Journal of Financial Market Infrastructures **5**(4), 77–103
DOI: 10.21314/JFMI.2017.078

Research Paper

Initial margin model sensitivity analysis and volatility estimation

Melanie Houllier[1] **and David Murphy**[2]

[1]The London Institute for Banking and Finance, 25 Lovat Lane,
London EC3R 8EB, UK; email: Mhoullier@libf.ac.uk

[2]Bank of England, Threadneedle Street, London EC2R 8AH, UK;
email: david.murphy@bankofengland.co.uk

(Received February 9, 2017; revised May 22, 2017; accepted May 23, 2017)

ABSTRACT

The advent of mandatory central clearing for certain types of over-the-counter derivatives, and margin requirements for others, means that margin is now the most important mitigation mechanism for many counterparty credit risks. Initial margin requirements are typically calculated using risk-based margin models, and these models must be tested to ensure they are prudent. However, two different margin models can calculate substantially different levels of margin and still pass the usual tests. This paper presents a new approach to parameter selection based on the statistical properties of the worst loss over a margin period of risk estimated by the margin model under scrutiny. This measure is related to risk estimated at a fixed confidence interval, but it leads to a more powerful test that is better able to justify the choice of parameters used in margin models. The test proposed is used on a variety of volatility estimation techniques applied to a long history of returns of the Standard & Poor's 500 index. Some well-known techniques, including exponentially weighted moving average volatility estimation and generalized autoregressive conditional heteroscedasticity, are considered, and novel approaches derived from signal processing are also analyzed. In

Corresponding author: D. Murphy

Print ISSN 2049-5404 | Online ISSN 2049-5412

each case, a range of model parameters that gives rise to acceptable risk estimates is identified.

Keywords: conditional volatility; filtered volatility; GARCH(1, 1); initial margin (IM) model; model backtesting; volatility estimation.

1 INTRODUCTION

The collection of margin to reduce counterparty credit risk is a key feature of postcrisis financial market reforms. This is because policy makers want to reduce the direct interconnectedness between financial institutions. Thus,

- the clearing of standardized over-the-counter (OTC) derivatives between large market participants, with the associated clearing house margin requirements, has been mandated by the G20 (2009), and
- margin requirements are being introduced for bilateral OTC derivatives between many market participants in a revision of the Basel Accords (Basel Committee on Banking Supervision 2013).

All of this means that some financial institutions hold tens of billions of pounds of initial margin (IM) against their exposures to their derivatives counterparties.[1]

The question of how IM is calculated is, therefore, commercially important. Often, a risk-based IM model is employed: this typically uses some representation of the portfolio's risk (some set of risk factors), as well as some information about how those risks have behaved in the past, to determine margin. For instance, a value-at-risk-based (VaR-based) margin model may determine an IM requirement based on the 99th percentile of the estimated loss distribution of the portfolio in question over an assumed liquidation horizon. The model thus targets a confidence interval, determining margin based on portfolio value changes that are as or less probable than this threshold.

1.1 Margin model design

The design of an IM model entails making a number of decisions. These can be summarized as follows.

- How should the risk of a portfolio be represented?
- What history of risk factors is to be selected, and how is it to be used?

[1] For instance, LCH.Clearnet Group Limited's 2015 consolidated financial statement states that the total margin liability of members at December 31, 2015 was €110 billion.

- What algorithm should be used to determine the portfolio return distribution?[2]
- How are the parameters of that algorithm to be selected?

Thus, the design of a simple margin model for portfolios of equity index futures may be based on the following decisions:

- the risk factors are positions in index futures of various maturities;
- a ten-year history of the prices of these futures will be used to calculate returns;
- a historical simulation (HS) VaR model will be used; and
- a ten-year window and a 99% confidence interval will be used.

The IM model is both the algorithm (HS VaR) and its calibration (ten-year window, 99% confidence interval).

Given the sums at stake as well as the potential consequences for the margin taker if there is too little margin, it is vital that robust margin models are used. One obvious requirement is that, if a model purports to calculate margin to some degree of confidence, it actually does so. Thus, backtesting of the calculated margin for a portfolio against the losses it would have experienced over some history of market movements is an important element of model validation, as, for instance, Berkowitz*et al* (2011) and Campbell (2007) discuss. However, as shown in Gurrola Perez (2015), the same algorithm with widely different parameters and giving different margin requirements can pass backtesting. Clearly, backtesting alone often does not provide enough discriminating power to justify all of the features of an IM model, or to fix its parameters within narrow ranges. Figure 1 illustrates this phenomenon, showing the amount of margin that two different filtered historical simulation (FHS) VaR models would demand for a position in the Standard & Poor's 500 (S&P 500) index. Both models pass the standard Kupiec (1995) test, yet the amount of margin they demand on the same date can differ substantially. Should the model that has, on average, higher margin requirements but lower margin variability (and, hence, lower liquidity burdens on margin posters) be preferred to the more reactive model?

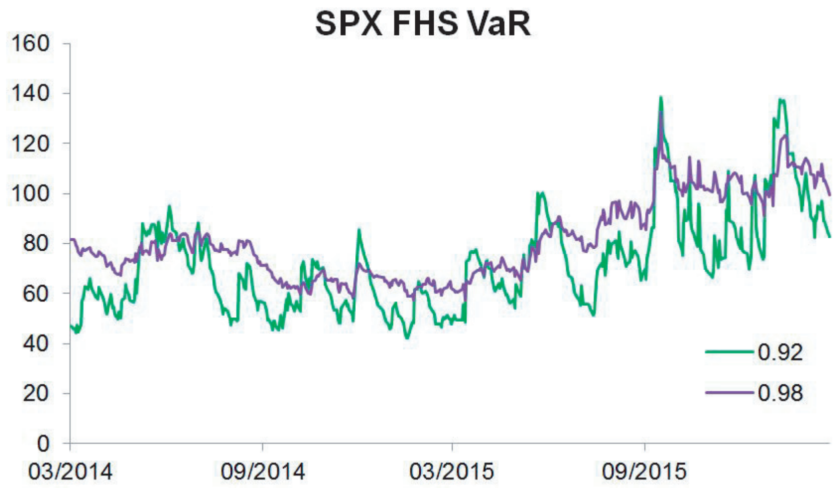
1.2 Sensitivity analysis

The importance of selecting good model parameters is recognized in derivatives policy. For instance, European regulation requires that central counterparties (CCPs) carry out sensitivity analysis to test model parameters:[3]

[2] There are sometimes two parts to this question: how is the return distribution for a single risk factor calculated, and how are these combined when more than one risk factor is relevant? The latter is the portfolio margining question.

[3] This quote is from the EMIR regulatory technical standards (European Union 2012b).

FIGURE 1 FHS differences example.



The VaR calculated by two different filtered historical simulation (FHS) models (one more variable with a decay constant of 0.92, the other less so with a decay constant of 0.98) for a position in the S&P 500 index.

> Sensitivity analysis shall be performed on a number of actual and representative clearing member portfolios. The representative portfolios shall be chosen based on their sensitivity to the material risk factors and correlations to which the CCP is exposed. Such sensitivity testing and analysis shall be designed to test the key parameters and assumptions of the IM model at a number of confidence intervals to determine the sensitivity of the system to errors in the calibration of such parameters and assumptions.

The work presented here is a contribution to the study of IM model design and sensitivity analysis. A new technique for comparing models is proposed based on the worst loss that could be experienced when liquidating a defaulter's portfolio over some margin period of risk (MPOR). This method is used to compare various algorithms used in IM modeling, and to select parameters for them that can be justified statistically. Our technique provides a new and discriminating way of selecting acceptable algorithms and parameters than conventional backtesting approaches. It maps a borderline that encompasses a narrower selection of good models than most techniques currently in use.

The rest of this paper is structured as follows. In the remainder of this section, related work is summarized. Section 2 discusses the model design and sensitivity testing problem. Section 3 shows how the statistical properties of empirical losses can be used for sensitivity testing, and Section 4 applies this approach to some volatility estimation algorithms. We consider popular techniques, such as exponentially weighted moving average (EWMA) volatility estimation, and less common approaches, such as half-kernel estimators. Section 5 sets out some nuances of and extensions to the sensitivity analysis technique proposed, and Section 6 concludes.

1.3 Related work

The need for good algorithm and parameter selection in risk modeling has been evident since such models were first used by banks. Kupiec's early and influential paper (Kupiec 1995) recognized the difficulties involved. It states:

> It does not appear possible for a bank or its supervisor to reliably verify the accuracy of an institution's internal model loss exposure estimates using standard statistical techniques.

