Viriato Correa Pahim

A METHODOLOGY OF ORDER ESTIMATION AND COMBINATION OF EGARCH MODELS IN ECONOMETRICS

Dissertação submetida ao Programa de Pós-Graduação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia Elétrica

Orientador: Prof. José Carlos Moreira Bermudez, Ph.D.

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Pahim, Viriato Correa
A methodology of order estimation and combination of egarch models in econometrics / Viriato Correa Pahim ; orientador, José Carlos Moreira Bermudez, 2017.
176 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Engenharia Elétrica, Florianópolis, 2017.

Inclui referências.

1. Engenharia Elétrica. 2. Predição de volatilidade. 3. Ponderação de modelos. 4. Seleção de ordem. 5. Séries financeiras. I. Bermudez, José Carlos Moreira. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

Viriato Correa Pahim

A METHODOLOGY OF ORDER ESTIMATION AND COMBINATION OF EGARCH MODELS IN ECONOMETRICS

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.

Florianópolis, 11 de dezembro de 2017.

_____
Prof. Marcelo Lobo Heldwein, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Santa Catarina

Banca Examinadora:

_____
Prof. José Carlos Moreira Bermudez, Ph.D.
Orientador
Universidade Federal de Santa Catarina

_____
Prof. André Alves Portela, Ph.D.
Universidade Federal de Santa Catarina

_____
Prof. Raimes Moraes, Ph.D.
Universidade Federal de Santa Catarina

_____
Prof. Márcio Holsbach Costa, Dr.
Universidade Federal de Santa Catarina

AGRADECIMENTOS

A meus pais e minha esposa, pelo apoio incondicional, dentre tantas outras coisas.

Aos professores e demais integrantes do LPDS pela receptividade e contribuição em diversos momentos, em especial ao colega Tales Imbiriba por seu tempo e valiosos ensinamentos de ordem prática. Agradeço também a Guillaume Barrault por sua ajuda e boa vontade nos passos iniciais do trabalho, ainda que tenha tomado rumos diversos, e ao professor José Carlos Moreira Bermudez, pela orientação, paciência e amizade.

Aos membros da banca, que gentilmente aceitaram o convite para avaliar e contribuir com o trabalho.

Never think that lack of variability is stability. Don’t confuse lack of volatility with stability, ever.

Nassim Nicholas Taleb

ABSTRACT

Financial series volatility forecasting is an important area of investment, since volatility, a general term often defined as standard deviation or variance, is strongly linked to the subjective concept of risk, which investors seek to minimize. Amongst the models that are used to forecast volatility, the parametric family of autoregressive conditional heteroscedasticity (ARCH) is one of the most important, due to stationarity (constant unconditional moments) and simultaneous characteristic of reproducing time-varying conditional variance, an important property of financial series. From the ARCH family, we chose to work with EGARCH due to its asymmetrical response to gains and losses, a property found in many financial series. In this work we propose a methodology to be applied to the volatility forecasting problem, which addresses the issue of order selection and generalizes it, evaluating different order models and strategies that beyond model or order selection opt for model averaging, in which it is used a combined forecast calculated as a weighted average of each individual model's forecast. In this methodology we use synthetic instead of real data to prevent model deficiencies to compromise statistical significance of the conclusions provided by evaluations of the strategies. We apply such a framework to compare several model selection and averaging techniques, under several different orders EGARCH models and data generating processes. Amongst these techniques, the best one was based on Schwarz Information Criterion (SIC), whereas Akaike Information Criterion (AIC) led to worse performances. We devise that these results were strongly influenced by under than correct order models displaying the best performances, a discussed effect that is possible in small samples. We exploit this effect suggesting a generalized version of SIC, in which a hyperparameter is inserted into SIC calculation to raise complexity penalties given to higher order models, and show that model averaging using this generalized SIC outperformances the other strategies examined. Moreover, the methodology proposed has significant flexibility to evaluate several different models, orders and selection or averaging strategies, and is also naturally able to compare generalized SIC averaging for different values of the suggested hyperparameter, thus addressing the issue of its choice.

Keywords: EGARCH, volatility forecasting, model averaging, order selection, maximum likelihood estimation, financial series, information criteria, AIC, SIC.

RESUMO EXPANDIDO

Introdução

Séries financeiras, correspondentes a retornos obtidos ao longo do tempo através do investimento em um dado ativo, são teoricamente caracterizadas de acordo com a hipótese dos mercados eficientes. Essa hipótese limita a previsibilidade dos retornos propriamente ditos, porém não compromete a previsibilidade de suas variâncias ou desvios padrão. Tais quantidades são associadas ao termo volatilidade, neste trabalho definida como o desvio padrão condicional do retorno (condicionado à observação dos retornos passados). A predição de volatilidade é de extrema importância para o campo dos investimentos, pois está associada ao conceito subjetivo de risco. Tal importância decorre do usual interesse em otimizar a relação risco-retorno dos investimentos (maximização de retorno para um dado nível de risco, ou minimização do risco para um dado nível de retorno). Uma das aplicações mais imediatas é a precificação de derivativos (como opções de compra ou venda de ações), instrumentos cujo valor e cuja função estão intrinsecamente ligados ao risco e à volatilidade dos ativos subjacentes.

Dentre os modelos paramétricos de predição de volatilidade, destaca-se na literatura a família de modelos autorregressivos com heterocedasticidade condicional (ARCH). Esses modelos conciliam a propriedade de estacionariedade com a modelagem da variância condicional dos retornos como um parâmetro variante no tempo. Essa propriedade é intrínseca das séries financeiras e compatível com a estacionariedade, pois a última requer apenas que os momentos incondicionais sejam constantes. Dentre os modelos da família ARCH, este trabalho utiliza exclusivamente o modelo EGARCH. Dentre suas vantagens em relação aos demais figura a resposta assimétrica de volatilidades futuras a retornos positivos (ganhos) e negativos (perdas), propriedade comumente associada a ativos financeiros.

Objetivos

Os objetivos gerais do trabalho são: 1 – estudar sob o aspecto de processamento de sinais os modelos econométricos e critérios de informação utilizados para mensuração da sua adequação, analisando estatisticamente as formulações e propriedades correspondentes; 2 – contribuir para a solução do problema de predição de volatilidade focando na escolha de ordem do modelo e em métodos baseados na ponderação de predições usando diferentes modelos. Tais objetivos gerais se desdobram nos seguintes objetivos específicos: 1 – apresentar

três modelos paramétricos de volatilidade relacionados entre si e justificar a escolha por um desses modelos (EGARCH) para uso neste trabalho; 2 – estudar dois critérios de informação para estimativa de adequação de modelos e comparar tais critérios em relação às implicações de suas propriedades para uso no contexto de ponderação de modelos; 3 – propor uma metodologia para combinar predições baseadas em modelos de diferentes ordens, analisando várias estratégias de ponderação, comparando-as estatisticamente em termos dos erros médios quadráticos de predição; 4 – propor uma nova estratégia de ponderação de modelos e mostrar seu melhor desempenho e erros médios quadráticos de predição.

Metodologia

É proposta uma metodologia de avaliação de modelos de diferentes ordens, a qual inclui a seleção de uma ordem particular dentre as várias que venham a ser cogitadas, bem como de estratégias mais gerais que englobam a ponderação de todos os modelos correspondentes. A avaliação é focada na predição da volatilidade uma amostra à frente, e o erro médio quadrático de predição é a figura de mérito escolhida. Para a composição do arcabouço metodológico proposto, são analisados a teoria de ponderação de predições de modelos e dois critérios de informação. Esses critérios se mostram úteis para o cálculo de pesos utilizados para ponderar predições de modelos individuais e obter assim uma predição ponderada composta por vários modelos. A metodologia proposta utiliza dados sintéticos gerados pelos modelos escolhidos. Isso é feito para isolar dificuldades desses modelos em capturar toda a complexidade dos dados reais. Assim procura-se impedir que tais imprecisões comprometam as conclusões obtidas acerca dos desempenhos relativos de cada estratégia de predição sendo considerada. Argumenta-se que dessa maneira as conclusões podem ser inferidas com significância estatística devido ao número arbitrariamente alto de realizações consideradas na análise e da validade das premissas acerca do modelo gerador.

Resultados e Discussão

Sob a abordagem metodológica sugerida, são avaliados modelos EGARCH de diversas ordens com parâmetros estimados por máxima verossimilhança. Esses modelos são utilizados para predição da volatilidade uma amostra à frente em mercados acionários de diversos países. Além dos modelos individuais, são avaliadas algumas técnicas existentes para ponderação de predições feitas usando diferentes

modelos. Dentre as técnicas consideradas, a de melhor desempenho foi a estratégia de ponderação baseada no critério de informação de Schwarz (SIC). Estratégias correspondentes de predição que utilizaram o critério de informação de Akaike (AIC) obtiveram desempenho inferior. A partir da análise dos desempenhos de predição, é destacada a presença de um interessante efeito contraintuitivo, em que modelos de ordem inferior à do modelo gerador dos dados obtiveram desempenho superior ao modelo estimado com ordem correta. Tal efeito, possível em cenários de "pequena" (não assintótica) amostra, se mostrou de tamanha magnitude que motivou a sugestão e aplicação de um novo critério de informação. Esse critério é uma generalização empírica do critério de Schwarz, obtida através da introdução de um hiperparâmetro capaz de modular incrementos da penalização a modelos mais complexos.

Considerações Finais

O critério de Schwarz generalizado utilizado como ferramenta para ponderação de modelos foi capaz de levar a desempenhos superiores aos das demais estratégias avaliadas. A metodologia proposta permite a avaliação dessa estratégia em relação às demais sendo consideradas, e endereça a questão da escolha do valor do hiperparâmetro sugerido.

Palavras-chave: EGARCH, predição de volatilidade, ponderação de modelos, seleção de ordem, estimação por máxima verossimilhança, séries financeiras, critérios de informação, AIC, SIC.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AIC	Akaike Information Criterion
ARCH	Autoregressive Conditional Heteroscedasticity
ARMA	Autoregressive, Moving-Average
AR	Autoregressive
CAPM	Capital Asset Pricing Model
DAX	Germany's stock index
DGP	Data Generating Process
EGARCH	Exponential Generalized Autoregressive Conditional Heteroscedasticity
EMH	Efficient Market Hypothesis
FIR	Finite Impulse Response
FTSE 100	England's stock index
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GED	Generalized Error Distribution
IBOV	Ibovespa (Brazil's stock index)
IID	Independent and Identically Distributed
IIR	Infinite Impulse Response
KL	Kullback-Leibler
MA	Moving-Average
MIN	Minimum value
MAX	Maximum value
ML	Maximum Likelihood
MSE	Mean Square Error
N225	Nikkei 225 (Japan's stock index)
OPT	In the subscript of a variable, denotes the optimum value of such a variable
PDF	Probability Density Function
SIC	Schwarz Information Criterion
S&P 500	Standard & Poor's 500 (USA stock index)
TIC	Takeuchi Information Criterion

LIST OF SYMBOLS

t	Discrete time instant
p_t	Price of an asset at time t
R_t	Return obtained from an asset held from t-1 to t
$R_t(k)$	Return obtained from an asset held from t-k to t
ln	Natural logarithm function
r_t	Logarithmic return obtained from an asset held from t-1 to t
$r_t(k)$	Logarithmic return obtained from an asset held from t-k to t
$\mathbf{\Psi}_t$	Vector of all conditioning information previous to time t relevant to the process r_t
z_t	Normalized innovations: an iid, zero-mean, unit-variance process
μ_t	The probabilistic mean of r_t
σ_t	The probabilistic standard deviation of r_t
$\hat{\sigma}_t$	Estimate of σ_t
ε_t	Perturbation, which corresponds to the demeaned logarithmic return
DoF	Degrees of Freedom (of a Student t distribution)
$E[\cdot]$	The expectation operator (probabilistic mean)
$Cov[\cdot]$	Covariance matrix of two multivariate random variables
$\sum$	Summation (of a sequence)
Π	Product (of a sequence)
Q	Moving-Average order parameter of an ARCH type model
P	Autoregressive order parameter of a GARCH type model
C	Parameter of a returns mean model which corresponds to the offset in the mean equation
κ	Parameter of an ARCH type model which corresponds to the offset in the variance equation
A_j	Parameter of an ARCH type model which corresponds to the moving-average summation, j-th lag coefficient
L_j	Parameter of an EGARCH model which corresponds to the moving-average leverage summation, j-th lag coefficient

G_i	Parameter of a GARCH type model which corresponds to the autoregressive summation, *i*-th lag coefficient
$\lvert \cdot \rvert$	Absolute value of a scalar, or determinant of a matrix
$\triangleq$	An assignment that defines a variable (on the left hand side)
$\sim$	Left hand side distributed according to the right hand side pdf.
$\propto$	Denotes proportionality: right-hand and left-hand sides ratio is a constant.
$\boldsymbol{\theta}$	Parameter vector of a model
$(\cdot)^{\mathrm{T}}$	Transpose operator
$\boldsymbol{r}$	A vector of successive logarithmic returns
p,f,g	Probability density functions
$I(g;f)$	Kullback-Leibler divergence between g and f probability density functions
$l(\boldsymbol{\theta})$	Log-likelihood function
$\hat{\boldsymbol{\theta}}$	Maximum likelihood estimate of $\boldsymbol{\theta}$
$\hat{l}$	Maximum log-likelihood
$\boldsymbol{\theta}_0$	Optimum parameter vector (in the KL sense)
b	Bias of a random estimate (the difference from the expected value of the estimate to the true value trying to be estimated)
$J(\boldsymbol{\theta})$	Fischer information matrix, composed of the second order derivatives of the likelihood of a model in respect to its parameter vector.
$\mathbf{N}(\cdot)$	The Normal (or Gaussian) distribution, parametrized by a mean and a variance
N	The number of logarithmic returns available in a data sample (number of observations available in the data).
M	The number of models being considered
p_m	Number of parameters of the *m*-th model being considered
tr	Trace operator (sum of diagonal elements of a matrix)
I_R	Identity matrix with R rows
∇	Gradient operator
w_m	Weight given for the *m*-th model forecast (in a model averaging context)
$\mathbf{w}$	Column vector of M weights (each element corresponding to one of the M models)

$\hat{\boldsymbol{\sigma}}$	Column vector of M one sample ahead standard deviation estimates (each element corresponding to one of the *M* models)
e	Column vector of *M* one sample ahead standard deviation estimation errors (each element corresponding to one of the *M* models)
e_m	The *m*-th element of **e**
$\mathbf{1}_M$	Column vector with M unitary elements
$\tilde{\mathbf{w}}$	Column vector given by **w** with last element suppressed
$\tilde{\boldsymbol{\sigma}}$	Column vector given by $\hat{\boldsymbol{\sigma}}$ with last element suppressed
$\tilde{\mathbf{e}}$	Column vector given by **e** with last element suppressed
AIC	The value attained by AIC criterion. In the subscript of a variable, indicates the evaluation of such a variable under AIC model selection approach
SIC	The value attained by SIC criterion. In the subscript of a variable, indicates the evaluation of such a variable under SIC model selection approach
A-S	In the subscript of a variable, indicates the evaluation of such a variable under simple mean model averaging approach
A-AIC	In the subscript of a variable, indicates the evaluation of such a variable under (linear) AIC model averaging approach
A-SIC	In the subscript of a variable, indicates the evaluation of such a variable under (linear) SIC model averaging approach
A-E-AIC	In the subscript of a variable, indicates the evaluation of such a variable under exponential AIC model averaging approach
A-E-SIC	In the subscript of a variable, indicates the evaluation of such a variable under exponential SIC model averaging approach
λ	Hyperparameter according to which the generalized SIC criterion increases the complexity penalty
SIC(λ)	Generalized SIC criterion value
A-SIC(λ)	In the subscript of a variable, indicates the evaluation of such a variable under (linear) generalized SIC model averaging approach

MSE_{RL}	Relative MSE loss, a value that represents how much a corresponding MSE is higher (in a relative basis) than an optimum MSE

CONTENTS

1 INTRODUCTION

The stock markets have great importance to the economies of their countries, and the investors are evidently concerned in forecasting their movements to be able to maximize their gains, obtaining higher than average returns and minimizing potential losses, seeking lower risks and thus the optimization of the portfolio.

To extract information from the markets, there is a large set of studies and mathematical models that aims to process the signals involved – historic prices of stocks and respective volumes of negotiation – intended to forecast the variations of prices, or returns (GRIFFIOEN, 2003). The evaluation of whichever technique is done in a probabilistic framework, using models in which stock price oscillations are modeled as random variables.

Financial returns series can be viewed as discrete time signals, whose analysis is made seeking advantage for the investor or information to support the investor's decision making.

The analysis techniques used by many market agents aim at forecasting rises and falls through mathematical indicators and graphs extracted from past returns series. Nevertheless, the most accepted paradigm in the academic community is the efficient market hypothesis (EMH), according to which all information available reflects itself instantaneously in the prices of the assets, which incorporate the expectations of risk and return. This paradigm limits the possibility of forecasting future returns (CLARKE; JANDIK; MANDELKER, 2001; FAMA, 1970). The relation between price and the variables return and risk is the scope of pricing models, from which the CAPM (Capital Asset Pricing Model) is one of the most referenced (FAMA; FRENCH, 2004).

1.1 RETURNS OF FINANCIAL ASSETS

Given a financial asset, being p_t the price of the asset in discrete time t, the return obtained by an investor holding that asset from instant $t-1$ to instant t (during a unit time period), is given by (RUPPERT, 2011):

$$R_t = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1 \qquad (1.1)$$

In the above definition, the numerator is the obtained profit, and the denominator the value paid for the asset, such that the return is the ratio of the profit to the invested value (normally shown percentually), so that the quantity of assets will not matter.

In the case the asset is held from instant *t-k* to instant *t* the total return obtained after *k* unit time periods can be calculated from the past unit time returns as

$$R_t(k) = \frac{p_t - p_{t-k}}{p_{t-k}} = \frac{p_t}{p_{t-k}} - 1 = \prod_{i=t-k+1}^{t} \left(R_i + 1\right) - 1 \qquad (1.2)$$

Notice that in (1.2) we omit the dependency of $R_t(k)$ on the number k of periods whenever k=1 to simplify the notation. Thus, we shall denote $R_t(1)$ as simply R_t as done in (1.1).

The last term of the equations above comes from the fact that, according to (1.1), $R_t + 1$ equals the ratio of immediate successive prices (p_t and p_{t-1}) so that the product in (1.2) is a product of successive prices, which in turn reduces itself to the ratio of the prices of concern (in this case p_t and p_{t-k}).

This makes evident that $R_t + 1$ is a quantity of potential interest, whose successive multiplication allows one to combine successive returns. An implication is that this quantity is often directly defined itself as a return, such as done by Ruppert (2011), although this will not be done in the present work.

1.2 MOTIVATION

While the EMH limits the possibility of forecasting returns (asset prices variations), the same does not apply to the volatility of those returns, whose forecasting has been the subject of several studies. It is possible to infer from these studies that the volatility, defined either as variance or standard deviation of the returns, is substantially predictable (POON; GRANGER, 2003).

The concern in forecasting volatility arises from the fact that the investor makes his decisions weighting two factors: return and risk. The measure of volatility is deeply related to the risk of the investment, which can be defined either as the volatility itself, or through maximum levels of expected loss, for which the volatility will be a proxy. Having a good volatility estimate, the investor can select his assets portfolio according to the risk he is willing to take, optionally adjusting such risk

by taking positions in derivative instruments such as options of buying or selling a given asset (known as the subjacent asset of the derivative instrument). Volatility estimation of subjacent assets plays a major role in the adequate valuation of such derivative instruments.

1.3 PROBLEM DEFINITION

The problem of volatility forecasting associated to a financial asset is defined from a variable whose variation is subjected to estimation. One of most employed such variables (TSAY, 2005), and the one that will be used in this work, is the logarithmic return, defined as:

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) = \ln\left(R_t + 1\right) \tag{1.3}$$

Similar to the notation employed for the standard or non-logarithmic return defined previously, the logarithmic return for the *k* past unit time periods is given by:

$$r_t(k) = \ln\left(\frac{p_t}{p_{t-k}}\right) = \ln\left(R_t(k) + 1\right) \tag{1.4}$$

The logarithmic return is usually employed because of its mathematical properties, which are well suited for analysis tools and can be modeled as a zero-mean Gaussian random variable in the simplest scenarios (RUPPERT, 2011).

The use of logarithmic return instead of the standard return does not complicate the interpretation of the gains and losses. Firstly, these variables are convertible to one another through equations (1.3) and (1.4). Secondly they tend to have very close magnitudes for usual value ranges, as can be seen from the graph in Figure 1.1.

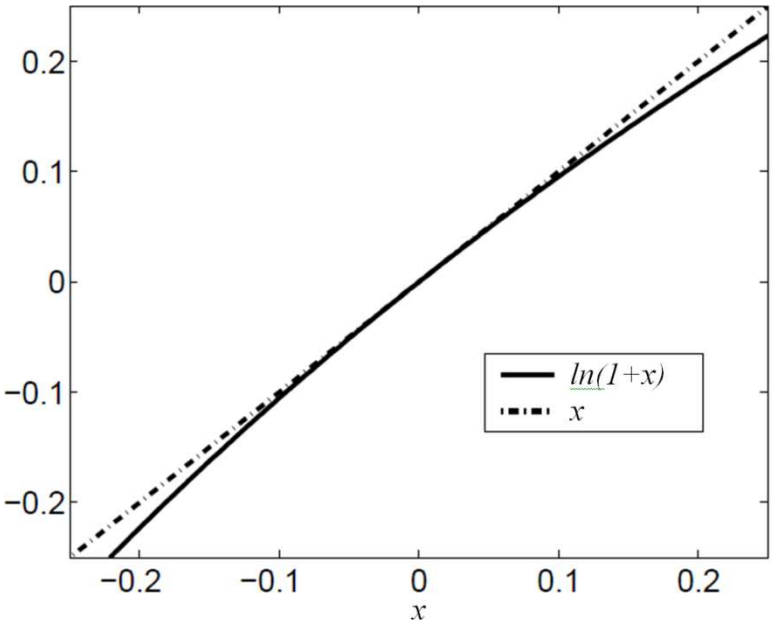

Figure 1.1 – Comparison of functions *ln(1+x)* and *x*. Source: (RUPPERT, 2011), adapted for notation.

Another important property of logarithmic returns is that multiple-period logarithmic returns are determined by adding single-period logarithmic returns. This is in contrast with the standard single-period returns, which must be multiplied to that end. This makes the expected value of the logarithmic return more directly relevant.

For the remaining of this work the term return will be used to refer to the logarithmic return, except when explicitly stated otherwise.

This work studies the forecasting of the standard deviation of the logarithmic return one sample ahead. Such standard deviation, conditioned on the past, is the definition of volatility used herein, to be more precisely stated in the next section. Since both return and volatility are discrete time stochastic signals, evaluation of their statistical properties in time (such as probabilistic moments, autocorrelation functions and stationarity) is well suited to signal processing techniques. This reasoning can thus be extended to volatility estimation, model fitting and residuals evaluation, amongst other signal processing techniques that can be used to formulate the volatility forecasting problem.

Nevertheless, because volatility forecasting frequently concerns finance and investment application areas, it is deeply studied in economics, more specifically in the econometrics field. Thus, the overlap of signal processing, econometrics and finance is extremely relevant in the analysis of financial time series (TSAY, 2005). Under such scope, this signal processing dissertation is devoted to the volatility forecasting application.

1.4 LITERATURE REVIEW

1.4.1 Returns probabilistic modeling – conditional moments

The return is generally modeled by a random variable r_t (TSAY, 2005) with conditional mean $\mu_t(\boldsymbol{\Psi}_t)$ and conditional standard deviation $\sigma_t(\boldsymbol{\Psi}_t)$ according to the following definitions:

$$r_t\left(\boldsymbol{\Psi}_t\right) = \mu_t\left(\boldsymbol{\Psi}_t\right) + \varepsilon_t\left(\boldsymbol{\Psi}_t\right) \tag{1.5}$$

$$\mu_t\left(\boldsymbol{\Psi}_t\right) = E\left[r_t \mid \boldsymbol{\Psi}_t\right] \tag{1.6}$$

$$\sigma_t{}^2\left(\boldsymbol{\Psi}_t\right) = E\left[\left(r_t - \mu_t\right)^2 \mid \boldsymbol{\Psi}_t\right] \tag{1.7}$$

$$\boldsymbol{\Psi}_t = [\mu_{t-1}, \sigma_{t-1}, z_{t-1}, \mu_{t-2}, \sigma_{t-2}, z_{t-2}, \ldots] \tag{1.8}$$

$$\varepsilon_t\left(\boldsymbol{\Psi}_t\right) = \sigma_t\left(\boldsymbol{\Psi}_t\right) \cdot z_t \tag{1.9}$$

where z_t (normalized innovations) are samples drawn from an independent and identically distributed (iid) stationary random process with zero mean and unit variance, ε_t is the corresponding perturbation (normalized innovation weighted by a conditional standard deviation) and $\boldsymbol{\Psi}_t$ is a vector containing all conditioning information previous to instant t that is relevant to the random process of returns r_t. The conditioning on $\boldsymbol{\Psi}_t$ will be omitted from now on to simplify the notation. It will be explicitly shown only when necessary to avoid confusion. Then, (1.5) and (1.9) yield

$$r_t = \mu_t + \varepsilon_t = \mu_t + \sigma_t \cdot z_t \tag{1.10}$$

The general model is multidimensional since in practice there are several assets in the market, so that r_t and μ_t are vectors of returns and its

conditional means, respectively, where each element refers to one particular asset. Since the assets bear correlation, σ_t is generalized to a covariance matrix. This higher dimensional model is employed when one is interested in applications such as portfolio management, where the overall performance of any given portfolio will depend on its allocations in individual assets, and thus the correlations between their returns must be considered (TSAY, 2005).

1.4.2 Stylized facts from financial series

Some properties of financial series were observed with such a frequency that they are normally referenced as “stylized facts” in the literature (FRANCQ; ZAKOIAN, 2010). Stylized facts are statistical attributes found in real data that, because of either their nature, frequency of observation or magnitude, are considered to be inherent to most financial series, and thus it is desirable that financial models be able to reproduce them.

A very important stylized fact, that has the theoretical support of the EMH, is the low autocorrelation of the returns. Linear parametric models usually specify the conditional mean as a constant added to a weighted sum of past returns and perturbations, as seen for example in the third chapter, page 100, of Tsay (2005). Given the cited stylized fact, the coefficients of the linear combination of past returns and perturbations are normally found to be close to zero, so that the conditional mean is often approximated simply by a constant (CLEMENTS; HENDRY, 1998; HANSEN; LUNDE, 2005).

The conditional mean of the returns, or expected value of the return of an asset, has its modeling as the scope of econometrical models of asset pricing. The Capital Asset Pricing Model (CAPM) is the most well-known among the existing models (FAMA; FRENCH, 2004).

This work focuses on the forecasting of the conditional volatility (defined as standard deviation) σ_t.

The modeling of conditional volatility has received much attention from the literature due to the following reasons:

a) High practical interest, since the conditional volatility of an asset is deeply linked, for instance, to the risk variable from which derivatives (such as calls and puts – options of buying or selling a stock, respectively) pricing can be highlighted (DUAN et al., 2006; FRANCQ; ZAKOIAN, 2010).

b) Theoretical support, as asset conditional volatility forecasting does not violate EMH. Moreover, the possibility of volatility forecasting

is another stylized fact of the financial series (POON; GRANGER, 2003).

1.4.3 The ARCH, GARCH and EGARCH conditional volatility models

Among parametric, semi-parametric and non-parametric models proposed to volatility forecasting, the first are the most studied ones. They include also the models whose application has led to the best practical results, in particular conditional volatility models. These started with the so-called ARCH (autoregressive conditional heteroscedasticity) model, proposed by Engle (1982) in his seminal paper. In Engle (1982) the conditional variance is modeled by a constant plus a linear combination of past perturbations whose weights are parameters of the model.

The ARCH model made possible to reproduce two other stylized facts of financial series (BERA; HIGGINS, 1993):

a) Time-varying conditional volatility, which is generated by the stationary ARCH model (unconditional statistical moments constant in time)

b) The phenomenon called “volatility clustering”, by which high (low) volatility sub-periods occur during a crisis (tranquility) epoch of a financial market. This stylized fact is also described as a statistical dependence of successive returns or, more specifically, as the existence of significant autocorrelation of squared returns.

The success of the ARCH model motivated the development of more general and sophisticated models, such as the GARCH model (BOLLERSLEV, 1986) and the EGARCH model (NELSON, 1991). These three models will be described in detail in the next chapter, and the work of this dissertation concentrates on the EGARCH model.

The GARCH (general autoregressive conditional heteroscedasticity) model generalized the ARCH, by introducing a feedback contribution from past variances. This modification allowed the modeling of financial series using less parameters (parsimony) (VRONTOS; DELLAPORTAS; POLITIS, 2000).

The EGARCH (exponential general autoregressive conditional heteroscedasticity) model cannot be considered a generalization of the ARCH or GARCH models. It is better viewed as a sophistication of these models which allows (at the cost of a larger number of parameters) the modeling of a stylized fact of financial series called “leveraging”. Leveraging occurs when negative returns (losses) tend to increase

volatilities more than same magnitude positive returns (gains) do (LI; LI, 2015). In other words, for given magnitudes of past returns, higher (lower) volatilities are expected in the future if those past returns were negative (positive).

After the introduction of the GARCH model, few studies continued adopting the ARCH model, which is presented here only for didactic and historical reasons, since it is the simplest (and first) of the models in the conditional heteroscedasticity family of models.

Even after the development of several models more sophisticated than the GARCH (EGARCH being one example), GARCH remains as one of the most popular models in its family, with relatively fewer studies being available using the EGARCH model. Moreover, most studies are limited to first order models (RUPPERT, 2011; ZHANG et al., 2013).

1.4.4 Order selection and information criteria

The choice of the parametric model to be used for conditional volatility prediction is not a trivial problem. One common approach is to choose a specific type of model, and then look for the order (to which the number of parameters is proportional) that leads to the best explanation of past observations. A second possible approach is to estimate the parameters of different types of models, and select one with the best prediction performance. The performances of different models are usually compared using some information criterion. In this work we follow the former approach, as it is the least complex of the two. Moreover, an efficient order selection methodology derived for a single model approach can be used for each of the distinct models in a multi-model approach.

Given a parametric model, its parameters are usually estimated using maximum likelihood (TSAY, 2005). However, the choice of the model order remains a less trivial question.

The simplest approach for order selection is to restrict the model to the lowest possible order (unitary). Although seemingly simplistic, this approach is frequently followed (FRANCQ; ZAKOIAN, 2010; WEI-MING; ZHONG-FU, 2012). However, this simplistic solution can hardly be assumed adequate without a more systematic exploration of the possibilities of employing higher order models. In this work we propose a systematic methodology to explore such possibilities, given that a type of model has been specified.

The most popular information criteria for order selection are the AIC (Akaike Information Criterion) and the SIC (Schwarz Information Criterion) (KONISHI; KITAGAWA, 2008). The former estimates the Kullback-Leibler divergence between each estimated model and the actual data generating process (DGP) (BURNHAM; ANDERSON, 2004). The latter estimates the probability of each estimated model to be the real DGP (KUHA, 2004), under the assumption that one of the estimated models is indeed the true DGP. Thus, choosing the model based on the AIC criterion corresponds to opt for the closest model to the true one in the Kullback-Leibler sense. Basing the model choice on the SIC criterion corresponds to opt for the model most likely to be the true DGP. Both criteria will be described in more detail in Chapter 3.

1.4.5 Forecasts weighting

An alternative approach for model selection besides selecting one among various alternative models is to combine all the available models or a subset of them and generate a forecast that corresponds to a weighted combination of the individual forecasts, as originally suggested by Bates and Granger (1969). This approach has been successfully employed to yield a lower forecast variance than that obtained by each of the individual model forecasts in the mean square error (MSE) sense (CHENG; ING; YU, 2015; JAMES; CHAN, 2011).

Forecast weighting (best known as model averaging) raises two important new issues that must be addressed (TIMMERMANN, 2006):

a) When averaging a subset of all the available models, a criterion for choosing such a subset must be defined.

b) When using a weighted average of a set of models, a criterion must be defined to determine the individual weights.

We will show in Chapter 4 that information criteria such as AIC and SIC are natural candidates to address these issues.

1.5 OBJECTIVES AND WORK OUTLINE

This work aims the following general objectives:

1) Discuss econometric models and information criteria in a signal processing framework, analyzing corresponding formulations and properties using a statistical signal processing approach.

2) Contribute to the solution of the volatility forecasting problem by focusing in the question of model order choice, under the general framework of model averaging.

The general objectives are further depicted into the following specific objectives:

1) Present three related parametric volatility models, their corresponding formulations and properties, and justify the choice for one of the models to be used in the work.
2) Present two information criteria to be used, their corresponding formulations, properties, differences and implications of their use as tools for model averaging.
3) Propose a methodology for combining different order models. Describe several model averaging strategies and, under the methodology proposed, statistically compare those strategies in terms of forecasting MSE.
4) Propose a new model averaging strategy and display its overperformance in terms of forecasting MSE.

Chapter 2 presents three of the most popular models (ARCH, GARCH and EGARCH) for autoregressive conditional heteroscedasticity (ARCH). The models are presented in increasing order of complexity, which corresponds also to the chronological order in which they have been proposed. The option of this work to concentrate on the EGARCH model is then justified. Finally, the application of the maximum likelihood approach to the estimation of the EGARCH parameters is described.

Chapter 3 presents the fundamentals and the formulations of the AIC and SIC information criteria, along with their application to the selection of competing models.

Chapter 4 describes the model averaging technique and details its application to EGARCH models of different orders. We propose the use of information criteria to support the choice of weights for the individual models, an approach that generalizes the special case where unselected models are assigned a weight equal to zero.

The majority of the studies in finances lead to performance results that vary widely, depending on the provided information. It is conjectured that these discrepancies are strongly due to the use of real data during the study. The availability of an amount of real data that is informationally insufficient to draw statistical conclusions leads to

results that are highly dependent on the specifics (assets, period, periodicity) of the data sample actually employed in the study. As examples we can reference Ezzat (2012) and Balaban (2004). The compromises of real data usage are further discussed in Chapter 4.

This work proposes a new methodology for the study of the one step ahead forecasting of the logarithmic return volatility. It is assumed, based on strong supporting evidence from the literature, that the behavior of the financial series follows an EGARCH volatility model. Under this assumption, a detailed study is realized which allows to draw statistical inferences regarding the performances of the forecasting approaches being compared. This approach isolates the effects of model inaccuracies from the choice of the most appropriate model for the data generated using a widely accepted model family. The results of the study are new methodologies for model order selection, and for the weighted averaging of different models. The chapter concludes with a proposal for model averaging using a modified version of the SIC to calculate the weights assigned to each model.

Chapter 5 presents a performance comparison of different approaches of model selection and averaging, using synthetic data generated by the EGARCH model.

Lastly, Chapter 6 exposes the conclusions of this work and the obtained results, and proposes future studies with the potential to wide or deepen the scope here presented.

2 MODELS

In this chapter we present three conditionally heteroscedastic models (ARCH) for the variance of time series. We present the models in chronological order of proposition, which coincides with the increasing order of complexity – ARCH, GARCH and EGARCH. These models can be viewed as successive sophistications of one another, enabling a contextualization until the point in which we decide to work exclusively with the EGARCH in the following chapters.

As was mentioned in the previous chapter, in the returns formulation (1.10) it is reasonable to approximate the conditional mean with an unconditional (constant) mean, because of the theoretical support granted by the efficient market hypothesis. Subtracting the mean from a return series yields:

$$r_t - \mu_t \approx \varepsilon_t = \sigma_t \cdot z_t \tag{2.1}$$

The above equation justifies proceeding the work with exclusive focus on forecasting the one step ahead standard deviation σ_t, since it corresponds to the volatility (defined as standard deviation) of the final signal of interest, the logarithmic return.

2.1 ARCH MODEL

The ARCH (Autoregressive Conditional Heteroscedastic) model expresses the variance of the logarithmic return of a given asset as a linear combination of past perturbations, added to a constant:

$$\sigma_t^2 = \kappa + \sum_{j=1}^{Q} A_j \cdot \varepsilon_{t-j}^2 = \kappa + \sum_{j=1}^{Q} A_j \cdot \sigma_{t-j}^2 \cdot z_{t-j}^2 \tag{2.2}$$

According to the ARCH model, the variance at a given instant, conditioned to the perturbations occurred at past instants, has a functional dependence on those perturbations. Hence, the model considers the statistical properties (variance and other moments in particular) of the random variable logarithmic return (conditioned to past observations) to be time-varying. This justifies the denomination of conditionally heteroscedastic assigned to the model.

$$Var\left(r_t \left| \varepsilon_{t-1}, \varepsilon_{t-2}, ... \right.\right) = \sigma_t^2 \left(\varepsilon_{t-1}, \varepsilon_{t-2}, ...\right) \tag{2.3}$$

The time-varying nature of the statistical moments of the return in (2.3) explicitly states that they depend on the values of the conditioning past perturbations (FRANCQ; ZAKOIAN, 2010).

To emphasize this important property, we can compare the expression of σ_t^2 with that of an autoregressive (AR) process of order Q, which could be described by:

$$\alpha_t = \kappa + \sum_{j=1}^{Q} A_j \cdot \alpha_{t-j} + z_t \tag{2.4}$$

The mean of the AR process conditioned to the past values is given by:

$$E\left(\alpha_t \mid \alpha_{t-1}, \ldots, \alpha_{t-Q}\right) = \kappa + \sum_{j=1}^{Q} A_j \cdot \alpha_{t-j} \tag{2.5}$$

This conditional mean is clearly time dependent. However, if the characteristic roots of the system defined by (2.4) are less than one in magnitude, the process is asymptotically stationary (MANOLAKIS; INGLE; KOGON, 2000), with an unconditional asymptotic mean given by:

$$E\left(\alpha_t\right) = \frac{\kappa}{1 - \sum_{j=1}^{Q} A_j} \tag{2.6}$$

Property (2.6) obviously applies to the ARCH, showing that the unconditional variance is asymptotically stationary with mean given by (2.6) if the coefficients A_j satisfy the stability conditions to be determined later in this section (RUPPERT, 2011).

The comparison between the signal σ_t^2 and the AR process was discussed to reinforce the duality between time-varying conditional mean and time-invariant unconditional mean. Since the ARCH models a signal $\varepsilon_t = \sigma_t \cdot z_t$ (according to (2.1)), the properties of the first order moment of σ_t^2 affect the second order moment (variance) of the ARCH

signal ε_t and, therefore, of the logarithmic return. Hence, differently from the AR process, the signal ARCH ε_t has a time-invariant unconditional variance (instead of mean), while the variance of the next sample conditioned on the past ones (the conditional variance) is time-varying according to equations (2.2) and (2.3). This property allows for the desirable nature of the ARCH model to be stationary, which is advantageous from an analytical point of view, while simultaneously displaying conditional heteroscedasticity, which is observed in financial data.

The stylized fact called volatility clustering, according to which high (low) magnitude returns tend to be followed by high (low) magnitude returns, is adequately reproduced by the ARCH family models due to the property of time-varying conditional variance discussed above.

One should note from (2.1) that the variance properties (either conditional or unconditional) of ARCH signals ε_t are reproduced in the final signal of interest r_t, as they differ only by a mean that is assumed to be constant.

For the variance to be non-negative, strictly stationary, and for the first and second order moments to be finite, the following restrictions on the model parameters values are sufficient, as demonstrated by Francq and Zakoian (2010) for the more general case of the GARCH model (to be also presented in the next section):

$$\sum_{i=1}^{Q} A_i < 1 \tag{2.7}$$

$$\kappa > 0; A_1, \ldots, A_Q \geq 0 \tag{2.8}$$

An ARCH model of order Q is denoted by ARCH(Q). The unconditional variance $E(\sigma_t^2)$ can be obtained by taking the expected value of equation (2.2) and using the property of stationarity (under the premise of the above restrictions being satisfied):

$$E\left(\sigma_t^2\right) = E\left(\kappa + \sum_{j=1}^{Q} A_j \cdot \sigma_{t-j}^2 \cdot z_{t-j}^2\right)$$
$$= \kappa + \sum_{j=1}^{Q} A_j \cdot E\left(\sigma_{t-j}^2\right) \quad (2.9)$$
$$= \frac{\kappa}{1 - \sum_{j=1}^{Q} A_j}$$

2.2 GARCH MODEL (GENERALIZED ARCH)

The GARCH model (BOLLERSLEV, 1986) generalizes the ARCH (GARCH means Generalized Autoregressive Conditional Heteroscedasticity), by adding extra autoregressive terms to the ARCH variance model to account for contributions of variance values at past time instants. The GARCH model is given by:

$$\sigma_t^2 = \kappa + \sum_{i=1}^{P} G_i \cdot \sigma_{t-i}^2 + \sum_{j=1}^{Q} A_j \cdot \varepsilon_{t-j}^2 \quad (2.10)$$

Note that the order of the GARCH model is defined by two parameters (P and Q), and not by a single parameter (Q) as in the ARCH model, which now becomes a particular case of GARCH for P = 0. The GARCH model is then denoted by GARCH(P,Q).

One important advantage of GARCH over ARCH is its capacity to reproduce financial series with a significantly smaller number of parameters (parsimony). This is due to the feedback added to the variance model through the terms ($G_i \cdot \sigma_{t-i}^2$). Moreover, this feedback makes GARCH able to model more persistent volatilities (long periods of higher or lower than average volatility), which are commonly found in real financial series (RUPPERT, 2011).

Indeed, since the introduction of GARCH, the ARCH model has almost ceased to be used (HANSEN; LUNDE, 2005). Despite the fact that several more sophisticated models have been formulated among the conditional heteroscedasticity family (POON; GRANGER, 2003; WEI-MING; ZHONG-FU, 2012), GARCH(1,1) is still one of the most found ARCH type models in the literature

The following parameter restrictions are needed to ensure variance non-negativity, stationarity and finite magnitude moments in a GARCH model (FRANCQ; ZAKOIAN, 2010).

$$\sum_{i=1}^{Q} A_i + \sum_{i=1}^{P} G_i < 1 \tag{2.11}$$

$$\kappa > 0, A_1, \ldots, A_Q, G_1, \ldots, G_P \geq 0 \tag{2.12}$$

These restrictions were particularized for the ARCH to (2.7) and (2.8) in Section 2.1.

Similar to the ARCH, the unconditional variance $E(\sigma_t^2)$ can be obtained by taking the expectation of equation (2.10) and using the stationarity property (which follows from the assumed premise that the restrictions above are satisfied):

$$\begin{aligned} E\left(\sigma_t^2\right) &= E\left(\kappa + \sum_{i=1}^{P} G_i \cdot \sigma_{t-i}^2 + \sum_{j=1}^{Q} A_j \cdot \sigma_{t-j}^2 \cdot z_{t-j}^2\right) \\ &= \kappa + \sum_{i=1}^{P} G_i \cdot E\left(\sigma_{t-i}^2\right) + \sum_{j=1}^{Q} A_j \cdot E\left(\sigma_{t-j}^2\right) \\ &= \frac{\kappa}{1 - \sum_{i=1}^{P} G_i - \sum_{j=1}^{Q} A_j} \end{aligned} \tag{2.13}$$

2.3 EGARCH MODEL (EXPONENTIAL GENERALIZED ARCH)

The EGARCH model (NELSON, 1991), which means Exponential Generalized Autoregressive Conditional Heteroscedasticity, is given by the expression:

$$\begin{aligned} \ln\left(\sigma_t^2\right) &= \kappa + \sum_{i=1}^{P} G_i \cdot \ln\left(\sigma_{t-i}^2\right) \\ &+ \sum_{j=1}^{Q} A_j \cdot \left(\left|z_{t-j}\right| - E\left(\left|z_{t-j}\right|\right)\right) + \sum_{j=1}^{Q} L_j \cdot z_{t-j} \end{aligned} \tag{2.14}$$

This model is not a generalization of GARCH, since the last one cannot be obtained from it by setting a subset of parameters to certain values (to zero for example). EGARCH can be considered to be a more sophisticated version of the ARCH family. It does not model the variance directly, but its logarithm. Moreover, it considers the positive and negative past innovations differently through two different summations. These properties and their consequences will be further discussed in the following.

