ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Bachelor Thesis Double Degree
Programme Econometrics & Economics

Backtesting VaR Estimates of HEAVY Models Using the Geometric-VaR Test

Name of Student:
S.R. Snijders

Name of Supervisor:
Prof. dr. D.J.C. van Dijk

Student ID Number:
414579

Name of Second Assessor:
S.H.L.C.G. Vermeulen

Date Final Version:
8 July 2018

Abstract

This paper examines the Geometric-VaR test of Pelletier and Wei (2016) as a framework for backtesting Value-at-Risk (VaR) estimates. This study confirms that the test provides good power properties against various forms of misspecification of VaR estimates, although slightly lower power is reported for smaller sample sizes compared to earlier research. The Geometric-VaR test is subsequently employed to investigate the HEAVY model of Shephard and Sheppard (2010) – an adaption of a standard GARCH model that incorporates realised measures – in the context of VaR estimation. Additionally, an asymmetric extension of the HEAVY model is introduced. 19 different models are tested using data of 21 equity indices over the period 2000-2017. A semi-parametric approach using Filtered Historical Simulation is found to provide better results than fully-parametric approaches. Additionally, this paper finds no evidence that the HEAVY models provide better VaR estimates than their GARCH counterparts over the entire sample period investigated. Notably, the HEAVY models perform significantly better during the global financial crisis of 2008, thus suggesting that they can be a valuable addition to a risk manager's toolkit during volatile periods.

1 Introduction

> *"Indeed, better risk management may be the only truly necessary element of success in banking."*
>
> — Alan Greenspan, *5 October 2004*

Since its introduction in the early 1990s and its consequent adoption in the Basel II accord, Value-at-Risk (VaR) has been the de facto industry standard for risk reporting in the financial sector (Jorion et al., 2007). Although its interpretation is clear-cut – daily VaR at coverage level p is defined such that the probability of one day's loss exceeding VaR equals p – the calculation of VaR brings several challenges with it and can be performed in various ways. Indeed, current bank regulations only provide rough guidelines on how VaR calculations should be implemented rather than a unified framework that should be adopted. However, current rulings prescribe the amount of capital banks are required to hold based on the daily reported VaR levels. As VaR estimates – and hence capital requirements – differ across estimation methods, model selection remains a relevant topic from a financial institution's perspective. In addition, backtesting the employed models is relevant for financial practitioners as well as regulators in order to verify whether risks are reported correctly.

The aim of this paper is twofold: (1) to examine the Geometric-VaR test of Pelletier and Wei (2016) as a framework for backtesting VaR estimates; (2) to establish whether utilising high-frequency return data by means of the HEAVY specification of Shephard and Sheppard (2010) can improve VaR estimation.

VaR backtests are statistical tests devised to reject VaR models that are unable to provide forecasts that satisfy certain conditions, such as correct unconditional coverage or serial independence.[1] Recently, Pelletier and Wei (2016), henceforth 'PW16', have added to the literature of VaR backtesting by devising a duration-based test that combines the ideas of the Geometric test of Berkowitz et al. (2011) and the CaViaR test of Engle and Manganelli (2004). The authors perform a series of simulation studies and claim increased power of their test compared to competing methods. Because of its recent publication, the Geometric-VaR test has not yet been applied in other research. Hence, this paper verifies the results obtained by PW16 and provides additional benchmarking before using it as a tool to discriminate between competing VaR estimation models

[1] Relevant conditions are defined in detail in further sections.

using a large set of equity indices. The results obtained in this research are similar to PW16, both quantitatively and qualitatively; indeed, it is found that the Geometric-VaR test provides increased power over competing backtesting procedures in general. However, this paper especially establishes the merits of the test in case of large sample sizes. For smaller sample sizes, the Geometric-VaR test does not always outperform; a fact remarked in PW16 but observed to a larger extent in the simulations of this research. This effect is especially apparent in case of negative leverage effects and low volatility persistence. However, it is evident that the Geometric-VaR test provides an integrated testing framework that is able to reliably detect various forms of VaR misspecification. Hence, this paper adopts the Geometric-VaR test as the approach to discriminate between specific VaR estimation models.

Clearly, the modelling of Value-at-Risk and thus volatility has been a major topic of interest amongst econometricians. Although volatility itself is not directly observable and hence no 'true' model for asset volatility has yet been established, the family of generalized autoregressive conditional heteroscedasticity (GARCH) models introduced by Bollerslev (1986) is widely considered as the leading specification in current academic practice. Still, in modelling the conditional variance process, GARCH models rely only on daily return information. The increased availability of financial intra-day data has however facilitated the advent of a new class of estimators, so-called 'realised measures' that gauge daily volatility using intra-day data leading to less noisy estimates of market volatility. As these measures provide more accurate estimates of volatility, it is hypothesised that they can improve conditional volatility modelling and thus VaR estimation. Although various attempts have been made to incorporate realised measures into conditional heteroscedasticity models, few do so in a direct fashion. Shephard and Sheppard (2010) however introduce the high frequency-based volatility (HEAVY) model, an adaption of the standard GARCH(1,1) model that directly incorporates realised measures into the conditional variance specification. Since the original paper demonstrates advantages of such a model in a general context, this research tests whether HEAVY models have any merit in the context of VaR estimation. Additionally, this paper contributes by generalising the concept of HEAVY to an asymmetric specification to allow for nonlinearities that are often observed in real-world asset returns.

The HEAVY models are investigated in a fully-parametric setting by varying distributional assumptions and a semi-parametric setting by means of Filtered Historical Simulation (FHS). As realised measures, both 'realised kernel' and 'realised variance' are examined. Various benchmark

models such as Historical Simulation and GARCH specifications are employed to verify whether indeed the addition of realised measures leads to better VaR estimates. Notably, the application of Filtered Historical Simulation seems to improve VaR estimates considerably compared to the fully-parametric approach for all specifications. However, no outperformance is observed for the HEAVY models compared to their GARCH counterparts over the entire period investigated. Overall, the GJR-GARCH model combined with FHS appears to perform best amongst all investigated specifications for the entire sample. Additionally, the HEAVY models do exhibit outperformance during the period of the global financial crisis.

The remainder of this paper is structured as follows: section 2 provides an overview of the existing body of literature on the topic. Section 3 discusses the methods employed to estimate VaR, while section 4 provides technical details on the Geometric-VaR test. Section 5 presents an overview of the data used in this research. Section 6 re-examines the Geometric-VaR test and its properties by means of simulation. Section 7 examines the performance of the HEAVY model using the Geometric-VaR test. Section 8 summarises and concludes. Lastly, section 9 provides a discussion and avenues for further research.

2 Literature

2.1 VaR Estimation

Because of the prominence of VaR as a risk management tool in the global financial system, a plethora of literature has been devoted to its estimation. In particular, the adoption of VaR as the standard for risk reporting in the Basel II accord has spurred academic research in the field. Clearly, one of the easiest and most widely applied methods to estimate Value-at-Risk is the non-parametric approach called Historical Simulation (HS) (Pérignon and Smith, 2010; Escanciano and Pei, 2012). The Historical Simulation method does not rely on any distributional assumptions with regards to the returns and estimates VaR by taking the appropriate percentile of the past N returns. Despite its easiness of use, the Historical Simulation method has several evident drawbacks. For instance, it does not take the predictability of volatility into consideration and the time-varying characteristics of volatility are only reflected by the shift in the rolling window. In addition, Pritsker (2006) demonstrates that Historical Simulation is severely under-responsive to changes in conditional risk.

More sophisticated approaches that attempt to account for the time-varying nature of volatility can be traced back to the influential work on autoregressive conditional heteroscedasticity (ARCH)

models by Engle (1982) and the subsequent GARCH generalisation by Bollerslev (1986). Both models account for time-varying volatility by modelling the conditional variance of the return process explicitly (as an AR or ARMA process respectively). Although other specifications have been developed, the GARCH model family is clearly established as the workhorse method to measure, model and forecast volatility in financial econometrics.

However, when looking at Value-at-Risk in a daily setting, GARCH models are designed such that their information set is based on returns at a daily level. This is problematic as a single return only provides a limited and noisy signal about the level of current volatility. This causes standard GARCH models to be ill-equipped in situations where rapid changes in volatility level might occur. Andersen et al. (2003) discuss that the standard GARCH model is slow in responding to sudden changes in volatility and takes many periods for its conditional variance to reach its desired level.

Recently, the increased availability of high-frequency return data and advancements in computational capabilities has led to a class of estimators of daily volatility based on intra-day prices called 'realised measures' which aim to provide improved signals of current levels of volatility. One of the earliest and most prominent examples of such realised measures was developed by Andersen et al. (2001b) which they name 'realised volatility'. Although the computation of the realised volatility is fairly straightforward (it averages intra-day squared returns at specified intervals), the authors and the concurrent research of Barndorff-Nielsen and Shephard (2002) show that under certain regularity conditions it is an unbiased and efficient estimator of return volatility. However, as pointed out by Hansen and Lunde (2006), high-frequency data often exhibits microstructure noise, which might cause the regularity conditions of the realised variance estimator to be violated. Hence, competing realised measures have been constructed in order to mitigate these issues, such as the 'realised kernel' approach of Barndorff-Nielsen et al. (2008). Alternatives are provided by the multiscale estimators of Zhang et al. (2005) and the pre-averaging method of Jacod et al. (2009).

Although attempts have been made to utilise high-frequency information to forecast volatility, most approaches have focused on fitting a standard time-series model on a sequence of daily squared returns, using realised measures as regressors (Andersen et al., 2001a,b, 2003, 2007). Also, estimation of GARCH models that additionally include information of realised measures (GARCH-X) has been attempted (see for instance Engle (2002) and Forsberg and Bollerslev (2002)).

In Shephard and Sheppard (2010), a parsimonious but novel approach to utilising high-frequency information is introduced by directly augmenting the original GARCH(1,1) model of Bollerslev

(1986) with realised measures. Specifically, the authors devise the HEAVY model, an abbreviation for 'High-frEquency-bAsed VolatilitY models', in which they replace the ARCH term, ϵ^2, of a standard GARCH(1,1) model by its realised measure counterpart. According to the authors, the HEAVY model is characterised by attractive momentum and mean reversion effects, as well as its ability to quickly adjust to structural breaks in the level of the volatility process, hence making it particularly interesting to investigate in the context of volatility and thus VaR modelling.

The above GARCH and HEAVY models can be used in a fully-parametric setting to forecast Value-at-Risk. However, models based on theoretical distributions often do not optimally reflect the empirical characteristics of asset returns, for instance, because of their inability to exhibit fat tails or volatility clustering (Nieto and Ruiz, 2016). In addition, as discussed, non-parametric methods such as Historical Simulation have shown to be poor estimates of VaR. An interesting semi-parametric method is provided by both Barone-Adesi et al. (1997) and Hull and White (1998) and is called Filtered Historical Simulation. The method uses explicit model estimates of the conditional variance to scale historical returns and thereby creates a normalised empirical distribution of asset returns that retains the distributional characteristics of the investigated asset. Subsequently, it uses the point forecast of the conditional volatility to 'scale up' the relevant percentile of the normalised empirical distribution to a VaR estimate.

Recent reviews of VaR estimation methodologies, such as Abad et al. (2014), have hinted at the potential of both realised measure based models, such as HEAVY, as well as Filtered Historical Simulation separately. Consequently, this paper adds to the existing body of knowledge by investigating the combination of the HEAVY model specification with the method of Filtered Historical Simulation in the context of estimating VaR — a variant that has not been rigorously investigated before to the knowledge of the author. In addition, a 'GJR inspired' asymmetric extension of HEAVY is introduced and investigated.

2.2 VaR Backtesting

Along with the proliferation of papers on VaR estimation methods in financial econometrics, backtesting VaR estimates has been an area of considerable attention. To ensure financial institutions are adequately capitalised in case of crises or other unexpected market events, validation of risk model estimates is crucial. This section provides an overview of the developments of literature on backtesting procedures, starting with tests based on the so-called 'violation' or 'hit' sequence. It is convenient to define a 'violation' or 'hit' as an event where a loss is observed that is greater

than the VaR estimate for that day. In particular, let us for the moment abuse notation[2] and define $I_t^q = 1(r_t < -\text{VaR}_t(q)),\ t = 1 \dots T$ with $1(\cdot)$ being the indicator function, r_t a daily return and q the VaR coverage level (i.e. 0.05 or 0.01). This paper follows convention by presenting VaR as a positive number. Engle and Manganelli (2004) argue that 'good' VaR forecasts should satisfy the following properties: (1) correct unconditional coverage should be provided; (2) violations should be independent; (3) violations should be uncorrelated with any information up to and including $t-1$. Condition (1) can be expressed as $\mathbb{E}[I_t^q] = q$. One of the first measures constructed to test VaR estimates was devised by Kupiec (1995) and is often referred to as the unconditional coverage test or simply the Kupiec test. Kupiec demonstrates that, if one assumes a constant probability of a violation occurring, the number of hits $\sum^N I_t^q$ follows a simple binomial distribution. Hence, the null hypothesis can be tested easily by means of likelihood ratio. However, this test evidently ignores property (2) and as a consequence VaR estimates that pass the test of Kupiec can exhibit violation clustering over time. As stressed by Lopez (1999), this is clearly not a desired situation – any VaR estimate that does not correct for information of increased hit probability in a subsequent period is by definition suboptimal. Additionally, Escanciano and Pei (2012) demonstrate that the test of Kupiec is always inconsistent in case of forecasts obtained using Historical Simulation.

