

Online Degradation Assessment and Adaptive Fault Detection Using Modified Hidden Markov Model

Seungchul Lee

Department of Mechanical Engineering,
University of Michigan-Ann Arbor,
1210 H. H. Dow, 2300 Hayward Street,
Ann Arbor, MI 48109-2136
e-mail: seunglee@umich.edu

Lin Li¹

Department of Mechanical Engineering,
University of Michigan-Ann Arbor,
1035 H. H. Dow, 2300 Hayward Street,
Ann Arbor, MI 48109-2136
e-mail: lilz@umich.edu

Jun Ni

e-mail: junni@umich.edu
Department of Mechanical Engineering,
University of Michigan-Ann Arbor,
1023 H. H. Dow, 2300 Hayward Street,
Ann Arbor, MI 48109-2136

Online condition monitoring and diagnosis systems play an important role in the modern manufacturing industry. This paper presents a novel method to diagnose the degradation processes of multiple failure modes using a modified hidden Markov model (MHMM) with variable state space. The proposed MHMM is combined with statistical process control to quickly detect the occurrence of an unknown fault. This method allows the state space of a hidden Markov model to be adjusted and updated with the identification of new states. Hence, the online degradation assessment and adaptive fault diagnosis can be simultaneously obtained. Experimental results in a turning process illustrate that the tool wear state can be successfully detected, and previously unknown tool wear processes can be identified at the early stages using the MHMM. [DOI: 10.1115/1.4001247]

Keywords: hidden Markov model, online degradation assessment, adaptive fault detection

1 Introduction

Condition-based maintenance (CBM) recommends maintenance plans based on the information collected through numerous condition monitoring techniques [1,2]. The basic principle behind CBM is that defects that gradually yield in machines can be detected through suitable monitoring techniques at the early stages so that appropriate maintenance plans can be scheduled accordingly. Because of the complexity of modern plants, CBM has become widely accepted as one of the key drivers to reduce maintenance costs and machine downtime of manufacturing systems [3].

Condition monitoring techniques for machine diagnosis have been studied extensively [1]. Many signal processing techniques that involve the analysis of the acquired data in time [4], frequency [5], and time-frequency domains [6] have been developed. Paya et al. [7] developed a condition monitoring method, which relied on wavelet transformation and artificial neural networks. A similar work that uses principal component analysis was reported in Ref. [8]. These methods are feature-based methods that use the statistical features of the signal. However, these methods require too much data and time to obtain the results of condition monitoring and diagnosis. In addition, they are data-based methods, which do not take the physical model of the system into consideration. On the other hand, model-based methods, under the assumption that measured information is stochastically correlated with the actual machine condition, take advantage of understanding the system structure [1].

This assumption leads to the application of a hidden Markov model (HMM) through a statistical approach in identifying the actual machine conditions from observable monitoring signals. Although HMMs were motivated by their successes in speech recognition [9], many applications of the HMM in machine process diagnosis have also been studied, demonstrating its effectiveness in online diagnosis. For example, Ertunc et al. [10] presented

a HMM approach for tool wear detection and prediction in a drilling process. A similar approach was also described for a turning process by Wang et al. [11]. Li et al. [12] used a HMM as a fault diagnosis tool in speed-up and speed-down processes for rotating machinery.

According to the literature review, most previous condition-based diagnosis models based on a HMM mainly focus on the online degradation assessment of a single failure mode system [9–12]. A HMM with a single failure mode system assumes that all possible system condition states are known a priori and that training data sets from associated states are available. In addition, training a HMM should be conducted offline. These assumptions significantly impede machine diagnosis applications when it is difficult to identify and train all of the possible states of the system in advance [13,14]. For instance, if a HMM that has been trained to model gradual tool wear in a drilling process does not have a state to represent a tool breakage or shortage of coolant, it is impossible for the HMM to estimate the correct state when these untrained states occur. The state structure of a conventional HMM will not be updated after the training stage. This inflexibility may cause serious estimation errors in the emergence of unknown or untrained faults that might provoke catastrophic damages to machining processes.

Therefore, it is necessary to introduce an anomaly detection algorithm into a HMM to trigger the HMM to adjust the number of hidden states or the hidden structure, and thus result in a more accurate model for the system. In this paper, the modified Hidden Markov Model (MHMM) with variable state space is developed to estimate the current state of system degradation as well as to detect the emergence of unknown faults at an early stage. The statistical process control (SPC) technique [15] is used in unknown fault detection and diagnostics in conjunction with the MHMM. By measuring the deviation of the current signal from a reference signal representing prior known states, the MHMM is able to see whether the current signal is within the control limits.

The rest of this paper is organized as follows: Sec. 2 introduces the principle of a HMM and the proposed MHMM for online degradation assessment and state update. In Sec. 3, case studies are performed to validate the effectiveness of the MHMM algo-

¹Corresponding author.

