Evaluating financial tail risk forecasts with the Model Confidence Set

Lukas Bauer[1]

February 26, 2024

Abstract

This paper is the first to provide results on the finite sample properties of the Model Confidence Set (MCS) by Hansen et al. (2011) applied to the asymmetric loss functions specific to financial tail risk forecasts, such as Value-at-Risk (VaR) and Expected Shortfall (ES). In this paper, we focus on statistical loss functions that are strictly consistent in the sense of Gneiting (2011a). Our comprehensive simulation results show that, first, the MCS test keeps the best model more frequently than the confidence level $1-\alpha$ in most settings. Second, it eliminates few inferior models for out-of-sample sizes of up to four years. Third, the MCS test shows little power against models that underestimate tail risk at the extreme quantile levels $p = 0.01$ and $p = 0.025$, while the power increases with the quantile level p. Our findings imply that the MCS test may be suitable to narrow down a set of competing models, but that it is not appropriate to test if a new model beats its competitors due to the lack of power.

Keywords: Model Confidence Set, VaR, ES, Simulation, Financial Risk, Forecast Evaluation

[1]Chair of Statistics and Econometrics, Institute of Economics, University of Freiburg, Rempartstr. 16, 79098, Freiburg, Germany; University of Freiburg, email: lukas.bauer@vwl.uni-freiburg.de, telephone: +49 761 203-2337

1 Introduction

Since the worldwide financial crisis of 2007 and 2008, both investors and regulators have become increasingly focused on predicting financial risk, in particular on rare but extreme events. Two important downside risk measures are the Value-at-Risk (VaR) and Expected Shortfall (ES). To assess VaR and ES forecasts, econometricians, institutions and regulators may choose among different measures. They may choose to backtest how often financial losses exceed the quantile forecasts, to employ economic objective functions or to use statistical loss functions.

In this paper, we restrict our attention to statistical losses and examine the finite sample properties of the model confidence set (MCS) by Hansen et al. (2011) when applied to evaluating VaR and ES forecasts. Furthermore, we empirically shed light on the underlying mechanisms that influence the finite sample properties of the MCS test.

The MCS is widely popular among applied econometricians and regularly used to assess the predictive ability of models that make VaR or ES forecasts. For instance, Bernardi et al. (2017) successfully use the MCS to determine model weights for combined forecasts. Both Taylor (2020) and Dimitriadis and Halbleib (2022) employ it to compare models that generate VaR and ES forecasts.

Unfortunately, little if anything is known about the finite sample properties of the MCS procedure when comparing VaR or ES forecasts. Yet, the losses that result from tail risk forecasts differ strongly from the symmetric losses that Hansen et al. (2011) consider. First, we find that the estimators of both mean and variance of the out-of-sample loss have huge variances, which may influence the MCS test's inference about the competing models' predictive ability. Frequently, the estimated loss differentials show a different sign than their expectation, i.e. they do not reflect the correct ranking of two models in terms of their expected losses. Second, we observe very high correlations between the losses of different models for both VaR and joint VaR and ES forecasts.

To study the finite sample properties of the MCS test under 'realistic' conditions, i.e. to examine losses that resemble those that we observe empirically, we perform Monte Carlo simulations and simulate forecasts. The forecasting targets of this paper are 1-day-ahead VaR and ES at the 1%, 2.5%, 5% and 10% level that result from location-scale models with time-varying conditional variances.

To evaluate the VaR and the joint VaR and ES forecasts, we use statistical loss functions that are *strictly consistent* as defined by Gneiting (2011a). A strictly consistent loss function ensures that the correct forecast of the target variable is

the unique minimizer of the expected loss. Appropriate loss functions to evaluate quantile forecasts live in the generalized piecewise linear (GPL) class. Though there is no strictly consistent loss function to evaluate ES forecasts on their own, Fissler and Ziegel (2016) provide a class of strictly consistent loss functions to jointly evaluate VaR and ES.

Realistically, we may regularly encounter situations when none of the forecasting methods consistently forecasts the true value of the target variable. Patton (2020) demonstrates that under model misspecification, estimation error or nonnested information sets of competing forecasting models, different consistent loss functions may induce different rankings of the models. Consequently, we understand the predictive ability (PA) of a forecasting method as the associated expected value of a specific loss function L.

Regardless the evaluation criterion, researchers must often choose from several or even many models that are available for forecasting. Over the last decades, researchers have thus proposed various statistical procedures to rank models according to their predictive ability. An early 'one-step' procedure is the Reality Check for data snooping by White (2000), developed to compare a baseline model to a number of other models. A rejection, however, only admits to conclude that there exists at least one superior alternative model. Hansen and Lunde (2005) adjusted his procedure by studentizing the test statistics to improve the power of the testing procedure and to make it more robust against irrelevant alternatives. Yet, both procedures do not deliver a complete subset of models with equal predictive ability (EPA). They only inform if there exists an alternative that produces more accurate forecasts. To address this issue, Romano and Wolf (2005) propose an iterative step-wise procedure to identify those superior alternatives. Their paper is a straightforward extension of White (2000) and requires the researcher to choose a certain benchmark.

Yet, the researcher may lack an obvious benchmark model or face several competing models that are asymptotically statistically equivalent. Thus, Hansen et al. (2011) choose a different approach with the Model Confidence Set (MCS) testing procedure. After a series of tests of EPA, the testing procedure returns a subset of models that contains the best model(s) with a given confidence level. For all models in the subset, we cannot reject the null of EPA at a specified level of the test α. The term 'best' refers to some suitable criterion, in this paper the smallest expected out-of-sample loss.[1][2]

[1] Alternative measures are e.g. Sharpe ratios or information criteria (Hansen et al. (2011)).

[2] As the MCS takes losses as primitives, the notion of the true model is less important. Yet, a strictly consistent loss function ensures that the true model is associated with the smallest expected loss.

In our simulations, we find that the MCS procedure usually keeps the model with the smallest expected loss - the best model - in the MCS more often than the specified confidence level $1 - \alpha$ while showing little power against inferior alternatives. The MCS test has better finite sample properties if the VaR and ES quantiles are larger, i.e. at for forecasts at the 5% and 10% level, and when markets are calmer. In particular, the MCS test shows little power against alternatives that underestimate financial tail risk at the 1% and 2.5% level. Our findings imply that the MCS test may be suitable to narrow down a set of competing models but that it is not appropriate to test if a new model beats its competitors due to the lack of power. The testing procedure's power benefits from long out-of-sample windows of two years and more. This is also the out-of-sample size for which evaluating joint VaR and ES forecasts becomes more informative than evaluating VaR forecasts. When using different parametrizations of the loss functions for both standalone VaR forecasts and joint VaR and ES forecasts, respectively, we do not find uniform patterns in the finite sample properties of the MCS.

The remainder of this paper is structured as follows. Section 2 introduces the criterion and procedure to evaluate forecasts. Next, we present the simulation and corresponding results in Section 3. Section 4 concludes.

2 Framework

In this paper, we focus on evaluating one-day-ahead VaR_p and ES_p forecasts. Formally, VaR_p and ES_p are defined as

$$VaR_p := F^{-1}(p) := \inf\{z \in \mathbb{R} : F(z) \geq p\}. \tag{2.1}$$

$$ES_p := \frac{1}{p}\int_0^p VaR_u(Y)du, \; p \in (0, 1], \tag{2.2}$$

where F is the cumulative distribution function of Y, and $ES_0(Y) = ess\inf Y$.

We assess competing forecasting methods through their global out-of-sample loss:

$$\bar{L} = \frac{1}{P}\sum_{t=1}^{P} L(x_t, y_t), \tag{2.3}$$

where $P \in \mathbb{N}$ denotes the number of forecasts, $x_1, \ldots, x_P$ are the forecasts and $y_1, \ldots, y_P$ the observations. L is a *strictly consistent* scoring or loss function.

Gneiting (2011b) characterizes a class of strictly consistent scoring or loss functions to evaluate quantile forecasts, the so called 'generalized piecewise linear'

(GPL) class. The loss functions take the form

$$L(x, y) = (\mathbb{1}\{y < x\} - p) \times (g(x) - g(y)), \tag{2.4}$$

where g is a strictly increasing function. This family nests a homogeneous parametric GPL family that Patton (2020) defines as:

$$L(x, y) = (\mathbb{1}\{y \leq x\} - p) \times (\text{sgn}(x)|x|^b - \text{sgn}(y)|y|^b)/b, \; b > 0. \tag{2.5}$$

For $b = 1$, the latter equation delivers the popular "tick" loss function.

$$L(x, y) = (\mathbb{1}\{y \leq x\} - p) \times (\text{sgn}(x)|x| - \text{sgn}(y)|y|). \tag{2.6}$$

We restrict our attention to the GPL family and set $b = 0.5, 1, 2$.

While no strictly consistent loss function exists to evaluate ES forecasts on their own (Gneiting 2011a), Fissler and Ziegel (2016) provide a class of loss functions that are consistent for joint VaR and ES forecasts:

$$\begin{aligned} L(x_1, x_2, y) =& (\mathbb{1}\{y \leq x_1\} - p) \times G_1(x_1) - \mathbb{1}\{y < x_1\} \times G_1(y) \\ &+ G_2(x_2) \times \left(x_2 - x_1 + \frac{1}{p}\mathbb{1}\{y \leq x_1\}(x_1 - y) \right) \\ &- \mathcal{G}_2(x_2) + a(y), \end{aligned} \tag{2.7}$$

where $G_1, G_2, \mathcal{G}_2, a : \mathbb{R} \to \mathbb{R}, \mathcal{G}_2' = G_2$, G_1 is increasing and $\mathcal{G}_2$ is increasing and convex. L is strictly consistent if $\mathcal{G}_2$ is strictly increasing and strictly convex.

We follow Taylor (2020) and use three different parametrizations of Equation (2.7) that we present in the table below. First, a scoring function based on the asymmetric Laplace (AL) density that Taylor (2019) proposes. Second, a parametrization that Nolde and Ziegel (2017) suggest (NZ). Third, the loss function put forward by Fissler and Ziegel (2016), slightly adjusted as in Taylor (2020) (FZG).[3]

Table 2.1: Parametrizations of joint loss function

	$G_1(x)$	$G_2(x)$	$\mathcal{G}_2(x)$	$a(y)$
AL	0	$-1/x$	$-ln(-x)$	$1 - ln(1-p)$
NZ	0	$1/2(-x)^{-1/2}$	$-(-x)^{1/2}$	0
FZG	x	$\exp(x)/(1+\exp(x))$	$ln(1+\exp(x))$	0

[3]These parametrizations give different weights to the conditional quantile forecast and the forecast of the conditional mean of the truncated distribution. For a discussion of these parametrizations see e.g. Taylor (2020).

2.1 Model Confidence Set

To compare the global out-of-sample loss of competing models, we use the MCS procedure by Hansen et al. (2011). The idea of the MCS is to reduce the set of candidate models to some smaller subset that contains the model(s) with the smallest expected loss, $\mathcal{M}^*$, with a given level of confidence $1-\alpha$. Initially, the MCS testing procedure starts with a set of models $\mathcal{M}^0$, on which it then performs a sequence of tests of equal predictive ability. Once the null hypothesis of equal predictive ability cannot be rejected anymore, the testing procedure halts. It delivers a subset $\widehat{\mathcal{M}}^*_{1-\alpha}$, the so-called *model confidence set*. This subset contains the models for which the testing procedure does not find statistically significant differences in the out-of-sample loss.

Below, we briefly outline the framework of the MCS and introduce the necess notation. The initial set $\mathcal{M}^0$ consists of $m_0 \in \mathbb{N}$ competing models, $\mathcal{M} \subset \mathcal{M}^0$ denotes the set that contains models $i = 1, \ldots, m,\ m \leq m_0$. We use a strictly consistent loss function L to evaluate the point forecasts. The loss at time $t = 1, \ldots, n$ corresponding to model i is denoted $l_{t,i}$. We consider the relative performance variables

$$d_{t,ij} \equiv l_{t,i} - l_{t,j} \text{ for } i, j \in \mathcal{M}, t = 1, \ldots, n, \tag{2.8}$$

and

$$d_{t,i\cdot} \equiv l_{t,i} - \frac{1}{m}\sum_{j}^{m} l_{t,j} \text{ for } i, j \in \mathcal{M}, t = 1, \ldots, n. \tag{2.9}$$

$\mu_{ij} \equiv \mathbb{E}[d_{t,ij}]$ and $\mu_{i\cdot} \equiv \mathbb{E}[d_{t,i\cdot}]$ denote the expected loss differentials. The competing models are ranked by their expected losses: model i is *superior* over model j if $\mu_{ij} < 0$. Under the null of equal predictive ability it holds that

$$H_{0,\mathcal{M}} : \mu_{ij} = 0 \text{ for all } i, j = 1, \ldots, m, \tag{2.10}$$

or equivalently[4],

$$H_{0,\mathcal{M}} : \mu_{i\cdot} = 0 \text{ for all } i = 1, \ldots, m. \tag{2.11}$$

To determine $\mathcal{M}^*$, the MCS testing procedure uses an equivalence test $\delta_{\mathcal{M}}$ and an elimination rule $\epsilon_{\mathcal{M}}$. At any testing step, it performs the test $\delta_{\mathcal{M}}$ to test $H_{0,\mathcal{M}}$. If the test reject $H_{0,\mathcal{M}}$, the elimination rule $\epsilon_{\mathcal{M}}$ selects the model that is eliminated from the current set of candidate models $\mathcal{M}$. Algorithm 1 below summarizes the MCS testing procedure.

Algorithm 1

- *Step 0: Initially set* $\mathcal{M} = \mathcal{M}^0$.

[4] Hansen et al. (2011) show this equivalence in Section 3.1.2.

- *Step 1: Test $H_{0,\mathcal{M}}$ using an equivalence test $\delta_{\mathcal{M}}$ at level α.*
- *Step 2: If $H_{0,\mathcal{M}}$ is not rejected, set $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$.*
  Otherwise, use an elimination rule $\epsilon_{\mathcal{M}}$ to eliminate an object from $\mathcal{M}$, repeat step 1 and step 2.

We focus on the testing procedure that uses the $T_{max,\mathcal{M}}$ statistic, which is based on the loss differential between model i and the average over all models. We defer results for the testing procedure that uses the $T_{R,\mathcal{M}}$ statistic to the appendix. The $T_{R,\mathcal{M}}$ statistic is based on the loss differentials between model i and model j. In both cases, the MCS testing procedure eliminates the model that has the largest standardized excess loss as compared to its competitors.

2.1.1 Finite sample properties of the MCS

In this paper we focus on the performance of the MCS procedure by means of *potency* and *power*.

Potency is a concept that Hendry and Doornik (2014) employ in the context of model selection, and Quaedvlieg (2021) uses it to describe the performance of a multi step MCS. While related to the usual notion of size, potency is defined as the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, i.e. as the frequency with which the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$ includes the model(s) with the smallest expected loss?

In their abstract, Hansen et al. (2011) put forward the following interpretation of the MCS. "A MCS is a set of models that is constructed such that it will contain the best model with a given level of confidence. The MCS is in this sense analogous to a confidence interval for a parameter." Thus, we discuss the relationship between potency, the level of the test α used in the MCS test and the level of confidence with which the MCS supposedly contains the best model.

We first consider the relationship between potency and the strong control of the familywise error rate[5], i.e. the probability of eliminating one or more models with the smallest expected loss. The familywise error rate is bounded by the level α used at every step of the MCS testing procedure. Yet, the level of the test α has different implications if $\mathcal{M}^*$ consists of a single best model as compared to if $\mathcal{M}^*$ consists of several best models, as also stated by Hansen et al. (2011).

First, assume that $\mathcal{M}^*$ consists of a single best model, i.e. exactly one model has the smallest expected loss. Then Corollary 1 of Hansen et al. (2011) states that potency approaches 1 for $P \to \infty$, i.e. the best model is included with

[5]'Strong' implies that the control of the familywise error rate holds for any $\mathcal{M}^* \subset \mathcal{M}^0$, while 'weak' control usually refers to the case that $\mathcal{M}^* = \mathcal{M}^0$, i.e. $\mu_{ij} = 0$ for all $i, j \in \{1, \dots, m\}$ (Lehmann and Romano (2022)).

probability 1 with the out-of-sample size approaching infinity. Thus, if there is a single best model, the level of the test does not imply the interpretation that we are $1-\alpha$ confident that $\widehat{\mathcal{M}}^*_{1-\alpha}$ contains the best model.

Second, assume that $\mathcal{M}^*$ consists of several best models, i.e. two or more models have the same, smallest expected loss. Then potency is asymptotically bounded from below by $1-\alpha$, the level of confidence that the MCS contains the best models. Intuitively, if the number of observations is large enough, the MCS test eliminates all models except those with the smallest expected loss, and then performs a test at level α under the null.

In reality, we do not know if $\mathcal{M}^*$ consists of one or more of the best models. Thus, we argue that the above interpretation of the MCS - that it contains the best model with a given level of confidence - is too strong. Yet, we argue that a desirable finite sample property of the MCS test is that potency does not fall below $1-\alpha$, as we could then interpret the level of the test α as an upper bound of eliminating at least one best model.

Power does not have a unique definition (e.g. Romano and Wolf (2005) discuss different notions of power). In general, we would like to know how many of the inferior models are eliminated from the $\widehat{\mathcal{M}^*_{1-\alpha}}$, and which models are eliminated how frequently. Thus, we stick to the definition of power that Hansen et al. (2011) use in their paper, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$. A more detailed way of capturing the power property would be: which distance of the expected losses between the best model(s) and an inferior model j is associated with which rate of rejection of the null of equal predictive ability? Unfortunately, if the losses have different variances, there is no straightforward way to describe the distance from the null of EPA such that it relates directly to the testing procedure and the elimination frequencies.

Consequently, we report the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}^*_{1-\alpha}}$, the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$ and discuss individual rejection frequencies. The latter indicates how frequently model j is removed from $\widehat{\mathcal{M}}^*_{1-\alpha}$. We argue that this is the most detailed and most informative measure. It helps understand the channels that determine if the testing procedure eliminates a model from the set of candidates. Those channels include the expected losses, variances and correlations of the losses.

2.1.2 Hypotheses on the predictive ability

The econometric literature considers two different ways of stating hypotheses about predictive ability. For simplicity, assume that the forecasts stem from parametric models with parameters β. In the first case, the hypotheses make a

statement about the expected losses at population values $\beta^* = \text{plim}\,\hat{\beta}_T = \lim_{T\to\infty} \hat{\beta}$, as e.g. in Diebold and Mariano (1995), West (1996) or White (2000). As West (1996) notes, this implies that we use estimated parameters $\hat{\beta}$ to infer about the population level predictive ability. We write $\mu(\beta^*)$ to denote the expected losses at population values, and $\mu(\hat{\beta}_T)$ to denote the expected loss at the estimated parameters $\hat{\beta}_T$. Likely, it holds that $\mu(\beta^*) \neq \mu(\hat{\beta}_T)$, and the ranking of two models in terms of their expected losses could be different for $\mu(\beta^*)$ and $\mu(\hat{\beta}_T)$. In the second case, the null is formulated in terms of $\hat{\beta}_T$, i.e. it depends on the sample size and the estimation method, as e.g. in Giacomini and White (2006), Clark and McCracken (2015). For a more detailed discussion of the two concepts, see e.g. Clark and McCracken (2013).

Our approach differs from the two above in the sense that we do not estimate the parameters of the models that we use in our simulations, but use fixed parameters $\bar{\beta}$ for the following reasons. Our focus is on evaluating how the losses impact the MCS test, and not how estimation error does. The goal of the simulation is to assess if the MCS test reveals differences the expected losses, when the losses that we use for testing are sufficiently realistic for VaR and ES forecasts. If the losses that we use in the simulations are realistic is an empirical question on which we shed light in Appendix A. The goal is not to assess if one or another estimation method is superior. By fixing the parameters of the competing models, we also ensure that the simulated samples correspond to our hypotheses: we use simulated losses with an expected value of $\mu(\bar{\beta})$ to test a hypothesis about $\mu(\bar{\beta})$.[6]

2.1.3 Related simulation study on the MCS's finite sample properties

Aparicio and López de Prado (2018) employ the MCS procedure to compare different trading strategies and use excess returns as loss functions. They inquire if the MCS procedure reveals the best model within a year of trading as a function of Sharpe ratios. Yet, they find that the associated Sharpe ratios are unrealistically large. They also find that the models selected by the MCS show significantly worse out-of-sample performance than in-sample performance and thus suffer from backtest overfitting. Yet, as the authors note, they assume *independent* strategies though correlations likely occur in practice, and although larger correlations lead to an easier distinction between models given the variances are held constant.

