Calibrating a new generation of initial margin models under the new regulatory framework

Pedro Gurrola

Systemic Risk in Over-The-Counter Markets
Third Annual Conference on Systemic Risk
Systemic Risk Centre, LSE
19th November 2015

Disclaimer

This is a work in progress. The views expressed in this paper are those of the author and not necessarily those of the Bank of England

Agenda

1. Background
2. Modelling using altered samples
   1. Stressed samples
   2. Filtered samples
3. The validation/calibration problem
   1. Backtesting (counting breaches)
   2. Score functions
4. Simulation exercise
5. Results

Table of Contents

- Calculation of margin requirements for central and non-centrally cleared trades generally involves estimation of some quantile-based risk measure, like Value-at-Risk (VaR) or Expected Shorfall (ES).

- Calculation of margin requirements for central and non-centrally cleared trades generally involves estimation of some quantile-based risk measure, like Value-at-Risk (VaR) or Expected Shorfall (ES).
- To estimate these measures, it is common to use an Historical Simulation (HS) approach. The assumption is that the historical sample is a good approximation of the forecast distribution

- Calculation of margin requirements for central and non-centrally cleared trades generally involves estimation of some quantile-based risk measure, like Value-at-Risk (VaR) or Expected Shorfall (ES).
- To estimate these measures, it is common to use an Historical Simulation (HS) approach. The assumption is that the historical sample is a good approximation of the forecast distribution
- The calibration of a basic HS VaR model reduces to choosing the length of historical window ("look-back period").

- Calculation of margin requirements for central and non-centrally cleared trades generally involves estimation of some quantile-based risk measure, like Value-at-Risk (VaR) or Expected Shorfall (ES).
- To estimate these measures, it is common to use an Historical Simulation (HS) approach. The assumption is that the historical sample is a good approximation of the forecast distribution
- The calibration of a basic HS VaR model reduces to choosing the length of historical window ("look-back period").
- To improve risk sensitivity and/or to meet new regulatory requirements, more recent HS approaches involve some alteration of the historical sample: adding periods of stress, scaling volatility, or both. The calibration then involves additional choices.

Historical Simulation

2-factor example

Historical sample

RF1 $\Delta S_0^1, \Delta S_1^1, \ldots, \Delta S_N^1$

RF2 $\Delta S_0^2, \Delta S_1^2, \ldots, \Delta S_N^2$

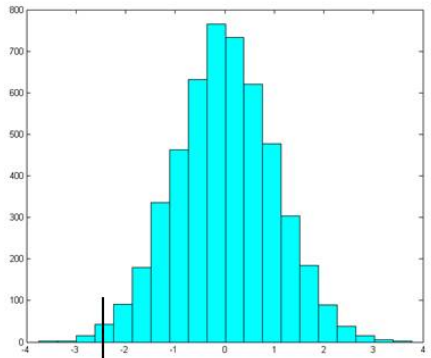
\+

Co-movement structure (implicit)

Evaluate portfolio

$\{\Delta\Pi_0, \Delta\Pi_1, \ldots, \Delta\Pi_N\}$

Estimate P&L distribution



Quantile-based risk measure $\hat{x}_p$

Table of Contents

Stressed by regulation: EMIR Art 28

A CCP shall ensure that its policy for selecting and revising (...) the lookback period deliver forward looking, stable and prudent margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected. This shall include avoiding when possible disruptive or big step changes in margin requirements and establishing transparent and predictable procedures for adjusting margin requirements in response to changing market conditions. In doing so, the CCP shall employ at least one of the following options:

Stressed by regulation: EMIR Art 28

A CCP shall ensure that its policy for selecting and revising (...) the lookback period deliver forward looking, stable and prudent margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected. This shall include avoiding when possible disruptive or big step changes in margin requirements and establishing transparent and predictable procedures for adjusting margin requirements in response to changing market conditions. In doing so, the CCP shall employ at least one of the following options:

1. *Applying a margin buffer at least equal to 25 % of the calculated margins which it allows to be temporarily exhausted in periods where calculated margin requirements are rising significantly;*

Stressed by regulation: EMIR Art 28

A CCP shall ensure that its policy for selecting and revising (...) the lookback period deliver forward looking, stable and prudent margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected. This shall include avoiding when possible disruptive or big step changes in margin requirements and establishing transparent and predictable procedures for adjusting margin requirements in response to changing market conditions. In doing so, the CCP shall employ at least one of the following options:

1. *Applying a margin buffer at least equal to 25 % of the calculated margins which it allows to be temporarily exhausted in periods where calculated margin requirements are rising significantly;*
2. *Ensuring that its margin requirements are not lower than those that would be calculated using volatility estimated over a 10 year historical lookback period;*

Stressed by regulation: EMIR Art 28

A CCP shall ensure that its policy for selecting and revising (...) the lookback period deliver forward looking, stable and prudent margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected. This shall include avoiding when possible disruptive or big step changes in margin requirements and establishing transparent and predictable procedures for adjusting margin requirements in response to changing market conditions. In doing so, the CCP shall employ at least one of the following options:

1. *Applying a margin buffer at least equal to 25 % of the calculated margins which it allows to be temporarily exhausted in periods where calculated margin requirements are rising significantly;*
2. *Ensuring that its margin requirements are not lower than those that would be calculated using volatility estimated over a 10 year historical lookback period;*
3. *Assigning at least 25 % weight to stressed observations in the lookback period calculated in accordance with Article 26*

Stressed by regulation: Trades not cleared by CCPs

Article 3 MRM - Calibration of the model: Initial margin models shall be calibrated based on historical data from a period of at least three years and not exceeding five years. The data used in initial margin models shall include the most recent continuous period from the calibration date and shall contain at least 25% of data representative of a period of significant financial stress (stressed data). Where the most recent data period does not contain at least 25% of stressed data, the least recent data in the time series shall be replaced by data from a period of significant financial stress, until the overall proportion of stressed data is at least 25% of the overall data set (...).

[Draft Regulatory Technical Standards (RTS) on risk-mitigation techniques for OTC-derivative contracts not cleared by a CCP under Article 11(15) of Regulation (EU) No 648/2012]

Benefits and challenges

Initial margin estimates will tend to be...

- ✓ *more conservative:* the stressed period will drag upwards the margin estimate,
- ✓ *more prudent*: the effect will be greater in times of low volatility,
- ✓ *more stable*: with the stressed period acting as a ballast.

Benefits and challenges

Initial margin estimates will tend to be...

- ✓ *more conservative:* the stressed period will drag upwards the margin estimate,
- ✓ *more prudent*: the effect will be greater in times of low volatility,
- ✓ *more stable*: with the stressed period acting as a ballast.

But the choice of the stressed period also brings some challenges:

Benefits and challenges

Initial margin estimates will tend to be...

- ✓ *more conservative:* the stressed period will drag upwards the margin estimate,
- ✓ *more prudent*: the effect will be greater in times of low volatility,
- ✓ *more stable*: with the stressed period acting as a ballast.

But the choice of the stressed period also brings some challenges:

- The increase in margin could be unnecessarily costly and economically inefficient.
- In a conditional setting, the timing of the stress affects the outcome
- Stress periods may not be consistent across risk factors

Filtered samples

Introducing a stressed period is not the only common alteration to the historical sample:

Filtered samples

Introducing a stressed period is not the only common alteration to the historical sample:

- To increase the sensitivity of margin models to the arrival of new information, it is frequent to incorporate a volatility updating scheme to better reflect current market conditions.

