Portfolio crash testing: making sense of extreme event exposures

Arcady Novosyolov

Siberian Federal University, Academgorodok 19-149, Krasnoyarsk, 660036, Russia;
email: arcady@novosyolov.ru

Daniel Satchkov

Rixtrema, Inc, 14–17 150th St, 2nd floor, Whitestone, NY 11358, USA;
email: dsatchkov@rixtrema.com

The topic of extreme events is becoming ever more important for risk management. Stress testing is a technique that is explicitly designed to deal with extreme shocks; however, its methodology and place in the risk process is often unclear to risk managers. This paper addresses some common misconceptions about stress testing and provides methodology for its incorporation into the risk process as a supplement to risk measures such as value-at-risk and tracking error. Two stress testing models are presented and empirically validated on actual extreme periods. Both are based on multivariate normal distributions conditional on a factor shock, differing only in the way that covariances are estimated. The first model uses temporal weighting commonly used in the risk model construction; the other uses event weighting, which assigns a higher weight to extreme events that are similar to the factor shock specified. The key conclusion is that the time weighted model performs better in moderate or semi-expected shocks, while the event weighted model performs better in more extreme and unexpected shocks like the Long-Term Capital Management crisis, 9/11 terrorist attacks, and the Fall 2008 financials-led meltdown. The event weighted model, which is designed to reflect the rise in correlations and variances during extreme markets, produces a more conservative estimate of return impacts. Our results support the conclusion that stress testing can be a very valuable addition to standard risk measures.

1 INTRODUCTION

Stress testing is a technique that is explicitly designed to validate risk models under extreme shocks; however, its motivation and methodology are not always clear. Before getting to the methodological issues and testing, we must understand what types of extreme events we need to consider in financial risk modeling, because this will determine the design of the tools needed to deal with the issue. Large losses have been occurring more frequently over the past two decades or so, with their intensity seemingly on the rise. The stock market crash of 1987, the Long-Term Capital Management (LTCM) debacle, the internet bubble, the terrorist attacks of 9/11, and the ongoing credit crunch all create impacts that were once seen as extremely rare, almost once-in-a-lifetime occurrences.

1.1 VaR and fat tails

In our view, the quantitative risk modeling process has to include two crucial components. The first is value-at-risk (VaR) analysis, which can be thought of as unconditional risk measure, because it is based on the model of the distribution of financial market fluctuations without postulating any particular market scenario, simply the best forecast based on today's data. The second is extreme impact risk measurement, which can be thought of as a quantification of exposures to particular extreme impacts, ie, a model of a portfolio response conditional on an occurrence of some financial market scenario. This is what is commonly known as stress testing, the topic of our article. Before we move on to our main theme, it is worth briefly discussing how VaR analysis is impacted by extreme events. The major problem with present VaR practices is that a normal distribution is assumed even when modeling short-term returns, where it is clearly inappropriate. For example, kurtosis of daily S&P 500 returns over the past 20 years is 4.6; over the past three years it is 7.8! That is, the data obviously exhibit significant fat tails not captured by the bell curve. It is virtually certain that any short-term VaR estimates based on the bell curve made over 2007 and 2008 failed miserably. We believe that the events of 2007 and 2008 will be a final wake up call to practitioners that the bell curve cannot be used and modelers will be forced to modify their assumptions to something that has a better chance of squaring with market reality. However, it should be made clear that this is not a problem of the VaR measure itself as is sometimes implied, see Taleb (2007), but a problem of unfit and poorly tested distributional assumptions underlying the VaR calculation. The problem should be dealt with by specifying a distribution that fits shorter-frequency data and not by getting rid of VaR, which is simply a way of characterizing future return distributions. There are few candidate distributions that better describe the "fat tail" behavior of short-term financial series, and we believe that this matter will get much more attention going forward. The topic is treated in detail in Rachev *et al* (2005).

1.2 Foundations of stress testing

An empirically tested mathematical apparatus for performing stress tests was laid down by Kupiec (1998). Yet the thinking of practitioners on stress testing remains ambiguous and, beyond running actual historical events, neither the justification nor the design and validation of such models seem to have developed much. This leads to stress testing being frequently dismissed as an arbitrary exercise that cannot provide actionable information. The objections usually center on the fact that there are an infinite number of possible mutually exclusive events and, therefore, each of them has a small enough probability to be considered irrelevant for practical purposes.

In other words, if we have no chance of predicting these events, then we should stop pretending that we can design models to prepare for them. This line of thinking is based on the fundamental misunderstanding of the practice of extreme impact risk management, namely that it is not concerned with reducing uncertainty by predicting the future events, but rather with the understanding of the sources of exposure – in our case exposure to market fluctuations. It can, of course, help reduce uncertainty, but that would be done by hedging, not by predicting particular future outcomes.

