Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

A SIMULATION ANALYTICS APPROACH TO DYNAMIC RISK MONITORING

Guangxin Jiang

Department of Economics and Finance
City University of Hong Kong
Kowloon, HONG KONG

L. Jeff Hong

Department of Economics and Finance and
Department of Management Sciences
City University of Hong Kong
Kowloon, HONG KONG

Barry L. Nelson

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208, USA

ABSTRACT

Simulation has been widely used as a tool to estimate risk measures of financial portfolios. However, the sample paths generated in the simulation study are often discarded after the estimate of the risk measure is obtained. In this article, we suggest to store the simulation data and propose a logistic regression based approach to mining them. We show that, at any time and conditioning on the market conditions at the time, we can quickly estimate the portfolio risk measures and classify the portfolio into either low risk or high risk categories. We call this problem dynamic risk monitoring. We study the properties of our estimators and classifiers, and demonstrate the effectiveness of our approach through numerical studies.

1 INTRODUCTION

In financial risk management, Monte Carlo simulation has been widely used as a tool to estimate risk measures of financial portfolios, e.g, exceeding probability, Value-at-Risk (VaR), and conditional VaR, etc. Many studies focus on how to estimate these risk measures accurately, see Hong et al. (2014) for a recent survey. For example, Glasserman et al. (2000) and Glasserman et al. (2002) studied how to estimate portfolio VaRs, and applied variance reduction techniques to improve the estimation accuracy. These approaches often require that the loss of the portfolio may be calculated easily through closed-form expressions or delta-gamma approximations. If portfolios include derivative products, whose prices need to be determined by additional simulation experiments, it is typically more challenging computationally. This type of problem is known as nested estimation problem. Lee and Glynn (2003) studied the general situations of nested simulation, and Gordy and Juneja (2010) applied it to portfolio risk measures estimation. Liu and Staum (2010) and Broadie et al. (2011) proposed using stochastic kriging and adaptive method to improve the estimation efficiency, respectively. To reduce simulation effort in the inner level, Broadie et al. (2015) proposed a regression method and Hong et al. (2015) proposed a kernel method to avoid nested simulations.

Almost all the risk measures estimation approaches are “one-off computation”, which means that they are interested only in estimating a risk measure at the current time point given the current values of all underlying risk factors. All simulated data are discarded after the estimation. Moreover, if the values of the underlying risk factors change, the risk measure needs to be estimated again. This introduces a heavy computational burden to risk management, which often requires the risk measure be re-evaluated frequently based on the values of the underlying risk factors. To solve this problem, we suggest to take a "simulation analytics" approach. The approach was proposed by Nelson (2016) and it uses big data analytics methods to explore the data generated by simulation studies and to uncover the conditional relationships hidden in simulation models.

In particular, we propose to use a logistic regression based approach to analyze the sample paths of the simulation study, and to develop a model of the exceeding probability and the values of the underlying risk factors at any time. We can then use the model to predict the risk of the portfolio at any time after observing the values of the underlying risk factors. The idea of reusing simulation data can be found in literatures of simulation on demand (Liu et al. 2010) and green simulation (Feng and Staum 2015). But distinct from those approaches, we apply data analytics and machine learning tools to mine the simulation data.

The rest of this article is organized as follows. In Section 2, we formulate the problem as estimating conditional probabilities and develop a logistic regression framework. In Section 3, We propose the maximum likelihood method to estimate the logistic regression model, and provide theoretical properties of the estimator including consistency, asymptotic normality, and large derivation. Section 4 presents numerical results. Section 5 concludes.

2 PROBLEM STATEMENT

Suppose that $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_m(t))'$ is a vector of the underlying risk factors, which may include prices of stocks and bonds, stochastic interest rates, etc. Let $\mathbf{S}(t)$ be a Markov process defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with a natural filtration $\mathscr{F}_t$ that governs the evolution of the process. Consider a portfolio with k financial products, e.g., stocks, bonds, derivatives, whose values at time t are denoted by $V_i(t), i = 1, \ldots, k$, which depend on the realizations of the underlying risk factors $\mathbf{S}(t)$. For the convenience of the notation, let $\mathbf{V}(t) = (V_1(t), \ldots, V_k(t))'$. Furthermore, suppose that the positions on the financial products are $\mathbf{w} = (w_1, \ldots, w_k)'$. Then the value of the portfolio at time t is

$$\Phi(t) = \sum_{i=1}^{k} w_i V_i(t) = \mathbf{w}^T \mathbf{V}(t). \tag{1}$$

Let $L(t) = \Phi(0) - \Phi(t)$. Then $L(t)$ is the loss of the portfolio at time t.

Suppose that there is a fixed future time T at which portfolio loss $L(T)$ needs to be evaluated and further actions need to be taken based on $L(T)$. For instance, T may be the end of the fiscal year at which portfolio loss needs to be reported to shareholders or T is the end of the investment cycle at which bonuses are distributed. Then, the managers of the portfolio may be interested in estimating the loss distribution and the portfolio risk measures at time T. In this article, we consider the estimation of the exceeding probability, i.e., $\Pr\{L(T) > y\}$ for some important threshold value y. Notice that, if $\Pr\{L(T) > y\}$ may be estimated for any y, we may use it to obtain other risk measures, such as VaRs and conditional VaRs (Glasserman et al. 2000). In the simulation literature, much has been done in estimating the unconditional probability, i.e., $\Pr\{L(T) > y\}$, but we are interested in estimating the conditional probability $\Pr\{L(T) > y | \mathscr{F}_u\}$, which denotes the exceeding probability given the information up to time $u \in [0, T]$. By the Markov property of the underlying risk factors $\mathbf{S}(t)$, we have $\Pr\{L(T) > y | \mathscr{F}_u\} = \Pr\{L(T) > y | \mathbf{S}(u)\}$. The conditional probability is useful because at any time $u \in [0, T]$, given the realization of $\mathbf{S}(u)$ observed in practice, we can tell the probability that the portfolio loss exceeds the threshold y at the important future time T and we may use the probability as a risk monitoring tool to determine whether the portfolio risk is under control.

