Reliability Engineering and System Safety 000 (2017) 1–14

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Hidden Markov model with auto-correlated observations for remaining useful life prediction and optimal maintenance policy

Zhen Chen[a], Yaping Li[a], Tangbin Xia[a,b], Ershun Pan[a,*]

[a] State Key Laboratory of Mechanical System and Vibration, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, China
[b] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Remaining useful life prediction
Preventive maintenance
Hidden Markov models
Auto-correlated observations

ABSTRACT

In this paper, a hidden Markov model with auto-correlated observations (HMM-AO) is developed to handle the degradation modeling of manufacturing systems. Unlike the standard hidden Markov models (HMMs), the current observation in the HMM-AO model not only depends on the corresponding hidden system state, but also on the previous observations. A novel algorithm using the expectation maximum is presented to estimate the unknown parameters. Furthermore, missing data and noise that accumulate over time are also considered by modifying the proposed model. Then two remaining useful life prediction methods based on the HMM-AO model are developed. Predictive values of more accuracy can be obtained, since the autocorrelation of observations has been considered and the temporal evolution of degradation processes has been described properly. A case study is illustrated to highlight the advantages of HMM-AO and demonstrate the accuracy and efficiency of the prediction methods. Furthermore, an improved maintenance policy is developed based on the results of remaining useful life prediction. Finally, a comparison with a conventional condition-based maintenance policy is provided to prove the performance of this proposed policy.

1. Introduction

Maintenance policies play important roles in improving the effectiveness of production systems' operation. In recent years, maintenance strategies combining data-driven reliability models with observed data of condition monitoring has gained more and more attention [1–3]. Diagnosis and prognosis are two important aspects in the framework of maintenance and have been largely researched [4]. The former focuses on the fault detection [5], while the latter devotes to evaluate the current health state of the system as well as to predict its remaining useful life (RUL). An accurate prediction of the RUL could help develop a more economical and effective maintenance policy.

The performance of a manufacturing system is typically affected by its degradation process. There are always some quality characteristics of systems (voltage, crack length, wear, etc.). Their values degrade over time and failures occur when their degradation paths exceed a predetermined critical threshold [6,7]. Monitoring system and collecting and analyzing the observations of one or more quality characteristics can inform the users about the state of the system and set the ground for RUL prediction and preventive maintenance policy planning. Beganovic and Söffker [8] developed modeling strategies capable to describe complex relations between measurable system variables, related system degradation and RUL. Zhao et al. [9] focused on remaining useful life prediction of aircraft engine in the same gradual degradation mode. Therefore, to obtain a precise RUL prediction, a system's degradation process should be thoroughly studied and properly modeled. If the degradation states can be directly identified and estimated by using discrete or continuous condition monitoring, we can make maintenance decisions based on the current and the predicted degradation state of the system. Consequently, the key part in RUL prediction and maintenance decision-making for a degrading system is degradation modeling.

The degradation phenomenon is commonly modeled by a kind of stochastic process [10], due to the variety and number of influencing factors (such as age, usage or environment). In previous studies, several probabilistic approaches have been adapted to capture the stochastic properties of degradation processes. Le et al. [11] used a non-homogeneous gamma process to model the system's degradation and estimate the remaining useful life distribution. Chen et al. [6] proposed a nonlinear generalized wiener process model and developed a joint multi-level classification and preventive maintenance model. However, these degradation models cannot precisely describe the temporal evolution of sequential data and give the real-time prediction combined with the latest collected data directly. Artificial neural network (ANN) [12] and Hidden Markov model (HMM) are two alternative approaches to make up these shortcomings. In contrast with HMMs, it is difficult to

[*] Corresponding author.
E-mail address: pes@sjtu.edu.cn (E. Pan).

https://doi.org/10.1016/j.ress.2017.09.002
Available online xxx
0951-8320/© 2017 Elsevier Ltd. All rights reserved.

find the degradation law and observe the health state transition when we employ ANN-based methods. Actually, we need this information to build a connection between mathematical models and the underlying physical degradation processes. Therefore, HMMs are utilized in this research.

HMM is a general statistical modeling technique for sequences or time series and has been successfully applied in temporal pattern recognition, such as speech and handwriting recognition [13,14]. A HMM normally consists of two stochastic processes: A Markov chain where the hidden states representing stochastic sequences cannot be observed directly, and an observation process which relates to the states by some probability distributions. Generally speaking, HMMs have finite numbers of hidden states with discrete or continuous observations [15]. Two assumptions are made by the model. The first, called the Markov assumption, states that the current state is dependent only on the previous state, this represents the memory of the model. The independence assumption states that the output observation at the current moment is dependent only on the current state, it is independent of previous observations and states. In recent years, HMMs in diagnosis and prognosis has gained increasing attention. HMMs can depict the system's health condition with several meaningful states, such as “healthy”, “good”, “normal”, “unhealthy” and “failure”. Thereby, it can give concise and straightforward explanations for maintenance [16,17]. Many studies have used HMMs as an efficient tool for degradation modeling, as well as the RUL prediction [18–20]. Wang and Wang [21] presented a continuous HMM method to solve the problems of tool condition monitoring and remaining useful life prediction. Le et al. [22] proposed a multi-branch HMM framework for remaining useful life estimation of systems under multiple deterioration modes. Ghasemi et al. [23] developed a method based on HMMs to calculate the reliability function and the mean residual life of a piece of equipment. Yu [24] proposed an adaptive-learning-based method for machine faulty detection and health degradation monitoring with an adaptive HMM. Cholette and Djurdjanovic [25] described a novel data-driven approach based on characterizing the degradation process via a set of operation-specific HMMs to monitoring of systems operating under variable operating conditions. Geramifard et al. [26] used multiple physically segmented HMM with continuous output for tool wear monitoring. Moreover, HMMs have also been used to degradation modeling and then for condition-based maintenance. Zhang et al. [27] demonstrated a Bayesian estimation scheme for the HMM parameters, as well as a method for condition-based monitoring and maintenance.

As for degradation modeling and prognosis, most existing applications of HMMs in the previous papers assume an independent scalar observation distribution associated with each state, namely conditional independence. That is, there are no correlations among the observations. However, this assumption is frequently invalid in many degradation processes. With the growth of automation in manufacturing, the quality characteristics of systems are being measured at higher rates and the degradation data is more likely to be auto-correlated [28]. For example, the crack propagation rate will be higher and the crack length will grow rapidly when the current crack length is larger. Due to this autocorrelation, traditional HMM-based methods cannot describe the degradation processes exactly and make an accurate real-time prediction. Therefore, it is realistic to take the autocorrelation of degradation data into account. Few papers have addressed this type of degradation process. Tang et al. [29] used an autoregressive model with time effect to describe the system degradation. This study combined both the system current age and the previous state observations. Adjengue et al. [30] dealt with independent as well as correlated maintenance observations by HMM. Though the autocorrelation has been investigated in those literatures, the studies about HMMs with auto-correlated observations for degradation modeling and the RUL prediction have not been deeply explored. Actually, they are urgently required for their potential importance. Moreover, missing data and noise occur frequently in various signal processing and statistical applications. Yu and Kobayashi [31] proposed a hidden semi-Markov model with missing observations and multiple observation sequences for mobility tracking. Palomäki et al. [32] described a binaural auditory model for recognition of speech in the presence of spatially separated noise intrusions, under small-room reverberation conditions. However, there is not much work considering this problem in HMM-based degradation models.

In this paper, we proposed an approach based on HMM with auto-correlated observations (HMM-AO) to model the degradation processes. Unlike standard HMMs, the HMM-AO model considers the probability of emitting an observation in real time. The observation not only depends on the corresponding hidden system state, but also on the previous observations. The autocorrelation property of the observations is characterized by coefficient matrices. Then these auto-correlated observations are used to feed our approach based on the HMM-AO model, and to produce the RUL predictions. A novel algorithm based on the Expectation maximum method is developed to estimate the unknown parameters. Furthermore, missing data and noise that accumulate over time are also considered by modifying the proposed model. The purpose of this paper is to present two RUL prediction methods based on the HMM-AO model. One is State-based RUL prediction method, which calculates the remaining number of time steps to reach the final state. The other is Observation-based prediction method, which estimates the residual time of the observation of degradation path first crossing the critical threshold. Thereby, we can obtain accurate predicted values for subsequent maintenance planning. The effectiveness of these proposed methods will be demonstrated by a real case study of a LED degradation dataset from Hamada et al. [33]. Furthermore, these data are used to illustrate the advantages and of the HMM-AO approach by comparing with the standard HMM. We note that HMMs with auto-correlated observations have not been utilized to model the system degradation and predict the remaining useful life. Besides, we propose a parameter estimation algorithm and two RUL prediction methods with this novel HMM-AO method.

In addition, in the framework of preventive maintenance, a RUL-based maintenance policy is developed with observations at equidistant time epochs to illustrate the application of the RUL prediction. If the predicted RUL of the system reaches a prefixed preventive maintenance threshold, the system would be preventively replaced. If its predicted RUL is higher than the threshold, the decision is postponed until next inspection [34]. Moreover, a corrective replacement is carried out, if the system's degradation value exceeds its failure threshold. The objective is to find the optimal preventive threshold to initiate a preventive replacement with the minimum expected maintenance cost per unit time [35]. Finally, using the fatigue-crack-growth data from Lu and Meeker [36], the proposed policy is compared with a conventional condition-based policy where only the current state of the system [37] is considered.

The major contributions and innovations of this study include:

(1) A hidden Markov model with auto-correlated observations (HMM-AO) is developed to handle the degradation modeling of manufacturing systems. In comparison with the standard HMMs, the HMM-AO is capable of higher fitting degree for degradation processes.
(2) A novel algorithm based on the expectation maximum is presented to estimate the unknown parameters of the proposed model. The rationality of this algorithm is proved.
(3) To illustrate the adaptability of HMM-AO, we discuss how to account for missing data and noise that accumulate over time in the proposed model.
(4) Two remaining useful life prediction methods based on the HMM-AO model are developed. The reliability function is derived. Predictive values of more accuracy can be obtained, since the autocorrelation of observations has been considered and the temporal evolution of degradation processes has been described well.
(5) An improved maintenance policy is developed based on the results of RUL prediction.

The remainder of this paper is organized as follows: Section 2 describes the degradation process by the HMM-AO model. Section 3 develops two RUL prediction methods. In Section 4, a real case study is provided to illustrate the effectiveness of the prediction methods. Section 5 develops a RUL-based maintenance policy and gives an example for comparison. Finally, Section 6 concludes the paper.

2. Degradation modeling

2.1. Hidden Markov models

An HMM is a probabilistic model which has a finite number of hidden states and a set of discrete or continuous observations. Each state is characterized by a transition probability set based on the first order Markov chain and by an emission probability distribution of observations. Formally, a standard HMM is defined by the following elements:

- A finite set of hidden states $\boldsymbol{S} = \{S_1, S_2, ..., S_N\}$. N is the number of the hidden states. It can be determined by a model selection technique such as cross validation. Note that q_t represents the hidden state at time t, where $q_t \in \boldsymbol{S}$.
- A state transition probability distribution, $\boldsymbol{A} = \{a_{ij}\}$, where

$$a_{ij} = P\left(q_{t+1} = S_j | q_t = S_i\right),\ \ 1 \le i, \mathrm{j} \le \mathrm{N},\ \ \sum_j a_{ij} = 1, \tag{1}$$

- An initial state probability distribution $\boldsymbol{\pi} = \{\pi_i\}$, where

$$\pi_i = P\{q_1 = S_i\},\ \ 1 \le i \le \mathrm{N},\ \ \sum_i \pi_i = 1. \tag{2}$$

- An observation sequence measured at regular time intervals,

$$O = \left(o_1, ..., o_T\right), \tag{3}$$

where T is the length of the observation sequence. It should be infinite if the observation space is continuous.
- An observation probability distribution related to the hidden states, $\boldsymbol{B} = \{b_j(\cdot)\}$.

