Risk journals

Volume 11 Number 4
Summer 2008

The Journal of Computational Finance

Volume 11 Number 4 Summer 2008

An incisivemedia publication

Incisive Media, Haymarket House, 28-29 Haymarket, London SW1Y 4RX

The Journal of
Computational Finance

The Journal of Computational Finance (1–31) Volume 11/Number 4, Summer 2008

An adaptive procedure for estimating coherent risk measures based on generalized scenarios

Vadim Lesnevski
Royal Bank of Scotland, Global Banking & Markets, 250 Bishopsgate, London EC2M 4AA, UK; email: vadim.lesnevski@rbos.com

Barry L. Nelson
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA; email: nelsonb@northwestern.edu

Jeremy Staum
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA; email: j-staum@northwestern.edu

Simulating coherent risk measures is potentially very computationally expensive. We present a procedure for generating a fixed-width confidence interval for a coherent risk measure based on a finite number of generalized scenarios. Computational experiments show that this procedure is much more efficient than standard methods, making simulation of coherent risk measures based on even a large number of generalized scenarios affordable. The procedure improves upon previous specialized methods by being reliably efficient when applied to simulation of generalized scenarios and portfolios with heterogeneous characteristics. We also show how robust the procedure's performance is to violations of the normality assumption under which its statistical validity is proved, and study the magnitude of estimation error.

1 INTRODUCTION

Coherent risk measures can improve the practice of risk management (Artzner *et al* (1999)) and pricing derivative securities (Jaschke and Küchler (2001); Staum (2004)). In some cases, coherent risk measures may need to be estimated by simulation. In such cases, especially for large firm-wide risk measurement problems, carrying out the simulation by standard methods could be much slower than simulations currently used in risk management and derivatives pricing, and too slow for routine use in practice.

This material is based upon work supported by the National Science Foundation under grants No. DMI-0217690, DMS-0202958 and DMI-0555485. Part of this material has been published in the *Proceedings of the 2006 Winter Simulation Conference*. We thank the Editor-in-Chief and an anonymous referee for their helpful comments that have led to an improved and expanded presentation.

To see why, consider that any coherent risk measure ρ has a representation of the form:

$$\rho(Y) = \sup_{\mathbf{P}\in\mathcal{P}} \mathrm{E}_{\mathbf{P}}[-Y/r] \tag{1}$$

where Y is the value of a portfolio at a future time horizon, r is a stochastic discount factor that represents the time value of money and $\mathcal{P}$ is a set of probability measures (Artzner *et al* (1999, Proposition 4.1)). Equations of a similar form hold for the related problems in derivative security pricing. We simplify the problem somewhat by limiting our analysis to the case where the set $\mathcal{P}$ has only a finite number k of elements $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$. The obvious way of estimating $\rho(Y)$ by simulation is to estimate $\mathrm{E}_{\mathbf{P}_i}[-Y/r]$ for each $i = 1, 2, \ldots, k$, which is typically about k times as expensive as estimating a single expectation by simulation. This may be impractically expensive when Y is the value of a portfolio that contains thousands of derivative securities and $\mathbf{P}_i$ represents a model governing hundreds of underlying risk factors.

The assumption that $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k\}$ holds, for instance, when the decision-maker designs the coherent risk measure (or the underlying acceptance set, in the case of derivative security pricing) by specifying these k generalized scenarios. The SPAN margin computation system of the Chicago Mercantile Exchange is closely related to such a risk measure. To simplify the example, we consider applying SPAN to a portfolio involving a futures contract for delivery in a single month and options on that contract. In this case, our risk measure has $k = 16$ generalized scenarios, which involve the changes over one day in the futures price and in the implied volatility of the options. They are based on a hypothetical moderate move δ or an extreme move Δ in the futures price and a hypothetical change σ in the implied volatility, as illustrated in Table 1. The first 14 probability measures are degenerate: each of them is a point mass on one scenario. The 15th and 16th probability measures place 35% probability on an extreme move in the futures price and 65% probability on no change; the expected loss under these probability measures is close to 35% of the loss in the case of an extreme move, which SPAN uses. The maximum expected loss over all generalized scenarios is used to compute margin requirements for the portfolio of futures and options.

The assumption that $\mathcal{P}$ is finite does not generally hold for worst conditional expectation (Artzner *et al* (1999, Section 5)) or related risk measures such as tail conditional expectation, conditional value-at-risk and expected shortfall (see Acerbi and Tasche (2002)). However, our procedure may provide a foundation for further work on efficiently simulating worst conditional expectation. For coherent risk measures such that $\mathcal{P}$ is infinite, it may also be possible to use our procedure by approximating $\mathcal{P}$ by the convex hull of k probability measures.

In Lesnevski *et al* (2007), we used tools from the ranking and selection literature to create efficient procedures that generate a fixed-width confidence interval for a coherent risk measure. Those procedures were usually no more than twice as expensive as estimating a single expectation by simulation, not k times as expensive; in some cases, they were as little as 5% more expensive. However, those procedure have a weakness: for some problems, that is, for some sets $\mathcal{P}$ and

TABLE 1 Generalized scenarios for SPAN.

Generalized scenario	Probability (%)	Futures price	Implied volatility
1	100	$+0$	$+\sigma$
2	100	$+0$	$-\sigma$
3	100	$+\delta/3$	$+\sigma$
4	100	$+\delta/3$	$-\sigma$
5	100	$-\delta/3$	$+\sigma$
6	100	$-\delta/3$	$-\sigma$
7	100	$+2\delta/3$	$+\sigma$
8	100	$+2\delta/3$	$-\sigma$
9	100	$-2\delta/3$	$+\sigma$
10	100	$-2\delta/3$	$-\sigma$
11	100	$+\delta$	$+\sigma$
12	100	$+\delta$	$-\sigma$
13	100	$-\delta$	$+\sigma$
14	100	$-\delta$	$-\sigma$
15	35	$+\Delta$	$+0$
	65	$+0$	$+0$
16	35	$-\Delta$	$+0$
	65	$+0$	$+0$

portfolio values Y, they could be substantially less efficient than standard methods. In this paper, we improve upon earlier procedures by creating an adaptive procedure whose efficiency is robust to variation in problem specification because it uses simulated data to determine when to employ efficiency improvement techniques. We also demonstrate that the adaptive procedure generates a confidence interval whose coverage is robust to violation of the normality assumption used to prove its statistical validity, and investigate the magnitude of estimation error in the rare event that the confidence interval does not include the true value of the risk measure.

To explain how our procedures enhance efficiency, we introduce some new notation and terminology. Let $X := -Y/r$ and $\mu_i := \mathrm{E}_{\mathbf{P}_i}[X]$. The risk measurement involves a single random variable X, which is a negative discounted portfolio value or a discounted loss, viewed under multiple probability measures. For clarity in discussing simulations, let X_i be a random variable whose distribution under the probability measure Pr is the same as that of X under $\mathbf{P}_i$, that is, such that $\Pr[X_i \le x] = \mathbf{P}_i[X \le x]$. Because of the parallel with ranking and selection, we refer to X_i as an observation of system i. In ranking and selection, we are interested in the best system, which is the one with the largest mean. In the risk measurement of Equation (1), the best system is the worst generalized scenario, the one with the largest expected loss.

Our procedures use screening to eliminate some systems that seem likely to be inferior after generating a small number of observations, instead of expensively getting precise estimates of the means of all k systems. To sharpen screening we employ common random numbers to induce positive correlation between the

systems and thereby reduce the variance of their differences: see Glasserman (2004, pp. 361, 380) or Law and Kelton (2000). To reduce the number of replications required for estimation, we employ control-variate estimators to exploit strong correlation between the response of interest, X, and a vector C of random variables with known expectations, called control variates: see Glasserman (2004, Section 4.1) or Law and Kelton (2000).

A disadvantage of the procedures presented in Lesnevski *et al* (2007) is that in some cases the user might need to choose the procedure or its parameters based on previous knowledge about the problem to gain efficiency. For example, having a large screening budget is usually good, as it allows the procedure to screen out most of the inferior systems. However, it might significantly decrease efficiency if more than one system has the maximum mean, or if some systems are nearly tied with the best. In such situations, screening might not be able to eliminate all systems but one, and the work done during screening might be more than is necessary to estimate the coherent risk measure accurately.

One of the procedures in Lesnevski *et al* (2007) uses the technique of "restarting", in which data that is used for screening are subsequently discarded so as to make it possible to reduce the required sample sizes for the systems that survive screening. The advantage of restarting is that the new data is statistically independent of the screening exercise, so one may ignore the measures that were screened out, and design for the smaller problem. Even though the procedure with restarting is usually preferable over other alternatives, if screening is ineffective, restarting is wasteful of data. Without restarting, information generated during screening is reused during estimation of the confidence interval, so the only danger of a large screening budget is that it might exceed the sample size required for accurate estimation. With restarting, information generated during screening is thrown away, so it is important to make sure that no excess work is done during screening. Before running the simulation, the user would have to decide whether or not to use restarting and how much data to allocate to the screening stage. Making a good decision without substantial experience with simulation problems of the same form is difficult. In this paper we develop an adaptive multi-stage procedure that is reliably efficient. It gains the benefits of restarting and of having a large budget to use for screening by restarting when simulated data suggests that restarting is worthwhile, rather than at a prespecified time that might be disadvantageous. With the adaptive procedure, the user does not have to guess whether or not to use restarting or what the screening budget should be.

In Section 2, we present motivating examples in which coherent risk measures are estimated. The computational experiments that illustrate the procedure's performance are carried out with these examples. Section 3 describes our adaptive procedure and gives a heuristic justification of its design, while the proof of its statistical validity is in Appendix A. Computational experiments demonstrating the procedure's efficiency are described in Section 4, while Sections 5 and 6 feature experiments that test the robustness of the confidence interval's coverage to non-normal data and explore the severity of error in the unlikely event that the confidence interval does not contain the true value. Section 7 concludes the paper.

2 MOTIVATING EXAMPLES

We will test the performance of our procedures on two examples, which were also used in Lesnevski *et al* (2007). We selected these examples because it is easy to find the true best mean, which we must do to study the coverage of the confidence interval that the procedure produces, but we believe that these examples have a structure similar to that of problems in which estimating the true best mean (worst expected loss) would be a significant challenge.

2.1 Basket put

The first problem is to price a basket put option, whose payout at a terminal time T is $\max\{0, K - w'S(T)\}$, where K is the strike price, w is a vector of weights and $S(T)$ is the vector of terminal prices of the securities in the basket. The underlying security price vector S obeys the Black–Scholes model, so the price of the basket put price is its risk-neutral expected discounted payout.