[6] Theoretically, we could try to formulate our hypothesis using $\mu(\hat{\beta}_T)$. Yet, the computational cost would be extreme, the results could be unstable and they could depend on parameters such as the starting values of our models.

3 Simulation

In this section, we conduct a simulation study to infer if the MCS reveals differences in predictive ability within a set of candidate models.

3.1 Simulation Design

This section describes the parameters of our simulations such as the DGP, competing forecasts and out-of-sample sizes.

3.1.1 DGPs

As DGP, we choose a process with zero conditional mean but with time varying variance that follows a TGARCH(1,1) (or GJR-GARCH) with standardized Student's t distributed innovations:

$$r_t = z_t\sqrt{\sigma_t^2} \tag{3.1}$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\mathbb{1}_{\{\varepsilon_t<0\}}\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{3.2}$$

$$z_t \sim sst_\nu, \tag{3.3}$$

where ν denotes the degrees of freedom, and sst_ν denotes the scaled Student's t distribution with unit variance.[7]

$$\sigma^2 = \frac{1}{1-\alpha-0.5\gamma-\beta} \tag{3.4}$$

denotes the unconditional variance of the returns.

The DGP captures stylized facts of financial returns such as fat tails, time-varying volatility, clustering and the leverage effect. The leverage effect is captured through the parameter γ. We consider three DGPs with different degrees of freedom $\nu \in \{3, 7, 12\}$. The motivation is that each DGP corresponds to a different state of the market, i.e. highly volatile, volatile and calm, respectively. We choose the model parameters close to parameters that we frequently observe on daily log-returns from closing prices adjusted for dividend and split. Specifically, we estimate the TGARCH on equally weighted indices composed of 30 randomly chosen stocks from the Dow Jones U.S. Small, Mid, and Large cap-indices.[8] The DGPs use the following parameters:

[7] If $T \sim t_\nu$, $\nu > 2$ then for the random variable $Z = \sqrt{\frac{\nu-2}{\nu}}T$ with unit variance we write $Z \sim sst_\nu$.

[8] We describe the data in more detail in Appendix A.

Table 3.1: Parameters DGPs

ω	α	γ	β	ν
0.04	0.03	0.1	0.9	$\{3, 7, 12\}$

Notes: This table displays the parameters of the DGPs (TGARCH-t, Equation 3.3) for the simulations that relate to the results in Section 3.2.

3.1.2 Other simulation parameters

We evaluate one-day-ahead daily forecasts for VaR and ES at quantile levels $p \in \{0.01, 0.025, 0.05, 0.1\}$ and consider six out-of-samples sizes $P \in \{63, 126, 251, 500, 1000, 2500\}$. These sample sizes correspond to one quarter, half a year, one year, two, four and ten years of daily observations, respectively. For the shorter out-of-sample sizes $P \leq 500$ and the smaller quantile levels $p \in \{0.01, 0.025\}$, we expect few returns that are smaller than the VaR forecasts, e.g. about 12.5 for an out-of-sample period of 500 observations and VaR forecast at quantile level 0.025. These are comparatively few observations to evaluate if a model's forecast of the conditional quantile or ES. Hence, we do not expect that the MCS test reveals the model with the smallest expected loss. Yet, we examine if the rejections that occur correctly indicate models that have larger expected losses. We set the number of bootstrap resamples in the MCS to $B = 5,000$ and perform $2,500$ Monte Carlo simulations for each DGP.

3.1.3 Sets of competing models

This section characterizes the two sets of competing models in our simulations. We use GARCH-type models due to their popularity to model financial returns (e.g. Li and Patton (2018), Patton (2020)).

In this paragraph, we characterize the first set of $m = 5$ competing models. The first model has a TGARCH conditional variance specification and sst_ν-distributed innovations. We refer to this model as the 'true' or 'correctly specified' model. Additionally, we consider the following four misspecified models that do not correctly model the return process. First, a TGARCH(1,1) with $z_t \sim \mathcal{N}(0, 1)$ (TGARCH-n). Second, a GARCH(1,1) with $z_t \sim sst_\nu$ (GARCH-t). Third, a GARCH(1,1) with $z_t \sim \mathcal{N}(0, 1)$. Fourth, a constant variance model with $z_t \sim st_\nu(0, \sigma^2)$ (constant variance model). The parameters of all models are set such that all models have the same unconditional variance σ^2 (Equation 3.4).

The second set of $m = 10$ competing forecasts is constructed by adding five models to the first set. These models are three Risk Metrics^{TM} models with $z \sim sst_{\nu'}$, $\nu' \in \{3, 7, 12\}$, and two additional GARCH(1,1) with $z \sim sst_{\nu'}$, $\nu' \in \{3, 7, 12\} \setminus \{\nu\}$. Depending on the DGP, we add models that show both large and

small expected losses relative to the initial models.

We examine the larger set of models for the following reasons: (1) including additional models may lead to fewer eliminations of inferior models as follows. Let $\mathcal{M}^0_{small}$ denote the initial set of competing models and $\mathcal{M}^0_{large} \supset \mathcal{M}^0_{small}$ the larger set of models. The control of the familywise error rate is achieved through adjusted p-values that monotonically increase with each testing step. Consequently, eliminations become less likely with each testing step. If we add models that have comparatively large expected losses to $\mathcal{M}^0_{small}$, we may thus observe that the MCS test eliminates inferior models from $\mathcal{M}^0_{large}$ less frequently than it eliminates them from $\mathcal{M}^0_{small}$. (2) including models with comparatively small expected losses may facilitate identifying the worst models. Assume that model $i = 1$ is inferior to models $j \in \{2, \dots, m\}$. If we increase m, we expect $var(D_{i\cdot})$ to decrease (cf. Equation D.2 in Appendix D.1) and $\bar{d}_{i\cdot}$ to become smaller. As model i is inferior, including more alternatives with smaller expected losses than model i leads to a larger t-statistic $t_{i\cdot} = \dfrac{\bar{d}_{i\cdot}}{\sqrt{\hat{var}(\bar{d}_{i\cdot})}}$ and we thus expect that the MCS test eliminates model i more frequently when we increase m.

3.1.4 Expected losses and correlations

Appendix B presents the results on expected losses and correlation matrices.

In general, it is not straightforward to characterize the size of the expected losses and the expected loss differentials. First, the individual losses $l_{t,i}$ associated with model i live on different scales as they have different variances in our simulations. This is opposed to the simulations that Hansen et al. (2011) perform to examine the finite sample properties of the MCS test, where all losses have the same variance. Yet, we observe empirically that the losses of different models have different variances (see Appendix A.0.1). Second, scaling the losses is not informative as we test about the absolute difference in expectation, i.e. model i is superior to model j if $\mu_{ij} < 0$ regardless the variances of the two models. Third, scaling the expected loss differentials μ_{ij} by the variance of $D_{t,ij}$ may help understand the first elimination. Yet, the scaled loss which is most relevant in each subsequent testing step depends on the testing path, i.e. it depends on which model(s) the MCS test eliminates in the previous step.

Instead, we may compare the expected loss differentials to those that we observe empirically in absolute values or relative to the model with the smallest average or expected loss. In our simulations, the size of the expected losses of misspecified models varies between 101 % and 110% of the expected loss of the true model. These ratios correspond to the ratios of the average sample losses in

Bernardi et al. (2017), Taylor (2020) and Dimitriadis and Halbleib (2022). We stress that'very similar' expected losses do not need to stem from 'very similar' forecasts: consider e.g. example 1 in Section 3.3. In this example, we illustrate how the tick loss function evaluates VaR forecast of two models if one model consistently makes forecasts that are closer to zero than the forecasts of the other model.

To illustrate how correlations affect the MCS test, assume that we compare only two models i and j and that the variances are fixed. It holds that $var(D_{t,ij}) = var(L_{t,i}) + var(L_{t,j}) - 2cov(L_{t,i}, L_{t,j})$. Thus, the higher the correlations between the losses of model i and j, the smaller the variances of the loss differential $D_{t,ij}$. The MCS test uses the t-statistic $t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\hat{var}(\bar{d}_{ij})}}$. Hence, high correlations lead to larger t-statistics, thus more rejections and imply that the MCS more easily detects differences in the expected losses.[9]

In our simulations, we observe correlations between 0.95 and 0.99 for the GARCH-type models, and correlations between 0.8 and 0.95 between the constant variance model and the GARCH-type models (see Appendix B.2). These correlations are in line with the correlations among the GARCH-type models that we estimate on the Dow Jones Indices data (see Appendix A.0.2).

3.2 Results

This section provides the results of Monte Carlo experiments that examine the finite sample properties of the MCS testing procedure. We discuss potency, power and individual rejection frequencies as a function of the quantile p, the out-of-sample size P, the number of models m, and the market conditions captured by the degrees of freedom ν of the DGP. We restrict the discussion to the level of the test $\alpha = 0.25$. We do not find strong differences between different parametrizations of the loss functions that we define in Equation 2.4 and Equation 2.7. In this section, we thus discuss results for the tick loss function (Equation 2.6) when evaluating VaR forecasts, and for the AL score (cf. Table 2.1) when evaluating joint VaR and ES forecast. Appendix C presents the results for alternative specifications such as the test statistic $T_{R,\mathcal{M}}$, different levels of the test α and different parametrizations of the loss functions.

[9] Hansen et al. (2011) also discuss this example when all losses have unit variance.

3.2.1 Results for the first set of competing models: $m = 5$.

Figure 3.1 below visualizes the VaR results. It shows potency in the upper row (Panels A, B, C) and power in the lower row (Panels D, E, F). The columns refer to the three choices of DGPs with different degrees of freedom $\nu \in \{3, 7, 12\}$.

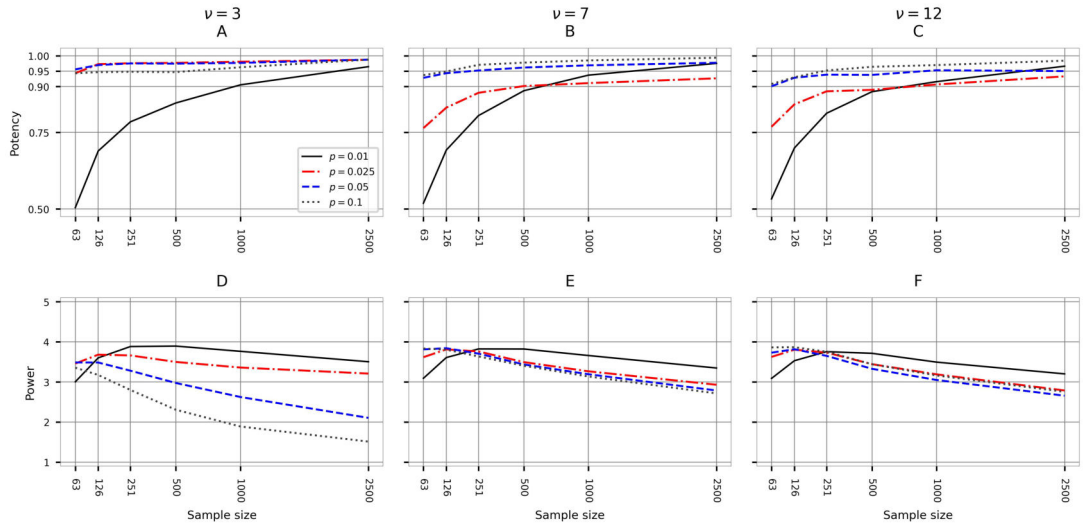

Figure 3.1: Finite sample properties for the tick loss

This figure displays the finite sample properties of the MCS procedure in the following setting: one-day-ahead VaR forecasts evaluated using the tick loss, true model included, number of models $m = 5$, level of the test $\alpha = 0.25$. The upper row displays the potency, i.e. the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$. The lower row displays the power property, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Panels A, B, C reveal that the MCS usually keeps the best model more often than the specified confidence level of 75% (1-α) even for the smallest out-of-sample size $P = 63$. Potency then increases with the quantile level p and with the out-of-sample size P. Its values are much lower for the extreme quantile level $p = 0.01$ (solid dark grey line) and out-of-sample sizes $P \leq 500$ than for the other quantiles $p \in \{0.025, 0.05, 0.1\}$ (red dash-dotted line, blue dashed line and grey dotted line, respectively). At the most extreme quantile level $p = 0.01$ (solid dark grey line), potency is below the specific confidence level $1 - \alpha$ for shorter out-of-sample sizes $P \leq 126$. Starting from out of-sample sizes $P \geq 251$, i.e. one year of daily data, the MCS test keeps the best model in around 80% of our simulations. For out-of-sample size $P = 1000$, potency increases to 90% for the DGPs with $\nu \in \{3, 12\}$ degrees of freedom (Panels A, C), and to 95% for the DGP with $\nu = 7$ degrees of freedom (Panel B).

At the quantile level $p = 0.025$, potency but still above the confidence level $1 - \alpha$: it increases from around 75% to around 85% from 63 observations to 500 observations for the DGPs with $\nu \in \{7, 12\}$, while it is above 90% for alle

out-of-sample sizes for volatile markets (Panel A). At the larger quantile levels $p \in \{0.05, 0.1\}$, the MCS keeps the best model more often than 90% for all out-of-sample sizes P.

To summarize, our findings suggest that - in comparable settings - potency of the MCS test at level $\alpha = 0.25$ exceeds the confidence level of 75% for half a year of daily data when we consider less extreme quantile levels $p \geq 0.025$, while the extreme quantile level $p = 0.01$ requires one year or more daily data.

Panels D, E and F in Figure 3.1 display the power as the average number of models in the MCS. In general, the power increases with the quantile level p and out-of-sample size P.

At the most extreme quantile level $p = 0.01$, we find the same patterns for all three DGPs, i.e. across the degrees of freedom ν: power decreases from 63 out-of sample observations to 251 out-of-sample observations and then increases with the out-of-sample size P. Simultaneously, we observe a steep rise in potency (Panels A, B, C) for out-of-sample sizes $P \leq 251$, and thus suspect that the MCS test is not reliable for this quantile level and out-of-sample sizes $P < 251$ (See Section 3.3 for a more detailed discussion).

For the DGP with $\nu = 3$ degrees of freedom (Panel A), we observe the most disperse measures of power across different quantile levels. Starting at 126 observations, i.e. half a year of data, power increases monotonically with the sample size for the quantile levels $p \in \{0.05, 0.1\}$. While for a sample size of 126, the MCS test keeps between 3.6 and 3.2 models for VaR forecasts at quantile level $p = 0.025$ and $p = 0.1$, respectively, it only keeps 3.5 and 2.2 models for a sample size of 500 observations, respectively.

In the settings that relate to calmer market conditions, i.e. when $\nu \in \{7, 12\}$, and for quantile levels $p \in \{0.05, 0.1\}$, power displays a homogeneous pattern: the number of models that the MCS test eliminates increases slightly with the out-of-sample size P. It eliminates 1 model for an out-of-sample size of $P = 126$, and between 1 and 1.6 models for a sample size of $P = 500$.

To summarize, we conjecture that one may use the MCS test to trim off some worst VaR models, but without expecting that the MCS test reveals the single best model.

Figure 3.2 presents potency and power of the MCS when we use the AL score to evaluate joint forecasts of VaR and ES - analogously to Figure 3.1 for VaR forecasts.

The potency at quantile levels $p \in \{0.05, 0.1\}$ is below the confidence level of 75% for 63 out-of-sample observations, while at the quantile level $p = 0.01$ it is

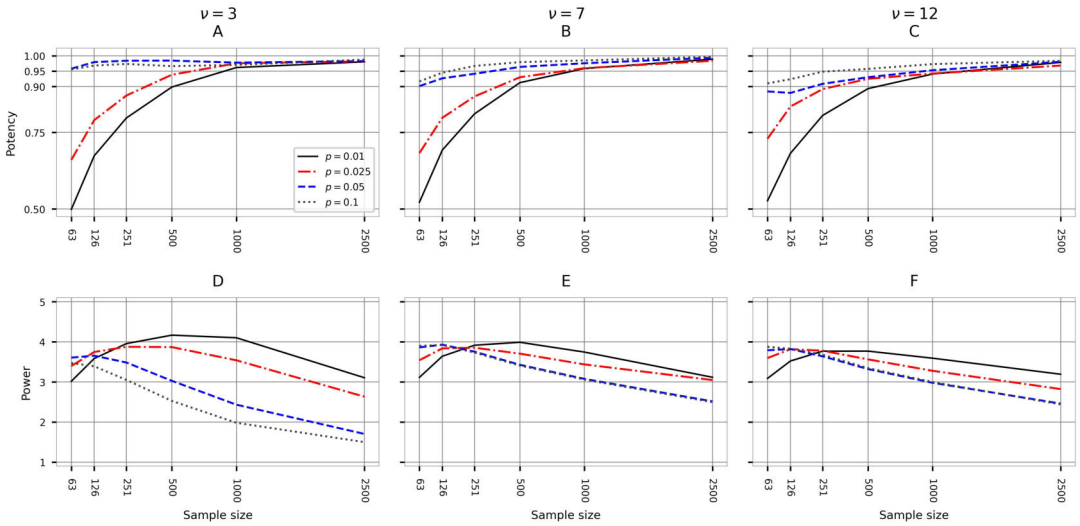

Figure 3.2: Finite sample properties for the AL score

This figure displays the finite sample properties of the MCS procedure in the following setting: one-day-ahead VaR and ES forecasts evaluated using the AL score, true model included, number of models $m = 5$, level of the test $\alpha = 0.25$. The upper panel displays the potency, i.e. the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$. The lower panel displays the power property, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

even below 75% for an out-of-sample size of 126 observations. For out-of-sample sizes $P \leq 500$, the MCS test consistently keeps the best model more frequently at the quantile level $p = 0.025$ than at the quantile level $p = 0.01$. This is analogous to the pattern which we observe for the tick loss in Figure 3.1 (Panels A, B, C).

For quantile levels $p \in \{0.05, 0.1\}$, potency is well above the confidence level of 75%. Panel A, which corresponds to the DGP with $\nu = 3$ degrees of freedom, e.g. shows that its values across all out-of-sample sizes exceed 95%.

The power is shown in Panels D, E, F of Figure 3.2. Our findings are very similar to those shown in Figure 3.1 (Panels D, E, F) above for VaR forecasts. Yet, overall power is slightly lower.

In the settings that relate to calmer market conditions, i.e. for $\nu \in \{7, 12\}$ degrees of freedom, and joint VaR and ES forecast at quantile level $p \in \{0.025, 0.05, 0.1\}$, we find that power increases slowly with the out-of-sample size P ($P \geq 126$) and the quantile level p: the MCS eliminates 1 model for 126 out-of-sample observations and between 1.4 and 1.6 models for an out-of-sample size of two years. For forecasts at the quantile level $p = 0.01$ (solid dark grey line), Panels D, E, F show that power increases with the out-of-sample size P only for $P \geq 251$, i.e. for more than one year of daily data.

To sum up, we find that the discriminatory power of the MCS increases with the quantile level p. Yet, as when we evaluate VaR forecasts, the MCS test does not reveal the best model for the out-of-sample sizes that we consider in these

simulations.

Figure 3.3 below show hows often the MCS test eliminates each model, to which we refer as individual rejection frequencies. The left and right panel provides individual rejection frequencies for standalone VaR forecasts and joint VaR and ES forecast, respectively. The results are for the DGP with $\nu = 3$ degrees of freedom and level of the test $\alpha = 0.25$.