Obviously, a standard HMM requires the specifications of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\pi}$. For convenience, the whole elements can be abbreviated as a triplet

$$\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}). \tag{4}$$

In practice, the degradation processes are usual irreversible. That is, a system cannot recover from its current state to a past state. Hence, the left–right HMM is a proper option for modeling degradation processes, whose failure rate increase as time passes. The state transition probability matrix of the model takes the following form:

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & ... & 0 \\ 0 & a_{22} & a_{23} & 0 & ... & 0 \\ ... & ... & ... & ... & & ... \\ 0 & 0 & 0 & ... & a_{(N-1)(N-1)} & a_{(N-1)N} \\ 0 & 0 & 0 & ... & 0 & 1 \end{bmatrix}. \tag{5}$$

2.2. Auto-correlated observations of degradation

The observations of an HMM might be discrete or continuous. The relation between the hidden states and observations is often built by some probabilistic links. When the HMMs are applied in degradation modeling, the observations are defined as the measured values of the degradation paths in this research. Since the degradation paths are usual continuous, the distribution of observations in the degradation processes can be specified using a parametric model family. Different from previous studies that generally assume independence among the observations, we consider auto-correlated Gaussian observations here. The probability of emitting an observation at the current time not only depends on the corresponding hide state, but also on the previous observations.

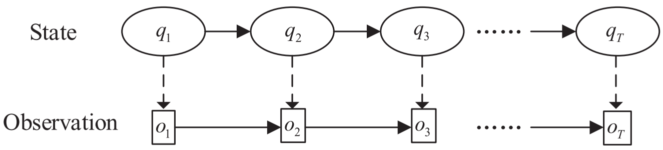

Fig. 1. Construction of a special case of the HMM-AO model.

Given the historical degradation of such a system and the state $q_t = S_i$, the mean of the conditional distribution for o_t (m-dimension) is a linear function of the d previous observations $o_{t-d}, ..., o_{t-1}$ added to a fixed offset. If m equals 1, the degradation process is associated with only one quality characteristic. If m is greater than or equal to two, there are two or more quality characteristics related to the degradation process which can jointly define the failure of the system [38,39]. The value of d can be determined by correlation analysis of the observation sequences. Therefore, the conditional distributions for observations are m-dimensional Gaussian as follows:

Mean: Let $\mu_i(t)$ denote the mean of the observation o_t in state S_i, then the lth component of $\mu_i(t)$ can be expressed as follows

$$\mu_{i,l}(t) = \varsigma_{i,l} + \sum_{\tau=1}^{d} c_{i,\tau,l}\, o_{t-\tau}(l), 1 \le l \le m, 1 \le i \le N, \tag{6}$$

where $o_{t-\tau}(l)$ is the lth component of $o_{t-\tau}$, $c_{i,\ \tau,\ l}$ and $\varsigma_{i,\ l}$ are the constant coefficients.

Covariance: The covariance is a state dependent symmetric positive definite $m \times m$ matrix Σ_i.

Construct a vector $x(t) = (o_{t-1}, ..., \mathrm{o}_{t-d}, 1)'$ consisting of d previous observations and a constant one. Eq. (6) can be rewritten as

$$\mu_i(t) = C_i x(t), \tag{7}$$

where C_i is an $m \times (md+1)$ matrix consisting of $\{c_{i,\ \tau,\ k}\}$ and $\{\varsigma_{i,\ k}\}$. At the conclusion of the above descriptions, the HMM-AO model has two key characteristics: the states follow a Markov process and the conditional observation distribution given the state S_i is multivariate Gaussian with mean $C_i x(t)$ and covariance Σ_i, i.e.,

$$P\left(q_{t+1} | q_1, ..., \mathrm{q}_t, \mathrm{o}_1, ..., \mathrm{o}_t\right) = P\left(q_{t+1} | \mathrm{q}_t\right), \tag{8}$$

$$\begin{aligned} b_i\left(o_t \middle| o_{(t-d):(t-1)}\right) &= P\left(o_t | \mathrm{o}_{t-d}, ..., \mathrm{o}_{t-1}, q_t = S_i\right) \\ &= \frac{1}{\sqrt{(2\pi)^m |\Sigma_i|}} \exp\left(-\frac{1}{2}\left(o_t - C_i x(t)\right)' \Sigma_i^{-1} \left(o_t - C_i x(t)\right)\right). \end{aligned} \tag{9}$$

The model demonstrates the correlation between an observation and its predecessors in two ways. The observation o_t is not only related to the d previous observations through the coefficient matrix C_i, but also related to all previous observations through the current state S_i. Fig. 1 shows a special case of the construction of HMM-AO when the observation o_t is only related to the previous observation o_{t-1}. Furthermore, if d is equal to zero, that is, the current observation has no relation to the previous observations, the HMM-AO model would degenerate into a standard HMM model. This is as expected since any properly developed model should cover the basic model as its special case.

Although Eq. (6) is a linear expression, the proposed auto-correlated Gaussian distribution can be extended to describe more complicated processes. In fact, many real degradation processes are nonlinear and non-monotonic. Usually the Gaussian mixture model (GMM) which can approximate, arbitrarily closely, any finite, continuous density function, are used to fit the observations [40]. Under the assumption of autocorrelation, the additional unknown parameters of GMM to be estimated would increase the computational complexity. The cost of the increased computation tends to make the techniques not worth using. Thus, an appropriate and cost-effective model for the fitting of these processes is preferred. To address this problem and improve the applicability of the proposed model, the following alternative manner is presented. The degradation processes are discretized into several segments within a "short time" observation interval, and are viewed as a direct concatenation of these smaller short time segments [4142]. Each such segment of observation is approximated as a linear path and can be individually modeled by Eq. (9). In other word, when the observation interval is short enough relative to the useful lifetime of the system, the sequential linear paths can be used to approximate the overall degradation process precisely. Besides, the stochastic properties of Gaussian can capture the uncertainty of the observations. The duration of a short-time interval is determined empirically in most physical systems. Since the systems' reliability and lifetime become more and more higher nowadays, a suitable observation interval can be fixed simply. The temporal variations between two sequential segments can be characterized by HMM. Unlike the autoregressive model, Eq. (6) in Gaussian distribution can also take random effects into account. Therefore, the HMM-AO models can describe more complex and diverse degradation processes and have a wider range of applicability.

2.3. Parameter estimation

The Baum–Welch formula is the common method for the parameter estimation of HMMs. However, since auto-correlated observations are cooperated into HMMs, the conventional algorithm needs to be improved. Here, a novel algorithm based on the Expectation maximum method (NEM) is proposed to estimate the unknown parameters $\lambda = (\boldsymbol{A}, C, \Sigma, \boldsymbol{\pi})$.

Before the implementation of the NEM algorithm, two probabilities called the forward variable $\alpha_t(i)$ and the backward variable $\beta_t(i)$ need to be calculated. The definitions and computation of the two variables are similar to those of standard HMMs in Le et al. [22]. Their initial values are $\alpha_{d+1}(i) = \pi_i b_i(o_{d+1}|o_{1:d})$ and $\beta_T(i) = 1$, respectively. Consequently, given the observation sequence O and the model parameter λ, two additional variables can be derived based on the forward and backward variables. One is the probability of being in state S_i at time t, denoted by $\gamma_t(i)$. The other is the probability of being in state S_i at time t and state S_{i+1} at time $t+1$, denoted by $\xi_t(i, j)$. Their explicit expressions in terms of the forward-backward variables are also similar to those of standard HMMs.

Consider a set of K observation sequences obtained from the historical degradation of K identical and available systems, where an observation sequence $O^k = (o_1^k, \ldots, o_{T_k}^k)$ is measured at regular time intervals. Hence, the training dataset can be denoted by $\boldsymbol{O} = \{O^k\}_k$, where the elements are independent and identically distributed. Using the definitions and dataset above, the NEM algorithm is presented to execute the following E-step and M-step iteratively.

E-step: Given the observed data $\boldsymbol{O}$ and the current estimates of the unknown parameters $\lambda = (\boldsymbol{A}, C, \Sigma, \boldsymbol{\pi})$, the values of $\gamma_t(i)$ and $\xi_t(i, j)$ can be obtained. For the re-estimation of C and Σ, the auxiliary function Q is given by taking the conditional expectation of the log-likelihood of all observation sequences

$$Q\left(\hat{\lambda}, \lambda\right) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T_k}\gamma_t^k(i)\Big[\ln\left(\left|\Sigma_i{}^{-1}\right|\right) - m\ln(2\pi) - \left(o_t^k - C_i x^k(t)\right)'\Sigma_i{}^{-1}\left(o_t^k - C_i x^k(t)\right)\Big]. \tag{10}$$

M-step:

Like the Baum–Welch algorithm, the state transition matrix and the initial state probabilities can be re-estimated by

$$\hat{a}_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\xi_t^k(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\gamma_t^k(i)}, \tag{11}$$

$$\hat{\pi}_i = \frac{\sum_{k=1}^{K}\gamma_1^k(i)}{K}. \tag{12}$$

Since the auxiliary function Q in Eq. (10) is complex, it is difficult to implement the maximization of Q with the common search algorithms. As for the re-estimation of the remaining unknown parameters, we can maximize Eq. (10) alternatively by using operations on vectors and matrices as the following steps:

1) Create matrices X_i^k with columns $x^k(t)\sqrt{\gamma_t^k(i)}$, where $\boldsymbol{x}^k(t)$ is defined in Eq. (7).
2) Create matrices O_i^k with columns $o_t^k\sqrt{\gamma_t^k(i)}$.
3) Solve

$$\hat{C}_i^k = \arg\min_W \left|O_t^k - W X_i^k\right|^2. \tag{13}$$

The singular value decomposition (SVD) methods are applied here due to its stability and ability in diagnosis. Suppose that X_i^k has an SVD, i.e., $X_i^k = U\Lambda V'$, where U is an $(md+1)\times(md+1)$ real or complex unitary matrix, Λ is an $(md+1)\times T_k$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and V is an $T_k \times T_k$ real or complex unitary matrix. The diagonal entries of Λ are the singular values of X_i^k. Then,

$$\begin{aligned}\hat{C}_i^k &= \arg\min_W \left|\left(O_t^k V - W U\Lambda\right)V'\right|^2 \\ &= \arg\min_W \left|O_t^k V - W U\Lambda\right|^2\end{aligned}. \tag{14}$$

Obviously, the value of $\hat{C}_i^k$ is the least square solution of Eq. (18), i.e.,

$$\hat{C}_i^k = O_t^k V \Lambda^+ U', \tag{15}$$

where Λ^+ is the generalized inverse matrix of Λ^+. Hence, the estimate of C_i can be expressed as follows

$$\hat{C}_i = \frac{\sum_{k=1}^{K}\hat{C}_i^k}{K}. \tag{16}$$

4) Calculate a new covariance matrix

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_t^k(i)\left(O_t^k - \hat{C}_i^k X_i^k\right)\left(O_t^k - \hat{C}_i^k X_i^k\right)'}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_t^k(i)}. \tag{17}$$

Then, return the estimates into the E-step and run the NEM algorithm iteratively until the solutions converge. The corresponding flowchart for the NEM algorithm is shown in Fig. 2.

To clarify the rationality and applicability of the re-estimation operations Eqs. (13)–(17) in the M-step, the following proposition is given.

