BANK OF ENGLAND

Working Paper No. 525

Filtered historical simulation Value-at-Risk models and their competitors

Pedro Gurrola-Perez and David Murphy

March 2015

Working papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee or Financial Policy Committee.

BANK OF ENGLAND

Working Paper No. 525

Filtered historical simulation Value-at-Risk models and their competitors

Pedro Gurrola-Perez[1] and David Murphy[2]

Abstract

Financial institutions have for many years sought measures which cogently summarise the diverse market risks in portfolios of financial instruments. This quest led institutions to develop Value-at-Risk (VaR) models for their trading portfolios in the 1990s. Subsequently, so-called filtered historical simulation VaR models have become popular tools due to their ability to incorporate information on recent market returns and thus produce risk estimates conditional on them. These estimates are often superior to the unconditional ones produced by the first generation of VaR models. This paper explores the properties of various filtered historical simulation models. We explain how these models are constructed and illustrate their performance, examining in particular how filtering transforms various properties of return distribution. The procyclicality of filtered historical simulation models is also discussed and compared to that of unfiltered VaR. A key consideration in the design of risk management models is whether the model's purpose is simply to estimate some percentile of the return distribution, or whether its aims are broader. We discuss this question and relate it to the design of the model testing framework. Finally, we discuss some recent developments in the filtered historical simulation paradigm and draw some conclusions about the use of models in this tradition for the estimation of initial margin requirements.

Key words: Value-at-Risk, filtered historical simulation, conditional volatility, volatility scaling, risk model backtesting.

JEL classification: C58, G18, G32.

(1) Bank of England. Email: pedro.gurrola-perez@bankofengland.co.uk
(2) Bank of England. Email: david.murphy@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England. We would like to thank the anonymous referee, Evangelos Benos, Michalis Vasios and Nick Vause for comments on prior versions of this paper. This paper was finalised on 12 February 2015.

The Bank of England's working paper series is externally refereed.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

Publications Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 Fax +44 (0)20 7601 3298 email publications@bankofengland.co.uk

© Bank of England 2015
ISSN 1749-9135 (on-line)

Summary

One of the standard approaches for measuring the risk of portfolios of financial instruments is a family of models collectively known as Value-at-Risk or 'VaR'. The most commonly used of the first generation of VaR models provide an unconditional risk measure, while later refinements estimated risk conditional on more recent market conditions. These second generation *filtered historical simulation* or 'FHS' models are the subject of this paper.

We begin by briefly setting out the historical development of VaR models and their use in financial risk measurement. We discuss the FHS approach in detail, showing how a new returns series is constructed in two steps: the first 'devolatilising' returns by dividing by an estimate of volatility on the day of the return; the second 'revolatilising' them by multiplying by an estimate of volatility on the day of the VaR measure. The performance of two models in the FHS family with different devolatilising methods is illustrated. This shows in particular how filtering modifies various properties of the return distribution such as its unconditional volatility, skewness, kurtosis, and autocorrelation. Filtering two return series separately also changes their correlation, as we illustrate. This in turn has consequences for portfolio risk measures, and hence its effects need to be understood by model users and model designers.

We present two ideas of risk measurement: one as a search for a particular risk measure, such as the 99^{th} percentile of the return distribution; the other as a search for a convincing account of the returns generating process which happens, as a side-product, to provide a variety of risk measures. This leads us to discuss the process for testing (and perhaps rejecting) a risk measure based on its performance both in backtesting and in capturing other features of the time series of returns.

A related issue is the calibration of risk models in general and FHS models in particular. We discuss some criteria for finding an optimal calibration, and the necessity of ensuring that models do not drift away from an acceptable calibration over time.

FHS models typically aim to react faster to changes in market conditions than first generation VaR models. A natural consequence of this reactivity is that if these models are used to calculate initial margin requirements (for instance at a central counterparty or by a party in the bilateral over the counter derivatives market), they place an increased liquidity burden on market participants. We analyse this *procyclicality* and illustrate the importance of calibration in this context.

The paper concludes with a discussion of various extensions to the FHS paradigm and some of the implications of this work for the application of FHS models in risk management.

1 Introduction

One of the standard approaches for measuring the risk of portfolios of financial instruments is a family of models collectively known as *Value-at-Risk* or 'VaR' [2, 31]. The most commonly used of the first generation of VaR models provide an unconditional risk measure based on some *window* of historical observations, while later refinements estimated risk conditional on more recent market conditions. These second generation *filtered historical simulation* or 'FHS' models are the subject of this paper. As such, it adds to the growing literature on the selection, construction and testing of VaR models [11, 28, 36, 39, 41, 46].

VaR models are increasingly used to calculate margin requirements on portfolios of financial instruments both between bilateral market participants and by central counterparties [42]. This use poses particular challenges for risk modelling, not least due to the potential for commercial pressure on margin model designers to keep risk estimates low. FHS VaR is an attractive model choice for a margin model in that it will often produce low margin estimates in calm markets and higher margin estimates during periods of elevated volatility. Currently regulatory requirements to post initial margin on both bilateral and cleared derivatives transactions are being phased in [7, 21], making risk-based initial margining ubiquitous across exchange-traded and OTC derivatives markets amongst others. Therefore a focus on the properties of FHS approaches is timely.

We begin by briefly setting out the historical development of VaR models and their use for estimating capital and margin requirements. Section 3 then discusses the filtered historical simulation approach in detail, while section 4 illustrates the performance of two models in the FHS family, showing in particular how filtering modifies various properties of the return distribution. The discussion here is deliberately slower than in most of the literature because we think that it is important for both model builders and model users to understand not just the final risk measure, but also the implications of some of the design decisions made in its construction.

This focus on the (sometimes implicit) assumptions used in FHS models leads in section 5 to a discussion of the aims of VaR modelling, the process for testing (and perhaps rejecting) a risk measure, and the calibration of FHS models.

A natural consequence of the reactivity of FHS models is the liquidity burden the margin requirements they produce might pose on market participants as conditions change, so section 6 investigates this issue. Then we turn to extensions to the FHS paradigm in section 7. Finally, section 8 concludes with some of the implications of this work for margin modelling.

2 A Short History of Value-at-Risk Modelling

VaR models have been deployed by financial institutions in their risk modelling process for some time. This section briefly sets out the key milestones in this use.

2.1 Value-at-Risk for Capital

The first VaR models[1] were used by investment banks to estimate market risk on portfolios of traded assets. At the time, the trading portfolios of large dealers often had return distributions

[1] The precise definition of a VaR model, and hence the first risk measure that deserves to be called 'Value-at-Risk' is contestable. As Holton points out [29], measures that foreshadow VaR can be identified as far back as the 1920s. However it was not until Garbade's work at Bankers Trust in the 1980s [23] that we find measures that a modern risk manager would unequivocally accept as VaR, so we begin our discussion with these proto-modern developments.

which were at least approximately normal most of the time. The maximum loss on these portfolios was not a useful risk measure as it was extremely unlikely. Instead, a measure that was in the (near) tail of the return distribution was used. VaR models answered the question 'how much could the portfolio lose with α% probability over a given time horizon?' Typically α was set at 99 and the holding period at ten days, so the question asked was often:

How much could the portfolio lose with 99% probability over ten days?

Here there was an assumption that the holder could remove all or nearly all of the risk in its portfolio over a ten day period – an assumption that may not have been too inaccurate at the time – and hence the 99% ten day VaR was a reasonable measure of the market risk of the portfolio.

The credibility of this measure received a boost in 1996 when the Basel Committee on Banking Supervision proposed that banks could use VaR to calculate the regulatory capital required for general market risk provided that they met certain standards [5].[2]

2.2 The Architecture of Value-at-Risk Models

A VaR model is constructed in the following steps:

- A set of *risk factors* which collectively drive changes in value of the portfolios we wish to analyse are identified. These might for instance be equity and commodity prices, FX rates, bond yields, swap rates, and various implied volatilities. A history of these risk factors is assembled over some *data window*.
- Some model of the joint returns of the risk factors over the data window is constructed.[3]
- Models of the sensitivity of each instrument in the portfolio to each risk factor are chosen. Thus for instance we might here answer the question 'if USD/JPY goes up 1% what is the change in value of the portfolio?'
- The model of the risk factors is used to construct a set of risk factor changes over the chosen time horizon.
- These risk factor changes are fed into the sensitivity models to estimate the change in portfolio value that would be expected were they to occur.
- These changes in value are gathered into a distribution, and the desired confidence interval is applied. Thus for instance we might pick the 10th worst loss out of 1,000 changes in value as the 99% VaR.

There are two divisions in the VaR model family:

- *Parametric* VaR models assume a particular functional form for the joint risk factor return distribution; while
- *Simulation* models do not, instead relying on historical data as a basis for constructing this distribution.

[2]Specifically, the capital requirement was set at three times the 99% ten day VaR for general market risk. The capital rules were extended to allow the use of VaR for calculating the capital requirement for specific risk in 1997.

[3]This 'model' might be the implicit one of assuming that future returns will be well characterised by some past sample.

In the first generation of VaR models in the late 1990s, examples from both families were constructed. RiskMetrics [32], for instance, was a popular parametric VaR approach based on the multivariate normal distribution; competing with this were *historical simulation* or 'HS' approaches which simply used some history of changes of the selected risk factors. A typical large bank's HS model from this period:

- Selected a set $j \in 1 \ldots J$ of risk factors, and gathered historical data of daily changes x_j in those risk factors over, perhaps, the last four years, so that the bank would have the daily change in each risk factor j on each day $i \in 1 \ldots 1000$, $x_j(i)$;
- Revalued the portfolio assuming that each day i's set of risk factor returns actually occurred; then
- Sorted the resulting 1,000 changes in portfolio value, and returned the 10th worst as the 99% 1 day VaR.

This measure was then typically scaled up to obtain a ten day VaR if needed for regulatory purposes.[4]

HS models like this were relatively simple to construct and understand, and they quickly gained popularity over parametric approaches. One advantage they had was that they made no assumptions about the tail of the joint risk factor return distribution. Moreover, provided that a period of stress had occurred in the data window, the model would incorporate it into its risk estimate.

These models were far from perfect, and an extensive literature criticising and comparing them and proposing modifications quickly developed [3, 13, 28, 35, 40]. One of the principal criticisms of the early HS models concerned their treatment of volatility clustering, so we address this next.

2.3 Conditional Volatility

There is strong empirical evidence that many financial risk factor returns are not well-described by processes with constant volatility [37, 45]: indeed, this has been known for over five decades [38]. This phenomenon means that risk factor changes are not necessarily independent over time. It also means that current market conditions contain some information about returns in the immediate future: if conditional volatility is elevated, then typically larger returns are to be expected than if they are not.

One way that the first generation of historical simulation models could respond to this challenge was to shorten their data windows to be more responsive to current conditions. This in turn reduced their accuracy, as estimating the tail of a return distribution is harder the less data you have.[5]

Relatively simple techniques to address these issues were tried by some banks, such as using the maximum of the VaR estimates from models with 100 day and 2,500 day data windows,

[4] Many models simply use the *square root of time* law [31] to scale up from a one day to a ten day holding period. In this approach one simply multiplies the one day VaR by $\sqrt{10}$. Various alternative methods are in use too. Some make direct use of ten day risk factor changes: however these tend to suffer either from data availability issues (as a data series ten times longer is needed if non-overlapping ten day periods are used), or from sampling issues (due to problems with the accuracy of estimates obtained from overlapping periods). Others fit a more nuanced model to the daily returns data, such as one of the extreme-value-theoretic models, then use either fully [18] or semi parametric [15] approaches to estimate the time-scaling factor. A longer account of the issues here can be found in [19] and the references therein.

[5] Shortening the data window of a HS VaR model will also make it more procyclical (in both peak-to-trough and n-day senses [43]) all other things being equal.

but these were not entirely satisfactory[6]. A more sophisticated solution is presented in the next section.

2.4 The Filtered Historical Simulation Concept

The question implicitly posed in the last section is:

> *How can we use the information in recent returns to estimate the current level of risk more accurately?*

A family of models which provide an answer to this question are the *scaled* or *filtered* historical simulation Value-at-Risk models. Broadly, these models scale the historical data based on some estimate of current conditions, so that if current conditions are less volatile then they 'damp down' the risk estimate, while if conditions are more volatile, they 'turn them up'. Models like this have demonstrated improved risk estimates compared to first generation VaR [3, 30] (although see also [46]), and hence thus have become widely used by financial institutions.