Under the Black–Scholes model, the price vector S follows multivariate geometric Brownian motion with risk-neutral drift r, the risk-free interest rate, and with covariance matrix Σ. That is, $\ln S_j(T) = \ln S_j(0) + (r - \|A_j\|^2/2)T + A_j Z\sqrt{T}$, where A is a matrix satisfying $AA' = \Sigma$, $\|A_j\|$ is the Euclidean norm of its jth row, ie, the volatility of the jth asset, and Z is a multivariate standard normal random vector. The short-term interest rate r is observable, and there are standard methods for calibrating the underlying securities' individual volatilities $\|A_j\|$, whether from historical data or by fitting to observable prices of market-traded options on the underlying securities: see Cont and Tankov (2004, Chs 7, 13) and Shiryaev (1999, Ch. IV). However, estimation of the non-diagonal elements of Σ poses a greater problem. For pricing the basket put, the crucial quantity is $\|w'A\|$, the volatility of the basket, and this depends strongly on the correlations between assets. There may be a range of plausible correlations and thus a range of plausible prices for the basket put.

In this example, the basket is a weighted average of three security prices with weights $w_1 = 0.5$, $w_2 = 0.3$ and $w_3 = 0.2$. The initial security prices are all 100, and the strike price is $K = 85$. The interest rate $r = 5\%$ and the volatilities are $\|A_1\| = 40\%$, $\|A_2\| = 30\%$ and $\|A_3\| = 20\%$. To account for uncertainty about correlations, we use the $k = 4^3 = 64$ probability measures produced by allowing each of the three pairwise correlations to be 0.2, 0.35, 0.55 or 0.75. Although the payout in this example is far from normally distributed, the sample averages are approximately normally distributed and the minimum coverage guarantees the confidence limits held in all our experiments.

The three control variates used in this example are the discounted payouts of put options with strike K on each individual asset in the basket. Their means are given by the Black–Scholes pricing formula, based on the known volatilities. The idea behind using them as control variates is that much of the (unknown) error in estimating the basket put's expected discounted payout can be explained as a linear function of the differences between the discounted payouts of the puts on individual

assets and their means, which are known; this makes possible a reduced-variance estimate of the basket put's price.

2.2 Options portfolio

In this example we assess the risk of a portfolio of European-style call and put options on three assets with initial prices of 100 and terminal prices $S_1(T)$, $S_2(T)$ and $S_3(T)$. All options in the portfolio expire at a terminal time T. We also consider a market index whose terminal level is $S_0(T)$. For each of $j = 0, 1, 2, 3$, $S_j(T)$ follows geometric Brownian motion with drift d_j and volatility σ_j, so $\ln S_j(T) = \ln S_j(0) + (d_j - \sigma_j^2/2)T + \sigma_j W_j \sqrt{T}$, where the W_j is standard normal. There is a one-factor model of dependence among the assets: under a probability measure $\mathbf{P}$, Z_0, Z_1, Z_2 and Z_3 are independent standard normal random variables, $W_0 = Z_0$, and $W_j = \lambda_j Z_0 + \sqrt{1 - \lambda_j^2} Z_j$ for $j = 1, 2, 3$. In this model, Z_0 corresponds to the market factor common to all assets, while Z_1, Z_2 and Z_3 are idiosyncratic factors corresponding to each individual asset.

The risk measure we consider in this setting is the maximum expected loss incurred while holding the portfolio, where the maximum is taken over $4^4 = 256$ conditional expectations given a generalized scenario. Of the probability measures $\mathbf{P}_i$ in Equation (1), 255 are defined by $\mathbf{P}_i[E] = \mathbf{P}[E|A_i]$ for some event A_i of probability $\mathbf{P}[A_i] = 1/20 = 5\%$, while the 256th probability measure is $\mathbf{P}$ itself. This risk measure is similar in spirit to worst conditional expectation (Artzner *et al* (1999, § 5)). We construct generalized scenarios by restricting some of the factors Z_0, Z_1, Z_2 and Z_3. Each of the factors can be "up" (corresponding to a large increase of the asset price), "down" (a large decrease), "middle" (not extreme) or "unrestricted". The probabilities of the restrictions on the restricted factors are always equal. For example, letting Φ be the standard normal distribution function, in the scenario "up-down-unrestricted-unrestricted", Z_0 is sampled conditional on exceeding $\Phi^{-1}(1 - 1/\sqrt{20})$, Z_1 is sampled conditional on being below $\Phi^{-1}(1/\sqrt{20})$, while Z_2 and Z_3 are not restricted. By independence among Z_0, Z_1, Z_2 and Z_3, the probability of this event is $1/20$. The time horizon T is one week, and the parameters were calibrated using three years of historical weekly data on the S&P500 index and shares of Intel (INTC), ExxonMobil (XOM) and Microsoft (MSFT). The result was the annualized volatilities $\sigma_1 = 39.8\%$, $\sigma_2 = 19.3\%$ and $\sigma_3 = 27.0\%$ and the factor loadings $\lambda_1 = 0.617$, $\lambda_2 = 0.368$, and $\lambda_3 = 0.785$ to match the observed correlations. Because one week is such a short period of time that the expected return is negligible, while mean returns are hard to estimate due to a high ratio of volatility to mean, we take each $d_j = 0$. Since we do not need to simulate S_0, the parameters d_0 and σ_0 are not relevant.

We investigated the performance of our procedures on several portfolios. The extent of the efficiency improvement depends on the portfolio, so here we present a portfolio yielding results that we consider typical. Table 2 lists the number of each type of option in this example portfolio. Each option is the right to buy or sell 100 shares. We do not use control variates in this example.

TABLE 2 Amounts of options in the portfolio.

Asset	Option type	Strike price						
		85	90	95	100	105	110	115
1	Put	−2,000	−2,000	−2,500	1,000	0	0	0
2	Put	2,500	−1,000	1,000	500	0	0	0
3	Put	1,500	1,000	2,500	−1,500	0	0	0
1	Call	0	0	0	−1,000	1,500	−500	−1,000
2	Call	0	0	0	1,500	−2,500	2,000	−2,000
3	Call	0	0	0	−2,000	−1,000	1,000	2,500

3 ADAPTIVE MULTI-STAGE PROCEDURE

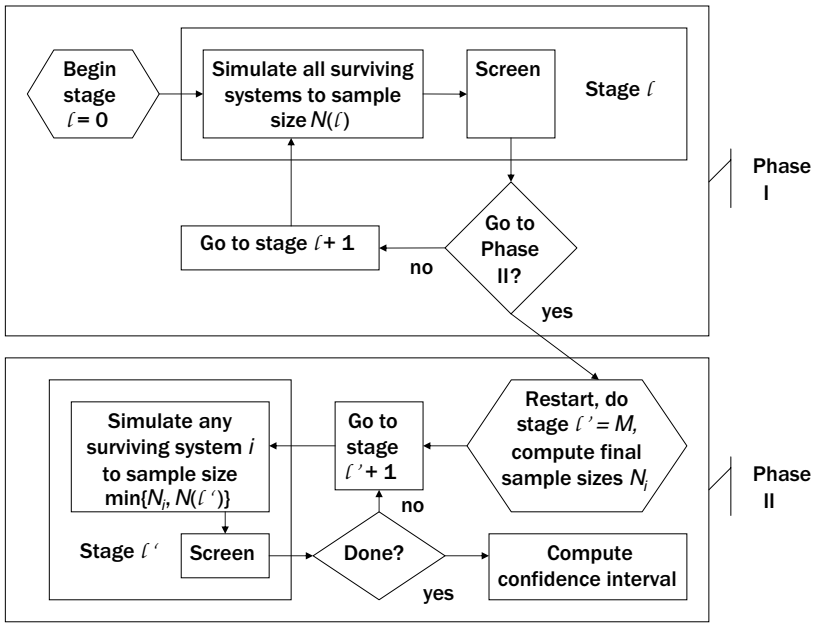
Our procedure produces a lower confidence limit that covers the coherent risk measure with probability at least $1 - \alpha_a$, and an upper confidence limit that covers with probability at least $1 - \alpha_b$, (see Appendix A for a proof). The procedure spends some of the allowable error α_a or α_b on screening (α_I), some on control variates (α_C) and the remainder on estimating the means of the systems that survive screening. We use the control variate C_i for the output X_i of system i to reduce the variance of estimating the mean μ_i of each system $i \in I$, where I is the set of systems that survive screening.

Our procedure updates the set I by screening in multiple stages. Each stage involves simulating a prespecified number of observations from each system that was in I at the beginning of the stage and then screening using all available observations. A system is screened out and removed from I if the sample average of all observations generated from that system is sufficiently far below some other system's sample average, relative to the estimated variance of the difference between the systems. We use common random numbers to reduce the variance of the differences between systems, but our procedure could be used without common random numbers. We do not use control variates in screening: we found little added benefit because common random numbers alone were so effective for the financial examples we considered, and using control variates in screening introduces technical complications (Nelson and Staum (2006)).

Once screening is completed, the procedure must estimate the means of all systems that have not been screened out. At some point, the procedure needs to determine for each system how many observations would be required to get a sufficiently precise estimate of the mean: this depends on the system's estimated variance and on the number of systems, which contributes to the natural bias of the problem of estimating the best. The purpose of restarting is to reduce this natural bias and thus the required sample sizes for surviving systems by throwing away data that was simulated when all of the systems were still in I.

Our adaptive procedure determines the required sample sizes when it restarts, which marks a transition between two phases. Phase I ("prescreening") consists of multi-stage screening whose purpose is, while controlling relative cost, to screen

FIGURE 1 A flowchart representing the adaptive multi-stage procedure.



out as many inferior systems as possible, so that they do not contribute to the critical values that determine the overall sample size for mean estimation. The procedure then restarts: no observations obtained during prescreening are used during Phase II. Phase II begins by determining required sample sizes for the remaining systems, continues to use multi-stage screening as it simulates the required number of observations and ends by constructing the confidence interval around the largest estimated mean of any system that has survived screening. Figure 1 shows a flowchart illustrating the procedure's phases, each of which contains multiple stages. In the figure, stage ℓ is a representative stage of Phase I and stage ℓ' is a representative stage of Phase II.

3.1 Phase I: prescreening

The sole purpose of the first phase is to reduce the number of systems and thus the natural bias of the estimation problem, making a fixed-width confidence interval attainable with fewer replications.