Christoffersen (1998) in turn developed the conditional coverage test (CC) which tests $H_0 : \mathbb{E}[I_t^q | I_{t-1}^q] = q$ and hence assesses property (1) and (2) simultaneously. Implementation of the test is done by means of likelihood ratio, of which the statistic is obtained by addition of two separate LR statistics; the unconditional (Kupiec) test statistic and the independence statistic, the latter of which is computed based on testing if I_t^q are i.i.d. Bern(q) distributed against the alternative of first-order Markov dependence. Importantly, Christoffersen (1998) only incorporates first-order autocorrelation of the hit sequence in its test. Christoffersen and Pelletier (2004) consequently show that the test is not suitable in a variety of regular settings and exhibits poor finite sample behaviour. Notwithstanding these shortcomings, Candelon et al. (2010) report it is still one of the most frequently used backtests in practice.

The aforementioned tests are both conducted by converting return information and the VaR estimates into a single binary hit sequence. In addition, the test of Kupiec only uses the number of violations as input, while the CC test of Christoffersen is limited by incorporating only infor-

[2]Technically, VaR is defined as the loss expressed in absolute dollar terms, rather than as a return. However, throughout this paper the convention of expressing VaR as an absolute return is adopted because of notational convenience.

mation regarding the first order autocorrelation of the binary hit sequence and thus ignores more general forms of violation dependence. Clearly, potential power is lost by such simplifications. Consequently, Christoffersen and Pelletier (2004) provide an alternative by devising a test that looks at the duration between violations. The intuition of such duration-based tests is that clustering of violations generates a large number of both relatively short and relatively long periods without violations. Let us define a no-hit duration variable D_i as $t_i - t_{i-1}$, the difference in time between two consecutive hits. Christoffersen and Pelletier (2004) postulate that under the null hypothesis of correct VaR specification, the no-hit durations should have no memory and a mean duration of $1/q$ observations. Hence, the durations under the null hypothesis are modelled as exponentially distributed, with the alternative hypothesis being a (continuous) Weibull specification. The test is then carried out by means of likelihood ratio but the authors mention that the small sample sizes that are often observed in practice as well as testing on the boundary of the parameter space warrant caution when using the asymptotic chi-squared critical values. Hence, they propose the application of the Monte Carlo method of Dufour (2006) in order to control test sizing under these conditions. Haas (2006) further examines the duration-based test of Christoffersen and Pelletier (2004) and notes the peculiarity of testing durations, discrete by nature, by means of continuous distributions. Consequently, Haas (2006) replaces the exponential distribution by the geometric distribution and applies the discretized Weibull variant of Nakagawa and Osaki (1975) as a replacement for the distribution under the alternative hypothesis. In all cases, an improvement in testing power is observed for the discretized version compared to the original continuous specification of Christoffersen and Pelletier (2004).

Note that all tests considered so far at maximum take into account (1) correct unconditional coverage and (2) violation independence. As mentioned, Engle and Manganelli (2004) contemplate that this is a necessary, but not a sufficient condition. Indeed, let us entertain the following thought experiment: generate a sequence of i.i.d. drawings $\{x_t\}_{t=1}^{T}$ with $x_t = K$ with probability q and $x_t = -K$ with probability $(1 - q)$ that serves as our VaR series. Setting K sufficiently large generates an adequate risk model for any real-world return sequence under requirements (1) and (2). The moment x_t is generated, the probability of a violation is known beforehand to be close to 0 or 1. Hence, Engle and Manganelli (2004) claim that adequate VaR forecasts should additionally satisfy a third property: violations should be uncorrelated with any information up to and including $t - 1$. They develop the regression-based 'dynamic quantile' (DQ) test, which is able to assess a wide range of violation dependence on past information. Berkowitz et al. (2011) implement the

DQ test in parsimonious form by using the one-period lagged VaR estimates and the lagged hit variable as regressors and assuming a logit distribution for the error terms. This specific form of the DQ test is referred to by the authors as the 'CaViaR' test. Berkowitz et al. (2011) also expand on previous duration-based testing literature by acknowledging that they can model durations as geometric variables, but with varying rather than constant probabilities to test the alternative hypothesis. The test is referred to as the 'Geometric test' and has the benefit of not having to resort to continuous distributions as in their previous works (Christoffersen and Pelletier (2004)). They compare their new Geometric test specification with the CaViaR test and find that the CaViaR test works best overall, but their duration-based test also performs well in many cases. Clearly, the Geometric test solves previous issues such as the discreteness/continuous mismatch. However, it still only tests requirement (1) and (2) indicated by Engle and Manganelli (2004).

Most recently, Pelletier and Wei (2016) have extended the existing duration-based testing literature by providing a combination of their Geometric test and the so-called 'VaR test' (an adaption inspired by the CaViaR test) which they appropriately call the 'Geometric-VaR test'. By incorporating the estimated Value-at-Risk of the relevant observation in the hazard function, they are able to test all three requirements of Engle and Manganelli (2004) simultaneously. The authors find that their Geometric-VaR test provides better power than other duration-based tests and regression-based tests such as the CaViaR test, and has power against various forms of misspecifications.

This paper recognises the potential of the Geometric-VaR test, but also notes that because of its recent introduction its application has thus far been limited to PW16. Hence, this research adds to the existing body of literature by implementing the test of PW16 independently of the authors and re-evaluating its merits. Specifically, its performance compared to other more established VaR backtesting procedures is examined. In addition, whereas other papers on VaR forecasting methods often use low-power tests such as those of Kupiec (1995) and Christoffersen (1998) to compare and contrast models, this paper uses the more powerful Geometric-VaR test to differentiate between competing model specifications.

3 Methodology for VaR Forecasting

To test the merit of realised measures in estimating VaR, this paper investigates several specifications, along with a number of benchmark models. An overview is provided in Table 1. In particular, the HEAVY model class of Shephard and Sheppard (2010) is tested against Historical Simulation

Table 1: Investigated models

	Non-Parametric	Normal	Student's t	Filtered Historical Simulation
Historical Simulation	HS			
HEAVY RV-based		HNv	HTv	HFv
HEAVY RK-based		HNk	HTk	HFk
GJR-HEAVY RV-based		gHNv	gHTv	gHFv
GJR-HEAVY RK-based		gHNk	gHTk	gHFk
GARCH(1,1)		GN	GT	GF
GJR-GARCH		gGN	gGT	gGF

Note. This table provides an overview of all model combinations investigated in the paper including acronyms. 'RV' denotes Realised Variance, 'RK' denotes Realised Kernel.

(HS), because of its prevalence amongst industry practitioners[3], and the workhorse GARCH(1,1) and GJR-GARCH models. HS computes VaR directly, whereas HEAVY and GARCH estimate the conditional variance of returns. Consequently, for HEAVY and GARCH, a translation from conditional variance to VaR is needed. This paper employs both a fully-parametric approach using Normal and Student's t distributions and a semi-parametric approach using Filtered Historical Simulation. This section describes the implementation of all forecasting methods investigated.

3.1 Preliminaries

Let us denote daily financial asset returns as $r_1, r_2, \dots, r_T$ and $\mathcal{F}_{t-1}^{LF}$ as the information set containing all available daily return information up to and excluding time t, which we call the 'low-frequency' dataset. In addition, define the conditional variance $\sigma_t^2 \equiv \mathbb{V}(r_t|\mathcal{F}_{t-1}^{LF})$. Next, let $X_{t_{j,t}}$ be the log-price of an asset with $t_{j,t}$ being the times of trades on day t. Consequently, define $x_{j,t} = X_{t_{j,t}} - X_{t_{j-1,t}}$ such that $\{x_{j,t}\}_{j=2}^{N}$ is a collection of intra-day 'returns'. We then call $\mathcal{F}_{t-1}^{HF}$ the 'high-frequency' dataset, which includes all intra-day returns $x_{j,i}$ and daily returns r_i up to and excluding time t. Analogously, define $h_t \equiv \mathbb{V}(r_t|\mathcal{F}_{t-1}^{HF})$.

Following similar literature, define VaR with coverage rate q as the q-th quantile of the conditional distribution of r_{t+1}:

$$\text{VaR}_{t+1}(q) \equiv -F_{t+1}^{-1}(q) \tag{1}$$

with F_{t+1} denoting the conditional distribution of r_{t+1}.

[3]In a survey, Pérignon and Smith (2010) find that 73% of the banks that disclosed their VaR forecast method used Historical Simulation.

3.2 Historical Simulation

Using the Historical Simulation method, VaR estimates are computed directly as an empirical estimate of F_{t+1} over a specified time window:

$$\text{VaR}^{\text{HS}}_{t+1}(q) = -\text{percentile}(\{r_s\}^t_{s=t-T_e+1}, 100 \times q) \tag{2}$$

with T_e being the size of the rolling window used to approximate the conditional distribution of the returns.

3.3 (Semi-)Parametric Models

3.3.1 GARCH(1,1)

As a useful benchmark, the family of GARCH models is employed to estimate the conditional variance of asset returns. Specifically, let us use the standard GARCH(1,1) model as introduced in the seminal paper by Bollerslev (1986):

$$r_t = \epsilon_t \tag{3}$$

$$\epsilon_t | \mathcal{F}^{LF}_{t-1} \sim \mathcal{N}(0, \sigma^2_t) \tag{4}$$

$$\sigma^2_t = \omega + \alpha \epsilon^2_{t-1} + \beta \sigma^2_{t-1} \tag{5}$$

with the parameter restrictions $\omega \geq 0, \alpha \geq 0, \beta \in [0, 1)$ to ensure positive variance and stability. As financial asset returns often exhibit fat-tails, this paper also employs a t-GARCH(1,1) model in which the distribution of the innovation is given by a Student's t distribution such that $\epsilon_t | \mathcal{F}^{LF}_{t-1} \sim t(\nu, \sigma^2_t)$, which denotes a distribution with ν degrees of freedom and a (scaled) variance of σ^2_t.

3.3.2 GJR-GARCH

To allow for an asymmetric leverage effect often found in asset returns (Cont, 2001), let us also examine the specification of Glosten et al. (1993):

$$\sigma^2_t = \omega + \alpha \epsilon^2_{t-1} + \gamma \epsilon^2_{t-1} I_{t-1} + \beta \sigma^2_{t-1} \tag{6}$$

in which $I_{t-1} = 0$ if $\epsilon_{t-1} \geq 0$, and $I_{t-1} = 1$ if $\epsilon_{t-1} < 0$. Again, both $\epsilon_t | \mathcal{F}^{LF}_{t-1} \sim \mathcal{N}(0, \sigma^2_t)$ and $\epsilon_t | \mathcal{F}^{LF}_{t-1} \sim t(\nu, \sigma^2_t)$ are employed. In addition, the following restrictions are imposed: $\omega \geq 0, \alpha \geq 0, \beta \geq 0$, $\gamma + \alpha > 0$ and $\alpha + \frac{1}{2}\gamma + \beta < 1$ to ensure positive variance and stability.

3.3.3 Models Based on Realised Measures

Although the GARCH approach to volatility modelling has been widely adopted, it is limited by the fact that it uses daily data and can hence be slow to adapt because the squared returns provide

noisy estimates of volatility. The HEAVY model of Shephard and Sheppard (2010) attempts to improve volatility forecasting by providing a parsimonious functional form to model the conditional variance that includes high-frequency intra-day data ($\mathcal{F}_t^{HF}$). Let us start by defining the 'realised measure' RM_t as an intra-day based estimate of volatility that will be employed in the modelling of conditional daily asset volatility. One of the most commonly utilised versions of the realised measure is the 'realised variance', which is simply computed as

$$\text{RM}_t^{\text{RV}} = \sum x_{j,t}^2 \tag{7}$$

However, high-frequency data often exhibits microstructure noise (Hansen and Lunde, 2006), which might cause consistency problems in case of the simple realised variance estimator. To counteract this, this paper replaces the tick-by-tick returns $x_{j,t}$ in (7) by intra-day 5-minute returns to compute the realised variance. Additionally, the statistic is sub-sampled every 30 seconds. That is, without loss of generality, let $s_1 = 0$ be the starting time in seconds of the estimation sample for the first realised variance measure $\text{RM}^{\text{RV},1}$. The set of intra-day 5-minute returns for this measure are then calculated by the log difference of the last tick prices at times $g_{1,w}$ and $g_{1,w-1}$ in seconds with $g_{1,w} = s_1 + (5 \cdot 60)w$ and $w = 1, \dots, M$ with M being the number of minutes in a trading day divided by 5. Now, define 10 of such measures: $\text{RM}^{\text{RV},i}, i \in \{1, \dots, 10\}$ and set $s_i = 30(i-1)$ such that the measures are spaced 30 seconds apart. Now calculate the subsampled version of the realised variance as $\text{RM}^{\text{RV}} = \frac{1}{10} \sum_{i=1}^{10} \text{RM}^{\text{RV,i}}$.

In addition to the realised variance, a more sophisticated approach is the 'realised kernel' estimator suggested by Barndorff-Nielsen et al. (2008) and further specified for practical purposes in Barndorff-Nielsen et al. (2009). Instead of subsampling, the realised kernel approach mitigates consistency issues due to microstructure noise by means of a kernel weighting function. The estimator is specified as follows:

$$\text{RM}_t^{\text{RK}} = \sum_{h=-H}^{H} k(h/(H+1))\gamma_h, \qquad \gamma_h = \sum_{j=|h|+1}^{n} x_{j,t} x_{j-|h|,t} \tag{8}$$

where $k(u)$ is a kernel weighting function, H is the bandwidth[4] and γ_h is the h-th realized autoco-

[4]In the computation of the realised kernel the exact approach of Barndorff-Nielsen et al. (2009) is employed in setting the appropriate values for H and the reader is referred to said paper for further details.

variance. Let the kernel weights be specified by the Parzen function:

$$k(u) = \begin{cases} 1 - 6u^2 + 6u^3, & 0 \leq u \leq 1/2 \\ 2(1-u)^3, & 1/2 < u \leq 1 \\ 0, & u > 1 \end{cases} \tag{9}$$

which benefits from various desirable properties over competing kernel weighting specifications in the setting of computing realised measures, such as non-negativity (Barndorff-Nielsen et al., 2009).