We examine the scaling process of these second generation models in more detail in the next section. First, though, we turn to the use of VaR models to calculate margin requirements on portfolios of derivatives.

2.5 Terminology

We use the term 'HS VaR' or 'historical simulation VaR' to refer to a first generation model, occasionally calling it 'unscaled' to contrast with the second generation 'scaled' or 'filtered historical simulation' models. The latter two terms are used interchangeably.

2.6 Value-at-Risk for Margin

The problem of calculating an initial margin requirement for a portfolio of financial instruments is in some ways similar to the problem of calculating a capital requirement: here too we want to estimate a high percentile of the loss distribution over some holding period. Market participants therefore turned to Value-at-Risk models for margin calculations. Indeed, the use of such models had become 'best practice' for prime brokers by 2007 [26]. Around the same time, some central counterparties began to use VaR models to determine initial margin requirements for cleared derivatives [42]. There are some key differences between the two applications however:

- Capital is the ultimate backstop, so it should cover possible losses to a high degree of confidence[7]: in contrast there are usually resources available beyond margin, so risk beyond margin is more acceptable than risk beyond capital.[8]
- Margin has to be funded. Moreover, margin calls are often made daily with the requirement that they are met the next morning. Thus increases in margin can create funding liquidity risk. We discuss margin increases in times of stress – *margin procyclicality* – further in section 6 below, noting here only that this liquidity impact is of nugatory concern for capital.

[6]One problem is that no first generation HS model, whatever its confidence interval, can ever provide a risk estimate that is higher than the loss on the worst day in the data window. This means that it cannot react as stress intensifies beyond the worst conditions it has 'seen'. Another problem is that long data windows require either that all risk factors have a long history – something that may not be the case – or that some necessarily arbitrary 'fill' technique is used to hypothesise risk factor returns where they are not available.

[7]This is one motivation for using expected shortfall rather than VaR as the basis for capital requirements [8].

[8]These include capital to cover residual risks for a bank, and both capital and default fund for a CCP.

- Capital requirements apply to a firm's whole portfolio, with diversification benefits often being given between different parts of it. This makes them more stable than the margin requirements on the portfolio of financial instruments a firm might have with a single bilateral counterparty or with a CCP.
- Hence, while the procyclicality of margin and of capital requirements for banks' loan books are issues [47], the procyclicality of market risk capital requirements is less of a concern.[9]
- Margin methodologies have to work for small portfolios as well as large ones, and highly directional portfolios as well as hedged ones. In contrast the key systemic issue for market risk capital is that it be sufficient for large bank trading books *in toto*.

3 Filtered Historical Simulation Value-at-Risk

We noted in the last section that VaR models which scale their risk estimates based on current conditions have become popular. Typically these models use historical returns, but multiply them by some *filter* or *scaling factor* based on current conditions. The model described by Boudoukh et al. [10] is an early example of this technique, and the FHS model proposed by Hull and White [30] soon afterwards was also influential. Various elaborations of these filtered historical simulation Value-at-Risk models have also been proposed: see for instance [3, 4, 13, 40]. In this section we describe the mechanics of FHS models and give two examples of members of the FHS family.

3.1 Historical Simulation Value-at-Risk

We fix notation by summarising the construction of a first generation HS VaR model. Suppose we have some set of risk factors $j \in 1 \dots J$, and a time series of the returns of each risk factor $x_j(i)$ for N days, $i \in 1 \dots N$.

An unfiltered historical simulation model assumes that the portfolio in question experiences each day i's returns tomorrow. Each of these possible repetitions of prior days generates a profit or loss ('P/L'). These P/Ls are then gathered into a distribution, and the desired percentile of this distribution is estimated. One key ingredient, then, is the time series of risk returns used to revalue the portfolio.

3.2 Filtered Historical Simulation: 'Devol' and 'Revol'

The question posed in section 2.4 can now be rephrased as

How can we use the prior returns $x_j(i)$, $i < t$ to improve our estimate of the day t VaR?

FHS models answer this question by identifying the volatilities of returns as a crucial property. For each day i, they calculate some estimate of the volatility of risk factor j on that day, $\sigma_j(i)$. All the historical returns in the data window are then *devolatilised* by dividing them by the relevant volatility estimate: this is termed the 'devol' process for short. The resulting series of devol'd returns are known as the *residuals*.[10]

[9] It should also be noted that the introduction of stressed VaR by the Basel Committee in 2009 [6] reduced the procyclicality of market risk capital requirements signifcantly.

[10] As Jorion [31] points out, any model that makes a prediction of conditional volatility can be used to devolatise returns, and thus (try to) handle non-stationarity of the returns process. We do not discuss these more general filtered simulation models here, concentrating instead on models where the volatility estimate is based on the historical returns in some data window.

FHS models then calculate the VaR for day t by scaling the residuals up by a current estimate of the volatility of each risk factor. That is, each residual $x_j(i)/\sigma_j(i)$ is scaled by a revolatilising or 'revol' volatility $\sigma_j(t)$.

3.3 Filtered Historical Simulation: A Very Simple Example

We illustrate this by giving an example of a very simple FHS model. This model makes two choices:

- It devols using an unconditional (i.e. long window) unweighted historical volatility; and
- It revols using an unweighted volatility estimated using a short window backwards from the date at which VaR is being calculated.

We write (for reasons that will become clear shortly), ${}^{1}_{N}\sigma_j$ for the unconditional devol volatility across the whole data period and ${}^{1}_{n}\sigma_j$ for the n-day current revol volatility, with $n \ll N$.

Our example model then defines a set of filtered returns to be used for a VaR calculation at time t by defining a new set of returns x^{UV} by

$$x_j^{UV}(i) = x_j(i)\frac{{}^{1}_{n}\sigma_j(t)}{{}^{1}_{N}\sigma_j}$$

The revol volatility ${}^{1}_{n}\sigma_j(t)$ is measured backwards from t, i.e. over the n returns $t-n+1, t-n+2, \ldots, t$.

The filtered returns x^{UV} can now be used in a historical simulation Value-at-Risk model in the usual way. Since this model uses an unconditional volatility to devol, we will call it the 'UV' model: this explains our notation.

It is worth unpicking the definition of x^{UV} a little. If the current period is less volatile than average for risk factor j, i.e. ${}^{1}_{n}\sigma_j(t) < {}^{1}_{N}\sigma_j$, then as expected we damp down the historical returns; while if it is more volatile, we 'turn up' the amplitude of returns.

3.4 A More Usual Filtered Historical Simulation Model

The UV scaling model described in the previous section, while simple, is not the first one proposed in the literature. Instead, the first papers [10, 30] suggested that a short term devol volatility is calculated for each day i's returns. This could be an equally weighted volatility estimate over n days, as above, or an exponentially weighted moving average ('EWMA') with some decay factor λ. The first papers suggest the latter. We write ${}^{\lambda}_{n}\sigma_j(i)$for this volatility estimate, recovering the equally weighted case when $\lambda = 1$.

We can then define a new model by making the following choices:

- Devol using the current volatility estimate for the return in question; and
- Revol using the volatility estimate for the day we are calculating value-at risk.

This defines a series of returns x^{CV} which we can use to calculate day t's VaR

$$x_j^{CV}(i) = x_j(i)\frac{{}^{\lambda}_{n}\sigma_j(t)}{{}^{\lambda}_{n}\sigma_j(i)}$$

The historical simulation VaR model based on the filtered returns x^{CV} employs the short term or conditional volatility to revol, hence the notation.[11]

[11] Clearly if λ is very close to 1 and n is large, ${}^{\lambda}_{n}\sigma_j$ is not a short term volatility estimate. More commonly though λ is in the range 0.99 to 0.94, justifying our terminology.

The well-known Hull and White model [30] is a particular case of this situation. It can readily be seen that CV models like Hull and White's attempt to improve upon the accuracy of ordinary historical simulation models by taking account of the volatility changes experienced during the data window. Equivalently, they assume that the distribution of residuals scaled by current volatility ${}^{\lambda}_{n}\sigma(i)$ is stationary (in the sense of having constant variance), an assumption we return to in section 4.5.

3.5 The Filter Factors

For a given day t's VaR, the UV model uses the same factor,

$$\frac{{}^{1}_{n}\sigma_j(t)}{{}^{1}_{N}\sigma_j}$$

to scale all of the returns in risk factor j for day t's VaR. The fact that the filter factor used in the UV model for a given day's VaR is constant means that the return series used to derive the final risk estimate have the same skewness and kurtosis as the unfiltered returns, and the same correlations with each other. Any drift (non-zero mean) in the returns is scaled by the factor. Moreover this is true for any model that uses a single devol volatility independent of the day i being scaled.

The CV model in contrast has a separate filter factor for each day i's return, viz.

$$\frac{{}^{\lambda}_{n}\sigma_j(t)}{{}^{\lambda}_{n}\sigma_j(i)}$$

This definition leads to different properties:

- The estimate of volatility ${}^{\lambda}_{n}\sigma_j$ is necessarily less precise than ${}^{1}_{N}\sigma_j$ if the data is stationary, as it is based on less information and hence will have more sample error. Moreover the short term filter factor will have larger swings than the long term one. The shorter term the volatility estimate is – the smaller λ is – the more pronounced this effect.
- Short term volatility scaling uses a different filter factor for each data point, so in general this approach will not preserve higher moments or covariances.
- Short term scaling converges to one, in the sense that the factor $\frac{{}^{\lambda}_{n}\sigma_j(t)}{{}^{\lambda}_{n}\sigma_j(i)}$ tends to 1 as i approaches t. This is not true for the long term approach, where even yesterday's data is scaled if ${}^{1}_{n}\sigma_j(t)$ is substantially different from ${}^{1}_{N}\sigma_j$.

Finally, we note that there are a range of FHS VaR models which lie between those with full scaling and the first generation models with no scaling. For instance, for any model which uses a scaling factor f, there are a range of models which scale proportionately less defined by setting the scaling factor equal to

$$\frac{1}{K+1}(K+f)$$

for $K \geq 0$. Thus for instance if we take $K = 1$, the filter factor is half way between f and 1.

3.6 General FHS Models

The notation in the prior sections suggests another dimension by which the family of FHS models can be classified. For a given set of returns, a member of the family can be specified by stating for each series j

- Which decay factor λ and time period P is used to define the devol volatility ${}^{\lambda}_{P}\sigma_j(t)$; and
- Which decay factor μ and time period Q is used to define the revol volatility ${}^{\mu}_{Q}\sigma_j(t)$.

The filtered returns used for the VaR on day t for this family member F are then

$$x_j^F(i) = x_j(i)\frac{{}^{\mu}_{Q}\sigma_j(i)}{{}^{\lambda}_{P}\sigma_j(t)}$$

In practice, we often find the same choice $\lambda = \mu$ (and $P = Q$) made for both devol and revol operations across all the data series, but this is not necessary.

4 Exploring Filtered Historical Simulation

Filtered historical simulation approaches were first introduced to attempt to address the problem that percentiles in the tail of observed return distributions were neither well characterised by simple functional forms such as the normal distribution nor sometimes by the returns in the models' data windows. As Pritsker puts it [46], (unfiltered) historical simulation models are "under-responsive to changes in conditional volatility". If conditional volatility changes, then we need to update our Value-at-Risk estimates based on this change. The simplest way to do this is to estimate the new conditional volatility and use this to scale the returns used for VaR calculation: this is the essence of the FHS idea. However, this approach is not without issues so in this section we explore some of the properties of the filtering process.

4.1 Illustration

The properties of our VaR models will be illustrated using data from the energy markets. There is no particular reason for this choice: energy market returns share many of the same features as returns from equity, FX, interest rate and credit markets, so we expect a broad read-across to these areas.

Specifically, we have taken 1,500 daily returns from three liquid futures contracts, Gasoil; UK natural gas ('Natgas'); and West Texas Intermediate ('WTI'), an important oil contract, all starting in September 2007. Risk factor returns were calculated from the times series of front month futures prices in the usual way.

