

**University of California, San Diego
Department of Economics**

Reality Check for Volatility Models

Ricardo Suganuma

Department of Economics, 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
rsuganum@weber.ucsd.edu

Abstract

Asset allocation decisions and value at risk calculations rely strongly on volatility estimates. Volatility measures such as rolling window, EWMA, GARCH and stochastic volatility are used in practice. GARCH and EWMA type models that incorporate the dynamic structure of volatility and are capable of forecasting future behavior of risk should perform better than constant, rolling window volatility models. For the same asset the model that is the 'best' according to some criterion can change from period to period. We use the reality check test* to verify if one model out-performs others over a class of re-sampled time-series data. The test is based on re-sampling the data using stationary bootstrapping. For each re-sample we check the 'best' model according to two criteria and analyze the distribution of the performance statistics. We compare constant volatility, EWMA and GARCH models using a quadratic utility function and a risk management measurement as comparison criteria. No model consistently out-performs the benchmark.

JEL codes: C.12, C.13, C.14, C.15, C.22, C.52, C.53 and G.11.

Keywords: bootstrap reality check, volatility models, utility-based performance measures and risk management.

This version: October, 2000

* Hal White (1997, 2000)

Reality Check for Volatility Models

Ricardo Sukanuma

1. INTRODUCTION

No financial analyst would disagree with the statement that volatility estimation is essential to asset allocation decisions. However if asked what model should be used to calculate volatility, several answers might be given. The reason for this lack of consensus on the subject is that there is a notion that different volatility models can result in considerably different estimates. One of the areas in which these different estimates can have serious effects is risk management.

Financial institutions are required to measure the risk of their portfolio to fulfill capital requirements. J. P. Morgan was the first institution to open its risk management methodology, *Riskmetrics*, to other financial institutions. Because of this, *Riskmetrics* became very influential and it is still considered a benchmark in risk management. However, two points were frequently criticized: the way *Riskmetrics* calculates volatility forecasts and the assumption that assets returns are Normally distributed. J. P. Morgan proposed using an exponential smoothing with the same smoothing constant for all assets to calculate volatility and covariances between assets. In the following years, several articles in practitioners' journals and academic papers discussed how appropriate the exponential weighting moving averages model is for calculating volatility and proposed other models that would be more suitable for estimating risk. Several of these papers presented one or more time series trying to demonstrate that an alternative model has a better performance in forecasting risk. The question that arises in this discussion is if these "good" forecasts are due to data snooping.

Until recently there was no simple way to test the hypothesis that the performance of the "best" model is not superior to the performance of the benchmark model. In a recent paper, White (2000) proposes a method for testing the null hypothesis that the best model encountered during a specification search has no predictive superiority over a benchmark.

The objective of this paper is to apply White's Bootstrap Reality Check to the problem of selecting the “best” volatility model based on utility and value-at-risk measures. In the next section we discuss volatility estimation in risk management. Section 3 describes the theory and methodology used in this paper. In Section 4, we explain the performance statistics used in the hypothesis testing and we apply the White’s Reality Check to a model with one risky asset, S&P500 and one risk free asset, the 3-month Treasury bill. Section 5 concludes.

2. VOLATILITY MODELS COMMONLY USED IN RISK MANAGEMENT

The notion that not only expected returns, but also risk are important to portfolio decisions is relatively old and goes back to Markowitz (1952). However only during the nineties did financial institutions begin the practice of having a daily report of how much they could loose in an adverse day. The recent experiences of financial failures such as Proctor and Gamble, Metallgesellschaft, Orange County and Barings and the regulation requirement to have a system to calculate risk started the discussion on how a financial institution should measure its exposure to risk.

Value-at-Risk is used to measure market risk, i.e. uncertainty of future earnings resulting from adverse changes in market conditions (prices of assets, exchange rates and interest rates) and can be defined as “the maximal loss for a given probability over a given period.” For a given horizon n and confidence level p , the Value-at-Risk is the loss in the market value over the time horizon n that is exceeded with probability $1-p$.

According to the *Riskmetrics* methodology, the Value-at-Risk (or VaR) for a single asset can be expressed as:

$$VaR = \text{Market value of position} \times \text{Sensitivity to price move per \$ market value} \\ \times \text{Adverse price move per period}$$

where adverse price move per period, ΔP , is

$$\Delta P = \gamma \sigma_t$$

and $\gamma = 1.65, 1.96$ or 2.33 for $p=10\%, 5\%$ and 1% , respectively if the asset return follows a Normal distribution, and σ_t is the volatility.

Considering that firms hold a portfolio of several assets and the correlations among the assets affect the loss in market value, the Value-at-Risk of a portfolio is

$$VaR = \sqrt{\bar{V} * C * \bar{V}^T}$$

where V is the n row vector of VaR 's for each single asset,

C is the $n \times n$ correlation matrix, and

V^T is the transpose of V .

Therefore, the estimation of volatility of the assets' returns and their correlations is essential to a correct calculation of this measure of market risk. The most common methods to estimate volatility are: *a)* the historical/moving average; *b)* the exponentially weighted moving average; *c)* GARCH models.

Historical/Rolling Window Moving Average Estimator (MA(n))

The historical or n -period rolling window moving average estimator of the volatility corresponds to the standard deviation and it is given by the square root of the expression

$$\hat{\sigma}_{t+1}^2 = \frac{1}{n} \sum_{s=t-n+1}^t (r_s - \mu)^2,$$

where r_s is the return of the asset at period s and μ is the mean return of the asset.

The advantages of this estimator are that it is very simple to calculate and that except for the window size it does not involve any kind of estimation. The size n is critical when one considers the effect of an extremely high or low observation in the sense that the smaller the size of the window, the bigger the effect on volatility.

Exponential Weighted Moving Averages Estimator (EWMA(λ))

This type of volatility estimator is commonly used in risk management calculations and is given by the square root of

$$\hat{\sigma}_{t+1}^2 = \lambda \hat{\sigma}_t^2 + (1 - \lambda)(r_t - \mu)^2$$

where λ is the decay factor, also known as the smoothing constant. In this method, the weights are geometrically declining, so the most recent observation has more weight compared to older ones. This weighting scheme helps to capture the dynamic properties of the data. *Riskmetrics* proposes the use of a common decay factor for all assets for a given periodicity. The smoothing constants are 0.94 for daily data and 0.97 for monthly data. By using the same constant decay factor, it simplifies the calculations of large-scale covariance matrices and eliminates the estimation aspect of problem. One drawback is that the h -period ahead ($h > 0$) forecast is the same as the 1-period ahead forecast.