Filtered samples

Introducing a stressed period is not the only common alteration to the historical sample:

- To increase the sensitivity of margin models to the arrival of new information, it is frequent to incorporate a volatility updating scheme to better reflect current market conditions.
- Common approaches are variants of the Filtered Historical Simulation (FHS) methods suggested by John Hull and Allan White (1998) and Barone-Adesi, Bourgoin and Giannopoulos (1998).
- Examples: initial margin methodologies for interest rate products used by LCH Swapclear, CME and Eurex [Gregory(2014)].

FHS (Hull-White)

Let $Y_T = \{r_1, r_2, ...r_{T-1}\}$ be the historical sample of EOD returns used to make a forecast for day T. The filtering process involves two steps:

1. Each historical return $r_i \in Y_T$ is divided by the volatility estimate σ_i for day i to obtain a sample of standardised residuals $\bar{r}_i = r_i/\sigma_i$ which is assumed to be approximately stationary (in volatility).

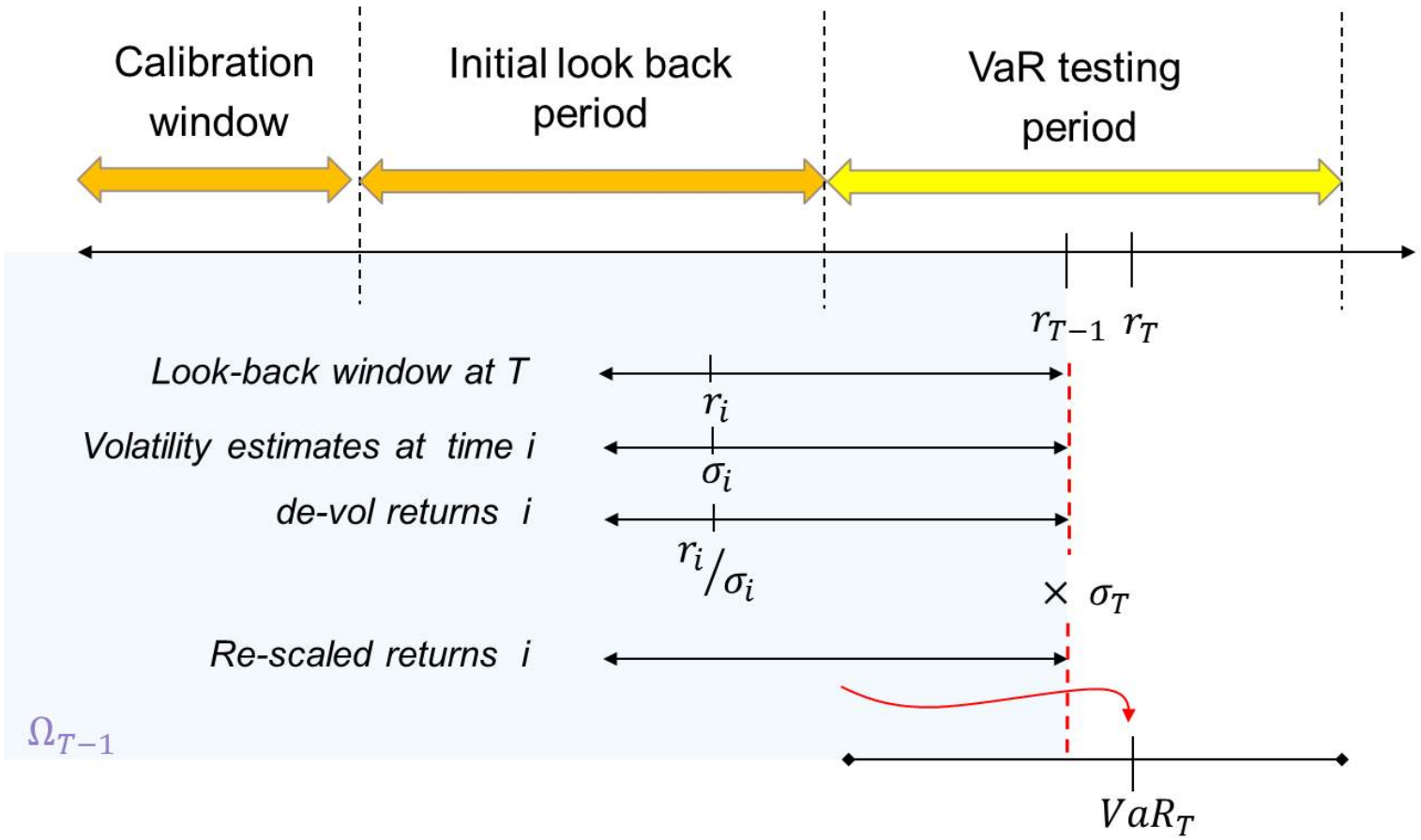
FHS (Hull-White)

Let $Y_T = \{r_1, r_2, ...r_{T-1}\}$ be the historical sample of EOD returns used to make a forecast for day T. The filtering process involves two steps:

1. Each historical return $r_i \in Y_T$ is divided by the volatility estimate σ_i for day i to obtain a sample of standardised residuals $\bar{r}_i = r_i/\sigma_i$ which is assumed to be approximately stationary (in volatility).
2. The residuals $\bar{r}_i$ are multiplied by day T volatility forecast σ_T to obtain a sample of rescaled returns

$$R_i = r_i \frac{\sigma_T}{\sigma_i}, \qquad 1 \leq i < T$$

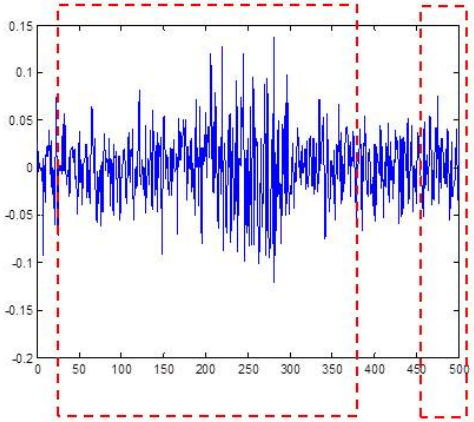
FHS (Hull-White)



FHS (Hull-White)

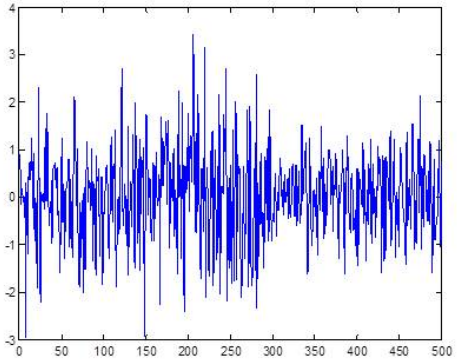
Original returns → "Standardised" returns → Updated (re-scaled) returns

r_i



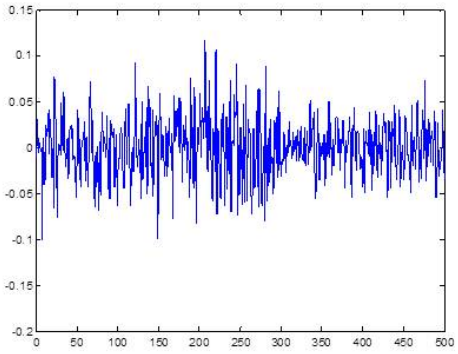
$\sigma_i \approx 0.04$ $\sigma_T \approx 0.034$

