Journal of Risk Model Validation **17**(2), 1–23
DOI: 10.21314/JRMV.2023.002

Research Paper

What can we expect from a good margin model? Observations from whole-distribution tests of risk-based initial margin models

David Murphy

Law School, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK; email: d.murphy3@lse.ac.uk

(Received November 29, 2022; revised January 19, 2023; accepted January 19, 2023)

ABSTRACT

Initial margin is typically calculated by applying a risk-sensitive model to a portfolio of derivatives with a counterparty. This paper presents an approach to testing initial margin models based on their predictions of the whole future distribution of returns of the relevant portfolio. This testing methodology is substantially more powerful than the usual "backtesting" approach based on returns in excess of margin estimates. The approach presented also provides a methodology for calibrating margin models via the examination of how test results vary as the model parameters change. We present the results of testing some popular classes of initial margin models for various calibrations. These give some insight into what it is reasonable to expect from an initial margin model. In particular, we find that margin models meet regulators' expectations that they are accurate around the 99th and 99.5th percentile of returns, but that they do not, for the examples studied, accurately model the far tails. Moreover, different models, all of which meet regulatory expectations, are shown to provide substantially different margin estimates in the far tails. The policy implications of these findings are discussed.

Keywords: backtesting; conditional volatility; filtered volatility; initial margin model; margin model testing; volatility estimation.

Print ISSN 1753-9579 | Online ISSN 1753-9587

1 INTRODUCTION

The margining of derivatives portfolios is a key element of the regulatory reforms that followed the 2007–9 global financial crisis. Many market participants are now required to post initial and variation margin on bilateral derivatives positions,[1] while central counterparties (CCPs), whose use is often mandated,[2] require margin on cleared positions.

Initial margin requirements for portfolios of derivatives are often estimated using a risk-based margin model. Due to the importance of margin as the first line of defense against counterparty credit risk, the behavior and prudence of margin models have come under increasing scrutiny.[3] This paper contributes to this growing body of literature.

The design of a modern initial margin model for derivatives portfolios typically relies on the assumption that portfolios can be liquidated over some fixed time period, known as the margin period of risk (MPOR). The model is designed to estimate how much a portfolio could change in value over this period.[4] This estimate is usually either a quantile of the distribution of portfolio value changes, such as the 99th, or a quantity closely related to this.[5]

There are a number of properties that are significant for margin models. For our purposes, the following three are important.

Portfolio risk sensitivity. This is a desirable property: initial margin should be sensitive to the return distribution of the portfolio, accurately calculating the target quantile.

Market conditions sensitivity. Some margin models are also sensitive to changes in market conditions, requiring more margin for the same portfolio as its underlying risk factors become more volatile. In particular, the margin models used by most CCPs are risk sensitive in this sense. Other margin models, notably the most common model for bilateral derivatives (see International Swaps and Derivatives Association 2021), calculate margin under the assumption that market conditions

[1] See Basel Committee on Banking Supervision (2013) for details of the regulatory requirements for bilateral margin.

[2] The regulatory requirement to clear certain over-the-counter (OTC) derivatives and for CCP initial margin models in Europe can be found in, for example, European Union (2012, 2013, 2015).

[3] See Basel Committee on Banking Supervision (2021) for a recent authoritative review of margining practices.

[4] Portfolios so big or risky that they cannot plausibly be liquidated over the MPOR are often subject to an additional margin charge known as concentration margin. This is often calculated outside the model.

[5] Some initial margin models estimate expected shortfall above a quantile rather than the quantile itself.

are experiencing (a fixed degree of) stress. This substantially reduces the variability of margin due to changes in market conditions, but produces higher margin requirements in placid markets.

Excess procyclicality. Margin estimates should not overreact to changes in market conditions, as this creates extra costs for market participants and increases funding liquidity risk.[6]

The accuracy of margin models is often tested by examining their exceedances; that is, the occasions when portfolio losses exceed margin estimates. Backtests, as these tests are known, are typically relatively low power, simply because exceedances are rare. Most days do not generate one, so new information on model performance only arrives occasionally.

Therefore, it is helpful to carry out tests of margin model performance in addition to backtesting. Many margin models, including the most popular ones, provide estimates of the whole distribution of portfolio returns one or more days hence. Tests based on comparing these estimates with the observed returns are often more powerful than backtesting, and they can give significant insights into model performance. This paper presents a whole-distribution approach to testing initial margin models. The approach is also useful for calibrating margin models, as test results across a range of calibration parameters can be examined.

The next two subsections give a short introduction to the approach presented here and its contribution to the literature, while the third discusses related work.

1.1 Whole-distribution tests of initial margin models

One way to describe an initial margin model is to begin with a portfolio and its returns over time. Suppose r_t, $1 \leqslant t \leqslant T$, is a time series of these log returns up to now (time $t = T$). The next return, r_{T+1}, is uncertain. The margin model predicts its distribution, say $\phi_{T+1}(r)$ (Section 2 discusses exactly how to extract these distribution predictions from a number of popular initial margin models). If the margin model targets a particular quantile of the return distribution, say α, then the margin is estimated as the α quantile of this distribution, $Q_\alpha(\phi_{T+1}(r))$.

Backtests examine situations in which there is a loss bigger than the margin $-r_{T+1} > Q_\alpha(\phi_{T+1}(r))$, or perhaps where there is also a sufficiently large profit (ie, $\text{abs}(r_{T+1}) > Q_\alpha(\phi_{T+1}(r))$).[7] However, as already noted, this approach, based on exceedances, discards a lot of information: since $\alpha \sim 0.99$, most of the time $\text{abs}(r_{T+1}) < Q_\alpha(\phi_{T+1}(r))$.

[6] See Murphy *et al* (2014), Maruyama and Cerezetti (2019) and Murphy and Vause (2022) for further discussions of initial margin model procyclicality.

[7] These are known, respectively, as one-sided or two-sided backtests.

The extra information available from knowing the realization r_{T+1} can be used by observing that the margin prediction $\phi_{T+1}(r)$ tells us how probable r_{T+1} is:

$$\Pr(r < r_{T+1} \mid \phi_{T+1}).$$

If the margin model makes good predictions, the time series of these cumulative probabilities of seeing r contingent on ϕ should be uniformly distributed on $[0, 1]$. Testing this allows us to determine how good the predictions $\phi_{T+1}(r)$ are.