We examine two members of the FHS family:

STV Here we use an EWMA volatility estimate with decay factor 0.97, ${}^{0.97}_{500}\sigma_j$, for both devol and revol steps with a 500 day data window;

LTV Here we use the same $\lambda = 0.97$ volatility to revol, but 500 day unweighted volatility ${}^{1}_{500}\sigma_j$ to devol. This model (like the UV models discussed above) uses the same filter factor for all the returns used to calculate a given day's VaR.

The first 500 days are used to calibrate the unweighted long term volatility estimates ${}^{1}_{500}\sigma_j$ and the EWMA volatilities[12]. We then calculate the filter factors and the VaR. Figure 2 illustrates the filter factor applied to the returns for WTI on the day shown on one axis when calculating the VaR for the day on another in each of the two models.

[12]That is, we begin the EWMA calculation on day 2 seeded at the long term average volatility ${}^{1}_{500}\sigma_j$. The 'run up' period of 500 days is long enough for this decay factor that the seed has an immaterial effect on the results.

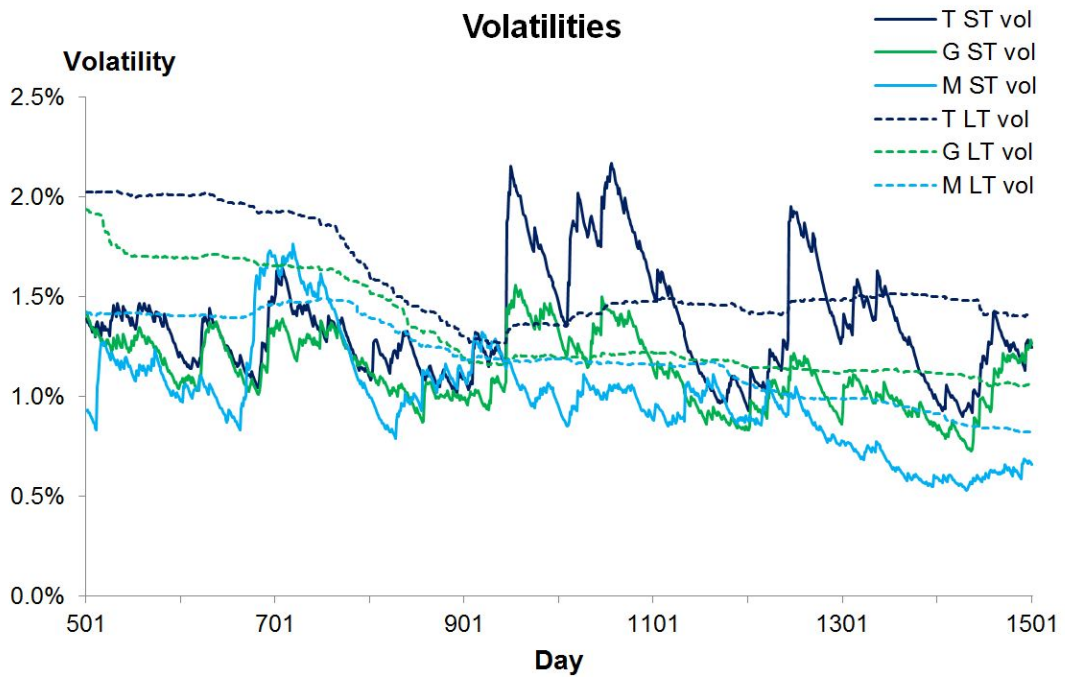

Figure 1: An illustration of short– and long term volatility estimates for the three data series made by the STV and LTV models for WTI ('T'), Gasoil ('G') and UK Natural Gas ('M')

4.2 Volatility Estimates

The short– and unweighted long-term volatilities $\frac{0.97}{500}\sigma_j(i)$ and $\frac{1}{500}\sigma_j$ of our chosen risk factors are illustrated in figure 1, while figures 2 and 3 show the filter factors for the STV and LTV models respectively.

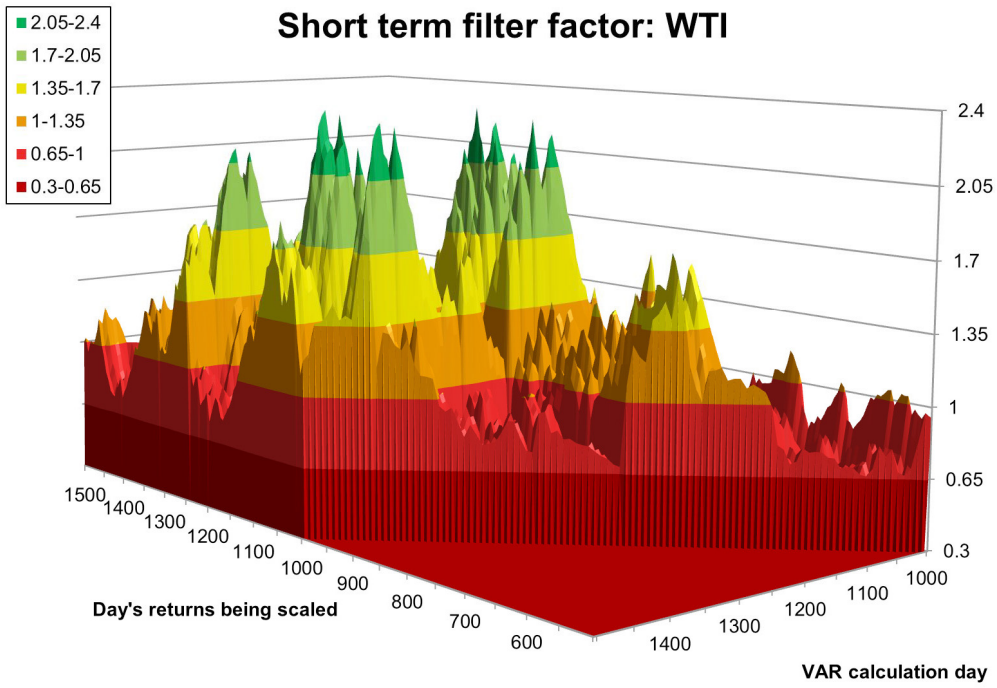

Figure 2: An illustration of short term filter factors for WTI

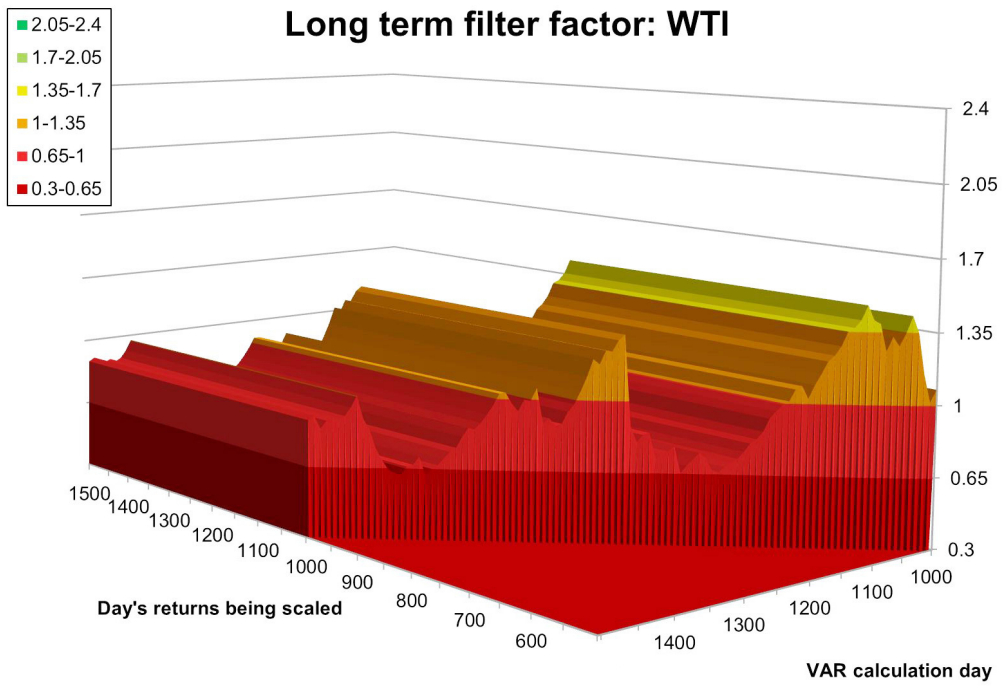

Figure 3: An illustration of the long term filter factors for WTI

Figure 1 explains the gross features of figure 2. For instance, $\frac{0.97}{500}\sigma_{WTI}(i)$ (the solid blue line) is mostly below $\frac{1}{500}\sigma_{WTI}(i)$ (the dotted blue line) on days 1,000 to 1,100, so the short term filter factor for WTI $\frac{\frac{0.97}{500}\sigma_{WTI}(i)}{\frac{1}{500}\sigma_{WTI}(i)}$ is less than one for these days. Turning to the right rear of figure 2, this is indeed the case.

The EWMA volatilities also vary through time significantly: this is reflected in a short term filter factor which varies substantially. For instance, for WTI (the solid dark blue line), the one day EWMA volatility is below 1.2% on day 940 but over 2.1% by day 950.

The conditional volatility trends in Figure 1 feed through into the Value-at-Risk estimates for the filtered models. Figure 4 illustrates this for a portfolio designed to have low variance: it is long Natural gas and Gasoil futures and short WTI. Thus for instance the spikes in EWMA volatility for WTI from days 900 to 1200 (shown in dark blue in figure 1) feed through into spikes in both FHS VaRs in the same period (shown in purple and pink in figure 4). It can also be seen that the VaR is relatively insensitive to the choice of the 'devol' volatility as LT and ST VaRs tend to follow broadly the same path.

4.3 The Impact of Scaling on Skewness and Kurtosis

In the case of a process $x_j(i)$ which follows a normal distribution $N(0,\sigma)$, the LTV filtered returns will also be normally distributed, with standard deviation equal to the most recent updated estimate $\sigma_j(t)$. The percentiles of the distribution will be rescaled accordingly. For example, if $f \times \sigma$ defined the α% percentile in the original distribution, then $f \times \sigma(t)$ will correspond to the same percentile in the distribution of the rescaled variable.

However, the impact of scaling is more complicated for non-normal returns with a variable filter factor. Insight into this can be obtained by comparing the descriptive statistics of the untreated data to the scaled.

Figure 5 reports the case of the scaled data used by the LTV model for day 1,500's VaR, i.e. with a constant filter factor. Here, as expected, the skewness and kurtosis of the scaled

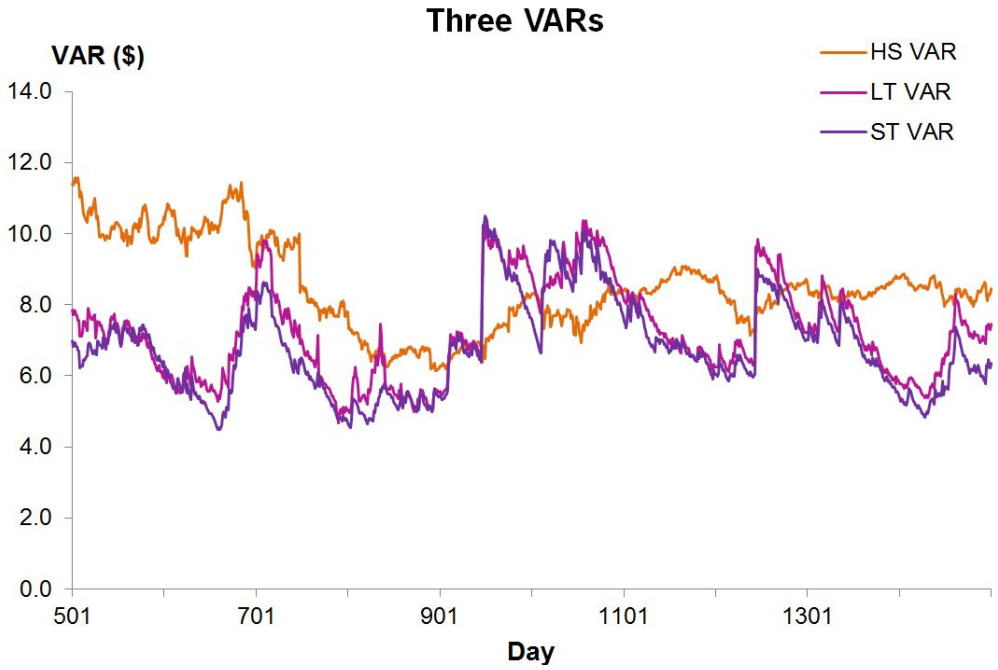

Figure 4: An illustration of ordinary, short– and long term filtered VaR estimates for a low variance portfolio

	Unfiltered Data			Long Term Scaled Data		
	WTI	Gasoil	Natgas	WTI	Gasoil	Natgas
Mean				-1.29E-4	-2.8E04	-1.12E-4
Mean × factor	-1.29E-4	-2.80E04	-1.12E-4			
SD				0.0124	0.0126	0.0066
SD × factor	0.0124	0.0126	0.0066			
Skewness	-0.215	-0.264	0.181	-0.215	-0.264	0.181
Kurtosis	3.56	0.971	1.27	3.56	0.971	1.27

Figure 5: The first four moments of the unfiltered and long term filtered risk factor returns

returns are identical to that of the originals and the means and volatilites are scaled by the appropriate factor.