r_i/σ_i



$\sigma \approx 1$

$r_i/\sigma_i \times \sigma_T$



$\sigma \approx 0.034$

FHS with EWMA

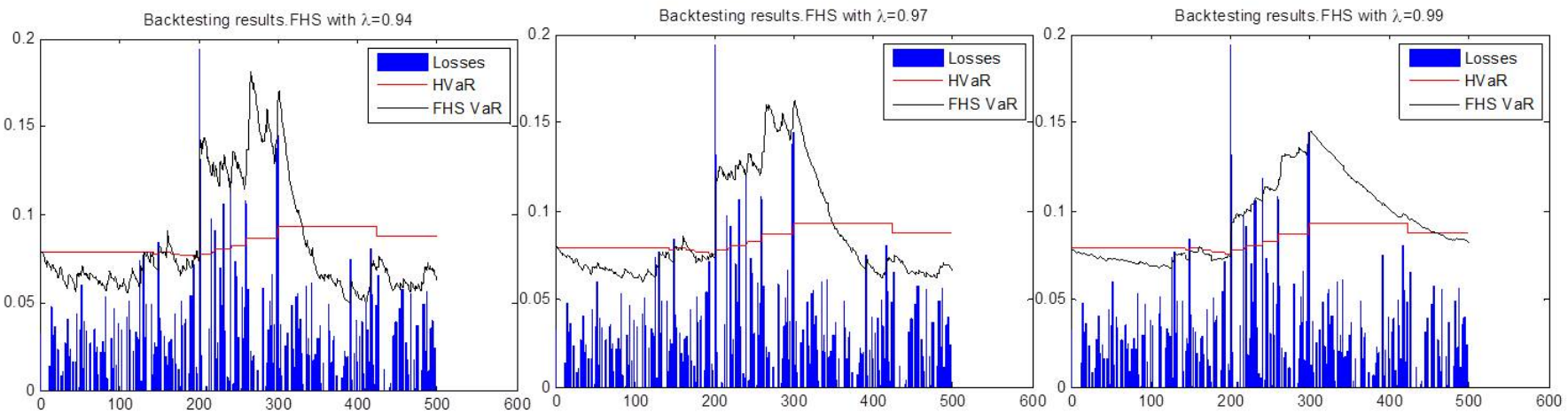
- The volatility estimates for the devol and revol steps can be derived, for example, from an EWMA volatility updating scheme or from a GARCH process.
- In the case of using an EWMA process, volatility estimates can be calculated using the recursive formula

$$\sigma_{t+1}^2 = \lambda \sigma_t^2 + (1 - \lambda) r_t^2 \tag{1}$$

The decay factor, $\lambda \in [0, 1]$, determines the responsiveness of the process to the arrival of new information.

FHS (Hull-White) with EWMA forecasts

The effect of lambda



$\lambda = 0.94$ $\lambda = 0.97$ $\lambda = 0.99$

More reactive to new information

Pasta data more persistent

Table of Contents

- The usual way to test the VaR forecasts is to observe the time series of past *ex-ante* VaR forecasts, compare it with the realized *ex-post* returns and count the number of times the observed returns breached the forecast (backtesting).
- The results of the backtesting exercise can be statistically assessed (e.g. hypothesis testing).
- Regulatory requirements are often linked and/or rely on backtesting as the main validation tool (Basel, EMIR, ...)
- This provides incentives to calibrate the model using backtesting requirements as objective. The question: *Is the model providing a realistic/plausible representation of the risk factors?* is replaced by the (weaker) question: *Is the model output historically consistent with a specified percentile?*

Let $q_{t+h}(\alpha)$ denote a portfolio's h-day VaR measure calculated on day t with a $(1-\alpha)\times 100\%$ coverage:

$$q_{t+h}(\alpha) = -F^{-1}(\alpha|\Omega_t)$$

For simplicity, we assume $h = 1$. Consider the hit function:

$$I_\alpha(t+1) = \begin{cases} 1 & \text{if } u_{t+1} < -q_{t+1}(\alpha) \\ 0 & \text{otherwise} \end{cases}$$

then

$$x = \sum_{i=1}^{N} I_\alpha(i)$$

is the total number of breaches on a sample of N observations and $f = \frac{x}{N}$ is the observed frequency of breaches.

The accuracy of VaR measure can be determined based on the hit sequence satisfying two properties [Christoffersen (1998)]:

- *Unconditional coverage*: The model meets its target $\alpha * 100\%$ of the time on average, that is

$$H_0 : \mathbb{E}[I_\alpha(t)] = \alpha$$

- *Independence:* Any two elements of the hit sequence must be independent from each other.

Equivalently: the hit sequence is identically and independently distributed as a Bernoulli random variable with probability α:

$$I_t(\alpha) \overset{i.i.d.}{\sim} B_t(\alpha)$$

Tests for unconditional coverage

- Proportion of failures (POF) test [Kupiec(1995)]: Apply a likelihood ratio LR_{uc} to test the null $H_0 : f = \alpha$.

$$LR_{uc} = -2 \ln \left(\frac{\alpha^x (1-\alpha)^{N-x}}{f^x (1-f)^{N-x}} \right), \qquad LR_{uc} \sim \chi^2(1) \qquad (2)$$

Well-known shortcomings:

- Low power in small samples (e.g. 1 year)[Kupiec(1995)]
- May fail to detect violation clustering.

Tests for conditional coverage

Tests for independence:

- Markov tests (using 1-day lags) [Christoffersen (1998)].
- Time between failures test [Haas (2001)]. For the i-th exception, consider the statistic

$$LR_i = -2 \ln \left(\frac{p(1-p)^{v_i - 1}}{\frac{1}{v_i}\left(1 - \frac{1}{v_i}\right)^{v_1 - 1}} \right)$$

where v_i is the time between exceptions i and $i-1$. Summing up the LR_i statistics yields the statistic LR_{ind}.

Joint tests (conditional coverage):

- "Mixed Kupiec-test" [Haas (2001)]:

$$LR_{mix} = LR_{POF} + LR_{ind}, \qquad LR_{mix} \sim \chi^2(x+1)$$

Drawbacks of backtesting based on counting breaches

- Poor predictive power in small samples
- Focused on a percentile, ignoring other distributional information.
- They do not fit well with conditional behaviour [Davis(2014)]
- Previous analysis shows that a single backtest is not good enough and that more than one test is necessary [Haas (2001)].

Drawbacks of backtesting based on counting breaches

- Poor predictive power in small samples
- Focused on a percentile, ignoring other distributional information.
- They do not fit well with conditional behaviour [Davis(2014)]
- Previous analysis shows that a single backtest is not good enough and that more than one test is necessary [Haas (2001)].

In a sample-modified model, the validation poses additional challenges...

Backtesting when using stressed samples

- Lack of consistency: Unless the period of stress genuinely happened during the lookback period, the historical sample is no longer consistent with the observations against which it is tested.
- Model misspecification: By calibrating it to a "stressed world", the model is deliberately misspecified (increased model risk).

Backtesting when using stressed samples

- Lack of consistency: Unless the period of stress genuinely happened during the lookback period, the historical sample is no longer consistent with the observations against which it is tested.
- Model misspecification: By calibrating it to a "stressed world", the model is deliberately misspecified (increased model risk).

$\Rightarrow$ Potentially misleading results: Backtesting will tend to produce less exceptions, leading to a false (too optimistic) conclusion about model's performance.

Backtesting when using stressed samples

Can we restore consistency?

Backtesting when using stressed samples

Can we restore consistency?