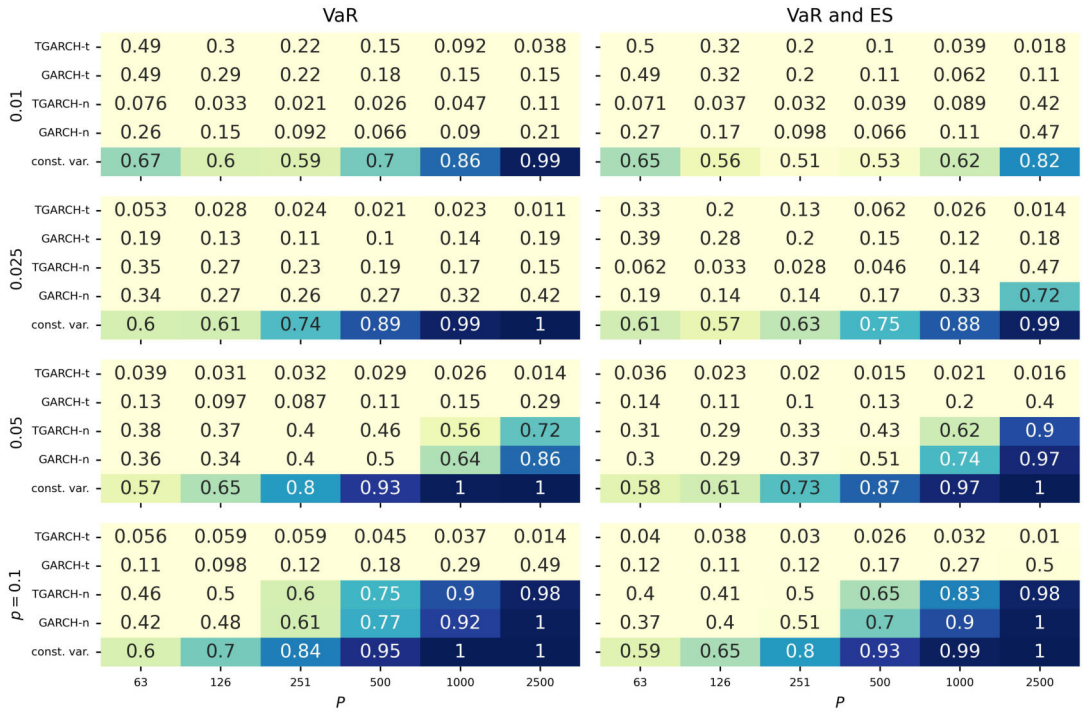

Figure 3.3: Individual rejection frequencies, $\alpha = 0.25$

This figure displays the individual rejection frequencies of the candidate models in the MCS, number of models $m = 5$, level of the test $\alpha = 0.25$. DGP with $\nu = 3$ degrees of freedom.

At the most extreme quantile level $p = 0.01$, the MCS test eliminates the constant variance model - which has by far the largest expected loss - most frequently. Starting from 251 observations, the rejection frequency increases monotonically with the number of observations until it reaches almost 1 for 2500 observations. We conclude that the MCS test reliably detects that this naive model has a larger expected loss than its competitors.

Yet, at the quantile level $p = 0.01$ we also observe patterns in the individual rejection frequencies that are not consistent with the ranking of the models in terms of the expected losses.

First, all individual rejection frequencies decrease from 63 to 251 out-of-sample observations. While Figure 3.2 reveals that some rejection frequencies have to decrease, as the number of models in the MCS increases, it is remarkable that

the MCS test eliminates *all* models less frequently when the out-of-sample size increases. Instead, we would like to observe this pattern for the best model only. As larger out-of-sample sizes provide a more reliable ranking of the competing forecasters, the MCS test should keep the model with the smallest expected loss more frequently, and simultaneously eliminate models with larger expected losses more frequently.

Second, for sample sizes of up to 4 years, i.e. 1000 observations, the MCS test eliminates the true model and the GARCH-t more frequently than the TGARCH-n and GARCH-n, although the latter models have larger expected losses. This discrepancy between the expected losses and the individual rejection frequencies decreases with the out-of-sample size P.

We suspect that our findings stem from imprecise estimates of the expected losses - see Figure D.3 - and the finite sample behavior of the tick loss, which we illustrate in Example 1 in Section 3.3 below.

Furthermore, the rejection frequencies of TGARCH-n and GARCH-n are much higher when we consider joint forecasts instead of VaR forecasts for an out-of-sample size of 2500. We conclude that for this out-of-sample size, the AL score and the MCS test can exploit information about the shape of the tail to identify a model that is too optimistic and underestimates the scale of financial losses when they exceed the respective VaR. We conjecture that statistical losses and the MCS test are not appropriate tools to detect VaR models that are too optimistic at the quantile level $p = 0.01$.

At the quantile levels $p \in \{0.025, 0.05, 0.1\}$, however, the MCS test consistently eliminates models with larger expected losses more frequently than models with smaller expected losses.

Consequently, the constant variance model is eliminated most frequently, followed by the GARCH-n, which has the second largest expected loss.

Yet, the MCS test shows little power against the second and third best GARCH-t and TGARCH-n model, respectively, for sample sizes of up to $P = 500$. On average, one of them becomes a candidate for elimination only once the MCS test eliminates the two other inferior models that have larger expected losses. Due to the control of the familywise error rate, such an elimination requires much stronger evidence than previous eliminations.

Interestingly, we observe more rejections for the tick loss than for the AL score for out-of-sample sizes of up to 500 observations, which reverses for larger out-of-sample sizes. Then, information about the tail of the conditional return distribution becomes valuable to identify the GARCH-n as inferior. We thus recommend to use joint forecast evaluation for sample sizes larger than and including

two years of data.

3.2.2 Second set of models - $m = 10$.

Next, we discuss potency, power and individual rejection frequencies for the larger set of models, which we describe in Section 3.1.3. Figure 3.4 presents potency and power for VaR forecasts, while we provide results for joint forecasts in Appendix C.3.

Potency - Panels A, B, C of Figure 3.4 - increases with the quantile level p and the out-of-sample size P. For the quantile levels $p \in \{0.025, 0.05, 0.1\}$, potency exceeds the confidence level of 75% even for the smallest out of sample size $P = 63$. For the most extreme quantile level $p = 0.01$, potency exceeds 75% starting from $P = 126$ observations.

Potency takes on values between 95% and 99% for out-of-sample sizes $P \geq 500$, i.e. starting from two years of daily data. For this larger set of $m = 10$ models, the MCS test keeps the model with the smallest expected loss more often than for the smaller set of $m = 5$ models (cf. Figure 3.1).

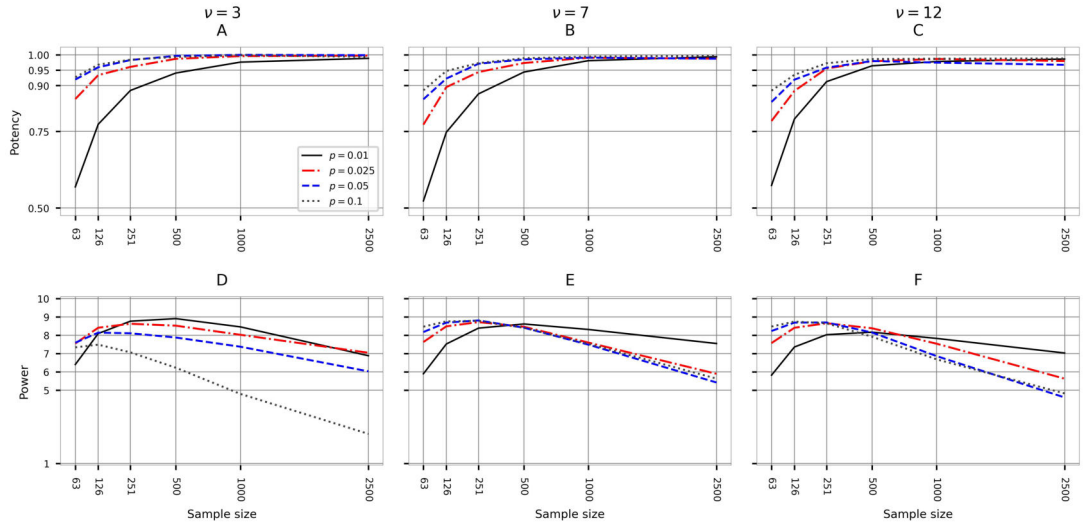

Figure 3.4: Finite sample properties for the tick loss

This figure displays the finite sample properties of the MCS procedure in the following setting: one-day-ahead VaR forecasts evaluated using the tick loss, true model included, number of models $m = 10$, level of the test $\alpha = 0.25$. The upper panel displays the potency, i.e. the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$. The lower panel displays the power property, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

The power - Panels D, E, F of Figure 3.4 - behaves similarly as in the case with $m = 5$ models depicted in Figure 3.1 above. We observe that the number of models in the MCS decreases starting from 126 out-of-sample observations for the quantile levels $p \in \{0.025, 0.05, 0.1\}$. For the most extreme quantile level

$p = 0.01$, the number of models in the MCS decreases for out-of-sample sizes $P \geq 251$. For the DGPs that relate to calmer markets conditions - panels E and F - the MCS test eliminates between 1.5 and 2 models for out-of-samples sizes of $P = 500$ and across all quantile levels p.

We find more variation in panel A which corresponds to the DGP with $\nu = 3$ degrees of freedom. For an out-of-sample size $P = 500$ and quantile levels $p \in \{0.01, 0.025, 0.05\}$, the MCS test eliminates 1, 1.5 and 2 models, respectively.

Overall, the MCS test still eliminates the constant variance model most often, though between 5-10% less frequently than for the set of $m = 5$ candidate models. At the quantile levels $p \in \{0.01, 0.025\}$, no strong pattern emerges among the other inferior models, which all show elimination frequencies of less than 10%.

At the larger quantile levels $p \in \{0.05, 0.1\}$, the models that the MCS test eliminates most frequently after the constant variance model, are the GARCH-n, TGARCH-n and GARCH-t_{12}. We note, however, that the elimination frequencies of the GARCH-n and the TGARCH-n are between 11% and 20% smaller for the larger set of $m = 10$ models as compared to the smaller set of $m = 5$ models (cf. Figure 3.3 above). To summarize, the MCS test has less power against inferior alternatives when we enlarge the set of competing models.

3.3 Detailed discussion

To explain our findings, we try to disentangle the properties of the MCS test and the loss functions. We narrow our focus to the setting with $m = 5$ models, $\nu = 3$ degrees of freedom and level of the test $\alpha = 0.25$. If we do not explicitly state otherwise, we refer to VaR forecasts that we evaluate using the tick loss.

First, we examine if the differences in predictive ability are large enough such that the MCS reveals them if the losses are symmetric. Thus, we directly simulate i.i.d. losses from a multivariate normal distribution $\mathcal{N} \sim (\mu, \Sigma)$ for which we set μ to the simulated mean and Σ to the simulated covariance matrix of the losses $l_{t,i}$ that we use in our initial simulations in Section 3.2.1.[10]

We compare the properties of the MCS test when we simulate normally distributed losses to its properties that we observe in our initial simulations in 3.2.1. This informs us how the finite sample properties of the realistic losses that we simulate affect the MCS test. This paragraph summarizes the results while Ap-

[10]For normally distributed data, we know that $t_{ij} = \dfrac{\bar{d}_{ij}}{\sqrt{\hat{var}(\bar{d}_{ij})}}$ follows a Student t distribution with $\nu = P - 1$ degrees of freedom, where P denotes the number of observations. In particular, the Student's t distribution is centered and approximately normal for sample sizes $P \geq 120$.

pendix D.2 provides the results tables when we simulate normally distributed losses.

When we simulate normally distributed losses, we still observe few eliminations of inferior models but higher potency. Furthermore, we observe that the individual rejection frequencies consistently increase with the expected loss μ_i. This implies that the rejections that we observe are informative: the MCS test uniformly eliminates inferior models more frequently than superior ones - which differs from our initial simulations in Section 3.2.1 at the quantile level $p = 0.01$.

Consequently, when we use the GPL or joint losses - in particular for the extreme quantile level $p = 0.01$ - we suspect that we obtain comparatively poor estimates $\bar{d}_{ij}, \bar{d}_{i\cdot}$ and their respective variance estimates $\widehat{var}(\bar{d}_{ij}), \widehat{var}(\bar{d}_{i\cdot})$

We thus proceed to examine those in more detail. Example 1 illustrates how the asymmetric loss functions may affect finite sample estimates of the mean.

Example 1 *Consider the tick loss defined in Equation 2.4 at quantile level* $p = 0.01$*. We compare the true, superior TGARCH-t (model* i*) and the inferior TGARCH-n (model* j*) that assumes normally distributed innovations* z_t*. Let* x_i *and* x_j *denote the VaR forecasts that stem from models* i *and* j*, respectively, while* y *denotes the daily financial return. We have* $x_i < x_j$ *for* $\nu \in \{3, 7, 12\}$ *and thus* $d_{ij}|x_j < y = p \times ((y - x_i) - (y - x_j)) = p \times (x_j - x_i) > 0$*. Each time the financial return is above the VaR forecast from the TGARCH-n, the loss associated with the TGARCH-n is smaller than the loss of the more conservative TGARCH-t. Consequently, we rely on sufficiently many returns that are smaller than the VaR forecast from the TGARCH-n to correctly rank the two models. Yet, such observations are few at level* $p = 0.01$ *and for sample sizes* $P \leq 500$*.*

Thus, we first consider how frequently the simulated evaluation samples correctly indicate the relative predictive ability between model i and j, i.e. how frequently do we observe that $\bar{d}_{ij} < 0$. Figures D.3 and D.4 in Appendix D.4 display $\frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\bar{d}_{ij} < 0\}$ across different levels of p and out-of-sample sizes P for the tick loss and joint AL loss, respectively, while we summarize the results below.

First, the constant variance model is consistently indicated as inferior. Seceond, when we evaluate VaR forecasts at the more extreme quantile levels $p \in \{0.01, 0.025\}$, many short samples $P \leq 500$ fail to provide evidence against the two models, TGARCH-n and GARCH-n, that are consistently too optimistic and underestimate risk. Third, evaluating joint VaR and ES forecasts helps providing evidence against too optimistic models for evaluation sample sizes $P \geq 500$ and $P \geq 251$, respectively, if we consider forecasts at quantile level $p = 0.01$ and $p = 0.025$, respectively.

However, wrongly estimating the sign of the expected loss differential - i.e. $\text{sgn}(\bar{d_{ij}}) \neq \text{sgn}(\mu_{ij})$ - is acceptable, if the MCS test does not indicate that $\bar{d_{ij}}$ is significant too often. If it does, the MCS test wrongly eliminates a model with a smaller expected loss but keeps a model with a larger expected loss. When we wrongly estimate $\text{sgn}(\bar{d_{ij}})$, few eliminations are desirable, as the MCS test then acknowledges that the data lack information (Hansen et al. 2011) and keeps most of the initial competing models.

Yet, we observe for $p = 0.01$ and the shorter out-of-sample sizes $P \leq 500$ that the rejection frequencies are not consistent with the expected losses: the true, superior model TGARCH-t is eliminated from the MCS more frequently than the inferior models TGARCH-n and GARCH-n. Furthermore, the MCS test using $T_{R,\mathcal{M}}$ (see Appendix C.1.1) eliminates the true model much more frequently than α, when we consider the quantile level $p = 0.01$ and small sample sizes $P \leq 251$.

Consequently, we assess the variance estimates of $\bar{d_{ij}}$ and $\bar{d}_{i\cdot}$ as an explanatory channel. To this end, we examine the bootstrapped estimates of $\widehat{var}(\bar{d}_{i\cdot})$ and the values of $var(\bar{d}_{i\cdot})$ that we obtain from simulating the covariance matrix of $\bar{l}$ for different sample sizes P and then using an analytical expression to calculate $var(\bar{d}_{i\cdot})$ (Equation D.2 in Appendix D.1). We refer to $var(\bar{d}_{i\cdot})$ as the simulated variance and consider it the true value. Yet, our simulations yield unstable results for forecasts at the extreme quantile level $p = 0.01$, which implies that the results based on the simulated variance are less reliable for this quantile level than for the others.

Histograms of $\widehat{var}(\bar{d}_{i\cdot})$ reveal considerable skewness in the variance estimate. While $\bar{d}_{i\cdot}$ seems to follow a normal distribution and is reasonably centered around its mean, $t_{i\cdot} = \dfrac{\bar{d}_{i\cdot}}{\sqrt{\widehat{var}(\bar{d}_{i\cdot})}}$ does not. We provide histograms of $t_{i\cdot}$ in Appendix D.3. The mechanism is that the distribution of $\widehat{var}(\bar{d}_{i\cdot})$ is severely skewed, we observe that the median and the mode are consistently smaller than the mean. Additionally, the mean of $\widehat{var}(\bar{d}_{i\cdot})$ is much smaller than its simulated equivalent, i.e. we systematically underestimate $var(\bar{d}_{i\cdot})$. The results are worse for extreme values of the quantile level such as $p \in \{0.01, 0.025\}$ and the shorter out-of-sample sizes P.

To assess if systematically underestimating $var(\bar{d}_{i\cdot})$ affects how often we reject both the best model and models with larger expected losses, we perform the initial Monte Carlo simulations from Section 3.2.1 but use the simulated values of $var(\bar{d}_{i\cdot})$ - which we consider correct.

Figure D.5 in Appendix D.5 provides the differences in the individual rejection frequencies when using the bootstrapped estimate of $\widehat{var}(\bar{d}_{i\cdot})$ and when using the

simulated values of $var(\bar{d}_{i\cdot})$, respectively. We summarize the results below.

First, underestimating the variance hardly affects how often the MCS test eliminates the model with the largest expected loss, which is the constant variance model.[11]

Second, for the larger quantile levels $p \in \{0.025, 0.05, 0.1\}$, underestimating the variance explains a relevant share of the rejection that we observe for the GARCH-n model. The GARCH-n has the second largest expected loss, but its forecasts are not as obviously worse than its competitors' as the constant variance model's forecasts.

Third, the inconsistencies between the rejection frequencies and the expected losses that we observe for the quantile level $p = 0.01$ stem mostly from wrongly estimating μ_{ij} and less so from underestimating the variance of $\bar{d}_{i\cdot}$.

4 Conclusion

This paper sheds light on the finite sample properties of the model confidence set (MCS) testing procedure by Hansen et al. (2011) applied to out-of-sample VaR and ES forecasts evaluated using asymmetric statistical loss functions.

In our comprehensive simulation results, first, we find that the MCS test usually keeps the model with the smallest expected loss more frequently than the specified confidence level $1-\alpha$. Second, it eliminates only a few models that have larger expected losses than the best model for out-of-sample sizes of up to four years. Third, the MCS test shows little power against models that underestimate tail risk at the extreme quantile levels $p = 0.01$ and $p = 0.025$, while the power increases with the quantile level p. Moreover, for sample sizes of two years and more of daily data, the MCS test displays more power when evaluating joint VaR and ES forecasts as compared to standalone VaR forecasts. Overall, we conclude that the MCS test needs long evaluation windows of several years to make reliable inference about the predictive ability of competing models.

For empirical studies, our findings imply that the MCS test may be an appropriate tool to select a subset of models for model averaging, but that it will not reveal a single best model. For instance, using the MCS test to infer if a new approach to model VaR and ES produces forecasts that are superior to existing alternatives' is likely uninformative due to the lack of power.

The distinct parametrizations of the generalized piecewise linear (GPL) loss to

[11] In this case, the data provide strong evidence that its associated loss is significantly larger than its competitors'. If the MCS test uses the simulated variance, it still indicates that the difference in the losses is significant.

evaluate VaR forecasts display no remarkable differences in potency and power. We neither find a uniform pattern between the different parametrizations of the loss functions to evaluate joint VaR and ES forecasts.

Acknowledgements

The author is grateful to Christian Gouriéroux, Roxana Halbleib, Ekaterina Kazak, Michael Massmann and Winfried Pohlmeier for helpful comments. All remaining errors are my own. This work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

References

Aparicio, Diego and Marcos López de Prado (2018). "How hard is it to pick the right model? MCS and backtest overfitting". In: *Algorithmic Finance* 7.1-2, pp. 53–61. ISSN: 21585571.

Bernardi, Mauro, Leopoldo Catania, and Lea Petrella (2017). "Are news important to predict the Value-at-Risk?" In: *The European Journal of Finance* 23.6, pp. 535–572. ISSN: 1351-847X.

Clark, Todd and Michael McCracken (2013). "Advances in Forecast Evaluation". In: vol. 2. Handbook of Economic Forecasting. Elsevier, pp. 1107–1201. ISBN: 9780444627315.

— (2015). "Nested forecast model comparisons: a new approach to testing equal accuracy". In: *Journal of Econometrics* 186.1, pp. 160–177. ISSN: 03044076.

