Assessing the Accuracy of Value at Risk

Joe H. Sullivan, Ph.D.
College of Business
Mississippi State University

Robert Brooks, Ph.D., CFA
Department of Finance
The University of Alabama

Zachary G. Stoumbos, Ph.D.
Department of Management Science and Information Systems and
Rutgers Center for Operations Research (RUTCOR)
Rutgers, The State University of New Jersey

Comments welcomed. Please do not quote without permission.

CONTACT:

Zachary G. Stoumbos
MS/IS Department, Rutgers Business School
Rutgers, The State University of New Jersey
Room 253 Levin Building
94 Rockafeller Road
Piscataway, NJ 08854-8054
732.445.6849
stoumbos@andromeda.rutgers.edu

Assessing the Accuracy of Value at Risk

An important question for corporate finance officers is whether risk assessments, such as Value at Risk (VaR), are *currently* accurate. In contrast to previous research on assessing the accuracy of VaR and related density estimates, which has focused on backtesting a large sample of fixed-size, we provide tools for real-time assessment, using a time window that varies adaptively with the data.

The VaR is determined by a single point of the estimated distribution of the portfolio gain. Previous literature has dichotomically tested the VaR forecast or the entire estimated distribution. A pure test converts each observed gain into a binary value indicating whether it was covered by the VaR forecast or not. A more powerful test results from using the entire distribution, by transforming the observed gain to a random variable that has a known distribution when the forecast is accurate, but this also detects errors unrelated to VaR or other measures of risk.

We propose the idea of a continuous compromise between detection power and purity, where power refers to quick detection of systematic bias and purity refers to insensitivity to errors irrelevant to VaR estimation.

Assessing the Accuracy of Value at Risk

I. Introduction

Banking, insurance, and corporate finance executives are implementing risk management systems to better understand and manage a firm's risks. These risk management systems often forecast the distribution of profit and loss, or "gain," in an asset portfolio and estimate summary risk measures, such as value at risk (VaR), volatility, and skewness. Our focus is on assessing accuracy, rather than making such estimates. In contrast with previous accuracy assessment research, which has focused on backtesting a large, fixed-size sample, often data for several years, our objective is to provide tools that evaluate the current accuracy of these systems. Waiting to accumulate hundreds or thousands of daily observations is not attractive in practical situations. Especially in markets where the risks are large and complex, it is crucial to establish the accuracy of the risk system and verify that there is no systematic bias currently present in the process.

The ability to certify that a risk management system is currently without significant bias is useful for building confidence in the risk-management process. Also, there is a relationship between theoretical market models and valuation models used within these systems and in certifying the systems. It is possible to identify superior market models and superior valuation models through the effort of system certification. Although we focus in this paper on VaR, the methodologies developed are directly applicable to other risk measures. The objective of this paper is to present a new method for monitoring the accuracy of the sequence of VaR values, a general framework for the rapid detection of the presence of a systematic bias.

Portfolio risk is comprehensively captured in the distribution of gain (profit and loss) over a specified time period, typically ending with the close of the next trading day. A common way of summarizing the risk implicit in such a distribution is the VaR, which

simplifies the complexities of making comparisons among distributions and decisions about the implications of the risk that is present. The VaR is determined by a single point of the distribution. For example, if the coverage probability is specified as 0.95 or 95 percent, then the VaR is the fifth percentile, or the 0.05 quantile (a negative value), of the estimated distribution, with the negative sign dropped. With a continuous distribution, there is exactly 5 percent probability that the observed gain will fall below the 0.05 quantile, which means the gain expressed as a loss would exceed the negative of the 0.05 quantile. Sometimes the VaR is expressed as the quantile divided by the initial portfolio value, again with the negative sign dropped. For more information on the basics of VaR, see Beder (1993), Beckstrom (1995), Jorion (1997), Dowd (1998), Best (1999), and Hull (2002).

The pure assessment of VaR accuracy hinges on whether it achieves the specified coverage probability, conditioned on the current information set. Under this philosophy, each observed gain is converted into a binary value, for example a 1 indicating coverage and a 0 indicating a violation. Christoffersen (1998) uses this approach, and points out that the resulting binary observations should be independent as well as have the specified unconditional coverage probability. He gives tests for both of these characteristics. The lack of independence is also called clustering, indicating that the conditional violation probability is elevated after a violation. Christoffersen's test for accurate *conditional* coverage is decomposed into subsidiary tests for *unconditional* coverage and *independence*. This approach has the advantage of being a pure test of the VaR estimates. The disadvantage is that when the specified coverage probability is large, such as the typical values of 95 percent or 99 percent, a large sample size is required to detect a conservative bias (an actual violation probability that is lower than specified). This is to be expected because many observations are required to show conclusively that there are too few instances of an event that is rare. Christoffersen (1998) gives examples involving the backtesting of 2000 daily observations. Berkowitz and O'Brien (2002) apply the testing framework of Christoffersen (1998) to 569 to 746 daily observations from six banks.

A more powerful test, in the sense of requiring fewer observations for the same detection probability, results from using the entire estimated distribution of the portfolio gain. Diebold, Gunther, and Tay (1998) use this approach, employing the probability integral transformation to convert the gains into random variables that are independent with uniform (0, 1) distribution under the null hypothesis that the density estimates are accurate. They opt for graphical tests for uniformity and autocorrelation but note that there are standard, objective tests for these characteristics. The authors also focus on backtesting a large historical sample.

