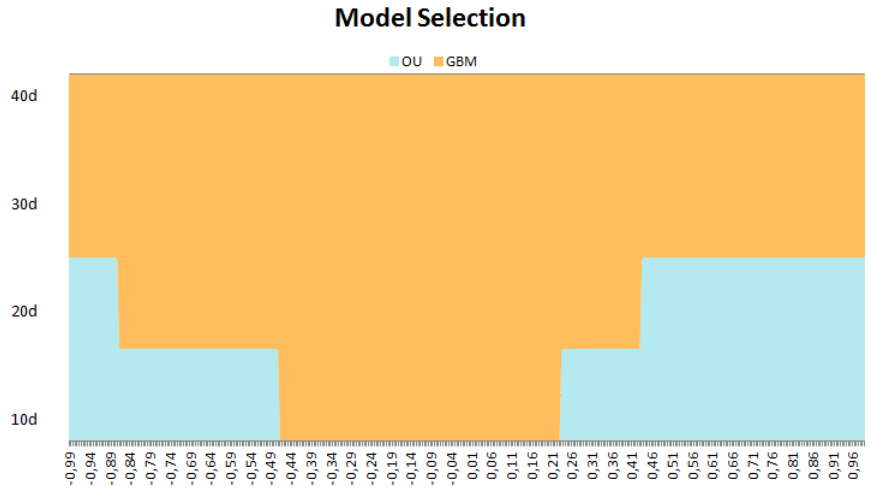


Figure 1: The coloured spaces highlight the less uncertain model for different states of nature $\bar{y}_{re}$ and time horizons s. The horizontal axis displays some possible scenarios for the predictable part $\bar{y}_{re}$, of the future logarithmic rate of return, while the vertical axis shows the time horizon s of the forecast.

you can notice that the two processes can be similarly expressed as we have done in formulas (7) (8).
Figure 11 shows what model should be selected according to PIC^{QMLE} (see (5)). In the light of our result reported below, we are confident that Ornstein–Uhlenbeck process provides better result in shorter time periods and just in the case of a strong trend of the crude oil price. The Geometric Brownian motion, despite its parsimony and simplicity, can perform better than OU for time horizons longer than 20 days and in the case of a flat or almost flat price trend.

a. With a 10 days time horizon $\forall\ \bar{y}_{re} \in (-0,465, 0,235)$ $PIC^{QMLE}_{GBM,OU} > 0 \Rightarrow$ OU has expected to be more uncertain with respect to GBM model

b. With a 20 days time horizon $\forall\ \bar{y}_{re} \in (-0,865, 0,43)$ $PIC^{QMLE}_{GBM,OU} < 0 \Rightarrow$ OU has expected to be more uncertain with respect to GBM model

c. With a 30 days time horizon $\forall\ \bar{y}_{re} \in (-0,99, 0,99)$ $PIC^{QMLE}_{GBM,OU} < 0 \Rightarrow$ OU has expected to be more uncertain with respect to GBM model

d. With a 40 days time horizon $\forall\ \bar{y}_{re} \in (-0,99, 0,99)$ $PIC^{QMLE}_{GBM,OU} < 0 \Rightarrow$ OU has expected to be more uncertain with respect to GBM model

12 Economic interpretation and final remarks

The nature of the underlying economic phenomena has broad implications for the choice of the stochastic model. For instance, in the work "A Microeconomic Approach to Diffusion Models for Stock Prices" (Föllmer and Schweizer (1993) [45]), it has been argued that, under some general conditions, the two analyzed processes of the example, GBM and OU, can both model the behavior of the equilibrium price. The key element is the concentration of two categories of agents in the market: 1) information traders (fundamentalists), who believe that the fundamentals drive the price; 2) noise traders, that instead respond to their own expectations.
In fact, the study suggests that the presence of different types of agents in the market has an effect on the resulting equilibrium price process. If only fundamentalist traders are active on the market, the price process, induced by information traders, behaves like an Ornstein-Uhlenbeck process around a time dependent level. If only noise traders are active on the market, the price process is induced by noise traders and it would be a geometric Brownian motion. In the light of this theoretical framework, for long-run forecasts, operators seem to behave predominantly as noise traders. Therefore GBM reasonably is a good choice. For short-run forecasts, operators appear to act mainly as informed traders, as long as it is expected a strong market change in price (high $|\bar{y}_{re}|$). Under more stable market expectations (low $|\bar{y}_{re}|$) operators seem to behave predominantly as noise traders and therefore geometric Brownian motion provides again better results. Strictly speaking, the informed traders mainly intervene on the market when expectations turn out to be too high or too low, meaning that mean reversion appears only when market expectation $\bar{y}_{re}$ exceeds certain thresholds. Deeper conclusions cannot be derived from the experiment since PIC does not ensure that the selected stochastic process is the right one.

13 Conclusion

In this paper, I provided a methodology to select alternative pricing models when prices, or log prices, are modelled as a stochastic process called semimartingale. I also recalled implicitly that, according to the general version of the Fundamental Theorem of Asset Pricing in continuous time (Delbaen and W. Schachermayer (1994)[33]), the proposed selection criterion is suitable when prices are in equilibrium in a market that does not admit arbitrage opportunities.
The predictability information criterion (PIC) is inspired by the Akaike's Information Criterion as it selects the model that most closely matches to that has generated the data by calculating and comparing distances in terms of Kullback-Leibler divergences. The PIC, unlike other selection criteria, is designed to compare models by their ability to simulate the only predictable part of the observed variable.
The proposed approach allows estimating the Kullback-Leibler divergence despite the fact that this predictable part, which has been named $\bar{y}_{re}$, is not directly observable. By parametrizing the estimate of the Kullback-Leibler distance to this value $\bar{y}_{re}$ and the time horizon s of the prediction, we are allowed to select a model for each possible scenario. I derived the predictability infor-

mation criterion (PIC) following three steps: 1) conditioning, for each model M, the employed probability density functions to Ω^M (see 12), in order to exclude that unpredictable factors can spoil the model selection procedure; 2) calculating, for each model M to compare, the Kullback-Leibler divergence between the theoretical model and the data generating process (as if it is known); 3) calculating an asymptotically unbiased estimator of this divergence. The paper can be ideally divided into three parts: the first part is an overview and the historical literature on modelling methodologies and pricing model selection. In particular, I explained the reasons why prices are modelled as semimartingale and what are the major types of information criteria that have been proposed over the years. The second part is devoted to the technical explanation of the proposed Predictability Information Criteria (PIC). Two versions are provided: 1) the first criterion is suitable for stochastic models whose parameters are estimated with the maximum likelihood technique; 2) a second version is presented for models estimated with a quasi-maximum likelihood procedure, as explained in Iacus (2011) [102]. The third part was, instead, dedicated to an application of the method to compare and select alternative stochastic models (Geometric Brownian Motion and Ornstein-Uhlenbeck) applied to data of the WTI spot prices. We made the selection for each possible scenario of $\bar{y}_{re}$ and time horizon of the forecast.
The advantages of the presented method are mainly three:
1) to allow a comparison between two models for each possible scenario $\bar{y}_{re}$ and for each possible time horizon s; 2) to compare the model uncertainty of predictions instead of the risks; 3) to evaluate models limiting the overfitting problem. The PIC can be a guide for understanding what model should be selected when the process can be partially predictable. Therefore it can be a tool particularly useful for practitioners, especially for pricing derivatives and risk quantification (e.g. in order to choose the best diffusion process for the quantification of VaR with Monte Carlo simulations).

COEFFICIENTS

10 days time horizon

Coefficients estimates of the GBM process

	μ	σ
Estimates	0.2326407	0.6281890
Std. Error	0.19675840	0.01633561

starting values

μ	1
σ	1

Coefficients estimates of the OU process

	μ	κ	σ
Estimates	1.5312540	0.4230507	1.0000000
Std. Error	1.66689142	0.46509013	0.09826974

starting values

μ	1
κ	1
σ	1

20 days time horizon

Coefficients estimates of the GBM process

	μ	σ
Estimates	0.2978370	0.7057621
Std. Error	0.24967004	0.02597167

starting values

μ	1
σ	1

Coefficients estimates of the OU process

	μ	κ	σ
Estimates	1.9642247	0.5430572	1.0000000
Std. Error	1.87462682	0.52325980	0.07505158

starting values

μ	1
κ	1
σ	1

30 days time horizon

Coefficients estimates of the GBM process

	μ	σ
Estimates	0.3576538	1.0000000
Std. Error	0.37993447	0.06869774

starting values	
μ	1
σ	1

Coefficients estimates of the OU process

	μ	κ	σ
Estimates	2.3919002	0.6613344	0.8761545
Std. Error	1.77896603	0.49600442	0.04854567

starting values	
μ	1
κ	1
σ	1

40 days time horizon

Coefficients estimates of the GBM process

	μ	σ
Estimates	0.4515039	0.8296112
Std. Error	0.33176734	0.04302128

starting values	
μ	1
σ	1

Coefficients estimates of the OU process

	μ	κ	σ
Estimates	2.8853790	0.7859244	0.9417678
Std. Error	2.00021662	0.55811807	0.06178878

starting values	
μ	1
κ	1
σ	1

APPENDIX

For the sake of clarity we redefine the theoretical variables and the information sets without specifying the model M they belong to:

$$\Omega = \{z = 0, \bar{y}_{re}, \bar{y}_{th}, \theta, f_{y_{re}}\}$$

$$\Psi = \{\bar{y}_{re}, \bar{y}_{th}, \theta, f_{y_{re}}\}$$

calculation of $f_{y_{re}}(y_{re}|\Psi)$

$$f_{y_{re}}(y_{re}|\Psi) = f_{e_{re}}(y_{re} - \bar{y}_{re}|\Psi)\left|1\right| \tag{33}$$

calculation of $f_{y_{th}}(y_{th}|\Psi)$

$$f_{y_{th}}(y_{th}|\Psi) = f_{e_{th}}(y_{th} - \bar{y}_{th}|\Psi)\left|1\right| \tag{34}$$

calculation of $f_Z(z|\Psi)$

Just by using the definition of conditional density function, we can calculate the density as follows

$$f_Z(z|\Psi) = \int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th} - z|\Psi) de_{th}$$

if $z = 0$ we get

$$f_Z(z = 0|\Psi) = \int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th} \tag{35}$$

calculation of $f_{y_{re}|\Omega}(y_{re}|\Omega)$

$$f_{y_{re}|z,\Psi}(y_{re}|z,\Psi) = \frac{f_{y_{re},z|\Psi}(y_{re}, z|\Psi)}{f_z(z|\Psi)} = \frac{f_{e_{re},e_{th}}(y_{re} - \bar{y}_{re}, z + y_{re} - \bar{y}_{re}|\Psi)\begin{vmatrix}1 & 0\\ 0 & 1\end{vmatrix}}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th} - z|\Psi) de_{th}}$$

if $z = 0$ we get

$$f_{y_{re}|\Omega}(y_{re}|\Omega) = f_{y_{re}|\Omega}(y_{re}|z = 0, \Psi) = \frac{f_{e_{re},e_{th}}(y_{re} - \bar{y}_{re}, y_{re} - \bar{y}_{re}|\Psi)}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}}$$

$$\frac{f_{e_{re}}(y_{re} - \bar{y}_{re}|\Psi) \cdot f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi)}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}} \tag{36}$$

calculation of $f_{y_{th}|\Omega}(y_{th}|\Omega)$

By using the same procedure as in formula (36) we get

$$f_{y_{th}|z}(y_{th}|z,\Psi) = \frac{f_{e_{re}}(y_{th} - \bar{y}_{th} - z|\Psi) \cdot f_{e_{th}}(y_{th} - \bar{y}_{th}|\Psi)\begin{vmatrix}0 & 1\\ 1 & 0\end{vmatrix}}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}}$$

if $z = 0$ we get

$$f_{y_{th}|\Omega}(y_{th}|\Omega) = f_{y_{th}|\Omega}(y_{th}|z=0,\Psi) = \frac{f_{e_{re}}(y_{th} - \bar{y}_{th}|\Psi) \cdot f_{e_{th}}(y_{th} - \bar{y}_{th}|\Psi)}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}} \tag{37}$$

calculation of $H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th})$

we want to calculate
$H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th}) = \mathbb{E}_{y_{re},y_{th}|\Omega}\left[log\left(\frac{p_{y_{re},y_{th}|\Omega}(y_{re},y_{th}|\Omega)}{p_{y_{re}|\Omega}(y_{re}|\Omega)p_{y_{th}|\Omega}(y_{th}|\Omega)}\right)\right]$
by discretization of the continuous variables ($\widetilde{y}_{re}$ and $\widetilde{y}_{th}$) and using Bayes formula we get the following expression:

$$p_{y_{re},y_{th}|\Omega}(y^i_{re}, y^j_{th}|\Omega) = p_{y_{re}|y_{th},\Omega}(y^i_{re}|y^j_{th},\Omega)p_{y_{th}|\Omega}(y^j_{th}|\Omega)$$

using the discretized variables y^i_{re} to y^j_{th} conditioned to the information set Ω we can write the terms defined in 7 8 as follows:

for every i and j

$$e^i_{re} = e^j_{th}$$

simplifying we get

$$y^i_{re} - \bar{y}_{re} = y^j_{th} - \bar{y}_{th}$$

$$y^i_{re} = \bar{y}_{re} + y^j_{th} - \bar{y}_{th}$$

this means that $i = j$ and the probability distribution $p_{y_{re}|y_{th},\Omega}(y^i_{re}|y^i_{th},\Omega)$ is totally concentrated at the point $y^i_{re} = \bar{y}_{re} + y^i_{th} - \bar{y}_{th}$

$p_{y_{re}|y_{th},\Omega}(y^i_{re}|y^i_{th},\Omega) = \int_{i\Delta}^{i\Delta+\Delta} \delta(y_{re} - (\bar{y}_{re} + y^i_{th} - \bar{y}_{th}))dy_{re} = \begin{cases} 1 \text{ if } y^i_{re} = \bar{y}_{re} + y^i_{th} - \bar{y}_{th} \\ 0 \text{ if } y^i_{re} \neq \bar{y}_{re} + y^i_{th} - \bar{y}_{th} \end{cases}$

where δ is the Dirac delta.
Thus imposing $y^i_{re} = \bar{y}_{re} + y^i_{th} - \bar{y}_{th}$ we can write
$p_{y_{re},y_{th}|\Omega}(y^i_{re}, y^i_{th}|\Omega) = p_{y_{re}|y_{th},\Omega}(y^i_{re}|y^i_{th},\Omega)p_{y_{th}|\Omega}(y^i_{th}|\Omega) = p_{y_{th}|\Omega}(y^i_{th}|\Omega)$
$= \int_{i\Delta}^{i\Delta+\Delta} f_{y_{th}|\Omega}(y_{th}|\Omega)dy_{th} = f_{y_{th}|\Omega}(y^i_{th}|\Omega)\Delta$
It follows that
$MI(y_{re}, y_{th}|\Omega) = \sum_i p_{y_{re},y_{th}|\Omega}(y^i_{re}, y^i_{th}|\Omega) log(\frac{p_{y_{re},y_{th}|\Omega}(y^i_{re}, y^i_{th}|\Omega)}{p_{y_{re}|\Omega}(y^i_{re}|\Omega)p_{y_{th}|\Omega}(y^i_{th}|\Omega)}) =$
$= \sum_i p_{y_{th}|\Omega}(y^i_{th}|\Omega) log(\frac{p_{y_{th}|\Omega}(y^i_{th}|\Omega)}{p_{y_{re}|\Omega}(y^i_{re}|\Omega)p_{y_{th}|\Omega}(y^i_{th}|\Omega)})$
$= \sum_i p_{y_{th}|\Omega}(y^i_{th}|\Omega) log(\frac{1}{p_{y_{re}|\Omega}(y^i_{re}|\Omega)})$
$= \sum_j p_{y_{th}|\Omega}(y^j_{th}|\Omega) log(\frac{1}{p_{y_{re}|\Omega}(\bar{y}_{re} + y^j_{th} - \bar{y}_{th}|\Omega)})$
$= \sum_i f_{y_{th}|\Omega}(y^i_{th}|\Omega)\Delta log(\frac{1}{f_{y_{re}|\Omega}(\bar{y}_{re} + y^i_{th} - \bar{y}_{th}|\Omega)\Delta})$

Letting $\Delta \to 0$ and considering both formulas (36), (37) we have
$= -\int_{-\infty}^{\infty} f_{y_{th}|\Omega}(y_{th}|\Omega) log(f_{y_{th}|\Omega}(y_{th}|\Omega)) dy_{th} - lim_{\Delta\to 0} log(\Delta) \sum_i f_{y_{th}|\Omega}(y_{th}^i|\Omega)\Delta$
$= -\int_{-\infty}^{\infty} f_{y_{th}|\Omega}(y_{th}|\Omega) log(f_{y_{th}|\Omega}(y_{th}|\Omega)) dy_{th} - \int_{-\infty}^{\infty} f_{y_{th}|\Omega}(y_{th}|\Omega) dy_{th} \cdot lim_{\Delta\to 0} log(\Delta)$

$$= -\int_{-\infty}^{\infty} f_{y_{th}|\Omega}(y_{th}|\Omega) log(f_{y_{th}|\Omega}(y_{th}|\Omega)) dy_{th} - lim_{\Delta\to 0} log(\Delta) \qquad (38)$$

Taking formula(37) we are allowed to write
$H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th}) =$
$-\int_{-\infty}^{\infty} \frac{f_{e_{re}}(y_{th} - \bar{y}_{th}|\Psi) \cdot f_{e_{th}}(y_{th} - \bar{y}_{th}|\Psi)}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}} log\left(\frac{f_{e_{re}}(y_{th} - \bar{y}_{th}|\Psi) \cdot f_{e_{th}}(y_{th} - \bar{y}_{th}|\Psi)}{\int_{-\infty}^{\infty} f_{e_{th}}(e_{th}|\Psi) \cdot f_{e_{re}}(e_{th}|\Psi) de_{th}}\right) dy_{th} - lim_{\Delta\to 0} log(\Delta)$
This formula can be computed for any generic theoretical model A and B coherently with the previous definitions (see equations 7 and 8).