Diebold, Francis and Roberto Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263. ISSN: 0735-0015.

Dimitriadis, Timo and Roxana Halbleib (2022). "Realized Quantiles". In: *Journal of Business & Economic Statistics* 40.3, pp. 1346–1361. ISSN: 0735-0015.

Fissler, Tobias and Johanna F. Ziegel (2016). "Higher order elicitability and Osband's principle". In: *The Annals of Statistics* 44.4, pp. 1680–1707. ISSN: 0090-5364.

Giacomini, Raffaella and Halbert White (2006). "Tests of Conditional Predictive Ability". In: *Econometrica* 74.6, pp. 1545–1578. ISSN: 0012-9682.

Gneiting, Tilmann (2011a). "Making and evaluating point forecasts". In: *Journal of the American Statistical Association* 106.494, pp. 746–762. ISSN: 0162-1459.

— (2011b). "Quantiles as optimal point forecasts". In: *International Journal of Forecasting* 27.2, pp. 197–207. ISSN: 01692070.

Hansen, Peter R. and Asger Lunde (2005). "A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?" In: *Journal of applied econometrics* 20.7, pp. 873–889. ISSN: 0883-7252.

Hansen, Peter R., Asger Lunde, and James M. Nason (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497. ISSN: 0012-9682.

Hendry, David F. and Jurgen A. Doornik (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. The MIT Press. ISBN: 9780262028356.

Lehmann, E. L. and Joseph P. Romano (2022). *Testing Statistical Hypotheses*. Cham: Springer International Publishing. ISBN: 978-3-030-70577-0.

Li, Jia and Andrew J. Patton (2018). "Asymptotic inference about predictive accuracy using high frequency data". In: *Journal of Econometrics* 203.2, pp. 223–240. ISSN: 03044076.

Nolde, Natalia and Johanna F. Ziegel (2017). "Elicitability and backtesting: Perspectives for banking regulation". In: *The Annals of Applied Statistics* 11.4, pp. 1833–1874, 42.

Patton, Andrew J. (2020). "Comparing Possibly Misspecified Forecasts". In: *Journal of Business & Economic Statistics* 38.4, pp. 796–809. ISSN: 0735-0015.

Quaedvlieg, Rogier (2021). "Multi-Horizon Forecast Comparison". In: *Journal of Business & Economic Statistics* 39.1, pp. 40–53. ISSN: 0735-0015.

Romano, Joseph P. and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping". In: *Econometrica* 73.4, pp. 1237–1282. ISSN: 0012-9682.

Taylor, James W. (2019). "Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution". In: *Journal of Business & Economic Statistics* 37.1, pp. 121–133. ISSN: 0735-0015.

— (2020). "Forecast combinations for value at risk and expected shortfall". In: *International Journal of Forecasting* 36.2, pp. 428–441. ISSN: 01692070.

West, Kenneth D. (1996). "Asymptotic inference about predictive ability". In: *Econometrica*, pp. 1067–1084. ISSN: 0012-9682.

White, Halbert (2000). "A reality check for data snooping". In: *Econometrica* 68.5, pp. 1097–1126. ISSN: 0012-9682.

A Data

The data on which we estimate the parameters of the DGPs in Section 3.1.1 consist of three equally weighted indices composed of 30 randomly chosen stocks from the Dow Jones U.S. Small, Mid, and Large cap-indices. The source of the data is the Thomson Reuters Data Stream. We compute daily log-returns from closing prices adjusted for dividend and split.

The symbols of the stocks included are given below. The small cap index contains the stocks with the following symbols: AGL, AIR, AMR, ASH, BDN, BEZ, BIG, BIO, BRE, BXS, CBRL, CBT, COO, CTX, CW, DLX, ESL, GAS, HXL, ITG, LIZ, LPX, MDP, NEU, PBY, PCH, PPD, RLI, TXI, UNS.

The mid cap index contains: ACV, ADSK, AMD, BCR, BDK, BMS, CBE, CCK, CEG, CSC, DBD, DOV, DTE, EK, DPL, GMGMQ, GR, GWW, HOT, MAS, MDC, MWV, NAV, NI, ROST, RSH, SWK, UNM, VFC, WEC.

The large cap index contains: ABT, ADBE, AMAT, APC, APD, AVP, BAC, BEN, BK, CA, CL, D, DD, EMR, FPL, ITW, JCI, JPM, LOW, MMM, MRK, OXY, PCAR, PEP, PFE, SO, SYY, TGT, WFT, XOM.

We estimate parameters from January 1987 to July 2009 using rolling windows of 4 years. Table A.1 below present the results for the small, mid and large cap index in the first, second and third panel, respectively.

Table A.1: Parameter estimates

	ω	α	γ	β	ν
Small					
mean	0.069	0.014	0.106	0.793	15.869
5% quantile	0.000	0.000	0.005	0.000	4.367
median	0.038	0.002	0.120	0.876	7.732
95% quantile	0.324	0.057	0.198	0.992	94.372
Mid					
mean	0.030	0.008	0.100	0.905	11.447
5% quantile	0.000	0.000	0.016	0.829	5.044
median	0.028	0.000	0.112	0.909	8.044
95% quantile	0.054	0.036	0.156	0.990	26.152
Large					
mean	0.038	0.010	0.106	0.884	12.215
5% quantile	0.003	0.000	0.029	0.836	4.440
median	0.020	0.000	0.104	0.918	8.011
95% quantile	0.068	0.036	0.143	0.970	34.728

Notes: This table displays parameter estimates for the TGARCH-t (Equation 3.3) that stem from a 4 year rolling window estimation from January 1987 to July 2009 on three equally weighted indices composed of 30 randomly chosen stocks from the Dow Jones U.S. Small, Mid, and Large cap-indices.

Table A.2: VCV matrix of quantile losses for the large cap index, p =0.01, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	103.277	104.596	115.682	116.404	106.031
garch-t	104.596	111.584	116.185	121.307	113.650
tgarch-n	115.682	116.185	137.535	136.173	117.566
garch-n	116.404	121.307	136.173	141.562	123.068
rmf	106.031	113.650	117.566	123.068	119.706

Notes: This table displays the variance-covariance matrix of the losses for VaR forecasts at quantile level $p = 0.01$ that correspond to the large cap index.

Table A.3: VCV matrix of quantile losses for the large cap index, p =0.025, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	228.371	230.164	232.105	232.409	225.478
garch-t	230.164	246.803	231.844	243.385	244.204
tgarch-n	232.105	231.844	238.341	237.122	224.562
garch-n	232.409	243.385	237.122	247.060	236.110
rmf	225.478	244.204	224.562	236.110	250.514

Notes: This table displays the variance-covariance matrix of the losses for VaR forecasts at quantile level $p = 0.025$ that correspond to the large cap index.

Table A.4: VCV matrix of quantile losses for the large cap index, p =0.05, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	423.469	432.114	402.798	407.604	421.841
garch-t	432.114	457.088	407.198	423.690	448.726
tgarch-n	402.798	407.198	386.988	388.536	395.292
garch-n	407.604	423.690	388.536	401.974	411.781
rmf	421.841	448.726	395.292	411.781	450.431

Notes: This table displays the variance-covariance matrix of the losses for VaR forecasts at quantile level $p = 0.05$ that correspond to the large cap index.

A.0.1 Empirical variances

Table A.5: VCV matrix of quantile losses for the large cap index, p =0.1, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	748.840	762.072	694.279	701.776	747.689
garch-t	762.072	790.992	701.404	720.861	779.155
tgarch-n	694.279	701.404	655.314	658.414	686.673
garch-n	701.776	720.861	658.414	674.704	706.683
rmf	747.689	779.155	686.673	706.683	778.333

Notes: This table displays the variance-covariance matrix of the losses for VaR forecasts at quantile level $p = 0.1$ that correspond to the large cap index.

A.0.2 Empirical correlations

Table A.6: Correlations of quantile losses for the large cap index, p =0.01, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	1.000	0.974	0.971	0.963	0.954
garch-t	0.974	1.000	0.938	0.965	0.983
tgarch-n	0.971	0.938	1.000	0.976	0.916
garch-n	0.963	0.965	0.976	1.000	0.945
rmf	0.954	0.983	0.916	0.945	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.01$ that correspond to the large cap index.

Table A.7: Correlations of quantile losses for the large cap index, p =0.025, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	1.000	0.969	0.995	0.978	0.943
garch-t	0.969	1.000	0.956	0.986	0.982
tgarch-n	0.995	0.956	1.000	0.977	0.919
garch-n	0.978	0.986	0.977	1.000	0.949
rmf	0.943	0.982	0.919	0.949	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.025$ that correspond to the large cap index.

Table A.8: Correlations of quantile losses for the large cap index, p =0.05, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	1.000	0.982	0.995	0.988	0.966
garch-t	0.982	1.000	0.968	0.988	0.989
tgarch-n	0.995	0.968	1.000	0.985	0.947
garch-n	0.988	0.988	0.985	1.000	0.968
rmf	0.966	0.989	0.947	0.968	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.05$ that correspond to the large cap index.

Table A.9: Correlations of quantile losses for the large cap index, p =0.1, b=1

	tgarch-t	garch-t	tgarch-n	garch-n	rmf
tgarch-t	1.000	0.990	0.991	0.987	0.979
garch-t	0.990	1.000	0.974	0.987	0.993
tgarch-n	0.991	0.974	1.000	0.990	0.961
garch-n	0.987	0.987	0.990	1.000	0.975
rmf	0.979	0.993	0.961	0.975	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.1$ that correspond to the large cap index.

B DGP

B.1 Expected losses

This section provides the expected losses of the competing models that we describe in Section 3.1.3.

B.1.1 Expected losses for VaR forecasts

Table B.1: Absolute expected losses for VaR; ν =3

b	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.500	0.010	4.040	4.061	4.064	4.071	4.075	4.094	4.494	4.452	4.196	4.168	4.232
1.000	0.010	4.414	4.456	4.460	4.473	4.471	4.510	5.295	5.216	4.642	4.606	4.689
2.000	0.010	15.636	15.835	15.849	15.894	15.819	16.012	19.238	18.958	16.242	16.186	16.327
0.500	0.025	8.555	8.595	8.628	8.624	8.572	8.615	9.383	9.490	8.713	8.786	8.716
1.000	0.025	7.943	8.010	8.057	8.052	7.968	8.039	9.273	9.447	8.167	8.235	8.169
2.000	0.025	20.752	20.973	21.092	21.079	20.816	21.043	24.468	24.936	21.353	21.349	21.347
0.500	0.050	14.969	15.030	15.211	15.245	15.193	15.270	16.202	16.890	15.195	15.300	15.203
1.000	0.050	12.213	12.301	12.537	12.583	12.508	12.618	13.897	14.902	12.532	12.577	12.550
2.000	0.050	25.241	25.459	25.927	26.027	25.844	26.101	28.629	30.777	26.056	25.804	26.128
0.500	0.100	25.841	25.923	26.423	26.575	26.613	26.744	27.507	29.525	26.241	26.251	26.316
1.000	0.100	18.353	18.451	19.013	19.191	19.236	19.395	20.227	22.769	18.861	18.734	18.965
2.000	0.100	29.905	30.074	30.903	31.191	31.266	31.531	32.446	36.500	30.915	30.327	31.149

Notes: This table displays the absolute expected losses for VaR forecasts that correspond to the DGP with $\nu = 3$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

Table B.2: Absolute expected losses for VaR; ν =7

b	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.500	0.010	3.990	4.026	4.022	4.027	4.020	4.051	4.519	4.574	4.067	4.058	4.086
1.000	0.010	4.143	4.212	4.205	4.214	4.196	4.260	5.210	5.311	4.280	4.270	4.308
2.000	0.010	10.415	10.705	10.688	10.732	10.608	10.907	15.038	15.382	10.937	10.939	10.998
0.500	0.025	9.015	9.118	9.072	9.072	9.017	9.074	9.863	9.862	9.149	9.250	9.152
1.000	0.025	8.480	8.656	8.580	8.580	8.484	8.583	9.978	9.975	8.696	8.833	8.700
2.000	0.025	17.681	18.265	18.028	18.029	17.695	18.043	22.665	22.659	18.342	18.602	18.345
0.500	0.050	16.500	16.831	16.583	16.588	16.507	16.594	17.650	17.685	16.688	17.074	16.680
1.000	0.050	14.254	14.737	14.385	14.391	14.265	14.400	16.048	16.101	14.528	15.004	14.522
2.000	0.050	25.412	26.559	25.770	25.781	25.436	25.799	30.068	30.190	26.088	26.877	26.097
0.500	0.100	29.672	30.367	29.777	29.801	29.731	29.854	31.072	31.293	29.902	30.688	29.894
1.000	0.100	23.203	24.044	23.344	23.374	23.279	23.442	25.064	25.356	23.495	24.331	23.497
2.000	0.100	34.718	36.104	35.007	35.055	34.859	35.175	38.221	38.734	35.265	36.311	35.317

Notes: This table displays the absolute expected losses for VaR forecasts that correspond to the DGP with $\nu = 7$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

Table B.3: Absolute expected losses for VaR; ν =12

b	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.500	0.010	3.882	3.934	3.921	3.916	3.894	3.928	4.420	4.456	3.950	3.952	3.958
1.000	0.010	3.893	3.990	3.966	3.958	3.915	3.982	4.959	5.026	4.019	4.028	4.029
2.000	0.010	8.842	9.215	9.130	9.112	8.919	9.212	13.274	13.497	9.344	9.413	9.343
0.500	0.025	8.912	9.019	8.973	8.973	8.914	8.973	9.751	9.753	9.040	9.136	9.042
1.000	0.025	8.225	8.408	8.331	8.330	8.228	8.333	9.702	9.705	8.435	8.568	8.438
2.000	0.025	15.959	16.554	16.310	16.310	15.968	16.321	20.808	20.815	16.603	16.865	16.605
0.500	0.050	16.497	16.918	16.584	16.584	16.499	16.586	17.615	17.626	16.689	17.144	16.677
1.000	0.050	14.137	14.750	14.274	14.272	14.139	14.276	15.893	15.910	14.416	15.000	14.403
2.000	0.050	23.911	25.351	24.284	24.276	23.915	24.281	28.506	28.544	24.588	25.639	24.580
0.500	0.100	29.978	30.929	30.089	30.083	29.993	30.109	31.314	31.404	30.226	31.226	30.193
1.000	0.100	23.479	24.638	23.632	23.623	23.500	23.655	25.287	25.407	23.791	24.902	23.760
2.000	0.100	33.865	35.769	34.184	34.160	33.903	34.212	37.367	37.580	34.425	35.944	34.409

Notes: This table displays the absolute expected losses for VaR forecasts that correspond to the DGP with $\nu = 12$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

B.1.2 Expected losses for joint VaR and ES forecasts

Table B.4: Absolute expected losses for ES and VaR; ν =3

loss	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.000	0.010	2.432	2.443	2.474	2.503	2.551	2.559	2.680	2.721	2.634	2.524	2.708
1.000	0.010	2.064	2.075	2.091	2.106	2.126	2.135	2.303	2.314	2.180	2.131	2.216
2.000	0.010	0.714	0.716	0.726	0.735	0.755	0.755	0.741	0.755	0.778	0.738	0.802
0.000	0.025	2.118	2.128	2.141	2.147	2.148	2.158	2.341	2.364	2.210	2.200	2.228
1.000	0.025	1.751	1.760	1.768	1.770	1.765	1.774	1.929	1.948	1.799	1.801	1.805
2.000	0.025	0.712	0.715	0.718	0.721	0.723	0.725	0.763	0.768	0.743	0.738	0.749
0.000	0.050	1.881	1.890	1.910	1.915	1.910	1.921	2.082	2.144	1.937	1.952	1.943
1.000	0.050	1.536	1.542	1.557	1.561	1.556	1.564	1.671	1.727	1.565	1.572	1.567
2.000	0.050	0.708	0.711	0.717	0.719	0.718	0.721	0.775	0.793	0.726	0.731	0.728
0.000	0.100	1.649	1.657	1.688	1.697	1.696	1.707	1.822	1.932	1.690	1.707	1.695
1.000	0.100	1.331	1.336	1.356	1.362	1.363	1.369	1.427	1.510	1.353	1.355	1.356
2.000	0.100	0.704	0.707	0.720	0.724	0.724	0.729	0.777	0.825	0.719	0.722	0.721

Notes: This table displays the absolute expected losses for joint VaR and ES forecasts that correspond to the DGP with $\nu = 3$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

Table B.5: Absolute expected losses for ES and VaR; ν =7

loss	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.000	0.010	2.377	2.421	2.394	2.399	2.409	2.425	2.672	2.746	2.423	2.428	2.444
1.000	0.010	2.006	2.038	2.022	2.026	2.028	2.044	2.289	2.339	2.045	2.049	2.058
2.000	0.010	0.707	0.715	0.709	0.710	0.715	0.716	0.742	0.758	0.715	0.715	0.721
0.000	0.025	2.192	2.219	2.206	2.207	2.199	2.213	2.413	2.428	2.229	2.248	2.235
1.000	0.025	1.815	1.837	1.827	1.828	1.818	1.830	1.999	2.006	1.843	1.860	1.845
2.000	0.025	0.728	0.735	0.731	0.732	0.730	0.733	0.776	0.781	0.738	0.741	0.740
0.000	0.050	2.044	2.080	2.056	2.057	2.047	2.059	2.218	2.226	2.074	2.115	2.075
1.000	0.050	1.664	1.693	1.673	1.674	1.666	1.675	1.792	1.797	1.684	1.715	1.684
2.000	0.050	0.759	0.770	0.763	0.763	0.760	0.764	0.814	0.817	0.769	0.783	0.770
0.000	0.100	1.892	1.934	1.902	1.903	1.896	1.906	2.024	2.036	1.915	1.967	1.914
1.000	0.100	1.501	1.531	1.507	1.508	1.504	1.510	1.583	1.593	1.514	1.547	1.514
2.000	0.100	0.809	0.828	0.813	0.814	0.811	0.815	0.865	0.871	0.819	0.840	0.818

Notes: This table displays the absolute expected losses for joint VaR and ES forecasts that correspond to the DGP with $\nu = 7$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

Table B.6: Absolute expected losses for ES and VaR; ν =12

loss	p	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
0.000	0.010	2.320	2.389	2.343	2.338	2.331	2.350	2.628	2.670	2.359	2.388	2.366
1.000	0.010	1.948	1.996	1.968	1.965	1.956	1.973	2.236	2.265	1.982	2.000	1.986
2.000	0.010	0.699	0.713	0.703	0.702	0.702	0.704	0.740	0.749	0.706	0.711	0.708
0.000	0.025	2.167	2.201	2.183	2.182	2.169	2.184	2.386	2.395	2.200	2.224	2.202
1.000	0.025	1.790	1.816	1.803	1.803	1.792	1.804	1.972	1.977	1.817	1.835	1.818
2.000	0.025	0.723	0.732	0.727	0.726	0.723	0.727	0.771	0.775	0.731	0.736	0.732
0.000	0.050	2.041	2.086	2.054	2.054	2.042	2.054	2.209	2.213	2.070	2.117	2.069
1.000	0.050	1.660	1.696	1.669	1.669	1.660	1.669	1.784	1.786	1.680	1.716	1.679
2.000	0.050	0.758	0.772	0.763	0.762	0.759	0.763	0.812	0.813	0.768	0.783	0.768
0.000	0.100	1.910	1.964	1.919	1.919	1.910	1.920	2.033	2.038	1.932	1.994	1.930
1.000	0.100	1.512	1.552	1.519	1.518	1.513	1.519	1.590	1.594	1.526	1.567	1.525
2.000	0.100	0.817	0.841	0.821	0.821	0.817	0.821	0.869	0.872	0.826	0.852	0.825

Notes: This table displays the absolute expected losses for joint VaR and ES forecasts that correspond to the DGP with $\nu = 12$ degrees of freedom. The first 2 columns indicate the parametrization of the loss function and the quantile level, respectively.

