International Journal of Production Research
2010, 1–18, iFirst

Optimal variability sensitive condition-based maintenance with a Cox PH model

Nan Chen[a], Yong Chen[b], Zhiguo Li[c], Shiyu Zhou[a]* and Cris Sievenpiper[d]

[a]*Department of Industrial and Systems Engineering, University of Wisconsin – Madison, USA;* [b]*Department of Mechanical and Industrial Engineering, University of Iowa, Iowa City, Iowa, USA;* [c]*Xerox R&D Center, Rochester, New York, USA;* [d]*GE Healthcare, Pewaukee, Wisconsin, USA*

(*Received 4 October 2009; final version received 10 February 2010*)

Condition based maintenance (CBM) is an important maintenance strategy in practice. In this paper, we propose a CBM method to effectively incorporate system health observations into maintenance decision making to minimise the total maintenance cost and cost variability. In this method, the system degradation process is described by a Cox PH model and the proposed framework includes simulation of failure process and maintenance policy optimisation using adaptive nested partition with sequential selection (ANP-SS) method, which can adaptively select or create the most promising region of candidates to improve the efficiency. Different from existing CBM strategies, the proposed method relaxes some restrictions on the system degradation model and taking the cost variation as one of the optimisation objectives. A real industry case study is used to demonstrate the effectiveness of our framework.

Keywords: condition-based maintenance; variability-sensitive decision making; simulation optimisation

1. Introduction

Despite the increasing quality and reliability, systems in production or service industries are still subject to deterioration and failures during their usage. Therefore, preventive maintenance remains necessary to reduce unexpected system failures, and thus attracts numerous research works in the literature (see e.g., Valdez-Flores and Feldman 1989, Wang 2002 for reviews).

Because of the rapid development of information and computer technology, a huge amount of data, such as in-process sensing signals (e.g., vibration, acoustic emission), usage patterns, and system event logs, are often collected electronically during the operation of many systems. It is generally believed that this data provides rich information regarding system working conditions. For example, a faulty detector in a computed tomography machine will eventually lead to a 'scan abort' failure. However, before the failure, a faulty detector can cause a series of other events such as analogue-to-digital

*Corresponding author. Email: szhou@engr.wisc.edu

ISSN 0020–7543 print/ISSN 1366–588X online

DOI: 10.1080/00207541003694811
http://www.informaworld.com

converter error, communication error, and software error. By observing these preceding events, we can predict the occurrence of the key failure and accordingly prevent its occurrence or minimise its damages.

Among various preventive maintenance policies, the one that suggests system inspection and maintenance actions based on the on-line observations of system conditions is called condition based maintenance (CBM). In many cases, CBM provides better performance than time based maintenance and the broad availability of data provides great opportunities for establishing optimal CBM policies. Thus, CBM has drawn significant attention in recent years. Jardine *et al.* (2006) provided an excellent review on the condition based maintenance. It was noted that a critical task in CBM is to identify a failure prognostic model to describe the system degradation process and as well impacts from maintenance policies.

In CBM, according to whether the health states of systems are directly observable, they can be classified as completely observable systems and partially observable systems (Jardine *et al.* 2006). In both categories, there is a large amount of literature introducing a variety of prognostic models. For completely observable systems, random coefficient models (e.g., Wang 2000), Markov chain models (e.g., Bloch-Mercier 2002, Chen *et al.* 2003), or Gamma processes (e.g., Grall *et al.* 2002b, Dieulle *et al.* 2003, Liao *et al.* 2006) are proposed to characterise the evolution of the system's health state from direct observations collected continuously or periodically. However, due to the increasing complexities of current sophisticated systems, it is extremely difficult if not impossible to directly observe the systems' health state. Most often, it is only possible to infer the health state from the observations or measurements of certain related characteristics, such as temperatures, pressures, etc. Generally, these characteristics can be classified as event data and condition monitoring data (Jardine *et al.* 2006), which refer to what happened in the system and the measurements related to the system health condition respectively. Many models for partially observable systems, such as hidden Markov models (e.g., Baruah and Chinnam 2005, Dong and He 2007), proportional hazard model (e.g., Makis and Jardine 1992, Kumar and Westberg 1997) were developed to accommodate the maintenance needs for complex systems. In this paper, we adopt the proportional hazard (PH) model, which is widely used (e.g., Jardine *et al.* 1997, Percy and Kobbacy 2000 and references therein) to characterise the health state of the system because of its flexibility in incorporating both event data and condition-monitoring data and its efficiency for statistical modelling.

There are some limitations in the existing works using a PH model for condition-based maintenance. First, many assumptions and restrictions are required to ensure the analytical tractability. For example, Makis and Jardine (1992) required the covariates follow a stochastically increasing Markov process and the coefficients of the PH model are non-negative to ensure the optimality of the policy. Therefore, in the situation where the assumption could not be fulfilled, simply using their results could end up with suboptimal maintenance policies. In this paper we try to relax some assumptions regarding the PH model, and make it extensible to a broader context. Second, in the literature the optimal maintenance policies are often derived with regard to the long run average maintenance cost, which is one of the most widely used optimisation objectives. However, as pointed out in Chen and Jin (2003), the cost variability is very important as well, and can lead to severe management crisis if not considered properly. For example, in risk-avert management, it is preferable to have steady and predictable cost in each month rather than large variant cost across different months. Although the variability-sensitive decision process in general has been studied by many researchers, few of them can be found for

maintenance policies. Tapiero and Venezia (1979) treated the variance of the cost as a risk factor to study a maintenance problem. Rangan and Grace (1988) used variance optimisation criterion to obtain the optimal replacement cycle for systems. However, both works only considered periodic replacement policy. Instead, Chen and Jin (2003) considered the cost-variability-sensitive criterion on different policies, such as age replacement and periodic replacement under minimal repair. They provided the conditions under which the variability-sensitive policies have a finite optimal solution. However, how to incorporate variability-sensitive policy into CBM is still an open question. It was realised that the optimal policy is more difficult to obtain compared with the variability-neutral policy due to the complexity introduced by the variance of cost in the objective function.