The great disadvantage of the more powerful tests of the entire distribution is that some of the increase in detection power comes from detecting errors that are not relevant to VaR estimation. For example, errors in the upper tail of the distribution would be detected, although the upper tail is not usually associated with any measure of risk, per se.

We propose the idea of tests with intermediate purity and power, rather than the previous all-or-nothing tests. The purity of a test refers to a lack of sensitivity to errors that are not related to VaR estimation, and power refers to the ability to detect errors. A powerful test offers quick detection of errors, while a test with high purity does not react to errors that do not affect VaR. Strictly speaking, since the VaR is determined by a single point of the estimated distribution, a test could be considered as either pure or not. However, we enlarge the concept of purity by defining a neighborhood containing the VaR. We define purity to be inversely related to the size of the neighborhood. The smallest size is zero, corresponding to the single point and the greatest purity, while the largest size includes the entire distribution and represents the least purity. In this way, the neighborhood size is a decision parameter, adjusting the compromise between purity and power. We show that a small relaxation in purity leads to large initial gains in power, suggesting a compromise useful in practical applications.

The remainder of the paper is organized as follows. In Section II, we describe the notation and model for testing a certain number of the most recent observations, the number tested being a function of the observed data. The key idea is to apply a sequence of sequential tests in which a test is conducted after each new observation.

Once a group of observations is judged to be without bias, it is set aside and not included in future tests. Thus, each of the sequential tests involves only the most recent observations, with the number of observations dependent on the recent history.

In Section III, we propose the idea of a continuous adjustment between purity and power by defining a neighborhood of the density forecast that will be used in the test. Finally, in Section IV, we show that the tests are effective when the true distribution is non-Gaussian and give concluding remarks in Section V.

II. Model for Real-Time Detection

Previously, the probability integral transformation has been used by Diebold, Gunther, and Tay (1998) and other authors to transform the observed portfolio gain to a random variable that has a uniform (0, 1) distribution and is independent of all other transformed variables when the conditional distribution estimate is accurate. The transformed variables can then be tested for independence and uniformity, and evidence against either characteristic is evidence of systematic bias in the series of conditional distribution estimates. However, Diebold, Gunther, and Tay (1998) criticize this approach, saying, “Alone, however, such tests are not likely to be of much value in the practical applications that we envision, because they are not constructive….” They adopt “less formal, but more revealing, graphical methods” that allow a subjective attribution of the nature of the bias as well as its historical existence. In proposing a test of the sequence of estimated conditional distributions, we use a related, continuous transformation rather than the uniform transformation. We transform each observed gain directly to a test statistic that indicates the presence or absence of a specific type of bias. We accumulate these individual statistics to determine if there is convincing evidence of such bias. To detect different types of bias, we run two (or possibly more) specialized tests in parallel. In this way, when a bias is detected, the source of the detection indicates the type of bias.

Let W_i represent the portfolio value observed at time i, and $X_i = W_i - W_{i-1}$ represent the gain in portfolio value from i - 1 to i, given Ψ_{i-1}, the data observed through time i - 1. Let the true cumulative distribution function (cdf) of X_i given Ψ_{i-1} be a continuous function G_i, meaning that $G_i(x) = P(X_i \le x \mid \Psi_{i-1})$, and let the estimate of G_i be F_i, also assumed to be continuous. Thus, the *true* lower limit of the one-sided (1 - α) * 100% confidence interval for X_i, conditioned on Ψ_{i-1}, is $G_i^{-1}(\alpha)$, where α is the specified violation probability. The *estimated* lower limit is $F_i^{-1}(\alpha)$, so that the VaR is $-F_i^{-1}(\alpha)$. Page (1954) proposed a powerful class of sequential tests based on the cumulative sum (CUSUM) of the score statistics

$$C_i = \log(g_i(X_i)/f_i(X_i)), \tag{1}$$

which is the natural logarithm of the likelihood ratio statistic for observation X_i. The function $f_i = \frac{d}{dx}F_i(x)$ is the conditional pdf under the null hypothesis that F_i is the true conditional cdf, while $g_i = \frac{d}{dx}G_i(x)$ is the alternative-hypothesis conditional pdf, with both distributions fully specified and free of unknown parameters. Lai (1995) summarized the CUSUM and its properties, noting that the test signals a regime shift at observation $M = \inf\left\{ m : \max_{1 \le k \le n} \left[\sum_{i=k}^{m} C_i \right] \ge c_\gamma \right\}$. In this signal rule, the value of k maximizing the summation that leads to a signal is an estimate for the location of the first shifted observation, although not the likelihood maximizing estimate. The critical value c_γ can be specified indirectly by requiring that the expected value of M, which is the average detection delay or average run length (ARL), be at least γ when there is no regime shift. The CUSUM test is optimal in the sense of requiring the smallest expected number of shifted observations to signal a regime shift. Moustakides (1986) and Ritov (1990) demonstrated this optimality, which can be summarized as follows. Let $\Im_\gamma$ be the class of all tests with ARL $\ge \gamma$, when all observations are from the sequence of null distributions $\{F_i\}$. When there is a regime shift beginning with observation $\nu < \infty$, the idea

is to consider the conditional expected detection delay $\overline{T}_{\nu,X} = E((T-\nu+1)^{+} \mid \Psi_{\eta-1})$ for a test $T \in \Im_{\gamma}$, conditioning on the shift location ν, on no false alarm prior to ν, and on $\Psi_{\eta-1} = \{X_1, \ldots X_{\eta-1}\}$. The CUSUM procedure is optimal in that it minimizes the worst-case conditional expected detection delay, $\overline{E}_1(T) = \sup_{1 \le \nu} \text{ess sup}(\overline{T}_{\nu,X})$. Stoumbos et al. (2000) also discuss the CUSUM, as well as Bayesian procedures that could be used if there is prior knowledge about the shift location.