In this work we shall use mainly the standard Gaussian as the distribution for the normalized innovations z_t. In this case, $E\left(\left|z_{t-j}\right|\right)=\sqrt{2/\pi}$. The EGARCH model can be used with other distributions (with zero mean and unit variance) for z_t, such as the Student t or the generalized error distribution GED (NELSON, 1991). In those cases the appropriate value for $E\left(\left|z_{t-j}\right|\right)$ must be set.

The choice of the distribution for z_t is related to the excess kurtosis (fatter than normal tails) normally found in financial series (RUPPERT, 2011). Using a Gaussian distribution for the normalized innovations still allows for the EGARCH signal ε_t to have fat tails, which is also true for other ARCH type signals, due to the dynamics of the model (FRANCQ; ZAKOIAN, 2010). However, the amount of excess kurtosis that can be obtained using the Gaussian distribution is limited. Hence, depending on the kurtosis of the financial series being modeled, higher kurtosis distributions for z_t might be needed (NELSON, 1991), such as the Student t distribution for example. The drawback of such a choice is the inclusion of extra parameters to be estimated (degrees of freedom of the t distribution, for example), which can lead to cumbersome convergence or out of sample performance losses due to overfitting (when comparing to a lower number of parameters choice). There are examples in the literature where choosing the Gaussian distribution (to model z_t) has led to a better performance than using the Student t distribution (LI; HUANG; ZHANG, 2013). However, there are also examples of the opposite, such as in Su (2010). In this work we will use mainly the Gaussian distribution as our reference scenario, but some examples will be repeated with Student t distribution for the normalized innovations so that the issues raised in this discussion can be further addressed.

Some observations can be made about expression (2.14), regarding to its direct comparison with the previous models. Firstly, the autoregressive terms (P terms) of the first summation preserve the

capacity to model stylized facts such as volatility clustering and persistent volatility. Secondly, (the logarithm of) the variance depends directly on the normalized innovations (z_t), instead of on the past perturbations (ε_t) used in the ARCH and GARCH models (in the Q terms summations). Since the normalized innovations are independent and identically distributed (iid), stationarity conditions are simplified. To clarify the last point, it is useful to see the EGARCH as a model that defines the process logarithm of variance as the output of a linear ARMA system, with the following inputs: 1 – the signal of normalized innovations and 2 – the modulus of this same signal subtracted from its statistical mean. Obviously, both inputs are iid and have zero mean. Thus, none of the Q-terms summations can make the system non-stationary, as they depend only on a finite number of past normalized innovations z_t (iid by construction). That does not happen in ARCH and GARCH, where the Q-terms summations depend on past perturbations ε_t, which are the normalized innovations multiplied by the standard deviation, thus incurring in a non-trivial variance feedback. Finally, the summation with P terms is a linear autoregressive component (dependence of the output on its past values). Hence, it is sufficient to restrict the system characteristic roots to lie inside the unit circle for stability:

$$1-\sum_{i=1}^{P} G_i \cdot \mathbb{Z}^{-i} = 0 \Rightarrow \left|\mathbb{Z}\right| < 1 \tag{2.15}$$

2.4 THE CHOICE OF THE EGARCH MODEL

In this work, we opted to work exclusively with the EGARCH model for two reasons. Firstly, the use of the logarithm automatically assures variance non-negativity. Hence, there is no need for additional parameter restrictions, differently from the previous models that demanded such restrictions through equations (2.8) or (2.12). It is interesting to note that the two restrictions required by the GARCH model (and therefore ARCH as well) to guarantee parameter non-negativity (2.12) and stationarity (2.11) imply the EGARCH stationarity restriction (2.15). However, the converse is not true (FRANCQ; ZAKOIAN, 2010). This reinforces the fact that the variance feedback introduced in GARCH through the past perturbations ε_t (instead of the past innovations z_t in the Q terms summations of EGARCH) imposes a more severe restriction to the GARCH model parameters to allow for stationarity.

The second reason for the adoption of EGARCH in this work is its desirable property of reproducing the leverage effect, another important stylized fact of financial series. Leveraging occurs when negative returns (losses) are followed by higher variances than positive returns (gains) of same magnitude (NELSON, 1991). Since volatility can be considered as a measure of uncertainty, the occurrence of leveraging means that losses of a given magnitude indicate an increased future uncertainty, or risk, when compared to the occurrence of gains of same magnitude.

The modeling of this marginal impact of a past return on the current variance can be taken from (2.14) and the fact that $E\left(\left|z_{t-j}\right|\right)$ is constant (a property of the iid distribution of the normalized innovations). Then, the dependence on past returns is determined by the terms:

$$A_j \cdot \left|z_{t-j}\right| + L_j \cdot z_{t-j} \tag{2.16}$$

It is clear from (2.16) that the marginal impact of past returns will depend on the sign of z_{t-j}. For a negative value (loss), (2.16) becomes

$$\left|z_{t-j}\right| \cdot \left(A_j - L_j\right), z_{t-j} \leq 0 \tag{2.17a}$$

while for a positive value (gain), it becomes:

$$\left|z_{t-j}\right| \cdot \left(A_j + L_j\right), z_{t-j} \geq 0 \tag{2.17b}$$

To better understand (2.17a) and (2.17b) and their impacts on model behavior, it is worth mentioning that the coefficients A_j tend to be positive. This consideration is supported by the property frequently found in financial series that larger magnitude innovations imply higher future volatilities (FRANCQ; ZAKOIAN, 2010). Then, assuming all A_j positive and any magnitude of the past normalized innovation, the occurrence of the leverage effect will require negative L_j's. This will lead to an increase in the magnitude of (2.17a) (losses) and to a reduction in the magnitude of (2.17b) (gains). Hence, losses are followed by higher variances, as expected from the model.

The discussion above indicates that, in general, we can expect positive A_j's, positive G_i's due to volatility clustering and negative L_j's

due to the leverage effect. The implications of the stylized facts on the signs of the coefficients do not need to be imposed in the estimation process, especially when using higher orders models. In fact, these signs can be different from what is expected from this simple discussion due to the dynamics of different lags involved. However, it is important to impose the stationarity and stability restriction (2.15).

The desirable ability of the EGARCH model to respond differently to magnitude and sign of past innovations to model the important leverage effect, usually comes at the cost of an increase in number of parameters when compared to GARCH, for instance. Considering models of the same order (P, Q), an increase in the value of Q will increase the number of EGARCH parameters twice as much as the number of GARCH parameters due to the extra summation.

2.5 MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

In the estimation of ARCH family models parameters, the maximum likelihood technique is almost ubiquitous (POON; GRANGER, 2003; STRAUMANN; MIKOSCH, 2006). Therefore, it is also used exclusively in this work. In this section, we describe the application of the maximum likelihood technique for the estimation of EGARCH parameters with arbitrary orders and standard Gaussian distributed innovations (z_t). Although maximum likelihood estimation is a well known parameter estimation technique, its application to the estimation of the EGARCH model parameters has some specifics that are worth detailing.

The vector $\boldsymbol{\theta}$ of parameters to be estimated can be readily obtained from (2.14) as

$$\boldsymbol{\theta} = \left[\kappa, A_1, \cdots, A_Q, L_1, \cdots, L_Q, G_1, \cdots, G_P \right]^T \tag{2.18}$$

Consider a series of T observed returns r_t, $t = 1, \ldots, T$, collected in a vector

$$\boldsymbol{r} = \left[r_1, \ldots, r_T \right]^T \tag{2.19}$$

The log-likelihood function of $\boldsymbol{\theta}$ is given by

$$l(\boldsymbol{\theta}) = \ln\left[\, p(\mathbf{r}|\boldsymbol{\theta})\right] = \ln\left[\prod_{t=1}^{T} p\left(r_t \,|\boldsymbol{\theta}, r_1, \ldots, r_{t-1}\right)\right] \\ = \sum_{t=1}^{T} \ln\left[\, p\left(r_t \,|\boldsymbol{\theta}, r_1, \ldots, r_{t-1}\right)\right] \tag{2.20}$$

where p denotes a conditional probability density function, and the dependence of future returns likelihoods on past returns is explicitly shown. This is because, due to the feedback nature of (2.14), the likelihood function associated with $\boldsymbol{\theta}$ depends also on the values of past and present standard deviations, past normalized innovations z_t and returns, of which only the returns can be assumed to be observed.

From (2.1), we can calculate z_t as $z_t = r_t/\sigma_t$. However, although the returns r_t are observed, the conditional standard deviations σ_t are not and need to be determined. This is accomplished recursively, from the EGARCH model equation (2.14), for a given set of assumed model parameters and past values of σ_t and z_t.

Such a recursive approach, however, requires the initialization of variables. A good initialization for σ_t can be obtained from the unconditional expected value of (2.14). Taking the expectation of (2.14) yields

$$E\left[\ln\left(\sigma_t^2\right)\right] = \kappa + \sum_{i=1}^{P} G_i \cdot E\left[\ln\left(\sigma_{t-i}^2\right)\right] \\ + E\left[\sum_{j=1}^{Q} A_j \cdot \left(\left|z_{t-j}\right| - E\left(\left|z_{t-j}\right|\right)\right) + \sum_{j=1}^{Q} L_j \cdot z_{t-j}\right] \\ = \kappa + \sum_{i=1}^{P} G_i \cdot E\left[\ln\left(\sigma_{t-i}^2\right)\right] = \frac{\kappa}{1 - \sum_{i=1}^{P} G_i} \tag{2.21}$$

where the last line has been obtained by noting that the innovations z_t and $\left|z_{t-j}\right| - E\left(\left|z_{t-j}\right|\right)$ are zero-mean, and the final result has been obtained by solving the equation for $E\left[\ln\left(\sigma_t^2\right)\right]$ using the stationarity of the unconditional variance. Using this expression, the values for σ_t, t

= 1, …, *T*, are determined recursively from the observed returns and assumed model parameters.

Particularizing the expression of $l(\boldsymbol{\theta})$ to Gaussian density function yields

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \ln\left[\frac{e^{-r_t^2 / 2\cdot\sigma_t^2}}{\sigma_t \cdot \sqrt{2\cdot\pi}} \right] \\ &= \sum_{t=1}^{T} \left[-\frac{r_t^2}{2\cdot\sigma_t^2} - \ln(\sigma_t) - \frac{1}{2}\cdot\ln(2\cdot\pi) \right] \\ &= -\frac{T}{2}\cdot\ln(2\cdot\pi) - \sum_{t=1}^{T}\left[\frac{r_t^2}{2\cdot\sigma_t^2} + \ln(\sigma_t) \right] \end{aligned} \tag{2.22}$$

which is now clearly a function of the observed returns and the determined conditional variances.

It should be noted that the log-likelihood expressions above do not display direct dependence on the model parameters nor on the model dynamics themselves. This is because (2.20) and (2.22) are valid to any ARCH type model (and other families as well) when Gaussian normalized innovations are considered. The dependence on the specific model exists through the values of σ_t, t = 1, …, *T*, whose calculations involve both the model dynamics and its parameters, as previously explained.

To summarize, we now state the steps required to estimate the parameter vector using maximum likelihood. It is assumed that the solution of the maximum likelihood estimation is obtained using an optimization routine. The vector $\boldsymbol{\theta}$ that maximizes $l(\boldsymbol{\theta})$ is then the maximum likelihood estimate, which we here denote by $\hat{\boldsymbol{\theta}}$.

1) Initialize the parameter vector estimate $\hat{\boldsymbol{\theta}}$.
2) Initialize $\ln\left(\sigma_{t-i}^2\right) = \kappa \Big/ 1 - \sum_{i=1}^{P} G_i$ for $t \le i$ according to (2.21).
3) From the vector $\boldsymbol{r}$ of observed returns r_t, t = 1, …, *T*, calculate recursively the corresponding values of σ_t, t = 1, …,

T, using the EGARCH volatility model (2.14). For Gaussian normalized innovations set $E\left(\left|z_{t-j}\right|\right)=\sqrt{2/\pi}$.

4) Use (2.22) to determine $l(\hat{\boldsymbol{\theta}})$.
5) Test the stopping criterion of the optimization routine and return to step 2 if required to update $\hat{\boldsymbol{\theta}}$.

Note that it is strongly recommended that the stability restriction(s) ((2.15) in EGARCH case) be imposed to the parameter vector iteratively estimated during optimization. Also, it may be useful to implement the optimization routine such that it aims at iteratively annihilate the derivatives of the log likelihood function, respective to each model parameter, instead of aiming at maximizing the log likelihood itself. The derivatives can also be used in the update of the parameter vector estimate, as in a Newton-Raphson approach.

Defining the two auxiliary variables

$$\alpha_t=\ln\left(\sigma_t^{\,2}\right) \tag{2.23}$$

$$\begin{aligned}\boldsymbol{x}_{\boldsymbol{\theta}t}=[1,&\left|r_{t-1}\right|/\sigma_{t-1}-\sqrt{2/\pi},\cdots,\left|r_{t-Q}\right|/\sigma_{t-Q}-\sqrt{2/\pi},\\ &r_{t-1}/\sigma_{t-1},\cdots,r_{t-Q}/\sigma_{t-Q},\alpha_{t-1},\cdots,\alpha_{t-P}]^T\end{aligned} \tag{2.24}$$

helps in the algebraic determination of the derivatives of the log likelihood. Firstly we calculate the derivative of α_t with respect to an arbitrary element θ of the parameter vector $\boldsymbol{\theta}$. This is done by applying the derivative operator to (2.14):

$$\begin{aligned}\frac{d\alpha_t}{d\theta} &= \frac{d\kappa}{d\theta} + \sum_{i=1}^{P} \frac{d\left(G_i \cdot \alpha_{t-i}\right)}{d\theta} \\ &+ \sum_{j=1}^{Q} \frac{d\left(A_j \cdot \left(\left|z_{t-j}\right| - \sqrt{2/\pi}\right)\right)}{d\theta} + \sum_{j=1}^{Q} \frac{d\left(L_j \cdot z_{t-j}\right)}{d\theta} \\ &= \frac{d\kappa}{d\theta} + \sum_{i=1}^{P} \frac{d\left(G_i \cdot \alpha_{t-i}\right)}{d\theta} \\ &+ \sum_{j=1}^{Q} \frac{d\left(A_j \cdot \left(\left|\frac{r_{t-j}}{\sigma_{t-j}}\right| - \sqrt{\frac{2}{\pi}}\right)\right)}{d\theta} + \sum_{j=1}^{Q} \frac{d\left(L_j \cdot \frac{r_{t-j}}{\sigma_{t-j}}\right)}{d\theta}\end{aligned} \tag{2.25}$$

To proceed from the expression above, we notice that the returns are the observed variables, and thus they can be treated as constants in the derivatives. Moreover, we calculate an auxiliary equation below to help dealing with the derivative of the inverse of the standard deviation:

$$\frac{d\sigma_t^{-1}}{d\theta} = -\sigma_t^{-2}\frac{d\sigma_t}{d\theta} = -\frac{1}{2\cdot\sigma_t}\frac{d\alpha_t}{d\theta} \tag{2.26}$$

where the last equality can be easily verified by replacing the derivative of α_t for its expression in function of the derivative of σ_t obtained from the differentiation of (2.23) with respect to θ.

Combining (2.25) and (2.26), straightforward algebraic manipulation results in the gradient of α_t with respect to the parameter vector $\boldsymbol{\theta}$.

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\alpha_t &= \boldsymbol{x}_{\boldsymbol{\theta}t} + \sum_{i=1}^{P} G_i \cdot \nabla_{\boldsymbol{\theta}}\alpha_{t-1} \\ &- \frac{1}{2}\cdot\sum_{j=1}^{Q}\frac{1}{\sigma_{t-i}}\cdot\left(A_i\cdot\left|r_{t-i}\right| + L_i\cdot r_{t-i}\right)\cdot\nabla_{\boldsymbol{\theta}}\alpha_{t-1}\end{aligned} \tag{2.27}$$

In each iteration of the maximum likelihood estimation, this gradient vector can be calculated for each time instant t recursively

using the expression (2.27), as done for calculation of σ_t. For initialization purposes, setting the vector to its unconditional mean for instants before first sample can be done using the following expressions:

$$E\left(\nabla_{\boldsymbol{\theta}}\alpha_t\right)=\begin{cases}\dfrac{1}{1-\sum_{i=1}^{P}G_i}, \theta=\kappa\\ \dfrac{\kappa}{\left(1-\sum_{i=1}^{P}G_i\right)^2}, \theta=G_i\\ 0, \theta=A_j, \theta=L_j\end{cases} \tag{2.28}$$

The results above can be reached by differentiating (2.14), applying the expectation operator and assuming stationarity, after some algebraic manipulation. A simpler alternative that leads to the same results is to differentiate (2.21) and interchange the orders of the expectation and derivative operators.

Differentiation of (2.22) combined with (2.26) yields the gradient of the log likelihood with respect to the model parameters:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}l\left(\boldsymbol{\theta}\right)&=-\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\left[\frac{r_t^{\,2}}{2\cdot\sigma_t^{\,2}}+\ln\left(\sigma_t\right)\right]\\ &=-\frac{1}{2}\cdot\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\left[r_t^{\,2}\cdot\sigma_t^{\,-2}+\alpha_t\right]\\ &=-\frac{1}{2}\cdot\sum_{t=1}^{T}\left[r_t^{\,2}\cdot 2\cdot\sigma_t^{\,-1}\cdot\left(-\frac{1}{2\cdot\sigma_t}\nabla_{\boldsymbol{\theta}}\alpha_t\right)+\nabla_{\boldsymbol{\theta}}\alpha_t\right]\\ &=\frac{1}{2}\cdot\sum_{t=1}^{T}\left[\frac{r_t^{\,2}}{\sigma_t^{\,2}}-1\right]\cdot\nabla_{\boldsymbol{\theta}}\alpha_t\end{aligned} \tag{2.29}$$

In this work, the maximum likelihood estimation was implemented using the “garchfit” function of the econometrics toolbox of Matlab version R2011a.

2.6 CHAPTER CONCLUSIONS

This chapter presented the conditional heteroscedasticity models, which are among the most used parametric models for describing financial returns series.

The ARCH, GARCH and EGARCH models were defined by equations (2.2), (2.10) and (2.14), respectively, and the latter was chosen to be used (exclusively) in this work for two of its advantages. Firstly, the less severe restrictions on the parameters needed to assure returns variance positiveness and stationarity. Secondly, EGARCH models are able to reproduce important properties of financial series, like the leverage effect in which negative returns (losses) are most likely to be succeeded by higher variances (which are related to risk), when compared to same magnitude positive returns (gains). This effect is extensively supported by the literature and market agents behavioral psychology (BALABAN, 2004; NELSON, 1991; SU, 2010).

Finally, a maximum likelihood framework for EGARCH model parameters estimation has been proposed, including equations for unobserved standard deviations, likelihood and its gradient calculations. The use of gradients was shown to facilitate the implementation of a numerical routine to iteratively determine the maximum likelihood estimate of the parameter vector.

3 EGARCH MODEL SELECTION CRITERIA

Having decided to work with the EGARCH model, the next steps are the determination of the orders (P and Q) and the estimation of the corresponding model parameters.

The estimation of the model parameters for given orders P and Q will be done by maximum likelihood, as typically done in the literature (BALABAN, 2004; NELSON, 1991; SU, 2010), and already detailed in the previous chapter. This work focuses on proposing a methodology for the choice of the model orders P and Q.

In this chapter, we review some of the most employed model selection criteria, also known as information criteria, which will be instrumental in constructing the methodology to be proposed in the next chapter.

The most used information criteria in the field of finances are the AIC (Akaike Information Criterion) and the SIC (Schwarz Information Criterion) (KUHA, 2004).

For notation purposes, in this work the acronyms AIC and SIC will denote the criteria and the underlying frameworks of model selection, whereas AIC_m and SIC_m, will denote, respectively, the values of the AIC and SIC information criteria for the m-th model.

3.1 THE KULLBACK-LEIBLER (KL) DIVERGENCE

The Kullback-Leibler (KL) divergence is the basis of the information criteria frequently used for model order selection. Let $g(\boldsymbol{x})$ be the true probability density function of some data vector $\boldsymbol{x}$, and let $f(\boldsymbol{x})$ denote the probability density function of a generic model for the data. The discrepancy between $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ can be expressed using the Kullback-Leibler divergence, given by (KONISHI; KITAGAWA, 2008):

$$I(g;f) = \int_{-\infty}^{\infty} \ln\left(\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right) g(\boldsymbol{x})\, d\boldsymbol{x} = E\left(\ln\left(\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right)\right) \tag{3.1}$$

which possesses the following properties (AKAIKE, 1974):

$$I(g;f) \geq 0 \tag{3.2}$$

$$I(g;f) = 0 \Leftrightarrow g(\boldsymbol{x}) = f(\boldsymbol{x}) \forall \boldsymbol{x} \tag{3.3}$$

From (3.1), the KL divergence can be written as the difference between the real expected log-likelihood and the model expected log-likelihood, both expectations being carried under the true data distribution $g(\boldsymbol{x})$:

$$I(g; f) = E(\ln(g(\boldsymbol{x}))) - E(\ln(f(\boldsymbol{x}))) \tag{3.4}$$

For a given vector $\boldsymbol{x}$, the first term on the right hand side of (3.4) is a constant (under the view of a model selection framework) since it does not depend on the candidate models being selected. Then, determining the model with smallest KL divergence is equivalent to determining the model with largest value of $E(\ln(f(\boldsymbol{x})))$, which is frequently referred to as the KL information associated to the model.

The KL divergence can be interpreted as a measure of the loss of information induced by the use of $f(\boldsymbol{x})$ in lieu of $g(\boldsymbol{x})$, and thus is often called an information function. Hence, the order selection rules derived from it are called information criteria.

The KL divergence cannot be directly used for model order selection because the probability density functions $f(\boldsymbol{x})$ of the data vector $\boldsymbol{x}$ under different hypotheses and the true probability density function of the data vector are usually unknown. Thus, practical model selection criteria rely on using estimates of the KL divergence.

3.2 AKAIKE INFORMATION CRITERION (AIC)

The AIC model selection approach is to select, from the set of all models under evaluation, the one with smallest KL divergence or, equivalently, the one with largest expected log-likelihood $E(\ln(f(\boldsymbol{x}|\boldsymbol{\theta})))$, where we have explicitly shown the dependence of the log-likelihood on the parameter vector $\boldsymbol{\theta}$ to be estimated. This expectation, however, should be evaluated with respect to the actual data probability density function $g(\boldsymbol{x})$, which is usually unknown. Moreover, the true pdf $f(\boldsymbol{x})$ under each hypothesis is also unknown. Hence, an estimate of this expectation must be used.

Let M be the number of models to be compared (hypotheses to be tested), and $f_m(\boldsymbol{x}|\boldsymbol{\theta})$, $m = 1,\ldots,M$ be the likelihood function of the m-th model. Then, for a given vector $\boldsymbol{x}$ of observed data and for each m in $[1,M]$ we define

$$l_{m,\boldsymbol{x}}(\boldsymbol{\theta}) \triangleq \ln(f_m(\boldsymbol{x} \mid \boldsymbol{\theta})) \tag{3.5}$$

$$\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} = \arg_{\boldsymbol{\theta}} \max l_{m,\boldsymbol{x}}(\boldsymbol{\theta}) \tag{3.6}$$

$$\hat{l}_{m,\boldsymbol{x}} = l_{m,\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}) \tag{3.7}$$

where $l_{m,\boldsymbol{x}}(\boldsymbol{\theta})$ is defined as the m-th model observed log-likelihood function (a function of the vector $\boldsymbol{\theta}$ of model parameters). We call it observed log-likelihood function to emphasize its dependence on the observed data $\boldsymbol{x}$, which is also explicit in the notation through the corresponding subscript. The vector $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ that maximizes $l_{m,\boldsymbol{x}}(\boldsymbol{\theta})$ is then the maximum likelihood parameter estimate, and $\hat{l}_{m,\boldsymbol{x}}$ is the maximum value of the m-th observed log-likelihood function.

Since the maximum likelihood estimate of $\boldsymbol{\theta}$ depends on the data, it is an error to confuse $E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right)$ with the KL information associated to the maximum likelihood model, which is aimed to be maximized. Given a vector $\boldsymbol{y}$ of fictitious data with the same distribution as $\boldsymbol{x}$ but independent from it ($\boldsymbol{x}$ and $\boldsymbol{y}$ are iid), the KL information for the data $\boldsymbol{y}$ of the model obtained through likelihood maximization using observation $\boldsymbol{x}$ is given by the left hand side of (3.8):

$$E_{\boldsymbol{X},\boldsymbol{Y}}\left(\ln(f_m(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right) \neq E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right) \tag{3.8}$$

as $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ is a function of the observation $\boldsymbol{x}$.

Notice that the KL information fairly evaluates the model with independent data (from that used to fit the model), and thus corresponds to the figure used in the KL divergence definition – the higher the KL information, the lower the KL divergence and the better the model is expected to be, in particular for forecasting. The right hand side of (3.8) will be referred to as naïve KL information (because it is a naïve approximation of the true KL information). It is also noted that the true KL information has an obvious cross validation interpretation.

Now, one possible approximation for the naïve KL information $E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right)$ is $\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))$, the latter being clearly an unbiased estimator of the former, although a biased estimator of the true KL information (KONISHI; KITAGAWA, 2008). Hence,

$$E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right)\approx\hat{l}_{m,\boldsymbol{x}} \tag{3.9}$$

Relying on this approximation alone for model evaluation would simply lead to the selection of the model with highest maximum observed log-likelihood, which is intuitively cumbersome since higher order models would be selected and the approach would be highly prone to overfitting. Indeed, as we will detail next, the orders of the models are key for the improvement of the approximation above and, since we are dealing with choosing the order of an EGARCH model (in other words, comparing different order models to select from), this question is of utmost importance. In practice, the simple approximation above without any correction would lead to the selection of the maximum order estimated EGARCH model.

The reason for this approximation to be inadequate is that the data used to estimate (the parameters of) the models is the same data used, afterwards, to evaluate these same models when simply $\hat{l}_{m,\boldsymbol{x}}$ is taken as the evaluation metric. In other words, the difference stated in (3.8) should be corrected since the naïve KL information inadequately leads to better evaluations for models that are overfitted to the available data.

The correction needed for (3.8) and (3.9) demands the estimation of the following bias, as derived by Konishi and Kitagawa (2008):

$$b_m \triangleq E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right)-E_{\boldsymbol{X},\boldsymbol{Y}}\left(\ln(f_m(\boldsymbol{y}\mid\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right) \tag{3.10}$$

For better understanding the discussion above, define the optimum parameter vector $\boldsymbol{\theta}_0$ as the value of $\boldsymbol{\theta}$ that minimizes the KL divergence between the model and the true DGP (data generating process) or, equivalently, maximizes the expected (under the true DGP) model log-likelihood (the true KL information). Then, the observed log-likelihood function and the expected log-likelihood vary with the model parameters as shown in Figure 3.1.

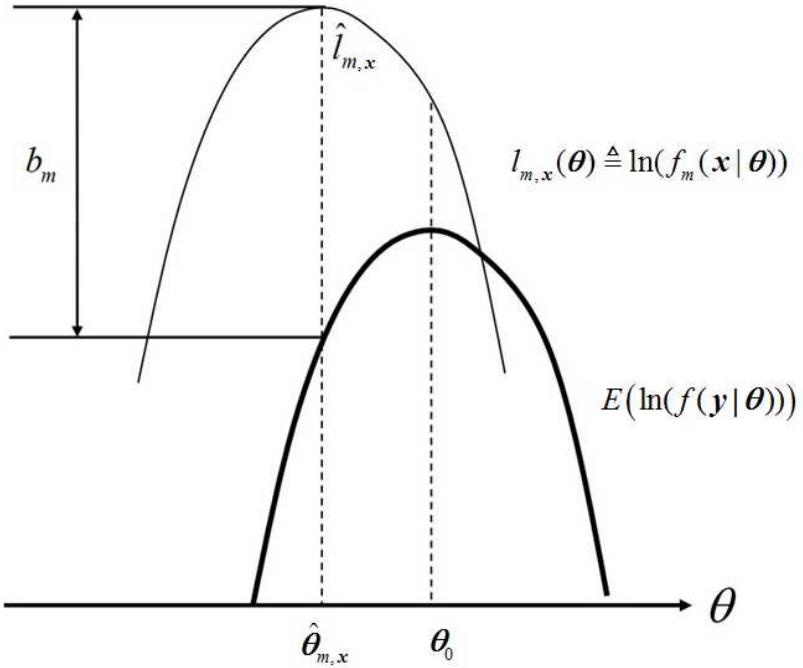

Figure 3.1 – Comparison of observed (at higher left) and expected (at lower right) log-likelihood. Source: (KONISHI; KITAGAWA, 2008), adapted for notation.

Notice that, according to the description of the right hand side terms of (3.10), the first is the expected value of the (inadequate) metric $\hat{l}_{m,\boldsymbol{x}}$, and the second is the true expected log-likelihood, conducted fairly on independent data (out-sample), which is adequate for matters of model evaluation and overfitting prevention.

Being b_m the expectation of the difference from the poor estimate $\hat{l}_{m,\boldsymbol{x}}$ to the real expected log-likelihood of the m-th model, we can correct (3.9) to obtain an unbiased estimate of the KL information:

$$E_{\boldsymbol{X},\boldsymbol{Y}}\left(\ln(f_m(\boldsymbol{y}\,|\,\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right) \approx \hat{l}_{m,\boldsymbol{x}} - b_m \tag{3.11}$$

The bias b_m can be estimated (AKAIKE, 1974) as being equal to the number of model parameters (that we will denote as p_m). This result is obtained under the assumption that the evaluated model will coincide

with the true DGP for some parameter vector $\boldsymbol{\theta}_0$ (optimal and unknown). In the following, we present a derivation of this result based on a second-order Taylor series expansion of the log-likelihood of the vector data $\boldsymbol{y}$ (STOICA; SELEN, 2004). To that end, we introduce some definitions below:

$$h(\boldsymbol{\theta}) \triangleq \ln(f_m(\boldsymbol{y} \mid \boldsymbol{\theta})) \tag{3.12}$$

$$J(\boldsymbol{\theta}) = E_{\boldsymbol{Y}}\left[-\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] \tag{3.13}$$

$$\hat{J}_y(\boldsymbol{\theta}) = -\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.14}$$

where the second order derivatives matrix $J(\boldsymbol{\theta})$ is the well-known Fischer information matrix associated to the m-th model and (3.14) defines its natural unbiased estimator, whose dependence on one single realization data vector $\boldsymbol{y}$ is explicit through the corresponding subscript.

From asymptotic maximum likelihood theory, we know that the maximum likelihood parameter vector estimate tends to a Gaussian distribution with mean $\boldsymbol{\theta}_0$ and covariance matrix equal to the inverse of the Fischer information matrix:

$$\lim_{N \to \infty} \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} \sim \mathbf{N}(\boldsymbol{\theta}_0, J(\boldsymbol{\theta}_0)^{-1}) \tag{3.15}$$

where ~ means that the left hand side random vector is distributed according to the right hand side pdf, and N on the left hand side is the number of samples (elements in data vector). **N** in the right hand side accounts for the Normal distribution, which should be clear from the context and is graphed differently from the former to avoid confusion.

Consider the function h evaluated at $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ to be approximated by its second-order Taylor series expansion around $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}$, where

$$\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}} = \arg_{\boldsymbol{\theta}} \max \ln(f_m(\boldsymbol{y} \mid \boldsymbol{\theta})) \tag{3.16}$$

yielding

$$
\begin{aligned}
h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}) &\approx h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}) + (\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \nabla_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}) \\
&\quad -\frac{1}{2}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}) \\
&= h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})
\end{aligned}
\tag{3.17}
$$

In (3.17) we used the fact that $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}$ is the point of maximum of the function h due to the maximum likelihood estimation, and thus the corresponding gradient is a zero vector. Combining (3.10) and (3.17):

$$
\begin{aligned}
b_m &= E_{\boldsymbol{X}}\left(\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right) - E_{\boldsymbol{X},\boldsymbol{Y}}\left(h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}})\right) \\
&\approx E_{\boldsymbol{X}}\left\{\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right\} - E_{\boldsymbol{X}}\left\{E_{\boldsymbol{Y}}\left[h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})\right]\right\} \\
&\quad + E_{\boldsymbol{X}}\left\{E_{\boldsymbol{Y}}\left[\frac{1}{2}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})\right]\right\} \\
&= E_{\boldsymbol{X}}\left\{\ln(f_m(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}))\right\} - E_{\boldsymbol{X}}\left\{E_{\boldsymbol{Y}}\left[\ln(f_m(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}))\right]\right\} \\
&\quad + E_{\boldsymbol{X}}\left\{E_{\boldsymbol{Y}}\left[\frac{1}{2}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})\right]\right\} \\
&= E_{\boldsymbol{X}}\left\{E_{\boldsymbol{Y}}\left[\frac{1}{2}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})\right]\right\}
\end{aligned}
\tag{3.18}
$$

Notice that to obtain the last line of (3.18) we cancelled out the terms referring to the naïve KL information evaluated through the data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, since these two vectors are i.i.d. The first term is the naïve KL information calculated using random vector $\boldsymbol{x}$ while the second one is the expectation, taken with respect to the pdf of $\boldsymbol{x}$, of the naïve KL information calculated using a random vector $\boldsymbol{y}$. As the latter is not a function of $\boldsymbol{x}$, the outer expectation reduces to its argument. Finally, since $\boldsymbol{x}$ and $\boldsymbol{y}$ are identically distributed, both naïve KL information measures are the same, from what the cancellation follows. Next, we proceed from (3.18), combine it with (3.15) and use the properties of the trace (sum of a matrix diagonal elements), abbreviated as *tr*:

$$
\begin{aligned}
b_m &= \frac{1}{2} E_{\boldsymbol{X},\boldsymbol{Y}} \{ tr[(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T \hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})] \} \\
&= \frac{1}{2} E_{\boldsymbol{X},\boldsymbol{Y}} \{ tr[\hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})^T] \} \\
&= \frac{1}{2} E_{\boldsymbol{X},\boldsymbol{Y}} \{ tr[\hat{J}_{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}})[(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \boldsymbol{\theta}_0) - (\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}} - \boldsymbol{\theta}_0)] \\
&\quad [(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}} - \boldsymbol{\theta}_0) - (\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}} - \boldsymbol{\theta}_0)]^T] \}
\end{aligned}
\tag{3.19}
$$

which asymptotically yields:

$$
\begin{aligned}
\lim_{N\to\infty} b_m = \frac{1}{2} tr[J(\boldsymbol{\theta}_0)(&Cov\{\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}\} \\
&-2Cov\{\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}\} + Cov\{\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}, \hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}\}]
\end{aligned}
\tag{3.20}
$$

Notice that since $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ and $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}$ are independent, their covariance matrix (abbreviated as *Cov*) is zero, and since $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ and $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{y}}$ are identically distributed, it follows from (3.15) that the autocovariance matrixes of both random vectors are equal to each other and tend asymptotically to the inverse of the Fischer information matrix. Hence, denoting the Identity matrix with an arbitrary number of rows *R* as I_R,

$$
\begin{aligned}
b_m \approx \lim_{N\to\infty} b_m &= \frac{1}{2} tr[J(\boldsymbol{\theta}_0)(J(\boldsymbol{\theta}_0)^{-1} + J(\boldsymbol{\theta}_0)^{-1})] \\
&= tr(I_{p_m}) = p_m
\end{aligned}
\tag{3.21}
$$

As stated, the Akaike criterion is based on the bias being estimated as the number of model parameters. Defining the AIC_m as the Akaike estimate for the expected log-likelihood (KL information) that evaluates (the higher the better) the *m*-th model, and combining equations (3.11) and (3.21), yields the Akaike Information Criterion:

$$
AIC_m \triangleq \hat{l}_{m,\boldsymbol{x}} - p_m \approx E_{\boldsymbol{X},\boldsymbol{Y}} \left(\ln(f_m(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}})) \right) \tag{3.22}
$$

For historical reasons (BURNHAM; ANDERSON, 2004) that are not relevant for the presentation of information criteria intended here,

the AIC is normally defined with a constant of -2 multiplying (3.22). Hence, the AIC would be minimized instead of maximized for the same model selection approach, due to the negative sign of such a constant.

If the premise according to which the evaluated model at $\boldsymbol{\theta}_0$ coincides with the true DGP is not used, the approach will lead to the Takeuchi criterion (TIC), in which the bias estimate depends on the derivatives of the model log-likelihood in respect to the individual parameters (KONISHI; KITAGAWA, 2008). However, the simplicity of the Akaike estimate, its independence to the real unknown DGP and the avoidance of errors in log-likelihood derivatives estimation, made the Akaike criterion one of the most used (MITCHELL; MCKENZIE, 2003).

Independently of adopting or not the simplifying premise that the parametric model includes in its parameters subspace the real DGP, leading respectively to the AIC or to TIC, it is important to emphasize that both are asymptotical approximations, strictly exact only when the number of observations tends to infinite.

Lastly, notice that the AIC reduces itself to the maximum attained observed log-likelihood obtained in the process of parameters estimation, subtracted from the number of parameters of the model (the dimension of the parameters vector). This subtraction accounts for a complexity penalty, compensating for the higher observed log-likelihood that higher order models are able to fit in-sample, thus providing a solution to the overfitting problem that demanded the bias correction in the first place.

3.3 SCHWARZ INFORMATION CRITERION (SIC)

The Schwarz criterion (frequently called bayesian information criterion) aims to select the model with *a posteriori* highest probability of being the correct one, among the set of models being evaluated. Keeping the notation, for a given random vector $\boldsymbol{x}$ of observed data, and considering M models indexed by subscript $m = 1,...,M$, the following expression denotes the probability (after data observation) that the m-th model be the true DGP:

$$P(g = f_m \mid \boldsymbol{x}) = \frac{P(g = f_m)\int f_m(\boldsymbol{x} \mid \boldsymbol{\theta})\pi_m(\boldsymbol{\theta})d\boldsymbol{\theta}}{\sum_{m'=1}^{M} P(g = f_{m'})\int f_{m'}(\boldsymbol{x} \mid \boldsymbol{\theta})\pi_{m'}(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{3.23}$$

Since g is the true probability density function, $P(g = f_m)$ is the *a priori* probability that the m-th model be the correct one. In the equation, π_m is the probability density function according to which the m-th model's parameter vector $\boldsymbol{\theta}$ is (*a priori*) distributed. This usage of Bayes theorem is only possible because of the same simplifying premise used to derive the Akaike criterion, the one that states that the true DGP corresponds to one of the models under evaluation.

By noticing that the denominator of (3.23) is identical for all models being considered, it can be disregarded in the model selection approach, since it corresponds to select the model with highest *a posteriori* probability of being correct.

The SIC approach also assumes that the *a priori* probabilities of each model being correct are equal to each other, so the only non-constant term left (among different models) of the *a posteriori* probability expression is the integral in the numerator of (3.23), which is then the quantity that should be maximized. Equivalently, one maximizes the logarithm of that integral and, as it is derived by Konishi and Kitagawa (2008), that can be asymptotically approximated as follows:

$$\ln\left(\int f_m(\boldsymbol{x} \mid \boldsymbol{\theta})\pi_m(\boldsymbol{\theta})d\boldsymbol{\theta}\right) \approx \hat{l}_{m,\boldsymbol{x}} - \frac{1}{2} p_m \ln(N) \tag{3.24}$$

In the equation above, N is the number of observations (the dimension of the observation vector $\boldsymbol{x}$), and, as was the case of the previous section, $\hat{l}_{m,\boldsymbol{x}}$ and p_m are the maximum observed log-likelihood and the number of model parameters, respectively.

In the following, we present a derivation of (3.24) based on Stoica and Selén (2004) and on Konishi and Kitagawa (2008), similar to the one used in the previous section for the AIC. We redefine the function h replacing its dependence on the data vector from $\boldsymbol{y}$ to $\boldsymbol{x}$. It is assumed that the definition being used for h in each case is clear from

the context avoiding confusion, since each one has its scope limited to the corresponding section. Here we define $h(\boldsymbol{\theta})$ as

$$h(\boldsymbol{\theta}) \triangleq \ln(f_m(\boldsymbol{x} \mid \boldsymbol{\theta})) \tag{3.25}$$

Now, consider the function h evaluated at an arbitrary $\boldsymbol{\theta}$ in the vicinity of $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$ to be approximated by its second-order Taylor series expansion about $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$. Hence, similarly to (3.17):

$$h(\boldsymbol{\theta}) \approx h(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}})^T \hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}) \tag{3.26}$$

Using the exponential of (3.26) and the first-order Taylor series expansion of π_m around $\hat{\boldsymbol{\theta}}_{m,\boldsymbol{x}}$, it follows that the integral aimed to be maximized can be approximated by the following expression, where the subscripts m will be dropped temporarily for concision purposes:

$$\begin{aligned}
&\int f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \approx \int e^{h(\boldsymbol{\theta})}\left(\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^T \nabla_{\boldsymbol{\theta}}\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right)d\boldsymbol{\theta} \\
&\approx \int e^{h(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^T \hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}\left(\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^T \nabla_{\boldsymbol{\theta}}\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right)d\boldsymbol{\theta} = \\
&\frac{f(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{(2\pi)^{-p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|^{1/2}} \times \\
&\int \frac{(\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^T \nabla_{\boldsymbol{\theta}}\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}))}{(2\pi)^{p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^{-1}\right|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^T \hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})} d\boldsymbol{\theta}
\end{aligned} \tag{3.27}$$

Now consider a random p-dimensional vector $\tilde{\boldsymbol{\theta}}$ whose distribution is Gaussian with mean $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}$ and covariance matrix $\hat{J}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}\right)^{-1}$. Then, the last expression reduces to the following, where the expectations are carried out for the distribution $\mathbf{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}, \hat{J}_{\boldsymbol{x}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}\right)^{-1}\right)$:

$$\int f(\boldsymbol{x}\mid\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\approx\frac{f(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{(2\pi)^{-p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|^{1/2}}\times$$
$$\left[\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})E(1)+\nabla_{\boldsymbol{\theta}}\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})^{T}E(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right]=\frac{f(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{(2\pi)^{-p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|^{1/2}} \tag{3.28}$$

Taking the logarithm of (3.28), the SIC approach is equivalent to maximize the following expression:

$$\ln\left(\frac{f(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{(2\pi)^{-p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|^{1/2}}\right)=$$
$$\hat{l}_{\boldsymbol{x}}+\ln(\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}))+\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln(\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|) \tag{3.29}$$

To evaluate the last term, we use the determinant property that $|k\mathrm{C}|=k^{n}|\mathrm{C}|$ for any given scalar k and n x n matric C. Then,

$$\begin{aligned}\ln\left(\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|\right)&=\ln\left(\left|\frac{N}{N}\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|\right)=\ln\left(N^{p}\left|\frac{1}{N}\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|\right)\\&=p\ln(N)+\ln\left(\left|\frac{1}{N}\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|\right)\end{aligned} \tag{3.30}$$

Combining (3.29) and (3.30), and disregarding the terms that are bounded as N tends to infinity:

$$\begin{aligned}&\lim_{N\to\infty}\ln\left(\frac{f(\boldsymbol{x}\mid\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})}{(2\pi)^{-p/2}\left|\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|^{1/2}}\right)=\lim_{N\to\infty}\left\{\hat{l}_{\boldsymbol{x}}-\frac{1}{2}p\ln(N)\right\}\\&+\lim_{N\to\infty}\left\{\ln(\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}))+\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\left(\left|\frac{1}{N}\hat{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\right|\right)\right\}\\&\approx\hat{l}_{\boldsymbol{x}}-\frac{1}{2}p\ln(N)\end{aligned} \tag{3.31}$$

Thus, the SIC approach seeks to maximize the last line of (3.31), which is the same as (3.24), previously stated. In short, the best model under this approach is the one with the highest Schwarz Information Criterion, defined below (returning to adopt the subscript *m* to index the models under evaluation):

$$SIC_m \triangleq \hat{l}_{m,\boldsymbol{x}} - \frac{1}{2} p_m \ln(N) \tag{3.32}$$

As the Akaike criterion, the Schwarz criterion reduces to the maximum observed log-likelihood subtracted from a complexity penalty that is proportional to the number of parameters of the model, thus avoiding overfitting. However, the scaling factor of the penalty is now proportional to the logarithm of the number of observations. Hence, the penalty will be higher than in Akaike criterion whenever the number of observations is higher than seven. Since this is the absolute rule, SIC tends to select simpler (more parsimonious) models than the AIC selected ones.