2.1 Nested Simulation and the Data

When the portfolio is formed at time 0, a thorough simulation study is typically conducted to analyze the risk profile of the portfolio and report the risk measures to relevant shareholders. In this article, we suppose that the unconditional exceeding probability $\Pr\{L(T) > y\}$ is estimated through a nested Monte Carlo simulation study. For this simulation study, one often has more time available to run simulation experiments, and the risk measures are often estimated accurately. Then, after the simulation study, we have n simulated sample paths of the underlying risk factors, denoted by $\mathbf{S}_1(t), \mathbf{S}_2(t), \ldots, \mathbf{S}_n(t)$ for $0 \le t \le T$. These sample paths are often simulated under the real probability measure and they are the output of the outer level simulation in a nested simulation study. Moreover, we also have the values of the financial products at time T evaluated based on each simulated realization of the underlying risk factors. We denote them as $\mathbf{V}_1(T), \mathbf{V}_2(T), \ldots, \mathbf{V}_n(T)$. Notice that these financial products may include complicated financial derivatives whose values do not have closed-form expressions. Then, an inner level simulation study under the risk neutral probability measure may need to be used to estimate the values. In this article, we assume that these values can be estimated accurately without estimation error. Then, given the weights $\mathbf{w}$ of the portfolio, we can easily calculate the portfolio loss at time T based on the simulated realizations of the underlyings, and we denote them as $L_1(T), L_2(T), \ldots, L_n(T)$.

2.2 Dynamic Risk Monitoring

Once the portfolio is constructed, the portfolio managers need to constantly monitor the risk of the portfolio. For instance, they may need to estimate the exceeding probability at any real time (instead of the simulated time) given the market conditions, i.e., the realization of the underlyings $\mathbf{S}(u)$, and decide whether the portfolio is safe or not. We call the first problem the "dynamic risk estimation problem" and the second problem the "dynamic risk classification problem". For both problems, we want to use the simulated data in Section 2.1 to avoid additional simulation efforts, so that both problems may be solved quickly to meet the practical requirements.

In the dynamic risk estimation problem, our goal is to estimate

$$p_u(\mathbf{S}(u)) = \Pr\{L(T) \ge y | \mathscr{F}_u\} = \Pr\{L(T) \ge y | \mathbf{S}(u)\} \tag{2}$$

for any $u \in [0,T]$. Notice that $p_u(\mathbf{S}(u))$ is a function of $\mathbf{S}(u)$. Therefore, our goal is to estimate a function, which is often called a regression problem in the field of machine learning (Hastie et al. 2009). If the function is estimated, we may then plug in $\mathbf{S}(u)$ observed at real time t to estimate the exceeding probability $p_u(\mathbf{S}(u))$. Notice that this may be done very quickly if the function has been estimated before hand.

In the dynamic risk classification problem, our goal is to classify dynamically the portfolio risk into two categories, safe and dangerous, based on the exceeding probability $p_u(\mathbf{S}(u))$. For instance, we may set $\alpha \in (0,1)$ as a threshold and classify the portfolio risk is dangerous if $p_u(\mathbf{S}(u)) > \alpha$ and safe otherwise. In practice, for instance, we may set $\alpha = 0.05$ or 0.1. The risk classification allows the portfolio managers to know immediately whether actions need to be taken to control the portfolio risk. One may further extend the number of categories from two to a higher number in the risk classification problem. This may lead to a risk ratings that resemble to the credit ratings, e.g., the AAA to D levels, used by credit rating agencies such as Standard & Poor's and Moody's. Notice that, once the function $p_u(\mathbf{S}(u))$ is estimated, the classification can also be solved immediately given the values of $\mathbf{S}(u)$.

3 LOGISTIC REGRESSION

As stated in Section 2, our goal is to estimate $p_u(\mathbf{S}(u)) = \Pr\{L(T) \ge y | \mathbf{S}(u)\}$ and use it for classification for any $u \in [0,T]$. To do that, notice that we have the simulated sample paths $\{\mathbf{S}_i(t), 0 \le t \le T\}$, $i = 1, \ldots, n$, and the corresponding portfolio loss $L_i(T)$ for each sample path i. To simplify the notation, we let $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n)$

denote $(\mathbf{S}_1(u),\mathbf{S}_2(u),\ldots,\mathbf{S}_n(u))$. Let

$$Y = \begin{cases} 1 & \text{if } L(T) \geq y, \\ 0 & \text{otherwise.} \end{cases}$$

We have $p_u(\mathbf{S}) = \mathrm{E}[Y|\mathbf{S}]$. Based on the simulated sample paths, we have n observations of Y, denoted by $Y_1,Y_2,\ldots,Y_n$. Then, to estimate the regression function $p_u(\mathbf{S}(u))$, we have n observations of the input-output pair, denoted by $\{(\mathbf{S}_1,Y_1),(\mathbf{S}_2,Y_2),\ldots,(\mathbf{S}_n,Y_n)\}$, where the inputs are also called predictors, input variables or features, and the output are also called dependent variables or responses or classes or labels in the areas of statistics and machining learning.