While the optimality of the CUSUM applies only to detecting a shift to one specific distribution, it is also very effective at detecting shifts to a broad class of other distributions. The class of distributions effectively detected can be defined based on the Kullback information concept. Suppose the CUSUM score statistic C_i is specified by an alternative likelihood (fully specified by the parameter vector θ_1) and a null likelihood (fully specified by the parameter vector θ_0). Then, the Kullback information content of a single observation X_i from a distribution completely specified by the parameter vector θ_2 is $K_i(\theta_2, \theta_1, \theta_0) = E_{\theta_2}(C_i)$. Basseville and Nikiforov (1993) give the asymptotic relationship between the ARL and the information per observation as $\ln(\overline{T}) / K_i(\theta_2, \theta_1, \theta_0)$, where $\overline{T}$ is the average delay between false alarms. The relationship is asymptotic as $\overline{T} \to \infty$. Thus, large information per observation leads to a small ARL, compared with the average delay between false alarms. Thus, a CUSUM would be effective, in the sense of having a small ARL to detection, provided the Kullback information is at least as great as the information for the distribution used in designing the CUSUM procedure. Furthermore, as long as the information is positive, the ARL will be less than the average delay between false alarms.

A special technique is used to expand the class of distributions that the CUSUM procedure effectively detects. A test effective in detecting an elevated violation probability can be run simultaneously with a test effective in detecting a decreased violation probability. A signal from either test indicates evidence of a systematic bias. While recognizing that the CUSUM test can be designed to be effective in detecting

changes in arbitrary distributions, for purposes of exposition, we first focus on the simpler situation of normal distributions with a mean of zero, so that the only issue lies in estimating the volatility. In that situation, there are only two types of bias, one of which being a violation probability that is less than what is specified, which we refer to as a "conservative" bias. The other type of bias, in which the violation probability is elevated, we refer to as a "liberal" bias. The score statistic simplifies to $C_i = \log(\sigma_F / \sigma_G) + 0.5(X_i / \sigma_F)^2\left(1-(\sigma_F / \sigma_G)^2\right)$, where σ_F is the estimated volatility and σ_G is the true value. Consider a CUSUM test for detecting a specific liberal bias, say $\sigma_G / \sigma_F = 1.3$. This test would be very effective, although not quite optimal, in detecting errors that resulted in even larger volatility ratios. The effectiveness would diminish with volatility ratios closer to 1, and the test would be ineffective in detecting any conservative bias (volatility ratio less than 1).

A reasonable companion would be a test optimized for a ratio of 1/1.3, a conservative bias. This test would be very effective in detecting even more severe conservative biases. Combining the two tests gives effectiveness in detecting both types of bias. The performance is measured by the worst-case ARL, which is evaluated by simulation. To illustrate the performance, two one-sided CUSUM tests were constructed. One test detects a liberal bias, optimized for a 1.3 ratio of actual to estimated volatility. The other test detects a conservative bias, optimized for an actual to estimated volatility of 1/1.3. Simulation was used to estimate a critical value of $c_\gamma = 2.6$ used in each one-sided CUSUM, which gave an overall average of 125 observations between false alarms for the combined test. Figure 1 gives a graph of the ARL versus σ_G / σ_F, the ratio of actual to estimated volatility.

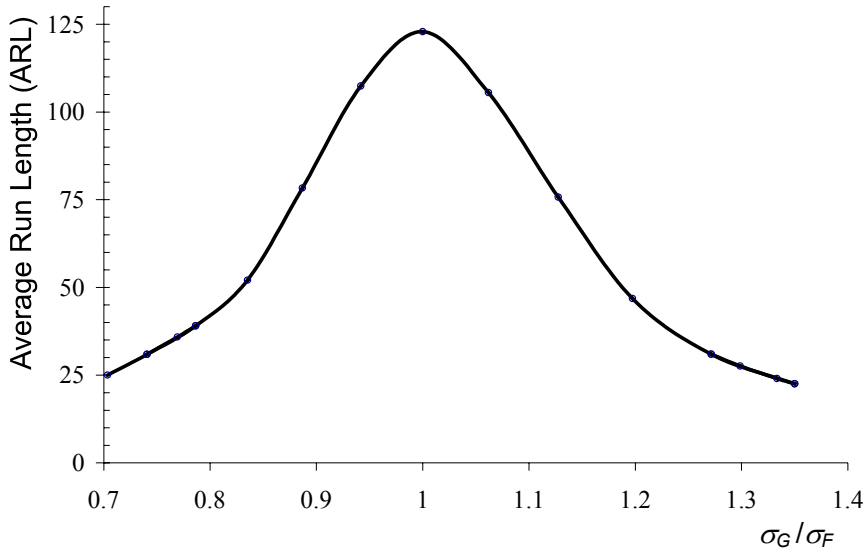

Figure 1: ARL vs. σ_G / σ_F for two one-sided CUSUM tests.