3.4 CHAPTER CONCLUSIONS

The information criteria AIC and SIC evaluate models with different metrics, so that the underlying difference between them accounts for what would theoretically be the “best” model.

These metrics are the Kullback-Leibler divergence (between probability density functions) for the AIC, and the *a posteriori* (after observations made) probability of each model being the correct data generating process (DGP) for the SIC.

The underlying quantities, AIC_m and SIC_m, used to select the corresponding *m*-th model (the larger the quantity value, the better the model), are defined in equations (3.22) and (3.32), respectively.

Both are given by the maximum observed log-likelihood subtracted by a complexity penalty term that is a function of the number of parameters of the model. Since our framework employs maximum likelihood estimation, which outputs the maximum observed log-likelihood, the calculations of the criteria are straightforward. The complexity penalty counters the tendency of overfitting that would happen otherwise, since higher order models will generally have higher maximized observed log-likelihood (in-sample). This is because the

availability of more parameters increases the tendency of fitting even the noise in the observations, which is undesirable.

In the end, the criteria differ in the magnitude of the complexity penalty, the SIC being more parsimonious than the AIC (the former selects lower order models due to a larger complexity penalty).

As it will be verified in the following chapters, the increased SIC parsimony was crucial for the present work, what led us to favor its use over the AIC.

4 PROPOSED METHOD

As previously stated, the aim of this work is to forecast the (one sample ahead) standard deviation of a time series of logarithmic returns, as defined in (2.14).

Based on the discussion in Chapter 2 and on the wide EGARCH model acceptance for financial series (POON; GRANGER, 2003; TSAY, 2005), the EGARCH will be used exclusively. Moreover, the parameters estimation will be done through maximum likelihood, carried out numerically.

The proposed solution includes the definition of the figure of merit to be optimized and the choice of the orders of the model. The former will be the mean squared error (MSE) of forecasting. Let N be the number of samples (from the logarithmic return) available for parameter estimation:

$$MSE \triangleq E(\hat{\sigma}_{N+1} - \sigma_{N+1})^2 \tag{4.1}$$

The choice of the standard deviation and not the variance as the variable whose mean square error will be minimized is somewhat arbitrary, although it is reasonable to suppose that it will not be critical to the final results. Moreover, this choice is supported by the literature where both options are widely used (HANSEN; LUNDE, 2005).

The work will be dedicated to the issue of model order choice, aimed to minimize the figure of merit MSE defined in (4.1).

However, as it will be depicted in a specific section of this chapter about model averaging, to rely solely in a single pair of orders (P and Q for EGARCH) can be excessively restrictive, since various EGARCH models (of different orders) can be used together for a better result.

Consider an arbitrary choice of M distinct orders (P_m,Q_m), $m = 1,\ldots, M$ corresponding to M EGARCH models, each one with its own forecast for the return standard deviation. This work studies the problem of determining the weights w_m to be attributed to each individual (model) forecast. To minimize the MSE associated with the combined forecast $\hat{\sigma}_{N+1}$ defined as

$$\hat{\sigma}_{N+1} = \sum_{m=1}^{M} w_m \hat{\sigma}_{N+1,m} \tag{4.2}$$

where $\hat{\sigma}_{N+1,m}$ is the forecast one sample ahead (N+1) obtained from the m-th model, EGARCH(P_m,Q_m).

It should be observed that the problem of obtaining the weights attributed to the individual model's forecasts does not exclude, but otherwise generalizes, the single model selection, since this setting is still possible through the attribution of a unitary weight for that selected model and zero weights for the disregarded ones.

4.1 RELATIONSHIP BETWEEN MODEL'S ORDER AND MSE

In the particular case of selecting only one from the M arbitrarily pre-chosen and estimated models, and assuming that the true DGP order lies in the model set, it is possible to choose the correct order, incur in underfitting or incur in overfitting.

When working with EGARCH, which has two distinct order parameters (P and Q), it is possible that neither of the three possibilities is strictly followed, since there is the possibility of choosing one order higher than the correct one and the other order lower than the correct one, for example. However, we will briefly discuss the three simpler scenarios mentioned previously, from which important conceptual support to the development of the work shall be drawn.

If the order of a model is increased, the estimates (of parameters and forecasts) tend to have higher variance (CLAESKENS; HJORT, 2008). This known effect should be intuitive as the information available in the data needs to be "shared" to estimate more parameters, reducing the "per parameter" amount of information. Conversely, there is a potential bias reduction, since higher biases happen in underfitting, where the disregarded parameters are therefore biased towards the value of zero.

In general, there should be an optimal number of parameters, that balances bias and variance in a minimal MSE; a point from which raising the number of parameters causes a higher variance increase than bias decrease, and from which reducing the number of parameters causes a higher bias increase than variance decrease, such that any change leads to a net increase in MSE.

In the scenario in which one of the candidate models has the exact same order than the true DGP, it is intuitive that such a candidate model corresponds to the choice of the optimal number of parameters (since it is the true number), leading to the best one sample ahead

forecasting MSE. However, this is not necessarily true when the number of samples used for estimation is finite.

Using a more complex model than the true DGP is always worse, since the extra parameters (nonexistent in the DGP) lead to higher variance but to no bias reduction, resulting in loss of MSE performance.

However, if a simpler than the true DGP is used, it leads both to bias increase due to disregarded parameters and variance decrease as well, due to fewer parameters being estimated. Which effect will offset the other is not possible to be claimed *a priori* for all situations, leaving open the possibility that a simpler than the true DGP model has better performance than a model with the correct order, in terms of MSE forecasting from a finite number of past observations.

It is important not to confuse a correct order model with the correct model itself, since the former has its parameter values estimated from the data. Therefore, they are not equal to their counterparts in the correct DGP. Obviously the correct model has zero (and consequently optimal) forecasting MSE, although it is never available for selection.

Not only the mentioned possibility exists but it was also actually observed in this work, in return standard deviation forecasting with EGARCH. As will be presented in greater detail in the results chapter, it was often the case that simpler than correct DGP models had better performance (lower MSE) than the correct order (estimated) model, although it has never happened that higher order (overfitted) models had better performance than correct order ones, in conformity with what was previously stated.

In such a scenario, it is important to emphasize that the correct order selection, although a valid objective in other situations, is not intended in this work, which otherwise aims at minimizing the forecasting MSE. It should therefore be clear from the previous discussion that those objectives are not equivalent.

On the contrary, since the models incurring in underfitting had superior performance, it happened to be desirable to look for an approach that favored lower than correct orders models. In the case of model averaging (and not only model selection), this is accomplished through attributing higher weights to such lower order models, without disregarding the need for adequacy of the models to the data, which is accounted for through the information criteria AIC or SIC described in the Chapter 3.

The better performance of simpler than the true DGP models is only possible in “small” samples, since under asymptotical premises (number of samples tending to infinite) the estimates of the parameters

tend to the true DGP parameters when the estimated model has the correct order or higher than correct order. In the latter case, the extra parameters will tend to zero, leading to better performance of such models when compared to the underfitted (simpler than DGP) ones (KONISHI; KITAGAWA, 2008). This should be intuitive since the described effect of cumbersome lower “per parameter” information due to higher number of parameters tends to disappear when the available information (data) tends to infinite.

Since the observed phenomenon is contrary to the asymptotical behavior, the number of samples must be considered “small”, making any asymptotical consideration to be inadequate for this study. In the absence of non-asymptotical (small sample) analytical expressions for the maximum likelihood estimation framework, this work will rely mostly on statistical observations do draw its conclusions.

To obtain results as described above, comparing the MSE performance of models with various orders (including the correct one), it was necessary to generate synthetic data using the EGARCH model. Moreover, synthetic data generation made available a sufficient number of realizations from which robust statistical conclusions were possible. The implementation and use of synthetic data will be better detailed in foregoing chapters, but its justification will be outlined in the next section.

4.2 SYNTHETIC DATA USAGE

To compare the models in the MSE sense, EGARCH(P,Q) synthetic data have been generated and, by means of several realizations (Monte Carlo simulations), each model forecasting MSEs were inferred statistically.

The use of synthetic data brings the advantage of allowing an arbitrarily large number of realizations and makes available the true conditional standard deviations of the returns, to be compared to model forecasts for MSE evaluation. Using real data, only the returns themselves are available, whereas the standard deviation is an unobservable variable whose estimation from the data compromises the evaluation of the quality of the obtained forecasts (POON; GRANGER, 2003; TSAY, 2005).

No model, EGARCH included, is a perfect description of asset returns dynamics. The most accepted point of view is that such a process is infinitely complex, and the broad set of models that have been proposed exhibit adequacies of description that strongly depend on the

specifics of the data. Indeed, comparisons of forecasting performances using different models widely vary in the literature, depending on several factors, such as the returns series analyzed (kind of financial asset, period, periodicity), kind of figure of merit (volatility, maximum level of expected loss, for example) and the exact form of such a figure of merit to be chosen (MSE of standard deviation, minimum absolute error of variance, to name a few of many possibilities) (EZZAT, 2012; HANSEN; LUNDE, 2005).

Adequacy of any given model (EGARCH for example) to real data is not only questionable and dependent of a large number of variables, but is also compromised from possible changes of markets behavior (CLEMENTS; HENDRY, 1998; HAMILTON; SUSMEL, 1994). Stationarity is thus a frail premise, particularly in longer periods, although necessary for most real data analysis. We claim that there is a compromise between quantity of data and (approximate) validity of premises such as stationarity and adequacy of any given model.

When the observation period is small, the amount of data can be insufficient do draw statistical conclusions. That same problem remains when the period is longer, because then the amount of data is larger but the premises fail, either because of real changes in the DGP or because the inadequacies of the model may change in behavior as the data evolves, which is possible even under a stationary DGP, since the model only approximates the behavior of a usually much more complex process.

Unfortunately, high frequency data, when available, is also limited to address these issues. With this respect, we quote from an important review on volatility forecasting: “shorter than five minutes returns are plagued by spurious serial correlation caused by various market microstructure effects”. It is also mentioned that higher frequency data is in some cases worse for forecasting over longer horizons, when compared to same period lower frequency data (POON; GRANGER, 2003).

Considering the compromises of using real data, which we understand to be the cause of the aforementioned wide variation of results among different studies, and the specifics of data and performance evaluations, we opted in this work to use synthetic data, so that unquestionable valid conclusions about the design of the EGARCH model can be drawn.

The extension of the obtained conclusions from synthetic data to practical relevance for the volatility forecasting problem (measured by the conditional standard deviations) in real financial series is supported

by the wide acceptance and use of the EGARCH model, already discussed previously. That adequacy is then considered to be a potentially valid premise, and not a conclusion we aim to validate, which is beyond the scope of this work.

Proceeding this way, we separate the effects of two aspects of financial time series analysis: the choice of an adequate parametric model (here assumed to be EGARCH), and a systematic procedure to estimate the model parameters for optimal volatility forecast.

The methodology proposed in this work to select or average out different order forecasts for a better expected out of sample performance can be applied to any parametric model estimated through maximum likelihood. Thus, it can be extended for the use with another model that happens to be more adequate than EGARCH in a given situation.

4.3 MODEL AVERAGING

Given a set of estimated models, and its forecasts, to select only one forecast and disregard the others is a particular case of usage for the set of all forecasts, but not necessarily the best choice. In Timmermann (2006), it is presented a revision of the potentials of combining individual forecasts in a forecast that takes all (or a subset) in consideration. For this work, in which the forecast is for the one sample ahead standard deviation, (4.2) is the expression that denotes such a combination.

Although the forecasts combination does not, in general, need to be linear as here considered, there are too few works that use successfully nonlinear combinations. The estimation errors of the individual forecasts weights, cumbersome in the linear scenario, are even more problematic to handle with in nonlinear strategies, making these approaches less reliable (TIMMERMANN, 2006).

When averaging different models, it is possible to diversify the combined forecast error, reaching a forecast whose variance (and MSE, consequently) is lower than the individual model forecast variances (BATES; GRANGER, 1969). Another reason to average model forecasts is to diversify among models that adapt rapidly to structural changes in data (nonstationarity due to, for example, a sudden change of DGP parameters) and models that are more precise in stationary scenarios. While the former class of models are better soon after the mentioned structural changes take place, the latter class are better in steady state periods. On average, the best to do may be to weight them out. Even in a stationary scenario, similar phenomena can occur since

the models are generally approximations of a much more complex reality, and as it evolves, it is unlikely that a same model remains to be the best all the time, so that model averaging can also be superior (TIMMERMANN, 2006).

Theoretically, the choice of optimum linear weights (in the MSE sense), depends on second order statistical moments of the forecast errors of individual models.

Let the column vectors of weights (applied to each forecast of the *M* models), individual forecasts (of return standard deviations), and corresponding forecasts errors, be defined respectively as:

$$\begin{aligned}
&\mathbf{w} \triangleq \left[w_1, ..., w_M \right]^T , \\
&\hat{\boldsymbol{\sigma}} \triangleq \left[\hat{\sigma}_{N+1,1}, ..., \hat{\sigma}_{N+1,M} \right]^T , \\
&\mathbf{e} = \left[e_1, ..., e_M \right] \triangleq \left[\hat{\sigma}_{N+1,1} - \sigma_{N+1}, ..., \hat{\sigma}_{N+1,M} - \sigma_{N+1} \right]^T \\
&\quad = \hat{\boldsymbol{\sigma}} - \sigma_{N+1} \mathbf{1}_M
\end{aligned} \tag{4.3}$$

where $\mathbf{1}_M$ denotes the *M* x 1 vector with all elements equal to unity. The temporal dependence of the vectors defined above has been omitted for concision, as it should be clear that they refer to one sample ahead standard deviation forecasting.

Considering the natural restriction that the weights sum is unitary, it is possible, without loss of generality, to force the *M*-th weight to be one subtracted from the other weights to incorporate this restriction in the equations. Therefore, we define auxiliary vectors that correspond to the ones defined in (4.3) with the last component suppressed, which relates to the *M*-th model whose weight is given by the unitary sum restriction as stated:

$$\begin{aligned}
&\tilde{\mathbf{w}} \triangleq \left[w_1, ..., w_{M-1} \right]^T \\
&\tilde{\boldsymbol{\sigma}} \triangleq \left[\hat{\sigma}_{N+1,1}, ..., \hat{\sigma}_{N+1,M-1} \right] \\
&\tilde{\mathbf{e}} \triangleq \left[\hat{\sigma}_{N+1,1} - \sigma_{N+1}, ..., \hat{\sigma}_{N+1,M-1} - \sigma_{N+1} \right]^T = \tilde{\boldsymbol{\sigma}} - \sigma_{N+1} \mathbf{1}_{M-1}
\end{aligned} \tag{4.4}$$

This way, the combined forecast is given by:

$$\hat{\sigma}_{N+1} = \mathbf{w}^T \hat{\boldsymbol{\sigma}} = \tilde{\mathbf{w}}^T \tilde{\boldsymbol{\sigma}} + \left(1 - \tilde{\mathbf{w}}^T \mathbf{1}_{M-1} \right) \hat{\sigma}_{N+1,M} \tag{4.5}$$

The MSE, from equations (4.1) to (4.5), is then given by:

$$\begin{aligned} &MSE = E\left[\tilde{\mathbf{w}}^T\tilde{\mathbf{e}} + w_M e_M\right]^2 = \tilde{\mathbf{w}}^T E\left[\tilde{\mathbf{e}}\tilde{\mathbf{e}}^T\right]\tilde{\mathbf{w}} + \\ &2\tilde{\mathbf{w}}^T\left(1-\mathbf{1}_{M-1}^T\tilde{\mathbf{w}}\right)E\left[\tilde{\mathbf{e}}e_M\right] + \left(1-\tilde{\mathbf{w}}^T\mathbf{1}_{M-1}\right)^2 E\left[e_M^2\right] = \\ &\tilde{\mathbf{w}}^T\left\{E\left[\tilde{\mathbf{e}}\tilde{\mathbf{e}}^T\right] - 2E\left[\tilde{\mathbf{e}}e_M\right]\mathbf{1}_{M-1}^T + \mathbf{1}_{M-1}E\left[e_M^2\right]\mathbf{1}_{M-1}^T\right\}\tilde{\mathbf{w}} \\ &+2\tilde{\mathbf{w}}^T\left(E\left[\tilde{\mathbf{e}}e_M\right] - E\left[e_M^2\right]\mathbf{1}_{M-1}\right) + E\left[e_M^2\right] \end{aligned} \tag{4.6}$$

Differentiating the MSE above with respect to $\tilde{\mathbf{w}}$ and equating the result to zero, we obtain below the expression for its optimum value $\tilde{\mathbf{w}}_{OPT}$:

$$\begin{aligned} &\frac{dMSE}{d\tilde{\mathbf{w}}} = (2E\left[\tilde{\mathbf{e}}\tilde{\mathbf{e}}^T\right] - 2E\left[\tilde{\mathbf{e}}e_M\right]\mathbf{1}_{M-1}^T - \\ &2\cdot\mathbf{1}_{M-1}E\left[\tilde{\mathbf{e}}^T e_M\right] + 2E\left[e_M^2\right]\mathbf{1}_{M-1}\mathbf{1}_{M-1}^T)\tilde{\mathbf{w}} \\ &+2E\left[\tilde{\mathbf{e}}e_M\right] - 2E\left[e_M^2\right]\mathbf{1}_{M-1} \end{aligned} \tag{4.7}$$

Solving for $\tilde{\mathbf{w}} = \tilde{\mathbf{w}}_{OPT}$ yields

$$\begin{aligned} &E\left[\left(\tilde{\mathbf{e}} - \mathbf{1}_{M-1}e_M\right)\left(\tilde{\mathbf{e}} - \mathbf{1}_{M-1}e_M\right)^T\right]\tilde{\mathbf{w}}_{OPT} = \\ &E\left[\left(\tilde{\mathbf{e}} - \mathbf{1}_{M-1}e_M\right)e_M\right] \end{aligned} \tag{4.8}$$

About notation, as previously stated, the forecasts errors depend, rigorously, on time, although the corresponding subscript has been dropped for concision. This is also convenient since, under stationarity assumption, the time dependence of these variables does not influence their statistical moments, in particular the second order ones, from which the optimum weights vector above is drawn.

In the synthetic data simulations, the statistical second order moments of individual forecasts errors will be estimated through the numerous realizations carried out, allowing for optimum weights vector and corresponding optimum forecasts combination MSE calculations.

They are used as theoretical references, since these statistical moments are not available to a realizable predictor with access to data from one realization only.

4.4 EXISTING MODEL AVERAGING METHODS

The practical estimation of optimum weights in (4.2) is rarely done in practice due to the absence of sufficient information to carry it out accurately. The most used averaging technique is the simple averaging (all weights equal to the inverse of the number of models averaged). Although seemingly simplistic, this strategy frequently leads to better results than optimal weight estimation approaches, due to the errors in these estimations (TIMMERMANN, 2006).

Although optimum weight estimation is not usually viable in practice, there are approaches that use model selection or information criteria, such as AIC and SIC described in the previous chapter. If such a criterion has useful information about the relative adequacy of the estimated models, it is natural to expect that the best evaluated ones should have higher weights in better performance averaging strategies, in particular when comparing to simple averaging.

In this work, we will use the information criteria (AIC and SIC) defined in the previous chapter, and the following three forecast combination strategies: 1) the selection of only one model (the best evaluated one according to each criterion), 2) the simple averaging forecast (with identical weights to each model), and 3) some more elaborate variations of averaging that use AIC or SIC to calculate the weights. These strategies will be presented in increasing order of complexity. For calculation of the weights from information criteria, both linear (4.12), (4.13) and exponential (4.14), (4.15) functions will be used, which should not be confused with nonlinear averaging since the functional relationship between the m-th model information criterion (AIC or SIC) and the corresponding weight w_m to be used in equation (4.2) does not influence the linear nature of the average (4.2) on the individual model forecasts. Nonlinear averaging will not be considered in this work due to lack of literature support, and hereafter expressions linear AIC (SIC) averaging and exponential AIC (SIC) averaging will be used to refer to the corresponding weight calculation strategies.

Model selection is the simplest approach, in which only one model is used. In (4.9) and (4.10), we define single model selection strategies based on AIC and SIC respectively, using the model averaging framework that consists of calculating the weights w_m to be

used into (4.2) to generate the final (combined, in general) forecast. For model selection, only the highest attained information criterion model is used, and the other model forecasts are disregarded.

$$w_{m,AIC} = \begin{cases} 1, AIC_m > AIC_j \forall j \neq m \\ 0, \exists j \mid AIC_m \leq AIC_j \end{cases} \tag{4.9}$$

$$w_{m,SIC} = \begin{cases} 1, SIC_m > SIC_j \forall j \neq m \\ 0, \exists j \mid SIC_m \leq SIC_j \end{cases} \tag{4.10}$$

Model selection, although contained in model averaging (which generalizes it), clearly represents a simpler paradigm. Thus, we compare it with model averaging strategies to quantify the benefits of the latter in our particular application.

Under true model averaging paradigm, the simplest strategy is simple averaging, which assigns to all models (and respective forecasts) the same weight:

$$w_{m,A-S} = 1/M \tag{4.11}$$

We recall that, although seemingly simplistic, simple averaging is in practice highly supported by model averaging literature, due to the absence of estimation errors found in more sophisticated weight calculations.

Based on the hypothesis that AIC or SIC have useful information about the relative merits of each model under evaluation, the next strategies seek to attribute higher weights to better evaluated models under each of these criteria. To combine this objective with the support of the simplest possible weight calculations granted by model averaging literature, (4.12) and (4.13) define linear AIC and SIC averaging schemes, respectively:

$$w_{m,A-AIC} = \frac{AIC_m - \min\limits_{1 \leq j \leq M} AIC_j}{\sum\limits_{i=1}^{M} \left(AIC_i - \min\limits_{1 \leq j \leq M} AIC_j \right)} \tag{4.12}$$

$$w_{m,A-SIC} = \frac{SIC_m - \min\limits_{1\leq j\leq M} SIC_j}{\sum\limits_{i=1}^{M}\left(SIC_i - \min\limits_{1\leq j\leq M} SIC_j \right)} \tag{4.13}$$

Note that the minimal value of each criterion (the value of the criterion obtained for the worst evaluated model) is subtracted in (4.12) and (4.13) from the values of that criterion obtained for all the models. This normalization amplifies the discrepancies among the models evaluated (within each criterion). This subtraction is not always performed in model averaging. It depends, for instance, on the variation of the information criterion values among the estimated models, relatively to the absolute criterion values. When models of different natures are combined and the relative values of the criterion highly differ, this minimum level subtraction may not be the most interesting approach (LI; LI, 2015). However, since we will use only EGARCH, we observe that the criteria calculations lead to values with small dispersion among the different order models. Then, the use of these absolute values would lead to almost identical weights (averaging equal to the simple average), justifying the convenience of the minimum value subtraction. For detailed information regarding the increase of the weight dispersion led by the offset applied to the information criterion values, we refer the reader to Appendix C.

There are also theoretical reasons for information criterion rescaling as described above. The absolute values attained by each criterion are not easily interpretable, contain arbitrary constants and are affected by sample size, whereas the relative differences obtained from the minimum attained value subtraction have more meaningful interpretations (BURNHAM; ANDERSON, 2004). Also notice that the subtraction, by construction, assigns zero weight to the worst evaluated model by the information criterion, which reduces the complexity of the problem, from determining M weights to determining M - 1 weights.

We also use the weights proportional to the exponential function applied to the (rescaled) attained information criteria. This is supported by the direct relationship between the criteria and log-likelihoods, so that the exponentiation makes the weights directly related to the likelihood itself. The functional dependence of the criteria on log-likelihoods comes from the use of the logarithm function in the derivations of both AIC and SIC. Exponential AIC and SIC averaging schemes are displayed in (4.14) and (4.15), respectively:

$$w_{m,A-E-AIC} = \frac{e^{AIC_m - \max\limits_{1\le j\le M} AIC_j}}{\sum\limits_{i=1}^{M} e^{AIC_i - \max\limits_{1\le j\le M} AIC_j}} \tag{4.14}$$

$$w_{m,A-E-SIC} = \frac{e^{SIC_m - \max\limits_{1\le j\le M} SIC_j}}{\sum\limits_{i=1}^{M} e^{SIC_i - \max\limits_{1\le j\le M} SIC_j}} \tag{4.15}$$

The argument of assigning weights proportional to the exponential of the information criteria due to their direct relationships with log-likelihoods can be best illustrated by the example of SIC. From (3.23), (3.24) and (3.32), it is clear that the exponential of SIC is approximately proportional to the *a posteriori* (given the data) probability that the corresponding model is the correct DGP. One of the simplifying assumptions underlying SIC framework is that one of the M models under evaluation is the correct DGP. Therefore, under this assumption, and given that the weights are always normalized to unitary sum, exponential of SIC weights are the ones that approximate the *a posteriori* probability of the model being the correct DGP. The following equations formalize this reasoning. Assume that,

$$P(g = f_m \mid \boldsymbol{x}) \propto e^{SIC_m} \tag{4.16}$$

$$w_m \propto e^{SIC_m} \tag{4.17}$$

$$\sum_{m=1}^{M} P(g = f_m \mid \boldsymbol{x}) = \sum_{m=1}^{M} w_m = 1 \tag{4.18}$$

where g and f_m are the probability density functions of the correct DGP and m-th model, respectively, according to which the data vector $\boldsymbol{x}$ is distributed, and $\propto$ denotes proportionality (right-hand and left-hand sides ratio is a constant). Notice that (4.16) is an approximation only, depending on simplifying premises assumed in SIC derivation. Then,

$$\begin{aligned}\hat{\sigma}_{N+1} &= \sum_{m=1}^{M} w_m \hat{\sigma}_{N+1,m} = \sum_{m=1}^{M} P\left(g = f_m \mid \boldsymbol{x}\right)\hat{\sigma}_{N+1,m} \\ &\approx \sum_{m=1}^{M} P\left(g = f_m \mid \boldsymbol{x}\right) E\left(\sigma_{N+1} \mid g = f_m, \boldsymbol{x}\right) \\ &= E\left(\sigma_{N+1} \mid \boldsymbol{x}\right)\end{aligned} \tag{4.19}$$

In other words, a weighting scheme that satisfies (4.16), (4.17), (4.18) and the corresponding assumptions generate, as its combined forecast, the conditional (given the data) expected value of the quantity being forecasted (in this case the one step ahead volatility), which is well-known to be the minimum MSE forecast (KAY, 1993). We notice that, to obtain the second line of (4.19), it is necessary the approximation according to which a given model maximum likelihood forecast is the expected value of the quantity being forecasted given the data and the hypothesis that the corresponding model matches the true DGP (STOICA; SELÉN; LI, 2004).

Due to Bayesian interpretations of AIC (different from the one presented in Chapter 3) and analogous penalized log-likelihood forms of both criteria, combined forecasts given by exponential AIC weights are similarly justified (BUCKLAND; BURNHAM; AUGUSTIN, 1997; BURNHAM; ANDERSON, 2004; STOICA; SELEN, 2004). Therefore, the choice of the information criterion to be used in the weighing scheme should depend mostly on the information criteria adequacies, given the application requirements, the same way as in a model selection framework.

Differently from (4.12) and (4.13), however, the rescaling of the criteria to remove offset constants in (4.14) and (4.15) are done subtracting the maximum (instead of the minimum) value of each criterion, so that the best model attains the rescaled criterion value of zero and the other models attain negative values. This is widely applied in practice (BURNHAM; ANDERSON, 2004; LI; LI, 2015) and avoids the exponentiation of too large positive values, corresponding to realizations where the discrepancy between the best and the worst model is too high. This exponentiation would lead to numerical problems prior to normalization of the weights, whereas the maximum criterion value subtraction leads to exponentiation of the corresponding high modulus difference with negative sign, which is numerically well handled through zero valued exponential. Notice that the mentioned rescaling

employed in (4.15) satisfies (4.17) due to the property of the exponential function according to which the exponential of an algebraic sum equals the corresponding product of exponentials. We also note that because of this property of the exponential function, the exponential of the maximum attained information criterion value is a common term to numerator and denominator in both (4.14) and (4.15), and thus the rescaling included in those expressions has practical relevance only due to the numerical issues previously described, since mathematically all instances of this term could be canceled out. In Appendix C we demonstrate that exponential functions are the only ones with this property.

Although it does not apply to this work, we mention that linear AIC and SIC averaging schemes (such as (4.12) and (4.13)) instead of exponentials AIC and SIC ones, tend to be the chosen ones when models from different families are used, due to frequently higher discrepancies among the criteria values (LI; LI, 2015).

Despite absence of theoretical support, (4.12) and (4.13) are the simplest ways to incorporate the information criteria into the model averaging approach, which in general favors simplicity (TIMMERMANN, 2006). That is one of the main reasons to include such strategies, as well as the simple averaging (4.11), the simplest model averaging strategy possible. This reasoning can be further depicted through the argument that linear AIC (or SIC) averaging can be seen as a form of shrinkage of their exponential counterparts towards simple averaging. More formally, suppose a given set of the information criterion values being considered, and that *m1* and *m2* index any two models contained in the model set, such that *m1* is better than *m2* in respect to the information criterion (we use the AIC as an example, without loss of generality). Thus,

$$\Delta AIC_m \triangleq AIC_m - \min_{1 \le j \le M} AIC_j \tag{4.20}$$

$$AIC_{m1} > AIC_{m2} \Leftrightarrow \Delta AIC_{m1} > \Delta AIC_{m2} \tag{4.21}$$

$$\begin{aligned} \frac{w_{m1,A-E-AIC}}{w_{m2,A-E-AIC}} &= \frac{e^{AIC_{m1} - \max\limits_{1 \le j \le M} AIC_j - \left(\min\limits_{1 \le j \le M} AIC_j - \min\limits_{1 \le j \le M} AIC_j \right)}}{e^{AIC_{m2} - \max\limits_{1 \le j \le M} AIC_j - \left(\min\limits_{1 \le j \le M} AIC_j - \min\limits_{1 \le j \le M} AIC_j \right)}} \\ &= \frac{e^{\Delta AIC_{m1}}}{e^{\Delta AIC_{m2}}} = \left(e^{\Delta AIC_{m1} / \Delta AIC_{m2} - 1} \right)^{\Delta AIC_{m2}} \end{aligned} \tag{4.22}$$

Next, assume that $\Delta AIC_{m2} > 1$ (generally true in practice) except when *m2* index is the worst evaluated model, thus ΔAIC_{m2} is clearly equal to zero. Plugging this assumption in (4.22):

$$\frac{w_{m1,A-E-AIC}}{w_{m2,A-E-AIC}} > e^{\Delta AIC_{m1} / \Delta AIC_{m2} - 1} > 1 + \left(\frac{\Delta AIC_{m1}}{\Delta AIC_{m2}} - 1 \right) = \frac{w_{m1,A-AIC}}{w_{m2,A-AIC}} \tag{4.23}$$

Thus, the ratio between the best to the worst evaluated model weights is higher when using the exponential function than the one corresponding to linear weights. Therefore, the linear weighing schemes (4.12) and (4.13) provide intermediary weight dispersions, larger than the zero dispersion of simple averaging and smaller than the dispersion of the strategies (4.14) and (4.15), in which the exponentials lead to combined forecasts more concentrated in the best evaluated models. In Appendix C, we further depict the dependence of the function used to calculate the weights from information criteria on the resulting weight dispersions

This reasoning provides support to the hypothesis that linear weighing schemes can be beneficial due to possible exploration of the compromise between simple averaging and exponential weighing ones. This hypothesis is reinforced by the fact that in Chapter 5, (4.11) and (4.13) were the best performing strategies for our application, from the existent ones evaluated.

As mentioned, the assumption that $\Delta AIC_{m2} > 1$ (and thus (4.23)as well) fails for *m2* corresponding to the worst evaluated model. However, for this model, the weight assigned should be negligible for all model averaging strategies anyway (besides simple averaging), and thus the overall conclusions should not be significantly affected.

It is convenient to notice that the denominators of the expressions that define the averaging strategies presented are common to all the corresponding weights and have normalizing purposes, such that the weights sum is always unitary. We remain omitting the time dependency for concision, reminding that this dependence will exist in the case of successive (model) estimations and volatility forecasts.

About notation, for clarity, we use two subscripts for the weights, separated with commas. The first one indexes the model to which the weight corresponds (in other words, the model whose forecast will be multiplied by that weight). The second subscript relates to the strategy of weight calculation, and is composed of letters and syllables separated with hyphens, each corresponding to an attribute of that strategy, as follows: AIC or SIC indicate the information criterion being used (if any), letter A indicates model averaging and its absence implies model selection, letter E indicates exponentiation of the information criterion and, finally, letter S indicates the simple averaging strategy.

4.5 PROPOSED METHODOLOGY

As previously discussed in this chapter, there was a tendency of simpler than true DGP models to exhibit better performance than correct order estimated models, in the MSE sense. When using information criteria, SIC then led to better performances than AIC, both in selection and averaging cases, due to the fact that SIC evaluates better the simplest models, when compared to AIC.

As expected from the model averaging background discussed, model averaging had better performance than model selection, in general. Consequently, from the existing forecast strategies considered, the best one resulted from averaging models with the use of SIC, as formulated in (4.13).

Taking these relative performances under consideration along with the observed advantages of underfitting, we propose a method of averaging based on (4.13), but with the following defined underlying SIC calculation generalized with the insertion of an “hyperparameter” λ (we use this word to distinguish from parameters of the models being averaged), which is the core innovation of this work. It is intended to raise the complexity penalty, which is already higher in SIC than in AIC:

$$SIC_m(\lambda) \triangleq \hat{l}_m - (1+\lambda)\frac{1}{2} p_m \ln(N) \tag{4.24}$$

In the equation above, we keep the model indexing through the subscript m and, as it is possible to observe due to (3.32), the original SIC is a particular case of our definition (4.24), for zero valued λ.

By using values of λ higher than zero, it is possible to overpenalize the most complex models (with more parameters). The corresponding averaging strategy then gives higher weights to simpler models, which possibly incur in better performance underfitting. The corresponding weights are given below, which employ identical mathematical formulation as (4.13) except for the underlying information criterion, which has been replaced by (4.24):

$$w_{m,A-SIC(\lambda)} = \frac{SIC_m(\lambda) - \min\limits_{1\leq j\leq M} SIC_j(\lambda)}{\sum\limits_{i=1}^{M}\left(SIC_i(\lambda) - \min\limits_{1\leq j\leq M} SIC_j(\lambda)\right)} \tag{4.25}$$

Notice in the left hand side that we replace *SIC* by *SIC*(λ) in the weights subscript notation when using generalized (rather than regular) SIC averaging. Also notice that the benefits of the proposed method come at the cost of the hyperparameter λ, which needs to be determined. Different values of λ consist of different model averaging strategies, and unsuitable values can lead to performance degradation.

We also notice that, in the frequent case where there is a model *m'* that has a number of parameters strictly lower than all other models under evaluation (*m'* then indexes the minimum model), then the higher the complexity penalty (or λ), the higher the weight assigned to the minimum model. In particular, it is clear that,

$$\lim_{\lambda\to\infty} w_{m',A-SIC(\lambda)} = 1 \tag{4.26}$$

Therefore, our proposed hyperparameter λ can be viewed as a form of shrinkage of the regular SIC averaging forecast towards the minimum model forecast, provided there is one. In this work, the minimum model is the EGARCH(1,1), and we remind that within ARCH family models literature in general, the unitary order model is the most widespread used.

Given the proposed averaging strategy summarized by (4.24) and (4.25), we depict the following methodology to apply it, based on synthetic data usage.

1) The first step is to have the best possible data, compatible models and estimation routines. Gather past real data and the

corresponding logarithmic returns. Naturally, the data need to be at least of similar nature of the one to be worked with, preferably past data from the same financial series. Also determine M parametric models (model families and different orders under evaluation) to be used for volatility forecasting. In this work we used only EGARCH family, with nine different (pairs of) orders, but from the methodology point of view, this is completely arbitrary and passible of extension. Maximum likelihood (ML) estimation routines compatible to the chosen models are also needed.

2) Apply ML routines to the models and data from previous step. The outputted estimated models are realistic models to be used to generate synthetic data, and thus are the next steps DGPs.
3) Determine the number of samples N to be used by the models to make their volatility forecasts. Data availability and non-stationarity tradeoffs must be considered for each individual application.
4) Generate, for each DGP, several independent realizations (Monte Carlo framework) of return series from each DGP obtained in step 2. Each return series needs to be N samples long, and synthetic volatility (return standard deviation) corresponding to N+1 sample also needs to be determined and recorded for each realization, since it is the (exact value of the) quantity to be forecasted. Notice that there are M (one to each DGP) sets of realizations, the sets differ in nature (each one corresponds to a different DGP) while different realizations from the same set are independent and identically distributed.
5) Apply ML routines to estimate each model for each synthetic data series. From an estimated model, determine its individual forecast for volatility at time N+1 for each synthetic data series. Due to practical reasons concerning memory usage, it may be useful to notice that from this point on, the return series and estimated models can be disregarded, being necessary to keep the volatilities at time N+1, both the estimated forecasts and the true ones (outputted by the corresponding DGPs). It is also necessary to keep the maximum attained log-likelihoods corresponding to all models estimations, for information criteria evaluation.

6) Generate, for each realization, the combined forecasts for each model averaging strategy under consideration. If our proposal is to be used alone, this means to apply (4.24), (4.25) and (4.2) for a set of values for λ in consideration. However, other model averaging strategies can also be included in this step. To the combined forecasts, apply (4.1) to obtain the forecast MSE for each model averaging strategy and DGP scenario, where the expected value is naturally computed through the mean across all the realizations corresponding to such a DGP.
7) From the previous step, choose the most suitable value for λ, depending on the MSEs attained for all DGPs scenarios. If other strategies other than the one based on generalized SIC were included, consider them as well to this step of strategy decision based on synthetic MSE.
8) Employ the methodology to real data in a particular application of interest: fit the models chosen in step 1 to such data through ML, input their individual forecasts to the model averaging strategy defined in step 7 and use the corresponding combined forecast.

Figure 4.1 summarizes the proposed methodology. For concision of the illustration, some details were omitted, such as the determination of the number of samples N and the generality of the model averaging strategy decision block (it corresponds to steps 6 and 7), which otherwise is displayed in the context that our proposed method for model averaging based on (4.24) and (4.25) is arbitrarily constrained to be the sole alternative.

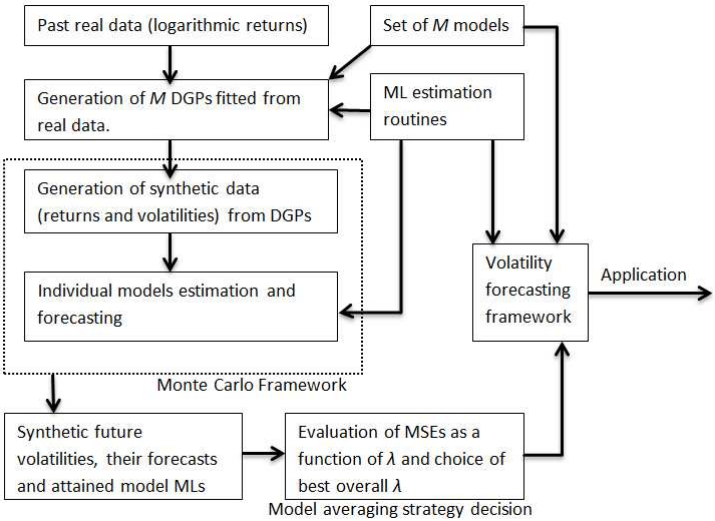

Figure 4.1 – Proposed methodology for volatility forecasting framework determination.

4.6 CHAPTER CONCLUSIONS

This chapter presented the technique of model averaging and, supported by its potentials, formulated the problem to which this work will be dedicated as the choice of weights to be given to each of the models under consideration. In our application, these are the different order EGARCH models arbitrarily defined as candidates to forecast the one sample ahead standard deviation of returns (one step ahead volatility). Thus, the weights lead to a combined forecast that is aimed to have the minimal forecast error, in the MSE sense.

Based on the problem so formulated, different existent strategies were presented, ranging from model selection to model averaging (being the former a particular case of the latter). The strategies also vary in respect to the use (if any) of information criteria discussed in the previous chapter.

Because of the particular results found in this work, which frequently favored simpler and underfitted models, a new method has been proposed to calculate model averaging weights, devised from SIC generalization that raises complexity penalties to exploit such simpler models overperformance. Therefore, the method increases simpler

models weights in the averaging strategy, ultimately aiming at (MSE) performance improvement.

We also propose a methodology based on synthetic data for implementing this model averaging strategy to our particular application of volatility forecasting application, which will be used with EGARCH models but can easily be extended to different families of parametric models estimated through maximum likelihood technique.

5 RESULTS

In this chapter, we present the statistical results obtained comparing the previously discussed one sample ahead standard deviation (volatility) forecast approaches.

The logarithmic returns of the following five stock indexes were used to support the simulations: Ibovespa or "IBOV" (Brazil's stock index), Standard & Poor's 500 or "SP500" (USA stock index), Nikkei 225 or "N225" (Japan's stock index), "DAX" (Germany's stock index) and "FTSE 100" (England's stock index). The data was taken from Yahoo! Finance from January 03, 2000 to April 09, 2015.

For each of those indexes, and for each pair (P,Q) of order parameters, a corresponding EGARCH model was fitted. We used P and Q ranging from one to three and the nine corresponding models (for each of the five indexes) are displayed in Appendix A. These nine EGARCH models fitted from the real data are the DGPs used to generate the synthetic data for the simulations. The normalized innovations were modeled as standard Gaussian random variables.