3.3.4 Standard HEAVY

Having defined the realised measures allows the specification of the 'standard' HEAVY model introduced by Shephard and Sheppard (2010) as follows:

$$h_t = \omega + \alpha \mathrm{RM}_{t-1} + \beta h_{t-1} \tag{10}$$

with the restrictions $\omega \geq 0, \alpha \geq 0$ and $\beta \in [0, 1)$, again for positive variance and stability. Observe that this specification is closely related to the GARCH(1,1) model, but ϵ_{t-1}^2 being replaced with the realised measure RM_{t-1}. Note that in this original specification, the model does not make any explicit assumptions with regards to the exact distribution of the return process and hence the normal-based maximum likelihood can be treated as a 'quasi' maximum likelihood estimator. However, the Student's t-distribution is also investigated as an alternative in this study.

3.3.5 GJR-HEAVY

Analog to the GJR-GARCH model, this paper introduces the 'GRJ-HEAVY' specification. It is closely related to its GARCH counterpart in terms of its ability to capture asymmetric leverage effects. However, as in the standard HEAVY model, our daily return information is replaced by realised measures. The model is formalised as follows:

$$h_t = \omega + \alpha \mathrm{RM}_{t-1} + \gamma \mathrm{RM}_{t-1} I_{t-1} + \beta h_{t-1} \tag{11}$$

in which $I_{t-1} = 0$ if $r_{t-1} \geq 0$, and $I_{t-1} = 1$ if $r_{t-1} < 0$. Additionally, let us impose restrictions similar to the GJR-GARCH specification: $\omega \geq 0, \alpha \geq 0, \beta \geq 0$, $\gamma + \alpha > 0$ and $\alpha + \frac{1}{2}\gamma + \beta < 1$.

3.3.6 Parameter and Conditional Variance Estimation

Parameter estimation for both the GARCH and HEAVY models is carried out by means of common (quasi) maximum likelihood procedures. In case of the GARCH models the following

log-likelihood function is utilised:

$$log\, Q_{\mathcal{N}}(\psi) = -\frac{T}{2} log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(log\, \sigma_t^2 + r_t^2/\sigma_t^2) \tag{12}$$

for the normal error distributions and

$$log\, Q_t(\psi, \nu) = T log\left[\frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}\right] - \frac{1}{2}\sum_{t=1}^{T} log\, \sigma_t^2 - \left(\frac{\nu+1}{2}\right)\sum_{t=1}^{T} log\left[1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right] \tag{13}$$

for the t-distributed cases, letting $\psi = (\omega, \alpha, \beta)$ for the symmetric specification and $\psi = (\omega, \alpha, \gamma, \beta)$ for GJR specification. Due to the similarities in model structure, it is also possible to employ (12) and (13) to estimate the parameters of the standard HEAVY and GJR-HEAVY specifications by replacing σ^2 with h_t. In all instances, this paper sets either σ_1^2 or h_1 to be $T^{-1/2}\sum_{t=1}^{\lfloor T \rfloor^{1/2}} r_t^2$.

For all models, estimation is done on a rolling window basis of size $T_e \in \{250, 1000, 1500\}$ (corresponding to 1, 4 and 6 years of data respectively) and hence one-day ahead forecasts of $\hat{\sigma}_{t+1}^2$ and $\hat{h}_{t+1}$ are easily computed with (5), (6), (10) and (11) using the estimated parameters $\widehat{\psi}$ in the normal cases or $\widehat{\psi}$ and $\hat{\nu}$ for the t-distributed cases.

3.3.7 VaR Estimation

Having estimated the conditional variance, there are two methods to compute the Value-at-Risk: the fully-parametric approach and the semi-parametric approach utilising Filtered Historical Simulation. The first method is carried out as follows: having obtained the estimated conditional variance series, the VaR estimates for $t = (T_e + 1), \ldots, T$ in case of normally distributed errors follow almost trivially using

$$\text{VaR}_{t+1}^{\text{LF}}(q) = -z_q \cdot \hat{\sigma}_{t+1} \qquad \text{and} \qquad \text{VaR}_{t+1}^{\text{HF}}(q) = -z_q\sqrt{\hat{h}_{t+1}} \tag{14}$$

with z_q being the critical value of the Normal distribution at the appropriate level q. This paper employs $q = 0.01$ and 0.05 as is usual in similar literature. For the models using a t-distribution, compute

$$\text{VaR}_{t+1}^{\text{LF}}(q) = -\hat{\sigma}_{t+1}\left(\frac{\hat{\nu}_{t+1}-2}{\hat{\nu}_{t+1}}\right)^{\frac{1}{2}} \tau_{q,\hat{\nu}_{t+1}} \quad \text{and} \quad \text{VaR}_{t+1}^{\text{HF}}(q) = -\sqrt{\hat{h}_{t+1}}\left(\frac{\hat{\nu}_{t+1}-2}{\hat{\nu}_{t+1}}\right)^{\frac{1}{2}} \tau_{q,\hat{\nu}_{t+1}} \tag{15}$$

with $\tau_{q,\hat{\nu}_{t+1}}$ the critical value of the t-distribution at coverage probability q and $\hat{\nu}_{t+1}$ degrees of freedom with $\hat{\nu}_{t+1}$ being estimated in the same maximum likelihood estimation as for $\widehat{\psi}$ with information of window size T_e up to and including t.

In addition to the parametric approach described above, let us combine our conditional variance

point forecasts $\hat{\sigma}^2_{t+1}$ and $\hat{h}_{t+1}$ of the GARCH and HEAVY models with the so-called 'Filtered Historical Simulation' technique inspired by Hull and White (1998). FHS constructs a standardised empirical distribution to compute critical values for VaR construction, rather than the theoretical Normal or t-distribution. Let $\{r_s\}^t_{s=t-T_e+1}$ be the set of relevant returns in the current rolling window up to and including time t. Computing $\widehat{\psi}$ using maximum likelihood yields an estimated conditional variance $\tilde{\sigma}^2_i$ or $\tilde{h}_i$ for $i = (t - T_e + 1) \dots t$ within the rolling window, as well as a point forecast $\hat{\sigma}^2_{t+1}$ or $\hat{h}^2_{t+1}$. Now let us compute the 'normalised' returns as $r^*_i = r_i/\tilde{\sigma}_i$ in the case of the GARCH models and $r^{**}_i = r_i/\sqrt{\tilde{h}_i}$ in case of HEAVY specifications, for $i = (t - T_e + 1), \dots, t$. VaR estimates follow from (16) and (17):

$$\text{VaR}^{\text{LF}}_{t+1}(q) = -\text{percentile}(\{r^*_s\}^t_{s=t-T_e+1}, 100 \times q) \cdot \hat{\sigma}_{t+1} \tag{16}$$

and

$$\text{VaR}^{\text{HF}}_{t+1}(q) = -\text{percentile}(\{r^{**}_s\}^t_{s=t-T_e+1}, 100 \times q) \cdot \sqrt{\hat{h}_{t+1}} \tag{17}$$

4 Methodology for VaR Evaluation

4.1 Geometric-VaR Test

To assess the VaR estimates of the various models, this paper makes use of the Geometric-VaR test introduced in PW16. The method is briefly summarised below, but the reader is referred to the original paper for further details and derivations. Let us define a violation or a hit as the event that a loss on a given day t exceeds the VaR estimate at level q generated by our model. Then, let I_t be an indicator function such that $I_t = 1$ when there is a violation, that is:

$$I_t = \begin{cases} 1, & \text{if } r_t < -\text{VaR}_t(q) \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

Let t_i denote the day of the i-th hit. The no-hit duration D_i is then simply constructed by $D_i = t_i - t_{i-1}$. Transformation of the hit sequence into a duration sequence by computing D_i for all hits allows the use of duration modelling techniques to perform backtesting. More specifically, specify the following hazard function:

$$\lambda^i_d = ad^{b-1}e^{-c \cdot \text{VaR}_{t_i+d}} \tag{19}$$

where $0 \leq a, 0 \leq b \leq 1$, and $c \geq 0$. Under the null hypothesis that VaR is correctly specified, violations follow an i.i.d. Bernoulli sequence, such that durations between violations follow a geometric

distribution with parameter q, in which case the null hypothesis corresponds to $a = q$, $b = 1$, and $c = 0$. The parameter a in the hazard rate captures the unconditional coverage. The second part in the hazard function, d^{b-1}, describes duration dependence or time dependence in violations. The exponential part $e^{-c \cdot \text{VaR}_{t_i+d}}$ controls for independence of violations on the estimated VaR levels.

Again, under the null hypothesis of correct VaR specification, D_i follows a geometric distribution with parameter q:

$$\Pr(D_i = d) = q(1-q)^{d-1} \tag{20}$$

which in combination with the hazard function λ_d^i allows us to write the probability of a specific duration occurring:

$$f^i(d) = \Pr(D_i = d) = \lambda_d^i \prod_{k=1}^{d-1}(1 - \lambda_k^i) \tag{21}$$

as well as a survival function:

$$S^i(d) = \Pr(D_i \geq d) = \prod_{k=1}^{d-1}(1 - \lambda_k^i) \tag{22}$$

To take into consideration data censoring, let us define a binary indicator sequence $\{C_i\}_{i=1}^N$ with $C_i = 1$ indicating that the corresponding duration D_i is censored. That is, C_1 equals 0 if the entire hit sequence starts with a violation and likewise C_N equals 0 if the hit sequence ends with a violation. Hence, if $C_i = 0$, the contribution of duration i to the log-likelihood function is given by $f^i(d)$. In case of censoring, $C_i = 1$, the contribution is simply given by the survival function of the duration, $S^i(d)$. Taking left and right censoring into account leads to the following log-likelihood function:

$$\begin{aligned} log\, L(D|\Theta) = C_1\, log\, S^1(D_1) + (1 - C_1)\, log\, f^1(D_1) \\ + \sum_{i=2}^{N-1} log\, f^i(D_i) + C_N\, log\, S^N(D_N) + (1 - C_N) log f^N(D_N) \end{aligned} \tag{23}$$

The full Geometric-VaR test is then performed by means of a standard likelihood ratio test:

$$\text{LR}^{\text{GV}} = -2\left[log\, L(D|\hat{a}, \hat{b}, \hat{c}) - log\, L(D|a = q, b = 1, c = 0)\right] \tag{24}$$

Note that reducing the test to the special case $c = 0$ yields the Geometric test of Berkowitz et al. (2011). Setting $b = 1$ corresponds with performing the test of Engle and Manganelli (2004). In PW16, the authors show that the Geometric-VaR test can be decomposed into separate parts. Specifically, let us define the following tests:

1. Unconditional coverage test (UC) (maintaining $b = 1$, $c = 0$):

$$H_0 : a = q \qquad\qquad H_a : a \neq q$$

2. Duration independence test (Dind) (maintaining $c = 0$):

$$H_0 : b = 1 \qquad\qquad H_a : b < 1$$

3. VaR independence test (Vind):

$$H_0 : c = 0 \qquad\qquad H_a : c > 0$$

4. Geometric test (Geom): unconditional coverage and duration independence (under the assumption that $c = 0$):

$$H_0 : a = q \text{ and } b = 1 \qquad\qquad H_a : a \neq q \text{ or } b < 1$$

5. VaR test (VaR): unconditional coverage and VaR independence (under the assumption that $b = 0$):

$$H_0 : a = q \text{ and } c = 0 \qquad\qquad H_a : a \neq q \text{ or } c > 0$$

6. Geometric-VaR test (GV): unconditional coverage, duration independence and VaR independence:

$$H_0 : a = q,\ b = 1 \text{ and } c = 0 \qquad\qquad H_a : a \neq q \text{ or } b < 1 \text{ or } c > 0$$

As mentioned, the design of the test allows for testing different hypothesis seperately as part of the Geometric-VaR test. In particular, note that we can write $\mathrm{LR}^{\mathrm{GV}} = \mathrm{LR}^{\mathrm{UC}} + \mathrm{LR}^{\mathrm{Dind}} + \mathrm{LR}^{\mathrm{Vind}}$.

Parameters are estimated using maximum likelihood after generating the duration sequence consisting of the different D_i's. As suggested in PW16, this paper employs the Monte Carlo technique of Dufour (2006) to generate p-values that are robust to limited sample size by constructing an empirical distribution of the likelihood ratio under the null hypothesis. Specifically, generate an i.i.d. Bernouilli sequence with the same length as the investigated sample size T_e and hit probability q to serve as the 'violation sequence'. In addition, construct independent VaR estimates by assuming that the returns follow a flexible nonlinear asymmetric GARCH process (NGARCH) of the type first introduced by Engle and Ng (1993), which allows the replication of stylised facts of

asset returns, such as asymmetries, heavy tails, volatility clustering and leverage effects:

$$r_{t+1} = \sigma_{t+1}((d-2)/d)^{\frac{1}{2}} z_{t+1} \tag{25}$$

$$\sigma^2_{t+1} = \omega + \alpha\sigma^2_t \left(\left(\frac{d-2}{d} \right)^{\frac{1}{2}} z_t - \theta \right)^2 + \beta\sigma^2_t \tag{26}$$

with z_t being drawn from a Student's $t(d)$ distribution. The 'true' values for VaR based on the conditional variance process follow easily by using

$$\text{VaR}_{t+1} = -\sigma_{t+1}((d-2)/d)^{\frac{1}{2}} \tau_{d,q} \tag{27}$$

with $\tau_{d,q}$ being the relevant critical value of a $t(d)$-distribution at VaR level q. The relevant parameters of the NGARCH process are estimated using maximum likelihood for the relevant return series for which a VaR model is being assessed. To generate the empirical distribution of LR statistics under the null, simulate K times such a Bernoulli violation sequence and NGARCH VaR sequence. Setting K sufficiently large, convert each Bernoulli hit sequence to a duration sequence and compute the LR statistics for the Geometric-VaR test and its sub-tests per (24) using the NGARCH VaR sequence as input for VaR_t in the hazard function. The p value is then calculated by means of comparing the LR statistic of the relevant VaR model being tested to the empirical distribution generated by the simulations. In doing so, this paper employs the exact procedure of Dufour (2006) and the reader is referred to said paper for further details.