GARCH(p,q) Models

Created by Engle (1982) and Bollerslev (1986), the GARCH family of models became very popular in financial applications. The GARCH(1,1) is the most successful specification. The general formulae for the conditional variance is

$$\hat{\sigma}_{t+1}^2 = \omega + \sum_{s=1}^q \alpha_s (r_{t+1-s} - \mu)^2 + \sum_{s=1}^p \beta_s \hat{\sigma}_{t+1-s}^2 .$$

As can be noted, the EWMA model is a special case of the IGARCH(1,1) model, with ω equal to zero. Even though GARCH is considered the model that best captures the dynamic nature of volatility, the computational complexity involved in its estimation, especially when dealing with covariance matrix estimates of several assets, is pointed to as the reason why it is less used in risk management than the exponential smoothing method.

Considering that the models above are used for estimating risk, the issue is that different ways to estimate volatility can lead to very different Value-at-Risk calculations.

Therefore, the natural thing to ask is which model is the “best” one. We present below some arguments used to answer this question:

"In the "historical model", all variations are due only to differences in samples. A smaller sample size yields a less precise estimate, the larger the sample the more accurate the estimate... Now, whether we are using 30-day or 60-day volatility, and whether we are taking the sample period to be in 1989 or in 1995, we are still estimating the *same thing*: the unconditional volatility of the time series. This is a number, a constant, underlying the whole series. So variation in the n-period historic volatility model, which we perceive as variation over time, is consigned to sample error alone... Hence the historical model is taking no account of the dynamic properties of the model." [Alexander (1996, p.236)].

"Concerning volatility estimation ... it doesn't make much sense to compare the models, for the simple reason that Arch estimation is much more complicated and unstable." [Longerstaeey and Zangani in *Risk Magazine*, January 1995, p.31].

"Given GARCH models' extra complexity and relative small predictive improvement for the majority of risk management users, **we have elected to calculate the volatility and correlations ... using exponential moving averages.**" [*Riskmetrics – Technical Document*(1995)].

"We should expect the Arch model to be superior to the exponential method because of the implicit weighting of the observations and the effective length of the moving window are chosen by the data." [Lawrence and Robinson, *Risk Magazine*, January 1995, p.26].

However, in order to select the best volatility model, the problem of data snooping possibly arises. For example, in the *Riskmetrics – Technical Document*, the statement

"[after] an extensive experimentation and analysis, we found the optimal decay factor for daily volatility to be 0.94 while for monthly volatility the optimal decay factor is 0.97."

is very suspicious.

The data snooping problem can be summarized in the following sentence

"Whenever a ‘good’ forecasting model is obtained by an extensive specification search, there is always the danger that the observed good performance results not from the actual forecasting ability, but is instead just luck." [White (2000), p.1097].

3. THEORY AND METHODOLOGY

The Bootstrap Principle

Suppose that, given a functional f_t , one wishes to determine that value t_0 of t that solves an equation such as

$$E\{f_t(F_0, F_1)|F_0\}=0,$$

where $F=F_0$ denotes the population distribution function and $\hat{F}=F_1$ is the distribution function of the sample. To obtain an approximate solution to this population equation, one can use

$$E\{f_t(F_1, F_2)|F_1\}=0,$$

where F_2 denotes the distribution function of a sample drawn from F_1 . Its solution \hat{t}_0 is a function of the sample values. \hat{t}_0 and the last expression are called the "bootstrap estimators" of t_0 and the population equation, respectively. In the bootstrap approach, inference is based on a sample of n random (i.i.d.) observations of the population. Samples drawn from the sample are used for inference.

Stationary Bootstrap

Assume that $\{X_n, n \in \mathbf{Z}\}$ is a strictly stationary and weakly dependent time series. Suppose that μ is a parameter of the joint distribution of this series and we are interested in making inferences about this parameter based on an estimator $T_N(\mathbf{X})$. The stationary bootstrap method of P&R(1994) allow us to approximate the distribution of T_N . Let X^* be randomly selected form the original series, $X^*_1 = X_{I_1}$. Let X^*_2 be randomly selected, with probability q from the original series and, with probability $1-q$ be the next observation of the original series¹, i.e., $X^*_2 = X_{I_1+1}$. We repeat this procedure until we obtain X^*_1, \dots, X^*_N .

¹ Note that this method is equivalent to the moving block bootstrap method proposed by Kuensch (1989), except that the block lengths are random with mean length $1/q$.

Proposition 1 (P&R (1994)): Conditional on $X_1, \dots, X_N, X_1^*, \dots, X_N^*$ is stationary.

Proof: see P&R (1994, p.1304).

White's Bootstrap Reality Check

Suppose that we want to make a forecast for P periods from R to T and our interest here is to test a hypothesis about a vector of moments, $E(f)$, where $f = f(Z, \beta^*)$, for a random vector Z and pseudo-true parameters β^* , and the statistic to be used is

$$\bar{f} = P^{-1} \sum_{t=R}^T f_{t+\tau}(\hat{\beta}_t)$$

$f_{t+\tau}(\beta) = f(Z_{t+\tau}, \beta)$ and the observed data is generated by $\{Z_t\}$, a strong α -mixing sequence with marginal distributions identical to the distribution of Z .

Now, assume that we have l different specification under consideration. Our task is to choose the model with the best value for the selection criteria and the null hypothesis is that no model has a superior performance than the benchmark model, that is

$$H_0: \max_{k=1, \dots, l} E(f_k) \leq 0$$

Under certain regularity condition [West (1996)],

$$P^{1/2}(\bar{f} - E(f)) \xrightarrow{d} N(0, \Omega)$$

assuming that $E\left[\frac{\partial}{\partial \beta} f(Z, \beta^*)\right] = 0$ or $\lim_{T \rightarrow \infty} n/R = 0$.

