On the accuracy of VaR estimates based on the Variance-Covariance approach

Rakhal D. Davé[1] and Gerhard Stahl[2]

[1] Olsen & Associates Research Institute for Applied Economics, Seefeldstrasse 233, 8008 Zürich, Switzerland (email: rakhal@olsen.ch)
[2] Federal Banking Supervisory Office, Gardeschützenweg 71-101, 12203 Berlin, Germany (The views expressed herein should not be construed as being endorsed by the Federal Banking Supervisory Office)

Abstract

We present a thorough empirical study (based on over 8 years of daily data) of candidate models for forecasting losses in relation to positions held against individual risk factors as well as losses in relation to a portfolio of risk factors. As part of the study, we also define various measures and visualization techniques to evaluate the performance of the candidate models in the context of risk management and introduce two innovations: 1) tail emphasized model optimization and 2) implied covariance forecasting. Finally, we highlight the important issue of the estimation error of the covariance matrix in relation to its dimension and the number of datum from which it is estimated and outline a framework for handling this problem.

1. Introduction

To measure the market risk of a portfolio of traded assets, many banks are increasingly employing internal models based on a methodology called Value-at-Risk (VaR). In an important regulatory innovation [1], the Basle Committee on Banking Supervision has proposed in 1996 that such models may be used in the determination of the capital requirements that banks must fulfill to back their trading activities. In addition, the Bank for International Settlements – Fisher Report [2] urged financial intermediaries to publicly disclose Values-at-Risk. The Group of thirty [3] and ISDA also advocate the VaR concept as a general measure of market risk.

The motivation behind the use of these models is to quantify risk for internal decision making and capital adequacy purposes. The quantification is in terms of the upper limit for loss inherent to a portfolio position with a given pre-specified probability, called *level of confidence*, and over a fixed period of time in the future *holding period*. An in depth description of the theory of VaR models is given in [4], [5], [6] and [7].

VaR models have to deal with 4 mathematical components:

1. Models – for predicting conditional distributions of returns for the underlying risk factors.
2. Stripping – of instruments so that they are expressed in terms of their underlying risk factors.
3. Value function of the portfolio – which may be non-linear with respect to the returns of the underlying.
4. Construction of profit and loss forecast distribution – based on the value function and class of conditional distribution of risk factors.

VaR models may differ from each other in different ways with respect to one or more of the above components. Components 3 and 4 are often mathematically interlinked – and together may be broadly addressed with the following popular methodologies:

– Variance-Covariance – or Δ approach.
– The $\Delta - \Gamma$ approach.
– Monte-Carlo simulation.
– Scenario simulation.

One of the objectives of this paper is to compare different models for predicting market risk. For this purpose we have chosen a candidate portfolio of spot foreign exchange and metal rates. Since the candidate portfolio is linear – the Variance-Covariance approach for the portfolio function is more than adequate for the purpose of this paper.

There is a considerable amount of work related to VaR models from academics, practitioners and regulators. Traditionally, a large volume of work from academics has been in the field of stochastic modeling of financial time series. This huge volume of work – review article [8] containing 230 references – while being extremely relevant is not directly presented in the context of risk management. Academic work directly addressing the field of risk management is less common but includes a broad range of topics of practical and theoretical relevance such as the formal properties and deficiencies of the VaR concept [9], the problem of actual versus risk neutral distributions [10], the discrepancies between the diversity of implementations of the same model – system risk [11], the determination of capital requirements with internal models [12], the appropriateness of VaR models as a management tool [13] and techniques for forecasting variances and covariances [14, 15, 16].

Important contributions to VaR models from practitioners include the breakthrough work of J. P. Morgan RiskMetrics giving a complete description of VaR models, criticisms, assumptions and methods [4], methods to judge the VaR model forecast quality [17], concepts that analyze the risk exposure of a portfolio by measuring the risk of the basic blocks of the total portfolio [18] and the probabilistic foundation of scenario simulation models [19].

The contributions by regulators in this field primarily focus on issues related to the integration of VaR models within BIS recommendations and the attendant problems faced in practical implementation of these recommendations. Consequently work by regulators focus on backtesting [20, 21] and on issues that relate VaR model performance to the BIS parameters [22], [23], [24] and [25].

This paper covers a cross section of issues focused on by academics, practitioners and regulators. From the academic perspective, we offer some innovations in stochastic modeling of financial time series - and offer a stringent testing framework for new models. From the regulators perspective, we introduce and discuss various measures for evaluating model performance in the context of risk management. From the practitioners perspective we carry out the entire evaluation using *real data* and try to construct performance measures which strike a balance between risk over-estimation and risk-under estimation.

In sections 2 through 6 of this paper, we evaluate the absolute and relative performance of VaR estimates based on historical simulation, the rectangular moving average, the exponential moving average, and the conditional variance forecasted by the GARCH process. In dealing with this issue we also discuss the meaning of *performance* in terms of different techniques to evaluate the success of the VaR model. In section 7, we discuss the issue of stochastic error in covariance matrix estimation and the inter-relation of this problem to the following issues:

- Possible singularities in the covariance matrix due to rank defects caused when the sample size is smaller than the dimension of the correlation matrix (which is generally true when using the popular figure of 250 days history).
- Systematic underestimation of risk arising out of the re-optimization of the portfolio (to minimize VaR) based on the estimated covariance matrix.

We end the discussion with some speculation on the innovations necessary to handle this problem. Although section 7 is directly unrelated to the earlier sections – we believe that it is a crucial point which must be emphasized for the completeness of this paper. We end the paper with a summary of innovations and recommendations for better overall handling of risk measurement.

2. General overview

Below, we present a general non-mathematical overview of this paper. The detailed mathematical description of the models and performance measures is in sections 3 and 4.

2.1 Data

This presentation is based on 2252 daily prices for the value of 1 United States Dollar (USD) expressed in the following 10 units:

- Swiss Francs (CHF)
- German Marks (DEM)
- French Francs (FRF)
- British Pounds (GBP)
- Italian Lira (ITL)
- Japanese Yen (JPY)
- Netherland Guilder (NLG)
- Swedish Kroner (SEK)
- Gold Ounces (XAU)
- Silver Ounces (XAG)

The 2252 prices $w_t^{(k)}$ corresponding to each series k are constructed by linearly interpolating the middle prices, constructed as a geometric mean of bid and ask [26], from the 2 nearest quotes which bracket 16:00 UKT. Of course, as a corollary to the linear interpolation – if a quote falls exactly on 16:00 UKT then the middle price from this quote is used. Only quotes which have been filtered using O&A filtering technology – and from an original high frequency database of O&A collected quotes from 3 real-time data vendors – Knight Ridder, Reuters and Telerate are used. The daily prices do not include Saturdays, Sundays, Christmas days, December 26 and New years day. The first price in this data set belongs to February 23, 1988 and the last price belongs to October 31, 1996.

2.2 Portfolio specification

Introduction: The performance measures of the risk models considered in this paper are presented in 2 ways:

Univariate: As an average performance over the 10 trivial portfolios that are defined by the 10 risk factors k – as being the sole asset in the portfolio.

Multivariate: In conjunction with a candidate portfolio P of 10 risk factors itemized in section 2.1 and having equal weights.

We shall hereafter use *univariate* and *multivariate* to describe the context of the performance measures as defined above.

The entire analysis is done assuming zero mean price change – but to account for any bias due to long term increases or decreases in one or more of the time series – the performance information is symmetrized by taking the mean performance over the 2 portfolio position vectors defined by:

Long on the US dollar with equal investment in each risk factor.

Short on the US dollar with equal investment in each risk factor.

The above *symmetrization* procedure applies equally to the *univariate* as well as the *multivariate* presentations.

Univariate Context: Taking into account the *long* and *short* positions, in the *univariate* case we introduce notation $\mu^{\pm(k)}$ to denote a generic measure, μ, associated with a specific portfolio and position. Using this notation we can than construct the mean measure, μ_u, in the *univariate* (u) context as

$$\mu_u = \frac{1}{20}\left(\sum_{k=1}^{10} \mu^{+(k)} + \sum_{k=1}^{10} \mu^{-(k)}\right) \tag{2.1}$$

where each of the $\mu^{\pm(k)}$ is an entirely univariate construct related to the 20 distinct *univariate* portfolios with respect to which loss is estimated by a candidate model.

Multivariate Context: Taking into account the *long* and *short* positions, in the *multivariate* case we introduce notation $\mu^{\pm(P)}$ to denote a generic measure, μ, associated with a specific portfolio and position. Using this notation, we can then construct the mean measure, μ_m, in the *multivariate* (m) context as

$$\mu_m = \frac{1}{2}\left(\mu^{+(P)} + \mu^{-(P)}\right) \tag{2.2}$$

where each $\mu^{\pm(P)}$ is an intrinsically multivariate construct related to the 2 distinct portfolios with respect to which loss is estimated by a candidate model.

Comment: The performance measures we present in this paper are all computed as an average over candidate portfolios. Such averaging is necessary so that the performance of the model is not biased by the choice of the portfolio. In particular, long term fall or rise in prices, can by chance make a fixed portfolio consistently profitable or lossy. To prevent bias of this nature we always average over pairs of portfolios with opposing positions.

2.3 Methodology

The 2252 prices $w_t^{(k)}$ for each of the risk factor series[1] k are then log differenced to produce 2251 consecutive log differenced price changes

$$x_{t+1}^{(k)} = \ln(w_{t+1}^{(k)}) - \ln(w_t^{(k)}). \tag{2.3}$$

In the *univariate* context, the primary focus of each candidate model is to predict the log difference series $x_{t+1}^{(k)}$ in terms of a conditional probability density function denoted by $p_t^{(k)}$. In the *multivariate* context, however, the primary focus of each candidate model is to predict $x_t^{(P)}$ defined by

$$x_t^{(P)} = \frac{1}{10}\sum_{k=1}^{10} x_t^{(k)} \tag{2.4}$$

in terms of a conditional probability density function denoted by $p_t^{(P)}$.

[1] Data set RISK97 is available from Olsen & Associates – http://www.olsen.ch

It is convenient at this point to establish the following nomenclature and notational guidelines.

- Price changes are numbered from -1249 to 1001. Specific, price change events are described as a subset of $t \in [-1249, 1001]$.
- The notations (k) and (P) are used as a superscript to reference the *univariate* portfolios k and the *multivariate* portfolio P respectively.
- The notation (k, P) is used as a superscript in an expression valid in both contexts.
- The superscripts $-$ and $+$ are used to reference the *long* and *short* positions of a portfolio.
- The superscript $\pm$ is used in an expression valid for both positions.
- The nomenclature *prediction-realization pair* refers to the ordered mathematical object $(p_t^{(k,P)}, x_{t+1}^{(k,P)})$. It should be kept in mind that by definition – $p_t^{(k,P)}$ are the forecast distributions for realizations $x_{t+1}^{(k,P)}$ and depend only on data until t.

The first 1250 of the price changes $x_t^{(k,P)}$, $(t \in [-1249, 0])$, constitute the in-sample period of those models requiring buildup and/or optimization. The final 1001 data points, $t \in [1, 1001]$ are used to construct 1000 out-of-sample *prediction-realization pairs*. This latter set of 1000 pairs are then analyzed in different ways to evaluate the performance of the models.

2.4 Zoo of models

We study the following 5 different models for generating the forecast distribution p_t over a *holding period* of 1 day:

1. Historical simulation with 250 day memory
2. 250 day rectangular moving average
3. J. P. Morgan RiskMetrics – Exponential moving average
4. GARCH(1, 1)
5. Tail emphasized GARCH(1, 1)

Model 1 is non-parametric and models 2 and 3 are pre-defined and do not require any optimization. Models 4 and 5 are optimized in an in-sample period of 1250 days (250 days for build up and 1000 days for optimization). The latter in-sample period precedes a 1000 day out-of-sample in which the performance of the models is gauged in the context of risk management.

The determination of the conditional distribution $p_t^{(k)}$ forecast by a model for each of the series k is of course an entirely univariate exercise. The multivariate aspect comes through the covariance matrix Σ_t which is determined by bivariate analysis – for all but model 1. In the case of models 2 and 3 the covariance is measured in a straightforward - intuitively obvious – bivariate generalization of the variance computation (details in sections 3.2 and 3.3). In models 4 and 5 however – a more sophisticated approach is presented - wherein the dynamics of the bivariate sum are also fit to the candidate model and the implied covariance is used (details in section 3.4). Another innovative feature (described in section 3.5) is introduced in model 5 – wherein the optimization of the model is performed using a specialized method which emphasizes optimization of the residual tails. This optimization methodology turns out to have dramatic impact when using these models in the context of risk-management – as the performance measures will show.

2.5 Performance measures

The purpose of this paper in not simply the study of the different candidate models itemized in section 2.4 but also to present the different subtleties in evaluation of models in the context of risk management. For better visualization of performance, the above measures are plotted for each model as a function of increasingly extreme classes of realizations. To construct the classes of extreme realizations we use two different parameters such that – as they increase – they define a class of more and more extreme observations. These are:

Volatility percentile: The candidate measure is plotted against the percentile level of $|x_t^{(k,P)}|$ – and for a given level the measure is evaluated as a count or average over those realizations which correspond to a higher percentile level.

Confidence level: The candidate measure is plotted against a confidence level – and for a given level the measure is evaluated as a count or average over those realizations which are outside the confidence level.

There is a subtle difference between the two kinds of plots in terms of how the X-axis refers to an increasingly extreme set of points. The first plot directly evaluates the performance of the model for large price changes. The second evaluates the performance of the model for movements which are regarded as increasingly improbable by the model itself.

From the latter, it should be clear that – when comparing plots for different models against the volatility percentile – a given point on the X-axis will always refer to the same data. On the other hand, for plots against the confidence level – a point on the X-axis will refer to differing data sets since this discrimination depends on the model itself.

The performance measures presented in this paper are:

1. Observed/Predicted exceedence ratio against confidence level
2. BIS Red, Yellow and Green Zone frequency against confidence level
3. Observed/Predicted serial exceedence against confidence level
4. Mean log likelihood against confidence level
5. Mean log likelihood against volatility percentile

As already mentioned in section 2.2, we have one further variation in the presentation of performance measures – the *univariate* context based on 10000 *prediction-realization pairs* and the *multivariate* context based on 1000 *prediction-realization pairs.* The two contexts allow us to distinguish between the performance of a model for forecasting a risk factor with the performance of the model for forecasting the portfolio. Knowledge of the relative performance of the model in the *univariate* and *multivariate* context highlights an important distinction. It is usually the case that the methodology used for the multivariate computation $p_t^{(P)}$ as prescribed by a given model can be refined without affecting the univariate computation $p_t^{(k)}$. (For all the models presented in this paper – barring historical simulation – this means that it is possible to modify the methodology for updating the off diagonal elements of the covariance matrix Σ_t without changing the prescription for the diagonal elements.) The relative performance of a given model in the *univariate* and *multivariate* context can indicate if the point of weakness for the model lies in the computation of $p_t^{(P)}$ and can therefore motivate further research in rectifying the point of weakness.

3. Model specification

Presented below are the detailed specifications for the construction of the forecast probability density $p_t^{(k)}$ and $p_t^{(P)}$ as prescribed by the different models included in this paper.