In this paper, we want to identify the optimal preventive maintenance policies that can minimise the average maintenance cost and cost variability of a given system where a PH model is used to describe its health evolution process. Because of the complexity introduced by relaxing assumptions on the PH model and adding cost variability in the objective function, it may be very difficult, if not impossible, to derive the objective function analytically. Consequently, classical numerical algorithms cannot be used to find the optimal solution. Therefore, we propose to use a simulation based methodology to optimise the maintenance policies. Compared with traditional optimisation on maintenance policies, simulation based methodology does not rely on restrictive (sometimes unrealistic) assumptions about system health evolution, and therefore can be applied in broader areas. In this paper, we propose to use a simulation model to replicate the evolution of the system health, based on which different maintenance policies are evaluated and compared. An improved optimisation algorithm is also presented to increase the convergence speed of the search process.

The rest of the paper is organised as follows. In Section 2, detailed formulation of the problem is given. In Section 3, the simulation model of system degradation and condition based maintenance is developed, and the optimisation framework ANP-SS is presented in detail. In Section 4, a case study based on real world data is presented to illustrate the effectiveness of our methods. Based on the case study, some general practical implications will be discussed. Finally, we conclude the paper in Section 5 and discuss potential future research directions.

2. Optimal variability sensitive condition-based maintenance

In this paper, we consider a system whose health state can be indirectly observed through collected event data or condition monitoring data. The goal is to develop a maintenance policy that can minimise the objective function including both average cost and cost variability of the CBM. The mathematical formulation is stated below:

$$\min_{\mathrm{P}\in\Gamma,\, I\in\mathbf{R}^+} \mathrm{E}(C)^2 + \gamma \cdot \mathrm{Var}(C), \quad \text{where } C = C_p \cdot N_{\mathrm{PR}} + C_f \cdot N_{\mathrm{ER}} + C_I \cdot N_{\mathrm{IP}}, \tag{1}$$

where P is the maintenance policy that will be optimised, Γ is the set of all feasible policies, and I is the inspection interval, taking positive values, that will be jointly optimised together with P; γ is the factor that adjusts the weight of cost variability in the objective function; C is the random variable denoting the total cost incurred during a pre-specified

time frame, say T. The expectation and variance of C are denoted by $\mathrm{E}(C)$ and $\mathrm{Var}(C)$, respectively. The cost for preventive maintenance is C_p; the cost for emergency replacement (when the system fails between two successive inspections) is C_f; the cost for inspection is C_I. Usually we have $C_f > C_p > C_I$. Furthermore, N_{PR}, N_{ER}, and N_{IP} are the random variables denoting the numbers of preventive maintenance, emergency replacement, and inspection completed within the total time T, respectively. In this paper, we consider the set of hazard rate control limit policies, i.e.:

$$\Gamma = \left\{ \mathrm{D}(t,g) \middle| \mathrm{D}(t,g) = \begin{cases} 1, & h(t|\mathbf{Z}(t)) > g \\ 0, & \text{otherwise} \end{cases}, \ t = kI \ (k = 1, 2, 3, \ldots); g \in \mathbf{R}^+ \right\}, \tag{2}$$

where $\mathrm{D}(t,g)$ is the decision made at each inspection time t, which equals 1 when immediate preventive maintenance action is taken, and equals 0 when no action is enforced; g is the hazard threshold; and $h(t|\mathbf{Z}(t))$ is the system hazard rate at time t with observed covariates $\mathbf{Z}(t)$. According to proportional hazard (PH) model (Cox 1972), we have:

$$h(t|\mathbf{Z}(t)) = h_0(t) \cdot \exp\left[\beta^T \mathbf{Z}(t)\right] = h_0(t) \cdot \exp\left[\sum_{k=1}^{p} \beta_k Z_k(t)\right], \tag{3}$$

where $h_0(t)$ is the baseline hazard rate function; $\mathbf{Z}(t)$ are the observations of system conditions; and the vector β is the coefficient vector. In this paper, we assume that the PH model is explicitly known, either from engineering knowledge or estimations from historical data.

Even though the observations $\mathbf{Z}(t)$ may contain some time-varying variables, it is often expensive and impractical to implement continuous monitoring (Jardine *et al.* 2006). Instead, we assume the system is inspected periodically at fixed interval I, where condition data $\mathbf{Z}(t)$ will be collected and system health $h(t|\mathbf{Z}(t))$ will be updated. Clearly, the frequency of inspection has some impact on the maintenance policies. For example, if the inspection frequency is below a certain level, then the probability that the system will fail between two inspections will increase, and thus the total emergency replacement cost will increase. On the other hand, if the inspection frequency is too high, although it can update the system condition promptly, the cost incurred by frequent inspection will increase. Obviously, there is a trade-off between the inspection cost and the emergency replacement cost; and it is desired to find a good inspection interval that can balance these two costs to achieve the optimal results. In fact, the problem of identifying the optimal inspection interval under some simple system degradation model has been investigated by some researchers (e.g., Hosseini *et al.* 2000, Grall *et al.* 2002a, b). Motivated by these observations, we will jointly optimise the inspection interval I and the maintenance policies P.

It is worth noting that, our methodology does not require the closed-form expression of the objective function. Thus, the proposed methodology can be extended easily to other more complicated maintenance policies. However, in this paper we limit our scope to the control limit policy for illustration purposes. The summary of the major assumptions and settings in our problem formulation are:

(1) The system degradation process can be described by a proportional hazard (PH) model, with covariates observable at inspection.

(2) Inspection is scheduled at fixed time intervals.
(3) Both preventive maintenance and emergency replacement can fully recover the system to as good as new condition.
(4) Preventive maintenance is conducted immediately after the decision of maintenance is made, and emergency replacements are conducted immediately after the system fails.

To solve this CBM problem, we develop a simulation model to evaluate the maintenance policies and use simulation based optimisation methods to direct the search towards optimality. In the next part, we will present our methodologies in detail.

3. Simulation based optimisation for identifying the optimal policy

3.1 *Simulation of the system degradation process based on PH model*

With the given PH model, the system degradation process and the maintenance actions can be simulated. The critical part is to simulate values of time-varying covariates in the PH model and the corresponding failure times that accurately follow the PH model. This means that the simulated sequence of the covariates should follow the same distribution as that in the real systems; and the failure times generated based on the sequence should follow the same distribution as indicated through the specified PH model.