B.2 Correlations

B.2.1 Correlations for VaR forecasts

Table B.7: Correlations of quantile losses for ν =3, p =0.01, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.994	0.993	0.992	0.996	0.989	0.878	0.876	0.981	0.983	0.979
garch-t_3	0.994	1.000	1.000	0.999	0.991	0.996	0.886	0.884	0.988	0.990	0.987
garch-t_7	0.993	1.000	1.000	1.000	0.993	0.998	0.888	0.887	0.990	0.991	0.989
garch-t_{12}	0.992	0.999	1.000	1.000	0.994	0.999	0.890	0.890	0.991	0.991	0.990
tgarch-n	0.996	0.991	0.993	0.994	1.000	0.994	0.885	0.886	0.985	0.984	0.984
garch-n	0.989	0.996	0.998	0.999	0.994	1.000	0.892	0.894	0.991	0.990	0.991
cv-t	0.878	0.886	0.888	0.890	0.885	0.892	1.000	0.997	0.860	0.859	0.862
cv-n	0.876	0.884	0.887	0.890	0.886	0.894	0.997	1.000	0.862	0.859	0.864
rmf	0.981	0.988	0.990	0.991	0.985	0.991	0.860	0.862	1.000	1.000	1.000
rmf-t_3	0.983	0.990	0.991	0.991	0.984	0.990	0.859	0.859	1.000	1.000	0.999
rmf-t_{12}	0.979	0.987	0.989	0.990	0.984	0.991	0.862	0.864	1.000	0.999	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.8: Correlations of quantile losses for ν =3, p =0.025, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.995	0.993	0.994	0.999	0.994	0.903	0.901	0.986	0.985	0.986
garch-t_3	0.995	1.000	0.998	0.998	0.993	0.999	0.908	0.907	0.992	0.991	0.992
garch-t_7	0.993	0.998	1.000	1.000	0.994	1.000	0.903	0.903	0.991	0.987	0.991
garch-t_{12}	0.994	0.998	1.000	1.000	0.994	1.000	0.904	0.903	0.991	0.987	0.991
tgarch-n	0.999	0.993	0.994	0.994	1.000	0.994	0.898	0.898	0.985	0.982	0.985
garch-n	0.994	0.999	1.000	1.000	0.994	1.000	0.905	0.904	0.991	0.988	0.991
cv-t	0.903	0.908	0.903	0.904	0.898	0.905	1.000	0.999	0.879	0.883	0.879
cv-n	0.901	0.907	0.903	0.903	0.898	0.904	0.999	1.000	0.877	0.881	0.878
rmf	0.986	0.992	0.991	0.991	0.985	0.991	0.879	0.877	1.000	0.998	1.000
rmf-t_3	0.985	0.991	0.987	0.987	0.982	0.988	0.883	0.881	0.998	1.000	0.998
rmf-t_{12}	0.986	0.992	0.991	0.991	0.985	0.991	0.879	0.878	1.000	0.998	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.9: Correlations of quantile losses for ν =3, p =0.05, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.996	0.990	0.988	0.991	0.988	0.922	0.913	0.986	0.988	0.985
garch-t_3	0.996	1.000	0.993	0.992	0.986	0.991	0.927	0.918	0.990	0.992	0.989
garch-t_7	0.990	0.993	1.000	1.000	0.995	1.000	0.916	0.914	0.991	0.982	0.992
garch-t_{12}	0.988	0.992	1.000	1.000	0.995	1.000	0.915	0.913	0.991	0.980	0.991
tgarch-n	0.991	0.986	0.995	0.995	1.000	0.995	0.909	0.908	0.985	0.974	0.985
garch-n	0.988	0.991	1.000	1.000	0.995	1.000	0.914	0.913	0.990	0.979	0.991
cv-t	0.922	0.927	0.916	0.915	0.909	0.914	1.000	0.993	0.895	0.903	0.894
cv-n	0.913	0.918	0.914	0.913	0.908	0.913	0.993	1.000	0.889	0.892	0.888
rmf	0.986	0.990	0.991	0.991	0.985	0.990	0.895	0.889	1.000	0.994	1.000
rmf-t_3	0.988	0.992	0.982	0.980	0.974	0.979	0.903	0.892	0.994	1.000	0.992
rmf-t_{12}	0.985	0.989	0.992	0.991	0.985	0.991	0.894	0.888	1.000	0.992	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.10: Correlations of quantile losses for ν =3, p =0.1, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.996	0.985	0.982	0.980	0.977	0.942	0.920	0.984	0.990	0.981
garch-t_3	0.996	1.000	0.988	0.984	0.976	0.980	0.945	0.923	0.987	0.994	0.984
garch-t_7	0.985	0.988	1.000	1.000	0.994	0.999	0.933	0.927	0.992	0.978	0.993
garch-t_{12}	0.982	0.984	1.000	1.000	0.995	1.000	0.929	0.926	0.991	0.974	0.992
tgarch-n	0.980	0.976	0.994	0.995	1.000	0.995	0.921	0.920	0.984	0.964	0.986
garch-n	0.977	0.980	0.999	1.000	0.995	1.000	0.925	0.925	0.989	0.969	0.991
cv-t	0.942	0.945	0.933	0.929	0.921	0.925	1.000	0.985	0.914	0.925	0.911
cv-n	0.920	0.923	0.927	0.926	0.920	0.925	0.985	1.000	0.902	0.900	0.901
rmf	0.984	0.987	0.992	0.991	0.984	0.989	0.914	0.902	1.000	0.989	1.000
rmf-t_3	0.990	0.994	0.978	0.974	0.964	0.969	0.925	0.900	0.989	1.000	0.985
rmf-t_{12}	0.981	0.984	0.993	0.992	0.986	0.991	0.911	0.901	1.000	0.985	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.11: Correlations of quantile losses for ν =7, p =0.01, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.972	0.972	0.970	0.989	0.961	0.657	0.660	0.947	0.948	0.944
garch-t_3	0.972	1.000	0.998	0.992	0.960	0.979	0.675	0.676	0.975	0.978	0.969
garch-t_7	0.972	0.998	1.000	0.998	0.968	0.989	0.686	0.688	0.978	0.978	0.975
garch-t_{12}	0.970	0.992	0.998	1.000	0.973	0.996	0.696	0.700	0.979	0.976	0.979
tgarch-n	0.989	0.960	0.968	0.973	1.000	0.975	0.680	0.688	0.948	0.943	0.952
garch-n	0.961	0.979	0.989	0.996	0.975	1.000	0.710	0.717	0.974	0.966	0.979
cv-t	0.657	0.675	0.686	0.696	0.680	0.710	1.000	0.995	0.646	0.636	0.655
cv-n	0.660	0.676	0.688	0.700	0.688	0.717	0.995	1.000	0.650	0.639	0.661
rmf	0.947	0.975	0.978	0.979	0.948	0.974	0.646	0.650	1.000	0.998	0.998
rmf-t_3	0.948	0.978	0.978	0.976	0.943	0.966	0.636	0.639	0.998	1.000	0.993
rmf-t_{12}	0.944	0.969	0.975	0.979	0.952	0.979	0.655	0.661	0.998	0.993	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.12: Correlations of quantile losses for ν =7, p =0.025, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.970	0.979	0.979	0.999	0.978	0.737	0.739	0.959	0.951	0.959
garch-t_3	0.970	1.000	0.991	0.992	0.974	0.995	0.781	0.784	0.978	0.984	0.978
garch-t_7	0.979	0.991	1.000	1.000	0.979	0.999	0.762	0.763	0.982	0.974	0.982
garch-t_{12}	0.979	0.992	1.000	1.000	0.979	1.000	0.763	0.765	0.982	0.975	0.982
tgarch-n	0.999	0.974	0.979	0.979	1.000	0.979	0.743	0.744	0.961	0.955	0.961
garch-n	0.978	0.995	0.999	1.000	0.979	1.000	0.767	0.768	0.982	0.977	0.983
cv-t	0.737	0.781	0.762	0.763	0.743	0.767	1.000	1.000	0.720	0.739	0.721
cv-n	0.739	0.784	0.763	0.765	0.744	0.768	1.000	1.000	0.722	0.741	0.723
rmf	0.959	0.978	0.982	0.982	0.961	0.982	0.720	0.722	1.000	0.992	1.000
rmf-t_3	0.951	0.984	0.974	0.975	0.955	0.977	0.739	0.741	0.992	1.000	0.993
rmf-t_{12}	0.959	0.978	0.982	0.982	0.961	0.983	0.721	0.723	1.000	0.993	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.13: Correlations of quantile losses for ν =7, p =0.05, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.962	0.984	0.984	0.999	0.983	0.802	0.800	0.969	0.948	0.969
garch-t_3	0.962	1.000	0.976	0.971	0.953	0.968	0.842	0.837	0.968	0.989	0.964
garch-t_7	0.984	0.976	1.000	1.000	0.982	0.999	0.822	0.820	0.986	0.964	0.986
garch-t_{12}	0.984	0.971	1.000	1.000	0.983	1.000	0.818	0.816	0.985	0.959	0.986
tgarch-n	0.999	0.953	0.982	0.983	1.000	0.983	0.795	0.794	0.967	0.939	0.968
garch-n	0.983	0.968	0.999	1.000	0.983	1.000	0.816	0.814	0.985	0.955	0.985
cv-t	0.802	0.842	0.822	0.818	0.795	0.816	1.000	1.000	0.784	0.807	0.780
cv-n	0.800	0.837	0.820	0.816	0.794	0.814	1.000	1.000	0.781	0.801	0.778
rmf	0.969	0.968	0.986	0.985	0.967	0.985	0.784	0.781	1.000	0.977	1.000
rmf-t_3	0.948	0.989	0.964	0.959	0.939	0.955	0.807	0.801	0.977	1.000	0.972
rmf-t_{12}	0.969	0.964	0.986	0.986	0.968	0.985	0.780	0.778	1.000	0.972	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.14: Correlations of quantile losses for ν =7, p =0.1, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.959	0.989	0.988	0.996	0.986	0.867	0.861	0.979	0.950	0.979
garch-t_3	0.959	1.000	0.968	0.957	0.934	0.944	0.889	0.873	0.964	0.994	0.953
garch-t_7	0.989	0.968	1.000	0.999	0.984	0.996	0.880	0.874	0.991	0.960	0.990
garch-t_{12}	0.988	0.957	0.999	1.000	0.987	0.999	0.875	0.871	0.989	0.949	0.990
tgarch-n	0.996	0.934	0.984	0.987	1.000	0.988	0.853	0.851	0.973	0.925	0.976
garch-n	0.986	0.944	0.996	0.999	0.988	1.000	0.868	0.866	0.986	0.936	0.989
cv-t	0.867	0.889	0.880	0.875	0.853	0.868	1.000	0.998	0.851	0.865	0.845
cv-n	0.861	0.873	0.874	0.871	0.851	0.866	0.998	1.000	0.844	0.848	0.839
rmf	0.979	0.964	0.991	0.989	0.973	0.986	0.851	0.844	1.000	0.968	0.999
rmf-t_3	0.950	0.994	0.960	0.949	0.925	0.936	0.865	0.848	0.968	1.000	0.957
rmf-t_{12}	0.979	0.953	0.990	0.990	0.976	0.989	0.845	0.839	0.999	0.957	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.15: Correlations of quantile losses for ν =12, p =0.01, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.949	0.955	0.957	0.993	0.950	0.578	0.583	0.923	0.920	0.923
garch-t_3	0.949	1.000	0.997	0.987	0.935	0.965	0.588	0.589	0.963	0.967	0.953
garch-t_7	0.955	0.997	1.000	0.997	0.948	0.981	0.603	0.606	0.969	0.968	0.964
garch-t_{12}	0.957	0.987	0.997	1.000	0.957	0.993	0.619	0.623	0.970	0.964	0.970
tgarch-n	0.993	0.935	0.948	0.957	1.000	0.960	0.597	0.604	0.922	0.912	0.927
garch-n	0.950	0.965	0.981	0.993	0.960	1.000	0.639	0.645	0.962	0.949	0.969
cv-t	0.578	0.588	0.603	0.619	0.597	0.639	1.000	0.997	0.557	0.542	0.571
cv-n	0.583	0.589	0.606	0.623	0.604	0.645	0.997	1.000	0.562	0.546	0.577
rmf	0.923	0.963	0.969	0.970	0.922	0.962	0.557	0.562	1.000	0.997	0.997
rmf-t_3	0.920	0.967	0.968	0.964	0.912	0.949	0.542	0.546	0.997	1.000	0.988
rmf-t_{12}	0.923	0.953	0.964	0.970	0.927	0.969	0.571	0.577	0.997	0.988	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.16: Correlations of quantile losses for ν =12, p =0.025, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.958	0.969	0.969	0.999	0.969	0.680	0.681	0.944	0.934	0.944
garch-t_3	0.958	1.000	0.987	0.988	0.963	0.992	0.739	0.741	0.969	0.980	0.971
garch-t_7	0.969	0.987	1.000	1.000	0.970	0.999	0.714	0.715	0.977	0.964	0.976
garch-t_{12}	0.969	0.988	1.000	1.000	0.970	1.000	0.715	0.717	0.977	0.966	0.977
tgarch-n	0.999	0.963	0.970	0.970	1.000	0.970	0.685	0.687	0.946	0.939	0.946
garch-n	0.969	0.992	0.999	1.000	0.970	1.000	0.720	0.722	0.977	0.970	0.977
cv-t	0.680	0.739	0.714	0.715	0.685	0.720	1.000	1.000	0.662	0.687	0.664
cv-n	0.681	0.741	0.715	0.717	0.687	0.722	1.000	1.000	0.664	0.690	0.666
rmf	0.944	0.969	0.977	0.977	0.946	0.977	0.662	0.664	1.000	0.987	1.000
rmf-t_3	0.934	0.980	0.964	0.966	0.939	0.970	0.687	0.690	0.987	1.000	0.989
rmf-t_{12}	0.944	0.971	0.976	0.977	0.946	0.977	0.664	0.666	1.000	0.989	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.17: Correlations of quantile losses for ν =12, p =0.05, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.942	0.978	0.978	1.000	0.978	0.762	0.761	0.960	0.926	0.960
garch-t_3	0.942	1.000	0.968	0.961	0.937	0.956	0.815	0.812	0.959	0.988	0.952
garch-t_7	0.978	0.968	1.000	1.000	0.977	0.999	0.793	0.792	0.983	0.954	0.983
garch-t_{12}	0.978	0.961	1.000	1.000	0.978	1.000	0.789	0.788	0.982	0.947	0.983
tgarch-n	1.000	0.937	0.977	0.978	1.000	0.978	0.759	0.758	0.958	0.921	0.959
garch-n	0.978	0.956	0.999	1.000	0.978	1.000	0.786	0.785	0.981	0.941	0.982
cv-t	0.762	0.815	0.793	0.789	0.759	0.786	1.000	1.000	0.748	0.775	0.744
cv-n	0.761	0.812	0.792	0.788	0.758	0.785	1.000	1.000	0.747	0.771	0.743
rmf	0.960	0.959	0.983	0.982	0.958	0.981	0.748	0.747	1.000	0.968	1.000
rmf-t_3	0.926	0.988	0.954	0.947	0.921	0.941	0.775	0.771	0.968	1.000	0.962
rmf-t_{12}	0.960	0.952	0.983	0.983	0.959	0.982	0.744	0.743	1.000	0.962	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.18: Correlations of quantile losses for ν =12, p =0.1, b=1

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.935	0.985	0.986	0.999	0.985	0.845	0.841	0.974	0.926	0.975
garch-t_3	0.935	1.000	0.960	0.946	0.918	0.929	0.868	0.857	0.956	0.994	0.942
garch-t_7	0.985	0.960	1.000	0.999	0.981	0.995	0.868	0.863	0.990	0.952	0.989
garch-t_{12}	0.986	0.946	0.999	1.000	0.984	0.999	0.862	0.859	0.988	0.938	0.990
tgarch-n	0.999	0.918	0.981	0.984	1.000	0.986	0.837	0.835	0.969	0.909	0.972
garch-n	0.985	0.929	0.995	0.999	0.986	1.000	0.855	0.853	0.984	0.921	0.988
cv-t	0.845	0.868	0.868	0.862	0.837	0.855	1.000	0.999	0.835	0.842	0.829
cv-n	0.841	0.857	0.863	0.859	0.835	0.853	0.999	1.000	0.830	0.831	0.825
rmf	0.974	0.956	0.990	0.988	0.969	0.984	0.835	0.830	1.000	0.960	0.999
rmf-t_3	0.926	0.994	0.952	0.938	0.909	0.921	0.842	0.831	0.960	1.000	0.946
rmf-t_{12}	0.975	0.942	0.989	0.990	0.972	0.988	0.829	0.825	0.999	0.946	1.000

Notes: This table displays the correlations of the losses for VaR forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

B.2.2 Correlations for joint VaR and ES forecasts

Table B.19: Correlations of joint losses for ν =3, p =, 0.01, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.988	0.987	0.987	0.996	0.984	0.708	0.717	0.947	0.950	0.944
garch-t_3	0.988	1.000	1.000	0.999	0.984	0.996	0.721	0.730	0.958	0.961	0.955
garch-t_7	0.987	1.000	1.000	1.000	0.986	0.998	0.719	0.729	0.962	0.964	0.959
garch-t_{12}	0.987	0.999	1.000	1.000	0.987	0.999	0.718	0.729	0.964	0.966	0.962
tgarch-n	0.996	0.984	0.986	0.987	1.000	0.988	0.702	0.714	0.956	0.957	0.954
garch-n	0.984	0.996	0.998	0.999	0.988	1.000	0.715	0.727	0.967	0.968	0.966
cv-t	0.708	0.721	0.719	0.718	0.702	0.715	1.000	0.997	0.621	0.623	0.619
cv-n	0.717	0.730	0.729	0.729	0.714	0.727	0.997	1.000	0.634	0.636	0.633
rmf	0.947	0.958	0.962	0.964	0.956	0.967	0.621	0.634	1.000	1.000	1.000
rmf-t_3	0.950	0.961	0.964	0.966	0.957	0.968	0.623	0.636	1.000	1.000	0.999
rmf-t_{12}	0.944	0.955	0.959	0.962	0.954	0.966	0.619	0.633	1.000	0.999	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.20: Correlations of joint losses for ν =3, p =, 0.025, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.989	0.987	0.987	0.999	0.988	0.734	0.727	0.958	0.955	0.958
garch-t_3	0.989	1.000	0.998	0.998	0.988	0.999	0.746	0.739	0.969	0.965	0.969
garch-t_7	0.987	0.998	1.000	1.000	0.988	1.000	0.748	0.742	0.964	0.958	0.964
garch-t_{12}	0.987	0.998	1.000	1.000	0.988	1.000	0.748	0.742	0.965	0.958	0.964
tgarch-n	0.999	0.988	0.988	0.988	1.000	0.989	0.735	0.728	0.955	0.950	0.955
garch-n	0.988	0.999	1.000	1.000	0.989	1.000	0.747	0.741	0.966	0.960	0.965
cv-t	0.734	0.746	0.748	0.748	0.735	0.747	1.000	0.999	0.655	0.652	0.654
cv-n	0.727	0.739	0.742	0.742	0.728	0.741	0.999	1.000	0.647	0.644	0.647
rmf	0.958	0.969	0.964	0.965	0.955	0.966	0.655	0.647	1.000	0.998	1.000
rmf-t_3	0.955	0.965	0.958	0.958	0.950	0.960	0.652	0.644	0.998	1.000	0.999
rmf-t_{12}	0.958	0.969	0.964	0.964	0.955	0.965	0.654	0.647	1.000	0.999	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.21: Correlations of joint losses for ν =3, p =, 0.05, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.990	0.984	0.983	0.992	0.982	0.758	0.734	0.963	0.959	0.963
garch-t_3	0.990	1.000	0.994	0.993	0.983	0.992	0.770	0.746	0.973	0.969	0.973
garch-t_7	0.984	0.994	1.000	1.000	0.989	1.000	0.773	0.755	0.967	0.954	0.968
garch-t_{12}	0.983	0.993	1.000	1.000	0.989	1.000	0.772	0.755	0.966	0.952	0.967
tgarch-n	0.992	0.983	0.989	0.989	1.000	0.989	0.761	0.744	0.955	0.942	0.956
garch-n	0.982	0.992	1.000	1.000	0.989	1.000	0.772	0.755	0.965	0.951	0.966
cv-t	0.758	0.770	0.773	0.772	0.761	0.772	1.000	0.994	0.685	0.680	0.685
cv-n	0.734	0.746	0.755	0.755	0.744	0.755	0.994	1.000	0.662	0.653	0.663
rmf	0.963	0.973	0.967	0.966	0.955	0.965	0.685	0.662	1.000	0.995	1.000
rmf-t_3	0.959	0.969	0.954	0.952	0.942	0.951	0.680	0.653	0.995	1.000	0.994
rmf-t_{12}	0.963	0.973	0.968	0.967	0.956	0.966	0.685	0.663	1.000	0.994	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.22: Correlations of joint losses for ν =3, p =, 0.1, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.991	0.982	0.979	0.985	0.976	0.788	0.749	0.966	0.965	0.965
garch-t_3	0.991	1.000	0.991	0.988	0.977	0.985	0.799	0.760	0.975	0.974	0.974
garch-t_7	0.982	0.991	1.000	1.000	0.990	0.999	0.803	0.774	0.971	0.955	0.972
garch-t_{12}	0.979	0.988	1.000	1.000	0.990	1.000	0.802	0.776	0.969	0.951	0.970
tgarch-n	0.985	0.977	0.990	0.990	1.000	0.990	0.792	0.766	0.957	0.940	0.959
garch-n	0.976	0.985	0.999	1.000	0.990	1.000	0.802	0.777	0.966	0.947	0.968
cv-t	0.788	0.799	0.803	0.802	0.792	0.802	1.000	0.989	0.722	0.716	0.722
cv-n	0.749	0.760	0.774	0.776	0.766	0.777	0.989	1.000	0.686	0.673	0.688
rmf	0.966	0.975	0.971	0.969	0.957	0.966	0.722	0.686	1.000	0.993	1.000
rmf-t_3	0.965	0.974	0.955	0.951	0.940	0.947	0.716	0.673	0.993	1.000	0.990
rmf-t_{12}	0.965	0.974	0.972	0.970	0.959	0.968	0.722	0.688	1.000	0.990	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 3$ degrees of freedom.