3.1 Logistic Regression Model and Maximum Likelihood Estimation

Because the response Y is a Bernoulli random variable, to model $p_u(\mathbf{S}(u))$, a natural choice is a logistic regression model (Hosmer Jr. and Lemeshow 2004). Let $\mathbf{X}(\cdot):\Re^m \to \Re^d$ denote a set of basis functions computed from $\mathbf{S}(u)$, and these basis functions may be the polynomials of $\mathbf{S}(u)$ or other transform like $\log(\mathbf{S}(u))$ or even the derived random variables of $\mathbf{S}(u)$. We then propose the following logistic regression model

$$\log\left(\frac{p_u(\mathbf{S}(u))}{1-p_u(\mathbf{S}(u))}\right) = \boldsymbol{\beta}^T\mathbf{X}(\mathbf{S}(u)), \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)'$ is the vector of coefficients. It is worthwhile noting that both $\mathbf{X}(\cdot)$ and $\boldsymbol{\beta}$ may depend on the time u. Therefore, for different time points $u \in (0,T)$, we may use different basis functions and obtain different coefficients.

The parameters of the logistic regression model Equation (3) are typically estimated by the maximum likelihood (ML) method. We let $\mathbf{X}(u)$ denote $\mathbf{X}(\mathbf{S}(u))$, and for fixed u

$$g(\mathbf{X}(u),\boldsymbol{\beta}(u)) = \exp(\boldsymbol{\beta}(u)^T\mathbf{X}(u))/(1+\exp(\boldsymbol{\beta}(u)^T\mathbf{X}(u))).$$

Then, the log likelihood function is

$$\log\ell(\boldsymbol{\beta}(u)|\mathbf{X}(u),Y) = Y\log\left(g(\mathbf{X}(u),\boldsymbol{\beta}(u))\right) + (1-Y)\log\left(1-g(\mathbf{X}(u),\boldsymbol{\beta}(u))\right). \tag{4}$$

Given the training data $\{(\mathbf{X}_i(u),Y_i), i=1,\ldots,n\}$, where $\mathbf{X}_i(u) = \mathbf{X}(\mathbf{S}_i(u))$,

$$\begin{aligned} L_n(\boldsymbol{\beta}(u)) &= \frac{1}{n}\sum_{i=1}^{n}\{Y_i\log\left(g(\mathbf{X}_i(u),\boldsymbol{\beta}(u))\right) + (1-Y_i)\log\left(1-g(\mathbf{X}_i(u),\boldsymbol{\beta}(u))\right)\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\boldsymbol{\beta}(u)^T\mathbf{X}_i(u) - \log\left(1+\exp(\boldsymbol{\beta}(u)^T\mathbf{X}_i(u))\right)\right\}, \end{aligned} \tag{5}$$

and the ML estimator $\hat{\boldsymbol{\beta}}_n(u)$ is given by

$$\hat{\boldsymbol{\beta}}_n(u) = \arg\max_{\boldsymbol{\beta}(u)\in\Re^d} L_n(\boldsymbol{\beta}(u)), \tag{6}$$

and the maximization problem may be solved numerically and efficiently by the coordinate descent algorithm (Hastie et al. 2009).

3.2 Dynamic Risk Estimation

In risk estimation problem our goal is to estimate $p_u(\mathbf{S}_r(u)) = \Pr\{L(T) \geq y | \mathbf{S}_r(u)\}$ given that we have observed $\mathbf{S}_r(u)$ at time u. We use the notation $\mathbf{S}_r(u)$ to denote the real-world observation of $\mathbf{S}(u)$, instead of a simulated observation. Nevertheless, we assume that it has the same distribution as the simulated observations. When using the logistic regression model, we estimate $p_u(\mathbf{S}_r(u))$ by

$$\hat{p}_u(\mathbf{S}_r(u)) = g(\mathbf{X}_r(u), \hat{\boldsymbol{\beta}}_n(u)) = \frac{\exp\left(\hat{\boldsymbol{\beta}}_n(u)^T \mathbf{X}_r(u)\right)}{1 + \exp\left(\hat{\boldsymbol{\beta}}_n(u)^T \mathbf{X}_r(u)\right)}, \tag{7}$$

where $\mathbf{X}_r(u) = \mathbf{X}(\mathbf{S}_r(u))$ is known at time u and $\hat{\beta}_n(u)$ is the ML estimator of the unknown parameter $\boldsymbol{\beta}(u)$ calculated using the training data, i.e., the simulated observations of $\{(\mathbf{X}_i(u), Y_i), i = 1, \ldots, n\}$ at time u. In this subsection, we assess the quality of the risk estimate $\hat{p}_u(\mathbf{S}_r(u))$ and analyze its asymptotic properties. To do that, we make the following assumptions. In the following, we drop the subscript u for convenience, and use $\mathbf{X}$ and $\boldsymbol{\beta}$ to denote $\mathbf{X}(u)$ and $\boldsymbol{\beta}(u)$, respectively.