The picture is quite different for the short term scaling approach with its variable filter factor, as figure 6 (again for the returns used for day 1,500's VaR) reports. Here skewness and kurtosis are not preserved, and there are no simple relationships between the means and standard deviations of the scaled data and those of the unfiltered returns. The scaling process has created a data series that reacts to short term changes in volatility, but it has done so at the cost of transforming the return distribution in an opaque way.

In general rescaling will not preserve the relation between the standard deviation and the α percentiles of the distribution. Moreover the effects tend to be greater as the decay parameter decreases.

We can gain more insight into this by examining rolling data windows. Figure 7 shows the skewness and kurtosis observed in 1,000 rolling samples from the Gasoil returns, with 500 observations each, with and without rescaling. As with the historic data, we see that scaling changes the higher moments, and scaling with a smaller decay factor produces larger changes.

	Unfiltered Data			Short Term Scaled Data		
	WTI	Gasoil	Natgas	WTI	Gasoil	Natgas
Mean × factor	-1.29E-4	-2.80E04	-1.12E-4			
SD				0.0185	0.0130	0.0099
SD × factor	0.0124	0.0126	0.0066			
Skewness	-0.215	-0.264	0.181	-0.053	-0.244	0.078
Kurtosis	3.56	0.971	1.27	7.35	3.04	2.70

Figure 6: The first four moments of the unfiltered and short term filtered risk factor returns

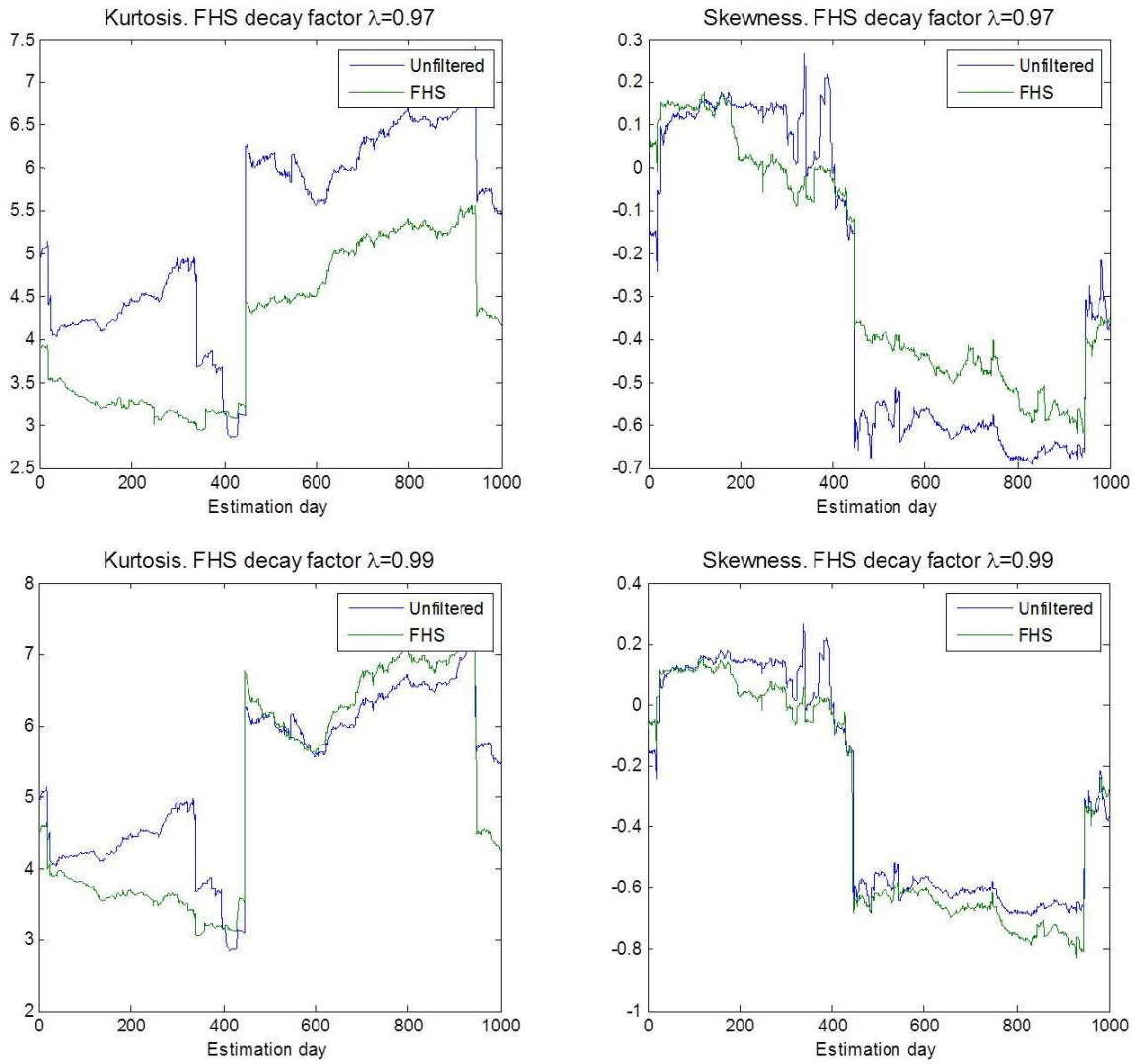

Figure 7: An illustration of the effect of volatility scaling on kurtosis and skewness estimates when using two different FHS decay factors $\lambda = 0.97$ (above) and 0.99 (below).

4.4 The Impact of Scaling on Autocorrelation

In the light of the previous discussion, it is also worth investigating whether the autocorrelation structure of the returns is also affected by the filtering process. For example, if a series has zero autocorrelation, then we would expect this condition to be preserved under rescaling. If the rescaled samples show autocorrelation but we assumed i.i.d. sampling, then the assumption would have to be revisited.

The empirical results seem to confirm that the rescaling process has a statistically significant impact on the autocorrelation structure. For example, when applying runs tests to 500 day series of WTI, Gasoil and UK Natural Gas returns,[13] we observe significant differences in the autocorrelation patterns. Figure 8 illustrates these differences by showing a vertical line at the points in which the results of the autocorrelation tests for unscaled and scaled samples were significantly different. In the case of UK Natural Gas, for example, the rescaled sample tends to eliminate most of the autocorrelation observed in the original series, an effect that could potentially translate into lower accuracy when backtesting the model.

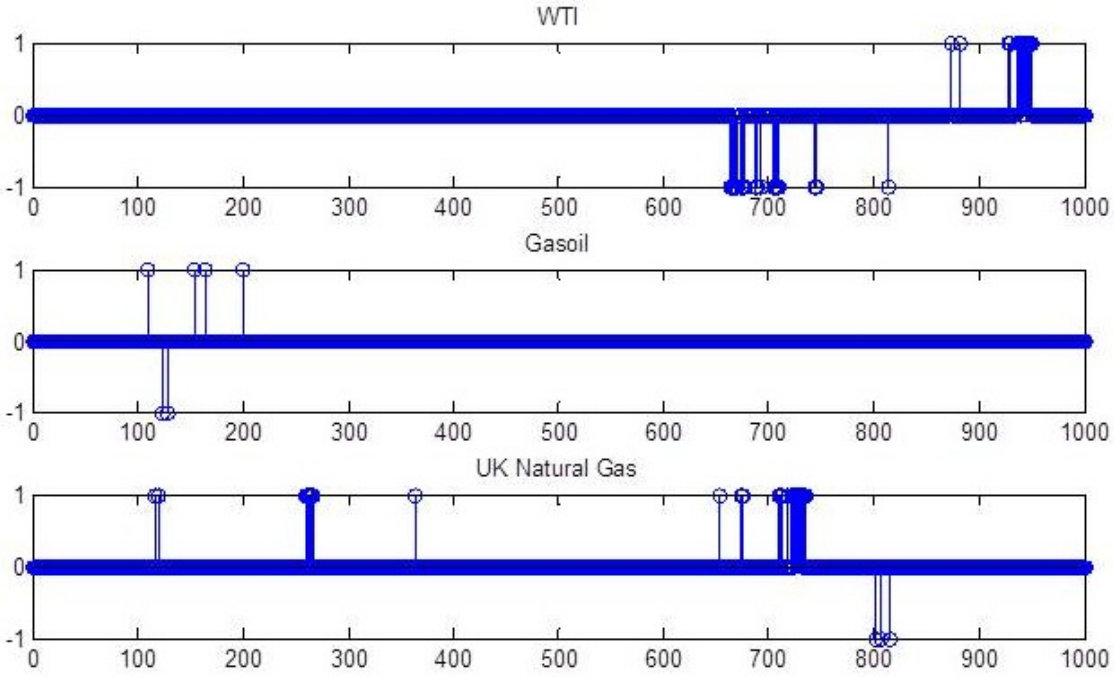

Figure 8: An illustration of the effect of volatility scaling on the autocorrelation of 1,000 rolling samples (of 500 days each). The plots show the difference between the run test results for unscaled and scaled samples. A value of 1 indicates that the null hypothesis of no autocorrelation is rejected at the 5% significance level for the unscaled sample but not for the the scaled one. A value of -1 indicates that the opposite holds, while 0 indicates coincidence of the test results.

4.5 The Tails of the Distribution of Filtered Returns

We are interested in calculating VaR, so the question of how the tails of the return distribution are affected by filtering is particularly pertinent. Recall that it had been argued:

[13]The runs test is based on counting the number of runs of consecutive values above or below the mean to test the null hypothesis that the values come in random order, against the alternative that they do not.

- There is information on previous near-tail events in a long term historical return series;
- However, historical returns from long ago may not be representative of conditions today;
- So we can keep some of the insights of the past while making it more relevant to today by scaling past returns to match current conditions.

It could be claimed that there is some circumlocution in this process. The obvious question is:

> *Why scale using volatility – which is predominantly determined by the centre of the return distribution – when what we care about is the 99th percentile?*

Pragmatically the answer may be 'because we need a lot more data to estimate the 99th percentile than to estimate the volatility'. After all, we can construct models which are a lot more reactive for the same accuracy if we scale based on volatility than if we scale based on the 99th percentile.[14] In this light, volatility scaling can be seen to rely on the relationship between the volatility and the 99th percentile of the conditional distribution remaining fixed.[15] If it does, historical data can indeed be scaled to provide a more up-to-date series with which to calculate VaR. If however the relationship between one standard deviation and the 99th percentile varies, perhaps because the tail of the distribution lengthens or contracts without matching changes in the centre of the distribution, then volatility scaling is more questionable.

In order to test this assumption, returns with a known VaR were simulated. Specifically, a normally-distributed return series was used, and an FHS VaR was calculated. If the FHS VaR differed from the true VaR only by a noise term, then we would expect that the ratio of FHS VaR to the standard deviation of unscaled returns would average 2.33. Instead, as Figure 9 illustrates, the ratio is biased lower than 2.33: this bias is moreover greater than the standard error expected (0.05).

This suggests that the claim of the stationarity of the residuals cannot always be taken at face value.