III. Adjusting Detection Purity and Power

Previously proposed tests have been a pure, direct test of VaR violations, such as Christoffersen (1998), and Berkowitz and O'Brien (2002) or tests of the entire forecast distribution, such as Diebold, Gunther, and Tay (1998). A disadvantage of a pure test is that many observations can be required to detect a substantial bias. For example, a series of daily 99 percent VaR estimates would be expected to have about 2.5 violations in the 250 or so trading days in a calendar year. The most extreme possible conservative bias would reduce the violation probability to zero. A sample of several years would be required to have adequate power in detecting even this most extreme possible bias. On the other hand, using the probability integral transformation applied to the observed portfolio gain increases the detection power but at the cost of detecting irrelevant errors that are misleading and can lead to an erroneous loss of confidence in the risk management system.

We propose an intermediate method that uses a part of the forecast distribution within a neighborhood of the point corresponding to the VaR. If $X_i \in [x_{Low}, x_{High}]$, with

$x_{Low} = F_i^{-1}(\max(0, \alpha - \delta))$, $x_{High} = F_i^{-1}(\min(1, \alpha + \delta))$, and $0 \le \delta$, the score statistic is C_i, the continuous transformation of the X_i value in Equation (1). Otherwise, we adjust the score statistic as follows:

$$C_{\delta,i} = \begin{cases} \log(G_i(x_{Low}) / F_i(x_{Low})), & X_i < x_{Low} \\ C_i, & x_{Low} \le X_i \le x_{High} \\ \log((1 - G_i(x_{High}))/(1 - F_i(x_{High}))), & x_{High} < X_i \end{cases} \tag{2}$$

For values of X_i in the neighborhood [x_{Low}, x_{High}], $C_{d,i} = C_i$. For values $X_i < x_{Low}$ the $C_{d,i}$ collapses the distributions for portfolio gain to a single discrete point, concentrated at x_{Low}, and likewise for $X_i > x_{High}$. These two discretizations render the adjusted statistic insensitive to the degree of VaR violation or coverage for gains outside the neighborhood, thus limiting the test to parts of the distribution that are relevant to risk assessment as specified by the decision parameter δ.

The pure VaR test using a binary transformation becomes a special case corresponding to $\delta = 0$. The special case $\delta \ge 1 - \alpha$ gives the continuous transformation. Intermediate values of δ represent a compromise between purity and power. An example of detecting a liberal bias, based on a simple design with zero-mean normal distributions, is shown in Figure 2. The actual volatility is 1 but is estimated to be 0.7, a liberal bias. The violation probability is 0.185, although specified to be 0.10. The figure shows a neighborhood based on $\delta = 0.05$ centered on α.

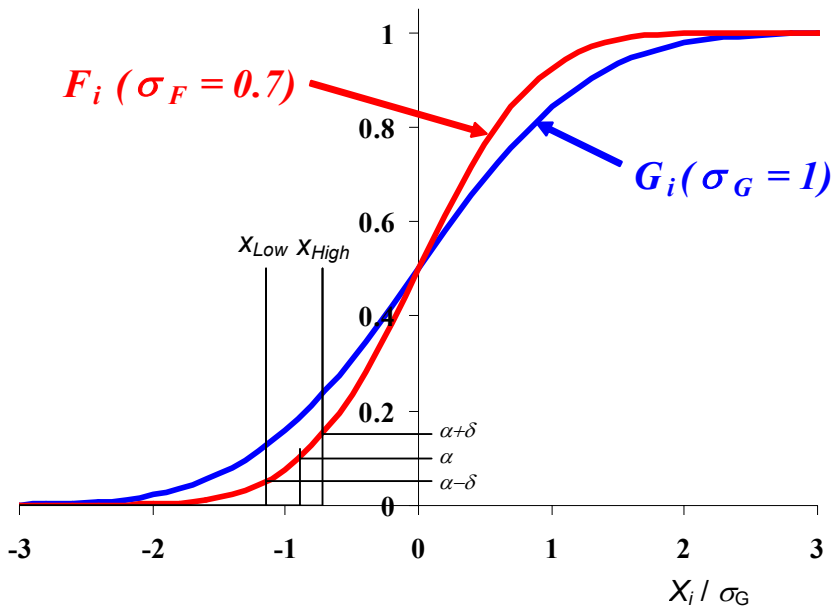

Figure 2: Defining a neighborhood.

Gains less than $x_{Low} = F_i^{-1}(0.10 - 0.05) = -1.514$ occur with probability 0.1248, rather than the estimated probability of 0.05, so the score statistic for all such observations is log(0.1248/0.05) = 0.9146, regardless of the value of X_i. These observations strongly support the alternative hypothesis and cause a rapid increase in the cumulative sum. Gains above x_{High} = $F_i^{-1}(0.10 + 0.05) = -0.7255$ occur with probability of 0.7659, rather than the estimated value of 0.85. All such observations have the score statistic of -0.1041, supporting the null hypothesis and decreasing the CUSUM.

In the neighborhood, the score statistic is $C_{\delta,i}$ = $-0.3567 + 0.5204 X_i^2$, which is positive for $X_i < -0.8279$ and negative otherwise. A plot of the score statistic versus the standardized portfolio gain is show in Figure 3.