Assumption 1 The observations $\{\mathbf{V}_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ are independent observations of $\mathbf{V} = (\mathbf{X}, Y)$ and, given $\mathbf{X}$, Y is a Bernoulli random variable with $\Pr(Y = 1 | \mathbf{X}) = g(\mathbf{X}, \boldsymbol{\beta}_0)$.

Assumption 1 is a typical assumption used to analyze the asymptotic properties of the ML estimators (see, for instance, Fahrmeir and Kaufmann (1985) and Newey and McFadden (1994)). Notice that the independence condition is easily satisfied because $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ are calculated based on sample paths that are simulated independently. Therefore, Assumption 1 basically assumes that the logistic regression model of Equation (3) is the true model and the true parameter is $\boldsymbol{\beta}_0$. This is a typical assumption that is made in parametric statistical estimations, and we can only build the properties of the estimators based on this assumption. Nevertheless, we have to keep in mind that models are just approximations and they may introduce bias that we may not be aware of.

Let $H(\boldsymbol{\beta})$ denote the Hessian matrix of the log-likelihood function $\log \ell(\boldsymbol{\beta} | \mathbf{X}, Y)$, defined in Equation (4). Notice that $H(\boldsymbol{\beta})$ is a function of $(\mathbf{X}, Y)$ as well. We make the following assumption on $H(\boldsymbol{\beta})$. In this assumption, the expectations $\mathrm{E}(\cdot)$ are taken with respect to the distribution of $(\mathbf{X}, Y)$.

Assumption 2 The Fisher information matrix $J = \mathrm{E}[-H(\boldsymbol{\beta}_0)]$ exists and is positive definite. Moreover, there exists a neighborhood $\mathscr{N}$ of $\boldsymbol{\beta}_0$ such that $\mathrm{E}\left[\sup_{\boldsymbol{\beta} \in \mathscr{N}} \|H(\boldsymbol{\beta})\|\right] < \infty$.

Assumption 2 is a typical assumption used to analyze the asymptotic properties of the ML estimators. In our problems, however, we show in the following lemma that it can be satisfied easily. All the proofs of the following lemmas and theorems can be found in Jiang et al. (2016).

Lemma 1 Assumption 2 holds if $\mathrm{E}\left(\|\mathbf{X}\|^2\right) < \infty$, where $\|\mathbf{X}\|$ denotes the Euclidean norm of the vector $\mathbf{X}$.

Based on the properties of the ML estimators of generalized linear models, we immediately have the following lemma on the consistency and the asymptotic normality of the ML estimator $\hat{\boldsymbol{\beta}}_n$.

Lemma 2 (Corollary 3, Fahrmeir and Kaufmann (1985)) Under Assumptions 1 and 2, as $n \to \infty$, $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ almost surely (a.s.), and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathbb{N}\left(\mathbf{0}, J^{-1}\right),$$

where "$\xrightarrow{d}$" denotes convergence in distribution and $\mathbb{N}\left(\mathbf{0}, J^{-1}\right)$ denotes a multivariate normal random vector with mean vector $\mathbf{0}$ and covariance matrix J^{-1}.

Based on Lemma 2, by the continuous mapping theorem (Shiryaev 1996) and the delta method Van der Vaart (2000), we can get the asymptotic properties of the dynamic risk estimator $\hat{p}_u(\mathbf{S}_r(u))$ defined in Equation (7). Again, we want to emphasize that, in the theorem, $\mathbf{S}_r(u)$ is known as it has been observed by time u, so it is a deterministic vector.

Theorem 1 Suppose that Assumptions 1 and 2 are satisfied. Then, as $n \to \infty$, $\hat{p}_u(\mathbf{S}_r(u)) \to p_u(\mathbf{S}_r(u))$ a.s., and

$$\sqrt{n}\,[\hat{p}_u(\mathbf{S}_r(u)) - p_u(\mathbf{S}_r(u))] \xrightarrow{d} N(0,D),$$

where $D = c\mathbf{X}_r J^{-1}\mathbf{X}_r^T$, $\mathbf{X}_r = \mathbf{X}(\mathbf{S}_r(u))$ and $c = \exp(2\boldsymbol{\beta}_0{}^T\mathbf{X}_r)/(1+\exp(\boldsymbol{\beta}_0{}^T\mathbf{X}_r))^4$.

Theorem 1 states that, under the assumption that the logistic regression model is the true model, the estimated conditional probability is a consistent estimator of the true conditional probability and it has an asymptotic normal distribution. As it has been shown earlier in the section that the logistic regression model is a good model for the conditional probability, Theorem 1 also shows that our proposed simulation analytic approach can be used to solve the dynamic risk estimation problem. Furthermore, as the logistic regression may be done very quickly given the sample paths or can even be done before $\mathbf{S}_r(u)$ is even observed, the proposed approach can be used for risk estimation in real time.

3.3 Accuracy of Risk Classification

Sometimes, we are concerned about the classification problems, i.e., at the given time, whether the portfolio is safe or not. Of course, an accurately estimated logistic regression model may lead to good classifications. However, a coarsely estimated logistic regression may also give acceptable classifications. For example in Figure 1, the true boundary is the dark blue line, and classifies the points into two categories. With a less accurate boundary, the light blue line, the classification accuracy is almost as good. This motivates us to consider the accuracy of classification instead of the accuracy of parameter estimation as in Section 3.2.