- Artificially "stress" the testing sample? *Not feasible in practice, for example, if backtesting is done on a daily basis.*

Backtesting when using stressed samples

Can we restore consistency?

- Artificially "stress" the testing sample? *Not feasible in practice, for example, if backtesting is done on a daily basis.*
- Backtest in parallel a model calibrated with the real historical data? *But, what confidence level could be used to calibrate the model?*

Backtesting when using stressed samples

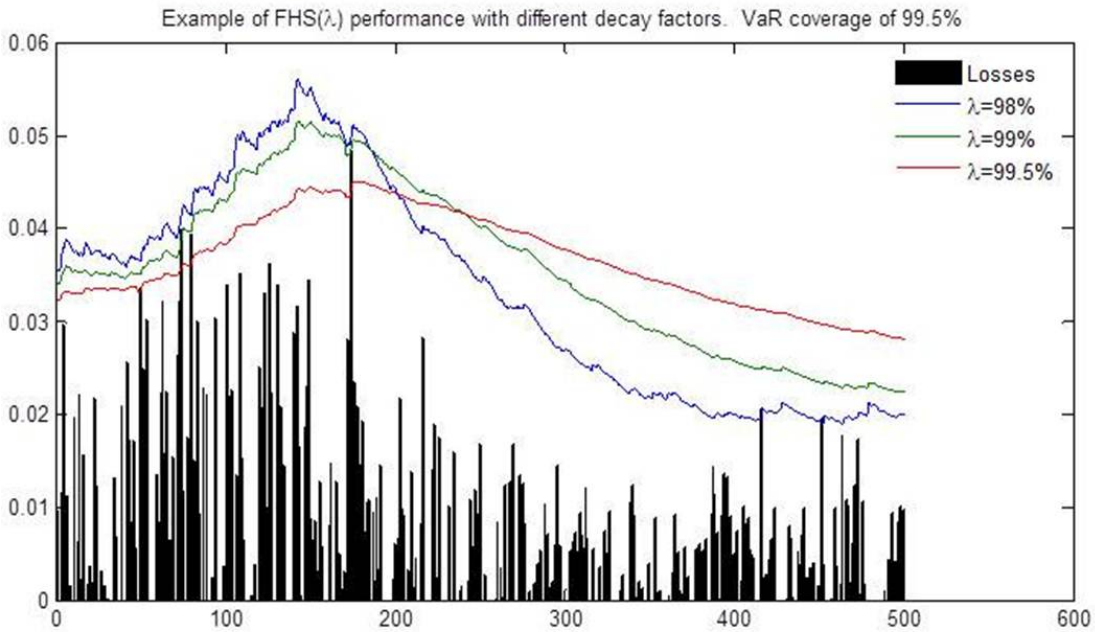
Can we restore consistency?

- Artificially "stress" the testing sample? *Not feasible in practice, for example, if backtesting is done on a daily basis.*
- Backtest in parallel a model calibrated with the real historical data? *But, what confidence level could be used to calibrate the model?*

These constitute a strong indication for the stressed element to be incorporated (along with any other add-ons) at the level of the risk measure and not within the model calibration (see also [Murphy, Vasios and Vause(2015)]).

Tweak the output, not the sample!

Backtesting when using filtered samples



λ	# breaches
0.9	4
0.92	3
0.94	3
0.96	2
0.97	2
0.98	2
0.99	1
0.995	4

The DGP is a randomly generated EWMA process with $\lambda_0 = 0.97$ and gaussian innovations.

- Kupiec's (POF) test at 95% confidence: 0 to 6 exceptions is acceptable.
- For an acceptability range between 1 and 4 the confidence should be 75%.

Using scores to choose between models

- Compare competing models by means of an error measure providing a score that reflects departures of the model from the expected behaviour.
- The calibration of a model could then be done by choosing the model that optimizes the score.
- These methodologies can incorporate specific evaluators concerns through the choice of a loss function.
- Loss functions can be specially attractive for backtesting with relatively small amount of observations or to incorporate additional distribution or specific tail information

Using scores to choose between models

Let $\mathcal{Q}$ denote a set of competing models and q_t^m the VaR forecast produced for time $t-1$ by model $m \in \mathcal{Q}$. A loss function

$$\ell_t^m = \begin{cases} g(u_t, q_t^m) & \text{if} \quad u_t < -q_t^m \\ f(u_t, q_t^m) & \text{if} \quad u_t \geq -q_t^m \end{cases}$$

(assume that $g(u_t, q_t^m) > f(u_t, q_t^m)$) can be used to define a score function:

$$S_t^m = S(\ell_t^m, p_t),$$

where p_t is a benchmark reflecting the expected (correct) model behaviour at time t. The final score for model m would be a function of the aggregated scores in time:

$$S(m) = \Sigma(S_t^m).$$

Scoring functions

The quadratic probability score (QPS) function can be used to measure the accuracy of probability forecasts over time [Lopez(1999)]:

$$\text{QPS}(m) = \frac{1}{N}\sum_{t=1}^{N} 2(\ell_t^m - p_t)^2$$

where p_t is the expected value of ℓ_t^m under the null hypothesis that the model is correct and N is the sample size.

- It is the analog of mean squared error for probability forecasts and implies a quadratic loss function.
- It is a strictly proper scoring rule; that is, forecasters must report their actual probability forecasts to minimize their expected QPS
- $\text{QPS}_m \in [0, 2]$ and has a negative orientation such that smaller values indicate more accurate forecasts.

Lopez loss function

The loss function can be specified in different ways. To take into account the size of the loss and penalize larger losses, it can take the following quadratic form [Lopez(1998)] :

$$\ell_t^m = \begin{cases} 1 + (u_t - q_t^m)^2 & \text{if } u_t < -q_t^m(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Although it has the disadvantage that there is no straightforward condition for the benchmark [Lopez(1998)], incorporating information about the size of the loss can help when comparing different models (everything else being equal).

Dowd loss function

A score can also be defined using the actual loss [Dowd(2005)] :

$$\ell_t^m = \begin{cases} u_t & \text{if } u_t < -q_t^m(\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In this case, the benchmark is the expected shortfall at time t, ES_t, and the scoring function can be defined as

$$\text{QPS}_m = \frac{1}{N} \sum_{t=1}^{N} 2(\ell_t^m - \text{ES}_t)^2. \tag{5}$$

Table of Contents

Simulation

- We consider FHS VaR models based on EWMA volatility estimates, with decay factors λ_m.
- First step would be to test the FHS models using an EWMA generated process as the true data generating process (DGP). In this way, the "true" decay factor λ_0 is known and we can assess which backtesting approach is optimal in solving the calibration problem.
- However, it will be convenient to set the problem in the more general context of Integrated GARCH processes (IGARCH) introduced by [Engle and Bollerslev (1986)].

IGARCH(1,1)

An IGARCH(1,1) model with normal innovations can be specified as follows:

$$\begin{aligned} r_t &= \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1) \\ \sigma_{t+1}^2 &= \omega + \lambda_0 \sigma_t^2 + (1-\lambda_0) r_t^2 \end{aligned}$$

so that σ_t^2 is the conditional variance of the returns r_t given the history of the system. The conditional expectation of σ_{t+k}^2 at time t is

$$\mathbb{E}(\sigma_{t+k}^2 | \sigma_t^2) = \sigma_t^2 + \omega \cdot k \tag{6}$$

In particular, when the drift component ω is zero, the variance follows an EWMA process.