The maximal number of Phase I stages, m, is specified in advance. The first stage of Phase I is stage 0 and the first stage of Phase II is stage M, where the random variable $M \leq m$. The decision to proceed to Phase II is based on the simulated data, when the cost of continuing and doing one more stage of Phase I is greater than the estimated approximate savings due to further prescreening. The growth rate R and

the initial sample size n_0 are also specified in advance, so that the total sample size during stage ℓ is $N(\ell) = \lceil n_0 R^\ell \rceil$.

The initial sample size n_0 should be chosen so that sample averages are approximately normal. In most cases, $n_0 = 30$ is adequate. The procedure is most efficient if the growth factor R is between 1.2 and 2.0, while m is such that the total budget available for prescreening is large. For example, if $R = 1.5$ and $m = 30$, the total budget available for Phase I is $\lceil n_0 R^{m-1} \rceil = 3{,}835{,}021$, which is large enough for most applications. We found that $R = 1.5$ and $m = 30$ worked well on all problems we consider. It was not possible to improve much on the performance by altering the parameters, as it was for the procedures presented in Lesnevski *et al* (2007).

Let I be the set of systems that have not been screened out. Initially, set $I \leftarrow \{1, \ldots, k\}$. Each stage $\ell = 0, \ldots, m-1$ of Phase I consists of the following steps.

1) *Simulation:* simulate (X_{ij}, C_{ij}) for $j = N(\ell-1)+1, \ldots, N(\ell)$ and all $i \in I$.
2) *Screening:* for each $h, i \in I$ such that $h \neq i$, set:

$$\bar{\bar{D}}_{hi} \leftarrow \frac{1}{N(\ell)} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij})$$

$$S_{hi}^2 \leftarrow \frac{1}{N(\ell)-1} \sum_{j=1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2$$

$$W_{hi} \leftarrow t_{N(\ell)-1,1-\alpha_I/(2m(k-1))} S_{hi} / \sqrt{N(\ell)}$$

where $t_{\nu,p}$ is the p quantile of the t distribution with ν degrees of freedom. Then set $I \leftarrow \{i \in I | \forall h \in I, \bar{\bar{D}}_{hi} \geq -W_{hi}\}$.

3) *Checking whether to proceed to Phase II:* for each $i \in I$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,1}, \ldots, X_{i,N(\ell)}$ on $C_{i,1}, \ldots, C_{i,N(\ell)}$ and define:

$$c_p := \frac{1}{L}(\Phi^{-1}(1 - \alpha_a/p + \alpha_C) + \Phi^{-1}(1 - \alpha_b + \alpha_I + \alpha_C)) \tag{2}$$

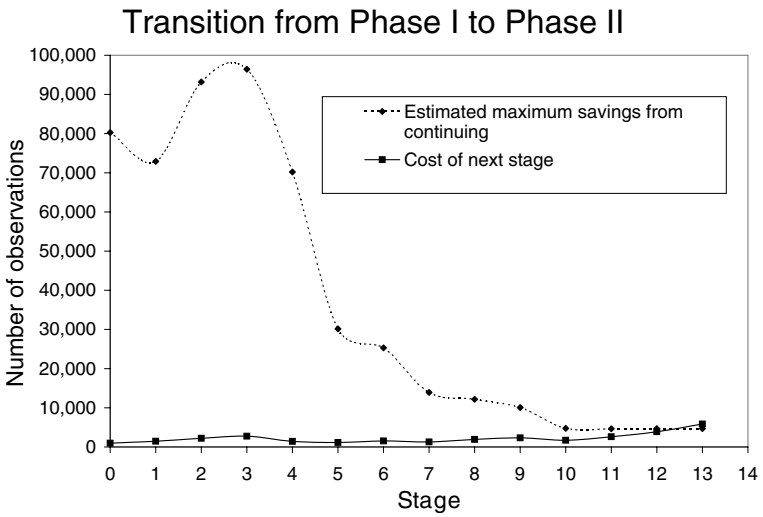
where Φ is the standard normal cumulative distribution function. If $\ell = m-1$ or:

$$|I| N(\ell)(R-1) > (c_{|I|}^2 - c_1^2) \max_{i \in I} \hat{\sigma}_i^2 \tag{3}$$

the procedure jumps to Phase II by setting $M \leftarrow \ell + 1$, which means that the next stage is the first stage of Phase II, and by setting $K \leftarrow |I|$, which is the number of systems left after prescreening and which will be used for determining final sample sizes. Otherwise, set $\ell \leftarrow \ell + 1$ and return to Step 1.

Under the transition rule given by inequality (3), prescreening stops when the cost of doing one more stage of prescreening exceeds an estimate of the maximum savings that could occur if prescreening continues. The estimate is computed under the assumption that after additional prescreening there will be only one system

FIGURE 2 Operation of the transition rule during one run of the adaptive procedure on the "2 best" configuration of the basket put example.

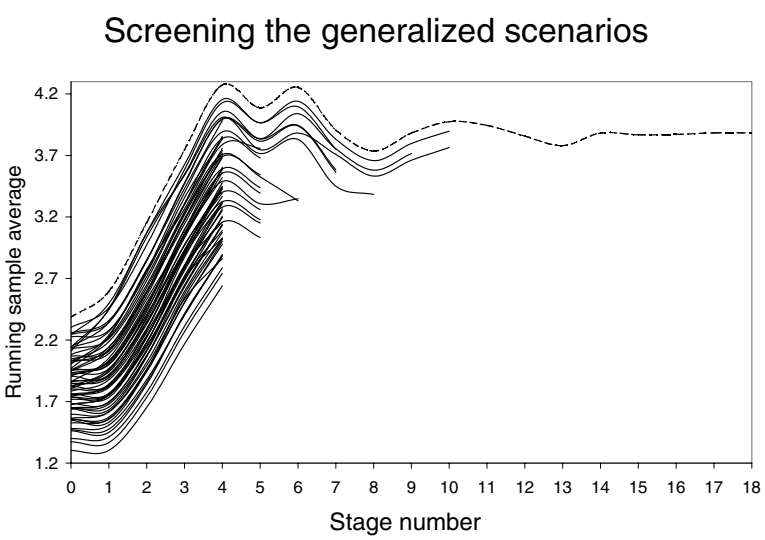


left and it will have the largest variance. For an explanation of the transition rule, see Section 3.3. Figure 2 illustrates how the transition rule works in one particular example. The example is the problem of pricing the basket put option to 5% precision, but with a duplicate of the best system – see Section 4 for details. For the same example, Figure 3 shows how the systems' sample averages change from stage to stage, and when screening eliminates the systems. The presence of the duplicate best system allows these figures to illustrate an important feature of the transition rule: although all systems but the two identical best systems (the top, superimposed lines in Figure 3) are eliminated after stage 10, Phase I does not end until after stage 13, when the cost of stage 14 finally exceeds the savings the procedure hopes to realize by eliminating one of the two remaining systems (Figure 2).

3.2 Phase II: screening and estimation

Phase II begins by restarting, that is, throwing out all the data obtained in Phase I. The only effect of Phase I on Phase II is that Phase I determines the subset I of systems that Phase II handles. The purpose of Phase II is to create a fixed-width confidence interval based on fresh data, uncontaminated by the selection bias caused by Phase I screening. If Phase II begins with more than one system, then the process of selecting one of them during Phase II will also create selection bias. However, if Phase II begins with fewer than k systems, the selection bias will be less than in Phase I. This reduction in selection bias is the purpose of restarting, and it allows a confidence interval of a fixed width to be created with fewer observations

FIGURE 3 Multi-stage screening during one run of the adaptive procedure on the "2 best" configuration of the basket put example.



than otherwise. Phase II contains three parts: first it determines the sample sizes required for all surviving systems, then it simulates observations up to this sample size while using screening and finally it produces a confidence interval based on the largest sample average of any surviving system.

In the initial stage M of Phase II, the procedure determines the required total sample sizes N_i for each of the systems in I and the maximal necessary number P of subsequent screening stages. Second, in stages $M, \ldots, M+P-1$, the procedure does more screening. It maintains two sets of systems: the set I contains systems that have survived screening and from which the procedure has simulated as many samples as are required to construct the fixed-width confidence interval, while the set $\hat{I}$ contains systems that have survived screening so far, but which still require more sampling. Finally, once the required sample size has been reached for all surviving systems, the procedure constructs a confidence interval.

Because M is the first stage after restarting, the procedure discards $\lceil n_0 R^{M-1} \rceil$ Phase I samples. To compensate for the discarded samples and keep the growth rate constant, during Phase II the procedure sets $N(\ell) \leftarrow \lceil n_0 R^{\ell-1}(R+1) \rceil$, $\ell \geq M$. This makes the total Phase II sample size grow at the rate R. It also makes the initial sample size of Phase II be $N(M) - N(M-1) \simeq n_0 R^M$, which is large enough to ensure high-quality variance estimates.

Initialize $\hat{I} \leftarrow I$ and then $I \leftarrow \emptyset$. Also initialize $N_i \leftarrow N(M)$ for all $i \in \hat{I}$. Each stage $\ell = M, \ldots, M+P$ consists of the following steps, except that only stage M contains Step 2, and Step 4 will not occur during stage $M+P$ because $\hat{I}$ will be empty.