3.1.1 *Simulation of covariates*

Covariates observed from the system can be classified into two major categories: event data and condition monitoring data. To simulate the sequence of predictor events, a typical method is to estimate the corresponding distribution of the time to the occurrences of each event, and sample from these distributions. Both parametric models (Leemis 1999) and nonparametric methods (e.g., Hormann and Leydold 2000) can be used to estimate the original input distribution for future sampling. On the other hand, simulation of condition monitoring data is more complicated. Based on the nature of the monitored covariates, different models could be used. Since these covariates are observed and updated at discrete time intervals, the Markov process is a common tool to model their evolution. Based on the transition probability distribution $G(\mathbf{Z}_{k+1}|\mathbf{Z}_k)$, we can obtain the sample observations of next inspection $\mathbf{z}_{k+1}$ based on current values of covariates $\mathbf{z}_k$, i.e., $\mathbf{z}_{k+1} \sim G(\cdot|\mathbf{z}_k)$.

3.1.2 *Simulation of failure time*

Compared with covariates generation, the generation of failure times is more involved. It has been noted that the cumulative hazard function:

$$H(t) = \int_0^t h_0(u) \cdot \exp[\mathbf{Z}(u)]\mathrm{d}u, \tag{4}$$

follows a unit exponential distribution (Leemis *et al*. 1990). Therefore, to generate the failure time, we can first generate a unit exponential distributed random variable u, then by solving the equation $H(t) = u$, we can obtain the corresponding failure time t. However, the baseline hazard function and covariates function can be very complex, making the

integral equation in (4) difficult to solve analytically. Therefore, numerical methods must be relied on for complicated models.

Fortunately, in many engineering applications, the baseline hazard function can be well approximated using the hazard function of a Weibull distribution with shape parameter λ and scale parameter α, then we have the baseline function as $h_0(t) = \lambda\alpha t^{\lambda-1}$. In this case, it is possible to generate the failure time more efficiently. Noticing the covariates are updated at fixed interval, and are considered as constant during two successive inspections, therefore according to (3), the exponent part of the hazard function only changes at inspections, and keeps constant otherwise. In other words, it has the form:

$$\exp\left(\sum_{k=1}^{p}\beta_k Z_k(t)\right) = \begin{cases} c_0, & 0 \le t < t_1 \\ c_1, & t_1 \le t < t_2 \\ \vdots & \vdots \\ c_n, & t_n \le t < t_{n+1} \\ \vdots & \vdots \end{cases}. \tag{5}$$

Consequently, we can obtain the cumulative hazard function correspondingly as:

$$\begin{aligned} H(t) &= \int_0^t h(x|\mathbf{Z}(x))\mathrm{d}x \\ &= \begin{cases} c_0\alpha t^{\lambda}, & 0 \le t < t_1 \\ \alpha[c_0 t_1^{\lambda} + c_1(t^{\lambda} - t_1^{\lambda})], & t_1 \le t < t_2 \\ \vdots & \vdots \\ \alpha\left[c_0 t_1^{\lambda} + c_1(t_2^{\lambda} - t_1^{\lambda}) + \cdots + c_{n-1}(t_n^{\lambda} - t_{n-1}^{\lambda}) + c_n(t^{\lambda} - t_n^{\lambda})\right], & t_n \le t < t_{n+1} \end{cases}. \end{aligned} \tag{6}$$

For illustration, a typical cumulative hazard function curve is shown as the dashed line in Figure 1. The baseline cumulative hazard function is also depicted as the solid line for comparison.

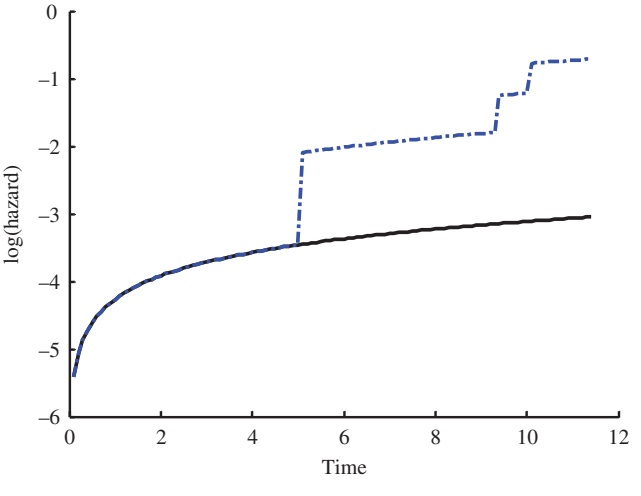

Figure 1. Total and baseline cumulative hazard function.

It can be noted that $H(t)$ in this case is a stepwise invertible function. By solving the equation $H(t) = u$, we can obtain:

$$t = \begin{cases} \left(\dfrac{u}{c_0\alpha}\right)^{1/\lambda}, & 0 \le u < c_0\alpha t_1^{\lambda} \\ \left(\dfrac{u - c_0\alpha t_1^{\lambda}}{\alpha c_1} + t_1^{\lambda}\right)^{1/\lambda}, & c_0\alpha t_1^{\lambda} \le u < \alpha[c_0 t_1^{\lambda} + c_1(t_2^{\lambda} - t_1^{\lambda})] \\ \vdots & \vdots \\ \left(\dfrac{u - \alpha[c_0 t_1^{\lambda} + c_1(t_2^{\lambda} - t_1^{\lambda}) + \cdots + c_{n-1}(t_n^{\lambda} - t_{n-1}^{\lambda})]}{\alpha c_n} + t_n^{\lambda}\right)^{1/\lambda}, & \alpha[c_0 t_1^{\lambda} + c_1(t_2^{\lambda} - t_1^{\lambda}) + \cdots + c_{n-1}(t_n^{\lambda} - t_{n-1}^{\lambda})] \\ & \le u < \\ & \alpha[c_0 t_1^{\lambda} + c_1(t_2^{\lambda} - t_1^{\lambda}) + \cdots + c_n(t_{n+1}^{\lambda} - t_n^{\lambda})] \end{cases}, \tag{7}$$

where u is a random variable following unit exponential distribution.

3.1.1 *Simulation of maintenance actions*

To identify the optimal CBM policy, we also need to incorporate the impact of the maintenance action in the simulation. In this paper, the maintenance policy is based on a hazard control limit: when the hazard rate exceeds the threshold at inspection, or the system fails anytime during the operation, the component will be replaced immediately, which corresponds to the termination of current simulation run and regeneration of the new covariates and failure time for the next run. During the same time, the corresponding cost and time elapsed are recorded for future evaluation.