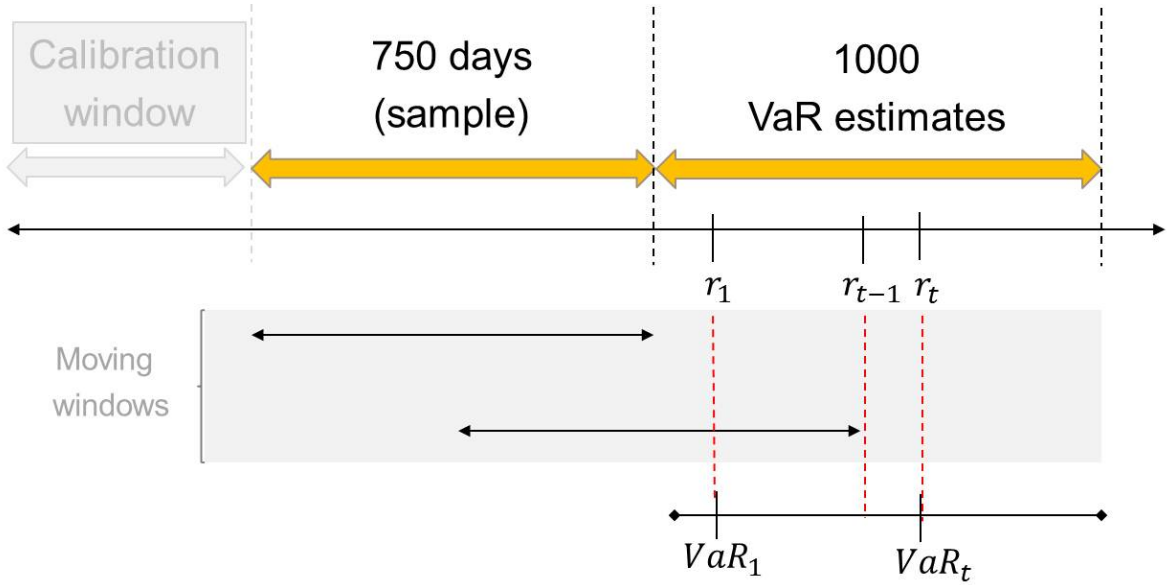
IGARCH(1,1)

- When the drift is zero, the distribution of σ_t^2 concentrates around zero with fatter tails [Engle and Bollerslev (1986)].
- An IGARCH process with zero drift converges almost surely to zero, while for $\omega > 0$ the process is strictly stationary and ergodic [Nelson(1990)].
- Moreover, in a volatility decreasing environment less sensitive models will tend to produce better backtesting results, which suggests that it may be more appropriate to test the models in the more general setting of non-zero drift IGARCH(1,1).
- To stay within the limits of a one-dimensional problem, one can assume the drift ω is a variable known to the modeler (and test the impact of the choice of ω, as robustness check).

For each simulation exercise:

- Daily returns are simulated using IGARCH(1,1)$[\omega_0, \lambda_0]$ as the data generating process (DGP).
- 1000 simulation runs, each one generating a sample of 1750 observations from the DGP.
- Eight different calibrations of an FHS VaR model with decay factor λ_m are tested.
- Each FHS$[\lambda_m]$ VaR is based on an IGARCH(1,1) with the same drift ω_0 but different decay factor λ_m.

By knowing the true conditional volatility σ_t and the decay speed λ_0 of the data generating process, we can compare the tests: the power of the test should be reflected in the number of times FHS$[\lambda_0]$ is selected as the optimal choice.



- Each run produces a set of 1000 daily VaR estimates obtained from 750 sample moving windows. These choices reflect common situations found when dealing with historical data.
- The set of VaR estimates is backtested at different VaR coverage levels.

Simulation Parameters

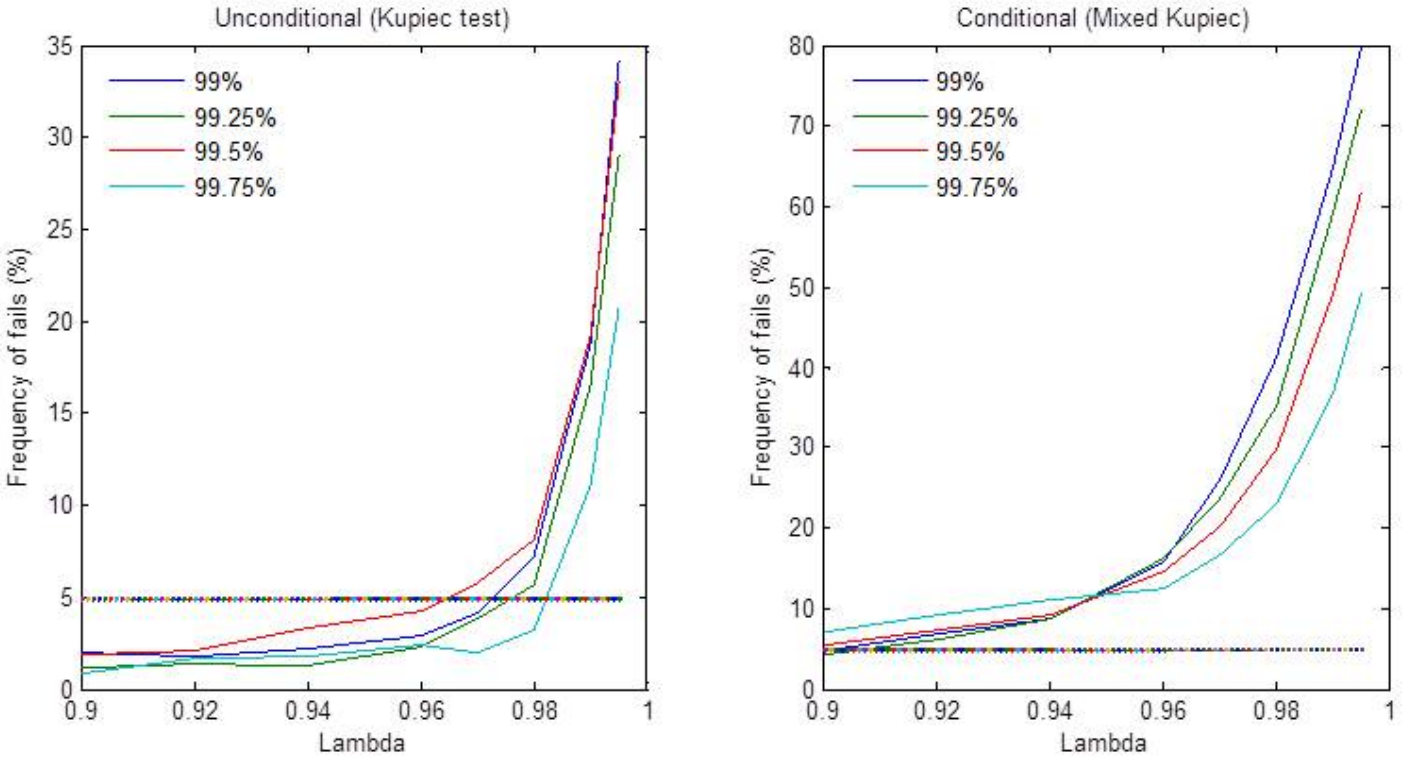
	Parameter	Values
DGP = IGARCH(1,1)	$\lambda_0 =$	0.94, 0.97, 0.99
	$\omega_0 =$	0.0179 (=volatility seed σ_0)
FHS model	$\lambda_m =$	0.9, 0.92 , 0.94 , 0.96 , 0.97 , 0.98 , 0.99 , 0.995
	$\omega_m =$	ω_0
VaR measure	coverage	99%, 99.25%, 99.5%, 99.75%

For each DGPs, the FHS[λ_m] model is tested using three sets of tests:

1. *Hypothesis testing approaches:* For each run and for each confidence α, determine which models fail or pass under Kupiec's POF and the Mixed Kupiec test (confidence level $\gamma = 0.95$). The number of times a model is rejected across simulations will produce a rejection ratio. We would expect the rejection ratio to increase and approach 0.95 as λ_m deviates from λ_0.
2. *Loss function scoring approaches:* On each run and for each VaR coverage level α, estimate the loss and choose the model that minimizes the given score. Then measure the number of times a model was assigned the lowest score.
3. IGARCH calibration using RMSE: Calculate the root mean squared error (RMSE) between realized and forecast volatility. Then measure the number of times a model was assigned the lowest error.

