Journal of Derivatives, **4:**3 (1997) 50-62

On the Covariance Matrices used in Value-at-Risk Models

C.O. Alexander, School of Mathematics, University of Sussex, UK and Algorithmics Inc.
and
C. T. Leigh, Risk Monitoring and Control, Robert Fleming, London, UK

This paper examines the covariance matrices that are often used for internal value-at-risk models. We first show how the large covariance matrices necessary for global risk management systems can be generated using orthogonalization procedures in conjunction with univariate volatility forecasting methods. We then examine the performance of three common volatility forecasting methods: the equally weighted average; the exponentially weighted average; and Generalised Autoregressive Conditional Heteroscedasticity (GARCH). Standard statistical evaluation criteria using equity and foreign exchange data with 1996 as the test period give mixed results, although generally favour the exponentially weighted moving average methodology for all but very short term holding periods. But these criteria assess the ability to model the centre of returns distributions, whereas value-at-risk models require accuracy in the tails. Operational evaluation takes the form of back testing volatility forecasts following the Bank for International Settlements (BIS) guidelines. For almost all major equity markets and US dollar exchange rates, both the equally weighted average and the GARCH models would be placed within the 'green zone'. However on most of the test data, and particularly for foreign exchange, exponentially weighted moving average models predict an unacceptably high number of outliers. Thus value-at-risk measures calculated using this method would be understated.

1. Introduction

The concept of value-at-risk is currently being adopted by regulators to assess market and credit risk capital requirements of banks. All banks in EEC countries should keep daily records of risk capital estimates for inspection by their central banks. Banks can base estimates either on regulators rules, or on a value-at-risk measure which is generated by an internal model. The risk capital requirements for banks that do not employ approved internal models by January 1988 are likely to be rather conservative, so the motivation to use mathematical models of value-at-risk is intense (see Bank for International Settlements, 1996a).

The purpose of this paper is to assess the different covariance data available for internal value-at-risk models. In section 2 we summarize two common types of internal models, one for cash/futures and the other for options portfolios. We show that an important determinant of their accuracy is the covariance matrix of risk factor returns[1]. Building large, representative covariance matrices for global risk management systems is a challenge of data modelling. In section 3 we outline how large positive definite covariance matrices can be generated using only volatility forecasts. Thus an evaluation of the accuracy of different univariate volatility forecasting methods provides a great deal of insight to the whole covariance matrix. The remaining sections deliver the main message of the paper on the predictive ability of volatility forecasting models. In section 4 the three most commonly used volatility forecasting models are outlined, and in section 5 we evaluate their accuracy using both statistical and operational procedures. Section 6 summarizes and concludes.

[1] An n-day covariance matrix is a matrix of forecasts of variances and covariances of n-day returns. The diagonal elements of this matrix are the variance forecasts, and the off diagonals are the covariance forecasts. Covariances may be converted to correlation forecasts on dividing by the square root of the product of the variances in the usual way. An n-day variance forecast v(n) is transformed to an annualized percentage volatility as $100\sqrt{(250v(n)/n)}$, assuming 250 trading days per year.

2 Value-at-risk Models

A 100α% h-period *value-at-risk* measure is the nominal amount C such that

$$\text{Prob}(\Delta P < -C) = \alpha \tag{1}$$

where ΔP denotes the change in portfolio value (P&L) over a pre-specified holding period h and α is a sufficiently small probability. Current recommendations by the Basle committee are for holding periods of at least 10 working days and a probability of 0.01. This definition shows that value-at-risk will be largely determined by the volatility of P&L over the holding period. It is for this volatility that we need accurate forecasts of the covariance matrix of factor returns.

A number of methods for determining value-at-risk by estimating the volatility of P&L use the covariance matrix of risk factor or asset returns as the major input. Firstly, the linear structure of non-options based portfolios makes them accessible to matrix methods based on the assumption that P&Ls are conditionally normally distributed. It has become standard to evaluate the volatility of P&L during the holding period as the square root of the quadratic form of the mark-to-market value vector with the covariance matrix of risk factor or asset returns. A second class of methods are necessary for options based portfolios. These require non-linear methods because of the significant gamma effects in these positions. Standard methods include historical or Monte Carlo simulation, and delta-gamma methods (which need not involve simulation). Structured Monte Carlo simulation uses the Cholesky decomposition of the covariance matrix to generate correlated factor returns over the holding period. The linear methods applicable to spot/futures/forwards positions are now described in more detail, followed by a comment on the use of covariance matrices in simulation methods.

2.1 Linear Methods

Standard methods of calculating value-at-risk for portfolios of cash, futures or forwards are based on the assumption that ΔP is conditionally normally distributed [2] with mean μ and variance σ^2. This gives

$$C = Z_\alpha \sigma - \mu \tag{2}$$

where Z_α denotes the appropriate critical value from the normal distribution. Unless the holding period is rather long, it can be best to ignore the possibility that risk capital calculations may be offset by a positive mean μ, and assume that $\mu=0$. Thus market risk capital is given by the nominal amount $C = Z_\alpha \sigma$ and, since Z_α is fixed, the accuracy of linear models depends upon only one thing: an accurate forecasts of the standard deviation of portfolio P&L over the holding period, σ.