Table B.23: Correlations of joint losses for ν =7, p =, 0.01, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.961	0.963	0.963	0.989	0.956	0.514	0.530	0.923	0.925	0.918
garch-t_3	0.961	1.000	0.998	0.992	0.945	0.979	0.548	0.562	0.949	0.956	0.940
garch-t_7	0.963	0.998	1.000	0.998	0.955	0.989	0.549	0.565	0.957	0.960	0.951
garch-t_{12}	0.963	0.992	0.998	1.000	0.962	0.996	0.551	0.568	0.961	0.961	0.958
tgarch-n	0.989	0.945	0.955	0.962	1.000	0.966	0.517	0.537	0.929	0.924	0.930
garch-n	0.956	0.979	0.989	0.996	0.966	1.000	0.553	0.573	0.962	0.958	0.963
cv-t	0.514	0.548	0.549	0.551	0.517	0.553	1.000	0.995	0.472	0.473	0.473
cv-n	0.530	0.562	0.565	0.568	0.537	0.573	0.995	1.000	0.490	0.489	0.492
rmf	0.923	0.949	0.957	0.961	0.929	0.962	0.472	0.490	1.000	0.998	0.998
rmf-t_3	0.925	0.956	0.960	0.961	0.924	0.958	0.473	0.489	0.998	1.000	0.993
rmf-t_{12}	0.918	0.940	0.951	0.958	0.930	0.963	0.473	0.492	0.998	0.993	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.24: Correlations of joint losses for ν =7, p =, 0.025, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.966	0.971	0.971	0.999	0.971	0.593	0.598	0.938	0.929	0.937
garch-t_3	0.966	1.000	0.993	0.994	0.969	0.996	0.629	0.634	0.968	0.968	0.968
garch-t_7	0.971	0.993	1.000	1.000	0.970	0.999	0.626	0.630	0.965	0.953	0.964
garch-t_{12}	0.971	0.994	1.000	1.000	0.970	1.000	0.626	0.630	0.966	0.955	0.965
tgarch-n	0.999	0.969	0.970	0.970	1.000	0.971	0.593	0.598	0.940	0.933	0.940
garch-n	0.971	0.996	0.999	1.000	0.971	1.000	0.625	0.630	0.967	0.958	0.967
cv-t	0.593	0.629	0.626	0.626	0.593	0.625	1.000	1.000	0.548	0.551	0.547
cv-n	0.598	0.634	0.630	0.630	0.598	0.630	1.000	1.000	0.552	0.556	0.552
rmf	0.938	0.968	0.965	0.966	0.940	0.967	0.548	0.552	1.000	0.994	1.000
rmf-t_3	0.929	0.968	0.953	0.955	0.933	0.958	0.551	0.556	0.994	1.000	0.994
rmf-t_{12}	0.937	0.968	0.964	0.965	0.940	0.967	0.547	0.552	1.000	0.994	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.25: Correlations of joint losses for ν =7, p =, 0.05, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.965	0.977	0.976	0.999	0.976	0.664	0.658	0.951	0.933	0.951
garch-t_3	0.965	1.000	0.983	0.979	0.959	0.977	0.688	0.680	0.969	0.976	0.967
garch-t_7	0.977	0.983	1.000	1.000	0.976	0.999	0.692	0.686	0.972	0.951	0.973
garch-t_{12}	0.976	0.979	1.000	1.000	0.976	1.000	0.692	0.685	0.971	0.946	0.972
tgarch-n	0.999	0.959	0.976	0.976	1.000	0.976	0.662	0.656	0.948	0.927	0.949
garch-n	0.976	0.977	0.999	1.000	0.976	1.000	0.691	0.685	0.970	0.944	0.971
cv-t	0.664	0.688	0.692	0.692	0.662	0.691	1.000	1.000	0.618	0.617	0.618
cv-n	0.658	0.680	0.686	0.685	0.656	0.685	1.000	1.000	0.611	0.608	0.611
rmf	0.951	0.969	0.972	0.971	0.948	0.970	0.618	0.611	1.000	0.984	1.000
rmf-t_3	0.933	0.976	0.951	0.946	0.927	0.944	0.617	0.608	0.984	1.000	0.981
rmf-t_{12}	0.951	0.967	0.973	0.972	0.949	0.971	0.618	0.611	1.000	0.981	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.26: Correlations of joint losses for ν =7, p =, 0.1, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.969	0.983	0.982	0.998	0.980	0.742	0.726	0.964	0.947	0.964
garch-t_3	0.969	1.000	0.981	0.974	0.956	0.967	0.752	0.732	0.972	0.984	0.967
garch-t_7	0.983	0.981	1.000	0.999	0.981	0.998	0.765	0.749	0.980	0.958	0.981
garch-t_{12}	0.982	0.974	0.999	1.000	0.982	0.999	0.765	0.750	0.978	0.951	0.979
tgarch-n	0.998	0.956	0.981	0.982	1.000	0.982	0.741	0.727	0.959	0.932	0.961
garch-n	0.980	0.967	0.998	0.999	0.982	1.000	0.764	0.750	0.975	0.943	0.977
cv-t	0.742	0.752	0.765	0.765	0.741	0.764	1.000	0.998	0.701	0.693	0.701
cv-n	0.726	0.732	0.749	0.750	0.727	0.750	0.998	1.000	0.685	0.673	0.685
rmf	0.964	0.972	0.980	0.978	0.959	0.975	0.701	0.685	1.000	0.983	0.999
rmf-t_3	0.947	0.984	0.958	0.951	0.932	0.943	0.693	0.673	0.983	1.000	0.977
rmf-t_{12}	0.964	0.967	0.981	0.979	0.961	0.977	0.701	0.685	0.999	0.977	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 7$ degrees of freedom.

Table B.27: Correlations of joint losses for ν =12, p =, 0.01, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.934	0.944	0.947	0.993	0.944	0.452	0.466	0.902	0.900	0.899
garch-t_3	0.934	1.000	0.996	0.986	0.917	0.965	0.491	0.501	0.933	0.943	0.921
garch-t_7	0.944	0.996	1.000	0.997	0.933	0.983	0.494	0.506	0.945	0.948	0.937
garch-t_{12}	0.947	0.986	0.997	1.000	0.944	0.994	0.498	0.511	0.951	0.949	0.947
tgarch-n	0.993	0.917	0.933	0.944	1.000	0.950	0.457	0.472	0.903	0.895	0.906
garch-n	0.944	0.965	0.983	0.994	0.950	1.000	0.503	0.518	0.950	0.942	0.953
cv-t	0.452	0.491	0.494	0.498	0.457	0.503	1.000	0.998	0.416	0.414	0.419
cv-n	0.466	0.501	0.506	0.511	0.472	0.518	0.998	1.000	0.428	0.426	0.432
rmf	0.902	0.933	0.945	0.951	0.903	0.950	0.416	0.428	1.000	0.996	0.997
rmf-t_3	0.900	0.943	0.948	0.949	0.895	0.942	0.414	0.426	0.996	1.000	0.988
rmf-t_{12}	0.899	0.921	0.937	0.947	0.906	0.953	0.419	0.432	0.997	0.988	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.01$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.28: Correlations of joint losses for ν =12, p =, 0.025, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.956	0.960	0.960	1.000	0.961	0.552	0.556	0.924	0.915	0.924
garch-t_3	0.956	1.000	0.990	0.991	0.959	0.994	0.599	0.604	0.963	0.964	0.963
garch-t_7	0.960	0.990	1.000	1.000	0.960	0.999	0.593	0.598	0.960	0.946	0.959
garch-t_{12}	0.960	0.991	1.000	1.000	0.960	1.000	0.593	0.597	0.961	0.947	0.960
tgarch-n	1.000	0.959	0.960	0.960	1.000	0.961	0.552	0.557	0.926	0.919	0.926
garch-n	0.961	0.994	0.999	1.000	0.961	1.000	0.593	0.597	0.962	0.952	0.962
cv-t	0.552	0.599	0.593	0.593	0.552	0.593	1.000	1.000	0.511	0.518	0.511
cv-n	0.556	0.604	0.598	0.597	0.557	0.597	1.000	1.000	0.516	0.523	0.515
rmf	0.924	0.963	0.960	0.961	0.926	0.962	0.511	0.516	1.000	0.991	1.000
rmf-t_3	0.915	0.964	0.946	0.947	0.919	0.952	0.518	0.523	0.991	1.000	0.992
rmf-t_{12}	0.924	0.963	0.959	0.960	0.926	0.962	0.511	0.515	1.000	0.992	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.025$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.29: Correlations of joint losses for ν =12, p =, 0.05, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.953	0.971	0.971	1.000	0.970	0.639	0.635	0.942	0.920	0.943
garch-t_3	0.953	1.000	0.978	0.973	0.949	0.970	0.669	0.665	0.964	0.976	0.961
garch-t_7	0.971	0.978	1.000	1.000	0.970	0.999	0.674	0.671	0.971	0.945	0.971
garch-t_{12}	0.971	0.973	1.000	1.000	0.970	1.000	0.673	0.670	0.969	0.940	0.970
tgarch-n	1.000	0.949	0.970	0.970	1.000	0.970	0.638	0.634	0.941	0.917	0.942
garch-n	0.970	0.970	0.999	1.000	0.970	1.000	0.672	0.669	0.968	0.937	0.969
cv-t	0.639	0.669	0.674	0.673	0.638	0.672	1.000	1.000	0.598	0.598	0.597
cv-n	0.635	0.665	0.671	0.670	0.634	0.669	1.000	1.000	0.595	0.594	0.593
rmf	0.942	0.964	0.971	0.969	0.941	0.968	0.598	0.595	1.000	0.980	1.000
rmf-t_3	0.920	0.976	0.945	0.940	0.917	0.937	0.598	0.594	0.980	1.000	0.976
rmf-t_{12}	0.943	0.961	0.971	0.970	0.942	0.969	0.597	0.593	1.000	0.976	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.05$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

Table B.30: Correlations of joint losses for ν =12, p =, 0.1, joint los =AL

	tgarch-t	garch-t_3	garch-t_7	garch-t_{12}	tgarch-n	garch-n	cv-t	cv-n	rmf	rmf-t_3	rmf-t_{12}
tgarch-t	1.000	0.958	0.981	0.980	0.999	0.978	0.733	0.724	0.961	0.937	0.961
garch-t_3	0.958	1.000	0.978	0.970	0.950	0.962	0.743	0.731	0.970	0.986	0.964
garch-t_7	0.981	0.978	1.000	0.999	0.979	0.997	0.759	0.750	0.981	0.957	0.981
garch-t_{12}	0.980	0.970	0.999	1.000	0.980	0.999	0.759	0.751	0.978	0.948	0.980
tgarch-n	0.999	0.950	0.979	0.980	1.000	0.979	0.732	0.724	0.957	0.928	0.959
garch-n	0.978	0.962	0.997	0.999	0.979	1.000	0.759	0.751	0.975	0.939	0.977
cv-t	0.733	0.743	0.759	0.759	0.732	0.759	1.000	1.000	0.697	0.687	0.696
cv-n	0.724	0.731	0.750	0.751	0.724	0.751	1.000	1.000	0.687	0.675	0.687
rmf	0.961	0.970	0.981	0.978	0.957	0.975	0.697	0.687	1.000	0.980	0.999
rmf-t_3	0.937	0.986	0.957	0.948	0.928	0.939	0.687	0.675	0.980	1.000	0.973
rmf-t_{12}	0.961	0.964	0.981	0.980	0.959	0.977	0.696	0.687	0.999	0.973	1.000

Notes: This table displays the correlations of the losses for joint VaR and ES forecasts at quantile level $p = 0.1$ that correspond to the DGP with $\nu = 12$ degrees of freedom.

C Additional results

C.1 Additional results - set of $m = 5$ models

This section provides power and potency of the MCS test for the first set of models that we describe in Section 3.1.3.

C.1.1 Results for $T_{R,\mathcal{M}}$

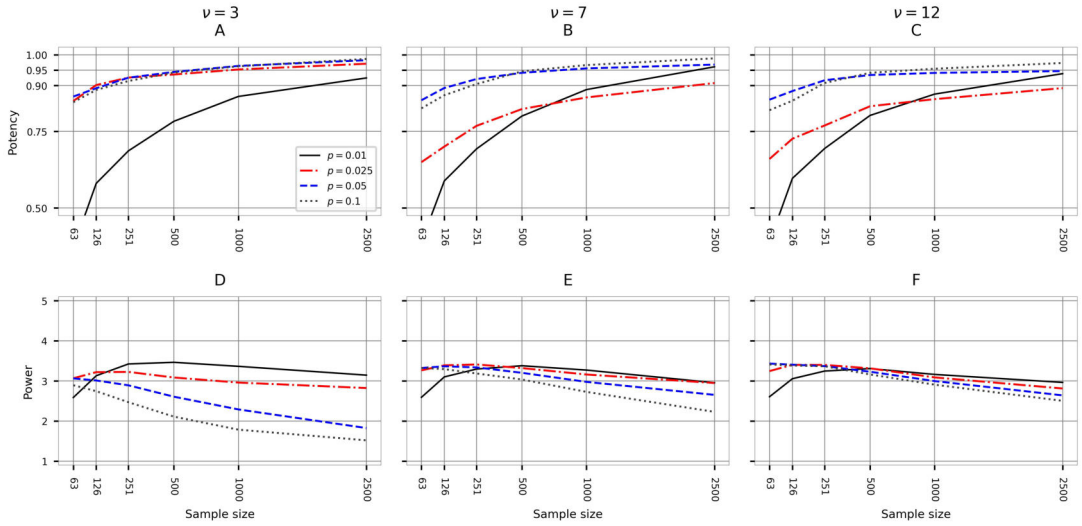

Figure C.1: Finite sample properties for the tick loss

This figure displays the finite sample properties of the MCS procedure using $T_{R,\mathcal{M}}$ in the following setting: one-day-ahead VaR forecasts evaluated using the tick loss, true model included, number of models $m = 5$, level of the test $\alpha = 0.25$. The upper row displays the potency, i.e. the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$. The lower row displays the power property, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

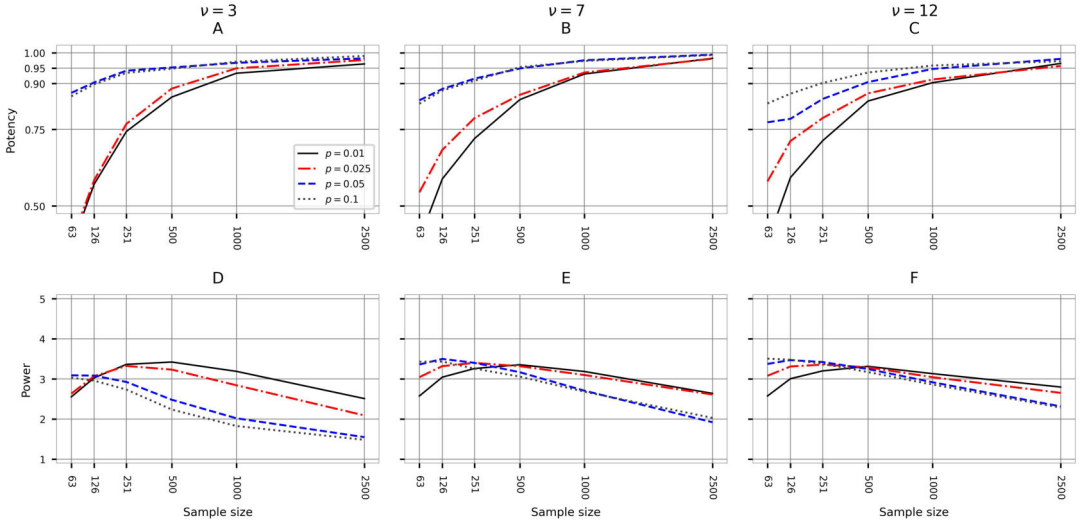

Figure C.2: Finite sample properties for the AL score

This figure displays the finite sample properties of the MCS procedure using $T_{R,\mathcal{M}}$ in the following setting: one-day-ahead VaR and ES forecasts evaluated using the AL score, true model included, number of models $m = 5$, level of the test $\alpha = 0.25$. The upper panel displays the potency, i.e. the frequency of $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$. The lower panel displays the power property, i.e. the average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

C.2 Results for varying level of the test α

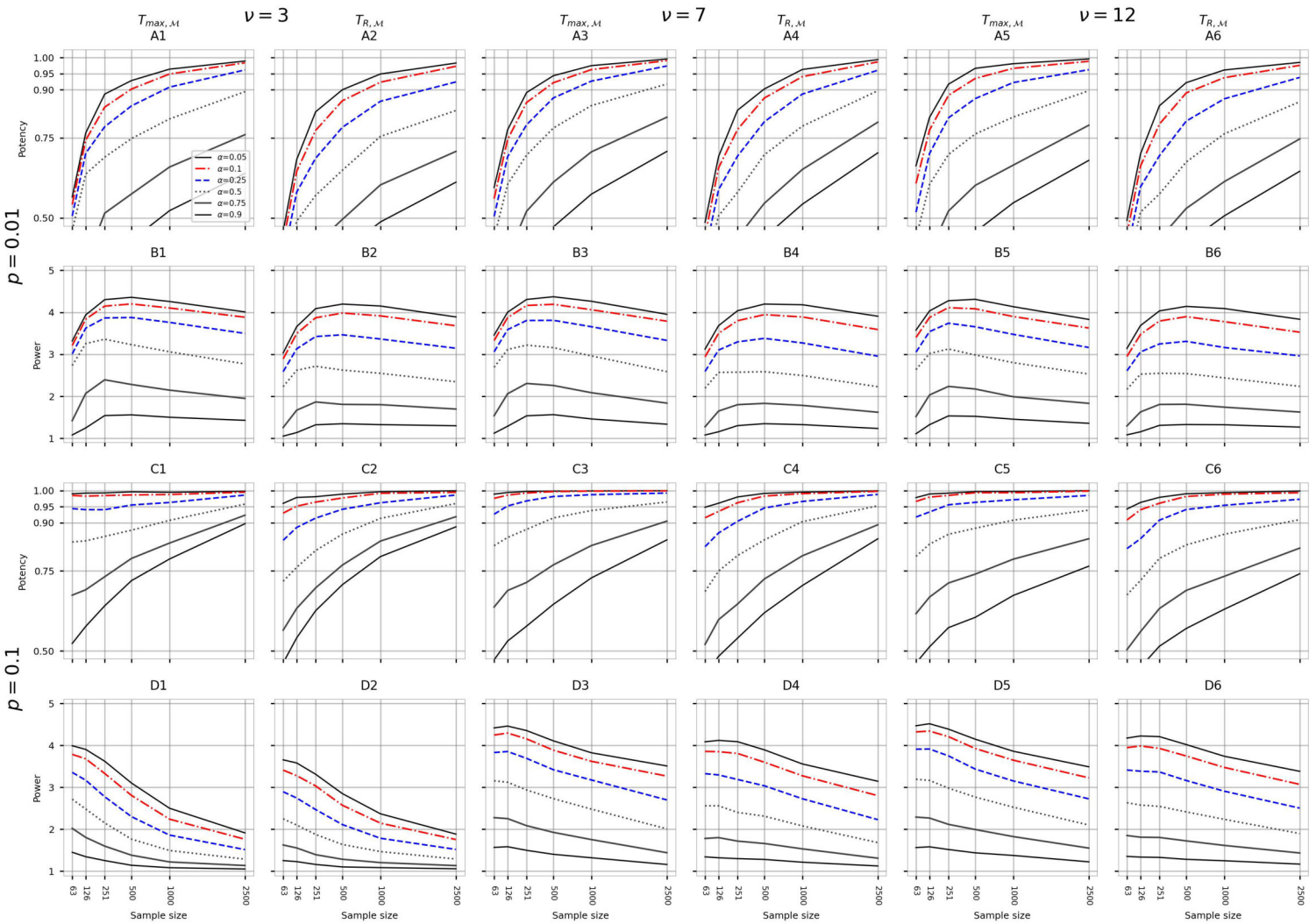

Figure C.3: Varying the level of the test α for the tick loss

This figure displays the finite sample properties of the MCS procedure for different levels of the test α in the following setting: one-day-ahead VaR forecasts evaluated using the tick loss, true model included, number of models $m = 5$. Columns 1 and 2 present the results for the DGP with $\nu = 3$, columns 3 and 4 for the DGP with $\nu = 7$ and columns 5 and 6 for the DGP with $\nu = 12$. Within each DGP, the two columns show the results for $T_{max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$, respectively. The two upper rows display the results for the quantile level $p = 0.01$ while the two lower rows show the results for the quantile level $p = 0.1$.

