On the choice of scoring functions for forecast comparisons

by

Jonathan Agyeman

B.A. Economics and Statistics, University of Ghana, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Statistics)

The University of British Columbia

(Vancouver)

April 2017

Abstract

Forecasting of risk measures is an important part of risk management for financial institutions. Value-at-Risk and Expected Shortfall are two commonly used risk measures and accurately predicting these risk measures enables financial institutions to plan adequately for possible losses. Point forecasts from different methods can be compared using consistent scoring functions, provided the underlying functional to be forecasted is elicitable. It has been shown that the choice of a scoring function from the family of consistent scoring functions does not influence the ranking of forecasting methods as long as the underlying model is correctly specified and nested information sets are used. However, in practice, these conditions do not hold, which may lead to discrepancies in the ranking of methods under different scoring functions.

We investigate the choice of scoring functions in the face of model misspecification, parameter estimation error and nonnested information sets. We concentrate on the family of homogeneous consistent scoring functions for Value-at-Risk and the pair of Value-at-Risk and Expected Shortfall and identify conditions required for existence of the expectation of these scoring functions. We also assess the finite-sample properties of the Diebold-Mario Test, as well as examine how these scoring functions penalize for over-prediction and under-prediction with the aid of simulation studies.

Preface

This dissertation is solely authored, unpublished work by the author, Jonathan Agyeman. The research topic was suggested by the supervisor Prof. Natalia Nolde, and the supporting simulation studies were designed and carried out by the author. The issue of the choice of scoring functions in the face of model misspecification, parameter estimation error or nonnested information sets of forecasting procedures was raised by Patton.

Table of Contents

List of Tables

List of Figures

Glossary

DGP	data generating process
D-M	Diebold-Mariano
ES	Expected Shortfall
FHS	filtered historic simulation
FPE	fully parametric estimation
GPL	generalized piecewise linear
VaR	Value-at-Risk

Acknowledgments

Firstly, I want to thank the Almighty God for seeing me through my Masters program at UBC. Being at UBC has been a great experience for me. I had the opportunity to learn from the greatest minds in statistics and also interact with very intelligent students from all over the world.

My deepest gratitude to my supervisor, Prof. Natalia Nolde. Natalia has been a great mentor and working with her has taught me a lot. Her dedication towards her work is second to none and this is a quality I want to emulate as I begin my career in this field. Her invaluable advice and ability to present complicated issues in a simple way have been very helpful throughout my research work. I am also grateful to Prof. Jiahua Chen for being my second reader and his helpful suggestions concerning my thesis.

A big thank you to my mother Juliana and my siblings Anne-Mary, Yvonne, David, Emmanuel, and Joshua for their tremendous support and prayers. Thank you to my wonderful nephews Gidon and Winston, and my lovely nieces Annabel and Valerie. I also wish to thank Rev.Dr. Enerstina Afriyie, Dr. Isaac Baidoo and Mr. Albert Owusu for their support throughout my education.

I wish to thank my friends Rowena, Kwasi, Yeng, Kojo, Andy, Farouk, Nayorm, Elsie, Naki and Ann for their support and sacrifices made for me. I also want to thank my ever supportive office mates, Creagh, Derek, Sohrab and Yijun. Thanks to my dear friends Qiong, Yiwei, Harry and Julian who have been of great help all through my program. Special thanks to Ali, Andrea and Peggy for their patience and willingness to answer all my questions and clarify any misunderstanding I had here. Lastly, I want to say a big thank you to my mean square terror mates Danny, David, Joe and Pablo. God bless you all.

Chapter 1

Introduction

Point forecasting plays an important role in many fields. To make accurate predictions about an uncertain future, there is a need to make desicions on the appropriate forecasting procedure as well as measures for evaluating forecasts. Competing forecasting procedures can be compared using scoring functions. Scoring functions play an important role in the theory and practice of forecasting [Gneiting, 2011a]. Gneiting [2011a] argues that in the situation where point forecasts are to be issued and evaluated, it is important to either specify a scoring function for evaluation from which the functional to be forecasted can be deduced or make the statistical functional (e.g., the mean, quantile or expectile) to be forecasted known. In finance, point forecasting plays an important role in decision making. Financial institutions rely on conditional forecasts of risk measures for the purposes of internal risk management as well as regulatory capital calculations [Nolde and Ziegel, 2016]. Most modern risk measures for a portfolio are statistical quantities which describe the (conditional) loss distribution of the portfolio [McNeil et al., 2005]. Value-at-Risk (VaR) and Expected Shortfall (ES) are examples of such risk measures. The decision on the forecasting procedure to use in estimating a chosen risk measure can be made using scoring functions. The choice of scoring function is therefore an important aspect of forecasting in finance.

A scoring function is an error measure which assesses how close or far-off point forecasts are from the verifying observations. When comparing two competing forecasting procedures using a scoring function, the forecast cases obtained from each of the forecasting procedures and the corresponding verifying observations are used to obtain score values for each forecast case. These score values are averaged to obtain a single estimate for each forecasting procedure as shown in (1.1)

$$\bar{S}^j = \frac{1}{n}\sum_{i=1}^{n} S(\hat{y}_i^j, y_i), \quad i = 1, ..., n \;\; j = 1, 2 \tag{1.1}$$

where $\hat{y}_i^j$ is the ith point forecast from forecasting procedure j and y_i is the correspoding verifying obeservation. Smaller average scores correspond to better forecasting procedures. Commonly used scoring functions for the mean and median forecasts include the squared error and the absolute error, respectively. A functional is a potentially set-valued mapping $T(F)$ from a class of probability distributions, $\mathscr{F}$, to the real line, $\mathbb{R}$. Examples include the mean and quantile. Gneiting [2011a] argues that it is important for a scoring function S$(\hat{y}, y)$, where $\hat{y}$ is the point forecast and y is the observed value, to be consistent for the functional T relative to the class $\mathscr{F}$. This means that the expectation of the scoring function with respect to the distribution F is minimized at a given functional compared to all other forecasts.

Statistical functionals for which a consistent scoring function exists are known as elicitable functionals. We discuss the concept of consistency and elicitability in Section 2.1. Patton argues that evaluating forecasts of a given functional using consistent scoring functions is a minimal requirement for sensible rankings of competing forecasts.

It is shown in Thomson [1979] and Saerens [2000] that quantiles are elicitable functionals. Hence, VAR is an elicitable functional since it is simply the quantile of a given distribution in probabilistic terms. On the other hand, ES is not elicitable. However, it is jointly elicitable with VAR; see [Acerbi and Szekely, 2014] and [Fissler and Ziegel, 2016]. This implies that there are scoring functions that are consistent for VAR and the pair (VAR, ES). Detailed review on these elicitable risk measures and their respective consistent scoring functions is presented in Section 2.2 and Section 2.3.

For a given elicitable functional, there are infinitely many consistent scoring functions that can be used in the evaluation of forecasts. Patton assesses the performance of consistent scoring functions for given functionals in the presence of misspecified models, parameter estimation error, or nonnested information sets. He concludes that the choice of the scoring function for ranking forecasts does not matter provided that models are correctly specified and information sets of competing forecasting procedures are nested. However, in practice, the model used for forecasting is uncertain, the parameters need to be estimated, and forecasting procedures do not have much information about competing forecast models. He argues that scoring functions are sensitive to model misspecification, parameter estimation error, or nonnested information sets and the ranking of forecasts may be inconsistent when different scoring functions are used in evaluating forecasts. Gneiting [2011a] also presents a simulation study where he shows that the choice of an arbitrary scoring function in the assessment of forecasts could produce misleading results. Due to this, there is the need to establish criteria for desirable scoring functions within the class of consistent scoring functions.

This research work seeks to identify the criteria for choosing desirable consistent scoring functions for the evaluation of VAR and the pair (VAR, ES) forecasts. We seek to identify the class of consistent scoring functions that satisfy the chosen criteria and may lead to accurate ranking of forecasting procedures in practice. The report is organized as follows: Chapter 2 presents key definitions and theoretical discussion of consistent scoring functions, elicitable functionals, risk measures VAR and ES, and the family of consistent scoring functions for VAR and the pair (VAR, ES). Chapter 3 presents an overview of the criteria to consider in choosing a consistent scoring function for evaluating the risk measures and introduces the class of homogeneous scoring functions for VAR and (VAR, ES). In Chapter 4, we look at the tail behaviour of GARCH processes, examining the conditions under which the expected value of chosen scoring function with respect to the random variable Y with distribution F will exist. In Chapter 5, we present simulation studies where the finite-sample size and power properties of the Diebold-Mariano test [Diebold and Mariano, 1995] are assessed for consistent scoring functions for VAR and the pair (VAR, ES). Chapter 6 assesses how consistent scoring functions penalize for under-prediction and over-prediction of the same magnitude with the aid of simulation studies. Finally, conclusions for our work and possible further studies are presented in Chapter 7.

Chapter 2

Preliminaries

A key component of point forecasting is the means by which competing forecasts are evaluated. Competing point forecasting procedures are typically compared based on a scoring function which is averaged over forecast cases. A scoring function, S, is any mapping $S : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$, where $S(\hat{y}, y)$ represents the score when a point forecast $\hat{y} \in \mathbb{R}^k$ is issued and the observation $y \in \mathbb{R}$ is realized. Decisions concerning the choice of scoring functions when evaluating competing forecasts is important in point forecasting since an arbitrary choice of the scoring function for the evaluation of forecasts may give misleading results.

2.1 Consistency and elicitability

Following Gneiting [2011a], a functional is a potentially set-valued mapping T(F) from a class of probability distributions, $\mathscr{F}$, to the real line $\mathbb{R}$. It is important that for the given functional relative to the class $\mathscr{F}$, the scoring function chosen for evaluation of forecasts be consistent. A scoring function is consistent for a given functional if the expected score is minimized under the true distribution compared to all other forecasts.

Definition 2.1.1. (*Gneiting [2011a]*) Let $\mathscr{F}$ a family of probability measures on $\mathbb{R}$, $\mathrm{T} : \mathscr{F} \to \mathbb{R}$ a functional, and $S : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ a scoring function. The scoring function S is *consistent* for the functional T relative to the class $\mathscr{F}$, if $\mathbb{E}_F S(\hat{y}, Y)$

exists and is finite for all $F \in \mathscr{F}$ and a given point forecast $\hat{y} \in \mathbb{R}$, and if

$$\mathbb{E}_F S(t,Y) \leq \mathbb{E}_F S(\hat{y},Y) \tag{2.1}$$

for all $F \in \mathscr{F}$, all $t \in T(F)$, and all point forecasts $\hat{y} \in \mathbb{R}$. It is *strictly consistent* if it is consistent and equality of expectations implies that $\hat{y} \in T(F)$.

This means that, for a given scoring function S, and a predictive distribution F, the functional $T(F)$ is the optimal point predictor if

$$T(F) = \text{argmin}_{\hat{y}} \mathbb{E}_F[S(\hat{y},Y)].$$

An example of a consistent scoring function for the mean functional is the squared error given by $S(\hat{y},y) = (\hat{y}-y)^2$. Even though it is desirable for a given functional to have a consistent scoring function for forecast evaluation. However, not all functionals have consistent scoring functions. One example is the variance. A functional for which a consistent scoring function exists is called an elicitable functional [Lambert et al., 2008]. Examples of elicitable functionals are the mean, quantile, and expectile functionals. A formal definition for an elicitable functional is given below.

Definition 2.1.2. (*Lambert et al. [2008]*) A functional T is elicitable relative to the class $\mathscr{F}$ if there exists a scoring function S that is strictly consistent for T relative to $\mathscr{F}$.

Misleading results about the quality of forecasts may be obtained when the scoring function used in the evaluation of point forecasts is not consistent for the given functional under the true distribution. Nolde and Ziegel [2016] illustrate with a simulation study how the methods used in forecasting VAR and the pair (VAR, ES) have reasonable rankings when consistent scoring functions for these functionals are used in evaluating forecasts. In particular, the optimal forecast with knowledge of the data generating process is ranked as “best” among competing forecasting procedures. Ziegel [2016] describes elicitable functionals as functionals for which meaningful point forecasts and forecast performance comparisons are possible. This report concentrates on forecasting of two risk measures, the VAR and the pair (VAR, ES), which are introduced next.

2.2 Value-at-Risk and Expected Shortfall

The use of VAR as a risk measure in risk management is very common. For a given confidence level $\alpha \in (0,1)$, VAR is the value x such that, the probability of observing a value greater than x is no larger than 1-α. Typical α values in risk management are α = 0.95 (internal risk management) and α = 0.99 (regulatory level). A formal definition for VAR is given below.

Definition 2.2.1. For a random variable X, VAR at confidence level $\alpha \in (0,1)$ is defined as:

$$\text{VAR}_{\alpha}(X) = inf\{x \in \mathbb{R} : P(X > x) \leq 1 - \alpha\}. \tag{2.2}$$

From the statistical perspective, VAR$_{\alpha}$ is simply an α-quantile of a given loss (or profit) distribution, provided the α-quantile is single-valued. As mentioned earlier, the α-quantile is an elicitable functional, and the family of consistent scoring functions for the α-quantile consists of generalized piecewise linear functions of order $\alpha \in (0,1)$. VAR as a risk measure has been criticized for structural drawbacks by a number of researchers over the years. ES, a coherent risk measure, was proposed by Artzner et al. [1999] as an alternative to VAR and there has been a growing consensus on the use of ES as a risk measure over the VAR; [Kusuoka, 2001] and [Acerbi and Tasche, 2002]. The Bank for International Settlements [2013] has been looking into the arguments for and against the change of the regulatory risk measure from VAR to ES. There is a close relation between ES and VAR and the two have been found to be jointly elicitable. This means that for there is a consistent scoring function for the pair (VAR, ES). The formal definition for ES is given below.

Definition 2.2.2. For a random variable X, ES at confidence levels $\nu \in (0,1)$ is defined as:

$$\text{ES}_{\nu}(X) = \frac{1}{1-\nu} \int_{\nu}^{1} \text{VAR}_{\alpha}(X) \mathrm{d}\alpha. \tag{2.3}$$

When X is a continuous random variable, ES can be written as

$$\mathrm{ES}_\nu(X) = \mathbb{E}(X|X > \mathrm{VAR}_\nu(X)). \tag{2.4}$$

As seen from the definition, the ES looks at the entire tail of the loss (or profit) distribution and averages the VAR over all levels of $\alpha \geq \nu$. Generally, $\mathrm{ES}_\alpha \geq \mathrm{VAR}_\alpha$ as illustrated in Figure 2.1. For α = 0.95, a VAR_α value of 3.41 is obtained while the ES_α is 4.77.

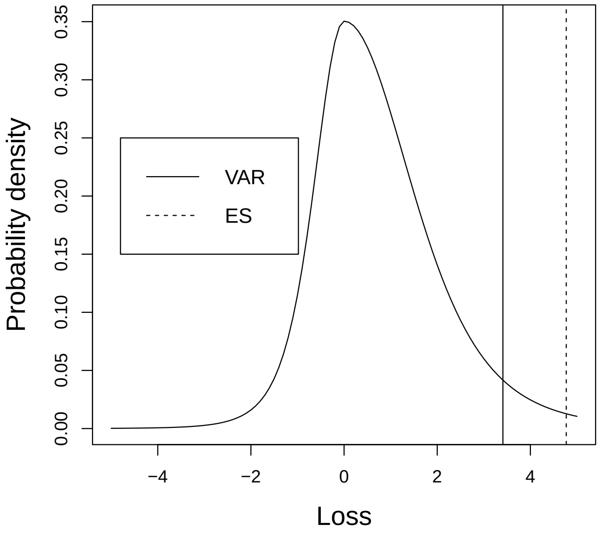


Figure 2.1: An example of a loss distribution with the 95% VAR marked as a vertical line and 95% ES shown with a dotted line.