1) *Simulation:* simulate (X_{ij}, C_{ij}) for $j = N(\ell-1)+1, \ldots, \min\{N_i, N(\ell)\}$ and all $i \in \hat{I}$.
   Set $n \leftarrow N(\ell) - N(M-1)$.
2) *Setting final sample sizes:* if $\ell > M$, skip this step.
   Set $\alpha_a'' \leftarrow \alpha_a/K - \alpha_C$ and $\alpha_b'' \leftarrow \alpha_b - \alpha_I - \alpha_C$, and set the scaling constant:

$$c \leftarrow \frac{1}{L}(t_{n-q-1,1-\alpha_a''} + t_{n-q-1,1-\alpha_b''}) \tag{4}$$

   where $q := \max_{i\in I} q_i$ and each q_i is the number of control variates in C_i.
   For each $i \in \hat{I}$, compute the residual variance $\hat{\sigma}_i^2$ of regressing $X_{i,N(M-1)+1}, \ldots, X_{i,N(M)}$ on $C_{i,N(M-1)+1}, \ldots, C_{i,N(M)}$, and from it the total sample size:

$$N_i \leftarrow \lceil c^2\hat{\sigma}_i^2 + \chi^2_{q_i,1-\alpha_C} \rceil + N(M-1) \tag{5}$$

   where $\chi^2_{\nu,p}$ is the p quantile of the chi-squared distribution with ν degrees of freedom.
   Set $P \leftarrow \lceil \log_R \max_{i\in I}(N_i/N(M)) \rceil$.
3) *Updating I and $\hat{I}$:* add systems that have reached their required sample sizes to I and remove them from $\hat{I}$: set $I \leftarrow I \cup \{i \in \hat{I} | N_i \le N(\ell)\}$ and $\hat{I} \leftarrow \hat{I} \backslash I$.
4) *Screening:* for each $h, i \in \hat{I}$ such that $h \neq i$, set:

$$\bar{\bar{D}}_{hi} \leftarrow \frac{1}{n} \sum_{j=N(M-1)+1}^{N(\ell)} (X_{hj} - X_{ij})$$

$$S_{hi}^2 \leftarrow \frac{1}{n-1} \sum_{j=N(M-1)+1}^{N(\ell)} (X_{hj} - X_{ij} - \bar{\bar{D}}_{hi})^2$$

$$W_{hi} \leftarrow t_{n-1,1-\alpha_I/(2P(K-1))} S_{hi}/\sqrt{n}$$

   Then set $\hat{I} \leftarrow \{i \in \hat{I} | \forall h \in I, \bar{\bar{D}}_{hi} \ge -W_{hi}\}$.
5) *Continue or compute confidence interval:* if $\hat{I} \neq \emptyset$, set $\ell \leftarrow \ell + 1$ and return to Step 1.
   Otherwise, for each $i \in I$, compute the estimate $\widehat{\mu}_i$ from the regression of $X_{i,N(M-1)+1}, \ldots, X_{i,N_i}$ on $C_{i,N(M-1)+1}, \ldots, C_{i,N_i}$. Set:

$$a \leftarrow \frac{1}{c} t_{N(M)-N(M-1)-q-1,1-\alpha_a''} \quad \text{and}$$

$$b \leftarrow \frac{1}{c} t_{N(M)-N(M-1)-q-1,1-\alpha_b''}$$

   The confidence interval is:

$$\left(\max_{i\in I} \widehat{\mu}_i - a, \max_{i\in I} \widehat{\mu}_i + b \right)$$

3.3 Efficiency of the rule for restarting

The adaptive procedure offers two significant improvements over our previous procedures.

First, we do not need to specify a screening budget in advance. Choosing the screening budget too small or too big could have a very significant effect on the performance of our previous procedures, in some configurations making a simulation dozens of times slower; see Table 6 in Section 4.1. The adaptive procedure solves this problem by trying to screen out a system in Phase II only until its required sample size is reached. In effect, this allows the screening budget to be arbitrarily large, to vary by system and to be determined adaptively by the required sample size.

Second, the adaptive procedure allows us to restart whatever the configuration of the means $\mu_1, \ldots, \mu_k$ may be. The effect of the decision whether or not to restart on performance is much less severe; as we will show below, usually we do not expect to save more than 40–80%. Restarting is usually beneficial because in a typical case there is only one best system. Having an adaptive prescreening phase identifying a good time to restart allows us to achieve very good performance in a typical case and reasonably good performance in all other cases.

How big are the benefits of prescreening in a typical case? To answer this question let us first estimate the maximal possible savings due to restarting.

In the following analysis we make several simplifying assumptions. First, we assume that the estimate of the residual variance $\hat{\sigma}_i^2$ of system i is always approximately equal to the true residual variance σ_i^2. Second, we ignore the effect of the number of degrees of freedom on the sample sizes for estimation. Third, we assume that the effort required for screening out an inferior system is always the same, whether in Phase I, Phase II or in an alternative procedure without prescreening and restarting (such as the multi-stage procedure with early stopping in Lesnevski *et al* (2007)).

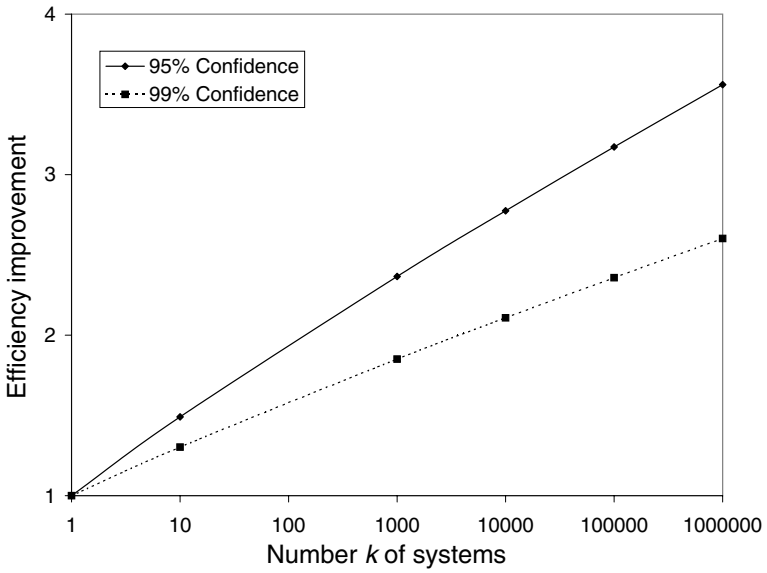
The total cost E of a simulation without prescreening is the sum of the cost E_s of screening out inferior systems and the cost E_e of estimation of the surviving systems: $E = E_s + E_e$.

The total cost $\tilde{E}$ of a simulation with prescreening is the sum of the prescreening cost $\tilde{E}_p$, the cost $\tilde{E}_s$ of screening out inferior systems in Phase II and the estimation cost $\tilde{E}_e$ of the surviving systems: $\tilde{E} = \tilde{E}_p + \tilde{E}_s + \tilde{E}_e$.

Under our assumptions, the sample size N_i in Equation (5) is approximately equal to $c^2\sigma_i^2$. Without prescreening, the constant c in Equation (4) is approximately equal to c_k defined in Equation (2), where k is the initial number of systems. With prescreening, c is approximately c_K, where K is the number of systems remaining after prescreening. The smaller K, the larger the benefit of prescreening, because smaller c_K leads to smaller sample sizes for estimation.

We will assume that, whether we simulate with prescreening or not, the set I of the surviving systems is the same. This is generally so when prescreening is stopped before the sample sizes for some systems exceed the sample sizes required for estimation, which is exactly the case when prescreening could be beneficial.

FIGURE 4 Maximal efficiency improvement due to restarting with $\alpha_a = 0.8\alpha$ and $\alpha_b = 0.2\alpha$.



A simulation without prescreening costs $E = E_s + c_k^2 \sum_{i \in I} \sigma_i^2$, and a simulation with prescreening costs $\tilde{E} = \tilde{E}_p + \tilde{E}_s + c_K^2 \sum_{i \in I} \sigma_i^2$. The latter is minimized when c_K^2 is as small as possible, which occurs when $K = 1$, ie, there is only one system left after prescreening. Also, under the assumptions we use in this section, the screening cost E_s is less than the total of the prescreening and screening costs $\tilde{E}_p + \tilde{E}_s$, so the maximal efficiency improvement $E/\tilde{E}$ is achieved when the prescreening and screening costs are negligible compared to estimation costs. This is a typical case in practice: prescreening and screening are very fast compared to estimation and they eliminate all but one system. Under our assumptions, and if prescreening and screening costs are negligible, the efficiency improvement due to restarting (ie, due to having a prescreening phase) is:

$$\frac{E}{\tilde{E}} \approx \frac{c_k^2 \sum_{i \in I} \sigma_i^2}{c_K^2 \sum_{i \in I} \sigma_i^2} = \frac{c_k^2}{c_K^2} \leq \frac{c_k^2}{c_1^2}$$

Figure 4 shows the maximal efficiency improvement c_k^2/c_1^2 as a function of the initial number of systems k. When the number of systems k is between 20 and 1,000, the savings in a typical case are 40–80% at $1 - \alpha = 99\%$ confidence and 60–140% at $1 - \alpha = 95\%$ confidence.

Recall that the transition rule given by inequality (3) chooses to restart when the cost of doing one more stage of prescreening is greater than the approximate maximal savings due to continuation, computed under the assumption that after

additional prescreening there will be only one system left and it will have the largest variance. A typical case indeed has one clear best system, so the effort required for screening out inferior systems is relatively small, the approximate maximal savings are relatively large and prescreening makes I a singleton.

How efficient is this transition rule in other situations? Let us consider a configuration when there are several systems that are tied for the best, while other systems are relatively easy to screen out. In this case we might worry that the cost of prescreening could get too high before the adaptive procedure proceeds to Phase II. Is our transition rule still efficient?

Because now we are concerned that prescreening may be too expensive, we assume that prescreening lasts a long time and eliminates all inferior systems: the set $I(M)$ of systems used in Phase II equals I, the set of systems that survive screening and reach their required sample sizes, and the Phase II cost of screening $\tilde{E}_s = 0$. Again we assume that I is the same whether we use prescreening or not: here we assume it contains only the systems that are tied. We now show how the transition rule in inequality (3) provides a bound on $\tilde{E}_p - E_s$, the excess cost of prescreening in the adaptive procedure over the cost of screening in a procedure without restarting. The effort required to screen out inferior systems is similar in either procedure, so $\tilde{E}_p - E_s \approx KN(M-1)$, the number of samples from the $K = |I|$ surviving systems that the adaptive procedure throws out by restarting.