Contributed by the Manufacturing Engineering Division of ASME for publication in the JOURNAL OF MANUFACTURING SCIENCE AND ENGINEERING. Manuscript received May 18, 2009; final manuscript received February 8, 2010; published online April 1, 2010. Assoc. Editor: Gloria Wiens.

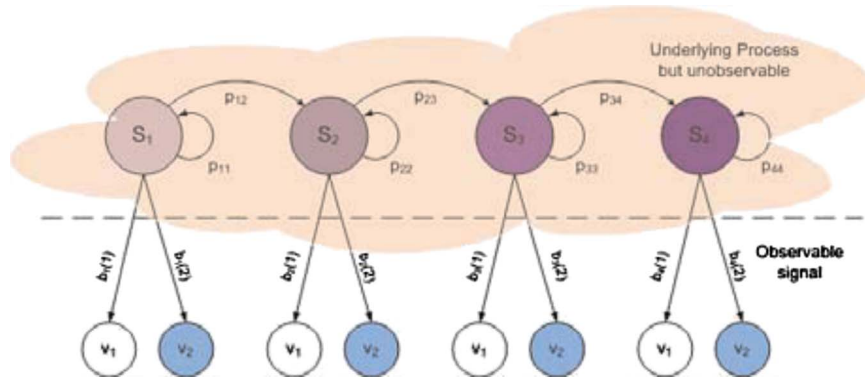


Fig. 1 Basic form of a HMM

rithm and to compare its performance with other methods using a turning process. The conclusions and future research directions are given in Sec. 4.

2 The Proposed MHMM With Variable State Space

2.1 Hidden Markov Model and State Estimation. Before introducing the MHMM with variable state space, we present the basic form of a traditional HMM with fixed state space, as shown in Fig. 1. The HMM $\lambda=(P, b, \pi)$ under consideration consists of

- a finite set of M states $\mathbf{S}=\{S_1, \dots, S_M\}$
- a state transition probability matrix $P=\{p_{ij}\}_{M \times M}$ ($1 \leq i, j \leq M$), where $p_{ij}=P\{q(n+1)=S_j|q(n)=S_i\}$ ($1 \leq i, j \leq M, 1 \leq n < N$)
- an observation symbol probability distribution $b_i(O(n))=P\{O(n)|q(n)=S_i\}$ ($1 \leq i \leq M, 1 \leq n < N$)
- an initial state probability distribution $\pi=\{\pi_i\}_M$, where $\pi_i=P\{q(1)=S_i\}$ ($1 \leq i \leq M$)

A HMM technique is applicable to a process that is assumed to possess homogeneous Markovian property [16] as follows:

$$p_{ij}=P\{q(n)=S_j|q(n-1)=S_i, \dots, q(1)=S_k\}=P\{q(n)=S_j|q(n-1)=S_i\} \quad (1)$$

Equation (1) implies that the conditional probability of the current state, given the knowledge of all previous states, is the same as the conditional probability of the current state, given the knowledge of the system state of one previous time unit.

The state transition probability matrix P encodes the uncertainty in the true underlying state evolution of the stochastic process, while each state emits observation symbols with the probability distribution $b_i(O(n))$, as shown in Fig. 1. Let $O_n=\{O(1), \dots, O(n)\}$ denote a sequence of all observation symbols up to time n , where an observed data point $O(n)$ is taken at time n . The actual state sequence up to time n can be represented as $q_n=\{q(1), \dots, q(n)\}$, where a state $q(k) \in S$ ($1 \leq k \leq n$). Then, we can find the maximum likelihood state sequence $\hat{q}_n=\{\hat{q}(1), \dots, \hat{q}(n)\}$ associated with a given sequence of observations $O_n=\{O(1), \dots, O(n)\}$ as well as a HMM model $\lambda=(P, b, \pi)$ through the Viterbi algorithm [17,18]. Furthermore, it is possible to adjust the HMM model parameters $\lambda=(P, b, \pi)$ to maximize the probability of the observation sequence using an iterative procedure such as the Baum–Welch method [19] or the expectation-maximization (EM) algorithm [20].

Since the primary purpose of a HMM in this paper is to estimate the system state as early as possible, the forward procedure [9] based on past and present measurements is employed. Consider the forward variable $\alpha_n(i)$ defined as $\alpha_n(i)=P\{O(1), \dots, O(n) \wedge q(n)=S_i|\lambda\}$, indicating the joint probability

of a series of observed symbols $O_n=\{O(1), \dots, O(n)\}$ and state S_i at time n , given the model λ . We can then calculate $\alpha_n(i)$ recursively, as follows:

- (1) Initialization

$$\alpha_1(i)=\pi_i b_i(O(1)), \quad 1 \leq i \leq M \quad (2)$$

- (2) Induction

$$\alpha_{n+1}(j)=\left[\sum_i p_{ij} \alpha_n(i)\right] b_j(O(n+1)), \quad 1 \leq n < N, \quad 1 \leq j \leq M \quad (3)$$