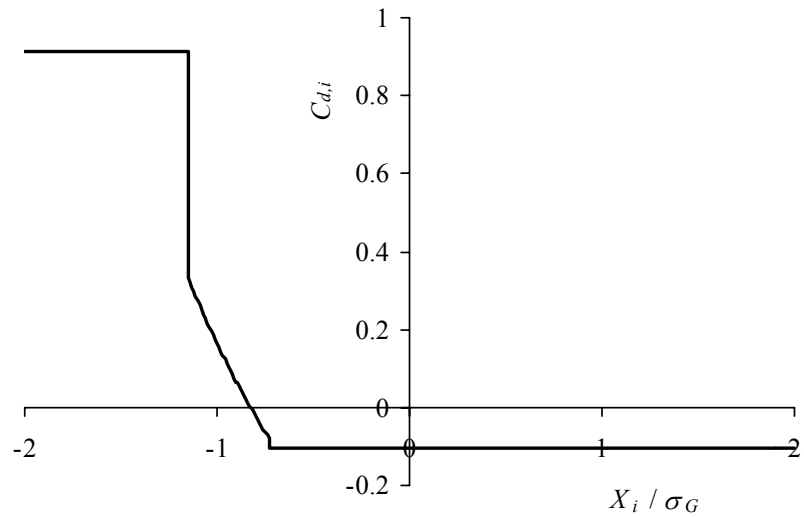

Figure 3: Score statistic.

Taking the expected value of the score statistic gives the Kullback information per observation, which is easily obtained for the simple case of normal observations. The information is smallest when the observations are converted to binary score statistics, which corresponds to classifying the observed gain as covered by the VaR or not. The information increases rapidly with neighborhood size as the tail of the distribution is included. Figure 4 shows a dramatic increase in the information with neighborhood size with a specified violation probability of 0.10 for a CUSUM designed to be most sensitive for a normal distribution with σ_G / σ_F = 1.43.

When the violation probability is smaller, using a neighborhood increases the information, but less dramatically. The information still increases steeply with neighborhood size, but the steep increase ends at $\delta = \alpha$, so the total increase is less for a reasonable neighborhood size. The situations for α = 0.05 and 0.01 are shown in Figure 5 and Figure 6.

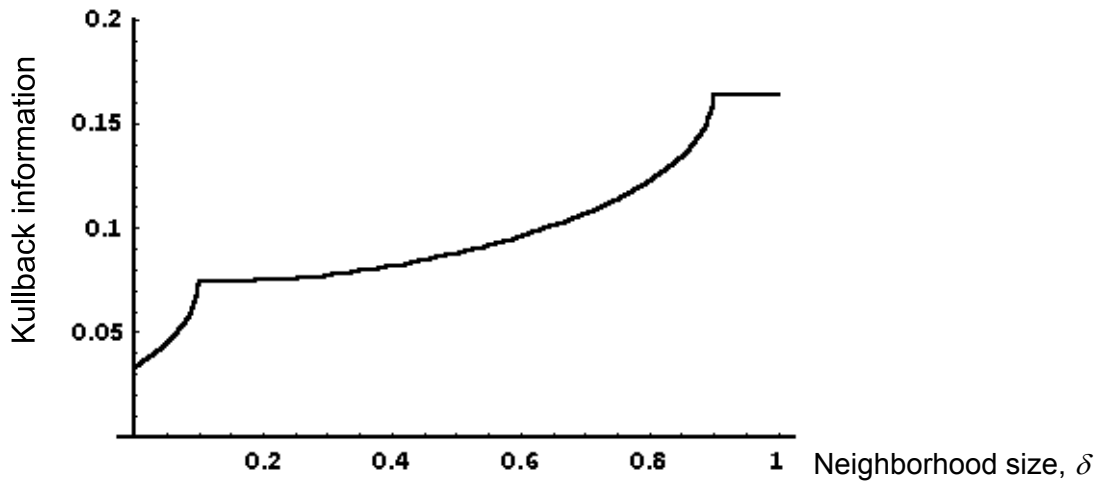

Figure 4: Variation in information with neighborhood size for α = 0.10.

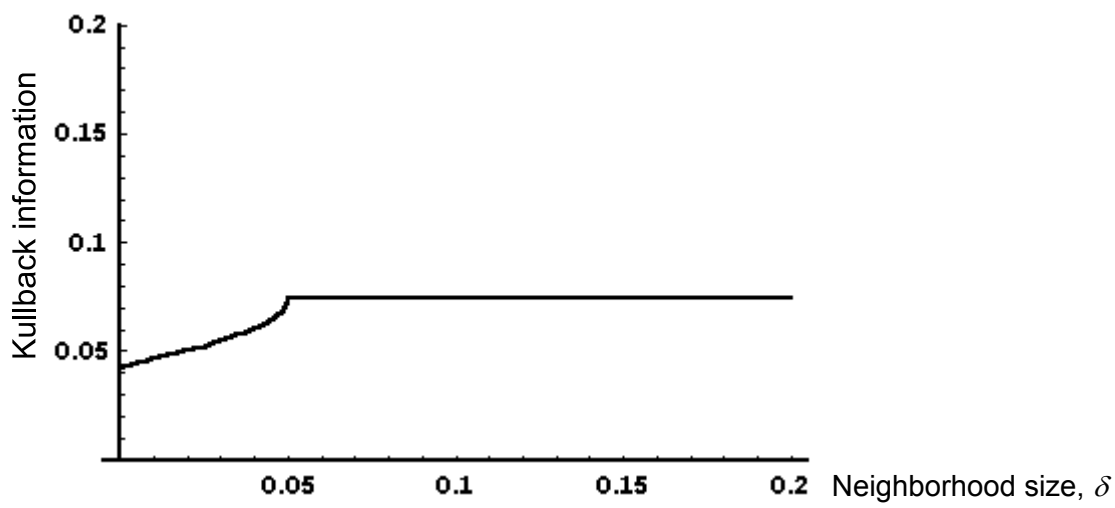

Figure 5. Effect of neighborhood size with α = 0.05.