5 Science and Carpentry

The key stylised facts characterising risk factor returns – conditional heavy tails, volatility clustering, and so on – have been known for many years: see [14] for a cogent summary. We do not however have a wholly satisfactory model of them which displays all of the important properties in the right amounts (and arguably we never will have). Moreover, as Davis [17] points out, the statement

> *The conditional distribution of the risk factor x_j, given data up to time t is φ.*

where φ is a specified distribution function is meaningless in the sense that it is not falsifiable. No subsequent data points $x_j(t+1)$, $x_j(t+2)$ can prove that φ was the wrong choice. Therefore instead of asking whether our model is correct, Davis suggests, we should ask whether

[14] Indeed, if we have a good estimate of the 99th percentile of the current return distribution then we know the VaR and hence we do not need to scale anything.

[15] This assumption is recognised in the early papers. For instance, Hull and White state in [30] (without providing evidence) that 'The probability distribution of a market variable, when scaled by an estimate of its volatility, is often found to be approximately stationary.' The issue is the extent to which that claim is true for the α percentile.

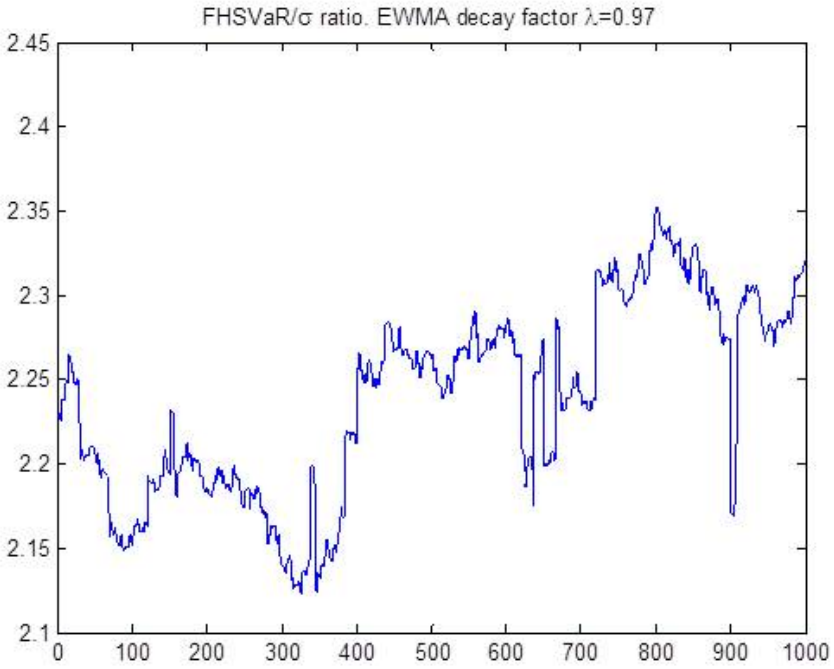

Figure 9: The ratio of the FHS VaR to the true standard deviation of returns for a simple simulated returns process

our objective in building the model has been achieved.[16] This then reduces to a question of whether we can falsify the statement "the prediction of the 99th percentile of the P/L distribution for all portfolios P sensitive to risk factors x_j is correct". This focusses attention squarely on backtesting procedures.

Our own view is that Davis' account, while fascinating, may be read too pessimistically. In general we may not be able to tell if a particular *conditional* distribution φ is wrong, but the evidence against some *unconditional* distributions is overwhelming. Equity index returns are not well described by the normal distribution, for instance. Thus while we endorse the view that the purpose of a VaR model is to predict some percentile of the return distribution, and it should be judged on how well it does that, we also suggest that the model's prediction of other properties of the return series, such as its higher moments or autocorrelation properties, is insightful. VaR model buiding is not pure science in the sense of finding the one true model of the return distribution, but equally it is not just carpentry in the sense of building something that is fit for one purpose. Thus at very least models which fail to provide a reasonably convincing account of the key properties of the historical return distributions should be subject to intense scrutiny of their VaR estimates. Chairs sometimes bear weight for a time, even if they are badly constructed: it is the job of the tester to determine if this is by design or by accident.

5.1 Backtesting FHS Models 1: General Remarks and Historical Results

There is now extensive literature on testing VaR models: see for instance [9, 11, 44]. In general models are tested by comparing their VaR estimates with the actual P/L that would have been experienced had a fixed portfolio been held on a given day in the historical past or hypothesised

[16] We are simplifying his argument here: see [17] for more details, and in particular for the important concept of 'elicitability' which plays the role of falsifiability in a stochastic setting.

future. A day when there is a loss bigger than the VaR is said to be *an exception*, and models are tested by examining the time series of exceptions. Thus for instance if a model claims to estimate one day VaR at the 99% confidence interval, but it has twenty exceptions in a year, then it may be suspect.

Three tests are in general use [11, 16]:

1. The *simple Kupiec* or Kupiec POF test compares the actual number of exceptions in a period with the expected number, given the target confidence interval.

2. The *Christoffersen* test also examines whether exceptions occur on neighbouring days. This gives it extra power to reject models which do not capture short term volatility clustering well.

3. The *mixed Kupiec* test extends this idea to examine volatility clustering over longer periods.

Each test requires the calculation of a different statistic with known distribution. The model can then be rejected if the statistic lies outside the acceptable bounds. In this case we say that it exceeds the *critical value*.

Several remarks should be made in this context:

- It is good practice to test a variety of different portfolios including outright positions in each risk factor, commonly traded spreads, and other well-known strategies. It is also important to test real-world portfolios (so that, in particular, realistically diversified portfolios are tested). There is value in testing portfolios where the first order risks have been fully hedged in order to understand the impact of higher order risks, too.

- There is a place for both historical testing using actual market data (including market data from stressed conditions) and simulation-based testing. The latter allows the effect of never-before-experienced conditions to be evaluated.

- It is also best practice to test the model at a range of different confidence intervals, as this can give insights beyond those available at a single confidence interval.

- The object to be tested is the model *together with its recalibration strategy*. That is, if in reality a key parameter of a model is reviewed and perhaps changed every month, then the backtest should be of the model with recalibration, not of the model with the parameter fixed.

Clearly a model that fails a number of the tests outlined above is questionable. However, it is important to understand the reasons for failure. As the number of tests increases, the probability of a good model failing one test increases, so rather than setting a standard which may be unrealistically high – such as pass all the tests – we prefer to suggest that model builders should be able to justify their failures. Consistent failure to handle elevated volatility is a serious problem, for instance; but a random and infrequent pattern of narrow failures consistent with the discriminating power of the test may not be.

We will present a small subset of backtests which give insight into the performance of FHS models compared to ordinary historical simulation models. Figure 10 gives the test results for our low variance portfolio using historical data: the STV model is least good for this portfolio, its relatively bad (although not unacceptable) performance is due in particular to four exceptions in a two month period starting on day 894.

More insight can be gained from the risk factor backtests and backtests of spread positions. These can show issues which arise due to a failure to handle a changing conditional correlation

VaR Model	HS	LTV	STV	Critical value
Number of exceptions	9	9	12	
Simple Kupiec Statistic	0.1	0.1	0.38	6.6
Christoffersen Statistic	0.27	0.27	0.67	9.2
Mixed Kupiec Statistic	14.6	11.5		23.2
			13.5	27.7

Figure 10: A summary of the backtesting performance of three 99% VaR models using 1,000 days historical data for a low variance portfolio. The critical value of the mixed Kupiec test is a function of the number of exceptions.

between the two risk factors, and so probe a different vulnerability from single factor backtests. As an illustration of the concern, consider figure 11. This shows the performance of all three models for a WTI/Brent spread position. Clearly the period from day 894 to day 1,106 is a challenge, and indeed the STV model fails all three backtests over the entire period.

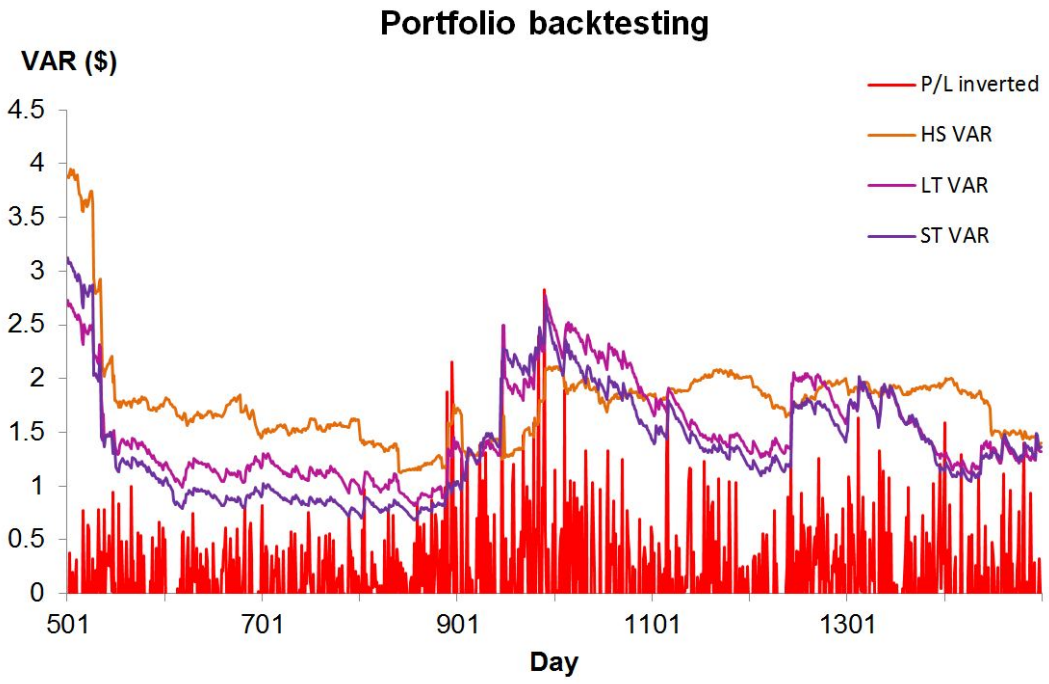

Figure 11: An illustration of model performance for WTI vs. Brent spread portfolio

This issue is illustrated not to criticise the models concerned, but rather as an illustration of the importance of understanding why a backtest failure occurs. Once we understand the cause of the problem, the model designer can investigate possible mitigations such as changing the volatility estimation procedure for devol, revol or both; adding a volatility floor to the model; or some other approach. These issues are discussed further in sections 6 and 7.

5.2 Backtesting FHS Models 2: Simulation Results

The phenomena we discuss can also be illustrated using simulation methods. Specifically we calibrate a well-known asymmetric GARCH model, GJR-GARCH [24], to the return series, then simulate returns generated by this model. This model allows conditional volatility to vary depending both on the size and sign of returns, and thus is a good fit to financial returns where

large negative returns tend to be associated with higher conditional volatility than smaller negative or even some larger positive returns.[17]

The GJR model we use was introduced by Glosten, Jagannathan and Runkle in 1993 (see [24]). The variance equation in a GJR-GARCH(1,1) model is defined as

$$\sigma_t^2 = \kappa + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma u_{t-1}^2 \cdot 1_{(\varepsilon_{t-1}<0)}$$

where $1_{(\varepsilon_{t-1}<0)}$ is the indicator function, γ is the leverage coefficient and the parameters satisfy the following constraints

$$\begin{aligned} \beta \geq 0, \alpha \geq 0, \kappa \geq 0, \\ \alpha + \gamma \geq 0 \\ \beta + \alpha + \frac{1}{2}\gamma < 1. \end{aligned}$$

This definition includes additional terms to capture the asymmetry in volatility clustering compared to the standard GARCH specification, allowing large negative changes to be more clustered than positive ones. We calibrate the GJR-GARCH(1,1) models to the three risk factors and we will capture their joint distribution using a Gaussian copula.

Simulation models are helpful because they allow us to test the performance of the FHS models against a sample of returns for which the joint dynamics are known in advance. In other words, by performing simulations where the 'real' dynamics of the returns are known in advance and testing the models against those simulated scenarios, we can measure to what extent the models succeed in capturing the characteristics of the underlying processes.

A standard analysis of the returns of WTI, Gasoil and UK Natural Gas over our data window indicates there is little autocorrelations in the returns but that the three series show significant heteroskedasticity. This suggests that the GJR model is indeed a reasonable choice for generating hypothetical scenarios for backtesting.

The performance of the models under volatility or correlation regime shifts will be illustrated using the WTI-Gasoil spread portfolio discussed above. We backtest the various VaR measures produced by the models under the following scenarios::

1. No changes in the GJR-GARCH(1,1) specification (baseline process)
2. Volatilities jump up by 50% and after 200 days they return to their previous level.
3. Correlations decrease by 50% in the second half of the observation window.