2 STRESS TESTING IS LIKE CAR CRASH TESTING

What, then, can we do about extreme events? We believe that the most useful analogy of our strategy in this regard is that of car crash testing. This analogy should clarify the goals and help deal with objections to the practice of stress testing. When a car is crash tested, the designers are not concerned with the precise scenario that may lead to a given accident (ie, they do not specify what type of animal running across the street may have caused the car to suddenly brake, nor do they differentiate between a pole and a tree or between a black wall and a beige one). Instead, they focus on a set of impacts such as frontal, side, and rollover. Every car must endure them to display its strengths and weaknesses. Similarly, when we are stress testing a portfolio model, we are never concerned with the infinity of possible causes, but rather with a limited number of impacts that are observable in the financial markets. Going forward we will use the names "impact" and "event" interchangeably with the understanding that we are only talking about financial impacts, not the infinity of possible causes of these impacts. These impacts could be separated into some of the following categories:

- market-wide impacts, eg, S&P 500;
- sector impacts, especially sectors that don't comprise a large part of portfolio holdings and therefore impact the portfolio in an uncertain manner;
- commodity and economic variable impacts.

This can bring the total number of factor stress tests to anywhere from 10 to 25. A well structured process will include tests from each of the above categories and will allow for decomposition of the portfolio's returns produced by these impacts. This decomposition must apportion the conditional return across groups and securities within the portfolio to facilitate understanding of its strengths and vulnerabilities. While the focus of our article is on validating the models, such decomposition is straightforward within the mathematical framework that we discuss in this article.

The analogy of car crash testing will at once help to deal with all of the "impossibility of prediction" types of objections. It can also provide ammunition to respond to a frequently lodged objection that "a stress test can always be designed to show a major loss, so it is not worth acting on it". This is no less true of a car, but who would want to drive a car that was not crash tested? Therefore, it is a question of a reasonable model design and validation, not of the flaw in the concept itself.

Before delving into the mechanics of implementation and empirical testing, it should be pointed out that it is possible in many cases to refine the process and discriminate between the impacts based on their plausibility. For a recent example, consider manifestations of the current credit crunch. Going back to the first six months of 2007 leading up to the major impact of July–August of that year, we would argue that a sound risk process had to have had a significant decline in Financials as one of the more plausible shocks throughout that period. This is not nearly the same thing as saying that timing of these events could have been predicted, only that some unspecified troubles were on the horizon for the Financials starting at the beginning of 2007 (and even possibly before then). After the events of July–August 2007, even those asleep had to wake up to the fact that financial companies were in danger and certainly could have implemented those tests in the fall of 2007. In October 2008, after a few more shocks to the system, stress tests around Financials stocks declines remain as relevant as ever. Therefore, tests are never run once or twice with some specific timing in mind, but should be updated periodically and run for significant periods of time with some getting more attention if conditions warrant.

3 STRESS TESTING MODELS

We are going to present and compare two models for calculating portfolio losses conditional on shock impacts. We will then compare the performance of the models out of a sample based on six actual extreme events observed between 1998 and 2008. There have been a few studies of stress testing technique such as Kupiec (1998), Kim and Finger (2000), Alexander and Sheedy (2007) and Wang *et al* (2003). The difference in our work is threefold. First, we consider longer horizon impacts: one month versus one day in past studies. Second, we test the models against actual global portfolios and not statistically constructed ones. Last, we propose a new method of estimating extreme event covariances, an extension of the commonly used exponentially-weighted moving average covariance estimators.

3.1 “Naive” model

In past papers on this topic, such as Kupiec (1998) and Kim and Finger (2000), methods tested also included the so-called “Naive” model, which consists of shocking one factor while leaving others untouched. Kupiec (1998) conclusively shows that this model grossly underestimates the loss potential and it is not difficult to see why. It completely disregards any relationships between the factor being shocked and the rest of the factors in the model. In other words, it implicitly assumes zero correlation between factors. It is safe to say that no matter what is observed in the markets during periods of turbulence, zero correlation between key factors is not. For this reason, we have not included tests of this method and focused on the methods which give us a chance to capture dynamics of co-movements of market variables.

3.2 "Time weighted" and "event weighted" models

We will call the model introduced by Kupiec (1998) "time weighted" (TW) because it uses correlation and volatility estimates obtained by using time decay, ie, it assumes that covariance structure during the extreme event is not different from such structure in today's environment. We will introduce our modification of the TW model, which we call the "event weighted" (EW) model, based on the fact that it modifies the weights assigned to each observation for calculation of covariance estimates to better mimic the extreme conditions being tested.

Both methods are based on conditional multivariate normal distribution:[1] let $x = \binom{x_1}{x_2}$ be a normal random vector of length with zero mean and covariance matrix:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

Then (Feller (1970)) the conditional distribution of x_1 given $x_2 = a$ is normal with mean:

$$C_{12} C_{22}^{-1} a \tag{1}$$

and variance:

$$C_{11} - C_{12} C_{22}^{-1} C_{21} \tag{2}$$

The random vector X_1 can be associated with factors within the model. The random vector X_2 can be associated with the factor(s) that we want to shock. Vector a is the size of the shock applied to X_2.