5 Data

This paper utilises data from the 'Realized Library (v0.2)' from the Oxford-Man Institute of Quantitative Finance[5], which contains daily return data for 21 indices spanning the period 3 January 2000 - 5 December 2017[6], as well as realised measures. The underlying source of the data is the Thomson Reuters DataScope Tick History database. Data points are cleaned and processed exactly based on the methodology specified in Shephard and Sheppard (2010). In addition, days on which exchanges are closed are excluded from the sample as in Opschoor et al. (2017). A wide variety of indices is included in the sample in order to test robustness of the VaR models under different market structures. Indeed, the summary statistics presented in Table B1 and B2 in Appendix B support the notion of variety in index characteristics. Specific attention has been paid to the

[5]See http://realized.oxford-man.ox.ac.uk/

[6]Some indices have different starting and ending dates because of holidays or data availability. The reader is referred to Table A1 in Appendix A for a full list of included indices, as well as the exact periods covered.

Table 2: Size of 10% duration-based tests applied to 5% VaR

Sample size	UC	Dind	Vind	Geom	VaR	GV
Using chi-squared asymptotic critical values						
250	0.148	0.028	0.094	0.058	0.098	0.060
500	0.120	0.033	0.101	0.068	0.105	0.066
750	0.121	0.035	0.094	0.062	0.096	0.065
1000	0.119	0.036	0.091	0.059	0.087	0.064
1250	0.114	0.038	0.089	0.061	0.086	0.064
1500	0.106	0.038	0.086	0.061	0.085	0.063
Chi-squared critical value	*2.706*	*2.706*	*2.706*	*4.605*	*4.605*	*6.251*
Using simulated test statistics computed with sample size 50,000						
250	0.139	0.063	0.193	0.108	0.152	0.123
500	0.115	0.071	0.187	0.102	0.156	0.130
750	0.121	0.076	0.178	0.099	0.143	0.134
1000	0.116	0.081	0.164	0.099	0.135	0.126
1250	0.112	0.077	0.162	0.094	0.130	0.118
1500	0.097	0.082	0.158	0.097	0.121	0.118
Simulated critical value	*2.788*	*1.510*	*1.669*	*3.759*	*3.838*	*4.750*

Note. This table assesses the size properties of the relevant tests by generating random VaR sequences that satisfy the null hypothesis. In the upper panel of this table, 'chi-squared asymptotic critical values', the size is calculated based on the rejection frequency of 10,000 replications using asymptotic critical values from a chi-squared distribution. In the lower panel, the same random VaR estimates are assessed using simulated critical values which are calculated using separate i.i.d. Bernoulli hit sequences and NGARCH VaR regressors that are independent of the simulated hit sequence. The empirical distribution of the asymptotic test statistic is generated using 10,000 replications of sample size 50,000. 'UC' stands for unconditional coverage test, 'Dind' for duration independence test, 'Vind' for VaR independence test, 'Geom' for Geometric test, 'VaR' for VaR test and 'GV' for Geometric-VaR test. See text for further details regarding procedures.

inclusion of the global financial crisis in the relevant sample as it is especially important that risk models are tested during challenging conditions.

6 Simulation Studies for the Geometric-VaR Test

6.1 Test Size

As the test of PW16 is relatively new and has hence not been adopted by other papers, it is of interest to reassess its effectiveness before applying it to test the HEAVY models. This research does so by attempting to replicate the results of the original paper. In addition, the test is benchmarked against more established backtesting procedures. This subsection will investigate the size[7] properties of the Geometric-VaR test, whereas the next subsection examines the power of the test. Size is investigated by testing how often correctly specified VaR estimates are rejected, whereas power is examined by means of rejections of falsely specified VaR estimates. Both exercises are conducted using Monte Carlo simulation.

[7]The term 'size' is used to denote the probability of a type I error throughout the rest of this paper. Hence, the terms 'oversized' and 'undersized' denote rejection frequencies that are above and below their theoretical values respectively.

To compute reliable p-values and control the size of the tests, PW16 relies on Dufour's (2006) Monte Carlo technique. This research validates the necessity of the approach by generating fictional returns by means of the NGARCH process described by (25) and (26). In particular, the same model parameters for the NGARCH return process as in PW16 are used: $\{d, \theta, \beta, \alpha, \omega\} = \{10, 0, 0.93, 0.05, 0.21\}$. Following PW16, the 'true' VaR estimates that perfectly satisfy the null hypothesis are computed using (27) with a coverage rate set to equal 1% and 5% respectively. The VaR 'estimates' are evaluated at a 10% significance level and the size of each test is hence given by the empirical rejection rate over 10,000 replications. The test size based on the asymptotic critical values of the chi-squared distribution is reported in the upper panel of Table 2 for 5% VaR. Even at 10,000 replications, notable variation in the obtained size values is observed over different trials. Taking this into account, the results are qualitatively and quantitatively similar to PW16. Indeed, basing the test on the chi-squared asymptotic values yields mostly undersized test statistics except for the test of unconditional coverage. Of interest are the values of the VaR independence test which show very similar values for the larger samples as in PW16. However, for smaller samples a slightly lower test sizing compared to the values reported in PW16 can be observed. Clearly, the exact implementation of the testing procedure has an effect on the values obtained. In addition, the initialisation of for instance the NGARCH process has a measurable effect on the size of the VaR independence test for small samples. Logically, the difference in VaR independence values for small samples feeds into the VaR and Geometric-VaR test.

As mentioned in PW16, the asymptotic critical values based on the chi-squared distribution do not take into account the effect of testing parameter values at the boundary of the parameter space. Hence, asymptotic critical values that are robust to this effect are also generated by means of simulation. In particular, simulate 10,000 VaR sequences by means of the NGARCH process, setting the sample size equal to 50,000. In addition, generate 10,000 i.i.d. Bernoulli 'hit sequences' of length 50,000. Empirical 10% critical values are then obtained using the 90% percentile of the empirical test statistic distribution over the 10,000 trials which are exhibited in Table 2. As expected, these simulated critical values are different from their chi-squared based counterparts and are in line with the values obtained by PW16 although the value found for duration independence is about 6% lower.

Again, let us assess the size of the duration-based tests, but now using the simulated critical values instead of the chi-squared critical values as in the upper panel of Table 2. Except for the test of

unconditional coverage, which was oversized in the chi-squared setting, all undersized tests increase in size. Again, taking into account variability across simulation runs, results are quantitatively similar to PW16. Interestingly, although the simulated critical value for duration independence is lower than reported in PW16, size remains unaffected and is close to the values found in PW16. As was the case for the chi-squared test values, lower statistics for small samples in the case of the VaR independence tests are observed when compared to PW16 which again transpires into the VaR and Geometric-VaR test statistics. Clearly, using asymptotic critical values for testing finite samples leads to incorrectly sized test values. The use of simulated asymptotical critical values does not solve this problem – all tests remain oversized, with the exception of duration independence which is undersized compared to the desirable 10% level. A similar exercise conducted for 1% VaR yields qualitatively similar results. Evidently, the test statistics are (1) not exactly chi-squared distributed and (2) test statistics obtained for finite samples differ considerably from their asymptotic values. This indeed supports the application of Dufour's (2006) Monte Carlo technique.

6.2 Test Power

Having established the necessity of the method of Dufour (2006), let us look at the power of the Geometric-VaR test in case of purposely misspecified VaR estimates. Following PW16, random return series of size $T + T_e$ with $T_e = 250$ and $T \in \{250, 500, 750, 1000, 1250, 1500\}$ are generated using the NGARCH process described in (25) and (26) and using parameters estimated from real-world business lines taken from PW16 and shown in Table 3. VaR forecasts are then computed based on the Historical Simulation method using a rolling window of size T_e which is known to provide incorrect estimates. Specifically, 5,000 series of returns with corresponding HS VaR estimates are generated. The relevant test statistics are computed and p-values are obtained using the empirical distribution of LR statistics and the method of Dufour (2006) using 9,999 trials. The power of each test is defined as the rejection frequency of the 5,000 replications at significance level p, which is chosen to be 10% for comparison purposes.

Table 3: Parameters of NGARCH simulations for the four business lines

Parameter	Business Line 1	Business Line 2	Business Line 3	Business Line 4
d	3.808	3.318	6.912	4.702
θ	-0.245	0.503	-0.962	0.093
β	0.749	0.928	0.873	0.915
α	0.155	0.052	0.026	0.072
ω	0.550	0.215	0.213	1.653

Table 4: Power of duration-based tests and the CaViaR test, 5% VaR, 10% significance

Sample size	UC	Dind	Vind	Geom	VaR	GV	CaViaR
Business Line 1							
250	0.147	0.473	0.321	0.351	0.290	0.442	0.441
500	0.064	0.671	0.650	0.482	0.470	0.686	0.514
750	0.036	0.791	0.796	0.615	0.605	0.838	0.597
1000	0.028	0.861	0.867	0.711	0.732	0.918	0.703
1250	0.018	0.907	0.902	0.781	0.781	0.957	0.772
1500	0.017	0.939	0.949	0.842	0.861	0.977	0.854
Business Line 2							
250	0.344	0.448	0.436	0.493	0.521	0.599	0.584
500	0.233	0.715	0.675	0.611	0.646	0.790	0.633
750	0.193	0.830	0.742	0.754	0.708	0.890	0.665
1000	0.182	0.899	0.779	0.813	0.772	0.940	0.735
1250	0.170	0.938	0.811	0.864	0.830	0.964	0.761
1500	0.171	0.955	0.830	0.905	0.845	0.980	0.832
Business Line 3							
250	0.053	0.136	0.407	0.066	0.120	0.141	0.295
500	0.012	0.144	0.713	0.047	0.324	0.264	0.403
750	0.007	0.159	0.861	0.055	0.513	0.475	0.445
1000	0.003	0.163	0.922	0.047	0.668	0.617	0.515
1250	0.003	0.160	0.962	0.045	0.802	0.743	0.593
1500	0.002	0.164	0.978	0.054	0.859	0.818	0.724
Business Line 4							
250	0.346	0.505	0.393	0.522	0.501	0.614	0.591
500	0.223	0.765	0.657	0.665	0.640	0.818	0.629
750	0.182	0.878	0.733	0.785	0.714	0.903	0.696
1000	0.174	0.934	0.797	0.855	0.783	0.952	0.741
1250	0.157	0.961	0.821	0.912	0.824	0.976	0.788
1500	0.166	0.979	0.859	0.940	0.863	0.989	0.859

Note. This table shows the power of each test at significance level 10% by generating random returns using NGARCH-t(d) models that have the same parameters as the four business lines specified in PW16 and shown in Table 3. 5% VaR estimates are then computed using Historical Simulation with a rolling window of size 250. The simulated power of each test is the rejection frequency from 5,000 replications. 'UC' stands for unconditional coverage test, 'Dind' for duration independence test, 'Vind' for 'VaR' independence test, 'Geom' for Geometric test, 'VaR' for VaR test and 'GV' for Geometric-VaR test. 'CaViaR' is a regression-based test. See text for further details regarding procedures.

In order to examine the performance of the Geometric-VaR test compared to other more established backtesting procedures, the power of the following tests is computed in similar fashion: the proportion of failures test of Kupiec (1995), the conditional coverage independence (CCI) test and conditional coverage mixed test (CC) of Christoffersen (1998), the time between failures independence test (TBFI) and time between failures test (TBF) of Haas (2001) and the CaViaR test of Engle and Manganelli (2004). Note that again the method of Dufour (2006) is employed to control the sizing of these tests for small samples. Due to spacing considerations, results for 5% VaR are presented in condensed fashion in Table 4 for the Geometric-VaR test and the CaViaR test. Table C1 and C2 in Appendix C show results for all investigated tests for 5% and 1% VaR respectively.

Again, this research finds quantitatively similar results to PW16. As in the original paper, the

power of the UC test proves to be low in practice for 5% VaR and 1% VaR – especially for the larger sample sizes. Specifically the power for Business Line 3 which exhibits relatively small volatility persistence is very low. PW16 claim performance of the UC test is expected to be similar to the POF test of Kupiec (1995) but leave aside the necessary evidence to back up this claim. Indeed, the results in Table C1 and C2 provide support for this assertion. Note also that although the power of the UC test presented in PW16 is similar, it is consistently lower than that of Kupiec's POF test in case of 5% VaR.

With regards to violation independence we can observe particularly good performance of the Duration independence test of PW16 compared to the CCI test of Christoffersen (1998) and the TBFI test of Haas (2001) for both 5% and 1% VaR. Although PW16 remark its relatively mediocre performance for Business Line 3 because of the opposite leverage effects in these series, the test still performs better than the CCI and TBFI test for this Business Line. Indeed, the power of the Duration Independence test of PW16 is higher in all business lines investigated.[8] Looking at the Geometric test of PW16, however, provides a notable contrast to this image. As the Geometric test is a combination of the UC and Duration independence tests, the test statistic is provided by the sum of the LR test statistics of the two tests. Hence, the low power of the UC test for Business Line 1 and Business Line 3 severely affects the power of the Geometric test for these business lines. Indeed, for 5% VaR, observe that the Geometric test of PW16 is outperformed in terms of power by the TBF test of Haas (2001) for Business Line 1 and 3 and by the CC test of Christoffersen (1998) in Business Line 3 for the majority of the sample sizes. In business lines 2 and 4, the Geometric test still provides superior performance for 5% as well as for 1% VaR.