Proposition 1. Given the current model parameters $\lambda = (\boldsymbol{A}, \mathrm{C}, \Sigma, \boldsymbol{\pi})$, the re-estimated parameters $\hat{\lambda} = (\hat{\boldsymbol{A}}, \hat{\mathrm{C}}, \hat{\Sigma}, \hat{\boldsymbol{\pi}})$ is computed by the right-hand sides of Eqs. (11)–(17). Then the model $\hat{\lambda}$ is more likely than model λ in the sense that $P(\boldsymbol{O}|\hat{\lambda}) > P(\boldsymbol{O}|\lambda)$, i.e., we have found a new model $\hat{\lambda}$ from which the observation sequences are more likely to have been produced.

Proof. With the definitions $\hat{C}_i^k = \arg\min_W |O_t^k - W X_i^k|^2$ as shown in Eq. (13), we can know that $|O_t^k - \hat{C}_i^k X_i^k| \le |O_t^k - C_i^k X_i^k|$. The following inequality is thereby obtained:

$$\left|\gamma_t^k(i)\left(O_t^k - \hat{C}_i^k X_i^k\right)\left(O_t^k - \hat{C}_i^k X_i^k\right)'\right| \le \left|\gamma_t^k(i)\left(O_t^k - C_i^k X_i^k\right)\left(O_t^k - C_i^k X_i^k\right)'\right|. \tag{18}$$

Then, we have $|\hat{\Sigma}_i| \le |\Sigma_i|$ and

$$\left(o_t^k - \hat{C}_i x^k(t)\right)'\hat{\Sigma}_i^{-1}\left(o_t^k - \hat{C}_i x^k(t)\right) \le \left(o_t^k - C_i x^k(t)\right)'\Sigma_i{}^{-1}\left(o_t^k - C_i x^k(t)\right). \tag{19}$$

Expanding Eq. (19) yields

$$Q\left(\hat{\lambda}, \lambda\right) > Q(\lambda, \lambda). \tag{20}$$

Now denote the log likelihood of the observed data given the model parameter λ as

$$L(\lambda) = \ln(P(\boldsymbol{O}|\lambda)), \tag{21}$$

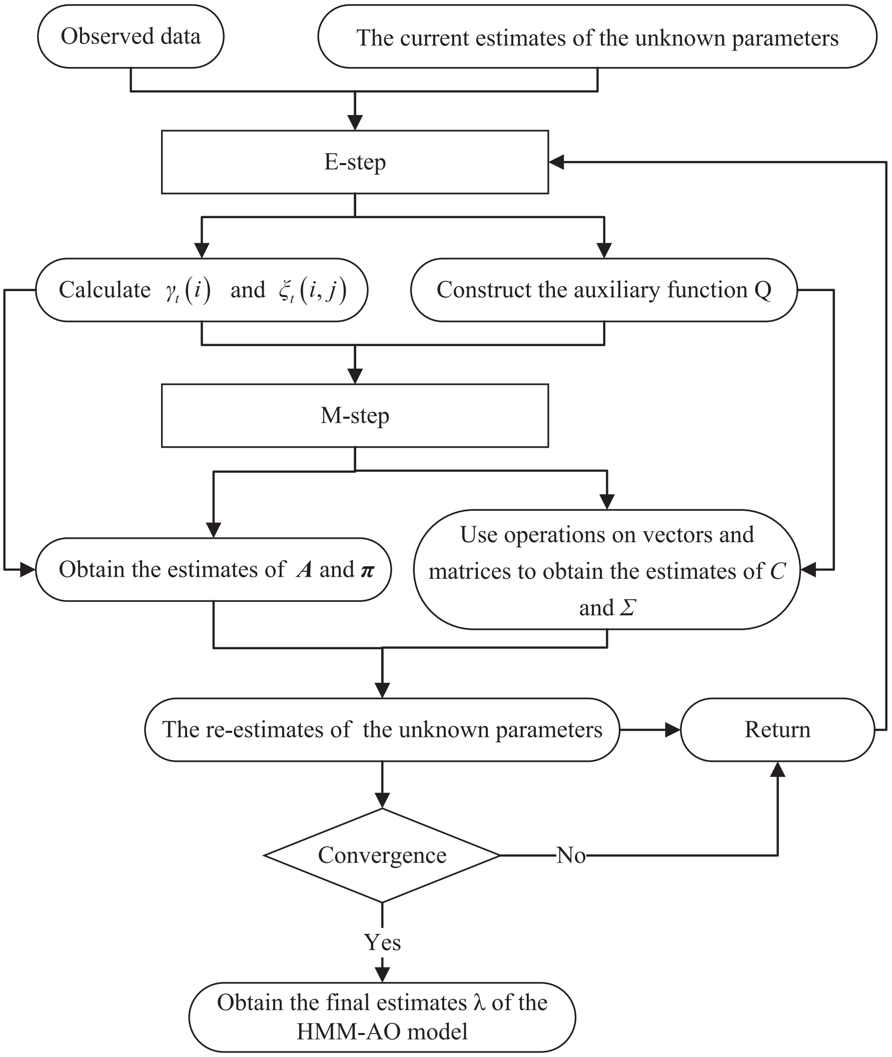
Observed data
The current estimates of the unknown parameters
E-step
Calculate $\gamma_t(i)$ and $\xi_t(i,j)$
Construct the auxiliary function Q
M-step
Obtain the estimates of A and π
Use operations on vectors and matrices to obtain the estimates of C and Σ
The re-estimates of the unknown parameters
Return
Convergence
No
Yes
Obtain the final estimates λ of the HMM-AO model

Fig. 2. Flowchart of the NEM algorithm.

and the cross entropy of the unobserved hidden states with respect to a model $\hat{\lambda}$ as

$$H\left(\hat{\lambda}\|\lambda\right) = -E_{P(\boldsymbol{S}|\boldsymbol{O},\lambda)}\left(\ln P\left(\boldsymbol{S}\middle|\boldsymbol{O},\hat{\lambda}\right)\right). \tag{22}$$

The auxiliary function Q can be rewritten as:

$$\begin{aligned} Q\left(\hat{\lambda},\lambda\right) &= E_{P(\boldsymbol{S}|\boldsymbol{O},\lambda)}\left(\ln P\left(\boldsymbol{S},\boldsymbol{O}\middle|\hat{\lambda}\right)\right) \\ &= E_{P(\boldsymbol{S}|\boldsymbol{O},\lambda)}\left(\ln P\left(\boldsymbol{S}\middle|\boldsymbol{O},\hat{\lambda}\right) + \ln P\left(\boldsymbol{O}\middle|\hat{\lambda}\right)\right), \\ &= L\left(\hat{\lambda}\right) - H\left(\hat{\lambda}\|\lambda\right) \end{aligned} \tag{23}$$

i.e., $L(\hat{\lambda}) = Q(\hat{\lambda},\lambda) + H(\hat{\lambda}\|\lambda)$. The fact that

$$H\left(\hat{\lambda}\|\lambda\right) \geq H(\lambda\|\lambda)\forall\hat{\lambda}, \tag{24}$$

with equality if $P(\boldsymbol{S}|\boldsymbol{O},\hat{\lambda}) = P(\boldsymbol{S}|\boldsymbol{O},\lambda)\forall\boldsymbol{S}$ is called the Gibbs inequality [43].

According to Eqs. (21)–(24), it is concluded that

$$\begin{aligned} L\left(\hat{\lambda}\right) - L(\lambda) &= \left(Q\left(\hat{\lambda},\lambda\right) + H\left(\hat{\lambda}\|\lambda\right)\right) - (Q(\lambda,\lambda) + H(\lambda\|\lambda)) \\ &= \left(Q\left(\hat{\lambda},\lambda\right) - Q(\lambda,\lambda)\right) - \left(H\left(\hat{\lambda}\|\lambda\right) - H(\lambda\|\lambda)\right) > 0 \end{aligned}. \tag{25}$$

Thus, we have $P(\boldsymbol{O}|\hat{\lambda}) > P(\boldsymbol{O}|\lambda)$ and the proposition follows. This implies that the NEM algorithm monotonically increases the likelihood.

2.4. Effects of missing data and noise

In practical engineering of degradation phenomenon, missing data and noise occur frequently. Due to their interference, the real degradation cannot be observed precisely and comprehensively. Accurate modeling of the observation process is therefore of great importance. To illustrate the adaptability of HMM-AO, we discuss how to account for missing data and noise that accumulate over time in this subsection.

Let us first introduce the modifications of the proposed model with the presence of missing data. If the observation o_t is missed, referring to Yu and Kobayashi [31] we can re-define the forward variable $\alpha_t(j)$ as

$$\begin{aligned} \alpha_t(j) &= \sum_z P\left(o_1,\ldots,o_{t-1},o_t = z, q_t = S_i|\lambda\right) \\ &= \left[\sum_{i=1}^N \alpha_{t-1}(i)a_{ij}\right]b_i\left(o_{t-1}\middle|o_{(t-d-1):(t-2)}\right) \end{aligned} \tag{26}$$

Similarly, we redefine the backward variable $\beta_t(i)$ as

$$\begin{aligned} \beta_t(i) &= P\left(o_{t+1},\ldots,o_T\middle|q_t = S_i,\lambda\right) \\ &= \sum_{j=1}^N a_{ij}b_j\left(o_{t+1}\middle|o_{(t-d):(t-1)}\right)\beta_{t+1}(j) \end{aligned}. \tag{27}$$

Next considering the noise effect on the observation process, the observation probability distribution in Section 2.2 should be modified. Denote y_t (m-dimension), o_t and e as the noisy observation at time t, the clean observation and noise, respectively. Then the noisy observation is given by

$$y_t = o_t + e, \tag{28}$$

where $e = (e_1,\ldots,e_m)$. As in Section 2.2, the conditional distributions for observation o_t given the state S_i is multivariate Gaussian distribution with mean $\mu_i(t)$ and covariance Σ_i. The components of noise e are regarded as white noise, i.e., e_l, $1 \le l \le m$, are Gaussian, independent, identically distributed random variables with zero mean and variance σ^2. Obviously, y_t is the sum of two multivariate Gaussian. For describing the distribution of y_t, we need to state the following proposition.

Proposition 2. If (1) o_t given the state S_i is multivariate Gaussian distribution with mean $\mu_i(t)$ and covariance Σ_i, (2) and the elements of e are Gaussian, independent, identically distributed random variables with zero mean and variance σ^2, then the conditional distribution of observation $y_t = o_t + e$ given the state S_i also follows multivariate Gaussian distribution with mean $\mu_i(t)$ and covariance $\Sigma_{yi} = \Sigma_i + \mathrm{diag}(\sigma^2, \ldots, \sigma^2)_{m\times m}$.