1.2 Our contribution

This paper presents a methodology for testing initial margin models based on their prediction of the future return distribution of margined portfolios and shows how to apply it to a number of popular classes of initial margin models, and for a range of portfolios whose return distributions have different characteristics (of, for instance, skewness, kurtosis and fatness of tails). The results give some insights into how accurately these models capture the return distribution, and hence what it is reasonable to expect from them. Our aim is not to advocate for or against a particular class of models, but rather to illustrate how to use the whole distribution of returns to test the performance of an initial margin model.[8]

Our approach allows us to examine how initial margin model performance varies with calibration. This is an important question, as margin takers regularly test and, if necessary, recalibrate initial margin models. Indeed, European regulation requires that CCPs conduct this "sensitivity analysis" regularly and submit their results to the CCP's risk committee (see European Union 2013, Article 50).

For some popular models, we find that there is a range of calibration such that the model provides a good estimate of nearly all of the return distribution ϕ_{T+1}, and in particular from the center out to the biggest regulatory minimum margin quantile $\alpha = 99.5\%$.[9]

However, we find that for no calibration none of the models studied is a good fit for the entire distribution.[10] This means in particular that the currently popular margin models do not give good estimates of far tail quantiles such as $Q_{99.9\%}$, and it should not be a surprise to see occasional returns substantially in excess of margin. This has policy implications: it means that resources beyond margin, such as the default funds currently required for CCPs, are likely necessary if robust protection against counterparty credit risk beyond 99.5%, or thereabouts, is desired.

[8] See Buczyński and Chlebus (2020), Hansen and Lunde (2005) and Stărică (2003) for a discussion of the accuracy of a wider range of risk models.

[9] Technically, the hypothesis that the estimate of the distribution within the quantiles [0.2%, 99.8%] is correct, using our chosen test, cannot be rejected at 95% confidence.

[10] The hypothesis that the estimate is correct, using our chosen test, is rejected at 95% confidence.

The far tails of the portfolio return distribution pose difficulties, so it is natural to examine how different margin models address them. A very stressful period associated with the onset of Covid-19 in March 2020 is chosen to explore this. The levels of margin estimated by different margin models are examined and found to vary significantly. This suggests that there is substantial uncertainty about the margin requirements in these conditions, and hence about how reactive a margin model should be. Criteria other than reactiveness, such as procyclicality, can therefore sometimes be used to choose between otherwise acceptable models.

1.3 Related work

The framework for backtesting margin models compares realized ex-post returns and assesses whether the number of times the observed returns exceeded or breached the forecast is consistent with the predicted percentile. The formalization of this approach through the use of a statistical test is discussed by Christoffersen (1998) and Kupiec (1995), while Campbell (2005) gives the regulatory perspective on the use of these tests. More sophisticated approaches to backtesting that consider properties of the exceedances, such as their size and correlation, are discussed in Escanciano and Pei (2012) and Haas (2001), while Gurrola-Perez (2018) discusses the particular question of testing an important class of initial margin models: filtered historical simulation (FHS) value-at-risk (VaR) models.

There is relatively little literature on other approaches to testing initial margin models. Diebold *et al* (1998) present a methodology for testing density forecasts that uses similar information to our whole-distribution tests. More recently, Houllier and Murphy (2017) discussed whole-distribution margin model testing using a similar approach to the one presented here. That paper was based on the properties of the worst loss experienced by the portfolio over the MPOR. The test it discusses does not rely on any assumptions about how the return distribution scales as the MPOR grows, so it is particularly suitable for settings with longer MPORs.[11] However, the test used is lower power than the one presented here, so our current approach scrutinizes models more closely.

2 METHODOLOGY AND EXAMPLE

This section outlines the initial margin models studied and the methodology for the tests performed on them.

[11] In particular, it does not assume $\sqrt{t}$ scaling, an important consideration, as Daníelsson and Zigrand (2006) discuss.

2.1 Notation and margin models

Let r_t, $1 \leqslant t \leqslant T$, be a time series of portfolio returns and let λ be a set of initial margin model calibration parameters. For a target margin quantile α, a margin model m with calibration λ is a function that estimates the distribution of r_{T+1}. We write $\phi_{T+1}(r)$ for this estimate.

A historical simulation VaR model with window size M estimates $\phi_{T+1}(r)$ as the empirical distribution of

$$\{r_{T-M+1}, r_{T-M+2}, \ldots, r_T\}.$$

The margin estimate is the α quantile of this distribution, as it is for all the models discussed here.

A parametric VaR model with window size M estimates $\phi_{T+1}(r)$ as the normal distribution $N(\mu, \sigma)$, where μ is the average of

$$\{r_{T-M+1}, r_{T-M+2}, \ldots, r_T\}$$

(or sometimes zero in practice), and σ is a volatility estimate for

$$\{r_{T-M+1}, r_{T-M+2}, \ldots, r_T\}.$$

An FHS VaR model with window size M estimates $\phi_{T+1}(r)$ as the scaled empirical distribution

$$\sigma_T\left(\frac{r_{T-M+1}}{\sigma_{T-M+1}}, \frac{r_{T-M+2}}{\sigma_{T-M+2}}, \ldots, \frac{r_T}{\sigma_T}\right),$$

where σ_t is a volatility estimate at t.[12] This is usually determined using an exponentially weighted moving average (EWMA) volatility estimator with decay parameter λ, so this margin model has two calibration parameters: M and λ.

2.2 Empirical distributions

From the above, it is clear that some margin models use finite series of returns $r_1, \ldots, r_M$ to model the distribution of future returns. It will be necessary to estimate the cumulative probability of seeing a given return r contingent on this discrete distribution. To do this, a continuous cumulative distribution function (CDF) is derived from $r_1, \ldots, r_M$ as follows.

(1) The empirical series is sorted, producing $r_{(1)}, \ldots, r_{(M)}$ with $r_{(i)} \leqslant r_{(i+1)}$ for all i.

[12] See Barone-Adesi and Giannopoulos (2001), Barone-Adesi *et al* (1999), Gurrola-Perez and Murphy (2015), Hull and White (1998) and Jorion (2006) for more details on the various VaR models.

(2) The cumulative probability of $r = r_{(i)}$ is set to $(1/M)(i - 1/2)$.

(3) For returns $r_{(i)} < r <_{(i+1)}$ between the observed ones, straight line interpolation is used, setting the cumulative probability as

$$\frac{1}{M}\left[\left(i - \frac{1}{2}\right) + \frac{r - r_{(i)}}{r_{(i+1)} - r_{(i)}}\right].$$

(4) For returns $r < r_{(1)}$ or $r > r_{(M)}$ beyond the empirical ones, the fact that the cumulative probability at $r_{(1)}$ is $1/(2M)$ is used, and the σ is found such that the CDF of the normal distribution of $N(\mu, \sigma)$ at $r_{(1)}$ is $1/(2M)$, where μ is the mean of $r_{(1)}, \ldots, r_{(M)}$. Then the CDF of this normal distribution is used in the left tail, following a symmetrical procedure in the right tail.