Two broad schools of model-testing approaches have appeared in response to this observation. Initially, the focus was on occasions when losses in excess of the risk predicted by the model at a fixed confidence were observed. The frequency of these exceptions forms the basis of the test proposed by Kupiec, while later tests also use information on the time between exceptions or their size. The tests proposed by Christoffersen (1998), the "mixed Kupiec" test of Haas (2001) and Pelletier and Wei's "Geometric VaR" (Pelletier and Wei 2016) are examples of the former, while Lopez (2001) uses the latter.

Subsequent work proposed tests that focussed on the extent to which a model's forecast return distribution could be said to be accurate, given the empirical quantiles observed. Crnkovic and Drachman (1997) and Diebold *et al* (1998) both suggested approaches of this type, and Berkowitz (2001) extended this idea to provide a pass/fail test.

More recently, the understanding of risk model testing has been informed by a closer study of data issues and small sample biases (see, for example, Daníelsson and Zhou 2015; Escanciano and Pei 2012).

2 THE MODEL-TESTING PROBLEM

The target confidence interval for IM models is typically high: for instance, European regulation sets a minimum level of 99% for cleared exchange traded derivatives and 99.5% for cleared OTC derivatives. Margin should be adequate to cover all but the most unlikely moves in portfolio value, making the margin taker fairly safe from counterparty credit risk. However, a high confidence interval can lead to problems with backtesting, as 1-in-100 or 1-in-200 events, by definition, do not happen often. Therefore, in some naive sense, in standard backtesting the model's performance on most days does not matter, and this means that the discriminating power of standard backtesting is often poor.[4]

Some of the approaches to backtesting discussed in Section 1.3 ameliorate this problem. However, there is another alternative: a risk-based margin model must predict (at least some properties of) the portfolio return distribution each day, so these predictions of a quantity closely related to the risk estimate can be used to evaluate the model. Therefore, in this paper, the accuracy of a risk model's prediction of (conditional) volatility will be the primary test object (see Houllier and Murphy (2017) for a treatment of a wider range of risk estimation techniques, including some that do not utilize explicit volatility estimates such as HS VaR).

2.1 Using volatility estimates directly

A simple approach to formalizing a test based on volatility estimates is to compare the realized squared return each day with the volatility estimated on the prior day. In particular, suppose we have some series of returns (either of single risk factors or portfolios) $\{r_t\}$. Then, given a prediction of volatility at time t, $\tilde{\sigma}_t$, it should be the case that the observed squared return r_t^2 is proportional to $\tilde{\sigma}_t^2$; so, a simple test of the accuracy of the volatility estimator is to look at the coefficient of determination of this relationship. However, this R^2 will typically be low; this is simply because, even if the expectation of r_t^2 is $\tilde{\sigma}_t^2$, there is a lot of variation around this average, so the approach will not be very discriminating.[5]

2.2 The margin period of risk

Another complication arises because it is not just one-day returns that are typically of interest. Instead, margin is usually calculated over some MPOR that is longer than a day: for instance, the regulatory minimum MPOR for exchange traded derivatives in the house account is two days, for cleared OTC derivatives it is five days and for bilateral OTCs in Europe it is ten days. Thus, ten-day returns are potentially of

[4] The fact that backtests can be performed on many different portfolios helps here: portfolios that depend sensitively on a particular part of the return distribution, or on particular properties of the returns (such as their autocorrelation), should be selected for testing, among others. However, the behavior of financial times series is so varied that identifying all of the relevant behaviors is difficult, and many portfolios will be required; this comes with the associated problem that, as the number of portfolios backtested rises, false positives become more likely.

[5] Using 1000 days of S&P 500 data, the best R^2 obtained for an EWMA volatility estimate using a range of decay constants from 0.94 to 0.99 was 0.15. This just shows that the squared return series is too noisy to be used directly to select the parameter(s) of a volatility estimation technique. This is an observation that goes back at least to Andersen and Bollerslev (1998), and which, as Poon and Granger (2003) point out, directly relates to the difficulty of estimating conditional kurtosis for fat-tailed distributions.

interest, but there are ten times less of them, compounding the problems with test discrimination for one-day returns.[6]

In the next section, a sensitivity analysis technique based on volatility estimates that does not suffer from the two problems identified above will be presented.

3 USING THE WORST LOSS OVER THE MARGIN PERIOD OF RISK

Suppose we have a time series of observations of the value of some portfolio $x_0, x_1, \dots, x_n$ at the close of each day. For an m-day MPOR, the worst loss assuming default at t, $\mathrm{WL}_m(t)$, is defined by

$$\mathrm{WL}_m(t) = x_t - \min\nolimits_{0 \leqslant u \leqslant m} x_{t+u}.$$

This can be thought of as the worst loss incurred in liquidating the portfolio, given the last successful margin call was based on prices at t, and a subsequent close-out after a default occurred at the close on the worst of the m days in the MPOR after t. As such, predicting the worst loss over an appropriate MPOR is the key task of an IM model; hence, the accuracy of this prediction is a natural thing to test.

3.1 The worst loss for conditionally lognormal distributions

It is common in risk factor modeling to assume conditional lognormality.[7] That is, the incremental change in a risk factor x from time t is assumed to be driven by a Brownian motion with a volatility σ_t, which varies only slowly with t and, hence, can be assumed constant over the MPOR. Under this assumption, the dependence of $\mathrm{WL}_m(t)$ on σ_t has been investigated in the literature. The standard source is Aitsahlia and Lai (1998), who consider the related problem of pricing discrete lookback options. From their results, it is straightforward to derive the probability density of $\mathrm{WL}_m(t)$ over an m-day period for a driftless process. That is,[8]

$$\Pr(\mathrm{WL}_m < k \mid \sigma = \tilde{\sigma}_t\} = \sum_{\nu=1}^{m} \alpha_{m-\nu}(\sigma_t) \int_0^k f_\nu(x, \sigma_t)\,\mathrm{d}x, \tag{3.1}$$

[6] Nonoverlapping returns are preferred to overlapping ones in the analysis of multiday MPORs due to the difficulty of interpreting an exception that comes from a single day but manifests in a number of overlapping MPOR-length returns if overlaps are used.

[7] For one-day returns on some asset classes, even conditional lognormality does not capture all of the features of the observed kurtosis (see, for example, Liesenfeld and Jung 2000). However, as the MPOR lengthens, returns become more conditionally lognormal. Moreover, many margin models use conditional lognormality to some extent. Hence, the assumption is not wholly unreasonable.

[8] For further details of this derivation, see Houllier and Murphy (2017).

where α and f are defined recursively by the convolutions

$$\begin{aligned}
\alpha_0 &= 1,\\
\alpha_k &= \int_{-\infty}^{0} g_k(x,\sigma_t)\,\mathrm{d}x, \quad k \geqslant 1,\\
f_1(x,\sigma_t) &= \psi(x,\sigma_t),\\
f_n(x,\sigma_t) &= \int_{0}^{\infty} f_{n-1}(y)\psi(x-y,\sigma_t)\,\mathrm{d}y, \quad 2 \leqslant n \leqslant m;
\end{aligned}$$

the auxiliary functions g and ψ are defined by

$$\begin{aligned}
g_1(x,\sigma_t) &= \psi(x,\sigma_t),\\
g_n(x,\sigma_t) &= \int_{-\infty}^{0} g_{n-1}(y,\sigma_t)\psi(x-y,\sigma_t)\,\mathrm{d}y, \quad 2 \leqslant n \leqslant m,\\
\psi(x,\sigma_t) &= \phi\biggl(0, \sqrt{m}\sigma_t, x + \frac{m\sigma_t^2}{2}\biggr),
\end{aligned}$$

and $\phi(\bar{x}, \mathrm{SD}, x)$ is the standard normal density function.