Assuming the DGPs to be realistic, since they were fitted from real data, we employed them to generate synthetic logarithmic returns. The parameters of each of these DGPs were held fixed during the simulations, such that for each DGP, all corresponding generated synthetic returns series were independent and identically distributed realizations (Monte Carlo framework). For each realization, the same $M = 9$ model structures (EGARCH with order parameters ranging from one to three) were fitted from the synthetic data and had their one step ahead volatility forecast evaluated and compared to the true one (outputted from the DGP), the differences being the forecast errors. The volatility forecasts of fixed order models and of model averaging strategies were compared in terms of MSE, given by the squared forecast errors averaged across the realizations. Through 50000 realizations of each DGP, the Monte Carlo simulations provided strong statistical conclusions. Moreover, the calculations of the squared forecast errors are exact, since the true volatilities are known in the simulations due to the synthetic data generation.

For each DGP of given orders P and Q the data consists of the logarithmic returns r_t with EGARCH variance, as defined by equations (2.1) and (2.14), repeated here for convenience:

$$r_t - \mu_t \approx \varepsilon_t = \sigma_t \cdot z_t \tag{5.1}$$

$$\ln\left(\sigma_t^2\right) = \kappa + \sum_{i=1}^{P} G_i \cdot \ln\left(\sigma_{t-i}^2\right) + \sum_{j=1}^{Q} A_j \cdot \left(\left|z_{t-j}\right| - E\left(\left|z_{t-j}\right|\right)\right) + \sum_{j=1}^{Q} L_j \cdot z_{t-j} \tag{5.2}$$

For the approximation in (5.1) to hold exactly, we first fit the DGPs from real data modeling μ_t as a constant C, jointly estimated with the EGARCH (variance model defined by (5.2)) parameters:

$$r_t = C + \sigma_t \cdot z_t \tag{5.3}$$

This is needed because in practice the returns are not zero mean and (5.1) approximation is only reasonable under the constant mean premise and assuming that the mean value of the data has been removed. We then set the constant mean C to zero both in the DGPs employed for the synthetic data generation and in the EGARCH models fitted to the synthetic data, so that the analysis is focused on the EGARCH parameters only.

Although practical data are not zero-mean, proceeding this way allows for fair comparisons of EGARCH models of different orders without the need to consider the possible effects of errors in the joint estimation of return means.

As the parameter of interest is the volatility, the applicability of the proposed methodology should not be compromised by this simplification.

In Section 5.1, we compare the performances of the existing forecasting strategies mentioned in the previous chapter for estimating the volatility, considering Gaussian innovations. The section is divided in subsections, one for each stock index (country). The conclusion is that averaging the forecasts with SIC is the best option in the majority of the scenarios.

Section 5.2 presents the performance of the proposed method corresponding to (4.24) and (4.25), in terms of its hyperparameter λ. Since it is a generalization of the SIC averaging strategy, which is the most promising of the existing strategies considered here, we concentrate on the comparison between the proposed method and regular SIC averaging.

Section 5.3 repeats the steps of the previous sections for Student t distributed normalized innovations. The corresponding models fitted from real data are found in Appendix B. The aim of this section is to determine if the same performance relationships hold for that scenario, which is capable of reproducing logarithmic returns with higher excess kurtosis than EGARCH with Gaussian normalized innovations.

5.1 STANDARD DEVIATION FORECASTS PERFORMANCES

Before comparing the forecast strategies among themselves, we first compare the performances of the individual fixed order EGARCH models, without any selection or averaging. This provides some insight about the relative performances of underfitting, correct fitting and overfitting.

In the following subsections, each devoted to a specific stock index, we present the MSE's and draw the corresponding conclusions for posterior consolidation.

5.1.1 Ibovespa or IBOV (Brazil's stock index)

Table 5.1 shows the MSE performance of fixed EGARCH order models. Each column stands for the EGARCH model (fitted from real data as previously discussed) used do generate the synthetic data for the 50000 realizations (true DGP). Each line corresponds to a fixed order meaning that in each realization an EGARCH model with that order was fitted from the synthetic data and its forecast evaluated. The MSE corresponding to each model was computed by averaging the forecasting squared errors of all the realizations, as defined by (4.1). In other words, the MSE in the cell corresponding to (P_m,Q_m) column and (P_n,Q_n) line is the mean squared error of forecasting using always an EGARCH(P_n,Q_n) fitted from the data generated by the EGARCH(P_m,Q_m) DGP. These DGPs were fitted once from real data and are available at Appendix A, therefore they were held constant over the realizations. The number of samples (N) was 250, therefore the forecast error corresponds to the 251^{th} sample.

(P,Q) of EGARCH fitted	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(1,1)	6.27E-06	6.17E-06	6.31E-06	1.10E-05	9.09E-06	9.04E-06	1.30E-05	9.50E-06	9.32E-06	8.87E-06
(2,1)	7.77E-06	7.70E-06	9.42E-06	1.50E-05	2.93E-05	1.02E-05	1.94E-05	3.19E-05	1.10E-05	1.58E-05
(3,1)	1.03E-05	1.01E-05	1.04E-05	1.87E-05	2.40E-05	2.47E-05	2.30E-05	2.39E-05	5.30E-05	2.20E-05
(1,2)	8.57E-06	8.38E-06	8.62E-06	8.03E-06	9.84E-06	1.00E-05	9.50E-06	9.19E-06	9.63E-06	9.09E-06
(2,2)	1.09E-05	1.10E-05	1.31E-05	1.00E-05	1.14E-05	1.26E-05	1.16E-05	1.19E-05	1.19E-05	1.16E-05
(3,2)	1.41E-05	1.45E-05	1.43E-05	1.34E-05	1.39E-05	1.38E-05	1.49E-05	1.47E-05	1.52E-05	1.43E-05
(1,3)	1.11E-05	1.09E-05	1.10E-05	1.06E-05	1.15E-05	1.15E-05	1.11E-05	1.12E-05	1.16E-05	1.12E-05
(2,3)	1.38E-05	1.38E-05	1.75E-05	1.34E-05	1.46E-05	1.45E-05	1.33E-05	1.42E-05	1.42E-05	1.44E-05
(3,3)	1.84E-05	1.80E-05	1.87E-05	1.75E-05	1.71E-05	1.73E-05	1.73E-05	1.64E-05	1.68E-05	1.75E-05

Table 5.1 – forecast MSEs from fixed order EGARCH models. IBOV index, Gaussian normalized innovations, $N = 250$.

Green cells correspond to the best model fit for each DGP. The yellow cells correspond to correct order fitting, where the fitted model has the same order as the true DGP. It is important not to confuse the latter with correct model forecasting since the fitted model has the correct orders but its parameters are estimated from the synthetic data and thus are subject to estimation errors. The last column averages the performance (MSE) of each EGARCH fitted model across all DGPs to provide an overall picture of each model performance. Simply put, the last column is the simple mean of the previous nine columns.

Notice that the number of EGARCH parameters, which can be inspected from (5.2), is given by $1+P+2\cdot Q$. Thus, the rows and columns of Table 5.1 are organized in a non-decreasing number of parameters per model sequence.

For a given DGP, the best model in the MSE sense is the one that achieves the optimum bias-variance tradeoff, where a lower than optimum number of parameters increases the bias (due to disregarded parameters) more than it decreases the variance (more data per estimated parameter), resulting in a net MSE increase. Analogously, a larger than optimum number of parameters increases the variance (less data per estimated parameter) more than the corresponding reduction in bias (if any), also resulting in a net MSE increase. Intuitively, this optimum should be the correct order. Indeed, overfitting is clearly worse than correct fitting. The overfitting models are the ones with both P and Q higher or equal than P and Q of the correct DGP, being at least one order parameter strictly higher. All overfitting models have worse performance (higher MSE) than the model with the same P and Q as the true DGP (correct fitting). In these cases, reducing the overfitting towards the correct order actually decrease variance without any

increase of bias due to disregarded parameters, since they are nonexistent in the true DGP.

However, in underfitting region, lower than correct orders models often display better results. Underfitting models are the ones with both P and Q lower or equal than P and Q of the correct DGP, being at least one order parameter strictly lower. For an EGARCH(2,1) DGP, EGARCH(1,1) was better than correct fitting; for an EGACRH(3,1) DGP, both EGARCH(2,1) and EGARCH(1,1) were better than correct fitting; for an EGARCH(1,2) DGP correct fitting was better than EGARCH(1,1) - the only underfitting model – this was the only case where underfitting was possible and correct fitting was best; for an EGARCH(2,2) DGP, two out of three underfitting models were better than correct fitting; for an EGARCH(3,2) DGP, four out of five underfitting models were better than correct fitting; for an EGARCH(1,3) DGP, one out of two underfitting models were better than correct fitting; for an EGARCH(2,3) DGP, four out of five underfitting models were better than correct fitting and finally for an EGARCH(3,3) DGP seven out of eight underfitting models were better than correct fitting.

The overall performance of underfitting was clearly better than correct fitting, with EGARCH(1,1) followed by EGARCH(1,2) with the best results. Thus, the bias introduced by disregarding parameters existing in the true DGP was more than compensated by the decrease in variance due to the lower number of parameters. This result is not in agreement with asymptotic theory, and thus this should be treated as a small sample problem.

In this subsection, we will repeat the results for $N = 500$ to look for changes in this aspect. It is important to remember that in financial series there is often a delicate tradeoff when deciding for an increase of the number of samples going further into the past, since this can weaken the data stationarity and thus the model validity assumptions. Our results show the importance of checking for small sample effects before any analysis using asymptotic theory is made. Moreover, they show that seeking for the correct order model can be misleading for forecasting applications.

The good performance of underfitting models, especially the minimum order model EGARCH(1,1) can also be regarded as the underlying cause of its good performance in practical applications, and thus of its widespread use over higher order models. This, however, is very different from the common assertion that real data is better described by EGARCH(1,1) than by higher order models, and we

consider confronting these hypotheses to be a promising topic for future research.

There are models that do not belong to underfitting, correct fitting nor overfitting categories. They have one order parameter (P or Q) higher than the correct and the other order parameter lower than the correct, being in some sense a mixture of underfitting and overfitting. As a result, these models displayed mixed results, which did not provide useful insights.

Next, in Table 5.2, we compare the forecasting strategies MSEs, but keep EGARCH(1,1) in the Table (first row) to compare the strategies to the best fixed order model. The columns correspond to the EGARCH DGPs as in Table 5.1 and the rows correspond to the model selection and averaging strategies depicted in Chapter 4. “BEST AIC” and “BEST SIC” rows correspond to model selection based on AIC and SIC criteria, “AVG AIC” and “AVG SIC” to model averaging with weights proportional to each (rescaled) criterion, “AVG-E AIC” and “AVG-E SIC” to model averaging with weights proportional to the exponential of each (rescaled) criterion, “SIMPLE AVG” averages the forecasts of all the models with equal weights ignoring the information criteria. These strategies correspond to model selection or averaging with weights calculated through (4.9), (4.10), (4.12), (4.13), (4.14), (4.15) and (4.11) respectively. Lastly, “OPT AVG” corresponds to optimum weights averaging using (4.7), which is not realizable since it requires the knowledge of correlations among forecast errors of different models, an information not available to a practical estimator with single realization data.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	6.27E-06	6.17E-06	6.31E-06	1.10E-05	9.09E-06	9.04E-06	1.30E-05	9.50E-06	9.32E-06	8.87E-06
BEST AIC	1.71E-05	1.70E-05	1.85E-05	1.74E-05	1.76E-05	1.79E-05	1.80E-05	1.78E-05	1.80E-05	1.77E-05
BEST SIC	9.62E-06	9.68E-06	1.07E-05	1.23E-05	1.22E-05	1.21E-05	1.40E-05	1.27E-05	1.23E-05	1.17E-05
AVG AIC	9.19E-06	9.05E-06	9.02E-06	9.74E-06	9.95E-06	1.01E-05	1.03E-05	9.95E-06	9.91E-06	9.69E-06
AVG SIC	7.01E-06	6.93E-06	7.00E-06	7.81E-06	8.14E-06	8.38E-06	8.78E-06	8.25E-06	8.48E-06	7.86E-06
AVG-E AIC	1.43E-05	1.41E-05	1.53E-05	1.44E-05	1.43E-05	1.48E-05	1.49E-05	1.46E-05	1.49E-05	1.46E-05
AVG-E SIC	8.51E-06	8.42E-06	9.25E-06	1.07E-05	1.07E-05	1.06E-05	1.22E-05	1.12E-05	1.09E-05	1.03E-05
SIMPLE AVG	7.42E-06	7.31E-06	7.18E-06	7.76E-06	8.21E-06	8.25E-06	8.51E-06	8.08E-06	8.51E-06	7.91E-06
OPT AVG	5.93E-06	5.80E-06	5.88E-06	6.99E-06	7.45E-06	7.61E-06	7.98E-06	7.36E-06	7.64E-06	6.96E-06

Table 5.2 – forecast MSEs from model selection and averaging strategies. IBOV index, Gaussian normalized innovations, $N = 250$.

There are no yellow cells because the concept of correct fitting (underfitting and overfitting as well) does not apply to the forecast strategies, as in general they use more than a single model forecast. The green cells were also dismissed because the optimum averaging (last row) leads always to the lowest MSE. However, we use the green cells again to identify the minimum value of each column in Table 5.3, where we display the same information of Table 5.2, but expressed in terms of relative MSE losses, defined in equation (5.4), which is how much each MSE is larger than the value of the optimum averaging MSE (whose row is thus omitted):

$$MSE_{RL} = \frac{MSE - MSE_{OPT}}{MSE_{OPT}} \tag{5.4}$$

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	5.81%	6.33%	7.35%	57.84%	21.95%	18.87%	63.39%	29.02%	21.92%	25.83%
BEST AIC	188.36%	192.58%	215.30%	148.12%	135.66%	134.99%	125.80%	141.81%	134.95%	157.51%
BEST SIC	62.18%	66.89%	82.11%	76.21%	63.93%	58.84%	75.35%	72.53%	60.32%	68.71%
AVG AIC	54.95%	55.97%	53.48%	39.27%	33.44%	32.18%	29.29%	35.17%	29.60%	40.37%
AVG SIC	18.19%	19.37%	19.07%	11.70%	9.13%	10.17%	10.03%	12.04%	10.94%	13.40%
AVG-E AIC	140.71%	143.40%	160.49%	105.55%	92.40%	94.10%	86.60%	97.93%	94.70%	112.88%
AVG-E SIC	43.51%	45.12%	57.30%	52.67%	43.49%	39.58%	52.53%	51.76%	43.20%	47.68%
SIMPLE AVG	25.19%	25.90%	22.19%	11.01%	10.07%	8.43%	6.58%	9.66%	11.30%	14.48%

Table 5.3 – model selection and averaging strategies relative MSE losses (MSE_{RL}). IBOV index, Gaussian normalized innovations, N = 250.

We notice that for Q = 1 a fixed EGARCH(1,1) dominates the strategies, although simple averaging and SIC linear averaging (which lead to similar results) dominate for Q > 1, being the best alternatives when all scenarios are averaged, with a slight advantage to SIC linear averaging. Moreover, although SIC linear averaging is the best performing strategy in only 2 cases, it is the second best approach in all the other seven scenarios, what reinforces its best overall performance.

To investigate the effect of a higher number of samples, we repeat the previous comparisons for N = 500. The results are shown in Table 5.4.

(P,Q) of EGARCH fitted	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(1,1)	2.35E-06	2.35E-06	2.55E-06	7.79E-06	5.47E-06	5.42E-06	1.11E-05	6.18E-06	6.10E-06	5.48E-06
(2,1)	3.01E-06	3.01E-06	3.57E-06	1.14E-05	2.26E-05	5.78E-06	1.54E-05	1.36E-05	6.83E-06	9.47E-06
(3,1)	4.39E-06	4.30E-06	4.60E-06	1.42E-05	1.38E-05	1.60E-05	1.74E-05	1.13E-05	2.12E-05	1.19E-05
(1,2)	3.33E-06	3.28E-06	3.52E-06	2.97E-06	4.87E-06	4.87E-06	4.35E-06	4.97E-06	5.29E-06	4.16E-06
(2,2)	4.26E-06	4.22E-06	4.86E-06	3.65E-06	4.23E-06	5.68E-06	4.92E-06	5.20E-06	6.13E-06	4.79E-06
(3,2)	6.20E-06	6.08E-06	6.25E-06	5.09E-06	5.25E-06	5.29E-06	6.23E-06	6.29E-06	6.68E-06	5.93E-06
(1,3)	4.30E-06	4.27E-06	4.37E-06	3.92E-06	4.97E-06	5.03E-06	4.31E-06	5.34E-06	5.52E-06	4.67E-06
(2,3)	5.31E-06	5.32E-06	6.38E-06	4.93E-06	5.46E-06	6.24E-06	5.10E-06	5.58E-06	6.60E-06	5.66E-06
(3,3)	7.80E-06	7.82E-06	8.06E-06	6.81E-06	6.68E-06	6.79E-06	6.75E-06	6.68E-06	6.69E-06	7.12E-06

Table 5.4 – forecast MSEs from fixed order EGARCH models. IBOV index, Gaussian normalized innovations, $N = 500$.

We notice that EGARCH(1,1), the minimum order model, is no longer the best overall model, while EGARCH(1,2) model takes that place. Overfitting remains always worse than correct fitting as expected, while the frequency of correct fitting being the best choice increases from once for $N = 250$ (EGARCH(1,2) DGP scenario) to three times for $N = 500$ (we disregard EGARCH(1,1) DGP scenario since underfitting is not possible in this case). From asymptotic theory, we know that when N tends to infinity, correct fitting will outperform all other choices, so the result is not surprising. However, for daily observations, 500 samples corresponds to two years of data (only available in trading days), which may be a long period of time to expect for absence of changes in regime, reinforcing the already mentioned delicate tradeoff between data amount and stationarity premise. Nevertheless, for $N = 500$, underfitting outperforming correct fitting remains a very significant effect, and the amount of data needed for it to be negligible is considered to be a promising topic for future research.

Table 5.5 and Table 5.6 show the results of the model selection and averaging strategies for $N = 500$, in the same way as done in Table 5.2 and Table 5.3, respectively.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	2.35E-06	2.35E-06	2.55E-06	7.79E-06	5.47E-06	5.42E-06	1.11E-05	6.18E-06	6.10E-06	5.48E-06
BEST AIC	7.31E-06	7.25E-06	7.94E-06	6.56E-06	6.91E-06	7.22E-06	7.36E-06	7.40E-06	7.86E-06	7.31E-06
BEST SIC	3.30E-06	3.18E-06	3.65E-06	4.84E-06	6.05E-06	5.97E-06	6.62E-06	6.73E-06	6.81E-06	5.24E-06
AVG AIC	3.94E-06	3.89E-06	4.03E-06	4.00E-06	4.52E-06	4.68E-06	4.53E-06	4.83E-06	5.11E-06	4.39E-06
AVG SIC	2.92E-06	2.89E-06	3.02E-06	3.37E-06	4.04E-06	4.27E-06	4.26E-06	4.57E-06	4.91E-06	3.81E-06
AVG-E AIC	6.14E-06	6.08E-06	6.64E-06	5.44E-06	5.79E-06	6.07E-06	6.04E-06	6.17E-06	6.58E-06	6.11E-06
AVG-E SIC	3.04E-06	2.97E-06	3.37E-06	4.28E-06	5.42E-06	5.41E-06	5.87E-06	6.00E-06	6.16E-06	4.72E-06
SIMPLE AVG	3.19E-06	3.16E-06	3.23E-06	3.49E-06	4.04E-06	4.11E-06	4.20E-06	4.30E-06	4.64E-06	3.82E-06
OPT AVG	2.33E-06	2.33E-06	2.50E-06	2.87E-06	3.58E-06	3.81E-06	3.69E-06	4.00E-06	4.16E-06	3.25E-06

Table 5.5 – forecast MSEs from model selection and averaging strategies. IBOV index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.71%	0.79%	2.19%	171.65%	52.81%	42.31%	201.83%	54.65%	46.72%	63.74%
BEST AIC	213.62%	211.65%	217.72%	128.70%	92.86%	89.34%	99.26%	85.09%	89.06%	136.37%
BEST SIC	41.31%	36.73%	46.11%	68.65%	68.79%	56.55%	79.39%	68.35%	63.81%	58.85%
AVG AIC	68.87%	66.99%	61.05%	39.54%	26.22%	22.70%	22.61%	20.89%	23.00%	39.10%
AVG SIC	25.33%	24.36%	20.84%	17.58%	12.66%	12.11%	15.48%	14.40%	18.04%	17.87%
AVG-E AIC	163.13%	161.38%	165.58%	89.67%	61.76%	59.33%	63.62%	54.41%	58.25%	97.46%
AVG-E SIC	30.49%	27.58%	34.66%	49.32%	51.28%	42.05%	58.91%	50.06%	48.27%	43.62%
SIMPLE AVG	36.91%	35.94%	29.20%	21.70%	12.65%	7.93%	13.61%	7.63%	11.58%	19.68%

Table 5.6 – model selection and averaging strategies relative MSE losses (MSE_{RL}). IBOV index, Gaussian normalized innovations, N = 500.

We again kept the EGARCH(1,1) row for fixed order model reference, even it being inferior to EGARCH(1,2), because of the widespread use of the minimum model. In the first three scenarios, EGARCH(1,1) dominates the strategies as happened for N = 250, but with lower relative MSE losses. However, it gets much worse for the other six scenarios, where simple averaging and linear SIC averaging dominate, with the former being the best more often but the latter having the lowest MSE average across all scenarios. SIC linear averaging relative MSEs losses lie roughly in the range from 12% to 25%, which seems a desirable feature when compared to other strategies that can go over 35% (simple averaging), or much higher (strategies other than simple averaging and SIC linear averaging).

All scenarios and number of samples considered, we conjecture that SIC averaging seems to be the best overall strategy. It combines the information of all the models, having better performances than fixed order models or model selection approaches, which agrees to model averaging literature. Compared to AIC averaging approaches, SIC ones provide better weights in general, since SIC number of parameters

penalty is higher and thus underfitting models weights tend to be higher, when compared to AIC based averaging. Since underfitting tended to be better even than correct fitting, this explains SIC approaches outperformance. Linear SIC averaging was better than exponential SIC averaging, which is in agreement to model averaging literature that favors simpler weighting strategies in general, and linear over non-linear ones, in particular. Outperformance of SIC averaging over simple averaging was not so evident, but still true. Model averaging literature indicates that simple averaging, although seemingly simplistic, is difficult to outperform, due to the estimation errors due to weights calculations, which may be the reason for the difference being so slight. The following subsections, with data from other markets, will help to answer the question of which of those two strategies is the one with best potential.

5.1.2 Standard & Poor's 500 or S&P 500 (USA stock index),

For the American index S&P 500, the performances of fixed order EGARCH models in each DGP scenario, as presented in the previous subsection for IBOV index, are displayed in Table 5.7, for N = 250:

		(P,Q) of EGARCH DGP used to generate the data									
		(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(P,Q) of EGARCH fitted	(1,1)	3.17E-06	3.08E-06	3.30E-06	9.34E-06	4.70E-06	4.96E-06	1.34E-05	5.42E-06	5.32E-06	5.85E-06
	(2,1)	3.69E-06	3.60E-06	4.06E-06	1.37E-05	1.97E-05	8.08E-06	1.60E-05	4.21E-05	5.73E-06	1.30E-05
	(3,1)	4.67E-06	4.60E-06	4.66E-06	1.87E-05	1.96E-05	1.63E-05	1.97E-05	4.98E-05	1.54E-05	1.71E-05
	(1,2)	4.23E-06	4.15E-06	4.46E-06	3.80E-06	4.51E-06	5.02E-06	5.10E-06	5.21E-06	5.40E-06	4.65E-06
	(2,2)	5.04E-06	4.97E-06	5.47E-06	4.64E-06	5.09E-06	7.04E-06	6.00E-06	5.70E-06	6.28E-06	5.58E-06
	(3,2)	6.41E-06	6.30E-06	6.26E-06	6.18E-06	6.46E-06	6.33E-06	7.67E-06	7.18E-06	2.12E-05	8.22E-06
	(1,3)	5.35E-06	5.24E-06	5.27E-06	5.05E-06	5.28E-06	5.41E-06	5.27E-06	5.46E-06	6.04E-06	5.38E-06
	(2,3)	6.32E-06	6.37E-06	6.83E-06	6.09E-06	6.66E-06	8.96E-06	6.13E-06	6.28E-06	6.40E-06	6.67E-06
	(3,3)	8.11E-06	8.25E-06	8.15E-06	8.10E-06	8.29E-06	8.30E-06	7.93E-06	8.03E-06	7.55E-06	8.08E-06

Table 5.7 – forecast MSEs from fixed order EGARCH models. S&P 500 index, Gaussian normalized innovations, N = 250.

The results above are similar to those of the Brazilian market for N = 500, in terms of order of magnitude. The MSEs lie roughly in the range of 3E-6 to 6E-6 in both cases, while Brazilian MSEs for N = 250 were approximately twice that magnitude. This relationship of Brazilian MSEs for N = 250 and N = 500 is to be expected, since doubling the data halved the MSEs. However, the same halving effect is observed when the data amount is kept at 250 samples but the Brazilian index

based DGPs are replaced with American based ones. This suggests that the market maturity (higher in USA) may lead to more predictable logarithmic returns variances.

Overfitting remained worse for every scenario, whereas underfitting was the best option in seven out of the eight DGP scenarios in which it is possible. However, the outperformance of underfitting over correct fitting was slighter than in IBOV. With N = 250, EGARCH(1,2) was the best overall model, as in IBOV, N = 500 case.

The performances of EGARCH(1,1), model selection and averaging strategies are summarized in Table 5.8 and Table 5.9.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	3.17E-06	3.08E-06	3.30E-06	9.34E-06	4.70E-06	4.96E-06	1.34E-05	5.42E-06	5.32E-06	5.85E-06
BEST AIC	7.47E-06	7.60E-06	7.66E-06	7.99E-06	8.35E-06	8.65E-06	8.45E-06	8.45E-06	8.56E-06	8.13E-06
BEST SIC	4.23E-06	4.20E-06	4.46E-06	6.03E-06	6.03E-06	6.23E-06	7.31E-06	6.39E-06	6.51E-06	5.71E-06
AVG AIC	4.32E-06	4.33E-06	4.21E-06	4.68E-06	4.82E-06	5.02E-06	4.99E-06	4.74E-06	5.00E-06	4.68E-06
AVG SIC	3.42E-06	3.38E-06	3.40E-06	3.95E-06	4.09E-06	4.26E-06	4.54E-06	4.28E-06	4.51E-06	3.98E-06
AVG-E AIC	6.19E-06	6.37E-06	6.31E-06	6.65E-06	6.90E-06	7.19E-06	7.05E-06	6.99E-06	7.00E-06	6.74E-06
AVG-E SIC	3.84E-06	3.81E-06	3.99E-06	5.32E-06	5.33E-06	5.55E-06	6.37E-06	5.73E-06	5.80E-06	5.08E-06
SIMPLE AVG	3.63E-06	3.57E-06	3.50E-06	4.04E-06	4.27E-06	4.25E-06	4.47E-06	4.95E-06	4.45E-06	4.13E-06
OPT AVG	3.01E-06	2.91E-06	3.03E-06	3.53E-06	3.62E-06	3.82E-06	4.12E-06	3.95E-06	4.07E-06	3.56E-06

Table 5.8 – forecast MSEs from model selection and averaging strategies. S&P 500 index, Gaussian normalized innovations, N = 250.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	5.52%	5.71%	9.02%	164.50%	29.96%	29.80%	225.51%	37.26%	30.72%	59.78%
BEST AIC	148.49%	160.94%	153.20%	126.20%	131.08%	126.49%	105.22%	114.19%	110.41%	130.69%
BEST SIC	40.76%	44.36%	47.25%	70.65%	66.84%	63.05%	77.53%	61.87%	60.02%	59.15%
AVG AIC	43.60%	48.73%	39.19%	32.43%	33.40%	31.45%	21.25%	20.11%	22.93%	32.57%
AVG SIC	13.81%	16.25%	12.19%	11.88%	13.12%	11.57%	10.34%	8.52%	10.84%	12.06%
AVG-E AIC	105.93%	118.87%	108.54%	88.29%	90.78%	88.19%	71.19%	77.19%	72.07%	91.23%
AVG-E SIC	27.81%	30.81%	31.97%	50.70%	47.32%	45.27%	54.83%	45.09%	42.48%	41.81%
SIMPLE AVG	20.91%	22.78%	15.62%	14.45%	18.03%	11.13%	8.54%	25.32%	9.37%	16.24%

Table 5.9 – model selection and averaging strategies relative MSE losses (MSE_{RL}). S&P 500 index, Gaussian normalized innovations, N = 250.

Once again EGARCH(1,1) is the best performing choice for Q = 1 (first three DGPs), while simple averaging and linear SIC averaging strategies are the best overall options, especially for Q > 1. Regarding overall comparison of these two best performing strategies, SIC linear averaging is again better due to lower average MSE across DGPs and maximum relative MSE losses.

The results for N = 500 are shown in Table 5.10.

(P,Q) of EGARCH fitted	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(1,1)	1.11E-06	1.10E-06	1.29E-06	1.36E-05	2.48E-06	2.65E-06	3.86E-05	3.08E-06	2.99E-06	7.43E-06
(2,1)	1.37E-06	1.38E-06	1.53E-06	1.69E-05	2.81E-05	3.61E-06	4.97E-05	5.38E-05	3.17E-06	1.77E-05
(3,1)	1.82E-06	1.80E-06	1.90E-06	2.39E-05	2.79E-05	2.57E-05	5.53E-05	3.28E-05	1.36E-05	2.05E-05
(1,2)	1.58E-06	1.53E-06	1.73E-06	1.30E-06	1.84E-06	2.17E-06	2.32E-06	2.44E-06	2.52E-06	1.94E-06
(2,2)	1.97E-06	1.93E-06	2.10E-06	1.58E-06	1.65E-06	3.25E-06	2.74E-06	2.75E-06	2.89E-06	2.32E-06
(3,2)	2.66E-06	2.59E-06	2.71E-06	2.15E-06	2.06E-06	2.23E-06	3.82E-06	2.68E-06	1.92E-05	4.46E-06
(1,3)	2.06E-06	2.01E-06	2.11E-06	1.73E-06	1.92E-06	2.01E-06	1.91E-06	2.12E-06	2.36E-06	2.03E-06
(2,3)	2.47E-06	2.45E-06	2.56E-06	2.17E-06	2.19E-06	3.23E-06	2.23E-06	2.31E-06	2.56E-06	2.46E-06
(3,3)	3.54E-06	3.60E-06	3.66E-06	2.98E-06	2.80E-06	2.97E-06	2.91E-06	2.89E-06	2.76E-06	3.12E-06

Table 5.10 – forecast MSEs from fixed order EGARCH models. S&P 500 index, Gaussian normalized innovations, $N = 500$.

The results for fixed EGARCH models for $N = 500$, when compared to results for $N = 250$, display a decrease in MSE magnitude and underfitting outperformance over correct fitting, both to be expected from the increased number of samples, as already found and discussed for the IBOV index case.

Although overfitting performance remained worse in general, it is noted that for an EGARCH(3,2), for the first time, the best choice was neither a correct fitting nor underfitting, but EGARCH(1,3) which is a mixture of overfitting (Q strictly higher than in DGP) and underfitting (P strictly lower than in DGP). The model that has the same underfitting attribute (P = 1) but no overfitting is EGARCH(1,2) which is the second best option. This shows that higher order effects (in this case, third order) caused by one order parameter (P in this case) can be reproduced through the other order parameter (Q in this case). The choice of the wrong order parameter causing better MSE performance can be due to its better behavior in terms of small sample parameter estimation errors and their impact on forecast MSE.

The results obtained using the model averaging strategies and $N = 500$ are shown in Table 5.11 and Table 5.12.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.11E-06	1.10E-06	1.29E-06	1.36E-05	2.48E-06	2.65E-06	3.86E-05	3.08E-06	2.99E-06	7.43E-06
BEST AIC	3.19E-06	3.21E-06	3.42E-06	2.73E-06	2.70E-06	2.99E-06	3.06E-06	3.07E-06	3.21E-06	3.06E-06
BEST SIC	1.35E-06	1.34E-06	1.56E-06	2.06E-06	2.55E-06	2.82E-06	2.98E-06	3.19E-06	3.20E-06	2.34E-06
AVG AIC	1.83E-06	1.81E-06	1.89E-06	1.69E-06	1.73E-06	1.94E-06	1.97E-06	1.94E-06	2.11E-06	1.88E-06
AVG SIC	1.37E-06	1.35E-06	1.43E-06	1.53E-06	1.63E-06	1.79E-06	1.95E-06	1.93E-06	2.16E-06	1.68E-06
AVG-E AIC	2.68E-06	2.68E-06	2.84E-06	2.30E-06	2.26E-06	2.53E-06	2.57E-06	2.60E-06	2.72E-06	2.57E-06
AVG-E SIC	1.29E-06	1.28E-06	1.48E-06	1.85E-06	2.22E-06	2.47E-06	2.58E-06	2.77E-06	2.83E-06	2.09E-06
SIMPLE AVG	1.52E-06	1.49E-06	1.52E-06	2.12E-06	2.29E-06	2.08E-06	3.81E-06	2.66E-06	2.26E-06	2.20E-06
OPT AVG	1.10E-06	1.09E-06	1.24E-06	1.28E-06	1.41E-06	1.60E-06	1.71E-06	1.75E-06	1.87E-06	1.45E-06

Table 5.11 – forecast MSEs from model selection and averaging strategies. S&P 500 index, Gaussian normalized innovations, $N = 500$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.43%	0.46%	4.23%	962.00%	76.26%	66.06%	2150.58%	76.55%	60.15%	377.41%
BEST AIC	189.51%	193.83%	175.97%	112.26%	92.28%	87.53%	78.35%	75.86%	71.79%	119.71%
BEST SIC	22.57%	22.97%	26.24%	60.50%	81.12%	76.92%	73.64%	82.63%	71.29%	57.54%
AVG AIC	65.80%	65.11%	52.71%	31.39%	22.86%	21.67%	14.90%	11.36%	12.87%	33.19%
AVG SIC	24.76%	23.42%	15.31%	18.78%	16.07%	11.84%	13.73%	10.32%	15.81%	16.67%
AVG-E AIC	143.17%	145.14%	129.48%	78.90%	60.47%	58.52%	50.04%	48.75%	45.48%	84.44%
AVG-E SIC	17.31%	17.04%	19.18%	43.78%	58.07%	54.90%	50.78%	58.62%	51.48%	41.24%
SIMPLE AVG	38.22%	36.61%	22.92%	65.19%	62.88%	30.59%	122.42%	52.18%	20.90%	50.21%

Table 5.12 – model selection and averaging strategies relative MSE losses (MSE_{RL}). S&P 500 index, Gaussian normalized innovations, $N = 500$.

These results favor once again the minimum model for Q = 1 and linear SIC averaging for overall performances. It is noted, however, that linear SIC averaging outperformance over simple averaging is clearly higher, which reinforces its best overall (across several indexes and number of samples) potential.

AIC linear averaging shows for the first time the second best overall performance (instead of simple averaging) and the best performance for a specific DGP – EGARCH(3,3). These can be regarded as consequences of weakening of underfitting outperformances as the number of samples increases, and as the market maturity increases, which seems to have a similar effect to an increase in the number of samples. When the benefits of underfitting are smaller, the outperformances of SIC or simple averaging over AIC are also less pronounced, since they stem from the smaller number of parameters penalty imposed by AIC.

5.1.3 Nikkei 225 or N225 (Japan's stock index)

For the Japanese index N225, the performances of fixed order EGARCH models in each DGP scenario, as presented in the previous subsections, are displayed in Table 5.13, for $N = 250$:

(P,Q) of EGARCH fitted	(P,Q) of EGARCH DGP used to generate the data: (1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(1,1)	3.74E-06	3.66E-06	3.81E-06	4.71E-06	5.83E-06	5.97E-06	5.67E-06	6.41E-06	6.40E-06	5.13E-06
(2,1)	5.08E-06	4.96E-06	5.57E-06	6.03E-06	1.41E-05	7.36E-06	6.99E-06	2.46E-05	7.80E-06	9.16E-06
(3,1)	7.17E-06	6.90E-06	7.13E-06	8.25E-06	1.12E-05	1.41E-05	9.34E-06	1.46E-05	1.45E-05	1.04E-05
(1,2)	5.43E-06	5.27E-06	5.41E-06	5.02E-06	6.41E-06	6.98E-06	6.19E-06	6.84E-06	6.88E-06	6.05E-06
(2,2)	7.34E-06	7.31E-06	7.82E-06	6.72E-06	8.06E-06	8.96E-06	7.87E-06	8.50E-06	8.75E-06	7.93E-06
(3,2)	9.97E-06	9.94E-06	1.03E-05	9.55E-06	1.03E-05	1.01E-05	1.06E-05	1.08E-05	1.10E-05	1.03E-05
(1,3)	7.18E-06	7.08E-06	7.10E-06	6.74E-06	7.70E-06	7.87E-06	6.94E-06	7.97E-06	8.01E-06	7.40E-06
(2,3)	9.43E-06	9.43E-06	1.06E-05	8.98E-06	1.07E-05	1.03E-05	8.90E-06	1.05E-05	1.01E-05	9.87E-06
(3,3)	1.30E-05	1.30E-05	1.31E-05	1.23E-05	1.33E-05	1.34E-05	1.23E-05	1.31E-05	1.31E-05	1.30E-05

Table 5.13 – forecast MSEs from fixed order EGARCH models. N225 index, Gaussian normalized innovations, $N = 250$.

For the Japanese index, the minimum order model shows the best performance for all DGPs, which is an exceptionally high underfitting outperformance scenario when compared to other indexes examined up to this point.

The ubiquitous outperformance of the minimum order model makes the N225 case not only a clear exception but also a stress scenario for the hypothesis of model averaging usefulness, since it could lead to performance losses over the widespread use of the minimum order model only. To analyze this question, the performances of EGARCH(1,1), model selection and averaging strategies are summarized in Table 5.14 and Table 5.15.

	(P,Q) of EGARCH DGP used to generate the data: (1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	3.74E-06	3.66E-06	3.81E-06	4.71E-06	5.83E-06	5.97E-06	5.67E-06	6.41E-06	6.40E-06	5.13E-06
BEST AIC	1.24E-05	1.23E-05	1.28E-05	1.22E-05	1.34E-05	1.38E-05	1.29E-05	1.38E-05	1.38E-05	1.30E-05
BEST SIC	6.41E-06	6.45E-06	6.90E-06	7.57E-06	8.71E-06	8.58E-06	8.47E-06	9.23E-06	8.96E-06	7.92E-06
AVG AIC	6.40E-06	6.35E-06	6.32E-06	6.48E-06	7.55E-06	7.47E-06	7.08E-06	7.68E-06	7.60E-06	6.99E-06
AVG SIC	4.72E-06	4.62E-06	4.68E-06	4.86E-06	5.84E-06	5.99E-06	5.60E-06	6.14E-06	6.21E-06	5.41E-06
AVG-E AIC	1.04E-05	1.04E-05	1.08E-05	1.02E-05	1.12E-05	1.15E-05	1.06E-05	1.14E-05	1.15E-05	1.09E-05
AVG-E SIC	5.66E-06	5.63E-06	5.97E-06	6.54E-06	7.66E-06	7.55E-06	7.47E-06	8.09E-06	7.93E-06	6.94E-06
SIMPLE AVG	5.02E-06	4.91E-06	4.88E-06	4.90E-06	5.76E-06	5.88E-06	5.42E-06	6.08E-06	6.03E-06	5.43E-06
OPT AVG	3.68E-06	3.57E-06	3.69E-06	4.05E-06	5.01E-06	5.15E-06	4.75E-06	5.35E-06	5.39E-06	4.52E-06

Table 5.14 – forecast MSEs from model selection and averaging strategies. N225 index, Gaussian normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.70%	2.42%	3.40%	16.15%	16.36%	15.88%	19.35%	19.93%	18.68%	12.65%
BEST AIC	236.22%	244.31%	248.03%	201.62%	166.54%	167.89%	170.77%	158.41%	155.97%	194.42%
BEST SIC	74.31%	80.43%	87.24%	86.71%	73.67%	66.38%	78.15%	72.58%	66.01%	76.16%
AVG AIC	74.07%	77.55%	71.53%	59.88%	50.55%	44.92%	48.98%	43.71%	40.88%	56.90%
AVG SIC	28.31%	29.25%	26.95%	19.87%	16.44%	16.22%	17.85%	14.82%	15.11%	20.54%
AVG-E AIC	182.76%	190.34%	191.60%	151.66%	123.60%	122.61%	123.96%	114.17%	112.66%	145.93%
AVG-E SIC	53.98%	57.42%	62.05%	61.32%	52.72%	46.48%	57.05%	51.30%	47.01%	54.37%
SIMPLE AVG	36.38%	37.30%	32.32%	20.97%	14.81%	14.13%	14.08%	13.74%	11.72%	21.72%

Table 5.15 – model selection and averaging strategies relative MSE losses (MSE_{RL}). N225 index, Gaussian normalized innovations, N = 250.

Once again EGARCH(1,1) is the best performing choice for Q = 1 (first three DGPs), while simple averaging is the best for Q > 1. Although the minimum model leads to the best average overall option, its underperformance over simple averaging for Q > 1 is surprising since it is the best fixed order model for all DGPs, which reinforces the model averaging potential. Therefore, when all indexes and number of samples are considered, this exception does not change the general conjecture that model averaging is better than any single model usage.

Regarding averaging strategies only, SIC linear averaging yielded the lowest average across DGPs relative MSE loss (20.54% against 21.72% of simple averaging), and the least maximum relative MSE loss (29.25% against 37.30% of simple averaging), which is a desirable consistency feature. Therefore, we conclude that the Japanese index with N = 250 did favor SIC linear averaging over simple averaging, although only slightly.

Table 5.16 shows the results for N = 500.

		(P,Q) of EGARCH DGP used to generate the data									
		(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(P,Q) of EGARCH fitted	(1,1)	1.39E-06	1.40E-06	1.53E-06	2.35E-06	3.34E-06	3.44E-06	3.19E-06	4.03E-06	4.01E-06	2.74E-06
	(2,1)	1.94E-06	1.93E-06	2.03E-06	2.91E-06	1.28E-05	4.04E-06	3.70E-06	2.02E-05	4.54E-06	6.01E-06
	(3,1)	2.85E-06	2.83E-06	2.91E-06	4.14E-06	7.00E-06	7.53E-06	4.80E-06	1.03E-05	8.90E-06	5.70E-06
	(1,2)	2.07E-06	2.04E-06	2.11E-06	1.89E-06	3.05E-06	3.45E-06	2.93E-06	3.68E-06	3.69E-06	2.77E-06
	(2,2)	2.76E-06	2.74E-06	2.81E-06	2.43E-06	2.74E-06	3.97E-06	3.29E-06	3.37E-06	4.28E-06	3.15E-06
	(3,2)	4.09E-06	4.04E-06	4.10E-06	3.56E-06	3.48E-06	3.74E-06	4.46E-06	4.17E-06	4.23E-06	3.99E-06
	(1,3)	2.75E-06	2.74E-06	2.73E-06	2.53E-06	3.19E-06	3.42E-06	2.72E-06	3.68E-06	3.64E-06	3.04E-06
	(2,3)	3.53E-06	3.53E-06	3.62E-06	3.26E-06	3.62E-06	4.28E-06	3.32E-06	3.76E-06	4.48E-06	3.71E-06
	(3,3)	5.39E-06	5.34E-06	5.48E-06	4.76E-06	4.65E-06	4.83E-06	4.68E-06	4.71E-06	4.74E-06	4.95E-06

Table 5.16 – forecast MSEs from fixed order EGARCH models. N225 index, Gaussian normalized innovations, N = 500.