2.3 The probability of a return contingent on an estimated distribution

Suppose a margin model produces an estimate of the distribution of returns at $T + 1$, $\phi_{T+1}(r)$, based on returns $r_1, \ldots, r_T$ and calibration λ. If ϕ is continuous, then $\Pr(r < r_{T+1} \mid \phi_{T+1})$ is written for the cumulative probability of observing a given return r_{T+1} at $T + 1$ given ϕ_{T+1}, that is, for

$$\int_{-\infty}^{r_{T+1}} \phi_{T+1}(r)\,\mathrm{d}r.$$

If ϕ_{T+1} is discrete, it is necessary to first smooth it as in Section 2.2; the definition immediately above can then be used. Each rolling window $[t = 1, \ldots, t = T]$, $[t = 2, \ldots, t = T + 1], \ldots$ gives a prediction of the return distribution the day after its end, $\phi_{T+n}(r)$, $n = 1, 2, \ldots$. Using the return observed at $T + n$, r_{T+n}, we calculate the cumulative probability of observing it given this predicted distribution: $\Pr(r < r_{T+n} \mid \phi_{T+n})$.

A single probability tells us very little, but the time series of cumulative probabilities is more informative. One criterion for the accuracy of the estimates ϕ_{T+n} is that the time series of cumulative probabilities $\{\Pr(r < r_{T+n} \mid \phi_{T+n})\}$, $n = 1, 2, \ldots$, is uniformly distributed: the r_{T+n} should randomly sample their respective distributions.

To see what happens when ϕ_{T+n} is biased, suppose first that it is systematically too wide. Then the cumulative probability $\Pr(r < r_{T+n})$ will be too low for returns close to zero, and too high for returns with higher absolute values. This will result in a somewhat $\vee$-shaped distribution of $\{\Pr(r < r_{T+n} \mid \phi_{T+n})\}$ rather than a uniform one. Similarly, if ϕ_{T+n} is too narrow, it will assign too low a probability to returns with a large absolute value, and hence a somewhat $\wedge$-shaped distribution will be obtained.

2.4 Testing uniformity

Unfortunately, direct tests of the uniformity of a time series are typically relatively low power. One promising approach is to transform the series being tested using the inverse cumulative normal distribution, then test the normality of the transformed series using the (typically high power) Shapiro–Wilk test.[13] If this test produces a p-value in excess of 0.05, the hypothesis that the original time series is uniform cannot be rejected at 95%. This is the principal test used in this paper.

2.5 Calibration

Let λ be a set of calibration parameters for a margin model. Then, a margin model is a function from (r_t, λ) to the space of distributions of $\phi_{T+1}(r)$. Acceptable calibrations λ can be studied by examining how the p-value of the test of $\Pr(r < r_{T+n} \mid \phi_{T+n})$, $n = 1, 2, \ldots$, varies with λ.

2.6 Margin model testing and calibration

In summary, the approach to testing and calibrating a margin model m taken in this paper is as follows.

(1) Select a portfolio with time series of returns r_i.

(2) Select a series of possible model calibrations λ_j.

(3) For some test period, $T + 1 \leqslant t \leqslant T + N$, calculate the model estimates of ϕ_t for each calibration.

(4) Calculate $\Pr(r < r_t \mid \phi_t(r))$ for each t in the test period and each calibration.

(5) For each calibration, test the uniformity of each time series of probabilities.

(6) Plot the p-value of the test as a function of the calibration parameter(s) λ_j and compare it with the critical value to determine the range of acceptable parameters, if any.

Testing a margin model is a special case of this where there is only one calibration.

2.7 Example and discussion

In order to illustrate this approach, consider a margin model based on historical simulation VaR with a 50-day window. An outright position in the Standard & Poor's 500

[13] See Razali and Yap (2011) and Shapiro and Wilk (1965) for discussions of the power of the Shapiro–Wilk test.

(S&P 500) index will be taken, and its log returns using index levels from 2020 will be calculated.

The last window's 50 sorted log returns begin $\{-3.59\%, -1.88\%, \ldots, \}$. The raw empirical distribution would thus assign a cumulative probability of zero to any return less than -3.59%, and of $1/50$ to any return between -3.59% and -1.88%. This step function in probabilities is undesirable and produces noise in the results.

Three design choices in the smoothing of empirical probabilities help to eliminate this noise.

(1) A probability of $(i - 1/2)/50$ is assigned to $r_{(i)}$ rather than $i/50$. This means that the total probability mass of the observed returns is $49/50$, giving us $1/50$ to assign to the tails beyond the observed returns.

(2) A straight line in cumulative probabilities is used between the observations. Thus, a return of -2.5%, which is 36% of the way between -3.59% (a cumulative probability of $1/100$) and -1.88% (a cumulative probability of $3/100$), is assigned a cumulative probability of $(1/100) + (36\% \times 2/100)$ or 0.0172.

(3) The $1/100$ available for each tail beyond the observed returns is used by assuming that the tails are cumulative normal with the same mean as the sample and a standard deviation that gives the right probability mass. For this sample, the mean is 0.17%, and the value of σ such that the cumulative normal of $N(\mu = 0.17\%, \sigma)$ evaluated at -3.59% is $1/100$ is $\sigma = 1.62\%$. This means that, instead of assigning a cumulative probability of zero to a return of -4%, the cumulative normal of $N(0.17\%, 1.62\%)$ evaluated at -4% is used, or roughly 0.005.

The first window of 2020 illustrates the issue with the tails beyond the empirical distribution. The first return after it (that is, the 51st log return in 2020) is -12.8% (this return is from March 2020, at the height of the Covid-19-induced market turmoil). Based on the smoothed CDF of the first window, it has a cumulative probability of 0.14%, rather than zero. The second is 5.8%, and this has a cumulative probability of 97.5%.

The Shapiro–Wilk test used examines the hypothesis that the next return is from the same distribution as the prior window's data (after smoothing of the CDF). In this case, the hypothesis is just rejected at 95% for this (very short) test period: historical simulation VaR with a 50-day window fails with a p-value of 0.055 for the S&P 500 index in 2020.

3 TEST RESULTS FOR SOME INITIAL MARGIN MODELS

In this section, we present the results for a long position in the S&P 500 index using a data series of 5488 returns from January 2000 to November 2021. The online appendix provides robustness checks by examining other portfolios.

Three classes of margin model are considered: historical simulation VaR, parametric VaR and FHS VaR. Each class of model is tested and the results presented. These results suggest tentative observations about the performance of the class of models, and of initial margin models more generally. These are given in boxes at the end of each relevant subsection.