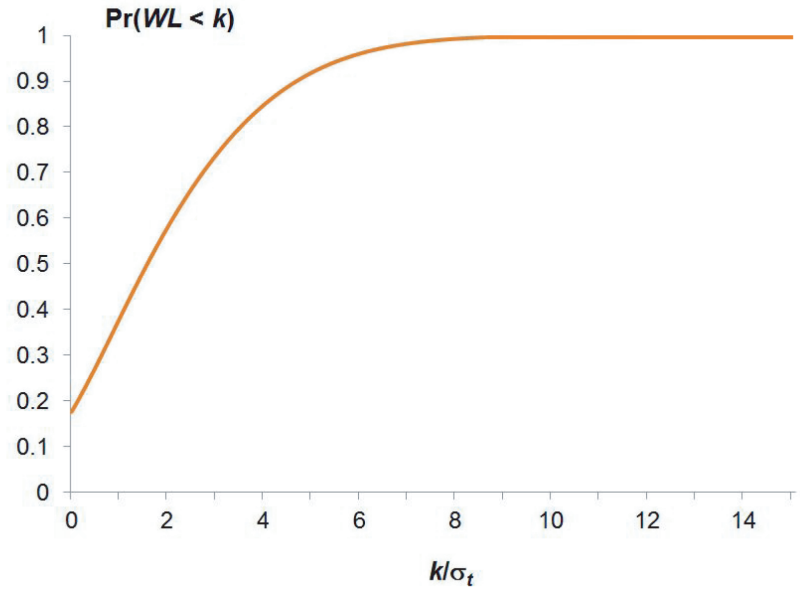
It is easiest to illustrate this if we express the worst loss as a multiple of the daily conditional volatility. The MPOR is fixed at ten days, and the subscript m is dropped going forward. The cumulative probability of a given worst loss as a fraction of the conditional volatility WL/σ_t is illustrated in Figure 2. The level of conditional volatility σ_t has a minor impact on the shape of this cumulative distribution. The illustration is for a daily volatility of 1%.

3.2 Using the probability of worst losses to compare volatility estimates

Suppose we observe, on consecutive days, events that we estimate as five, three, four, six and two standard deviation occurrences. We may reasonably conclude that either unusual things are happening a lot or our estimate of the standard deviation is wrong. If the pattern of these five days is repeated in many subsequent periods, the latter conclusion becomes more and more likely, assuming that events on consecutive days are independent. This insight is key to using the cumulative conditional probabilities (3.1) to test the quality of the prediction $\tilde{\sigma}_t$.

The data needed here is illustrated in Table 1. First, a time series of risk factors is used to generate worst losses in each nonoverlapping MPOR. Second, a model estimates the conditional volatility for the relevant period (based on data up to the start of it). Finally, the cumulative probability of each worst loss conditional on that volatility is calculated. That is, a (likely different) process is estimated for each window, and this process gives a conditional volatility estimate that is used for an MPOR-length

FIGURE 2 Cumulative probability.



The cumulative probability of seeing a given worst loss for a ten-day MPOR, measured as a multiple of conditional volatility. A worst loss of zero happens about 17% of the time, and the 99th percentile worst loss is approximately 7.4 times the conditional volatility.

period following the window. It is convenient to index these results by the time the window ends (t), so $\tilde{\sigma}_t$ is the conditional volatility estimate using the window stretching back from t, and $\mathrm{WL}(t)$ is the worst loss for the MPOR $t+1,\dots,t+\mathrm{MPOR}$. Given a sequence of these worst losses $\mathrm{WL}(t), \mathrm{WL}(t'), \dots$ and conditional volatility estimates $\tilde{\sigma}_t, \tilde{\sigma}_{t'}, \dots$, the collection of all the cumulative probabilities should be uniformly distributed if the volatility estimates are unbiased.

A simple way to test this is to choose a binning scheme for the cumulative probabilities $[0, b_1), [b_1, b_2), \dots, [b_n, 1]$ and use a χ^2 test versus the expected uniform distribution.[9]

Figure 3 illustrates the empirical distribution of cumulative probabilities from the data in Table 1 and the expected uniform distribution. A big first bucket is used to allow for the fact that $\mathrm{WL} = 0$ occurs with expected probability $\approx 17\%$. It can be seen that for the volatility estimation technique used (in this case, an EWMA estimate

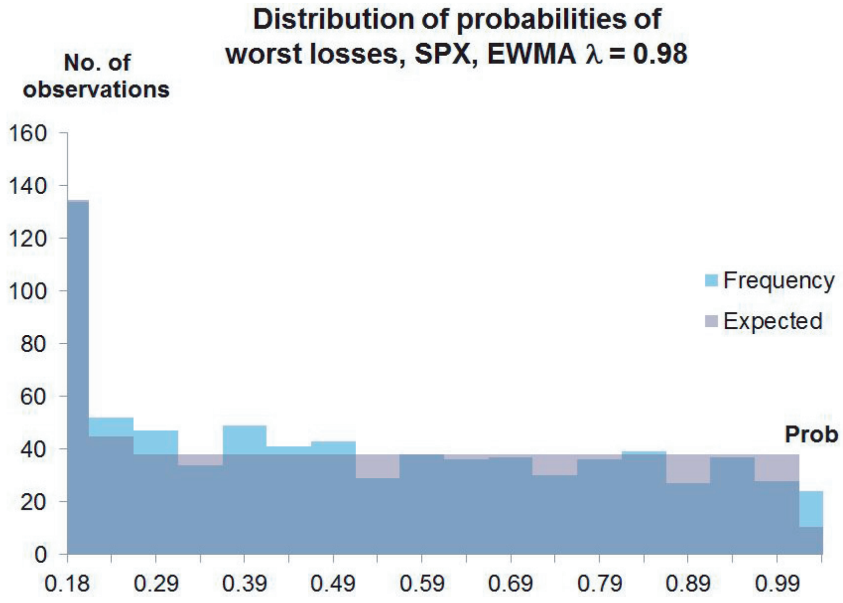
[9] For the standard χ^2 test, the bin widths b_i should be chosen so that the expected population of each bin is at least five; otherwise, slight modifications, such as the use of Fisher's exact test, are needed. See also Section 5.2 for a discussion of a more sophisticated approach.

TABLE 1 Cumulative probabilities example.

Date of estimate	Volatility estimate	Worst loss in subsequent MPOR	Cumulative probability
01/13/1986	0.98%	3.81	0.648
01/20/1986	0.94%	0	0.176
⋮			
03/07/2016	1.28%	22.8	0.380

The worst loss for nonoverlapping ten-day MPORs and the conditional cumulative probabilities of observing them, given a particular estimate of conditional volatility.

FIGURE 3 Cumulative probabilities example.



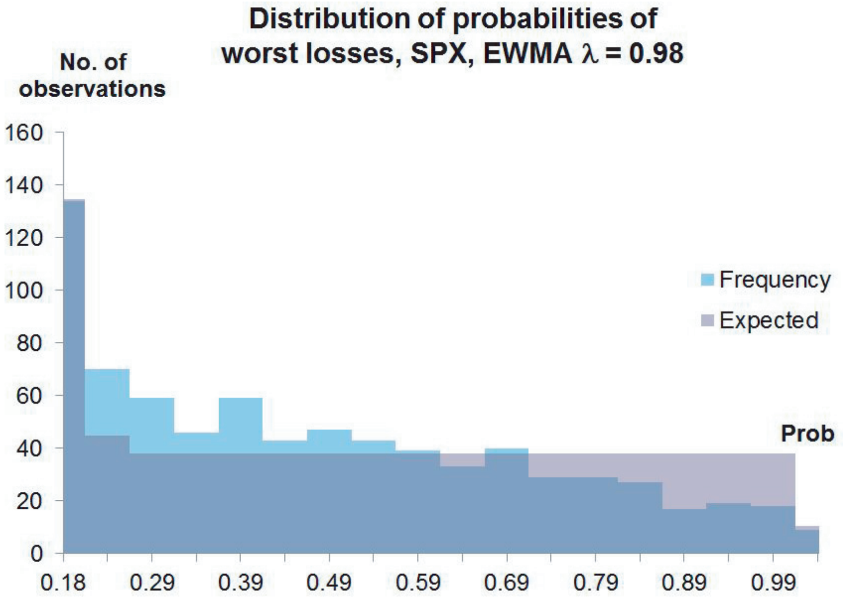
The distribution of conditional cumulative probabilities for ten-day worst losses given a particular method for estimating conditional volatility.

with a decay constant of 0.98), the empirical distribution of cumulative probabilities is close to the expected one.

In contrast, Figures 4 and 5 show the distributions for volatility estimates that have been scaled up and down by 30%.

Clearly, if volatility is overestimated, an excess of low cumulative probability events are observed, while if it is underestimated, an excess of high cumulative probability

FIGURE 4 Cumulative probabilities example with increased volatility estimates.