Once $\alpha_n(i)$ are obtained, the posterior probabilities $P(q(n)=S_i|O(1), \dots, O(n) \wedge \lambda)$ that the current state $q(n)$ is in state S_i , given the observed symbols $O_n=\{O(1), \dots, O(n)\}$ can be calculated by the Bayes' rule

$$P\{q(n)=S_i|O(1), \dots, O(n) \wedge \lambda\} = \frac{P\{q(n)=S_i \wedge O(1), \dots, O(n)|\lambda\}}{P\{O(1), \dots, O(n)|\lambda\}} = \frac{\alpha_n(i)}{\sum_j \alpha_n(j)}, \quad 1 \leq i \leq M \quad (4)$$

Hence, we can estimate the state $\hat{q}(n)$, which maximizes the posterior probability as

$$\hat{q}(n)=\operatorname{argmax}_i \{P\{q(n)=S_i|O(1), \dots, O(n) \wedge \lambda\}\} \quad (5)$$

Furthermore, the EM algorithm is used to find the maximum likelihood HMM parameters $\lambda=(P, b, \pi)$ that could have produced the sequence of observations $O_N=\{O(1), \dots, O(N)\}$. Define $\xi_n(i, j)$ and $\gamma_n(i)$ as follows:

$$\xi_n(i, j)=P\{q(n)=S_i \wedge q(n+1)=S_j|O_N \wedge \lambda\} \quad (6)$$

$$\gamma_n(i)=P\{q(n)=S_i|O_N \wedge \lambda\} \quad (7)$$

where $\xi_n(i, j)$ in Eq. (6) is the probability of being in state S_i at time n and in state S_j at time $n+1$, given the model λ and the observation sequence O_N up to time N . Note that $\gamma_n(i)$ in Eq. (7) is the probability of being in S_i at time n , given the model and the observation sequence O_N . Thus, a set of re-estimation for $\lambda=(P, b, \pi)$ would be expressed as

$$\hat{\pi}_i=\gamma_1(i) \quad (8)$$

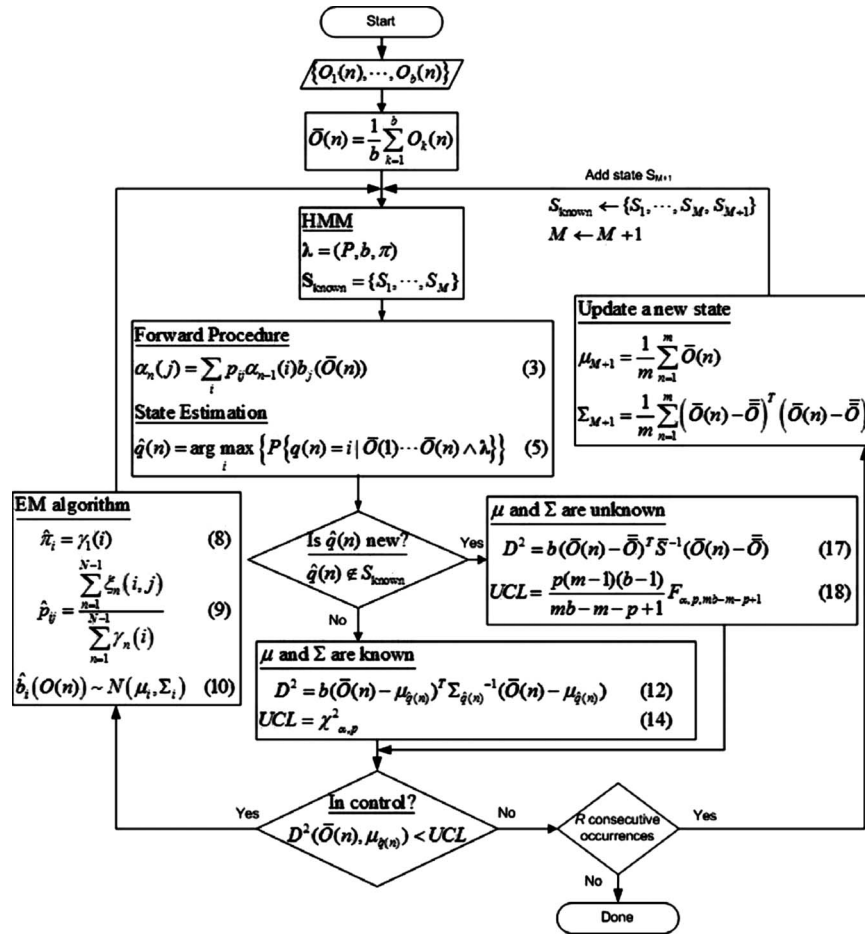


Fig. 2 Block diagram of the proposed modified HMM algorithm

$$\hat{p}_{ij} = \frac{\sum_{n=1}^{N-1} \xi_n(i, j)}{\sum_{n=1}^{N-1} \gamma_n(i)}, \quad 1 \leq i, j \leq M \quad (9)$$

$$\hat{b}_i(O(n)) \sim N(\mu_i, \Sigma_i) \quad (10)$$

We will use Eqs. (8)–(10) to update a HMM. It should be noted that the number of discrete states and the selection of training data sets have a great influence on the HMM performance for the state estimation. Therefore, states have to be selected in such a way that maximizes the discrepancies among the states. In addition, the size of the training data set has to be large enough to ensure observation symbol probability distributions to be statistically significant.