3.1 Historical simulation VaR

Consider a historical simulation VaR model with window length M. Such a model is based on the idea that the distribution of the prior M returns provide an acceptable model of this distribution of the current return. In order to examine this claim, a reasonable choice of window, $M = 300$, is used. The Q–Q plot of the next return versus the smoothed CDF of the prior M returns is presented in Figure 1.

Visual inspection of Figure 1 hints at a good fit over most of the distribution (from roughly 0.3% to 99.7%, or $\pm 2.75\sigma$) but also suggests that the probability of large positive or negative returns is not well estimated by the prior window. This is confirmed by the p-values. Figure 2 shows the results (red diamonds) of the Shapiro–Wilk test for the hypothesis that the next return comes from the same distribution as the returns in a prior window, as a function of window length. The critical value is shown as a dashed line. It can be seen that the hypothesis is rejected for all window lengths.

In order to probe the hypothesis that the issue arises in the tails, the far tails outside $(0.2\%, 99.8\%)$ were removed and the test rerun. The results (blue triangles) support the intuition that it is the far tails that are causing the prior failures. For the truncated distribution, window lengths from 300 to 400 observations pass. Since the typical quantiles for margin, 99% and 99.5%, are well within $(0.2\%, 99.8\%)$, this implies that the hypothesis that historical simulation VaR can estimate margin accurately cannot be rejected, at least for some calibrations. The results presented in the online appendix confirm that this is not an idiosyncratic feature of this portfolio or risk factor. Therefore, the following observation is suggested.

OBSERVATION 3.1 Historical simulation VaR can, for some calibrations, produce accurate estimates of initial margin at 99% and 99.5%. It does not provide good estimates of the far tails of the one-day forward return distribution, and hence it does not estimate expected shortfall accurately.

FIGURE 1 Q–Q plot of the next return versus the empirical distribution of the prior 300 returns.

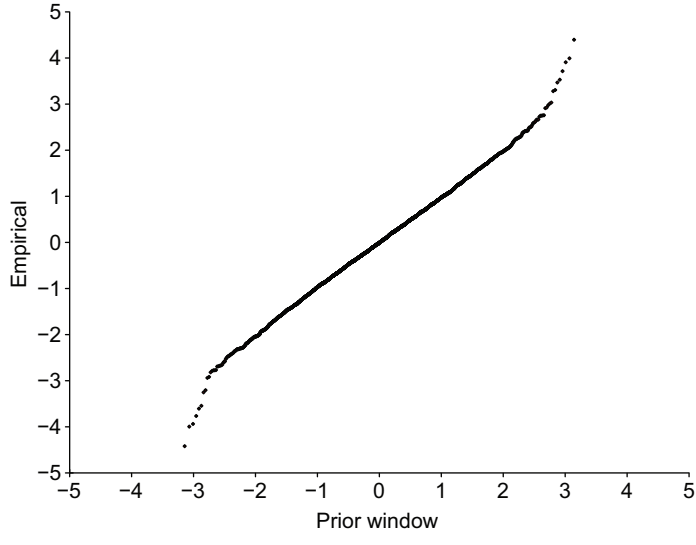


FIGURE 2 Test results for historical simulation VaR.

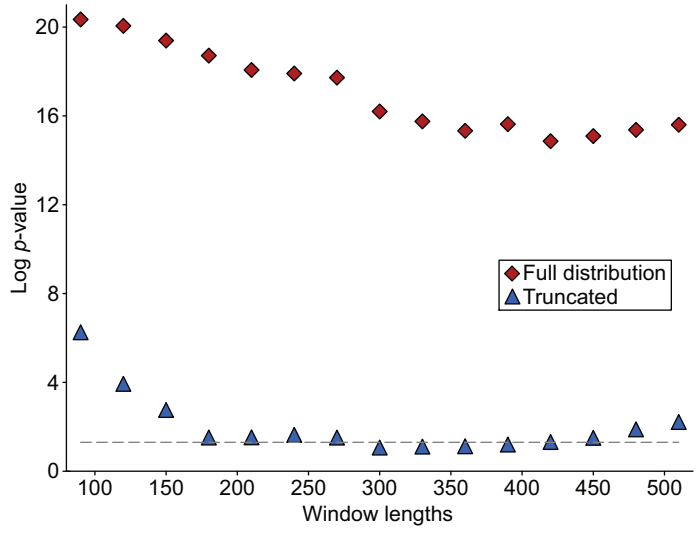
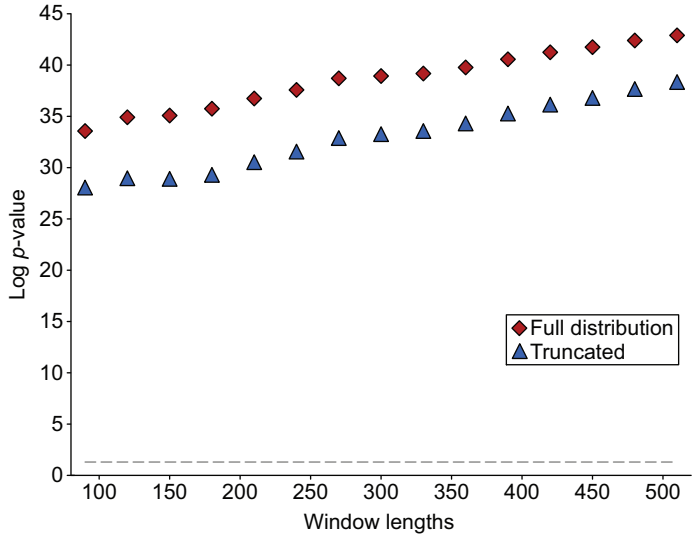


FIGURE 3 Test results for parametric VaR using an unweighted volatility estimate.



3.2 Parametric VaR

Figure 3 presents the test results for parametric VaR using an unweighted volatility estimate. This model of the whole return distribution is, unsurprisingly, worse than the prior window empirical distribution. Truncation helps, but not enough to allow it to pass the test: the hypothesis that the next return comes from the normal distribution with the mean and standard deviation of the prior window is rejected at 95% for all window lengths studied.

3.3 Parametric VaR using conditional volatility

It would be reasonable to conjecture that the problem with the parametric VaR studied in the previous section is that the volatility estimator is too primitive. In order to investigate this, two more sophisticated volatility estimators are considered: EWMA volatility and the conditional volatility from a type of generalized autoregressive conditional heteroscedasticity (GARCH) model.

EWMA volatility estimators start with some long-term volatility estimate, then estimate conditional volatility recursively via

$$\sigma_T^2 = (1-\lambda) r_{T-1}^2 + \lambda \sigma_{T-1}^2.$$

The decay parameter λ controls how quickly data falls off: if it is close to one, the estimator has a long memory, while smaller lambdas, such as 0.96, forget the volatility contribution of older returns more quickly.