The distribution of conditional cumulative probabilities of ten-day worst losses when volatility estimates are scaled up by 30%.

events are seen. Hence, the departure of the empirical distribution from the uniformity expected can be used to test the performance of the volatility estimation technique being considered.

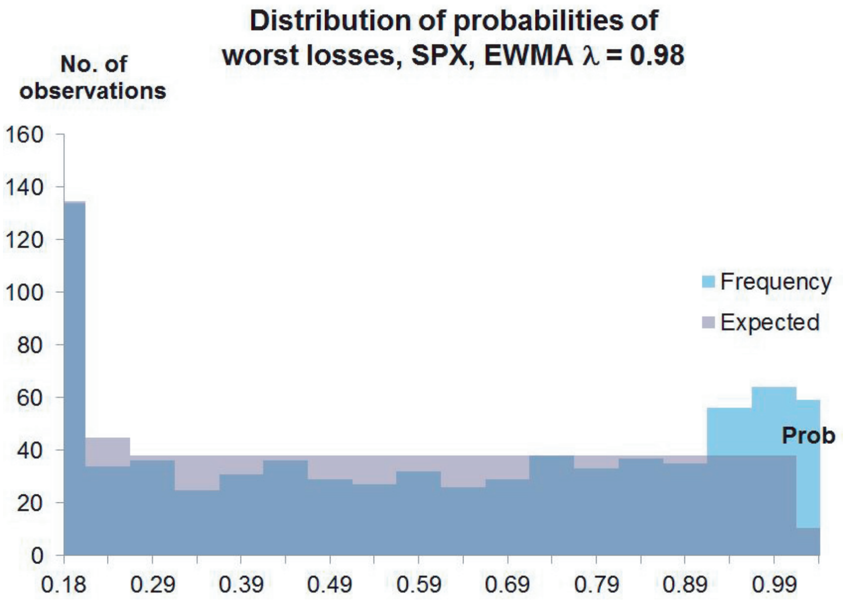
3.3 Example: ARCH volatility

The proposed approach works even when conditional volatility varies significantly over time. To see this, consider an autoregressive conditional heteroscedasticity, ARCH(1), process: in this well-known Engle (1982) model, the assumption is that detrended daily returns of a risk factor x still follow a discretized version of $\mathrm{d}x = \sigma_t x \,\mathrm{d}Z$, but the current volatility σ_t depends on the previous return according to

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2. \tag{3.2}$$

The returns in this model demonstrate autocorrelation, but this is entirely driven by the volatility process: Z is still white noise. Thus, if the ARCH parameters ω and α are correctly estimated, the probabilities of seeing two worst losses $\mathrm{WL}(t)$, $\mathrm{WL}(t')$ at two different dates t, t', conditional on σ_t and $\sigma_{t'}$, respectively, will be independent

FIGURE 5 Cumulative probabilities example with decreased volatility estimates.



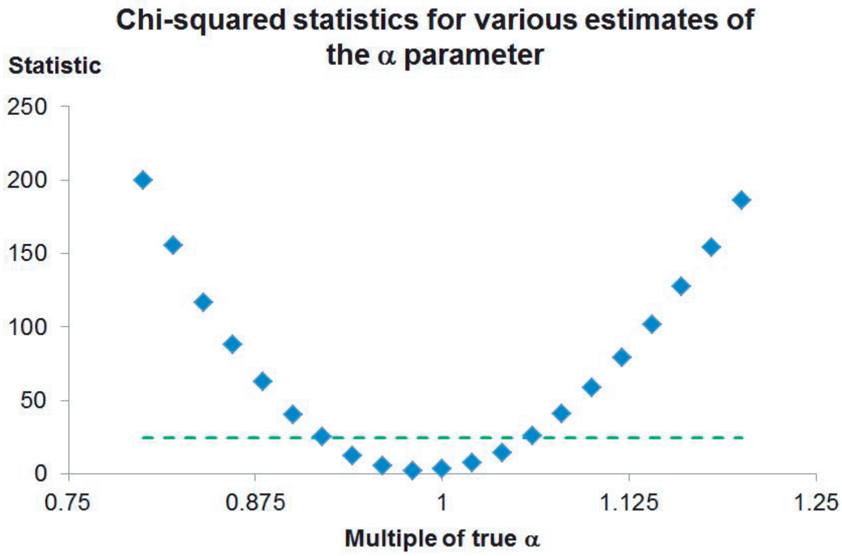
The distribution of conditional cumulative probabilities of ten-day worst losses when volatility estimates are scaled down by 30%.

of each other, and the theory above will apply. Therefore, the worst-loss test can be applied to volatility estimates backed out from (3.2). Figure 6 illustrates this: here, an ARCH process has been simulated with $\omega = 10^{-5}$ and $\alpha = 0.9$, and ARCH volatility estimates with a range of α around the true value are tested. Higher values of the test statistic are worse, and models above the dotted line are rejected at 99%. Of course, the test proposed is not the optimal way to calibrate an ARCH model, but the results presented do at least demonstrate that it can pick out model parameters accurately, even in the presence of volatility clustering (and thus fat tails).

4 TESTING VOLATILITY ESTIMATION TECHNIQUES USING WORST LOSS

This section presents some of the results obtained using worst-loss tests on various techniques employed in margin modeling. In all cases, we use returns from the S&P 500 index from January 3, 1984 to March 24, 2016 and calculate margin over a ten-day MPOR. This period gives a total of 761 nonoverlapping ten-day periods.

FIGURE 6 Testing ARCH.



The χ^2 statistic for a range of hypothesized α parameters around the true value for an ARCH process. The dashed green line shows the critical value of the statistic, so models above the line are 99% likely to be wrong.

4.1 EWMA volatility estimates

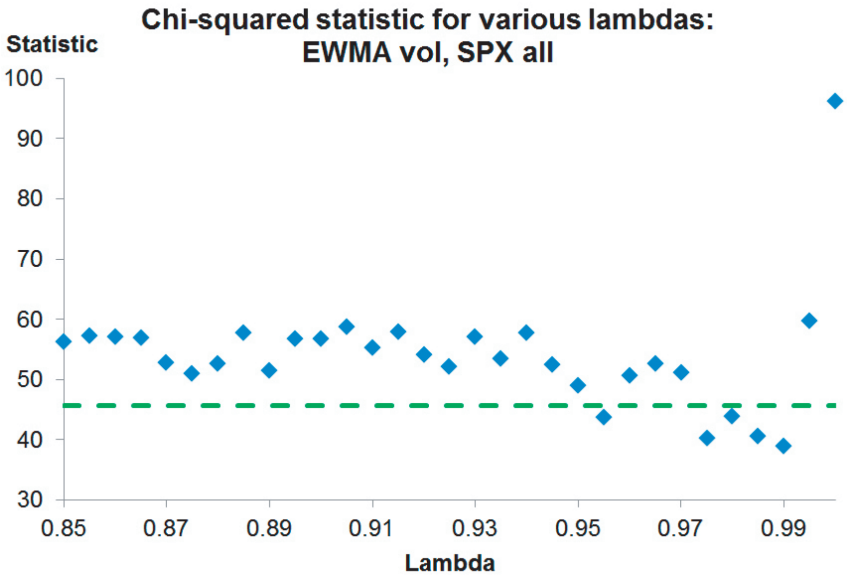
EWMA volatility estimators are widely used in risk models (see Gijbels *et al* (1999) for a perspective on their use). In EWMA, the current volatility estimate $\tilde{\sigma}_t$ is updated based on the previous estimate, $\tilde{\sigma}_{t-1}$, and the previous return, r_{t-1}, using the recurrence

$$\tilde{\sigma}_t^2 = \lambda\tilde{\sigma}_{t-1}^2 + (1-\lambda)r_{t-1}^2.$$

The key issue in designing volatility estimation models of this class is selecting an appropriate decay constant, λ. If λ is too low, the volatility estimate overreacts to changes in conditions. If, however, it is too high, it reacts too slowly. Figure 7 show the results from a worst-loss-based sensitivity analysis of EWMA models with various λ.

These results suggest the hypothesis that “the S&P 500 is well-described by a locally-constant volatility which can be estimated using an EWMA technique” is only true for EWMA models with a narrow range of decay constants around 0.98; models with other λ fail the test, including (at least for the window length studied) the unweighted volatility estimator.

FIGURE 7 Testing EWMA (part 1).



The χ^2 statistic for EWMA volatility estimates on S&P 500 index returns as a function of the decay constant λ.