Let uss consider a case where we want to stress a single factor.[2] Matrix C_{11} can be thought of as the covariance matrix of the factors that exist in the factor risk model used by the investment firm for VaR or tracking error estimation, ie, it includes only the permanent model factors and does not necessarily include the factor that we want to stress.[3] Its dimension is equal to n, the number of factors in the model. The element of matrix C located in row j and column k is usually calculated by using the exponential weighting of observations, such as:

$$\sigma_{jk} = (1-\lambda) \sum_{i=1}^{T} \left[\frac{\lambda^{i-1} x_{ji} x_{ki}}{\sum_{s=1}^{T} (1-\lambda)\lambda^{s-1}} \right] = \frac{1}{\sum_{s=1}^{T} \lambda\lambda^{s-1}} \sum_{i=1}^{T} \lambda^{i-1} x_{ji} x_{ki} \tag{3}$$

[1]There are two reasons why normality is acceptable here, despite our earlier discussion of VaR models. First, we are dealing with monthly impacts, where returns resemble the bell curve much more than on a daily basis. Second, assuming conditional normal distribution is quite different from the unconditional one and, in the case of mixture of normal distributions, can produce fat-tail effects.

[2]In principle, this method allows us to stress multiple factors as well since X_2 and a can be vectors.

[3]In commercially available risk models, the decision of factor specification is made outside the investment firm.

TABLE 1 Example of temporal weighting (TW model) as of 7/31/1998.

Date	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Oil	Weight
7/31/1998	1.18	−1.96	−0.38	2.89	1.82	−2.00	0.06
6/30/1998	3.18	2.16	−0.59	−5.63	−1.49	−4.61	0.06
5/29/1998	−1.75	1.26	2.14	−1.39	−3.77	−3.76	0.05
4/30/1998	−1.75	1.26	2.14	−1.39	−3.77	−3.76	0.05
3/31/1998	6.62	5.12	7.80	6.56	−1.59	8.55	0.05
2/27/1998	4.28	1.27	2.61	−10.05	3.54	−1.26	0.04
1/30/1998	6.43	2.48	1.55	2.75	−3.26	−23.05	0.04
12/31/1997	1.38	4.06	2.56	3.25	2.79	2.51	0.04
11/28/1997	−6.35	−3.90	−7.88	−10.84	−11.47	−7.38	0.04
10/31/1997	7.99	3.14	2.45	6.36	3.20	−8.95	0.03
9/30/1997	5.92	6.05	6.79	6.56	−1.96	16.06	0.03
8/29/1997	9.48	7.07	6.64	5.88	0.58	3.68	0.03

where λ is the exponential decay factor, and x_{ji} denotes the return of factor $j = 1, \dots, n$ at time $i = 1, \dots, T$.

The exponential decay factor λ assigns higher values to those observations that are deemed to have the most valuable information content. This is where the crux of the difference between the TW and EW models lies. Under the TW method, a typical risk model estimation method of assigning higher weights to the more recent observations is used. This is the same method that Kupiec (1998) implicitly uses in his study. It basically assumes that the more recent observations carry the most valuable information for the purpose of analysis. While this assumption can be valid when the objective is to predict a risk profile of a portfolio in the present environment, as we mentioned above, there is considerable evidence that today's correlations are not going to remain relevant when a major impact occurs;[4] see, for example, Folmpers and de Rijke (2008) or Campbell *et al* (2002). With the EW method we propose sticking to the same basic idea of assigning weights based on the value of the information, but tailor it to the particular task at hand, namely to estimating how our portfolio will behave in a particular extreme environment. The way we will achieve this is by simply changing the importance order of the observations. In the temporal weighting scheme employed in the TW model, the observations are ordered by time and progressively lower weights are applied as we get further and further from today. Table 1 shows an example of temporal weighting applied as of 7/31/1998.

On the other hand, in EW stress tests we will order the observations based on their similarity to the event we are trying to model. The logic here is quite straightforward. If we are trying to model a 30% monthly decline in NASDAQ, past events when it fell by 31%, 30%, or 29% contain more valuable information than the event when

[4]Unless, of course, we are in the midst of extreme impacts which make today's covariance estimates the same as extreme period covariances.

TABLE 2 Example of event weighting (EW model).