As in the earlier size investigation, we observe a lower rejection frequency for the VaR independence test for the smaller samples compared to PW16. With regards to testing all hypotheses at once, it is worth noting that the Geometric-VaR test outperforms the competing CaViaR test in most, but not all cases − specifically for the smaller sample sizes for 1% VaR in general and Business Line 3 for 5% VaR. As seen before, it appears that the Geometric test component provides suboptimal power in case of negative leverage effects which feeds into the Geometric-VaR test. Clearly, one must exercise caution when relying on the Geometric-VaR test in case return processes exhibit low volatility persistence and negative leverage effects in small sample settings. However, in line with PW16, the results obtained illustrate that the Geometric-VaR test provides good power

[8] One exception is the observation for 1% VaR in Business Line 3 for sample size 1500, which can most likely be attributed to randomness in simulations.

properties against various forms of misspecification that might be present in VaR estimates.

7 Empirical Results

7.1 Full Sample - Geometric-VaR Test

Having revisited the properties of the Geometric-VaR test, let us proceed by investigating the performance of the HEAVY models. The next sections provide empirical results of model performance. In particular, this subsection investigates the merit of models the over the full sample using the Geometric-VaR test. The next subsection provides an economic interpretation. The last subsection covers empirical performance specifically during the period of the global financial crisis.

For all models, 5% VaR estimates[9] are obtained for all 21 equities over the period January 2000 - December 2017 (for further details, see Section 5: Data). In particular, VaR estimates are computed using a rolling window size T_e of 250, 1000 and 1500 respectively. Note that for $T_e = 250$ and $T_e = 1000$ the first 1,250 and 500 VaR estimates are truncated respectively to get the same effective sample period across the different rolling window sizes (January 2006 - December 2017). The Geometric-VaR test is then utilised to evaluate the estimates. The null hypothesis of all tests is evaluated at a 10% level and consequently the number of equity indices for which the null hypothesis of correct model performance is not rejected is computed for each relevant test. Results are presented in Table D1 in Appendix D.

First, notice that model performance is poor for the smallest rolling window size. Although a window size of 250 days is sufficient to provide correct unconditional coverage (7 models achieve correct coverage for either 20 or 21 out of 21 equity indices), independence of VaR estimates is lacking with the best performing model (gGN) not rejecting the null hypothesis of independence in only 8 out of 21 cases. Logically, this transpires in the Geometric-VaR test which it does not reject the null hypothesis of correct VaR estimates for only 5 out of 21 indices for the best performing model. Clearly, the Geometric-VaR test suggests a rolling window size of 250 days is insufficient to provide VaR estimates that satisfy the relevant conditions. The Historical Simulation method which is prevalent amongst practitioners scores well in terms of unconditional coverage, but fails to pass the Geometric-VaR test in all cases and is hence clearly insufficient – especially with respect to duration independence.

For a rolling window size of 1,000 days, a substantial increase in performance can be observed

[9]Because of computational limitations, only the case of 5% VaR is investigated. Assessing model performance in case of 1% VaR is left as an avenue for further research.

for most models compared to the case of $T_e = 250$. In particular, the increased estimation window causes the p-values for duration independence and VaR independence to increase substantially. In terms of overall performance, notice that the models that incorporate the Filtered Historical Simulation approach outperform their fully-parametric counterparts. Although some HEAVY models perform quite well (notably HFk and gHFk), they are outperformed by asymmetric GARCH specifications. Most notably, the GJR-GARCH model combined with Filtered Historical Simulation (gGF) passes the Geometric-VaR test for all but one index investigated and is clearly the best performing model, both in terms of number of passed tests as well as average p-value. Although this model appears to score best by means of the Geometric-VaR test, it must be noted that it does not always pass its sub-tests. Most notably, the gGF model seems to have relatively mediocre performance in terms of VaR independence – a test competing parametric specifications pass to a considerably higher degree. Again, notice the particularly poor performance of Historical Simulation.

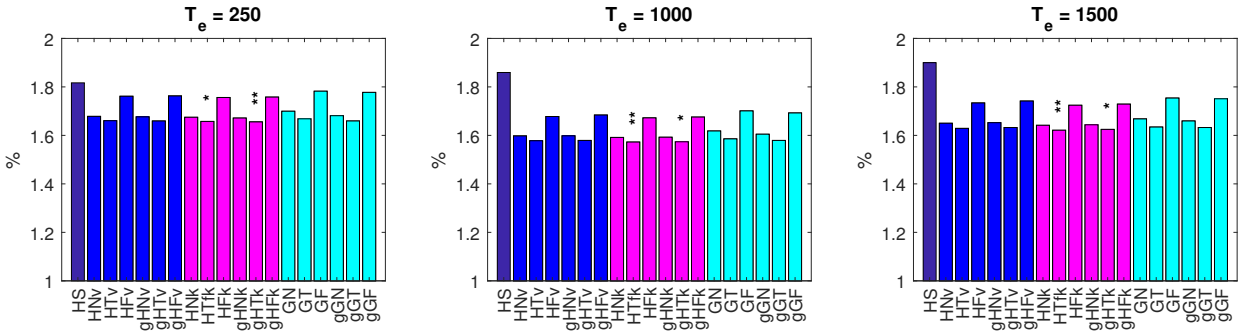
Increasing the rolling window size to $T_e = 1500$ does not seem to yield a particular increase in performance and underscores the issue of VaR independence for the gGF model. In fact, model performance decreases in terms of the number of indices where it passes the test. Indeed, increased rolling window sizes might prohibit flexibility in case of structural breaks which leads to worse model performance. However, increasing the rolling window size does appear to improve VaR independence performance for some of the models investigated.

Note that for the HEAVY models, the fully parametric approach seems to work best with realised variance data, whereas the FHS method performs better with the realised kernel estimator. In addition, the adoption of an asymmetric specification in the form of GJR-HEAVY does not necessarily improve performance.

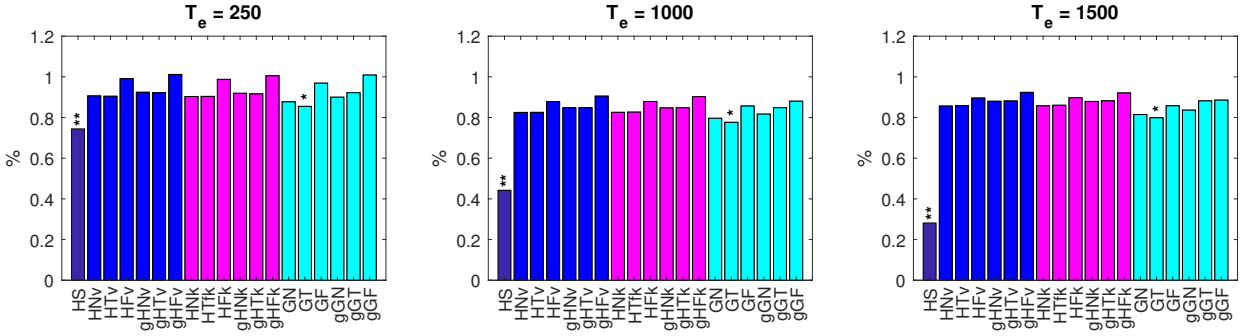
Two things become clear: (1) The HEAVY models do not outperform their asymmetric GARCH counterparts; (2) the Filtered Historical Simulation approach leads to better performance than a fully-parametric approach for all models investigated.

Considering the models across all rolling window sizes, it is clear that the GJR-GARCH model in combination with Filtered Historical Simulation provides superior performance over competing specifications. Minimum rolling window sizes of 1,000 must however be maintained in order for the specification to deliver adequate estimates as especially violation independence of VaR estimates seems to be an issue for smaller rolling window sizes.

Figure 1: Economic statistics of VaR estimates over 24 January 2006 - 5 December 2017



(a) Mean Value-at-Risk across indices



(b) Standard deviation of Value-at-Risk across indices

Panel (a) shows the mean VaR estimates in % across time and indices per model. Panel (b) shows the mean of the standard deviation of the VaR estimates in % points across indices. Colors correspond to different model classes (HS, realised variance based HEAVY, realised kernel based HEAVY and GARCH respectively). ** denotes the best performing model in the respective rolling window length. * denotes second best performing model.

7.2 Full Sample - Economic Evaluation

In addition to having theoretically sound VaR estimates in the sense that they satisfy the conditions of (1) correct unconditional coverage; (2) independent violations; (3) violations independent of information up to and including $t-1$, practitioners need to make sure that the VaR estimates are optimal from an economic perspective as well. That is, as capital requirements are determined based on estimated VaR, it is of interest to keep them as low as possible in order to minimise opportunity costs presented by retaining excess reserve capital.[10] Let us assess the economic implications of the VaR models by means of two simple statistics: the average level of VaR and the variability of VaR (as measured by its standard deviation). Although capital requirements in practice depend in a nonlinear fashion on the estimated levels of VaR, let us assume for feasibility that they are strictly

[10] Although regulations are migrating from using VaR as a means to calculate capital requirements to Expected Shortfall (Basel Committee on Banking Supervision, 2013), the minimisation of VaR remains relevant as Expected Shortfall is defined as the expected loss beyond VaR and hence VaR serves as a boundary from which to calculate the Expected Shortfall.

increasing with VaR. The statistics are computed as an average over time and across models. Note that again for $T_e = 250$ and $T_e = 1000$ the first 1,250 and 500 VaR estimates are truncated respectively to get the same effective sample period across the different rolling window sizes. Statistics are presented in Figure 1 and Table D2. A few important observations can be made. In addition to providing the best VaR estimates from a Geometric-VaR testing perspective, using a rolling window of 1,000 days leads on average to lower average VaR estimates compared to both $T_e = 250$ and $T_e = 1500$. Moreover, the variability of the estimates is lowest for this length of estimation window, making it desirable from a capital allocation perspective as frequent changes in the level of capital reserves bring along extra costs. An obvious exception is the Historical Simulation method which by construction produces higher average VaR estimates when estimation windows increase, while the variability of these estimates decreases. When comparing competing model specifications, we observe that the choice of GARCH versus HEAVY does not have a notable effect on the average VaR levels estimated. However, the variability of the estimates appears to be slightly lower for some GARCH specifications compared to their HEAVY counterparts. Across the board, Filtered Historical Simulation produces VaR estimates that are both higher and more volatile in nature than the fully-parametric approaches. Apparently, the superior performance as measured by the Geometric-VaR test for the FHS approach comes at the economic expense of higher capital buffers, which presents a clear trade-off for practitioners.

7.3 Global Financial Crisis of 2008 - Geometric VaR Test

Clearly, having adequate risk models in place is especially important during adverse market conditions. However, historically banks tend to underestimate Value-at-Risk in periods of recessions (Bank for International Settlements, 2009). Hence, examining model performance during periods of financial distress is of particular interest. Consequently, let us re-examine the performance of the HEAVY models during the global financial crisis of 2008. As the exact starting point of the recession is open to interpretation, this paper takes a safe testing window of two years spanning July 2007 - June 2009. An illustration of the relevant period including several risk models is provided in Figure 2. Using the same approach as in earlier sections, the Geometric-VaR test is carried out for all models on the relevant window. Testing outcomes are presented in Table D3. The results differ markedly from those obtained for the full sample. For all rolling window sizes, the HEAVY specifications outperform their GARCH counterparts considerably, suggesting that the addition of realised measures is of value in this period of distress. Note that again, the use of Filtered Historical Simulation provides better estimates than the fully-parametric approach for all models investigated.

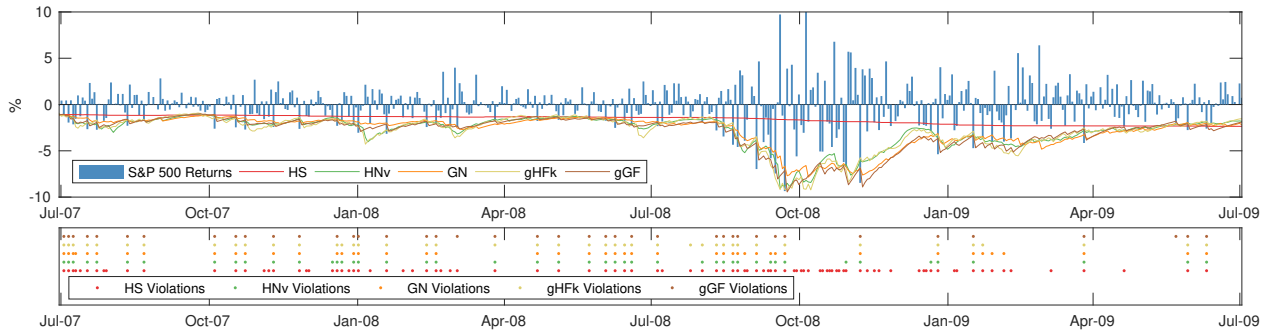


Figure 2: S&P 500 Returns with various VaR estimates and VaR violations during the financial crisis of 2008 ($T_e = 1000$)

Surprisingly, the asymmetric gHFk model for $T_e = 250$ and symmetric HFk model for $T_e = 1500$ perform similar in terms of rejection rates. Rather than for the full sample a clearly favoured rolling window size is not apparent. Indeed, the short estimation size of 250 days that was found to be inadequate in the full sample may help in providing essential flexibility during this sudden period of aberrant volatility. Examining Figure 2 provides some further insights. As found in Shephard and Sheppard (2010), there is some evidence that at the height of the financial crisis, the GARCH models are behind the curve, while the HEAVY models adjust rapidly. Additionally, it appears as if the GARCH models estimate volatility as unnecessarily elevated compared to the HEAVY models during December 2008 and the model does not allow for conditional volatility to fall rapidly enough.