For model A we write:
$H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th}^A) =$
$-\int_{-\infty}^{\infty} \frac{f_{e_{re}}(y_{th}^A - \bar{y}_{th}^A|\Psi^A) \cdot f_{e_{th}^A}(y_{th}^A - \bar{y}_{th}^A|\Psi^A)}{\int_{-\infty}^{\infty} f_{e_{th}^A}(e_{th}^A|\Psi^A) \cdot f_{e_{re}}(e_{th}^A|\Psi^A) de_{th}^A} log\left(\frac{f_{e_{re}}(y_{th}^A - \bar{y}_{th}^A|\Psi^A) \cdot f_{e_{th}^A}(y_{th}^A - \bar{y}_{th}^A|\Psi^A)}{\int_{-\infty}^{\infty} f_{e_{th}^A}(e_{th}^A|\Psi^A) \cdot f_{e_{re}}(e_{th}^A|\Psi^A) de_{th}^A}\right) dy_{th}^A - lim_{\Delta\to 0} log(\Delta)$
For model B we write:
$H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th}^B) =$
$-\int_{-\infty}^{\infty} \frac{f_{e_{re}}(y_{th}^B - \bar{y}_{th}^B|\Psi^B) \cdot f_{e_{th}^B}(y_{th}^B - \bar{y}_{th}^B|\Psi^B)}{\int_{-\infty}^{\infty} f_{e_{th}^B}(e_{th}^B|\Psi^B) \cdot f_{e_{re}}(e_{th}^B|\Psi^B) de_{th}^B} log\left(\frac{f_{e_{re}}(y_{th}^B - \bar{y}_{th}^B|\Psi^B) \cdot f_{e_{th}^B}(y_{th}^B - \bar{y}_{th}^B|\Psi^B)}{\int_{-\infty}^{\infty} f_{e_{th}^B}(e_{th}^B|\Psi^B) \cdot f_{e_{re}}(e_{th}^B|\Psi^B) de_{th}^B}\right) dy_{th}^B - lim_{\Delta\to 0} log(\Delta)$
Therefore using the expectation operator we get for the generic model M the following expression

$$H_{res}^M(\widetilde{y}_{re}, \widetilde{y}_{th}^M) = -\,\mathrm{E}_{y_{th}^M|\Omega^M}\left[log\left(f_{y_{th}^M|\Omega^M}(\widetilde{y}_{th}^M|\Omega^M)\right)\right] - lim_{\Delta\to 0} log(\Delta)$$

with

$$f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M) = \frac{f_{e_{re}}(y_{th}^M - \bar{y}_{th}^M|\Psi^M) \cdot f_{e_{th}^M}(y_{th}^M - \bar{y}_{th}^M|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^M}(e_{th}^M|\Psi^M) \cdot f_{e_{re}}(e_{th}^M|\Psi^M) de_{th}^M}$$

Study of the density $f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M)$

We wonder whether it is possible to make inference about the parameter θ^M of the probability density function $f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M)$ despite the fact that we don't own data from such distribution.
We can work around the problem by finding a sufficient statistic which can allows to make inference with the available data $\widetilde{y}_{re}|\Psi^M$. In this regard we study the relations between $\widetilde{y}_{th}^M|\Omega^M$, $\widetilde{y}_{re}|\Omega^M$ and $\widetilde{y}_{re}|\Psi^M$.
We define the following random variables first:
be

- $\widetilde{y}_{th}^M|\Omega^M$ a random variable with density function $f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M)$
- $\widetilde{y}_{th}^M|\Psi^M$ a random variable with density function $f_{y_{th}^M}(y_{th}^M|\Psi^M)$
- $\widetilde{y}_{re}|\Omega^M$ a random variable with density function $f_{y_{re}|\Omega^M}(y_{re}|\Omega^M)$
- $\widetilde{y}_{re}|\Psi^M$ a random variable with density function $f_{y_{re}}(y_{re}|\Psi^M)$

relation between $\widetilde{y}_{th}^{M}|\Omega^M$ and $\widetilde{y}_{re}|\Omega^M$

Now we wonder if there exist a continuous function g() such that we are allowed to write:

$$\widetilde{y}_{th}^{M}|\Omega^M = g(\widetilde{y}_{re}|\Omega^M)$$

and how can we calculate the density

$$f_{y_{th}^{M}|\Omega^M}(g(y_{re}|\Omega^M))$$

We first notice that under Ω^M

$$\widetilde{y}_{re}|\Omega^M - \bar{y}_{re} = \widetilde{y}_{th}^{M}|\Omega^M - \bar{y}_{th}^{M}$$

That is sufficient to demonstrate the existence of the continuous function g().

$$\widetilde{y}_{th}^{M}|\Omega^M = g(\widetilde{y}_{re}|\Omega^M) = \widetilde{y}_{re}|\Omega^M - \bar{y}_{re} + \bar{y}_{th}^{M}$$

The existence and the linearity of the function g() in $y_{re}|\Omega^M$ let us transform the original density function such that

$$f_{y_{th}^{M}|\Omega^M}(y_{th}^{M}|\Omega^M) = f_{y_{th}^{M}|\Omega^M}(g(y_{re}|\Omega^M)) = f_{y_{re}|\Omega^M}(y_{re}|\Omega^M)$$

however by formula (13) we can also write

$$f_{y_{th}^{M}|\Omega^M}(y_{th}^{M}|\Omega^M) = f_{y_{th}^{M}|\Omega^M}(g(y_{re}|\Omega^M)) = \frac{f_{e_{re}}(y_{re} - \bar{y}_{re}) \cdot f_{e_{th}^{M}}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^{M}}(e_{th}|\Psi^M) \cdot f_{e_{re}}(e_{th}) de_{th}}$$

relation between $\widetilde{y}_{th}^{M}|\Omega^M$, $\widetilde{y}_{re}|\Omega^M$ and $\widetilde{y}_{re}|\Psi^M$

Now we wonder if there exist a continuous function h() such that we are allowed to write :

$$\widetilde{y}_{th}^{M}|\Omega^M = g(\widetilde{y}_{re}|\Omega^M) = h(\widetilde{y}_{re}|\Psi^M) \tag{39}$$

and how can we calculate the density

$$f_{y_{th}^{M}}(h(y_{re}|\Psi^M))$$

Once again, by transformation of the original density function we get

$$f_{y_{th}^{M}|\Omega^M}(h(y_{re}|\Psi^M)) = f_{y_{re}}(y_{re}|\Psi^M) \cdot \frac{1}{\left|\frac{d}{dy_{re}}(h(y_{re}|\Psi^M))\right|}$$

By formula (33) we can write

$$f_{y_{th}^{M}|\Omega^M}(h(y_{re}|\Psi^M)) = f_{e_{re}}(y_{re} - \bar{y}_{re}|\Psi^M) \cdot \frac{1}{\left|\frac{d}{dy_{re}}(h(y_{re}|\Psi^M))\right|}$$

imposing

$$\frac{1}{\left|\frac{d}{dy_{re}}(h(y_{re}|\Psi^M))\right|} = \frac{f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^M}(e_{th}|\Psi^M) \cdot f_{e_{re}}(e_{th}|\Psi^M) de_{th}}$$

we get the following result

$$f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M) = f_{y_{th}^M|\Omega^M}(g(y_{re}|\Omega^M)) = f_{y_{th}^M|\Omega^M}(h(y_{re}|\Psi^M)) = f_{e_{re}}(y_{re} - \bar{y}_{re}|\Psi^M) \cdot \frac{f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^M}(e_{th}|\Psi^M) \cdot f_{e_{re}}(e_{th}|\Psi^M) de_{th}} \tag{40}$$

Consequences on the calculation of $H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th}^M)$

By focusing our attention on the expectation contained inside $H_{res}(\widetilde{y}_{re}, \widetilde{y}_{th})$ (see formula (14)) we notice that the relation 39 together with the law of the unconscious statistician let us simplify the formula as follows:

$$\mathbb{E}_{f_{y_{th}^M|\Omega^M}}\left[log\left(\frac{1}{f_{y_{th}^M|\Omega^M}(y_{th}^M|\Omega^M)}\right)\right] dy_{th}^M =$$

$$= \mathbb{E}_{\widetilde{y}_{re}}\left[log\left(\frac{1}{f_{y_{th}^M|\Omega^M}(h(y_{re}|\Psi^M))}\right)\right] dy_{re} =$$

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}|\Psi^M) log\left(\frac{1}{f_{y_{th}^M|\Omega^M}(h(y_{re}|\Psi^M))}\right) dy_{re} =$$

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}|\Psi^M) log\left(\frac{1}{\frac{f_{e_{th}^M}(e_{th}|\Psi^M) \cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{e_{th}^M}(e_{th}|\Psi^M) \cdot f_{e_{re}}(e_{th}|\Psi^M) de_{th}}}\right) dy_{re} =$$

using formulas (33) and (34)

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) log\left(\frac{1}{\frac{f_{y_{re}}(y_{re}) \cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M) dy_{re}}}\right) dy_{re}$$

notice that
$f_{y_{re}}(y_{re}|\Psi^M) = f_{y_{re}}(y_{re})\ \forall M$ and
$f_{y_{re}|\Omega^M}(y_{re}|\Omega^M) = \frac{f_{y_{re}}(y_{re}) \cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M)}{\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \cdot f_{e_{th}^M}(y_{re} - \bar{y}_{re}|\Psi^M) dy_{re}}$

Building the Estimator of $H_{A,B}$

The Maximum Likelihood Estimator (MLE) $\hat{\theta}$

The Maximum Likelihood estimator $\hat{\theta}_n$ for the parameter of the model M converges in distribution to a normal with expected value θ_o and variance $\frac{\Sigma}{n}$

$$\hat{\theta}_n \xrightarrow{d} N\left[\theta_o, \frac{\Sigma}{n}\right] \tag{41}$$