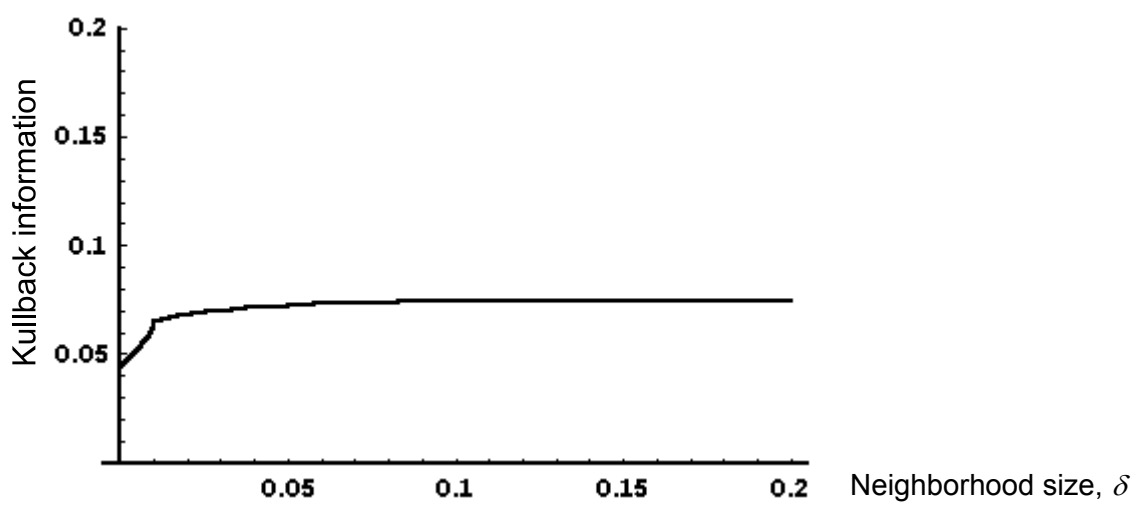

Figure 6: Effect of neighborhood size with α = 0.01.

The benefit of using a positive neighborhood also depends on whether the bias is conservative or liberal. When α is small, say 0.01, a conservative bias is more difficult to detect, as reflected in the smaller information content per observation. This is reasonable, since the most extreme conservative bias only reduces the violation probability from 0.01 to 0, so a large number of observations would be required to accumulate sufficient information to demonstrate such was the case. Figure 7 shows the much smaller information per observation with $\delta = 0$ for a CUSUM designed to detect a conservative bias, which is optimized for $\sigma_G / \sigma_F = 0.7$. There is so little information with $\delta = 0$ that the increase is proportionately much greater than in detecting a liberal bias. With $\delta = 0$, the information is only 0.00822, which increases to 0.030139 for $\delta = 0.05$. Since the asymptotic detection delay is proportional to the inverse of the information, using $\delta = 0.05$ reduces the asymptotic detection delay to about 27 percent of the delay with $\delta = 0$.

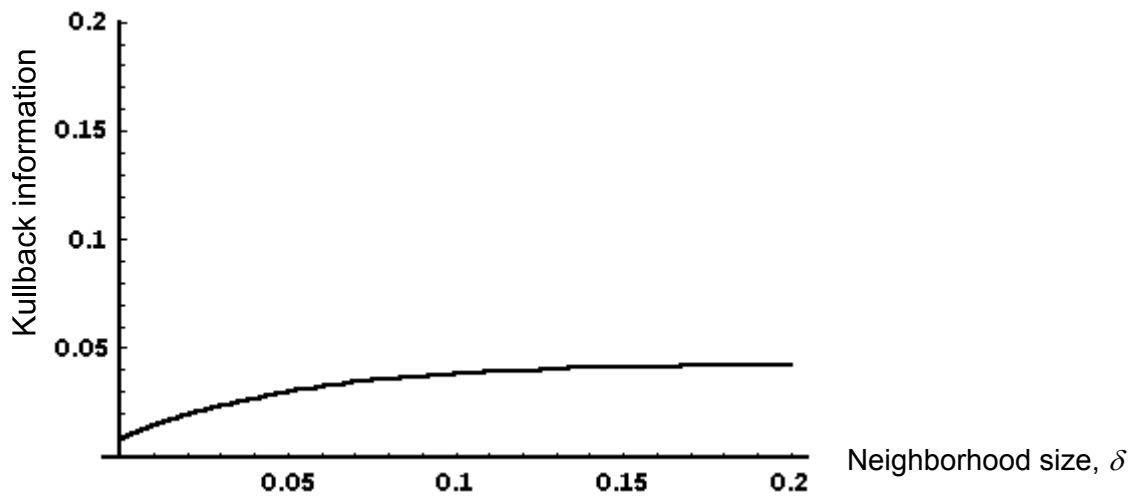

Figure 7: Effect of neighborhood size with conservative bias and $\alpha = 0.01$.

The variation in Kullback information with neighborhood size is tabulated for several combinations of VaR violation probability and neighborhood size in Table 1. This table shows the neighborhood size in the first column. The next three columns give the information per observation with $\sigma_G / \sigma_F = 1/0.7$, a liberal bias, and a CUSUM optimized

for this bias. The final three columns correspond to σ_G / σ_F = 0.7, a conservative bias, and a CUSUM optimized for the same bias.