The results for fixed EGARCH models for N = 500, when compared to N = 250, display diminishing MSE magnitudes and underfitting outperformances over correct fitting, both to be expected from the larger number of samples, as already found and discussed for IBOV and S&P 500 indexes.

For EGARCH(3,2) DGP, the best choice was neither correct fitting nor underfitting, but EGARCH(1,3), which is a mixture of overfitting (Q strictly higher than in DGP) and underfitting (P strictly lower than in DGP). That is exactly the same phenomenon observed and discussed for S&P 500, N = 500.

The results obtained using model averaging for N = 500 are shown in Table 5.17 and Table 5.18.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.39E-06	1.40E-06	1.53E-06	2.35E-06	3.34E-06	3.44E-06	3.19E-06	4.03E-06	4.01E-06	2.74E-06
BEST AIC	4.92E-06	4.91E-06	5.20E-06	4.56E-06	4.67E-06	5.16E-06	5.11E-06	5.17E-06	5.35E-06	5.00E-06
BEST SIC	1.91E-06	1.98E-06	2.14E-06	2.82E-06	3.78E-06	3.91E-06	3.79E-06	4.44E-06	4.44E-06	3.25E-06
AVG AIC	2.55E-06	2.55E-06	2.59E-06	2.60E-06	2.96E-06	3.21E-06	3.04E-06	3.33E-06	3.44E-06	2.92E-06
AVG SIC	1.85E-06	1.83E-06	1.86E-06	2.03E-06	2.59E-06	2.92E-06	2.64E-06	3.05E-06	3.23E-06	2.45E-06
AVG-E AIC	4.11E-06	4.06E-06	4.29E-06	3.70E-06	3.86E-06	4.30E-06	4.11E-06	4.28E-06	4.46E-06	4.13E-06
AVG-E SIC	1.78E-06	1.82E-06	1.98E-06	2.52E-06	3.39E-06	3.59E-06	3.40E-06	4.01E-06	4.05E-06	2.95E-06
SIMPLE AVG	2.06E-06	2.05E-06	2.03E-06	2.05E-06	2.57E-06	2.77E-06	2.47E-06	3.02E-06	3.02E-06	2.45E-06
OPT AVG	1.39E-06	1.39E-06	1.49E-06	1.63E-06	2.19E-06	2.40E-06	2.08E-06	2.65E-06	2.66E-06	1.99E-06

Table 5.17 – forecast MSEs from model selection and averaging strategies. N225 index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.17%	0.50%	2.43%	44.00%	52.58%	43.51%	53.61%	52.15%	50.67%	33.29%
BEST AIC	254.54%	253.11%	248.03%	179.70%	113.38%	114.93%	145.69%	95.23%	100.98%	167.29%
BEST SIC	37.39%	42.35%	43.10%	72.81%	72.83%	62.89%	82.44%	67.78%	66.87%	60.94%
AVG AIC	83.98%	83.75%	73.34%	59.61%	35.43%	33.69%	46.08%	25.85%	29.29%	52.34%
AVG SIC	33.21%	31.93%	24.26%	24.32%	18.57%	21.78%	27.25%	15.32%	21.40%	24.23%
AVG-E AIC	196.00%	191.78%	187.08%	126.88%	76.37%	79.18%	97.93%	61.57%	67.43%	120.47%
AVG-E SIC	28.56%	30.73%	32.37%	54.19%	54.89%	49.73%	63.81%	51.32%	52.01%	46.40%
SIMPLE AVG	48.60%	47.35%	35.62%	25.55%	17.61%	15.46%	18.63%	13.91%	13.54%	26.25%

Table 5.18 – model selection and averaging strategies relative MSE losses (MSE_{RL}). N225 index, Gaussian normalized innovations, N = 500.

These results favor the minimum model for Q = 1 and simple averaging for Q > 1. However, regarding overall performance, linear SIC averaging is best with lowest mean and maximum across DGPs relative MSE losses. It is the best option just once (EGARCH(1,2) DGP) but its consistent second best performance (losing to EGARCH(1,1)

when Q = 1 and for simple averaging when Q > 1) makes it the best overall strategy.

5.1.4 DAX (Germany's stock index)

For the German index DAX, the performances of fixed order EGARCH models in each DGP scenario, as presented in the previous subsections, are displayed in Table 5.19, for $N = 250$. The performances of EGARCH(1,1), model selection and averaging strategies are compared in Table 5.20 and Table 5.21, also for $N = 250$.

		(P,Q) of EGARCH DGP used to generate the data									
		(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(P,Q) of EGARCH fitted	(1,1)	4.32E-06	4.29E-06	4.43E-06	8.01E-06	6.26E-06	6.27E-06	8.44E-06	6.37E-06	6.08E-06	6.05E-06
	(2,1)	5.17E-06	5.14E-06	5.89E-06	1.14E-05	2.84E-05	3.21E-05	1.43E-05	3.80E-05	7.12E-06	1.64E-05
	(3,1)	6.81E-06	6.67E-06	6.88E-06	1.50E-05	2.30E-05	2.73E-05	1.62E-05	2.81E-05	3.04E-05	1.78E-05
	(1,2)	5.84E-06	5.80E-06	5.92E-06	5.47E-06	6.86E-06	7.05E-06	6.32E-06	6.46E-06	6.70E-06	6.27E-06
	(2,2)	7.21E-06	7.09E-06	8.10E-06	6.79E-06	7.68E-06	7.80E-06	7.65E-06	7.82E-06	8.37E-06	7.61E-06
	(3,2)	9.20E-06	9.21E-06	9.25E-06	8.97E-06	9.17E-06	9.22E-06	9.88E-06	9.54E-06	9.97E-06	9.38E-06
	(1,3)	7.52E-06	7.47E-06	7.39E-06	7.17E-06	8.07E-06	8.20E-06	7.32E-06	7.89E-06	8.10E-06	7.68E-06
	(2,3)	9.09E-06	9.06E-06	1.09E-05	8.82E-06	9.70E-06	9.94E-06	8.78E-06	9.52E-06	9.76E-06	9.51E-06
	(3,3)	1.16E-05	1.19E-05	1.20E-05	1.17E-05	1.16E-05	1.17E-05	1.14E-05	1.10E-05	1.14E-05	1.16E-05

Table 5.19 – forecast MSEs from fixed order EGARCH models. DAX index, Gaussian normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	4.32E-06	4.29E-06	4.43E-06	8.01E-06	6.26E-06	6.27E-06	8.44E-06	6.37E-06	6.08E-06	6.05E-06
BEST AIC	1.10E-05	1.11E-05	1.15E-05	1.16E-05	1.18E-05	1.18E-05	1.20E-05	1.18E-05	1.21E-05	1.16E-05
BEST SIC	6.16E-06	6.16E-06	6.55E-06	8.28E-06	8.16E-06	8.21E-06	8.87E-06	8.33E-06	7.89E-06	7.62E-06
AVG AIC	6.12E-06	6.14E-06	5.99E-06	6.65E-06	6.96E-06	6.95E-06	6.90E-06	6.77E-06	6.81E-06	6.59E-06
AVG SIC	4.79E-06	4.77E-06	4.75E-06	5.45E-06	5.73E-06	5.76E-06	5.84E-06	5.68E-06	5.85E-06	5.40E-06
AVG-E AIC	9.01E-06	9.18E-06	9.54E-06	9.61E-06	9.74E-06	9.76E-06	9.87E-06	9.65E-06	9.95E-06	9.59E-06
AVG-E SIC	5.53E-06	5.54E-06	5.79E-06	7.29E-06	7.23E-06	7.29E-06	7.83E-06	7.40E-06	7.11E-06	6.78E-06
SIMPLE AVG	5.07E-06	5.03E-06	4.91E-06	5.42E-06	5.88E-06	5.97E-06	5.70E-06	5.93E-06	5.92E-06	5.54E-06
OPT AVG	4.11E-06	4.07E-06	4.13E-06	4.87E-06	5.18E-06	5.21E-06	5.34E-06	5.09E-06	5.25E-06	4.81E-06

Table 5.20 – forecast MSEs from model selection and averaging strategies. DAX index, Gaussian normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	5.12%	5.42%	7.41%	64.45%	20.78%	20.28%	58.16%	25.02%	15.84%	24.72%
BEST AIC	166.81%	172.26%	179.81%	137.72%	126.84%	126.86%	125.47%	131.33%	130.26%	144.15%
BEST SIC	49.70%	51.43%	58.72%	70.00%	57.46%	57.60%	66.20%	63.69%	50.23%	58.34%
AVG AIC	48.73%	50.95%	45.33%	36.59%	34.28%	33.26%	29.20%	32.94%	29.71%	37.89%
AVG SIC	16.44%	17.15%	15.18%	11.98%	10.47%	10.58%	9.35%	11.61%	11.41%	12.69%
AVG-E AIC	118.99%	125.78%	131.23%	97.39%	88.01%	87.33%	84.88%	89.43%	89.48%	101.39%
AVG-E SIC	34.44%	36.27%	40.41%	49.75%	39.44%	39.89%	46.61%	45.30%	35.46%	40.84%
SIMPLE AVG	23.37%	23.58%	18.94%	11.28%	13.55%	14.60%	6.73%	16.45%	12.84%	15.70%

Table 5.21 – model selection and averaging strategies relative MSE losses (MSE_{RL}). DAX index, Gaussian normalized innovations, N = 250.

The results for N = 500 are shown in Table 5.22 to Table 5.24.

		(P,Q) of EGARCH DGP used to generate the data									
		(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(P,Q) of EGARCH fitted	(1,1)	1.52E-06	1.55E-06	1.71E-06	7.00E-06	3.51E-06	3.56E-06	6.10E-06	3.75E-06	3.60E-06	3.59E-06
	(2,1)	1.93E-06	1.97E-06	2.13E-06	9.89E-06	2.73E-05	3.33E-05	1.08E-05	1.24E-05	4.01E-06	1.15E-05
	(3,1)	2.73E-06	2.65E-06	2.78E-06	1.54E-05	1.92E-05	2.22E-05	1.67E-05	8.85E-06	4.73E-05	1.53E-05
	(1,2)	2.24E-06	2.18E-06	2.32E-06	1.94E-06	3.29E-06	3.41E-06	2.70E-06	3.19E-06	3.39E-06	2.74E-06
	(2,2)	2.82E-06	2.75E-06	2.93E-06	2.36E-06	2.69E-06	2.76E-06	3.04E-06	3.16E-06	3.99E-06	2.95E-06
	(3,2)	3.91E-06	3.87E-06	3.91E-06	3.22E-06	3.28E-06	3.39E-06	3.90E-06	3.79E-06	4.09E-06	3.71E-06
	(1,3)	2.86E-06	2.86E-06	2.90E-06	2.57E-06	3.41E-06	3.54E-06	2.73E-06	3.51E-06	3.71E-06	3.12E-06
	(2,3)	3.48E-06	3.47E-06	3.71E-06	3.22E-06	3.50E-06	3.60E-06	3.27E-06	3.55E-06	4.35E-06	3.57E-06
	(3,3)	5.13E-06	5.14E-06	5.22E-06	4.48E-06	4.27E-06	4.33E-06	4.38E-06	4.21E-06	4.27E-06	4.60E-06

Table 5.22 – forecast MSEs from fixed order EGARCH models. DAX index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.52E-06	1.55E-06	1.71E-06	7.00E-06	3.51E-06	3.56E-06	6.10E-06	3.75E-06	3.60E-06	3.59E-06
BEST AIC	4.70E-06	4.65E-06	4.95E-06	4.25E-06	4.29E-06	4.44E-06	4.69E-06	4.53E-06	4.80E-06	4.59E-06
BEST SIC	1.95E-06	1.99E-06	2.15E-06	3.07E-06	3.78E-06	3.91E-06	3.90E-06	4.11E-06	3.93E-06	3.20E-06
AVG AIC	2.59E-06	2.57E-06	2.61E-06	2.62E-06	2.91E-06	2.97E-06	2.89E-06	3.06E-06	3.16E-06	2.82E-06
AVG SIC	1.93E-06	1.92E-06	1.95E-06	2.22E-06	2.60E-06	2.68E-06	2.62E-06	2.87E-06	3.03E-06	2.42E-06
AVG-E AIC	3.93E-06	3.89E-06	4.10E-06	3.55E-06	3.62E-06	3.74E-06	3.84E-06	3.80E-06	4.07E-06	3.84E-06
AVG-E SIC	1.84E-06	1.86E-06	2.02E-06	2.75E-06	3.42E-06	3.51E-06	3.48E-06	3.68E-06	3.64E-06	2.91E-06
SIMPLE AVG	2.14E-06	2.12E-06	2.09E-06	2.44E-06	2.81E-06	2.97E-06	2.72E-06	2.79E-06	3.38E-06	2.61E-06
OPT AVG	1.52E-06	1.55E-06	1.66E-06	1.90E-06	2.28E-06	2.34E-06	2.32E-06	2.51E-06	2.62E-06	2.08E-06

Table 5.23 – forecast MSEs from model selection and averaging strategies. DAX index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.33%	0.44%	2.93%	269.29%	54.05%	52.19%	162.59%	49.73%	37.49%	69.89%
BEST AIC	209.40%	201.14%	198.47%	124.29%	88.29%	89.59%	102.12%	80.57%	83.60%	130.83%
BEST SIC	28.43%	28.90%	29.65%	61.75%	66.13%	66.97%	68.10%	63.97%	50.27%	51.57%
AVG AIC	70.81%	66.50%	57.70%	37.93%	27.93%	26.84%	24.68%	22.17%	20.76%	39.48%
AVG SIC	27.25%	24.01%	17.63%	17.05%	13.93%	14.43%	13.02%	14.41%	15.86%	17.51%
AVG-E AIC	158.60%	151.56%	147.18%	87.28%	58.81%	59.81%	65.27%	51.57%	55.42%	92.83%
AVG-E SIC	21.36%	20.27%	22.16%	44.83%	50.13%	49.91%	50.06%	46.59%	38.99%	38.26%
SIMPLE AVG	40.63%	36.93%	26.32%	28.81%	23.48%	26.91%	16.99%	11.29%	29.33%	26.74%

Table 5.24 – model selection and averaging strategies relative MSE losses (MSE_{RL}). DAX index, Gaussian normalized innovations, N = 500.

The German index situation displays basically the same behavior already depicted in previous subsections: worse performances of overfitting, benefits of underfitting (although diminishing with an increase in the number of samples) and MSE magnitude reduction as the number of samples increases. Regarding the forecasting strategies, for both number of samples (250 and 500), the dominance of linear SIC averaging is clear. It has the lowest overall (across DGPs) MSE losses, with both the lowest overall MSE losses means and maximums.

5.1.5 FTSE 100 (England's stock index).

For the British index FTSE 100, the performances of fixed order EGARCH models in each DGP scenario, as presented in the previous subsections, are displayed in Table 5.25, for N = 250. The performances of EGARCH(1,1), model selection and averaging strategies are compared in Table 5.26 and Table 5.27, also for N = 250.

		(P,Q) of EGARCH DGP used to generate the data									
		(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(P,Q) of EGARCH fitted	(1,1)	3.04E-06	2.99E-06	3.06E-06	3.14E-06	3.27E-06	3.50E-06	3.33E-06	3.35E-06	3.40E-06	3.23E-06
	(2,1)	3.56E-06	3.53E-06	3.78E-06	3.85E-06	4.78E-06	4.11E-06	3.95E-06	4.29E-06	3.93E-06	3.98E-06
	(3,1)	4.58E-06	4.51E-06	4.67E-06	4.78E-06	4.93E-06	2.21E-05	5.01E-06	4.83E-06	4.93E-06	6.70E-06
	(1,2)	4.09E-06	3.96E-06	4.02E-06	3.79E-06	4.49E-06	4.37E-06	4.00E-06	4.15E-06	4.21E-06	4.12E-06
	(2,2)	4.89E-06	4.87E-06	5.04E-06	4.66E-06	4.99E-06	5.47E-06	4.81E-06	4.86E-06	4.98E-06	4.95E-06
	(3,2)	6.22E-06	6.25E-06	6.33E-06	6.12E-06	5.90E-06	5.54E-06	6.22E-06	5.95E-06	6.10E-06	6.07E-06
	(1,3)	5.25E-06	5.00E-06	5.11E-06	4.93E-06	5.56E-06	5.36E-06	5.03E-06	5.23E-06	5.35E-06	5.20E-06
	(2,3)	6.17E-06	6.06E-06	6.47E-06	5.96E-06	6.11E-06	6.58E-06	5.96E-06	5.82E-06	5.88E-06	6.11E-06
	(3,3)	7.84E-06	7.82E-06	7.97E-06	7.76E-06	7.39E-06	6.78E-06	7.80E-06	7.03E-06	7.77E-06	7.57E-06

Table 5.25 – forecast MSEs from fixed order EGARCH models. FTSE 100 index, Gaussian normalized innovations, N = 250.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	3.04E-06	2.99E-06	3.06E-06	3.14E-06	3.27E-06	3.50E-06	3.33E-06	3.35E-06	3.40E-06	3.23E-06
BEST AIC	7.29E-06	7.23E-06	7.53E-06	7.27E-06	6.79E-06	7.14E-06	7.48E-06	6.78E-06	7.37E-06	7.21E-06
BEST SIC	4.12E-06	4.16E-06	4.35E-06	4.33E-06	4.33E-06	4.47E-06	4.53E-06	4.36E-06	4.45E-06	4.34E-06
AVG AIC	4.15E-06	4.09E-06	4.15E-06	4.11E-06	3.95E-06	4.11E-06	4.23E-06	3.90E-06	4.00E-06	4.08E-06
AVG SIC	3.29E-06	3.24E-06	3.27E-06	3.22E-06	3.23E-06	3.50E-06	3.37E-06	3.21E-06	3.27E-06	3.29E-06
AVG-E AIC	6.03E-06	5.94E-06	6.24E-06	5.99E-06	5.58E-06	5.79E-06	6.11E-06	5.53E-06	6.07E-06	5.92E-06
AVG-E SIC	3.75E-06	3.75E-06	3.90E-06	3.84E-06	3.82E-06	4.06E-06	4.05E-06	3.89E-06	3.96E-06	3.89E-06
SIMPLE AVG	3.49E-06	3.41E-06	3.43E-06	3.37E-06	3.31E-06	3.59E-06	3.47E-06	3.26E-06	3.33E-06	3.41E-06
OPT AVG	2.87E-06	2.81E-06	2.87E-06	2.86E-06	2.82E-06	3.10E-06	3.01E-06	2.87E-06	2.99E-06	2.91E-06

Table 5.26 – forecast MSEs from model selection and averaging strategies. FTSE 100 index, Gaussian normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	5.87%	6.56%	6.68%	9.70%	15.82%	12.80%	10.54%	16.70%	13.73%	10.93%
BEST AIC	153.48%	157.11%	162.56%	153.83%	140.71%	130.26%	148.26%	136.07%	146.27%	147.62%
BEST SIC	43.39%	48.21%	51.57%	51.27%	53.66%	43.96%	50.25%	51.84%	48.64%	49.20%
AVG AIC	44.53%	45.65%	44.67%	43.66%	40.03%	32.53%	40.40%	35.76%	33.57%	40.09%
AVG SIC	14.29%	15.13%	14.11%	12.65%	14.45%	12.89%	11.73%	11.86%	9.28%	12.93%
AVG-E AIC	109.71%	111.33%	117.50%	109.40%	97.74%	86.74%	102.95%	92.65%	103.04%	103.45%
AVG-E SIC	30.37%	33.31%	35.91%	34.07%	35.32%	30.98%	34.40%	35.46%	32.30%	33.57%
SIMPLE AVG	21.44%	21.24%	19.43%	17.76%	17.23%	15.71%	15.14%	13.51%	11.27%	16.97%

Table 5.27 – model selection and averaging strategies relative MSE losses (MSE_{RL}). FTSE 100 index, Gaussian normalized innovations, $N = 250$.

The results are similar to the Japanese case, showing clear dominance of the minimum order model over all other fixed order models for every DGP scenario and over the other forecast strategies (model selection and averaging) for six out of nine DGPs, with the lowest average across DPGs relative MSE loss (10.93% against 12.93% of second best strategy, linear SIC averaging). Nevertheless SIC linear averaging yields the lowest maximum MSE loss across DGPs (15.13% against 16.70% of EGARCH(1,1)).

Table 5.28 to Table 5.30 show the results for $N = 500$.

(P,Q) of EGARCH fitted	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
(1,1)	1.11E-06	1.11E-06	1.13E-06	1.27E-06	1.16E-06	1.91E-06	1.42E-06	1.32E-06	1.36E-06	1.31E-06
(2,1)	1.37E-06	1.38E-06	1.38E-06	1.57E-06	2.10E-06	2.08E-06	1.75E-06	1.87E-06	1.62E-06	1.68E-06
(3,1)	1.86E-06	1.85E-06	1.86E-06	2.16E-06	2.09E-06	2.32E-05	2.29E-06	2.19E-06	2.18E-06	4.41E-06
(1,2)	1.58E-06	1.53E-06	1.54E-06	1.42E-06	1.59E-06	2.22E-06	1.57E-06	1.49E-06	1.53E-06	1.61E-06
(2,2)	1.97E-06	1.94E-06	1.93E-06	1.72E-06	1.92E-06	2.63E-06	1.91E-06	1.96E-06	2.05E-06	2.00E-06
(3,2)	2.65E-06	2.62E-06	2.62E-06	2.44E-06	2.36E-06	2.44E-06	2.57E-06	2.52E-06	2.68E-06	2.54E-06
(1,3)	2.06E-06	2.01E-06	1.98E-06	1.85E-06	2.08E-06	2.54E-06	1.96E-06	1.97E-06	2.03E-06	2.05E-06
(2,3)	2.46E-06	2.42E-06	2.40E-06	2.25E-06	2.40E-06	3.05E-06	2.33E-06	2.36E-06	2.28E-06	2.44E-06
(3,3)	3.45E-06	3.54E-06	3.46E-06	3.24E-06	3.18E-06	3.02E-06	3.29E-06	3.02E-06	3.33E-06	3.28E-06

Table 5.28 – forecast MSEs from fixed order EGARCH models. FTSE 100 index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.11E-06	1.11E-06	1.13E-06	1.27E-06	1.16E-06	1.91E-06	1.42E-06	1.32E-06	1.36E-06	1.31E-06
BEST AIC	3.14E-06	3.19E-06	3.23E-06	3.06E-06	2.95E-06	3.19E-06	3.17E-06	2.97E-06	3.30E-06	3.13E-06
BEST SIC	1.39E-06	1.39E-06	1.41E-06	1.52E-06	1.46E-06	2.15E-06	1.69E-06	1.62E-06	1.64E-06	1.59E-06
AVG AIC	1.81E-06	1.80E-06	1.76E-06	1.75E-06	1.71E-06	2.04E-06	1.85E-06	1.76E-06	1.82E-06	1.81E-06
AVG SIC	1.37E-06	1.36E-06	1.33E-06	1.34E-06	1.35E-06	1.82E-06	1.44E-06	1.40E-06	1.42E-06	1.43E-06
AVG-E AIC	2.64E-06	2.68E-06	2.67E-06	2.51E-06	2.45E-06	2.64E-06	2.62E-06	2.44E-06	2.73E-06	2.60E-06
AVG-E SIC	1.31E-06	1.31E-06	1.33E-06	1.43E-06	1.36E-06	2.02E-06	1.58E-06	1.50E-06	1.54E-06	1.49E-06
SIMPLE AVG	1.51E-06	1.49E-06	1.44E-06	1.42E-06	1.43E-06	1.96E-06	1.50E-06	1.45E-06	1.46E-06	1.52E-06
OPT AVG	1.11E-06	1.11E-06	1.12E-06	1.16E-06	1.13E-06	1.60E-06	1.28E-06	1.21E-06	1.25E-06	1.22E-06

Table 5.29 – forecast MSEs from model selection and averaging strategies. FTSE 100 index, Gaussian normalized innovations, N = 500.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.51%	0.54%	1.01%	9.40%	2.73%	19.24%	10.44%	8.57%	9.40%	6.87%
BEST AIC	183.71%	188.14%	189.18%	163.00%	161.38%	99.13%	146.63%	144.68%	164.15%	160.00%
BEST SIC	25.65%	25.27%	26.16%	30.81%	29.66%	34.37%	31.81%	33.38%	31.71%	29.87%
AVG AIC	63.13%	62.80%	57.26%	50.08%	51.45%	27.36%	43.97%	45.02%	45.65%	49.64%
AVG SIC	23.74%	22.82%	18.90%	15.51%	19.73%	13.89%	11.88%	15.21%	14.09%	17.31%
AVG-E AIC	137.99%	142.45%	139.40%	115.70%	117.42%	65.03%	104.05%	101.27%	118.45%	115.75%
AVG-E SIC	18.61%	18.53%	19.19%	23.28%	20.61%	26.43%	23.28%	23.57%	23.49%	21.89%
SIMPLE AVG	36.19%	34.99%	29.08%	22.23%	27.03%	22.37%	16.63%	19.73%	17.30%	25.06%

Table 5.30 – model selection and averaging strategies relative MSE losses (MSE_{RL}). FTSE 100 index, Gaussian normalized innovations, N = 500.

The British index situation for N = 500 displays almost ubiquitous dominance of EGARCH(1,1) over any other strategy, being the SIC linear averaging the clear second best option.

5.1.6 Consolidation of best performing strategies

In this subsection, we aim to summarize the best performing forecast strategies within each of the ten general scenarios (five indexes, two numbers of samples). This is displayed in Table 5.31.

Index	$N = 250$	$N = 500$
IBOV	Linear SIC and simple averaging	Linear SIC and simple averaging
S&P 500	Linear SIC averaging with slight advantage over simple averaging	Linear SIC averaging
N225	EGARCH(1,1)	Linear SIC averaging with slight advantage over simple averaging
DAX	Linear SIC averaging with slight advantage over simple averaging	Linear SIC averaging
FTSE 100	EGARCH(1,1) and linear SIC averaging	EGARCH(1,1)

Table 5.31 – Consolidation of best performing strategies across indexes and numbers of samples

The most relevant strategies are thus the fixed EGARCH(1,1) model, linear SIC and simple averaging. From the latter two, linear SIC averaging is clearly the better overall option, since either it clearly outperforms simple averaging or both display similar performances.

Comparison of EGARCH(1,1) and linear SIC weighting in extreme underfitting outperforming scenarios such as Japanese (N = 250) and British indexes (N = 250 and N = 500) favored EGARCH(1,1) once with minor advantages (British index for N = 250), and twice with significant advantage (Japanese index for N = 250 and British index for N = 500). We conclude that these occurrences are largely compensated by the linear SIC averaging benefits, higher in both magnitude and frequency, observed in the other seven scenarios analyzed.

Hence, our conclusion is that linear SIC averaging is the best option, as it yields higher and more likely MSE performance potential over the most often used minimum order model, at the cost of higher complexity, since it demands several models to be estimated. Under the scope in which this computational cost is acceptable (or even negligible)

and the MSE gains are significant, we proceed in the next section with linear SIC averaging as our reference strategy, aiming to generalize and outperform it.

5.2 GENERALIZED SIC AVERAGING STRATEGY

In this section, we introduce the performances of generalized SIC averaging, which corresponds to linear SIC averaging exploited in the previous section replacing regular SIC criterion by generalized SIC criterion, proposed in Chapter 4 and defined by (4.24), and restated below for convenience:

$$SIC_m(\lambda) \triangleq \hat{l}_m - (1+\lambda)\frac{1}{2} p_m \ln(N)$$

where $SIC_m(0)$ corresponds to regular SIC, and positive values of λ allow for higher complexity penalties and thus higher weights for simpler (possibly underfitting) models in the context of model averaging.

Since the extra penalty magnitude that brings improvement is expected to decrease with the number of samples, we present the results separately for each value of N investigated (250 and 500).

5.2.1 Results for N = 250

Firstly, we assign to each one of the DGPs, a sequential number according to the following table:

(P,Q) of EGARCH DGP to generate the data	Sequential number assigned to DGP, or DGP number
(1,1)	1
(2,1)	2
(3,1)	3
(1,2)	4
(2,2)	5
(3,2)	6
(1,3)	7
(2,3)	8
(3,3)	9

Table 5.32 – Sequential numbers assigned to each DGP

The ordering presented in Table 5.32 is the same used in the previous section to order the columns of the Tables therein. It features a non-descending number of parameters order and constant Q for each subsequence of three DGPs: Q = 1 for DGPs 1-3, Q = 2 for DGPs 4-6 and Q = 3 for DGPs 7-9.

Next, we present five sets of two graphs, each set referring to one of the five markets analyzed. The first graph of a given set (market) plots the relative MSE losses (MSE_{RL}) as defined by (5.4) in function of DGP number, each curve corresponding to generalized SIC averaging strategy for a particular λ. The second graph of each set plots the average relative MSE losses across all DGPs as a function of λ, so that the minimum value corresponds to the optimum value of λ in the sense of minimum MSE_{RL} under the (somewhat arbitrary) hypothesis of *a priori* equal probabilities of each DGP being the correct one.

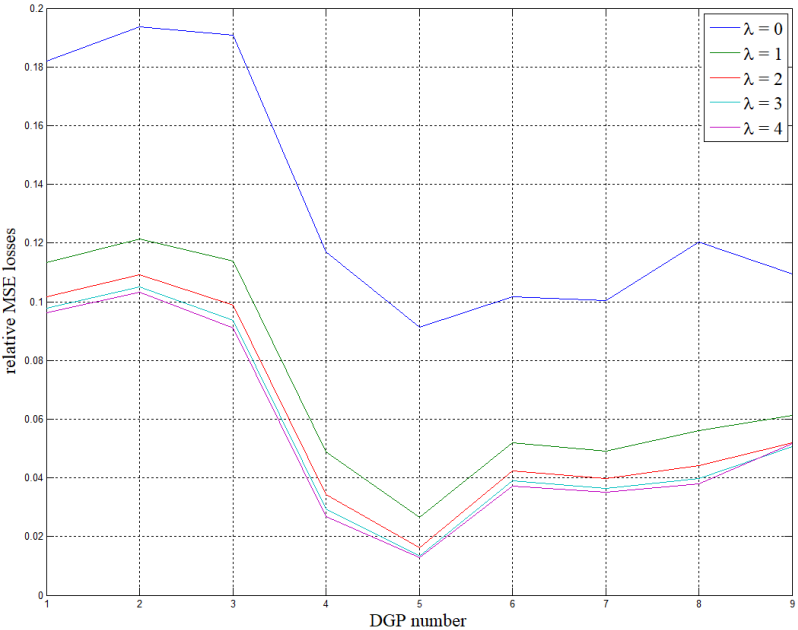

Figure 5.1 – relative MSE losses (y-axis) versus DGP number (x-axis), for IBOV index, Gaussian normalized innovations and $N = 250$.

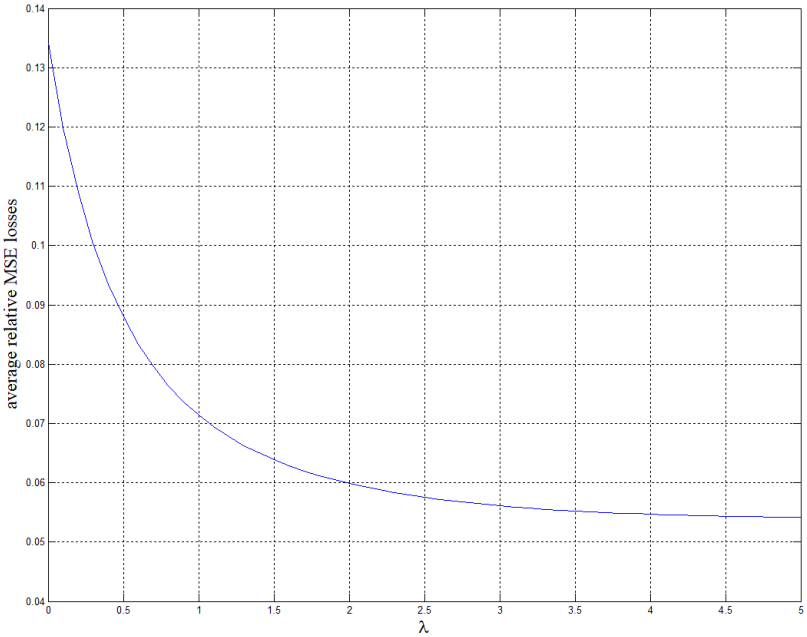

Figure 5.2 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for IBOV index, Gaussian normalized innovations and $N = 250$.

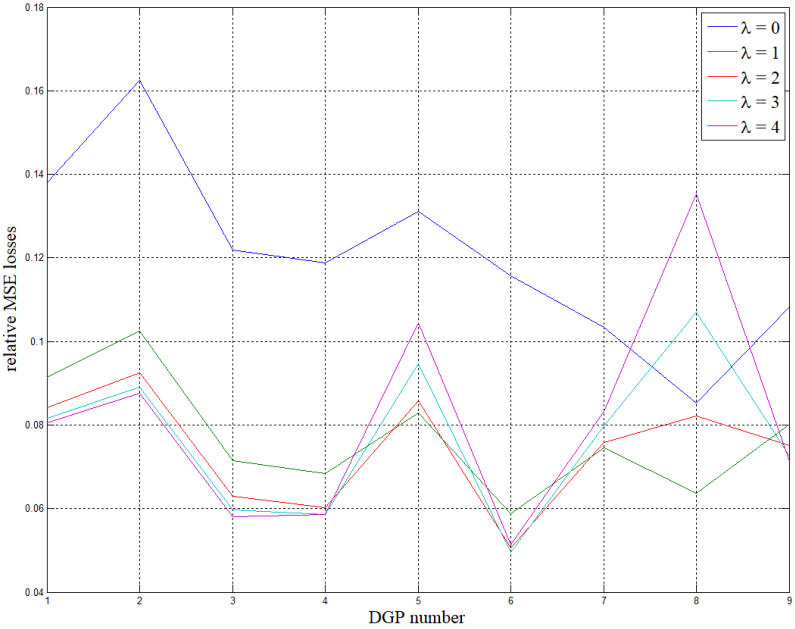

Figure 5.3 – relative MSE losses (y-axis) versus DGP number (x-axis), for S&P 500 index, Gaussian normalized innovations and $N = 250$.

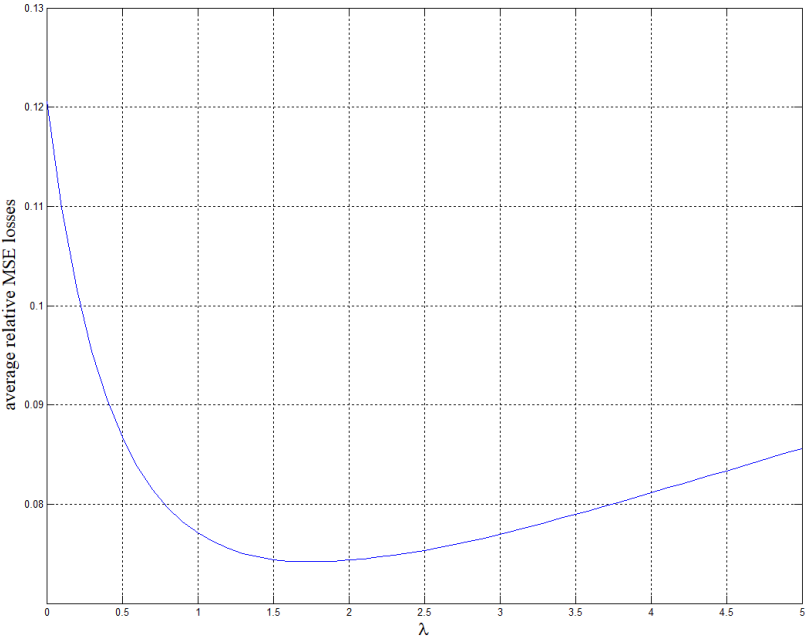

Figure 5.4 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for S&P 500 index, Gaussian normalized innovations and $N = 250$.

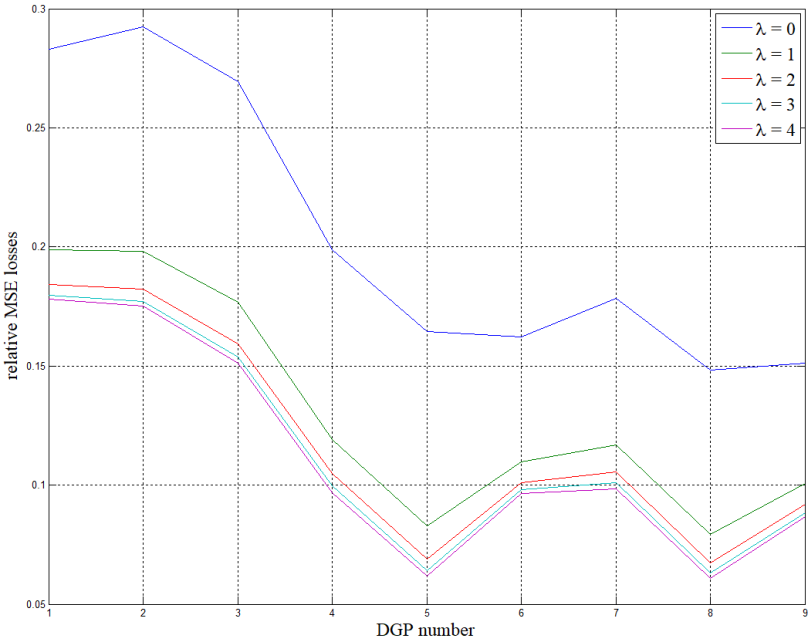

Figure 5.5 – relative MSE losses (y-axis) versus DGP number (x-axis), for N225 index, Gaussian normalized innovations and $N = 250$.

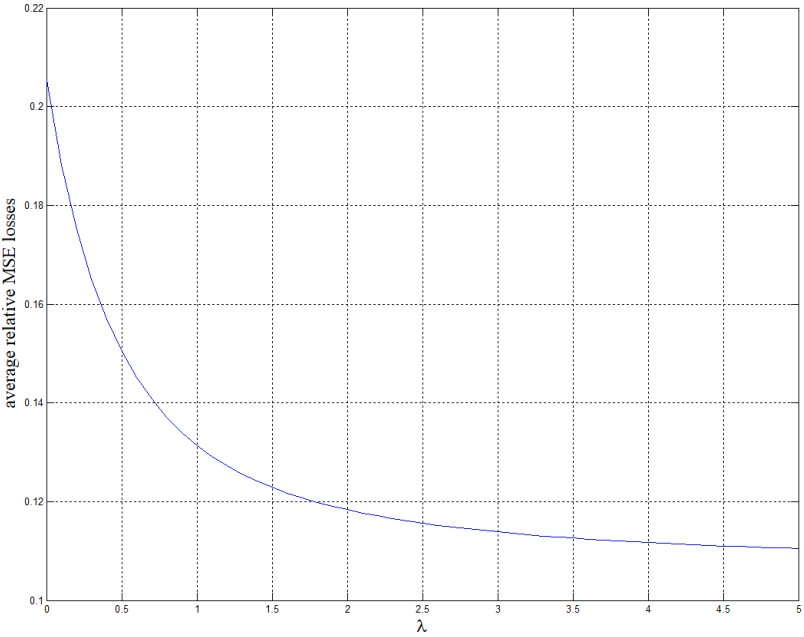

Figure 5.6 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for N225 index, Gaussian normalized innovations and $N = 250$.

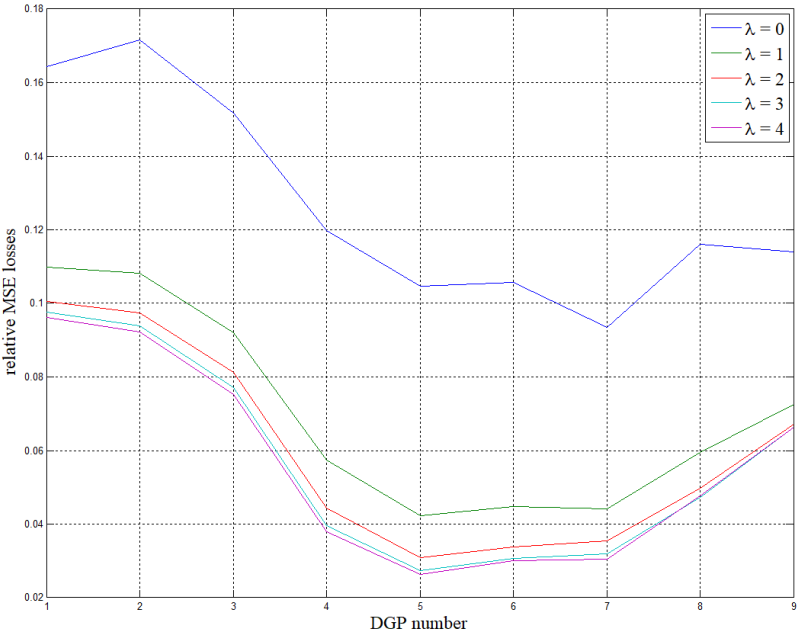

Figure 5.7 – relative MSE losses (y-axis) versus DGP number (x-axis), for DAX index, Gaussian normalized innovations and $N = 250$.

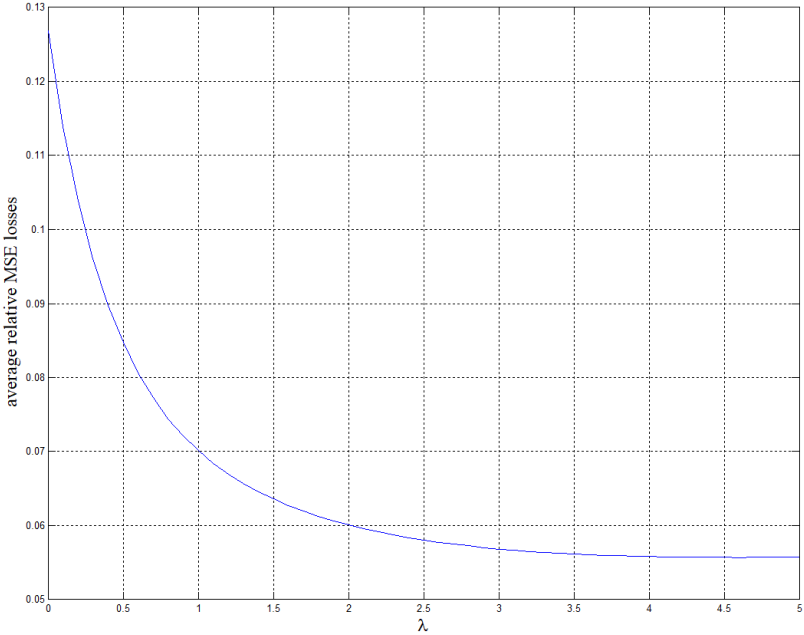

Figure 5.8 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for DAX index, Gaussian normalized innovations and $N = 250$.