Date	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Oil	Weight
9/28/1990	−14.79	−14.02	−11.99	−10.35	−10.70	36.33	0.060
3/31/1999	1.05	5.89	2.16	6.78	3.74	35.62	0.056
4/30/2001	5.49	7.67	5.80	9.95	4.14	28.38	0.053
4/28/2000	−0.17	4.77	2.24	−4.87	4.61	25.41	0.050
11/30/1999	6.39	5.88	4.02	8.45	3.68	23.54	0.047
8/31/2000	1.79	3.88	0.82	8.52	3.85	18.45	0.044
5/31/2000	−0.56	−6.14	−2.57	−5.46	−1.74	17.91	0.041
9/30/2004	2.52	2.89	0.83	1.63	1.24	17.85	0.039
7/31/1990	0.71	−0.89	0.13	−2.76	−2.94	17.40	0.037
6/30/1999	3.49	9.29	4.21	5.25	4.68	16.92	0.034
9/30/1997	5.92	2.78	4.50	4.99	5.46	16.06	0.032
12/31/2002	−7.56	−5.68	−5.32	−9.03	−4.87	16.03	0.030
10/30/1998	−9.67	−0.52	−3.95	8.24	−0.52	15.85	0.029
9/30/1998	−9.67	−0.52	−3.95	8.24	−0.52	15.85	0.027
10/31/2007	3.34	3.23	2.60	3.11	3.93	15.76	0.025

it fell three or rose five percent, regardless of when they occurred. Table 2 shows an example of such dynamic weighting for a 40% upward shock in oil prices. Higher weights are applied to those events that most resemble the proposed shock. Instead of time, the similarity of impact becomes the axis of importance.

It is not difficult to see that we are making an assumption that covariances in similar extreme periods are similar to each other, but are different from normal periods. Heuristically, the EW model can be thought of as being based on an assumption of a mixture of two normal distributions as the underlying model.[5] This is in principle similar to the Kim and Finger (2000) approach, but our method does not require an explicit estimation of the underlying mixture of normal distributions, which makes the method easier to deploy in the real-time risk measurement process.

4 TESTING METHODOLOGY

We believe that one of the biggest problems of risk management today is a lack of meaningful testing and validating of the predictive models. This absence of testing makes some models too theoretical and leads to complexity for complexity's sake. On the other hand, it allows others to make sensational claims dismissing all of risk management as fraud. Therefore, procedures for validating and accountability in the

[5] A mixture of normal distributions assumes that there are two distinct environments, both bell curved, but with drastically different variances and possibly covariances. One can be thought of as the normal environment, which manifests itself most of the time, the other as the extreme environment with much higher volatility. In fact, combining the two can produce the fat tail behavior observed in the markets, but we are only concerned with the idea that there are some distinct extreme covariances, which we attempt to estimate with the EW method.

risk estimates should be one of the priorities of the industry. Testing predictions is difficult work, but if the model is not testable it becomes useless, since it cannot support decision-making or, worse, may lead to erroneous risk taking. As discussed above, the issue of prediction for financial risk managers performing stress testing is simplified by the fact that they do not have to predict actual events, but rather responses of the financial system to broad classes of events, which produce similar consequences.

Kupiec (1995) deals with this issue and backtesting of VaR is outlined by the Basel Committee on Banking Supervisions guidelines. Binomial testing of VaR is a fairly established and simple technique, but what about empirical verification of stress testing models? First, we must remember that the testing of any prediction is a comparison of a prediction made before the fact with the actual outcome observed subsequent to established prediction. The actual in this case is quite simple: it is a return of some portfolio over the period containing the actual extreme impact that we use for verification. The predicted is a result of a stress test run before that extreme impact was observed, ie, without looking ahead. Kupiec (1998)'s backtesting of results was limited to factors within the risk model used for estimation of VaR. Since he was dealing only with such pre-defined factors, he was able to go through every outcome for every factor within the model and used all observations that fell outside the 3.5 unconditional standard deviation band.

Because our approach is explicitly designed to go beyond the pre-defined factors in a given covariance matrix and allows flexibility in shock definition, we potentially have an infinite number of ways to specify the impacts. To narrow the number of tests, we must again use the fact that various extreme shocks manifest themselves through a limited number of observed indices and metrics, which we call impacts. For example, market events dubbed the credit crunch started primarily with shocks to the Financials sector and then spread to other areas. S&P 500 Diversified Financials declined 14.4% between July 15 and August 15 of 2007. If we define such a shock to the Financials sector on July 14 and use the factor model data available prior to that date, we can use Equation (1) to make a prediction of a monthly return for a portfolio. We can then compare it with the way that the portfolio actually performed over that subsequent month.

To be sure, nothing in this testing implies that the cause could have been known or that the actual event could have been predicted with respect to timing, only that the possibility for a decline in Financials at some point in 2007 was plausible enough to make it onto the stress testing agenda. This is certainly not a heroic assumption.

Repeating this procedure over many portfolios and over available history of actual extreme impacts would give us a good idea of the predictive power of the model. This method of validation obviously assumes that the shock was correctly anticipated in order to test the validity of the stress testing engine predictions and not the design skills of the risk manager. Designing stress tests is an art that deserves separate detailed treatment, but we would argue that it is not as complex as it might seem. As we will show, applying plausible shocks to broad indices like the S&P 500 or S&P

Diversified Financials would suffice to produce predictions that would have given valuable information about loss potential under difficult and uncertain conditions.