Proof. The mean of y_t given the state S_i can be computed as

$$E(y_t|S_i) = E(o_t|S_i) + E(e) = \mu_i(t). \tag{29}$$

The covariance is calculated as follows:

$$\begin{aligned}\Sigma_{yi} &= E\left(y_t - \mu_i(t)\right)'\left(y_t - \mu_i(t)\right)\\ &= E\left(\begin{bmatrix} o_t(1) + e_1 - \mu_{i,1}(t) \\ o_t(2) + e_2 - \mu_{i,2}(t) \\ \ldots \\ o_t(m) + e_m - \mu_{i,\mathrm{m}}(t)\end{bmatrix}\left[o_t(1) + e_1 - \mu_{i,1}(t), \ldots, o_t(m) + e_m - \mu_{i,\mathrm{m}}(t)\right]\right).\end{aligned} \tag{30}$$

Since the elements of e are Gaussian, independent, identically distributed random variables with zero mean and variance σ^2, we have

$$E(e_l e_j) = \begin{cases} 0, 1 \le \mathrm{l} \ne j \le m \\ \sigma^2, 1 \le \mathrm{l} = j \le m \end{cases}. \tag{31}$$

The element of Σ_{yi} at lth row and jth column, $1 \le l$, $j \le m$, can be simplified as

$$\begin{aligned}\Sigma_{yi}(l,\mathrm{j}) &= E\left[\left(o_t(l) + e_l - \mu_{i,1}(t)\right)\left(o_t(j) + e_j - \mu_{i,\mathrm{j}}(t)\right)\right]\\ &= E\left[\left(o_t(l) - \mu_{i,l}(t)\right)\left(o_t(l) - \mu_{i,1}(t)\right)\right] + E(e_l e_j)\\ &= \Sigma_i(l,j) + E(e_l e_j)\end{aligned}. \tag{32}$$

Then the covariance of y_t is ultimately expressed as follows

$$\Sigma_{yi} = \Sigma_i + \mathrm{diag}\left(\sigma^2, \ldots, \sigma^2\right)_{m\times m}. \tag{33}$$

Therefore, the result follows immediately. That is, the conditional observation distribution given the state S_i can be rewritten as follows

$$b_i\left(y_t\middle|y_{(t-d):(t-1)}\right) = \frac{1}{\sqrt{(2\pi)^m\left|\Sigma_{yi}\right|}}\exp\left(-\frac{1}{2}\left(y_t - C_i x(t)\right)'{\Sigma_{yi}}^{-1}\left(y_t - C_i x(t)\right)\right). \tag{34}$$

From Eq. (34), we can see that the observation distribution is still multivariate Gaussian despite considering noise. Then the unknown parameters of HMM-AO in presence of missing data and noise can be estimated using the NEM algorithm with the above redefinitions.

3. Two methods for remaining useful life prediction

Typically, a system is regarded as in a failure when its degradation value crosses the critical threshold. In details, the lifetime of the system usually evolves through several distinct health states prior to reaching failure. We can identify N distinct sequential hidden states for a failure mechanism. That is, the research classifies its health states into $(N-1)$ levels: no-defect (state S_1), level-one defect (state S_2), …, level – $(N-2)$ defect (state S_{N-1}). The final state S_N means a failure.

Based on the discussions above, we develop two methods of the RUL prediction: (1) calculating the remaining number of time steps to reach the final state S_N and (2) estimating the residual time that the observation of degradation path first crosses the critical threshold ω. We shall label the former as "State-based RUL prediction" and the latter as "Observation-based RUL prediction". Fig. 3 shows the flowchart of the RUL prediction.

3.1. State-based RUL prediction method

Given a new observation sequence $O_{1:T} = (o_1, \ldots, o_T)$ and a trained HMM-AO model λ, we estimate the remaining number of time steps τ to first arrive the failure state S_N from the current time T. The remaining useful life can be considered as a random variable with a conditional probability distribution. Accordingly, the RUL at time T can be defined as follows

$$RUL(T) = \inf\left\{\tau > 0 : q_{T+\tau} = S_N\middle|O_{1:T}, \lambda\right\}. \tag{35}$$

The definition is reasonable under the assumption of "short time" observation interval as in Section 2.2. The time the failed system already spent in state S_N can be neglected. Namely, every failure occurs at the observation moment approximately. From Eq. (35), the RUL is regarded as a discrete random variable and its conditional probability on the current state is

$$\begin{aligned}r_i^\tau &= P\left(RUL(T) = \tau\middle|q_T = S_i, i < N\right)\\ &= P\left(q_{T+\tau} = S_N, q_{T+\tau-1} \ne S_N, \ldots, q_{T+1} \ne S_N\middle|q_T = S_i, i < N\right)\end{aligned}. \tag{36}$$

Under the assumption of Eq. (5), the RUL can be recursively calculated as follows:

- When $q_T = S_{N-1}$,

$$r_{N-1}^\tau = \begin{cases} \mathrm{a}_{(N-1)N}, & \tau = 1 \\ \mathrm{a}_{(N-1)(N-1)} r_{N-1}^{\tau-1}, & \tau \ge 2 \end{cases}. \tag{37}$$

- When $q_T = S_i$, i ≤ N – 2,

$$r_i^\tau = \begin{cases} 0, & \tau = 1 \\ a_{ii} r_i^{\tau-1} + a_{i(i+1)} r_{i+1}^{\tau-1}, & \tau \ge 2 \end{cases}. \tag{38}$$

Then, the predicted value of the RUL at time T is given by

$$RUL(T) = \sum_{\tau=1}^{\infty}\sum_{i=1}^{N-1} r_i^\tau \cdot \gamma_T(i) \cdot \tau. \tag{39}$$

Proposition 3. Given a new observation sequence $O_{1:T} = (o_1, \ldots, o_T)$ and a trained HMM-AO model λ, the predicted value of the RUL at time T can be can be demonstrated as shown in (39).

Proof. According to the definition of system failure, the reliability function at time T is given by

$$R(T) = P(t > T) = P\left(q_T = S_i, \mathrm{i} < \mathrm{N}\middle|\lambda\right). \tag{40}$$

For the system which has not failed by time T, if $O_{1:T}$ is known, the conditional reliability function of surviving beyond time $T + \tau$ is given by

$$\begin{aligned}R\left(T + \tau\middle|O_{1:T}, \lambda\right) &= P\left(t > T + \tau\middle|t > T, O_{1:T}, \lambda\right)\\ &= \sum_{i=1}^{N-1} P\left(t > T + \tau\middle|q_T = S_i, \mathrm{i} < \mathrm{N}\right) \cdot P\left(q_T = S_i, \mathrm{i} < \mathrm{N}\middle|O_{1:T}, \lambda\right)\\ &= \sum_{i=1}^{N-1} r_i^\tau \cdot \gamma_T(i).\end{aligned} \tag{41}$$

By deriving the expectation of the discrete random variable τ, the RUL at time T can be calculated as follows:

$$\begin{aligned}RUL(T) &= E(\tau) = \sum_{\tau=1}^{\infty}\sum_{i=1}^{N-1} R\left(T + \tau\middle|O_{1:T}, \lambda\right) \cdot \tau\\ &= \sum_{\tau=1}^{\infty}\sum_{i=1}^{N-1} r_i^\tau \cdot \gamma_T(i) \cdot \tau.\end{aligned} \tag{42}$$

Based on Eq. (35), τ should be a positive integer and its lower bound is one-time step. Therefore, the result follows immediately.

Remark 1. In the calculation of Eq. (39), the range of τ need to be prefixed. With the increase of τ, the conditional probabilities r_i^τ ($i = 1, 2, \ldots, \mathrm{N}$) decrease gradually until close to zero. If we determine a reasonable upper limit of τ which makes r_i^τ equal to zero under a given relative tolerance, a precise prediction result can be obtained.

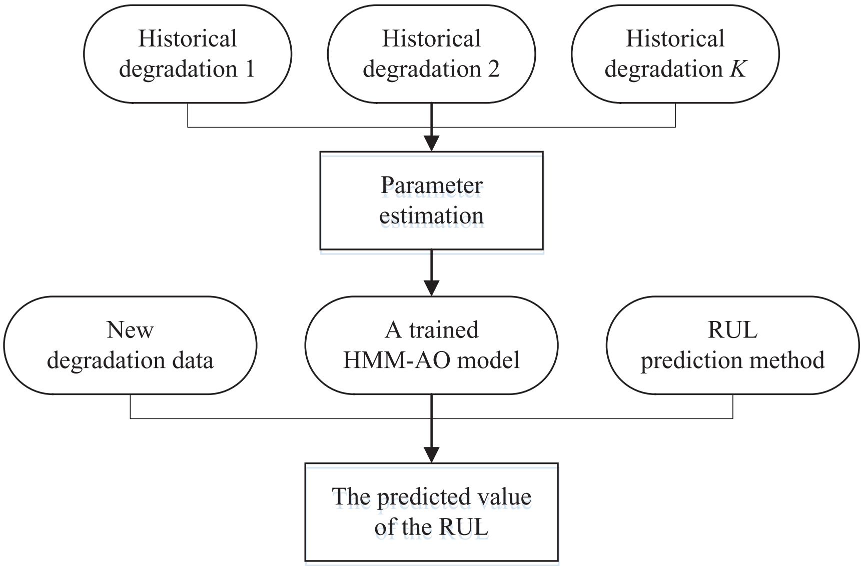

Fig. 3. Flowchart of the prediction of the RUL.

3.2. Observation-based RUL prediction method

Similarly, given a new observation sequence $O_{1:T} = (o_1, ..., o_T)$ and a trained HMM-AO model λ, the "Observation-based RUL prediction" method is to calculate the residual time τ that the degradation path first passes the critical threshold ω. Therefore, the remaining useful life at time T can be expressed as follows

$$RUL(T) = \inf\left\{\tau > 0 : o_{T+\tau} \ge \omega \middle| O_{1:T}, \lambda\right\}. \tag{43}$$

To obtain the predicted value of $RUL(T)$, the estimates of observations $o_{T+\tau}$ should be firstly calculated. A recursive method based on the expectations of the observations is developed here to derive the expressions as follows:

- Initialization: the conditional probability distribution of the observation o_{T+1} can be given by

$$\begin{aligned} P\left(o_{T+1}\middle|o_1, ..., o_T, \lambda\right) &= \sum_{i=1}^{N} P\left(o_{T+1}\middle|q_T = S_i, o_{1:\mathrm{T}}, \lambda\right) P\left(q_T = S_i\middle|o_{1:T}, \lambda\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} P\left(o_{T+1}\middle|q_{T+1} = S_j, o_{(T+1-d):T}, \lambda\right) P\left(q_{T+1} = S_j\middle|q_T = S_i\right)\gamma_i(T). \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \gamma_i(T) a_{ij} b_j\left(o_{T+1}\middle|o_{(T+1-d):T}\right) \end{aligned} \tag{44}$$

Thus, the estimate of o_{T+1} can be expressed by its expectation as follows

$$\hat{o}_{T+1} = \int o_{T+1} P\left(o_{T+1}\middle|o_1, ..., o_T, \lambda\right) do_{T+1}. \tag{45}$$

- Induction: the conditional probability distribution of the observation $o_{T+\tau}$ can be given by

$$P\left(o_{T+\tau}\middle|o_1, ..., o_{T+\tau-1}, \lambda\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} \gamma_i(T+\tau-1) a_{ij} b_j\left(o_{T+\tau}\middle|o_{(T+\tau-d):(T+\tau-1)}\right). \tag{46}$$

Based on the estimates of the τ previous observations, the estimate of $o_{T+\tau}$ can be expressed by its expectation as follows

$$\hat{o}_{T+\tau} = \int o_{T+\tau} P\left(o_{T+\tau}\middle|o_1, ..., o_T, \hat{o}_{T+1}, ..., \hat{o}_{T+\tau-1}, \lambda\right) do_{T+\tau}. \tag{47}$$

Therefore, the predicted value of $RUL(T)$ is estimated by

$$RUL(T) = \inf\left\{\tau > 0 : \hat{o}_{T+\tau} \ge \omega, \hat{o}_{T+\tau-1} < \omega\right\}. \tag{48}$$

Remark 2. The bounds of integration in Eq. (47) should be ascertained according to the possible span of observations. The predicted value of $RUL(T)$ is the estimated number of time steps that the degradation path first passes the critical threshold. Furthermore, the time steps are discrete and the degradation paths and the residual life are continuous. Hence, the "short time" observation interval can make the estimation accuracy higher. Compared with the "State-based RUL prediction" method, the "Observation-based RUL prediction" method is more accurate. However, due to the calculation of the expectations of observations, the computational complexity of the latter method is also higher. This means that we need to compromise between accuracy and complexity in the choice of these two prediction methods.