GARCH models, as described by Bollerslev (1986), are a popular class of stochastic volatility model. In the simplest version of these models, GARCH(1, 1), volatility evolves as

$$\sigma_T^2 = \omega + \alpha r_{T-1}^2 + \beta \sigma_{T-1}^2.$$

Thus, today's conditional volatility, σ_T^2, depends on yesterday's conditional volatility, σ_{T-1}^2, and yesterday's return, r_{T-1}^2. Typically, $\alpha, \beta > 0$ while $\alpha + \beta < 1$ is required, so large returns yesterday produce increases in volatility today. The parameters of these models (α, β and ω) are fitted using quasi-maximum likelihood estimation.

The extension of GARCH(1, 1) proposed by Glosten *et al* (1993) allows a differential (and in practice usually larger) impact on volatility from negative returns than from positive ones. An extra parameter, γ, is introduced and volatility evolves as

$$\sigma_T^2 = \omega + (\alpha + \mathbf{1}_{T-1}\gamma) r_{T-1}^2 + \beta \sigma_{T-1}^2, \quad \text{where } \mathbf{1}_{T-1} = \begin{cases} 0 & \text{if } r_{T-1} > 0, \\ 1 & \text{otherwise.} \end{cases}$$

We calculated EWMA and Glosten–Jagannathan–Runkle-GARCH (GJR-GARCH) conditional volatilities for our data.[14] Figure 4 illustrates the conditional volatilities calculated by an EWMA volatility estimator with $\lambda = 0.97$ and by a GJR-GARCH volatility estimator using rolling 2000-day windows.

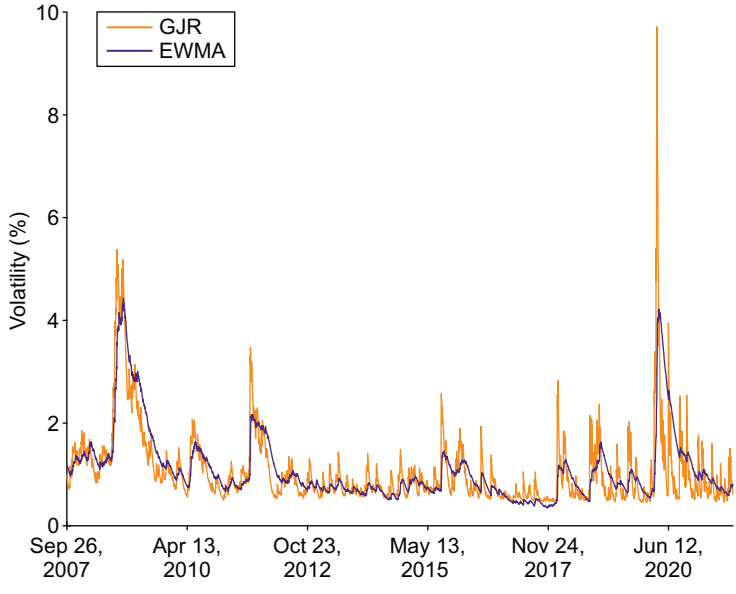
It can be seen that the GJR-GARCH estimator is somewhat more reactive than the EWMA one: it goes higher in stressed markets, such as those of March 2020, and falls lower during placid ones, such as those in mid-2017.

The parameter for the EWMA volatility estimator is λ, the decay parameter; for GJR-GARCH, it is the window length. Therefore, it makes sense to study the performance of these two volatility estimators as these parameters vary. Figure 5 presents the results for both the full distribution of returns and the one with the far tails truncated. It can be seen that EWMA performance is flat for a range of lambdas, but then declines as reactiveness decreases. For GJR-GARCH, performance is not strongly determined by window length.[15] However, parametric VaR derived from

[14] This is not a trivial procedure, in that GJR-GARCH fitting, like many forms of GARCH model fitting, requires care, as Hill and McCullough (2019) discuss. Following their recommendation, the `rugarch` package in R with the hybrid solver and control of the numerical differentiation parameters is used to manage convergence.

[15] GJR-GARCH, in common with many GARCH approaches, requires substantial amounts of data for good fitting (see Stărică (2003) for a discussion). However, the extent of this might be surprising: even moving from 8 years of data (a 2000-return window) to 12 years increases performance.

FIGURE 4 EWMA and GARCH conditional volatility estimates.



neither volatility estimate passes the test for any parameter setting: all the points are well above the critical value.

Both volatility estimators improve for the truncated distribution. However, neither of them are close to the critical value for any calibration. This suggests the following observation.

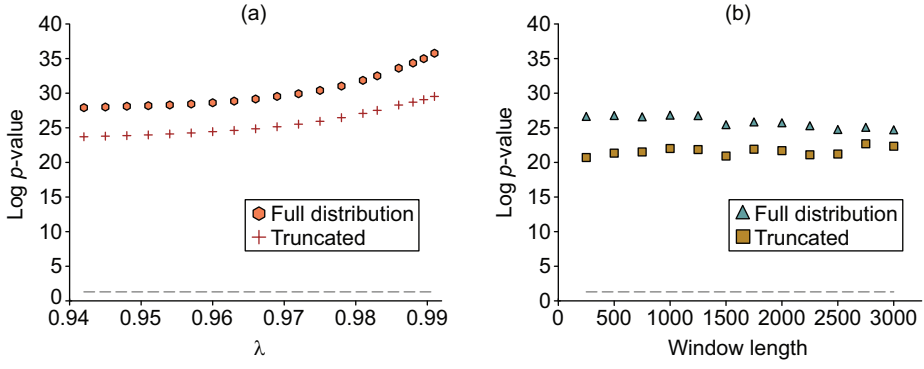
OBSERVATION 3.2 Parametric VaR using unweighted EWMA and GJR-GARCH volatility estimates with normal innovations does not provide accurate estimates of initial margin at 99% and 99.5%, nor does it provide good estimates of the far tails of the one-day forward return distribution.