Density $\hat{f}^n_{y_{re}}$

Let $(\widetilde{y}^1_{re}, ..., \widetilde{y}^n_{re})$ be a random sample of i.i.d. random variables drawn from an unknown distribution with unknown density function $f_{y_{re}}(y_{re})$ or a weakly dependent (see proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]).
It is proposed to employ a uniform kernel density estimator

$$\hat{f}^n_{y_{re}} = \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{1}\Big(|\frac{\widetilde{y}^i_{re} - y_{re}}{h}| \leq 1\Big)$$

where:

- $\mathbf{1}\Big(|\frac{\widetilde{y}^i_{re} - y_{re}}{h}| \leq 1\Big)$ is an indicator function
- $h > 0$ is a smoothing parameter called the bandwidth

Study of the Asymptotic properties of $L(\hat{\theta}_n)$

Definition of $L(\hat{\theta}_n)$

We denote by $L(\hat{\theta}_n)$ the expected value of the log likelihood function with a random parameter $\hat{\theta}_n$. We shall consider it to be a function of the aforementioned estimator $\hat{\theta}_n$.

$$\begin{aligned} L(\hat{\theta}_n) &= \mathbb{E}_{y^M_{th}|\Omega^M}\left[log\left(f_{y^M_{th}|\Omega}(y_{th}|\Omega)\right)\right] \\ &= \mathbb{E}_{y_{re}}\left[log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right)\right] \\ &\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right) dy_{re} \end{aligned} \tag{42}$$

with

$$\Omega^M = \{z = 0, \bar{y}_{re}, \bar{y}^M_{th}, \hat{\theta}_n, f_{y_{re}}\}$$

Approximation of $L(\hat{\theta}_n)$ around θ_o

$$L(\hat{\theta}_n) \approx L(\theta_o) + \left(\hat{\theta}_n - \theta_o\right)' D_\theta \left\{L(\theta_o)\right\} + \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' D_\theta^2 \left\{L(\theta_o)\right\} \left(\hat{\theta}_n - \theta_o\right)$$

$$= L(\theta_o) - \frac{1}{2}\left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \left(\hat{\theta}_n - \theta_o\right) \tag{43}$$

Expectation of $L(\hat{\theta}_n)$

$$\mathbb{E}_\Theta \left[L(\hat{\theta}_n)\right] = L(\theta_o) - \frac{1}{2n}\left[Trace\left(I(\theta_o) \cdot \Sigma\right) + 0\right]$$

If the unbiased estimator $\hat{\theta}_n$ achieves the Cramér–Rao lower bound (CRLB) then $\Sigma = [I(\theta_o)]^{-1}$. Therefore the previous formula is simplified as follows

$$\mathbb{E}_\Theta \left[L(\hat{\theta}_n)\right] = L(\theta_o) - \frac{d}{2n} \tag{44}$$

Calculation of the Gradient $D_\theta \left\{L(\theta_o)\right\}$

$$D_\theta \left\{L(\theta_o)\right\}_i = \frac{\partial L(\theta_o, f_{y_{re}})}{\partial \theta_i}$$

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \frac{\partial log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right)}{\partial \theta_i} dy_{re} \tag{45}$$

exploiting the law of the unconscious statistician we get

$$= \int_{-\infty}^{\infty} f_{y_{re}|\Omega}(y_{re}|\Omega) \frac{\partial log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right)}{\partial \theta_i} dy_{re}$$

$$= \int_{-\infty}^{\infty} \frac{\partial f_{y_{re}|\Omega}(y_{re}|\Omega)}{\partial \theta_i} dy_{re}$$

under some regularity conditions we can invert the integral and the derivative operators

$$= \frac{\partial}{\partial \theta_i} \int_{-\infty}^{\infty} f_{y_{re}|\Omega}(y_{re}|\Omega) dy_{re}$$

$$= \frac{\partial}{\partial \theta_i} 1 = 0 \forall i$$

the explicit formula of (45) is

$$D_\theta \left\{L(\theta_o)\right\}_k = \mathbb{E}_{\widetilde{y}_{re}} \left(\frac{\frac{\partial}{\partial \theta_k} f_{e_{th}}(\widetilde{y}_{re} - \bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re} - \bar{y}_{re}|\Psi)} \right) - \frac{\mathbb{E}_{\widetilde{y}_{re}} \left(\frac{\partial}{\partial \theta_k} f_{e_{th}}(\widetilde{y}_{re} - \bar{y}_{re}|\Psi) \right)}{\mathbb{E}_{\widetilde{y}_{re}} \left(f_{e_{th}}(\widetilde{y}_{re} - \bar{y}_{re}|\Psi) \right)} = 0 \tag{46}$$

Calculation of the Hessian $D_\theta^2 \{L(\theta_o)\}$

$$D_\theta^2 \{L(\theta_o)\}_{ij} = \frac{\partial^2 L(\theta_o, f_{y_{re}})}{\partial\theta_i \partial\theta_j}$$

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \frac{\partial^2 log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right)}{\partial\theta_i \partial\theta_j} dy_{re}$$

by the law of the unconscious statistician we get

$$= \int_{-\infty}^{\infty} f_{y_{re}|\Omega}(y_{re}|\Omega) \frac{\partial^2 log\left(f_{y_{re}|\Omega}(y_{re}|\Omega)\right)}{\partial\theta_i \partial\theta_j} dy_{re}$$

Which is equal to the $ij-th$ element of the Fisher Information Matrix

$$= -I(\theta_o)_{ij} \tag{47}$$

The explicit formula is

$$= \int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \left(\frac{\frac{\partial^2}{\partial\theta_i \partial\theta_j} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi)}{f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi)} - \frac{\frac{\partial}{\partial\theta_i} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) \cdot \frac{\partial}{\partial\theta_j} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi)}{f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi)^2} \right) dy_{re} - \cdots$$

$$\cdots \left[\frac{\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) dy_{re}}{\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) dy_{re}} - \frac{\left(\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \frac{\partial}{\partial\theta_i} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) dy_{re}\right) \cdot \left(\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) \frac{\partial}{\partial\theta_j} f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) dy_{re}\right)}{\left(\int_{-\infty}^{\infty} f_{y_{re}}(y_{re}) f_{e_{th}}(y_{re} - \bar{y}_{re}|\Psi) dy_{re}\right)^2} \right]$$

. (48)

Approximation of $D_\theta \left\{ L(\hat{\theta}_n) \right\}$ around θ_o

$$D_\theta \left\{ L(\hat{\theta}_n) \right\} \approx D_\theta \{L(\theta_o)\} + \left(\hat{\theta}_n - \theta_o\right)' D_\theta^2 \{L(\theta_o)\}$$

using (46) and (47)

$$= -\left(\hat{\theta}_n - \theta_o\right)' I(\theta_o) \tag{49}$$

Expectation of $D_\theta \left\{ L(\hat{\theta}_n) \right\}$

$$\mathbb{E}_\Theta \left[D_\theta \left\{ L(\hat{\theta}_n) \right\} \right] \approx D_\theta \{L(\theta_o)\} + \mathbb{E}_\Theta \left[\left(\hat{\theta}_n - \theta_o\right)' \right] D_\theta^2 \{L(\theta_o)\}$$

On the base of formula (41) and of formula (46) we have approximately a null expectation

$$\mathbb{E}_\Theta \left[D_\theta \left\{ L(\hat{\theta}_n) \right\} \right] \approx 0 \tag{50}$$

Variance of $D_\theta\left\{L(\hat{\theta}_n)\right\}$

$$\mathbb{VAR}_\Theta\left[D_\theta\left\{L(\hat{\theta}_n)\right\}\right] \approx \mathbb{VAR}_\Theta\left[D_\theta\left\{L(\theta_o)\right\} + \left(\hat{\theta}_n - \theta_o\right)' D_\theta^2\left\{L(\theta_o)\right\}\right]$$

$$= \mathbb{VAR}_\Theta\left[\hat{\theta}'_n D_\theta^2\left\{L(\theta_o)\right\}\right]$$

$$= D_\theta^2\left\{L(\theta_o)\right\}\mathbb{VAR}_\Theta\left[\hat{\theta}_n\right] D_\theta^2\left\{L(\theta_o)\right\}$$

$$= D_\theta^2\left\{L(\theta_o)\right\}\mathbb{VAR}_\Theta\left[\hat{\theta}_n\right] D_\theta^2\left\{L(\theta_o)\right\}$$

$$= I(\theta_o)\frac{\Sigma}{n}I(\theta_o)$$

where Σ is the variance matrix of the estimator $\hat{\theta}_n$.