We test our methodologies on six historical shocks:

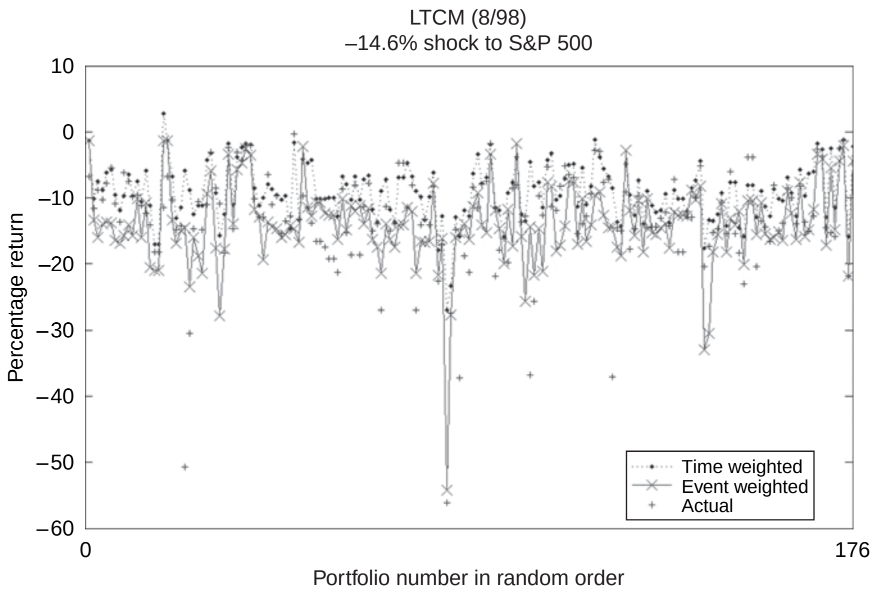
1) LTCM crisis: the application of −14.6% shock to S&P 500 on 7/31/1998 and comparison of the predictions with returns realized over August 1998.
2) The end of the internet bubble: application of −22.9% shock to NASDAQ on 10/31/2000 and comparison with the returns realized over November 2000.
3) September 2001: application of −8.2% shock to S&P 500 on 8/31/2001 and comparison with the returns realized over September 2001. There is no implication of any anticipation of a precise event, only testing of a market-wide impact sensitivity of a portfolio.
4) The first major impact of the credit crunch: application of −14.4% shock to S&P 500 Diversified Financials on 7/14/2007 and comparison with the returns realized 7/15/2007–8/15/2007.
5) The continuing manifestations of the credit crunch: application of −19.5% shock to S&P 500 Diversified Financials index on 5/31/2008 and comparison with the actual returns realized over June 2008.
6) The credit crunch continued: application of −17.3% shock to the S&P Diversified Financials index on 8/16/2008 and comparison of predictions with returns realized between 8/17/2008 and 9/17/2008.

We performed this analysis on 176 global portfolios randomly chosen from the actual institutional portfolios. The holdings came from publicly available filings of institutional fund positions.

5 EMPIRICAL RESULTS

The most obvious pattern that emerges from the charts displayed in Figures 1–6 (see pages 62–66) is that the EW model produces more conservative estimates of conditional average portfolio losses. Since the EW model assigns higher weights to similar extreme events in the past, it makes sense that its correlation and variance estimates are mostly higher than those produced by the TW model, which uses temporal weighting. Therefore, more extreme losses predicted by the EW model are in line with what we would expect given its assumptions. To start, consider Figure 1, which contains the result of testing our methodologies during the period known as the LTCM crisis. This is a prime example of a manifestation of what is referred to as "increased correlations", which is math-speak for panic in the marketplace. It is apparent from the figure that the EW model does considerably better at estimating conditional losses in that scenario. It produces a line that is generally closer than that produced by the TW model and, even more importantly, it produces far fewer significant underestimations, which constitute one of the key dangers in a stress testing process. Such underestimation is seen a bit more frequently in Figure 3 (September 2001) even for the EW method, which nevertheless remains

FIGURE 1 LTCM 1998 crisis; application of −14.6% shock on S&P 500 on 7/31/1998.



more conservative in that period as well. Apart from those few instances, we again see that the EW model performs much better and rarely undershoots the actual performance by much.