3.1 Historical Simulation

Historical simulation is the method of prediction in terms of which the forecast probability density depends directly on the past empirical distribution. In general the length of the past used for forecasting can be a parameter – which is adjusted for best performance in the in-sample period. The Bank of International Settlements in Basle (BIS), Switzerland, has however recommended [1] the use of at least 250 day memory. In view of this we present this model with a 250 day memory since the comparative performance of the BIS recommendation is likely to be of general topical interest.

Since no optimization is involved, the 10000 *univariate* and 1000 *multivariate prediction realization pairs* are constructed using the last 1250 data points $t \in [-248, 1001]$ in the 10 log differenced price change series k.

To construct $p_t^{(k,P)}$ we first define

$$X_t^{(k,P)} = \{x_t^{(k,P)}, x_{t-1}^{(k,P)}, \ldots, x_{t-249}^{(k,P)}\} \tag{3.1}$$

as the set of 250 points to be used in constructing $p_t^{(k,P)}$ at time t. Then we introduce ranking notation using parenthetical subscripts (to be distinguished from $x_t^{(k,P)}$) such that

$$x_{(i,t)}^{(k,P)} \leq x_{(i+1,t)}^{(k,P)}, \quad i \in [1, 250], \quad x_{(i,t)}^{(k,P)} \in X_t^{(k,P)}. \tag{3.2}$$

Thus $x_{(1,t)}^{(k,P)}$ is the minimum and $x_{(250,t)}^{(k,P)}$ is the maximum element in $X_t^{(k,P)}$.

We then construct $p_t^{(k,P)}$ to be the probability density function such that:

$$\int_{-\infty}^{x_{(i,t)}^{(k,P)}} p_t^{(k,P)}(x)dx = \frac{i-\frac{1}{2}}{250} \quad \forall i \in [1,250]. \tag{3.3}$$

The above density function is assumed to remain constant between $x_{(i,t)}^{(k,P)}$ and $x_{(i+1,t)}^{(k,P)}$ so that:

$$\int_{x_{(i,t)}^{(k,P)}}^{x_{(i+1,t)}^{(k,P)}} p_t^{(k,P)}(x)dx = \frac{1}{250}. \tag{3.4}$$

If $x_{(1,t)}^{(k,P)} < x_{t+1}^{(k,P)} < x_{(250,t)}^{(k,P)}$ then the probability density and fractile value associated with $x_{t+1}^{(k,P)}$ is well defined. If however $x_{t+1}^{(k,P)} < x_{(1,t)}^{(k,P)}$ or $x_{t+1}^{(k,P)} > x_{(250,t)}^{(k,P)}$ the probability density function is not defined and the fractile value of such an event cannot be determined. To resolve this – we extend this model by asserting that $p_t^{(k,P)}$ has Gaussian tails and hence the form implied by a normal distribution with mean $\bar{x} = \bar{x}_t^{(k,P)}$ estimated from set $X_t^{(k,P)}$ and where $\sigma = \sigma_t^{l(k,P)}$ for the left side of the distribution which satisfies

$$\int_{-\infty}^{x_{(1,t)}^{(k,P)}} \frac{1}{\sqrt{2\pi}\ \sigma_t^{l(k,P)}} \exp\left(-\frac{1}{2}\left[\frac{x-\bar{x}}{\sigma_t^{l(k,P)}}\right]^2\right) dx = \frac{1}{500} \tag{3.5}$$

and with $\sigma = \sigma_t^{r(k,P)}$ for the right side of the distribution which satisfies

$$\int_{x_{(250,t)}^{(k,P)}}^{\infty} \frac{1}{\sqrt{2\pi}\ \sigma_t^{r(k,P)}} \exp\left(-\frac{1}{2}\left[\frac{x-\bar{x}}{\sigma_t^{r(k,P)}}\right]^2\right) dx = \frac{1}{500} \tag{3.6}$$

This completely defines the construction of the forecast distribution $p_t^{(k,P)}$ and through this the out-of-sample *prediction-realization pairs* that form the starting point for performance assessment of this model.

3.2 Rectangular Moving Average

Since the horizon is already fixed by the data (and the *holding period*) at 1 day, the only remaining parameter one can associate with this method is the size of the memory. Once again – because of the commonly used 250 day memory (consistent with the BIS recommendation) we do not undertake any optimization and use the model with this pre-defined value. As in the case of the historical simulation model – the 1000 *univariate* and *multivariate prediction realization pairs* are constructed using the last 1251 data points in the 10 log differenced price change series k.

Univariate context: In this model, as in all subsequent models described below, the conditional distribution $p_t^{(k)}$ is predefined to be a Gaussian distribution with variance ${\sigma_t^{(k)}}^2$. The difference between this and subsequent models lies only in the manner $\sigma_t^{(k)}$ is generated from the price change history of a given series. For this model:

$$ {\sigma_t^{(k)}}^2 = \frac{1}{250} \sum_{i=0}^{249} {x_{t-i}^{(k)}}^2 \tag{3.7} $$

and this along with $x_t^{(k)}$ constitutes the full specification of the *prediction-realization pairs* that form the starting point of the *univariate* performance analysis of the model.

Multivariate context: In this case we define a multivariate object, the covariance matrix Σ_t, determined from the moving past 250 day history of the 10 series k. The elements of Σ_t are determined by:

$$ s_t^{(jk)} = s_t^{(kj)} = \frac{1}{250} \sum_{i=0}^{249} x_{t-i}^{(j)} x_{t-i}^{(k)}. \tag{3.8} $$

Note that the diagonal elements ($j = k$) are exactly the variances ${\sigma_t^{(k)}}^2$ obtained in the *univariate context.*

Then, using standard multivariate probability theory the forecast distribution $p_t^{(P)}$ for $x_t^{(P)}$ is the Gaussian distribution with variance:

$$ {\sigma_t^{(P)}}^2 = \frac{1}{100} \left(\sum_{k=1}^{10} s_t^{(kk)} + 2 \sum_{k=1}^{10} \sum_{j<k}^{10} s_t^{(jk)} \right). \tag{3.9} $$

The determination of $\sigma_t^{(P)}$ as above along with $x_t^{(P)}$ constitutes the full specification of the *prediction-realization pairs* that form the starting point of the *multivariate* performance analysis of the model.

3.3 J. P. Morgan RiskMetrics

This model is basically a non-stationary GARCH$(1, 1)$ model. The GARCH coefficients are pre-defined by the RiskMetrics technology – and so no optimization is necessary. To be consistent with the previous 2 models – we choose a build up period of 250 days – although the effective model memory is much shorter due to the exponentially decaying weight of the moving average that the GARCH$(1, 1)$ process emulates. As in the previous 2 models – the analysis is based on the 1000 *prediction-realization pairs* constructed from the last 1251 data points in the 10 log differenced price change series k. For convenience we define $t \in [-249, 0]$ as representing the build up period.

Univariate context: In this model the variance of the Gaussian distribution $p_t^{(k)}$ is given by the recursive formula:

$$\sigma_t^{(k)\,2} = 0.94\sigma_{t-1}^{(k)\;2} + 0.06x_t^{(k)\,2}, \quad \sigma_{-249}^{(k)} = |x_{-249}^{(k)}|. \tag{3.10}$$

The value of $\sigma_t^{(k)}$ is not sensitive to the seed value $\sigma_{-249}^{(k)}$ once the recursion has been applied 250 times or more. Of course, for the latter to be true, the seed value must be a reasonable value like the one used in expression 3.10. The repeated application of the recursive expression 3.10 will allow the specification of 1000 *prediction-realization pairs* starting with $(p_1^{(k)}, x_2^{(k)})$ and ending with $(p_{1000}^{(k)}, x_{1001}^{(k)})$. These pairs are the starting point of the *univariate* performance analysis of the model.

Multivariate context: As in the previous model, here again we construct the covariance matrix Σ_t. The matrix elements are generated with the intuitively obvious bivariate generalization of the univariate case:

$$s_t^{(jk)} = s_t^{(kj)} = 0.94s_{t-1}^{(jk)} + 0.06x_t^{(j)}x_t^{(k)}. \tag{3.11}$$

Then, using standard multivariate probability theory the forecast distribution $p_t^{(P)}$ for $x_{t+1}^{(P)}$ is constructed from Σ_t in the same manner as expressed at the end of section 3.2 (expression 3.9).

3.4 GARCH(1, 1)

The GARCH process is a popular stochastic process which has been fairly successful in modeling financial time series [27]. In general the GARCH(p, q) process has $p+q+1$ parameters which must be fit to the data. GARCH$(1, 1)$ is the simplest of this class with 3 parameters.

The GARCH$(1, 1)$ based risk model is the first model we present for which optimization is involved – and we make full use of all 2251 data points available to us from each of the 10 log differenced price change series k. For convenience we define $t \in [-1249, -1001]$ as representing the build up period, $t \in [-1000, 0]$ as representing the in-sample period and $t \in [1, 1001]$ as the period over which we construct the 1000 *prediction-realization pairs* associated with this model.

Univariate context: In this model the variance of the Gaussian distribution p_t is given by the recursive formula:

$$\sigma_t^{(k)\,2} = \alpha_0^{(k)} + \alpha_1^{(k)}x_t^{(k)\,2} + \beta_1^{(k)}\sigma_{t-1}^{(k)\;2}, \quad \sigma_{-249}^{(k)} = |x_{-249}^{(k)}|. \tag{3.12}$$

The recursion is initialized using the same seed value as for the J. P. Morgan RiskMetrics case. As discussed in section 3.3, we emphasize that the built up values $\sigma_t^{(k)}$ for $t \in [1, 1000]$ are not sensitive to this arbitrary but reasonable choice.

We have already stated that the J. P. Morgan RiskMetrics model is a GARCH$(1, 1)$ model. What is different here, however, is that we have separately optimized the GARCH$(1, 1)$ for each of the k risk-factors in the in-sample period. Unlike the J. P. Morgan prescription of using $\alpha_0^{(k)} = 0$, $\alpha_1^{(k)} = 0.06$ and $\beta_1^{(k)} = 0.94$ for all k risk factors, in our case we used the values reported in Table 1.

The optimization – reported in Table 1 – was done with the objective of maximizing the average log-likelihood of the 1000 *prediction-realization pairs* in the in-sample period – corresponding to the events $t \in [-1000, 0]$. Mathematically, the average log-likelihood may be expressed as:

$$\bar{\mathcal{L}}^{(k)} = \frac{1}{1000} \sum_{t=-1000}^{-1} \mathcal{L}_t^{(k)} = \frac{1}{1000} \sum_{t=-1000}^{-1} \ln(p_t^{(k)}(x_{t+1}^{(k)})) \tag{3.13}$$

where, $p_t^{(k)}(x_{t+1}^{(k)}) = p_t(\sigma_t^{(k)}, x_{t+1}^{(k)})$ is the probability density of the realized event in terms of the forecast distribution and $\sigma_t^{(k)} = \sigma_t^{(k)}(\alpha_0^{(k)}, \alpha_1^{(k)}, \beta_1^{(k)})$ is the variance of the forecast distribution in accordance with the GARCH$(1,1)$ expression 3.12. For stationarity of the GARCH$(1,1)$ process, the optimization is done with constraint $\alpha_0^{(k)} + \alpha_1^{(k)} + \beta_1^{(k)} < 1$. (We note here that the J. P. Morgan RiskMetrics process equation violates this stationarity condition). The solution space was searched using the genetic algorithms method – but convergence was aided by strategic use of the BHHH algorithm [28] at the end of every generation[29].

With Table 1 and the GARCH$(1,1)$ recursion formula for the variance $\sigma_t^{(k)}$ of the Gaussian distribution $p_t^{(k)}$, we construct the 1000 *prediction-realization pairs* over the period $t \in [1, 1001]$ that is the starting point for the evaluation of this model in the *univariate* context.

Multivariate context Here – once again – we proceed via the construction of the covariance matrix Σ_t but in this case we introduce a different methodology. Instead of making the straightforward bivariate generalization of the GARCH$(1,1)$, we proceed by fitting the 45 sum series,

$$x_t^{(jk)} = x_t^{(j)} + x_t^{(k)}, \tag{3.14}$$

(note new notation to reference sum series) to the GARCH$(1,1)$. The fit is done in the same in-sample period and using the same mix of the genetics algorithm and BHHH approach used in the *univariate* context. The fit parameters for the 45 series are presented in Table 2.

The $\sigma_t^{(jk)}$ are then determined for the sum series using the same recursion expression and build up criteria used in the *univariate* context. The off-diagonal elements of Σ_t are then computed as an implied covariance:

$$s_t^{(jk)} = s_t^{(kj)} = \frac{\sigma_t^{(jk)} - \sigma_t^{(j)} - \sigma_t^{(k)}}{2}, \quad i \neq j. \tag{3.15}$$

The following additional specification of diagonal elements:

$$s_t^{(kk)} = {\sigma_t^{(k)}}^2, \tag{3.16}$$

where $\sigma_t^{(k)}$ is determined in the *univariate* context, concludes the full prescription on how Σ_t is generated in this model.

Finally, using standard multivariate probability theory the forecast distribution $p_t^{(P)}$ for $x_{t+1}^{(P)}$ is constructed from Σ_t in the same manner as expressed at the end of section 3.2 (expression 3.9). The 1000 *prediction-realization pairs*, $(p_t^{(P)}, x_{t+1}^{(P)})$, over the period $t \in [1, 1001]$ is the starting point for the evaluation of this model in the *multivariate* context.

	Garch(1,1)				Tail emphasized Garch(1,1)			
Series	$\alpha_0 \times 10^5$	α_1	β_1	LL	$\alpha_0 \times 10^5$	α_1	β_1	LL
CHF	0.1887	0.04185	0.9279	3.4410	0.6404	0.01669	0.9361	3.1113
DEM	0.1806	0.06171	0.9053	3.5187	0.9698	0.02959	0.8880	3.1732
FRF	0.1488	0.06580	0.9050	3.5659	0.9808	0.05182	0.8646	3.2155
GBP	0.1912	0.07187	0.8925	3.5427	0.9695	0.02959	0.8864	3.1809
ITL	0.1601	0.08730	0.8836	3.5648	0.9808	0.05182	0.8658	3.1927
JPY	0.1074	0.05991	0.9132	3.6761	0.9748	0.05153	0.8431	3.3196
NLG	0.1803	0.06020	0.9061	3.5275	0.9715	0.05133	0.8728	3.1813
SEK	0.0069	0.04783	0.9558	3.6714	0.9626	0.02959	0.8670	3.2376
XAG	1.2128	0.06719	0.8441	3.0388	0.8269	0.00749	0.9653	2.6915
XAU	0.0212	0.02793	0.9680	3.4656	0.4518	0.00840	0.9582	3.0685

Table 1: Fit parameters (α_0, α_1, β_1) for variance forecasts and the mean log-likelihood (LL) over the in-sample for GARCH$(1,1)$ (using 1000 points) and tail emphasized GARCH$(1,1)$ processes (using 500 points). See sections 3.4 and 3.5 for additional details. The series are as defined in section 2.1.