Prescreening stops after stage $\ell = M-1$, the first time that the cost $(R-1)$ $|I(\ell+1)|N(\ell)$ of the next stage exceeds $(c^2_{|I(\ell+1)|} - c_1^2)\max_{i\in I(\ell+1)} \hat{\sigma}_i^2(\ell)$. Under our present assumption that the residual variance estimates are approximately correct, this yields the approximate upper bound:

$$\begin{aligned} N(M-2) &\le \frac{(c^2_{|I(M-1)|} - c_1^2)\max_{i\in I(M-1)} \sigma_i^2}{(R-1)|I(M-1)|} \\ &\le \frac{(c_K^2 - c_1^2)\max_{i\in I} \sigma_i^2}{(R-1)K} \end{aligned}$$

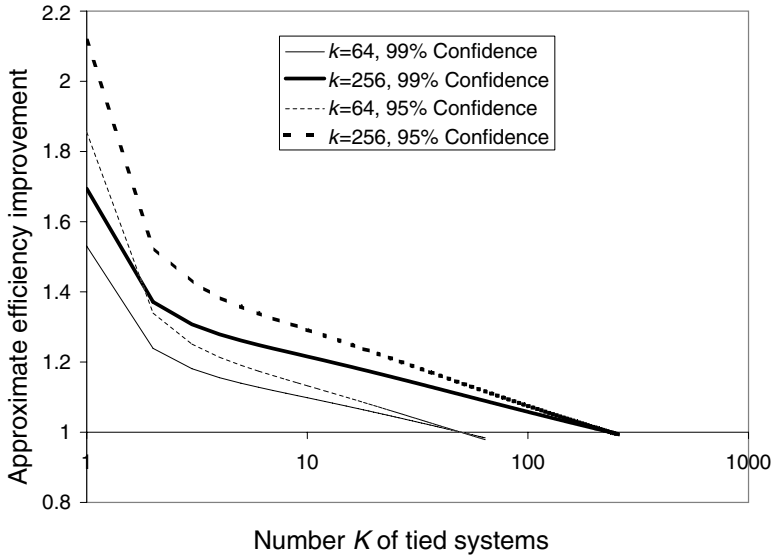
because $I(M-1)$ contains $I(M) = I$ whose size is K, and c_p^2 defined in Equation (2) increases in p at a rate that is less than linear. Thus:

$$\begin{aligned} KN(M-1) &\le KRN(M-2) \\ &\le \frac{R(c_K^2 - c_1^2)\max_{i\in I} \sigma_i^2}{R-1} \end{aligned}$$

For $R = 1.5$, $R/(R-1) = 3$, and the relative efficiency improvement is:

$$\begin{aligned} \frac{E}{\tilde{E}} &= \frac{E_s + c_k^2 \sum_{i\in I} \sigma_i^2}{\tilde{E}_p + \tilde{E}_s + c_K^2 \sum_{i\in I} \sigma_i^2} \\ &= \frac{E_s + c_k^2 \sum_{i\in I} \sigma_i^2}{E_s + 3(c_K^2 - c_1^2)\max_{i\in I} \sigma_i^2 + c_K^2 \sum_{i\in I} \sigma_i^2} \\ &\approx \frac{c_k^2 \sum_{i\in I} \sigma_i^2}{3(c_K^2 - c_1^2)\max_{i\in I} \sigma_i^2 + c_K^2 \sum_{i\in I} \sigma_i^2} \end{aligned}$$

FIGURE 5 Effect of ties on approximate efficiency improvement due to restarting with $\alpha_a = 0.8\alpha$ and $\alpha_b = 0.2\alpha$.



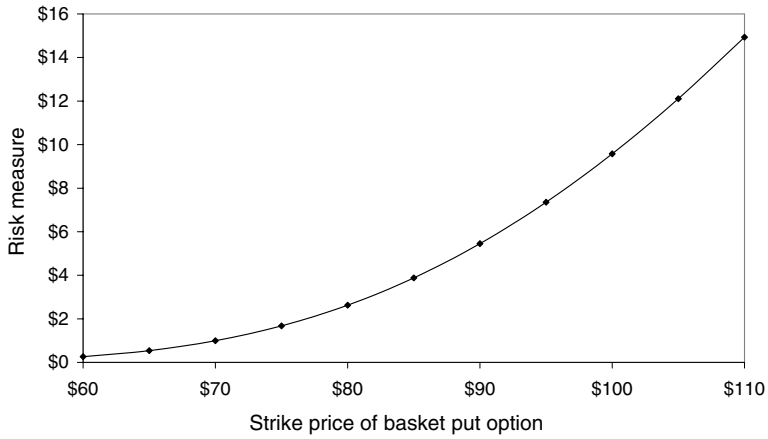
approximately, if the cost E_s of screening is small. If the variances of the tied systems are approximately equal, this simplifies to:

$$\frac{Kc_k^2}{3(c_K^2 - c_1^2) + Kc_K^2}$$

For $k = 256$ and $k = 64$ the efficiency improvements as a function of the number K of tied systems are shown in Figure 5. A value less than 1 represents a loss of efficiency. We see that even when some systems are tied, restarting with our transition rule can still produce substantial benefits. Even when all the systems are tied, the loss of efficiency is very slight.

The transition rule we have presented is heuristic and is one of many similar rules that all work well. This rule is advantageous because of its simplicity and because it allows us to reap most of the benefits of restarting, without causing significant inefficiencies when restarting could be harmful. More efficient transition rules could be designed that take into account not only the sample variances of the systems, but also their sample means. However, such rules are complicated, and in most cases provide either small or no savings. Because the benefits seem insufficient to justify the additional complexity, we do not consider this approach here.

FIGURE 6 Basket put option example: dependence of risk measure (maximum expectation of discounted payout) on strike price.

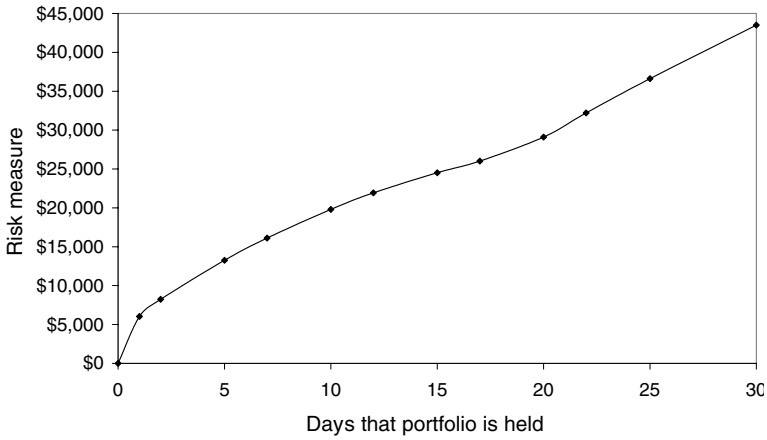


4 PERFORMANCE OF THE ADAPTIVE MULTI-STAGE PROCEDURE

In this section we illustrate our procedure's performance on the basket put and options portfolio examples discussed in Section 2. The risk measure in the basket put example, which can be interpreted as an upper bound for the value of the basket put option (Staum (2004)), is US$3.877. Figure 6 shows how this risk measure depends on the strike price of the basket put option. The confidence interval widths may be compared to the variation in the risk measure depicted by Figure 6. The risk measure in the options portfolio example is US$16,107. Figure 7 shows how it depends on the number T of days until the time horizon at which the potential loss is measured. The curve shows the usual concave behavior for time horizons between zero and 20 days: risk increases as the portfolio is held for a longer period of time, but at a rate that decreases as a function of the time horizon. The change in slope at about $T = 20$ occurs because the greatest conditional expectation of discounted loss is produced by different generalized scenarios for time horizons of less than 20 days than for time horizons of more than 20 days. That is, the generalized scenario most to be feared over short time horizons is different from that most to be feared over longer time horizons.

To test the adaptiveness of the procedure, in addition to the ordinary configuration with one best system, we also consider configurations "2 best" (obtained by adding a duplicate of the best system), "4 best" (by adding three duplicates) and "16 best" (by adding 15 duplicates), so that configuration "2 best" in the basket put example has $64 + 1 = 65$ systems in total, while configuration "16 best" has $64 + 15 = 79$ systems. This is not the same as in Figure 5, where the total number k of systems remains constant while the number K that are tied varies.

FIGURE 7 Options portfolio example: dependence of risk measure (maximum conditional expectation of discounted loss) on number of days elapsed while portfolio is held.



We split the $1-\alpha = 1\%$ allowable error into components $\alpha_a = 0.8\%$ for the lower confidence limit and $\alpha_b = 0.2\%$ for the upper confidence limit. The error allocated to screening is $\alpha_I = 0.04\%$ and, when using control variates, $\alpha_C = 0.002\%$ is allocated to controlling them. We choose the initial sample size n_0 and the maximal number m of Phase I stages to be 30, and the growth factor R to be 1.5. We use common random numbers in all examples.

For ease of interpretation, we specify the fixed confidence interval width L as a percentage of a quantity that provides a natural scale for the example. For the basket put example, this quantity is the true value, the largest mean, which is US$3.877. For the options portfolio example, this quantity is the portfolio's standard deviation, which is US$6,012. Confidence intervals, averaged over many independent runs of the adaptive procedure, are reported in Table 3. The number of independent runs varies over entries in the table: we used between 30 and 400, depending on how many were required to attain the desired statistical accuracy for the averages presented.

In Section 4.1, we show that the adaptive procedure is efficient: it can be hundreds of times faster than the standard procedure and, in many examples, only generates 10–20% more samples than the minimum that could possibly be needed to generate a confidence interval of the required width. We also show that it is adaptive: because the transition rule picks a good time to restart, the adaptive procedure works well on all the problem instances we looked at, whereas each particular parametrization of a procedure from Lesnevski *et al* (2007) works well for some problem instances and not for others. Section 4.2 contains an empirical analysis of the rate at which the confidence interval width decreases as the computational resources required by the adaptive procedure increase. We

TABLE 3 Average 99% confidence intervals (in US dollars) produced by the adaptive procedure.

	Example					
	Options portfolio Precision			Basket put Precision		
Confidence limit	**0.3%**	**1%**	**5%**	**0.3%**	**1%**	**5%**
Upper	16,117	16,140	16,271	3.884	3.900	3.968
Lower	16,099	16,080	15,974	3.872	3.861	3.776

TABLE 4 Efficiency relative to the standard procedure at 99% confidence.

	Example					
	Options portfolio Precision			Basket put Precision		
Configuration	**0.3%**	**1%**	**5%**	**0.3%**	**1%**	**5%**
1 best	**252**	**244**	**154**	**208**	**158**	**22**
2 best	104	98	81	85	76	19
4 best	51	48	43	40	38	15
16 best	12	12	12	11	10	6.7

tentatively find that the adaptive procedure shows the typical Monte Carlo behavior for sufficiently narrow confidence intervals, and that confidence interval width decreases more rapidly for larger widths. In Section 4.3, we show that the efficiency of the adaptive procedure is not impaired even when there are several systems very similar to the best system, a case that might seem to be difficult for the adaptive procedure because of the difficulty of screening out systems that are close to the best.

4.1 Relative efficiency and adaptiveness

We report efficiency as a speed improvement relative to a standard procedure that is a modification of the two-stage procedure of Chen and Dudewicz (1976), as explained in Lesnevski *et al* (2007). That is, we report the ratio of the average number of samples required by the standard procedure to the average number of samples required by the adaptive multi-stage procedure. The results are summarized in Table 4. Recall that efficiency improvement can be larger than the number of systems k, which is 64 for the ordinary configuration of the basket put and 256 for that of the options portfolio. The reason for this is that the improvement depends not only on k, but also on the size of the best system's standard deviation relative to the standard deviations of other systems.