Figure 2 (11/2000) shows quite a different story. In it we see that the EW model is far too conservative for most of the portfolios and that TW provides a much better estimate of portfolio losses. This was the end of the burst of the internet bubble – a unique period that was characterized by both extreme sector impact (NASDAQ shocks) and non-rising (and even reduced) correlations. Therefore, it would make sense in this case that the method which assumes that covariances stay constant would perform better than the one that estimates them from extreme periods. We believe that the current high degree of global interconnectedness and presence of significant leverage makes such decoupling of asset returns in crisis far less probable. Figures 4–6 show three different stretches of the ongoing turmoil (July 15, 2007 to August 15, 2007, June 2008, and August 17, 2008 to September 17, 2008). In the first two, we see that the dynamic estimation of covariances employed by the EW method does not improve on the static covariance assumption, since both methods show very similar results. However, in these events, as elsewhere, we observe that EW results are more conservative in prospective loss estimation. What seems unintuitive is that the extreme impact of 2007 (Figure 4), almost universally characterized as panic,

FIGURE 2 End of internet bubble; application of −22.9% shock to NASDAQ on 10/31/2000.

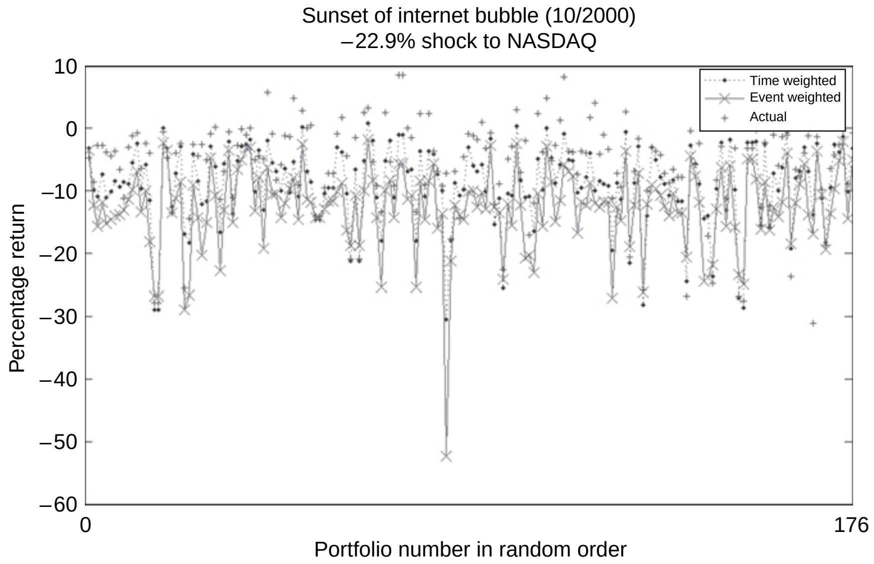


FIGURE 3 September 2001; application of −8.2% shock to S&P 500 on 8/31/2001.

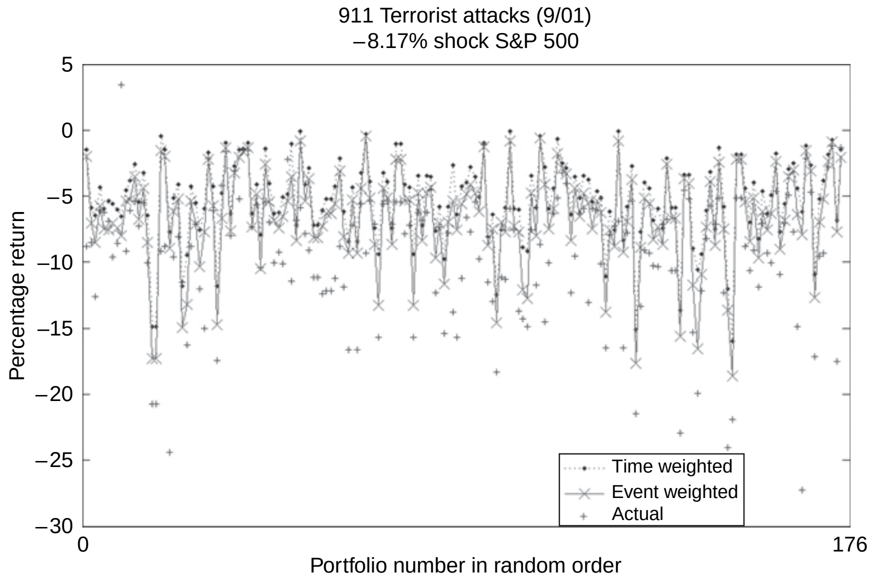
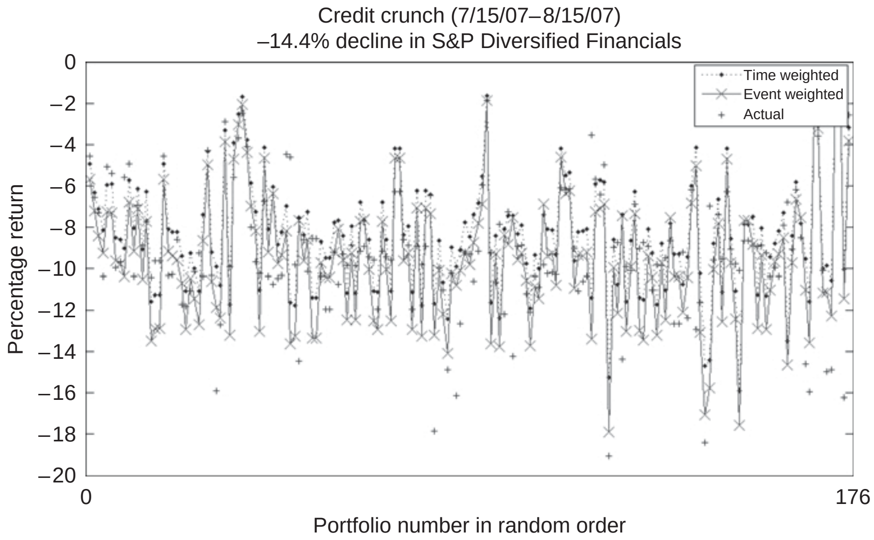