With the simulation of covariates, failure times, and maintenance actions, we can establish a complete simulation flow for the CBM process, as shown in Figure 2. Under this simulation logic, we can estimate and compare the objective function under different maintenance policies and use an optimisation technique to select the optimal one.

3.2 *Optimising the maintenance policy based on simulation*

Different from deterministic optimisation, simulation based optimisation is often a stochastic problem because the objective function is estimated based on random outputs from multiple simulation runs. It is often required that the estimation of the objective

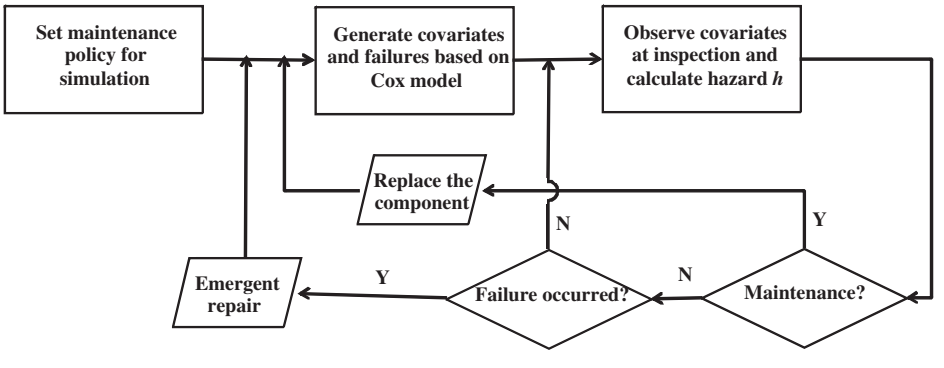

Figure 2. Simulation flow chart for maintenance policies.

function should be unbiased to ensure convergence. Suppose under a given CBM policy, we run the simulation n times, and obtain the total costs $C_1, C_2, \ldots, C_n$. Then the objective function can be estimated by:

$$\hat{f} = \frac{1}{n}\sum_{i=1}^{n} C_i^2 + \frac{\gamma - 1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C})^2, \quad \text{where } \bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i. \tag{8}$$

We can show that it is unbiased by noting:

$$\begin{aligned} \mathrm{E}(\hat{f}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(C_i^2) + \mathrm{E}\left[\frac{\gamma-1}{n-1}\sum_{i=1}^{n}(C_i - \bar{C})^2\right] \\ &= \mathrm{E}(C^2) + (\gamma - 1)\mathrm{Var}(C) \\ &= [\mathrm{E}(C)]^2 + \gamma \cdot \mathrm{Var}(C). \end{aligned} \tag{9}$$

To avoid gradient estimation, which is often very time-consuming for simulation-based procedures, we adopt and improve the gradient free optimisation method nested partition (NP) (Shi and Chen 2000) in this paper. The idea of NP is as follows. In each iteration, a region is selected as the most promising region. Then this region is partitioned into M subregions; all the other regions are aggregated into one region. Each of these $M+1$ disjoint regions are sampled and evaluated through some performance function. The region with the highest score will be selected as the most promising region in the next iteration. A brief description of the procedure is given in the Appendix.

It is also worth noting that the NP method is most effective with finite or countable sample space. In our application, we first discretise the continuous sample space to a discrete and countable sample space at a given precision before applying the optimisation methods. We also propose some improvements on the original NP framework. The method we use here is called adaptive nested partition with sequential selection, or ANP-SS for short. Simply speaking, we improve the estimation of the promising index, and choose the most promising region more efficiently in each iteration. In the following, we will focus on describing the changes we made on the original NP framework.

3.2.1 *Partitioning and sampling*

Denote the dimension of the sample space as n. Then at each partitioning step, one dimension is chosen to be partitioned. In this paper, we choose the dimension which has maximum cardinality as partition dimension. In other words, the dimension with the largest range is chosen to reduce the likelihood of incorrectly selecting the most promising region during the sequential selection stage. Suppose the sth dimension has been chosen to partition, and its upper bound and lower bound of the region is u_s and l_s, respectively. Then, it is sufficient to find a set of cut-points which can divide the range between l_s and u_s into approximately equal intervals. Denote these cut-points as $\{r_1 = l_s, r_2, r_3, \ldots, r_{M+1} = u_s\}$, and the region to be partitioned can be expressed as $\sigma(k) = \{\mathbf{x} | l_j \le x_j \le u_j, 1 \le j \le n\}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the point in n-dimensional space. Then the subregions obtained by partitioning dimension s can be expressed as:

$$\sigma_i(k) = \left\{\mathbf{x} | l_j \le x_j \le u_j, j \ne s, r_i \le x_s \le r_{i+1}\right\}, \quad \forall i = 1, 2, \ldots, M. \tag{10}$$

In this way, each promising region can be partitioned into M subregions. The next step would be drawing random samples from these regions.

In the general case, the feasible region in the sample space may have a complex shape. In the case where the feasible region is convex and defined by a set of linear constraints, a procedure called MIX-D can be used to draw samples approximately uniformly from the region (Pichitlamken and Nelson 2003). In this paper, we use stratified sampling, which takes samples at each dimension separately, and then combines them together to obtain the final samples from solution space. To be specific, we denote $\sigma(k) = \{\mathbf{x}|l_i \le x_i \le u_i, 1 \le i \le n\}$ as the space to be sampled. For dimension i, we will draw $m(k)$ random samples uniformly from range $l_i \le x_i \le u_i$, denoted as x_{ij}, $j = 1, 2, \ldots, m(k)$. After obtaining the samples in each dimension, we can combine them together to obtain the samples in the original space by:

$$E = \left\{\mathbf{x}|(x_{1j}, x_{2j}, \ldots, x_{nj}), \quad j = 1, 2, \ldots, m(k)\right\}.$$

Since samples in each dimension are independent with that in other dimensions, the uniform sampling in each dimension guarantees the uniform distribution of $\mathbf{x}$ in the original space.

3.2.2 *Sequential selection and adaptive partitioning*

After partitioning and drawing $m(k)$ samples from each subregion, we will obtain $(M+1) \cdot m(k)$ samples. However, since our sample space is discrete, we will only keep distinct samples, and index these samples from 1 to q as $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q\}$. The problem we are trying to solve here is to find the sample with the minimum objective function value. However, since the objective function can only be estimated through simulations, the problem is not as trivial as it sounds. The sequential selection procedure for ranking and selection problems (Swisher and Jacobson 1999) can guarantee to select the sample whose expected objective value is larger than others by an amount $\delta > 0$ with probability at least $1 - \varepsilon$, where δ is called the indifference-zone parameter, which specifies by which amount we think one solution is better than another.