Moreover,

$$\max_{l=1, \dots, k} P^{1/2}(\bar{f} - E(f)) \xrightarrow{d} V \equiv \max_{l=1, \dots, k} \{Z_k\},$$

where Z is a vector with components Z_k , $k = 1, \dots, l$, distributed as $N(0, \Omega)$.

Therefore, we can approximate the P -value for the test of the null hypothesis based on the value of our predictive model selection criterion (White's Reality Check P -Value). The problem is that the distribution of the extreme value of a vector of correlated Normals for the general case is not known. White (2000) proposes the use of the stationary bootstrap to handle dependent processes.

The purpose of this paper is to test if some volatility model performs better than the model used in the benchmark methodology Value-at-Risk, *i.e.*, EWMA with decay factor $\lambda = 0.94$ for daily returns. To do this, we need to define our performance statistics, which are based on a quadratic utility function and VaR coverage.

Utility Based Performance Statistics

The idea of using utility functions to evaluate models is not new. West *et al* (1996) use a quadratic utility function to compare forecasts of exchange rate volatility and Engle and Mezrich (1997) propose a quadratic utility function to evaluate different volatility models used in Value-at-Risk calculations.

We assume that the investor has a quadratic time-independent additive utility function and there are only two assets, a risky asset and a risk free asset. Therefore, the agent will solve the following problem:

$$\text{Max } E_t U_{t+1} \equiv E_t (W_{t+1} - 0.5 \lambda W_{t+1}^2), \quad (3.1)$$

subject to

$$W_{t+1} = \omega_{t+1} r_{t+1} + (1-\omega_{t+1}) r_{ft+1}, \quad (3.2)$$

where E_t is the expectation at $t+1$, U_{t+1} is the utility at $t+1$, W_{t+1} is the return of the portfolio at $t+1$, λ is a constant associated with the degree of relative risk aversion, ω_{t+1} is the weight of the risky asset in the portfolio, r_{t+1} is the return of this asset at $t+1$ and r_{ft+1} is the return on the risk free asset in $t+1$.

From the first order conditions, we obtain

$$\omega_{t+1} = \frac{E_t \{y_{t+1} - \lambda y_{t+1} r_{f,t+1}\}}{E_t y_{t+1}^2} \quad (3.3)$$

where $y_{t+1} = r_{t+1} - r_{f,t+1}$, i.e. the excess return of the risky asset.

Assuming that $E_t y_{t+1} r_{f,t+1} = 0$, as in West *et al* (1996), and that $r_f = r_{f,t+1}$, i.e. the risk free is constant, and denoting $\mu_{t+1} = E_t y_{t+1}$ and h_{t+1} the conditional variance of y_{t+1} given the information at t , we have

$$\omega_{t+1} = \frac{1}{\lambda} \frac{\mu_{t+1}}{\mu_{t+1}^2 + h_{t+1}} \quad (3.4)$$

Thus, the realized utility in period $t+1$ is given by

$$U_{t+1} = \frac{1}{\lambda} \frac{\mu_{t+1}}{\mu_{t+1}^2 + h_{t+1}} (1 - \lambda r_f) y_{t+1} - 0.5 \frac{1}{\lambda} \frac{\mu_{t+1}^2}{(\mu_{t+1}^2 + h_{t+1})^2} y_{t+1}^2 + r_f - 0.5 \lambda r_f^2 \quad (3.6)$$

or,

$$U_{t+1} = \frac{1}{\lambda} \left\{ \frac{\mu_{t+1} y_{t+1} (1 - \lambda r_f)}{\mu_{t+1}^2 + h_{t+1}} - \frac{1}{2} \frac{\mu_{t+1}^2 y_{t+1}^2}{(\mu_{t+1}^2 + h_{t+1})^2} \right\} + r_f + 0.5 \lambda r_f^2. \quad (3.7)$$

As can be seen, it is sufficient for our analysis to use only the term in brackets,

$$u_{t+1} \equiv \frac{\mu_{t+1} y_{t+1} (1 - \lambda r_f)}{\mu_{t+1}^2 + h_{t+1}} - 0.5 \frac{\mu_{t+1}^2}{(\mu_{t+1}^2 + h_{t+1})^2} y_{t+1}^2 \quad (3.8)$$

since the other terms are common to all volatility estimators.

The expression above can be simplified by assuming that the investor faces a mean-variance utility maximization problem, i.e., $\max E_t U_{t+1} \equiv E_t (W_{t+1}) - 0.5 \lambda E_t (W_{t+1} - E_t(W_{t+1}))^2$ subject to constraint (2.2). In this case, the optimal portfolio weight of the risky asset is given by

$$\omega_{t+1} = \frac{1}{\lambda} \frac{\mu_{t+1}}{h_{t+1}} \quad (3.4')$$

and the realized utility is

$$U_{t+1} = \frac{1}{\lambda} \left\{ \frac{\mu_{t+1} y_{t+1}}{h_{t+1}} - \frac{1}{2} \frac{\mu_{t+1}^2 (y_{t+1} - \mu_{t+1})^2}{h_{t+1}^2} \right\} + r_f. \quad (3.7')$$

Again, it is sufficient for our analysis to use only the following term

$$u_{t+1} \equiv \frac{\mu_{t+1} y_{t+1}}{h_{t+1}} - 0.5 \frac{\mu_{t+1}^2}{(h_{t+1})^2} (y_{t+1} - \mu_{t+1})^2. \quad (3.8)$$

Our utility-based performance statistic is

$$\bar{f}_k \equiv \frac{1}{P} \sum_{t=R}^T f_{t+\tau}(\hat{\beta}_t)$$

where

$$f_{t+1} \equiv u_{t+1}(y_{t+1}, \hat{\beta}_{k,t}) - u_{t+1}(y_{t+1}, \hat{\beta}_{0,t}).$$

VaR Coverage Based Performance Statistics

The idea behind the coverage statistics is that given a series of volatility forecasts, the Value-at-Risk estimates should ideally cover $(1-p)$ percent of the observed losses. If we assume that the investor is holding a portfolio with only one asset, the one-day holding period *VaR* would be calculated using

$$\begin{aligned} VaR = & \text{Market value of position} \times \text{Sensitivity to price move per \$ market value} \\ & \times \text{Adverse price move per day} \end{aligned}$$

To simplify the exposition, suppose that the asset in consideration is a stock or a market index implying that the sensitivity to a price move is one and we normalize the

value of the portfolio so that the market value of the portfolio is equal to one. In this case, the *VaR* for one day is the same as the adverse price move per day

$$\Delta P = \gamma \sigma_t.$$

Assuming Normality, γ should be 1.645 for $p = 5\%$, say. However, given the strong empirical evidence that returns do not follow a Gaussian distribution, we propose the following method to calculate the value of γ using the last d days of the in-sample period:

$$\inf \gamma \quad s.t. \quad d^{-1} \sum_{t=T-P-d+1}^{T-P} 1\{y_t < -\gamma \hat{\sigma}_{j,t}\} = p$$

where $1\{\cdot\}$ is the indicator function.