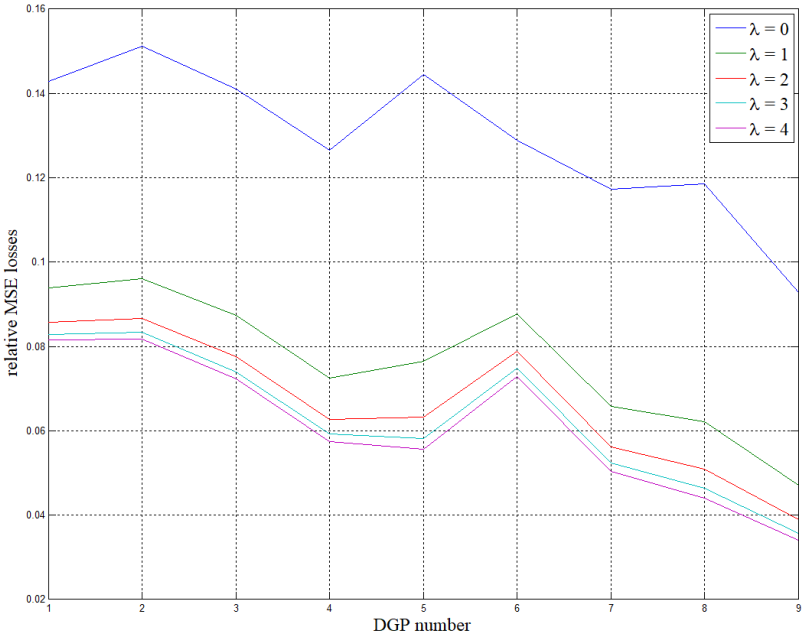

Figure 5.9 – relative MSE losses (y-axis) versus DGP number (x-axis), for FTSE 100 index, Gaussian normalized innovations and $N = 250$.

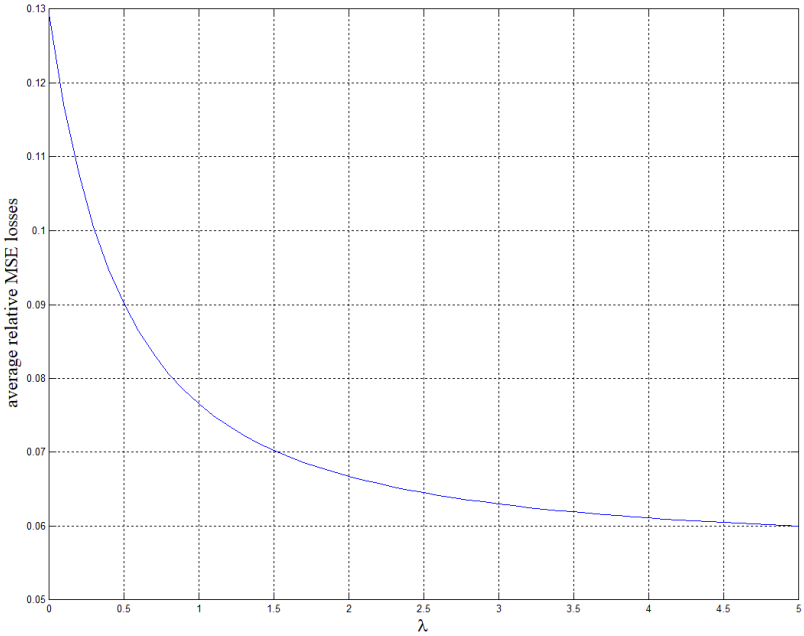

Figure 5.10 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for FTSE 100 index, Gaussian normalized innovations and $N = 250$.

In the above graphs, it is clear that increasing λ brings significant performance gains over regular SIC averaging ($\lambda = 0$). The relative MSE losses are generally decreasing with λ, although the outperformances almost flatten out for $\lambda = 4$. However, in two scenarios (S&P 500 index, DGPs 5 and 8) the results displayed opposite behavior.

In Figure 5.11 we plot the relative MSE losses as a function of λ, but averaged not only over the DGPs (for each given market, as done beforehand) but also over all the five markets, to provide an overall picture.

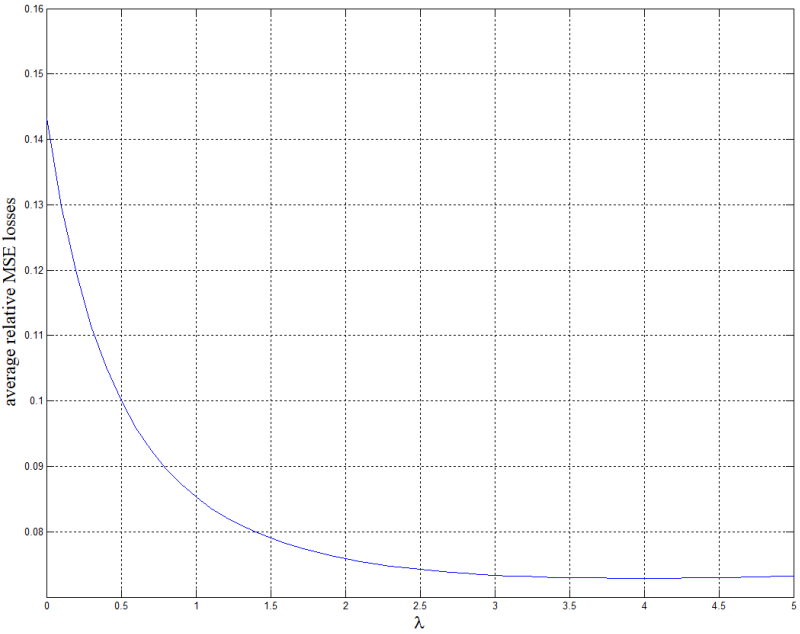

Figure 5.11 – average (across DGPs and markets) relative MSE losses (y-axis) versus λ (x-axis), for Gaussian normalized innovations and $N = 250$.

For all indexes but S&P 500, the average relative MSE losses were strictly decreasing with λ (until 5, the maximum value displayed, except DAX for which a bottom was reached at $\lambda = 4.5$). However, due to the exceptional scenarios for which increasing λ resulted in worse performances (S&P 500 index, DGPs 5 and 8), it is noticed that the optimum λ value (averaging over all DGPs) was 1.8 for the S&P 500 and 4 when all markets are averaged.

Although values of λ until 4 can be defended, especially under the overall average relative MSE loss figure, we would instead recommend a more conservative value of 2, since it is enough to capture most of the benefits provided by the generalized SIC strategy while limits the potential performance degradation in scenarios such as the American index (DGPs 5 or 8) where higher values are worse.

Undoubtedly, our main proposal is the application of the methodology here presented to available data to evaluate each scenario and then choose the most suitable value of λ. Nevertheless, we secondarily suggest the heuristic value of $\lambda = 2$ for $N = 250$.

5.2.2 Generalized SIC versus the minimum model

It is interesting to check if the generalized SIC strategy outperforms the minimal model - EGARCH(1,1) - on the average over all DGPs when regular SIC averaging did not. For $N = 250$, this was the case for N225 and FTSE 100 markets (see Table 5.31). From Table 5.15 (Japanese market), the average MSE loss over all DGPs is 12.65% for the minimal model, which is outperformed by generalized SIC averaging for λ larger than 1.25 (see Figure 5.6). Similarly, from Table 5.27 and Figure 5.10, we conclude that for the FTSE 100 index the minimal model average relative MSE loss of 10.93% is outperformed for λ larger than 0.2. Thus, differently from regular SIC averaging, the generalized SIC averaging can outperform the minimal model for every market analyzed, in the sense of average relative MSE performance over all DGPs.

5.2.3 Results for $N = 500$

In this subsection, we proceed exactly as done in the previous one, replacing the graphs with the ones corresponding to $N = 500$.

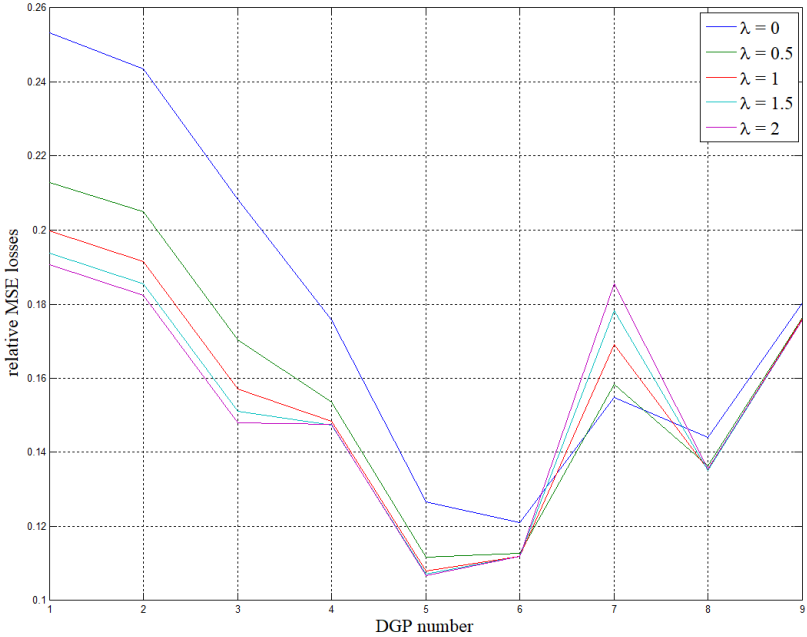

Figure 5.12 – relative MSE losses (y-axis) versus DGP number (x-axis), for IBOV index, Gaussian normalized innovations and $N = 500$.

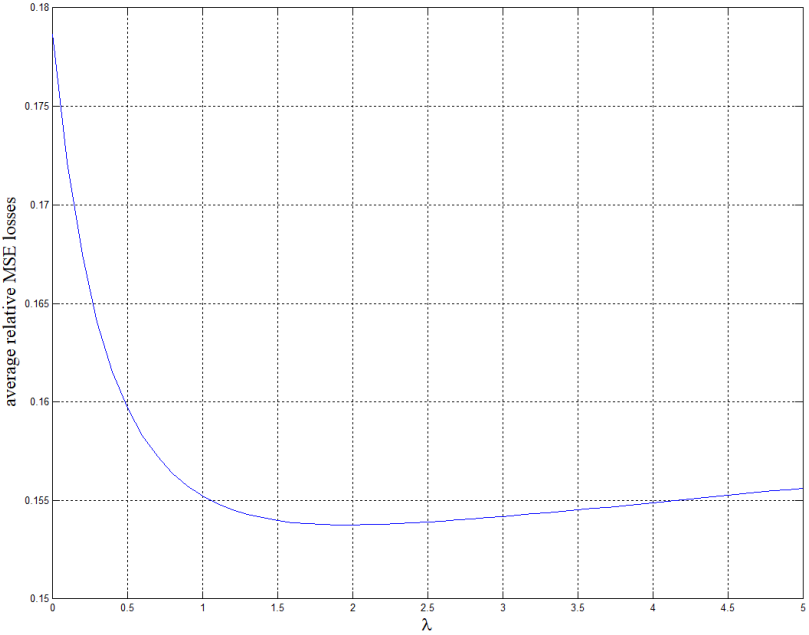

Figure 5.13 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for IBOV index, Gaussian normalized innovations and $N = 500$.

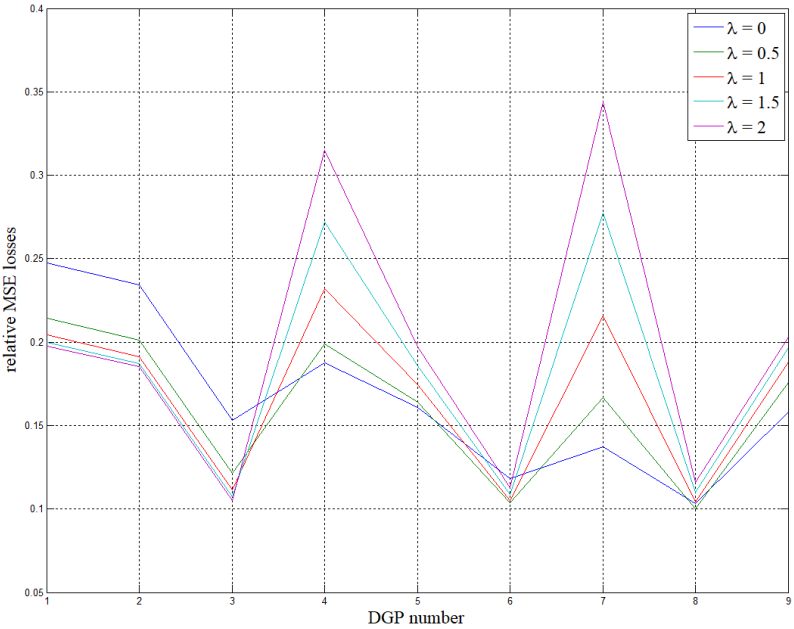

Figure 5.14 – relative MSE losses (y-axis) versus DGP number (x-axis), for S&P 500 index, Gaussian normalized innovations and $N = 500$.

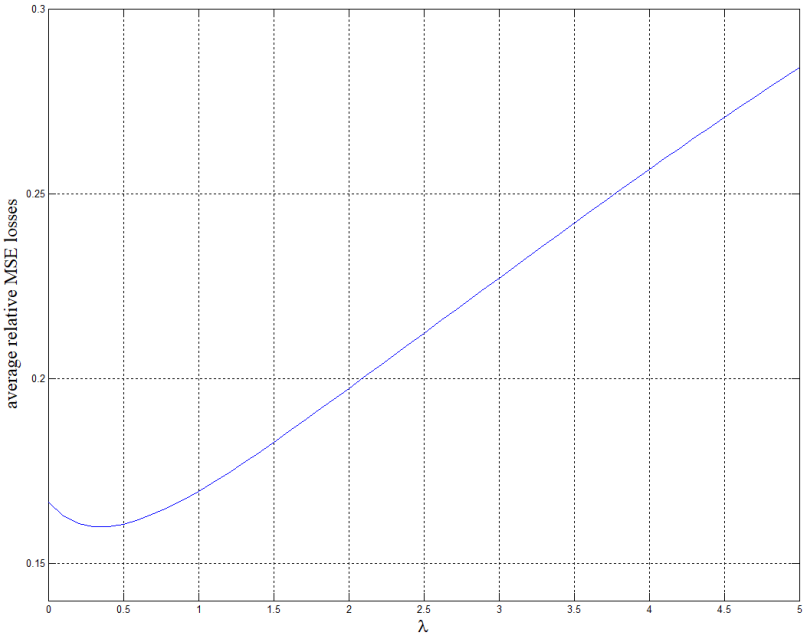

Figure 5.15 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for S&P 500 index, Gaussian normalized innovations and $N = 500$.

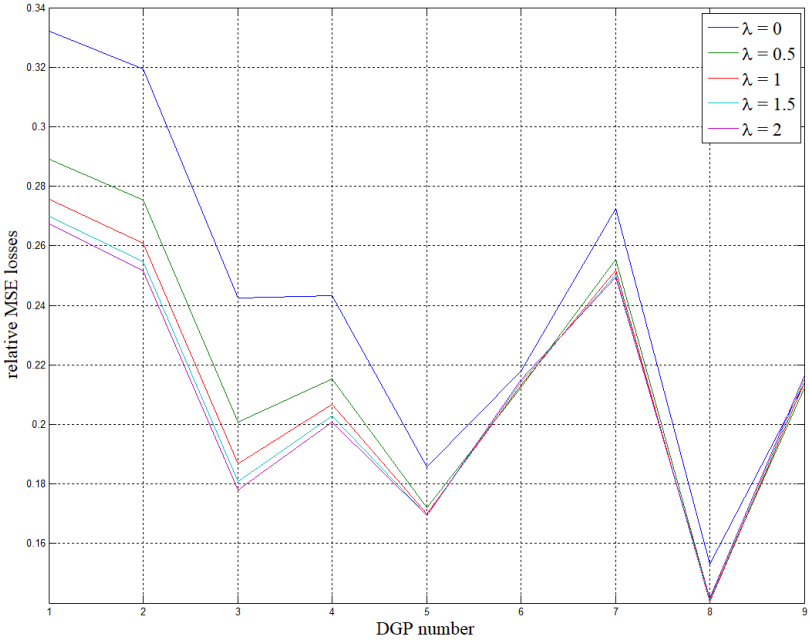

Figure 5.16 – relative MSE losses (y-axis) versus DGP number (x-axis), for N225 index, Gaussian normalized innovations and $N = 500$.

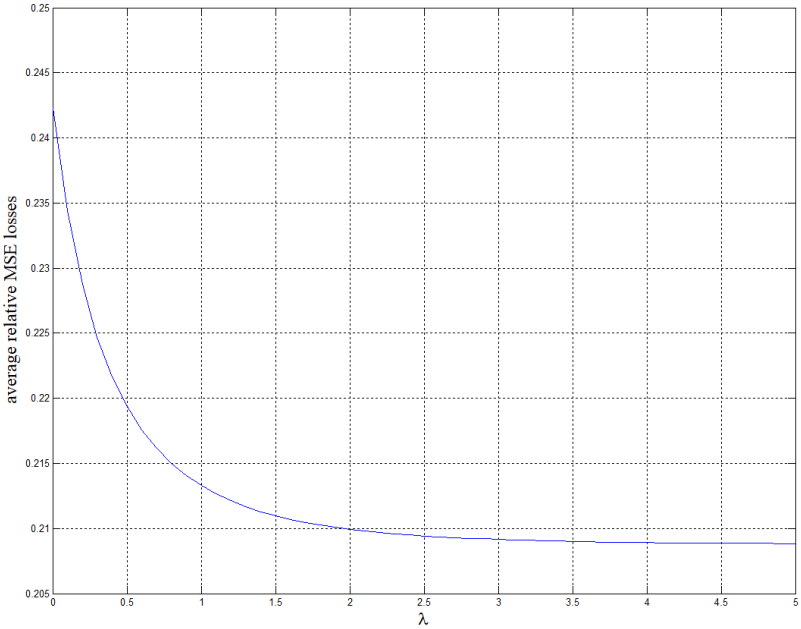

Figure 5.17 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for N225 index, Gaussian normalized innovations and $N = 500$.

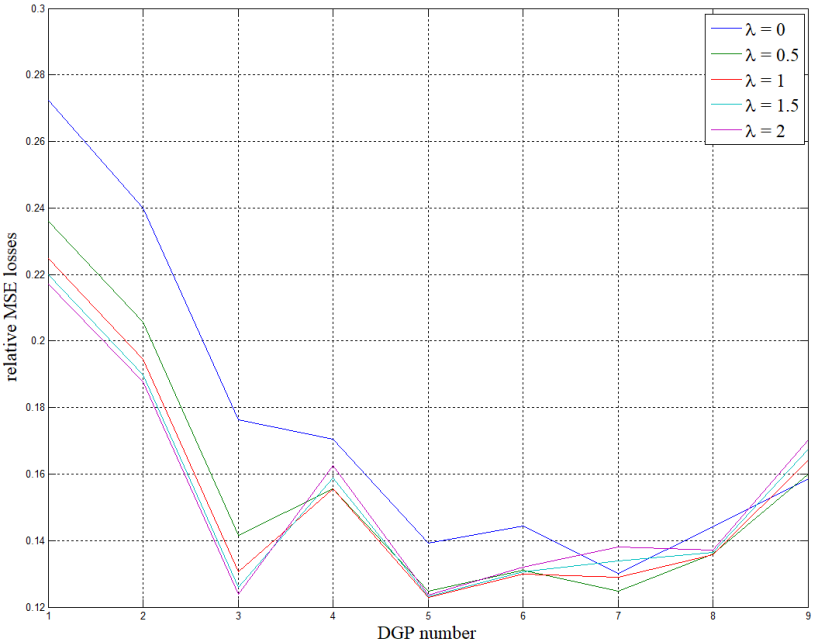

Figure 5.18 – relative MSE losses (y-axis) versus DGP number (x-axis), for DAX index, Gaussian normalized innovations and $N = 500$.

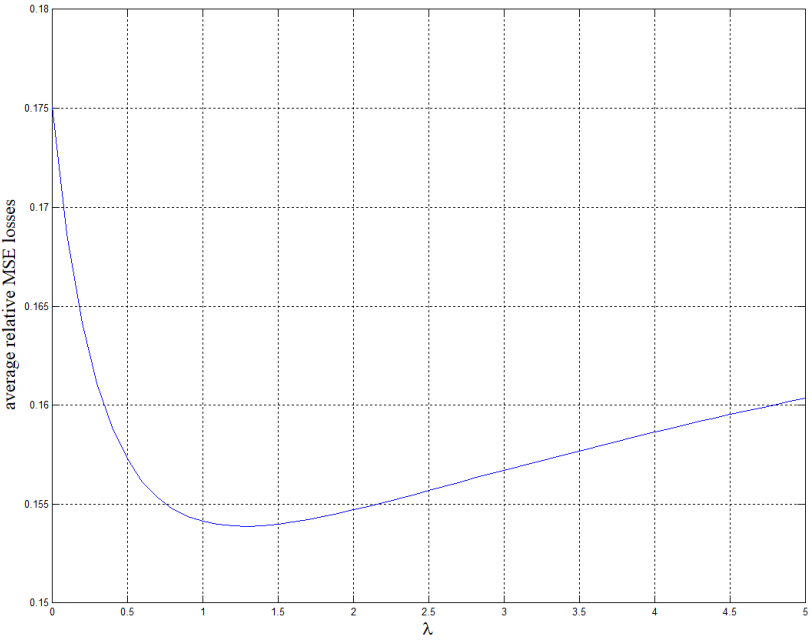

Figure 5.19 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for DAX index, Gaussian normalized innovations and $N = 500$.

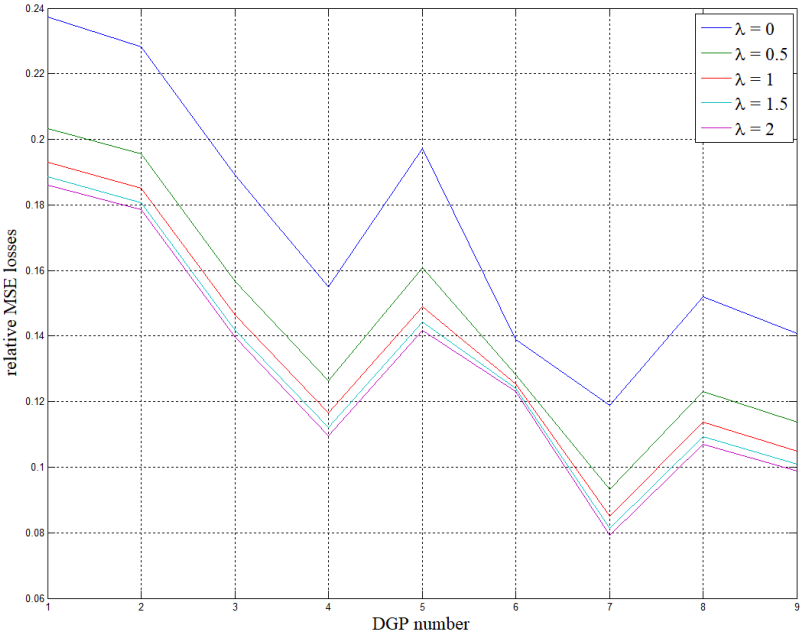

Figure 5.20 – relative MSE losses (y-axis) versus DGP number (x-axis), for FTSE 100 index, Gaussian normalized innovations and $N = 500$.

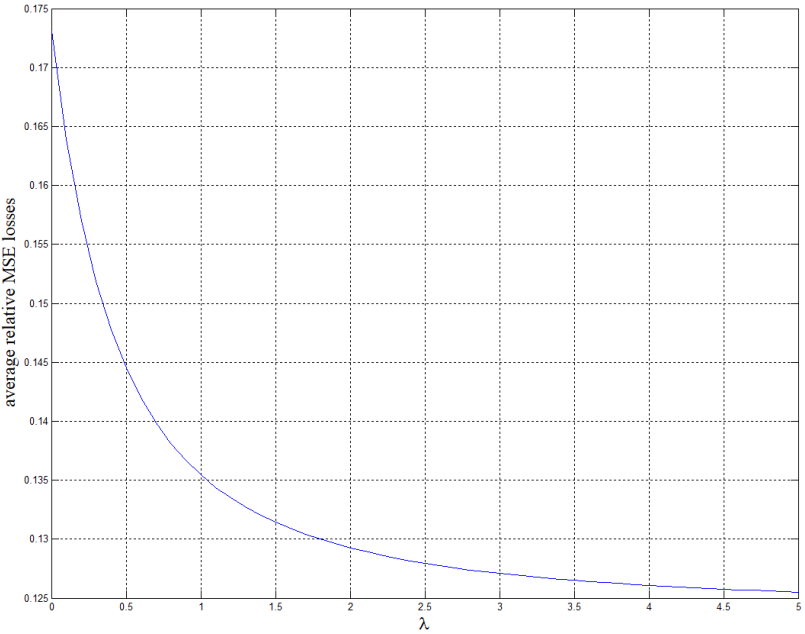

Figure 5.21 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for FTSE 100 index, Gaussian normalized innovations and $N = 500$.

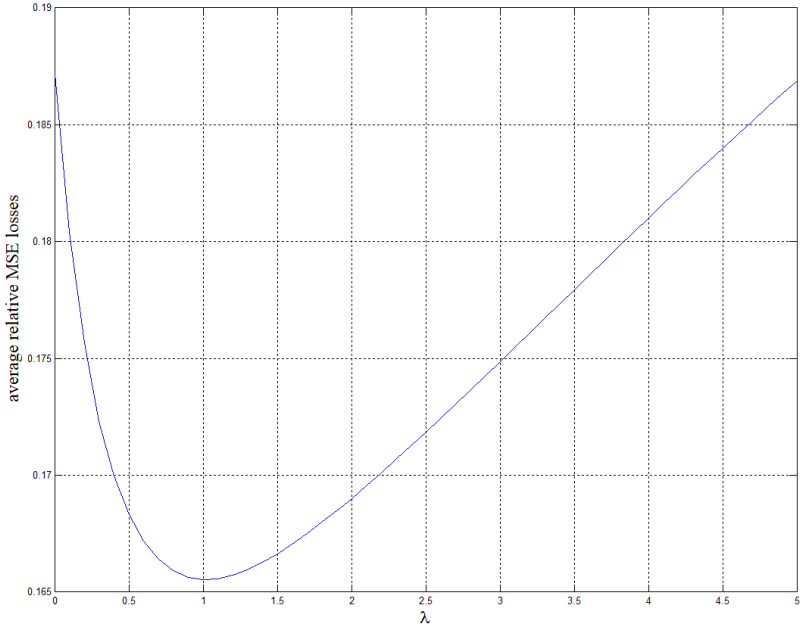

Figure 5.22 – average (across DGPs and markets) relative MSE losses (y-axis) versus λ (x-axis), for Gaussian normalized innovations and $N = 500$.

From Figure 5.22, we see that the optimum value of λ is 1, in overall (across all DGPs and markets) average relative MSE losses sense. However, it is worth noticing the significant variability of scenarios. For the Japanese and British markets (N225 and FTSE 100 indexes, respectively), the largest values of λ led to better performances, as happened for $N = 250$, while for the other indexes there were several scenarios for which the larger the λ, the worse the performance. On the average, these other indexes – IBOV, S&P 500 and DAX – had as optimum values $\lambda = 2$, $\lambda = 0.3$ and $\lambda = 1.3$, respectively. The different behavior of Japanese and British indexes stems from the fact that they are the ones for which the minimal model (EGARCH(1,1)) had the best performances. Since larger values of λ represent higher complexity penalties, increasing λ is expected to be better when lower complexity models are outperforming.

The variability of scenarios reinforces our suggestion of applying the methodology to choose a suitable value for λ depending on the data particularities. Nevertheless, similarly to the previous subsection, where we heuristically suggested the value of $\lambda = 2$ for $N = 250$, we suggest $\lambda =$

1 for N = 500, since it seems to provide the best balance between losses and gains across the scenarios analyzed, as also indicated by the attained minimum average relative MSE losses.

Lastly, we check if the generalized SIC strategy outperforms the minimal model (EGARCH(1,1)) on the average over all DGPs when regular SIC averaging did not. For N = 500, this was the case only for FTSE 100 market (see Table 5.31). From Table 5.30 and Figure 5.21, we conclude that for FTSE 100 market the minimal model average relative MSE loss of 6.87% outperforms generalized SIC averaging for all λ up to 5. Thus, this market is an exceptional scenario in which the minimal model is recommended over higher order models averaging.

5.3 STUDENT T NORMALIZED INNOVATIONS CASE

Although our focus has been on Gaussian normalized innovations, in this section we provide some results using Student t innovations, since it is known from the literature that there is a wide set of applications in which volatility data have higher kurtosis (fatter tails) than EGARCH models with Gaussian normalized innovations are able to provide.

Firstly, we establish that linear SIC averaging is the best strategy from the existing ones (and depicted in this work). To that end, the following tables display the relative MSE losses obtained for each strategy as done in Section 5.1, for the same five indexes (IBOV, S&P 500, N225, DAX and FTSE 100), two number of samples (N = 250 and N = 500), and nine DGPs (all combinations of order parameters P and Q ranging from 1 to 3).

The DGPs with t Student normalized innovations are given in Appendix B and were fitted using the same data used in Appendix A to fit the models for Gaussian innovations.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	8.29%	8.24%	11.73%	723.08%	21.03%	19.79%	667.90%	25.99%	22.63%	167.63%
BEST AIC	191.36%	200.39%	193.01%	152.30%	157.00%	151.21%	138.67%	166.17%	138.22%	165.37%
BEST SIC	78.12%	80.75%	83.71%	84.41%	77.37%	70.06%	85.15%	88.39%	66.81%	79.42%
AVG AIC	48.39%	49.92%	46.91%	35.02%	36.12%	33.59%	28.72%	37.19%	31.62%	38.61%
AVG SIC	15.36%	16.14%	14.03%	8.79%	10.38%	8.88%	8.21%	12.80%	10.91%	11.72%
AVG-E AIC	144.56%	149.95%	145.89%	110.33%	114.36%	111.00%	96.37%	120.99%	97.41%	121.21%
AVG-E SIC	52.47%	54.81%	57.55%	56.42%	51.81%	46.06%	57.13%	60.25%	45.75%	53.58%
SIMPLE AVG	19.77%	19.94%	18.13%	14.04%	19.06%	59.21%	10.05%	31.64%	43.36%	26.13%

Table 5.33 – model selection and averaging strategies relative MSE losses (MSE_{RL}). IBOV index, Student t normalized innovations, N = 250.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.38%	1.34%	5.08%	1768.38%	45.03%	47.42%	1874.79%	52.69%	45.93%	426.89%
BEST AIC	230.15%	238.90%	195.32%	150.82%	134.85%	124.97%	129.93%	122.67%	92.68%	157.81%
BEST SIC	53.55%	55.63%	55.36%	88.56%	72.89%	67.83%	102.70%	78.05%	64.50%	71.01%
AVG AIC	65.47%	65.99%	51.25%	44.30%	34.93%	32.39%	31.36%	29.66%	23.69%	42.12%
AVG SIC	21.05%	21.19%	14.03%	18.87%	15.16%	13.94%	19.03%	17.86%	17.06%	17.58%
AVG-E AIC	180.71%	187.28%	150.75%	109.46%	99.72%	90.08%	92.35%	85.99%	61.82%	117.57%
AVG-E SIC	38.71%	40.10%	39.92%	64.19%	53.59%	50.22%	73.86%	58.05%	48.58%	51.91%
SIMPLE AVG	28.86%	28.98%	20.71%	31.34%	36.65%	132.01%	25.98%	56.99%	54.70%	46.25%

Table 5.34 – model selection and averaging strategies relative MSE losses (MSE_{RL}). IBOV index, Student t normalized innovations, $N = 500$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	9.81%	9.70%	11.98%	704.93%	28.85%	32.82%	817.89%	38.14%	33.52%	187.52%
BEST AIC	155.55%	149.94%	141.44%	123.94%	127.44%	119.35%	107.76%	118.44%	89.14%	125.89%
BEST SIC	58.00%	54.65%	50.88%	93.50%	73.64%	63.81%	87.12%	73.82%	55.76%	67.91%
AVG AIC	38.40%	36.98%	33.31%	28.70%	27.25%	26.32%	20.67%	25.57%	19.23%	28.49%
AVG SIC	11.22%	10.82%	9.77%	11.12%	10.97%	10.16%	8.96%	11.82%	8.83%	10.41%
AVG-E AIC	114.41%	109.09%	103.72%	88.56%	90.24%	85.81%	72.55%	82.78%	58.35%	89.50%
AVG-E SIC	38.58%	37.52%	34.81%	61.94%	49.41%	45.69%	57.40%	51.04%	39.13%	46.17%
SIMPLE AVG	14.25%	13.79%	11.99%	10.19%	22.42%	24.32%	8.01%	17.73%	14.79%	15.28%

Table 5.35 – model selection and averaging strategies relative MSE losses (MSE_{RL}). S&P 500 index, Student t normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.10%	0.61%	6.18%	4751.36%	63.67%	65.34%	4124.94%	69.25%	62.62%	1016.12%
BEST AIC	181.41%	183.57%	168.95%	121.50%	123.83%	98.06%	95.20%	94.26%	74.93%	126.86%
BEST SIC	32.31%	29.54%	34.43%	81.67%	83.36%	76.22%	87.28%	85.72%	78.17%	65.41%
AVG AIC	54.53%	55.66%	48.24%	31.81%	33.37%	21.08%	17.74%	25.43%	17.83%	33.97%
AVG SIC	17.37%	16.58%	13.65%	21.95%	20.27%	11.46%	15.12%	22.67%	18.61%	17.52%
AVG-E AIC	139.34%	143.67%	129.24%	86.58%	89.64%	68.85%	64.39%	67.57%	49.88%	93.24%
AVG-E SIC	22.41%	20.88%	25.01%	56.69%	60.64%	55.37%	59.61%	61.46%	56.00%	46.45%
SIMPLE AVG	24.03%	24.59%	17.87%	58.76%	77.79%	88.24%	46.65%	45.74%	25.69%	45.48%

Table 5.36 – model selection and averaging strategies relative MSE losses (MSE_{RL}). S&P 500 index, Student t normalized innovations, $N = 500$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	5.37%	5.02%	7.70%	335.55%	17.89%	16.53%	309.98%	25.13%	24.36%	83.06%
BEST AIC	223.17%	219.03%	210.43%	170.74%	179.20%	167.95%	154.44%	166.82%	145.05%	181.87%
BEST SIC	88.66%	90.67%	92.35%	89.43%	88.15%	75.91%	82.99%	95.36%	74.25%	86.42%
AVG AIC	59.13%	58.56%	55.48%	43.33%	44.33%	41.86%	35.59%	35.10%	31.27%	44.96%
AVG SIC	20.07%	20.19%	17.45%	10.90%	13.49%	13.92%	9.81%	15.65%	13.59%	15.01%
AVG-E AIC	174.11%	169.53%	164.10%	128.04%	133.60%	128.13%	111.41%	124.10%	106.27%	137.70%
AVG-E SIC	61.86%	63.18%	64.51%	60.20%	61.00%	53.82%	57.32%	67.19%	52.58%	60.18%
SIMPLE AVG	24.08%	23.36%	21.01%	11.39%	15.64%	18.96%	7.65%	39.96%	50.82%	23.65%

Table 5.37 – model selection and averaging strategies relative MSE losses (MSE_{RL}). N225 index, Student t normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.36%	1.07%	4.54%	898.74%	44.59%	40.62%	905.83%	54.48%	57.11%	223.04%
BEST AIC	263.22%	263.38%	228.93%	157.58%	152.84%	148.79%	123.56%	117.68%	97.22%	172.58%
BEST SIC	56.56%	58.76%	58.37%	82.03%	79.66%	68.53%	79.91%	78.69%	74.69%	70.80%
AVG AIC	79.02%	78.31%	64.72%	47.57%	44.78%	40.05%	31.96%	22.02%	22.29%	47.86%
AVG SIC	27.90%	25.70%	19.26%	19.43%	19.86%	20.04%	16.49%	16.12%	17.22%	20.22%
AVG-E AIC	211.15%	211.30%	178.76%	114.05%	115.21%	108.04%	87.28%	82.97%	66.42%	130.58%
AVG-E SIC	42.48%	43.05%	41.51%	58.51%	58.99%	52.12%	56.68%	57.61%	56.13%	51.90%
SIMPLE AVG	37.59%	34.66%	26.93%	24.28%	30.06%	39.99%	17.93%	124.54%	135.57%	52.39%

Table 5.38 – model selection and averaging strategies relative MSE losses (MSE_{RL}). N225 index, Student t normalized innovations, $N = 500$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	7.94%	7.69%	8.69%	791.51%	25.42%	24.21%	671.42%	30.19%	21.91%	176.55%
BEST AIC	179.65%	168.86%	164.57%	141.48%	143.14%	144.77%	131.59%	147.88%	128.35%	150.03%
BEST SIC	63.55%	60.33%	63.49%	90.91%	79.22%	67.42%	86.50%	77.80%	60.26%	72.16%
AVG AIC	44.28%	40.97%	40.00%	34.64%	32.49%	34.78%	29.29%	32.73%	31.72%	35.66%
AVG SIC	13.76%	12.29%	12.45%	11.12%	10.87%	11.52%	10.75%	12.06%	12.28%	11.90%
AVG-E AIC	135.08%	124.00%	120.07%	102.77%	104.62%	104.48%	93.24%	106.16%	89.77%	108.91%
AVG-E SIC	43.31%	40.53%	42.76%	61.12%	52.04%	47.22%	59.10%	53.87%	42.32%	49.14%
SIMPLE AVG	16.85%	15.05%	15.90%	13.10%	18.76%	27.57%	9.42%	18.73%	30.42%	18.42%

Table 5.39 – model selection and averaging strategies relative MSE losses (MSE_{RL}). DAX index, Student t normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	0.74%	0.49%	3.03%	2743.69%	58.50%	55.96%	2173.93%	57.05%	46.84%	571.14%
BEST AIC	216.53%	215.47%	197.44%	124.85%	130.60%	117.64%	110.99%	110.96%	87.66%	145.79%
BEST SIC	37.01%	37.60%	37.46%	83.46%	82.75%	73.22%	87.19%	74.70%	60.74%	63.79%
AVG AIC	63.80%	65.14%	52.77%	34.73%	34.14%	29.43%	23.42%	28.51%	23.86%	39.53%
AVG SIC	21.10%	21.56%	14.29%	21.29%	20.49%	13.63%	17.20%	18.70%	17.23%	18.39%
AVG-E AIC	171.01%	172.21%	152.50%	89.07%	94.98%	85.37%	74.25%	78.02%	58.71%	108.46%
AVG-E SIC	27.64%	28.21%	26.92%	59.54%	60.14%	53.09%	61.15%	53.16%	45.23%	46.12%
SIMPLE AVG	29.29%	29.53%	20.30%	40.37%	63.56%	66.48%	29.82%	27.72%	44.40%	39.05%

Table 5.40 – model selection and averaging strategies relative MSE losses (MSE_{RL}). DAX index, Student t normalized innovations, $N = 500$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	11.31%	7.43%	8.53%	50.37%	16.69%	14.80%	62.35%	17.94%	12.27%	22.41%
BEST AIC	159.49%	169.14%	171.68%	142.22%	158.48%	151.92%	138.46%	152.36%	121.84%	151.73%
BEST SIC	56.13%	61.19%	59.89%	56.37%	68.13%	53.51%	57.42%	65.79%	45.65%	58.23%
AVG AIC	37.28%	44.23%	42.20%	33.53%	36.25%	34.08%	29.55%	29.76%	29.12%	35.11%
AVG SIC	10.02%	13.89%	12.82%	6.76%	11.24%	10.31%	5.16%	11.19%	10.73%	10.24%
AVG-E AIC	117.50%	126.56%	126.46%	102.87%	117.14%	111.97%	98.73%	111.91%	83.10%	110.69%
AVG-E SIC	37.27%	41.17%	41.37%	37.03%	44.95%	35.77%	37.39%	46.32%	31.85%	39.24%
SIMPLE AVG	12.33%	17.28%	16.13%	8.28%	27.27%	42.09%	5.62%	41.91%	29.27%	22.24%

Table 5.41 – model selection and averaging strategies relative MSE losses (MSE_{RL}). FTSE 100 index, Student t normalized innovations, $N = 250$.

	(P,Q) of EGARCH DGP used to generate the data									
	(1,1)	(2,1)	(3,1)	(1,2)	(2,2)	(3,2)	(1,3)	(2,3)	(3,3)	Mean
EGARCH(1,1)	1.14%	1.08%	1.62%	198.46%	21.27%	22.15%	158.27%	27.51%	16.52%	49.78%
BEST AIC	191.28%	206.55%	199.56%	152.95%	146.37%	137.92%	140.11%	138.46%	93.90%	156.34%
BEST SIC	31.21%	34.25%	34.74%	37.79%	47.81%	44.65%	38.36%	53.70%	32.27%	39.42%
AVG AIC	55.37%	61.36%	55.75%	39.61%	41.13%	34.31%	35.62%	26.98%	24.63%	41.64%
AVG SIC	17.48%	19.25%	16.01%	7.04%	14.43%	11.13%	6.56%	13.70%	11.40%	13.00%
AVG-E AIC	150.20%	163.85%	155.25%	112.04%	112.50%	105.44%	103.62%	101.59%	63.03%	118.61%
AVG-E SIC	22.71%	25.09%	24.40%	24.90%	35.16%	32.75%	26.57%	40.78%	24.56%	28.55%
SIMPLE AVG	24.45%	27.02%	23.06%	11.72%	59.63%	94.61%	8.77%	111.29%	57.11%	46.41%

Table 5.42 – model selection and averaging strategies relative MSE losses (MSE_{RL}). FTSE 100 index, Student t normalized innovations, $N = 500$.

As in the Gaussian case, linear SIC averaging was the overall best strategy in all ten situations (five indexes, two numbers of samples), always displaying the lowest relative average MSE loss across DGPs. For Student t distributions, it is possible to conclude that this advantageous nature of linear SIC averaging is much more evident.

Next, we proceed as in Section 5.2 to examine if the Generalized SIC averaging can provide even better results than regular SIC averaging. We repeat the procedure of devoting one subsection to each

number of samples, each beginning with five sets of two graphs (10 graphs), each set referring to one of the five markets analyzed. The first graph of a given set (market) plots the relative MSE losses as a function of DGP number, each curve corresponding to generalized SIC averaging strategy for a particular λ. The second graph within each set plots the average relative MSE losses across all DGPs as a function of λ, for that particular market.

5.3.1 Results for N = 250

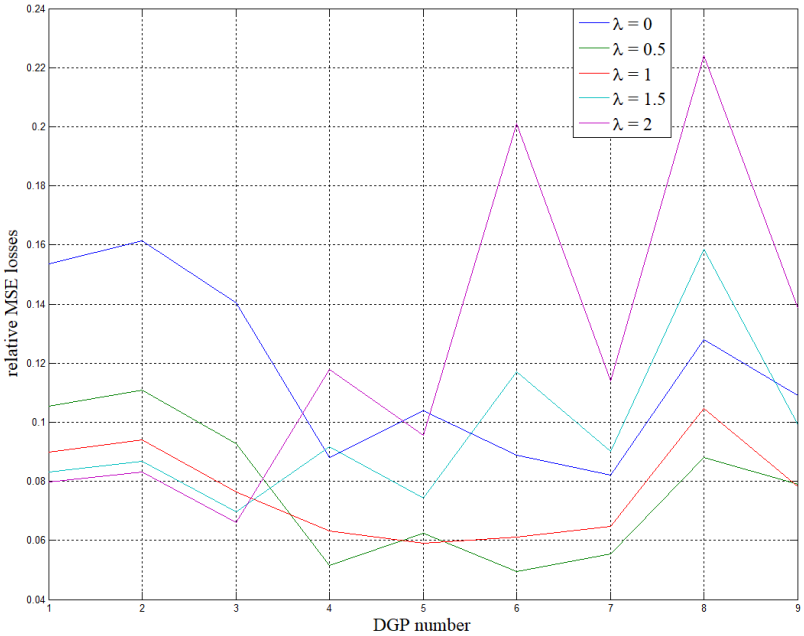

Figure 5.23 – relative MSE losses (y-axis) versus DGP number (x-axis), for IBOV index, N = 250 and Student t normalized innovations.