In identifying the $\mathbf{x}_k$ with minimum objective function value, we can utilise the sequential selection method to improve the probability of correct selection efficiently. Since we can only obtain estimates of the objective function, it is necessary to have several independent runs. To efficiently allocate the simulation budgets, we can decide the number of replications adaptively based on the function estimates and their variability. The following procedure indicates how the replication number should be determined:

Step 1: For each $\mathbf{x}_i$, $i = 1, 2, \ldots, q$, take n_0 observations by simulation to obtain the objective function estimate Y_{iv}, $v = 1, 2, \ldots, n_0$.

Step 2: Compute the variance estimator of $\mathrm{Var}(Y_i - Y_j)$ by:

$$S_{ij}^2 = \frac{1}{n_0 - 1} \sum_{v=1}^{n_0} \left[Y_{iv} - Y_{jv} - \bar{Y}_i(n_0) + \bar{Y}_j(n_0)\right]^2, \tag{11}$$

where $\bar{Y}_i(\kappa)$ is the sample mean using the first κ samples from the simulation results Y_{iv}. With the initial estimates of variance, we can determine the procedure

parameters Δ and a_{ij} as:

$$\Delta = \frac{\delta}{2} \quad \text{and} \quad a_{ij} = \frac{(n_0 - 1)S_{ij}^2}{4(\delta - \Delta)}\left[\left(\frac{q-1}{2\varepsilon}\right)^{\frac{2}{n_0-1}} - 1\right], \tag{12}$$

and the number of observations N_i needed to be taken to assure the correct selection probability $N_i = \max_{j \neq i}\{|a_{ij}/\Delta|\}$. Also set $Q = \{1, 2, \ldots, q\}$ as the index set for candidate selections, and $\kappa = n_0$ as the initial number of observations for the following iterative screening.

Step 3: Screening. Set $Q^{\text{old}} = Q$, and update Q as:

$$Q = \left\{ i : i \in Q^{\text{old}} \text{ and } \sum_{v=1}^{\kappa} Y_{iv} \leq \max_{j \in Q^{\text{old}}, j \neq i}\left\{\sum_{v=1}^{\kappa} Y_{jv} + a_{ij}\right\} - \kappa\Delta \right\}. \tag{13}$$

Step 4: Stopping rule. If any of the following three criteria is satisfied, then the sequential selection is terminated, and the most promising region is returned:

- If $n_0 > \max\{N_i\}$, then select the solution with smallest $\bar{Y}_i(n_0)$, and corresponding subregion as the most promising region $\sigma(k+1)$.
- Adaptive partitioning. If $\mathbf{x}_i \in \sigma(k) = \{\mathbf{x} | l_w \leq x_w \leq u_w, 1 \leq w \leq n\}, \forall i \in Q$, and:

$$\max_{i,j \in Q}\{|x_{is} - x_{js}|\} \leq \max_{w=1,2,\ldots,M}\{r_{w+1} - r_w\},$$

where r_w are the subregion boundary defined in (10), then the most promising region is constructed as $\sigma(k+1) = \{\mathbf{x} | l_w \leq x_w \leq u_w, w \neq s \text{ and } l'_s \leq x_s \leq u'_s\}$, where s is the dimension chosen as the partitioning dimension, and:

$$\begin{aligned} l'_s &= \frac{1}{|Q|}\sum_{j \in Q} x_{js} - \frac{1}{2}\max_{w=1,2,\ldots,M}\{r_{w+1} - r_w\} \\ \text{and } u'_s &= \frac{1}{|Q|}\sum_{j \in Q} x_{js} + \frac{1}{2}\max_{w=1,2,\ldots,M}\{r_{w+1} - r_w\}, \end{aligned} \tag{14}$$

where $|Q|$ is the number of remaining elements in the set Q, and x_{js} is the value of $\mathbf{x}_j$ in sth dimension.

- Run an additional simulation for each $\mathbf{x}_i$, $\forall i \in Q$, and set $\kappa = \kappa + 1$. If $\kappa = \max\{N_i\} + 1$, then select the solution with smallest $\bar{Y}_i(\kappa)$, and the corresponding subregion as most promising region $\sigma(k+1)$; otherwise, go to Step 3 for further screening

The first three steps in the above procedure are the same as that provided in Pichitlamken (2002). However, the adaptive partition in the stopping rule is added in Step 4 in order to improve the original method. In the original method, the sequential selection is stopped when all the remaining samples are in the same subregion. However, in the method we proposed, the sequential selection is stopped when the remaining samples are close enough to each other to form a new subregion as the most promising region. The reason for adaptive partitioning is that when we select cut-points to define the subregions, the selection is arbitrary, without any consideration on the objective function structure.

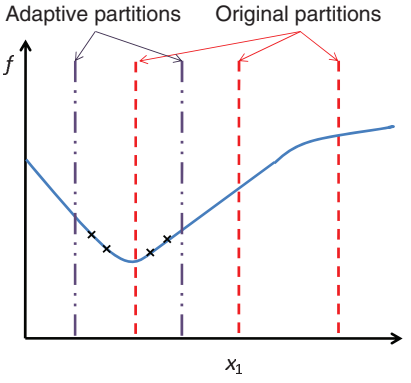

Figure 3. Illustration of advantages of adaptive partitioning.

However, it is possible that the subregion selection is inappropriate which can lead to incorrect selection of the most promising region, as demonstrated in Figure 3.

From Figure 3, we can observe that when the original partitioning line is close to the minimum solution we try to find, and if we select the most promising region from the original subregions, then it is very likely we will miss the optimal solution in our most promising region, which will cause inefficiency. Instead, if we find from the remaining samples that they are close to each other, as illustrated by the black crosses in the figure, although not in one subregion, we can repartition the original region to have one subregion cover this area, and accordingly this subregion will become the most promising region in the next iteration. Through this strategy, we can reduce the number of iterations and evaluations of the objective function, and thus increase the efficiency and effectiveness of the optimisation scheme.