3.5 Tail emphasized GARCH(1,1)

The implementation of this model differs from the GARCH$(1,1)$ implementation only in the manner of optimization of the process over the 10 series $x_t^{(k)}$ (needed in the *univariate* context) and the additional 45 sum series $x_t^{(jk)}$ (needed in the *multivariate context*). Instead of maximizing the average of all 1000 $\mathcal{L}_t^{(k)}$ or $\mathcal{L}_t^{(jk)}$ (for the sum series) in the in-sample – we maximize only the average of the most adverse 500. In describing this method as *tail emphasized* GARCH$(1,1)$, we note that the tail referred to in this context is not the tail of the distributions of $x_t^{(k)}$ and $x_t^{(jk)}$, but the tail of the distributions of $\mathcal{L}_t^{(k)}$ and $\mathcal{L}_t^{(jk)}$. We however do expect considerable overlap of events ($t \in [-1000, 0]$) in the tails of the 2 distributions.

Since the conditional distributions $p_t^{(k,jk)}$ (notation (k, jk) indicating reference to the series k and the sum series jk) is Gaussian with variance $\sigma_t^{(k,jk)\,2}$ we can expand $\mathcal{L}_t^{(k,jk)}$, $t \in [-1000, -1]$, more concretely as:

$$\mathcal{L}_t^{(k,jk)} = -\frac{1}{2}\ln(2\pi) - \ln(\sigma_t^{(k,jk)}) - \frac{x_{t+1}^{(k,jk)\,2}}{2\sigma_t^{(k,jk)\,2}}. \tag{3.17}$$

It is clear from the above expression that the dominant contribution to adverse (low) values of $\mathcal{L}_t^{(k,jk)}$ come from the third term when the model underestimates the risk and predicts a small variance compared to the market move on the next day. It is expected that by optimizing with the objective

	Garch(1,1)				Tail emphasized Garch(1,1)			
Series	$\alpha_0 \times 10^5$	α_1	β_1	LL	$\alpha_0 \times 10^5$	α_1	β_1	LL
DEM+CHF	0.6877	0.05098	0.9184	2.8053	0.9193	0.01906	0.9675	2.4652
FRF+CHF	0.6354	0.05278	0.9178	2.8283	0.9927	0.01411	0.9675	2.4890
FRF+DEM	0.6543	0.06359	0.9053	2.8509	0.9926	0.01434	0.9665	2.4990
GBP+CHF	0.5459	0.05642	0.9178	2.8513	0.9925	0.01406	0.9667	2.5069
GBP+DEM	0.5400	0.06301	0.9102	2.8814	0.9913	0.01360	0.9657	2.5257
GBP+FRF	0.4976	0.06509	0.9095	2.9031	0.9884	0.01360	0.9647	2.5450
GBP+ITL	0.5374	0.07466	0.8984	2.9057	0.9901	0.01360	0.9647	2.5344
GBP+JPY	0.5536	0.05651	0.9056	3.0168	0.8425	0.00992	0.9644	2.6733
GBP+NLG	0.5330	0.06163	0.9116	2.8857	0.9914	0.01360	0.9654	2.5303
GBP+SEK	0.2070	0.05930	0.9320	2.9429	0.9432	0.01360	0.9649	2.5612
ITL+CHF	0.6234	0.05523	0.9157	2.8355	0.9914	0.01383	0.9673	2.4921
ITL+DEM	0.6732	0.06993	0.8979	2.8600	0.9912	0.01383	0.9665	2.5014
ITL+FRF	0.6211	0.07343	0.8960	2.8824	0.9926	0.01406	0.9652	2.5211
ITL+JPY	0.6008	0.05949	0.8991	3.0180	0.8425	0.00992	0.9642	2.6728
JPY+CHF	0.5489	0.03985	0.9261	2.9578	0.8897	0.01096	0.9648	2.6301
JPY+DEM	0.6172	0.04760	0.9114	2.9923	0.8582	0.00980	0.9649	2.6573
JPY+FRF	0.5591	0.05111	0.9099	3.0193	0.8426	0.00992	0.9639	2.6824
NLG+CHF	0.6784	0.04999	0.9196	2.8093	0.9194	0.01907	0.9673	2.4691
NLG+DEM	0.7250	0.06102	0.9055	2.8306	0.9915	0.01383	0.9678	2.4809
NLG+FRF	0.6535	0.06272	0.9058	2.8553	0.9927	0.01406	0.9665	2.5039
NLG+ITL	0.6812	0.06939	0.8977	2.8639	0.9915	0.01359	0.9664	2.5058
NLG+JPY	0.5970	0.04709	0.9128	2.9973	0.8583	0.00992	0.9646	2.6617
SEK+CHF	0.2163	0.04709	0.9438	2.8810	0.9912	0.01360	0.9660	2.5219
SEK+DEM	0.1637	0.05014	0.9437	2.9063	0.9802	0.01360	0.9654	2.5314
SEK+FRF	0.1435	0.05186	0.9429	2.9313	0.9649	0.01360	0.9646	2.5528
SEK+ITL	0.1439	0.05694	0.9385	2.9369	0.9432	0.01360	0.9652	2.5476
SEK+JPY	0.3178	0.04254	0.9349	3.0461	0.8165	0.00583	0.9658	2.7015
SEK+NLG	0.1574	0.04853	0.9454	2.9106	0.9866	0.01360	0.9650	2.5359
XAG+CHF	1.3111	0.03395	0.9057	2.7977	0.9193	0.01906	0.9680	2.4649
XAG+DEM	1.1672	0.04607	0.8969	2.8316	0.9928	0.01410	0.9671	2.4979
XAG+FRF	1.1351	0.04487	0.8978	2.8472	0.9911	0.01371	0.9666	2.5136
XAG+GBP	0.6276	0.03225	0.9375	2.8249	0.9914	0.01383	0.9678	2.4904
XAG+ITL	1.2375	0.04736	0.8904	2.8464	0.9913	0.01371	0.9667	2.5127
XAG+JPY	0.3264	0.02016	0.9616	2.8848	0.9900	0.01360	0.9648	2.5500
XAG+NLG	1.0297	0.04256	0.9068	2.8339	0.9914	0.01383	0.9672	2.5001
XAG+SEK	1.4667	0.05619	0.8673	2.8667	0.9912	0.01359	0.9660	2.5266
XAU+CHF	0.4364	0.02503	0.9462	2.9736	0.8897	0.00922	0.9650	2.6454
XAU+DEM	0.5821	0.03243	0.9261	3.0152	0.8269	0.00749	0.9655	2.6902
XAU+FRF	0.5088	0.03200	0.9298	3.0406	0.8198	0.00672	0.9647	2.7147
XAU+GBP	0.3598	0.03306	0.9406	3.0339	0.8267	0.00749	0.9649	2.6998
XAU+ITL	0.6591	0.04087	0.9104	3.0367	0.8199	0.00672	0.9651	2.7082
XAU+JPY	0.0177	0.01519	0.9824	3.1589	0.6367	0.00840	0.9655	2.8000
XAU+NLG	0.4568	0.02875	0.9383	3.0203	0.8267	0.00749	0.9652	2.6951
XAU+SEK	11.2385	0.09932	0.0000	3.0703	0.8132	0.00495	0.9642	2.7376
XAU+XAG	0.1431	0.01737	0.9776	2.6099	0.9908	0.01565	0.9756	2.2537

Table 2: Fit parameters (α_0, α_1, β_1) for variance forecasts and the mean log-likelihood (LL) over the in-sample for GARCH$(1,1)$ (using 1000 points) and tail emphasized GARCH$(1,1)$ processes (using 500 points). See sections 3.4 and 3.5 for additional details. The series are the sum $(+)$ of the series defined in section 2.1. This optimization is needed in the *multivariate* context to compute the implied covariances used in these methods.

of maximizing the average of the 500 most adverse $\mathcal{L}_t^{(k,jk)}$, the model will provide better day to day consistency in forecasting performance than the traditional method which includes all events with equal emphasis.

4. Performance measures

Whether for the *univariate* case or the *multivariate* case, each model specification ends with a clear recipe for constructing 1000 *prediction-realization pairs*, $(p_t^{(k,P)}, x_{t+1}^{(k,P)})$, $t \in [1, 1000]$. These *prediction-realization pairs* are the starting point for performance measurement of the models in the context of risk management.

4.1 Exceedence ratio against confidence level

Introduction: Basically this measure is a count of the events for which the loss of assets exceeds the loss predicted by the model as a function of the confidence level c. The count is normalized by the theoretical expectation of the exceedence count – and hence we refer to this as the exceedence ratio. This measure is a relatively coarse measure since it is not sensitive in distinguishing models which have the same exceedence count but a different *degree* of exceedence.

Specification: As a function of confidence level c (in %), it is convenient to first define the exceedence limits $\chi_t^{+(k,P)}(c)$ and $-\chi_t^{-(k,P)}(c)$ so that:

$$\int_{-\infty}^{\chi_t^{+(k,P)}(c)} p_t^{(k,P)}(x)dx = \int_{-\chi_t^{-(k,P)}(c)}^{\infty} p_t^{(k,P)}(x)dx = \frac{c}{100}. \tag{4.1}$$

It should be clear from expression 4.1 that $\chi_t^{+(k,P)}(c)$ is the loss that will be exceeded by a long (+) position on the US Dollar against series k or portfolio P with probability $(1 - \frac{c}{100})$. Similarly the exceedence limit $\chi_t^{-(k,P)}(c)$ is the correspondingly probable loss for a short position on the US Dollar against series k or portfolio P. We note that for all the models except historical simulation – the conditional distribution $p_t^{(k,P)}$ is symmetric around zero and $\chi_t^{+(k,P)}(c) = \chi_t^{-(k,P)}(c)$.

Next, we define the set of loss exceedence events $\Phi^{+(k,P)}(c)$ and $\Phi^{-(k,P)}(c)$ so that:

$$\Phi^{\pm(k,P)}(c) = \{t | \pm x_t^{(k,P)} > \chi_{t-1}^{\pm(k,P)}(c), t \in [2, 1001]\}. \tag{4.2}$$

Using sets $\Phi^{\pm(k,P)}(c)$ we can finally specify the exceedence ratio measure

$$r^{\pm(k,P)}(c) = \frac{\#(\Phi^{\pm(k,P)}(c))}{1000(1 - \frac{c}{100})}. \tag{4.3}$$

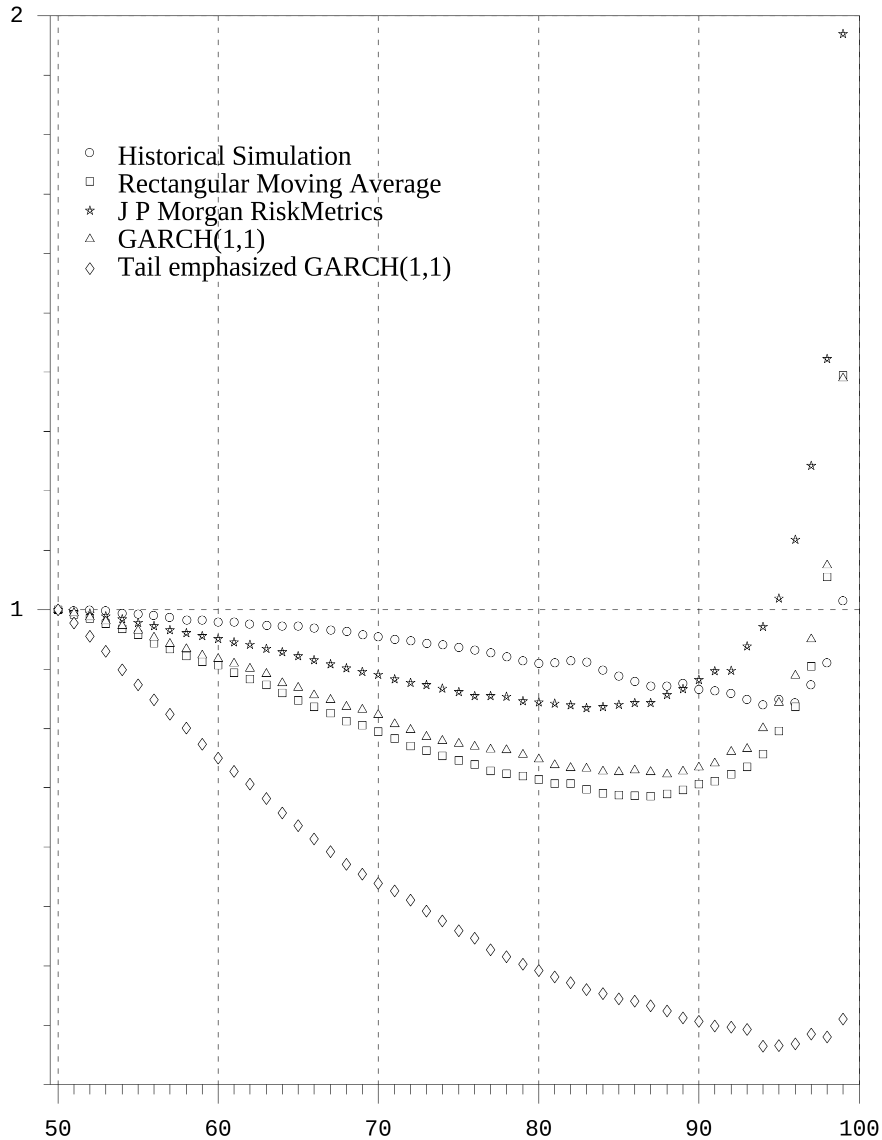

Figure 1u: Ratio of observed and expected loss exceedence count against confidence level in the *univariate* context for all candidate models. This plot shows the ratio (X axis) of the frequency of exceedences and the expected exceedence count against confidence level c (Y axis) for all the 5 models considered in this paper. Ratio values larger than unity indicate that the model underestimates risk at the confidence level for which this occurs. The value plotted for a given model is the average of the ratios computed over 20 trivial portfolios – each consisting of a *long* or *short* position of the US Dollar against the 10 series k. The curves are based on the analysis of 10000 *prediction-realization pairs* in the out-of-sample period. See section 6.3 for additional comments.