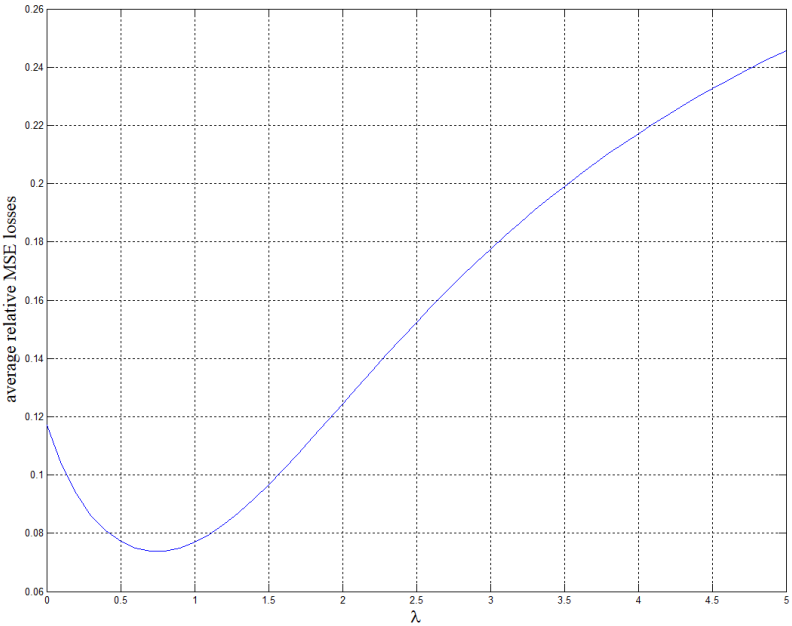

Figure 5.24 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for IBOV index, N = 250 and Student t normalized innovations.

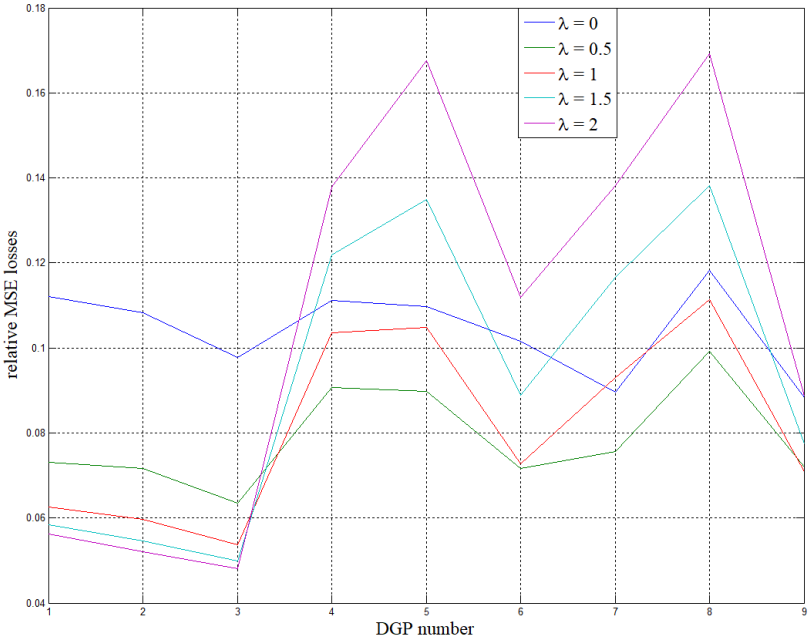

Figure 5.25 – relative MSE losses (y-axis) versus DGP number (x-axis), for S&P 500 index, $N = 250$ and Student t normalized innovations.

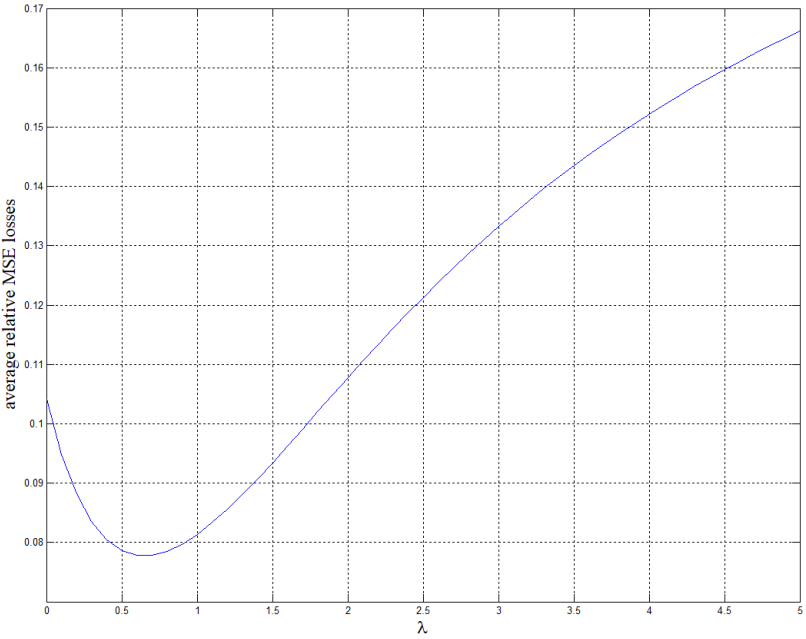

Figure 5.26 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for S&P 500 index, $N = 250$ and Student t normalized innovations.

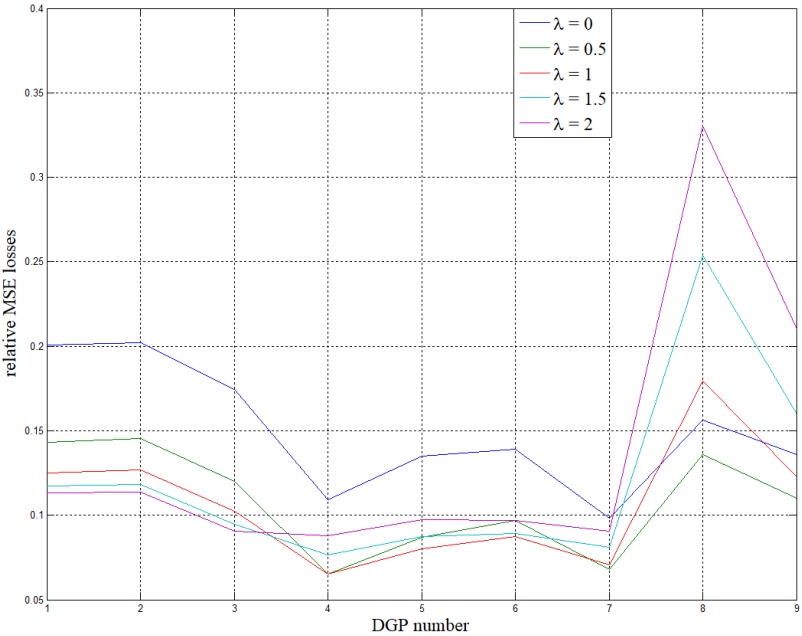

Figure 5.27 – relative MSE losses (y-axis) versus DGP number (x-axis), for N225 index, $N = 250$ and Student t normalized innovations.

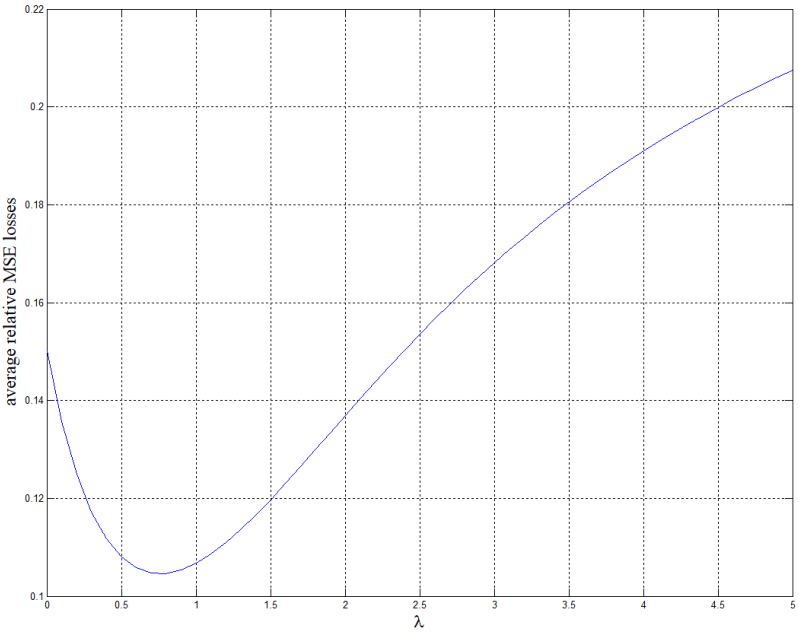

Figure 5.28 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for N225 index, N = 250 and Student t normalized innovations.

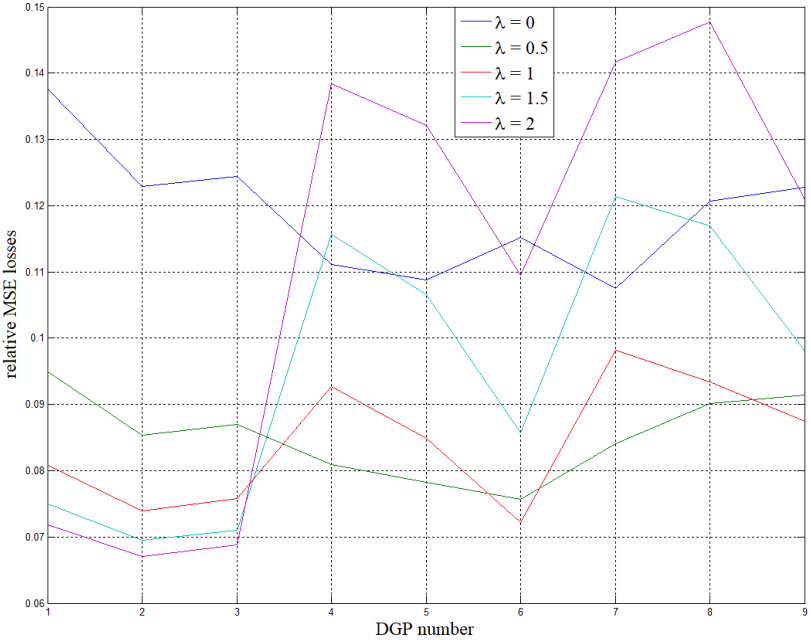

Figure 5.29 – relative MSE losses (y-axis) versus DGP number (x-axis), for DAX index, $N = 250$ and Student t normalized innovations.

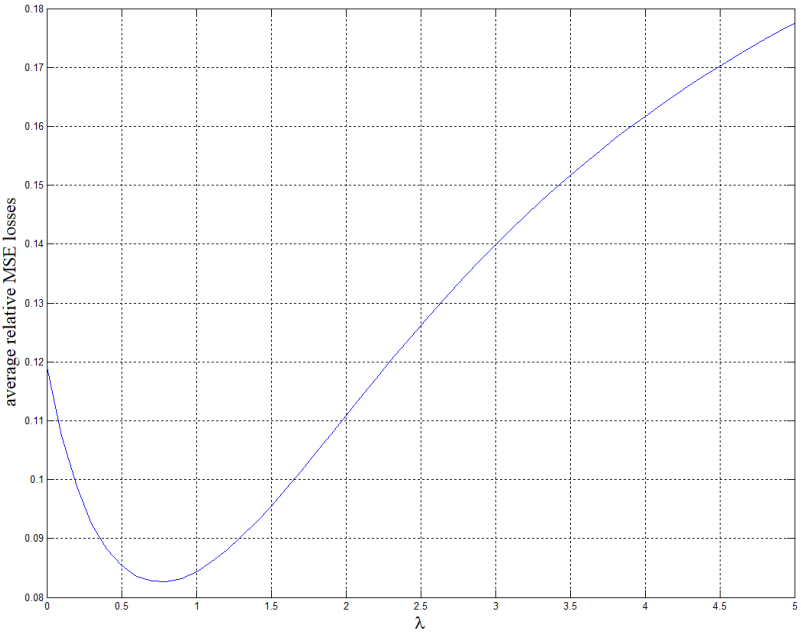

Figure 5.30 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for DAX index, N = 250 and Student t normalized innovations.

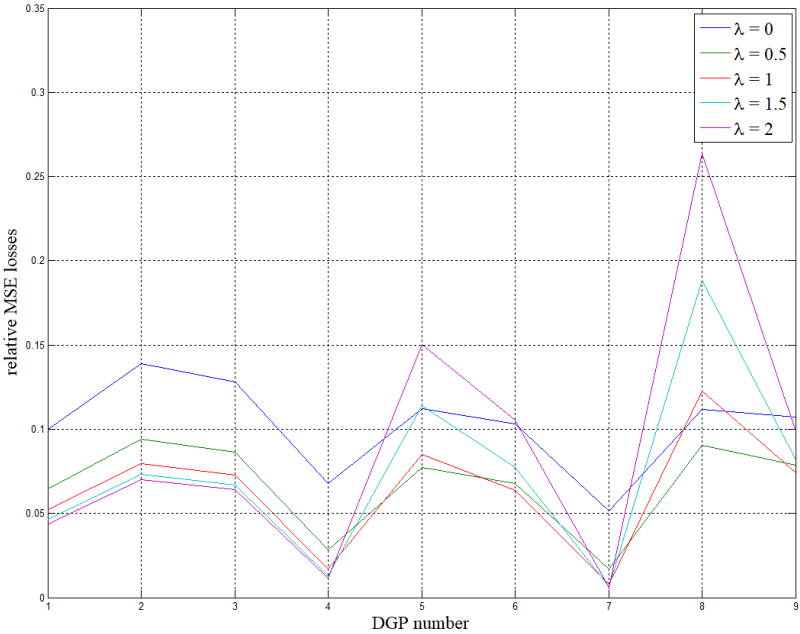

Figure 5.31 – relative MSE losses (y-axis) versus DGP number (x-axis), for FTSE 100 index, N = 250 and Student t normalized innovations.

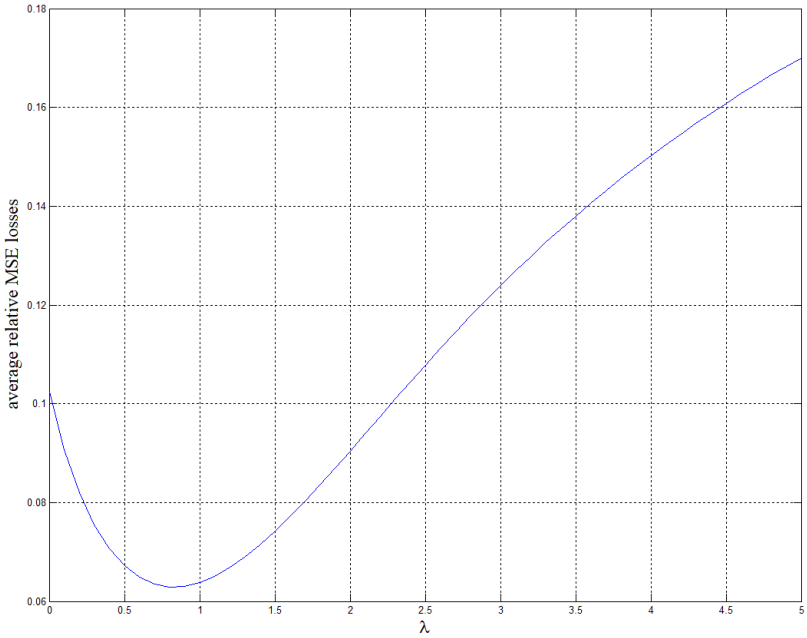

Figure 5.32 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for FTSE 100 index, N = 250 and Student t normalized innovations.

Lastly, it follows the relative MSE losses as functions of λ when averaged over all DGPs and all markets considered.

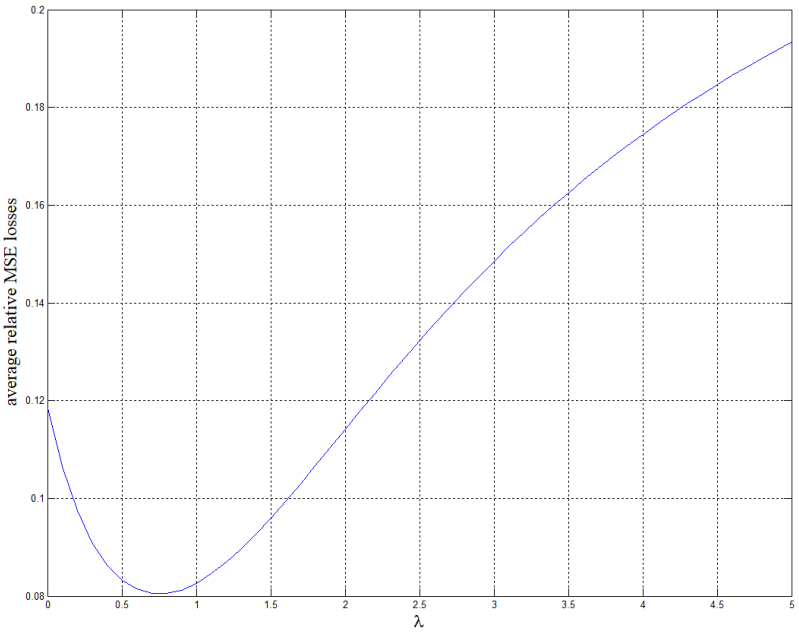

Figure 5.33 – average (across DGPs and markets) relative MSE losses (y-axis) versus λ (x-axis), for N = 250 and Student t normalized innovations.

Differently from the Gaussian case, these results display very similar behavior for all markets. The outperformances potentials of generalized SIC (compared do regular SIC) are less significant than the ones corresponding to the Gaussian scenarios, although still significant.

Small values of λ provide benefits, and performance degradation begins to occur for values of λ larger than approximately 2 (the exact value depending on the specific market). The optimal values of λ for the five indexes – IBOV, S&P 500, N225, DAX and FTSE 100 – were very similar one to another, being equal to 0.7, 0.6, 0.8, 0.8 and 0.8, respectively. Only multiples of 0.1 have been evaluated, what yields a precision error strictly lower than 0.1 in the optimal λ values determined.

When not only the DGP's but also all market's MSE relative losses are averaged, the mean sense optimal λ is 0.8.

5.3.2 Results for $N = 500$

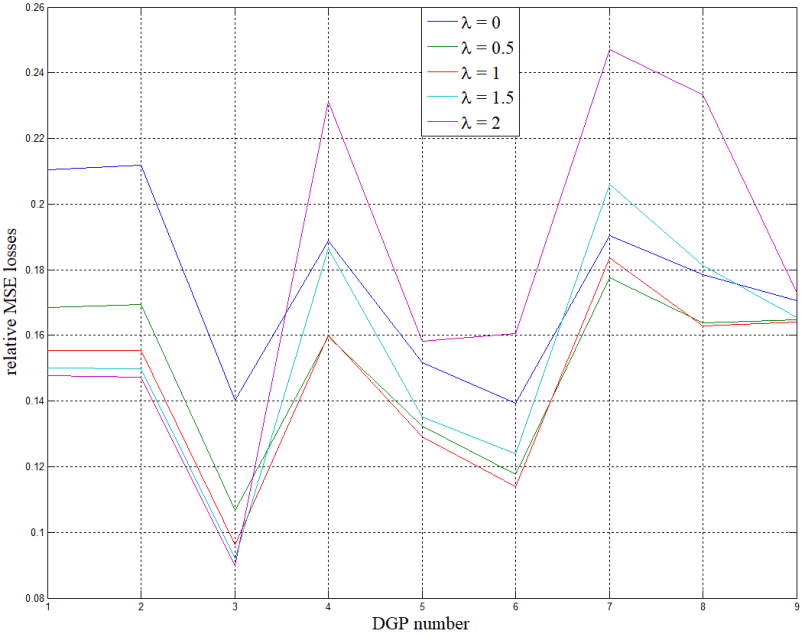

Figure 5.34 – relative MSE losses (y-axis) versus DGP number (x-axis), for IBOV index, $N = 500$ and Student t normalized innovations.

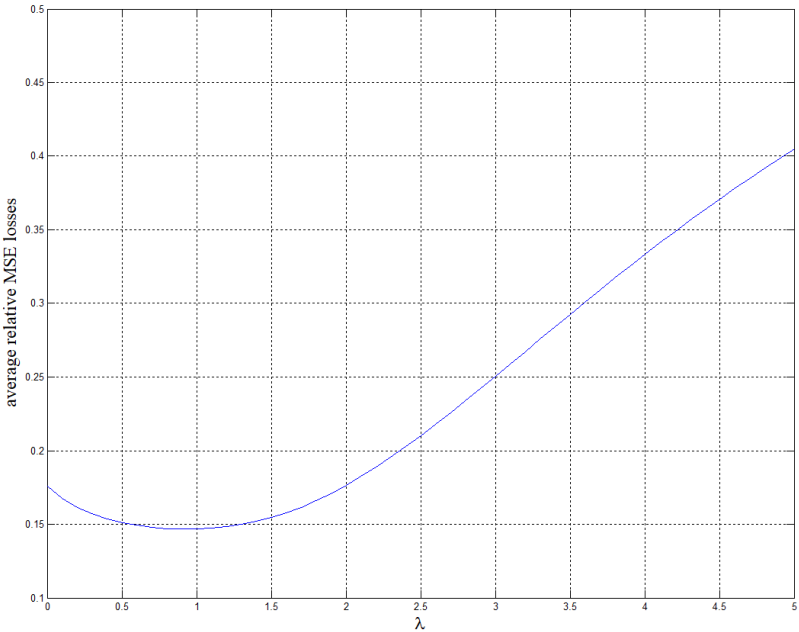

Figure 5.35 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for IBOV index, N = 500 and Student t normalized innovations.

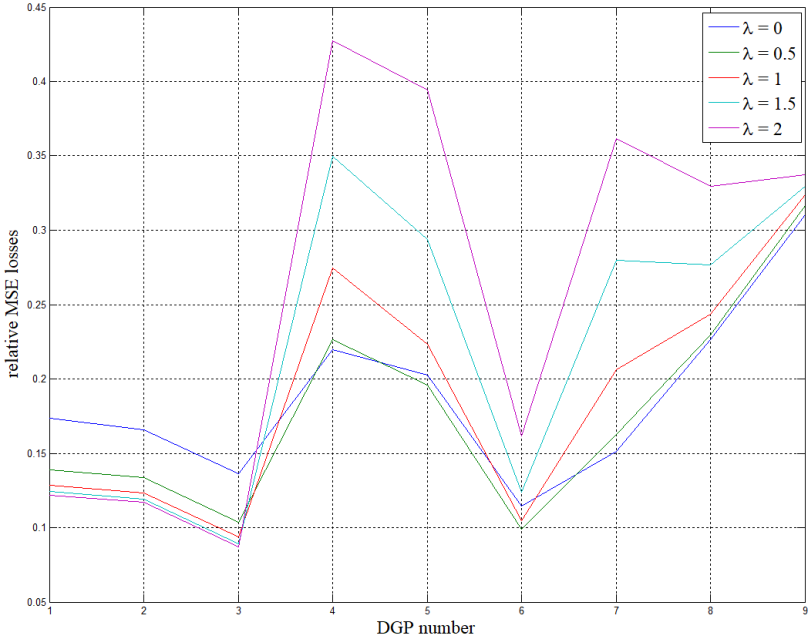

Figure 5.36 – relative MSE losses (y-axis) versus DGP number (x-axis), for S&P 500 index, N = 500 and Student t normalized innovations.

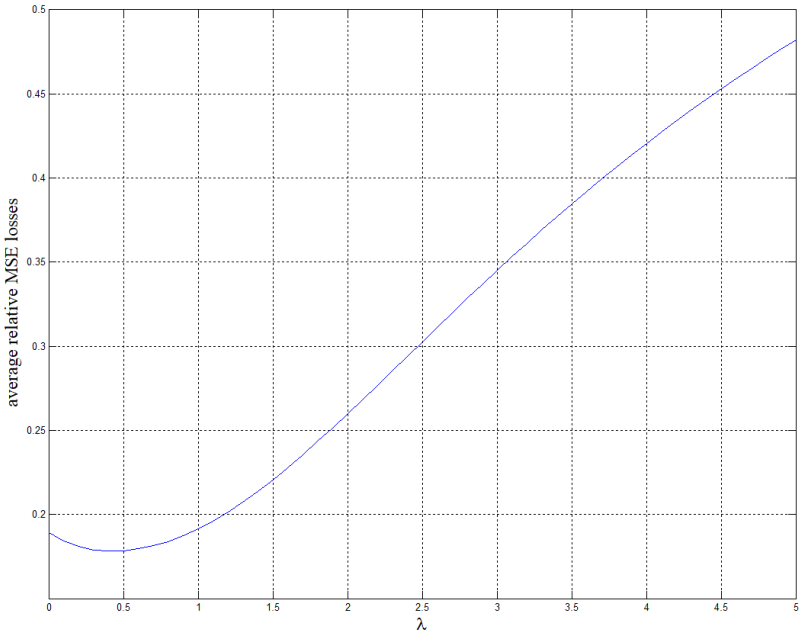

Figure 5.37 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for S&P 500 index, N = 500 and Student t normalized innovations.

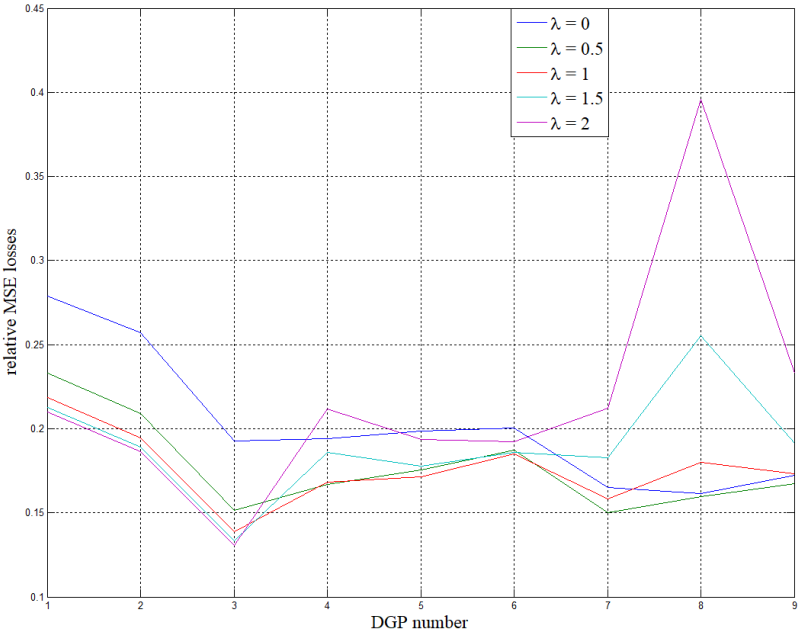

Figure 5.38 – relative MSE losses (y-axis) versus DGP number (x-axis), for N225 index, N = 500 and Student t normalized innovations.

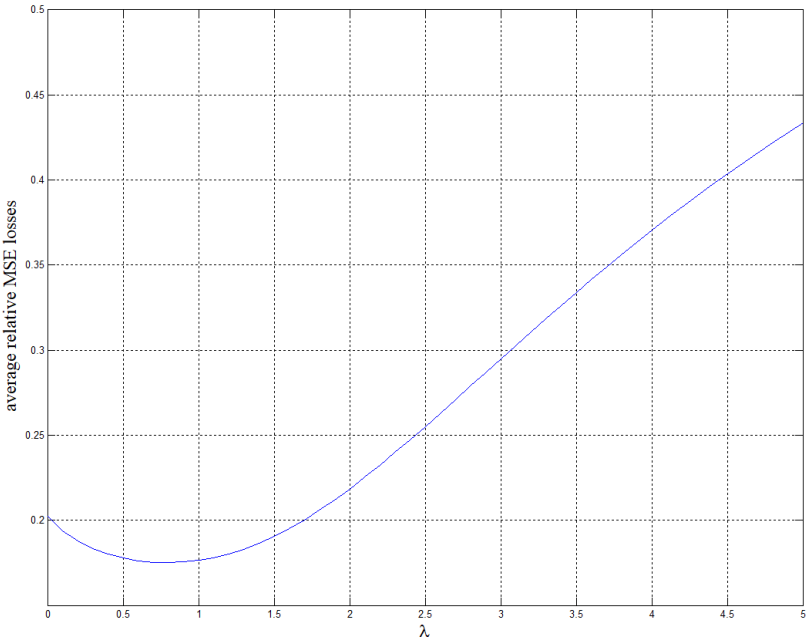

Figure 5.39 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for N225 index, N = 500 and Student t normalized innovations.

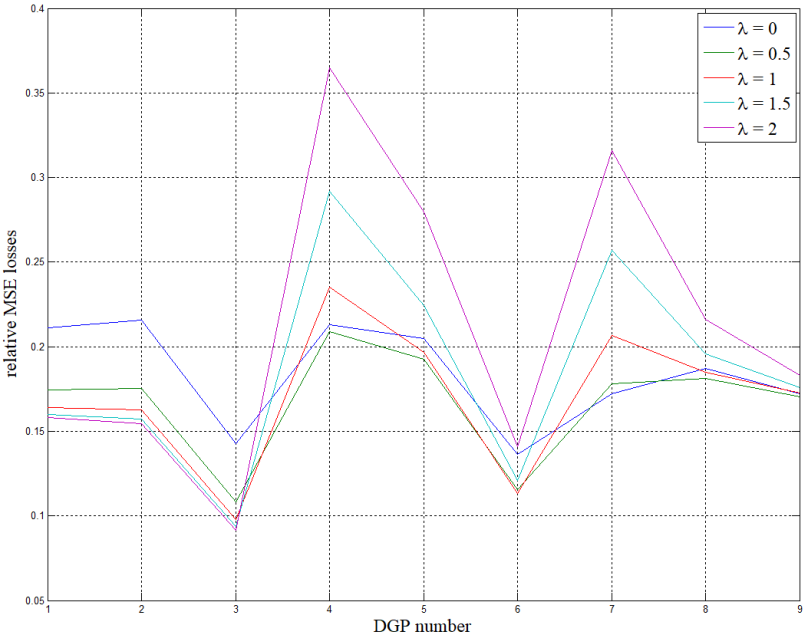

Figure 5.40 – relative MSE losses (y-axis) versus DGP number (x-axis), for DAX index, N = 500 and Student t normalized innovations.

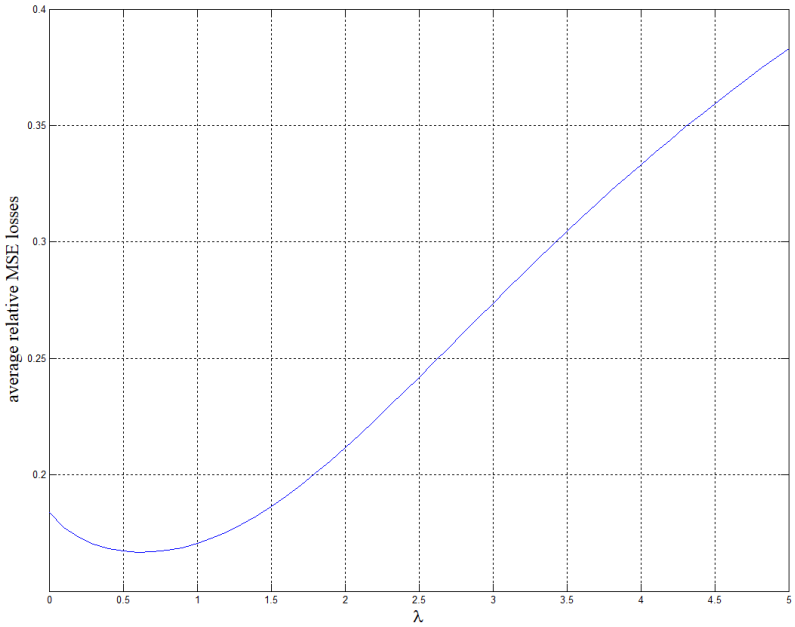

Figure 5.41 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for DAX index, N = 500 and Student t normalized innovations.

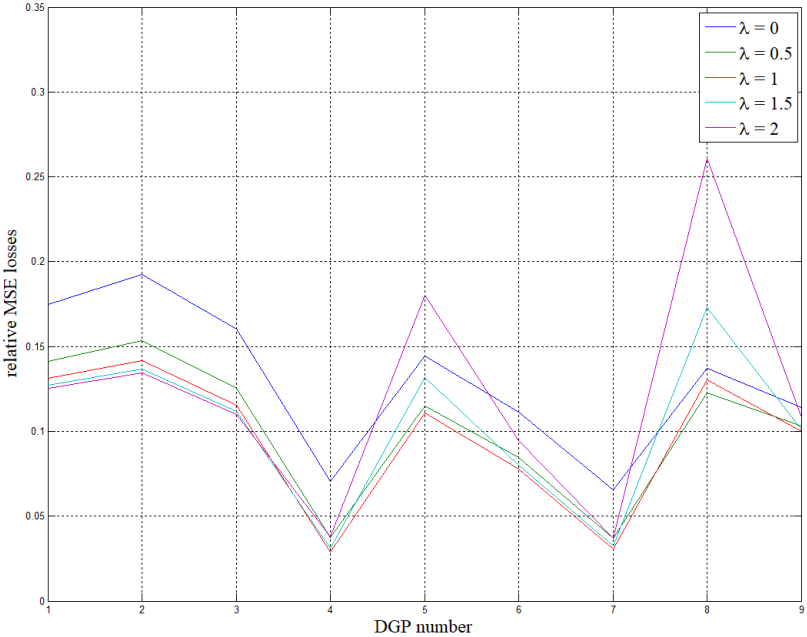

Figure 5.42 – relative MSE losses (y-axis) versus DGP number (x-axis), for FTSE 100 index, N = 500 and Student t normalized innovations.

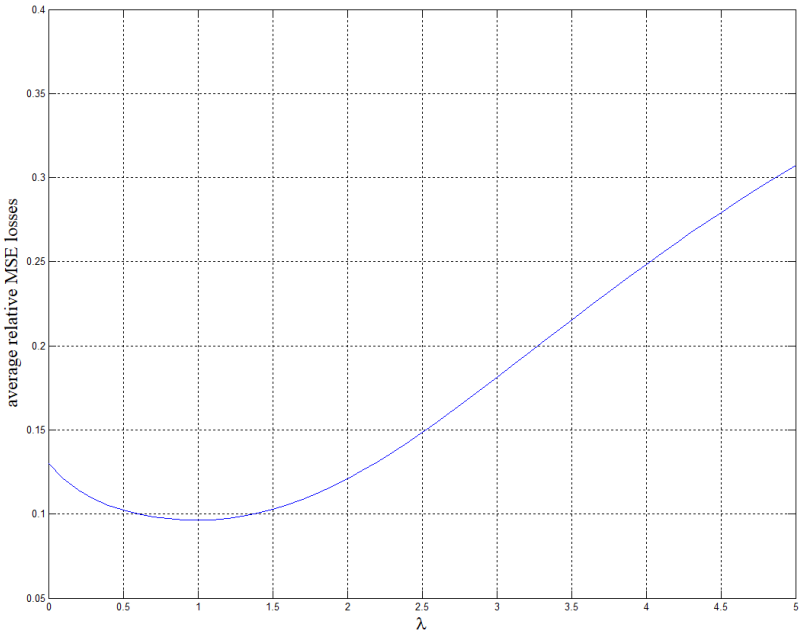

Figure 5.43 – average (across DGPs) relative MSE losses (y-axis) versus λ (x-axis), for FTSE 100 index, N = 500 and Student t normalized innovations.

Lastly, it follows the relative MSE losses as functions of λ when averaged over all DGPs and all markets.

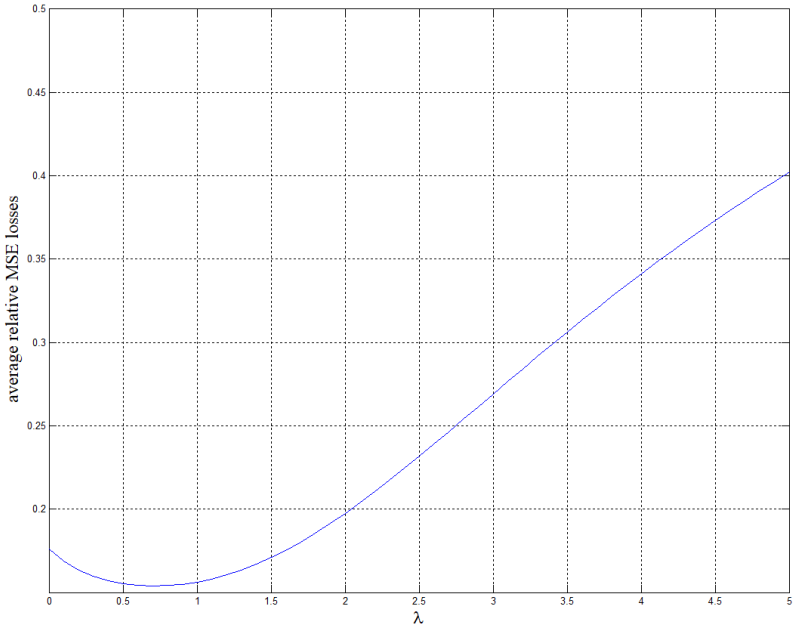

Figure 5.44 – average (across DGPs and markets) relative MSE losses (y-axis) versus λ (x-axis), for N = 500 and Student t normalized innovations.

The above results are very similar to the previous obtained for N = 250, with smaller (compared to Gaussian case) range of benefit for λ, which brings performance degradation when higher than a threshold between 1 and 2 (the exact threshold depending on the specific case). The optimal point for λ for the five indexes – IBOV, S&P 500, N225, DAX and FTSE 100 – were very similar one to another, being equal to 0.9, 0.4, 0.8, 0.8 and 1, respectively. Here again only multiples of 0.1 have been evaluated, yielding aa precision error strictly lower than 0.1 for the optimal λ values determined.

When not only the DGP's but also all market's MSE relative losses are averaged, the mean sense optimal λ is 0.7.

Thus, the most noticeable qualitative difference relative to the Gaussian innovation scenario is the reduced dependence of optimal λ on the number of samples (at least for the range examined), since the increase from N = 250 to N = 500 caused optimal λ value to be much smaller, whereas they are roughly the same when normalized innovations are modeled as Student t distributions.

5.4 CHAPTER CONCLUSIONS

In this chapter, the performances of individual EGARCH models were explored, regarding the MSE of one sample ahead returns standard deviations forecasts. Moreover, those same performances regarding the model selection and averaging strategies depicted in the previous chapter were presented. Extensive results including nine DGPs for each of the five markets analyzed and two sample sizes, allowed for several conclusions, with high statistical significance due to the synthetic data usage.

Firstly, while overfitting severely underperformed as expected, underfitting was found to be often beneficial, and the causes for that phenomenon were discussed in terms of bias-variance tradeoff for a small sample (rather than asymptotic) scenario.

Secondly, model averaging strategies were shown to be able to deliver outperformances even over the best forecasting fixed models in general, with the SIC linear averaging being the best overall strategy.

From the existing strategies, our analysis moved to the approach of replacing the regular SIC criterion with the generalized one, where positive values of the hyperparameter λ allow for extra complexity penalties. Regarding this strategy, its overall potential for outperforming regular SIC was clear, although strongly dependent on the data specifics such as number of data samples (N), complexity of DGP (order parameters P and Q), and market particularities (reflected in the DGPs fitted from real data).

Thus, our main proposals are the use of the generalized SIC criterion introduced in this work for model averaging, combined with the methodology suggested for data simulation and forecast performance evaluation to help in the choice of the most suitable value for λ. This methodology can be applied to other model families (different from EGARCH) and to other sets of orders to compose the models being averaged, being thus useful for other applications, markets and figures of merit, rather than just to the ones exploited in this work.

As a secondary contribution, when not strictly following the above advice for whatever reason, we suggest heuristic values for λ for each N analyzed, which can be generalized to $\lambda = 500/N$, so that it is asymptotically zero as $N \to \infty$, and thus converges to the regular SIC.

As an example of different scenario, we also showed the results corresponding to replacing the Gaussian distribution for the normalized innovations in the DGP models with a Student t distribution. The benefits of generalized SIC averaging were still significant, mainly for λ

in the 0.4 to 1 range, whereas the heuristic $\lambda = 500/N$ failed to provide better than regular SIC ($\lambda = 0$) performances, which reinforces the superiority of the methodology proposed over the heuristic suggested.

6 CONCLUSIONS

This work examined strategies for one sample ahead standard deviation forecasting, in financial series for logarithmic returns. The study considered EGARCH models estimated from maximum likelihood and concentrated on the problem of selection or averaging of different order models.

The high sensitivity of conclusions found in the literature to particularities of the real data has led us to propose a study based on a synthetic data generation methodology for evaluation of forecast strategies. Under this approach, deficiencies of the chosen model to represent some set of real data were isolated and thus unable to prevent the conclusions to be statistically significant. The conclusions of our study are derived based on data generated from the chosen model, which is assumed to be adequate from its support as a good model in the literature. Moreover, the methodology is extendable to any other model to be considered to be more adequate for a given application, or even to any set comprised of different models.

In our application of the proposed methodology, using EGARCH models of orders ranging from one to three, we found that the best forecast strategy was to average the model's forecasts, using weights proportional to the SIC (Schwarz Information Criterion). This result was shown to be due to the combination of higher complexity penalties imposed by SIC, and to an interesting, counterintuitive effect, the underfitting outperformance for forecasting purposes, possible for the small sample case, as opposed to the asymptotic behavior hypothesis.

Moreover, under this scenario of benefits of using lower than correct orders, we suggested the insertion of a hyperparameter λ into SIC calculation, which is able to modulate higher complexity penalties. The use of the corresponding modified criterion, the generalized SIC, combined with the model averaging strategy, was able to deliver the intended outperformances.

Thus, the methodology proposed includes not only the synthetic data usage to provide strong conclusions about the forecasting performances of different models and selection or averaging strategies, but also a new criterion, the generalized SIC, whose hyperparameter λ can be devised in such a framework.

In the scenarios exploited in this work, the application of the proposed methodology provided different ranges of suitable values of λ depending on data specifics, such as the market analyzed and number of samples available, which reinforces our interpretation of the variability

of conclusions in the literature and the value of the proposed methodology, due to its flexibility attribute.

Nevertheless, we secondarily suggested a simple heuristic of $\lambda = 500/N$, based on the results of the scenarios analyzed. However, its value is potentially restricted to the particularities of the data used, and this hypothesis is reinforced by its bad performance in the examples discussed in Section 5.3 in which the normalized Gaussian innovations were replaced with Student t innovations. In that scenario, for all markets and numbers of samples, the most suitable values for λ were mainly in the interval ranging from 0.4 to 1 (the exact optimal point depending on the data specifics).