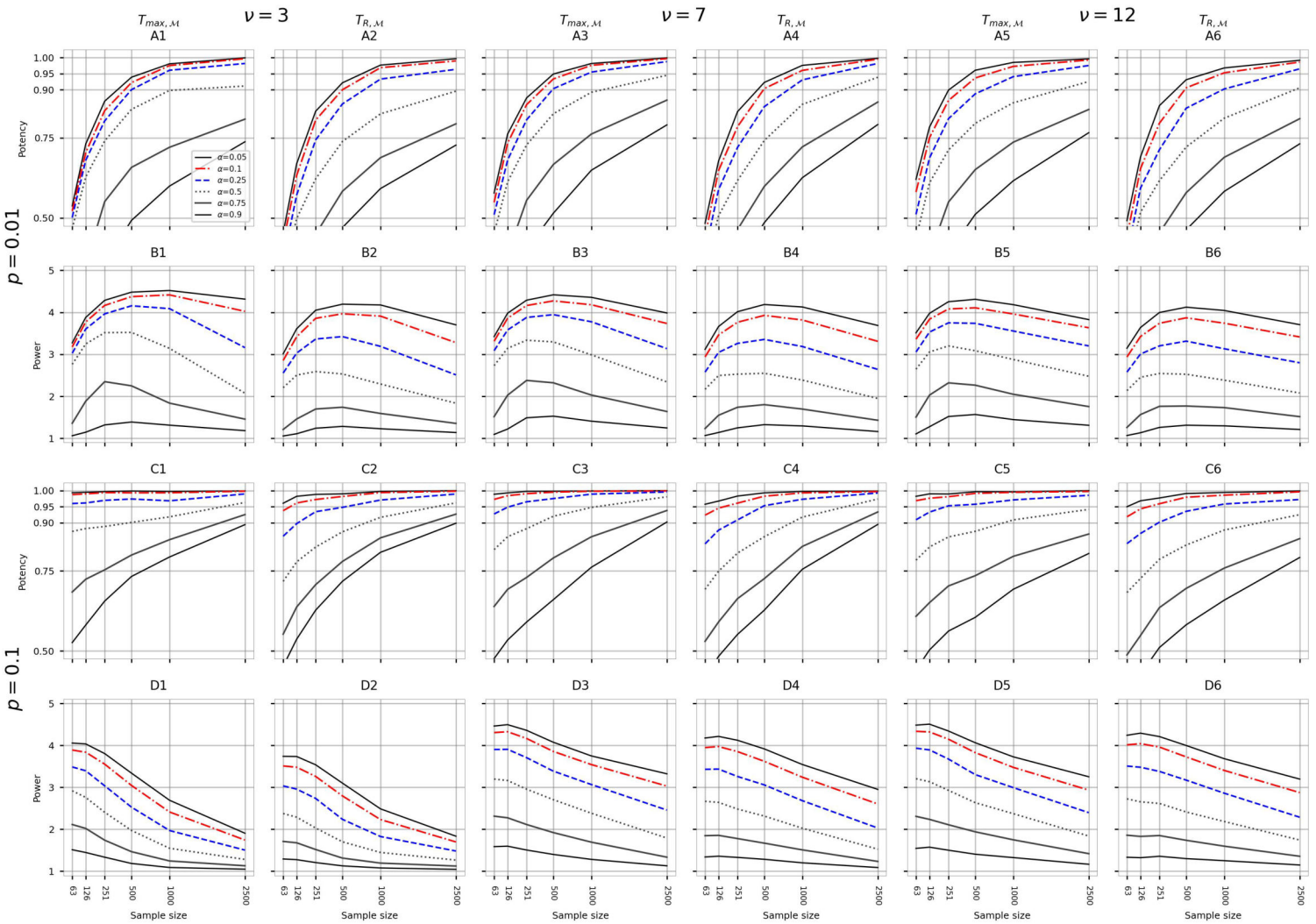

Figure C.4: Varying the level of the test α for the AL score

This figure displays the finite sample properties of the MCS procedure for different levels of the test α in the following setting: one-day-ahead VaR and ES forecasts evaluated using the AL score, true model included, number of models $m = 5$. Columns 1 and 2 present the results for the DGP with $\nu = 3$, columns 3 and 4 for the DGP with $\nu = 7$ and columns 5 and 6 for the DGP with $\nu = 12$. Within each DGP, the two columns show the results for $T_{max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$, respectively. The two upper rows display the results for the quantile level $p = 0.01$ while the two lower rows show the results for the quantile level $p = 0.1$.

C.2.1 Results for $T_{max,\mathcal{M}}$ and differently parametrized loss functions

Table C.1: Simulation results for the set of $m = 5$ models: VaR forecasts evaluated using the GPL loss with $b = 0.5$.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.503	0.699	0.798	0.879	0.926	0.968	3.018	3.636	3.911	3.974	3.813	3.482
	0.025	0.947	0.972	0.978	0.976	0.983	0.990	3.485	3.716	3.692	3.553	3.360	3.182
	0.050	0.963	0.970	0.975	0.976	0.980	0.988	3.580	3.543	3.340	3.018	2.626	1.976
	0.100	0.947	0.941	0.942	0.952	0.963	0.990	3.416	3.201	2.766	2.260	1.806	1.461
$\nu = 7$	0.010	0.525	0.702	0.823	0.882	0.933	0.982	3.127	3.649	3.896	3.876	3.650	3.322
	0.025	0.766	0.827	0.893	0.906	0.919	0.938	3.673	3.824	3.797	3.502	3.248	2.924
	0.050	0.928	0.944	0.962	0.965	0.972	0.982	3.870	3.895	3.734	3.425	3.187	2.736
	0.100	0.928	0.949	0.967	0.974	0.987	0.996	3.856	3.846	3.680	3.391	3.146	2.628
$\nu = 12$	0.010	0.514	0.691	0.816	0.881	0.934	0.971	3.063	3.537	3.752	3.670	3.493	3.200
	0.025	0.781	0.838	0.890	0.904	0.921	0.930	3.652	3.788	3.718	3.447	3.169	2.801
	0.050	0.908	0.928	0.949	0.952	0.956	0.955	3.812	3.835	3.662	3.365	3.084	2.610
	0.100	0.913	0.935	0.955	0.966	0.976	0.985	3.947	3.918	3.760	3.483	3.173	2.663

Notes: This table displays the finite sample properties of the MCS for VaR forecasts evaluated using the GPL loss with $b = 0.5$ and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.2: Simulation results for the set of $m = 5$ models: VaR forecasts evaluated using the tick loss.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.504	0.689	0.784	0.846	0.905	0.964	3.004	3.601	3.883	3.894	3.764	3.504
	0.025	0.943	0.973	0.976	0.977	0.981	0.987	3.458	3.680	3.661	3.499	3.359	3.209
	0.050	0.956	0.970	0.975	0.974	0.977	0.987	3.486	3.485	3.281	2.974	2.625	2.105
	0.100	0.943	0.947	0.948	0.947	0.962	0.987	3.356	3.174	2.800	2.308	1.890	1.514
$\nu = 7$	0.010	0.519	0.692	0.805	0.886	0.937	0.975	3.091	3.608	3.823	3.821	3.659	3.347
	0.025	0.764	0.831	0.879	0.901	0.911	0.926	3.618	3.812	3.760	3.494	3.265	2.930
	0.050	0.928	0.943	0.952	0.962	0.969	0.977	3.811	3.841	3.706	3.435	3.192	2.787
	0.100	0.937	0.949	0.970	0.978	0.985	0.994	3.834	3.805	3.636	3.403	3.136	2.715
$\nu = 12$	0.010	0.533	0.699	0.812	0.883	0.916	0.966	3.088	3.530	3.756	3.714	3.494	3.201
	0.025	0.768	0.842	0.884	0.889	0.906	0.933	3.625	3.805	3.733	3.443	3.190	2.788
	0.050	0.900	0.928	0.938	0.938	0.953	0.950	3.730	3.815	3.650	3.328	3.049	2.657
	0.100	0.908	0.930	0.952	0.964	0.970	0.984	3.861	3.866	3.753	3.442	3.162	2.754

Notes: This table displays the finite sample properties of the MCS for VaR forecasts evaluated using the tick loss and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.3: Simulation results for the set of $m = 5$ models: VaR forecasts evaluated using the GPL loss with $b = 2$.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.478	0.655	0.746	0.815	0.844	0.910	2.921	3.503	3.761	3.818	3.655	3.555
	0.025	0.933	0.966	0.972	0.971	0.978	0.982	3.364	3.623	3.629	3.513	3.322	3.194
	0.050	0.952	0.961	0.958	0.964	0.966	0.980	3.411	3.387	3.234	3.006	2.692	2.380
	0.100	0.935	0.928	0.933	0.944	0.948	0.973	3.238	3.087	2.814	2.520	2.134	1.797
$\nu = 7$	0.010	0.500	0.667	0.775	0.825	0.876	0.943	3.053	3.573	3.804	3.766	3.617	3.429
	0.025	0.740	0.796	0.866	0.868	0.881	0.903	3.588	3.756	3.790	3.520	3.298	3.094
	0.050	0.912	0.934	0.954	0.954	0.965	0.983	3.779	3.844	3.754	3.469	3.254	3.003
	0.100	0.922	0.944	0.970	0.973	0.982	0.990	3.778	3.804	3.733	3.432	3.221	2.929
$\nu = 12$	0.010	0.504	0.675	0.786	0.838	0.887	0.942	3.020	3.510	3.733	3.656	3.480	3.292
	0.025	0.749	0.823	0.870	0.882	0.894	0.908	3.578	3.768	3.735	3.508	3.258	2.987
	0.050	0.898	0.919	0.938	0.951	0.964	0.973	3.754	3.799	3.686	3.448	3.219	2.901
	0.100	0.906	0.930	0.953	0.964	0.976	0.985	3.873	3.891	3.806	3.560	3.316	2.982

Notes: This table displays the finite sample properties of the MCS for VaR forecasts evaluated using the GPL loss with $b = 2$ and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.4: Simulation results for the set of $m = 5$ models: Var and ES forecasts evaluated using the AL score.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.499	0.674	0.797	0.898	0.962	0.980	3.017	3.584	3.956	4.166	4.101	3.106
	0.025	0.660	0.790	0.870	0.938	0.975	0.982	3.398	3.747	3.876	3.869	3.540	2.633
	0.050	0.958	0.979	0.984	0.984	0.978	0.983	3.607	3.655	3.484	3.028	2.433	1.705
	0.100	0.955	0.968	0.973	0.966	0.970	0.987	3.472	3.390	3.051	2.528	1.983	1.501
$\nu = 7$	0.010	0.522	0.692	0.810	0.912	0.958	0.989	3.114	3.641	3.915	3.988	3.744	3.116
	0.025	0.682	0.798	0.867	0.930	0.960	0.984	3.542	3.836	3.853	3.705	3.437	3.047
	0.050	0.901	0.926	0.941	0.964	0.976	0.994	3.861	3.930	3.754	3.427	3.076	2.516
	0.100	0.916	0.945	0.967	0.979	0.985	0.997	3.901	3.920	3.733	3.407	3.059	2.496
$\nu = 12$	0.010	0.527	0.682	0.805	0.893	0.940	0.978	3.088	3.523	3.767	3.768	3.592	3.193
	0.025	0.729	0.834	0.891	0.925	0.942	0.968	3.596	3.817	3.779	3.563	3.278	2.823
	0.050	0.884	0.879	0.908	0.930	0.953	0.980	3.790	3.820	3.639	3.322	2.980	2.461
	0.100	0.910	0.924	0.948	0.957	0.973	0.984	3.876	3.829	3.679	3.345	2.999	2.441

Notes: This table displays the finite sample properties of the MCS for Var and ES forecasts evaluated using the AL score and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.5: Simulation results for the set of $m = 5$ models: Var and ES forecasts evaluated using the NZ score.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.500	0.679	0.790	0.890	0.955	0.980	3.022	3.598	3.926	4.091	3.969	3.087
	0.025	0.922	0.954	0.965	0.974	0.985	0.990	3.525	3.830	3.829	3.635	3.286	2.666
	0.050	0.956	0.976	0.980	0.981	0.977	0.985	3.549	3.572	3.365	2.991	2.489	1.843
	0.100	0.952	0.958	0.963	0.957	0.967	0.990	3.427	3.286	2.911	2.395	1.896	1.483
$\nu = 7$	0.010	0.525	0.693	0.808	0.904	0.953	0.985	3.102	3.616	3.887	3.934	3.714	3.200
	0.025	0.709	0.817	0.872	0.924	0.949	0.979	3.571	3.833	3.811	3.614	3.353	2.999
	0.050	0.910	0.930	0.952	0.968	0.973	0.992	3.820	3.890	3.716	3.387	3.064	2.576
	0.100	0.918	0.948	0.970	0.976	0.985	0.995	3.872	3.871	3.724	3.390	3.081	2.568
$\nu = 12$	0.010	0.529	0.687	0.805	0.886	0.931	0.976	3.084	3.514	3.746	3.731	3.548	3.191
	0.025	0.748	0.841	0.890	0.912	0.928	0.960	3.611	3.815	3.758	3.490	3.226	2.786
	0.050	0.902	0.923	0.940	0.944	0.963	0.982	3.772	3.820	3.626	3.299	2.961	2.479
	0.100	0.911	0.928	0.952	0.962	0.970	0.981	3.880	3.848	3.704	3.366	3.041	2.540

Notes: This table displays the finite sample properties of the MCS for Var and ES forecasts evaluated using the NZ score and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.6: Simulation results for the set of $m = 5$ models: Var and ES forecasts evaluated using the FZG score.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.462	0.617	0.732	0.809	0.886	0.954	2.892	3.411	3.746	3.862	3.838	3.606
	0.025	0.896	0.940	0.951	0.960	0.971	0.983	3.419	3.733	3.753	3.623	3.396	3.122
	0.050	0.949	0.972	0.972	0.972	0.973	0.977	3.382	3.431	3.314	3.067	2.712	2.314
	0.100	0.938	0.948	0.949	0.951	0.955	0.979	3.227	3.155	2.923	2.604	2.195	1.861
$\nu = 7$	0.010	0.492	0.654	0.763	0.844	0.898	0.960	3.026	3.557	3.781	3.832	3.675	3.446
	0.025	0.690	0.796	0.849	0.890	0.914	0.953	3.511	3.796	3.788	3.612	3.378	3.172
	0.050	0.902	0.923	0.946	0.956	0.966	0.983	3.731	3.851	3.750	3.471	3.211	2.929
	0.100	0.914	0.944	0.966	0.976	0.982	0.994	3.768	3.837	3.758	3.471	3.212	2.933
$\nu = 12$	0.010	0.510	0.674	0.784	0.847	0.891	0.940	3.031	3.494	3.741	3.713	3.530	3.286
	0.025	0.726	0.824	0.874	0.875	0.901	0.924	3.547	3.798	3.774	3.509	3.287	3.000
	0.050	0.892	0.914	0.931	0.933	0.950	0.972	3.688	3.787	3.698	3.381	3.108	2.794
	0.100	0.899	0.924	0.945	0.956	0.966	0.982	3.792	3.834	3.753	3.465	3.172	2.868

Notes: This table displays the finite sample properties of the MCS for Var and ES forecasts evaluated using the FZG score and the set of $m = 5$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

C.3 Additional results - set of $m = 10$ models

Table C.7: Simulation results for the set of $m = 10$ models: VaR forecasts evaluated using the tick loss.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.569	0.773	0.884	0.940	0.976	0.988	6.399	8.090	8.763	8.903	8.452	6.877
	0.025	0.856	0.934	0.960	0.987	0.996	0.997	7.544	8.405	8.623	8.518	8.020	7.040
	0.050	0.919	0.959	0.983	0.996	1.000	0.999	7.587	8.138	8.103	7.871	7.370	6.024
	0.100	0.927	0.968	0.984	0.993	0.998	0.993	7.335	7.482	7.062	6.223	4.794	2.611
$\nu = 7$	0.010	0.523	0.747	0.872	0.944	0.981	0.994	5.893	7.518	8.384	8.611	8.310	7.544
	0.025	0.772	0.894	0.944	0.973	0.990	0.987	7.625	8.482	8.713	8.459	7.595	5.889
	0.050	0.855	0.922	0.971	0.985	0.991	0.987	8.172	8.686	8.804	8.408	7.480	5.416
	0.100	0.884	0.947	0.973	0.989	0.994	0.997	8.464	8.751	8.785	8.401	7.527	5.628
$\nu = 12$	0.010	0.574	0.790	0.912	0.964	0.978	0.986	5.814	7.354	8.026	8.163	7.823	7.026
	0.025	0.784	0.882	0.955	0.979	0.986	0.980	7.568	8.404	8.651	8.377	7.534	5.625
	0.050	0.846	0.919	0.958	0.980	0.974	0.967	8.227	8.686	8.693	8.145	6.858	4.600
	0.100	0.883	0.935	0.972	0.986	0.987	0.988	8.469	8.741	8.646	7.914	6.680	4.819

Notes: This table displays the finite sample properties of the MCS for VaR forecasts evaluated using the tick loss and the set of $m = 10$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

Table C.8: Simulation results for the set of $m = 10$ models: Var and ES forecasts evaluated using the AL score.