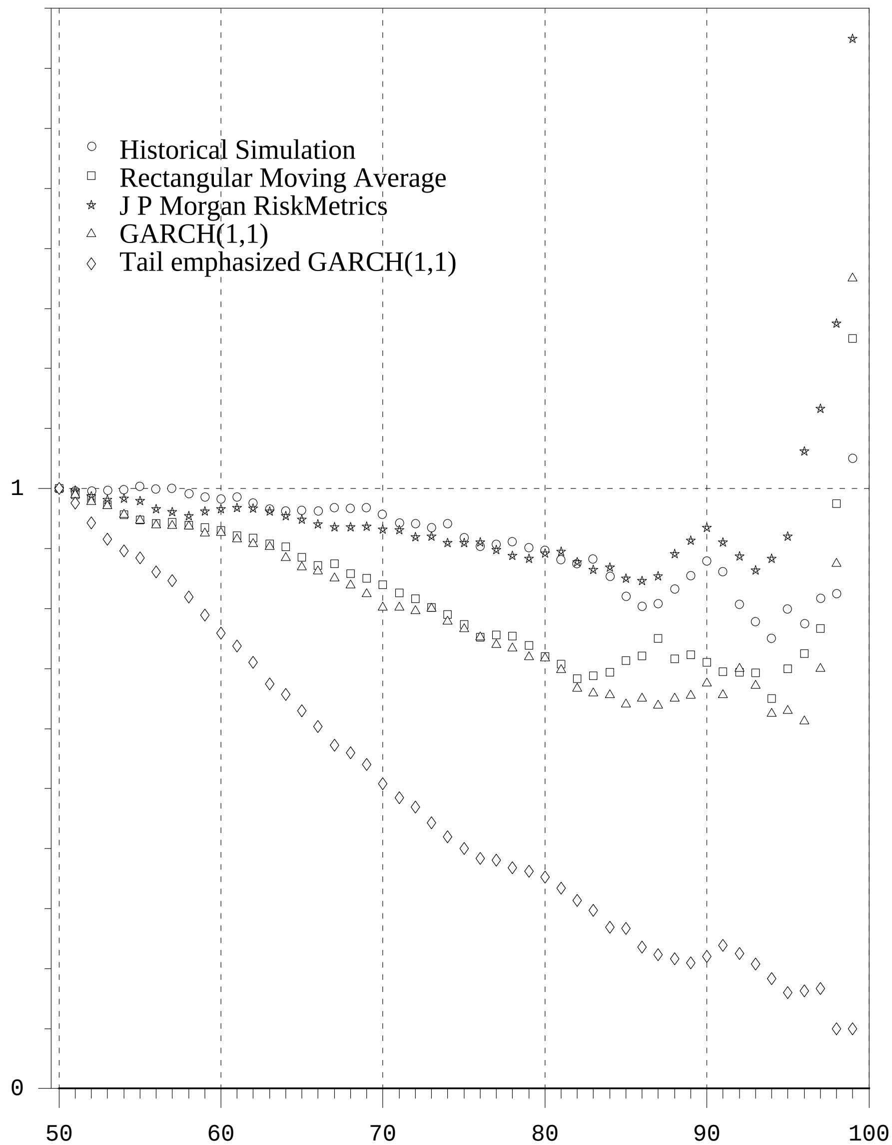

Figure 1m: Ratio of observed and expected loss exceedence count against confidence level in the *multivariate* context for all candidate models. This plot shows the ratio (X axis) of the frequency of exceedences and the expected exceedence count against confidence level c (Y axis) for all the 5 models considered in this paper. Ratio values larger than unity indicate that the model underestimates risk at the confidence level for which this occurs. The value plotted for a given model is the average of the ratios computed over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 1000 *prediction-realization pairs* in the out-of-sample period. See section 6.3 for additional comments.

Finally, as specified by expressions 2.1 and 2.2, we construct the *univariate* average exceedence ratio

$$r_u(c) = \frac{1}{20}\left(\sum_{k=1}^{10} r^{+(k)}(c) + \sum_{k=1}^{10} r^{-(k)}(c)\right) \tag{4.4}$$

and the *multivariate* average exceedence ratio

$$r_m(c) = \frac{1}{2}\left(r^{+(P)}(c) + r^{-(P)}(c)\right) \tag{4.5}$$

plotted in figures 1u and 1m respectively for all candidate models.

Comment: This is a simple and absolute measure which shows general success of a risk model as a function of the confidence level. The region above unity represents risk under-estimation and the region below unity represents risk over-estimation by the model.

Comparing figures 1u and 1m for the same model – can give an insight into the performance of the covariance forecast by the model. If the performance in the *multivariate* context drops drastically compared to the *univariate* context, this may indicate that the model may be salvaged by refining the methodology for covariance forecasting in conjunction with the model. (This is of course generally true when comparing any u figure with a corresponding m figure for the same model.)

4.2 BIS Colour frequency

Introduction: This measure is inspired by the BIS Green, Yellow and Red zone designations associated with a risk model defined as follows:

Green zone: A risk model is stated to be in this zone on day t if there are 4 or fewer events of losses beyond the 99% confidence level in the last 250 days – which is precisely the set $X_t^{(k,P)}$ defined in expression 3.1.

Yellow zone: A risk model is stated to be in this zone on day t if there are between 5 and 8 events of losses beyond the 99% confidence level in set $X_t^{(k,P)}$.

Red zone: A risk model is stated to be in this zone on day t if there are 9 or more events of losses beyond the 99% confidence level in set $X_t^{(k,P)}$.

The BIS uses the above zones as the basis for assessing the reliability of VaR models.

To construct our measure we first allow the prescribed confidence (99%) to become a variable c and then determine the frequency of the *red* and *green* zone designations associated with the model in the out-of-sample period as a function of c. Since our out-of-sample has 1000 *prediction-realization pairs*, we have 4 non-overlapping periods of 250 days but 751 distinct but overlapping moving samples, $X_t^{(k,P)}$, $t \in [251, 1001]$, in which to determine the frequency of the BIS colour designations. (We omit the *yellow* frequency from the presentation since the frequency of this middle designation leads to no clarity in the comparison of models.)

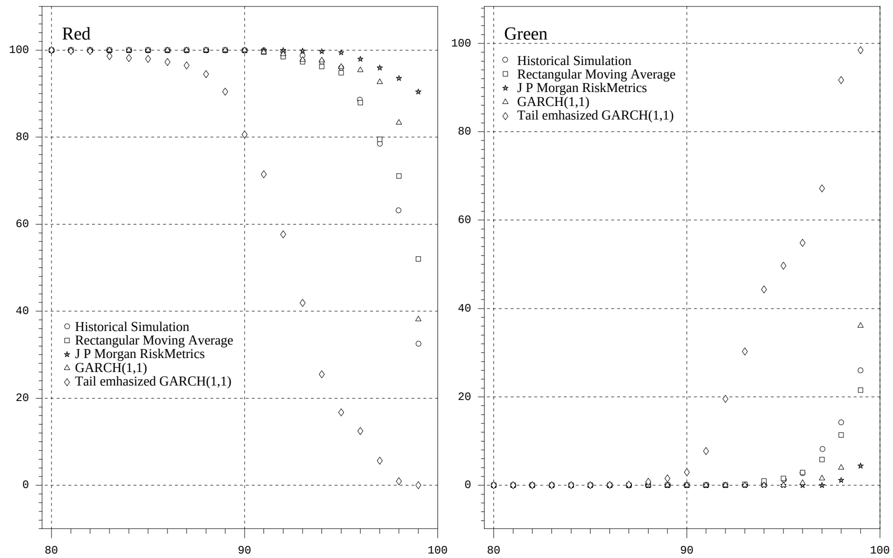

Figure 2u: BIS *red* and *green* zone frequency (Y axis) against confidence level (X axis) in the *univariate* context for all candidate models. The frequency is expressed as a fraction of the 751 days in the 1000 day out-of-sample period for which the the BIS *red* and *green* designations can be assigned to a model (249 days are needed for build up of this measure). The varying confidence level generalizes the BIS definition which is fixed at the 99% level. It is expected that the better model will approach the 100% green and 0% red frequencies faster (at lower confidence levels). The frequency plotted for a given model is the mean over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 10000 *prediction-realization pairs* in the out-of-sample period. See section 6.4 for additional comments.

Specification: Re-using expression 4.1 defining the exceedence level $\chi_t^{\pm(k,P)}$ and expression 3.1 defining the set $X_t^{(k,P)}$ we extend definition 4.2 so that

$$\Phi_t^{\pm(k,P)}(c) = \{t| \pm x_t^{(k,P)} > \chi_{t-1}^{\pm(k,P)}(c), t \in X_t^{(k,P)}\}. \tag{4.6}$$

Note that here we have defined sets Φ against a universal set of a moving sample of events rather than the entire out-of-sample set of events as expressed in 4.2. Next, using $\Phi_t^{\pm(k,P)}(c)$ defined as above, we can construct *booleans* $R_t^{\pm(k,P)}(c)$ and $G_t^{\pm(k,P)}(c)$ ($\in \{0,1\}$) such that:

$$(G_t^{\pm(k,P)}(c), R_t^{\pm(k,P)}(c)) = \begin{cases} (1,0) & \text{if } \#(\Phi_t^{\pm(k,P)}(c)) \leq 4 \\ (0,0) & \text{if } 5 \leq \#(\Phi_t^{\pm(k,P)}(c)) \leq 8 \\ (0,1) & \text{if } \#(\Phi_t^{\pm(k,P)}(c)) \geq 9 \end{cases} \tag{4.7}$$

Clearly, through expression 4.7, one can identify that $G_t^{\pm(k,P)}(99) = 1$ and $R_t^{\pm(k,P)}(99) = 1$ are the BIS defined colour designation for a risk model with

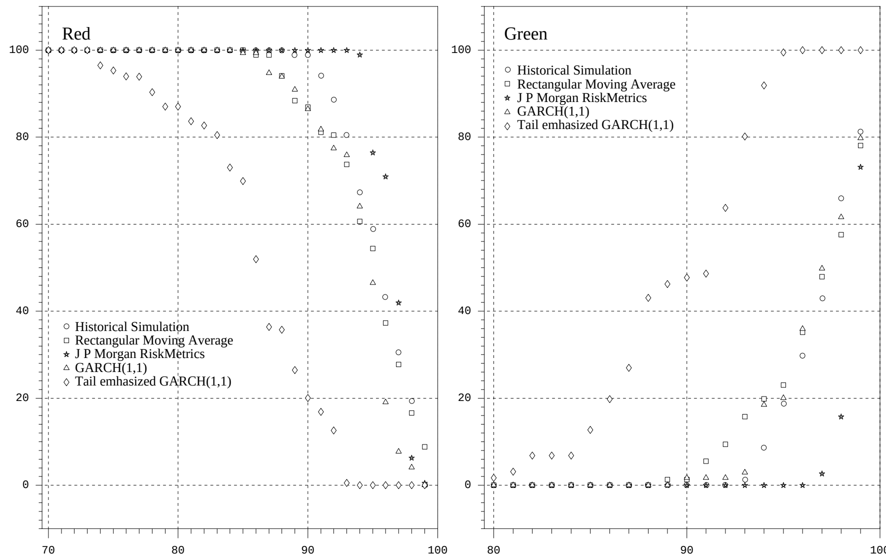

Figure 2m: BIS *red* and *green* zone frequency (Y axis) against confidence level (X axis) in the *multivariate* context for all candidate models. The frequency is expressed as a fraction of the 751 days in the 1000 day out-of-sample period for which the the BIS *red* and *green* designations can be assigned to a model (249 days are needed for build up of this measure). The varying confidence level generalizes the BIS definition which is fixed at the 99% level. It is expected that the better model will approach the 100% green and 0% red frequencies faster (at lower confidence levels). The frequency plotted for a given model is the mean over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 1000 *prediction-realization pairs* in the out-of-sample period. See section 6.4 for additional comments.

respect to the assets on day t. As usual, the portfolio represented by the superscripts covers the cases of the long (+) or short (−) position in the US Dollar with respect to the series k or portfolio P. Using the *booleans* defined by expression 4.7 we can construct the colour frequency measures

$$(G,R)^{\pm(k,P)}(c) = \frac{1}{751} \sum_{t\in[251,1001]} (G,R)_t^{\pm(k,P)}(c) \tag{4.8}$$

where the notation (G,R) transparently refers to 2 expressions – one each for $G^{\pm(k,P)}(c)$ and $R^{\pm(k,P)}(c)$ respectively. $(G,R)^{\pm(k,P)}(c)$ is of course the fraction of events in the out-of-sample period when the *green* and *red* designation is applicable. Note again that we have extended the BIS definition for this measure and allowed confidence level c to be a variable.

Finally, as specified by expressions 2.1 and 2.2, we construct the *univariate* average BIS colour frequency

$$(G,R)_u(c) = \frac{1}{20}\left(\sum_{k=1}^{10}(G,R)^{+(k)}(c) + \sum_{k=1}^{10}(G,R)^{-(k)}(c)\right) \tag{4.9}$$

and the *multivariate* average BIS colour frequency

$$(G,R)_m(c) = \frac{1}{2}\left((G,R)^{+(P)}(c) + (G,R)^{-(P)}(c)\right) \tag{4.10}$$

plotted in figures 2u and 2m respectively for all candidate models.

Comment: When c is low, it is naturally expected that the models will be predominantly 100% in the *red*. As c increases, *red* frequency drops and the *green* frequency increases. The *yellow* frequency – not plotted – peaks somewhere between 50% and 100% – and is different for different models. The X-axis of the plots are restricted to high c to highlight the portions which are most interesting for comparing the models.

This absolute measure does not carry much more information than the exceedence ratio presented in the last section. It also inherits the *digital* property of the exceedence ratio – in that it is not sensitive to the degree of exceedence. The main motivation of this measure is only to present exceedence information in the more popular framework of the BIS colour designations.

4.3 Serial exceedence ratio against confidence level

Introduction: Neither the exceedence ratio of section 4.1 nor the BIS colour frequency of the last section measure the clustering of confidence level exceedences. We now present the *serial exceedence ratio* which tests the propensity of the model for consecutive prediction failures.

Specification: In order to construct the measure we define the set of consecutive loss exceedence events $\Psi^{\pm(k,P)}(c)$ so that:

$$\Psi^{\pm(k,P)}(c) = \{t|\{t, t-1\} \subset \Phi^{\pm(k,P)}(c), t \in [3, 1001]\}. \tag{4.11}$$

Using sets $\Psi^{\pm(k,P)}(c)$ we can finally specify the serial exceedence ratio measure

$$s^{\pm(k,P)}(c) = \frac{\#(\Psi^{\pm(k,P)}(c))}{1000(1-\frac{c}{100})^2} \tag{4.12}$$

Finally, as specified by expressions 2.1 and 2.2, we construct the *univariate* average exceedence ratio

$$s_u(c) = \frac{1}{20}\left(\sum_{k=1}^{10} s^{+(k)}(c) + \sum_{k=1}^{10} s^{-(k)}(c)\right) \tag{4.13}$$

and the *multivariate* average exceedence ratio

$$s_m(c) = \frac{1}{2}\left(s^{+(P)}(c) + s^{-(P)}(c)\right) \tag{4.14}$$

plotted in figures 3u and 3m respectively for all candidate models.