As suggestions for future works, we mention the following:

- Analysis of returns mean parameter models and evaluation of their joint estimation effects;
- Real data scrutiny aiming to confront two hypothesis that could explain the better results of minimum order models – if it happens because the data is better described by these models or because even being better described by higher order models, the forecasting performances of the formers are better due to underfitting benefits exploited in this work;
- Analysis of the forecasting benefits of underfitting as a function of N, considering its decay rate and number of samples necessary to the benefits become negligible and asymptotic premises acceptable;
- Quality evaluation of the $\lambda = 500/N$ heuristic to other numbers of samples, markets, kinds of data and models under consideration (averaging).

REFERENCES

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

BALABAN, E. Comparative forecasting performance of symmetric and asymmetric conditional volatility models of an exchange rate. **Economics Letters**, v. 83, n. 1, p. 99–105, 2004.

BATES, J. M.; GRANGER, C. W. J. The Combination of Forecasts. **Journal of the Operational Research Society**, v. 20, n. 4, p. 451–468, 1969.

BERA, A. K.; HIGGINS, M. L. ARCH Models: Properties, Estimation and Testing. **Journal of Economic Surveys**, v. 7, n. 4, p. 305–366, 1993.

BOLLERSLEV, T. Generalized Autoregressive Conditional Heteroskedasticity. **Journal of Econometrics**, v. 31, p. 307–327, 1986.

BUCKLAND, S. T.; BURNHAM, K. P.; AUGUSTIN, N. H. Model Selection: An Integral Part of Inference. **Biometrics**, v. 53, n. 2, p. 603–618, 1997.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. **Sociological Methods & Research**, v. 33, n. 2, p. 261–304, 2004.

CHENG, T.-C. F.; ING, C.-K.; YU, S.-H. Toward optimal model averaging in regression models with time series errors. **Journal of Econometrics**, v. 189, n. 2, p. 321–334, 2015.

CLAESKENS, G.; HJORT, N. L. **Model selection and model averaging**. Cambridge: Cambridge University Press, 2008.

CLARKE, J.; JANDIK, T.; MANDELKER, G. The Efficient Markets Hypothesis. In: ARFFA, R. C. (Ed.). . **Expert financial planning: Advice from industry leaders**. New York: John Wiley & Sons, 2001. p. 126–141.

CLEMENTS, M. P.; HENDRY, D. F. Forecasting economic processes. **International Journal of Forecasting**, v. 14, n. 1, p. 111–131, 1998.

DUAN, J.-C. et al. Approximating the GJR-GARCH and EGARCH Option Pricing Models Analytically. **Journal of Computational Finance**, v. 9, n. 3, p. 41, 2006.

ENGLE. Autoregressive Conditional Heteroscedacity with Estimates of variance of United Kingdom Inflation. **Econometrica**, v. 50, n. 4, p. 987–1008, 1982.

EZZAT, H. The Application of GARCH and EGARCH in Modeling the Volatility of Daily Stock Returns During Massive Shocks: The Empirical Case of Egypt. **International Research Journal of Finance and Economics**, n. 96, p. 143–154, 2012.

FAMA, E. F. Efficient Capital Markets: a Review of Theory and Empirical Work. **Journal of Finance**, v. 25, n. 2, p. 383–417, 1970.

FAMA, E. F.; FRENCH, K. R. The Capital Asset Pricing Model : Theory and Evidence. **Journal of Economic Perspectives**, v. 18, n. 3, p. 25–46, 2004.

FRANCQ, C.; ZAKOIAN, J.-M. **GARCH Models: Structure, Statistical Inference and Financial Applications**. [s.l.] John Wiley & Sons, 2010.

GRIFFIOEN, G. A. W. **Technical Analysis in Financial Markets**. Ph.D. thesis. University of Amsterdam, 2003.

HAMILTON, J. D.; SUSMEL, R. Autoregressive conditional heteroskedasticity and changes in regime. **Journal of Econometrics**, v. 64, n. 1, p. 307–333, 1994.

HANSEN, P. R.; LUNDE, A. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? **Journal of Applied Econometrics**, v. 20, n. 7, p. 873–889, 2005.

JAMES, A.; CHAN, F. Application of Forecast Combination in Volatility Modelling. In: 19TH INTERNATIONAL CONGRESS OF MODELLING AND SIMULATION**,** 19, 2011, Perth. **Proceedings…** Perth: Modeling and Simulation Society of Australia and New Zealand, 2011. p. 1610–1616.

KAY, S. M. **Fundamentals of statistical signal processing: estimation theory**. New Jersey: Prentice-Hall, 1993.

KONISHI, S.; KITAGAWA, G. **Information Criteria and Statistical Modeling**. New York: Springer, 2008.

KUHA, J. AIC and BIC: Comparisons of Assumptions and Performance. **Sociological Methods & Research**, v. 33, n. 2, p. 188–229, 2004.

LI, G.; LI, Y. Forecasting copper futures volatility under model uncertainty. **Resources Policy**, v. 46, p. 167–176, 2015.

LI, Y.; HUANG, W.-P.; ZHANG, J. Forecasting volatility in the Chinese stock market under model uncertainty. **Economic Modelling**, v. 35, p. 231–234, 2013.

MANOLAKIS, D. G.; INGLE, V. K.; KOGON, S. M. **Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing**. Boston: McGraw-Hill, 2000.

MITCHELL, H.; MCKENZIE, M. D. GARCH model selection criteria. **Quantitative Finance**, v. 3, p. 262–284, 2003.

NELSON, D. B. Conditional Heteroskedasticity in Asset Returns: A New Approach. **Econometrica**, v. 59, n. 2, p. 347–370, 1991.

POON, S.-H.; GRANGER, C. W. J. Forecasting Volatility in Financial Markets: A review. **Journal of Economic Literature**, v. 41, n. 2, p. 478–539, 2003.

RUPPERT, D. **Statistics and Data Analysis for Financial Engineering**. New York: Springer, 2011.

STOICA, P.; SELEN, Y. Model-order selection: a review of information criterion rules. **IEEE Signal Processing Magazine**, v. 21, n. 4, p. 36–47, 2004.

STOICA, P.; SELÉN, Y.; LI, J. Multi-model approach to model selection. **Digital Signal Processing**, v. 14, p. 399–412, 2004.

STRAUMANN, D.; MIKOSCH, T. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. **The Annals of Statistics**, v. 34, n. 5, p. 2449–2495, 2006.

SU, C. **Application of EGARCH Model to Estimate Financial Volatility of Daily Returns: The empirical case of China**. Master Degree Project No.2010:142. University of Gothenburg, 2010.

TIMMERMANN, A. Forecast Combinations. In: ELLIOTT, G.; GRANGER, C. W. J.; TIMMERMANN, A. (Eds.). . **Handbook of Economic Forecasting**. 1. ed. Amsterdam: Elsevier, 2006. p. 135–196.

TSAY, R. S. **Analysis of Financial Time Series**. 2. ed. Hoboken: John Wiley & Sons, 2005.

VRONTOS, I. D.; DELLAPORTAS, P.; POLITIS, D. N. Full Bayesian Inference for GARCH and EGARCH Models. **Journal of business & Economics Statistics**, v. 18, n. 2, p. 187–198, 2000.

WEI-MING, H.; ZHONG-FU, L. A comparison of Forecast Models of REIT volatility: GARCH Model, AFIMA Model, Markov Switching Model. In: INTERNATIONAL CONFERENCE ON MANAGEMENT SCIENCE & ENGINEERING, 19, 2012, Dallas. **Proceedings…** Dallas: IEEE, 2012. p. 270–275.

ZHANG, K. et al. Bridging Information Criteria and Parameter Shrinkage for Model Selection. **arXiv:1307.2307v1 [stat.ML]**, 2013. Available in < https://arxiv.org/abs/1307.2307>.

APPENDIX A – EGARCH models with Gaussian normalized innovations fitted from real data

In this appendix we list the EGARCH models that were fitted to real data. The real data were the logarithmic returns of five major stock indexes: Ibovespa or "IBOV" (Brazil's stock index), Standard & Poor's 500 or "SP500" (USA stock index), Nikkei 225 or "N225" (Japan's stock index), "DAX" (Germany's stock index) and "FTSE 100" (England's stock index), all taken from Yahoo! Finance from January 03, 2000 to April 09, 2015.

These models differ by the order parameters P and Q and are listed as follows for each stock index. All the models have Gaussian normalized innovations, such that z_t are independent, zero-mean, unit-variance Gaussian distributions.

The general model is

$$r_t = C + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = \kappa + \sum_{i=1}^{P} G_i \cdot \ln\left(\sigma_{t-i}^{\,2}\right) + \sum_{j=1}^{Q} A_j \cdot \left(\left| z_{t-j} \right| - \sqrt{\frac{2}{\pi}} \right) + \sum_{j=1}^{Q} L_j \cdot z_{t-j}$$

A.1 IBOV models

P=Q=1 model:

$$r_t = 1.4526 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1776 + 0.9781 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1218 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$-0.0751 \cdot z_{t-1}$$

P=1, Q=2 model:

$$r_t = 1.9783 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1853 + 0.9772 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0595 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$+0.1886 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.1883 \cdot z_{t-1} + 0.1119 \cdot z_{t-2}$$

P=1, Q=3 model:

$$r_t = 1.8892 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1320 + 0.9837 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0635 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.1972 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.0166 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1864 \cdot z_{t-1}$$

$$+0.0201 \cdot z_{t-2} + 0.0989 \cdot z_{t-3}$$

P=2, Q=1 model:

$$r_t = 1.4285 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1704 + 1.0402 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0612 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.1163 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.0708 \cdot z_{t-1}$$

P=2, Q=2 model:

$$r_t = 1.5269 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.0404 + 1.6479 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6529 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.0152 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.0249 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1932 \cdot z_{t-1}$$

$$+0.1757 \cdot z_{t-2}$$

P=2, Q=3 model:

$$r_t = 1.4385 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.0249 + 1.7188 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7219 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$-0.0826 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.2698 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$-0.1600 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1806 \cdot z_{t-1} + 0.1661 \cdot z_{t-2} + 0.0030 \cdot z_{t-3}$$

P=3, Q=1 model:

$$r_t = 1.3608 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1517 + 1.3814 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7325 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.3324 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1105 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.0671 \cdot z_{t-1}$$

P=3, Q=2 model:

$$r_t = 1.5244 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0387 + 1.6483 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6408 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0123 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0171 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$+0.0213 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1929 \cdot z_{t-1} + 0.1764 \cdot z_{t-2}$$

P=3, Q=3 model:

$$r_t = 1.3773 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0326 + 1.3408 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0385 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.3063 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0634 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.2436 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$-0.1443 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1720 \cdot z_{t-1} + 0.0806 \cdot z_{t-2} + 0.0773 \cdot z_{t-3}$$

A.2 S&P 500 models

P=Q=1 model:

$$r_t = 8.6519 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1874 + 0.9795 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1105 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$-0.1437 \cdot z_{t-1}$$

P=1, Q=2 model:

$$r_t = 2.9735 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.2258 + 0.9752 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1147 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.2492 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2309 \cdot z_{t-1} + 0.0799 \cdot z_{t-2}$$

P=1, Q=3 model:

$$r_t = -1.5375 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1836 + 0.9798 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1255 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.2053 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.0435 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2342 \cdot z_{t-1}$$

$$-0.0349 \cdot z_{t-2} + 0.1361 \cdot z_{t-3}$$

P=2, Q=1 model:

$$r_t = 8.7417 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1824 + 1.0182 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0382 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.1081 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1385 \cdot z_{t-1}$$

P=2, Q=2 model:

$$r_t = -3.8744 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0958 + 1.5062 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.5167 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0794 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right) + 0.1465 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2679 \cdot z_{t-1}$$
$$+0.2040 \cdot z_{t-2}$$

P=2, Q=3 model:

$$r_t = -2.0236 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0998 + 1.3680 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.3790 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.1257 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right) + 0.2495 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right)$$
$$-0.0505 \cdot \left(\left| z_{t-3} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2336 \cdot z_{t-1} + 0.0624 \cdot z_{t-2} + 0.0960 \cdot z_{t-3}$$

P=3, Q=1 model:

$$r_t = 8.6512 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1787 + 1.2328 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.5648 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.3125 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1054 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.1472 \cdot z_{t-1}$$

P=3, Q=2 model:

$$r_t = -3.0041 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1000 + 1.6378 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.8393 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1904 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0839 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$
$$+0.1532 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2534 \cdot z_{t-1} + 0.1819 \cdot z_{t-2}$$

P=3, Q=3 model:

$$r_t = -3.0148 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1814 + 0.6288 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.7582 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$-0.4070 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0863 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.0952 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.1240 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2377 \cdot z_{t-1} - 0.1054 \cdot z_{t-2} + 0.2213 \cdot z_{t-3}$$

A.3 N225 models

P=Q=1 model:

$$r_t = 9.7515 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.2717 + 0.9679 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1927 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$-0.0822 \cdot z_{t-1}$$

P=1, Q=2 model:

$$r_t = 1.0877 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.2943 + 0.9653 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0342 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.1700 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1609 \cdot z_{t-1} + 0.0801 \cdot z_{t-2}$$

P=1, Q=3 model:

$$r_t = 1.3263 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^2\right) = -0.2526 + 0.9702 \cdot \ln\left(\sigma_{t-1}^2\right) + 0.0356 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right)$$

$$+0.1094 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}} \right) + 0.0553 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.1557 \cdot z_{t-1}$$

$$-0.0249 \cdot z_{t-2} + 0.1075 \cdot z_{t-3}$$

P=2, Q=1 model:

$$r_t = 9.4590 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^2\right) = -0.2430 + 1.1302 \cdot \ln\left(\sigma_{t-1}^2\right) - 0.1588 \cdot \ln\left(\sigma_{t-2}^2\right)$$

$$+0.1694 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.0707 \cdot z_{t-1}$$

P=2, Q=2 model:

$$r_t = 1.4709 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^2\right) = -0.0707 + 1.6669 \cdot \ln\left(\sigma_{t-1}^2\right) - 0.6752 \cdot \ln\left(\sigma_{t-2}^2\right)$$

$$+0.0835 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.0177 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.1841 \cdot z_{t-1}$$

$$+0.1672 \cdot z_{t-2}$$

P=2, Q=3 model:

$$r_t = 1.3864 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^2\right) = -0.0579 + 1.6669 \cdot \ln\left(\sigma_{t-1}^2\right) - 0.6737 \cdot \ln\left(\sigma_{t-2}^2\right)$$

$$+0.0228 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right) + 0.1331 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}} \right)$$

$$-0.0965 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.1615 \cdot z_{t-1} + 0.1030 \cdot z_{t-2} + 0.0445 \cdot z_{t-3}$$

P=3, Q=1 model:

$$r_t = 8.9061 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.2235 + 1.4127 \cdot \ln\left(\sigma_{t-1}{}^2\right) - 0.7039 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$+0.2648 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.1586 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.0677 \cdot z_{t-1}$$

P=3, Q=2 model:

$$r_t = 1.5273 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.0657 + 1.8613 \cdot \ln\left(\sigma_{t-1}{}^2\right) - 1.0594 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$+0.1904 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.0751 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right)$$
$$-0.0118 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.1663 \cdot z_{t-1} + 0.1492 \cdot z_{t-2}$$

P=3, Q=3 model:

$$r_t = 1.3906 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.0644 + 1.5617 \cdot \ln\left(\sigma_{t-1}{}^2\right) - 0.4942 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$-0.0752 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.0272 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}} \right) + 0.1352 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}} \right)$$
$$-0.0965 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}} \right) - 0.1603 \cdot z_{t-1} + 0.0818 \cdot z_{t-2} + 0.0630 \cdot z_{t-3}$$

A.4 DAX models

P=Q=1 model:

$$r_t = 2.0702 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1814 + 0.9791 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1223 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$-0.1182 \cdot z_{t-1}$$

P=1, Q=2 model:

$$r_t = 1.2706 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1940 + 0.9776 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0710 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$+0.2061 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2082 \cdot z_{t-1} + 0.0956 \cdot z_{t-2}$$

P=1, Q=3 model:

$$r_t = 1.1165 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1679 + 0.9805 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0634 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$+0.1575 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) + 0.0381 \cdot \left(\left| z_{t-3} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2085 \cdot z_{t-1}$$

$$-0.0082 \cdot z_{t-2} + 0.0981 \cdot z_{t-3}$$

P=2, Q=1 model:

$$r_t = 2.0593 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1800 + 0.9891 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0099 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.1214 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.1170 \cdot z_{t-1}$$

P=2, Q=2 model:

$$r_t = 1.0399 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0386 + 1.7083 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7128 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.0213 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.0157 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2252 \cdot z_{t-1}$$
$$+0.2026 \cdot z_{t-2}$$

P=2, Q=3 model:

$$r_t = 1.0491 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0269 + 1.7628 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7659 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0603 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.2208 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$-0.1332 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2130 \cdot z_{t-1} + 0.1962 \cdot z_{t-2} + 0.0007 \cdot z_{t-3}$$

P=3, Q=1 model:

$$r_t = 2.0870 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1570 + 1.3856 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7302 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.3265 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1100 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1047 \cdot z_{t-1}$$

P=3, Q=2 model:

$$r_t = 1.0203 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0365 + 1.7162 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7170 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0034 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0226 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$+0.0128 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.2249 \cdot z_{t-1} + 0.2035 \cdot z_{t-2}$$

P=3, Q=3 model:

$$r_t = 1.2129 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.0326 + 1.3955 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0846 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$-0.3147 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0281 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right) + 0.1638 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$-0.1012 \cdot \left(\left| z_{t-3} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.2021 \cdot z_{t-1} + 0.0986 \cdot z_{t-2} + 0.0847 \cdot z_{t-3}$$

A.5 FTSE 100 models

P=Q=1 model:

$$r_t = -7.0491 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1582 + 0.9826 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1139 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$-0.1209 \cdot z_{t-1}$$

P=1, Q=2 model:

$$r_t = -8.3244 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1612 + 0.9823 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0200 \cdot \left(\left| z_{t-1} \right| - \sqrt{\frac{2}{\pi}} \right)$$

$$+0.0995 \cdot \left(\left| z_{t-2} \right| - \sqrt{\frac{2}{\pi}} \right) - 0.1695 \cdot z_{t-1} + 0.0546 \cdot z_{t-2}$$

P=1, Q=3 model:

$$r_t = -8.7090 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1487 + 0.9837 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0179 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.0985 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.0006 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1704 \cdot z_{t-1}$$

$$+0.0150 \cdot z_{t-2} + 0.0457 \cdot z_{t-3}$$

P=2, Q=1 model:

$$r_t = -7.0732 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1597 + 0.9700 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0125 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.1151 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1223 \cdot z_{t-1}$$

P=2, Q=2 model:

$$r_t = -6.7163 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.3141 + 0.0115 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.9541 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.1102 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.1150 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1295 \cdot z_{t-1}$$

$$-0.1084 \cdot z_{t-2}$$

P=2, Q=3 model:

$$r_t = -7.8434 \cdot 10^{-5} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.3232 + 0.0061 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.9584 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$

$$+0.0207 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.1198 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$

$$+0.0964 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1701 \cdot z_{t-1} - 0.1055 \cdot z_{t-2} + 0.0473 \cdot z_{t-3}$$

P=3, Q=1 model:

$$r_t = -6.6684 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.1458 + 1.2079 \cdot \ln\left(\sigma_{t-1}{}^2\right) - 0.3944 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$+0.1705 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.1066 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1129 \cdot z_{t-1}$$

P=3, Q=2 model:

$$r_t = -7.8762 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.0122 + 1.8342 \cdot \ln\left(\sigma_{t-1}{}^2\right) - 0.8223 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$-0.0132 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.0997 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$-0.0857 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1911 \cdot z_{t-1} + 0.1818 \cdot z_{t-2}$$

P=3, Q=3 model:

$$r_t = -8.1681 \cdot 10^{-5} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t{}^2\right) = -0.3104 + 0.3331 \cdot \ln\left(\sigma_{t-1}{}^2\right) + 0.3645 \cdot \ln\left(\sigma_{t-2}{}^2\right)$$
$$+0.2684 \cdot \ln\left(\sigma_{t-3}{}^2\right) + 0.0179 \cdot \left(\left|z_{t-1}\right| - \sqrt{\frac{2}{\pi}}\right) + 0.1011 \cdot \left(\left|z_{t-2}\right| - \sqrt{\frac{2}{\pi}}\right)$$
$$+0.1125 \cdot \left(\left|z_{t-3}\right| - \sqrt{\frac{2}{\pi}}\right) - 0.1629 \cdot z_{t-1} - 0.0783 \cdot z_{t-2} + 0.0191 \cdot z_{t-3}$$

APPENDIX B – EGARCH models with Student t normalized innovations fitted from real data

In this appendix we list the EGARCH models that were fitted to real data. The real data were the logarithmic returns of five major stock indexes: Ibovespa or "IBOV" (Brazil's stock index), Standard & Poor's 500 or "SP500" (USA stock index), Nikkei 225 or "N225" (Japan's stock index), "DAX" (Germany's stock index) and "FTSE 100" (England's stock index), all taken from Yahoo! Finance from January 03, 2000 to April 09, 2015.

These models differ by the order parameters P and Q and are listed as follows for each stock index. All the models have Student t normalized innovations, such that z_t are independent, zero-mean, unit-variance Student t distributions whose degrees of freedom were jointly fitted from the data and as such this parameter is also indicated for each model and labeled as *DoF*.

The general model is

$$r_t = C + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = \kappa + \sum_{i=1}^{P} G_i \cdot \ln\left(\sigma_{t-i}^{\,2}\right) + \sum_{j=1}^{Q} A_j \cdot \left(\left|z_{t-j}\right| - E\left(\left|z_t\right|\right)\right) + \sum_{j=1}^{Q} L_j \cdot z_{t-j}$$

$$E\left(\left|z_t\right|\right) = \sqrt{\frac{DoF-2}{\pi}} \cdot \frac{\Gamma\left(DoF - 1\!\!\diagup\!\!2\right)}{\Gamma\left(DoF\!\!\diagup\!\!2\right)}$$

B.1 IBOV models

P=Q=1 model (*DoF* = 14.9924):

$$r_t = 2.8192 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1607 + 0.9804 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1203 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.0746 \cdot z_{t-1}$$

P=1, Q=2 model (*DoF* = 16.0436):

$$r_t = 3.1250 \cdot 10^{-4} + \sigma_t \cdot z_t$$

$$\ln\left(\sigma_t^{\,2}\right) = -0.1708 + 0.9793 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0704 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.1985 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.1870 \cdot z_{t-1} + 0.1111 \cdot z_{t-2}$$

P=1, Q=3 model (*DoF* = 16.8197):

$$r_t = 2.9325 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1269 + 0.9846 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0747 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2054 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.0133 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1812 \cdot z_{t-1}$$
$$+0.0190 \cdot z_{t-2} + 0.0983 \cdot z_{t-3}$$

P=2, Q=1 model (*DoF* = 14.9543):

$$r_t = 2.8016 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1517 + 1.0658 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0843 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1128 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.0687 \cdot z_{t-1}$$

P=2, Q=2 model (*DoF* = 16.7677):

$$r_t = 2.6948 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0400 + 1.6295 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6343 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.0070 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.0350 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.1972 \cdot z_{t-1}$$
$$+0.1786 \cdot z_{t-2}$$

P=2, Q=3 model (*DoF* = 17.0855):

$$r_t = 2.6273 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0245 + 1.7034 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7064 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0867 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2794 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1640 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1809 \cdot z_{t-1} + 0.1589 \cdot z_{t-2} + 0.0098 \cdot z_{t-3}$$

P=3, Q=1 model (*DoF* = 14.5818):

$$r_t = 2.6674 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0977 + 1.8757 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 1.4476 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.5600 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0787 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.0458 \cdot z_{t-1}$$

P=3, Q=2 model (*DoF* = 16.8503):

$$r_t = 2.7019 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0408 + 1.6279 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6348 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0019 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0048 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.0379 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.1980 \cdot z_{t-1} + 0.1791 \cdot z_{t-2}$$

P=3, Q=3 model (*DoF* = 16.9327):

$$r_t = 2.5348 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0288 + 1.4057 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1545 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.2547 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0758 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2664 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1564 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1789 \cdot z_{t-1} + 0.1044 \cdot z_{t-2} + 0.0608 \cdot z_{t-3}$$

B.2 S&P 500 models

P=Q=1 model (*DoF* = 8.7234):

$$r_t = 3.0678 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1482 + 0.9844 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1017 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1521 \cdot z_{t-1}$$

P=1, Q=2 model (*DoF* = 8.5448):

$$r_t = 2.7687 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1792 + 0.9811 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1505 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2726 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2775 \cdot z_{t-1} + 0.1202 \cdot z_{t-2}$$

P=1, Q=3 model (*DoF* = 8.8785):

$$r_t = 2.3348 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1431 + 0.9849 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1521 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2316 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) + 0.0298 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2791 \cdot z_{t-1}$$
$$-0.0005 \cdot z_{t-2} + 0.1406 \cdot z_{t-3}$$

P=2, Q=1 model (*DoF* = 8.7114):

$$r_t = 3.0669 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1503 + 0.9632 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0210 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1028 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1552 \cdot z_{t-1}$$

P=2, Q=2 model (*DoF* = 8.7371):

$$r_t = 2.1548 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0736 + 1.4806 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.4884 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0980 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.1567 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.3129 \cdot z_{t-1}$$
$$+0.2435 \cdot z_{t-2}$$

P=2, Q=3 model (*DoF* = 8.7609):

$$r_t = 2.4997 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0256 + 1.6957 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6983 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.1605 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.3775 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1937 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2820 \cdot z_{t-1} + 0.2316 \cdot z_{t-2} + 0.0187 \cdot z_{t-3}$$

P=3, Q=1 model (*DoF* = 8.7275):

$$r_t = 3.0663 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1529 + 1.1628 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.5108 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.3321 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1020 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1711 \cdot z_{t-1}$$

P=3, Q=2 model (*DoF* = 8.8285):

$$r_t = 2.2421 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0905 + 1.5699 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7830 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.2035 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.1093 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.1792 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2917 \cdot z_{t-1} + 0.2039 \cdot z_{t-2}$$

P=3, Q=3 model (*DoF* = 8.9503):

$$r_t = 2.3925 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0517 + 1.3096 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1347 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.1803 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.1398 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2922 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1058 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2774 \cdot z_{t-1} + 0.0948 \cdot z_{t-2} + 0.1244 \cdot z_{t-3}$$

B.3 N225 models

P=Q=1 model (*DoF* = 11.8059):

$$r_t = 2.6006 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.2501 + 0.9711 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1670 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.0900 \cdot z_{t-1}$$

P=1, Q=2 model (*DoF* = 11.3458):

$$r_t = 2.4840 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.2854 + 0.9670 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0543 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2359 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.1788 \cdot z_{t-1} + 0.0837 \cdot z_{t-2}$$

P=1, Q=3 model (*DoF* = 11.4295):

$$r_t = 2.5312 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.2389 + 0.9724 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0507 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.1867 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) + 0.0388 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1713 \cdot z_{t-1}$$
$$-0.0194 \cdot z_{t-2} + 0.1072 \cdot z_{t-3}$$

P=2, Q=1 model (*DoF* = 11.8154):

$$r_t = 2.5855 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.2145 + 1.1733 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1981 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1415 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.0734 \cdot z_{t-1}$$

P=2, Q=2 model (*DoF* = 11.8537):

$$r_t = 2.5358 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0911 + 1.5738 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.5843 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.0231 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.0516 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2036 \cdot z_{t-1}$$
$$+0.1746 \cdot z_{t-2}$$

P=2, Q=3 model (*DoF* = 11.4231):

$$r_t = 2.5354 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0532 + 1.6525 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6587 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0610 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2680 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1548 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1760 \cdot z_{t-1} + 0.1129 \cdot z_{t-2} + 0.0457 \cdot z_{t-3}$$

P=3, Q=1 model (*DoF* = 11.5331):

$$r_t = 2.6547 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1700 + 1.6823 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1092 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.4072 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1172 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.0613 \cdot z_{t-1}$$

P=3, Q=2 model (*DoF* = 11.8916):

$$r_t = 2.6674 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0837 + 1.8509 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 1.1341 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.2736 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0252 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.0470 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.1754 \cdot z_{t-1} + 0.1466 \cdot z_{t-2}$$

P=3, Q=3 model (*DoF* = 11.3649):

$$r_t = 2.5326 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0525 + 1.6445 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6414 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0092 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0615 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2699 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1562 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.1753 \cdot z_{t-1} + 0.1105 \cdot z_{t-2} + 0.0476 \cdot z_{t-3}$$

B.4 DAX models

P=Q=1 model (*DoF* = 12.2587):

$$r_t = 4.0050 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1604 + 0.9820 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1256 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1266 \cdot z_{t-1}$$

P=1, Q=2 model (*DoF* = 11.5244):

$$r_t = 3.3346 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1712 + 0.9807 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1088 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2490 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2406 \cdot z_{t-1} + 0.1188 \cdot z_{t-2}$$

P=1, Q=3 model (*DoF* = 11.8340):

$$r_t = 3.1979 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1459 + 0.9835 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.1006 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.2057 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) + 0.0309 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2379 \cdot z_{t-1}$$
$$+0.0269 \cdot z_{t-2} + 0.1014 \cdot z_{t-3}$$

P=2, Q=1 model (*DoF* = 12.2581):

$$r_t = 4.0456 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1586 + 0.9949 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0127 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1244 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1250 \cdot z_{t-1}$$

P=2, Q=2 model (*DoF* = 12.1417):

$$r_t = 3.0324 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0596 + 1.5585 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.5653 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0321 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.0920 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2594 \cdot z_{t-1}$$
$$+0.2163 \cdot z_{t-2}$$

P=2, Q=3 model (*DoF* = 12.2879):

$$r_t = 3.2359 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0203 + 1.7857 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7880 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.1049 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.3091 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1799 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2417 \cdot z_{t-1} + 0.2484 \cdot z_{t-2} - 0.0224 \cdot z_{t-3}$$

P=3, Q=1 model (*DoF* = 12.2911):

$$r_t = 4.1488 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1416 + 1.3554 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7000 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.3287 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1159 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1163 \cdot z_{t-1}$$

P=3, Q=2 model (*DoF* = 12.1247):

$$r_t = 3.1174 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0776 + 1.5549 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6694 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1058 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0514 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.1260 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2507 \cdot z_{t-1} + 0.1936 \cdot z_{t-2}$$

P=3, Q=3 model (*DoF* = 12.4064):

$$r_t = 3.3115 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0240 + 1.5056 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.2665 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.2418 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) - 0.0796 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.2627 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1532 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2296 \cdot z_{t-1} + 0.1632 \cdot z_{t-2} + 0.0486 \cdot z_{t-3}$$

B.5 FTSE 100 models

P=Q=1 model (*DoF* = 10.9434):

$$r_t = 1.3941 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1521 + 0.9839 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.1151 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1340 \cdot z_{t-1}$$

P=1, Q=2 model (*DoF* = 10.8992):

$$r_t = 1.3918 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1537 + 0.9837 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0064 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.1281 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2014 \cdot z_{t-1} + 0.0734 \cdot z_{t-2}$$

P=1, Q=3 model (*DoF* = 10.9406):

$$r_t = 1.3112 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1450 + 0.9847 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0052 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$+0.0986 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) + 0.0258 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2024 \cdot z_{t-1}$$
$$+0.0297 \cdot z_{t-2} + 0.0494 \cdot z_{t-3}$$

P=2, Q=1 model (*DoF* = 10.9241):

$$r_t = 1.3889 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1581 + 0.9296 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) + 0.0536 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1197 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1406 \cdot z_{t-1}$$

P=2, Q=2 model (*DoF* = 11.6006):

$$r_t = 1.2493 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0196 + 1.7908 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7929 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.0694 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.0486 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2154 \cdot z_{t-1}$$
$$+0.1977 \cdot z_{t-2}$$

P=2, Q=3 model (*DoF* = 11.4773):

$$r_t = 1.3584 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0170 + 1.7942 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.7960 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0180 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.1519 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1156 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2114 \cdot z_{t-1} + 0.1903 \cdot z_{t-2} + 0.0043 \cdot z_{t-3}$$

P=3, Q=1 model (*DoF* = 11.0604):

$$r_t = 1.3829 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.1509 + 1.0724 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.2027 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$+0.1142 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.1148 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) - 0.1355 \cdot z_{t-1}$$

P=3, Q=2 model (*DoF* = 11.2925):

$$r_t = 1.2938 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0166 + 1.7127 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.6162 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.0982 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0789 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.0607 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right) - 0.2261 \cdot z_{t-1} + 0.2111 \cdot z_{t-2}$$

P=3, Q=3 model (*DoF* = 11.9261):

$$r_t = 1.3601 \cdot 10^{-4} + \sigma_t \cdot z_t$$
$$\ln\left(\sigma_t^{\,2}\right) = -0.0167 + 1.3887 \cdot \ln\left(\sigma_{t-1}^{\,2}\right) - 0.0003 \cdot \ln\left(\sigma_{t-2}^{\,2}\right)$$
$$-0.3902 \cdot \ln\left(\sigma_{t-3}^{\,2}\right) + 0.0219 \cdot \left(\left|z_{t-1}\right| - E\left(\left|z_t\right|\right)\right) + 0.1060 \cdot \left(\left|z_{t-2}\right| - E\left(\left|z_t\right|\right)\right)$$
$$-0.1083 \cdot \left(\left|z_{t-3}\right| - E\left(\left|z_t\right|\right)\right) - 0.2037 \cdot z_{t-1} + 0.0984 \cdot z_{t-2} + 0.0894 \cdot z_{t-3}$$

APPENDIX C – weight calculation function and resulting weight dispersion

Linear model averaging framework summarized by (4.2) demands a weight calculation function to determine the weights w_i, $i = 1,\ldots,M$ from a given model evaluation metric, which in this appendix we represent by a generic value x_i. In particular, this work used as such a metric the information criteria AIC and SIC. In this appendix only, consider f to be any given function to be used in the mapping from x_i to w_i, as stated in the next equation:

$$w_i = \frac{f\left(x_i\right)}{\sum_{j=1}^{M} f\left(x_j\right)} \tag{C.1}$$

In (C.1) we include the normalizing denominator to assure weights unitary sum, and thus f can be any function with the following two properties: 1) be non-negative in the domain in which the values of x_i lie, and 2) be a non-decreasing function. These properties ensure non-negativity of the weights and absence of smaller weights for better evaluated models.

The choice for f impacts the dispersion of the weights or, in other words, how much larger the weight of a better evaluated model is relatively to a worse evaluated one. As a metric to those concepts, it is defined below the relative dispersion D_f. For that matter, it depends on two given models indexed by *m1* and *m2*, for which the former is better evaluated than the latter ($x_{m1} > x_{m2}$).

$$D_f \triangleq \frac{w_{m1}}{w_{m2}} - 1 \tag{C.2}$$

Although D_f is obviously dependent on x_{m1} and x_{m2}, we omit those dependences and explicit only the dependence on f for notation concision, since the choice of the latter is the focus of the analysis carried out hereafter.

The lower bound for D_f is zero (since f is non-decreasing), which is reached when the better evaluated model attains the same weight as the worse evaluated one (this is the case for simple averaging). On the other hand, there is no upper bound, since the better model can attain an infinitely larger weight (than the worse model one), as occurs in model selection, when the best model (unitary weight) is compared to any other model (zero weight). The relative dispersion can be further depicted:

$$D_f = \frac{f(x_{m1})}{f(x_{m2})} - 1 = \frac{f(x_{m2}) + f'(x^*)(x_{m1} - x_{m2})}{f(x_{m2})} - 1 = \frac{f'(x^*)(x_{m1} - x_{m2})}{f(x_{m2})} \quad \text{(C.3)}$$

In (C.3), x^* is an intermediary point between x_{m2} and x_{m1}, for which the derivative of f attains its medium value over the $[x_{m2}, x_{m1}]$ interval. Geometrically, x^* in the point for which the line tangent to f at $f(x^*)$ is parallel to the secant line that crosses the points $(x_{m2}, f(x_{m2}))$ and $(x_{m1}, f(x_{m1}))$.

Regarding the choice of f, the last expression of (C.3) indicates the relevance of the function given by the ratio of the derivative f' to f, denoted by C_f in (C.4):

$$C_f(x) \triangleq \frac{f'(x)}{f(x)} \quad \text{(C.4)}$$

It is noticed, however, that C_f is not clearly present in (C.3), because f and f' are evaluated at different points therein. For that matter we define the following quantity B_f in equation (C.5):

$$B_f \triangleq \frac{f'(x^*)}{f'(x_{m2})} - 1 \quad \text{(C.5)}$$

The dependence of x_{m2} and x_{m1} on B_f is omitted for notation concision, as done for D_f. Combining (C.3) to (C.5):

$$D_f = C_f(x_{m2})(1 + B_f)(x_{m1} - x_{m2}) \quad \text{(C.6)}$$

From (C.6) the dependence of function f on the relative dispersion of the weights is due through C_f and B_f terms. The latter has a complicated relationship with the behavior of higher order derivatives of f, especially due to its dependence of x^*. However it is noticed that B_f and the second derivative of f have the same signs (provided that the second derivative f'' does not change sign in $[x_{m2}, x_{m1}]$ interval). This is because, since $x^* > x_{m2}$, a non-negative (non-positive) second derivative implies a non-decreasing (non-increasing) first derivative, and thus a non-negative (non-positive) B_f. Therefore, B_f represents a second-order effect of f on D_f, according to which the larger the convexity of f, the higher the relative dispersion of the weights (all else kept constant). However, the first order effect of f on D_f represented by the simple C_f function will be of higher practical relevance in the analysis that follows.

We next determine f, D_f and C_f for the model averaging schemes used in this work. For that matter, k will denote arbitrary (possibly zero) offset applied to the input x_i. For model selection:

$$f-sel(x)=\begin{cases}1, x=\max\limits_j x_j\\ 0, otherwise\end{cases} \tag{C.7}$$

$$D_{f-sel}=\begin{cases}\infty, x_{m1}=\max\limits_j x_j\\ 0, otherwise\end{cases} \tag{C.8}$$

$$C_{f-sel}(x)=\begin{cases}\infty, x=\max\limits_j x_j\\ 0, otherwise\end{cases} \tag{C.9}$$

For the weights calculated through the exponential function:

$$f-exp(x)=e^{x-k} \tag{C.10}$$

$$D_{f-exp}=e^{x_{m1}-x_{m2}}-1 \tag{C.11}$$

$$C_{f-exp}(x)=1 \tag{C.12}$$

For the weights calculated through the linear function:

$$f-lin(x)=x-k \tag{C.13}$$

$$D_{f-lin}=\frac{x_{m1}-k}{x_{m2}-k}-1=\frac{x_{m1}-x_{m2}}{x_{m2}-k} \tag{C.14}$$

$$C_{f-lin}(x)=\frac{1}{x-k} \tag{C.15}$$

For the weights calculated through the simple averaging:

$$f-simple(x)=1 \tag{C.16}$$

$$D_{f-simple}=0 \tag{C.17}$$

$$C_{f-simple}(x) = 0 \tag{C.18}$$

It should be clear that, from model selection to simple averaging, the averaging strategies above were presented from the highest (potentially infinite) to the lowest (zero) relative dispersions of weights. The same applies to C_f, which reinforces the practical relevance of the relationship of both quantities. For the exponential function, the offset k does not affect D_f nor C_f, the former dependent only on the relative differences of the evaluation metric (which are the only theoretical valuable quantities regarding the information criteria), and the latter a constant equal to one. More generally, the absolute magnitude of the evaluation metric x_i (changeable by the offset k) does not influence the weight dispersion. On the other hand, for the linear function, the denominators of the right hand sides of both (C.14) and (C.15) (D_f and C_f, respectively) show that the dispersion of the weights is inversely proportional to the absolute magnitude of the evaluation metric.

In the particular case of this work, the magnitudes of the information criteria were large enough to make the weight dispersion close to zero, and thus the strategy close to simple averaging (zero dispersion). The use of an offset equal to the smallest attained value (among the different order models) of the information criterion was responsible to increase the dispersion of the weights and make the strategy differ from simple averaging. However, since the differences between the values of the information criterion were generally higher than one, the absolute magnitude of the evaluation metric values remained higher than one even after the minimal information criterion value offset was applied, which corresponds to lower than unitary C_f (the constant value of C_f for the exponential). Thus, the higher than unitary differences among the values of the information criterion are the cause of the lower weight dispersion provided by the linear function, when compared to the exponential one. This statement was algebraically demonstrated in equation (4.23), for which these higher than unitary differences of the information criterion were explicitly assumed. Nonetheless, the still lower than exponential weight dispersion attained by the linear function was the maximum possible to be reached through an offset, since an offset k higher than the one used (minimal information criterion value) would clearly violate the restriction on f to be non-negative.

The model averaging literature suggests that simple averaging (and thus zero weight dispersion) is difficult to outperform with more complex strategies. On the other hand, model selection tends to be highly outperformed by model averaging. Thus, from a zero to infinite weight dispersion, the optimum point should generally be close to the former, and this could be the reason that in this work the linear function performed better than the exponential function regarding the weight calculation strategy, while the outperformance of the former over simple averaging was not so evident.

Lastly, we answer the question of the conditions needed by a function *f* to display weight dispersion invariant to offset, as it was shown to be the case of the exponential function. In particular, we answer if there is another function with this property, which is formally given below:

$$\frac{f(x)}{f(y)} = \frac{f(x+k)}{f(y+k)} \Rightarrow \frac{f(y+k)}{f(y)} = \frac{f(x+k)}{f(x)} = h(k) \tag{C.19}$$

Since for an invariant to offset function *f* the left hand side has to hold true for all *x* and *y*, the right hand side follows, which means that the ratio of the function evaluated at a given point with the offset to the function evaluated at the same point (whatever it is) without the offset cannot depend on this given point, only on the offset itself. Thus, we differentiate the last equality of (C.19) in respect to *x* and *k*, obtaining (C.20) and (C.21), respectively:

$$\frac{\partial}{\partial x} f(x+k) = h(k) \frac{\partial f(x)}{\partial x} \tag{C.20}$$

$$\frac{\partial}{\partial k} f(x+k) = f(x) \frac{\partial h(k)}{\partial k} \tag{C.21}$$

The left hand sides of both (C.20) and (C.21) are equal to the derivative of *f* evaluated at point $x + k$, and therefore we subtract those equations leading to:

$$h(k) \frac{\partial f(x)}{\partial x} - \frac{\partial h(k)}{\partial k} f(x) = 0 \tag{C.22}$$

For any given *k* and *h*, (C.22) is a differential equation with constant coefficients that allows one to determine its sole solution as:

$$f(x) = Ae^{\beta x} \tag{C.23}$$

where

$$\beta = \frac{\partial h(k)}{\partial k} \left(h(k) \right)^{-1} \tag{C.24}$$

Therefore, the exponential function is the only possible one invariant to offset, in the constant weight dispersion sense depicted here. This means that (C.23) is a necessary condition to (C.19) whereas sufficiency can be

easily be shown by inspection of the latter combined with the former and $h(k) = e^{\beta k}$.

It is noticed that (C.22) to (C.24) are also equivalent to a constant ratio of the derivative of f to the function f itself, which is the definition of C_f given by (C.4). Thus, the invariance to offset explored here is determined by a constant C_f, that constant being β and the corresponding function given by (C.23).