Figure 12 below shows the average number of breaches observed after 100 simulations, together with the significance of those breaches according to the simple Kupiec test.

These results show that both LTV and STV consistently outperform the ordinary HS VaR which, according to the Kupiec test, fails to correctly capture the model dynamics even in the baseline case. On the other hand, STV calibrated to a $\lambda = 0.97$ presents a better performance, although LTV tends to be more conservative.

As expected, both LTV and STV react to changes in volatility, while the HS model fails to capture it. On the other hand, the three models fail to capture the change in correlation, confirming the need to introduce a correlation updating mechanism into the model. We will come back to this in the final sections.

[17] The question of which GARCH-type model 'best' models financial returns is a complex and nuanced one which we will not address, noting simply that GJR-GARCH is a reasonable performer which captures the essential phenomena of volatility clustering, skewness, and (differential) fatness of the tails.

[17] The statistic for each individual test is χ-squared distributed, so the statistic for the average is gamma-distributed. The critical value is 1.358.

Model	Unfiltered	$LTV_{0.97}$	$LTV_{0.99}$	$STV_{0.97}$	$STV_{0.99}$
Baseline returns (no change)	**12.20**	8.83	10.20	10.09	11.11
	1.82	*0.94*	*0.74*	*0.51*	*0.73*
50% volatility increase	**13.00**	**8.14**	10.57	9.59	10.70
	1.88	***1.39***	*0.87*	*0.63*	*1.02*
50% correlation decrease	**17.61**	**15.22**	**16.01**	**16.48**	**16.33**
	6.18	***3.38***	***3.75***	***4.09***	***3.91***

Figure 12: Backtesting results for the WTI/Gasoil spread using simulated returns over 1,000 days. The baseline returns are modeled using a GJR-GARCH(1,1) process. Numbers in italics are the simple Kupiec test statistics. The figures in bold indicate failure under the Kupiec test at 99% confidence.

5.3 The Ratchet Model

The difficulty of modelling risk factor returns brings model risk. Moreover, as risk models become more complex, and the associated calibration procedure becomes more intricate, the danger of over-fitting increases. This is particularly so when 'success' is measured using an undemanding test such as the simple Kupiec. This is perhaps best illustrated by a cynical model: one that simply aims to pass its Kupiec test at all costs.

Recall that the historical simulation VaR is completely determined by the set of portfolio P/Ls given the input risk factor changes. Each day's risk factor changes, whether scaled or not, give us a P/L, and the VaR is one of these P/Ls.[18] Specifically, if we have 500 days of risk factor changes then the 99% VaR is the fifth worst loss. A 99% VaR model passes its Kupiec test if the number of days in which the actual loss on a portfolio is bigger than the VaR in some period is below some threshold. Thus for instance the 'Basel red' version of a Kupiec test for a 1 day VaR requires that there are less than ten days that the loss is bigger than the 99% VaR in a year. This kind of test is known to have relatively low power to distinguish good from bad models. Moreover, despite fat tails and variable conditional volatility, it is relatively easy to build a model that tries to pass its Kupiec test as follows:

- Use a 500 day (unfiltered) historical simulation VaR model, and set the day used for VaR as the 99th percentile worst, $d = 5$, initially as for an ordinary 99% unfiltered VaR.
- Each day, calculate the number of backtest exceptions in the last two years.
- If the number of exceptions is greater than some threshold VaR_u, 'ratchet up' by using a bigger loss for today's VaR, setting $d = d + 1$.
- If the number of exceptions is less than some threshold VaR_d, 'ratchet down' by using a smaller loss for today's VaR, setting $d = d - 1$.

We call this a *ratchet VaR* model.[19] Ratchet VaR with $\text{VaR}_u = 3$ and $\text{VaR}_d = 1$ will increase the risk estimate if we have had three backtest exceptions or more in the past two years, and decrease it if we have had one or none. This model often passes risk sensitivity tests as, just like volatility scaling models, it increases the risk estimate in volatile periods. Indeed, provided the historical returns used for this model contain a stressed period and the VaR_u threshold is

[18] Some VaR models use a few rather than one point in the tail to estimate the VaR, for instance fitting a curve to the fourth, fifth and sixth worth losses. The point remains however that only a few losses completely determine the VaR.

[19] Ratchet VaR is a kind of crude empirical quantile estimation model [34]. It could be improved upon in a number of ways, notably by keeping track of exceptions at other confidence intervals and using that information to improve on the decision of when to ratchet up or down.

set low enough, only stress much more intense than anything in the data windows typically causes the model to have too many backtest exceptions.

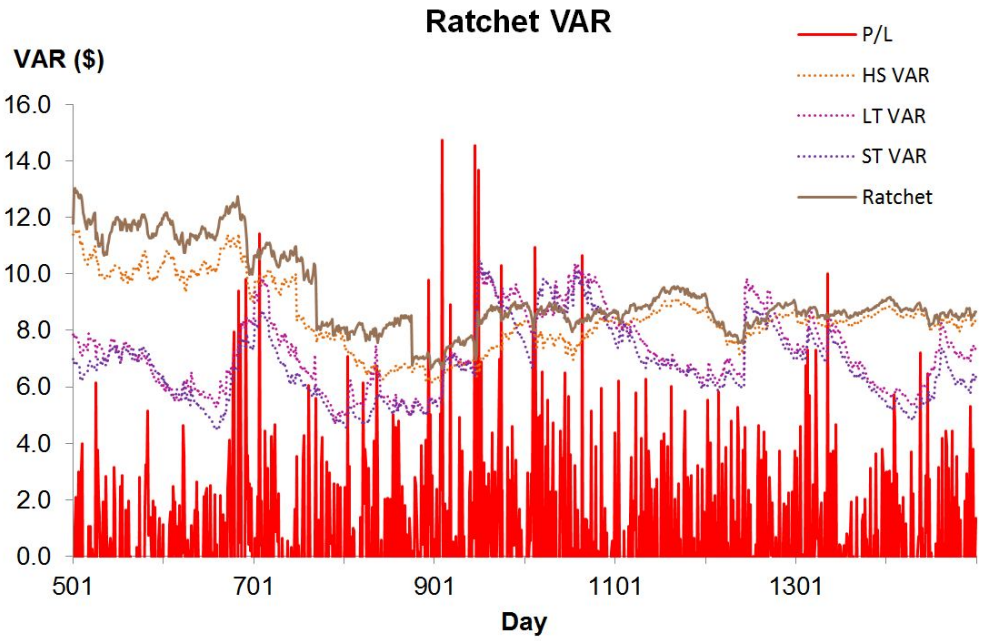

Figure 13: Ratchet VaR compared to various other VaRs and realised P/L for the low variance portfolio

Figure 13 illustrates the ratchet VaR model and compares it with an unfiltered VaR.[20] It can be seen that the ratchet VaR with these triggers is (slightly) more conservative than the historical simulation VaR: it has 5 exceptions rather than 7 for this portfolio, for instance. It also reacts to the increased volatility around day 894 faster than any of the other VaRs; this leaves it relatively well placed to handle the further bout of volatility starting around day 980 – although, to be fair, the LTV model handles this period well too.

This example is instructive as it shows that a model that reacts to local volatility conditions well does not necessarily have to have a credible model of the underlying returns process: it can be a good deal simpler. As we discussed above, the extent to which this matters depends on one's view of the modelling process. If it is seen as purely outcome-based – derive an accurate risk measure however you like – then ratchet VaR might be a useful innovation. If however one can only be confident in a risk measure if it has an accurate model of the underlying returns-generating process, then the ratchet VaR approach might be seen as cheating.

5.4 What Would It Mean For FHS To 'Explain' Returns?

The distinction made above is between :

- a model which is a compelling model of returns, and as a by-product, estimates VaR correctly; and
- a model which calculates VaR well enough to pass its backtests.

[20] Ratchet models tend to perform better with longer data windows as firstly this provides smaller jumps when the ratchet is hit, and second it reduces the likelihood that the data window will not contain events which are sufficiently stressful. A 1,000 day ratchet VaR will therefore often out-perform the 500 day one we present here. It also allows us finer control over VaR_u and VaR_d as there are more points to chose from. We could also elaborate the model by separating the VaR data window from the backtesting window used to determine whether to ratchet up or down.

It is worth exploring what it would mean for an FHS model to fall into the former class.

The underlying theory of FHS is that the devolatilisation process gives rise to residuals which are N(0,1) with no autocorrelation. At least for our data, this is not true, as Figure 14 illustrates: of the long term devolatilised residuals, for instance, only WTI is anywhere close to normality based on the Jarque Bera test. Therefore FHS VaR estimates are often accurate *not always* because the model assumptions hold, but perhaps in part because the 'ghost' of non-i.i.d. normality survives the devolatilising process, and thus can inform the VaR estimate.

	WTI	Gasoil	Natgas
Long term	11	91	65
Short term	395	5.0	2.3

Figure 14: The Jarque Bera Statistic for the devolatilised returns used for day 1,500's VaR using $\lambda = 0.97$. The asymptotically critical value of the statistic at 99% confidence is 9.2.

The normality of the residuals can be improved slightly by optimising the EWMA decay parameter separately for each return series, as figure 15 shows.[21] However in some sense this only emphasises the problem: why should different λs be needed to produce optimal residuals for different risk factors?

	WTI	Gasoil	Natgas
Optimal λ	1.00	0.958	0.973
Jarque Bera Statistic at this λ	11	12	5.2

Figure 15: The decay factor which minimises the Jarque Bera statistic for each series of devolatilised returns used for day 1,500's ST VaR

Our aim in presenting these results is two-fold. First it shows that different FHS family members perform differently, and thus there is the need for model builders to carefully calibrate parameters to achieve optimal performance. Second it illustrates the need for ongoing monitoring of the appropriateness of the chosen parameterisation. If we are not 'explaining returns' but rather estimating VaR, it is incumbent upon us to show that the estimation process continues to be relevant as the properties of the risk factor returns change.[22]

6 The Procyclicality Of FHS Models

The variability of risk estimates is an important consideration for margin models as a model which over-reacts to current conditions can place liquidity burdens on the parties margined. In extreme conditions, these burdens can contribute to systemic risk.

This issue has been recognised in regulation, so that for instance the European Union's EMIR regulation [21] requires that CCPs

> *should adopt initial margin models and parameters that are risk-based . . . [and these should] to the extent practicable and prudent, limit the need for destabilising, procyclical changes.*

[21] We are not claiming here that the Jarque Bera statistic is the only relevant measure here, or even that it is the best measure of departures from normality. Calculating which λ minimises it is insightful in that it suggests that different returns series could have substantially different conditional volatility dynamics.

[22] Cont [14] makes the points that 'even after correcting returns for volatility clustering (e.g. via GARCH-type models), the residual time series *still* exhibits heavy tails' (emphasis ours). This contradicts the assumption of stationarity of the distribution of residuals, and emphasises the need to control for failure of this assumption.

In previous work [43], one of us has identified a number of measures of procyclicality which can be used to compare the performance of margin models. This section applies two of these measures to the models discussed here in order to shed light onto the procyclicality of FHS models.

6.1 Peak-to-Trough Procyclicality

The peak-to-trough procyclicality of a margin model is the ratio of the maximum initial margin required for a constant portfolio to the minimum margin required over a fixed observation period. This is therefore an 'across the cycle' measure of procyclicality.

The time series of VaR shown in figure 4 suggest that the peak-to-trough procyclicality of the unfiltered models is smaller than that of the filtered ones, and indeed this is the case, as figure 16 reports. Here we show both the usual LTV model discussed above with filter factor 0.97, and two variants $\text{LTV}_{0.95}$ and $\text{LTV}_{0.99}$ using the same devol but with revol volatilities calculated using a higher and a lower lambda, ${}^{0.95}_{500}\sigma_j$ and ${}^{0.99}_{500}\sigma_j$ respectively.