Table 5 shows how much work the procedure does in excess of the work required by the "clairvoyant" procedure, the procedure that knows in advance which systems

TABLE 5 Sample size relative to the clairvoyant procedure at 99% confidence.

	Example					
	Options portfolio Precision			Basket put Precision		
Configuration	0.3%	1%	5%	0.3%	1%	5%
1 best	**1.0**	**1.1**	**1.7**	**1.1**	**1.4**	**10**
2 best	1.2	1.2	1.5	1.2	1.3	5.4
4 best	1.1	1.2	1.3	1.2	1.2	3.2
16 best	1.1	1.1	1.1	1.1	1.1	1.7

are tied for the best, and applies the standard procedure to only these systems in isolation. That is, the clairvoyant procedure screens out all inferior systems by guessing right with no work.

Like the multi-stage procedure with restarting analyzed in Lesnevski *et al* (2007), the adaptive procedure is less than 10% more expensive than estimating a single mean in the "1 best" configuration when a precise estimate is required. If there are ties the procedure first tries to break them, but when this becomes too expensive it proceeds to estimation: this is its advantage over the multi-stage procedure with restarting. Table 5 demonstrates the robustness of the adaptive procedure's performance to configuration.

As we see from the last column of Table 5, in the configuration with no ties at 5% precision the adaptive procedure looks relatively inefficient compared to the clairvoyant procedure (10 times slower), but adding ties can make the adaptive procedure look more favorable. This is because 5% is a low precision, so the final sample size is not very large relative to the sample size required for screening. At 5% precision the clairvoyant procedure has a big advantage in screening perfectly for free.

Table 6 shows the efficiency improvement of the adaptive procedure relative to the most efficient procedure of Lesnevski *et al* (2007): the multi-stage procedure with restarting. (In all cases reported in Table 6, the multi-stage procedure with early stopping was somewhat more expensive than the multi-stage procedure with restarting.) In some cases, the efficiency is slightly less than 1, ie, the adaptive procedure required slightly more samples than the multi-stage procedure with restarting: the adaptive procedure does not always pick the best possible time to restart, but it picks a good time.

The efficiency of both of the procedures depends heavily on the actual configuration of the means and the total screening budget of $n_0 R^{m-1}$ observations per system. We tested these procedures with $n_0 = 30$ and $R = 1.5$ while varying the maximal number of stages available for screening from 5 to 30, so that the total budget available for screening varied from 152 to 3, 835, 022 observations per system. We set $R = 1.5$, not $R = 2$ as in Lesnevski *et al* (2007), as this choice of the growth factor makes all procedures more efficient when there are ties.

TABLE 6 Efficiency relative to the multi-stage procedure with restarting at 99% confidence.

Configuration and precision		Options portfolio Number of screening stages m						Basket put Number of screening stages m					
		5	10	15	20	25	30	5	10	15	20	25	30
1 best	0.3%	1.0	1.0	1.0	1.0	1.0	1.0	41	2.8	1.0	1.0	1.0	1.0
	1%	1.0	1.0	1.0	1.0	1.0	1.0	30	2.3	1.0	1.0	1.0	1.0
	5%	1.0	1.0	1.0	1.0	1.0	1.0	4.7	1.0	0.9	0.9	1.0	1.0
2 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.8	17	1.6	0.9	1.0	1.6	6.1
	1%	0.9	0.9	0.9	1.0	2.0	10	16	1.5	1.0	1.8	7.8	54
	5%	0.9	0.9	1.2	4.0	25	205	4.6	1.0	1.6	6.5	44	328
4 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.8	8.3	1.3	0.9	1.0	1.5	5.8
	1%	0.9	0.9	0.9	1.0	2.1	10	7.8	1.2	1.0	1.8	7.7	53
	5%	0.9	0.9	1.2	4.1	27	197	3.4	1.1	2.0	10	68	509
16 best	0.3%	0.9	0.9	0.9	0.9	1.0	1.7	3.1	1.0	1.0	1.0	1.6	5.7
	1%	0.9	0.9	1.0	1.1	2.0	9.2	3.0	1.1	1.1	1.8	7.7	52
	5%	0.9	1.0	1.2	4.3	27	201	2.1	1.1	2.7	15	110	833

The results in Table 6 illustrate the danger for our previous multi-stage procedures of choosing the budget for screening either too small or too large. What constitutes too small or too large depends on the actual configuration, whereas the adaptive procedure works well in all of them.

4.2 Rates of convergence

Figure 8 shows the average number of simulation observations required by the adaptive procedure for the experiments reported in Table 4. (The average number of observations required by the standard procedure is the product of corresponding data points in Table 4 and Figure 8.) Figure 9 shows a similar plot, but here computational effort is measured in time instead of observations. Both figures are log–log plots on which a typical Monte Carlo procedure's performance yields a straight line of slope $-1/2$, meaning that confidence interval width is proportional to the reciprocal of the square root of the computational resources used. Figure 8 shows approximately this typical behavior for the options portfolio examples: the performance curves are nearly straight and have slopes close to $-1/2$. For the basket put examples with a small number of systems tied for the best, the performance curves are at first steeper and then they have slopes near $-1/2$. This phenomenon has already been discussed while interpreting Table 5. The number of observations used that are then thrown out when restarting is negligible when a very narrow confidence interval is required, but non-negligible when the demand for precision is sufficiently low. Thus we expect that, viewed over a wide enough range of confidence interval widths, the performance curve for the adaptive procedure on any example would be convex, with a slope approaching $-1/2$ from below as more precision is required.

FIGURE 8 A log–log plot of the average number of samples required by the adaptive procedure to attain a fixed confidence interval width.

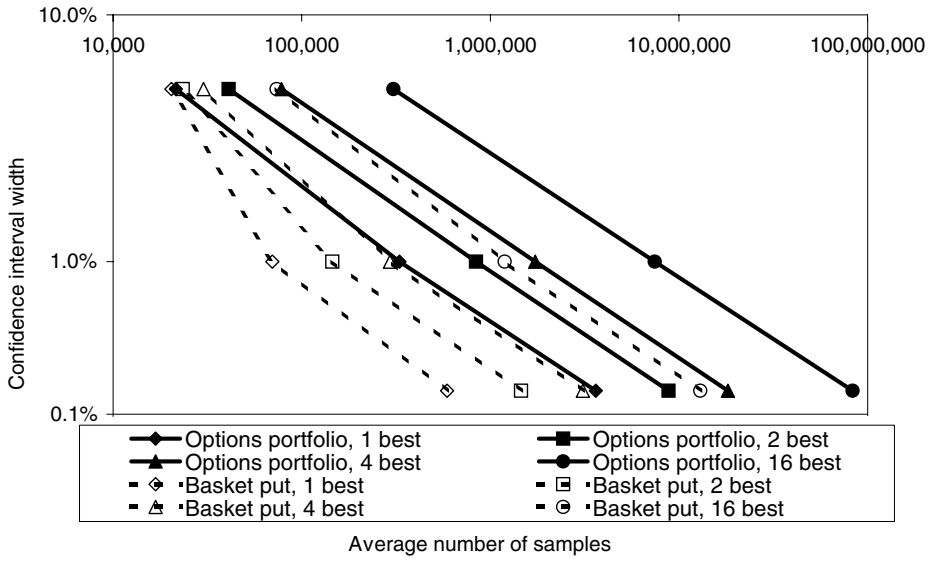
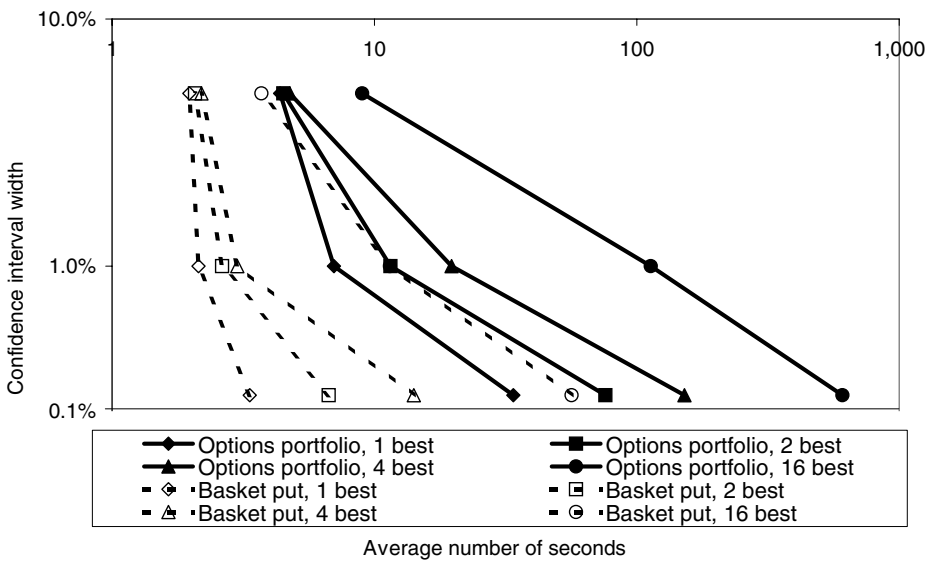


FIGURE 9 A log–log plot of the average number of seconds required by a MATLAB implementation of the adaptive procedure to attain a fixed confidence interval width.



When computational cost is measured in seconds, in Figure 9 the performance curves have a qualitatively similar shape and behavior. The performance curves for examples with a larger number of systems tied for the best are closer to being straight lines, because more of the computation in those examples is performed for systems that are never screened out. In the range of precision investigated, many of the slopes were much steeper than $-1/2$, and none were very close to it. This may have to do with a non-negligible computational cost of screening in these examples. The results presented in Figure 9 are elapsed times of simulation experiments in which the adaptive procedure was run in MATLAB 7.0 under Windows XP on a 3.0 GHz Pentium IV processor with 1 GB of RAM. Caution must be used in interpreting these results: in particular, because of the impact of loops on running times of MATLAB programs, the absolute and relative times reported in Figure 9 are not necessarily indicative of results that would be obtained with an implementation of the adaptive procedure in another language, such as C.

This discussion of convergence does not address how the procedures' computational requirements grow as the number k of generalized scenarios grows. It would be difficult to address this topic briefly because the answer depends on how the set $\mathcal{P}$ of generalized scenarios is enlarged: for example, adding generalized scenarios that are much worse than the best and have low variance causes a slight increase in the number of samples that the adaptive procedure requires, while a large increase results from adding generalized scenarios that are nearly as good as the best, independent of it, and have high variance.