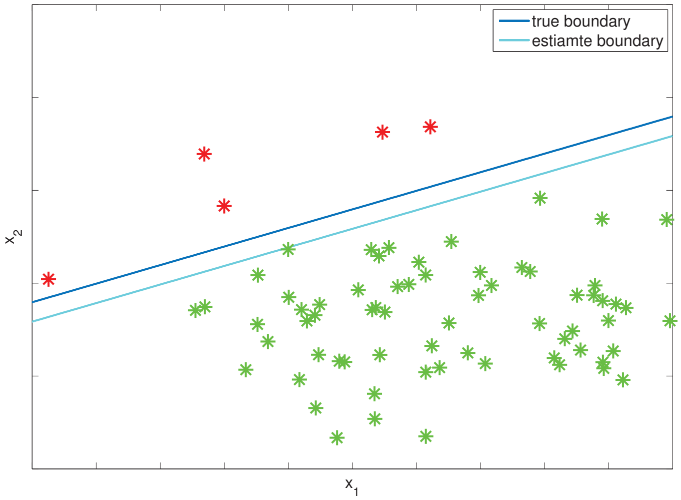

Figure 1: Classification boundary.

Suppose that p_0 is the threshold probability of safe/dangerous zone. If the probability is smaller or equal to p_0, we say the portfolio is in the safe zone. Otherwise, the portfolio is in the dangerous zone. We may estimate $p_u(\mathbf{S}_r(u))$ by $\hat{p}_u(\mathbf{S}_r(u))$. Let I and $\hat{I}$ denote the safe/dangerous indictor random variable for true probability and estimated probability, respectively, i.e.,

$$I = \begin{cases} 1 & \text{if } p_u(\mathbf{x}) \le p_0 \\ 0 & \text{if } p_u(\mathbf{x}) > p_0 \end{cases} \qquad \hat{I} = \begin{cases} 1 & \text{if } \hat{p}_u(\mathbf{x}) \le p_0, \\ 0 & \text{if } \hat{p}_u(\mathbf{x}) > p_0. \end{cases}$$

Then, $\hat{I} \neq I$ indicates a misclassification. In this subsection, we are interested in understanding how the misclassification probability, i.e., $\Pr\{\hat{I} \neq I\}$, converges to zero as the sample size $n \to \infty$. To study this, we need some additional assumptions.

Assumption 3 Suppose that

$$\Pi_0 \triangleq \mathbb{E}\left[\frac{\exp\left(\boldsymbol{\beta}_0^T \mathbf{X}\right)}{\left(1+\exp\left(\boldsymbol{\beta}_0^T \mathbf{X}\right)\right)^2}\mathbf{X}\mathbf{X}^T\right] \tag{8}$$

is positive definite.

Assumption 3 is reasonable, because for any $\mathbf{t} \in \Re^d$,

$$\mathbf{t}^T \Pi_0 \mathbf{t} = \mathbf{t}^T \mathbb{E}\left[\frac{e^{\boldsymbol{\beta}_0^T \mathbf{X}}}{\left(1+e^{\boldsymbol{\beta}_0^T \mathbf{X}}\right)^2}\mathbf{X}\mathbf{X}^T\right]\mathbf{t} = \mathbb{E}\left[\frac{e^{\boldsymbol{\beta}_0^T \mathbf{X}}}{\left(1+e^{\boldsymbol{\beta}_0^T \mathbf{X}}\right)^2}(\mathbf{t}^T\mathbf{X})(\mathbf{t}^T\mathbf{X})^T\right].$$

If $\Pr\{\mathbf{t}^T \mathbf{X} = 0\} < 1$ for any $\mathbf{t} \neq \mathbf{0}$, $\mathbf{t}^T \Pi_0 \mathbf{t} > 0$, so Π_0 is postive definite.

Assumption 4 Suppose that for each X_i, $\mathbb{E}[\exp(t_i|X_i|)] < \infty$ for some $t_i > 0$.

Assumption 4 basically requires that all risk factors are light-tailed. Then, we have the following theorem on the rate of convergence of the misclassification probability. The proof of the theorem is omitted in this paper due to space limit.

Theorem 2 (Vanishing of misclassification) Suppose that Assumptions 1-4 are satisfied and that $p_u(\mathbf{x}_r) \neq p_0$. Then

$$\lim_{n\to\infty} -\frac{1}{n}\log \Pr\left\{I \neq \hat{I}\right\} \geq c$$

for some constant $c > 0$.

Theorem 2 indicates that the misclassification probability converges to zero with an exponential rate, i.e., it satisfies a large derivative principle. This shows that the risk classification problem is in general an easier problem than the risk estimation problem. If one is only interested in risk classification, it may require a smaller amount of samples.

4 NUMERICAL EXAMPLES

4.1 One-dimension Example

We starts with a simple example. The portfolio longs one European call option V^c and short one European put option V^p, which are based on the same underlying asset driven by a geometric Brownian motion,

$$S(t) = S(0)e^{(\mu-\frac{1}{2}\sigma^2)t+\sigma W(t)}, \tag{9}$$

and the maturities are both τ. By put-call parity formula, $V^c(t) - V^p(t) = S_t - Ke^{-r(\tau-t)}$, and we can calculate the analytical default probabilities and the default bounds.