Data	Unfiltered	$\text{LTV}_{0.95}$	LTV	$\text{LTV}_{0.99}$	STV
Historical	1.69	2.97	2.53	2.02	2.82
Simulated	1.54	3.92	3.21	2.10	3.23

Figure 16: The Peak-to-Trough procyclicality of five VaR models for a position in the WTI risk factor

It is evident that all the FHS models are more procyclical on this measure than the unfiltered historical simulation model. Moreover, the reactivity of a smaller lambda comes at a significant cost in procyclicality: the $\lambda = 0.95$ LTV model will adapt more quickly (and noisily) to conditional volatility than the ones with larger λs, but that very reactivity causes the risk estimate to vary more across the cycle.

6.2 n-day Procyclicality

The n-day procyclicality of a margin model is the largest increase in margin over an n-day period for a constant typical portfolio over a fixed observation period. For small n, such as 5- or 30-days, this measure captures the short term liquidity stress caused to a market participant by changes in market conditions creating higher risk estimates and thus margin calls. Figure 17 reports the 5- or 30-day procyclicality measures for our three models as a percentage of the notional of the position.

Data	Unfiltered	$\text{LTV}_{0.95}$	LTV	$\text{LTV}_{0.99}$	STV
Historical	0.39	3.18	2.27	1.00	2.38
Simulated	0.50	2.30	1.77	0.94	1.79

Figure 17: Two n-day procyclicality measures of five VaR models for a position in the WTI risk factor: above, the 5-day measure; below, the 30-day

Data	Unfiltered	$\text{LTV}_{0.95}$	LTV	$\text{LTV}_{0.99}$	STV
Historical	0.71	3.74	2.60	1.26	2.88
Simulated	0.83	4.05	3.24	1.87	3.20

The same pattern can be seen here as for the P-T measure; filtering increases procyclicality, and using a smaller decay factor increases the procyclicality further.

6.3 Mitigating Procyclicality

It is clear that FHS models can be substantially more procyclical in both P-T and n-day measures than unfiltered models. They may therefore require procyclicality mitigation. As discussed above, EMIR [21] requires that initial margin models used by CCPs include one of three forms of procyclicality mitigation. The most interesting of these for our purposes is the use of a ten year unweighted VaR floor. Figure 3 gives some insight into this: the effect of a such a floor, roughly[23], would be to take the max of the orange and the pink or purple lines. The impact of this flooring would clearly be to stop the FHS VaR from falling 'too low' in quiet markets, such as those from day 560 to day 850.

7 Extensions

There are various responses to the issues identified in prior sections. For instance:

- Different definitions of filtering could be explored;
- We could estimate more sophisticated conditional distributions, such as location-scale models [40] or generalised lambda distributions [12];
- We could use a non-parametric or semi-parametric approach, for instance using Kernel estimators [1], or explicitly fitting the higher moments of the distribution [22].
- We could explore a more elaborate function to determine the capital or margin requirement from the VaR. The use of a floor, discussed in the previous subsection, is a simple example of this approach.

Some of these possibilities are explored in a little more detail below.

7.1 The 'Filtered' in 'Filtered Historical Simulation'

The EWMA volatility updating scheme usually used in financial applications provides an estimate of the volatility on day t conditional on the information available at day $t-1$.

$$ {}^{\lambda}_{n}\sigma_j(t)^2 = \lambda \cdot {}^{\lambda}_{n}\sigma_j(t-1)^2 + (1-\lambda)x(t-1,j)^2 \tag{1} $$

This is consistent with the interpretation of a volatility estimate as a forecast. However, this may not be the only way of estimating current volatility through an EWMA process. For example, in control theory applications it is frequent to find an updating scheme where today's volatility at t is estimated using today's information available at the same day t:

$$ {}^{\lambda}\tilde{\sigma}_j(t)^2 = \lambda \cdot {}^{\lambda}\tilde{\sigma}_j(t-1)^2 + (1-\lambda)x(t,j)^2 \tag{2} $$

Although this last specification cannot be interpreted as a forecast for day t volatility, it could still be understood as a forecast for the volatility from $t+1$ onwards.

Even if the differences from different indexing conventions may be small, it may be worth observing that they effectively lead to two different FHS outcomes. To see this, it is sufficient to observe that as λ decreases, the filtered returns defined at time T using (1) will converge to $\{\frac{x(i)}{x(i-1)}x(T)\}_{t<T}$, while under (2) they will converge to a constant return series $\{x(T)\}$.

Control theory also suggests a much wider range of filters than are the ones commonly in use in FHS models and discussed above. Some of these approaches are not unknown in time

[23] The HS VaR illustrated has a shorter lookback period than ten years, so the result is not precise.

series analysis: the venerable textbook by Kendall and Ord [33] for instance treats high pass and low pass filters. The general setting here is:

- Fix the quantity that exhibits mean-reverting variation: for us this would be some volatility estimate ${}^{\lambda}_{n}\sigma_j(i)$;
- Perform a decomposition of a complete cycle of the quantity into Fourier components;
- Apply some weighting scheme to the components, so that for instance in a low pass filter the low frequency components would be unaffected and the highest frequency ones attenuated;
- Calculate a filtered quantity from the weighted components.

Obviously a very wide range of filters can be defined depending on the precise weighting scheme chosen (and how a 'complete cycle' is defined). In the risk measurement setting it would be natural to start with a low pass filter which discarded some of the highest frequency variation in volatility estimates. This might well remove much of the noise even for quite low lambda volatility estimates.

7.2 The Impact of FHS on Correlations

In a FHS process, the volatility of each risk factor is re-scaled without any reference to other risk factors. In fact, using filtered returns to model the joint behaviour relies on the implicit assumption that the relationships between those risk factors do not depend on the re-scaling of the returns. Such assumption may be wrong when we move away from processes with unconditional (constant) volatility.

Suppose that we use linear correlation as a measure of the relationship between the risk factors. From the definition of x^{STV} and the properties of the correlation coefficient, it follows that for FHS to preserve correlations between risk factors the following must hold

$$\rho\left(\frac{x_1(i)}{{}^{\lambda}_{n}\sigma_1(i)}, \frac{x_2(i)}{{}^{\lambda}_{n}\sigma_2(i)}\right) = \rho((x_1(i), x_2(i)) \tag{3}$$

If we assume constant volatility then the equality (3) will hold. However, we cannot rely on this assumption (which would mean that any differences observed between sample volatility estimates at different points in time can be solely attributable to sampling error), because it contradicts the whole purpose of FHS. Therefore, we cannot expect that the volatility rescaling will preserve the correlations between factors. The following example illustrates how the divergence between the correlations observed in the rescaled series and the original ones could materialize under an EWMA specification.

Example: Consider two risk factors, each one generated from a normal distribution and with a dependence structure defined by the following covariance matrix:

$$\Sigma = \begin{pmatrix} 0.01 & 0.004 \\ 0.004 & 0.0025 \end{pmatrix}$$

In other words, the volatilities are $\sigma_1 = 0.1$, $\sigma_2 = 0.05$ and the correlations are $\rho_{12} = 0.8$. For each factor we generate 1,500 observations and we analyze 500 VaR estimations, each one obtained from a 1,000 day window.

Figure 18 shows the effect of the choice of λ on the sample correlation for each one of the 500 rescaled samples. In particular, there is a gap between the correlations of non-scaled and of the scaled samples and this gap increases as lambda decreases.

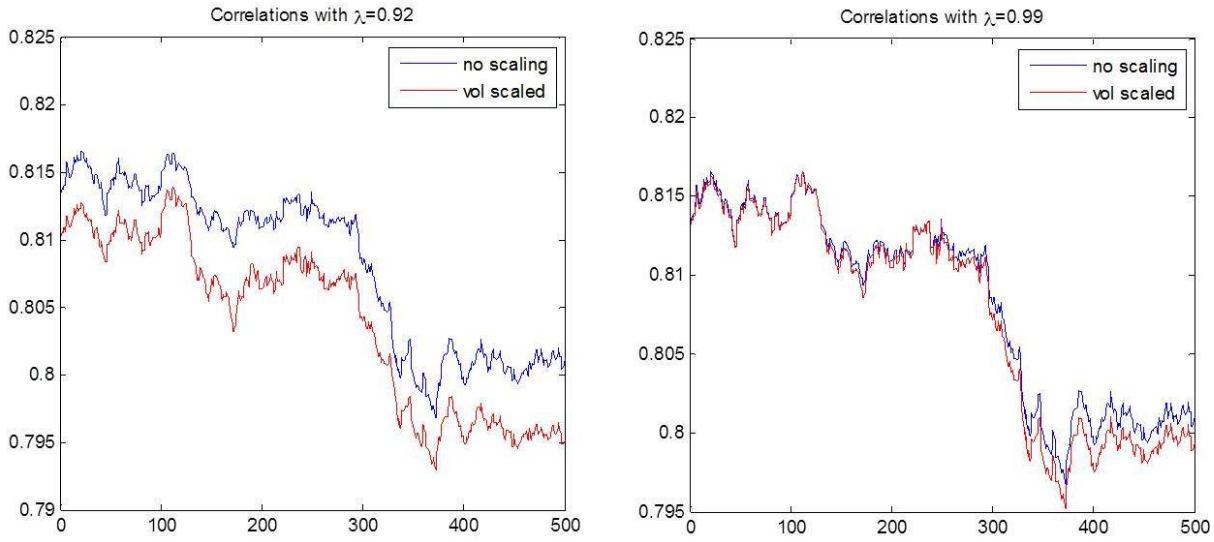

Figure 18: An illustration of the divergence of correlation estimates for each of 500 rolling windows (of 1,000 days each) when the samples are rescaled. The gap widens as decay factor decreases.

The fact that a correlation-consistent FHS transformation implicitly relies on the assumption of constant volatility is an important shortcoming of the methodology, as this assumption does not hold in most cases. In fact, FHS is justified when differences in volatility estimates at different points in time may are not only be a consequence of sampling error but may are due to a structural change in the underlying processes.

7.3 Correlation Updating

The above observation that FHS distorts correlations confirms the potential importance of rescaling the volatilities and the correlations (and therefore, of the whole covariance matrix) in a consistent way. In fact, in a context of varying correlations this approach could potentially improve the model performance as the historical multivariate sample will be rescaled in such a way that the covariance of the rescaled sample will better reflect the current covariance structure. This process could involve two steps which could be seen as a direct generalization of the devol and revol steps used in FHS. However, this process may be not uniquely defined and there may be different ways for rescaling covariances.

Duffie and Pan [20] suggested a way of rescaling covariances by considering the square root of the covariance matrix. More precisely, if Σ denotes the historical covariance matrix and $\hat{\Sigma}$ is an updated covariance estimate then the historical returns distribution can be updated for volatility and correlation by replacing each vector $x_j(i)$, at each past date i, with

$$\hat{x}_j(i) = \hat{\Sigma}^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}x_j(i) \tag{4}$$

where $\Sigma^{1/2}$ denotes the matrix square root of Σ.[24] Since the covariance of Mx is $M\Sigma M^T$ for any $j \times j$ matrix M and any j-vector x_j, then the covariance of $\hat{x}_j(i)$ is

$$\hat{\Sigma}^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma\left[\hat{\Sigma}^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right]^T = \hat{\Sigma} \tag{5}$$

To test the effectiveness of the process of filtering covariances in capturing changes in correlation, we run the simulations described in section 4.2 but now using the covariance filtering (CF)

[24] A similar argument can be applied if instead of the square root of Σ, the Cholesky decomposition A is used.

scheme defined by equation 5. The VaR with filtered correlation and a $\lambda = 0.97$ achieved on average 11 exceptions (with a Kupiec LR of 0.7095). A comparison of this result with the outcomes shown in Figure 12, shows that the covariance filtering approach outperforms the other models. Figure 19 confirms that this is indeed the case at different values of the parameter λ.