2.3 Scoring functions for VAR and (VAR, ES)

There are many possibilities from which to choose a consistent scoring function for an elicitable functional. Gneiting [2011b] showcases a number of consistent scoring functions for VAR that arises in different fields where different weights are assigned in penalizing over-prediction and under-prediction in the evaluation of forecasts. Nolde and Ziegel [2016] use the homogeneous consistent scoring functions for the pair (VAR, ES) for comparative backtesting of the risk measures. There exists an entire family of consistent scoring functions for VAR and the pair (VAR, ES). We next review the characterization of the family of consistent scoring functions for VAR and the pair (VAR, ES).

Theorem 2.3.1. (*Thomson [1979]: Saerens [2000]*) Up to equivalence and mild regularity conditions, a consistent scoring function for VAR$_\alpha$ (α-quantile), $\alpha \in (0,1)$ relative to the class of compactly supported probability measures on I $\subseteq \mathbb{R}$ is given by

$$S(\hat{y},y) = (\mathbb{1}(\hat{y} \geq y) - \alpha)(G(\hat{y}) - G(y)), \tag{2.5}$$

where G is a non-decreasing function on I, $\hat{y}$ is the point forecast and y is the observed value. Scoring functions in (2.5) form the family of generalized piecewise linear (GPL) functions.

Theorem 2.3.2. (*Fissler and Ziegel [2016]*) Up to equivalence and mild regularity conditions, all consistent scoring functions for (VAR$_\nu$, ES$_\nu$) , $\nu \in (0,1)$ relative to the class of probability measures on I $\subseteq \mathbb{R}$ are of the form

$$\begin{aligned} S(\hat{y}_1,\hat{y}_2,y) = (\mathbb{1}\{\hat{y}_1 \geq y\} - \nu)(G_1(\hat{y}_1) - G_1(y)) + \frac{1}{\nu} G_2(\hat{y}_2)\mathbb{1}\{\hat{y}_1 \geq y\}(\hat{y}_1 - y) \\ + G_2(\hat{y}_2)(\hat{y}_2 - \hat{y}_1) - \mathscr{G}_2(\hat{y}_2), \end{aligned} \tag{2.6}$$

where $\hat{y}_1$ is the VAR forecast and $\hat{y}_2$ is the ES forecast with G_1 and G_2 being strictly increasing continuously differentiable functions such that the expectation $\mathbb{E}[G_1(X)]$ exists, $\lim_{x\to-\infty} G_2(x) = 0$ and $\mathscr{G}_2^{'} = G_2$; (see Fissler and Ziegel [2016], corollary 5.5).

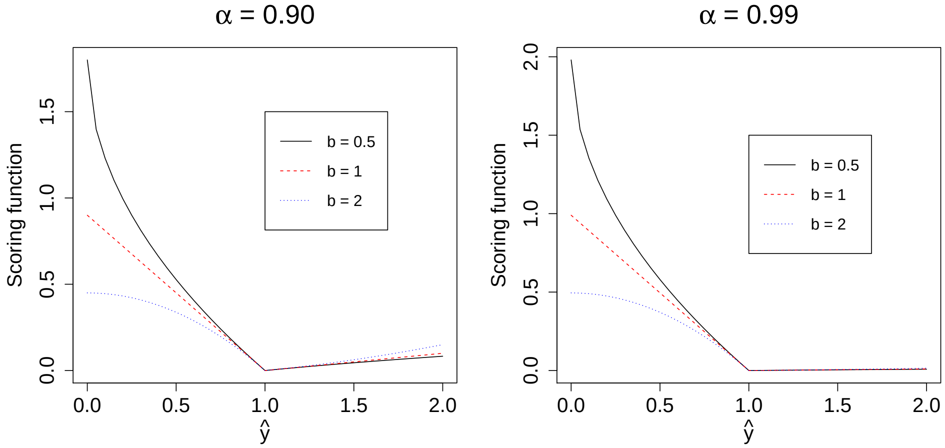
From Theorem 2.3.1 and Theorem 2.3.2 above, it is seen that the choice of the consistent scoring functions for the VAR and the pair (VAR, ES) depends on the choice G in (2.5) and G_1 and G_2 in (2.6), respectively. A well known example of a GPL scoring function is the *Lin-Lin* (or *tick*) scoring function which is obtained when G in (2.5) is the identity function:

$$S(\hat{y},y) = (\mathbb{1}(\hat{y} \geq y) - \alpha)(\hat{y} - y). \tag{2.7}$$

Patton defines a special case of GPL scoring functions which is obtained when $G(y) = sgn(y)\frac{|y|^b}{b}$ for b > 0. The scoring function is then given by

$$S(\hat{y}, y) = (\mathbb{1}\{\hat{y} \geq y\} - \alpha)(sgn(\hat{y})|\hat{y}|^b - sgn(y)|y|^b)/b \quad , b > 0. \tag{2.8}$$

This family of GPL scoring functions is a homogeneous family of scoring functions. The *Lin-Lin* score function arises when $b = 1$. Figure 2.2 presents the plot of $S(\hat{y}, y)$ for a range of values of $\hat{y}$ for the homogeneous GPL scoring functions obtained for varying values of b and selected values of α. We examine the behaviour of the scoring functions for $\alpha = 0.9$ and $\alpha = 0.99$ as we are concerned with the tails of loss distributions. It can be seen that the shape of the different homogeneous GPL differs even though they all assign the same score at the optimal value for a given α level.



(a) **(b)**

Figure 2.2: Various homogeneous GPL scoring function $S(\hat{y}, y)$ in (2.8), with $\alpha = 0.90$ (a) and $\alpha = 0.99$ (b). The *Lin-Lin* scoring function is obtained when b = 1. The value of y is set at 1.

For the case where $b = 0$, Nolde and Ziegel [2016] present the 0-homogeneous score differences for VAR which is given as

$$S(\hat{y}, y) = (1 - \alpha - \mathbb{1}\{y > \hat{y}\})\log\hat{y} + \mathbb{1}\{y > \hat{y}\}\log y \quad , \hat{y} > 0. \tag{2.9}$$

Another example of a GPL scoring function is the exponential GPL scoring function which is obtained when $G(y) = \exp(y)$.

For consistent scoring functions for the pair (VAR, ES), Nolde and Ziegel [2016] present two homogeneous scoring functions for score differences. The (1/2)-homogeneous scoring function is obtained when $G_1(y) = 0$ and $G_2(y) = \frac{1}{2\sqrt{y}}$, $y > 0$ and is given by

$$S(\hat{y}_1, \hat{y}_2, y) = \big(\frac{1}{\nu}\big)\frac{1}{2\sqrt{\hat{y}_2}}\mathbb{1}\{\hat{y}_1 \geq y\} + \frac{1}{2\sqrt{\hat{y}_2}}(\hat{y}_2 - \hat{y}_1) - \sqrt{\hat{y}_2} \tag{2.10}$$

where $\hat{y}_1$ is the VAR$_\nu$ forecast, $\hat{y}_2$ is the ES$_\nu$ forecast and y is the observed value.

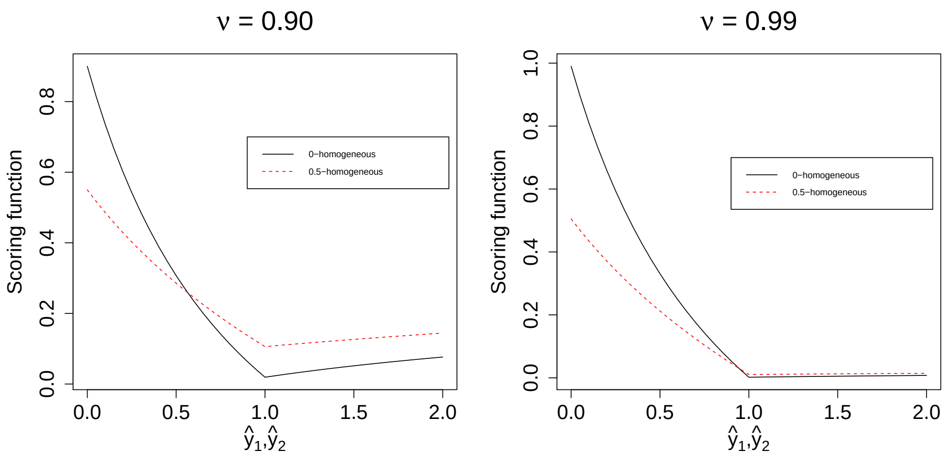
The 0-homogeneous scoring function is obtained when $G_1(y) = 0$ and $G_2(y) = \frac{1}{y}$, $y > 0$ and is given by

$$S(\hat{y}_1, \hat{y}_2, y) = \big(\frac{1}{\nu}\big)\frac{1}{\hat{y}_2}\mathbb{1}\{\hat{y}_1 \geq y\} + \frac{1}{\hat{y}_2}(\hat{y}_2 - \hat{y}_1) - \log(\hat{y}_2) \tag{2.11}$$

Figure 2.3 displays the behaviour of the two homogeneous scoring functions for a range of values of $\hat{y}_1$ and $\hat{y}_2$ when $\nu = 0.90$ and $\nu = 0.99$. It is seen that both scoring functions assign the lowest score at the optimal forecast. Other examples of consistent scoring functions for the pair (VAR, ES) presented by Fissler et al. [2015] include the case where $G_1(y) = y$ and $G_2(y) = \exp(y)$ and the case where $G_1(y) = y$ and $G_2(y) = \frac{\exp(y)}{1+\exp(y)}$. It can be seen from Figure 2.2 and Figure 2.3 that the scoring functions for evaluating these risk measures have varying shapes with different scores assigned to under-prediction and over-prediction of the same magnitude. We assess the behaviour of the consistent scoring function when penalizing for under-prediction and over-prediction in chapter 6 of this report.

Ehm et al. [2016] present a Choquet type mixture representation of GPL scoring functions. This mixture representation is a reduction of the infinite number of GPL scoring functions to a one-dimensional family of elementary scoring functions, in the sense that every GPL scoring function admits a representation as a mixture of elementary elements.

Theorem 2.3.3. (*Ehm et al. [2016]*) Any member of the class of GPL scoring



(a) **(b)**

Figure 2.3: Homogeneous scoring functions $S(\hat{y}_1, \hat{y}_2, y)$ in (2.10) and (2.11) for the pair (VAR_ν, ES_ν), with $\nu = 0.90$ (a) and $\nu = 0.99$ (b). The value of y is set at 1.

functions (S^Q_α) admits a representation of the form

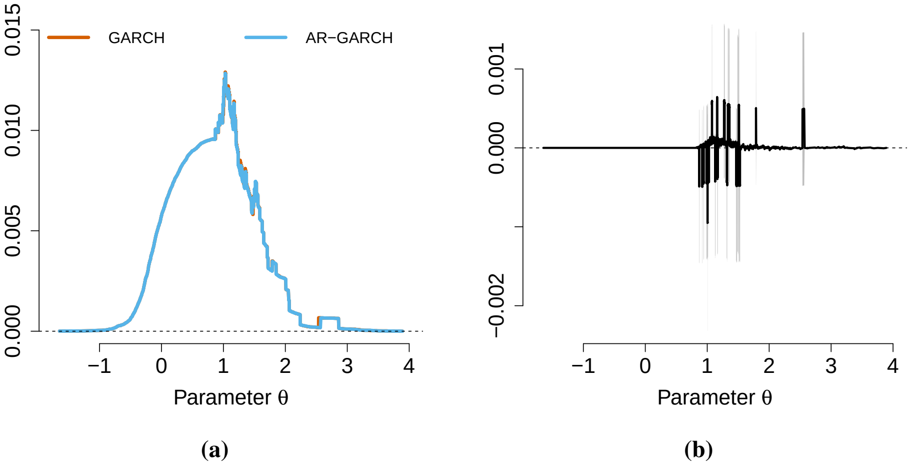
$$S(\hat{y}, y) = \int_{-\infty}^{+\infty} S^Q_{\alpha,\theta}(\hat{y}, y) \, \mathrm{d}H(\theta) \quad (\hat{y}, y \in \mathbb{R}), \tag{2.12}$$

where H is a nonnegative measure and

$$\begin{aligned} S^Q_{\alpha,\theta}(\hat{y}, y) &= (\mathbb{1}(y < \hat{y}) - \alpha)\,(\mathbb{1}(\theta < \hat{y}) - \mathbb{1}(\theta < y)) \\ &= \begin{cases} 1-\alpha, & y \le \theta < \hat{y}, \\ \alpha, & \hat{y} \le \theta < y, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.13}$$

The mixing measure H is unique and satisfies $\mathrm{d}H(\theta) = \mathrm{d}G(\theta)$ for $\theta \in \mathbb{R}$, where G is the non-decreasing function in the representation (2.5).

For the mixture representations, plots of the average scores obtained based on $S^Q_{\alpha,\theta}(\hat{y}, y)$ are used in comparing competing forecasts. These plots are called Murphy diagrams. Examples of Murphy diagrams are illustrated in Figure 2.4 and Figure 2.5. The Murphy diagrams display the plot of the scores for two competing forecasting procedures obtained from the elementary scoring function in (2.13) against the parameter θ and the plot of the score differences against the parameter θ. When there is no difference in the performance of competing forecasting procedures, the plot of the scores overlap and the score differences are close to zero (Figure 2.4). However, when the first listed forecasting procedure performs better than the second, there is a distinction in the plot of the scores over a range of θ values and the score differences are negative (Figure 2.5).



(a)

(b)

Figure 2.4: Murphy diagrams for simulation study comparing forecasts for 0.99-quantile. Data generating process is a GARCH (1, 1) model with skew-t innovations. The first forecasting procedure uses a GARCH (1, 1) model and the second forecasting procedure uses an AR (1)-GARCH (1, 1) model. (a) the plot of the scores from the elementary scoring function against the θ. (b) the plot of the scores differences against θ. No difference in the performance of forecasting procedures.

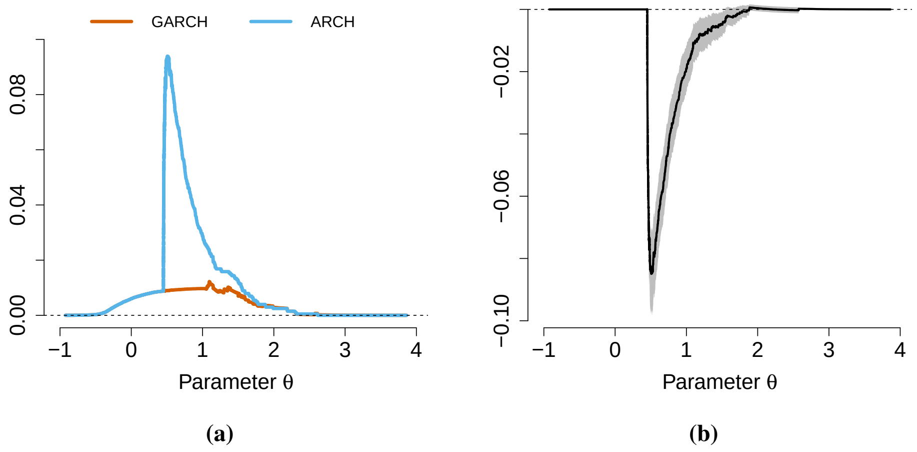


Figure 2.5: Murphy diagrams for simulation study comparing forecasts for 0.99-quantile. Data generating process is a GARCH (1, 1) model with skew-t innovations. The first forecasting procedure uses a GARCH (1, 1) model with information on the conditional variance and the second forecasting procedure uses an ARCH (1) model with no information on the conditional variance. (a) the plot of the scores from the elementary scoring function against the θ. (b) the plot of the scores differences against θ. Forecasting procedure with GARCH model performs better than forecasting procedure with ARCH model.