4.2 Blended volatility estimates

EWMA models with decay constants close to 1 react relatively slowly to new data. Another way to produce a volatility estimate that reacts slowly is to keep a reactive decay constant but "blend in" some long-term average volatility (such as a ten-year unweighted volatility estimate) σ_{LT} by estimating variance at t as

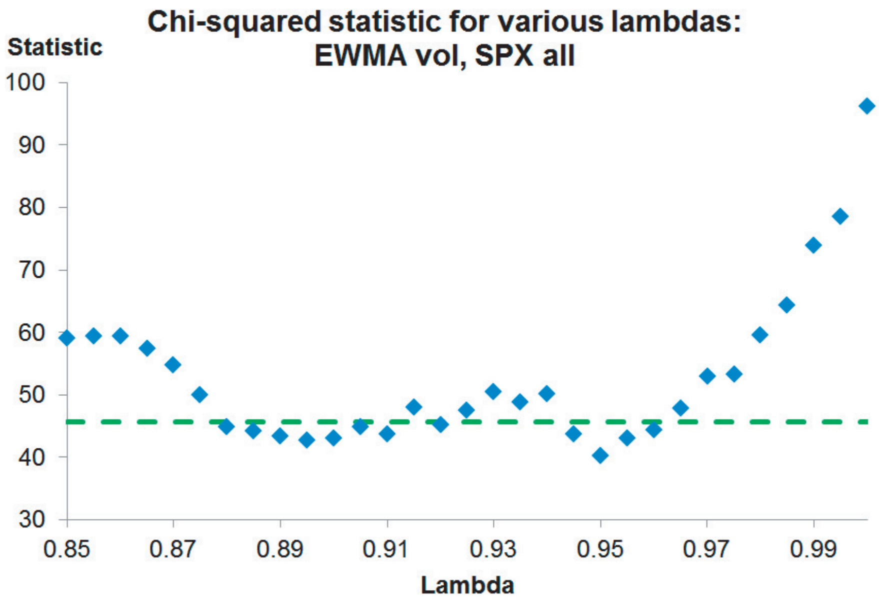
$$\tfrac{1}{2}(\tilde{\sigma}_t^2 + \sigma_{\text{LT}}^2),$$

where $\tilde{\sigma}_t$ is an EWMA volatility estimate.

Figure 8 shows the sensitivity analysis for the blended volatility estimates technique.[10] The aim of blending in a long-term average volatility is clearly to tame the reactiveness of the smaller decay constants: the resulting volatility estimates are acceptable, or close to acceptable, for a wide range of λ.

[10] There is no theoretical justification for this volatility estimator, but it is of practical interest. There is a connection here with the requirement in the EMIR regulatory technical standards that "the data used for calculating historical volatility capture[s] a full range of market conditions, including periods of stress" (European Union 2012b). This requirement is sometimes met by "blending in" a stressed volatility with a current volatility estimate.

FIGURE 8 Testing EWMA (part 2).



The χ^2 statistic for a blend of an EWMA volatility estimate and long-term average volatility.

4.3 The normal half-kernel estimator

Exponential smoothing is a common model-free means of forecasting a future realization of a time series, but it is not the only one. As Stărică (2003) points out, the general question is to select some (likely declining) weight function w_i such that the w-weighted variance estimate

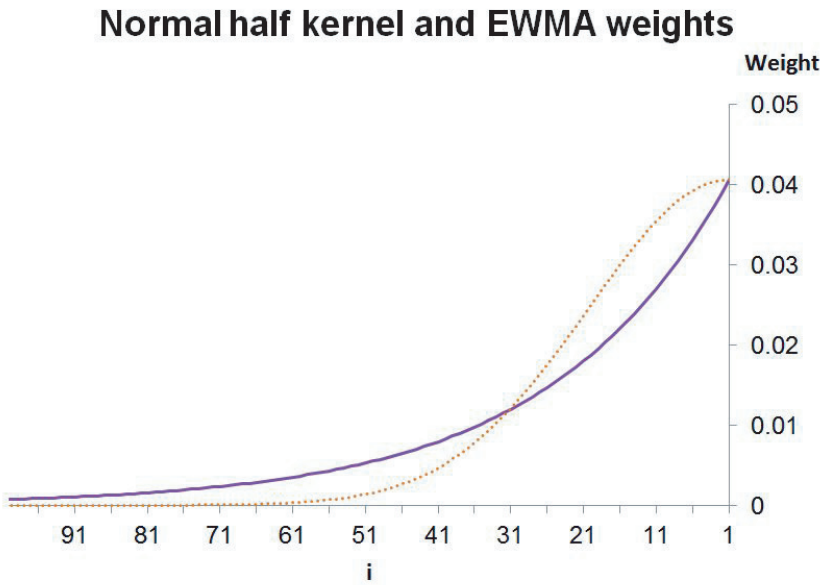
$$\frac{\sum_{i=1}^{m} w_i r_{t-i}^2}{\sum_{i=1}^{m} w_i}$$

is optimal. Clearly EWMA with decay constant λ is an example of this approach, with $w_i = (1-\lambda)^i$. The function w is known as a half kernel.

The literature suggests that the normal half kernel (where the weights are defined by negative half of the normal probability density function (PDF)) is often of interest (see Wand and Jones (1995) for a discussion of this and other half kernels, and Figure 9 for a comparison of the normal half-kernel weighting function with that used in the EWMA approach).

The probability of worst loss approach was used to test a number of normal half-kernel volatility estimators with different widths. Figure 10 shows the performance

FIGURE 9 Weight functions.



The weight functions for a normal half kernel (orange dotted line) and an EWMA approach (blue solid line).

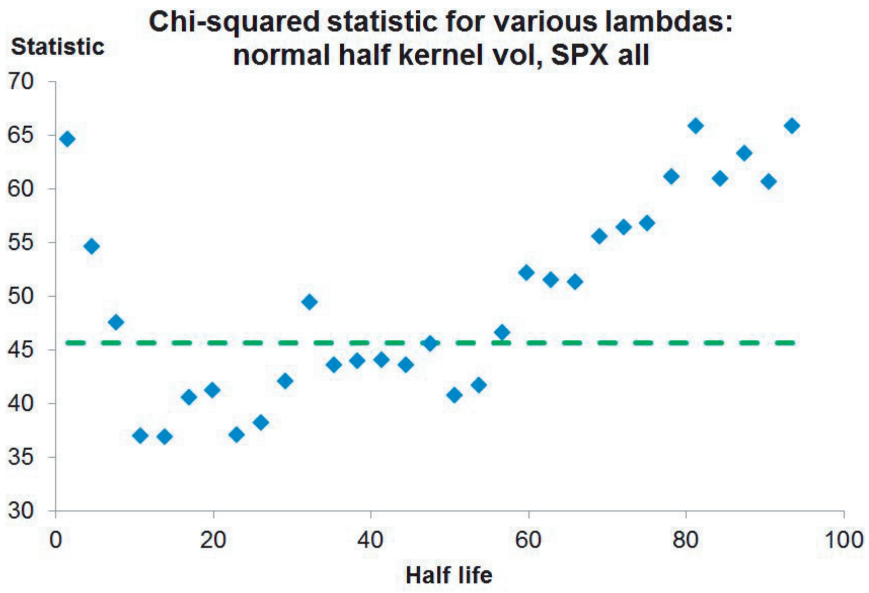
of these volatility estimators as a function of the width parameter. For comparability with other approaches, width is measured in "half-life", ie, the number of days before the weighting function falls to 50% of its peak value.

The best-performing normal half kernel is one with a half-life of twenty-three days, roughly corresponding to a λ of 0.97 (in the sense that $0.97^{23} \approx 0.5$); this again suggests that weighting schemes that fall to half strength over roughly twenty to sixty days tend to produce acceptable volatility estimates for the S&P 500 index. There are obviously many more half kernels that could be evaluated at this point, and this might be a fruitful area for future work.

4.4 Volatility estimation via signal processing

The philosophy of the half-kernel approach is that it is not a priori known what the right weighting scheme for calculating volatility from squared returns is, so one should be selected based on performance. This problem specification suggests we treat volatility estimation as a signal-processing problem, with the squared returns as the input and the volatility estimates as the output. This stance is particularly productive, as there is a large literature on signal processing that can potentially be

FIGURE 10 Testing half-kernel volatility.



The χ^2 statistic for normal half-kernel volatility estimates of various half-lives.

drawn upon (see, for instance, Proakis and Manolakis (2006) for an introduction to this literature).