Let the initial value $S(0) = 50$, the drift $\mu = 0.04$, the volatility $\sigma = 0.2$, the risk-free interest rate $r = 0.02$, the risk maturity $T = 0.3$, discretion points for time horizon $N = 30$, and the derivatives maturities $\tau = 1$. Even though the derivatives are the standard European options, we still use the nested simulation to price them. According to Gordy and Juneja (2010), let the outer simulation samples $M_O = 10000$ and inner simulation samples $M_I = 100$. If the portfolio loses 20% of its initial value at the maturity T, i.e., $L_T < 0.8L_0$, we say that the portfolio is in default. First, consider dynamic risk estimation. For the estimated probability, we fit the logistic regression model with predictor S_t, i.e., using the training set to estimated the intercept and the coefficient of S_t via maximum likelihood estimation. Then, we generate a testing set from the original underlying asset with sample size $M_t = 1000$, and calculate the true probabilities and estimated probabilities, respectively. In the left panel of Figure 2, each subfigure corresponds to a time point. In these figures, the x-axis are the true probabilities and y-axis are the estimated probabilities. The points in all subfigures are along the 45° lines, indicating that the estimated probabilities are close to the true probabilities.

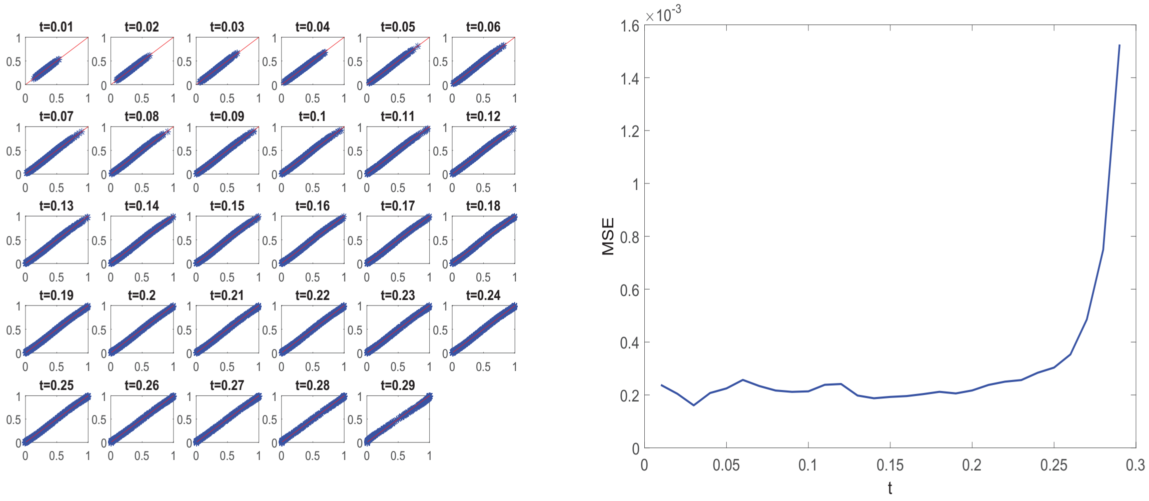

Figure 2: Left: the true default probabilities with respect to the estimated probabilities for time from 0.01 to 0.29; Right: the MSEs for time from 0.01 to 0.29.

Further, we consider the mean squared error (MSE) for the estimated probabilities for different time point. The MSE is calculated as

$$\text{MSE} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{J}\sum_{j=1}^{J}(\hat{p}_l(X_j) - p(X_j))^2,$$

where L is the number of training sets and J is the number of testing sets for each training set, i.e., for each training set l, we get an estimated probability, and we generate J testing set to calculate the MSE for this estimated probability, then we replicate L times to calculate the average of the MSEs. Let $L = J = 20$. Then, we get the right panel of Figure 2. The MSEs are basically small, and increasing when t increases. This is because when t approaching to T, most paths starting at $\mathbf{S}_t$ have probabilities either close to 0 or 1, leading the parameters of the logistics regression models to very large values, thus increasing the variances of the estimators and decreasing the estimate accuracy.

Then, we study dynamic risk classification. First, we consider the classification boundaries for different probabilities. The left panel of Figure 3 indicates that the logistic regression provides a good estimate for default boundaries. Moreover, to evaluate classification clearly, we use the probability of correct classification (PCC) as a criterion. The right panel of Figure 3 shows the PCCs are closed to 0, i.e., almost all the testing data are classified in right categories.

4.2 High-dimension Example

We further consider a high-dimension example. Suppose that a portfolio has two parts, the first part longs 3 call options and 2 put options based on 5 different stocks, which are mutually independent and driven by GBMs. The second part shorts 20 call options, longs 20 put options and the corresponding stocks, which make this part be perfectly hedged. More specifically, Let $\Phi(t) = \Phi_1(t) + \Phi_2(t)$, where $\Phi_1(t) = V^c_{1,1}(t) + V^c_{1,2}(t) + V^c_{1,3}(t) + V^p_{1,4}(t) + V^p_{1,5}(t)$, and $\Phi_2(t) = \sum_{i=1}^{20} V^p_{2,i}(t) - V^c_{2,i}(t) + S_{2,i}(t)$.