4. Case study

To evaluate the performance of the proposed approaches, a case study is carried out to train the HMM-AO models and predict the RULs. The case is based on real LED degradation data presented by Hamada et al. [33]. A part of the data is used in this study. The advantages of the "State-based RUL prediction method" and "Observation-based RUL prediction method" are demonstrated in comparison with each other. Moreover, other comparisons between the HMM-AO models and standard HMM are also conducted. The point is that we can use these contrasts to illustrate the autocorrelation of the dataset and the fitting accuracy of the HMM-AO models.

4.1. Validation of the RUL predication methods

Generally, the whole dataset is divided into training and test dataset. The unknown parameters of the HMM-AO models are estimated based on the training dataset. The test dataset is used to predict the RUL of each sample and then verify the predictive accuracy of the HMM-AO model. For the training dataset, the measurements are complete from the starting time until failure, and the trajectories are truncated once the failures occur. This is to ensure that the final state is the failure state. Moreover, to evaluate the performance of the proposed prediction methods, the real failure times are given for the comparison with the predicted values of RUL. The dataset used here consists of 19 degradation historical trajectories of LEDs. The relative luminosity (proportion of initial luminosity for LEDs) degrading over time was measured at every h = 336 h up to $T = 9744$ h. A failure is defined when the relative luminosity drops to 0.55, that is, 55% of initial luminosity. Fig. 4 shows the degradation

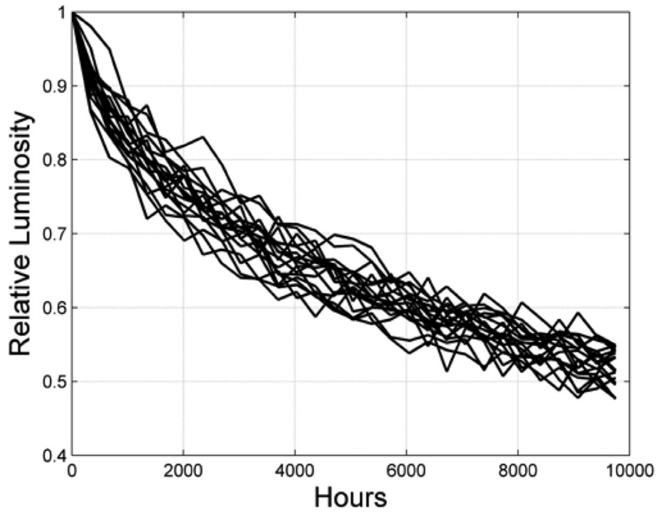

Fig. 4. Degradation paths in terms of the relative luminosity of LEDs.

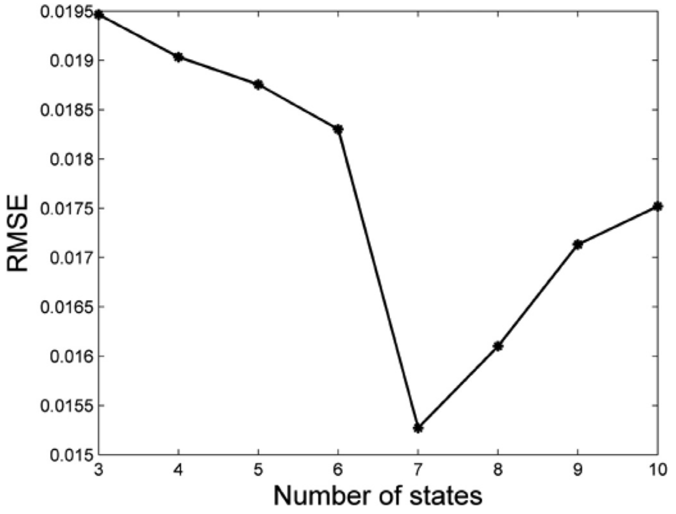

Fig. 5. The RMSE between the predicted and real observations versus number of states.

paths of the LEDs. These trajectories need to be truncated for the model training.

To illustrate the application of the proposed methods, the value of d is assumed to be one for convenience throughout the section. This is reasonable because the observation o_t is related to the previous observation o_{t-1} and then o_{t-1} is related to the observation o_{t-2}. Therefore, the current observation can relate to all previous observations through the coefficient matrices even d is equal to one. What we need to do then is to find the most appropriate parameters. Before the training, the number of hidden states of the HMM-AO model, which affects the generalization of the models, should be fixed first. There is no appropriate formula to the determination of the number of states. In this research, we select cross-validation method to optimize it. The root mean square error (RMSE) between the estimated observation of HMM-AO and the real observation is used as the criterion

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} \left(o_t - \hat{o}_t\right)^2}{T}}. \tag{49}$$

Normally, the smaller RMSE, the better the corresponding estimation. Fig. 5 shows a plot of prediction RMSE versus number of states. Here, it can be seen that the RMSE is somewhat insensitive to the number of states, achieving a local minimum at N=7 from this optimization.

Table 1
RMSE of the observation sequence prediction of HMM-AO considering the estimation bias.

ρ_1	ρ_2	ρ_3	RMSE	ρ_1	ρ_2	ρ_3	RMSE
5%	5%	5%	0.017	0	−5%	5%	0.044
5%	0	0	0.016	−5%	5%	−5%	0.016
5%	−5%	−5%	0.043	−5%	0	5%	0.016
0	5%	0	0.017	−5%	−5%	0	0.044
0	0	−5%	0.016	0	0	0	0.016

After the number of states is determined, the HMM-AO model is trained and then used to predict the RULs of samples in the test dataset.

Fig. 6 gives examples of the predicted RUL for two data histories by using the proposed prediction methods. The real RUL is also displayed in Fig. 6 to show the efficiency of these two methods. From the results, it is clear that the predictive accuracy of the State-based RUL prediction method is lower than that of the Observation-based RUL prediction method. This is because the autocorrelation is considered in the predictions of observations and the uncertainty of the event that the degradation path first crosses the critical threshold is lower. The calculation of the remaining number of time steps to reach the final state only take the state transition but without autocorrelation between observations into account. Fig. 6 also suggests that with more degradation observations available, the predicted values of RUL obtained by the State-based RUL prediction method become more accurate.

However, another phenomenon in Fig. 6 is equally worthy of attention: the predicted RULs from the Observation-based RUL prediction method show jagged appearance and overestimate the real RULs sometimes to some extent. Furthermore, the closer the failure, the higher the volatility of the predicted RUL. Due to the jagged degradation paths in Fig. 4, the predictions of observations are not smooth and leading to a jagged predicted RUL curve. Actually, if the degradation path was strictly monotone, the predicted values of the RUL should be strictly monotone decreasing and thereby the accuracy would be improved. In this case, we can use an alternative method called ε- approximation to reduce volatility when using Eq. (26) to define a failure in the Observation-based RUL prediction method. Given a relative tolerance ε, modified Eq. (26) can be rewritten as follows

$$RUL(T) = \inf\left\{\tau > 0 : \hat{o}_{T+\tau} - \varepsilon \geq \omega, \hat{o}_{T+\tau-1} < \omega+\varepsilon\right\}. \tag{50}$$

In addition, the predicted RULs from State-based RUL prediction seem to smaller than the real RULs. This result highlights the efficiency of this method and shows the interest of its use in preventive maintenance where it is necessary to plan maintenance actions before the failure.

4.2. Sensitivity analysis in parameter estimation

In practice, due to insufficient training data, the estimated parameters $\hat{\lambda} = (\hat{\boldsymbol{A}}, \hat{C}, \hat{\Sigma}, \hat{\boldsymbol{\pi}})$ would depart from the true parameters $\lambda = (\boldsymbol{A}, C, \Sigma, \boldsymbol{\pi})$. Without loss of generality, we assume that ρ_1, ρ_2, ρ_3 denote the estimation bias for a_{ii}, C, π_1 (as the values of Σ and other parameters are too small, we do not consider their estimation bias). Table 1 displays the RMSE of the observation sequence prediction of HMM-AO under various combinations of $(1+\rho_1)\hat{a}_{ii}$, $(1+\rho_2)\hat{C}$, and $(1+\rho_3)\pi_1$. From these results, we can find that the prediction results tend to be robust to estimation bias, given that the bias is not too large.

4.3. Comparison between HMM-AO and standard HMM

In previous studies, the standard HMMs without considering the autocorrelation of observations are usually adopted to model the degradation processes. However, the previous degradation degree could affect the current degradation trend and thereby the correlation is produced between the observations. Hence, we use the HMM-AO model here to describe the auto-correlated degradation process. In order to demonstrate

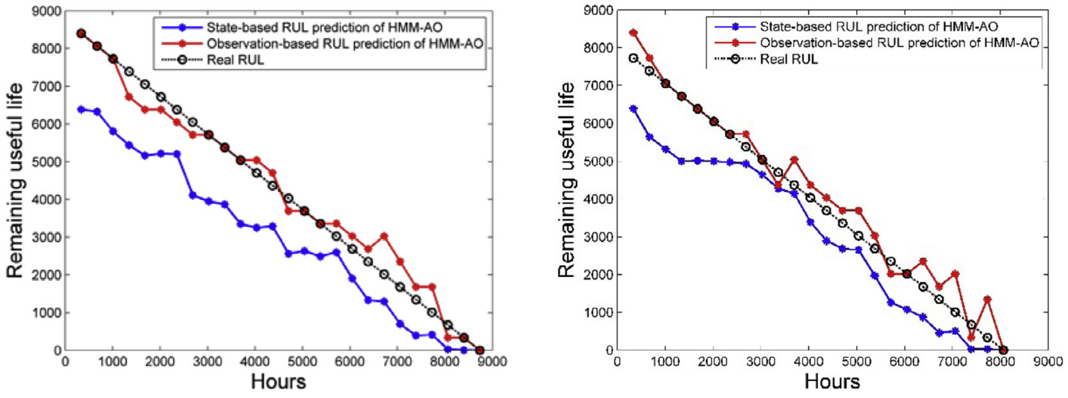

Fig. 6. The real remaining time to observe failure and the predicted values of the remaining useful life for two data histories by using two proposed prediction methods.

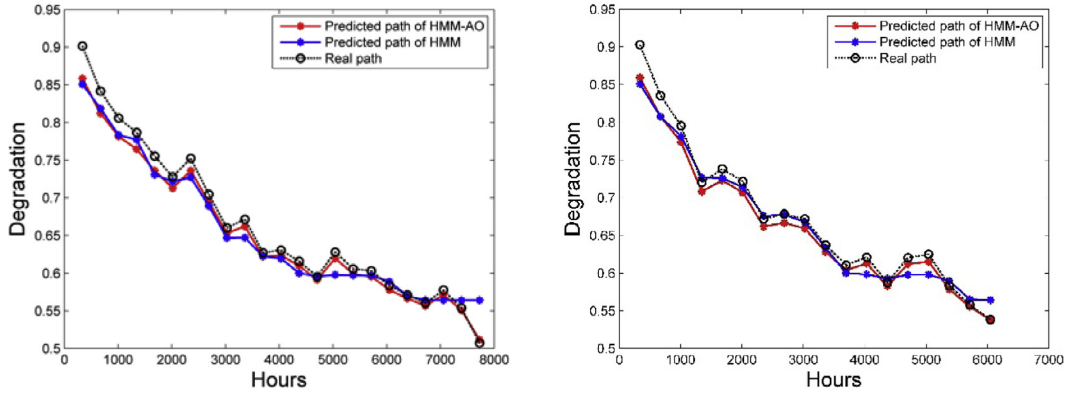

Fig. 7. The real degradation paths and the predicted degradation paths calculated by HMM-AO and HMM for two data histories.

the advantages and fitting accuracy of the proposed model, a series of comparisons between HMM-AO and HMM are provided.