4. Numerical case study

In this section, we present a numerical case study to illustrate the effectiveness of the proposed method. We use historical event logs from a computed tomography (CT) system to fit the PH model to describe the failure occurrences. After model fitting, simulation and optimisation based on the estimated model are conducted to identify the optimal maintenance strategy. In the following, we will illustrate these procedures step by step.

4.1 *PH model estimation from log files*

In the CT system log files, a large amount of historical events are recorded during the period of monitoring. After preprocessing, there are 7199 events of as many as 179 different types. For the critical failure we are interested in, which requires immediate attention and repair, there are 107 occurrences in this data set. In this paper, we encode the events as varying binary variables, i.e.:

$$Z_A(t) = \begin{cases} 0, & 0 \le t < \text{the occurence time of } A \\ 1, & \text{the occurence time of } A \le t \le \text{the end of the TIBF} \end{cases}. \tag{15}$$

After pattern identification and model selection, we can select several important events as predictor events, and use them as covariates to predict the failure distribution (Li *et al.* 2007). The model estimated from the data is shown below:

$$h(t) = 0.049 \cdot t^{0.558} \cdot \exp(0.86Z_A + 1.22Z_B + 0.64Z_C - 1.55Z_A \times Z_B). \quad (16)$$

It is noted that the interaction effect of events A and B is negative, which means it will decrease the hazard rate. In many theoretical analyses, an important assumption is the non-decreasing hazard rate along the time line (Makis and Jardine 1992). Thus, those theoretical analysis results cannot be applied to this model. However, the simulation based method proposed in this paper does not need this assumption, and can handle this situation effectively.

With the selected statistically significant predictor events in the PH model, we also need to estimate the distributions of their occurrence time. It is observed that, not all the predictor events would happen before the critical failures. In other words, some predictor event occurrence times would be censored by the failures. Therefore, it is necessary to take this censored data into consideration when estimating the distribution of the occurrence time of predictor events. Figure 4 illustrates the empirical survival function of the predictor events (considering censored data) and their corresponding estimated survival function using exponential distribution.

From Figure 4, we can find that the censoring indeed has a large influence on the estimation of the distribution. If we ignore the censored data, and only use the completely observed data to test the goodness of fit of the estimated distribution, the test may reject the hypothesis that the data comes from the tested distribution. However, by considering the censoring in the data, the p-value of the goodness of fit tests would be improved, as shown in Table 1.

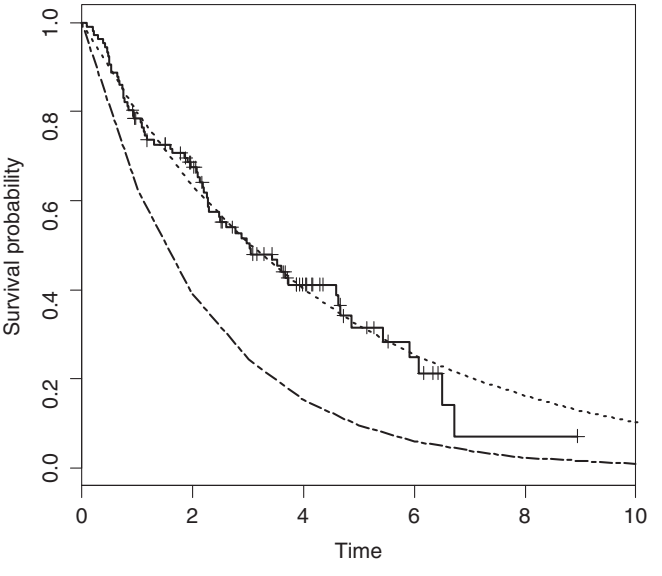

Figure 4. Comparison of the empirical distribution and the estimated exponential distribution.

From Table 1, we can observe that the goodness of fit χ^2 tests do not reject the hypothesis that the data is from these estimated distributions. Therefore, we will use these estimated distributions to generate the occurrence time of predictor events during simulation.

With the distribution of the covariates and the PH model as shown in (16), we can use the methodology described in Section 3.1 to generate the failure time based on (7). Additionally, with the estimated baseline hazard function parameter $\lambda = 1.558$, and $\alpha = 0.0315$, we can plug them in (7) to obtain the failure time according to the distribution implied by the PH model.

4.2 *Optimal variability sensitive policies*

After identifying the PH model, we can use simulation optimisation to find the optimal policies. Suppose we use $\gamma = 20$ in the objective function as the weight of cost variance, and corresponding costs for preventive maintenance and emergency replacement are $C_p = 200$ and $C_f = 800$. To illustrate the effectiveness of our simulation optimisation framework, we first set the inspection interval to 1 (month), and use our method to find the optimal hazard threshold. The simulation length is 100 (months), and no inspection cost is considered during this validation process. For comparison, we also use 10,000 replications to estimate the objective function under different hazard thresholds, and the result is shown in Figure 5.

From Figure 5, we can find that the variability sensitive policy is more conservative and results in smaller hazard threshold. It can also be observed that with little sacrifice in mean cost, the variability of the maintenance cost can be greatly reduced. The optimal hazard threshold for variability sensitive policy is identified as 0.11 from the graph. Alternatively, if we use the optimisation technique introduced in Section 3.2, we can quickly find the optimal value as 0.11, which is consistent with Figure 5. Our framework is more efficient when the decision variables are multi-dimensional, in which case the computation load is exponentially increased for grid evaluations. As an example, the proposed optimisation can find the solution for the two-dimensional problem in around 15 hours, while the grid evaluation will take more than 7 days on the same computer with the same problem settings. To illustrate the advantages of condition based maintenance over time based maintenance, we also compare our optimal policy with the optimal periodical maintenance policy. Numerical results show that the optimal periodical maintenance policy has about 10% higher value in objective function than optimal CBM policy. Since in general, the degradation process of the CT system does not follow the Markov process, many maintenance policies based on the Markovian property are not applicable here.

For the multi-dimensional problem of finding the optimal inspection interval and hazard threshold combination to minimise the objective function, we can still achieve satisfactory results by applying the optimisation algorithm we proposed. We change the

Table 1. Estimated parameters of exponential distribution.