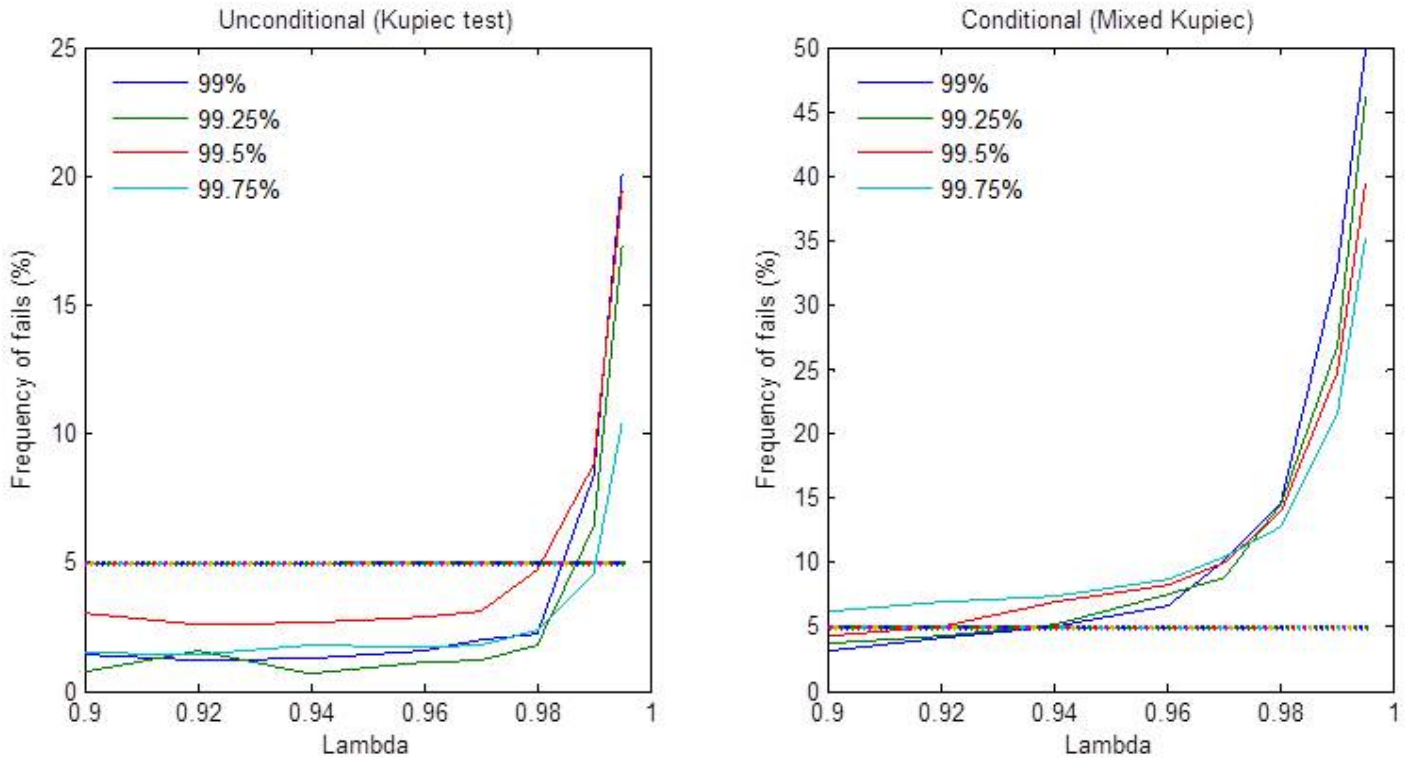
Table of Contents

DGP with decay factor $\lambda_0 = 0.94$



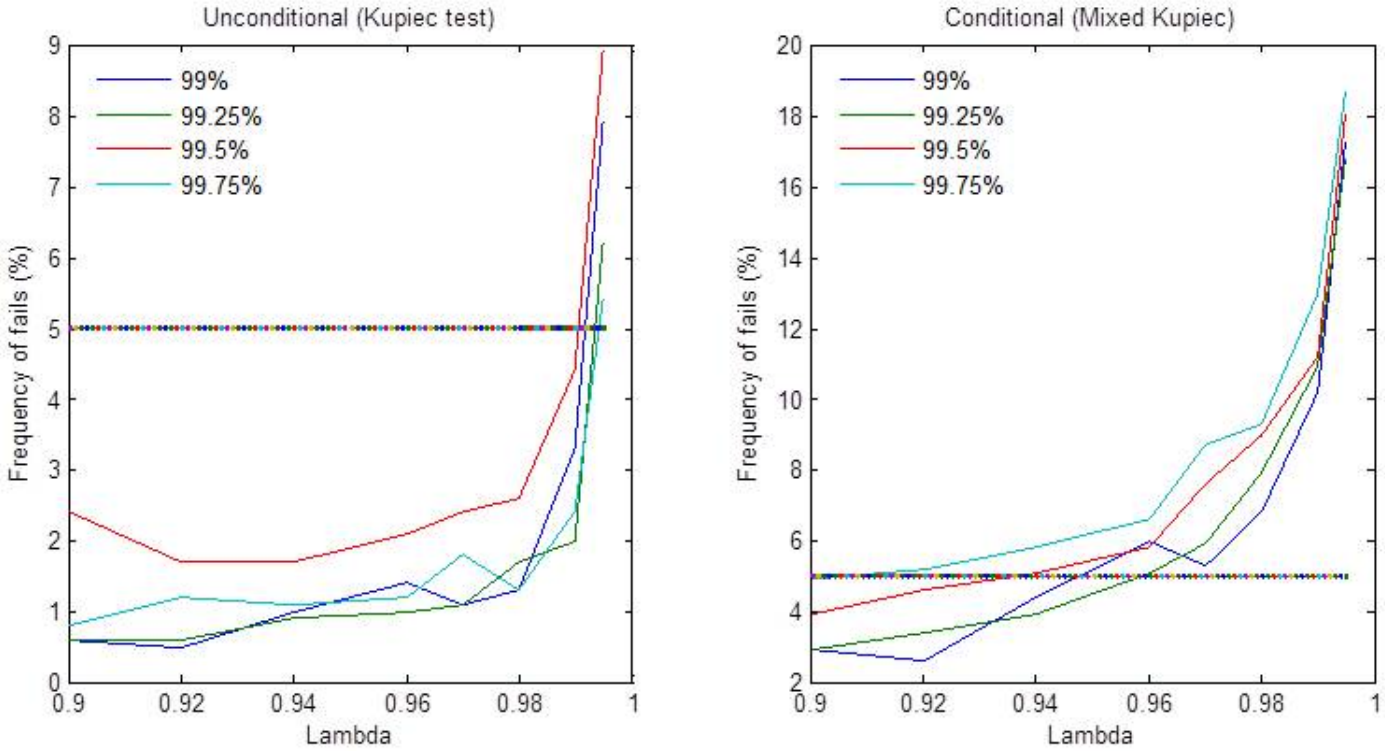
Percentage of rejections for different decay factors λ after 1000 simulations. Tests at 95% confidence.