Chapter 3

Choice of scoring functions

The choice of a consistent scoring function for evaluating VAR and pair (VAR, ES) forecasts is a challenging task due to issues that arise in forecasting. Holzmann and Eulert [2014] show that increasing the information sets, which is the amount of information forecasting procedures have access to in making forecasts, leads to better point forecasts and smaller average scores when strictly consistent scoring functions are used in evaluating the forecasts provided the underlying model is correctly specified. Patton examines the case of forecasting procedures using correctly specified models and competing forecasting procedures with nested information sets compared to forecasting procedures with nonnested information sets or using misspecified models. He argues that the choice of consistent scoring functions for a given functional is of no concern when models are correctly specified, and if information sets are nested. However, when the models are misspecified or forecasting procedures' information sets are nonnested, care should be taken in choosing consistent scoring functions for evaluating forecasts. He gives the following propositions for VAR forecasts.

Proposition 1. *(Patton) Assume that (i) The information sets of two forecasting procedures are nested, so $\mathscr{F}_t^B \subseteq \mathscr{F}_t^A$ or $\mathscr{F}_t^A \subseteq \mathscr{F}_t^B$, and (ii) Forecasts A and B are optimal under some GPL scoring function. Then the ranking of these forecasts by expected Lin-Lin scoring function is sufficient for their ranking by any GPL scoring function.*

Proposition 2. *Assume that (a) The information sets of two forecasting procedures are nonnested, so $\mathscr{F}_t^B \nsubseteq \mathscr{F}_t^A$ and $\mathscr{F}_t^A \nsubseteq \mathscr{F}_t^B$, for some t, but Forecasts A and B are optimal under some GPL scoring function, or (b) one or both of the α-quantile forecasts are based on misspecified models. Then the ranking of these forecasts is, in general, sensitive to the choice of GPL scoring function.*

He presents a realistic set-up where he compares the ranking of forecasts by two forecasting procedures. Forecasting procedure A has knowledge of only the conditional mean of the time series process while forecasting procedure B has knowledge of only the conditional variance of the process. In assessing the performance of quantile forecasts from the two forecasting procedures using a homogeneous GPL scoring function, inconsistent results are obtained with forecasting procedure B performing better than forecasting procedure A for quantiles close to the lower tail while for quantiles between the lower tail and the center, forecasting procedure A performs better than forecasting procedure B.

Inconsistency in the results when ranking forecasts with different consistent scoring functions poses a possible problem in forecasting. Since the issue of parameter estimation error, model misspecification and nonnested information sets are often encountered in practice, it is desirable to identify which of the consistent scoring functions are better at identifying forecasting procedures that give more accurate forecasts. That is, since two competing forecasting procedures can be compared based on the average score obtained from the scoring functions as shown in (1.1), we seek to identify consistent scoring functions that may lead to making the right decision on forecasting procedures. From Theorem 2.3.1 and Theorem 2.3.2, we realize that there is a large number of consistent scoring functions that can be used in assessing VAR and pair (VAR, ES) forecasts. The choice of a consistent scoring function depends on the choice of G in (2.5) and G_1 and G_2 in (2.6). We begin by identifying some criteria for the selection of consistent scoring functions to assess forecasts of the risk measures.

A desirable property of scoring functions is the property of homogeneity. Patton [2011] argues that for homogeneous scoring functions, the ranking of forecasts is invariant to re-scaling of data and this is very useful in economic and financial applications where the choice of units of measurements is arbitrary (e.g., measuring in US dollars versus Canadian dollars).

A scoring function is positive homogeneous of order b if

$$S(c\hat{y},cy) = c^b S(\hat{y},y) \text{ for all } \hat{y} = (\hat{y}_1,...,\hat{y}_k), \text{ all } y \text{ and for } c > 0.$$

We concentrate on the homogeneous scoring functions for VAR and pair (VAR, ES) presented in (2.8), (2.9), (2.10), and (2.11) and identify the scoring functions that yield more accurate ranking of forecasts based on the criteria presented in following paragraphs.

The use of consistent scoring functions for a given functional is a minimal requirement in forecast evaluation, and hence one criterion for a desirable scoring function is that the expected value of the scoring function with respect to the distribution of random variable Y (the observed value) should exist under the weakest conditions. In finance, the GARCH model is commonly used in modeling log-returns data and the stationary solutions of GARCH processes are known to have heavy tails which may lead to restrictions on the choice of scoring functions as the $\mathbb{E}_F S(\hat{y},Y)$ may not exist. Following Sun and Zhou [2014], we examine the tail behaviour of a GARCH (1, 1) model and what implications it has on the range of values of b, the homogeneity order.

Next, Patton mentions that the use of multiple scoring functions in evaluating forecasts can lead to clouded results since a forecaster may be the best under one scoring function and the worst under the other. He states that consistent scoring functions have different sampling properties and a careful choice of the scoring function to use in forecast evaluation may result in improved efficiency. We assess the finite-sample properties of the Diebold-Mariano (D-M) test which is used to test the significance of the score difference for two forecasting procedures. The D-M test makes use of score values obtained from scoring functions. We assess the performance of the different homogeneous consistent scoring functions for VAR and the pair (VAR, ES) when used for the D-M test.

Lastly, the ability of scoring functions to penalize more for under-prediction than over-prediction is a desirable property in financial applications. To enable financial institutions make informed decisions in risk management, risk measures are estimated and the predicted amount is put aside to cater for any possible future losses. Forecasts may over-predict or under-predict the possible loss and this may have some negative effect on the operations of financial institutions. The magnitude of under-prediction could lead to institutions making big losses. We examine the homogeneous scoring functions to identify the scoring functions that penalize more for under-prediction than for over-prediction of VAR and ES. We expand on the above listed criteria in the following chapters.

Chapter 4

Scoring functions and tail behaviour

4.1 Expectation of homogeneous scoring function

The choice of a consistent scoring function from the family of consistent scoring functions for VAR and pair (VAR, ES) respectively, depends on the choice of G for the consistent scoring functions for VAR and the choice of G_1 and G_2 for the consistent scoring functions for the pair (VAR, ES). The $\mathbb{E}_F G(Y)$, for a chosen G, should exist and be finite for the expectation of the scoring function to exist and hence be consistent. The homogeneous scoring functions with homogeneity order b in (2.8) is consistent for VAR$_\alpha$ if $\mathbb{E}_F(|Y|^b)$ exists and is finite.

The use of GARCH processes to model returns is very common in finance. We seek to identify the conditions under which the expected value of a return Y, modeled as a GARCH process, will exist. This section looks at the tail behaviour of GARCH processes. In particular, we concentrate on the range of values for the parameters of a GARCH (1, 1) model for which the expectation of a homogeneous GPL scoring function exists.

4.1.1 GARCH Processes in QRM

When looking at VAR and ES in quantitative risk management, we concentrate on the tail behaviour of the loss distribution. Various models have been proposed when describing common features of financial returns. Amongst these are the models of the GARCH and ARCH family. One popular model used is the GARCH (1, 1) model. The stationary GARCH (1, 1) model is believed to capture various empirical observed properties of financial returns despite its simplicity [Mikosch and Starcia, 2000].

The marginal distributions of GARCH models are known to have heavy tails. This may lead to restrictions on the choice of a consistent scoring function as the $\mathbb{E}[S(\hat{y}_t, Y_t)]$ may not exist. We next look at the tail behaviour of a GARCH (1, 1) model presented by [Sun and Zhou, 2014] and what implications it has on the range of values of b, the homogeneity order.

Consider a GARCH (1, 1) model Y_t,

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned} \tag{4.1}$$

where $\{\varepsilon_t\}$ are independent and identically distributed (i.i.d) innovations with zero mean and unit variance, and the parameters α_0, α_1 and β_1 are non-negative. Suppose κ is the non-zero solution to the equation [1]

$$\mathbb{E}[(\alpha_1 \varepsilon_{t-1}^2 + \beta_1)^\kappa] = 1. \tag{4.2}$$

The stationary solution of $\{Y_t\}$ follows a heavy-tailed distribution with tail index 2κ:

$$P(|Y_t| > x) \sim C x^{-2\kappa}, \text{as} \quad x \to \infty. \tag{4.3}$$

From (4.3) it follows that $\mathbb{E}(|Y_t|^b) < \infty$ as long as $b < 2\kappa$. This leads to the following proposition.

[1] the solution to (4.2) exists provided that $\alpha_1 + \beta_1 < 1$ and $E[\varepsilon_t^{2\kappa_0}] = +\infty$ for $\kappa_0 := sup\{m : E[\varepsilon_t^{2m}] < +\infty\}$ (see Davis and Mikosh, 2009).

Proposition 3. *For a GARCH (1, 1) model in (4.1), the homogeneous scoring function with homogeneity order b in (2.8) is strictly consistent for VAR_α provided $b < 2\kappa$, where κ is the solution to (4.2).*

For the pair (VAR, ES), the expected value for the 0-homogeneous scoring function and the (1/2)-homogeneous scoring functions for the score differences is finite and exists for given values of $\hat{y}_1$ and $\hat{y}_2$ and information set up to time t-1 as long as $\kappa > \frac{1}{2}$.

We present in Figure 4.1 the plot of tail index (2κ) values against selected parameter values of α_1 and β_1 for the GARCH (1, 1) model to show the upper bound of the homogeneity order b for which the expected value of the scoring function will exist. We consider the case of the standard normal innovations, t-distributed and skewed t innovations for 3, 4, 5, 6, 7, and 8 degrees of freedom. With all tail indices falling above 2 for the different parameter value combinations, the different distributions of the innovations, and the different degrees of freedom for the t-distribution and skewed t-distribution, it indicates that the homogeneous scoring function is consistent for VAR in most cases where the homogeneity order ranges from 0 to 2 and a GARCH (1, 1) model with $\alpha_1 + \beta_1 < 1$ is used to model financial returns. More care has to be taken in choosing homogeneous scoring functions with $b > 2$ since the moments might not exist.

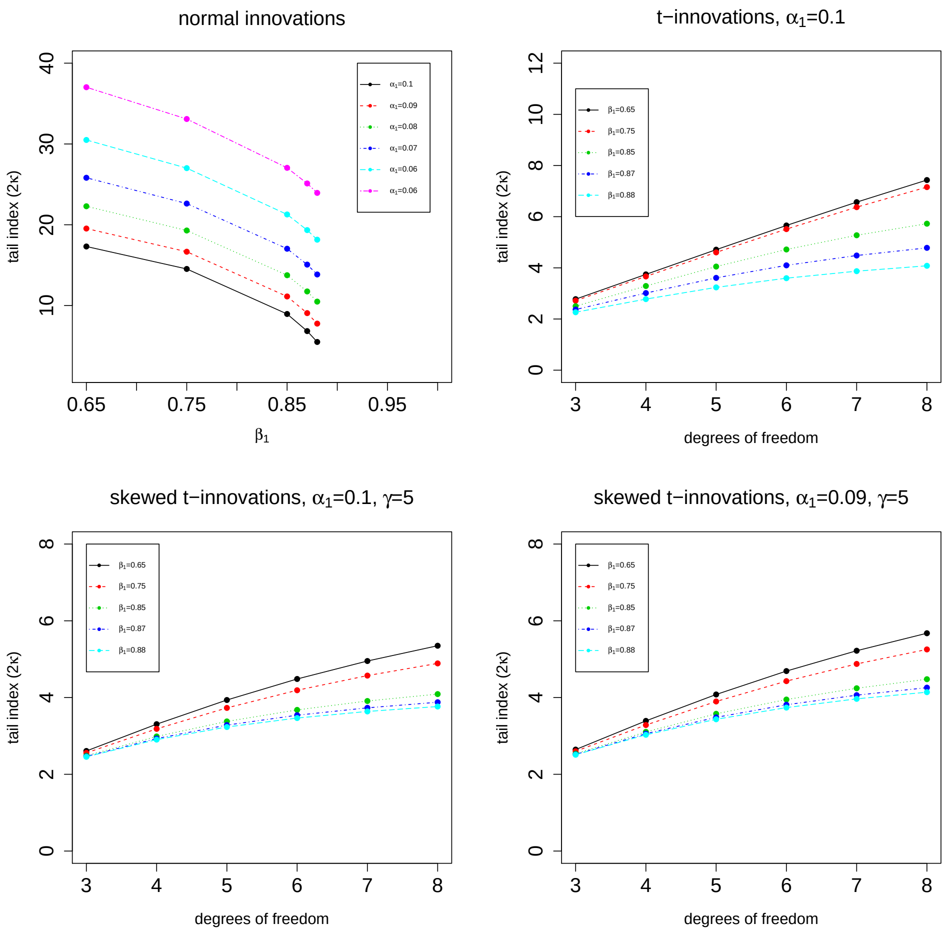


Figure 4.1: Upper bounds of the homogeneity order b for the expectation of the b-homogeneous scoring functions for VAR_α to exist.

Chapter 5

Finite-Sample Properties of the Diebold-Mariano test

In this chapter, we assess the finite-sample size and power properties of the Diebold-Mariano (D-M) test when the homogeneous consistent scoring functions are used in assessing the VAR and the pair (VAR, ES) forecasts from misspecified models, forecasting procedures with nonnested information sets, and models with parameter estimation errors.

For the simulation study, data is generated from an ARMA-GARCH process with different model specifications. We consider the three cases: parameter estimation error, wrongly specified models, and forecasting procedures with nonnested information sets in forecasting the risk measures. Table 5.1 and Table 5.2 summarize the models for the data generating process (DGP), the model used by the first forecaster in making predictions and the model used by the second forecaster in making predictions for the various cases when looking at the size and power properties of the D-M test for the homogeneous consistent scoring functions for VAR and the pair (VAR, ES).