		Potency						Power					
	$p\backslash P$	63	126	251	500	1000	2500	63	126	251	500	1000	2500
$\nu = 3$	0.010	0.548	0.736	0.871	0.946	0.989	0.999	6.394	7.986	8.812	9.099	8.782	6.618
	0.025	0.755	0.869	0.936	0.983	0.996	0.998	7.632	8.522	8.938	8.989	8.362	5.725
	0.050	0.884	0.939	0.974	0.995	0.999	0.999	7.888	8.589	8.794	8.668	8.016	5.822
	0.100	0.920	0.967	0.986	0.997	0.999	0.995	7.850	8.287	8.288	8.022	7.286	4.212
$\nu = 7$	0.010	0.520	0.743	0.863	0.940	0.984	0.998	5.820	7.338	8.208	8.596	8.516	7.865
	0.025	0.774	0.893	0.953	0.980	0.993	0.995	7.619	8.452	8.711	8.363	7.477	5.524
	0.050	0.900	0.946	0.979	0.989	0.994	0.998	8.412	8.806	8.826	8.138	6.690	4.335
	0.100	0.905	0.956	0.979	0.992	0.994	0.998	8.577	8.842	8.776	8.075	6.794	4.624
$\nu = 12$	0.010	0.570	0.761	0.891	0.958	0.976	0.989	5.668	7.003	7.705	7.934	7.611	6.591
	0.025	0.811	0.914	0.961	0.978	0.983	0.982	7.533	8.287	8.401	7.826	6.609	4.607
	0.050	0.897	0.945	0.976	0.982	0.978	0.990	8.428	8.780	8.625	7.723	6.016	3.734
	0.100	0.901	0.949	0.980	0.987	0.987	0.986	8.520	8.753	8.579	7.572	6.000	3.894

Notes: This table displays the finite sample properties of the MCS for Var and ES forecasts evaluated using the AL score and the set of $m = 10$ models. The level of the test is α=0.25, the number of simulations is 2,500. Potency is the frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}$, power is defined as the average of $|\widehat{\mathcal{M}}^*_{1-\alpha}|$.

D Detailed discussion

D.1 Variance Formula

For random variables $X_i, i = 1, ..., m$ and real numbers $a_i, i = 1, ..., m$ it holds that

$$var(\sum a_i X_i) = \sum_{i=1}^{m} a_i^2 var(X_i) + 2 \sum_{i,j=1, i \neq j}^{m} a_i a_j cov(X_i, X_j). \tag{D.1}$$

For $var(D_{i\cdot}) = var(L_i - \frac{1}{m}\sum_{j=1}^{m} L_j)$, we get

$$\begin{aligned} var(D_{i\cdot}) &= var(L_i - \frac{1}{m}\sum_{j=1}^{m} L_j) \\ &= var\left(\frac{m-1}{m} L_i - \frac{1}{m} \sum_{j=1, j\neq i}^{m} L_j\right) \\ &= \left(\frac{m-1}{m}\right)^2 var(L_i) + \frac{1}{m^2} var(\sum_{j=1, j\neq i}^{m} L_j) - \frac{2(m-1)}{m^2} \sum_{j=1, j\neq i}^{m} cov(L_i, L_j) \\ &= \left(\frac{m-1}{m}\right)^2 var(L_i) + \frac{1}{m^2} \sum_{j=1, j\neq i}^{m} var(L_j) \\ &+ \frac{2}{m^2} \sum_{j=1, j\neq i, k=j+1, k\neq i}^{m} cov(L_j, L_k) - \frac{2(m-1)}{m^2} \sum_{j=1, j\neq i}^{m} cov(L_i, L_j). \end{aligned} \tag{D.2}$$

D.2 Simulating normally distributed losses

Figure D.1 below shows the individual rejection frequencies that relate to the simulations that we discuss in Section 3.3, when we simulate normally distributed losses.

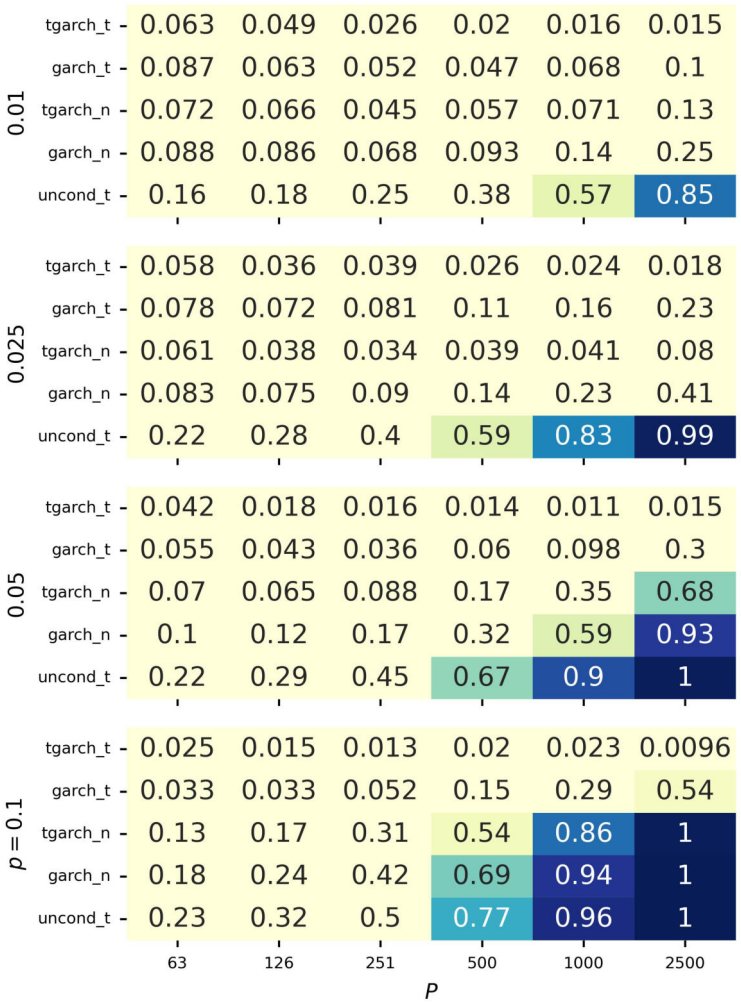

Figure D.1: Individual rejection frequencies, normally distributed losses, $\alpha = 0.25$

This figure displays the individual rejection frequencies of the candidate models in the MCS, number of models $m = 5$, level of the test $\alpha = 0.25$. We directly simulate normally distributed losses for which the expected value and covariance matrix correspond to the DGP with $\nu = 3$ degrees of freedom and VaR forecasts evaluated using the tick loss function. The results are for the MCS test using the $T_{max,\mathcal{M}}$ test.

D.3 Plots of $t_{i\cdot}$

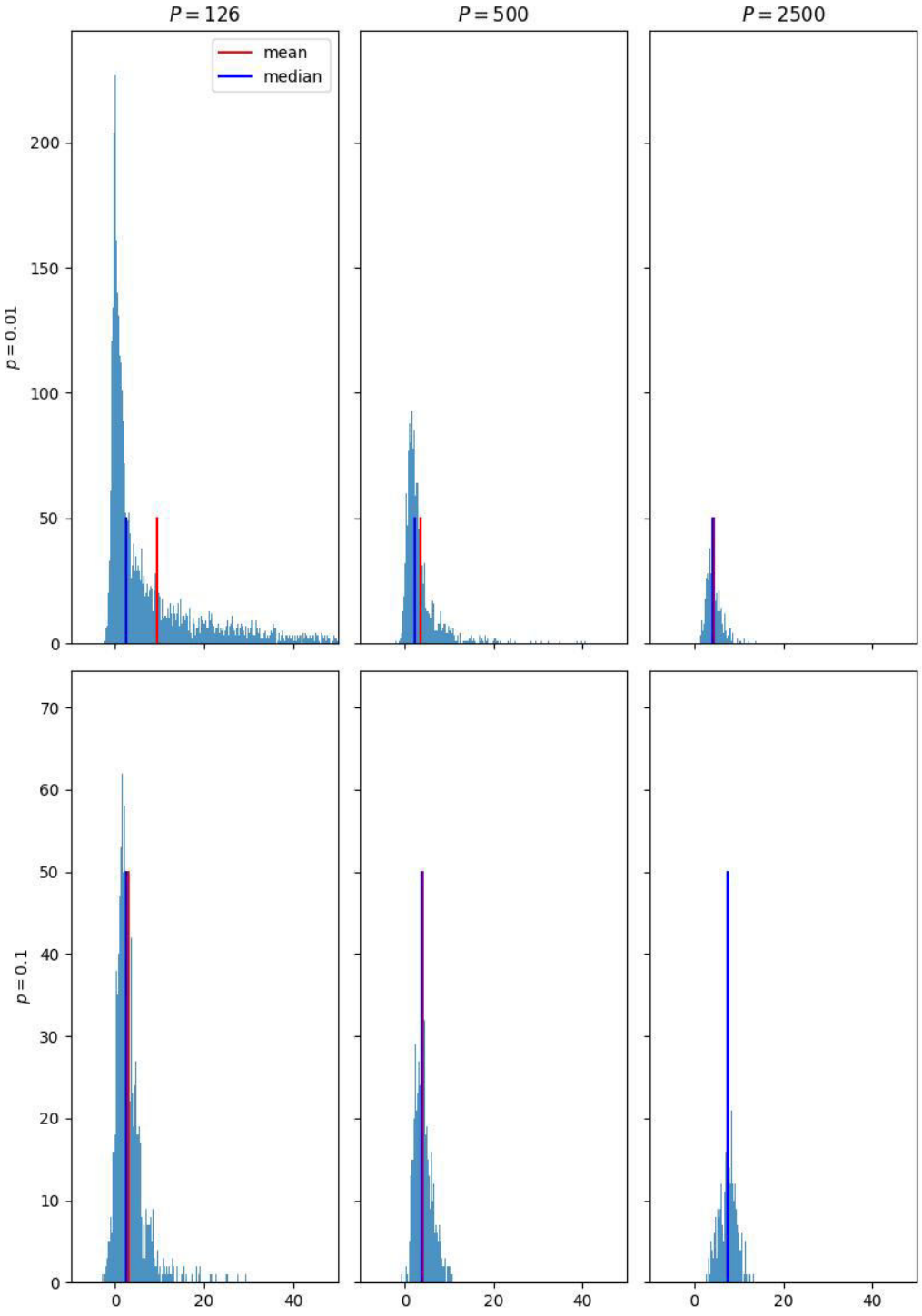

Figure D.2: Density plot of $t_{i\cdot}$

This figure shows density plots of $t_{i\cdot}$ based on the bootstrapped estimate $\widehat{var}(\bar{d}_{i\cdot})$ that the MCS test uses. Model i is the constant variance model. Scenario where $\nu = 3$, VaR forecasts evaluated using the tick loss function. The out-of-sample size P is displayed on top, the quantile level p on the left. Mind the different scales on the y-axes.

D.4 Sign of d_{ij}

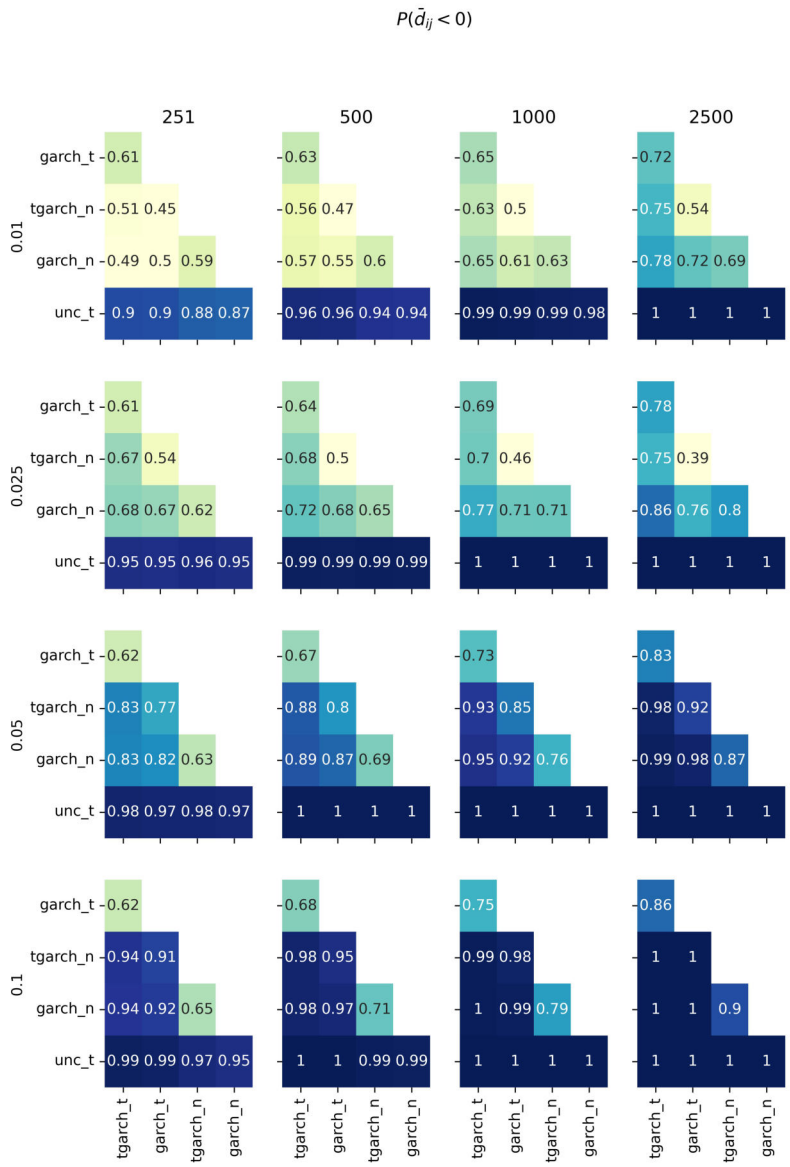

Figure D.3: Sign of the $\bar{d}_{ij}$

This figure displays how often we observe samples for which the average sample loss differential $\bar{d}_{ij} < 0.$, i.e. how often we find evidence that model i is superior to model j. Model i is on the x-axis, model j on the y-axis. Scenario where $\nu = 3$, VaR forecasts evaluated using the tick loss. The out-of-sample size P is displayed on top, the VaR level p on the left.

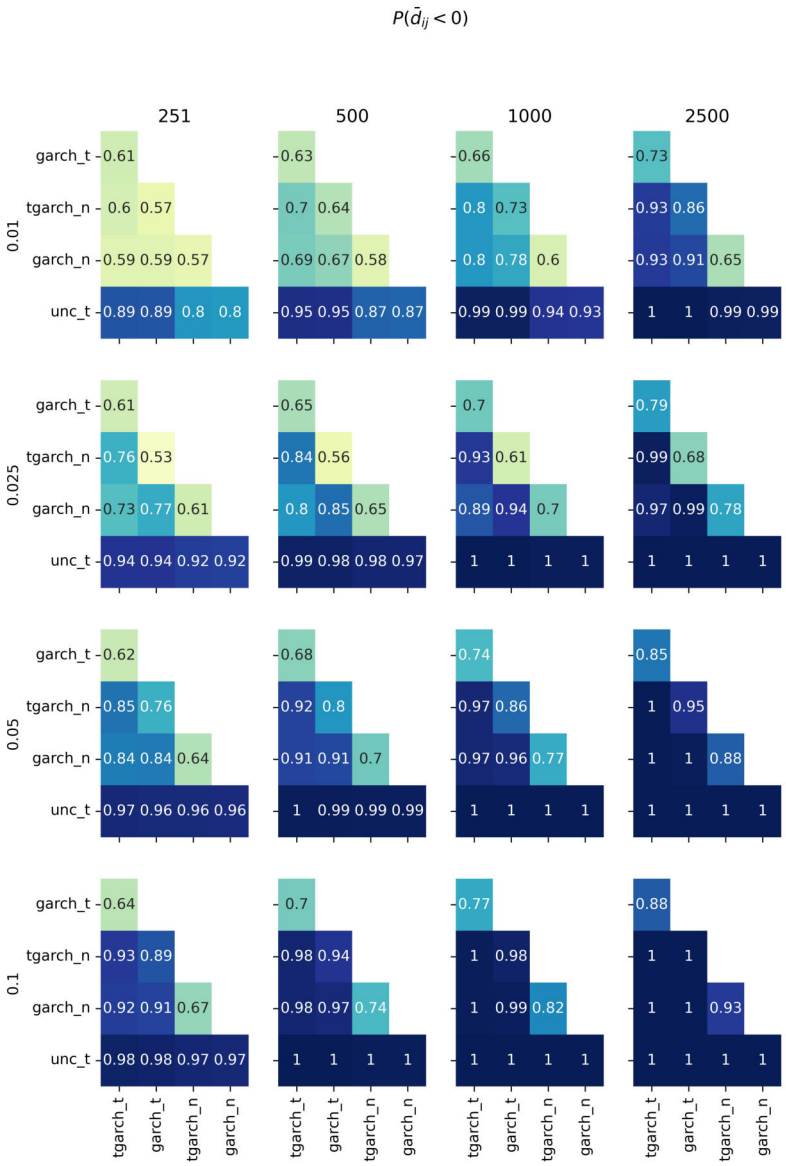

Figure D.4: Sign of the $\bar{d}_{ij}$

This figure displays how often we observe samples for which the average sample loss differential $\bar{d}_{ij} < 0.$, i.e. how often we find evidence that model i is superior to model j. Model i is on the x-axis, model j on the y-axis. Scenario where $\nu = 3$, VaR and ES forecasts evaluated using the AL loss. The out-of-sample size P is displayed on top, the VaR level p on the left.

D.5 Differences in rejection frequencies

Figure D.5 below displays ΔIRF and - in brackets - IRF_{cbb}, to put the difference into perspective. Each subplot relates to a different quantile p, shown on the left-hand side, while out-of-sample sizes P are on the x-axis, and the models are on the y-axis.

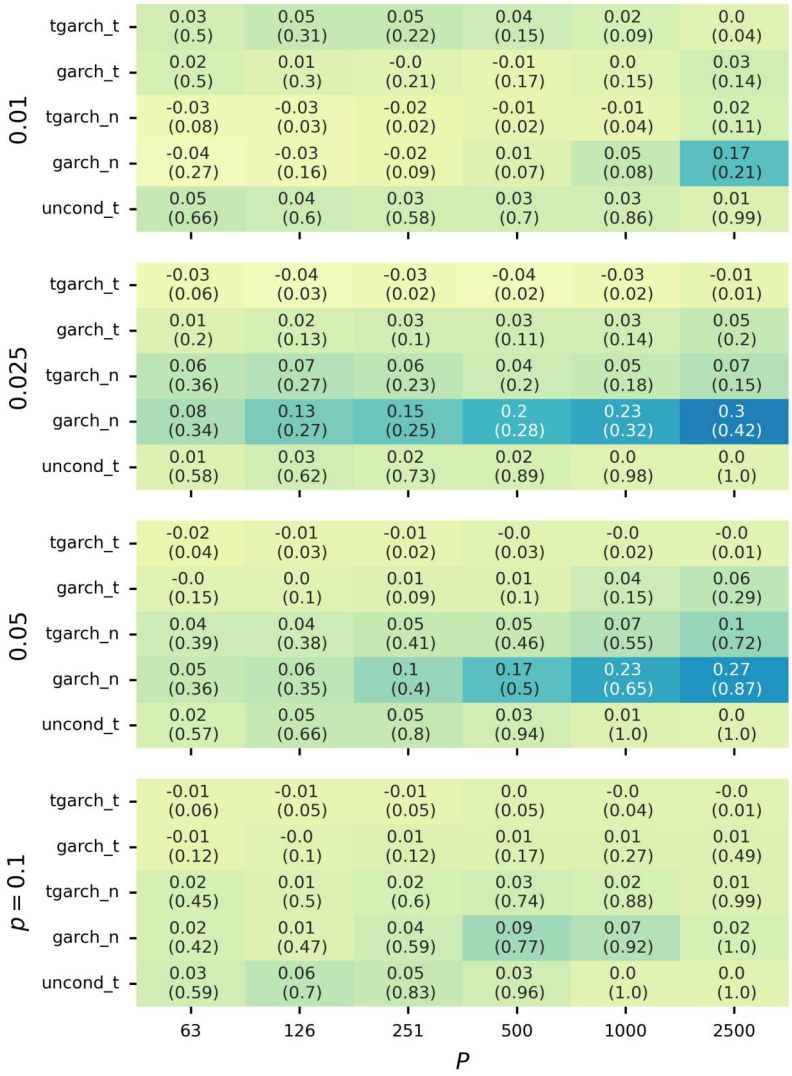

Figure D.5: Differences in individual rejection frequencies

This figure displays the differences in the individual rejection frequencies when using the bootstrapped estimate of $\widehat{var}(\bar{d}_{i\cdot})$ vs. using the expressions of $var(\bar{d}_{i\cdot})$ based on the simulated covariance matrix of $\bar{l}$. The numbers in brackets below show the rejection frequency when using the bootstrapped estimate of $\widehat{var}(\bar{d}_{i\cdot})$. Tick loss, number of models $m = 5$, level of the test $\alpha = 0.25$.