4.3 Similar systems nearly tied for the best

When simulating a coherent risk measure with common random numbers, several highly correlated systems may have nearly the largest mean. Such situations can occur when one or several factors that are usually important in computation of a risk measure turn out to be insignificant in a particular instance, or when parameters differ only slightly for some systems. For example, an equity derivative may have very similar values in generalized scenarios that differ only in interest rates. One might worry that simulation in this case is expensive and relatively inefficient, similar to what we see in Table 4.

However, even if the variances of the systems are large, the variances of the differences of the means of such systems will tend to be small. Unless some systems are identical, which is easy to recognize when carrying out simulations with common random numbers and in which case the duplicates should be taken out, small variances of the differences allow even very small differences in performance to be quickly detected, and even slightly inferior systems will be screened out relatively quickly.

For example, in the case of the basket put, the best system is the one that has a pairwise correlation of 0.75 between the assets. Table 7 shows the effect of adding a system that has a pairwise correlation of 0.74 between assets (configuration "2 similar"), adding three systems that have two out of three pairwise correlations of

TABLE 7 Increase in average sample size due to adding systems similar to the best at 99% confidence.

	Example					
	Options portfolio Precision			Basket put Precision		
Configuration	0.3%	1%	5%	0.3%	1%	5%
2 similar	<1%	<1%	<1%	<1%	<1%	1%
4 similar	<1%	<1%	1%	<1%	2%	7%
16 similar	<1%	<1%	7%	1%	7%	28%

0.74 and one pairwise correlation of 0.75 ("4 similar") and adding 15 similar systems that have pairwise correlations of 0.75, 0.74 and 0.73 in various combinations ("16 similar"). In the case of the options portfolio, the best system (scenario) is the one that has the first and the fourth factors "up", while the other two factors are unrestricted. We can add a similar system by assigning to one of the "up" events a probability of $9/(10\sqrt{20})$ (in place of $1/\sqrt{20}$ in the best system) and the other a probability of $10/(9\sqrt{20})$ (configuration "2 similar"). In configuration "4 similar" we add two more systems by assigning to one of the "up" events a probability of $99/(100\sqrt{20})$ and to the other a probability of $98/(100\sqrt{20})$, while in configuration "16 similar" we add 12 more similar systems of this form.

From Table 7 we see that the increase in the average sample size due to adding similar systems is usually small. It is also not very sensitive to the similarity parameter, such as the pairwise correlation in the basket put example: in configuration "2 similar" it stays roughly the same whether we use a correlation of 0.73 or 0.7499. Even though the two systems have almost exactly the same mean, the variance of the difference is so small that it is easily detected with common random numbers. The correlation between the best system and the similar system that we have added in configuration "2 similar" is 99.99% in the basket put example and 99.83% in the options portfolio example. Adding more such systems does not increase the sample size by much, as the correlation is so high that the procedure will quickly screen out systems with smaller means. This increase is mostly due to the larger number of systems that need to be screened out, while the total sample size per system stays roughly the same.

This allows us to conclude that efficiency loss due to closeness of the best means should not in general be significant in financial applications, and that in most cases we will have a clear best.

5 ROBUSTNESS TO NON-NORMALITY

Under normal-theory assumptions, our procedures are exact, ie, they deliver at least the nominal coverage probability. Although these assumptions are reasonable in many situations, they are usually not precisely correct. Our screening procedures

TABLE 8 Effect of strike price and initial sample size n_0 on error rates at 90% confidence and 5% precision in the basket put example.

Strike price (zero payout probability)	n_0	Error rate		
		Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
$K = 85$	5	16%	4%	0%
($\approx$71%)	7	7%	2%	0%
	10	5%	1%	0%
$K = 65$	30	7%	6%	$\ll$1%
($\approx$92%)	50	5%	5%	0%
	100	4%	5%	0%
$K = 55$	30	42%	7%	3%
($\approx$98%)	100	7%	5%	0%
	300	4%	5%	0%

use sample averages when the sample sizes are still small, and since the distributions of sample averages might be very far from normal, one might worry that screening errors might occur much more often than if distributions were normal.

It is comforting to know that the screening procedures are protected by the use of very conservative probability inequalities (such as the Bonferroni inequality) in their derivation. Error is allocated to pairwise comparisons between all systems during the maximal possible number of stages, but many of these comparisons are never performed. Because of this, we can expect screening to be very robust to non-normality. In fact, in most of our experiments, all of which included 5,000 independent replications, screening errors never occurred.

On the other hand, our estimation procedure will typically require large sample sizes. As we become more demanding, requiring a smaller confidence interval width or higher confidence, the final sample size becomes larger, making normality of mean estimators more plausible. For this reason moderate non-normality does not seem to be a problem for the final estimator.

However, if non-normality is extreme and the initial sample size is not adequate, the sample sizes after Phase I might be too small and the estimates of the variances that are used to compute final sample sizes could have a distribution that is far from (scaled) χ^2.

Let us consider the basket put example (see Table 8). In the ordinary configuration the strike price is 85 and the probability of a zero payout is approximately 71%. If the probability of a zero payout is 98% and n_0 is smaller than 200–300, estimates of the variances are so poor that coverage is inadequate. When the probability of a zero payout is 92%, this can happen if n_0 is smaller than 50–100. When non-normality is not so extreme, such as in the case of when the probability of a zero payout is 90% or less, coverage is adequate as long as n_0 is larger than 10–20.

Importance sampling might be a way to improve the coverage when the payout is highly non-normal. Importance sampling is usually used as a variance reduction

TABLE 9 Effect of initial sample size n_0 on error rates at 90% confidence and 5% precision in the options portfolio example.

	Error rate		
n_0	Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
5	4%	5%	0%
10	4%	5%	0%
30	4%	5%	0%

TABLE 10 Error rates with log-t returns in the basket put example at 90% confidence and 5% precision ($n_0 = 30$).

Error rate		
Upper (5% nominal)	Lower (5% nominal)	Screening (1% nominal)
4%	5%	0%

technique, to make the variance of the product of likelihood ratio and payout under a new probability measure lower than the variance of the payout under the original probability measure (see Glasserman (2004, Section 4.6)). Here importance sampling could also be used to make the distribution of the product of likelihood ratio and payout under a new probability measure closer to normal. For example, a standard form of importance sampling applied to the out-of-the-money basket put changes the mean asset returns in a way that decreases the probability of a zero payout and could thus reduce skewness and produce a more normal distribution.

In the options portfolio example non-normality is not very significant, so the coverage is adequate even when n_0 is very small (see Table 9).

The coverage is also adequate when distributions are heavy-tailed. For example, if in the basket put example logarithmic returns are not normal, but rather have the t distribution with three degrees of freedom, the coverage is adequate (see Table 10).

For our experiments in this section we chose relatively low 5% precision and 90% confidence. Because in this case the total sample sizes are smaller and therefore the sample averages are less normal, this should represent the hardest test for our procedure.

6 EMPIRICAL ANALYSIS OF RARE ERRORS

In this section we analyze the event of probability at most $1 - \alpha_a - \alpha_b$, in which the confidence interval does not contain the true value. Because screening is so conservative and screening errors are so extremely rare, the error event consists primarily of estimation errors.

In Table 11 we present the relative root-mean-squared distances from the true largest mean to the nearest confidence limit: to the upper limit when the true value

TABLE 11 Root-mean-squared distance from true value to confidence interval as a percentage of its width and error rates at 90% confidence.

	Options portfolio Precision			Basket put Precision		
	0.3%	**1%**	**5%**	**0.3%**	**1%**	**5%**
$n_0 = 30$						
Upper distance	17	17	18	16	16	13
Lower distance	18	18	18	18	16	14
Upper error (5% nominal)	4%	4%	4%	4%	4%	5%
Lower error (5% nominal)	5%	5%	5%	5%	5%	1%
$n_0 = 10$						
Upper distance	17	16	17	900	92	91
Lower distance	17	17	17	55	64	21
Upper error (5% nominal)	4%	4%	4%	4%	4%	6%
Lower error (5% nominal)	5%	5%	5%	5%	5%	1%

lies above the confidence interval and to the lower limit when the true value lies below the confidence interval, as a percentage of its width. These are conditional on the error event, ie, they are distances given that the true value is above or below the confidence interval:

$$\frac{1}{L}\sqrt{\mathrm{E}\Big[\Big(\mu_k - \Big(\max_{i\in I}\widehat{\mu}_i + b\Big)\Big)^2 \mid \mu_k > \max_{i\in I}\widehat{\mu}_i + b\Big]}$$

for the upper distance and:

$$\frac{1}{L}\sqrt{\mathrm{E}\Big[\Big(\Big(\max_{i\in I}\widehat{\mu}_i - a\Big) - \mu_k\Big)^2 \mid \mu_k < \max_{i\in I}\widehat{\mu}_i - a\Big]}$$

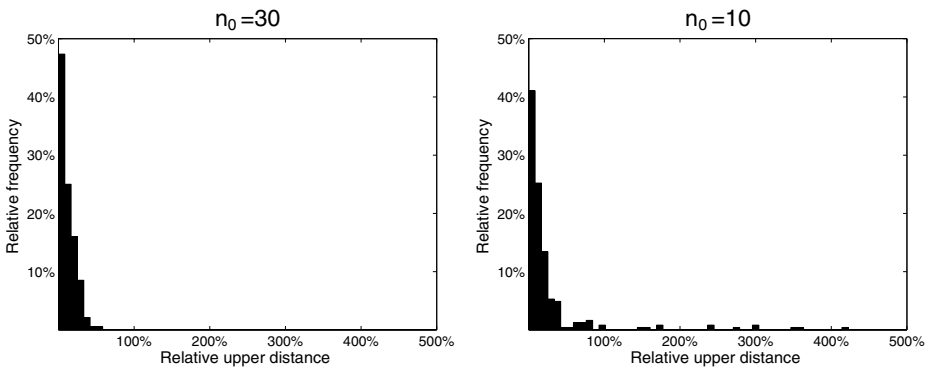
for the lower distance. If we were estimating the mean of just one system in isolation, which has the same mean and variance as the best system and which is normally distributed, we would have a relative root-mean-squared distance of approximately 17% for both the upper and the lower confidence limits at 90% confidence:

$$0.17 \approx \frac{1}{2z_{.95}}\sqrt{\int_{z_{.95}}^{\infty}(x - z_{.95})^2\frac{\phi(x)}{0.05}\,\mathrm{d}x}$$

where ϕ is the probability density function and $z_{.95} = \Phi^{-1}(0.95)$ is the 95%-quantile of the standard normal distribution. Table 11 shows that when non-normality of sample averages is not extreme the errors of the adaptive procedure on average are no more severe than the errors that happen when estimating a mean of a normal population.