5.1 Diebold-Mariano test

When predictions are made by competing forecasting procedures, differences are observed in these predictions and based on the observed value of the underlying

	DGP	Forecasting procedure A	Forecasting procedure B
Parameter estimation error			
Case 1	GARCH (1, 1),	GARCH (1, 1),	AR (1)-GARCH (1, 1),
(FPE & FHS)	$\varepsilon \sim Skew\,t(0,1,5,1.5)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$
Case 2	GARCH (1, 1),	GARCH (1, 1),	AR (1)-GARCH (1, 1),
(FPE)	$\varepsilon \sim t(0,1,\nu=5)$	$\varepsilon \sim t(0,1,\nu)$	$\varepsilon \sim t(0,1,\nu)$
Model misspecification			
Case 1	GARCH (1, 1),	GARCH (1, 1),	AR (1)-GARCH (1, 1),
(FPE)	$\varepsilon \sim Skew\,t(0,1,5,1.5)$	$\varepsilon \sim t(0,1,\nu)$	$\varepsilon \sim t(0,1,\nu)$
Case 2	GARCH (1, 1),	GARCH (1, 1),	AR (1)-GARCH (1, 1),
(FPE)	$\varepsilon \sim Skew\,t(0,1,5,1.5)$	$\varepsilon \sim\sim \mathcal{N}(0,1)$	$\varepsilon \sim \mathcal{N}(0,1)$
Nonnested information sets			
(FPE)	GARCH (1, 1),	GARCH (1, 1),	GARCH (1, 1),
	$\varepsilon \sim Skew\,t(0,1,5,1.5)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$
		Known	Unknown
		conditional variance	conditional variance

Table 5.1: Models for parameter estimation error, model misspecification and nonnested information sets in the simulation study for the finite-sample size property of the D-M test. Models are used in making quantile and expected shortfall forecasts and evaluated using selected consistent scoring functions.

	DGP	Forecasting procedure A	Forecasting procedure B
Parameter estimation error			
(FPE)	AR (1)-GARCH (1, 1),	GARCH (1, 1),	AR (1)-GARCH (1, 1),
	$\varepsilon \sim Skew\,t(0,1,3,3.5)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$	$\varepsilon \sim \mathcal{N}(0,1)$
Model misspecification			
(FPE)	AR (1)-GARCH (1, 1),	GARCH (1, 1),	ARCH (1),
	$\varepsilon \sim Skew\,t(0,1,4,3.5)$	$\varepsilon \sim t(0,1,\nu)$	$\varepsilon \sim \mathcal{N}(0,1)$
Nonnested information sets			
(FPE)	GARCH (1, 1),	GARCH (1, 1),	ARCH (1),
	$\varepsilon \sim Skew\,t(0,1,5,3)$	$\varepsilon \sim Skew\,t(0,1,\nu,\gamma)$	$\varepsilon \sim \mathcal{N}(0,1)$
		Known	Unknown
		conditional variance	conditional variance

Table 5.2: Models for the three scenarios (parameter estimation error, model misspecification and nonnested information sets) in the simulation study for the finite-sample power property of the D-M test. Models are used in making quantile and expected shortfall forecasts and evaluated using selected consistent scoring functions.

process, we would like to assess which forecast is the most accurate. There is the need for formal tests to compare predictive accuracy. Diebold and Mariano [1995] proposed the Diebold-Mariano test which is now widely used in the field of econometrics in comparing the predictive accuracy of competing forecasts. The D-M test makes use of (consistent) scoring functions in assessing the accuracy of forecasts. The scores for point forecasts from competing forecasting procedures are obtained from a scoring function and the differences in the score values are used in the computation of the test statistic. This indicates that the ability of the scoring function to accurately distinguish between forecasting procedures is key to the outcome of the test. Patton and Sheppard [2009] assess the size and power properties of the D-M test for homogeneous scoring functions used in evaluation volatility proxies and identify the 0-homogeneous scoring function as a desirable scoring function for evaluation of volatility forecasts.

The D-M test is for pairwise comparison of forecasts with a null hypothesis of equal expected score value against a one-sided or a two-sided alternative hypothesis. The D-M test is defined under the following null hypothesis:

$$H_0 : E[S(\hat{R}_t^A, Y_t)] = E[S(\hat{R}_t^B, Y_t)] \quad \forall t, \tag{5.1}$$

where $\hat{R}_t^A$ and $\hat{R}_t^B$ are point forecasts from two competing forecasting procedures at time t, Y_t is the observed value at time t, and $S : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$, is a consistent scoring function. For the computation of the test statistic, we first define the difference in a series of values from a scoring function for forecasts $\{d_t : t = 1, 2, ..., T\}$ as:

$$d_t = S(\hat{R}_t^A, Y_t) - S(\hat{R}_t^B, Y_t) \tag{5.2}$$

and the test statistic is given as:

$$DM_T = \frac{\bar{d}_T}{\sqrt{\widehat{avar}(\bar{d}_T)}}, \tag{5.3}$$

$$\text{where} \quad \bar{d}_T \equiv \frac{1}{T} \sum_{t=1}^{T} d_t$$

and $\widehat{avar}(\bar{d}_T)$ is a consistent estimator of the asymptotic variance of the average difference. The asymptotic variance of the average difference can be computed using the Newey-West variance estimator with the number of lags set to $\mathrm{T}^{\frac{1}{3}}$ for h step ahead forecasts where $h > 1$ [Patton and Sheppard, 2009]. For $h = 1$ it is possible to show that the score differential series is covariance stationary and hence we can estimate the asymptotic variance of the average difference using the sample variance. Under the null hypothesis, the test statistic is asymptotically normally distributed [Diebold and Mariano, 1995].

5.2 Simulation-based result for size properties

We begin by assessing the size properties of the D-M test when the respective homogeneous consistent scoring functions for VAR and pair (VAR, ES) are used in assessing forecasts under realistic scenarios.

One-step ahead forecasts for the 0.90, 0.95 and 0.99 quantiles using a moving window of $w = 1000$ observations are obtained for four different out-of-sample sizes, $T = \{500, 1000, 1500, 2000\}$. The D-M test is repeated 2000 times for each T at a 5% level of significance. For the homogeneous GPL scoring functions for VAR, we consider $b \in \{0, 0.1, 0.5, 1, 1.5, 2, 3, 5\}$.
We test the hypotheses

$$\begin{aligned} H_0 &: E[S(\hat{R}_t^A, Y_t)] = E[S(\hat{R}_t^B, Y_t)] \quad \forall t \\ H_1 &: E[S(\hat{R}_t^A, Y_t)] \neq E[S(\hat{R}_t^B, Y_t)] \quad \forall t. \end{aligned}$$

The first scenario considered is the situation of parameter estimation error. For this example, we generate data $\{Y_t\}_{t\in\mathbb{Z}}$ from a GARCH (1, 1) model first with skewed t innovations and then with t-distributed innovations:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= 0.05 + 0.10 Y_{t-1}^2 + 0.85 \sigma_{t-1}^2. \\ \text{Case 1: } \varepsilon_t &\overset{iid}{\sim} Skew\, t(0, 1, 5, 1.5) \\ \text{Case 2: } \varepsilon_t &\overset{iid}{\sim} t(0, 1, 5) \end{aligned} \tag{5.4}$$

The two forecasting procedures compared are based on correctly specified models that are subject to estimation error. The fully parametric estimation (FPE) and filtered historic simulation (FHS) procedures [2] are used in parameter estimations. forecasting procedure A uses a GARCH (1, 1) model while forecasting procedure B uses an AR (1)-GARCH (1, 1) model with zero mean:

$$\begin{aligned} \widehat{VAR}_\alpha(Y_t^A | \mathscr{F}_{t-1}) &= \widehat{\sigma}_t \widehat{VAR}_\alpha(\varepsilon_t) \\ \widehat{VAR}_\alpha(Y_t^B | \mathscr{F}_{t-1}) &= \widehat{\mu}_t + \widehat{\sigma}_t \widehat{VAR}_\alpha(\varepsilon_t) \end{aligned} \tag{5.5}$$

[2] description of the estimation methods is given in the Appendix.

where $\widehat{VAR}_\alpha(\varepsilon_t) = F_\varepsilon^{-1}(\alpha)$, with $F_\varepsilon(.)$ denoting the cumulative distribution function of the distribution of the innovations. For case 1, both forecasting procedures specify a skewed t distribution for the innovations while both forecasting procedures also specify a t-distribution with mean zero and variance one for the innovations in the second case.

Table 5.3 shows the results of the size of the D-M test for the case where the fully parametric approach is used in the estimation of model parameters. The size of the test when using the homogeneous GPL scoring function with b less than 2 is close to the theoretical value of 0.05 for the chosen α levels and out-of-sample sizes (T). For higher values of b, e.g., b = 5, the size of the test is higher than the theoretical value. Table 5.4 shows the size of the test when forecasting procedures use the FHS in estimating parameters and once again, homogeneous scoring functions with lower homogeneity order perform better compared to the higher homogeneity order values.

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.059	0.054	0.063	0.058	0.046	0.060	0.053	0.057
b=0.1	0.059	0.057	0.063	0.059	0.047	0.061	0.054	0.056
b=0.5	0.061	0.055	0.059	0.058	0.050	0.065	0.055	0.059
b=1.0	0.061	0.056	0.059	0.058	0.056	0.064	0.057	0.062
b=1.5	0.063	0.066	0.063	0.057	0.062	0.072	0.059	0.067
b=2.0	0.063	0.073	0.071	0.063	0.075	0.083	0.074	0.080
b=3.0	0.073	0.086	0.083	0.078	0.097	0.125	0.106	0.116
b=5.0	0.077	0.104	0.092	0.096	0.118	0.150	0.142	0.150
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.076	0.072	0.076	0.076				
b=0.1	0.078	0.072	0.074	0.075				
b=0.5	0.079	0.075	0.078	0.076				
b=1.0	0.080	0.079	0.080	0.079				
b=1.5	0.093	0.090	0.091	0.093				
b=2.0	0.113	0.115	0.111	0.110				
b=3.0	0.159	0.183	0.177	0.174				
b=5.0	0.259	0.312	0.308	0.303				

Table 5.3: Parameter estimation error: (Case 1) Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

We now consider the case where the innovations follow a t-distribution with mean zero and variance one rather than a skewed t-distribution using the same models in (5.5) for forecasting procedure A and forecasting procedure B respec-

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.071	0.078	0.094	0.092	0.086	0.091	0.109	0.131
b=0.1	0.070	0.078	0.093	0.094	0.088	0.089	0.108	0.131
b=0.5	0.077	0.076	0.084	0.089	0.091	0.091	0.105	0.124
b=1.0	0.077	0.076	0.082	0.090	0.092	0.092	0.106	0.118
b=1.5	0.084	0.086	0.081	0.083	0.098	0.095	0.106	0.114
b=2.0	0.088	0.092	0.084	0.088	0.106	0.109	0.114	0.117
b=3.0	0.091	0.105	0.092	0.107	0.144	0.141	0.138	0.137
b=5.0	0.102	0.112	0.102	0.119	0.178	0.184	0.164	0.172
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.145	0.124	0.141	0.152				
b=0.1	0.145	0.121	0.140	0.152				
b=0.5	0.152	0.123	0.146	0.147				
b=1.0	0.167	0.134	0.148	0.155				
b=1.5	0.181	0.149	0.159	0.155				
b=2.0	0.202	0.169	0.181	0.165				
b=3.0	0.248	0.216	0.227	0.216				
b=5.0	0.339	0.340	0.359	0.345				

Table 5.4: Parameter estimation error: (Case 1) Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The filtered historic simulation is used in the estimation of the model parameters in this case.

tively. Estimation of parameters is done using the fully parametric approach. The results presented in Table 5.5 show that the size of the test is close to the theoretical value for the homogeneous GPL scoring function where b is 2 or less for the three α levels. For higher values of b (e.g., b = 3), the size is slightly lower than 0.05 for the 0.90 and 0.95 quantiles but gets close to the theoretical value for 0.99 quantile. Overall, the homogeneous GPL scoring functions with b less than 2 perform well in the cases considered under the parameter estimation error scenario.

Next, we examine the size of the D-M test for the situation where the two forecasting procedures which are based on similar but misspecified models. We look at the ability of the scoring function to assess similar forecasts. Just as in the first scenario, we generate data $\{Y_t\}_{t\in\mathbb{Z}}$ from a GARCH (1, 1) model with skewed t innovations:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= 0.05 + 0.10 Y_{t-1}^2 + 0.85 \sigma_{t-1}^2. \\ \varepsilon_t &\overset{iid}{\sim} Skew\ t(0,1,5,1.5) \end{aligned} \tag{5.6}$$

For the first case, the forecasting procedures specify a t-distribution with mean

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.044	0.055	0.062	0.060	0.051	0.061	0.060	0.061
b=0.1	0.045	0.056	0.063	0.061	0.048	0.055	0.061	0.064
b=0.5	0.046	0.059	0.060	0.061	0.048	0.057	0.060	0.061
b=1.0	0.047	0.056	0.056	0.061	0.049	0.059	0.058	0.058
b=1.5	0.049	0.056	0.055	0.060	0.044	0.055	0.058	0.056
b=2.0	0.049	0.054	0.054	0.055	0.041	0.055	0.056	0.049
b=3.0	0.046	0.047	0.049	0.052	0.037	0.050	0.049	0.044
b=5.0	0.040	0.040	0.038	0.036	0.038	0.040	0.044	0.038
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.051	0.060	0.051	0.062				
b=0.1	0.050	0.059	0.051	0.062				
b=0.5	0.049	0.053	0.048	0.062				
b=1.0	0.046	0.052	0.047	0.061				
b=1.5	0.043	0.049	0.047	0.055				
b=2.0	0.042	0.047	0.042	0.053				
b=3.0	0.044	0.046	0.046	0.052				
b=5.0	0.054	0.046	0.052	0.055				

Table 5.5: Parameter estimation error: (Case 2) Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

zero and variance one for the innovations of their models, while they both specify a standard normal distribution for the innovations in the second case. Forecasting procedure A uses a GARCH (1, 1) model while forecasting procedure B uses an AR (1)-GARCH (1, 1) model.

The results presented in Table 5.6 show that the size of the test is close to the theoretical value for lower values of b of the homogeneous scoring function. The size of the test is, however, higher for larger values of b(e.g., $b = 5$) as we get closer to the upper tail. Table 5.7 shows results of the second example where both forecasting procedures specify a standard normal distribution for the innovations of their models. The size of the test is generally close to the theoretical value for most values of b. Some cases of higher size values are observed for the homogeneous scoring function when $b > 2$ and T increases. Overall, the homogeneous GPL scoring functions with $b < 2$ do perform quite well in assessing the similar forecasts.

The last scenario we consider for the size of the D-M test is when the forecasting procedures have nonnested information sets. Forecasting procedure A has information about the conditional variance of the returns while forecasting procedure

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.064	0.061	0.065	0.063	0.061	0.065	0.067	0.059
b=0.1	0.063	0.060	0.065	0.063	0.059	0.065	0.065	0.059
b=0.5	0.063	0.060	0.061	0.064	0.059	0.064	0.066	0.057
b=1.0	0.064	0.068	0.063	0.062	0.060	0.066	0.067	0.058
b=1.5	0.064	0.074	0.066	0.063	0.060	0.066	0.067	0.064
b=2.0	0.063	0.074	0.068	0.066	0.061	0.071	0.072	0.067
b=3.0	0.068	0.082	0.071	0.066	0.074	0.083	0.076	0.072
b=5.0	0.064	0.083	0.070	0.069	0.081	0.090	0.080	0.085
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.045	0.060	0.063	0.061				
b=0.1	0.046	0.060	0.060	0.060				
b=0.5	0.047	0.061	0.061	0.061				
b=1.0	0.050	0.057	0.058	0.058				
b=1.5	0.050	0.056	0.057	0.057				
b=2.0	0.054	0.064	0.062	0.062				
b=3.0	0.073	0.076	0.065	0.065				
b=5.0	0.119	0.116	0.107	0.107				

Table 5.6: Model misspecification: (Case 1) Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

B has no information on the conditional variance of the process. Both forecasting procedures use a GARCH (1, 1) model with skewed-t innovations. The data is generated from the same model used for the first case of the parameter estimation example. Table 5.8 shows that the size of the test is close to the theoretical value for lower values of b of the homogeneous GPL scoring function across the different out-of-sample sizes. The size values are higher for all homogeneous GPL scoring functions for the case where $\alpha = 0.99$. The homogeneous GPL scoring function with $b = 0$ seems to perform well amongst all the scoring functions with size values relatively close to the theoretical value of the test in most of the cases considered.