	Liberal Bias, Increased Violation Probability			Conservative Bias, Decreased Violation Probability		
δ =	α = 0.01	α = 0.05	α = 0.10	α = 0.01	α = 0.05	α = 0.10
0.000	0.0442	0.0424	0.0328	0.0082	0.0258	0.0322
0.005	0.0525	0.0443	0.0339	0.0116	0.0275	0.0332
0.010	0.0648	0.0463	0.0350	0.0146	0.0291	0.0341
0.025	0.0670	0.0526	0.0383	0.0219	0.0331	0.0366
0.050	0.0727	0.0741	0.0448	0.0301	0.0374	0.0395
0.100	0.0742	0.0743	0.0744	0.0385	0.0411	0.0421
0.25 - α	0.0746	0.0746	0.0746	0.0422	0.0422	0.0422
0.50 - α	0.0819	0.0819	0.0819	0.0508	0.0508	0.0508
1.00	0.1637	0.1637	0.1637	0.1017	0.1017	0.1017

Table 1: Kullback information per observation.

As mentioned earlier, it is difficult to detect a conservative bias when α is small, which is confirmed by the small information per observation when δ = 0 in this situation. In all the combinations shown in the table, the information is greater with a liberal bias, which is equivalent to the detection of the liberal bias with a smaller average delay. For example, with α = 0.01 and δ = 0, the information under the liberal bias is 5.39 times that under the conservative bias, so the asymptotic expected detection delay would be about 5.39 times greater for detecting the conservative bias. A conservative bias is especially difficult to detect when α is small, such as 0.01. Using a neighborhood δ = α increases the Kullback information by 50 percent to 75 percent for α = 0.01 and 0.05 and seems to be a reasonable, practical compromise between purity and power.

IV. Skewed and Other Non-Gaussian Distributions

The CUSUM approach can be designed to be optimal for detecting non-Gaussian distributions as well. Even if optimized for detecting a Gaussian distribution, the test is very effective in detecting a systematic bias when the true distribution is non-Gaussian, as we demonstrate next.

Generally, one will want to detect a systematic bias that results in an increased or decreased violation probability. This can be accomplished with certainty if the CUSUM is designed with a neighborhood of size zero, a pure test. To see that this is true, consider a $\delta = 0$ CUSUM designed for a specified violation probability $\alpha \in (0,1)$ and a true violation probability $p_d \in (0,1)$. The score statistic is

$$C_{\delta,i} = \begin{cases} \text{Log}(p_d / \alpha), & X_i < F_i^{-1}(\alpha) \\ \text{Log}((1-p_d)/(1-\alpha)), & X_i > F_i^{-1}(\alpha) \end{cases}$$

If the true violation probability is $p_V \in (0,1)$, the Kullback information per observation is

$K = p_V \text{Log}\left(\frac{p_d}{\alpha}\right) + (1-p_v)\text{Log}\left(\frac{1-p_d}{1-\alpha}\right) = p_V \text{Log}\left(\frac{p_d(1-\alpha)}{\alpha(1-p_d)}\right) + \text{Log}\left(\frac{1-p_d}{1-\alpha}\right)$. Consider first the liberal case, $p_V > p_d > \alpha > 0$. The partial derivative with respect to p_V is positive for $1-p_d > 0$, so the information increases with p_V, demonstrating that the test will be effective with any bias more severe than that for which it is designed. For the conservative case, $0 < p_V < p_d < \alpha$, the partial derivative is negative, so the information increase as p_V decreases and the bias becomes more severe. Thus, the pure test will always be effective in detecting the type of bias for which it is designed.

When a positive neighborhood is specified, the expected number of observations to a signal will be decreased if the actual distribution results in an increase in the Kullback information per observation. It seems that most distributions that shift the violation probability in the direction for which the test is sensitive will increase the information, although some unrealistic exceptions can be found, as demonstrated later in this section. We will now consider some non-Gaussian distributions and show that the CUSUM test is effective in detecting a bias. We will consider a CUSUM designed with

Gaussian distributions using the score function of Equation (2), with α = 0.1 and σ_G / σ_F = 1.52. For this specification, the violation probability is exactly double the specified value. This results in information of 0.1231 per observation when the neighborhood size is 0.05. The ability to detect bias when the true distribution is non-Gaussian can be assessed by comparing this information with that produced by other distributions when the violation probability is also doubled.

Suppose the true distribution is Student's *t* with 5 degrees of freedom and a volatility of 1.291. If the estimated distribution is normal with volatility of 0.7175, the volatility ratio is about 1.7992, and the violation probability is exactly double the specified value. How quickly would this bias be detected? The information per observation is 0.1443, which is about 17 percent greater than that with the normal case, so the bias will be detected quicker, on average, with an average detection delay about 85 percent of that with the normal distribution. The higher kurtosis of the Student's *t* distribution gives greater probability for gains that correspond to the maximum score statistic.

Similarly, when the distribution is skewed, it is also possible to detect the bias quicker. For example, suppose that the true distribution is left skewed, with a density function of $f_{\Gamma,5,1}(5 - X_i)$, where $f_{\Gamma,5,1}(X)$ represents the density function for a gamma distribution with shape parameter of 5 and scale parameter of 1. Using the negative of the gain gives a distribution skewed to the left, and adding the shape parameter to the argument gives a distribution with mean of zero. This distribution has a skewness of -0.8944, excess kurtosis of 1.2, and volatility of 2.236. If the estimated distribution is normal with volatility of 1.343, then the volatility ratio is 1.6651 and the violation probability will be exactly twice the specified value. Keeping the same neighborhood size of 0.05 gives information per observation of 0.28776, about 133 percent more than that with the normal distribution. This bias would be detected with an average delay of about 43 percent of that with a normal distribution. This shows that with many non-normal distributions a similar bias is detected with even less delay.