2.2 The Modified Hidden Markov Model. We propose to use the MHMM with variable state space to detect the emergence of the different failure modes at early stages as well as to estimate the current state of system degradation. The technique of SPC is combined with a HMM to detect the different failure modes and diagnostics. The MHMM can check whether the current signals

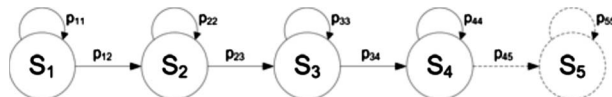


Fig. 3 Markov chain with an unknown state S_5

are emitted from unknown failure modes that have not been observed by calculating the deviation of the current signal from the reference signals representing prior known states.

Suppose that there are m observations available from the process, each of size b and the observation symbol probability distributions $b_i(O(n)) = P\{O(n)|q(n)=S_i\}$ follow a p -jointly Gaussian density distribution. This assumption is reasonable in many applications because of the central limit theorem, which states that the sum of independently distributed random variables is approximately Gaussian-distributed regardless of the distributions of the individual variables as the number of samples becomes large [16]. Then $b_i(O(n))$ can be expressed as

$$b_i(O(n)) = P\{O(n)|q(n)=S_i\} = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(O(n) - \mu_i)^T \Sigma_i^{-1} (O(n) - \mu_i)\right) \quad (11)$$

where $\mu^N = [\mu_1, \dots, \mu_N]$ is the mean vector and Σ is the covariance matrix of the distribution.

Then the weighted distance of $\bar{O}(n)$ from μ_i , known as the Mahalanobis distance [21], can be calculated as

$$D^2(\bar{O}(n), \mu_i) = b(\bar{O}(n) - \mu_i)^T \Sigma_i^{-1} (\bar{O}(n) - \mu_i) \quad (12)$$

where $\bar{O}(n)$ is the vector of the observation symbol mean of the n th observation

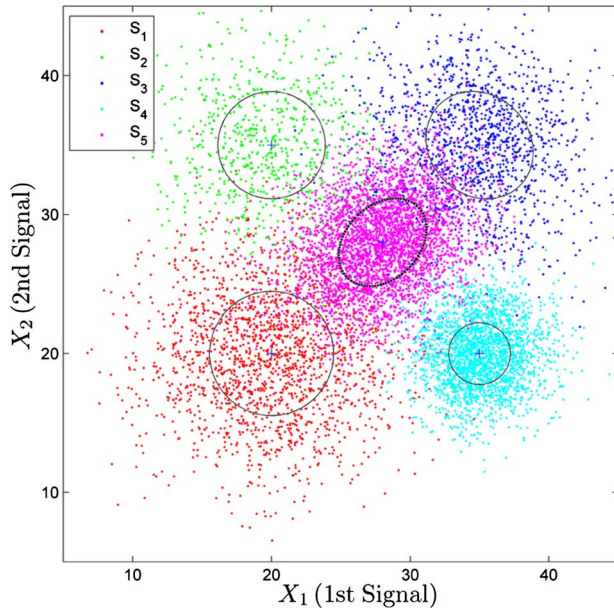


Fig. 4 Original observable signals and the HMM with the four known states

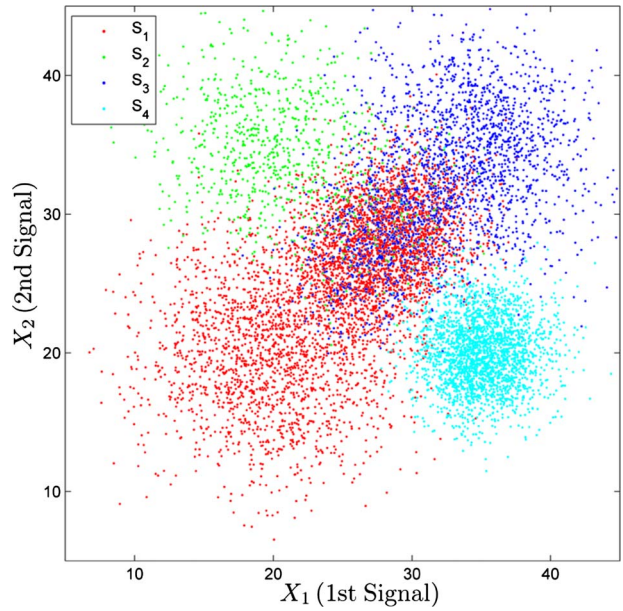


Fig. 5 HMM algorithm serving to estimate the states

$$\bar{O}(n) = \frac{1}{b} \sum_{k=1}^b O_k(n)$$

We use this statistic to detect an unknown state in the MHMM because $D^2(\bar{O}(n), \mu_i)$ can represent a dissimilarity distance when the number of monitoring signals is more than one [22]. The most familiar multivariate process monitoring technique is the Hotelling multivariate control chart [23]. We use the Hotelling multivariate control chart technique for an anomaly detection algorithm in the MHMM because this method can deal with multiple monitoring signals, make an online decision based on current monitoring signals, and has shown effectiveness, especially in the manu-