3.4 FHS VaR

The results for FHS VaR are presented in Figure 6 as a function of the EWMA decay parameter, λ, with the window length, M, fixed at 300. The p-values are

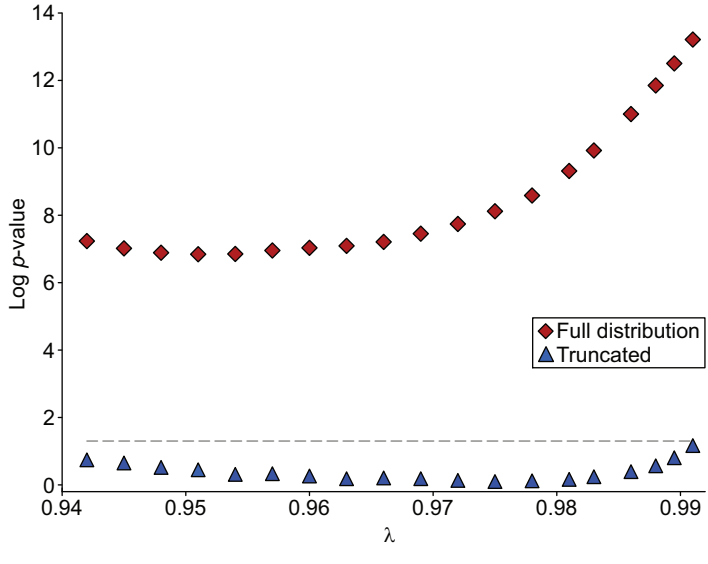
This might perhaps be seen in the light of the results of Daníelsson and Zhou (2015), Reghenzani *et al* (2019) and Stărică (2003), which showed that very long data series are needed to fit the far tails accurately in various distributional situations relevant to financial risk modeling. This of course means that there are rather few financial data series for which these fits are possible.

FIGURE 5 Test results for parametric VaR with conditional volatility estimators.



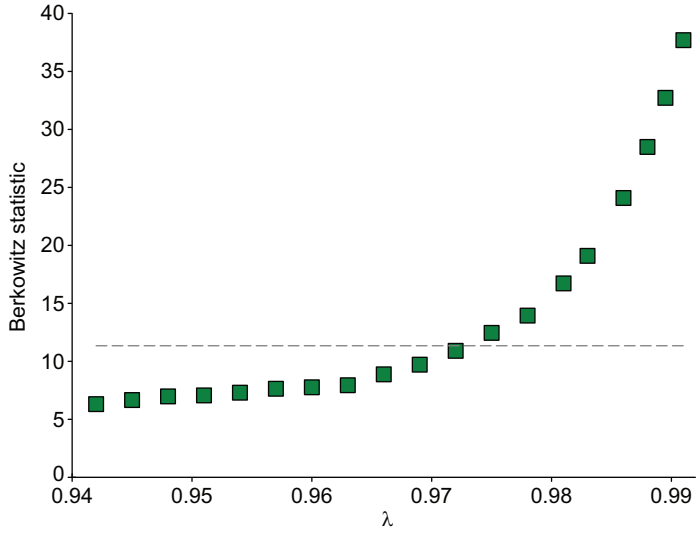
(a) EWMA volatility estimates. (b) GJR-GARCH volatility estimates.

FIGURE 6 Test results for FHS VaR using the Shapiro–Wilk test.



better than those for the unscaled distribution in both the whole and truncated cases. For the truncated case, a wide range of lambdas cannot be rejected, meaning that

FIGURE 7 Test results for FHS VaR using the Berkowitz test on truncated returns.



FHS is correctly estimating margin for these calibrations and is least based on the Shapiro–Wilk test.

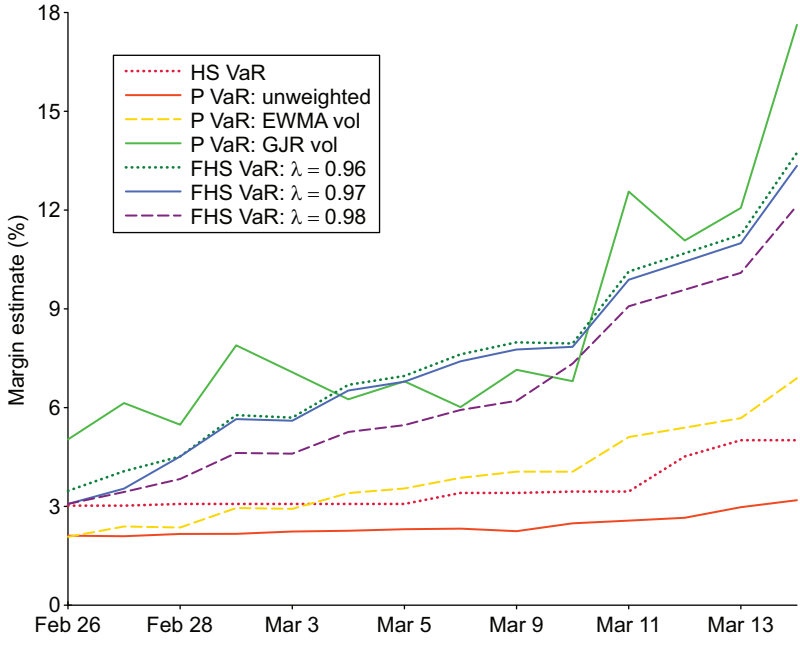
Of course, margin backtesting would be needed in addition to this. This process is relatively well understood, so we present another aspect of whole-distribution testing: the autocorrelation of returns. It would be expected not only that the time series of probabilities $\Pr(r < r_t \mid \phi_t^{\mathrm{FHS}})$ is uniform, but also that it does not display excess autocorrelation. An ARCH process can be fitted to the transformed probabilities, and the significance of the variation of its coefficients from Markovian normality can be tested.

This approach is known as the Berkowitz test.[16] It is less sensitive than the Shapiro–Wilk test to the shape of the distribution, so it should not be used alone, but it is a useful adjunct to our principal test due to its sensitivity to autocorrelation. Figure 7 presents the results, which imply that higher λ FHS VaR models fail due to excess autocorrelation, even though they match most of the distribution well. Together, these two tests suggest the following.

OBSERVATION 3.3 FHS VaR can, for some calibrations, produce accurate estimates of initial margin at 99% and 99.5%. It does not always provide good estimates

[16] See Berkowitz (2001), Berkowitz *et al* (2011) and Hamerle and Plank (2009) for a discussion of this test and its application.

FIGURE 8 Margin estimates at 99% for the S&P 500 index for a number of selected margin models from February 27, 2020 to March 17, 2020 (the period of Covid-19 stress).



of the far tails of the one-day forward return distribution, and hence does not estimate expected shortfall accurately.

4 INITIAL MARGIN ESTIMATES IN HIGH STRESS

In this section, the variability of margin estimates in high stress is examined. Following Gurrola-Perez (2021), the Covid-19-related market events of March 2020 are used as a natural test. Figure 8 presents the margin estimates from a number of the models discussed above for the last two days of February and the stressed period of early March, and Table 1 summarizes some key properties of the margins estimated.

These returns are among the largest absolute size ones in the whole series: -12.8% is the worst, and so lies at 99.98%, while -7.9%, which occurred on March 9, is the seventh worst, and hence lies at roughly 99.86%. The figures show that different models whose quantile estimates are acceptable for nearly all of the distribution produce quite different risk estimates for this period.