If the unbiased estimator $\hat{\theta}_n$ achieves the Cramér–Rao lower bound (CRLB) then $\Sigma = [I(\theta_o)]^{-1}$. Therefore the previous formula is simplified as follows

$$= I(\theta_o)\frac{[I(\theta_o)]^{-1}}{n}I(\theta_o)$$

$$= \frac{I(\theta_o)}{n} \tag{51}$$

Study of the asymptotic distribution of $D_\theta\left\{L(\hat{\theta}_n)\right\}$

Using the results (50) (51) we deduce its convergence in distribution

$$D_\theta\left\{L(\hat{\theta}_n)\right\} \xrightarrow{d} N\left[0,\ \frac{I(\theta_o)}{n}\right] \tag{52}$$

Study of the Asymptotic properties of $L_n(\hat{\theta}_n)$

Definition of $L_n(\hat{\theta}_n)$

We denote by $L_n(\hat{\theta}_n)$ the sample expected value of the log likelihood function with a random parameter $\hat{\theta}_n$. Let $\left\{\widetilde{y}^i_{re}\right\}_{i=1,\ldots,n}$ and $\left\{\widetilde{y}^j_{re}\right\}_{j=1,\ldots,n}$ be two random samples of i.i.d. random variables drawn from the same density distribution $f_{y_{re}}(y_{re})$, or weakly dependent (see proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11])random variables.
We define

$$L_n(\hat{\theta}_n) = = \int_{-\infty}^{\infty} \frac{1}{2nh}\sum_{i=1}^n \mathbf{1}\left(|\frac{\widetilde{y}^i_{re}-y_{re}}{h}| \leq 1\right) log\left(\frac{\frac{1}{2nh}\sum_{j=1}^n \mathbf{1}\left(|\frac{\widetilde{y}^j_{re}-y_{re}}{h}| \leq 1\right) f_{e_{th}}(y_{re}-\bar{y}_{re}|\Psi)}{\int_{-\infty}^{\infty}\frac{1}{2nh}\sum_{i=1}^n \mathbf{1}\left(|\frac{\widetilde{y}^i_{re}-y_{re}}{h}| \leq 1\right) f_{e_{th}}(y_{re}-\bar{y}_{re}|\Psi)dy_{re}}\right) dy_{re} \tag{53}$$

and simplifying we get

$$= \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{\frac{1}{2nh}\sum_{j=1}^{n}\mathbf{1}\left(|\frac{\widetilde{y}_{re}^{j}-\widetilde{y}_{re}^{i}}{h}|\leq 1\right)f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}\right) \tag{54}$$

2nd order approximation of $L_n(\hat{\theta}_n)$ around θ_o

$$L_n(\hat{\theta}_n) \approx L_n(\theta_o)+\left(\hat{\theta}_n-\theta_o\right)' D_\theta\left\{L_n(\theta_o)\right\}+\frac{1}{2}\left(\hat{\theta}_n-\theta_o\right)' D_\theta^2\left\{L_n(\theta_o)\right\}\left(\hat{\theta}_n-\theta_o\right) \tag{55}$$

by using formula (59)

$$= L_n(\theta_o)-\left(\hat{\theta}_n-\theta_o\right)' I(\theta_o)\left(\hat{\theta}_n-\theta_o\right)+\frac{1}{2}\left(\hat{\theta}_n-\theta_o\right)' D_\theta^2\left\{L_n(\theta_o)\right\}\left(\hat{\theta}_n-\theta_o\right)$$

Calculation of the Gradient $D_\theta\left\{L_n(\theta_o)\right\}$

The k-th element of the gradient is

$$D_\theta\left\{L_n(\theta_o)\right\}_k = \frac{\partial L_n(\theta_o, f_{y_{re}})}{\partial\theta_k}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k} log\left(\frac{\frac{1}{2nh}\sum_{j=1}^{n}\mathbf{1}\left(|\frac{\widetilde{y}_{re}^{j}-\widetilde{y}_{re}^{i}}{h}|\leq 1\right)f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}\right)$$

$$. = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}\right)-\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)} \tag{56}$$

Calculation of the Hessian $D_\theta^2\left\{L_n(\theta_o)\right\}$

the kj-th element of the hessian is

$$D_\theta^2\left\{L_n(\theta_o)\right\}_{kj} = \frac{\partial^2 L_n(\hat{\theta}_n, f_{y_{re}})}{\partial\theta_k\partial\theta_j}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{\partial^2}{\partial\theta_k\partial\theta_j}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}-\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)\cdot\frac{\partial}{\partial\theta_j}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)^2}\right)-\cdots$$

$$\cdots\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_k\partial\theta_j}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}-\frac{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)\right)^2}\right]$$

looking at formula (48) it's straightforward to see that $D_\theta^2\left\{L_n(\theta_o)\right\}$ converges in probability to $D_\theta^2\left\{L(\theta_o)\right\}$ as n goes to infinity, by the law of large numbers (LLN).

Study of the asymptotic distribution of $D_\theta \{L_n(\theta_o)\}$

In order to study the asymptotic distribution of $D_\theta \{L_n(\theta_o)\}$ we first focus our attention on the first addend in formula (56). If its mean and variance are constant and known, using the Central Limit Theorem (CLT)[11], we can immediately write its asymptotic distribution.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}\right) \xrightarrow{d} N\left[\mathbb{E}_{\widetilde{y}_{re}}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}\right), \frac{\mathbb{VAR}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}\right)}{n}\right]$$

Notice that, using the Weak Law of Large Numbers (LLN)[12] a property of convergence in probability [13] and the Slutsky's theorem [14], the second addend converge in probability to a constant as n tends to infinity

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)}{\frac{1}{n}\sum_{i=1}^{n}f_{e_{th}}(\widetilde{y}_{re}^{i}-\bar{y}_{re}|\Psi)} \xrightarrow{p} \frac{\mathbb{E}_{\widetilde{y}_{re}}\left(\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)\right)}{\mathbb{E}_{\widetilde{y}_{re}}\left(f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)\right)}$$

Therefore again by Slutsky's theorem[15] we can state that

$$D_\theta \{L_n(\theta_o)\}_k \xrightarrow{d} N\left[\mathbb{E}_{\widetilde{y}_{re}}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}\right) - \frac{\mathbb{E}_{\widetilde{y}_{re}}\left(\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)\right)}{\mathbb{E}_{\widetilde{y}_{re}}\left(f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)\right)}, \frac{\mathbb{VAR}\left(\frac{\frac{\partial}{\partial\theta_k}f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}{f_{e_{th}}(\widetilde{y}_{re}-\bar{y}_{re}|\Psi)}\right)}{n}\right]$$

using (46) we simplify as follows

$$D_\theta \{L_n(\theta_o)\} \xrightarrow{d} N\left[0, \frac{I(\theta_o)}{n}\right] \tag{57}$$

[11](Lindeberg–Lévy CLT) : given a sequence of i.i.d. random variables (or weakly dependent random variables. See proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]) $\widetilde{x}_1, \widetilde{x}_2, ...$ with $E[\widetilde{x}_i] = \mu$ and $Var[\widetilde{x}_n] = \sigma^2 < \infty$, the random variable $\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i\right) - \mu\right) \xrightarrow{d} N(0,\ \sigma^2)$ converges in distribution to a normal $N(0, \sigma^2)$ as n approaches infinity

[12](Weak Law of Large Numbers) Given a sequence of i.i.d. random variables (or weakly dependent random variables. See proposition 4.2. in Bardet, J. M., Doukhan, P., Lang, G., & Ragache, N. (2008) [11]) $\widetilde{x}_1, \widetilde{x}_2, ...$ with $E[\widetilde{x}_i] = \mu$ and $Var[\widetilde{x}_i] = \sigma^2 < \infty$, then the random variable $\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_n \xrightarrow{p} \mu$ as n approaches infinity

[13]$plim\,(\widetilde{x}_n\widetilde{y}_n) = plim\,(\widetilde{x}_n) \cdot plim\,(\widetilde{y}_n)$

[14]Given two random variables $\widetilde{y}_n$ and $\widetilde{z}_n$ and a function $g()$. If $\widetilde{y}_n = g(\widetilde{z}_n)$ and if the function g is continuous at $plim\,(\widetilde{z}_n)$ therefore $plim\,(\widetilde{y}_n) = g(plim\,(\widetilde{z}_n))$.

[15]we define $\widetilde{x}_n$ to be a statistic, $\widetilde{x}$ a random variable, c a constant if $\widetilde{x}_n \xrightarrow{d} \widetilde{x}$ and $\widetilde{y}_n \xrightarrow{p} c$. Then $\widetilde{x}_n + \widetilde{y}_n \xrightarrow{d} \widetilde{x} + c$.

Some Relevant Asymptotic Results

We have previously studied the asymptotic behavior of $D_\theta \left\{ L(\hat{\theta}_n) \right\}$ and $D_\theta \{L_n(\theta_o)\}$ the two results (see (52) (57)) allow us to identify an identity between random variables [16]):

$$D_\theta \{L_n(\theta_o)\} \approx D_\theta \left\{ L(\hat{\theta}_n) \right\} \tag{58}$$

Therefore using (49)

$$D_\theta \{L_n(\theta_o)\} \approx - \left(\hat{\theta}_n - \theta_o \right)' I(\theta_o) \tag{59}$$

Moreover, the previous result let us find a second relation.

$$D^2_\theta \{L_n(\theta_o)\} \approx D^2_\theta \left\{ L(\hat{\theta}_n) \right\} \tag{60}$$

Thus applying formula (59) we have

$$D^2_\theta \{L_n(\theta_o)\} \approx -I(\theta_o) \tag{61}$$

[16] Notice that $D_\theta \{L_n(\theta_o)\}$ is a random variable in $\hat{y}_{re}$ and $D_\theta \left\{ L(\hat{\theta}_n) \right\}$ is a random variable in $\hat{\theta}_n$

References

[1] Hirotogu Akaike. *Information theory and an extension of the maximum likelihood principle.* In: ***Selected Papers of Hirotugu Akaike***. Springer, 1998, pp. 199–213.