Covariate	$\hat{\mu}$	p-value
Z_A	2.0783	0.9073
Z_B	4.3755	0.2251
Z_C	7.275	0.4615

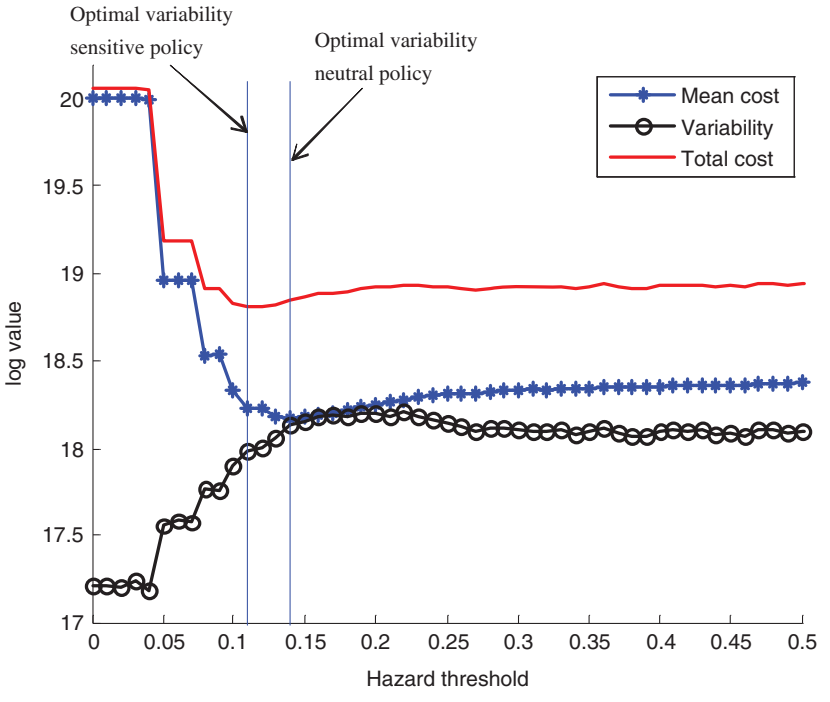

Figure 5. Estimates of objective function values for different thresholds.

inspection cost to $C_I = 20$ to avoid triviality (otherwise, if $C_I = 0$, the optimal inspection interval is always 1). The solution found by our optimisation algorithm is: inspection interval equals 5 (months) and hazard threshold equals 0.96. The estimated minimum value of the objective function is 19.86 (logarithm value).

4.3 *Impact of inspection cost*

From the previous results, we can find that the inspection cost plays an important role in determining the optimal inspection interval. When the cost tends to be small, the inspection would like to be more frequent to lower the risk of emergent failure. On the other hand, when the inspection cost is high, the cost saving from frequent inspection could not compensate the high inspection costs, and therefore the optimal solution is more likely to have a longer inspection interval. To illustrate this idea, we change the inspection cost from 0 to 100, and run our simulation optimisation algorithm to find the optimal solution under each condition. The inspection interval is in the range of $[1, 10]$ with integer values, and the hazard threshold is in the range of $[0, 1]$ with real values. The result is summarised in Table 2.

From Table 2, we can find that as the inspection cost increases, the optimal inspection interval consistently increases until reaching the upper bound. When the inspection cost is zero, the optimal inspection interval is the smallest possible value. Also, as the inspection cost increases, the optimal objective function value also increases because: first at each inspection, the operation cost increases; second the longer inspection interval increases the

Table 2. Optimal maintenance policies with different inspection costs.

Inspection cost	Minimal function	Optimal inspection	Optimal hazard threshold
0	19.37	1	0.19
10	19.49	1	0.19
20	19.60	5	0.86
30	19.61	7	0.94
40	19.63	10	0.91
50	19.64	10	0.98
60	19.65	10	0.87
70	19.66	10	0.93
80	19.67	10	0.96
90	19.68	10	0.91
100	19.69	10	0.86

risk of unexpected failure and thus increases the overall maintenance cost. In contrast, the optimal hazard threshold also increases as a general trend, but with small fluctuations. The main reason is that during the optimisation, if the difference between objective function values is within the indifference zone, the search will be terminated, and the solution which may not necessarily be the precisely optimal one, will be returned. Therefore, if the hazard thresholds have their objective function values very close to each other within a certain range, the solutions found by the optimisation algorithm are likely to fluctuate within this range.

4.4 *Impact of variance of cost*

A major difference between the variability-sensitive policy and the policies considered in the existing literature is that the variance of the maintenance cost is taken into consideration. Therefore, how the variance of the cost would change the optimal threshold is interesting to study.

Intuitively, as γ in the objective function increases, the optimal policy should become more conservative to make the total cost more predictable. In other words, it may reduce the inspection interval or hazard threshold to make the preventive maintenance more frequent. Although this may increase the average cost slightly, it can reduce the variance of the cost considerably, and thus reduce the overall objective function value, as illustrated in Section 4.2. To further demonstrate this point, we change the γ values from 0 to 25 with a step of 5, and use our optimisation algorithm to find the optimal solution in each case. In this simulation, we set the inspection cost to 20. The results are depicted in Table 3.

From Table 3, we can find that, as γ increases, the optimal inspection interval decreases consistently and under the same inspection interval, the optimal hazard threshold decreases. These trends demonstrate the optimal policy is moving toward the conservative direction. Also, it can be observed that as γ increases, the optimal objective function value also increases. Through the simulation, we can see that by including the variability of cost into consideration, the maintenance policy tends to be more conservative, and thus the risk can be controlled.

Table 3. Optimal maintenance policies with different weight of cost variances.

Cost variance weight	Minimal function value	Optimal inspection interval	Optimal hazard threshold
0	19.59	7.5	0.85
5	19.67	7.5	0.71
10	19.74	7.5	0.64
15	19.81	6	0.94
20	19.86	5	0.96
25	19.92	5	0.48

5. Concluding remarks and future work

In this paper, we presented a method to find the optimal variability sensitive condition-based maintenance policy based on system health monitoring data. Different from existing works, we used simulation to characterise the system degradation process and maintenance actions and apply simulation based optimisation to find the optimal CBM policy. This method does not require strict assumptions and is very flexible to handle different settings. A case study from a real system illustrates the effectiveness of our method.