facturing industry [22]. The Hotelling multivariate control chart signals that a statistically significant shift in the mean has occurred when

$$D^2(\bar{O}(n), \mu_i) > UCL \tag{13}$$

where $UCL > 0$ is a specified upper control limit (UCL).

The calculation of the UCL depends on whether the values of μ and Σ are known or not in advance. If μ and Σ are known, the D^2 statistic follows the χ^2 -distribution with p degrees of freedom [24]. Thus, the UCL can be obtained as

$$UCL = \chi^2_{\alpha, p} \tag{14}$$

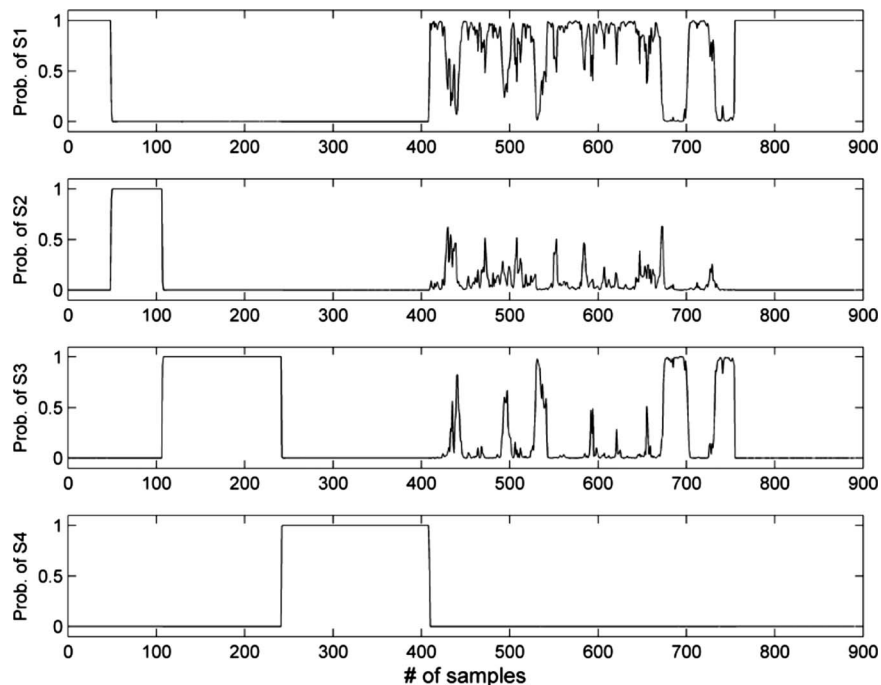


Fig. 6 Posterior probability of $P\{q(n)=S_i | O(1), \dots, O(n)\}$

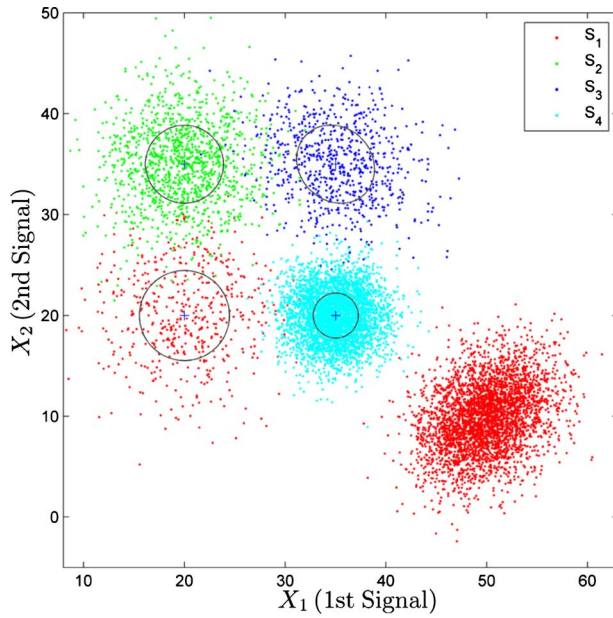


Fig. 7 Result of a wrong state estimation

where α is the risk level.