A key insight in signal processing is that it is sometimes helpful to work in the frequency domain. Thus, many signal-processing techniques first transform the input series into an equivalent representation in the frequency domain using a Fourier transform, manipulate this representation and then transform it back. For instance, a low-pass filter retains the low-frequency components of a signal while discarding the high-frequency ones. If the signal is $\{r_n\}$, $n \in 0, \dots, N-1$, the simplest low-pass filtering approach would

(1) calculate the (complex) coefficients of the kth frequency representation

$$X_k = \sum_{n=0}^{N-1} r_n \exp\left(\frac{-2\pi \mathrm{i} k n}{N}\right)$$

for each $k \in 0, \dots, N-1$;

(2) cut off the frequencies above some threshold υ, applying a filter F by defining

$$Y_k = F_k X_k \quad \text{where } F_k = \begin{cases} 1 & \text{if } k \leqslant \upsilon, \\ 0 & \text{otherwise;} \end{cases} \tag{4.1}$$

(3) rebuild the filtered return series r_i' using the inverse Fourier transform on the Y_k,

$$r_n' = \frac{1}{N} \sum_{k=0}^{N-1} Y_k \exp\left(\frac{2\pi \mathrm{i} k n}{N}\right);$$

and

(4) estimate volatility using the low-pass filtered returns $\{r_n'\}$.

For a well-chosen υ, this approach would filter out high-frequency noise but retain the information in lower-frequency variation in squared returns. More sophisticated versions would use a more gradual cut-off than the step function used in (4.1).

Techniques based on Fourier analysis have been used in volatility estimation by various authors, including Malliavin and Mancino (2002), Mancino and Recchioni (2015) and Barucci and Renò (2002). These show significant promise in being able to offer a highly customizable (and, thus, optimizable) framework. In order to assess their potential in our setting, we first compare various low-pass filters.

The Hann half window was found to perform well using 512 days of returns.[11] This filter is defined by two parameters: the center point, υ (ie, the point where the filter is attenuating by 50%), and the width, w:

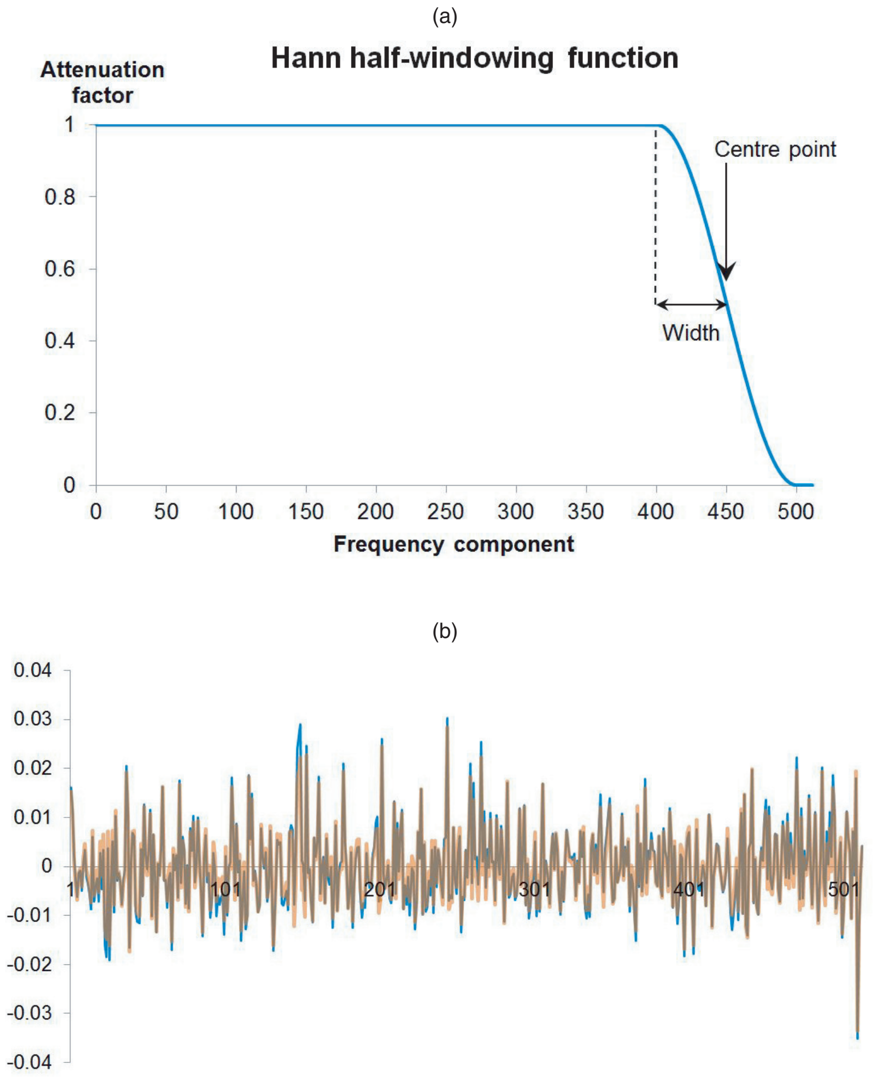
$$F_k = \begin{cases} 1 & \text{if } k \leqslant \upsilon - w, \\ \dfrac{1}{2}\left(1 - \cos \pi \dfrac{k + 3w - \upsilon}{2w}\right) & \text{if } \upsilon - w < k < \upsilon + w, \\ 0 & \text{if } k \geqslant \upsilon + w. \end{cases}$$

Figure 11 illustrates the Hann windowing function with a center point of 450 and a width of 40 applied to a window of 512 days.

The data length was fixed and various center points and widths for the Hann half window were then tested. The optimal approach was found to be a low-pass filter with a high center point and low width, so that only the highest frequencies were attenuated. With 512 days of data, the highest possible frequency was number 511, and the optimal filter was found to have a center point of 509 and a width of 3. The performance of this approach is shown in Figure 12.

[11] Window functions are traditionally applied before the Fourier transform to improve properties of the filter such as spectral leakage or dynamic range; here, however, the window is being applied after the transform.

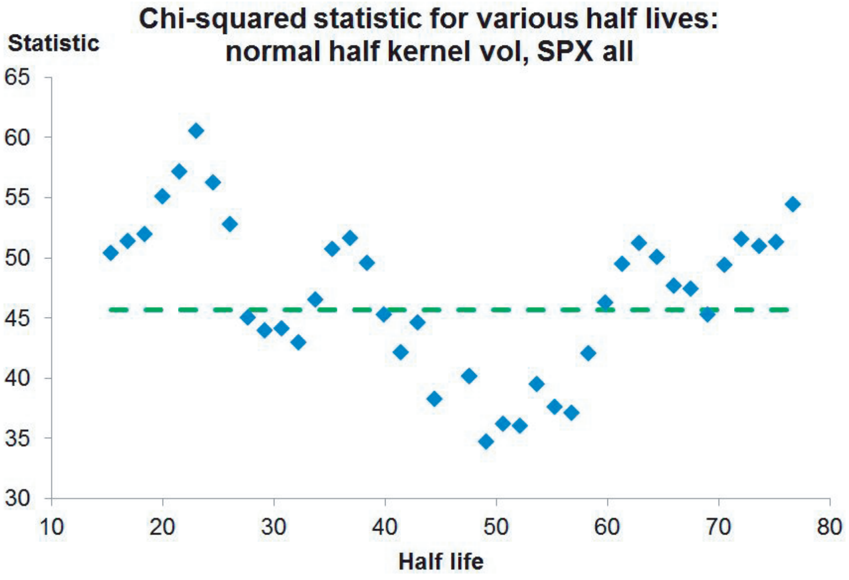
FIGURE 11 Hann half-windowing function.



(a) The Hann half-windowing function with center point 450 and width 40 for a low-pass filter. (b) A time series of returns before (in blue) and after (in orange) the application of the filter.

Using a normal half-kernel volatility estimator without filtering, a range of half-lives between twenty-eight and fifty-eight days was found to be acceptable. If Figure 10 is compared with Figure 12, it can be seen that the low-pass filter increases

FIGURE 12 Fourier performance.



The χ^2 statistic for various normal half-kernel volatility estimators using returns filtered with a low-pass Hann half window.

the acceptable half-life: once the high-frequency variation in returns is attenuated, longer-term volatility estimates perform better.