[2] Hirotugu Akaike. *Information theory and an extension of the maximum likelihood principle.* In: ***Second International Symposium on Information Theory***. Akademinai Kiado. 1973, pp. 267–281.

[3] Hirotugu Akaike. *A new look at the statistical model identification.* In: ***Automatic Control, IEEE Transactions on*** 19.6 (1974), pp. 716–723.

[4] Hirotugu Akaike. *Likelihood of a model and information criteria.* In: ***Journal of Econometrics*** 16.1 (1981), pp. 3–14. URL: `https://ideas.repec.org/a/eee/econom/v16y1981i1p3-14.html`.

[5] Hirotugu Akaike. *Information measures and model selection.* In: ***Bulletin of the International Statistical Institute*** 50.1 (1983), pp. 277–291.

[6] Hirotugu Akaike. *Prediction and entropy.* In: ***Selected Papers of Hirotugu Akaike***. Springer, 1985, pp. 387–410.

[7] S Amari, N Murata, and S Yoshizawa. *A criterion for determining the number of parameters in an artificial neural model.* In: ***Proc. ICANN'91*** (1991), pp. 9–14.

[8] Jean-Pascal Ansel and Christophe Stricker. *Lois de martingale, densités et décomposition de Föllmer Schweizer.* In: ***Annales de l'IHP Probabilités et statistiques***. Vol. 28. 3. 1992, pp. 375–392.

[9] Anthony C Atkinson. *Likelihood ratios, posterior odds and information criteria.* In: ***Journal of Econometrics*** 16.1 (1981), pp. 15–20.

[10] Malcolm P Baker, E Scott Mayfield, and John E Parsons. *Alternative models of uncertain commodity prices for use with modern asset pricing methods.* In: ***The Energy Journal*** (1998), pp. 115–148.

[11] Jean-Marc Bardet et al. *Dependent Lindeberg central limit theorem and some applications.* In: ***ESAIM: Probability and Statistics*** 12 (2008), pp. 154–172.

[12] Stephen Hope Begg, Nea Smit, et al. *Sensitivity of Project Economics to Uncertainty in Type and Parameters of Price Models.* In: ***SPE Annual Technical Conference and Exhibition***. Society of Petroleum Engineers. 2007.

[13] Jonathan B Berk and Richard C Green. *Mutual fund flows and performance in rational markets.* Tech. rep. National Bureau of Economic Research, 2002.

[14] Werner FM Bondt and Richard Thaler. *Does the stock market overreact?* In: ***The Journal of finance*** 40.3 (1985), pp. 793–805.

[15] Peter Bossaerts and Richard C Green. *A general equilibrium model of changing risk premia: Theory and tests.* In: ***Review of Financial Studies*** 2.4 (1989), pp. 467–493.

[16] Peter Bossaerts and Pierre Hillion. *Implementing statistical criteria to select return forecasting models: what do we learn?* In: ***Review of Financial Studies*** 12.2 (1999), pp. 405–428.

[17] Jacob Boudoukh, Matthew Richardson, and Robert F Whitelaw. *The myth of long-horizon predictability.* In: ***Review of Financial Studies*** 21.4 (2008), pp. 1577–1605.

[18] Michael J Brennan and Eduardo S Schwartz. *Evaluating natural resource investments.* In: ***Journal of business*** (1985), pp. 135–157.

[19] Jean-Bernard Brissaud. *The meanings of entropy.* In: ***Entropy*** 7.1 (2005), pp. 68–96.

[20] Alexandre Brouste et al. *The yuima project: A computational framework for simulation and inference of stochastic differential equations.* In: ***Journal of Statistical Software*** 57.4 (2014), pp. 1–51.

[21] Kenneth P Burnham and David R Anderson. *Multimodel inference understanding AIC and BIC in model selection.* In: ***Sociological methods & research*** 33.2 (2004), pp. 261–304.

[22] John Y Campbell and Robert J Shiller. *Stock prices, earnings, and expected dividends.* In: ***The Journal of Finance*** 43.3 (1988), pp. 661–676.

[23] José Pinto Casquilho. *Discussing an expected utility and weighted entropy framework.* In: ***Natural Science*** 2014 (2014).

[24] Gregory C Chow. *A comparison of the information and posterior probability criteria for model selection.* In: ***Journal of Econometrics*** 16.1 (1981), pp. 21–33.

[25] John H Cochrane. *Asset Pricing:(Revised Edition).* Princeton university press, 2009.

[26] Rama Cont. *Model uncertainty and its impact on the pricing of derivative instruments.* In: ***Mathematical finance*** 16.3 (2006), pp. 519–547.

[27] Gonzalo Cortazar and Eduardo S Schwartz. *The valuation of commodity contingent claims.* In: ***The Journal of Derivatives*** 1.4 (1994), pp. 27–39.

[28] Gonzalo Cortazar and Eduardo S Schwartz. *Implementing a stochastic model for oil futures prices.* In: ***Energy Economics*** 25.3 (2003), pp. 215–238.

[29] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[30] John C Cox and Chi-fu Huang. *Optimal consumption and portfolio policies when asset prices follow a diffusion process.* In: ***Journal of economic theory*** 49.1 (1989), pp. 33–83.

[31] John C Cox and Chi-Fu Huang. *A variational problem arising in financial economics.* In: ***Journal of Mathematical Economics*** 20.5 (1991), pp. 465–487.

[32] Alessandro De Gregorio and Stefano M Iacus. *Adaptive Lasso-type estimation for multivariate diffusion processes.* In: ***Econometric Theory*** 28.04 (2012), pp. 838–860.

[33] Freddy Delbaen and Walter Schachermayer. *A general version of the fundamental theorem of asset pricing*. In: ***Mathematische annalen*** 300.1 (1994), pp. 463–520.

[34] Eugene Demidenko. *Mixed models: theory and applications with R*. John Wiley & Sons, 2013.

[35] Marco Antonio Guimaraes Dias. *Valuation of exploration and production assets: an overview of real options models*. In: ***Journal of Petroleum Science and Engineering*** 44.1 (2004), pp. 93–114.

[36] Avinash K Dixit and Robert S Pindyck. *Investment under uncertainty*. Princeton university press, 1994.

[37] Darrell Duffie. *Stochastic Equilibria: Existence, Spanning Number, and theNo Expected Financial Gain from Trade'Hypothesis*. In: ***Econometrica: Journal of the Econometric Society*** (1986), pp. 1161–1183.

[38] Darrell Duffie and Chi-Fu Huang. *Implementing Arrow-Debreu equilibria by continuous trading of few long-lived securities*. In: ***Econometrica: Journal of the Econometric Society*** (1985), pp. 1337–1356.

[39] Darrell Duffie and Chi-fu Huang. *Multiperiod security markets with differential information: martingales and resolution times*. In: ***Journal of Mathematical Economics*** 15.3 (1986), pp. 283–303.

[40] Eugene F Fama. *Mandelbrot and the stable Paretian hypothesis*. In: ***The journal of business*** 36.4 (1963), pp. 420–429.

[41] Eugene F Fama. *The behavior of stock-market prices*. In: ***The journal of Business*** 38.1 (1965), pp. 34–105.

[42] Eugene F Fama. *Random walks in stock market prices*. In: ***Financial analysts journal*** 51.1 (1995), pp. 75–80.

[43] Eugene F Fama and Kenneth R French. *Dividend yields and expected stock returns*. In: ***Journal of financial economics*** 22.1 (1988), pp. 3–25.

[44] Hans Föllmer, Alexander Schied, et al. *Probabilistic aspects of finance*. In: ***Bernoulli*** 19.4 (2013), pp. 1306–1326.

[45] Hans Föllmer and Martin Schweizer. *A microeconomic approach to diffusion models for stock prices*. In: ***Mathematical Finance*** 3.1 (1993), pp. 1–23.

[46] Helyette Geman. *Commodities and commodity derivatives*. In: ***Modeling and Pricing for Agriculturals, Metals and Energy*** (2005).

[47] Valentine Genon-Catalot and Jean Jacod. *On the estimation of the diffusion coefficient for multi-dimensional diffusion processes*. In: ***Annales de l'IHP Probabilités et statistiques***. Vol. 29. 1. 1993, pp. 119–151.

[48] Rajna Gibson and Eduardo S Schwartz. *Stochastic convenience yield and the pricing of oil contingent claims*. In: ***The Journal of Finance*** 45.3 (1990), pp. 959–976.

[49] Amit Goyal and Ivo Welch. *Predicting the equity premium with dividend ratios*. In: ***Management Science*** 49.5 (2003), pp. 639–654.

[50] J Michael Harrison and David M Kreps. *Martingales and arbitrage in multiperiod securities markets*. In: ***Journal of Economic theory*** 20.3 (1979), pp. 381–408.

[51] J Michael Harrison and Stanley R Pliska. *Martingales and stochastic integrals in the theory of continuous trading*. In: ***Stochastic processes and their applications*** 11.3 (1981), pp. 215–260.

[52] J Michael Harrison and Stanley R Pliska. *A stochastic calculus model of continuous trading: complete markets*. In: ***Stochastic processes and their applications*** 15.3 (1983), pp. 313–316.

[53] Ralph VL Hartley. *Transmission of information1*. In: ***Bell System technical journal*** 7.3 (1928), pp. 535–563.