If μ and Σ are not known, the m observation subgroups of each size b must be used to estimate μ with \bar{O} , the overall mean vector, and Σ with \bar{S} , the covariance matrix. \bar{O} and \bar{S} can be calculated:

$$\bar{O} = \frac{1}{m} \sum_{n=1}^m \bar{O}(n) \quad (15)$$

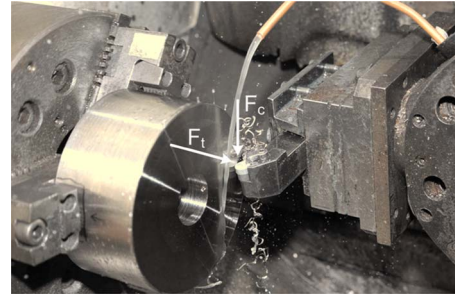


Fig. 9 Test bed of the turning process with coolant supply (F_t : thrust force; F_c : cutting force)

$$\bar{S} = \frac{1}{m} \sum_{n=1}^m (\bar{O}(n) - \bar{O})^T (\bar{O}(n) - \bar{O}) \quad (16)$$

It has been shown that \bar{O} and \bar{S} are the maximum likelihood estimates of μ and Σ , respectively [25]. In this case, the D^2 statistics and the UCL for the Hotelling multivariate control chart are defined as follows:

$$D^2(\bar{O}(n), \bar{O}) = b(\bar{O}(n) - \bar{O})^T \bar{S}^{-1} (\bar{O}(n) - \bar{O}) \quad (17)$$

$$UCL = \frac{p(m-1)(b-1)}{mb-m-p+1} F_{\alpha, p, mb-m-p+1} \quad (18)$$

Equation (18) is based on the fact that the $D^2(\bar{O}(n), \bar{O})$ statistic follows an F -distribution with p and $(mb-m-p+1)$ degrees of freedom when its mean and covariance are not known [26].

Therefore, we can claim that the process of interest is experiencing a statistically significant shift in the mean if the Mahalanobis distance D^2 becomes larger than the UCL. The MHMM makes use of this characteristic of the SPC for the purpose of detecting the unknown states. The MHMM with variable state space will

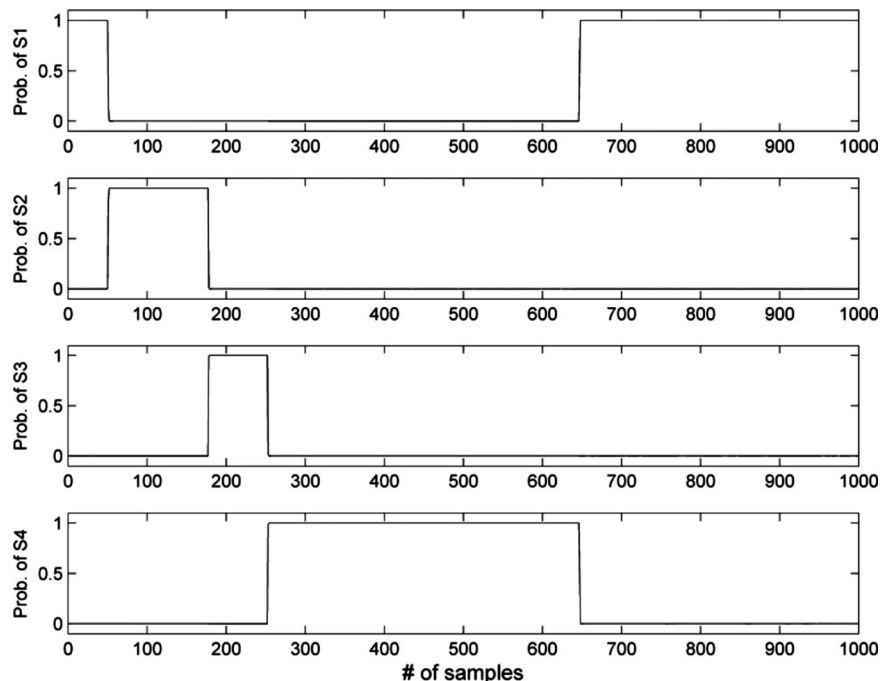


Fig. 8 Posterior probability, but no wiggling is shown

Table 1 The cutting conditions

Depth of cut (μm)	Feed rate ($\mu\text{m}/\text{rev}$)	Cutting speed (m/min)
228.6	228.6	152.4

adjust the number of hidden states or the hidden structure based on the result of the Hotelling multivariate control chart. A summary of the MHMM algorithm is shown in Fig. 2.