Estimating volatility in forward windows (ie, those whose first day is not the day after the estimation day) is more difficult than estimating it for spot windows, which is what we have done thus far. The low-pass technique works well here. It turns out that no EWMA model is acceptable for predicting ten-day worst losses ten days forward, for instance; however, several low-pass normal half-kernel estimates are, including the best model for the MPOR starting tomorrow (filter center point 509 and width 3; half-kernel, half-life 49 days). This performance illustrates that filtering can be a useful technique in volatility estimation.

4.5 GARCH models

Generalized autoregressive conditional heteroscedasticity (GARCH) models, as described by Bollerslev (1986) and extended in various ways, for instance, by Glosten *et al* (1993), are popular conditional volatility modeling approaches in the literature. In the simplest version of these models, GARCH(1,1), volatility evolves as

$$\tilde{\sigma}_t^2 = \omega + \alpha r_{t-1}^2 + \beta \tilde{\sigma}_{t-1}^2.$$

There is a range of epistemological positions that can be taken regarding models such as these. At one extreme, the modeler views the chosen process as a true description of the returns process, and, hence, their job is to find the "right" model parameters ω, α and β. Given a long enough history, estimates of these parameters will be accurate in this paradigm, so they should not change materially on recalibration. At the other extreme, the view is that the model's local dynamics are "close enough" in pertinent respects, and, hence, the model's predictions may be useful even if the "true" process is not the modeled one. In this setting, there is no reason to expect that best-fit parameters will not drift over time, and, hence, that periodic recalibration will be necessary. Indeed, it will not be surprising if, here, recalibration results in substantial swings in model parameters. The distinction between the two views is well articulated by Stărică (2003): GARCH(1,1) with a particular set of parameters is "the true data generating process", or it is "a local stationary approximation of the data" (and, thus, he implies that the best locally stationary approximation of the data for a different window may well have different parameters).

The "true description" paradigm can be tested by taking a long history of returns, fitting a GARCH(1,1) model and then testing the out-of-sample performance of this model at volatility prediction. It is found that the hypothesis that the GARCH(1,1) predictions of worst losses are correct is rejected at 99%: the critical statistic is over 100 versus a critical value of about 45.[12]

The "locally true" paradigm requires a recalibration strategy. A window of 512 returns is selected, and a standard GARCH(1,1) model is fitted in each window using a quasi-maximum likelihood estimator with variance targeting.[13] This gives a single volatility prediction for the ten-day period starting at the end of the window, which we can use together with the worst loss in this period. The window is then rolled on ten days and the model is refitted. Figure 13 illustrates the fitted parameters as a function of the starting point of the window.

The hypothesis that the volatility estimates generated by this procedure are correct is accepted at 99%. Nevertheless, the unstable model parameters suggest that the recalibration burden of this approach can be substantial.[14]

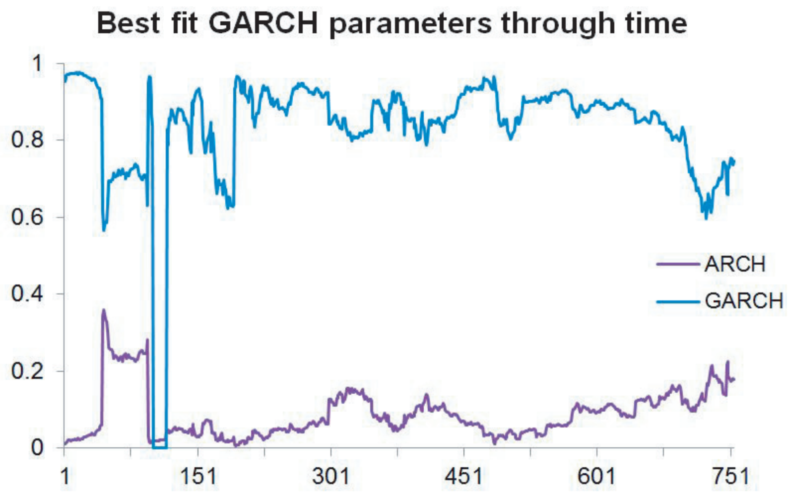
Models with more parameters can sometimes give a better account of returns. As an example, consider the Glosten, Jagannathan and Runkle GJR–GARCH(1,1) model

[12] This rejection of GARCH predictions is not supported by some other studies based on different data sets (see Andersen and Bollerslev (1998) or Hansen and Lunde (2005)).

[13] Variance targeting constrains the fit so that the asymptotic variance of the process is set to the unconditional variance. This obviates the need to estimate ω, and so reduces the complexity of the fitting problem.

[14] The big fall in the GARCH parameter on day 101 of Figure 13 corresponds to the October 1987 crash. This is worrying, as it tends to suggest that just when we need model parameter (and, hence, margin) stability (when the market is crashing), we do not have it.

FIGURE 13 GARCH parameters.



The best fit α (ARCH) and β (GARCH) parameters for each data window in our history of S&P 500 returns.

(Glosten *et al* 1993). This contains an additional parameter, γ, which allows volatility to increase for large negative returns:

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

The additional parameter γ sometimes allows GJR–GARCH(1,1) to outperform GARCH(1,1) models on skewed return series, as Liu and Hung (2010) report. However, it makes model estimation more difficult and thus does not necessarily improve the problem of the fitted parameters being unstable. Indeed, the issue was worse for the data used here, with negative γ occasionally being returned during the same period of stressed conditions that caused unstable GARCH(1,1) parameter fits. The resulting GJR–GARCH volatility estimates also (just) failed the worst-loss test. Thus, this is an interesting counterexample in which the additional freedom of another parameter does not improve volatility estimation.

These results are not intended as a general critique of GJR–GARCH(1,1); it may be that this technique, or the related threshold GARCH models described by Li and Lam (1995), could perform well for different windows or return horizons. Rather, they illustrate that having better explanatory power is not always useful in a technique employed for margin modeling if this power comes with occasional calibration problems. It is unlikely that margin posters will agree to post margin that is mostly calculated using a complex model but is sometimes calculated using something simpler because the complex model has a bad calibration using a window going backward from yesterday's data.

5 NUANCES AND EXTENSIONS

This section describes some complexities that should be recognized before using the probability of worst loss approach for sensitivity analysis in practice.

5.1 Data requirements

Techniques that discriminate between risk models, and especially those that discriminate between identical algorithms with similar parameters, tend to require a lot of data. Indeed, Daníelsson and Zhou (2015, p. 5) state:

> A sample size of half a century of daily observations is needed for the empirical estimators [of VaR] to achieve their asymptotic properties.

Our approach is not quite this constraining, but it is nearly so: even with thirty years' worth of data, Figure 7 is not smooth in the model parameter, which indicates some noise in the estimates of the χ^2 statistic. In this case, it might be helpful to reduce the number of probability buckets. The lack of smoothness also suggests that, in practice, there will not be many risk factors where enough data is available to use bucketing with confidence.

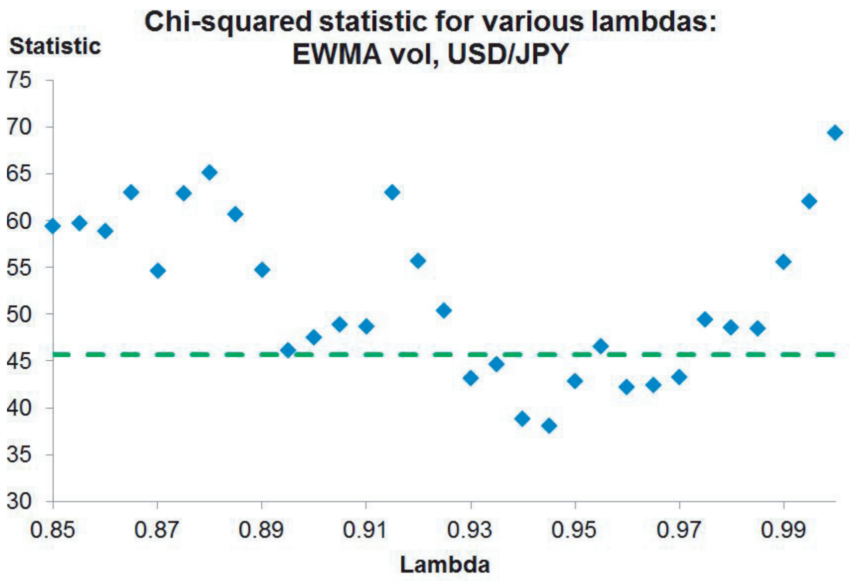
5.2 A better test for uniformity

The χ^2 approach to normality testing, while appealingly simple, has relatively low power. A more powerful approach would be to use Berkowitz's probability integral transforms (Berkowitz 2001; Hamerle and Plank 2009). Here, the assumed uniform distribution is transformed using the inverse cumulative normal distribution, giving data z_t. A first-order autoregressive process

$$z_t - \mu = \rho(z_{t-1} - \mu) + \epsilon_t$$

is then fitted. If the origination distribution is uniform, then μ and ρ should be close to zero and the variance of the ϵ should be close to 1. Likelihood ratios can be used to test this. We hope to explore the probability integral transform approach in future work. It would be natural to combine an even more discriminating test such as this with a study of conditionally heavy-tailed processes (as discussed in Section 3.1).