The estimated parameters of both HMM-AO and HMM are reported in the Appendix. Through the contrast of the estimates of transition matrices, we can see that the transition probabilities from the current states to the next states of HMM-AO are higher than that of HMM. The variances of HMM-AO for observations are much smaller than that of HMM. In addition, the initial state probability of HMM-AO is more dispersed. These suggest that the HMM-AO model considering autocorrelation can distinguish states and observations more precisely and thereby lead a better modeling effect. Fig. 7 presents two examples of the predicted degradation paths obtained by HMM-AO and HMM which are compared with the real degradation paths. These examples show that the predicted paths of the HMM-AO model coincide with the real path basically and even the trend is also captured precisely. In contrast, the accuracy of the prediction of HMM is endless flexibly.

Therefore, the HMM approach with auto-correlated observations is reasonable and appropriate in degradation modeling. Specifically, the relative errors of the predicted observations corresponding to these two data histories above are plotted against time in Fig. 8. It is can be found that the relative error of HMM-AO becomes smaller and smaller over time while the relative error of HMM oscillates substantially until gets larger. With more observations available, the autocorrelation of observations becomes more significant and the prediction accuracy of HMM-AO becomes higher and higher. However, the relative error of HMM eventually becomes larger without considering autocorrelation. This suggests that the prediction of HMM-AO corresponds better with the behavior of the actual degradation process than that of HMM.

Furthermore, by checking all the nineteen degradation trajectories corresponding to 19 samples, it can be noticed in Fig. 9 that almost for all samples the RMSE of the observation prediction of HMM-AO is smaller than that of HMM. Hence, the applicability of the HMM-AO model is extensive. To further illustrate the effect of autocorrelation on the prediction accuracy, we only consider the prediction of the second half observation sequences of all samples. As the autocorrelation of observations becomes more and more significant over time, Fig. 10 reveals that the gap between the RMSE of the second half observation sequence prediction of HMM-AO and HMM becomes larger. This fully demonstrates the excellent performance of HMM-AO when we model auto-correlated degradation processes.

Come back to the RUL prediction. Fig. 11 gives examples of the predicted RUL for two data histories by using HMM-AO and HMM. The results show that the accuracy of the RUL prediction of HMM-AO is significantly higher than that of HMM. In order to illustrate the efficiency of the proposed HMM-AO approach, comparison is applied to the predicted RUL of each sample. The RMSE of each sample is presented in Fig. 12. For most of the samples the RMSE of the predicted RUL of HMM-AO is smaller than the RMSE of HMM. This result highlights the efficiency of our method again and shows the interest of its use in preventive maintenance where it is necessary to make maintenance policies before the failure.

In sum, the performance of the HMM-AO model is better than HMM. The autocorrelation does exist in degradation data. When using HMM for degradation processes, we cannot assume that the observations are independent of each other and cannot ignore the autocorrelation. Therefore, the HMM-AO model is well adapted for the degradation modeling and remaining life prediction. After having studied the impact of the proposed approach on a real case, we try to explore the use and the interest of RUL prediction in engineering applications. The preventive maintenance is an essential application area of the RUL. This problem is described in the next section.

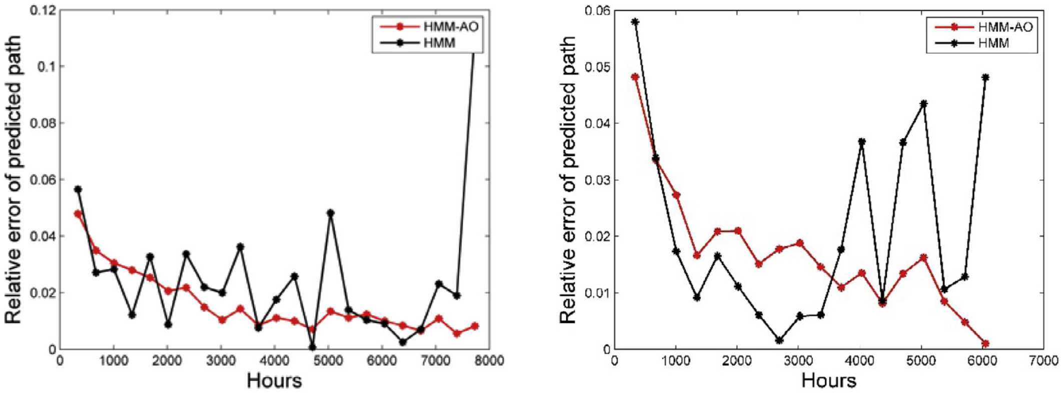

Fig. 8. The relative error of the predicted observations corresponding to the two data histories used in Fig. 7.

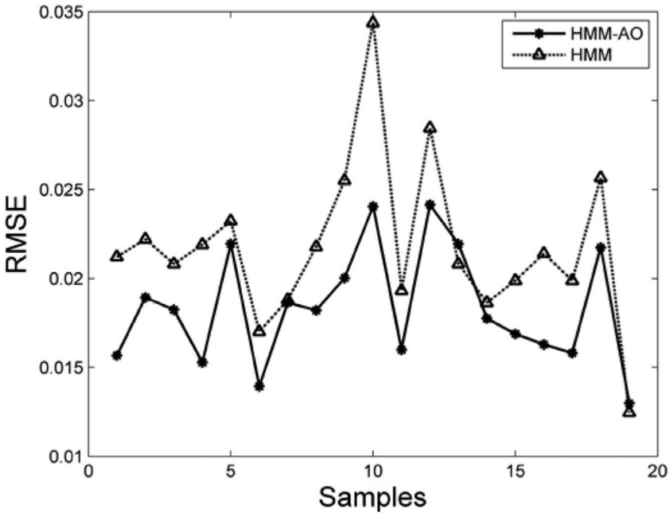

Fig. 9. RMSE of the observation sequence prediction of HMM-AO and HMM for samples.

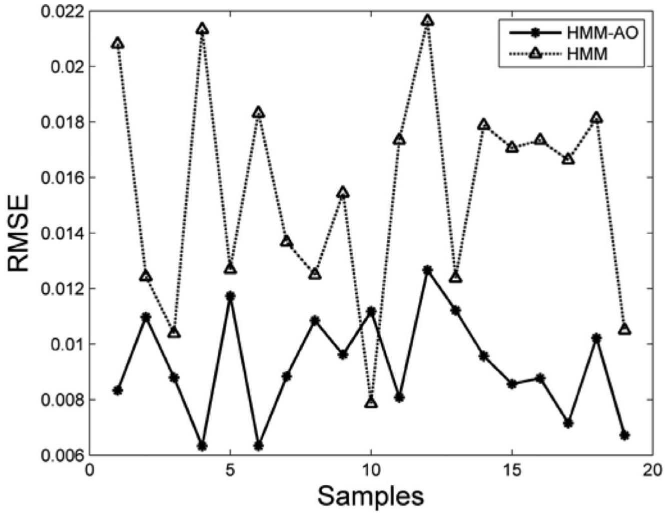

Fig. 10. RMSE of the second half observation sequence prediction of HMM-AO and HMM for samples.

5. Optimal maintenance policy

In this section, to illustrate the application of the RUL in maintenance scheduling, a RUL-based maintenance policy is developed with observations at equidistant time epochs. The objective is to find the optimal preventive threshold to initiate a preventive replacement, by minimizing the expected maintenance cost per unit time. Once the optimal threshold is obtained, the maintenance policy is determined. Then the policy is compared with a condition-based policy where only the current state of the system is considered.

The RUL-based maintenance policy proposed here is based on the predicted values of RUL. In contrast to previous policies, the policy based on expectation and probability not only can capture the random

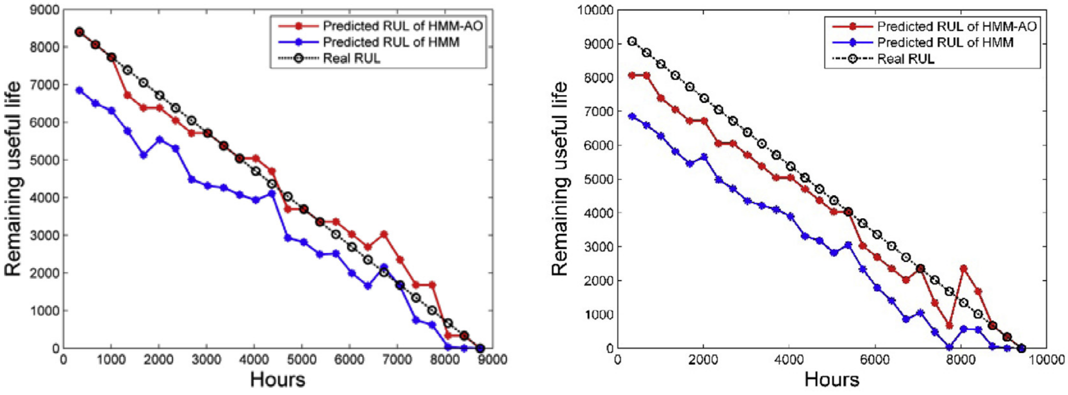

Fig. 11. The real RUL and the predicted RULs calculated by HMM-AO and HMM for two data histories.

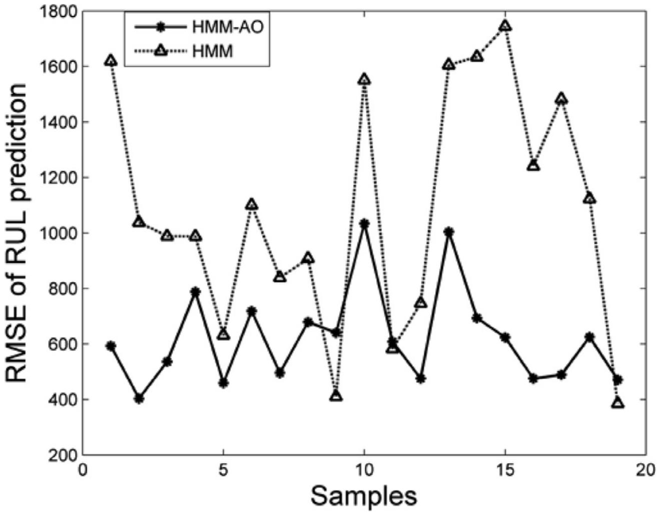

Fig. 12. RMSE of the RUL prediction of HMM-AO and HMM for samples.

properties of degradation more accurately, but also can take the following degradation trend into account. Thus, sound economical and operational decisions are made. Throughout this section, it is assumed that the system is observed periodically at inspection time t_1, t_2, ..., where $\mathrm{t}_\upsilon = \upsilon h$ with $\upsilon \in \mathbb{N}$ and $h \in \mathbb{R}$ is the inspection interval. That is, the failure can only be observed at inspections. If the current degradation of the system is below the preventive threshold, we are only allowed to take maintenance actions at next inspection.

A preventive replacement with cost C_p and a corrective replacement with cost C_r are considered as two possible maintenance actions in this research. The cost per inspection is C_s and these costs satisfy $C_s < C_p < C_r$. Moreover, we suppose that the model parameters are known and the system is as good as new after replacements. According to renewal theory, the expected cycle cost per unit time used as the evaluation criterion of the maintenance policies can be given by

$$CC = \frac{C_E}{T_E}, \tag{51}$$

where C_E and T_E denote the expected maintenance cost and the cycle time, respectively. To be specific, the expected maintenance costs consist of the replacement cost and the inspection cost during the cycle time. We also define an integer δ as the total number of inspections for the system, and thereby the cycle time $T_E = \delta h$. In the following subsections, the maintenance policies are presented and evaluated.