When such a portfolio can be written as a weighted sum of the individual assets, the portfolio P&L over the next h days is related to asset returns over the next h days:

$$\Delta P_t = P_{t+h} - P_t = \mathbf{P_t}' \mathbf{R_t}$$

where $\mathbf{P_t}$ denotes the vector of mark-to-market values in the portfolio and $\mathbf{R_t}$ denotes the vector of returns (one to each asset in the portfolio) over the next h days. So the forecast of the quantity σ at time t to use in the formula (2) will be the square root of the quadratic form

$$\mathbf{P_t}' \mathbf{V(R_t)} \mathbf{P_t} \tag{3}$$

where $\mathbf{V(R_t)}$ denotes the covariance matrix of asset returns over the next h days. More generally, linear portfolios are written as a sum of risk factors weighted by the net sensitivities to these factors. In this case the formula (3) is used but now $\mathbf{P_t}$ denotes the vector of mark-to-market

[2]If we assume instead that portfolio *returns* are normally distributed gives $C = (Z_\alpha \sigma - \mu)P$ where σ is now the standard deviation of portfolio returns. Although this assumption is fine for options portfolios, where simulation methods are commonly employed, it does not lead to the usual quadratic form method for cash portfolios. Neither does the more usual assumption that log returns are normally distributed lead to this quadratic form, so value-at-risk models are usually based on the assumption that portfolio *P&Ls* are normally distributed.

values of the exposure to each risk factor (i.e. price x weight x factor sensitivity) and $\mathbf{V(R_t)}$ denotes the forecast covariance matrix of the risk factor returns.

2.2 Non-Linear Methods

There are many value-at-risk models for options portfolios which take into account the non-linear response to large movements in underlying risk factors (a useful survey of these methods may be found in Coleman, 1996). Two of the most common methods use direct simulation of portfolio P&L, either on historical data or on risk factor returns over the holding period, and the value-at-risk is read-off directly as the lower $100\alpha\%$ quantile of this distribution, as in (1). Historical simulation is employed by a number of major institutions, but since it does not use the returns forecast covariance matrix we do not discuss it at length here.[3]

Many banks and other financial institutions now rely on some sort of Monte Carlo simulation of portfolio P&L over the holding period. Structured Monte Carlo applies the Cholesky decomposition of $\mathbf{V(R_t)}$ to a vector of simulated, uncorrelated risk factors to a vector with covariance matrix $\mathbf{V(R_t)}$. This generates a terminal vector of risk factors at the end of the holding period. The price functional is applied to this vector to get one simulated value of the portfolio in h periods time, and hence one simulated profit or loss. The process is repeated thousands of times, to generate a representative P&L distribution, the lower $100\alpha\%$ quantile of which is the value-at-risk number.

[3] Historical simulation uses the past few years of market data - BIS recommendations are for between 3 and 5 years - for all risk factors in the portfolio. An artificial price history of the portfolio is generated, by applying the price functionals with current parameters, to every day in the historic data set. This is a time consuming exercise, but it does enable the value-at-risk to be read off from the historic P&L distribution without making any distributional assumptions other than those inherent in the pricing models. However there are some disadvantages with using this method. Value-at-risk is a measure of *everyday* capital requirements. To investigate what sort of capital allowances need to be made in extreme circumstances such as Black Monday, the model should be used to stress test the portfolio and the results of stress tests should be reported separately from everyday value-at-risk calculations. But historical simulation tends to mix the two together - if extreme events occur during the historic data period these will contaminate the everyday value-at-risk measures. Another problem with historic simulation is that the use of current parameters for pricing models during the whole historic period is very unrealistic: volatility in particular tends to change significantly during the course of several years. Finally, the BIS recommend that the past 250 days of historic data be used for backtesting the value-at-risk model (see section 5.3). But the same data cannot be used both to generate and to test results.

In order to obtain the Cholesky decomposition the covariance matrix must be positive definite.[4] The same condition is necessary to guarantee that linear value-at-risk models always give a positive value-at-risk measure. Positive definiteness is easy enough to ensure for small matrices relevant to only a few positions, but firm-wide risk management requirements are for very large covariance matrices indeed, and it is more difficult to develop good methods for generating the very large positive definite matrices required.

3. Methods for Generating Large Positive Definite Covariance Matrices

Moving average methods do not always give positive definite covariance matrices. Equally weighted moving averages of past squared returns and cross products of returns will only give positive definite matrices if the number of risk factors is less than the number of data points. Under the same conditions exponentially weighted moving averages will give positive semi-definite matrices[5] but only if the same smoothing constant is applied to all series. In both moving average methods the covariance matrix can have very low rank, depending on the data and parameters. If data are linearly interpolated[6] and if the smoothing constant for the exponential method is sufficiently low[7] the matrix will have zero eigenvalues which are often estimated as negative in Cholesky decomposition algorithms - so the algorithm will not work.[8] These difficulties are small compared with the challenge of using GARCH models to generate covariance matrices. Direct estimation of the large multivariate GARCH models necessary for global risk systems is an insurmountable computational problem.

---

[4] A square, symmetric matrix **V** is positive definite iff **x'Vx**>0 for all non-zero vectors **x**.

[5] So some risk positions would have a zero value-at-risk

[6] Such as the RiskMetrics yield curve data

[7] For example with a smoothing constant of 0.94, effectively only 74 data points are used. So models with more than 74 risk factors have covariance matrices of less than full rank.

[8] Many thanks to Michael Zerbs, Dan Rosen and Alex Krenin of algorithmics Inc for helping me explore reasons for the failure of Cholsky decompositions.