[54] Jimmy E Hilliard and Jorge Reis. *Valuation of commodity futures and options under stochastic convenience yields, interest rates, and jump diffusions in the spot*. In: ***Journal of Financial and Quantitative Analysis*** 33.01 (1998), pp. 61–86.

[55] Robert J Hodrick. *Dividend yields and expected stock returns: Alternative procedures for inference and measurement*. In: ***Review of Financial studies*** 5.3 (1992), pp. 357–386.

[56] Jennifer A Hoeting et al. *Bayesian model averaging: a tutorial*. In: ***Statistical science*** (1999), pp. 382–401.

[57] Chi-Fu Huang. *Information structures and viable price systems*. In: ***Journal of Mathematical Economics*** 14.3 (1985), pp. 215–240.

[58] Chi-Fu Huang. *An intertemporal general equilibrium asset pricing model: The case of diffusion information*. In: ***Econometrica: Journal of the Econometric Society*** (1987), pp. 117–142.

[59] Clifford M Hurvich and Chih-Ling Tsai. *Regression and time series model selection in small samples*. In: ***Biometrika*** 76.2 (1989), pp. 297–307.

[60] Maintainer Stefano M Iacus. *Package 'yuima'*. In: (2016).

[61] Makio Ishiguro, Yosiyuki Sakamoto, and Genshiro Kitagawa. *Bootstrapping log likelihood and EIC, an extension of AIC*. In: ***Annals of the Institute of Statistical Mathematics*** 49.3 (1997), pp. 411–434.

[62] Babak Jafarizadeh, Reidar Brumer Bratvold, et al. *A Two-Factor Price Process for Modeling Uncertainty in Oil Prices*. In: ***SPE Hydrocarbon Economics and Evaluation Symposium***. Society of Petroleum Engineers. 2012.

[63] Edwin T Jaynes. *Information theory and statistical mechanics*. In: ***Physical review*** 106.4 (1957), p. 620.

[64] Constantinos Kardaras and Eckhard Platen. *On the semimartingale property of discounted asset-price processes*. In: ***Stochastic processes and their Applications*** 121.11 (2011), pp. 2678–2691.

[65] Donald B Keim and Robert F Stambaugh. *Predicting returns in the stock and bond markets*. In: ***Journal of financial Economics*** 17.2 (1986), pp. 357–390.

[66] Mathieu Kessler. *Estimation of an ergodic diffusion from discrete observations.* In: ***Scandinavian Journal of Statistics*** 24.2 (1997), pp. 211–229.

[67] George J Klir. *Uncertainty and information: foundations of generalized information theory.* John Wiley & Sons, 2005.

[68] Andrei N Kolmogorov. *Three approaches to the quantitative definition ofinformation'.* In: ***Problems of information transmission*** 1.1 (1965), pp. 1–7.

[69] David M Kreps. *Arbitrage and equilibrium in economies with infinitely many commodities.* In: ***Journal of Mathematical Economics*** 8.1 (1981), pp. 15–35.

[70] Jouni Kuha. *AIC and BIC comparisons of assumptions and performance.* In: ***Sociological Methods & Research*** 33.2 (2004), pp. 188–229.

[71] Carol C Kuhlthau. *A principle of uncertainty for information seeking.* In: ***Journal of documentation*** 49.4 (1993), pp. 339–355.

[72] David G Laughton and Henry D Jacoby. *Reversion, timing options, and long-term decision-making.* In: ***Financial Management*** (1993), pp. 225–240.

[73] David G Laughton and Henry D Jacoby. *The effects of reversion on commodity projects of different length.* In: ***Real options in capital investments: Models, strategies, and applications.*** (1995), pp. 185–205.

[74] Edward E Leamer. *Information criteria for choice of regression models: A comment.* In: ***Econometrica: Journal of the Econometric Society*** (1979), pp. 507–510.

[75] Stephen F LeRoy. *Risk aversion and the martingale property of stock prices.* In: ***International Economic Review*** (1973), pp. 436–446.

[76] Stephen F LeRoy and Larry D Singell. *Knight on risk and uncertainty.* In: ***The Journal of Political Economy*** (1987), pp. 394–406.

[77] Jonathan Lewellen. *Predicting returns with financial ratios.* In: ***Journal of Financial Economics*** 74.2 (2004), pp. 209–235.

[78] Angelika Linde. *DIC in variable selection.* In: ***Statistica Neerlandica*** 59.1 (2005), pp. 45–56.

[79] Dennis V Lindley. *On a measure of the information provided by an experiment.* In: ***The Annals of Mathematical Statistics*** (1956), pp. 986–1005.

[80] Robert E Lucas Jr. *Asset prices in an exchange economy.* In: ***Econometrica: Journal of the Econometric Society*** (1978), pp. 1429–1445.

[81] Burton G Malkiel and Eugene F Fama. *Efficient capital markets: A review of theory and empirical work.* In: ***The journal of Finance*** 25.2 (1970), pp. 383–417.

[82] Robert L McDonald and Daniel R Siegel. *Investment and the valuation of firms when there is an option to shut down.* In: ***International economic review*** (1985), pp. 331–349.

[83] Ryuei Nishii. *Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression*. In: ***Ann. Statist.*** 12.2 (June 1984), pp. 758–765. DOI: 10.1214/aos/1176346522. URL: http://dx.doi.org/10.1214/aos/1176346522.

[84] Ryuei Nishii et al. *Asymptotic properties of criteria for selection of variables in multiple regression*. In: ***The Annals of Statistics*** 12.2 (1984), pp. 758–765.

[85] James L Paddock, Daniel R Siegel, and James L Smith. *Option valuation of claims on real assets: The case of offshore petroleum leases*. In: ***The Quarterly Journal of Economics*** (1988), pp. 479–508.

[86] Dragana Pilipovic. *Energy Risk: Valuing and Managing Energy Derivatives*. McGraw Hill Professional, 2007.

[87] Robert S Pindyck. *The long-run evolution of energy prices*. In: ***The Energy Journal*** (1999), pp. 1–27.

[88] Robert S Pindyck. *The dynamics of commodity spot and futures markets: a primer*. In: ***The Energy Journal*** (2001), pp. 1–29.

[89] Stanley R Pliska. *A stochastic calculus model of continuous trading: optimal portfolios*. In: ***Mathematics of Operations Research*** 11.2 (1986), pp. 371–382.

[90] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. *Akaike information criterion statistics*. In: ***Dordrecht, The Netherlands: D. Reidel*** (1986).

[91] Paul A Samuelson. *Proof that properly anticipated prices fluctuate randomly*. 1965.

[92] Eduardo Schwartz and James E Smith. *Short-term variations and long-term dynamics in commodity prices*. In: ***Management Science*** 46.7 (2000), pp. 893–911.

[93] Eduardo S Schwartz. *The stochastic behavior of commodity prices: Implications for valuation and hedging*. In: ***The Journal of Finance*** 52.3 (1997), pp. 923–973.

[94] Gideon Schwarz et al. *Estimating the dimension of a model*. In: ***The annals of statistics*** 6.2 (1978), pp. 461–464.

[95] *Security Markets—Stochastic Models*. In: ().

[96] Claude Elwood Shannon. *A mathematical theory of communication*. In: ***ACM SIGMOBILE Mobile Computing and Communications Review*** 5.1 (2001), pp. 3–55.

[97] Ritei Shibata. *An optimal selection of regression variables*. In: ***Biometrika*** 68.1 (1981), pp. 45–54.

[98] Christopher A Sims. *MARTINGALE-LIKE BEHAVIOR OF PRICES AND INTEREST RATES*. In: (1990).

[99] James E Smith and Kevin F McCardle. *Options in the real world: Lessons learned in evaluating oil and gas investments*. In: ***Operations Research*** 47.1 (1999), pp. 1–15.

[100] David J Spiegelhalter et al. *Bayesian measures of model complexity and fit*. In: ***Journal of the Royal Statistical Society: Series B (Statistical Methodology)*** 64.4 (2002), pp. 583–639.

[101] Robert F Stambaugh. *Predictive regressions*. In: ***Journal of Financial Economics*** 54.3 (1999), pp. 375–421.

[102] M Iacus Stefano. *Option Pricing and Estimation of Financial Models with R*. 2011.

[103] M Stone. *Comments on model selection criteria of Akaike and Schwarz*. In: ***Journal of the Royal Statistical Society. Series B (Methodological)*** (1979), pp. 276–278.

[104] Nariaki Sugiura. *Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's*. In: ***Communications in Statistics-Theory and Methods*** 7.1 (1978), pp. 13–26.

[105] K Takeuchi. *Distribution of informational statistics and a criterion of model fitting*. 1976.

[106] Allan Timmermann and Clive WJ Granger. *Efficient market hypothesis and forecasting*. In: ***International Journal of forecasting*** 20.1 (2004), pp. 15–27.

[107] Rossen Valkanov. *Long-horizon regressions: theoretical results and applications*. In: ***Journal of Financial Economics*** 68.2 (2003), pp. 201–232.

[108] Ching-Zong Wei. *On predictive least squares principles*. In: ***The Annals of Statistics*** (1992), pp. 1–42.

[109] Bart JA Willigers, Reidar B Bratvold, et al. *Valuing oil and gas options by least-squares Monte Carlo simulation*. In: ***SPE Projects, Facilities & Construction*** 4.04 (2009), pp. 146–155.