We now assess the performance of consistent scoring functions used in evaluating forecasts of the pair (VAR, ES) using the same scenarios presented for VAR. A one-step ahead forecast is estimated for the ES at level $\nu = \{0.754, 0.875, 0.975\}$ which should yield similar magnitude of risk as VAR$_{0.90}$, VAR$_{0.95}$, and VAR$_{0.99}$ respectively [Nolde and Ziegel, 2016].

For the size of the D-M test, the (1/2)-homogeneous and 0-homogeneous scoring functions presented in (2.10) and (2.11) respectively, produce size values which are close to the theoretical value of 0.05 for the different values of T at the dif-

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.062	0.063	0.063	0.066	0.058	0.066	0.077	0.069
b=0.1	0.063	0.064	0.063	0.067	0.059	0.068	0.077	0.067
b=0.5	0.062	0.065	0.067	0.067	0.056	0.071	0.079	0.071
b=1.0	0.062	0.068	0.068	0.065	0.058	0.070	0.077	0.073
b=1.5	0.061	0.074	0.070	0.069	0.061	0.072	0.080	0.074
b=2.0	0.061	0.085	0.077	0.078	0.065	0.078	0.085	0.080
b=3.0	0.071	0.090	0.083	0.079	0.072	0.084	0.085	0.084
b=5.0	0.063	0.087	0.080	0.077	0.077	0.089	0.082	0.095
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.054	0.068	0.069	0.078				
b=0.1	0.054	0.070	0.071	0.078				
b=0.5	0.053	0.070	0.068	0.074				
b=1.0	0.052	0.070	0.067	0.079				
b=1.5	0.051	0.070	0.064	0.073				
b=2.0	0.055	0.068	0.059	0.073				
b=3.0	0.055	0.065	0.057	0.067				
b=5.0	0.064	0.069	0.066	0.069				

Table 5.7: Model misspecification: (Case 2) Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

ferent levels of ν. Comparing the size values of the pair (VAR, ES) at level $\nu = \{0.754, 0.875, 0.975\}$ with the size values obtained for VAR at level $\alpha = \{0.90, 0.95, 0.99\}$, it is seen that the values obtained for the 0-homogeneous and (1/2)-homogeneous scoring functions used in assessing the pair (VAR, ES) forecasts are similar to that of the 0-homogeneous and (1/2)-homogeneous GPL scoring functions used in assessing the VAR forecasts. Overall, the size values for the D-M test in the evaluation of forecasts of the pair (VAR, ES) are close to the theoretical value of 0.05. The results for size of the D-M test for the pair (VAR$_\nu$, ES$_\nu$) under the different scenarios are presented in the Tables 5.9, 5.10, 5.11, 5.12, and 5.13.

5.3 Simulation-based result for power properties

For the power of the D-M test, we look at the ability of the scoring function to distinguish between forecasts and assign smaller scoring values to forecasts close

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.058	0.064	0.074	0.073	0.059	0.064	0.073	0.077
b=0.1	0.056	0.064	0.074	0.074	0.060	0.066	0.071	0.078
b=0.5	0.054	0.066	0.067	0.076	0.060	0.061	0.072	0.078
b=1.0	0.052	0.064	0.066	0.076	0.065	0.064	0.071	0.078
b=1.5	0.056	0.066	0.065	0.079	0.071	0.070	0.072	0.077
b=2.0	0.060	0.076	0.065	0.086	0.083	0.083	0.083	0.082
b=3.0	0.079	0.103	0.094	0.108	0.131	0.130	0.123	0.124
b=5.0	0.159	0.179	0.165	0.162	0.243	0.243	0.237	0.244
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.130	0.108	0.097	0.109				
b=0.1	0.133	0.108	0.100	0.112				
b=0.5	0.138	0.111	0.105	0.114				
b=1.0	0.144	0.122	0.112	0.117				
b=1.5	0.164	0.136	0.120	0.136				
b=2.0	0.185	0.156	0.144	0.155				
b=3.0	0.259	0.225	0.227	0.235				
b=5.0	0.439	0.426	0.437	0.441				

Table 5.8: Nonnested information sets: Size values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

	$\nu = 0.754$				$\nu = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.061	0.059	0.070	0.063	0.054	0.060	0.054	0.052
b=0.5	0.059	0.064	0.071	0.067	0.059	0.057	0.061	0.056
	$\nu = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.053	0.064	0.066	0.063				
b=0.5	0.059	0.061	0.059	0.062				

Table 5.9: Parameter estimation error: (Case 1) Size values of the D-M test for the one-step ahead forecast of (VAR, ES) for ν values 0.754, 0.875 and 0.975 with various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

to the realized value. We test the hypotheses

$$H_0 : E[S(\hat{R}_t^A, Y_t)] = E[S(\hat{R}_t^B, Y_t)] \ \ \forall t$$
$$H_1 : E[S(\hat{R}_t^A, Y_t)] < E[S(\hat{R}_t^B, Y_t)] \ \ \forall t.$$

We examine the three scenarios examined under the size of the D-M test in this section. We begin with the parameter estimation error scenario for VAR forecasts where we generate data from an AR (1)-GARCH (1, 1) model with a zero mean

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.060	0.063	0.070	0.080	0.050	0.063	0.067	0.075
b=0.5	0.055	0.066	0.069	0.079	0.049	0.058	0.068	0.075
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.042	0.048	0.058	0.058				
b=0.5	0.041	0.046	0.057	0.056				

Table 5.10: Parameter estimation error: (Case 2) Size values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.056	0.067	0.067	0.064	0.059	0.061	0.064	0.061
b=0.5	0.056	0.069	0.068	0.065	0.059	0.063	0.070	0.060
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.051	0.058	0.069	0.069				
b=0.5	0.058	0.061	0.064	0.069				

Table 5.11: Model misspecification: (Case 1) Size values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 with various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

and a right-skewed t-distribution for the innovations:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \varepsilon_t, \quad \mu_t = 0.05Y_{t-1} \\ \sigma_t^2 &= 0.05 + 0.10Y_{t-1}^2 + 0.85\sigma_{t-1}^2. \\ \varepsilon_t &\overset{iid}{\sim} Skew\ t(0,1,3,3.5) \end{aligned} \tag{5.7}$$

Forecasting procedure A uses a GARCH (1, 1) model with innovations following a skewed-*t* distribution while forecasting procedure B uses an AR (1)-GARCH (1, 1) model with standard normal innovations. The results in Table 5.14 show that the homogeneous GPL scoring function with lower b values are able to discern superior forecast performance for all the quantile levels recording higher power values. Figure 5.1 shows the pattern plot of the power values for all $T \in \{500, 1000, 1500, 2000\}$ across the values b. The plots for the 0.90, 0.95 and 0.99 quantile show a general drop in the power of the test as the value of b increases for all levels of T. The highest power is mostly obtained when b = 0.

For the second scenario, which looks at model misspecification, data is gener-

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.058	0.068	0.075	0.068	0.060	0.063	0.071	0.065
b=0.5	0.061	0.077	0.073	0.078	0.064	0.068	0.068	0.069
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.056	0.070	0.080	0.076				
b=0.5	0.058	0.072	0.071	0.075				

Table 5.12: Model misspecification: (Case 2) Size values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 with various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.063	0.062	0.073	0.074	0.066	0.070	0.082	0.081
b=0.5	0.056	0.054	0.073	0.073	0.057	0.066	0.075	0.075
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.108	0.099	0.105	0.121				
b=0.5	0.092	0.090	0.097	0.108				

Table 5.13: Nonnested information sets: Size values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

ated from an AR (1)-GARCH (1, 1) model with a zero mean and a right-skewed t distribution for the innovations:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \varepsilon_t, \quad \mu_t = 0.05 Y_{t-1} \\ \sigma_t^2 &= 0.05 + 0.10 Y_{t-1}^2 + 0.85 \sigma_{t-1}^2. \\ \varepsilon_t &\overset{iid}{\sim} Skew\ t(0,1,4,3.5) \end{aligned} \tag{5.8}$$

Forecasting procedure A uses a GARCH (1, 1) model but wrongly specifies the distribution of the innovations by choosing a t-distribution. forecasting procedure B on the other hand wrongly specifies the model and innovation using an ARCH (1) model with standard normal innovations. In this case, we expect forecasting procedure A to perform better than forecasting procedure B. Table 5.15 shows high power values for lower b of the homogeneous GPL scoring function with the lowest power obtained at b = 5. Higher power values for the lower values of b indicates the ability of these homogeneous GPL scoring functions to distinguish between forecasting procedures. Figure 5.2 shows the pattern plot of the power

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.389	0.513	0.634	0.701	0.210	0.276	0.347	0.409
b=0.1	0.368	0.511	0.635	0.698	0.189	0.264	0.339	0.397
b=0.5	0.379	0.522	0.638	0.694	0.187	0.255	0.336	0.385
b=1.0	0.389	0.521	0.646	0.668	0.188	0.242	0.322	0.363
b=1.5	0.386	0.499	0.579	0.625	0.192	0.242	0.292	0.331
b=2.0	0.403	0.478	0.542	0.576	0.203	0.227	0.263	0.297
b=3.0	0.394	0.436	0.467	0.471	0.206	0.221	0.222	0.243
b=5.0	0.370	0.350	0.344	0.340	0.202	0.190	0.205	0.206
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.477	0.715	0.836	0.888				
b=0.1	0.459	0.710	0.830	0.884				
b=0.5	0.423	0.681	0.801	0.863				
b=1.0	0.360	0.612	0.742	0.813				
b=1.5	0.269	0.490	0.628	0.714				
b=2.0	0.171	0.336	0.461	0.547				
b=3.0	0.046	0.109	0.139	0.171				
b=5.0	0.009	0.016	0.012	0.018				

Table 5.14: Parameter estimation error: Power values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

values for all $T \in \{500, 1000, 1500, 2000\}$ across the values b. The plots for the 0.90 and 0.95 quantile show a peak around b = 1.5 with general drop in the power of the test as the value of b increases.

For the third scenario, we look at the case of forecasting procedures with nonnested information sets with data generated from a GARCH (1, 1) model with innovations from a right-skewed t-distribution with mean zero and unit variance:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \varepsilon_t, \quad \mu_t = 0.05Y_{t-1} \\ \sigma_t^2 &= 0.05 + 0.10Y_{t-1}^2 + 0.85\sigma_{t-1}^2. \\ \varepsilon_t &\overset{iid}{\sim} Skew\ t(0,1,5,3) \end{aligned} \tag{5.9}$$

Forecasting procedure A uses a GARCH (1, 1) model with skewed t-distributed innovations while forecasting procedure B uses an ARCH (1) model with standard normal innovations. Forecasting procedure A has knowledge of the conditional variance while forecasting procedure B has no knowledge of the conditional variance. The results presented in Table 5.16 show high power values for lower values of b of the homogeneous GPL scoring function with the power value decreasing as

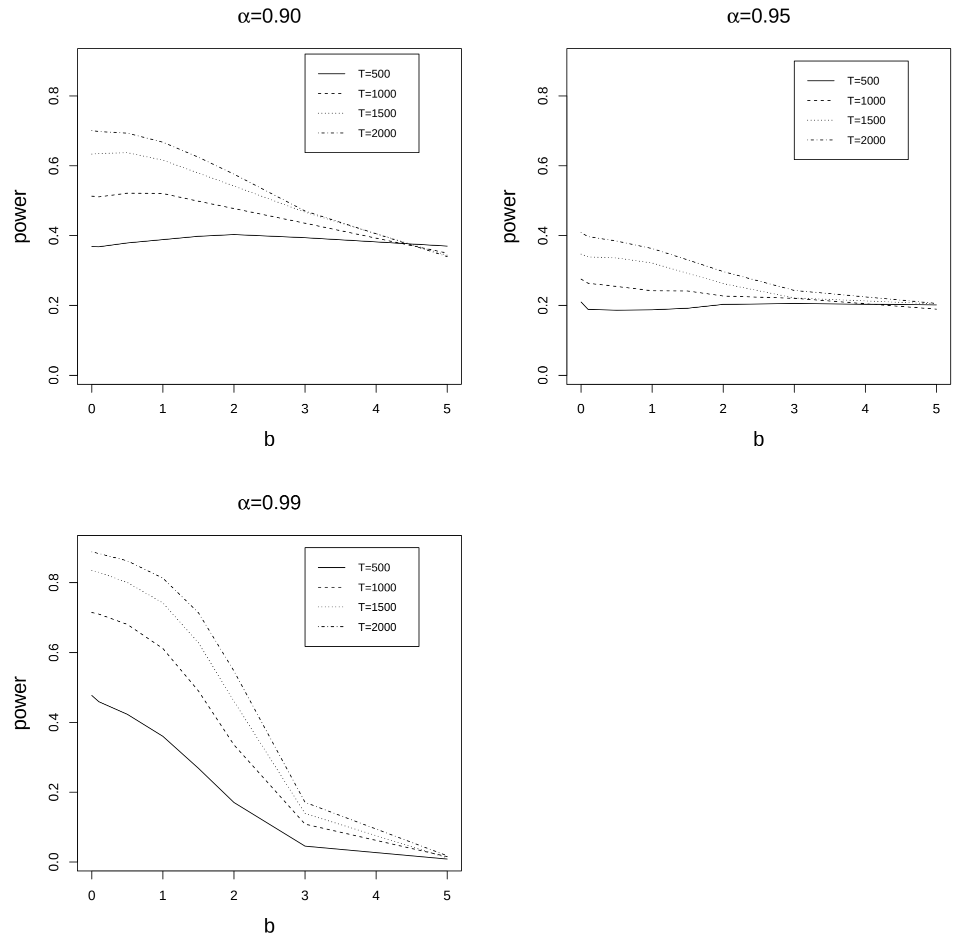


Figure 5.1: Plot of power of the D-M test against homogeneity order *b* of the homogeneous GPL for the different out-of-sample sizes.

b increases. From Figure 5.3, we see a general drop in the power values for all T $\in \{500, 1000, 1500, 2000\}$ as the value of *b* increases. The highest power values of the D-M test are obtained when the homogeneous GPL scoring function with b = 0 is used in assessing forecasting procedures.