There are still some open issues. In this paper, we only considered the simple hazard rate control limit policy. Although this policy is widely used in practice, other more sophisticated policies may lead to larger cost savings and smaller management risks. Another potential improvement lies in failure time modelling. Currently, we use a PH model to characterise the relationship between condition variables and the failure time distribution. When the proportional hazard assumption is not valid, however, the model accuracy would be affected. Other more general models could be developed to handle more complicated situations, and thus could lead to better predictions and lower costs. In the future, we will focus on these two directions, and will report the results in the near future.

Acknowledgement

The financial support of this work is provided by NSF grants #0757683 and #0758178, and GE Healthcare.

References

Baruah, P. and Chinnam, R.B., 2005. HMMs for diagnostics and prognostics in machining processes. *International Journal of Production Research*, 43 (6), 1275–1293.

Bloch-Mercier, S., 2002. A preventive maintenance policy with sequential checking procedure for a Markov deteriorating system. *European Journal of Operational Research*, 142 (3), 548–576.

Chen, C.T., Chen, Y.W., and Yuan, J., 2003. On a dynamic preventive maintenance policy for a system under inspection. *Reliability Engineering and System Safety*, 80 (1), 41–47.

Chen, Y. and Jin, J., 2003. Cost-variability-sensitive preventive maintenance considering management risk. *IIE Transactions*, 35 (12), 1091–1101.

Cox, D.R., 1972. Regression models and life-tables. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 34 (2), 187–220.

Dieulle, L., *et al*., 2003. Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of Operational Research*, 150 (2), 451–461.

Dong, M. and He, D., 2007. Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research*, 178 (3), 858–878.

Grall, A., Berenguer, C., and Dieulle, L., 2002a. A condition-based maintenance policy for stochastically deteriorating systems. *Reliability Engineering and System Safety*, 76 (2), 167–180.

Grall, A., *et al*., 2002b. Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE Transactions on Reliability*, 51 (2), 141–150.

Hormann, W. and Leydold, J., 2000. Automatic random variate generation for simulation input. *In*: *Proceedings of the 2000 winter simulation conference, vol. 1*, 10–13 December, Orlando, Florida, 675–682.

Hosseini, M.M., Kerr, R.M., and Randall, R.B., 2000. An inspection model with minimal and major maintenance for a system with deterioration and Poisson failure. *IEEE Transactions on Reliability*, 49 (1), 88–98.

Jardine, A.K.S., Banjevic, D., and Makis, V., 1997. Optimal replacement policy and the structure of software for condition-based maintenance. *Journal of Quality in Maintenance Engineering*, 3 (2), 109–119.

Jardine, A.K.S., Lin, D., and Banjevic, D., 2006. A review on machinary diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20 (7), 1483–1510.

Kumar, D. and Westberg, U., 1997. Maintenance scheduling under age replacement policy using proportional hazards model and TTT-ploting. *European Journal of Operations Research*, 99 (3), 507–515.

Leemis, L., Shih, L.H., and Keynertson, K., 1990. Variate generation for accelerated life and proportional hazards models with time dependent cases. *Statistics and Probability Letters*, 10 (4), 335–339.

Leemis, L., 1999. Simulation input modelling. *In*: *Proceedings of the 31st conference on winter simulation: simulation – a bridge to the future. vol. 1*, 5–8 December, Phoenix, Arizona, 14–23.

Li, Z., *et al*., 2007. Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. *IIE Transactions*, 39 (3), 303–315.

Liao, H.T., Elsayed, E.A., and Chan, L.Y., 2006. Maintenance of continuously monitored degrading systems. *European Journal of Operational Research*, 175 (2), 821–835.

Makis, V. and Jardine, A.K.S., 1992. Optimal replacement in the proportional hazards model. *INFOR*, 30 (1), 172–183.

Percy, D.F. and Kobbacy, A.H., 2000. Determining economical maintenance intervals. *International Journal of Production Economics*, 67 (1), 87–94.

Pichitlamken, J., 2002. *A combined procedure for optimization via simulation*. Dissertation (PhD). Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Pichitlamken, J. and Nelson, B., 2003. A combined procedure for optimization via simulation. *ACM Transactions on Modeling and Computer Simulation*, 13 (2), 155–179.

Rangan, A. and Grace, R.E., 1988. A non-Markov model for the optimum replacement of self-repairing systems subject to shocks. *Journal of Applied Probability*, 25 (2), 375–382.

Shi, L. and Chen, C., 2000. A new algorithm for stochastic discrete resource allocation optimization. *Discrete Event Dynamic Systems*, 10 (3), 271–294.

Swisher, J.R. and Jacobson, S.H., 1999. A survey of ranking, selection, and multiple comparison procedures for discrete-event simulation. *In*: *Proceedings of the 31st conference on*

winter simulation: simulation – a bridge to the future, vol. 1, 5–8 December, Phoenix, Arizona, 492–501.

Tapiero, C.S. and Venezia, I., 1979. A mean variance approach to the optimal machine maintenance and replacement. *The Journal of the Operational Research Society*, 30 (5), 457–466.

Valdez-Flores, C. and Feldman, R.M., 1989. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Naval Research Logistics*, 36 (4), 419–446.

Wang, H., 2002. A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research*, 139 (3), 469–489.

Wang, W., 2000. A model to determine the optimal critical level and the monitoring intervals in condition-based maintenance. *International Journal of Production Research*, 38 (6), 1425–1436.

Appendix

The framework of the NP method proposed by Shi and Chen (2000) is summarised. Denote Ω as the feasible region, and $\sigma(k)$ as the most promising region in the kth iteration.

Stage 1: Initialisation.
Set $k = 0$, choose the whole sample space as the most promising region $\sigma(0) = \Omega$.

Stage 2: Partitioning.
Partition $\sigma(k)$ into M subregions: $\sigma_1(k)$, $\sigma_2(k), \ldots,$ $\sigma_M(k)$, and aggregate all the other regions into one region $\sigma_{M+1}(k)$.

Stage 3: Sampling.
Randomly draw $m(k)$ samples in each region.

Stage 4: Evaluation and selection.
Evaluate and estimate the objective function at these samples through simulation. Based on the estimates, choose the most promising region for the next step $\sigma(k+1)$.

Stage 5: If $\sigma(k+1)$ is not fully contained in $\sigma(k)$, then backtracking is needed, and $\sigma(k+1)$ is set to $\sigma(k-1)$, which is the super region of $\sigma(k)$. Otherwise, Stage 2 to Stage 4 will be repeated until $\sigma(k+1)$ is a singleton, which cannot be further partitioned.