Although the HEAVY models incorporating FHS perform well compared to their GARCH counterparts during the financial crisis, it must be noted that the best model still only fails to be rejected for 15 out of 21 indices, suggesting evident room for further improvement.

8 Conclusion

This research studies how Value-at-Risk (VaR) models can be backtested and subsequently applies this knowledge to study VaR models based on high-frequency data. The starting point of this paper is to replicate the work of Pelletier and Wei (2016) (PW16), who introduce the Geometric-VaR test. By taking a duration-based approach to testing VaR, they show that VaR estimates can be backtested by modelling the duration between violations as geometrically distributed with a flexible hazard function. The framework allows simultaneous, as well as separate testing of unconditional coverage, duration independence and VaR independence. This paper has four particular contributions to existing literature. Firstly, it replicates results of PW16 and thus validates the

conclusions drawn by PW16 about the power of the test. This is of essence as the test has not yet been adopted in other research. Secondly, benchmarking of the Geometric-VaR test is conducted by comparing the power of the test to not only the CaViaR test of Engle and Manganelli (2004), as is done in PW16, but a wide range of commonly employed backtesting frameworks. Thirdly, this paper is the first to investigate the high-frequency based volatility (HEAVY) models of Shephard and Sheppard (2010) in the context of VaR estimation using a Filtered Historical Simulation (FHS) approach. Lastly, an asymmetric adaptation of the standard HEAVY model is introduced to the literature.

The lion's share of the replication results is similar to those presented in PW16. Indeed, examining the Geometric-VaR test with simulated data shows that its test statistic is not asymptotically chi-squared distributed and that small samples lead to sizing issues. Both problems are mitigated using the method of Dufour (2006) which provides robust p-values for the test. Simulations based on real business line returns demonstrate good power properties against various forms of misspecification of VaR estimates. However, this paper finds slightly lower power for smaller sample sizes compared to PW16, in particular for the VaR independence component of the test. Likely reasons for this deviation include the exact technical implementation of the test as well as the initialisation of the NGARCH processes. Detailed considerations are provided in the next section, 'Discussion'. The Geometric-VaR test mostly provides better power than competing established backtesting procedures, but it appears that one must exercise caution if return processes exhibit low volatility persistence and negative leverage effects as in this case the Geometric test component provides suboptimal power for small sample sizes which feeds into the Geometric-VaR test.

Subsequently, the Geometric-VaR test is employed to investigate the HEAVY model in the context of VaR estimation, using Historical Simulation and a family of GARCH models as a benchmark. Both a fully-parametric approach using Normal and Student's t distributions, as well as a semi-parametric approach based on Filtered Historical Simulation are examined, which, in total, leads to 19 different specifications. VaR estimates are computed for a broad universe of 21 equity indices with data ranging from 2000 to 2017. Using the full sample, this paper finds no evidence that the HEAVY models outperform their GARCH counterparts in terms of being able to pass the Geometric-VaR test. Indeed, the GJR-GARCH model combined with FHS provides the best results. This result is surprising as realised measures should theoretically improve conditional variance forecasts and thus VaR estimates. Hence, one could expect the HEAVY models, which are

based on realised measures, to provide better performance than their GARCH counterparts. A clear explanation for this unexpected result is not readily apparent. An important observation is the fact that the application of FHS improves VaR estimates considerably for both HEAVY and GARCH models compared to fully-parametric approaches, but leads to a higher level of VaR estimates on average, which can be a disadvantage for practitioners.

This paper also considers the HEAVY models specifically during the global financial crisis as having adequate risk model performance is especially important during adverse market conditions. The results provide a stark contrast to those of the full sample. Interestingly, the HEAVY specification appears to work particularly well during this period of unusual volatility, beating the GARCH counterparts. Again, the application of FHS works best for all specifications considered.

In conclusion, this paper assesses the merit of the Geometric-VaR backtesting procedure of PW16 and provides insights about VaR estimates based on high-frequency data. The Geometric-VaR test is confirmed to be a valuable addition to VaR backtesting methods. Contradictory findings between normal times and the global financial crisis challenge a conclusive answer as to whether HEAVY models provide accurate VaR estimates. However, results indicate that HEAVY models are adept in periods of unusual volatility and could hence be a valuable addition to a risk manager's toolkit during times of recession.

9 Discussion

The fact that this study was bounded by time and computational capabilities gives rise to several limitations that are discussed in detail in this section. First, the results obtained in Table 2, 4, C1 and C2 are based on simulated processes that are subject to a degree of precision varying with trial size. In particular, note that the values in Table 2 are obtained using 10,000 replications as in PW16. Although it is difficult to give an exact quantification of the degree of variation, several runs with replication size 10,000 demonstrate that the obtained results can still vary up to the second decimal figure. In particular, this appears to be the case for the power replications of the different business lines where only 5,000 replications are used. Hence, caution must be exercised when interpreting the results in PW16 and this paper as 'true values'. Taking this uncertainty into account, however, there are differences in the results obtained that cannot be ascribed to randomness introduced by a finite number of simulations. In particular, the VaR independence test provides lower rejection frequencies than reported in PW16 for the smaller sample sizes. A logical

explanation for this observed difference is the exact implementation of the testing procedure.[11] In addition, the initialisation of the NGARCH process has a measurable effect on the sizing of the VaR independence test for small sample sizes. This paper initialises the NGARCH process at its unconditional mean and in addition truncates the first 1,000 simulations as a 'burn-in' sample. The approach taken by Pelletier and Wei (2016) remains unspecified and is hence difficult to replicate. Logically, the difference in VaR independence values for small samples feeds into the VaR and Geometric-VaR test. Although this paper must conclude that the results of PW16 for smaller sample sizes cannot be replicated exactly, it does not affect the results obtained in the rest of this research as the figures obtained in the empirical part of the paper rely on sample sizes that are sufficiently large. In addition, it is found that the test provides mediocre power properties in case of low volatility persistence and negative leverage effects. Estimating NGARCH parameters on the 21 indices provided no evidence of such effects being present in the data used for the empirical study and hence this shortcoming of the test is not expected to affect the outcomes.

Secondly, the fact that the HEAVY models are found to underperform their GARCH counterparts over the full sample period is surprising as it contradicts results presented in Shephard and Sheppard (2010). However, different sample periods, as well as the fact that Shephard and Sheppard (2010) assess conditional volatility directly, rather than VaR – one specific point on the left-tailed distribution of the returns – may help to explain differences.

Further research could investigate specifically under which conditions HEAVY models generate good and bad VaR estimates respectively. Additionally, combining forecasts in order to exploit particular model strengths could be investigated in, for instance, a Markov switching setting. Moreover, this paper recognises that estimation risk is a relevant issue in the context of VaR backtesting. Although the critical values obtained by the method of Dufour (2006) were found robust to estimation risk presented by the NGARCH process[12], estimation risk within the VaR models themselves provides further uncertainty. A resampling approach such as suggested by Escanciano and Olmo (2010) could mitigate such issues. Adoption of such an approach was considered, but this proved to be computationally unfeasible to conduct for all models. Hence, this is left for future research.

[11] A clear answer on any exact difference of test implementation can come from the authors of PW16. Although the author of this paper has politely reached out to both Denis Pelletier and Wei Wei, they have not yet responded at the time of writing this paper. Clearly, this case reiterates the merit of providing transparent code along with one's publication for replication purposes.

[12] Specifically, a Bayesian approach with non-informative priors was adopted to estimate the NGARCH parameters, rather than maximum likelihood. The empirical parameter distribution was then used as input to simulate the NGARCH return series in the Dufour process. No significant difference was observed in the critical values obtained by adopting such an approach.

References

Abad, P., Benito, S., and López, C. (2014). A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, 12(1):15–32.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The review of economics and statistics*, 89(4):701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001a). The distribution of realized stock return volatility. *Journal of financial economics*, 61(1):43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001b). The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 96(453):42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Bank for International Settlements (2009). Revisions to the Basel II market risk framework.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3).

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280.

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (1997). Var without correlations for nonlinear portfolios. *Journal of Futures Markets*.

Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: A revised market risk framework.

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, 57(12):2213–2227.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S. (2010). Backtesting value-at-risk: a gmm duration-based test. *Journal of Financial Econometrics*, 9(2):314–343.

Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2(1):84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, pages 841–862.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues.

Dufour, J.-M. (2006). Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 133(2):443–477.

Engle, R. (2002). New frontiers for arch models. *Journal of Applied Econometrics*, 17(5):425–446.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The journal of finance*, 48(5):1749–1778.

Escanciano, J. C. and Olmo, J. (2010). Robust backtesting tests for value-at-risk models. *Journal of Financial Econometrics*, 9(1):132–161.

Escanciano, J. C. and Pei, P. (2012). Pitfalls in backtesting historical simulation var models. *Journal of Banking & Finance*, 36(8):2233–2244.

Forsberg, L. and Bollerslev, T. (2002). Bridging the gap between the distribution of realized (ecu) volatility and arch modelling (of the euro): the garch-nig model. *Journal of Applied Econometrics*, 17(5):535–548.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801.

Haas, M. (2001). New methods in backtesting. *Financial Engineering Research Center, Bonn*.

Haas, M. (2006). Improved duration-based backtesting of value-at-risk. *The Journal of Risk*, 8(2):17–38. Copyright - Copyright Risk Waters Group Winter 2005/2006; Document feature - graphs; tables; references; equations; Last updated - 2012-02-09.

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161.

Hull, J. and White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of risk*, 1(1):5–19.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications*, 119(7):2249–2276.

Jorion, P. et al. (2007). *Financial risk manager handbook*, volume 406. John Wiley & Sons.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models.

Lopez, J. A. (1999). Regulatory evaluation of value-at-risk models. *The Journal of Risk*, 1(2):37–64.

Nakagawa, T. and Osaki, S. (1975). The discrete weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301.

Nieto, M. R. and Ruiz, E. (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting*, 32(2):475–501.

Opschoor, A., Van Dijk, D., and van der Wel, M. (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics*, 32(7):1298–1313.

Pelletier, D. and Wei, W. (2016). The geometric-var backtesting method. *Journal of financial econometrics*, 14(4):725–745.

Pérignon, C. and Smith, D. R. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, 34(2):362–377.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, 30(2):561–582.

Shephard, N. and Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (heavy) models. *Journal of Applied Econometrics*, 25(2):197–231.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Appendix A: Included Indices in the Dataset

Table A1: Included Indices in the Dataset

Symbol	Name	Earliest Available	Latest Available
.AEX	AEX index	1/3/2000	12/05/2017
.AORD	All Ordinaries	1/4/2000	12/05/2017
.BVSP	BVSP BOVESPA Index	1/3/2000	12/05/2017
.DJI	Dow Jones Industrial Average	1/3/2000	12/05/2017
.FCHI	CAC 40	1/3/2000	12/05/2017
.FTMIB	FTSE MIB	6/1/2009	12/05/2017
.FTSE	FTSE 100	1/4/2000	12/05/2017
.GDAXI	DAX	1/3/2000	12/05/2017
.GSPTSE	S&P/TSX Composite index	5/2/2002	12/05/2017
.HSI	HANG SENG Index	1/3/2000	12/05/2017
.IBEX	IBEX 35 Index	1/3/2000	12/05/2017
.IXIC	Nasdaq 100	1/3/2000	12/05/2017
.KS11	Korea Composite Stock Price Index (KOSPI)	1/4/2000	12/05/2017
.MXX	IPC Mexico	1/3/2000	12/05/2017
.N225	Nikkei 225	2/2/2000	12/05/2017
.NSEI	NIFTY 50	1/3/2000	12/05/2017
.RUT	Russel 2000	1/3/2000	12/05/2017
.SPX	S&P 500 Index	1/3/2000	12/04/2017
.SSMI	Swiss Stock Market Index	1/4/2000	12/05/2017
.STI	Straits Times Index	1/3/2000	12/05/2017
.STOXX50E	EURO STOXX 50	1/3/2000	12/05/2017

Note. All dates are shown using American timing notation convention (mm/dd/yyyy)

Appendix B: Summary Statistics for the Employed Indices

Table B1: Summary statistics for the employed indices

	SP500	FTSE2	N2252	GDAXI2	RUT2	AORD2	DJI2	IXIC2	FCHI2	HSI2	KS11
						Realised Kernel					
Mean	0.011	0.008	0.011	0.017	0.010	0.005	0.010	0.013	0.014	0.009	0.013
Std	0.025	0.015	0.019	0.031	0.021	0.008	0.025	0.026	0.023	0.016	0.024
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.931	0.326	0.384	0.643	0.643	0.156	0.913	0.667	0.455	0.446	0.648
Kurt	441.7	107.0	115.9	120.7	223.0	90.7	422.0	153.3	99.2	312.6	188.4
Skew	15.2	8.0	8.8	8.5	11.1	7.6	15.4	9.1	7.8	14.2	9.9
#obs	4481	4505	4346	4538	4484	4484	4484	4487	4568	4150	4410
						Realised Variance					
Mean	0.011	0.008	0.011	0.017	0.011	0.005	0.011	0.013	0.014	0.009	0.013
Std	0.025	0.016	0.017	0.030	0.021	0.008	0.027	0.023	0.023	0.016	0.023
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
Max	0.775	0.463	0.323	0.588	0.585	0.153	0.862	0.430	0.512	0.437	0.594
Kurt	233.17	202.97	104.10	98.39	179.43	94.66	310.48	86.12	133.74	257.38	153.53
Skew	11.32	10.69	8.28	7.57	10.02	7.67	13.59	7.17	8.85	12.68	9.07
#obs	4481	4505	4346	4538	4484	4484	4484	4487	4568	4150	4410
						Returns					
Mean	0.010	-0.035	-0.033	-0.029	0.009	0.000	0.022	-0.018	-0.037	-0.047	-0.043
Std	1.159	0.929	1.157	1.297	1.400	0.802	1.114	1.353	1.198	0.994	1.176
Min	-9.351	-5.760	-10.563	-9.412	-11.053	-6.438	-8.405	-8.046	-8.124	-11.616	-11.779
Max	10.220	7.044	11.658	9.993	7.776	3.891	10.754	14.908	7.282	12.155	8.758
Kurt	11.16	7.54	14.03	8.07	7.46	6.88	11.84	10.70	7.37	16.13	9.11
Skew	-0.17	-0.15	-0.56	-0.10	-0.26	-0.49	-0.01	0.11	-0.15	0.07	-0.35
#obs	4481	4505	4346	4538	4484	4484	4484	4487	4568	4150	4410

Note. This table shows summary statistics for all employed indices. Full names and date ranges for the index tickers can be found in appendix A.