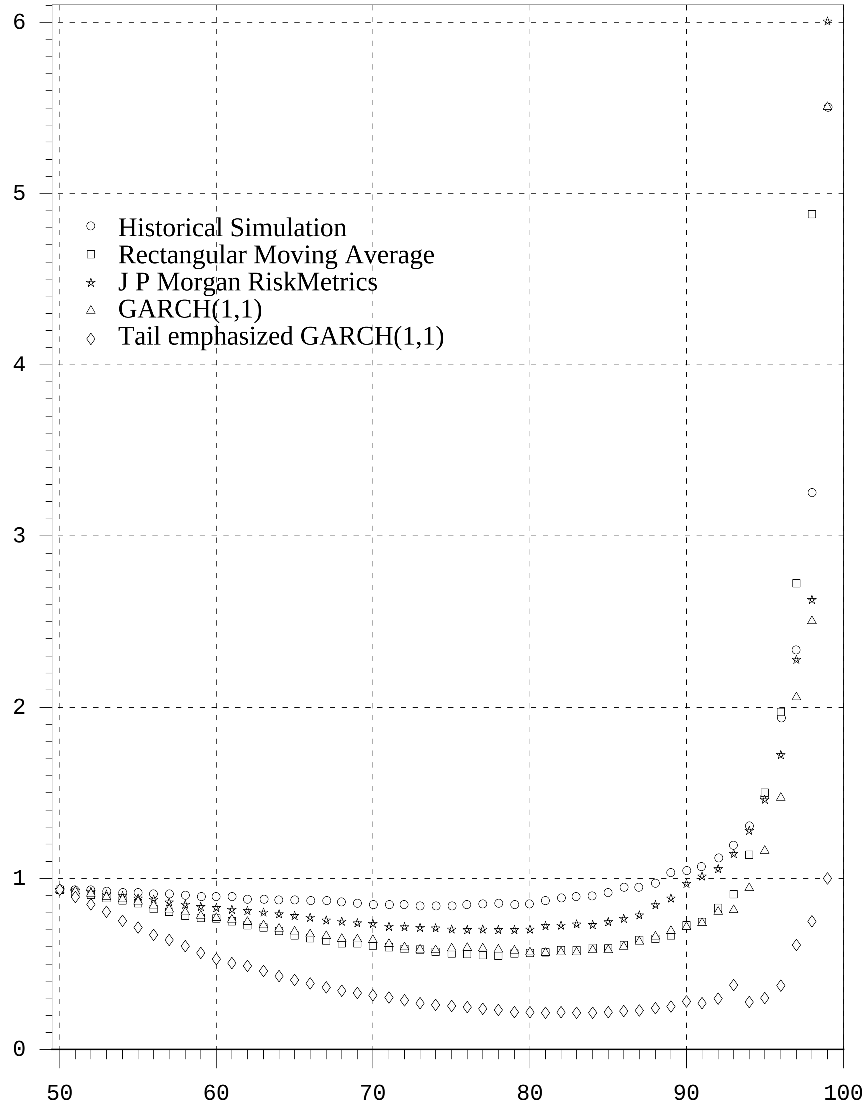

Figure 3u: Ratio of observed and expected consecutive loss exceedence counts against confidence level in the *univariate* context for all candidate models. This plot shows the ratio (X axis) of the frequency of consecutive exceedence events and the expected frequency of such events against the confidence level c (Y axis) for which the events are defined – for all the 5 models considered in this paper. Ratio values larger than unity indicate that the model is prone to greater than expected consecutive prediction failures at the confidence level for which this occurs. The value plotted for a given model is the average of the ratios computed over 20 trivial portfolios – each consisting of a *long* or *short* position of the US Dollar against the 10 series k. The curves are based on the analysis of 10000 *prediction-realization pairs* in the out-of-sample period. See section 6.5 for additional comments.

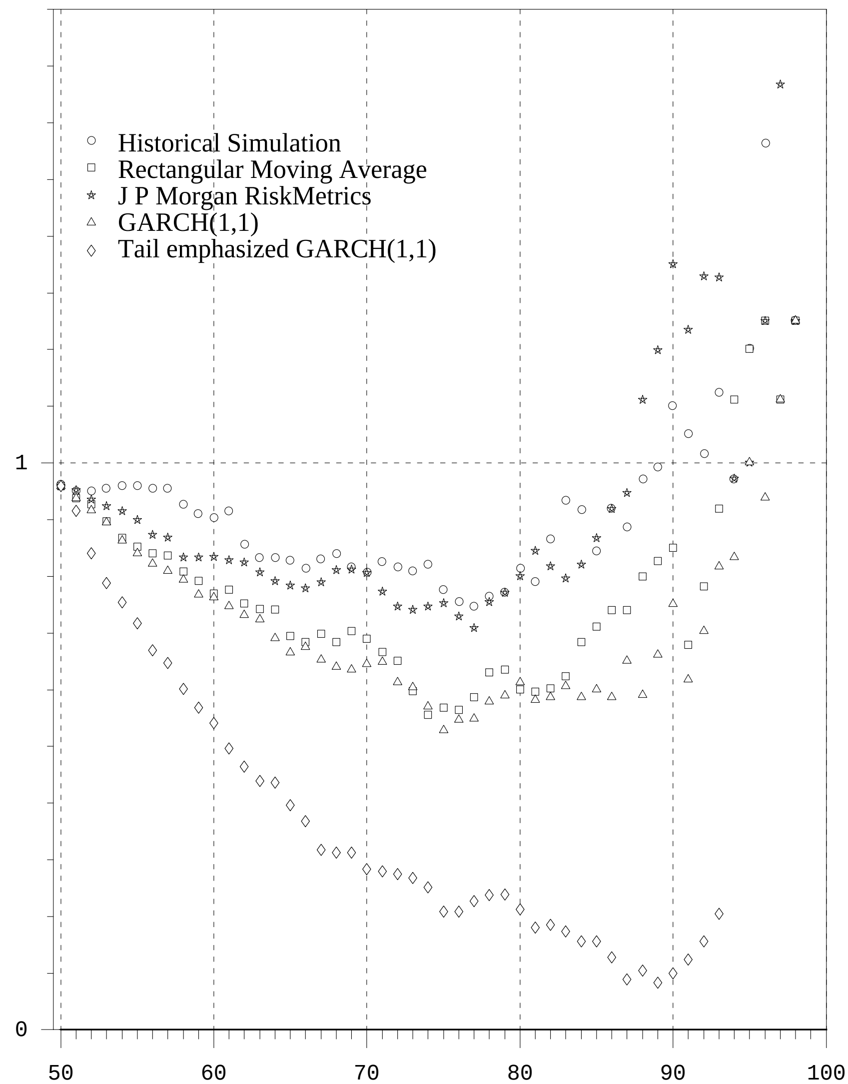

Figure 3m: Ratio of observed and expected consecutive loss exceedence counts against confidence level in the *multivariate* context for all candidate models. This plot shows the ratio (X axis) of the frequency of consecutive exceedence events and the expected frequency of such events against the confidence level c (Y axis) for which the events are defined – for all the 5 models considered in this paper. Ratio values larger than unity indicate that the model is prone greater than expected to consecutive prediction failures at the confidence level for which this occurs. The value plotted for a given model is the average of the ratios computed over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 1000 *prediction-realization pairs* in the out-of-sample period. See section 6.5 for additional comments.

Comment: This is a simple and absolute measure which shows general success of a risk model in capturing the heteroskedasticity of the market – as a function of the confidence level. The region above unity represents a higher than expected rate of successive exceedences.

4.4 Mean log-likelihood against confidence level

Introduction: The main drawback of the exceedence ratio as a performance measure is that it is only sensitive to the frequency and not the *degree* with which the loss exceeds that predicted at a certain confidence level. A natural measure for the *degree* of exceedence is the log-likelihood contribution

$$\mathcal{L}_t^{(k,P)} = \ln(p_t^{(k,P)}(x_{t+1}^{(k,P)})) \tag{4.15}$$

which is the logarithm of the probability density of the realized event in terms of the forecast probability distribution $p_t^{(k,P)}$. While we have used the mean log-likelihood as a criteria for optimization in the in-sample, we now extend it to measure performance out-of-sample. Although the expression 3.17 refers only to a Gaussian conditional distribution – it does qualitatively show through its third term that it is appropriate to regard the log-likelihood contribution $\mathcal{L}_t^{(k,P)}$ of a single event as a comparative (between models) measure of the *degree* of exceedence.

Specification: Using expressions 4.2 and 4.15 we can directly construct

$$\ell^{\pm(k,P)}(c) = \frac{\sum_{t\in\Phi^{\pm(k,P)}(c)} \mathcal{L}_{t-1}^{(k,P)}.}{\#(\Phi^{\pm(k,P)}(c))} \tag{4.16}$$

which is the mean log-likelihood of all loss exceedences over confidence level c for a long (+) or short (−) position in the US Dollar with respect to the series k or portfolio P.

Finally, as specified by expressions 2.1 and 2.2, we construct the *univariate* average mean log-likelihood

$$\ell_u(c) = \frac{1}{20}\left(\sum_{k=1}^{10} \ell^{+(k)}(c) + \sum_{k=1}^{10} \ell^{-(k)}(c)\right) \tag{4.17}$$

and the *multivariate* average mean log-likelihood

$$\ell_m(c) = \frac{1}{2}\left(\ell^{+(P)}(c) + \ell^{-(P)}(c)\right) \tag{4.18}$$

plotted in figures 4u and 4m respectively for all candidate models. Plots are always for the confidence range of 50% and higher, since this defines the division between loss making and profitable events. Due to *symmetrization* (consideration of both *long* and *short*, positions) the value of $\ell_u(c=50)$ is based on all 10000 out-of-sample events $x_t^{(k)}$ $(t \in [2,1001])$ and the value $\ell_m(c=50)$ is based on all 1000 out-of-sample events $x_t^{(P)}$ $t \in [2,1001]$ $(t \in [2,1001])$.

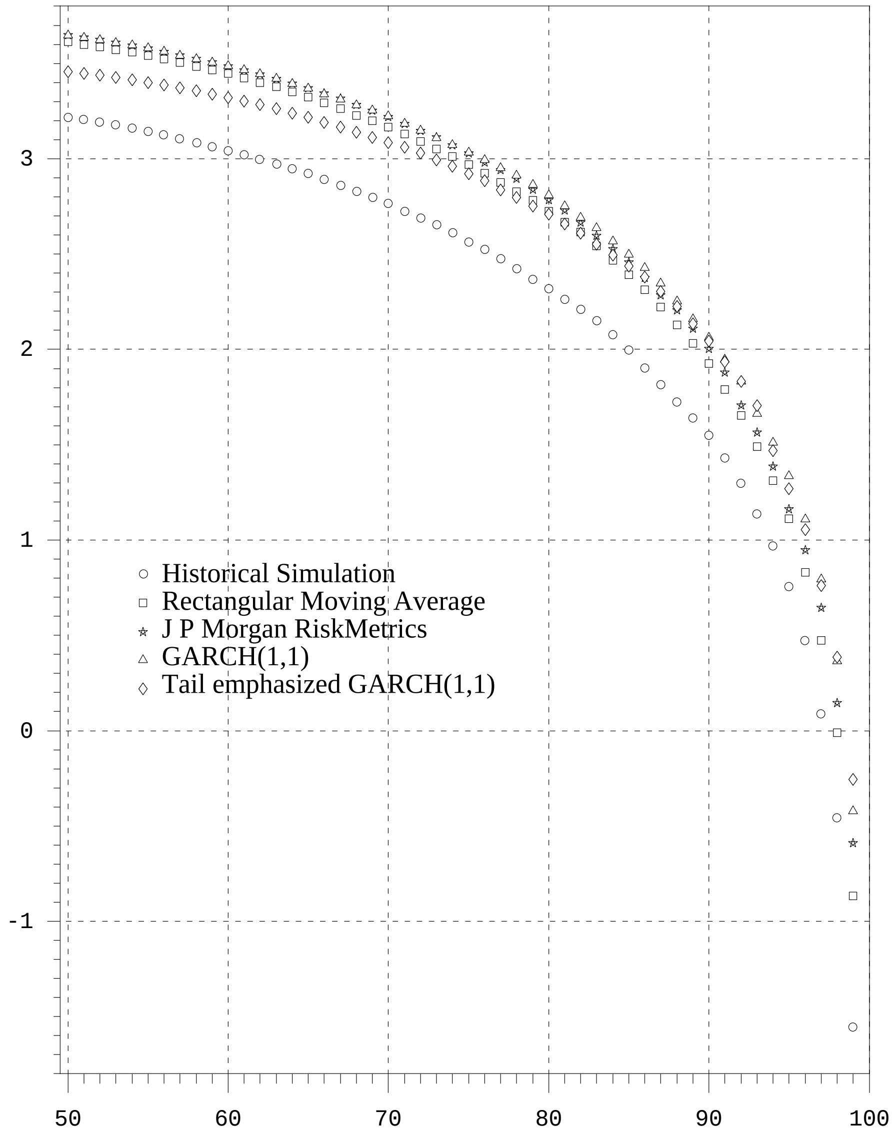

Figure 4u: Mean log-likelihood of exceedence events against confidence level in the *univariate* context for all candidate models. This plot shows the mean log-likelihood (X axis) of all exceedence events against the confidence level c (Y axis) for which the events are defined – for all the 5 models considered in this paper. Higher mean log-likelihood values at a given confidence level indicate better models and lower degrees of exceedence at that confidence level. The value plotted for a given model is the average of the mean log-likelihoods of exceedence events computed over 20 trivial portfolios – each consisting of a *long* or *short* position of the US Dollar against the 10 series k. The curves are based on the analysis of 10000 *prediction-realization pairs* in the out-of-sample period. See section 6.6 for additional comments.

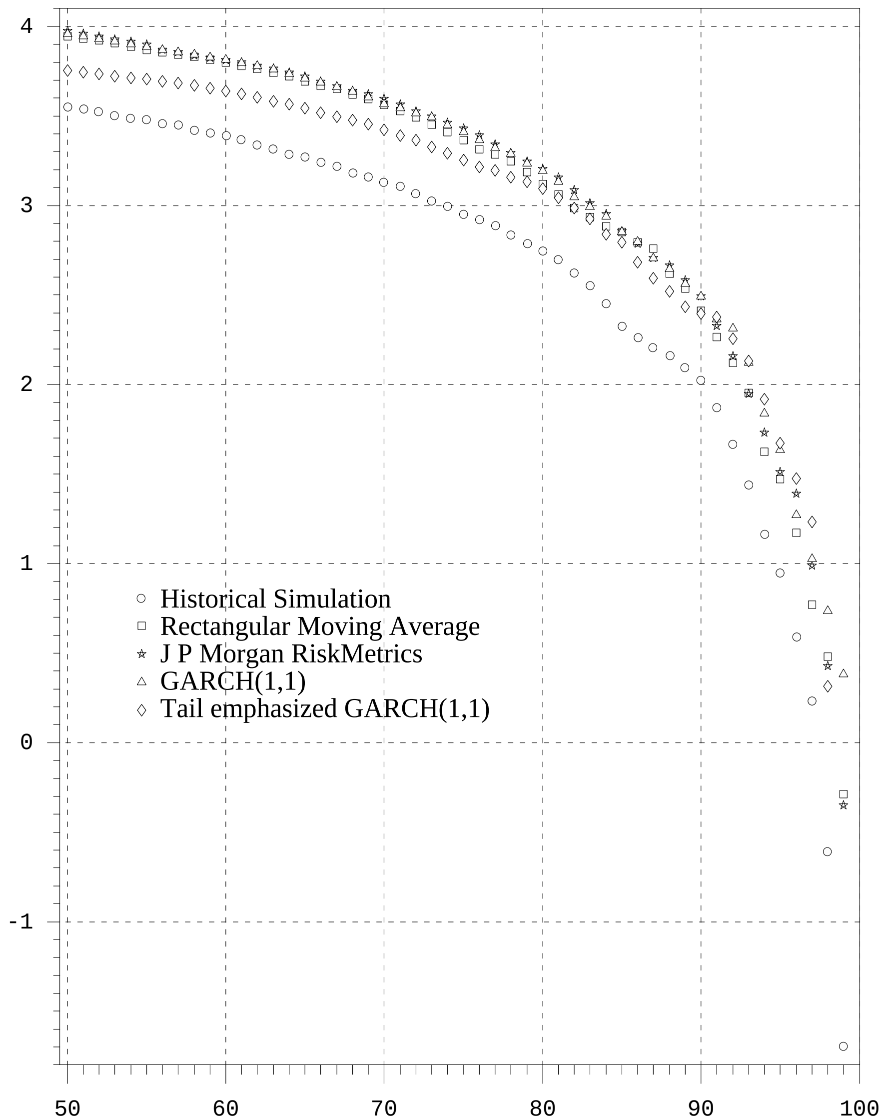

Figure 4m: Mean log-likelihood of exceedence events against confidence level in the *multivariate* context for all candidate models. This plot shows the mean log-likelihood (X axis) of all exceedence events against the confidence level c (Y axis) for which the events are defined – for all the 5 models considered in this paper. Higher mean log-likelihood values at a given confidence level indicate better models and lower degrees of exceedence at that confidence level. The value plotted for a given model is the average of the mean log-likelihoods of exceedence events computed over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 1000 *prediction-realization pairs* in the out-of-sample See section 6.6 for additional comments.