[110] Nakahiro Yoshida. *Estimation for diffusion processes from discrete observation*. In: ***Journal of Multivariate Analysis*** 41.2 (1992), pp. 220–242.

R Code

```
Dati.per.R <- read.delim2("E:/Dati␣per␣R.txt")
View(Dati.per.R)
library("yuima", lib.loc="~/R/win-library/3.2")
library("fBasics", lib.loc="~/R/win-library/3.2")

# price frequency 10 days
p10daysna<-is.na(Dati.per.R$X10ggp)
p10days<-Dati.per.R$X10ggp[!p10daysna]

# price frequency 20 days
p20daysna<-is.na(Dati.per.R$X20ggp)
p20days<-Dati.per.R$X20ggp[!p20daysna]

# price frequency 30 days
p30daysna<-is.na(Dati.per.R$X30ggp)
p30days<-Dati.per.R$X30ggp[!p30daysna]

# price frequency 40 days
p40daysna<-is.na(Dati.per.R$X40ggp)
p40days<-Dati.per.R$X40ggp[!p40daysna]

# logprice frequency 10 days
lnp10days<-log(p10days)

# logprice frequency 20 days
lnp20days<-log(p20days)

# logprice frequency 30 days
lnp30days<-log(p30days)

# logprice frequency 40 days
lnp40days<-log(p40days)

# Geometric Brwnian Motion 10days
ydata<-setData(p10days, delta=1/(length(p10days)^0.65))
diff.matrix4 <- matrix(c("sigma*x"), 1, 1)
ymodel4 <- setModel(drift = c("(mu*x)"), diffusion = diff.matrix4, time.
    variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mle10days <- qmle(yuima4, start=list(mu=1, sigma=1),method="L-BFGS-B",
    lower=list(mu=0,sigma=0))
coef(mle10days)
summary(mle10days)
prof10 <- profile (mle10days)
vcov ( mle10days)

# Geometric Brwnian Motion 20days
ydata<-setData(p20days, delta=1/(length(p20days)^0.65))
diff.matrix4 <- matrix(c("sigma*x"), 1, 1)
ymodel4 <- setModel(drift = c("(mu*x)"), diffusion = diff.matrix4,time.
    variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
```

```
mle20days <- qmle(yuima4, start=list(mu=1, sigma=1),method="L-BFGS-B",
    lower=list(mu=0,sigma=0))
coef(mle20days)
summary(mle20days)
prof20 <- profile (mle20days)
vcov ( mle20days)


# Geometric Brwnian Motion 30days
ydata<-setData(p30days, delta=1/(length(p30days)^0.65))
diff.matrix4 <- matrix(c("sigma*x"), 1, 1)
ymodel4 <- setModel(drift = c("(mu*x)"), diffusion = diff.matrix4,time.
    variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mle30days <- qmle(yuima4, start=list(mu=1, sigma=1), method="L-BFGS-B",
    lower=list(mu=0,sigma=0))
coef(mle30days)
summary(mle30days)
prof30 <- profile (mle30days)
vcov ( mle30days)

# Geometric Brwnian Motion 40days
ydata<-setData(p40days, delta=1/(length(p40days)^0.65))
diff.matrix4 <- matrix(c("sigma*x"), 1, 1)
ymodel4 <- setModel(drift = c("(mu*x)"), diffusion = diff.matrix4, time.
    variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mle40days <- qmle(yuima4, start=list(mu=1, sigma=1), method="L-BFGS-B",
    lower=list(mu=0,sigma=0))
coef(mle40days)
summary(mle40days)
prof40 <- profile (mle40days)
vcov ( mle40days)

# Ornstein-Uhlenbeck process 10days
ydata<-setData(lnp10days, delta=1/(length(lnp10days)^0.65))
diff.matrix4 <- matrix(c("sigma"), 1, 1)
ymodel4 <- setModel(drift = c("(mu-kappa*x)"), diffusion = diff.matrix4,
    time.variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mleln10days <- qmle(yuima4, start=list(mu=1, kappa = 1, sigma =1), method
    ="L-BFGS-B", lower=list(mu=0,kappa=0,sigma=1))
coef(mleln10days)
summary(mleln10days)
prof10 <- profile (mleln10days)
vcov ( mleln10days)


# Ornstein-Uhlenbeck process 20days
ydata<-setData(lnp20days, delta=1/(length(lnp20days)^0.65))
diff.matrix4 <- matrix(c("sigma"), 1, 1)
ymodel4 <- setModel(drift = c("(mu-kappa*x)"), diffusion = diff.matrix4,
    time.variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
```

```
mleln20days <- qmle(yuima4, start=list(mu=1, kappa = 1, sigma =1), method
    ="L-BFGS-B", lower=list(mu=0,kappa=0,sigma=1))
coef(mleln20days)
summary(mleln20days)
prof20 <- profile (mleln20days)
vcov ( mleln20days)


# Ornstein-Uhlenbeck process 30days
ydata<-setData(lnp30days, delta=1/(length(lnp30days)^0.65))
diff.matrix4 <- matrix(c("sigma"), 1, 1)
ymodel4 <- setModel(drift = c("(mu-kappa*x)"), diffusion = diff.matrix4,
    time.variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mleln30days <- qmle(yuima4, start=list(mu=1, kappa =1, sigma =1), method=
    "L-BFGS-B", lower=list(mu=0,kappa=0,sigma=1))
coef(mleln30days)
summary(mleln30days)
prof30 <- profile (mleln30days)
vcov ( mleln30days)


# Ornstein-Uhlenbeck process 40days
ydata<-setData(lnp40days, delta=1/(length(lnp40days)^0.65))
diff.matrix4 <- matrix(c("sigma"), 1, 1)
ymodel4 <- setModel(drift = c("(mu-kappa*x)"), diffusion = diff.matrix4,
    time.variable = "t", state.variable = "x", solve.variable = "x")
yuima4 <- setYuima(data=ydata, model = ymodel4)
mleln40days <- qmle(yuima4, start=list(mu=1, kappa =1, sigma =1), method=
    "L-BFGS-B", lower=list(mu=0,kappa=0,sigma=1))
coef(mleln40days)
summary(mleln40days)
prof40 <- profile (mleln40days)
vcov ( mleln40days)

# Coefficients


coef(mle10days)
coef(mle20days)
coef(mle30days)
coef(mle40days)
coef(mleln10days)
coef(mleln20days)
coef(mleln30days)
coef(mleln40days)

# h

1/(length(p10days)^0.65)
1/(length(p20days)^0.65)
1/(length(p30days)^0.65)
1/(length(p40days)^0.65)
```

```
# last value

lnp10days[end(lnp10days)[1]]
lnp20days[end(lnp20days)[1]]
lnp30days[end(lnp30days)[1]]
lnp40days[end(lnp40days)[1]]

# lengths

length(lnp10days)
length(lnp20days)
length(lnp30days)
length(lnp40days)


install.packages("Box.test")

# Autocorrelograms

acfPlot(diff(lnp10days, lag=1), lag.max = 20)
acfPlot(diff(lnp20days, lag=1), lag.max = 20)
acfPlot(diff(lnp30days, lag=1), lag.max = 20)
acfPlot(diff(lnp40days, lag=1), lag.max = 20)


# Partial Autocorrelograms

pacfPlot(diff(lnp10days, lag=1), lag.max = 20)
pacfPlot(diff(lnp20days, lag=1), lag.max = 20)
pacfPlot(diff(lnp30days, lag=1), lag.max = 20)
pacfPlot(diff(lnp40days, lag=1), lag.max = 20)


# Partial Ljung-Box Test

install.packages("FitAR")
library("FitAR", lib.loc="~/R/win-library/3.2")

LBQPlot(diff(lnp10days, lag=1), lag.max = 10)
LBQPlot(diff(lnp20days, lag=1), lag.max = 10)
LBQPlot(diff(lnp30days, lag=1), lag.max = 10)
LBQPlot(diff(lnp40days, lag=1), lag.max = 10)


LjungBoxTest(diff(lnp10days, lag=1), lag.max = 2)
LjungBoxTest(diff(lnp20days, lag=1), lag.max = 2)
LjungBoxTest(diff(lnp30days, lag=1), lag.max = 2)
LjungBoxTest(diff(lnp40days, lag=1), lag.max = 2)

# ACF PACF

par(mfrow=c(1,3))
acfPlot(diff(lnp10days, lag=1), lag.max = 20)
```

```
pacfPlot(diff(lnp10days, lag=1), lag.max = 20)
LBQPlot(diff(lnp10days, lag=1), lag.max = 20)

par(mfrow=c(1,3))
acfPlot(diff(lnp20days, lag=1), lag.max = 20)
pacfPlot(diff(lnp20days, lag=1), lag.max = 20)
LBQPlot(diff(lnp20days, lag=1), lag.max = 20)

par(mfrow=c(1,3))
acfPlot(diff(lnp30days, lag=1), lag.max = 20)
pacfPlot(diff(lnp30days, lag=1), lag.max = 20)
LBQPlot(diff(lnp30days, lag=1), lag.max = 20)

par(mfrow=c(1,3))
acfPlot(diff(lnp40days, lag=1), lag.max = 20)
pacfPlot(diff(lnp40days, lag=1), lag.max = 20)
LBQPlot(diff(lnp40days, lag=1), lag.max = 20)
```