However we can employ a general framework which uses principal components analysis to orthogonalize the risk factors, and then generate the full covariance matrix of the original risk factors from the volatilities of all the orthogonal factors.[9] In the orthogonal method, firstly risk factors are sub-divided into correlated categories, and then univariate variance forecasts are made for each of the principal components in a sub-division. Since the principal components are uncorrelated their covariance matrix is diagonal, so only volatility forecasts are required for the covariance matrix forecasts. Then the factor weights matrices (one per risk category sub-division) are used to transform the diagonal covariance matrix of principal components into the full covariance matrix of original system as follows: (a) apply a standard similarity transform using the factor weights of each risk category separately. This gives within risk factor category covariances and a block diagonal covariance matrix of the full system; then (b) apply a transform using factor weights from two different categories to get the cross factor category covariances. The full covariance matrix which accounts for correlations between all risk positions will be positive definite under certain conditions on cross-correlations between principal components.[10]

Orthogonalization methods allow properties of the full covariance matrix to be deduced from volatility forecasting methods alone. This comes as something of a relief. Volatility forecasts are difficult enough to evaluate without having to use multivariate distributions to evaluate

---

[9] It is not necessary to use GARCH volatility models on the principal components - equally or exponentially weighted moving average methods could be used instead. However, there is always the problem of which smoothing constant to use for the exponentially weighted moving average. One of the advantages of using GARCH is that the parameters are chosen optimally (to maximize the likelihood of the data used). The strength of the orthogonalization technique is the generation of large positive definite covariance matrices from volatility forecasts alone, and not the particular method employed to produce these forecasts. Whether GARCH or moving averages are used for the volatility forecasts of the principal components, the method is best applied to a set of risk factors which is reasonably highly correlated. The full set of risk factors should be classified not just according to risk factor category. For example, within equities or foreign exchange it might be best to have sub-divisions according to geographical location or market capitalization.

[10] A general method of using orthogonal components to 'splice' together positive definite matrices - such as covariance matrices of different risk factors - takes a particularly easy form when orthogonal components of the original system have been obtained. Suppose $\mathbf{P} = (P_1 , ... P_n)$ are the PCs of the first system (n risk factors) and let $\mathbf{Q} = (Q_1 ,...Q_m)$ be the PCs of the second system (m risk factors). Denote by **A** (nxn) and **B** (mxm) the factor weights matrices of the first and second systems. Then cross factor covariances are **ACB'** where **C** denotes the mxn matrix of covariances of principal components. Within factor covariances are given by **AV(P)A'** and **BV(Q)B'** respectively as explained in Alexander and Chibumba (1996). Positive definiteness of the full covariance matrix of both risk factor systems depends on the cross covariances of principal components (see Alexander and Ledermann, in prep.).

covariances, which are often unstable. In this paper we assess the accuracy of three types of volatility forecasting methods which we term 'regulatory' (an equally weighted average of the past 250 squared returns), 'EWMA' (an exponentially weighted moving average of squared returns) and GARCH (a normal GARCH(1,1) model). The three methods are fully described and critically discussed in the next section.

4. The Variance Forecasts

4.1 Regulators recommendations

One of the requirements of the Bank for International Settlements (BIS) for internal value-at-risk models is that at least one year of historic data be used. Following JP Morgan (1995) we call the covariance matrix which is based on an equally weighted average of squared returns over the past year the 'regulatory' matrix. The 'regulatory' variance forecasts at time T are therefore given by

$$s_T^2 = \sum_{t=T-n}^{t=T-1} r_t^2 \ / n$$

where n = 250 and r_t denotes the daily return at time t. Since returns are usually stationary, this is the unbiased sample estimate of the variance of the returns distribution if they have zero mean.[11]

The regulatory forecasts can have some rather undesirable qualities[12]. Firstly, the BIS recommend that forecasts for the entire holding period be calculated by applying the 'square root of time' rule. This rule simply calculates h-day standard deviations as √h times the daily standard deviation. It is based on the assumption that daily log returns are normally, independently and identically distributed, so the variance of h-day returns is just h times the variance of daily

[11]It was found that post sample predictive performance (according to the criteria described in section 5) deteriorated considerably when forecasts are computed around a non-zero sample mean. This finding concords with those of Figlewski (1994). Thus this paper assumes mean of zero, both in (2) and in the variance and covariance forecasting models.

[12]A discussion of the problems associated with equally and exponentially weighted variance estimates is given in Alexander (1996)

returns. But since volatility is just an annualised form of the standard deviation, and since the annualising factor is - assuming 250 days per year - √250 for daily returns but √(250/h) for h-day returns, this rule is equivalent to the Black-Scholes assumption that current levels of volatility remain the same.

The second problem with equally weighted averages is that if there is even just *one* unusual return during the past year it will continue to keep volatility estimates high for exactly one year following that day, even though the underlying volatility will have long ago returned to normal levels. Generally speaking there may be a number of extreme market movements during the course of the past year, and these will keep volatility estimates artificially high in periods of tranquillity. By the same token they will be lower than they should be during the short bursts of volatility which characterise financial markets. The problem with equally weighted averages is that extreme events are just as important to current estimates whether they occurred yesterday or a long time ago.

Figure 1: Historic Volatilities of the FTSE from 1984 to 1995, showing 'ghost features' of Black Monday and other extreme events