Comment: This relative measure, $\ell_{u,m}$, is maximized when the distribution of events in $\Phi^{\pm(k,P)}(c)$ are predicted with the correct probability by the model. Of course this applies strictly only when $\#(\Phi^{\pm(k,P)}(c))$ is large. In figures 4u and 4m we have plotted data in steps of 1% starting from 50% up to 99% – which is for increasing confidence levels but decreasing population of contributing events. It is precisely because the robustness of the statistic is in question for very high levels of confidence, that we have taken the approach of visualizing the measure as a function of the confidence level. When one model shows consistently superior performance than another in the entire high confidence level range – can we be sure that it is a superior model in the context of risk management.

The figures 4u and 4m inform us about the relative performance of the models for events deemed improbable by the model itself. As already discussed in section 2.5 the sets $\Phi^{\pm(k,P)}(c)$ for different models do not refer to the same subset of $t \in [2, 1001]$. The motivation of the next measure is to evaluate how the model performs explicitly for large price moves and to allow comparison of models over exactly the same data.

4.5 Mean log-likelihood against volatility percentile

Introduction: Instead of computing the mean log-likelihood over all events above a certain confidence level – we now compute the mean log-likelihood over all events wherein the predicted movement exceeds a certain absolute size – parametrized by the percentile level of the event size. (The nomenclature *volatility percentile* is motivated by the popular identification of the movement size $|x_t^{(k,P)}|$ with volatility.) In this case the X-axis is independent of the model – so the points plotted for every model which correspond to the same volatility percentile – refer to the same extreme events. This measure tests the conformity of the model predictions to the tail of the empirical *unconditional* distribution instead of the tail of the forecast *conditional* distribution (presented in the last section).

Specification: The empirical unconditional distribution of out-of-sample events may not be symmetric around zero – especially because of long term increases or decreases in one or more series k. We must however still be able to put losses with respect to the *long* and *short* positions in the US Dollar against series k and portfolio P on equal footing when assigning percentile levels to large movements on both extremes. To impose symmetry in the percentile assignments we first construct the event sets:

$$X^{\pm(k,P)}(\chi) = \left\{ t \mid \pm x_t^{(k,P)} > \chi, t \in [2, 1001] \right\}. \tag{4.19}$$

Using sets 4.19 we can implicitly define exceedence levels $\chi^{\pm(k,P)}(\varphi)$

$$\frac{\#(X^{\pm(k,P)}(\chi^{\pm(k,P)}(\varphi)) - \frac{1}{2}}{\#(X^{\pm(k,P)}(0))} = \frac{1}{2} + \frac{1}{2}\left(\frac{\varphi}{100}\right), \quad \varphi > 50, \tag{4.20}$$

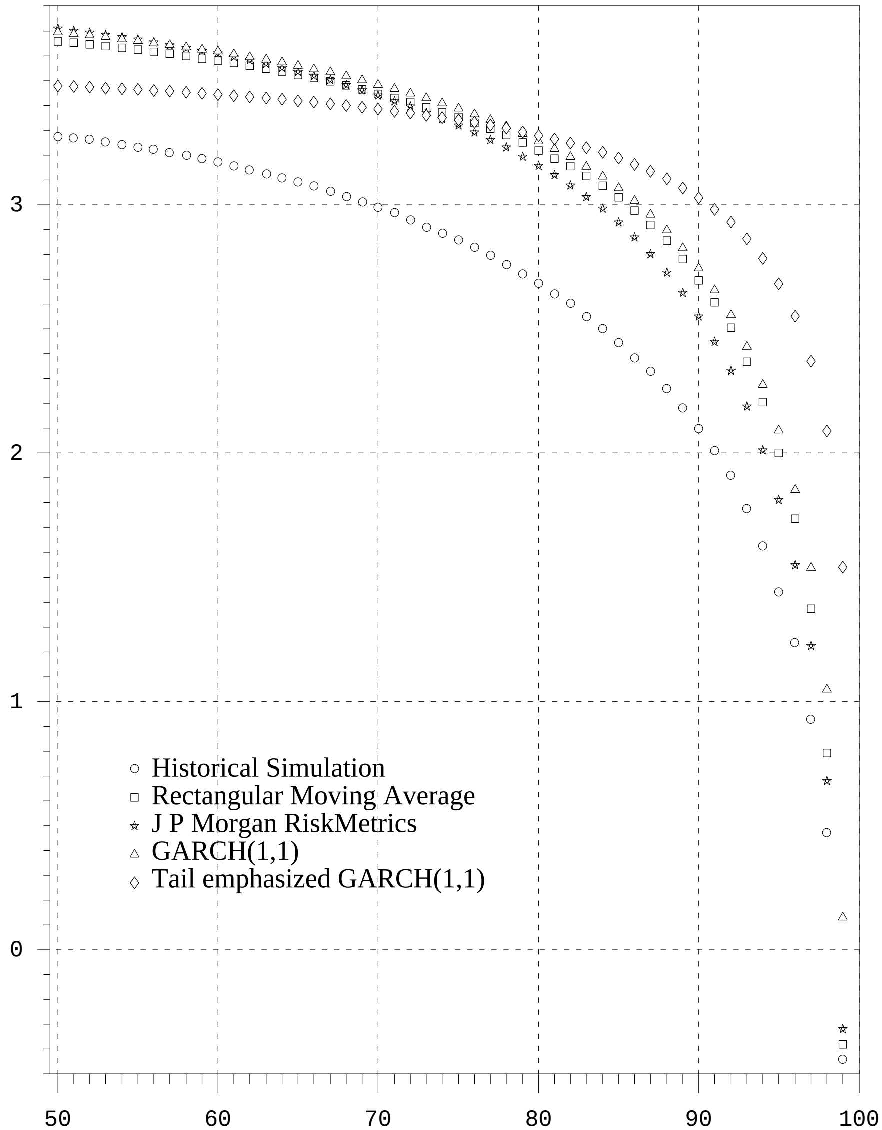

Figure 5u: Mean log-likelihood of loss making events against exceedence percentile of events in the *univariate* context for all candidate models. This plot shows the mean log-likelihood (X axis) of all loss inducing price changes above a certain percentile level in magnitude. High mean log-likelihood values at large percentile levels indicate models which have better captured the dynamics of large financial movements. The value plotted for a given model is the average of the mean log-likelihoods of events beyond a certain percentile level computed over 20 trivial portfolios - each consisting of a *long* or *short* position of the US Dollar against the 10 series k. The curves are based on the analysis of 10000 *prediction-realization pairs* in the out-of-sample period. See section 6.7 for additional comments.

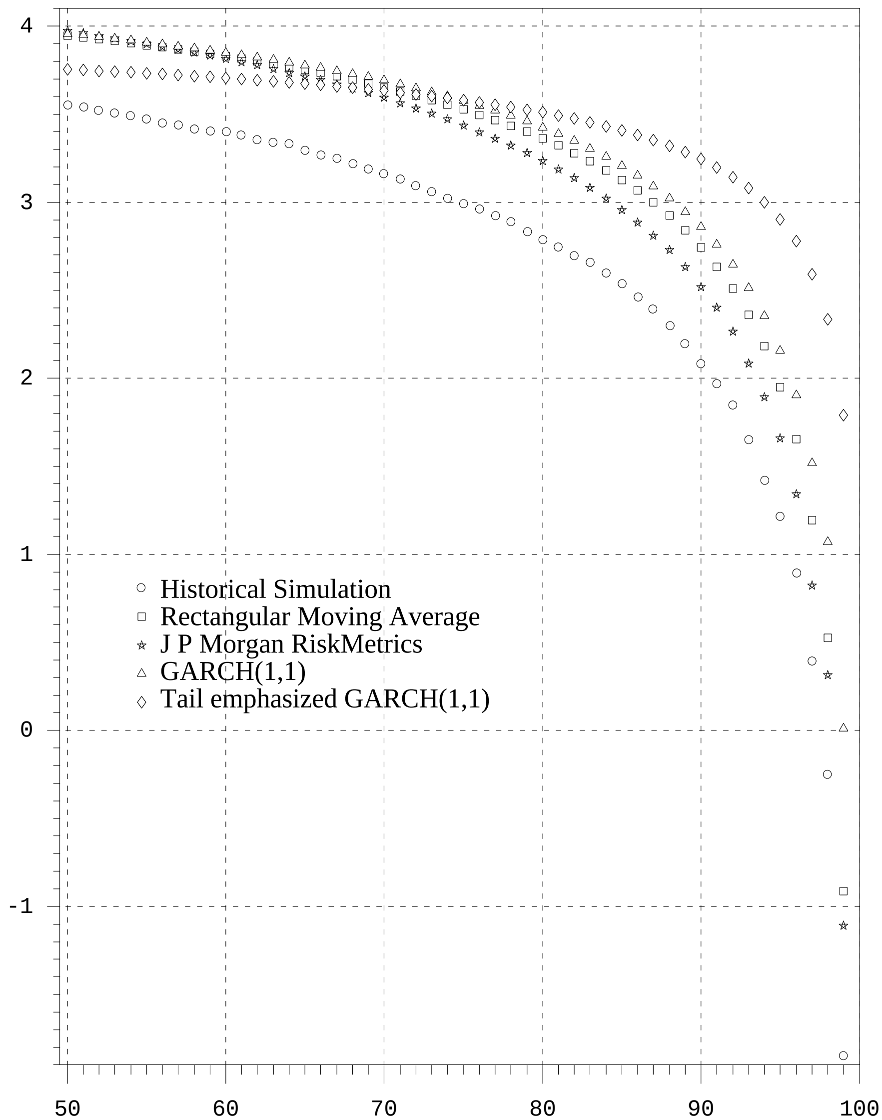

Figure 5m: Mean log-likelihood of loss making events against exceedence percentile of events in the *multivariate* context for all candidate models. This plot shows the mean log-likelihood (X axis) of all loss inducing price changes above a certain percentile level in magnitude. High mean log-likelihood values at large percentile levels indicate models which have better captured the dynamics of large financial movements. The value plotted for a given model is the average of the mean log-likelihoods of events beyond a certain percentile level computed over 2 portfolios – each consisting of uniformly *long* or uniformly *short* positions of the US Dollar against the 10 series k. The curves are based on the analysis of 1000 *prediction-realization pairs* in the out-of-sample See section 6.7 for additional comments.

where the exceedence percentile is φ. The right hand side of expression 4.20 imposes a mapping so that the loss making data (relative to any portfolio and position thereof) will always cover a 50% range of percentile values. (Cases of $x_t^{(k,P)} = 0$ may be ignored or assigned arbitrarily to either of sets $X^{\pm(k,P)}(0)$, without substantially affecting the results.)

Next, using expressions 4.2 and 4.15 we can directly construct

$$\ell^{\pm(k,P)}(\varphi) = \frac{\sum_{t \in X^{\pm(k,P)}(\varphi)} \mathcal{L}_{t-1}^{(k,P)}.}{\#(X^{\pm(k,P)}(\varphi))} \tag{4.21}$$

which is the mean log-likelihood of all loss exceedences over volatility percentile φ for a long (+) or short (−) position in the US Dollar with respect to the series k or portfolio P.

Finally, as specified by expressions 2.1 and 2.2, we construct the *univariate* average mean log-likelihood

$$\ell_u(\varphi) = \frac{1}{20}\left(\sum_{k=1}^{10} \ell^{+(k)}(\varphi) + \sum_{k=1}^{10} \ell^{-(k)}(\varphi)\right) \tag{4.22}$$

and the *multivariate* average mean log-likelihood

$$\ell_m(\varphi) = \frac{1}{2}\left(\ell^{+(P)}(\varphi) + \ell^{-(P)}(\varphi)\right) \tag{4.23}$$

plotted in figures 5u and 5m respectively for all candidate models. The construction of expression 4.20 guarantees that what is true naturally for the 50% confidence level (based on the forecast distribution) is also true for the 50% percentile level (of the empirical out-of-sample distribution) – that this level marks the division of profitable and loss making events in the out-of-sample. As a consequence of this and *symmetrization* (consideration of both *long* and *short*, positions) the value of $\ell_u(\varphi = 50)$ is based on all 10000 out-of-sample events $x_t^{(k)}$ $(t \in [2, 1001])$ and the value $\ell_m(\varphi = 50)$ is based on all 1000 out-of-sample events $x_t^{(P)}$ $(t \in [2, 1001])$ – as was the case for $\ell_u(c = 50)$ and $\ell_m(c = 50)$ specified in the last section.

Comment: This relative measure clearly tests the success of models when large movements are encountered. Figures 5u and 5m show that the *tail emphasized* GARCH(1, 1) stands out in this respect. In figures 5u and 5m the mean log-likelihood value at the 50% percentile level is the mean log-likelihood over all out-of-sample events. As we move to higher percentile levels, small movements are ignored. From this it is clear that the *tail emphasized* GARCH(1, 1) slightly sacrifices performance in predicting small movements for better performance when movements are large.

5. Salient features of methodology

5.1 Modeling

Implied covariances: In the GARCH based models 4 and 5 we construct covariance forecasts for series i and j by looking at the variance forecasts for the series i and j as well as the forecast for the sum series of i and j (expression 3.15). While applying this procedure we are well aware that formally the sum of two GARCH processes is not a GARCH process. This fact however does not intrinsically pose any caveat to the methodology – since the GARCH parameters for the three series (i, j and the sum thereof) are separately optimized in the in sample for all three – and no implicit constructions are used extending the parameters for the series i and j to their sum series.

What can however be of serious concern is that we have no guarantee that Σ_t has positive eigen-values. If Σ_t does have one or more negative eigen-values - this implies that there exists a portfolio with a forecast variance which is negative. Unless the dynamics of the market changes drastically between the in-sample and the out-of-sample, it is extremely unlikely that the method of implied covariances will lead to this pathological condition. We can confirm that in our analysis, with the fixed portfolios, we did not run into this situation – but this does not imply that Σ_t never had negative eigen-values. In section 7 we again discuss this issues of pathology in Σ_t.

Tail emphasis: In model 5 we introduce the idea of tail emphasized fitting. As explained in section 3.5 – in this method, for every candidate parameter set the mean log-likelihood is computed only over the adverse half of the total number of contributing events in the in sample. To be precise, the log-likelihood contributions of all events in the in-sample are sorted and the mean is computed over the lowest half of the set. The consequent fitness landscape is considerably fractured since small changes in parameters can lead to the inclusion and exclusion of one or more events in the computation of the mean. Without the use of genetic algorithms [29] – for example by using only the BHHH algorithm – it would be quite impractical to find the optimal solution for such a fitness landscape.