5.1. RUL-based maintenance policy

The observation and the RUL of the system at the inspection time t_υ are denoted by o_υ and $RUL(\mathrm{t}_\upsilon)$, respectively. In the predicted value-based maintenance policy, a fixed maintenance threshold η_R is defined and at each inspection time t_υ during the cycle:

- If $\mathrm{o}_\upsilon \geq \omega$, the system has failed, and a corrective replacement is performed.
- If $\mathrm{o}_\upsilon < \omega$ and $RUL(\mathrm{t}_\upsilon) \leq \eta_R$, a preventive replacement is carried out to renew the system.
- If $\mathrm{o}_\upsilon < \omega$ and $RUL(\mathrm{t}_\upsilon) > \eta_R$, the system runs normally, and the maintenance decision should be made at the next inspection time $\mathrm{t}_{\upsilon+1}$.
- If the system does not reach the maintenance threshold and still operates without failure when the last inspection is performed, a preventive replacement will be enforced.

Note that the threshold η_R is lower than the cycle time. The system will surely be replaced before the last inspection if δ is large enough. From Eq. (26), the probability distribution of $RUL(\mathrm{t}_\upsilon)$ can be given by

$$P\left(RUL\left(t_\upsilon\right) = \tau\right) = \sum_i P\left(RUL\left(t_\upsilon\right) = \tau \middle| q_T = S_i\right) \times P\left(q_T = S_i \middle| o_1, ..., o_\upsilon, \lambda\right). \tag{52}$$

Since the RUL is regarded as a discrete random variable and Eq. (34) is the probability mass function, we can use an approximate method, such as the artificial neural network, to fit the probability density function of the RUL. Let $f_\upsilon(RUL)$ denotes the probability density function of the RUL at the inspection time t_υ.

Given the policy above, the probability that the system is preventively replaced at the inspection time t_υ can be given by

$$\begin{aligned} P_{pr}\left(t_\upsilon\right) &= P\left(RUL\left(t_\upsilon\right) \leq \eta_R, \mathrm{o}_\upsilon < \omega \middle| RUL\left(\mathrm{t}_{\upsilon-1}\right) > \eta_R, \mathrm{o}_{\upsilon-1} < \omega\right) \\ &= P\left(RUL\left(t_\upsilon\right) \leq \eta_R, \mathrm{o}_\upsilon < \omega \middle| RUL\left(\mathrm{t}_{\upsilon-1}\right) - h > \eta_R - h, \mathrm{o}_{\upsilon-1} < \omega\right), \\ &= \textstyle\int_{\eta_R-h}^{\eta_R} f_\upsilon(\tau) d\tau \int_0^\omega P\left(o_\upsilon \middle| o_1, ..., \mathrm{o}_{\upsilon-1} < \omega, \lambda\right) do_\upsilon \end{aligned} \tag{53}$$

where an approximation formula $RUL(\mathrm{t}_\upsilon) = RUL(\mathrm{t}_{\upsilon-1}) - h$ is used for the inference.

The probability that the system is correctively replaced at the inspection t_υ is given by

$$P_{cr}\left(t_\upsilon\right) = P\left(\mathrm{o}_\upsilon \geq \omega \middle| \mathrm{o}_{\upsilon-1} < \omega\right) = \int_\omega^{+\infty} P\left(o_\upsilon \middle| o_1, ..., \mathrm{o}_{\upsilon-1} < \omega, \lambda\right) do_{\upsilon+1}. \tag{54}$$

The usage time of a system between two sequential replacements is denoted by L_T. A cycle can contain several replacements. For simplicity, L_T is calculated as follows

$$L_T = \textstyle\sum_\upsilon t_\upsilon\left(P_{pr}\left(t_\upsilon\right) + P_{cr}\left(t_\upsilon\right)\right). \tag{55}$$

Hence, the cumulated maintenance cost C_E during the cycle T_E can be calculated by

$$C_E = \lfloor T_E/L_T \rfloor \textstyle\sum_{\upsilon=1}^{\lfloor L_T/h \rfloor} \left[C_p P_{pr}\left(t_\upsilon\right) + C_r P_{cr}\left(t_\upsilon\right)\right] + C_p, \tag{56}$$

where $\lfloor z \rfloor$ denotes the integer part of the real number. The preventive threshold is the only decision variable for the maintenance policy. The optimal solution $\eta_R{}^*$ can be obtained as follows:

$$\eta_R{}^* = \arg\min \left\{ CC = \frac{\lfloor T_E/L_T \rfloor \sum_{\upsilon=1}^{\lfloor L_T/h \rfloor} \left[C_p P_{pr}\left(t_\upsilon\right) + C_r P_{cr}\left(t_\upsilon\right)\right] + C_p}{T_E} \right\}. \tag{57}$$

Remark 3. The decision variable η_R in Eq. (58) is the maintenance threshold and should be smaller than the useful lifetime, i.e., $RUL(\mathrm{t}_0)$. Since the RUL is regarded as a discrete random variable, the feasible solutions of η_R consist of finite positive integers. Then, we can use traversal method to search the optimal solution of Eq. (58). Moreover, for the integral of Eqs. (54) and (55), given the upper and lower limit values numerical integration methods can be adopted. Hence, the useful lifetime and the inspection interval determine the computational complexity. Once an appropriate interval is fixed, Eq. (58) can be solved quickly and directly.

5.2. State-based maintenance policy

We compare the performance of our proposed maintenance policies with a conventional condition-based maintenance policy that was obtained in Le et al. [11]. A fixed maintenance threshold $\eta_o < \omega$ is defined in the condition-based maintenance policy and at each inspection time t_υ:

- If $\mathrm{o}_\upsilon \geq \omega$, the system has failed, and a corrective replacement is performed.

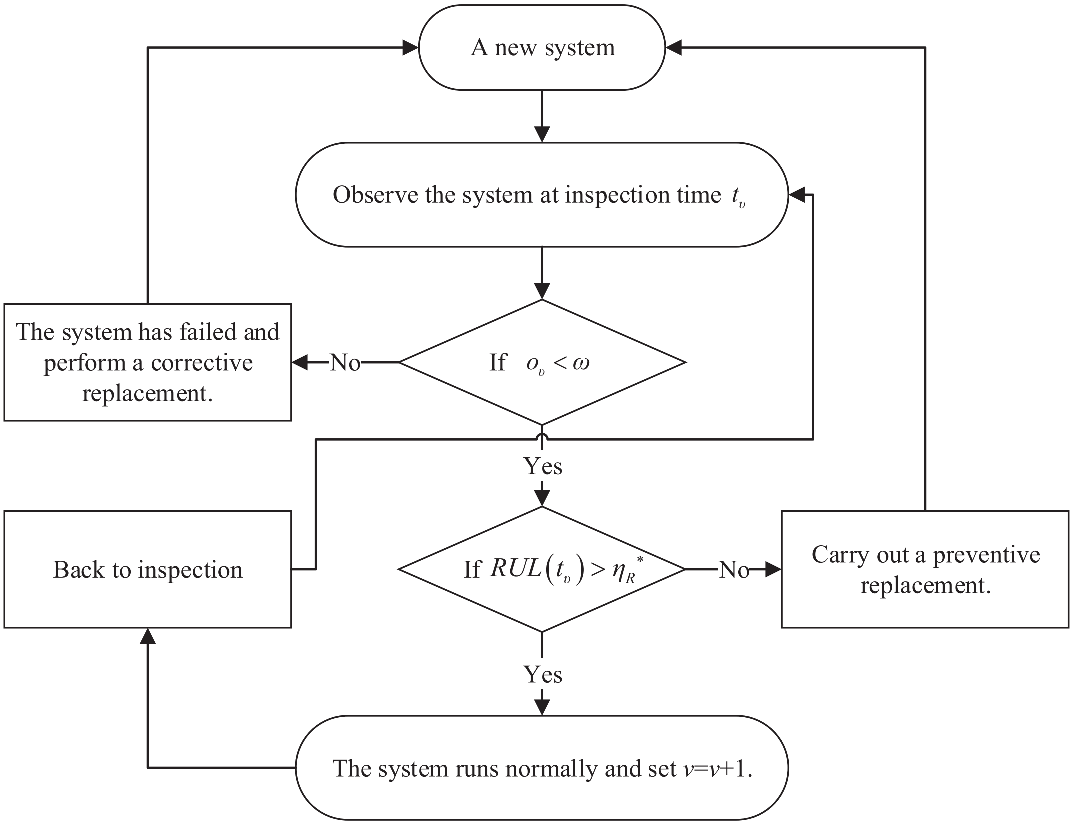
A new system
Observe the system at inspection time t_υ
If $o_\upsilon < \omega$
No
The system has failed and perform a corrective replacement.
Yes
Back to inspection
If $RUL(t_\upsilon) > \eta_R^*$
No
Carry out a preventive replacement.
Yes
The system runs normally and set v=v+1.

Fig. 13. The flowchart for the application of the proposed RUL-based maintenance policy.

- If $\eta_o \leq o_\upsilon < \omega$, a preventive replacement is carried out to renew the system.
- If $o_\upsilon < \eta_o$, the system runs normally, and the maintenance decision should be made at the next inspection time $t_{\upsilon+1}$.

The preventive threshold η_o which is the decision variable to be optimized can be obtained as follows

$$\eta_o{}^* = \arg\min\left\{CC = \frac{C_E(\eta_o)}{T_E}\right\}. \tag{58}$$

5.3. Comparison

We use the fatigue-crack-growth data from Lu and Meeker [36] to illustrate the efficiency of the proposed method and compare the performances of the maintenance policies. There are 21 sample paths of degradation. Since "Time" in the proposed model and method could be real time or some other measure like miles for automobile tires or cycles in fatigue tests, we take ten thousand cycles as a unit "time" here. The cracks of these samples are observed every ten thousand cycles. We define a critical crack length of 1.6 inches to be the failure threshold. The corresponding flowchart for the application of the RUL-based maintenance policy is given in Fig. 13. Suppose that the cost parameters are set as C_s=2, C_p=50 and C_r=500. The total running time is set as one billion cycles and thereby the number of replacements must be greater or equal to one.

The settings of the optimal RUL-based and condition-based maintenance policies with different inspection intervals are displayed in Table 2. We obtained the minimum expected cycle cost rate CC^*=6.523 with the optimal preventive maintenance threshold $\eta_R{}^*$=3×10^4 cycles under the RUL-based maintenance policy when the inspection interval is 10^4 cycles. This means that when the predicted RUL of the sample is less than 3×10^4 cycles, a preventive replacement need to be performed. By comparison of the optimal results of the two policies in Table 2, it can be found that the RUL-based maintenance policy is more cost-effective than the conventional condition-based maintenance policy. This demonstrates that when we consider not only the current degradation state but also the future degradation trend, more economical and effective decisions can be made. In addition, when the inspection interval increases, the maintenance thresholds of both policies become harsher and the expected cycle cost rates get higher. As a shorter interval in the same total running time means more inspections and thereby more data can be obtained. It can be concluded that more information available can make it possible to develop a practical operational policy.

Besides, sensitivity analysis is performed to show the effects of cost parameters on optimal solutions. The key parameter considered here is the ration between the cost of a preventive replacement and the cost of a corrective replacement, i.e., C_r/C_p. The results are shown in Fig. 14. It can be observed that how the solutions change with the parameter changing. Both the maintenance threshold and expected cost per unit time increase with the increasing of the ratio. The increasing of the maintenance threshold means the replacement frequency becomes larger and thereby leads to increased cost. Since the corrective cost keeps getting higher, we need to carry out more preventive replacements to avoid the occurrence of potential failure.