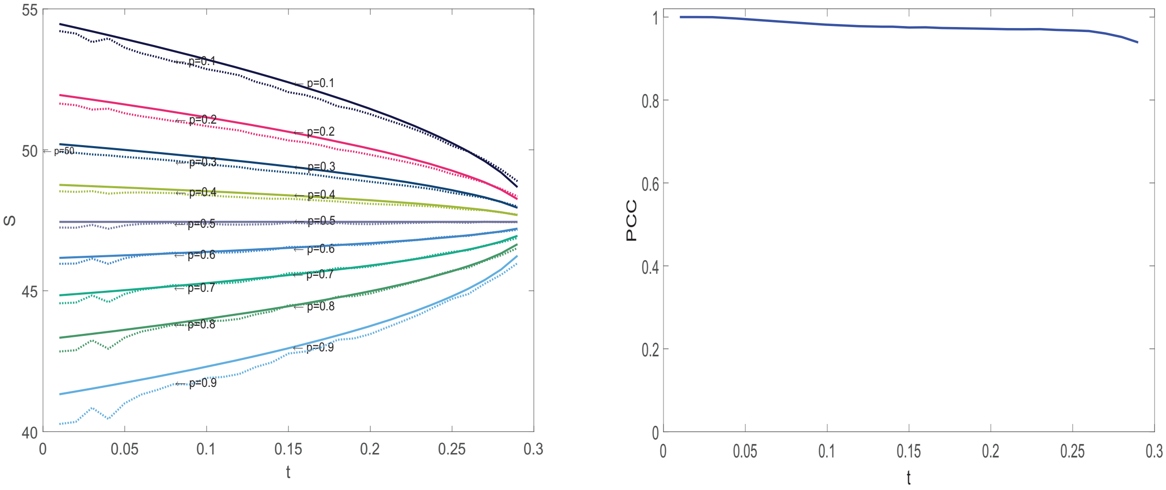

Figure 3: Left: the boundaries for different default probabilities. The solid lines are the true boundaries, and the dotted lines are the estimated boundaries via logistic regression; Right: the PCCs for different time points.

Let the initial values of the five stocks in the first part $\mathbf{S}_1(0) = (50,50,60,60,70)$, the drift $\mu_1 = (0.05,0.06,0.07,0.06,0.05)$, the volatility $\sigma_1 = (0.1,0.1,0.1,0.1,0.1)$, $\mathbf{K}_1 = (40,40,45,80,85)$. Let the initial values of the twenty stocks in the second part $S_{2,i}(0) = 50, i = 1,\ldots,20$, $\mu_{2,i} = 0.02, i = 1,\ldots,20$, $\sigma_{2,i} = 0.05, i = 1,\ldots,10$ and $\sigma_{2,j} = 0.2, j = 11,\ldots,20$, $K_{2,i} = 55, i = 1,\ldots,20$. Let the risk-free interest rate $r = 0.02$, the risk maturity $T = 0.3$, discretion points for time horizon $N = 30$. Let the maturities of all the derivatives $\tau = 1$. Let $Q_1 = 65$, and $Q_2 = e^{-r\tau}\sum_{i=1}^{20} K_{2,i}$. If the portfolio value at maturity T $\Phi(T) = \Phi_1(T) + \Phi_2(T) \leq Q_1 + Q_2$, we say that the portfolio is in default. Let the number of the outer simulation samples $M_O = 10000$, and the number of inner simulation samples $M_I = 100$.

By put-call parity, for the second part, $\sum_{i=1}^{20}\{V_{2,i}^{p}(t) - V_{2,i}^{c}(t) + S_{2,i}(t)\} = e^{-r(\tau-t)}\sum_{i=1}^{20} K_{2,i}, i = 1,\ldots,20$, so the randomness of the portfolio governs by $\mathbf{S}_1(t)$, i.e.,

$$\Pr\{\Phi(T) \leq Q_1 + Q_2 | \mathbf{S}_1(t), \mathbf{S}_2(t)\} = \Pr\{\Phi_1(T) \leq Q_1 | \mathbf{S}_1(t)\}. \tag{10}$$

We then simulate 10^6 scenarios of $\mathbf{S}_1(T)$, and the corresponding derivatives are calculated by the Black-Scholes formula, so that we can calculate $\Phi_1(T)$ and the default probability, which is regarded as the true default probability. Notice that, for fitting the logistic regression model, we still use all the underlying stocks $\mathbf{S}_1(t)$ and $\mathbf{S}_2(t)$ as predictors.

To evaluate the dynamic risk estimation and classification, we use the MSE and PCC as criteria, and let $L = J = 5$, and obtain the following figure.