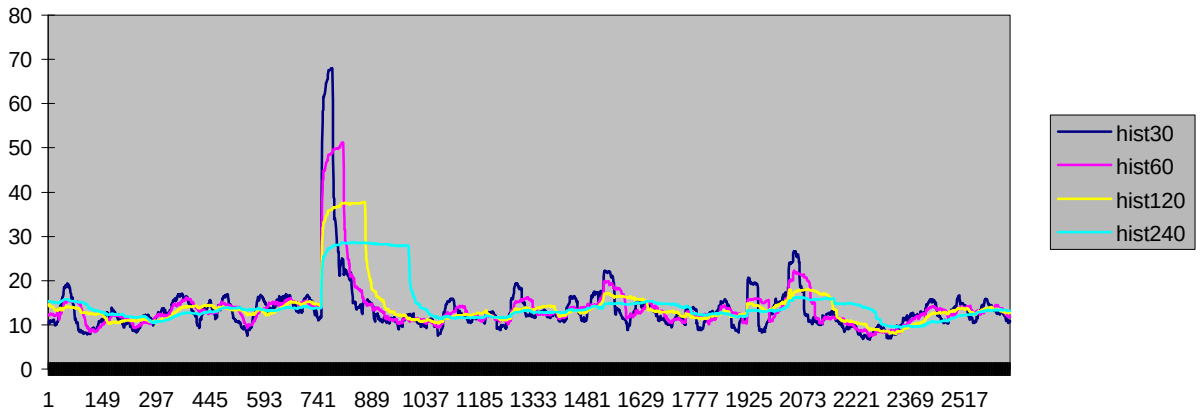

Figure 1 illustrates the problem with equally weighted averages of different lengths on squared returns to the FTSE. Daily squared returns are averaged over the last n observations, and this variance is transformed to an annualized in figure 1. Note that the one-year volatility of the FTSE jumped up to 26% the day after Black Monday and it stayed at that level for a whole year because that one, huge squared return had exactly the same weight in the average. Exactly one year after the event the large return falls outs of the moving average, and so the volatility forecast returned to its normal level of around 13%. In shorter term equally weighted averages this 'ghost feature'

will be much bigger because it will be averaged over fewer observations, but it will last for a shorter period of time.

4.2 Exponentially Weighted Moving Averages

The 'ghost features' problem of equally weighted moving averages has motivated the extensive use of infinite exponentially weighted moving averages (EWMA) in covariance matrices of financial returns. These place less and less weight on observations as they move further into the past, by using a 'smoothing' parameter λ. The larger the value of λ the more weight is placed on past observations and so the smoother the series becomes. An n-period EWMA of a time series x_t is defined as

$$\frac{x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \ldots\ldots + \lambda^n x_{t-n}}{1 + \lambda + \lambda^2 + \ldots\ldots + \lambda^n}$$

where $0 < \lambda < 1$. The denominator converges to $1/(1-\lambda)$ as $n \to \infty$, so an infinite EWMA may be written

$$\hat{s}_T^2 = (1-\lambda)\sum_{i=0}^{\infty} \lambda^{i-1} x_{T-i} = (1-\lambda)x_T + \lambda \hat{s}_{T-1}^2 \qquad (4)$$

Comparing (4) and (5) reveals that an infinite EWMA on squared returns is equivalent to an Integrated GARCH model with no constant term (see Engle and Mezrich, 1995).[13] In an Integrated GARCH model n-step ahead forecasts do not converge to the long-term average volatility level so an alternative method should be found to generate forecasts from volatility estimates. It is standard to assume, just as with equally weighted averages, that variances are proportional to time. In the empirical work of section 5 we take one-day forecasts to be EWMA

[13] Since Integrated GARCH volatility estimates are rather too persistent for many markets, this explains why many RiskMetrics™ daily forecasts of volatility do not 'die-away' as rapidly as the equivalent GARCH forecasts. The Third Edition of JP Morgans RiskMetrics™ uses an infinite EWMA with $\lambda = 0.94$ for all markets and x_t to be the squared daily return.

estimates with $\lambda = 0.94$, and employ the square root of time rule to produce 5, 10 and 25 day variance forecasts.[14]

4.3 Generalized Autoregressive Conditional Heteroscedasticity

The normal GARCH(1,1) model of Bollerslev (1986) is a generalisation of the ARCH model introduced by Engle (1982) which has a more parsimonious parameterization and better convergence properties. The simple GARCH(1,1) model is

$$\begin{aligned} r_t &= c + e_t \\ s_t^2 &= w + a\, e_{t-1}^2 + b\, s_{t-1}^2 \end{aligned} \tag{5}$$

where r_t denotes the daily return and σ_t denotes the conditional variance of ε_t , for t = 1, … T. In this 'plain vanilla' GARCH model the conditional variance is assumed to be normal with mean zero. Non-negativity constraints on the parameters are necessary to ensure that the conditional variance estimates are always positive, and parameters are estimated using constrained maximum likelihood as explained in Bollerslev (1986). Forecasts of variance over any future holding period, denoted $\hat{s}_{T,h}^2$, may be calculated from the estimated model as follows:

$$\begin{aligned} \hat{s}_{T+1}^2 &= \hat{w} + \hat{a}\, e_T^2 + \hat{b}\, \hat{s}_T^2 \\ \hat{s}_{T+s}^2 &= \hat{w} + (\hat{a} + \hat{b})\, \hat{s}_{T+s-1}^2 \qquad s > 1 \\ \hat{s}_{T,h}^2 &= \sum_{s=1}^{h} \hat{s}_{T+s}^2 \end{aligned} \tag{6}$$

---

[14] We do not use the 25-day forecasts which are produced by RiskMetrics™ because there are significant problems with the interpretation of these. To construct their 25-day forecasts, RiskMetrics™ have taken $\lambda = 0.97$ and x_t to be the 25-day historic variance series. Unfortunately this yields monthly forecasts with the undesirable property that they achieve their maximum 25 days after a major market event. It is easy to show why this happens: the monthly variance forecast is

$$\hat{s}_t^2 = (1-l)\, s_t^2 + l\, \hat{s}_{t-1}^2,$$

so clearly

$$\hat{s}_t^2 > \hat{s}_{t-1}^2 \quad \Leftrightarrow \quad s_t^2 > \hat{s}_{t-1}^2 .$$

The third equation gives the forecast of the variance of returns over the next h days. If $\alpha+\beta=1$ then the instantaneous forecasts given by the second equation will grow by a constant amount each day, and the h-period variance forecasts will never converge. This is the Integrated GARCH model. But when $\alpha+\beta<1$ the forecasts converge to the unconditional variance $\omega/(1-(\alpha+\beta))$ and the GARCH forward volatility term structure has the intuitive shape: upwards sloping in tranquil times and downwards sloping in volatile times.[15]

5. Evaluating the Volatility Forecasts

There is an extensive literature on evaluating the accuracy of volatility forecasts for financial markets. It is a notoriously difficult task, for several reasons. The results of operational evaluation, for example by using a trading metric, will depend on the metric chosen and not just the data employed. But even the more objective statistical evaluation procedures have produced very conflicting results.[16] The problem is that a volatility prediction cannot be validated by direct comparison with the returns data - this is only applicable to the mean prediction - and indirect means need to be used. In this paper we use both statistical and operational evaluation procedures, but none of the chosen methods is without its problems: Likelihood methods assume a known distribution for returns (normal is assumed, but is it realistic?); Root-mean-square-error measures need a benchmark against which to measure error (which?); Both statistical methods focus on the accuracy of the centre of the predicted returns distribution, whereas value-at-risk models require accuracy in the tails of the distribution; Operational evaluation focusses on the lower tail, but statistical errors in the procedure can be significant.