FIGURE 4 Credit crunch; application of −14.4% shock to S&P 500 on 7/14/2007.

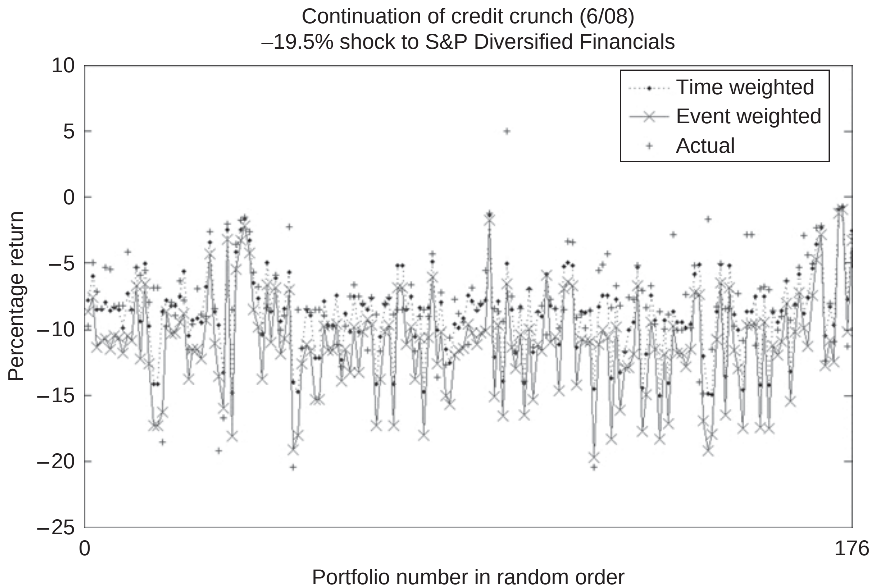


appears to be well captured by the static covariance assumption. This point requires further investigation through both qualitative and quantitative tools. The latter could be done by examining the likelihoods of the actual returns under the then-current TW covariance matrix. Good performance of the TW model in Figure 5 on the other hand seems in line with the experience of June 2008, which seemed more like an orderly retreat, rather than a shock impact. Figure 6 shows the period containing the unprecedented crisis in the US financial sector in the early fall of 2008, which quickly spread to all sectors worldwide. This one did not leave any doubt as to the presence of panic correlations. The graph clearly shows that the EW method was much better at predicting the losses in this panic environment, much like it was during the LTCM period and September 11 terrorist attacks.

Table 3 (see page 66) summarizes much of what we have already observed in the charts. First, consider the root mean square error (RMSE)[6] column. We see that the EW method has a clear edge in 8/1998 (7.47 for TW versus 6.45 for EW), 9/2001 and 9/2008 stress tests, while TW does better in 11/2000 and 6/2008. During 7/2007 we observe roughly similar results. RMSE gives us some idea of average proximity

[6]We are using the abbreviation because the calculation is familiar to the practitioners and we found it useful as a summary statistic. However, the E (error) in RMSE is a bit of a misnomer, since we are using both methods to estimate the conditional averages. Therefore, even with perfect estimation some variation is inevitable, and it is not really an error as much as a metric that gives a general idea of the degree of dispersion of our calculated averages around the actual losses.

FIGURE 5 Continuing manifestation of the credit crunch; application of −19.5% shock to S&P 500 on 5/31/2008.



of estimates, but we must also consider the major failures to underestimate the conditional loss, since such failures represent the greatest danger to the risk process. The three rightmost columns show the proportion of portfolios (out of 176) for which a given method underestimated the actual loss by the percentage indicated in the name of the column.

Focusing on the third column from the right, which shows the proportion of underestimations of the actual by at least two-fold (ie, by 100% or more), we see that the EW method is steadily better under this metric. This confirms the hypothesis that it is a more conservative method that gives the ability to capture potential for rising variances and correlations that occur during the extreme periods.