Suppose the initial MHMM is trained only with prior known states $\mathbf{S}_{\text{known}}=\{S_1, S_2, \dots, S_M\}$ and associated training data sets. This MHMM receives a set of data $\{O_1(n), \dots, O_b(n)\}$ with a sample size b at time n . The sample mean of each set $\bar{O}(n)$ is calculated, i.e., $\bar{O}(n)=1/b\sum_{k=1}^b O_k(n)$, and fed to the HMM state estimation algorithm, as shown in Eqs. (2) and (3), to estimate the current state $\hat{q}(n)$ from the sequence of observation symbols $\bar{O}_n=\{\bar{O}(1), \dots, \bar{O}(n)\}$. If $\hat{q}(n)$ belongs to the prior known state set $\mathbf{S}_{\text{known}}$, then the distance $D^2(\bar{O}(n), \mu_{\hat{q}})$ and UCL are obtained by means of Eqs. (12) and (14), respectively. This is possible because the corresponding μ and Σ of $\hat{q}(n)$ are known. If $\hat{q}(n)$ does not belong to the prior known state space $\mathbf{S}_{\text{known}}$, Eqs. (17) and (18) can be used instead. If any anomalous behavior has not been detected via the control chart (i.e., $D^2 < UCL$), the sequence of observation symbols $\bar{O}_n=\{\bar{O}(1), \dots, \bar{O}(n)\}$ will be used to update the MHMM through the EM algorithm. On the other hand, if $D^2 > UCL$ occurs R consecutive times, a new state S_{M+1} needs to be introduced to the MHMM to model an unknown state of the system with μ_{M+1} and Σ_{M+1} , as shown in Eqs. (15) and (16). The number R can be used to control the sensitivity of the unknown detection algorithm. For instance, if R is increased, the unknown detection algorithm may become more robust against false detections caused by process randomness itself. On the other hand, the MHMM may respond more slowly to the unknown state. Thus, the number R needs to be chosen with considerable caution [27].

3 Case Studies

3.1 Inability of a Conventional HMM With Unknown State. In this section, we illustrate the effectiveness and outperformance of the MHMM with comparison to a conventional HMM using a numerically generated case study. To study the numerical cases where some of the hidden states of a HMM are not known, we consider the HMM which is trained initially with the four states $\mathbf{S}_{\text{known}}=\{S_1, S_2, S_3, S_4\}$, while the true system actu-

ally contains another unknown state S_5 , as shown in Fig. 3.

Suppose that two signals (X_1, X_2) are monitored. The observation symbol probabilities from each state have two jointly Gaussian density distributions, and the HMM has the transition probability matrix P , summarized as follows:

$$\mu_1 = \begin{bmatrix} 20 \\ 20 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 20 \\ 35 \end{bmatrix}, \quad \mu_3 = \begin{bmatrix} 35 \\ 35 \end{bmatrix}, \quad \mu_4 = \begin{bmatrix} 35 \\ 20 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 15 & -2 \\ -2 & 15 \end{bmatrix},$$

$$\Sigma_4 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

$$P = \begin{bmatrix} 0.99 & 0.01 & 0 & 0 \\ 0 & 0.99 & 0.01 & 0 \\ 0 & 0 & 0.99 & 0.01 \\ 0.01 & 0 & 0 & 0.99 \end{bmatrix}$$

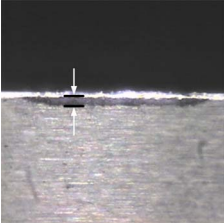
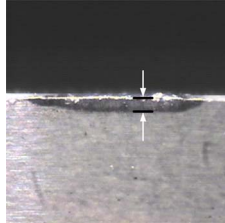
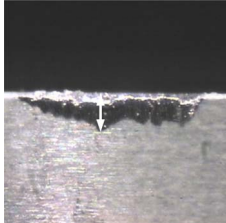
One possible result of the observable signals is illustrated in Fig. 4 if samples of size $b=10$ (i.e., one subgroup consists of ten samples) are taken. Note that these signals are abstract and are not linked to any specific physical meaning.

However, the estimated states obtained from the sequence of observable signals by means of the conventional HMM algorithm are different from the true states of the system, as shown in Fig. 5. This is because the conventional HMM has to assign each observation to one of the known states $\mathbf{S}_{\text{known}}=\{S_1, S_2, S_3, S_4\}$ according to the posterior probability calculation via Eq. (4) even when an observation signal is emitted from the unknown state S_5 , where

$$\mu_5 = \begin{bmatrix} 28 \\ 28 \end{bmatrix}, \quad \Sigma_5 = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$$

The posterior probabilities of being in each state given the sequence of the observation symbols up to the current time are obtained and illustrated in Fig. 6. The wiggling in the posterior probabilities happens after approximately the 400th sample, since the conventional HMM does not account for the emergence of the unknown state. In this case, the conventional HMM is unable to estimate the correct states. On the other hand, the wiggling in the posterior probabilities represents the presence of an unknown state from the observation symbols. From the result shown in Fig. 6, we might conclude that the posterior probability is the key criterion in determining the detection of unknown states, as explained in Ref. [13]. However, the following case study shows that this conclu-

Table 2 Three states of HMM based on different tool flank wears

States	S_1	S_2	S_3
Pictures			
Tool flank wear (μm)	79.05 ± 0.005	103.70 ± 0.005	151.80 ± 0.005
Mean of the two forces N	$[108.5 \ 124.6]$	$[166.7 \ 251.1]$	$[230.4 \ 404.8]$
Covariance matrix of the two forces	$\begin{bmatrix} 226.0 & 199.1 \\ 199.1 & 242.0 \end{bmatrix}$	$\begin{bmatrix} 151.1 & 547.7 \\ 547.7 & 2198.7 \end{bmatrix}$	$\begin{bmatrix} 159.8 & 234.2 \\ 234.2 & 538.4 \end{bmatrix}$