For the power of the D-M test for the pair (VAR, ES) forecasts, the 0-homogeneous and (1/2)-homogeneous scoring functions perform well under the three cases presented. The power values increase as T gets larger for the different ν levels. Comparing the power of (VAR, ES) at levels of $\nu = \{0.754, 0.875, 0.975\}$ to the power

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.394	0.520	0.604	0.652	0.287	0.327	0.351	0.358
b=0.1	0.385	0.520	0.612	0.659	0.261	0.293	0.330	0.341
b=0.5	0.408	0.546	0.636	0.695	0.273	0.318	0.363	0.383
b=1.0	0.430	0.560	0.654	0.710	0.283	0.351	0.404	0.437
b=1.5	0.441	0.543	0.648	0.700	0.305	0.374	0.449	0.485
b=2.0	0.460	0.533	0.620	0.658	0.332	0.389	0.461	0.503
b=3.0	0.439	0.486	0.535	0.562	0.349	0.398	0.463	0.473
b=5.0	0.326	0.346	0.361	0.360	0.295	0.328	0.348	0.346
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.494	0.613	0.708	0.773				
b=0.1	0.476	0.604	0.698	0.763				
b=0.5	0.468	0.604	0.706	0.775				
b=1.0	0.460	0.598	0.706	0.775				
b=1.5	0.443	0.589	0.706	0.767				
b=2.0	0.427	0.574	0.689	0.742				
b=3.0	0.399	0.523	0.622	0.688				
b=5.0	0.297	0.360	0.417	0.431				

Table 5.15: Model misspecification: Power values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

obtained for VAR at levels of $\alpha = \{0.90, 0.95, 0.99\}$ it is seen that the power values of the D-M test for the pair (VAR, ES) forecasts are higher than that of the VAR in most cases. Results are presented in Tables 5.17, 5.18 and 5.19. We also present plots of the power of the test against the out-of-sample sizes for the two homogeneous scoring functions considered. The plots indicate that the power of the test increases as T increases. The (1/2)-homogeneous scoring function has higher power values for v values close to the center, however, the 0-homogeneous scoring function produces higher power values as we get closer to the upper tail.

From the simulation results, we observe that the homogenous consistent scoring functions with lower homogeneity order values have desirable size and power properties for the D-M test. In particular, the 0-homogeneous scoring function for evaluating the VAR forecasts and pair (VAR, ES) forecasts respectively, have better finite-sample size and power properties in the presence of model uncertainty, parameter estimation error and nonnested information sets. It is also observed that the when looking at the extreme tails of the distribution, the size values tend to increase while the performance of the scoring function for the power of the test switches for the pair (VAR, ES) forecasts.

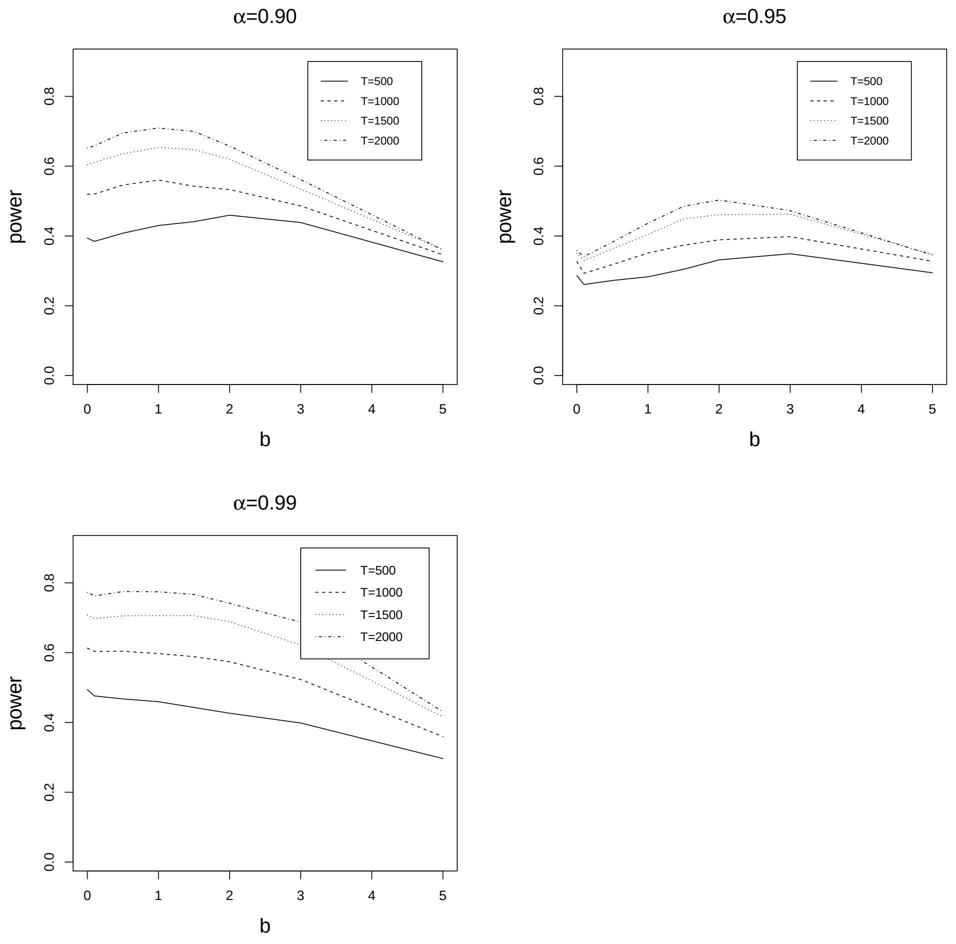


Figure 5.2: Plot of power of the D-M test against homogeneity order *b* of the homogeneous GPL for the different out-of-sample sizes.

	$\alpha = 0.90$				$\alpha = 0.95$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.402	0.570	0.702	0.779	0.649	0.871	0.938	0.970
b=0.1	0.395	0.566	0.701	0.778	0.638	0.869	0.936	0.968
b=0.5	0.395	0.555	0.694	0.771	0.619	0.850	0.927	0.963
b=1.0	0.381	0.526	0.657	0.742	0.578	0.812	0.904	0.952
b=1.5	0.371	0.488	0.593	0.682	0.511	0.753	0.860	0.927
b=2.0	0.360	0.441	0.526	0.600	0.422	0.656	0.784	0.862
b=3.0	0.333	0.369	0.415	0.462	0.240	0.401	0.504	0.576
b=5.0	0.263	0.266	0.274	0.286	0.094	0.117	0.115	0.120
	$\alpha = 0.99$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.999	0.999	0.999	0.999				
b=0.1	0.996	0.999	0.999	0.999				
b=0.5	0.993	0.999	0.999	0.999				
b=1.0	0.988	0.999	0.999	0.999				
b=1.5	0.979	0.998	0.999	0.999				
b=2.0	0.941	0.994	0.999	0.999				
b=3.0	0.672	0.873	0.923	0.937				
b=5.0	0.002	0.007	0.014	0.017				

Table 5.16: Nonnested information sets: Power values of the D-M test for the one-step ahead forecast of the 0.90, 0.95 and 0.99 quantiles at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.890	0.989	0.995	0.998	0.682	0.909	0.972	0.994
b=0.5	0.921	0.994	0.995	0.998	0.726	0.912	0.978	0.995
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.411	0.676	0.802	0.872				
b=0.5	0.356	0.604	0.741	0.815				

Table 5.17: Parameter estimation error: Power values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

	$v = 0.754$				$v = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.529	0.725	0.823	0.892	0.378	0.491	0.587	0.656
b=0.5	0.697	0.874	0.942	0.976	0.455	0.609	0.723	0.790
	$v = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.413	0.531	0.630	0.689				
b=0.5	0.375	0.487	0.581	0.635				

Table 5.18: Model misspecification: Power values of the D-M test for the one-step ahead forecast of (VAR, ES) for v values 0.754, 0.875 and 0.975 at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

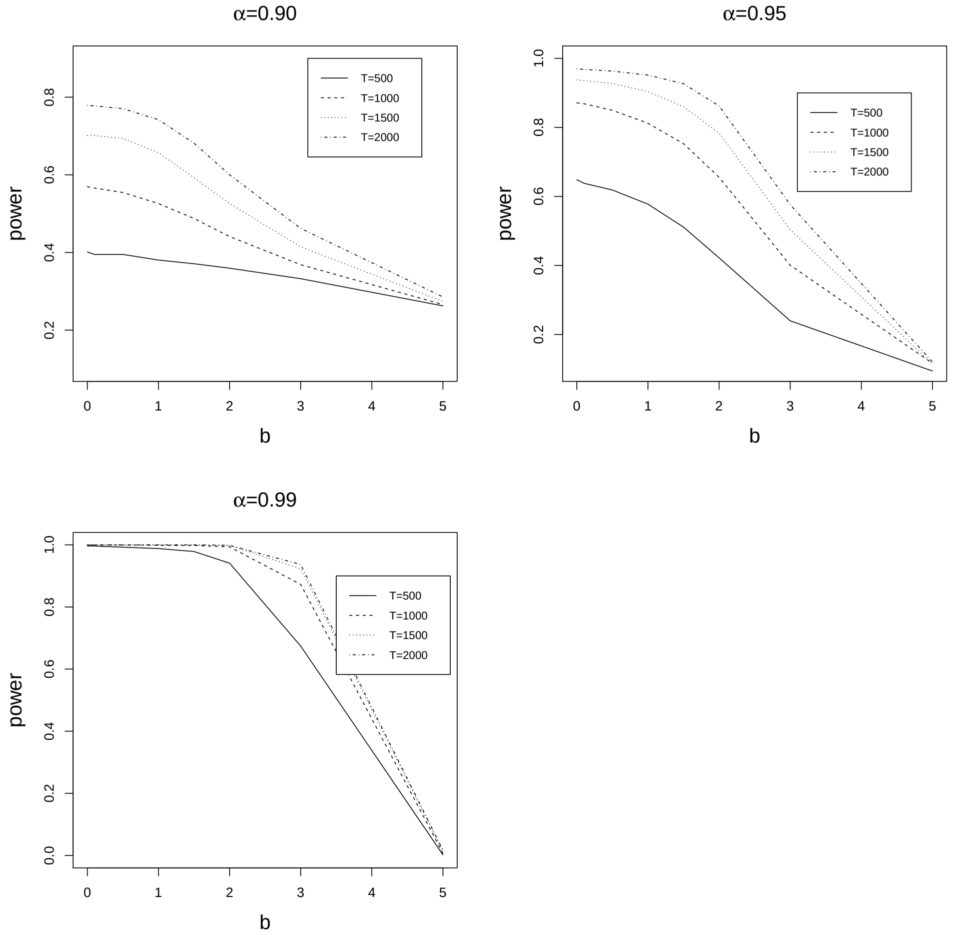


Figure 5.3: Plot of power of the D-M test against homogeneity order *b* of the homogeneous GPL for the different out-of-sample sizes.

	$\nu = 0.754$				$\nu = 0.875$			
	T= 500	T=1000	T=1500	T=2000	T= 500	T=1000	T=1500	T=2000
b=0	0.708	0.922	0.983	0.998	0.560	0.819	0.925	0.975
b=0.5	0.751	0.943	0.992	0.998	0.537	0.774	0.903	0.952
	$\nu = 0.975$							
	T= 500	T=1000	T=1500	T=2000				
b=0	0.541	0.777	0.900	0.946				
b=0.5	0.489	0.728	0.854	0.924				

Table 5.19: Nonnested information sets: Power values of the D-M test for the one-step ahead forecast of (VAR, ES) for ν values 0.754, 0.875 and 0.975 at various out-of-sample sizes. The fully parametric approach is used in the estimation of the model parameters in this case.

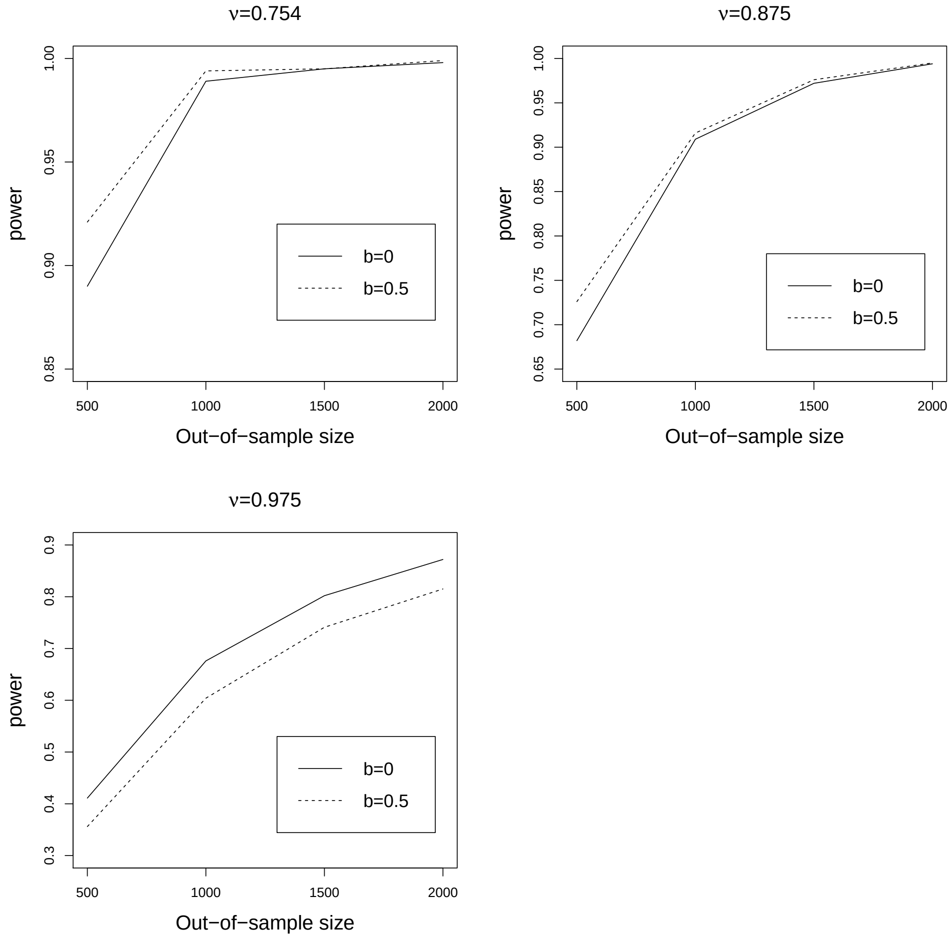


Figure 5.4: Parameter estimation Error: Plot of power of the D-M test against out-of-sample size (T) for the homogeneous scoring functions for the pair (VAR_ν, ES_ν).