DGP with decay factor $\lambda_0 = 0.97$



Percentage of rejections for different decay factors λ after 1000 simulations. Tests at 95% confidence.

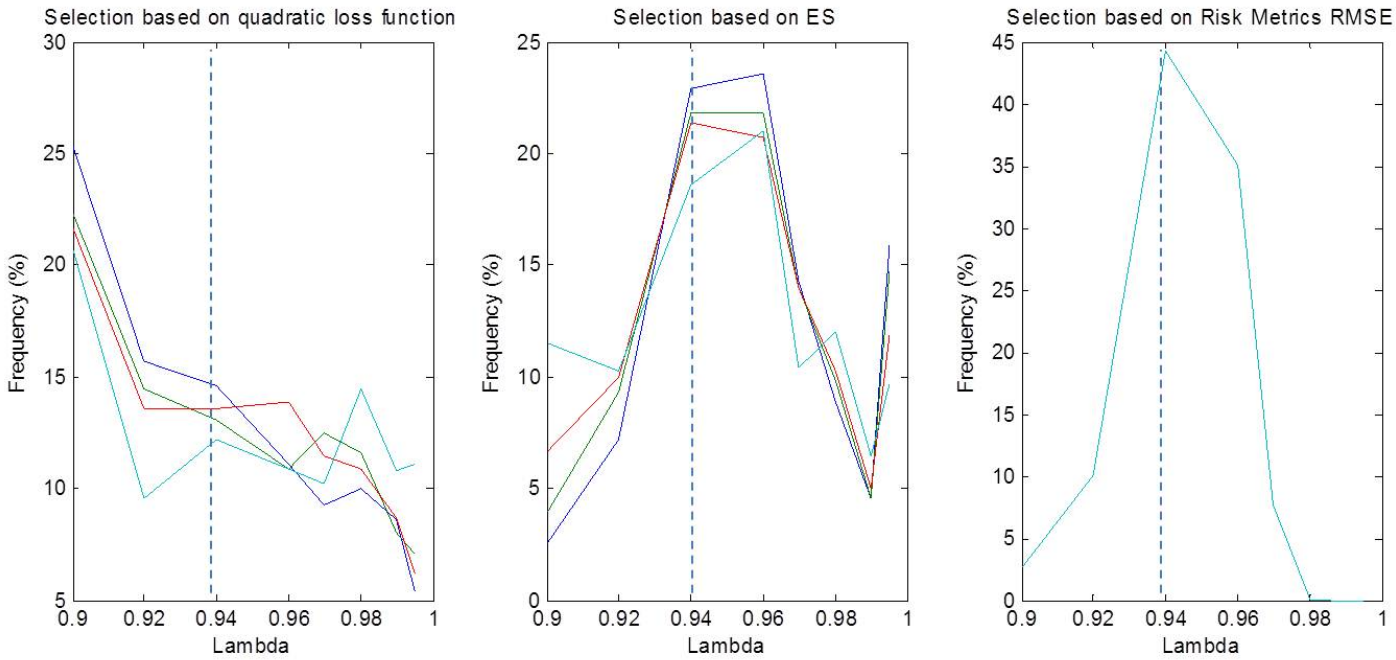
DGP with decay factor $\lambda_0 = 0.99$



Percentage of rejections for different decay factors λ after 1000 simulations. Tests at 95% confidence.

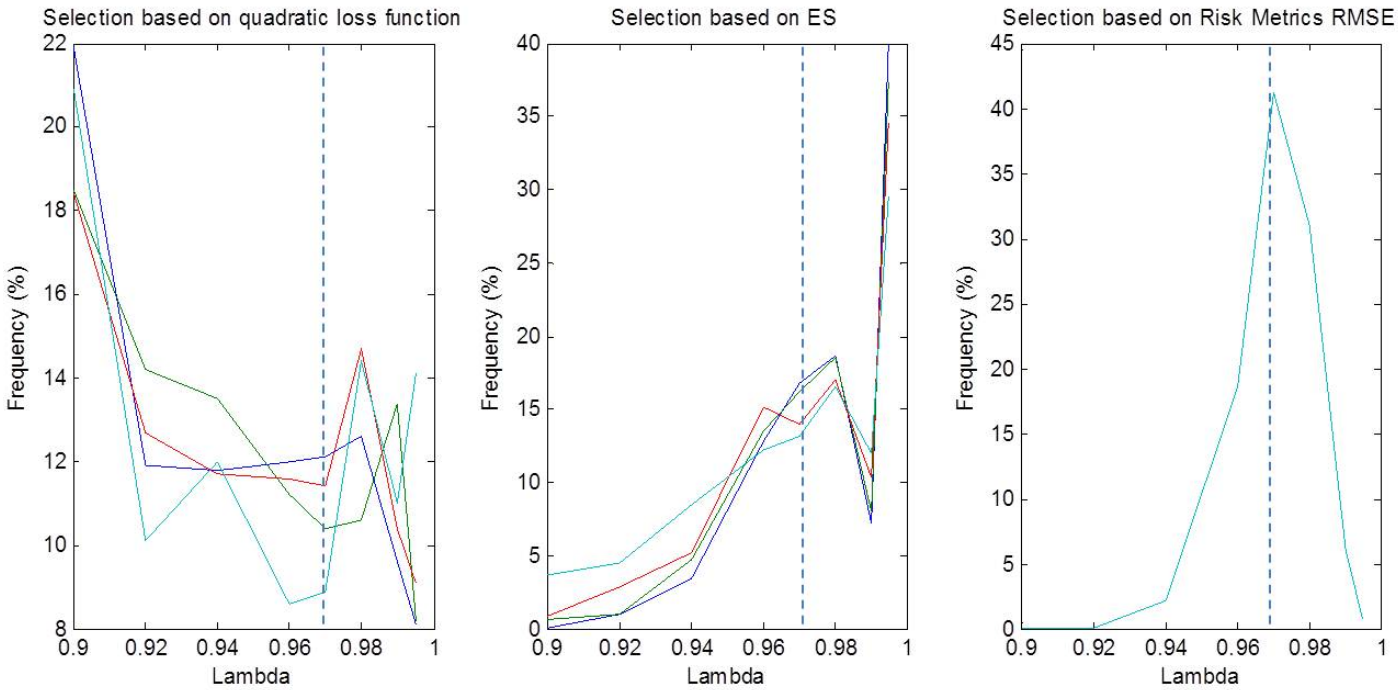
- For $\lambda_m > \lambda_0$ the tests tend to correctly reject the wrong models in more than 5% of the cases, but they systematically fail to reject wrong models when for $\lambda_m < \lambda_0$.
- Since higher decay factors mean more stable EWMA processes, this asymmetry seems to suggest that a model that underreacts to underlying volatility changes will attract more breaches (and hence more rejections) compared with a model which overreacts.
- When considering higher λ_0, the tests provide poorer results suggesting that the power of the tests increases as the underlying process moves away from a constant volatility process.