This means that the RiskMetrics™ variance estimate will continue to rise while the 25-day equally weighted average remain artificially high during the 'ghost feature'. But exactly 25 days after the extreme event which caused the feature, s_t^2 will drop dramatically, and so the maximum value of $\hat{s}_t^2$ will occur at this point.

[15] Some GARCH models fit the implied volatility term structure from market data better than others (see Engle and Mezrich, 1995, Duan, 1996). GARCH(1,1) give a montonically convergent term structure, but more advanced GARCH models can have interesting, non-monotonic term structures which better reflect market behaviour.

[16] See for example Brailsford and Faff (1996), Dimson and Marsh (1990), Figlewski (1994), Tse and Tung (1992), and West and Cho (1995).

5.1 Description of data

Daily closing prices on the five major equity indices, and the four corresponding US dollar exchange rates from 1-Jan-93 to 6-Oct-96 were used in this study. [17] GARCH volatility forecasts were made for the period 1-Jan-1996 to 6-Oct-1996 using a rolling three year window, starting on 1-Jan-1993, as a training data set. The steps in the GARCH calculation are as follows:

- In the GARCH model we set ε_t. = $R_t = \log(P_t / P_{t-1})$ where P_t is the index price or exchange rate at day t.

- The coefficients ω, α and β are optimised for the training dataset R_1 to R_{781}.
  Time t = 0 represents the day 1-Jan-93 and t = 781 represents the day 29-Dec-95.

- The first variance estimate, $_{782}^2$, for 1-Jan-96 is calculated using (5). The sequence is initialised with $_1^2$=0. The GARCH term structure forecasts are then made using (6). Term structure forecasts used in this study were 1-day, 5-day, 10-day and 25-day.

- For the following day a new set of optimised coefficients, ω, α and β are calculated for the dataset R_2 to R_{782}. The next variance estimate, $_{783}^2$ is calculated using equation (5), and term structure forecasts again made using (6).

- The procedure is repeated, rolling the estimation period one day at a time, to obtain values of $_{784}^2$ to $_{983}^2$.

EWMA variance estimates were also made for the period 1-Jan-96 to 6-Oct-96. Using (4) the smoothing constant λ was taken as 0.94, following JP Morgan's RiskMetrics. For the 'regulatory' forecasts a 250 day moving average of the squares of returns was taken for the variance estimates. So for example the one-day variance estimate for 1-Jan-1996 was calculated as the mean of the squares of returns for the 250 days from 16-Jan-1995 to 31-Dec-1995. Both exponentially and equally weighted estimates were scaled using the 'square-root-of-time' rule to obtain (constant) volatility term structure forecasts.

---

[17] Identical tests were performed on Sterling exchange rates, with very similar results.

5.2 Statistical Evaluation

If we assume conditional normality and zero means, forecasting the variance is equivalent to forecasting the probability density functions of returns, and we can evaluate their accuracy by measuring how well the forecasted distribution fits the actual data. This is exactly what likelihood methods do.[18] Assuming returns are normal with zero mean the density function is

$$f_t(r_t) = \frac{1}{\sqrt{2p s_t^2}} \exp\{-\frac{1}{2}\left(\frac{r_t^2}{s_t^2}\right)\}$$

and the likelihood of getting a set of returns r_1 , $r_{2,}$ .... r_N , is

$$L(s_1^2, s_2^2, ..., s_n^2 \mid r_1, r_2, ... , r_N) = \prod_{t=1}^{N} f_t(r_t)$$

The log likelihood is easier to work with, in fact if a multiple of log 2π is excluded, -2logL is given by

$$\sum_{t=1}^{N}((r_t^2 / s_t^2) + \log(s_t^2)) \qquad (7)$$

Likelihood methods are used to distinguish between the different models for variance forecasts by saving a test set from 1-Jan-96 to 6-Oct-96 as described above. For each of these points we make three h-day variance predictions for h = 1, 5, 10 and 25 using the 'regulatory', EWMA and GARCH methods. For each of these methods we compute the quantity (7) using the actual h-day returns data and putting σ_t equal to the volatility forecast pertaining to for the return r_t. The lower the quantity (7) the better is the forecast evaluation. Results for equity indices and US dollar exchange rates are given in table 1.