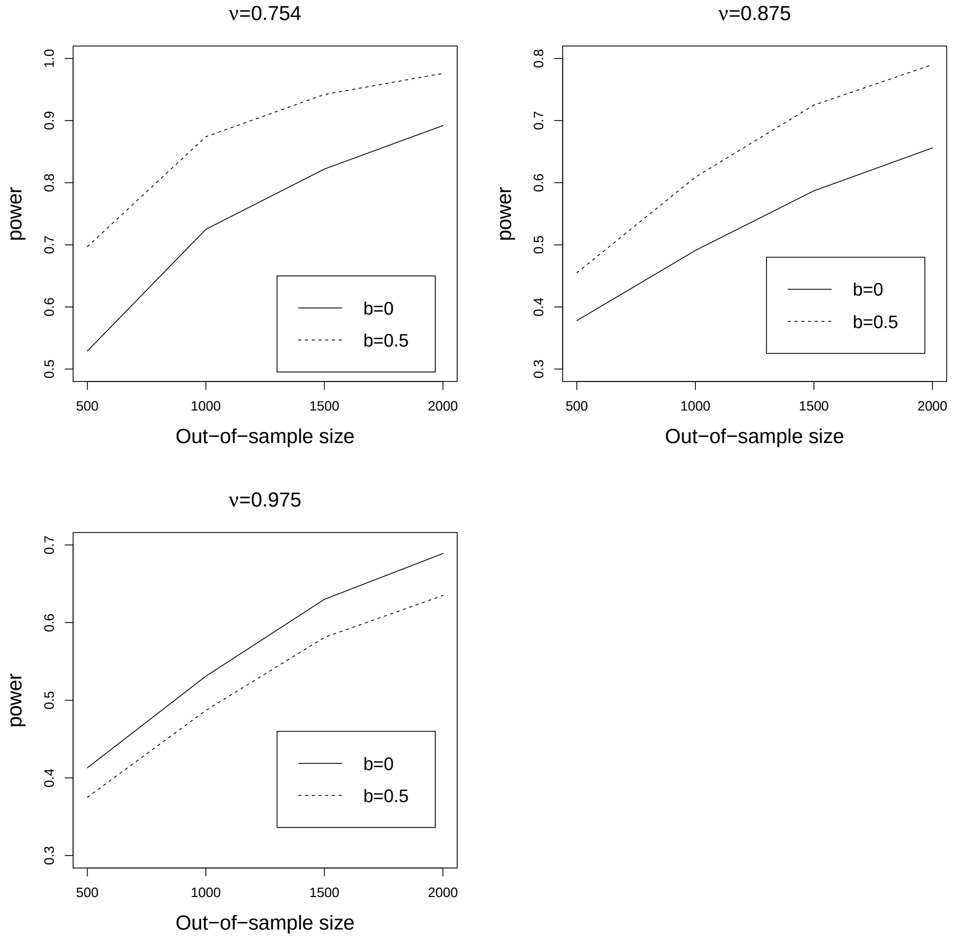


Figure 5.5: Model misspecification: Plot of power of the D-M test against out-of-sample size (T) for the homogeneous scoring functions for the pair (VAR_ν, ES_ν).

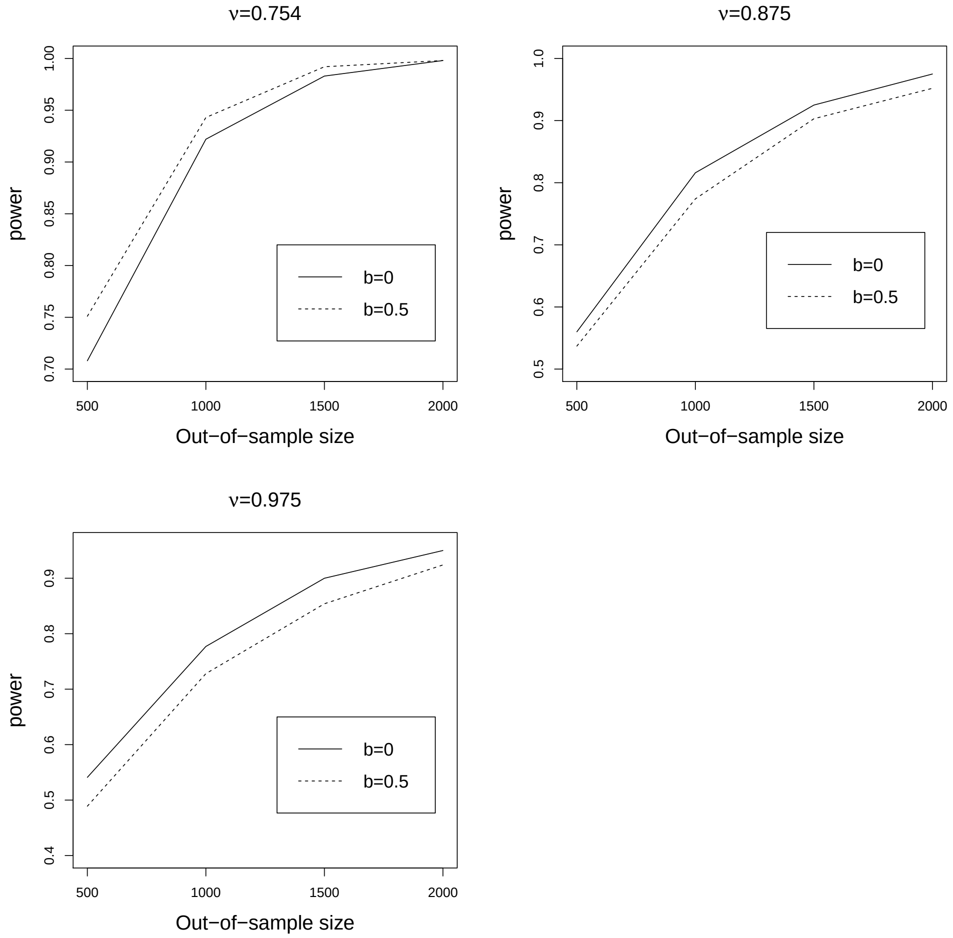


Figure 5.6: Nonnested information sets: Plot of power of the D-M test against out-of-sample size (T) for the homogeneous scoring functions for the pair (VAR_ν, ES_ν).

Chapter 6

Over-prediction and Under-prediction in finance

Financial institutions usually set aside an amount of money to deal with worst case-scenarios in the financial markets. VAR and ES are often used as the underlying risk measure to calculate the amount of capital to set aside to cover potential losses. Since institutions lose interest on the capital stored, it is their desire to make accurate predictions of the VAR and ES. If financial institutions over-predict the VAR or ES, they end up storing more capital than needed and hence lose potential interest income that they could have earned on the extra amount. However, under-prediction risk measures may lead to banks not having enough capital to absorb large losses and hence makes them crises-prone.

In this chapter, we look at how the homogeneous scoring functions penalize for over-prediction and under-prediction. Since firms only have to deal with the loss of profit that would have been made on extra money stored for over-predicted VAR or ES, it is desirable for a scoring function to penalize more for under-prediction than over-prediction.

6.1 Simulation Study

We begin by illustrating how the homogeneous scoring functions penalize for over-prediction and under-prediction of VAR$_\alpha$ of the same magnitude. We assume our returns follow a standard normal distribution and select the optimal forecast (R_{opt}) as the 0.95 quantile of the underlying distribution. The 0.99 and 0.8325 quantiles are set as the over-predicted and under-predicted VAR$_\alpha$ forecasts respectively. The over-predicted VAR$_\alpha$ forecast ($\hat{R}_{ov}$) and under-predicted VAR$_\alpha$ forecast ($\hat{R}_{ud}$) are chosen such that

$$\hat{R}_{ov} - R_{opt} = R_{opt} - \hat{R}_{ud}$$

We compute the expected score ($\mathbb{E}_F S(\hat{R}, Y)$) for chosen values of the homogeneity order (b) of our scoring function and rank the expected scores for the optimal, over-predicted and under-predicted VAR$_\alpha$ forecasts. Table 6.1 below shows the ranks for different values of the homogeneity order. It is seen that the optimal forecast has the lowest expected score and is hence ranked as the best forecast for all values b. The over-predicted VAR$_\alpha$ forecast has the second lowest expected score for the homogeneous scoring functions with lower homogeneity order. The ranking of the over-predicted and under-predicted forecasts however changes for the case where b = 3 or higher.

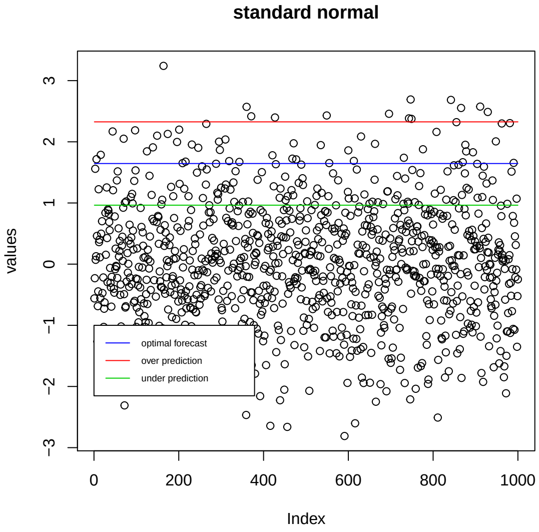


Figure 6.1: Plot depicting the optimal forecast value, over-predicted and under-predicted VAR forecasts for a case where returns are assumed to follow a standard normal distribution

	$\mathbb{E}(S_{op})$[Rank]	$\mathbb{E}(S_{ov})$[Rank]	$\mathbb{E}(S_{ud})$[Rank]
b=0	0.0372 [1]	0.0434 [2]	0.0668 [3]
b=0.1	0.5368 [1]	0.5455 [2]	0.5670 [3]
b=0.5	0.1431 [1]	0.1546 [2]	0.1750 [3]
b=1.0	0.1031 [1]	0.1197 [2]	0.1375 [3]
b=1.5	0.0999 [1]	0.1238 [2]	0.1370 [3]
b=2.0	0.1098 [1]	0.1442 [2]	0.1501 [3]
b=3.0	0.1618 [1]	0.2337 [3]	0.2098 [2]
b=5.0	0.5392 [1]	0.8592 [3]	0.6112 [2]

Table 6.1: Expected scores and corresponding forecasts for selected values of the homogeneity order b used in the scoring function for $\text{VAR}_{0.95}$ forecasts. S_{op}, S_{ov} and S_{ud} indicate scoring functions when the optimal forecast, over-predicted forecast and under-predicted forecasts, respectively, are used.

We further explore how the homogeneous scoring functions penalize for over-prediction and under-prediction as the magnitude of the difference (d) between the optimal forecast and the over-/under-predicted forecasts increases. We assess the behaviour of the scoring functions using data generated from a skewed-normal distribution with skewness parameter of 3 and a skewed t-distribution with 5 degrees of freedom and a skewness parameter of 3. The expectation of the homogeneous scoring function under the given distribution of the random variable Y is computed at various magnitudes of difference.

Figure 6.2 and Figure 6.3 display the plot of the expected score of the scoring function against the magnitude of the difference. For the two cases studied, it is seen that for lower homogeneity order, the expected score for the under-predicted value is greater than that of the over-predicted value. However, for higher values of b (e.g., $b = 3$), over-prediction is penalized more than under prediction as the magnitude of the difference (d) increases.

For the pair (VAR, ES) forecasts, we look at the 0-homogeneous and the (1/2)-homogeneous scoring functions and assess how they penalize for over-prediction and under-prediction of the same magnitude. The homogeneous scoring functions for the pair (VAR, ES) assign negative score values and hence assessment of how they penalize for under-prediction and over-prediction is done by measuring the distance from the expected score at the optimal forecast to the expected score for the under-predicted and over-predicted value at a given d. Figure 6.4 illustrates how the distance from the expected score at the optimal forecast to the expected score of the under-predicted value is higher than that for the over-predicted value

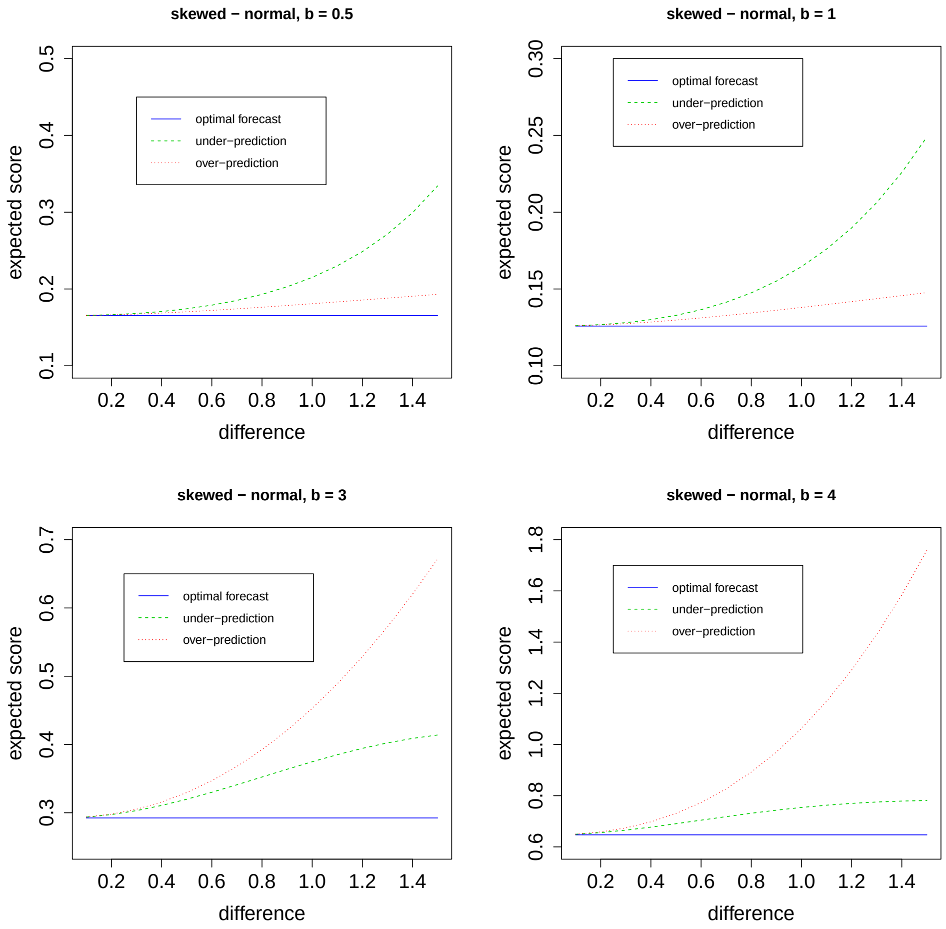


Figure 6.2: Plot of expectation of homogeneous scoring function for $\text{VAR}_{0.95}$ forecasts against the magnitude of difference for over-prediction and under-prediction. Top panel shows the case for the lower homogeneous order *b*, where under-prediction is penalized more than over-prediction. The lower panel shows the case for higher values of *b* where over-prediction is penalized more than under-prediction. Data is generated from a skewed-normal distribution with mean zero, unit variance and skewness parameter of 3.

as d increases.