It is important to note that adverse log-likelihood values are not necessarily contributed by the events of largest magnitude in the in-sample data. The tail referred to in the current context – is the tail in terms of bad predictions by a model – the tail of log-likelihood contributions. The tail emphasized optimization allows the model to find a parameter set which alleviates the problem with its worst predictions by compromising on its better predictions. It makes the model more homogeneous in the quality of its predictions and as a by product of this effect this model also shows better performance for large movements.

In-Sample and Out-Of-Sample: We again emphasize that all optimization – wherever applicable – was done in an in-sample period $t \in [-1249, 0]$ and all performance evaluation was done in the out-of-sample period $t \in [1, 1001]$.

HARCH and EMA-HARCH: Although the HARCH [30] and EMA-HARCH [29] processes were constructed in order to capture the information from high frequency data, we have also studied these models (in addition to the 5 presented in detail) in the present context with daily data. The results we obtained with the 4 parameter processes HARCH(3) and EMA-HARCH(3) and the 5 parameter processes HARCH(4) and EMA-HARCH(4) (all implemented with tail emphasized optimization) showed marginal improvement when compared with tail emphasized GARCH(1, 1) (model 5). We have however not emphasized these models in the present study because we did not see the kind of improvement per degree of freedom to warrant their use with daily data. The comparative improvement of these models over tail emphasized GARCH(1, 1) when run with high frequency data – which they were designed for – is currently under investigation to be presented in a separate paper.

5.2 Performance measures

Measuring performance over extreme subsets: One of the main problems of evaluating models in the context of risk management is that the performance is gauged by looking at the few adverse events beyond the 99% confidence level or associated with the few strongest adverse price movements in relation to some portfolio. Even with a back testing history of 1000 days (as used in this paper) we expect only 10 events beyond the 99% confidence level. Statistics and performance measures based solely on these few events cannot help clearly distinguish good models from bad. Our approach of plotting model performance against confidence level or volatility percentile is motivated by exactly this problem. As discussed in section 2.5, the confidence level or the volatility percentile (plotted on the X axis of all figures) parametrize a class of increasingly extreme partial sets of events by differing definitions. By probing model performance as a *function* of increasingly extreme (but at the same time increasingly depopulated) event classes – we can get a better feel for the behaviour in the limit of extreme behaviour for that class type. We propose that the consistent superiority of one model over another over an entire range of confidence levels or volatility percentiles – approaching the 99% level (say from 75% to 99%) – increases confidence in the model. In addition by looking at the extrapolated approach to the 99% level from the less depopulated classes – we can better appreciate the error in the empirically determined measure at this level.

Separation of variance and covariance forecasting: Another important issue we cover is the separate analysis of the model with regard to variance and covariance forecasting. As already discussed in section 2.5 the results plotted in the *univariate* context determine the quality of the model for variance forecasting. The measures plotted in the *multivariate* context determine the quality of the model for variance and covariance forecasting. If it has been determined that the model has good performance in the *univariate* context but bad performance in the *multivariate* context – this must mean that some modifications are necessary only in the covariance forecasting aspect of the model. Aside from the latter important information – the analysis in the *univariate* context is based on far more data and helps give more robust estimates of variance forecasting performance by a model even at the 99% level where exceedence data is sparse. (Note how the figures m are noisier than figures u.)

Symmetrization: As part of our performance measurement methodology we propose that measures must always be presented as a mean over opposing portfolios. As discussed in section 2.2 this procedure gets rid of any bias in the measure arising out of long term increases or decreases in the series k.

Increasing robustness – averaging over more portfolios: For completeness we mention here that we have evaluated the models in the *multivariate context* as an average over 2 portfolios with opposing position vectors – equally weighted and *long* on the US Dollar against all series k and equally weighted and *short* on the US Dollar against all series k. For a more complete analysis it is viable to extend the methodology by averaging over a larger set of position vectors (such as for example on a sphere of dimension k with some appropriate radius), always ensuring that the symmetrically opposed portfolio is also included.

6. Discussion and Results

In this section we present a number of possibly disjoint but important comments with the purpose of achieving two objectives. One is to discuss the results of how the various models performed in comparison with each other. The second is to discuss the details of what the various performance measures convey, how they complement each other and what their drawbacks are. For convenience we shall refer to the models by the numbers associated with them in section 2.4 and to the performance measures by the numbers associated with them in section 2.5. The numbers are also consistent with the order in which we presented the models and the measures.

6.1 Measures sensitive to number of exceedence

Performance measures 1, 2 and 3 are *digital* measures sensitive to the concept of exceedence as defined by the event set Φ (expression 4.2) and are entirely based on the cardinality of these or similar (Ψ for measure 3) sets as a function of confidence level c. As already remarked (in the detailed specification) – these measures are oblivious to the *degree* of exceedence – and hence lacking in this respect. It is quite possible that two models may have the same number of exceedences – even at the same points in time – but one model barely overshoots the exceedence limit $\chi(c)$ whereas another goes well beyond. Clearly measures 1, 2 and 3 will be unable to discriminate between two such models.

6.2 Measures sensitive to degree of exceedence

To capture the degree of exceedence – we have presented performance measures 4 and 5. The use of the mean log-likelihood of exceedent or large events (measures 4 and 5 respectively) in the out of sample is a sophisticated measure. For the conditional Gaussian distributions used in models 2 through 5, the probability density is strictly monotonically decreasing on both sides of the zero mean – and we are thus guaranteed that as desired $\ln(p_t(x_{t+1}))$ for each exceedent event becomes increasingly negative as the degree of exceedence of x_{t+1} increases. In the case of historical simulation however (model 1), this is not strictly true – since p_t can be and is most likely multi-modal. If according to the model – a large event is more likely than a smaller event of lower size (as in the case of a multi-modal distribution) – and if the out-of-sample data confirms this distribution – then this is a truth which is captured by the model and the model should not be penalized for such an occurrence. On the other hand, if the conditional distribution is multi-modal, but the realizations do not fit this distribution – then the model should be penalized. This is exactly what is achieved by the mean log-likelihood measures 4 and 5 – by expressing exceedence as a function of the probability density of the event rather than simply looking at the size of the exceedence. In general, for large populations, the mean log-likelihood maximizes when the predicted and realized distributions of an event class match up.

6.3 Measure 1

It is interesting to note that when viewed in terms of measure 1, historical simulation does appears to be a reasonable model which neither significantly overestimates or underestimates risk over the full range of confidence levels. In fact, on the basis of this measure, one could perhaps reject model 5 (tail emphasized GARCH$(1, 1)$) as being too conservative – in the sense of always over-estimating risk. In general the message from figures 1u and 1m is that models 1 and 3 have similar behaviours at low confidence level but in the regime beyond the 95% level – tables are turned and – model 1 becomes far superior than all models except model 5. We also find that there is not much too choose between models 2 and 4.

6.4 Measure 2

Measure 2 is interesting only because of its relationship to BIS recommendations. In terms of this measure, if we looked only at the 99% level – which is the legally defined relevant level – the best model is clearly model 5 – based on figure 2m for the *green* zone frequency. This is however not unexpected in view of the appearance that model 5 is extremely conservative in terms of figures 1u and 1m. While model 5 is the only one which is 100% in the *green*, figure 2m shows that none of the models – barring model 2 – are ever significantly in the *red* zone.

6.5 Measure 3

In spite of the fact that measure 3 is what we refer to as a *digital* measure (like measures 1 and 2) – it brings brings out an important point which is missed by measures 1 and 2. This is the clustering property of exceedences. This measure determines if the frequency of consecutive exceedences is higher than expected at a given confidence level. It is expected that models 1 and 2 which make no attempt to capture the autocorrelation of volatility should behave the worst in terms of this measure. It is therefore surprising to find that model 3 fares worse than model 1 with respect to this measure. (The consecutive loss exceedence ratio of model 2 in the *univariate* context at the 99% confidence level is 11.01 – not plotted in figure 3u for clarity of presentation.)

We have already pointed out in section 5.2 the serious caveat with regard to low number of exceedence events when focusing on the 99% confidence levels. We again emphasize that while it is legitimate that any regulatory authority should be looking at these high levels for the evaluation of risk on a day to day basis once a model is accepted – the criteria of accepting this model should be based on performance of the model over an entire range of extreme levels starting from say the 75% level. Figure 3m provides an example of this problem. In figure 3m we have only plotted data when there is at least one exceedence event beyond a certain confidence level (contributed by at least one of the two portfolios over which the measurement is averaged). Thus at the 99% level the absence of any entry in figure 3m indicates that none of the models admit a single event of consecutive exceedences over a 1000 day out-of-sample period. It is only when viewing the approach to the 99% level that we can pick out the relative performance of the models and the consistent superiority of model 5.

6.6 Measure 4

Measure 4 looks at the mean log-likelihood of all exceedent events beyond the confidence level plotted on the X axis. We present figures 4u and 4m only to illustrate the difference in relation to measure 5 which we strongly recommend as a solid performance measure. While the mean log-likelihood against confidence level does help in establishing the *degree* of exceedence – it fails as a valid comparative measure between models. This is because the exceedent events referred to by a specific choice on the X axis do not correspond to the same set of events for different models. Differing sets of exceedent events at the same confidence level imply differing theoretical maximums for the mean log-likelihood of exceedence – making direct comparison fallacious. (In addition we note that figure 4m has no expression of mean log-likelihood for model 5 at the 99% level because both contributing portfolios do not contribute an exceedent event. This again highlights the problem of passing legislation based entirely on evaluations at the 99% level.)

6.7 Measure 5

What is most important in the context of risk management is *not* whether a large number of run-of-the-mill movements were predicted well at the expense of a few large movements – but exactly the contrary. In fact, we can well afford a general over-estimate or under-estimate of risk with regard to small movements in exchange for better performance in the tail. Measure 5 presents the mean log-likelihood of all movements beyond a certain percentile level rather than all movements exceeding a certain confidence level. As discussed in section 5.1 the results of model 5 in figures 5u and 5m clearly show that tail emphasized optimization used in model 5 achieves decreased sensitivity of forecasting performance to the size of the event as desired. In addition, figures 5u and 5m are the only figures presented where the X and Y axis are not both model dependent. The X axis depends only on the data and the corresponding mean log-likelihoods refer to the same set of events – making comparison of models meaningful.

6.8 Model 1

In terms of the measures presented it turns out that historical simulation is not as bad as one would expect for such a simplistic model – which makes no attempt to capture the dynamics of the market and which is based entirely on the series $x^{(P)}$ without referring to the characteristics of the individual risk factors $x^{(k)}$. It has reasonable results in terms of the *digital* measures but has very bad quality in terms of the *degree* of exceedence as reported in figures 4 and 5. We believe that by introducing smoothing in the conditional distribution determined from the 250 day empirical distribution – with the constraint that the probability density must monotonically decrease away from the center – the performance of this model can be considerably improved. If in addition the smoothed 250 day empirical distribution based on X_t (expression 3.1) could be modulated so as to confirm to the variance of a GARCH$(1,1)$ process (optimized with tail emphasis) then the resultant model may be extremely powerful. The caveat of such a scheme however is the complexity introduced in the portfolio function if this methodology is employed in the usual way – where the stochastic process predicts each series $x^{(k)}$ rather than directly predicting the portfolio series $x^{(P)}$. For now, we leave this proposal of such a hybrid approach – mixing historical simulation and stochastic methods – as food for thought for future investigations.

6.9 Model 2

Rectangular simulation is the simplest of the models employing a covariance matrix at the basis of the VaR forecast and completely satisfies the BIS framework. The individual analysis of each series and series pair for the determination of the covariance matrix is the most important element distinguishing this and other models from historical simulation (model 1). This model (like historical simulation) makes no attempt to capture the dynamics of the financial markets – particularly with regard to the observed autocorrelation of volatility.

6.10 Model 3

We have already pointed out that the J. P. Morgan RiskMetrics process is a non-stationary GARCH$(1,1)$ process. The generally poor performance of this process which is fully multivariate in the sense of being based on the construction of a covariance matrix and which does also attempt to capture the autocorrelation of volatility is rather surprising. The main drawback of this process is *possibly* the en-masse assignment of the same parameter set for all financial series. In defense of the en-masse RiskMetrics setting of $\beta_1 = 0.94$ we would like to point out that the parameters for tail emphasized GARCH$(1,1)$ in table 1 may very well be approximated by an en-masse setting of $\beta_1 \approx 0.96$ but with non-zero and diverse minimum variance α_0. It would seem that adding a minimum variance ($\alpha_0 \neq 0$) to the RiskMetrics variance prediction might go a long way in improving its performance. We believe that the difference in values for β_1 may arise simply out of our portfolio choice which does not have the variations in instrument class which are attempted to be captured by the RiskMetrics settings.

6.11 Models 4 and 5

Table 1 shows the comparison between the parameters fit for GARCH$(1,1)$ and tail emphasized GARCH$(1,1)$. The main observation is that α_0 for tail emphasised GARCH$(1,1)$ – which represents a minimal variance – is considerably larger than regular GARCH$(1,1)$. Also, by and large, β_1 is comparatively smaller for the individual series but comparatively larger for the sum series for tail emphasised GARCH$(1,1)$. This opposing movement of β_1 on the individual and sum series as a consequence of tail emphasis may point to possible volatility conditional effects on covariance to be discussed in paper [31].

While increasing the size of the minimal variance α_0 may appear as a panacea for bad performance from the regulatory perspective – it should be noted that the mean log-likelihood will begin to deteriorate rapidly – if α_0 is increased further – causing the model to predominantly over-estimate risk. Thus, in our view, the mean log-likelihood function achieves the right balance for penalizing risk underestimation and risk overestimation. In the beginning of this section we had commented that figures 1u and 1m make model 5 appear too conservative. Through figure 5u and 5m we can appreciate however that model 5 is reasonable and particularly so in the context of large movements. This highlights the deficiency of *digital* measures like exceedence count as the sole criteria for comparative evaluation of model performance in the context of risk management.

7. Problems of pathological VaR estimates

In this section we comment on various sources of pathological estimates of VaR and finally speculate on a framework for defining and handling pathologically underestimated risk.

7.1 Rank defects

In typical implementation of VaR methodologies it is not uncommon that the number of risk factors involved, d, is far greater than the number of data, n, in the history used for forecasting the covariance matrix Σ_t (of dimension d). In particular, if $n < d$ and Σ_t is constructed with the rectangular moving average model (model 2) of section 3.2 or the RiskMetrics model (model 3) of section 3.3 it is guaranteed that the covariance matrix will have a rank defect. The existence of a rank defect implies that Σ_t admits the existence of a portfolio (or a set of portfolios) $\lambda = \lambda(\Sigma_t)$ such that $\lambda^T \Sigma_t \lambda = 0$ (or $\mathrm{VaR}(\lambda, \Sigma_t, c) = 0$ for all c, where c is the confidence level used – usually 99%).