Appendix B: Summary Statistics for the Employed Indices Cont.

Table B2: Summary statistics for the employed indices cont.

	AEX	SSMI	IBEX2	NSEI	MXX	BVSP	GSPTSE	STOXX50E	FTSTI	FTSEMIB
						Realised Kernel				
Mean	0.012	0.008	0.015	0.014	0.006	0.023	0.005	0.016	0.006	0.013
Std	0.021	0.014	0.021	0.032	0.011	0.038	0.014	0.032	0.009	0.020
Min	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000
Max	0.396	0.280	0.477	1.011	0.301	0.836	0.355	1.092	0.277	0.430
Kurt	70.8	86.7	116.4	326.3	202.1	137.1	185.1	354.5	283.4	113.6
Skew	6.6	7.3	8.1	14.1	10.7	9.5	11.2	13.8	12.0	7.9
#obs	4567	4490	4533	3901	4486	4388	3897	4543	3879	4524
						Realised Variance				
Mean	0.012	0.009	0.015	0.015	0.009	0.022	0.006	0.017	0.006	0.013
Std	0.020	0.016	0.021	0.043	0.018	0.035	0.015	0.033	0.009	0.021
Min	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000
Max	0.362	0.420	0.551	1.883	0.521	0.676	0.360	1.083	0.211	0.527
Kurt	65.55	159.41	155.21	926.50	259.78	126.35	193.61	312.79	180.63	140.69
Skew	6.40	9.53	9.14	24.42	12.55	9.29	11.63	13.08	10.29	8.69
#obs	4567	4490	4533	3901	4486	4388	3897	4543	3879	4524
						Returns				
Mean	-0.042	-0.019	-0.050	0.023	0.037	0.003	-0.018	-0.026	-0.034	-0.056
Std	1.150	0.971	1.255	1.194	1.279	1.711	0.867	1.326	0.908	1.250
Min	-8.416	-9.731	-7.585	-13.382	-8.262	-15.921	-7.717	-9.346	-7.706	-9.194
Max	9.237	8.681	13.037	7.130	9.953	13.251	6.476	8.267	9.472	8.231
Kurt	9.88	11.62	8.48	14.01	8.29	8.16	12.25	7.83	12.11	7.01
Skew	-0.21	-0.31	-0.04	-1.01	0.00	-0.17	-0.62	-0.20	0.38	-0.27
#obs	4567	4490	4533	3901	4486	4388	3897	4543	3879	4524

Note. This table shows summary statistics for all employed indices. Full names and date ranges for the index tickers can be found in appendix A.

Appendix C: Test Performance Benchmarking

Table C1: Power of Geometric-VaR test across business lines compared to competing tests. 5% VaR, 10% significance

	Unconditional Coverage (1)		Violation Independence (2)			Unconditional Coverage & Violation Ind. (1&2)			VaR Ind. (3)	UCC & VaR Ind. (1&3)	All Hypotheses (1,2,&3)	
	Kupiec POF	PW16 UC	Christ. CCI	Haas TBFI	PW16 Dind	Christ. CC	Haas TBF	PW16 Geom	PW16 Vind	PW16 Var	Engle CaViaR	PW16 GV
Business Line 1												
250	**0.166**	0.147	0.227	0.383	**0.473**	0.252	**0.373**	0.351	0.321	0.290	0.441	**0.442**
500	**0.067**	0.064	0.274	0.566	**0.671**	0.227	**0.545**	0.482	0.650	0.470	0.514	**0.686**
750	**0.046**	0.036	0.307	0.666	**0.791**	0.267	**0.648**	0.615	0.796	0.605	0.597	**0.838**
1000	**0.044**	0.028	0.355	0.789	**0.861**	0.288	**0.774**	0.711	0.867	0.732	0.703	**0.918**
1250	**0.033**	0.018	0.453	0.856	**0.907**	0.317	**0.842**	0.781	0.902	0.781	0.772	**0.957**
1500	**0.027**	0.017	0.512	0.904	**0.939**	0.389	**0.897**	0.842	0.949	0.861	0.854	**0.977**
Business Line 2												
250	**0.363**	0.344	0.185	0.396	**0.448**	0.379	0.411	**0.493**	0.436	0.521	0.584	**0.599**
500	**0.258**	0.233	0.219	0.560	**0.715**	0.335	0.554	**0.611**	0.675	0.646	0.633	**0.790**
750	**0.219**	0.193	0.256	0.714	**0.830**	0.334	0.701	**0.754**	0.742	0.708	0.665	**0.890**
1000	**0.203**	0.182	0.283	0.794	**0.899**	0.296	0.786	**0.813**	0.779	0.772	0.735	**0.940**
1250	**0.195**	0.170	0.364	0.859	**0.938**	0.331	0.854	**0.864**	0.811	0.830	0.761	**0.964**
1500	**0.196**	0.171	0.428	0.902	**0.955**	0.400	0.895	**0.905**	0.830	0.845	0.832	**0.980**
Business Line 3												
250	**0.062**	0.053	0.108	0.115	**0.136**	0.079	**0.111**	0.066	0.407	0.120	**0.295**	0.141
500	**0.018**	0.012	0.123	0.120	**0.144**	0.050	**0.112**	0.047	0.713	0.324	**0.403**	0.264
750	**0.009**	0.007	0.121	0.135	**0.159**	0.042	**0.125**	0.055	0.861	0.513	0.445	**0.475**
1000	**0.010**	0.003	0.098	0.137	**0.163**	0.060	**0.128**	0.047	0.922	0.668	0.515	**0.617**
1250	**0.005**	0.003	0.087	0.138	**0.160**	0.046	**0.126**	0.045	0.962	0.802	0.593	**0.743**
1500	**0.003**	0.002	0.098	0.132	**0.164**	0.042	**0.125**	0.054	0.978	0.859	0.724	**0.818**
Business Line 4												
250	**0.350**	0.346	0.205	0.421	**0.505**	0.386	0.435	**0.522**	0.393	0.501	0.591	**0.614**
500	**0.240**	0.223	0.236	0.625	**0.765**	0.326	0.618	**0.665**	0.657	0.640	0.629	**0.818**
750	**0.206**	0.182	0.279	0.757	**0.878**	0.345	0.746	**0.785**	0.733	0.714	0.696	**0.903**
1000	**0.194**	0.174	0.309	0.842	**0.934**	0.321	0.836	**0.855**	0.797	0.783	0.741	**0.952**
1250	**0.197**	0.157	0.389	0.908	**0.961**	0.345	0.901	**0.912**	0.821	0.824	0.788	**0.976**
1500	**0.183**	0.166	0.458	0.930	**0.979**	0.426	0.925	**0.940**	0.859	0.863	0.859	**0.989**

Note. This table shows the testing power at significance level 10% by generating random returns using NGARCH-t(d) models with parameters of the four business lines specified in PW16. 5% VaR estimates are computed using Historical Simulation with a rolling window of 250 observations. The simulated power is the rejection frequency of 5,000 replications. 'Ind.' and 'UCC' are shorthand notation for independence and unconditional coverage respectively. For category 'PW16', 'UC' denotes unconditional coverage test, 'Dind' duration independence test, 'Vind' VaR independence test, 'Geom' Geometric test, 'VaR' VaR test and 'GV' Geometric-VaR test. 'Kupiec POF' refers to the test of Kupiec (1995). 'Christ. CCI' and 'Christ. CC' refer to the conditional coverage independence test and conditional coverage mixed test of Christoffersen (1998) respectively. 'Haas TBFI' and 'Haas TBF' correspond to the time between failures independence test and time between failures test of Haas (2001). 'Engle CaViaR' denotes the test of Engle and Manganelli (2004). The power for all tests is computed using the method of Dufour (2006) with 9,999 trials to control sizing for smaller sample sizes. Boldfaced figures indicate the highest power per hypothesis combination in case of competing backtests.

Appendix C: Test Performance Benchmarking Cont.

Table C2: Power of Geometric-VaR test across business lines compared to competing tests. 1% VaR, 10% significance

	Unconditional Coverage (1)		Violation Independence (2)			Unconditional Coverage & Violation Ind. (1&2)			VaR Ind. (3)	UCC & VaR Ind. (1&3)	All Hypotheses (1,2,&3)	
	Kupiec POF	PW16 UC	Christ. CCI	Haas TBFI	PW16 Dind	Christ. CC	Haas TBF	PW16 Geom	PW16 Vind	PW16 Var	Engle CaViaR	PW16 GV
Business Line 1												
250	0.119	**0.124**	0.168	0.172	**0.319**	0.133	0.130	**0.185**	0.172	0.120	**0.399**	0.208
500	0.060	**0.062**	0.229	0.237	**0.440**	0.078	0.110	**0.232**	0.469	0.260	**0.425**	0.411
750	0.038	**0.039**	0.297	0.282	**0.508**	0.053	0.097	**0.313**	0.643	0.349	**0.538**	0.501
1000	**0.029**	0.025	0.365	0.324	**0.571**	0.051	0.103	**0.321**	0.758	0.487	0.619	**0.661**
1250	**0.026**	0.018	0.427	0.371	**0.617**	0.039	0.110	**0.351**	0.823	0.605	0.681	**0.771**
1500	**0.021**	0.021	0.469	0.420	**0.656**	0.033	0.117	**0.414**	0.886	0.678	0.743	**0.831**
Business Line 2												
250	0.149	**0.178**	0.187	0.189	**0.269**	0.159	0.165	**0.167**	0.188	0.147	**0.430**	0.198
500	0.083	**0.104**	0.220	0.208	**0.424**	0.088	0.098	**0.249**	0.431	0.289	**0.425**	0.408
750	0.053	**0.071**	0.245	0.242	**0.543**	0.059	0.091	**0.342**	0.572	0.348	0.477	**0.526**
1000	0.033	**0.067**	0.279	0.271	**0.600**	0.044	0.084	**0.361**	0.671	0.453	0.534	**0.660**
1250	0.022	**0.067**	0.304	0.298	**0.668**	0.026	0.069	**0.430**	0.720	0.530	0.587	**0.744**
1500	0.020	**0.057**	0.292	0.309	**0.706**	0.021	0.061	**0.474**	0.780	0.575	0.622	**0.792**
Business Line 3												
250	**0.078**	0.073	0.073	0.090	**0.146**	0.083	0.089	**0.105**	0.284	0.096	**0.328**	0.109
500	0.039	**0.042**	0.065	0.099	**0.151**	0.043	0.055	**0.057**	0.606	0.147	**0.330**	0.171
750	**0.032**	0.027	0.090	0.094	**0.154**	0.035	**0.054**	0.048	0.767	0.394	**0.408**	0.336
1000	**0.029**	0.013	0.118	0.096	**0.142**	0.030	**0.040**	0.037	0.832	0.488	**0.525**	0.488
1250	**0.027**	0.011	0.133	0.079	**0.139**	0.027	0.033	**0.039**	0.899	0.648	0.611	**0.633**
1500	**0.026**	0.009	**0.143**	0.081	0.140	0.028	0.037	**0.037**	0.936	0.751	0.688	**0.714**
Business Line 4												
250	0.145	**0.186**	0.212	0.196	**0.319**	0.149	0.156	**0.184**	0.172	0.154	**0.460**	0.221
500	0.065	**0.087**	0.242	0.218	**0.523**	0.069	0.095	**0.291**	0.421	0.270	**0.448**	0.435
750	0.042	**0.066**	0.299	0.275	**0.603**	0.056	0.082	**0.389**	0.573	0.349	0.507	**0.547**
1000	0.026	**0.080**	0.331	0.301	**0.679**	0.031	0.074	**0.435**	0.676	0.462	0.572	**0.698**
1250	0.019	**0.070**	0.340	0.341	**0.742**	0.024	0.066	**0.481**	0.735	0.524	0.611	**0.773**
1500	0.015	**0.072**	0.336	0.390	**0.786**	0.017	0.073	**0.574**	0.783	0.609	0.662	**0.836**

Note. This table shows the testing power at significance level 10% by generating random returns using NGARCH-t(d) models with parameters of the four business lines specified in PW16. 1% VaR estimates are computed using Historical Simulation with a rolling window of 250 observations. The simulated power is the rejection frequency of 5,000 replications. 'Ind.' and 'UCC' are shorthand notation for independence and unconditional coverage respectively. For category 'PW16', 'UC' denotes unconditional coverage test, 'Dind' duration independence test, 'Vind' VaR independence test, 'Geom' Geometric test, 'VaR' VaR test and 'GV' Geometric-VaR test. 'Kupiec POF' refers to the test of Kupiec (1995). 'Christ. CCI' and 'Christ. CC' refer to the conditional coverage independence test and conditional coverage mixed test of Christoffersen (1998) respectively. 'Haas TBFI' and 'Haas TBF' correspond to the time between failures independence test and time between failures test of Haas (2001). 'Engle CaViaR' denotes the test of Engle and Manganelli (2004). The power for all tests is computed using the method of Dufour (2006) with 9,999 trials to control sizing for smaller sample sizes. Boldfaced figures indicate the highest power per hypothesis combination in case of competing backtests.