Table 2 reports the root mean squared error (RMSE) between squared h-day returns and the h-day variance forecasts over the test set. The RMSE is given by

[18] For more information on how the likelihood function can be used as a means of validating a volatility model Magdon-Ismail and Abu-Mostafa (1996).

$$\sqrt{(1/N)\sum_{t=1}^{N}(r_t^2 - \hat{s}_t^2)^2} \qquad (8)$$

Again, the smaller this quantity the better the forecast. It provides a metric by which to measure deviations between two series, but it has no other statistical foundation for evaluating accuracy of variance forecasts.[19] We state these results because likelihood methods assume conditional normality, but this assumption may be violated.[20]

Table 1: RMSE (*1000) for international equity markets and US dollar rates in 1996

	1 day			*5 day*			*10 day*			*25 day*		
	Reg	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**
DEM_SE	*0.1027*	0.1037	0.1042	0.2874	*0.2457*	0.3208	0.5967	*0.5036*	0.6775	1.5492	*1.3069*	1.8173
FRF_SE	0.0822	*0.0756*	0.0849	0.4182	*0.2843*	0.4448	0.8837	*0.6087*	0.9425	2.2889	*1.5956*	2.4647
GBP_SE	*0.0421*	0.0428	0.0427	0.1448	*0.1389*	0.1674	0.3118	*0.2983*	0.3626	0.8212	*0.7856*	0.9775
JPY_SE	0.1707	*0.1627*	0.1638	0.6392	*0.3855*	0.5074	1.3280	*0.7919*	1.0912	3.4206	*2.0587*	3.0542
USD_SE	*0.1127*	0.1128	0.1118	*0.1940*	0.2788	0.2118	*0.3878*	0.5753	0.4159	1.0159	1.4956	*1.0108*
DEM_XS	0.0405	*0.0348*	0.0357	0.1825	*0.0821*	0.1243	0.3774	*0.1680*	0.2667	0.9661	*0.4358*	0.7412
FRF_XS	0.0349	*0.0302*	0.0313	0.1499	*0.0696*	0.1086	0.3089	*0.1411*	0.2316	0.7897	*0.3645*	0.6396
GBP_XS	0.0180	*0.0164*	0.0172	0.0755	*0.0446*	0.0682	0.1571	*0.0929*	0.1461	0.4042	*0.2420*	0.4017
JPY_XS	0.0533	*0.0438*	0.0443	0.2562	*0.1094*	0.1525	0.5303	*0.2266*	0.3321	1.3573	*0.5903*	0.9439

[19] The RMSE is related to the normal likelihood when variance is fixed and *means* are forecast, not variances

[20] It is easy to test for unconditional normality (using 'QQ' plots or the Jarques-Bera normality test). We have found significant excess kurtosis in the historic unconditional returns distribution. But this does not in itself contradict the assumption of conditional normality: if volatility is stochastic outliers in the unconditional distribution can still be captured with time varying volatilities in the conditional distributions.

Table 2: -2logL/1000 for international equity markets and US dollar rates in 1996

	1 day			*5 day*			*10 day*			*25 day*		
	Reg	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**	**Reg**	**EWMA**	**GARCH**
DEM_SE	*-1.5502*	-1.5445	-1.5469	-1.3752	*-1.4129*	-1.3608	-1.2670	*-1.3093*	-1.2485	-1.1138	*-1.1554*	-1.0879
FRF_SE	-1.5149	*-1.5273*	-1.5094	-1.3188	*-1.3865*	-1.3066	-1.2068	*-1.2799*	-1.1934	-1.0477	*-1.1194*	-1.0348
GBP_SE	*-1.6245*	-1.6190	-1.6221	-1.4776	*-1.4864*	-1.4592	-1.3731	*-1.3831*	-1.3511	-1.2223	*-1.2343*	-1.1950
JPY_SE	-1.4463	-1.4580	*-1.4638*	-1.2582	*-1.3271*	-1.2867	-1.1456	*-1.2266*	-1.1729	-0.9815	*-1.0799*	-1.0049
USD_SE	-1.5095	-1.5231	*-1.5247*	*-1.4226*	-1.3864	-1.4176	-1.3288	-1.2812	*-1.3214*	*-1.1930*	-1.1316	-1.1861
DEM_XS	-1.7161	-1.7267	*-1.7334*	-1.5001	*-1.6056*	-1.5365	-1.3821	*-1.5012*	-1.4187	-1.2117	*-1.3486*	-1.2466
FRF_XS	-1.7509	-1.7634	*-1.7686*	-1.5333	*-1.6445*	-1.5654	-1.4153	*-1.5399*	-1.4466	-1.2459	*-1.3827*	-1.2720
GBP_XS	-1.8387	*-1.8407*	-1.8373	-1.6419	*-1.7084*	-1.6371	-1.5275	*-1.6029*	-1.5211	-1.3606	*-1.4462*	-1.3530
JPY_XS	-1.6589	-1.6912	*-1.6914*	-1.4363	*-1.5608*	-1.5027	-1.3175	*-1.4564*	-1.3839	-1.1469	*-1.2969*	-1.2044

The italic type in tables 1 and 2 denotes the model which performs best according to these data and statistical criteria. For the 5, 10 and 25 day forecasts of all series except US equities the EWMA has predicted returns in the test set with the greatest accuracy, according to both RMSE and likelihood results. The out-performance of EWMA over 'vanilla' GARCH for longer holding periods comes as no surprise. It is well known that normal GARCH(1,1) models do not fit market term volatility structures as well as asymmetric and/or components GARCH models, particularly in equity markets (see Engle and Mezrich, 1995, and Duan, 1996). It is more surprising that US equities do not seem to favour EWMA methods above either of the alternative models.

The statistical results for the evaluation of one-day forecasts is very mixed with no clear pattern emerging. Not only do the RMSE and likelihood results often conflict, but the results can change depending on the timing of the test set employed.[21] In the next section the one-day forecasts are evaluated operationally, and a much clearer picture emerges.

[21] Results available from authors on request.

5.3 Operational Evaluation

There is a problem with the use of RMSE or likelihoods to evaluate covariance matrices for value-at-risk models: these criteria assess the ability of the model to forecast the centre of returns distributions, but it is the accurate prediction of outliers which is necessary for value-at-risk modelling. A volatility forecasting model will have a high likelihood/low RMSE if most of the returns on the test set lie in the normal range of the predicted distribution. But since value-at-risk models attempt to predict worst case scenarios, it is really the lower percentiles of the predicted distributions that we should examine.