6 CONCLUSIONS

Stress testing should be thought of as analogous to car crash testing, where instead of chasing after an infinite number of possible events, a risk manager must consider a limited number of impacts. The manager may pay more attention to certain impacts that appear more plausible but should not make specific timing predictions. We tested two models of performing stress tests, time weighted and event weighted, both of which are based on the conditional multivariate normal distribution (Kupiec (1998)). Both methods provide valuable insight, but the EW model performs better

FIGURE 6 The credit crunch continued; application of −17.3% shock to S&P 500 on 8/16/2008.

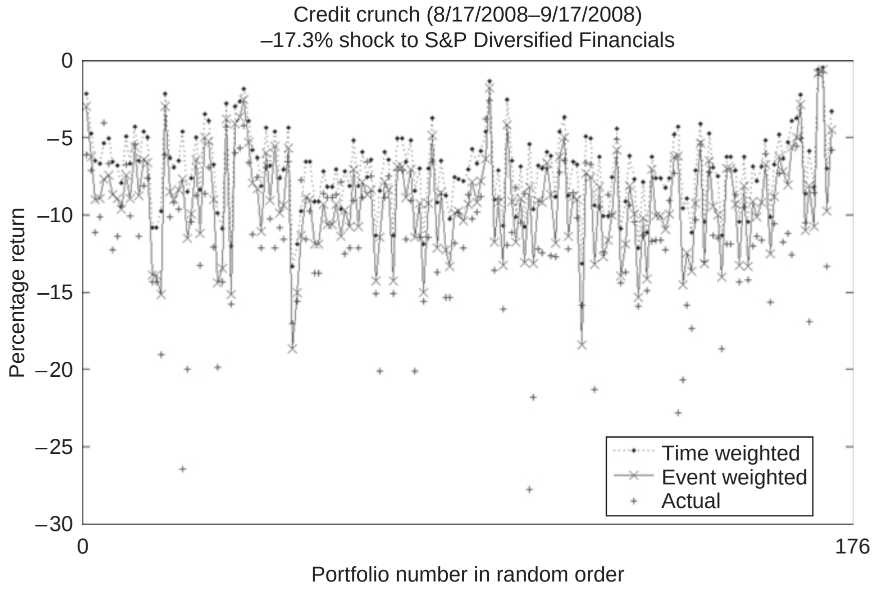


TABLE 3 Summary statistics for empirical testing of TW and EW models.

EVENT	RMSE TW (EW)	Underestimation of actual returns (fraction of portfolios observed) >100% TW (EW)	>50% TW (EW)	>25% TW (EW)
LTCM (S&P 500 −14.6%)	7.47 (6.45)	0.15 (0.06)	0.3 (0.11)	0.53 (0.19)
9/11 (S&P 500 −8.17%)	6.8 (5.76)	0.29 (0.18)	0.65 (0.38)	0.89 (0.58)
Internet bubble (NASDAQ −22.9%)	5.17 (8.95)	0.04 (0)	0.06 (0.01)	0.1 (0.02)
Credit Crunch 07 (S&P Div Financials −15%)	3.14 (3.31)	0.01 (0.01)	0.09 (0.03)	0.3 (0.13)
Credit Crunch 6/08 (S&P Div Financials −19.5%)	3.23 (4.55)	0.01 (0)	0.05 (0.01)	0.11 (0.03)
Credit Crunch 9/08 (S&P Div Financials −17.3%)	5.3 (3.68)	0.13 (.05)	0.53 (0.12)	0.81 (0.32)

under more extreme impacts characterized by panic in the market. The EW model is consistently more conservative and produces higher estimates of portfolio losses. This property is likely due to the design of the EW model, which estimates covariances by overweighting the more extreme observations. The TW model should be used when correlations are not expected to move; otherwise the EW model presents a more powerful alternative.

REFERENCES

Alexander, C., and Sheedy, E. (2007). Model-based stress tests: linking stress tests to var for market risk. ICMA Centre Discussion Papers in Finance, The Business School for Financial Markets, University of Reading.

Campbell, R., Koedijk, K., and Kofman, P. (2002). Increased correlation in bear markets. *Financial Analysts Journal* **58**(1), 87–94.

Feller, W. (1970). *An Introduction to Probability Theory and Its Applications*, Vol. 2. 2nd ed. Wiley, p. 704.

Folmpers, M., and de Rijke, P. (2008). Correlations in a downturn economy: implications for systematic risk and capital adequacy. *GARP Risk Review* (43, August/September), 17–21.

Kim, J., and Finger, C. (2000). A stress test to incorporate correlation breakdown. *The Journal of Risk* **2**(3, Spring), 5–19.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* Winter 73–84.

Kupiec, P. (1998). Stress testing in a value at risk framework. *Journal of Derivatives* Fall 7–24.

Rachev, S., Frank, F. F., and Menn, C. (2005). *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*. Wiley.

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House, New York.

Wang, S., Mun, C., and Dollery, B. (2003). An analysis of stress testing for Asian stock portfolios. Working Paper Series in Economics, University of New England School of Economics.