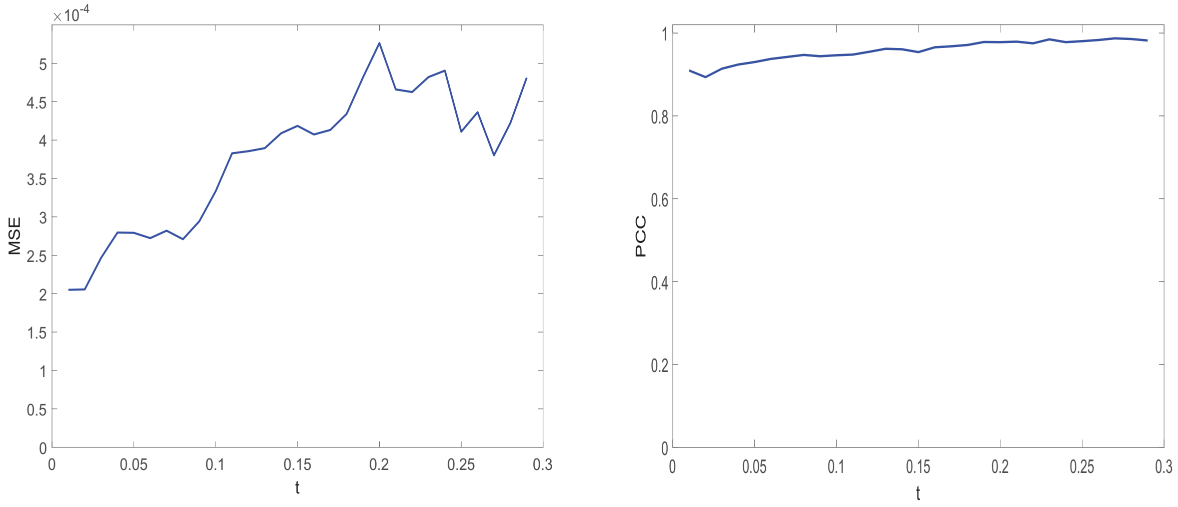

Figure 4: MSEs (left penal) and PCCs (right penal) for different time in Experiment 2.

Figure 4 shows that the MSE curve is close to zero and the PCC curve is close to 1 for all time points, indicating both the risk estimators and the risk classifiers work well in this example.

5 CONCLUSION

In this article, we propose a simulation analytics approach to dynamic risk estimation and classification, which are more meaningful than static risk measurements because they can tell the risk of the portfolio in real time. Our approach illustrates how simulation analytics works, that is, how to use data analytics methods to process the simulation data to uncover the conditional relationship hidden in simulation models, and to make dynamic risk monitoring possible and practical. We also show that the risk classification is in general easier than the risk estimation, and we can obtain good classifications with coarse logistic regression models.

ACKNOWLEDGMENTS

The research of L. Jeff Hong was partially supported by Hong Kong Research Grants Council under projects GRF613213 and GRF16203214. The research of Barry L. Nelson was partially supported by the National Science Foundation under Grant Number CMMI-1537060 and GOALI co-sponsor SAS Institute.

REFERENCES

Broadie, M., Y. Du, and C. C. Moallemi. 2011. “Efficient risk estimation via nested sequential simulation”. *Management Science* 57 (6): 1172–1194.

Broadie, M., Y. Du, and C. C. Moallemi. 2015. “Risk estimation via regression”. *Operations Research* 63 (5): 1077–1097.

Fahrmeir, L., and H. Kaufmann. 1985. “Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models”. *The Annals of Statistics* 13 (1): 342–368.

Feng, M., and J. Staum. 2015. “Green simulation designs for repeated experiments”. In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 403–413. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glasserman, P., P. Heidelberger, and P. Shahabuddin. 2000. “Variance reduction techniques for estimating Value-at-Risk”. *Management Science* 46:1349–1364.

Glasserman, P., P. Heidelberger, and P. Shahabuddin. 2002. "Portfolio value-at-risk with heavy-tailed risk factors". *Mathematical Finance* 12:239–269.

Gordy, M. B., and S. Juneja. 2010. "Nested simulation in portfolio risk measurement". *Management Science* 56 (10): 1833–1848.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Second ed. New York: Springer-Verlag.

Hong, L. J., Z. Hu, and G. Liu. 2014. "Monte Carlo methods for Value-at-Risk and conditional Value-at-Risk: A review". *ACM Transactions on Modeling and Computer Simulation* 24 (4): Article 22.

Hong, L. J., S. Juneja, and G. Liu. 2015. "Kernel smoothing for nested estimation with application to portfolio risk measurement". under review.

Jiang, G., L. J. Hong, and B. L. Nelson. 2016. "Mining Simulation Data for Dynamic Risk Monitoring". working paper.

Lee, S.-H., and P. W. Glynn. 2003. "Computing the distribution function of a conditional expectation via monte carlo: Discrete conditioning spaces". *ACM Transactions on Modeling and Computer Simulation* 13 (3): 238–258.

Liu, M., B. L. Nelson, and J. Staum. 2010. "Simulation on demand for pricing many securities". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, 2782–2789. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Liu, M., and J. Staum. 2010. "Stochastic kriging for efficient nested simulation of expected shortfall". *Journal of Risk* 12 (3): 3–27.

Nelson, B. L. 2016. "'Some Tactical problems in digital simulation' for the next ten years". *Journal of Simulation* 10 (1): 2–11.

Newey, W. K., and D. McFadden. 1994. "Large sample estimation and hypothesis testing". In *Handbook of Econometrics*, edited by R. Engle and D. McFadden, Volume 4, Chapter 36, 2111–2245. New York: Elsevier.

Shiryaev, A. N. 1996. *Probability*. Second ed. Berlin: Springer.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*. Cambridge, MA: Cambridge University Press.

AUTHOR BIOGRAPHIES

GUANGXIN JIANG is a postdoctoral fellow in the Department of Economics and Finance at the City University of Hong Kong. His research interests lie in simulation methodology, modeling, analytics, and optimization, with applications in financial engineering. His email address is guajiang@cityu.edu.hk.

L. JEFF HONG is a chair professor in the Department of Economics and Finance, and the Department of Management Sciences, at the City University of Hong Kong. His research interests include Monte Carlo methods, financial engineering, and stochastic optimization. He is currently an associate editor for *Operations Research, Naval Research Logistics and ACM Transactions on Modeling and Computer Simulation*. His email address is jeffhong@cityu.edu.hk.

BARRY L. NELSON is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*, from Springer. His e-mail address is nelsonb@northwestern.edu.