That is, find the *inf* of γ such that if we compare the realized returns and the estimated adverse price move, given by product of the factor γ and the estimated volatility, only in 5% of the times the return is less than estimated adverse move. In other words, we select the smallest γ necessary to “cover” the loss 95% of the times.

Once we have estimated γ we can use it to define a mean square error type performance criterion²:

$$\hat{f}_{k,t+1} = -(1\{y_{t+1} < -\hat{\gamma} \hat{\sigma}_{k,t+1}\} - p)^2 + (1\{y_{t+1} < -\hat{\gamma} \hat{\sigma}_{0,t+1}\} - p)^2.$$

4. Testing the Performance of Volatility Models

The data are the daily closing prices of the S&P500 Index and the 3-month T-bills from January 1, 1988 to April 8, 1998. The stock index data is transformed into rate of returns using the log-differences. Figure 1 presents the closing price and daily returns on S&P500 and figure 2 the interest rate on 3-month Treasury Bills.

² To the best of our knowledge, this type of criterion, mean square error of the coverage rate, has not been discussed in the literature.

Figure 1

S&P500 Index and Returns

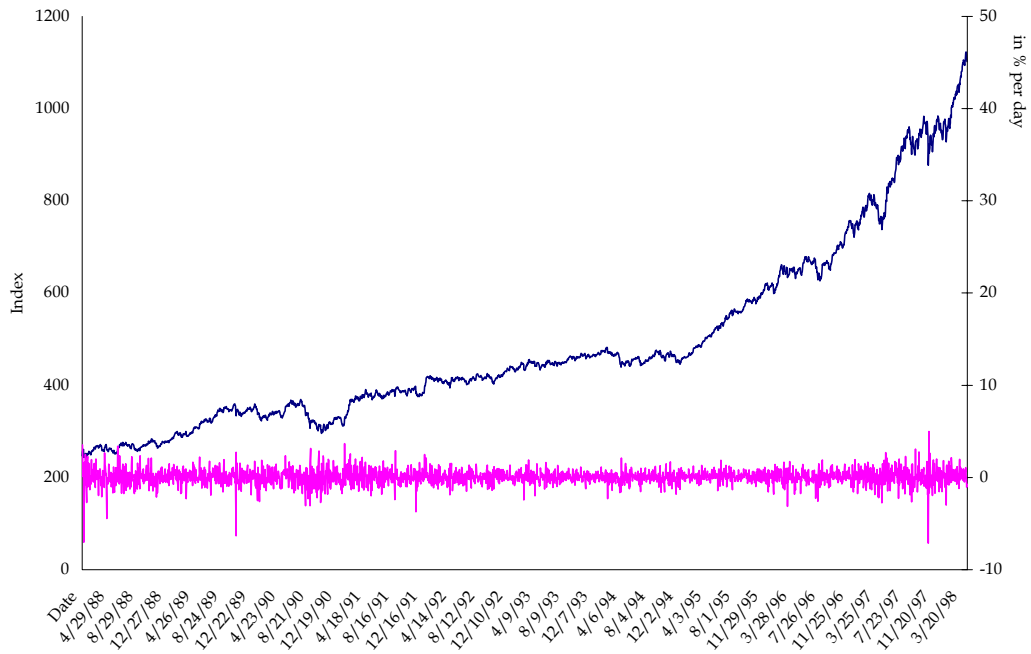


Figure 2

Treasury Bill - 3 Months



The prediction window used is $P = 504$ days, and the number of resamples is 1000. We use eleven different models: EWMA with $\lambda = 0.94, 0.97$ and 0.99 , MA with 5, 10, 22, 43, 126, 252 and 504 day windows and a GARCH(1,1). All the models were estimated assuming the mean return to be zero³, as in White(1997b). However, in order to use our utility-based performance measure, we need later to assume that the mean was equal to its unconditional value of the in-sample period. This procedure was followed in order to minimize the effect of the estimation of the mean, so our focus is only on the volatility forecasts. We use the EWMA ($\lambda = 0.94$) model as our benchmark. The parameters for the GARCH(1,1) model are kept constant and equal to their in-sample period values, removing any estimation aspect of the volatility forecasts. The probability q of the stationary bootstrap was set equal to 0.1. The out-of-sample volatility forecasts are presented in figures 3 to 6. As expected, EWMA models with a higher decay factor are smoother than the one with small λ (Figure 4). The same is true for MA models with greater n (Figure5). The similarity of some models can be noticed in figure 6. The similarity between EWMA(0.94) and GARCH(1,1) is not surprising, since the estimated value for β in the GARCH equation is close to 0.94 and the sum of α and β is close to 1.

³ For daily returns the mean of the excess return on the S&P500 index is close to zero. By assuming that it is zero we avoid the uncertainty related to its estimation, which is not likely to improve our analysis in terms of reduction in bias.