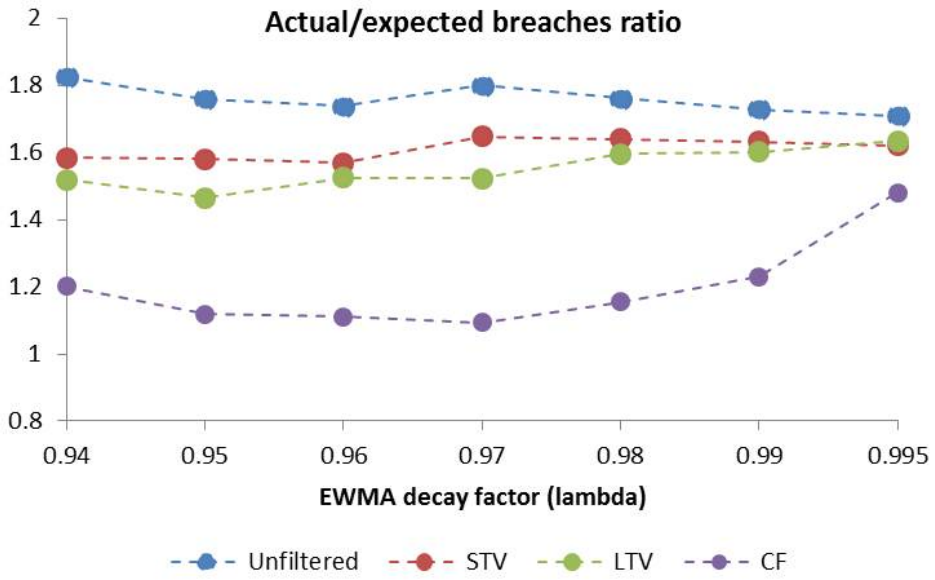

Figure 19: An illustration of the average ratios between observed and expected breaches for different models and for different values of the parameter λ. The averages reflect backtesting results for the WTI/Gasoil spread using simulated returns over 1,000 days.

Note that this approach assumes that at least in each rolling window the processes $x_j(i)$ have constant covariance. However, this condition may not hold in general. This may lead to consideration of a short term covariance scaling process similar in spirit to the STV models we have already analyzed. Just as EWMA volatility estimates may be preferred when more reactive volatility forecasts are needed, and because they avoid the 'ghost' effects of equally weighted moving averages, the covariance filtering process can be implemented using EWMA covariance estimators. However, covariances based on EWMA must have the same λ for all the variance and covariance estimates to ensure a positive definitive covariance matrix. Moreover, because EWMA correlation estimates tend to be more unstable than unweighted ones, this filtering approach may lead to more unstable VaR measures.

7.4 Higher Moment Scaling

We have seen that FHS relies on the assumption that volatility is the only factor determining the shape of the distribution and, in particular, of the tail percentiles. When this is not the case, for example in the presence of skewness or fat tails, VaR estimates may be inadequate. This lack of sensitivity to higher moments could be addressed by extending the updating process to include skewness and kurtosis. One way of implementing this approach could be to estimate VaR using the Cornish-Fisher expansion series for the first few moments of the distribution together with some updating scheme that allows the modeller to jointly capture the dynamics of, for instance, volatility, skewness and kurtosis.

If z_α is a standard normal variate for a confidence level α, then the Cornish-Fisher expansion for the first four moments is:

$$z_\alpha + \frac{s_t}{3!}(z_\alpha^2 - 1) + \frac{k_t}{4!}(z_\alpha^3 - 3z_\alpha) + \frac{s_t^2}{36}(2z_\alpha^3 - 5z_\alpha)$$

where s_t and k_t denote the distribution's skewness and kurtosis coefficients, respectively. The Cornish-Fisher expansion allows estimation of a percentile adjusting the normal variate z_α for skewness and kurtosis. On the other hand, one can introduce an updating scheme that jointly estimates time-varying volatility, skewness and kurtosis using a modified EWMA process, as suggested in [22]. In this context, a FHS process could potentially react more adequately to changes in the shape of the distribution. However, the performance of this approach is mixed, suggesting the model will not always provide a significant improvement over the traditional FHS models. Moreover, the 4th order Cornish Fisher approach has some drawbacks due to the fact that the expansion provides good approximations only if the deviations from normal are small. Its use will significantly increase the sensitivity of the model to data problems, complicating its calibration: in bad cases this can lead to unstable VaR estimates.

Another approach is to replace the EWMA variance estimator by a more general exponentially weighted maximum likelihood (EWML) procedure that potentially allows for time variation in the variance and in the higher moments of the distribution. This approach was suggested in [25] and it was applied to forecast VaR allowing for time-variation in both the variance and the kurtosis of returns. Again, though, issues of stability of calibration often arise and their mitigation may necessitate the use of smoother, less reactive moment estimates.

8 Conclusions and Further Work

The FHS model family contains many members. Undoubtedly some of them are useful and interesting risk models in some situations. However our results have shown that care is needed to select the right family member, to calibrate it effectively, and to test it comprehensively. Moreover this is not a 'one off' process; regular re-calibration and re-testing is needed to ensure that the model remains relevant.

We have also shown that the filtering process changes the return distribution in ways that may not be intuitive. This may not matter if the only concern is the calculation of a conditional VaR estimate at a fixed confidence interval for simple portfolios. In most applications, though, careful testing is needed to verify the accuracy of the risk estimates of FHS models, and to understand the circumstances under which they fail. This is especially so when the portfolios of interest include those which are sensitive to the far tail, or to higher moments of the return distribution. Examining the properties of the residuals and the scaled distribution may be helpful here.

There are many possible extensions to the FHS paradigm, of which we have outlined a few. One promising line of further work here may be the application of filtering techniques familiar in signal processing to financial time series. For instance, a *low pass* filter might add substantial procyclicality mitigation to a FHS model with a low decay factor without overly compromising its risk sensitivity. We hope to examine these issues further in forthcoming work.

References

[1] R. Alemany, C. Bolancé, M. Guillén, *Nonparametric estimation of Value-at-Risk*. Working Paper, Department of Econometrics, Riskcenter-IREA, University of Barcelona, 2012

[2] C. Alexander, *Handbook of Risk Management and Analysis*. Wiley, 1996

[3] G. Barone-Adesi, K. Giannopoulos, *Non-parametric VaR techniques: myths and realities*. Economic Notes, Volume 30, Banca Monte dei Paschi di Siena, Siena, 2001

[4] G. Barone-Adesi, K. Giannopoulos, L. Vosper *VaR without Correlations for Portfolios for Portfolios of Derivative Securities*. Journal of Futures Markets, Volume 19, 1999

[5] Basle Committee on Banking Supervision, *Overview of the Amendment to the Capital Accord to incorporate market risks*, BCBS 23, BIS 1996

[6] Basle Committee on Banking Supervision, *Revisions to the Basel II market risk framework*, BCBS 158, BIS 2013

[7] Basle Committee on Banking Supervision, *Margin requirements for non-centrally cleared derivatives*, BCBS 261, BIS 2013

[8] Basle Committee on Banking Supervision, *Fundamental review of the trading book - second consultative document*, BCBS 265, BIS 2013

[9] J. Berkowitz, P. Christoffersen, D. Pelletier, *Evaluating Value-at-Risk Models with Desk-Level Data*. Management Science, Volume 57 Issue 12, 2011

[10] J. Boudoukh, M. Richardson, R. Whitelaw, *The Best of Both Worlds*. Risk Magazine, May 1998

[11] S. Campbell, *A Review of Backtesting and Backtesting Procedures*. Finance and Economics Discussion Series Divisions of Research and Statistics and Monetary Affairs, Board of Governors of the Federal Reserve System, 2005

[12] N. Cecchinato, *Forecasting Time-Varying Value-at-Risk*. Thesis, Queensland University of Technology, 2010

[13] Y. Chen, A. Tu, *Portfolio Value-at-Risk Estimation with a Time-varying Copula Approach: An Illustration of Model Risk*, in the proceedings of the 5th International Conference on Risk Management, 2008

[14] R. Cont, *Empirical properties of asset returns: stylized facts and statistical issues*. Quantitative Finance, Volume 1, 2001

[15] J. Cotter, *Scaling conditional tail probability and quantile estimators*. Risk Magazine, 2009

[16] P. Christoffersen, D. Pelletier, *Backtesting Value-at-Risk: A Duration-Based Approach*. Journal of Financial Econometrics, Volume 2, Issue 1, 2004

[17] M. Davis, *Consistency of risk measure estimates*, Working Paper, Imperial College, 2014

[18] M. Dacorogna, U. Müller, O. Pictet, C. de Vries, *Extremal Forex Returns in Extremely Large Data Sets*. Extremes, Volume 4, Number 2, 2001

[19] J. Danielsson, J-P. Zigrand, *On time-scaling of risk and the squareâĂŞrootâĂŞofâĂŞtime rule*. Discussion paper Number 439. Financial Markets Group, London School of Economics and Political Science.

[20] D. Duffie, J. Pan, *An Overview of Value at Risk*. The Journal of Derivatives, Volume 4, Number 3, 1997

[21] European Union, *Regulation (EU) No. 648/2012 on OTC derivatives, central counterparties and trade repositories*, July 2012

[22] A. Gabrielsena, P. Zagagliab, A. Kirchnerc, Z. Liud, *Forecasting Value-at-Risk with Time-Varying Variance, Skewness and Kurtosis in an Exponential Weighted Moving Average Framework*, June 2012

[23] K. Garbade, *Assessing risk and capital adequacy for Treasury securities*, Topics in Money and Securities Markets, Volume 22, 1986

[24] L. Glosten, R. Jagannathan, D. Runkle, *On The Relation between The Expected Value and The Volatility of Nominal Excess Return on stocks*. Journal of Finance, Volume 48, 1993

[25] C. Guermat, R. Harris, *Forecasting value at risk allowing for time variation in the variance and kurtosis of portfolio returns*, International Journal of Forecasting 18, 2002

[26] P. Hildebrand, *Hedge funds and prime broker dealers: steps towards a "best practice proposal"*. Banque de France Financial Stability Review, Number 10, April 2007

[27] D. Heller, N. Vause, *Collateral requirements for mandatory central clearing of over-the-counter derivatives*, BIS Working Paper Number 373, 2012

[28] D. Hendricks, *Evaluation of Value at Risk Models Using Historical Data*. Economic Policy Review, Federal Reserve Bank of New York, April 1996

[29] G. Holton, *History of Value-at-Risk: 1922-1998*, Working Paper, July 2002

[30] J. Hull, A. White, *Incorporating volatility updating into the historical simulation method for Value-at-Risk*. Journal of Risk, Volume 1, 1998

[31] P. Jorion, *Value at risk*, IRWIN, 1997 (3rd edition, McGraw-Hill 2006)

[32] JPMorgan/Reuters, *RiskMetrics Technical Manual* , 1996

[33] M. Kendall, J. Ord, *Time series*, Oxford University Press, 1989

[34] R. Koenker, G. Bassett, *Regression Quantiles*, Econometrica Volume 46 Number 1, 1978

[35] K. Kuester, S. Mittnik, M. Paolella, *Value-at-Risk Prediction: A Comparison of Alternative Strategies*, Journal of Financial Econometrics, Volume 4, Number 1, 2006

[36] L. Kalvyas, N. Dritsakis, C. Siriopoulos, C. Grose, *Selecting Value at Risk Methods According to their Hidden Characteristics*. Operational Research, Volume 4, Number 2, 2004

[37] A. McNeil, R. Frey, *Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach*, Journal of Empirical Finance, Volume 7, 2000

[38] B. Mandelbrot, *The variation of certain speculative prices*, IBM External Research Report NC-87, 1962

[39] S. Manganelli, R. Engle *Value at risk models in finance*. Working paper series 75, European Central Bank, 2001

[40] C. Martins-Filho, F. Yao, M. Torero, *Nonparametric estimation of conditional Value-at-Risk and expected shortfall based on extreme value theory*. Working Paper, University of Colorado, 2012

[41] R. Miura, S. Oue, *Statistical Methodologies for Market Risk Measurement*, Asia Pacific Financial Markets, 2001

[42] D. Murphy, *OTC Derivatives: Bilateral Trading and Central Clearing*, Palgrave Macmillan, 2013

[43] D. Murphy, M. Vasios, N. Vause, *An investigation into the procyclicality of risk-based initial margin models*, Bank of England Financial Stability Paper Number 29, 2014

[44] O. Nieppola, *Backtesting Value-at-Risk Models*, Master's Thesis, Helsinki School of Economics, 2009

[45] A. Pagan, *The econometrics of financial markets*, Journal of Empirical Finance, Volume 3, Number 1, 1996

[46] M. Pritsker, *The Hidden Dangers of Historical Simulation*, The Journal of Banking and Finance, Volume 30, 2006

[47] R. Repullo, J. Suarez, *The Procyclical Effects of Basel II*. Center for Monetary and Financial Studies Working Paper Number 0809, 2008