This can be attempted with an 'operational' evaluation procedure such as that proposed by the Bank for International Settlements. The BIS (1996b) have proposed a supervisory framework for operational evaluation by 'back testing' one-day value-at-risk measures.[22] The recommended framework is open to two interpretations which we call 'back' and 'forward' testing respectively. In 'back' tests the current 1% one-day value-at-risk measure is compared with the daily P&L which would have accrued if the portfolio had been held static over the past 250 days. In the 'forward' tests a value-at-risk measure was calculated for each of the past 250 days, and compared with the observed P&L for that day.

Over a one year period a 1% daily risk measure should cover, on average, 247 of the 250 outcomes, leaving three exceptions. Since type 1 statistical errors from the test 'reject the model if more than three exceptions occur' are far too large, the BIS have constructed three 'zones' within which internal value-at-risk models can lie. Models fall into the 'green zone' if the average number of exceptions is less than five; five to nine exceptions constitutes the 'yellow zone'; and if there are ten or more exceptions when a 1% model is compared to the last year of daily P&L the model falls into the 'red zone'. Models which fall into the yellow zone may be subject to an increase in the scaling factor applied when using the value-at-risk measure to allocate risk capital from 3 to between 3.4 and 3.85, whilst red zone models may be disallowed altogether since they are thought to seriously underestimate 1% value-at-risk.

[22] Back testing of static portfolios for longer holding periods is thought to be less meaningful, since it is common that major trading institutions will change portfolio compositions on a daily basis.

The thresholds have been chosen to maximize the probabilities that accurate models will fall into the green zone, and that greatly inaccurate models will be red zone. With the red zone threshold set at 10 exceptions there is only a very small probability of a type one error, so it is very unlikely that accurate models will fall into the red zone. But both accurate and inaccurate models may be categorised as yellow zone, since both type one and type two statistical errors occur. The yellow zone thresholds of 5-9 have been set so that outcomes which fall into this range are more likely to have come from inaccurate than from accurate models.[23]

Table 3 reports the results of back tests on equity indices of the three different types of volatility forecasts. The test could be run by comparing the historical distribution of the daily change in price of the index during the last 250 days with the lower 1%-ile predicted by multiplying the current one-day returns standard deviation forecast by 2.33 times the current price. However if markets have been trending up/down this can lead to over/under estimating value-at-risk. So we use the historical distribution of returns, rather than price changes, and count the number of observations in the tail cut off by -2.33 times the one-day returns forecast. For each of the 200 days in the test set (1-Jan-96 to 6-Oct-96) we generate the historical empirical distribution of returns over the last 250 days, and count the number of exceptions according to the current one-

[23] It would be imprudent to already reject a model into the yellow zone if it predicts four exceptions in the back testing sample, since accurate models have 24.2% chance of generating four or more exceptions. If the null hypothesis is that 'the model is accurate' and the decision rule is 'reject the null hypothesis if the number of exceptions is $\geq x$', then a 'type one' statistical error consists of rejecting an accurate model. So put another way, the probability of a type one error is 0.242 if we set $x = 4$. This probability is also the significance level associated with the test: if the threshold for green/yellow zone models were set at $x = 4$ the significance level of the test would be only 24.2% - we would have only 75.6% confidence in the results! The threshold is therefore raised to five, which reduces the probability of a type one error to 0.108, and gives a test with a higher significance level: accurate models have a 10.8% chance of being erroneously categorised as yellow zone and we are almost 90% confident that the conclusion will be accurate. To raise the significance level of back tests to 1% the BIS would have to accept models into the green zone if they generate as many as 7 exceptions, but this increases the probability of a 'type two' statistical error. In a fixed sample size (250) there is a trade-off between type one and type two statistical errors: it is impossible to simultaneously decrease the probability of both. The 'type two' error is to erroneously accept an inaccurate model, and this will depend on the degree of inaccuracy. For example with the rule 'reject the null hypothesis if there are seven or more exceptions in the sample size 250', an inaccurate model which is really capturing 2% rather than 1% of the exceptional P&Ls would have a type two error of 0.764 , that is it would have a 76.4% chance of being classified in the green zone. A 3% value-at-risk model would have a 37.5% chance of being erroneously accepted and a 4% model would be accepted 12.5% of the time. To reduce these probabilities of type two errors, the green zone threshold is set at $x = 5$.

day returns standard deviation forecast. In table 3 we report the average number of exceptions, over all 200 back tests of each volatility forecast.

Table 3: Average number of exceptions in BIS back tests during 1996

	Regulatory	**EWMA**	**GARCH**
DEM_SE	4.241206	6.778894	3.758794
FRF_SE	3.492462	6.2864	3.301508
GBP_SE	2.256281	3.527638	1.768844
JPY_SE	1.537688	8.256281	4.834171
USD_SE	5.668342	3.653266	5.135678
DEM_XS	5.366834	14.40704	9.361809
FRF_XS	4.763819	15.70854	7.819095
GBP_XS	4.386935	9.366834	4.763819
JPY_XS	3.482412	12.1005	6.81407

Instead of comparing the current forecast with the last 250 returns observed over the previous year, and averaging results over the whole test set we can compare the one-day volatility forecasts made for each day in our test set with the observed P&L for that day. If the change in value of the equity index or exchange rate falls below the lower 1%-ile of the predicted P&L distribution for that day, it is counted as an outlier. This type of 'forward' testing is done over for each of the 200 days in the test set and the total number of outliers recorded in table 4. An accurate 1% value-at-risk model would give 2 exceptions from a total of 200 comparisons, but to allow for type 1 and type 2 errors as in the 'back' testing procedure just outlined, models should be classified as 'green zone' if they yield less than four exceptions, 'yellow zone' if they give 4-8 exceptions, and 'red zone' if more than 8 exceptions are recorded. The results are reported in table 4.