Figure 3

Volatility Forecasts

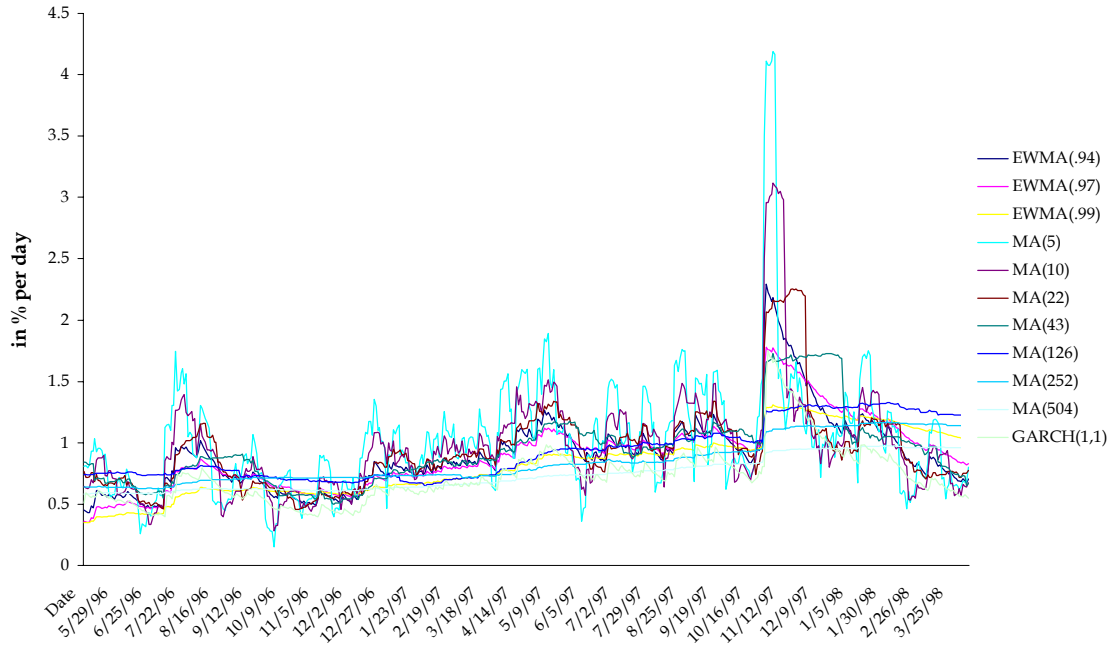


Figure 4

Exponential Moving Averages Forecasts

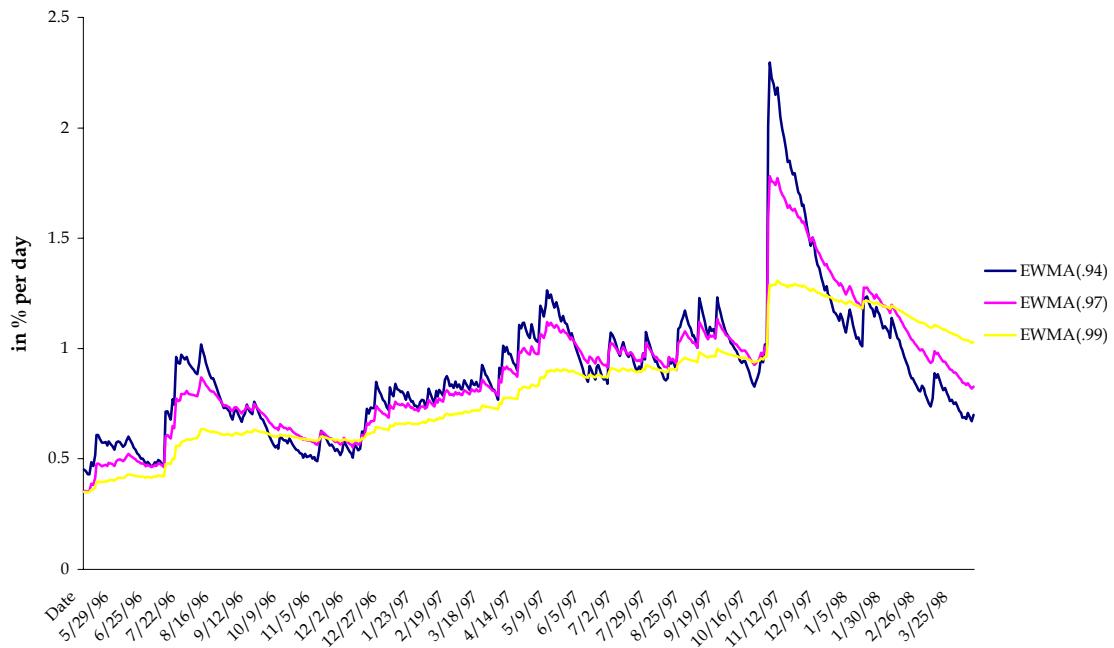


Figure 5