DGP with decay factor $\lambda_0 = 0.94$



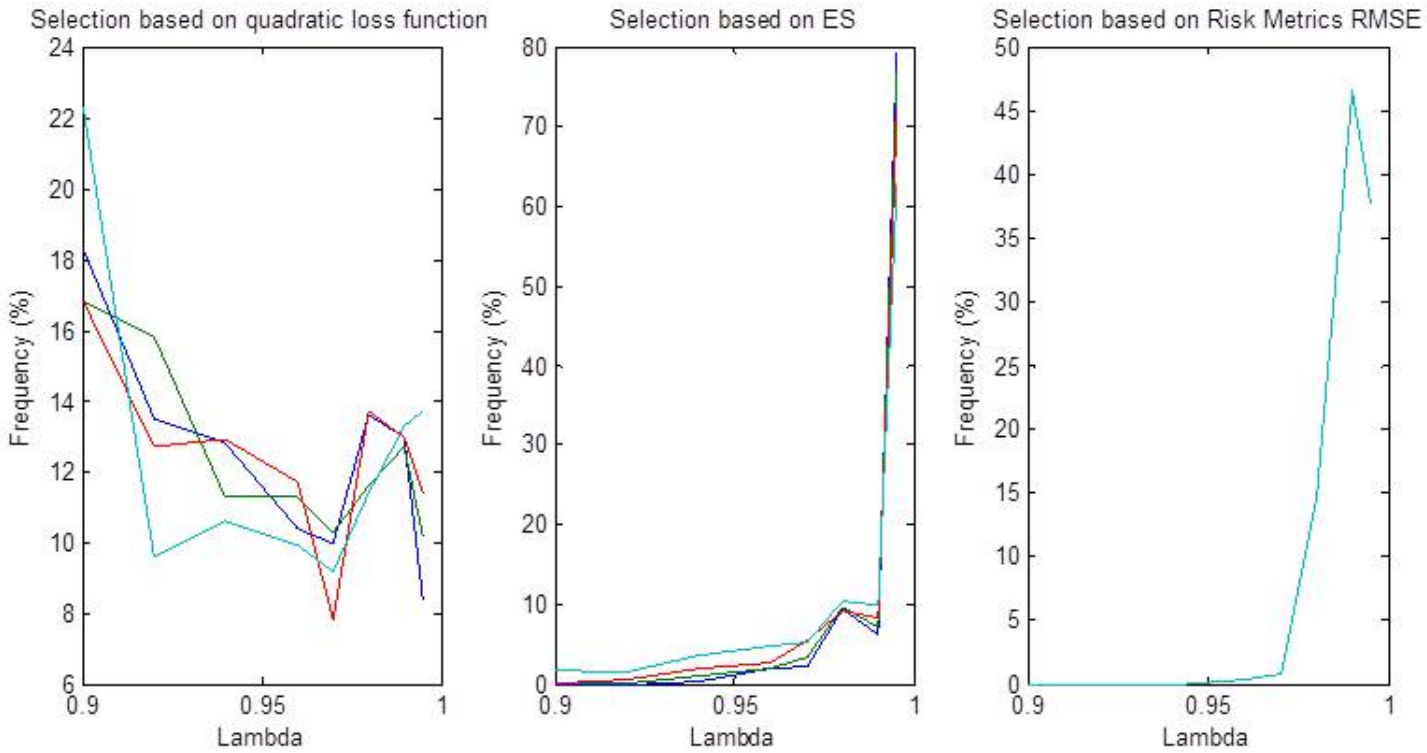
Percentage of time each model was selected in 1000 simulation runs.

DGP with decay factor $\lambda_0 = 0.97$



Percentage of time each model was selected in 1000 simulation runs.

DGP with decay factor $\lambda_0 = 0.99$



Percentage of time each model was selected in 1000 simulation runs.

- When we turn to the loss-function test suggested by [Lopez(1998)], the results do not seem to improve. Models that overrreact to underlying volatility tend to attract higher scores (this is to be expected, as the quadratic term penalizes larger breaches).
- The test based on excess losses from the expected shortfall shows some improvement although results deteriorate for larger decay factors.
- If, instead of backtesting, we aim at minimizing volatility forecast error (RMSE), the calibration is significantly more accurate.

Final remarks

- Modifying the historical sample (whether by introducing artificial stresses, by rescaling volatility, or both) poses additional challenges to the correct calibration/validation of initial margin models.
- There is strong case for *not* introducing a stressed component into the margin calculation even if it leads to more conservative results: it is preferable to introduce alterations at the level of the risk measure.
- Backtesting based on exception counting tends to favour overreacting FHS calibrations, which is an undesirable outcome in terms of the procyclicality.

Final remarks (cont'd)

- Backtesting seems to be inadequate to calibrate an FHS VaR model, even when incorporating the size of the losses.
- In contrast, when calibration is based on minimizing the forecasting errors, results are significantly more accurate. This underscores the importance for FHS validation to take into account the model's response to the underlying dynamics.
- Focussing on the question *"Is the model producing the right percentile?"* is insufficient when dealing with models that aim at responding to the underlying dynamics (and not only at forecasting a distribution). A FHS validation/calibration framework should also consider the model's dynamic response.


Barone-Adesi, G., Bourgoin,F., Giannopoulos,K. (1998), Don't look back. *Risk*, August.


Campbell, S. (2005), A Review of Backtesting and Backtesting Procedure, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington D.C.(2005)


Christoffersen, P.F.(1998). Evaluating Interval Forecasts. *International Economic Re- view*, 39, 841-862.


Davis, M. (2014). Consistency of risk measure estimates, Working Paper, Imperial College, 2014


Dowd, K.(2005). Measuring market risk, 2nd Ed. Wiley Finance, London.


Gregory, J. (2014), Central Counterparties. Mandatory clearing and bilateral margin requirements for OTC derivatives. Wiley.


Gurrola, P., Murphy, D.(2015). Filtered historical simulation Value-at-Risk models and their competitors . *Bank of England, Working Paper Series*, No. 525.


Haas, M. (2001), New Methods in Backtesting, Financial Engineering, Research Center Caesar, Bonn.


Hull J., and White A. (1998). Incorporating Volatility Updating into the Historical Simulation Method for Value at Risk. Journal of Risk (Fall); Vol. 1, No. 1; Pages: 5-19.


Engle, R.F. and T. Bollerslev (1986), Modeling the Persistence of Conditional Variances, *Econometric Reviews*, 5, 1-50.


Kupiec, P. (1994). The performance of S&P 500 futures margins under the SPAN margining system, *The Journal of Futures Markets*, Vol 14, No 7, 789-811.


Kupiec, P. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models, *The Journal of Derivatives*, 3, 73-84.


Lopez, J. (1998), Methods for Evaluating Value-at-Risk Estimates, *FRBNY Economic Policy Review*, October 1998, 119-64.


Lopez, J. (1999), Regulatory Evaluation of Value-at-Risk Models, *Journal of Risk* 1, 37-64.


Murphy, D., Vasios, M., Vause, N. (2015), A comparative analysis of tools to limit the procyclicality of initial margin requirements, paper presented at the International Conference on Payments and Settlement, Deutsche Bundesbank, Eltville, Germany, September.


Nelson, D.B. (1990), Stationarity and Persistence in the GARCH(1,1) Model, *Econometric Theory*, 6, 318-334.


Zangari, P.,(1996), Estimation and Forecast, RiskMetrics Technical Document, 4th Edition, J.P. Morgan, New York.