However, when non-normality is extreme and n_0 has not been chosen adequately large, the estimation errors can be much more severe. For example, when using

FIGURE 10 Distances from the upper confidence limit at 90% confidence and 1% precision as percentages of confidence interval width, for initial sample size $n_0 = 30$ and $n_0 = 10$.



$n_0 = 10$ in the case of the basket put, we found that the coverage was adequate, but the root-mean-squared distance from the upper confidence limit was approximately equal to the confidence interval width when using 1% precision, and it was about nine times that width when using 0.3% precision. (Recall that the confidence interval width is proportional to the precision.) Because non-coverage is a rare event, these large root-mean-squared distance estimates are not very precise, even though we used more than 5,000 replications to estimate them. This indicates that when non-normality is extreme the procedure might significantly under- or overestimate the risk measure: see Figure 10, representing the non-coverage events in a representative batch of 5,000 replications.

7 CONCLUSIONS

The adaptive procedure proposed here generates a two-sided, fixed-width confidence interval for a coherent risk measure based on a finite number of generalized scenarios.

Under normal-theory assumptions, the coverage of the confidence interval can be guaranteed. Unless non-normality is extreme and the first-stage sample size n_0 is too small, the procedure is very robust: the coverage meets or exceeds the nominal level, and even when the confidence interval does not contain the true value the errors are usually not severe. In extreme cases we have to make sure that n_0 is large enough to get variance estimates with the right statistical properties. Generally $n_0 = 30$ should be sufficient, but in some cases a preliminary assessment of normality of sample averages might be necessary in order to pick an appropriate initial sample size. In financial applications, similar simulation problems tend to arise from day-to-day (eg, risk measurement of a portfolio when the portfolio's composition and the market environment have changed only slightly since yesterday), in which case

extra simulation effort would not be required to determine an appropriate value of n_0. The problems due to non-normality can also be mitigated by importance sampling. It might also be possible to improve robustness to non-normality by changing the screening procedure. We have carried out screening on the basis of a two-sample t test with paired observations. Screening could also be carried out based on a non-parametric test. One relevant approach is the nonparametric subset selection procedure of Hsu (1980).

The adaptive procedure improves upon previous procedures, which might decrease efficiency when applied to the wrong problem or with the wrong parameters, by being more reliably efficient and easier to use. One might fear that the time required to estimate a coherent risk measure based on k generalized scenarios would be of the order of k times as long as the time required to estimate a single mean, and this is true for standard methods. The adaptive procedure is typically dozens to hundreds of times faster than standard methods for $k = 64$ or $k = 256$, even for challenging problems in which some of the generalized scenario means are very similar or even tied. Even for these challenging problems, the adaptive procedure was only a few percent to 40% more expensive than estimating a single mean for the examples we investigated of generating a moderately precise 99% confidence interval for a coherent risk measure (Tables 5 and 7, precision 1% or less).

It seems difficult, then, to construct a procedure that is much more efficient for these problems while providing a valid confidence interval. However, it may be possible to improve efficiency for problems where a wide confidence interval is acceptable. We tried using fully sequential screening instead of multi-stage screening, but found that this seldom improved efficiency. Another possibility is to avoid using the conservative probability inequalities we used to guarantee the confidence interval's coverage probability. Pursuing this line of thought, one might also develop a procedure that provides only a point estimate for the risk measure and not a confidence interval: it may be possible to get a point estimator of comparable quality faster than our confidence interval by screening more aggressively than our probability inequalities permit or by avoiding restarting. Furthermore, for applications that require a point estimator, it could be valuable to apply bias-reduction techniques to get a point estimator superior to the straightforward one, which is the maximum surviving sample average $\max_{i \in I} \widehat{\mu}_i$ around which our confidence interval is constructed.

APPENDIX A VALIDITY OF THE PROCEDURE

While I is the set of systems that survives screening after Phase II, let $[k]$ be the index of the best system, the one with the largest mean. Let $I(M)$ be the set of systems that survives screening in Phase I, and let $[k]_M$ be the best system in $I(M)$.

PROPOSITION A.1 *If, for each* $i = 1, 2, \ldots, k$, $X_{ij} = \mu_i + (C_{ij} - \xi_i)'\beta_i + \eta_{ij}$, *where the residuals* $\{\eta_{ij},\ j = 1, 2, \ldots\}$ *and controls* $\{C_{ij},\ j = 1, 2, \ldots\}$ *are independent sets of independently and identically distributed normal random variables,* β_i *is an unknown constant vector,* $E[C_{i1}] = \xi_i$ *and* $E[\eta_{i1}] = 0$, *then the procedure*

satisfies:

$$\Pr\left\{\mu_{[k]} \geq \max_{i\in I} \widehat{\mu}_i - a\right\} \geq 1-\alpha_a \tag{A.1}$$

and:

$$\Pr\left\{\mu_{[k]} \leq \max_{i\in I} \widehat{\mu}_i + b\right\} \geq 1-\alpha_b \tag{A.2}$$

PROOF We decompose the screening error α_I in the following way. Allocate $\alpha_I/2$ to Phase I and $\alpha_I/2$ to Phase II.

Phase I has at most m stages and there are at most k systems during any stage, so there are at most $m(k-1)$ comparisons with system $[k]$ during screening in Phase I. Therefore, during Phase I we use screening thresholds:

$$W_{hi}(\ell) = \frac{S_{hi}(\ell)t_{N(\ell)-1,1-\alpha_I/(2m(k-1))}}{\sqrt{N(\ell)}}$$

at stage ℓ for differences of sample averages of observations generated during stages 1 to ℓ. Phase II has at most P screening stages and there are at most K systems during any stage, so there are at most $P(K-1)$ comparisons with system $[k]_M$ during screening. Although P and K are random, they are determined before Phase II begins by data from Phase I, which is not used for estimation in Phase II, so this randomness does not pose a difficulty. Therefore, during Phase II we use thresholds:

$$W_{hi}(\ell) = \frac{S_{hi}(\ell)t_{N(\ell)-N(M-1)-1,1-\alpha_I/(2P(K-1))}}{\sqrt{N(\ell)-N(M-1)}}$$

at stage ℓ for differences of sample averages generated during stages M to ℓ. By the Bonferroni inequality, $\Pr[[k] \notin I(M)] \leq \alpha_I/2$ and $\Pr[[k]_M \notin I] \leq \alpha_I/2$.

Applying Proposition 4 of Nelson and Staum (2006) to the randomly generated problem of estimating the value of the best system in $I(M)$ shows that:

$$\Pr\{\widehat{\mu}_i - \mu_i > x\} \leq 1 - G_a(cx) \quad \text{and} \quad \Pr\{\widehat{\mu}_i - \mu_i < -x\} \leq 1 - G_b(cx)$$

holds with $G_a(x) = G_b(x) = F_{t_{N(M)-N(M-1)-q-1}}(x) - \alpha_C$ for each $i \in I(M)$. This statement is true even for a system $i \in I(M) \setminus I$ that is screened out during Phase II: although the procedure does not compute the estimate $\widehat{\mu}_i$, this random variable exists on the probability space under consideration and satisfies these inequalities. Proposition 3.1 of Lesnevski *et al* (2007) then implies that:

$$\Pr\left\{\max_{i\in I(M)} \mu_i \geq \max_{i\in I} \widehat{\mu}_i - a\right\} \geq 1-\alpha_a$$

and:

$$\Pr\left\{\max_{i\in I(M)} \mu_i \leq \max_{i\in I} \widehat{\mu}_i + b\right\} \geq 1-\alpha_b+\alpha_I/2$$

Consider the lower confidence limit and notice that $\mu_{[k]} = \max_{i=1,2,\ldots,k} \mu_i \geq \max_{i\in I(M)} \mu_i$, whatever the subset $I(M) \subseteq \{1, 2, \ldots, k\}$ generated after Phase I may be. Consequently:

$$\Pr\left\{\mu_{[k]} \geq \max_{i\in I} \widehat{\mu}_i - a\right\} \geq \Pr\left\{\max_{i\in I(M)} \mu_i \geq \max_{i\in I} \widehat{\mu}_i - a\right\} \geq 1-\alpha_a$$

which verifies inequality (A.1). Next consider the upper confidence limit and notice that if $[k] \in I(M)$, then $\mu_{[k]} = \max_{i \in I(M)} \mu_i$. Consequently:

$$\begin{aligned}
\Pr\left\{\mu_{[k]} \le \max_{i \in I} \widehat{\mu}_i + b\right\} &\ge \Pr\left\{[k] \in I(M),\ \mu_{[k]} \le \max_{i \in I} \widehat{\mu}_i + b\right\} \\
&= \Pr\left\{[k] \in I(M),\ \max_{i \in I(M)} \mu_i \le \max_{i \in I} \widehat{\mu}_i + b\right\} \\
&\ge 1 - \Pr\{[k] \notin I(M)\} - \Pr\left\{\max_{i \in I(M)} \mu_i > \max_{i \in I} \widehat{\mu}_i + b\right\} \\
&\ge 1 - \alpha_I/2 - (\alpha_b - \alpha_I/2) = 1 - \alpha_b
\end{aligned}$$

which verifies inequality (A.2). □

REFERENCES

Acerbi, C., and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance* **26**, 1487–1503.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* **9**, 203–228.

Cont, R., and Tankov, P. (2004). *Financial Modelling with Jump Processes (Financial Mathematics Series)*. Chapman & Hall/CRC, London.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer, New York.

Hsu, J. C. (1980). Robust and nonparametric subset selection procedures. *Communications in Statistics—Theory and Methods* **9**(14), 1439–1459.

Jaschke, S., and Küchler, U. (2001). Coherent risk measures and good-deal bounds. *Finance and Stochastics* **5**, 181–200.

Law, A. M., and Kelton, W. D. (2000). *Simulation Modeling and Analysis*, 3rd edn. McGraw-Hill, New York.

Lesnevski, V., Nelson, B. L., and Staum, J. (2007). Simulation of coherent risk measures based on generalized scenarios. *Management Science* **53**(11), 1756–1769.

Nelson, B. L., and Staum, J. (2006). Control variates for screening, selection, and estimation of the best. *ACM Transactions on Modeling and Computer Simulation* **16**(1), 1–24.

Shiryaev, A. N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory (Advanced Series on Statistical Science & Applied Probability, No. 3)*. World Scientific, Singapore.

Staum, J. (2004). Fundamental theorems of asset pricing for good deal bounds. *Mathematical Finance* **14**(2), 141–161.