6. Conclusion

In this paper, we propose an approach based on the HMM-AO model to reflect the degradation process. The autocorrelation property of the observations is characterized by coefficient matrices. A novel algorithm based on the Expectation maximum method is developed to estimate the unknown parameters. Missing data and noise that accumulate over time are taken into account by modifying the proposed model. Then two RUL prediction methods based on the HMM-AO models are presented. The effectiveness of the proposed methods is demonstrated by a real case study with a LED degradation dataset. Furthermore, these data are

Table 2
Comparison of RUL-based maintenance and state-based maintenance policies.

Policy	Inspection interval	Maintenance threshold	The expected cycle cost rate
RUL-based	10^4 cycles	3×10^4 cycles	6.523
	2×10^5 cycles	4×10^4 cycles	10.277
State-based	10^4 cycles	1.52 inches	9.692
	2×10^4 cycles	1.40 inches	16.285

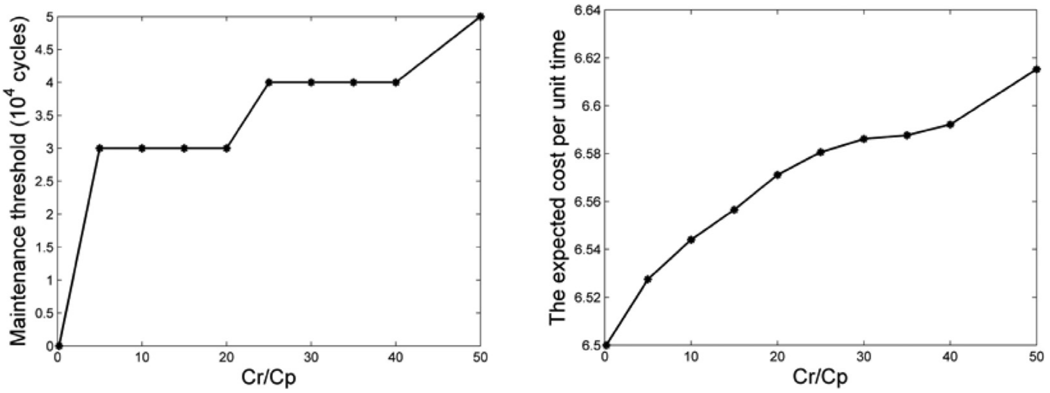

Fig. 14. Sensitivity analysis of the maintenance threshold and expected cost per unit time on C_r/C_p.

then used to illustrate the advantages of the HMM-AO model by comparing with the standard HMM. To illustrate the application of the RUL in the framework of the maintenance, a RUL-based maintenance policy is developed with observations at equidistant time epochs. Using the fatigue-crack-growth data, the proposed policy is also compared with a conventional condition-based policy.

Further extension of the HMM-AO model to fault diagnosis of production systems is a suitable topic for future research. Another interesting topic could be the development of the degradation models and the maintenance policies based on the proposed model. Moreover, HMMs state that the current state is dependent only on the previous state, this represents the distributions of the state durations are constant or geometric. Modeling the time duration of the hidden states is necessary in the furfure and therefore the corresponding model is more capable of real degradation processes.

Acknowledgments

The authors would like to thank anonymous referees for their remarkable comments. This research is supported by National Natural Science Foundation of China [grant number 51475289], [grant number 51505288]; and China Postdoctoral Science Foundation [grant number 2016M600320].

Appendix

The estimation of unknown parameters of HMM-AO in the case study
A = [0.571 0.429 0 0 0 0 0;
0 0.419 0.581 0 0 0 0;
0 0 0.637 0.363 0 0 0;
0 0 0 0.685 0.315 0 0;
0 0 0 0 0.713 0.287 0;
0 0 0 0 0 0.712 0.288;
0 0 0 0 0 0 1.0]
C = [0.573 0.331;
1.025 – 0.050;
0.740 0.184;
0.821 0.113;
0.883 0.065;
0.754 0.150;
0.836 0.087]
Σ = [2.970 0.424 6.436 5.527 2.741 6.107 3.322]/10ˆ4;
π = [0.781, 0.052, 0.167, 0, 0, 0, 0];
The estimation of unknown parameters of HMM in the case study
A = [0.544 0.456 0 0 0 0 0;
0 0.599 0.401 0 0 0 0;
0 0 0.641 0.359 0 0 0;
0 0 0 0.713 0.287 0 0;
0 0 0 0 0.680 0.32 0;
0 0 0 0 0 0.804 0.196;
0 0 0 0 0 0 1.0]
B = [0.887 0.034;
0.807 0.021;
0.745 0.019;
0.696 0.017;
0.649 0.014;
0.606 0.017;
0.564 0.020];
π = [1 0 0 0 0 0 0];

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Alaswad S, Xiang Y. A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliab Eng Syst Saf 2017;157:54–63.
[2] Jafari L, Makis V. Joint optimization of lot-sizing and maintenance policy for a partially observable two-unit system. Int J Adv Manuf Technol 2016;87(5-8):1621–39.
[3] Son J, Zhou S, Sankavaram C, Du X, Zhang Y. Remaining useful life prediction based on noisy condition monitoring signals using constrained Kalman filter. Reliab Eng Syst Saf 2016;152:38–50.
[4] Dong M, He D. A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. Mech Syst Signal Process 2007;21(5):2248–66.
[5] Gómez MJ, Castejón C, García-Prada JC. Automatic condition monitoring system for crack detection in rotating machinery. Reliab Eng Syst Saf 2016;152:239–47.
[6] Chen Z, Xia T, Pan E. Optimal multi-level classification and preventive maintenance policy for highly-reliable products. Int J Prod Res 2017;55(8):2232–50.
[7] Xia T, Jin X, Xi L, Ni J. Production-driven opportunistic maintenance for batch production based on MAM-APB scheduling. Eur J Oper Res 2015;240(3):781–90.
[8] Beganovic N, Söffker D. Remaining lifetime modeling using State-of-Health estimation. Mech Syst Signal Process 2017;92:107–23.

[9] Zhao Z, Liang B, Wang X, Lu W. Remaining useful life prediction of aircraft engine based on degradation pattern learning. Reliab Eng Syst Saf 2017;164:74–83.
[10] Khorasgani H, Biswas G, Sankararaman S. Methodologies for system-level remaining useful life prediction. Reliab Eng Syst Saf 2016;154:8–18.
[11] Le Son K, Fouladirad M, Barros A. Remaining useful lifetime estimation and noisy gamma deterioration process. Reliab Eng Syst Saf 2016;149:76–87.
[12] Mazhar MI, Kara S, Kaebernick H. Remaining life estimation of used components in consumer products: life cycle data analysis by Weibull and artificial neural networks. J Oper Manage 2007;25(6):1184–93.
[13] Juang BH, Rabiner LR. Hidden Markov models for speech recognition. Technometrics 1991;33(3):251–72.
[14] Hu J, Brown MK, Turin W. HMM based online handwriting recognition. IEEE Trans Pattern Anal Mach Intell 1996;18(10):1039–45.
[15] Yu SZ. Hidden semi-Markov models. Artif Intell 2010;174(2):215–43.
[16] Liao W, Li D, Cui S. A heuristic optimization algorithm for HMM based on SA and EM in machinery diagnosis. J Intell Manuf 2016:1–13.
[17] Vrignat P, Avila M, Duculty F, Kratz F. Failure event prediction using hidden Markov model approaches. IEEE Trans Reliab 2015;64(3):1038–48.
[18] Soualhi A, Clerc G, Razik H, Guillet F. Hidden Markov models for the prediction of impending faults. IEEE Trans Ind Electron 2016;63(5):3271–81.
[19] Fort A, Mugnaini M, Vignoli V. Hidden Markov models approach used for life parameters estimations. Reliab Eng Syst Saf 2015;136:85–91.
[20] Wang F, Tan S, Yang Y, Shi H. Hidden Markov model-based fault detection approach for a multimode process. Ind Eng Chem Res 2016;55(16):4613–21.
[21] Wang M, Wang J. CHMM for tool condition monitoring and remaining useful life prediction. Int J Adv Manuf Technol 2012;59(5-8):463–71.
[22] Le TT, Chatelain F, Bérenguer C. Multi-branch hidden Markov models for remaining useful life estimation of systems under multiple deterioration modes. Proc Inst Mech Eng, Part O: J Risk Reliab 2016;230(5):473–84.
[23] Ghasemi A, Yacout S, Ouali MS. Evaluating the reliability function and the mean residual life for equipment with unobservable states. IEEE Trans Reliab 2010;59(1):45–54.
[24] Yu J. Adaptive hidden Markov model-based online learning framework for bearing faulty detection and performance degradation monitoring. Mech Sys Signal Process 2017;83:149–62.
[25] Cholette ME, Djurdjanovic D. Degradation modeling and monitoring of machines using operation-specific hidden Markov models. IIE Trans 2014;46(10):1107–23.
[26] Geramifard O, Xu JX, Zhou JH, Li X. Multimodal hidden Markov model-based approach for tool wear monitoring. IEEE Trans Ind Electron 2014;61(6):2900–11.
[27] Zhang D, Bailey AD, Djurdjanovic D. Bayesian identification of hidden Markov models and their use for condition-based monitoring. IEEE Trans Reliab 2016;65(3):1471–82.
[28] Pacella M, Semeraro Q. Using recurrent neural networks to detect changes in autocorrelated processes for quality monitoring. Comput Ind Eng 2007;52(4):502–20.
[29] Tang D, Makis V, Jafari L, Yu J. Optimal maintenance policy and residual life estimation for a slowly degrading system subject to condition monitoring. Reliab Eng Syst Saf 2015;134:198–207.
[30] Adjengue L, Yacout S, Ilk O. Parameters estimation for condition based maintenance with uncorrelated and correlated observations. Qual Eng 2007;19(3):197–206.
[31] Yu SZ, Kobayashi H. A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. Signal Process 2003;83(2):235–50.
[32] Palomäki KJ, Brown GJ, Wang DL. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Commun 2004;43(4):361–78.
[33] Hamada MS, Wilson A, Reese CS, Martz H. Bayesian Reliability. Berlin, Germany: Springer; 2008.
[34] Hamidi M, Szidarovszky F, Szidarovszky M. New one cycle criteria for optimizing preventive replacement policies. Reliab Eng Syst Saf 2016;154:42–8.
[35] Xia T, Xi L, Pan E, Ni J. Reconfiguration-oriented opportunistic maintenance policy for reconfigurable manufacturing systems. Reliab Eng Syst Saf 2016 Online. doi:10.1016/j.ress.2016.09.001.
[36] Lu CJ, Meeker WO. Using degradation measures to estimate a time-to-failure distribution. Technometrics 1993;35(2):161–74.
[37] Kurt M, Kharoufeh JP. Monotone optimal replacement policies for a Markovian deteriorating system in a controllable environment. Oper Res Lett 2010;38(4):273–9.
[38] Peng W, Li YF, Mi J, Yu L, Huang HZ. Reliability of complex systems under dynamic conditions: a Bayesian multivariate degradation perspective. Reliab Eng Syst Saf 2016;153:75–87.
[39] Xu D, Wei Q, Elsayed EA, Chen Y, Kang R. Multivariate degradation modeling of smart electricity meter with multiple performance characteristics via vine copulas. Qual Reliab Eng Int 2016 Online. doi:10.1002/qre.2058.
[40] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 1989;77(2):257–86.
[41] Rabiner L, Juang B. An introduction to hidden Markov models. IEEE ASSP Mag 1986;3(1):4–16.
[42] Safi K, Mohammed S, Attal F, Amirat Y, Oukhellou L, Khalil M, et al. Automatic segmentation of stabilometric signals using hidden Markov model regression. IEEE Trans Autom Sci Eng 2017 online. doi:10.1109/TASE.2016.2637165.
[43] Weeks JD. External fields, density functionals, and the Gibbs inequality. J Stat Phys 2003;110(3-6):1209–18.