FIGURE 14 US$/¥ performance.



The χ^2 statistic for various EWMA volatility estimators applied to US$/¥ returns.

5.3 Risk factor differences

There is no reason to believe that the same volatility estimator will be optimal, or even acceptable, for all the risk factors in a large portfolio. To study this, the various EWMA estimators were retested using the US$/¥ spot rate as the risk factor instead of the S&P 500 index. Figure 14 illustrates the results. There are some differences: $\lambda = 0.93$ is acceptable for US$/¥ but not for the S&P 500, for instance. Clearly there is some danger that, as the number of risk factors grows, there will be no model that is acceptable for all of them.[15]

6 CONCLUSIONS

A key part of IM model validation will always be backtesting model margin requirements at the target confidence interval. However, this is not enough for optimal model design or to perform sensitivity analysis: other techniques are necessary to verify the

[15] Once the estimation of covariances is included in the problem, this problem clearly becomes worse.

choice of model parameters. One approach to this problem has been presented, which focusses on the statistical accuracy of the models' predictions of worst loss over the MPOR. Techniques such as this can also help by giving early warnings of a model that is providing accurate high-quantile risk estimates for a limited period of time, but which will not prove to be robust as market conditions evolve.

Various algorithms have been tested using the technique proposed, and acceptable parameterizations have been presented. The algorithms range from the well known and commonly used, such as EWMA volatility estimation, to techniques inspired by the idea that volatility estimation can be thought of as a signal-processing problem. These latter approaches are potentially interesting in that they open up a large stock of filtering techniques that can assist in separating out uninformative high-frequency noise in returns from valuable information about volatility trends.

One interesting feature of the analysis presented is that most models tested had some acceptable range of parameters, and that, for decay-constant-like parameters, the acceptable range tended to include models with a half-life of between twenty and sixty days. This suggests that persistence on this time scale is a somewhat model-independent feature of the S&P 500.

It can sometimes be found that the performance of a model is insensitive enough to parameter choices that a wide range of them is acceptable. The blended volatility model illustrated in Figure 8 is an example of this. When this happens, it is helpful to have an additional criterion for preferring one parameter setting over another. The obvious choice is procyclicality: margin stability is valuable, so less procyclical models are preferred over more reactive ones. Murphy *et al* (2014) discuss measures of margin model procyclicality that can be used here. In other cases, when a model only performs acceptably for a narrow range of parameters, the choice is easy. Neither outcome, however, absolves the model user from continued diligence: it is important to perform sensitivity analysis regularly to ensure that model design and parameters remain appropriate.

DECLARATION OF INTEREST

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England, or any other institution.

ACKNOWLEDGEMENTS

The authors would like to thank Paul Bedford, Fernando Cerezetti and Michalis Vasios for comments on previous versions of this paper. Melanie Houllier conducted work while employed at the Bank of England.

REFERENCES

Aitsahlia, F., and Lai, T. (1998). Random walk duality and the valuation of discrete lookback options. *Applied Mathematical Finance* **5**, 227–240.

Andersen, T., and Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review* **39**(4), 885–905.

Barucci, E., and Renò, R. (2002). On measuring volatility and the GARCH forecasting performance. *Journal of International Financial Markets, Institutions and Money* **12**(3), 183–200.

Basel Committee on Banking Supervision (2013). Margin requirements for non-centrally cleared derivatives. Report 261, Bank for International Settlements.

Berkowitz, J. (2001). Testing density forecasts with applications to risk management. *Journal of Business and Economics Statistics* **19**(4), 465–474.

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

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

Campbell, S. D. (2007). A review of backtesting and backtesting procedures. *The Journal of Risk* **9**(2), 1–17 (http://doi.org/b7v5).

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review* **39**(4), 841–862.

Crnkovic, C., and Drachman, J. (1997). *"Quality Control" in VaR: Understanding and Applying Value-at-Risk*. Risk Publications.

Daníelsson, J., and Zhou, C. (2015). Why risk is so hard to measure. Systemic Risk Centre Discussion Paper 36, London School of Economics.

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

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of United Kingdom inflation. *Econometrica* **50**(4), 987–1007.

Escanciano, J., and Pei, P. (2012). Pitfalls in backtesting historical simulation VAR models. *Journal of Banking & Finance* **36**(8), 2233–2244 (http://doi.org/b7v6).

European Union (2012a). 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 [commonly known as EMIR]. *Official Journal of the European Union* **L 201**, 1–59.

European Union (2012b). 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 [commonly known as the EMIR RTS]. *Official Journal of the European Union* **L 52**, 41ff.

G20 (2009). G20 leaders statement: the Pittsburgh summit [September 24–25, 2009, Pittsburgh]. Report, G20 Research Group. URL: www.g20.utoronto.ca/2009/2009communique0925.html.

Gijbels, I., Pope, A., and Wand, M. (1999). Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society* B **61**(1), 39–50.

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

Gurrola Perez, P. (2015). Calibrating a new generation of initial margin models under the new regulatory framework. In *Proceedings of Systemic Risk in Over-The-Counter Markets: The Third Annual LSE Conference on Systemic Risk, Systemic Risk Centre*.

Haas, M. (2001). New methods in backtesting. Financial Engineering Note, 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 (http://doi.org/b7v7).

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 (http://doi.org/cqjnjj).

Houllier, M., and Murphy, D. (2017). Borderline: judging the adequacy of return distribution estimation techniques in initial margin models. Staff Working Paper, Bank of England, forthcoming.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* **3**(2), 73–84 (http://doi.org/b846ct).

Liu, H.-C., and Hung, J.-C. (2010). Forecasting S&P-100 stock index volatility: the role of volatility asymmetry and distributional assumption in GARCH models. *Expert Systems with Applications* **37**, 4928–4934 (http://doi.org/dk7brt).

Liesenfeld, R., and Jung, R. (2000). Stochastic volatility models: conditional normality versus heavy-tailed distributions. *Journal of Applied Econometrics* **15**(2), 137–160 (http://doi.org/dqnp2q).

Li, W., and Lam, K. (1995). Modelling asymmetry in stock returns by a threshold autoregressive conditional heteroscedastic model. *Journal of the Royal Statistical Society* **44**(3), 333–341 (http://doi.org/c9x7wg).

Lopez, J. (2001). Evaluating the predictive accuracy of variance models. *Journal of Forecasting* **20**(2), 87–109 (http://doi.org/fjs7p3).

Malliavin, P., and Mancino, M. (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics* **6**(1), 49–61.

Mancino, M., and Recchioni, M. (2015). Fourier spot volatility estimator: asymptotic normality and efficiency with liquid and illiquid high-frequency data. *PLOS ONE* **10**(9), e0139041 (http://doi.org/f759nh).

Murphy, D., Vasios, M., and Vause, N. (2014). An investigation into the procyclicality of risk-based initial margin models. Financial Stability Paper 29, Bank of England.

Pelletier, D., and Wei, W. (2016). The geometric–VaR backtesting method. *Journal of Financial Econometrics* **14**(4), 725–745 (http://doi.org/f88gmm).

Poon, S., and Granger, C. (2003). Forecasting volatility in financial markets: a review. *Journal of Economic Literature* **41**(2), 478–539.

Proakis, J., and Manolakis, D. (2006). *Digital Signal Processing*, 4th edn. Prentice Hall.

Stărică, C. (2003). Is GARCH(1,1) as good a model as the Nobel prize accolades would imply? Working Paper, Chalmers University of Technology. URL: http://ssrn.com/abstract=637322.

Wand, M., and Jones, M. (1995). *Kernel Smoothing*. Chapman & Hall, London (http://doi.org/b7v4).