Results from both 'back' and 'forward' tests paint the same general picture: with the exception of US equities, both GARCH and the equally weighted 'regulatory' model would be classified as 'green zone' by the BIS and their value-at-risk measures would be multiplied by 3.0 to calculate risk capitial requirements. But for the US equity index the GARCH and 'regulatory' models would be 'yellow zone', and therefore subject to a capital requirement multiplier of between 3.4 and 3.8.

Table 4: Number of exceptions in BIS forward tests during 1996

	Regulatory	EWMA	GARCH
DEM_SE	2	4	2
FRF_SE	0	1	0
GBP_SE	2	3	2
JPY_SE	1	4	2
USD_SE	8	6	8
DEM_XS	1	6	1
FRF_XS	1	5	1
GBP_XS	2	4	2
JPY_XS	0	5	2

Operational evaluation of the prediction of lower tails of one-day returns distributions gives results which contrast the success of the EWMA method in predicting the centre of the distribution: for all but US equities the EWMA model would give ‘yellow zone’ results at best. Indeed in many cases, and for exchange rates in particular, an EWMA model would be classified as ‘red zone’, since value-at-risk measures appear to be rather too low. However for US equities EWMA methods seem better at predicting the tails than the centre, and ‘back’ tests (but not forward tests) on US equities would imply a ‘green zone’ value-a-risk model.

6. Summary and Conclusions

This paper examines the covariance matrix of risk factor returns forecasts which is often used in value-at-risk models. Common methods of measuring value-at-risk which are crucially dependent on accurate covariance matrices are described and a general framework for building large positive definite covariance matrices is proposed. This method requires only univariate volatility forecasting procedures, so the paper attempts to assess the accuracy of the three most common methods of volatility forecasting: equally and exponentially weighted moving averages, and ‘plain vanilla’ GARCH.

Data on major equity markets and US dollar exchange rates are employed, with a test set running from 1-Jan-96 to 6-Oct-96, a total of 200 data points. The results show that whilst EWMA methods are better at predicting the centre of longer-term returns distributions, their prediction of the lower 1%-iles are too high. Thus value-at-risk measures may be too low, at least according to regulators recommendations. On the other hand, the standard normal GARCH(1,1) model (which makes no allowance for the asymmetry of returns distributions) does not perform well according to statistical criteria which measure the centre of the distribution, although it would generally give 'green zone' models in operational 'back tests'. Thus GARCH models give more conservative risk capital estimates which more accurately reflect a 1% value-at-risk measure. The one exception in these general statements is the US equity market, and there the results are reversed: either a one-year equally weighted average or a vanilla GARCH model performs better in the statistical tests, whilst the EWMA model out-performs both of these in the operational evaluation.

The paper has focused on the inherent difficulties of evaluating volatility forecasts, be it for trading or for value-at-risk purposes. There are many different statistical or operational criteria which could be used to evaluate a volatility forecasting model, and test results may also depend on the data period employed. This investigation has not attempted any general statement that one method is universally superior to another, such a conclusion would seems fallacious given the complexity of the evaluation process. Rather, we would like to highlight the need for value-at-risk scenario analysis which perturbs covariance matrices by small amounts to reflect the inaccuracies which one normally expects in standard statistical volatility forecasting methods.

REFERENCES

Alexander, C.O. (1996) “Evaluating RiskMetrics as a risk measurement tool for your operation: What are its advantages and limitations?” *Derivatives: Use, Trading and Regulation* **2** No. 3 pp277-284.

Alexander C.O. and Chibumba (1996) “Multivariate orthogonal factor GARCH” *University of Sussex, Mathematics Dept. discussion paper*

Bank for International Settlements (1996a) “Amendment to the Capital Accord to incorporate market risks”

Bank for International Settlements (1996b) “Supervisory framework for the use of ‘backtesting’ in conjunction with the internal models approach to market risk capital requirements”

Bollerslev, T. (1986) "Generalised autoregressive conditional heteroscedasticity". *Journal of Econometrics* **31** pp 307-327.

Bollerslev, T, RF Engle and D Nelson (1994) "ARCH models" in *Handbook of Econometrics* volume 4 (North Holland) pp2959-3038

Brailsford T.J. and Faff R.W. (1996) “ An evaluation of volatility forecasting techniques” *Journal of Banking and Finance* **20** pp419-438

Clemen, R.T. (1989) “ Combining forecasts: a review and annotated bibliography” *International Journal of Forecastsing* **5** pp559-583

Dimson, E. and P. Marsh (1990) “Volatility forecasting without data-snooping” *Journal of Banking and Finance* **14** pp399-421

Duan, JC (1996) ‘Cracking the smile’ *RISK* **9** pp 55- 59

Engle, R.F. (1982) "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". *Econometrica* **50**:4 pp 987-1007.

Engle, RF and J Mezrich (1995) "Grappling with GARCH" *RISK* **8** No9 pp 112-117

Figlewski, S. (1994) "Forecasting volatility using historical data" *New York University Salomon Center (Leonard N. Stern School of Business) Working Paper Series* no. S-94-13

Magdon-Ismail, M. and Y.S. Abu-Mostafa (1996) “Validation of volatility models” *Caltech discussion paper*

JP Morgan (1995) “RiskMetrics™ third edition” *http://www.jpmorgan.com/RiskManagement/ RiskMetrics/pubs.html*

Tse, Y.K. and S.H. Tung (1992) “Forecasting volatility in the Singapore stock market” *Asia Pacific Journal of Management* **9**, pp1-13

West, K. D. and D. Cho (1995) “The predictive ability of several models of exchange rate volatility” *Journal of Econometrics* **69** pp367-391

Acknowledgements

Many thanks to Professor Walter Ledermann and Dr Peter Williams of the University of Sussex for very useful discussions, and to the referees of this paper for their careful, critical and constructive comments.