Moving Averages Forecasts

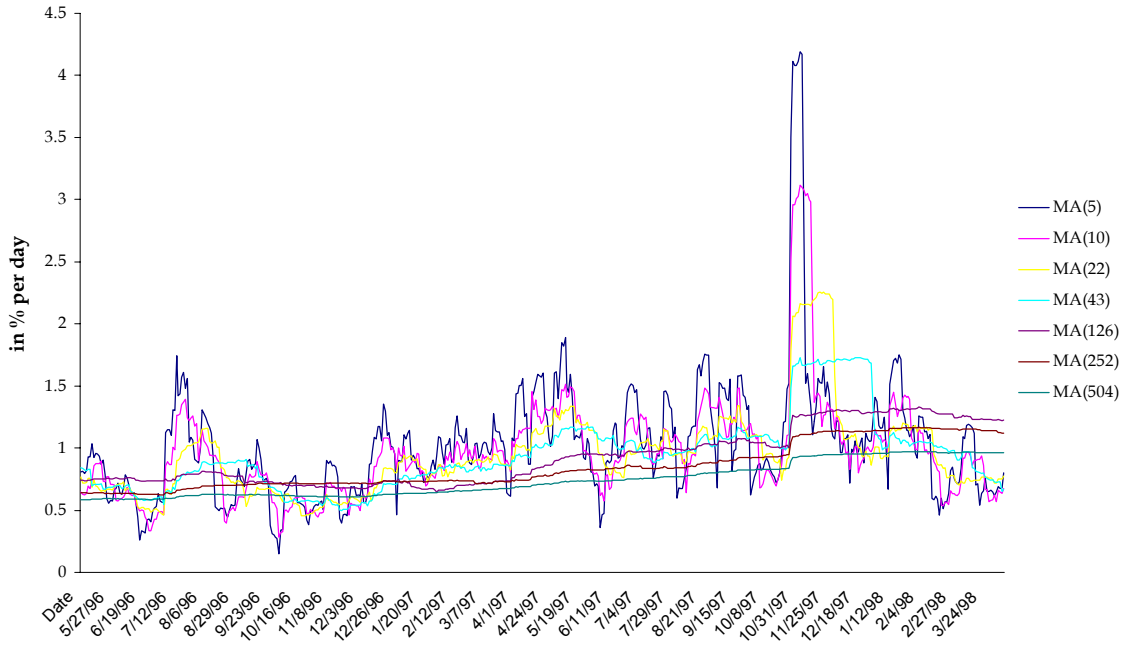


Figure 6

MA, EWMA and GARCH Forecasts

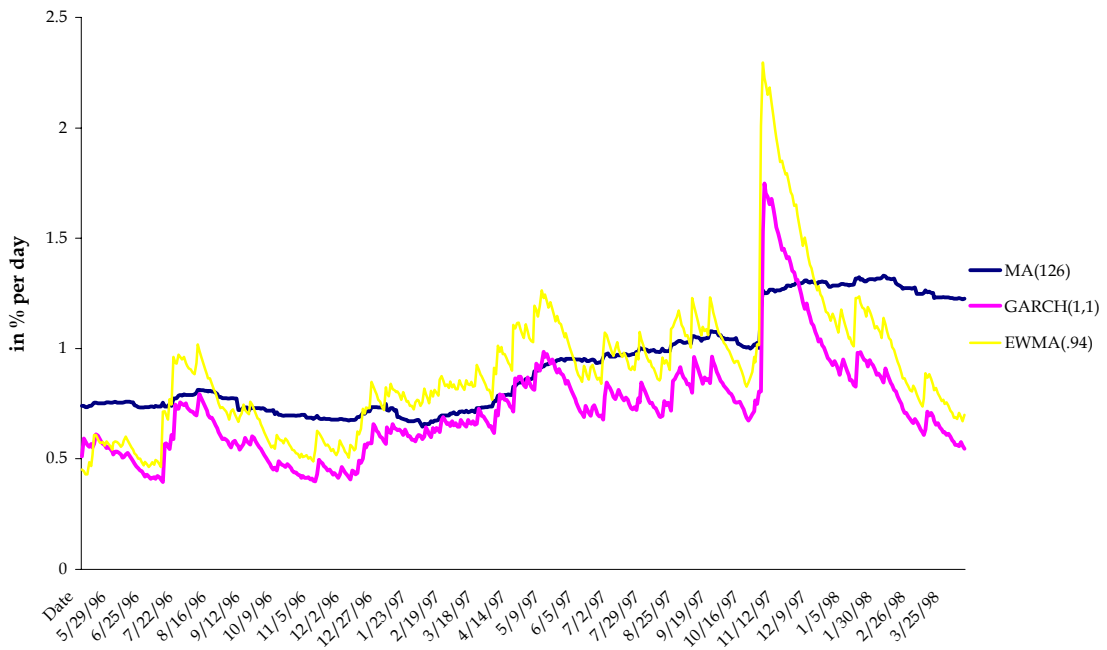
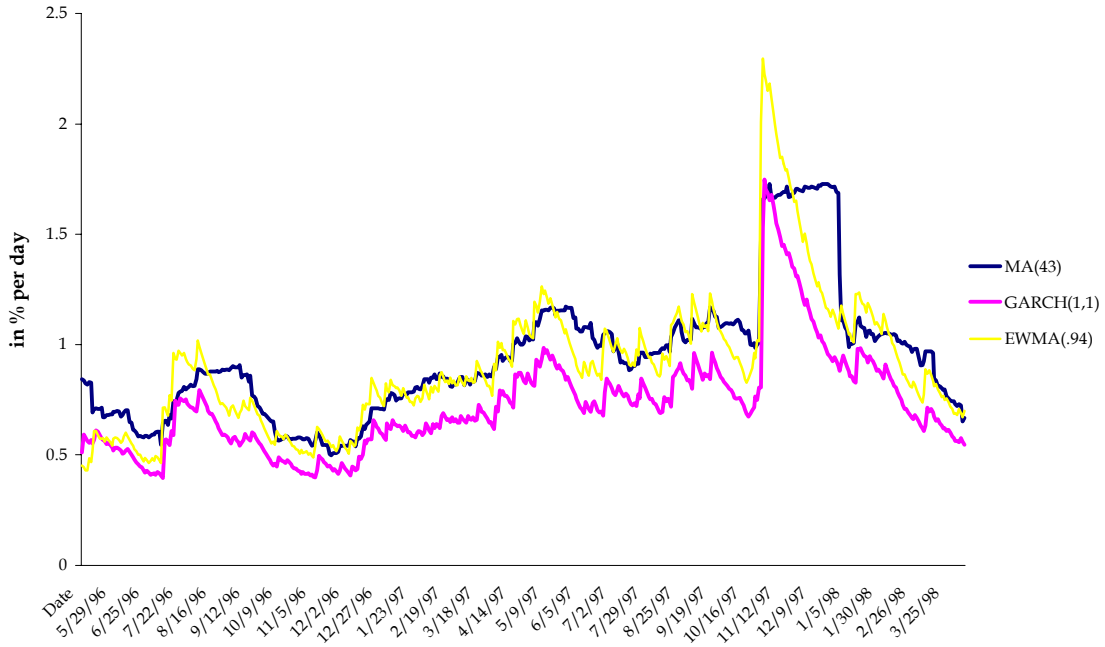