TABLE 1 Properties of S&P 500 returns and margin estimates at 99% for selected models from February 27 to March 17, 2020.

(a) Properties of S&P 500 log returns

	Max (%)	On date	Min (%)	On date	Max drawdown (%)
SPX log returns	8.9	Mar 13	−12.8	Mar 16	21.8

(b) Properties of margin estimates

Model	Max (%)	On date	Min (%)	On date	Max drawdown (%)
Historic simulation VaR, $M = 300$	3.0	Feb 27	7.9	Mar 17	4.9
Unweighted parametric VaR	2.1	Feb 27	3.6	Mar 17	1.5
EWMA parametric VaR, $\lambda = 0.97$	2.4	Feb 28	7.7	Mar 17	5.3
GJR parametric VaR	5.5	Feb 28	17.6	Mar 16	12.1
FHS VaR, $\lambda = 0.96$	4.1	Feb 27	15.3	Mar 17	11.2
FHS VaR, $\lambda = 0.97$	3.5	Feb 27	14.9	Mar 17	11.3
FHS VaR, $\lambda = 0.98$	3.4	Feb 27	13.5	Mar 17	10.0

Even if the parametric VaR estimates are discarded as unreliable based on the results above, the historic VaR estimates still suggest that parametric VaR with GJR volatility and FHS at $\lambda = 0.98$ and below may be overreacting. It cannot be determined with certainty from this evidence: all that is certain is that there are substantial differences between the conditional distribution estimates from different models over this period.[17] All models start the period with lower estimates and end it with higher ones, but the degree of reactiveness to the market turmoil varies considerably. This suggests the following.

OBSERVATION 4.1 In periods of high stress, margin is quite uncertain, even among models whose performance cannot be rejected on the basis of statistical tests. A margin model cannot be expected to unambiguously tell us the right level of margin in these periods; nor is there a correct degree of margin reactiveness to them.

Table 1 also presents the maximum drawdown associated with the margin model over the period (ie, the difference between the highest and lowest levels of margin).

[17] This is consistent with the results of Daníelsson *et al* (2016), which showed that, while the model risk of risk models "is typically quite moderate, it sharply increases during crisis periods".

Clearly, some models impose much more liquidity stress on margin posters than others.

5 POLICY DISCUSSION

The minimum quantile that margin models are required by European regulation to target is 99% for exchange-traded derivatives or 99.5% for OTC derivatives. Our analysis suggests that those are sensible choices, as it is certainly possible to produce a margin model that performs well at those quantiles. At least for many economically important portfolios, these models can be relatively simple, reducing the risk of overfitting and increasing transparency for margin posters and the wider market.[18]

Margin estimates at substantially higher confidence intervals and in periods of high stress are uncertain. It is difficult to be sure that a model is performing well in the far tails of the return distribution. This suggests that if regulators desire a greater overall level of safety than 99.5%, it should be achieved by means other than setting a higher confidence interval for a risk-based margin model. The current regulatory design does that, using default fund and bank capital as an additional protection above margin for cleared and bilateral portfolios, respectively. Our analysis validates the regulation approach to sizing CCP default funds using "extreme but plausible" stresses (see Committee on Payment and Settlement Systems and Technical Committee of the International Organization of Securities Commissions 2012, Principle 4; European Union 2012, Article 42) rather than targeting a far tail quantile of the portfolio return distribution.

The results shown suggest that tests of initial margin models should assess their performance for most of the portfolio return distribution; that is, out to, and a little past, the target quantile.[19] The far tails remain terra incognita.[20] Given that different, but equally acceptable, models perform differently under, and have different degrees of reactiveness to, high stress, other criteria can be used to select between them. One

[18] This point is important: increasing model complexity can often increase in-sample model performance. GJR-GARCH with Student t innovations, for instance, often performs better than the GJR with normal innovations studied here. However, the complexity of the fitting process is somewhat increased and it can be more unstable for some asset classes/product types, while working well for others. Also, the resulting margin may be less predictable and more opaque to less sophisticated market participants.

[19] Indeed, the approach proposed could be used to estimate how far out into the tails a margin model could be robustly used, at least for linear products.

[20] One interesting approach to this issue, discussed by Daníelsson and Zhou (2015), is to use a multiple of a lower confidence interval as a regulatory standard: if the 99th percentile can be reliably estimated but the 99.5th cannot, then perhaps setting margin at, say, $1.5\times$ the 99th percentile (with an add-on for far out of the money options) might be sensible. Figure 1 suggests that for this return series, quantiles out to about 2.75σ are relatively safe, but model risk rises significantly after that.

possibility would be to use measures of procyclicality, so that models that impose less liquidity stress in turbulent markets are preferred.

The test proposed in this paper based on most or all of the conditional return distribution is an aid to model calibration. While the job of a margin model is to predict a quantile, not the whole distribution of returns, its predictions of the whole distribution can be insightful. In particular, given the relatively low power of backtesting, if a model cannot be rejected based on its prediction at the margin confidence interval, but it is problematic based on the whole distribution, then the question arises as to whether sufficient information is available to see the problem at the margin quantile. Whole-distribution tests therefore have a role as indicators of problems that might appear later with margin calculations, while also providing useful insights into calibration and the issues with estimating return distributions in the far tails.

ACKNOWLEDGEMENTS

The author thanks Fernando Cerezetti, Melanie Houllier, Pedro Gurrola-Perez, Dmitrij Senko, the two anonymous referees and the participants in the World Federation of Exchanges' Clearing and Derivatives Conference 2022 for their comments on earlier versions of this paper.

DECLARATION OF INTEREST

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the paper.

REFERENCES

Barone-Adesi, G., and Giannopoulos, K. (2001). Non-parametric VaR techniques: myths and realities. *Economic Notes* **30**(2), 167–181 (https://doi.org/10.1111/j.0391-5026.2001.00052.x).

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets* **19**(5), 583–602. URL: https://doi.org/bn24q9.

Basel Committee on Banking Supervision (2013). Margin requirements for non-centrally cleared derivatives. Standards Document, September, Bank for International Settlements. URL: www.bis.org/publ/bcbs261.pdf.

Basel Committee on Banking Supervision, Committee on Payments and Market Infrastructures and Board of the International Organization of Securities Commissions (2021). Review of margining practices. Consultative Report, October, Bank for International Settlements. URL: www.bis.org/bcbs/publ/d526.htm.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* **19**(4), 465–474 (https://doi.org/10.1198/07350010152596718).

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science* **57**(12), 2213–2227 (https://doi.org/10.1287/mnsc.1080.0964).

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**(3), 307–327 (https://doi.org/10.1016/0304-4076(86)90063-1).