Appendix D: Model Performance

Table D1: Geometric-VaR test model performance over 24 January 2006 - 5 December 2017 for various rolling windows (5% VaR)

		HS	HEAVY Realised Variance Based						HEAVY Realised Kernel Based						GARCH					
		HS	HNv	HTv	HFv	gHNv	gHTv	gHFv	HNk	HTk	HFk	gHNk	gHTk	gHFk	GN	GT	GF	gGN	gGT	gGF
$T_e = 250$																				
$\bar{p}$	UC	0.60	0.10	0.04	0.48	0.07	0.04	0.40	0.11	0.06	0.50	0.08	0.03	0.38	0.11	0.03	0.53	0.06	0.02	0.43
	Dind	0.00	0.28	0.34	0.23	0.35	0.39	0.29	0.37	0.35	0.24	0.38	0.41	0.30	0.05	0.05	0.05	0.44	0.47	0.44
	Vind	0.07	0.08	0.08	0.03	0.06	0.07	0.01	0.09	0.08	0.02	0.10	0.08	0.02	0.04	0.05	0.01	0.09	0.07	0.01
	GV	0.00	0.05	0.01	0.08	0.02	0.00	0.03	0.04	0.02	0.06	0.01	0.00	0.04	0.01	0.01	0.01	0.03	0.01	0.04
$\#p > 0.1$	UC	**21**	5	1	20	4	1	20	6	2	20	3	2	20	6	3	20	5	1	20
	Dind	0	14	13	11	16	17	14	16	15	12	15	16	15	2	3	2	**19**	**19**	18
	Vind	5	4	4	2	5	4	0	5	4	1	4	4	2	2	2	0	**8**	6	0
	GV	0	2	0	**5**	1	0	2	3	0	3	1	0	1	1	1	0	1	0	3
$T_e = 1000$																				
$\bar{p}$	UC	0.22	0.27	0.15	0.64	0.25	0.16	0.57	0.21	0.12	0.60	0.21	0.11	0.60	0.34	0.11	0.64	0.24	0.08	0.62
	Dind	0.00	0.28	0.24	0.19	0.33	0.39	0.29	0.30	0.33	0.31	0.34	0.41	0.35	0.11	0.09	0.12	0.52	0.56	0.54
	Vind	0.12	0.44	0.47	0.33	0.36	0.34	0.27	0.46	0.43	0.31	0.37	0.36	0.32	0.26	0.26	0.23	0.43	0.44	0.31
	GV	0.00	0.20	0.12	0.33	0.18	0.12	0.33	0.17	0.10	0.39	0.19	0.10	0.41	0.17	0.06	0.20	0.26	0.13	0.53
$\#p > 0.1$	UC	14	14	6	**21**	12	6	**21**	10	5	**21**	11	5	**21**	15	9	**21**	11	7	**21**
	Dind	0	13	14	13	16	16	18	14	15	16	16	16	16	9	4	7	19	**20**	**20**
	Vind	5	19	**20**	15	18	17	14	**20**	19	17	19	19	17	16	18	12	19	17	17
	GV	0	13	9	15	11	7	16	9	6	18	11	7	18	10	6	12	15	10	**20**
$T_e = 1500$																				
$\bar{p}$	UC	0.19	0.24	0.16	0.63	0.14	0.06	0.61	0.16	0.13	0.67	0.11	0.03	0.70	0.37	0.07	0.63	0.26	0.04	0.66
	Dind	0.00	0.46	0.39	0.33	0.47	0.50	0.37	0.60	0.47	0.51	0.50	0.45	0.54	0.29	0.21	0.21	0.65	0.58	0.61
	Vind	0.17	0.43	0.46	0.40	0.43	0.38	0.38	0.37	0.33	0.32	0.37	0.35	0.37	0.34	0.28	0.31	0.43	0.61	0.57
	GV	0.00	0.23	0.12	0.39	0.15	0.10	0.47	0.20	0.11	0.51	0.19	0.06	0.61	0.31	0.09	0.37	0.35	0.10	0.59
$\#p > 0.1$	UC	10	14	9	18	11	6	18	10	6	17	9	5	18	14	10	**21**	13	9	20
	Dind	0	14	14	14	14	14	14	14	14	15	15	17	16	9	8	11	19	**20**	19
	Vind	5	19	18	18	17	18	16	19	18	19	**21**	19	19	16	16	13	16	16	14
	GV	0	12	9	18	11	8	16	10	7	18	11	5	16	11	7	15	12	11	**20**

Note. This table shows the performance of various models tested with the method of PW16. Statistics are given for three rolling window sizes, $T_e \in \{250, 500, 1500\}$, with $\bar{p}$ denoting the average p-value of the test and $\#p > 0.1$ the number of cases (out of 21) that the null hypothesis of correct VaR estimates was not rejected. 'UC' denotes the 'Unconditional Coverage' test, 'Dind' the 'Duration Independence' test, 'Vind' the 'VaR Independence' test and 'GV' the 'Geometric-VaR' test. Boldfaced figures denote the best performing model(s) per particular test.

Appendix D: Model Performance Cont.

Table D2: Economic model statistics over 24 January 2006 - 5 December 2017 for various rolling windows (5% VaR)

	HS	HEAVY Realised Variance Based						HEAVY Realised Kernel Based						GARCH					
	HS	HNv	HTv	HFv	gHNv	gHTv	gHFv	HNk	HTk	HFk	gHNk	gHTk	gHFk	GN	GT	GF	gGN	gGT	gGF
$T_e = 250$																			
VaR	1.82	1.68	1.66	1.76	1.68	1.66	1.76	1.67	1.66	1.76	1.67	**1.66**	1.76	1.70	1.67	1.78	1.68	1.66	1.78
Cum. Rank	394	195	93	317	177	76	319	167	63	296	150	**57**	301	249	132	359	202	97	346
σ(VaR)	**0.74**	0.91	0.90	0.99	0.92	0.92	1.01	0.90	0.90	0.99	0.92	0.92	1.01	0.88	0.85	0.97	0.90	0.92	1.01
Cum. Rank	**21**	166	150	323	210	199	361	154	142	328	194	193	350	99	71	315	138	220	356
$T_e = 1000$																			
VaR	1.86	1.60	1.58	1.68	1.60	1.58	1.68	1.59	**1.57**	1.67	1.59	1.57	1.68	1.62	1.59	1.70	1.60	1.58	1.69
Cum. Rank	394	187	98	316	191	90	316	152	**64**	294	154	72	296	238	147	344	196	111	330
σ(VaR)	**0.44**	0.83	0.83	0.88	0.85	0.85	0.91	0.83	0.83	0.88	0.85	0.85	0.90	0.80	0.78	0.86	0.82	0.85	0.88
Cum. Rank	**21**	180	164	300	230	219	344	177	165	294	225	223	335	108	101	232	148	240	284
$T_e = 1500$																			
VaR	1.90	1.65	1.63	1.73	1.65	1.63	1.74	1.64	**1.62**	1.72	1.64	1.62	1.73	1.67	1.63	1.75	1.66	1.63	1.75
Cum. Rank	376	185	94	303	200	99	326	143	**57**	290	155	72	304	234	144	341	209	120	338
σ(VaR)	**0.28**	0.86	0.86	0.90	0.88	0.88	0.92	0.86	0.86	0.90	0.88	0.88	0.92	0.81	0.80	0.86	0.84	0.88	0.89
Cum. Rank	**21**	180	180	286	237	230	335	175	179	283	236	250	343	119	104	178	147	251	256

Note. This table shows economic statistics of the VaR models investigated. Statistics are given for three rolling window sizes, $T_e \in \{250, 500, 1500\}$, with 'VaR' denoting the average Value-at-Risk of the model over the sample period and across all 21 indices and σ(VaR) the average standard deviation of the Value-at-Risk estimates over the sample period and across indices. Note that for $T_e = 250$ and $T_e = 1000$ the first 1250 and 500 VaR estimates are truncated to get the same effective sample period across the rolling window sizes. Per statistic, 'Cum. Rank' indicates the cumulative rank of the mean VaR and standard deviation of VaR respectively over all indices. Lower indicates better performance. The best performing model per rolling window size is indicated in bold.

Appendix D: Model Performance Cont.

Table D3: Geometric-VaR test model performance over 2 July 2007 - 30 June 2009 for various rolling windows (5% VaR)

		HS	HEAVY Realised Variance Based						HEAVY Realised Kernel Based						GARCH					
		HS	HNv	HTv	HFv	gHNv	gHTv	gHFv	HNk	HTk	HFk	gHNk	gHTk	gHFk	GN	GT	GF	gGN	gGT	gGF
									$T_e = 250$											
$\bar{p}$	UC	0.12	0.27	0.27	0.58	0.15	0.14	0.56	0.34	0.33	0.55	0.14	0.17	0.55	0.08	0.10	0.35	0.12	0.11	0.40
	Dind	0.00	0.50	0.47	0.56	0.45	0.46	0.36	0.61	0.58	0.45	0.40	0.51	0.31	0.45	0.44	0.33	0.60	0.62	0.49
	Vind	0.07	0.29	0.31	0.19	0.22	0.22	0.21	0.22	0.30	0.24	0.19	0.20	0.20	0.21	0.24	0.20	0.16	0.26	0.26
	GV	0.00	0.21	0.25	0.29	0.18	0.17	0.31	0.25	0.27	0.30	0.18	0.19	0.28	0.10	0.12	0.14	0.14	0.16	0.15
$\#p > 0.1$	UC	4	10	10	**20**	9	7	**20**	11	10	**20**	8	8	**20**	8	6	14	6	4	18
	Dind	0	19	19	17	21	21	18	**21**	20	18	21	20	18	18	18	17	**21**	**21**	**21**
	Vind	7	11	**12**	9	9	8	7	11	9	7	10	10	7	6	7	6	7	8	4
	GV	0	6	9	12	7	7	14	7	8	11	7	6	**15**	6	5	9	5	4	10
									$T_e = 1000$											
$\bar{p}$	UC	0.00	0.13	0.15	0.37	0.16	0.20	0.40	0.26	0.21	0.44	0.21	0.17	0.34	0.13	0.12	0.17	0.14	0.13	0.23
	Dind	0.03	0.57	0.36	0.44	0.48	0.50	0.53	0.49	0.50	0.48	0.54	0.53	0.42	0.43	0.36	0.28	0.72	0.64	0.59
	Vind	0.34	0.34	0.37	0.26	0.35	0.35	0.33	0.35	0.43	0.35	0.31	0.34	0.31	0.25	0.23	0.25	0.27	0.28	0.24
	GV	0.00	0.12	0.11	0.20	0.17	0.19	0.24	0.26	0.25	0.29	0.23	0.22	0.27	0.11	0.11	0.14	0.13	0.13	0.20
$\#p > 0.1$	UC	0	8	7	14	7	8	14	7	7	**15**	8	7	14	3	3	8	5	3	7
	Dind	7	18	17	17	18	18	19	19	19	17	19	19	18	19	18	19	**21**	**21**	**21**
	Vind	18	15	**16**	14	12	11	12	14	13	15	12	11	12	8	10	8	10	10	10
	GV	0	9	8	11	7	8	12	9	8	**13**	7	7	12	4	4	5	4	4	5
									$T_e = 1500$											
$\bar{p}$	UC	0.00	0.33	0.36	0.53	0.24	0.27	0.45	0.36	0.27	0.53	0.31	0.25	0.49	0.17	0.12	0.25	0.14	0.10	0.33
	Dind	0.03	0.48	0.46	0.56	0.52	0.43	0.46	0.46	0.52	0.40	0.42	0.50	0.49	0.37	0.31	0.37	0.59	0.61	0.60
	Vind	0.08	0.34	0.39	0.36	0.33	0.34	0.31	0.40	0.46	0.38	0.36	0.39	0.38	0.23	0.21	0.23	0.23	0.28	0.26
	GV	0.00	0.24	0.30	0.26	0.22	0.21	0.26	0.33	0.29	0.34	0.31	0.28	0.40	0.12	0.13	0.18	0.11	0.12	0.26
$\#p > 0.1$	UC	0	13	10	16	10	12	15	14	13	**19**	10	9	16	7	5	10	6	5	11
	Dind	4	16	17	16	15	15	18	17	17	17	18	18	18	20	20	17	20	20	**21**
	Vind	11	**16**	**16**	**16**	14	14	15	17	15	15	14	14	13	12	13	11	13	12	11
	GV	0	12	11	14	10	9	12	12	11	**15**	9	10	12	6	6	7	6	5	8

Note. This table shows the performance of various models tested with the method of PW16 during the global financial crisis. Statistics are given for three rolling window sizes, $T_e \in \{250, 500, 1500\}$, with $\bar{p}$ denoting the average p-value of the test and $\#p > 0.1$ the number of cases (out of 21) that the null hypothesis of correct VaR estimates was not rejected. 'UC' denotes the 'Unconditional Coverage' test, 'Dind' the 'Duration Independence' test, 'Vind' the 'VaR Independence' test and 'GV' the 'Geometric-VaR' test. Boldfaced figures denote the best performing model(s) per particular test.