Figure 7

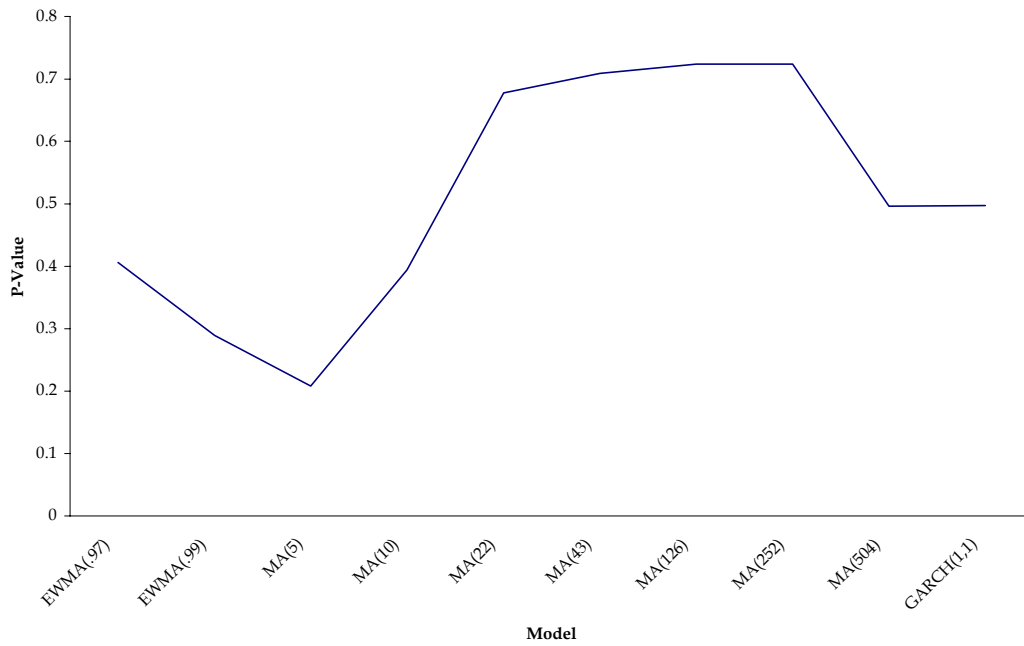
MA, EWMA and GARCH Forecasts



The best model according to our utility-based performance measure presents a nominal or naive P -value of 0.013. This P -value corresponds to the situation where the Bootstrap Reality Check is applied to the best model only. For the White's Reality Check, the P -value is 0.8409 and, therefore, the best model according to our performance measure does not outperform the benchmark. If we had based our inference on the naive P -value, we would have accepted the hypothesis that it has a better predictive ability than the benchmark. The evolution of the P -value as we add new models is presented in figure 8.

Figure 8

Evolution of Reality Check P-Values



We use the last 504 observations of the in-sample period to find the coverage factor γ . Table 1 presents the values for each different volatility estimation method.

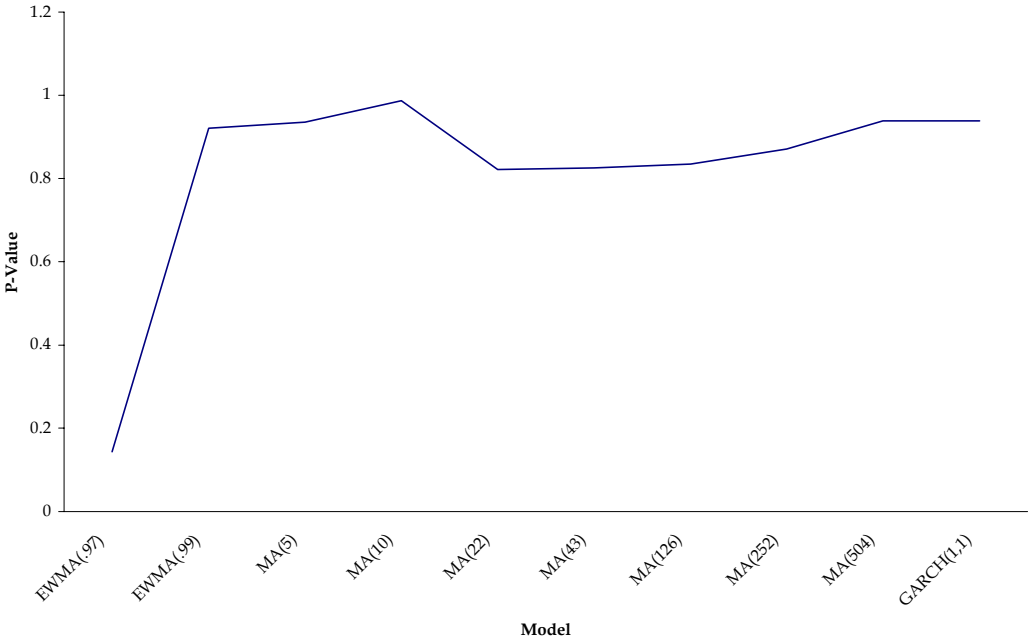
Table 1

Model	γ
EWMA(0.94)	1.599
EWMA(0.97)	1.643
EWMA(0.99)	1.588
MA(5)	1.648
MA(10)	1.578
MA(22)	1.725
MA(43)	1.693
MA(126)	1.584

MA(256)	1.573
MA(504)	1.6
GARCH(1,1)	1.403

The White's Bootstrap Reality Check P -value is 0.939 and the best model according to the VaR coverage criterion has no predictive superiority over the benchmark. The naive P -value of the best model is 0.08. Again, if we had considered the “best” model by itself the result would be misleading. The evolution of the P -value as we add new models is presented in figure 9.

Evolution of Reality Check P-Values



5. FINAL CONSIDERATIONS

Asset allocation decisions and Value-at-Risk calculations rely strongly on volatility estimates. Volatility measures such as rolling window, EWMA and GARCH are commonly used in practice. We have used the White’s Bootstrap Reality Check to verify if one model out-performs the benchmark over a class of re-sampled time-series

data. The test was based on re-sampling the data using stationary bootstrapping. We compared constant volatility, EWMA and GARCH models using a quadratic utility function and a risk management measurement as comparison criteria. No model consistently out-performs the benchmark. This helps to explain the observation that practitioners seem to prefer simple models like constant volatility rather more complex models such as GARCH.

REFERENCES

Alexander, C. (1996). **Handbook of Risk Management and Analysis**. New York, John Wiley and Sons.

Basel Committee on Banking Supervision (1995). **An Internal Model-Based Approach to Market Risk Capital Requirements**. Basle, Switzerland: Basle Committee on Banking Supervision.

Basel Committee on Banking Supervision (1996). **Amendment to the Capital Accord to Incorporate Market Risk**. Basle, Switzerland: Basle Committee on Banking Supervision.

Bertail, P., D. N. Politis, and H. White (undated). "A Subsampling Approach to Estimating the Distribution of Diverging Statistics with Applications to Assessing Financial Market Risks". *Manuscript*.

Danielson, J., P. Hartman and C. G. de Vries (1998). "The Cost of Conservatism: Extreme Return, Value at Risk and the Basle Multiplication Factor". *Manuscript*.