Buczyński, M., and Chlebus, M. (2020). Old-fashioned parametric models are still the best: a comparison of value-at-risk approaches in several volatility states. *The Journal of Risk Model Validation* **14**(2), 1–20 (https://doi.org/10.21314/jrmv.2020.222).

Campbell, S. (2005). A review of backtesting and backtesting procedures. Working Paper, April, Finance and Economics Discussion Series, Divisions of Research and Statistics and Monetary Affairs, Board of Governors of the Federal Reserve System, Washington, DC (https://doi.org/10.17016/feds.2005.21).

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review* **39**(4), 841–862 (https://doi.org/10.2307/2527341).

Committee on Payment and Settlement Systems and Technical Committee of the International Organization of Securities Commissions (2012). Principles for financial market infrastructures. Standards Document, April, Bank for International Settlements. URL: www.bis.org/cpmi/publ/d101a.pdf.

Daníelsson, J., and Zhou, C. (2015). Why risk is so hard to measure. Working Paper 494, De Nederlandsche Bank (https://doi.org/10.2139/ssrn.2597563).

Daníelsson, J., and Zigrand, J.-P. (2006). On time-scaling of risk and the square-root-of-time rule. *Journal of Banking and Finance* **30**(10), 2701–2713 (https://doi.org/10.1016/j.jbankfin.2005.10.002).

Daníelsson, J., James, K., Valenzuela, M., and Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability* **23**, 79–91 (https://doi.org/10.1016/j.jfs.2016.02.002).

Diebold, F., Gunther, T., and Tay, A. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**(4), 863–883 (https://doi.org/10.2307/2527342).

Escanciano, J., and Pei, P. (2012). Pitfalls in backtesting historical simulation VaR models. *Journal of Banking and Finance* **36**(8), 2233–2244 (https://doi.org/10.1016/j.jbankfin.2012.04.004).

European Union (2012). Regulation (EU) No. 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories. *Official Journal of the European Union* **55**(L201), 1–59. URL: https://data.europa.eu/eli/reg/2012/648/oj.

European Union (2013). Commission Delegated Regulation (EU) No. 153/2013 of 19 December 2012 supplementing Regulation (EU) No. 648/2012 of the European Parliament and of the Council with regard to regulatory technical standards on requirements for central counterparties. *Official Journal of the European Union* **56**(L52), 41–74. URL: https://data.europa.eu/eli/reg_del/2013/153/oj.

European Union (2015). Commission Delegated Regulations (EU) No. 2015/2205 of 6 August 2015 supplementing Regulation (EU) No. 648/2012 of the European Parliament and of the Council with regard to regulatory technical standards on the clearing obligation. *Official Journal of the European Union* **58**(L314), 13–21. URL: https://data.europa.eu/eli/reg_del/2015/2205/oj.

Glosten, L., Jagannathan, R., and Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* **48**(5), 1779–1801 (https://doi.org/10.1111/j.1540-6261.1993.tb05128.x).

Gurrola-Perez, P. (2018). The validation of filtered historical value-at-risk models. *The Journal of Risk Model Validation* **12**(1), 85–112 (https://doi.org/10.21314/jrmv.2018.185).

Gurrola-Perez, P. (2021). Procyclicality of central counterparty margin models: systemic problems need systemic approaches. *The Journal of Financial Market Infrastructures* **10**(1), 23–55 (https://doi.org/10.21314/jfmi.2022.002).

Gurrola-Perez, P., and Murphy, D. (2015). Filtered historical simulation value-at-risk models and their competitors. Working Paper 525, Bank of England, London (https://doi.org/10.2139/ssrn.2574769).

Haas, M. (2001). New methods in backtesting. Unpublished Note, February, Financial Engineering, Research Center Caesar, Bonn. URL: www.ime.usp.br/∼rvicente/risco/haas.pdf.

Hamerle, A., and Plank, K. (2009). A note on the Berkowitz test with discrete distributions. *The Journal of Risk Model Validation* **3**(2), 3–10 (https://doi.org/10.21314/jrmv.2009.038).

Hansen, P., and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? *Journal of Applied Econometrics* **20**(7), 873–889 (https://doi.org/10.1002/jae.800).

Hill, C., and McCullough, B. (2019). On the accuracy of GARCH estimation in R packages. *Econometric Research in Finance* **4**(2), 133–156. URL: www.erfin.org/journal/index.php/erfin/article/view/64/40.

Houllier, M., and Murphy, D. (2017). Initial margin model sensitivity analysis and volatility estimation. *The Journal of Financial Market Infrastructures* **5**(4), 77–103 (https://doi.org/10.21314/JFMI.2017.078).

Hull, J., and White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *The Journal of Risk* **1**(1), 5–19 (https://doi.org/10.21314/JOR.1998.001).

International Swaps and Derivatives Association (2021). ISDA SIMM methodology, version 2.4. Technical Document, December, ISDA. URL: www.isda.org/a/CeggE/ISDA-SIMM-v2.4-PUBLIC.pdf.

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd edn. McGraw-Hill.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* **3**(2), 73–84 (https://doi.org/10.3905/jod.1995.407942).

Lopez, J. (1998). Methods for evaluating value-at-risk estimates. *Federal Reserve Bank of San Francisco Economic Policy Review* **4**(3), 3–17. URL: www.frbsf.org/economic-research/wp-content/uploads/sites/4/3-17.pdf.

Maruyama, A., and Cerezetti, F. (2019). Central counterparty anti-procyclicality tools: a closer assessment. *The Journal of Financial Market Infrastructures* **7**(4), 1–25 (https://doi.org/10.21314/JFMI.2018.110).

Murphy, D., and Vause, N. (2022). A cost–benefit analysis of anti-procyclicality: analyzing approaches to procyclicality reduction in central counterparty initial margin models.

The Journal of Financial Market Infrastructures **9**(4), 27–50 (https://doi.org/10.21314/jfmi.2021.013).

Murphy, D., Vasios, M., and Vause, N. (2014). An investigation into the procyclicality of risk-based initial margin models. Working Paper 525, Bank of England, London (https://doi.org/10.2139/ssrn.2437916).

Razali, N., and Yap, B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics* **2**(1), 21–33.

Reghenzani, F., Massari, G., Santinelli, L., and Fornaciari, W. (2019). Statistical power estimation dataset for external validation GoF tests on EVT distribution. *Data in Brief* **25**, Paper 104071 (https://doi.org/10.1016/j.dib.2019.104071).

Shapiro, S., and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**(3–4), 591–611 (https://doi.org/10.2307/2333709).

Stărică, C. (2003). Is GARCH$(1, 1)$ as good a model as the accolades of the Nobel Prize would imply? Working Paper, Social Science Research Network (https://doi.org/10.2139/ssrn.637322).