We move now to an example of an extreme situation mentioned above such that a liberal bias is not detected by a test with a non-zero neighborhood. We make this point to show the potential drawback of choosing a neighborhood if an unusual distribution, very dissimilar to the anticipated distribution, is encountered. We do not believe such situations are likely to arise in practice with a small neighborhood. Consider again the CUSUM for α = 0.1 and σ_G / σ_F = 1.52. Let the true distribution be binary, or a bimodal continuous distribution that is highly concentrated at the two values. Let one of the values be x_1, say - 0.9, an arbitrarily small amount to the left of VaR, and having probability 2 α = 0.2. Let the other value be $\frac{-2\alpha}{1-2\alpha}\ x_1$, say 0.2250, with probability 1 – 2 α = 0.8. The expected value is zero, and the violation probability is twice the specified value. The score statistic for x_1 is 0.0649, and –0.1041 for the "well covered" value. Thus, the information per observation is $0.2 * 0.0649 - 0.8 * 0.1041 = -0.0703$, which is less than the information under the null distribution, -0.0346. Thus, this example of liberal bias would not be detected by the CUSUM, since the ARL would be greater than the average delay between false alarms.

V. Conclusion

Previous accuracy assessment literature has focused on backtesting a fixed number of observations, addressing the question of historical accuracy over some specified period. We propose a method suitable for determining if the VaR process is *currently* without systematic bias, based on testing a window of recent observations where the window size varies adaptively according to the data. The window size adjusts to exclude observations once the suspicion of bias is eliminated for those observations. In this way, the testing involves only the most recent observations for which the issue of bias remains open. This technique is optimal in the sense of minimizing the expected detection delay for a specified bias situation. It is also effective in detecting similar biases.

Generally, bias conditions can be grouped as those that systematically decrease the violation probability below the value specified, a conservative bias, and those that increase the violation probability, a liberal bias. Since a technique effective in detecting one type of bias may not be effective in detecting the other type, we propose to run instances of each type of test in parallel. In this way, either type of bias can be detected. Using this approach the average detection delay would be about 25 observations for a conservative bias that decreased the violation probability to 19 percent of the specified value or for a liberal bias that increased the violation probability to 214 percent of the specified value, based on a specified violation probability of 0.10 and a normal distribution for the actual distribution. The test would be effective in detecting systematic bias when the actual distribution is non-Gaussian, in some cases with less average detection delay.

We also addressed the issue of purity and detection power. Previous studies have viewed this as an all-or-nothing decision. A pure test of VaR converts the observed gain into a binary variable indicating whether it was covered by the VaR or not. However, pure tests typically require a large number of observations to detect a systematic bias in the VaR process. Testing the entire estimated distribution increases the detection power, but detects errors irrelevant to VaR and other measures of risk, such as errors in the upper tail of the distribution. Detecting these irrelevant errors can lead to chasing ghosts and to an erroneous lack of confidence in the risk management system. We have developed the idea of a continuous adjustment between purity and power by testing only the part of the estimated distribution within a specified neighborhood of the point used to calculate the VaR. The two cases treated in the literature emerge as special cases. A neighborhood size of zero gives the pure test, and we show that increasing the size slightly greatly increases the power with only a modest decrease in purity. This suggests a compromise that would be useful in practical applications.

References

Basseville, M. and Nikiforov, I. V. Detection of Abrupt Changes: Theory and Application. Englewood Cliffs, NJ: P T R Prentice Hall (1993).

Beckstrom, Rod A. "Value at Risk: Theoretical Foundations." In *Capital Market Strategies*. London, UK: IFR Publishing (1995).

Beder, Tanya. "VaR: Seductive but Dangerous." *Financial Analysts Journal*. (September/October 1993), 12 - 24.

Berkowitz, Jeremy, and James O'Brien. "How Accurate Are Value-at-Risk Models at Commercial Banks?" *Journal of Finance* 57 (3), (June 2002), 1093 – 1111.

Best, Philip. *Implementing Value at Risk*. Chichester, England: John Wiley & Sons, Inc. (1999).

Christoffersen, Peter F. "Evaluating Interval Forecasts." *International Economic Review*, (November 1998) 39 (4), 841-862.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review*, (November 1998) 39 (4), 863-883.

Dowd, Kevin. *Beyond Value at Risk*. New York, NY: John Wiley & Sons, Inc. (1998).

Hull, John C. *Options, Futures, & Other Derivatives, fifth edition*. Upper Saddle River, NJ: Prentice Hall (2002)

Jorion, Philippe. *Value At Risk: The New Benchmark for Controlling Market Risk*. Chicago, IL: Irwin Professional Publishing (1997)

Lai, Tze Leung. "Sequential Changepoint Detection in Quality Control and Dynamical Systems." *Journal of the Royal Statistical Society B*, 57 (4) (1995), 613-658.

Moustakides, G. V. "Optimal Stopping Times for Detecting Changes in Distribution." *Annals of Statistics*, 14 (1986), 1379-1387.

Page, E. S. "Continuous Inspection Schemes." *Biometrika*, 41 (1954), 100-114.

Ritov, Y. "Decision Theoretic Optimality of the CUSUM Procedure." *Annals of Statistics*, 18 (1990), 1464-1469.

Stoumbos, Z. G., M. R. Reynolds, Jr., T. P. Ryan, and W. H. Woodall. "The State of Statistical Process Control as We Proceed into the 21st Century." *Journal of the American Statistical Association*, 95 (2000), 992-998.