Davidson, A. C. and D. V. Hinkley (1997). **Bootstrap Methods and Their Applications**. Cambridge, Cambridge University Press.

Duffie, D. and J. Pan (1997). "An Overview of Value at Risk". *The Journal of Derivatives*, Spring, p. 7-49.

Engle, R. F (1995). **ARCH: Selected Readings**. Oxford, Oxford University Press.

Engle, R. F and J. Mezrich (1997). "Fast Acting Asset Allocation and Tracking with Time Varying Covariances". *Manuscript*.

Galambos, J. (1978). **The Asymptotic Theory of Extreme Order Statistics**. New York, John Wiley and Sons.

- Gonçalves, S. (2000). "The Bootstrap for Heterogeneous Processes". Doctoral Dissertation, Department of Economics, University of California, San Diego.
- Gumbel, E. J. (1958). **Statistic of Extremes**. New York, Columbia University Press.
- Hall, P. (1992). **The Bootstrap and Edgeworth Expansion**. New York, Springer Verlag.
- Hansen, B. E. (1991). "GARCH(1,1) Processes are Near Epoch Dependent". *Economic Letters*, 36, p. 181-186.
- J. P. Morgan (1995). **RiskMetrics – Technical Document, Third edition**. New York, J. P. Morgan.
- Kuensch, H. R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations". *Annals of Statistics*, 17, p. 1217-1241.
- Longin, F. (1997). "From Value at Risk to Stress Testing: The Extreme Value Approach". CERESSEC Working Paper 97-004, February. *Manuscript*.
- _____ (1997) "The Asymptotic Distribution of Extreme Market Returns". *Journal of Business*, 63, p. 383-408.
- Markowitz H. (1952) "Portfolio Selection". *Journal of Finance*, 7, p. 77-91.
- Politis, D. R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations". *Annals of Statistics*, 17, p. 1217-1241.
- Risk Magazine (1996). **Value at Risk, Special Supplement of Risk Magazine**. London, UK, Risk Publications.
- Sullivan, R., A. Timmermann and H. White (1998). "Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns". UCSD Department of Economics Discussion Paper 98-31.
- _____ (1999). "Data Snooping, Technical Trading Rule Performance and the Bootstrap". *Journal of Finance*, 54, 1647-1692.
- West, K. D., H. J. Edison and D. Cho(1993). "A Utility-Based Comparison of Some Models of Exchange Rate Volatility". *Journal of International Economics*, 35, p. 23-45.
- White, H. (2000). "A Reality Check for Data Snooping". *Econometrica*, 68, p. 1097-1126. *Manuscript* version, 1997.

_____ (1997b). "Various Methods for Measuring Value at Risk". *Manuscript*.

Appendix A Specification Search Using the Bootstrap Reality Check

1. Compute parameters estimates and performance values for the benchmark model, for example, in the case of the utility-based performance measure, $\hat{h}_{0,t+1} \equiv u_0(y_{t+1}|\mathfrak{I}_t)$. Then, calculate the parameters estimates and performance values for the first model, $\hat{h}_{1,t+1} \equiv u_1(y_{t+1}|\mathfrak{I}_t)$. From this calculate $\hat{f}_{1,t+1} \equiv \hat{h}_{1,t+1} - \hat{h}_{0,t+1}$ and $\bar{f}_{1,t+1} \equiv P^{-1} \sum_{t=R}^T \hat{f}_{1,t+1}$. Using the stationary bootstrap, compute $\bar{f}_{1,i}^* \equiv P^{-1} \sum_{t=R}^T \hat{f}_{1,i}^*, i = 1, \dots, B$. Set $\bar{V}_1 \equiv P^{-1} \bar{f}_1$ and $\bar{V}_{1,i}^* \equiv P^{-1}(\bar{f}_{1,i}^* - \bar{f}_1), i = 1, \dots, B$. Compare the sample value of \bar{V}_1 to the percentiles of $\bar{V}_{1,i}^*$.
2. Compute, for the second model, $\hat{h}_{2,t+1} \equiv u_2(y_{t+1}|\mathfrak{I}_t)$. From this form $\hat{f}_{2,t+1} \equiv \hat{h}_{2,t+1} - \hat{h}_{0,t+1}$ and $\bar{f}_{2,t+1} \equiv P^{-1} \sum_{t=R}^T \hat{f}_{2,t+1}$. and $\bar{f}_{2,i}^* \equiv P^{-1} \sum_{t=R}^T \hat{f}_{2,i}^*, i = 1, \dots, B$. Set $\bar{V}_2 \equiv \max\{P^{-1} \bar{f}_2, \bar{V}_1\}$ and $\bar{V}_{2,i}^* \equiv \max\{P^{-1}(\bar{f}_{2,i}^* - \bar{f}_1), \bar{V}_{1,i}^*\}, i = 1, \dots, B$. Now, compare the sample value of \bar{V}_2 to the percentiles of $\bar{V}_{2,i}^*$ to test if the best of the two models outperforms the benchmark.
3. Repeat the procedure in 2 for $k=3, \dots, l$, to test if the best of the model analysed so far beats the benchmark by comparing the sample value of $\bar{V}_k \equiv \max\{P^{-1} \bar{f}_k, \bar{V}_{k-1}\}$ to $\bar{V}_{k,i}^* \equiv \max\{P^{-1}(\bar{f}_{k,i}^* - \bar{f}_1), \bar{V}_{k-1,i}^*\}$.
4. Sort the values of of $\bar{V}_{l,i}^*$ and denote them as $\bar{V}_{l,(1)}^*, \bar{V}_{l,(2)}^*, \dots, \bar{V}_{l,(B)}^*$. Find M such that $\bar{V}_{l,(M)}^* \leq \bar{V}_l < \bar{V}_{l,(B)}^*$. The Bootstrap Reality P -Value is $P_{RC} = 1-M/B$.

Appendix B Notation and Terms

In the description of the Reality Check test, P represents the prediction window and, in the description of Value-at-Risk, ΔP represents the adverse price move.

In the utility function, λ represents the coefficient of risk tolerance, whereas, in the EWMA model, λ is the smoothing constant.

In the GARCH model, ω corresponds to the constant in the variance equation, whereas, in the portfolio allocation problem, ω represents the optimal weight of risky asset.