In the rectangular moving average model every $x_{t-i}^{(j)}$ entering the computation 3.8 of $s^{(jk)}$ for any k does so with the same weight – unity. In the J. P. Morgan RiskMetrics case the computation 3.11 expresses $s^{(jk)}$ in recursive form - but when expanded in terms of the data history it is stll true that every participation of $x_{t-i}^{(j)}$ in the construction of the covariance matrix occurs with the same weight or coefficient (although not unity).

In general - any covariance matrix construction which can be expressed in the form:

$$\Sigma_t = W_t^T W_t \quad \text{where}$$

$$W_t^T = \begin{pmatrix} w_t^{(1)} x_t^{(1)} & w_{t-1}^{(1)} x_{t-1}^{(1)} & \cdots & w_{t-n+1}^{(1)} x_{t-n+1}^{(1)} \\ w_t^{(2)} x_t^{(2)} & w_{t-1}^{(2)} x_{t-1}^{(2)} & \cdots & w_{t-n+1}^{(2)} x_{t-n+1}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ w_t^{(d)} x_t^{(d)} & w_{t-1}^{(d)} x_{t-1}^{(d)} & \cdots & w_{t-n+1}^{(d)} x_{t-n+1}^{(d)} \end{pmatrix} \tag{7.1}$$

(W^T is a matrix of the weighted data history) will suffer from a rank defect if $n < d$. (Theoretically, the rank of Σ_t is the the rank of W_t – which is the minimum of d and n. If $n < d$, then the rank n of Σ_t is smaller than its dimension d and hence Σ_t must have a rank defect.)

7.2 Near singularity

We have illustrated above that the rectangular moving average model satisfying the BIS parameters (model 2) as well as the J. P. Morgan RiskMetrics model (model 3) necessarily have rank defects if the number of risk factors d exceeds the number of historic data n from which the correlation matrix is constructed. This *overt* existence of the rank defect however points to a deeper problem. Even when the methodology is more complex – such as when we are using model 4 or 5 with implied covariances – while it is unlikely that Σ_t is exactly singular (and hence admits a rank defect), it can very well be *near* singular. This means that while there will not exist a portfolio λ (aside from the trivial null portfolio $\lambda_i = 0$) such that $\mathrm{VaR}(\lambda, \Sigma_t, c) = 0$ (for all c) there will still exist portfolios for which risk is pathologically underestimated (as defined in section 7.6).

7.3 Stochastic errors

Even if we have very good models (and even if $n > d$), simply due to inevitable stochastic errors in the estimation of Σ_t it is possible for the covariance matrix to become near singular and admit portfolios for which risk is pathologically underestimated. More often however, stochastic errors will lead to spurious minima in the VaR – which can also lead to a systematic underestimation of risk as explained in the next subsection.

7.4 Bias due to portfolio optimization

In our entire analysis of models and measures in the earlier sections, we worked with fixed, pre-defined portfolios. During this analysis, spurious minima could have occasionally appeared in the vicinity of the fixed portfolios leading to risk underestimation but spurious maxima at other times would have largely balanced this effect in the overall performance evaluation of the model.

In typical scenarios of the application of VaR methodology, however, the actual portfolio is determined with the help of the *estimated* covariance matrix Σ_t. This means that, on a frequent basis, investors are likely to take advantage of rank defects, near singularities and spurious minima due to stochastic errors – and tend to select portfolios in that portion of the portfolio space which necessarily underestimates risk.

7.5 Negative variance

We have already mentioned in section 5.1 the possibility for Σ_t having negative eigen-values and hence admitting the existence of portfolios for which the forecast variance of assets is negative. This is clearly a pathological situation that requires to be addressed. Below we present a preliminary framework for addressing all of the above problems which also serves to partially address the problem of negative variance.

7.6 Treatment of pathological VaR estimates

All the problems listed above deal with the existence of unreliable regions of portfolio space in which risk is underestimated due to estimation errors in the covariance matrix. Our proposal to handle this problem is to determine $\kappa(\lambda, \Sigma_t, c)$ as the minimum acceptable risk for a portfolio at confidence level c. The reported risk (at level c) is then the larger of $\mathrm{VaR}(\lambda, \Sigma_t, c)$ and $\kappa(\lambda, \Sigma_t, c)$. As a corollary, we assert that $\mathrm{VaR}(\lambda, \Sigma_t, c)$ is defined as pathologically underestimated if $\mathrm{VaR}(\lambda, \Sigma_t, c) < \kappa(\lambda, \Sigma_t, c)$.

To build this machinery we first introduce the function:

$$\nu(\lambda) = \sum_{i=1}^{d} |\lambda_i| \tag{7.2}$$

and refer to this as the *collateral function* – since this function approximates the collateral needed to support the investment. (The picture here is consistent with the view that trading is conducted without leverage and via a broker who allows one to take long or short positions on paper – as long as the value of the collateral function is deposited with the broker. The true collateral may be different due to several reasons – but modeling the true collateral is tangential to our objective.) Our purpose is simply to define a meaningful norm, $||\lambda|| = \nu(\lambda)$, against which the size of $\mathrm{VaR}(\lambda, \Sigma_t, c)$ may be compared, and which satisfies the condition that the ratio

$$\frac{\mathrm{VaR}(\lambda, \Sigma_t, c)}{\nu(\lambda)} \tag{7.3}$$

is invariant under scalar multiplication of λ (true because both the numerator and the denominator satisfy the property, $f(\alpha\lambda) = \alpha f(\lambda)$).

The scale invariance of the ratio 7.3 has the important repercussion that distinct values must correspond to distinct *directions* in the portfolio space of dimension d (from the origin). Since only *direction* matters, for further analysis it is convenient to define the surface S of normalized portfolios, as the locus of all portfolios for which $\nu = 1$. In this notation, the components of $\lambda^{(S)} = \frac{\lambda}{\|\lambda\|} \in \mathcal{S}$ may be denoted:

$$\lambda_i^S = \frac{\lambda_i}{\sum_{i=1}^d |\lambda_i|}. \tag{7.4}$$

(For example, if $d = 3$, then $\mathcal{S}$ is a regular octahedron joining the points on each axis at positive and negative unity.)

We still need two more mathematical objects to define the methodology for determining $\kappa(\lambda, \Sigma_t, c)$. One is the function $\mathcal{P}$ defined by:

$$\mathcal{P}(\lambda^{(1)}, \lambda^{(2)}) = \frac{1}{d} \sum_{i=1}^{d} \log \left(\frac{\lambda_i^{(1)}}{\lambda_i^{(2)}} \right)^2 \tag{7.5}$$

which measures the strength of a perturbation that transforms $\lambda^{(1)}$ to $\lambda^{(2)}$. The other is the more complicated object, λ_o^S, determined as a function of λ^S. This object has different definition depending on whether $\mathrm{VaR}(\lambda^S, \Sigma_t, c)$ is negative or positive.

- If $\mathrm{VaR}(\lambda^S, \Sigma_t, c) > 0$ then λ_o^S is the local minimum for the VaR on surface $\mathcal{S}$ which can be reached from λ^S via a continuous path of decreasing VaR. If the local minima is negative – λ_o^S is the first point on the steepest descent where VaR becomes zero.
- If $\mathrm{VaR}(\lambda^S, \Sigma_t, c) < 0$ then λ_o^S is the point on the surface $\mathcal{S}$ such that $\mathrm{VaR}(\lambda_o^S, \Sigma_t, c) = 0$ which requires the smallest perturbation $\mathcal{P}(\lambda^S, \lambda_o^S)$.
- If $\mathrm{VaR}(\lambda^S, \Sigma_t, c) = 0$ then $\lambda_o^S = \lambda^S$.

We are now in a position to describe a procedure for reporting VaR which largely handles the underestimation problems when reporting risk:

1. From λ, determine $\nu(\lambda)$ and $\lambda^{(S)} = \frac{\lambda}{\|\lambda\|} \in \mathcal{S}$.
2. From $\lambda^{(S)}$, determine λ_o^S.
3. From λ_o^S follow the path of steepest ascent (of VaR) – constrained to the surface $\mathcal{S}$ – until $\tilde{\lambda}^S$ such that $\mathcal{P}(\lambda_o^S, \tilde{\lambda}^S) = p$, where p is a pre-established (but ad-hoc) perturbation strength.
4. From $\tilde{\lambda}^S$, define:
$$\kappa(\lambda, \Sigma_t, c) = \nu(\lambda) \mathrm{VaR}(\tilde{\lambda}^S, \Sigma_t, c) \tag{7.6}$$
5. The reported VaR at confidence level c is then the larger of $\mathrm{Var}(\lambda, c)$ and $\kappa(\lambda, \Sigma_t, c)$.

Admittedly, our skeletal presentation above, is not an exhaustive consideration of all possible topological complexities that can arise – particularly when Σ_t admits negative eigen-values. Our motivation, in this section, is only to introduce the concept of minimum acceptable risk for a portfolio λ with respect to perturbation level p – as a useful concept which can go a long way in handling the various problems discussed in this section.

8. Conclusions

- Model judgement is dramatically improved by visualization of model performance over an adequately large range of high confidence levels.
- A sophisticated measurement strategy should include the following in addition to the conventional measures based on exceedence counts:
  - Measures (or variants thereof) sensitive to degree of exceedence (such as measures 4 and 5).
  - Measures discriminating model behaviour for large and small movements (such as measure 5).
  - Measures which assess model behaviour with respect to market dynamics (such as measure 3).
- The comparative performance of models in the *univariate* and *multivariate* contexts provide a systematic approach to model building which allows discrimination between the quality of the diagonal and off-diagonal elements of the covariance matrix.
- The method of tail-emphasized optimization brings striking improvement in model performance – particularly for the large movements to which VaR methodologies are most vulnerable.
- The usage of *implied covariances* in models 4 and 5 to estimate Σ_t has shown remarkable stability and performance.
- The issues of underestimated VaR due to rank defects or near-singularity of Σ_t, and stochastic errors in estimation of Σ_t – particularly in conjunction with frequent portfolio adjustments – are non trivial.
- Underestimated VaR may be improved by analyzing VaR as a function of portfolio space in the neighbourhood of the candidate portfolio.

Acknowledgements

The authors acknowledge useful discussions with Giuseppe Ballocchi, the software support provided by Olivier V. Pictet and the careful review of the manuscript by Michel M. Dacorogna.

References

1. Basle Committee on Banking Supervision. Proposal to issue a supplement to the basle capital accord to cover market risks. 1995.

2. Basle Euro-currency standing committee of the central banks of ten countries, Bank of International Settlements. Public disclosure of market and credit risks by financial intermediaries. *Fisher report*, 1994.
3. Group of thirty. Derivatives: Practices and principals. New York, 1993.
4. J. P. Morgan (Risk management advisory) and Reuters Ltd. (International marketing). RiskMetrics – technical document, fourth edition. 1996.
5. P. Jorion. Value at risk. *IRWIN, Chicago*, 1997.
6. C. Alexander. *Handbook of Risk Management and Analysis*. Wiley, 1996.
7. R. A. Beckström and A. R. Campbell, editors. *An Introduction to VaR*. Cats Software, Palo Alto, CA, 1995.
8. E. Ghysels, A. Harvey and E. Renault. Stochastic volatility. *CORE discussion paper No. 9569*, 1995.
9. P. Artzner, F. Delbaen, J. M. Eber and D. Heath. A characterization of measures of risk. *Preprint*, 1996.
10. E. Eberlein, U. Keller and K. Prause. New insights into smile – mispricing and value at risk: the hyperbolic model. *Preprint, Freiburger Zentrum für Datenanalyse und Modellbildung*, 1997.
11. C. Marshall and M. Siegel. Value at risk: Implementing a risk measurement standard. *Preprint, MIT Finance Research Center*, 1996.
12. W. Bühler, O. Korn and A. Schmidt. Ermittlung von Eigenkapitalanforderungen mit internen Modellen. *Preprint, University of Mannheim, Germany*, 1997.
13. T. S. Beder. VaR: Seductive but dangerous. *Financial Analysts Journal*, 12–24, 1995.
14. C. Alexander. History debunked. *RISK*, 1994.
15. C. Alexander. After the event: A review of the volatility and correlation forecasts in the third edition of RiskMetrics. *Risk*, 1995.
16. C. Alexander and A. M. Chidumba. Multivariate orthogonal factor garch. *Preprint*, 1995.
17. C. Crnkovic and J. Drachman. A universal tool to discriminate between risk management techniques. *Risk*, 1996.
18. M. Chen T. S. Y. Ho and F. Eng. VaR analytics: Portfolio structure, key rate convexities and var betas. *Journal of Portfolio Management*, 89–98, Fall 1996.
19. F. Jamshidian and Y. Zhu. Scenario simulation: Theory and methodology. *Journal of Finance and Stochastics*, 1:43–68, 1997.
20. P. Kupiec. Techniques for verifying the accuracy of risk management models. *Risk measurement and systemic risk: Proceedings of a joint central bank research conference, Washington D. C, November 16-17, 1995.*
21. José A. Lopez. Regulatory evaluation of value-at-risk models. *Research Paper # 9710, Federal Reserve Bank of New York*, 1997.
22. D. Hendricks. Evaluation of value-at-risk models using historical data. *Risk measurement and systemic risk: Proceedings of a joint central bank research conference, Washington D. C, November 16-17, 1995*, 1995.
23. P. Jackson, D. Maude and W. Perraudin. Value-at-risk techniques: An empirical study. *Risk measurement and systemic risk: Proceedings of a joint central bank research conference, Washington D. C, November 16-17, 1995.*
24. J. M. Mahoney. Empirical based versus model-based approaches to value-at-risk: an examination of foreign exchange and global equity portfolios. *Risk measurement and systemic risk: Proceedings of a joint central bank research conference, Washington D. C, November 16-17, 1995.*
25. F. Drudi, A. Generale and G. Majoni. Sensitivity of VaR measures to different risk models. *Preprint, Banca d'Italia*, 1996.

26. U. A. Müller, M. M. Dacorogna, R. B. Olsen, O. V. Pictet, M. Schwarz, and C. Morgenegg. Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking and Finance*, 14:1189–1208, 12, 1990.
27. Tim Bollerslev, Ray Y. Chou, and Kenneth F. Kroner. ARCH modeling in finance. *Journal of Econometrics*, 52:5–59, January 1992.
28. E.K. Berndt, B.H. Hall, R.E. Hall, and J.A. Hausman. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3:653–665, 1974.
29. Michel M. Dacorogna, Ulrich A. Müller, Richard B. Olsen, and Olivier V. Pictet. *Modeling Short-Term Volatility with GARCH and HARCH Models.* Internal document MMD.1997-01-08, Olsen & Associates, Seefeldstrasse 233, 8008 Zürich, Switzerland, January 08 1997.
30. Ulrich A. Müller, Michel M. Dacorogna, Rakhal D. Davé, Richard B. Olsen, Olivier V. Pictet, and Jakob E. von Weizsäcker. Volatilities of different time resolutions – analyzing the dynamics of market components. *Forthcoming in the Journal of Empirical Finance*, 1997.
31. Rakhal D. Davé, Gerhard Stahl, Giuseppe Ballocchi, Olivier V. Pictet and Richard B. Olsen. Volatility conditional correlation between financial markets. *Paper presented at 6th International Workshop on Parallel Applications in Statistics and Economics, PASE-97, Mariaánské Lázně, November 9-12, 1997*, 1997.