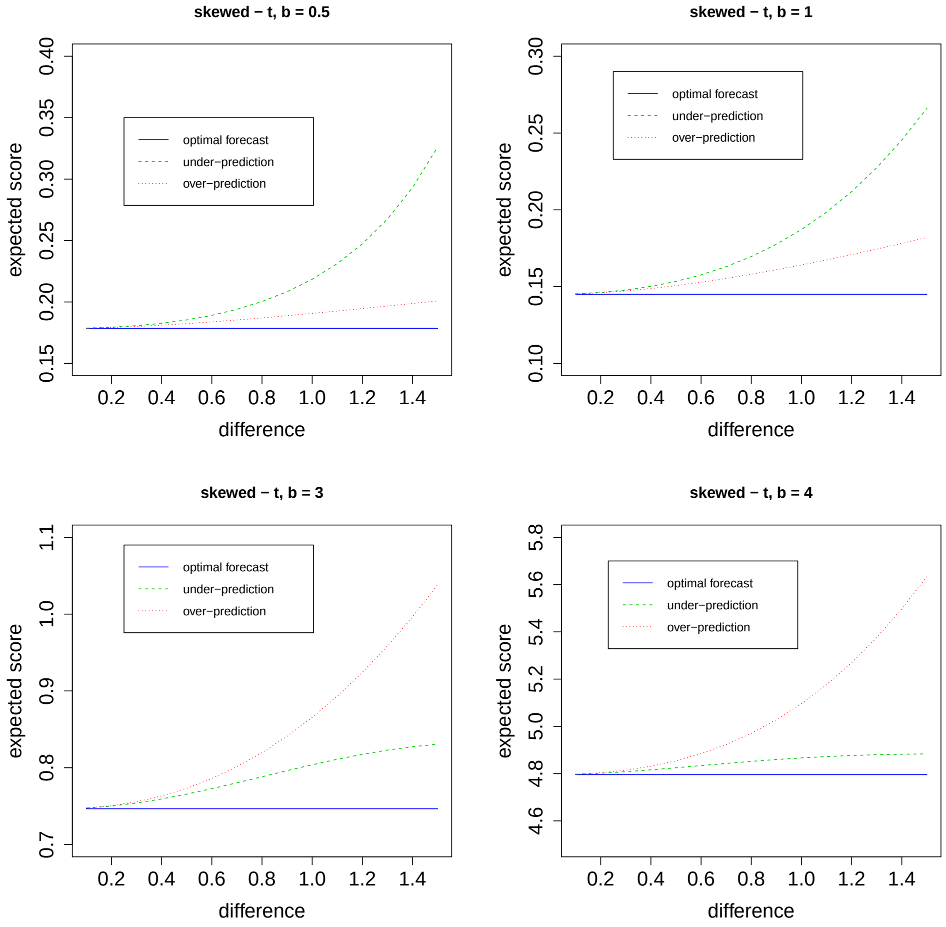


Figure 6.3: Plot of expectation of homogeneous scoring function for $\text{VAR}_{0.95}$ forecasts against the magnitude of difference for over-prediction and under-prediction. Top panel shows the case for the lower homogeneous order *b*, where under-prediction is penalized more than over-prediction. The lower panel shows the case for higher values of *b* where over-prediction is penalized more than under-prediction. Data is generated from a skewed-t distribution with mean zero, unit variance, 5 degrees of freedom and skewness parameter of 3.

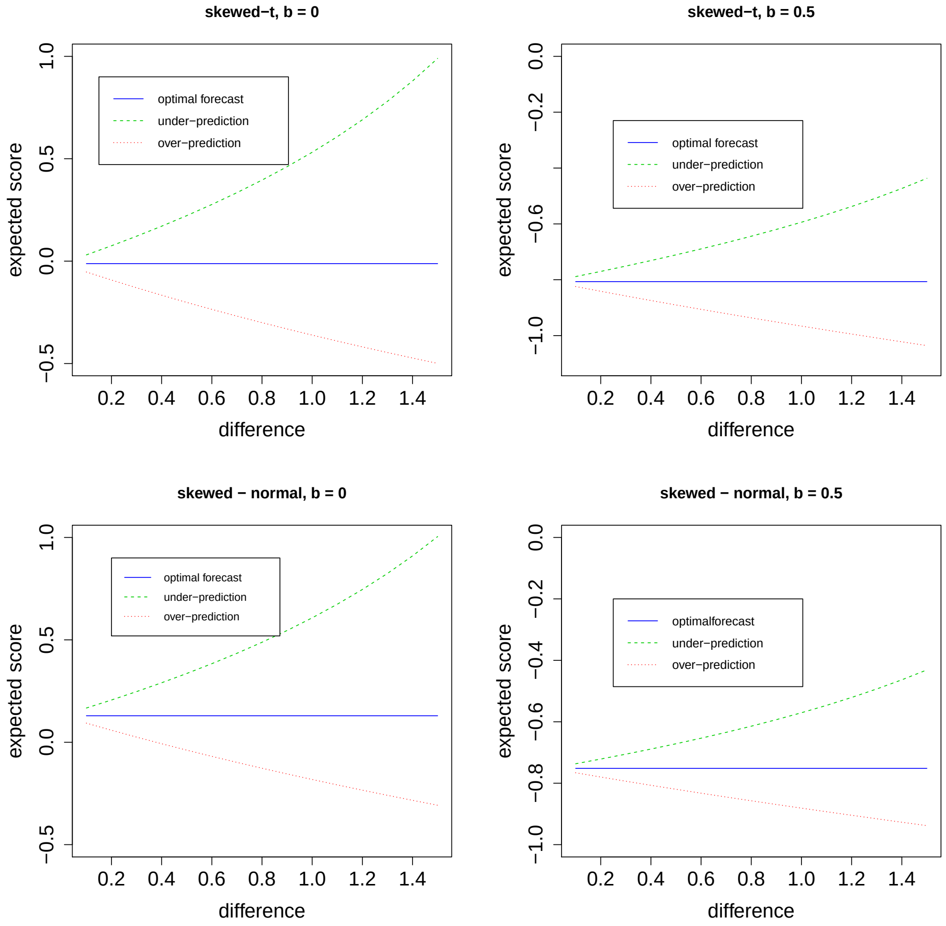


Figure 6.4: Plot of expectation of homogeneous scoring function for($\mathrm{VAR}_{0.95}$, $\mathrm{ES}_{0.95}$) forecasts against the magnitude of difference for over-prediction and under-prediction. Top panel shows the case where data is generated from a skewed-t distribution with mean zero, unit variance, 5 degrees of freedom and skewness parameter of 3. The lower panel shows the case where data is generated from a skewed-normal distribution with mean zero, unit variance and skewness parameter of 3.

Chapter 7

Conclusion

This thesis investigates the criteria for choosing desirable scoring functions for evaluating forecasting procedures for Value-at-Risk and Expected Shortfall. Gneiting [2011a] argues that consistent scoring functions lead to reasonable ranking of forecasts. Patton illustrates that in the face of model misspecification, parameter estimation error and nonnested information sets, ranking of forecasts may change depending on the consistent scoring function used and hence care should be taken in selecting a consistent scoring function for the evaluation of forecasts. We identify homogeneity of scoring functions as the first criterion since ranking of forecasts is invariant under change of units when homogeneous scoring functions are used. We concentrate on the family of homogeneous scoring functions for evaluating VAR and pair (VAR, ES) forecasts. We show that for a heavy-tailed GARCH (1, 1) model with tail index 2κ, the expectation of the homogeneous scoring functions exists as long as the homogeneity order (b) is less than the tail index. We assess the finite-sample properties of the Diebold-Mariano test [Diebold and Mariano, 1995] which is used in testing for significance of the difference between competing forecasts. With the aid of simulations, we show that homogeneous scoring functions with lower homogeneity order have better size and power properties for the D-M test and hence should be considered in evaluating forecasts of VAR and pair (VAR, ES). Lastly we show that the homogeneous scoring functions with lower homogeneity order penalize more for under-prediction than over-prediction of the same magnitude which is a desirable property of a scoring function for financial institutions when predicting risk measures.

Some future work can be done building on the criteria we have established for choosing scoring functions. Firstly, further research can be done to study the behaviour of the homogeneous scoring functions as we examine the extreme upper tails of a given distribution to help understand why the size and power values for the extreme upper tails are different from the less extreme cases. Secondly, research can be done on other criteria to consider in choosing a consistent scoring function for evaluating the VAR and the pair (VAR, ES) forecasts. Furthermore, research can be done on the criteria for choosing among non-homogeneous consistent scoring functions for evaluating VAR and pair (VAR, ES). Lastly, this work concentrates on the criteria for selecting scoring functions for evaluating forecasts for quantiles and expected shortfall in risk management. This can be extended to identify the criteria for the selection of consistent scoring functions for the evaluation of forecasts for functionals such as the mean and expectiles in different applications and fields where forecasts of these functionals are issued and evaluated.

Bibliography

C. Acerbi and B. Szekely. Backtesting expected shortfall. *Risk Magazine*, 2014. → pages 2

C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking and Finance*, 26(7):1487–1503, 2002. → pages 7

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9:203–228, 1999. → pages 7

Bank for International Settlements. Consultative Document: Fundamental review of the trading book: A revised marked risk framework. 2013. → pages 7

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *J. Bus. Econ. Stat.*, 13:253–263, 1995. → pages 3, 24, 25, 50

F.X. Diebold, T. Schuermann, and J.D. Stroughair. Pitfalls and opportunities in the use of extreme value theory in risk management. *Journal of Risk Finance*, 1:30–35, 2000. → pages 56

W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: Consistent scoring functions, choquet representations, and forecast rankings. *J. Roy. Statist. Soc. Ser. B*, 2016. To appear. → pages 11

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Ann. Statist.*, 2016. To appear. → pages 2, 9

T. Fissler, J. F. Ziegel, and T. Gneiting. Expected shortfall is jointly elicitable with value at risk – implications for backtesting. *Risk Magazine*, December, 2015. → pages 11

T. Gneiting. Making and evaluating point forecasts. *J. Am. Statist. Assoc.*, 106: 746–762, 2011a. → pages 1, 2, 3, 5, 50

T. Gneiting. Quantiles as optimal point forecasts. *Journal of Banking and Finance*, 27:197–207, 2011b. → pages 8

H. Holzmann and M. Eulert. The role of the information set for forecasting – with applications to risk management. *Ann. Stat.*, 8:595–621, 2014. → pages 15

K. Kuester, S. Mittnik, and M.S. Paolella. Value-at-risk prediction: a comparison of alternative strategies. *J. Financial Econometrics*, 4:53–89, 2006. → pages 55

S. Kusuoka. On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3:83–95, 2001. → pages 7

N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, Chicago, Il, USA, 2008. extended abstract. → pages 6

A. J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *J. Empir. Financ.*, 7:271–300, 2000. → pages 56

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, Tools*. Princeton University Press, Princeton, 2005. → pages 1

T. Mikosch and C. Starcia. Limit theory for the sample autocorrelations and extreme of a garch(1,1) process. *Ann. Stat.*, pages 1427–1451, 2000. → pages 20

N. Nolde and J. F. Ziegel. Elicitability and backtesting. *Ann. Appl. Statist.*, 2016. To appear. → pages 1, 6, 9, 10, 11, 30, 56

A. J. Patton. Evaluating and comparing possibley misspecified forecasts, https://public.econ.duke.edu/~ap172/Patton_bregman_comparison_27mar15.pdf. Accessed: 2017-04-20. → pages iii, 2, 3, 10, 15, 17, 50

A. J. Patton. Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics*, 160:246–256, 2011. → pages 16

A. J. Patton and K. Sheppard. Evaluating volatility and correlation forecasts. In T. Mikosch, J.-P. Kreiss, R. A. Davis, and T. G. Andersen, editors, *Handbook of Financial Time Series*, pages 801–838. Springer, Berlin, 2009. → pages 25

M. Saerens. Building cost functions minimizing to some summary statistics. *IEEE Trans. Neural Netw.*, 11:1263–1271, 2000. → pages 2, 9

P. Sun and C. Zhou. Diagnosing the distribution of garch innovations. *J. Empir. Financ.*, 29:287–303, 2014. → pages 17, 20

W. Thomson. Eliciting production possibilities from a well-informed manager. *J. Econ. Theory*, 20(3):360–380, 1979. → pages 2, 9

J. F. Ziegel. Coherence and elicitability. *Math. Finance*, 26(4):901–918, 2016. → pages 6

Appendix A

Forecasting Risk Measures

The aim of this report is to examine the performance of consistent scoring functions used in assessing VAR and pair (VAR, ES) forecasts. There are various methods used in producing point forecasts for risk measures. Kuester et al. [2006] review a number of approaches including the fully parametric estimation, historical simulation, quantile regression approach, among other methods. To illustrate how the estimation methods for risk measures are used, we assume the series of negated log-returns $\{Y_t\}_{t\in\mathbb{N}}$ can be modeled as

$$Y_t = \mu_t + \sigma_t \varepsilon_t, \tag{A.1}$$

where $\{\varepsilon_t\}_{t\in\mathbb{N}}$ is a sequence of independent and identically distributed (i.i.d) random variables with zero mean and unit variance, and μ_t and σ_t are measurable with respect to sigma algebra $\mathscr{F}_{t-1}$, which represents information about the process $\{Y_t\}$ available up to time t-1. To capture time dynamics of financial time series, we can assume that the conditional mean μ_t follows an ARMA process. An ARMA process of order (p,q) is given as

$$\mu_t = c + \sum_{i=1}^{p} a_i Y_{t-i} + z_t + \sum_{j=1}^{q} b_j z_{t-j}, \;\; t \in \mathbb{Z}, \tag{A.2}$$

where (z_t) is the linear innovation process of (Y_t).
The conditional variance σ_t^2 is assumed to evolve according to a GARCH (p,q)

model specification given as

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i Y_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \;\; t \in \mathbb{Z}. \tag{A.3}$$

Let $\mathscr{R}$ be a generic risk measure, a rea-valued map from a space of random variables. Then based on (A.1), the conditional one step ahead forecast of a risk measure $\mathscr{R}$ is

$$\mathscr{R}(Y_t|\mathscr{F}_{t-1}) = \mu_t + \sigma_t \mathscr{R}(\varepsilon), \tag{A.4}$$

where ε is used to denote a generic random variable with the same distribution as the ε_t's. We estimate μ_t, σ_t via the maximum likelihood procedure under a specific assumption on the distribution of innovations ε_t in (A.1) and use the fully parametric estimation (FPE) and the (filtered) historic simulation (FHS) to estimate $\mathscr{R}(\varepsilon)$ based on the sample of standardized residuals

$$\{\hat{\varepsilon}_t = (y_t - \hat{\mu}_t)/\hat{\sigma}_t\}. \tag{A.5}$$

This two-stage estimation procedure from that of McNeil and Frey [2000] and Diebold et al. [2000].

A.1 Fully parametric estimation

For the fully parametric approach, a parametric model is assumed for the innovations with parameters of the distribution estimated based on standardized residuals of the model in (A.1). From the fitted distribution, we estimate the given risk measure. For example, if the innovations follow a standardized t-distribution, then $\text{VAR}_\alpha(\varepsilon)$ is given by $t_{\hat{d}}^{-1}(\alpha)$, where $\hat{d}$ is the estimated degree of freedom and $t_{\hat{d}}^{-1}(\alpha)$ is the α-quantile of the distribution. The ES can be computed as

$$ES_\nu(\varepsilon) = \mathbb{E}(\varepsilon|\varepsilon \geq VAR_\nu(\varepsilon))$$

where numeric integration is used to evaluate the conditional expectation [Nolde and Ziegel, 2016].

A.2 Filtered historic simulation

The historic simulation makes use of non-parametric estimation of VAR$_\alpha$ based on the standardized residuals $\{\hat{\varepsilon}_t\}$ in (A.5), which can be seen as representing a filtered time series. A sample $\{\hat{\varepsilon}_t^*; 1 \leq t \leq m\}$ of a large size m (e.g., $m = 10,000$) is drawn from the estimated standardized residuals $\{\hat{\varepsilon}_t; 1 \leq t \leq n\}$ and then the empirical estimate of the α-quantile is taken which gives $\widehat{VAR}_\alpha^{FHS}(\varepsilon)$, the VAR estimate. The ES is estimated using the empirical version of the conditional expectation given that the residual exceeds the corresponding VAR estimate:

$$\widehat{ES}_\nu^{FHS}(\varepsilon) = \frac{1}{\#\{i : i = 1, ..., m, \hat{\varepsilon}_i^* > \widehat{VAR}_\alpha^{FHS}(\varepsilon)\}} \sum_{i=1}^{m} \hat{\varepsilon}_i^* \mathbb{1}\{\hat{\varepsilon}_i^* > \widehat{VAR}_\alpha^{FHS}(\varepsilon)\}.$$