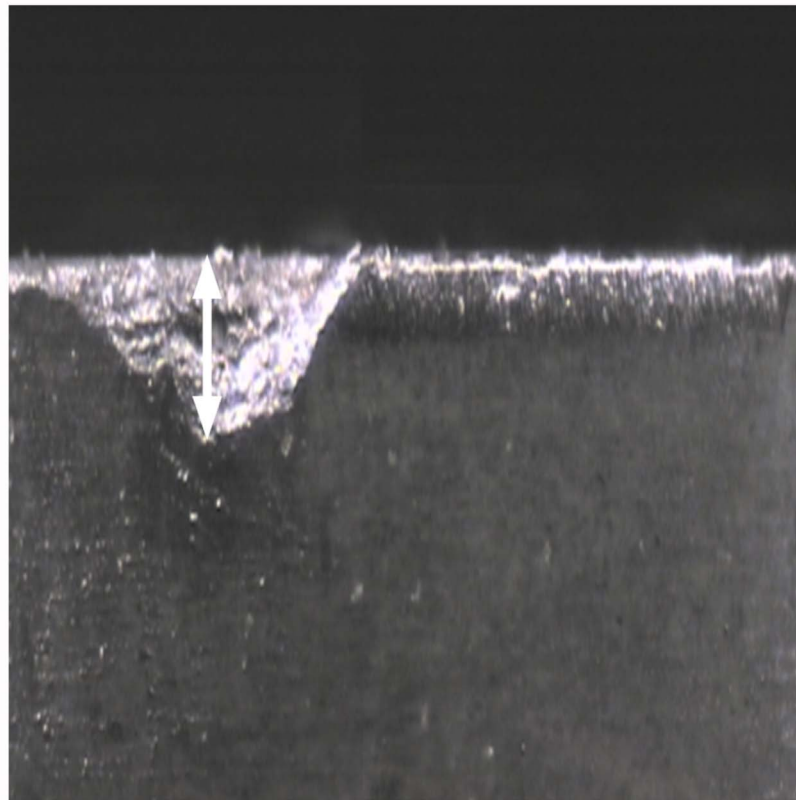
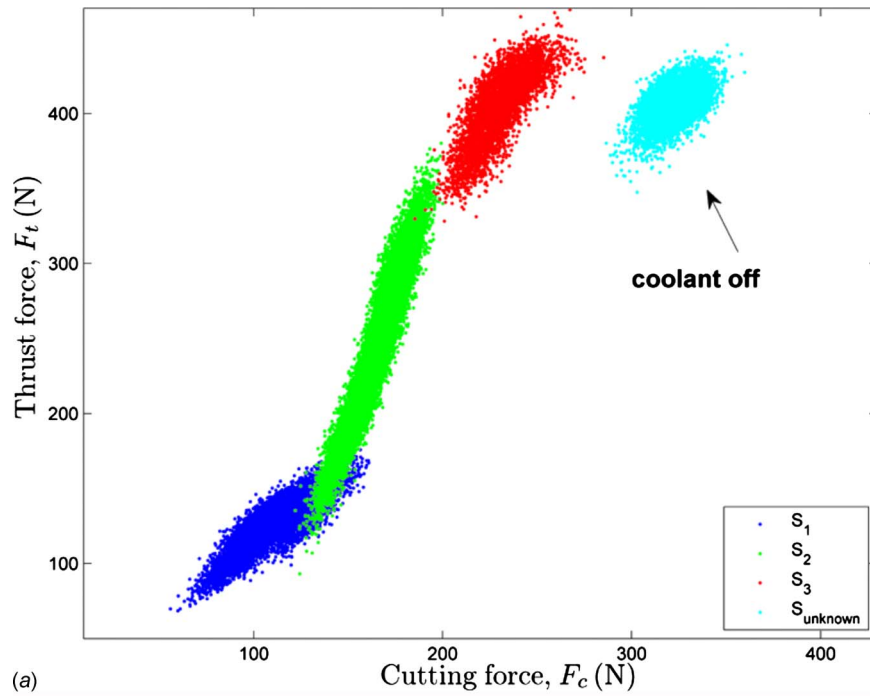


Fig. 10 (a) Normal turning process with coolant and (b) different tool wear modes without coolant

sion may not always be true.

The trained HMM $\lambda=(P, b, \pi)$ is the same as in the previous example. However, in this case, it turns out that an unknown state S_5 has the following Gaussian observation symbol density distribution

$$\mu_5 = \begin{bmatrix} 50 \\ 10 \end{bmatrix}, \quad \Sigma_5 = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$$

Instead of being in the middle of the other states, the unknown state is far away from the other four known states, as shown in

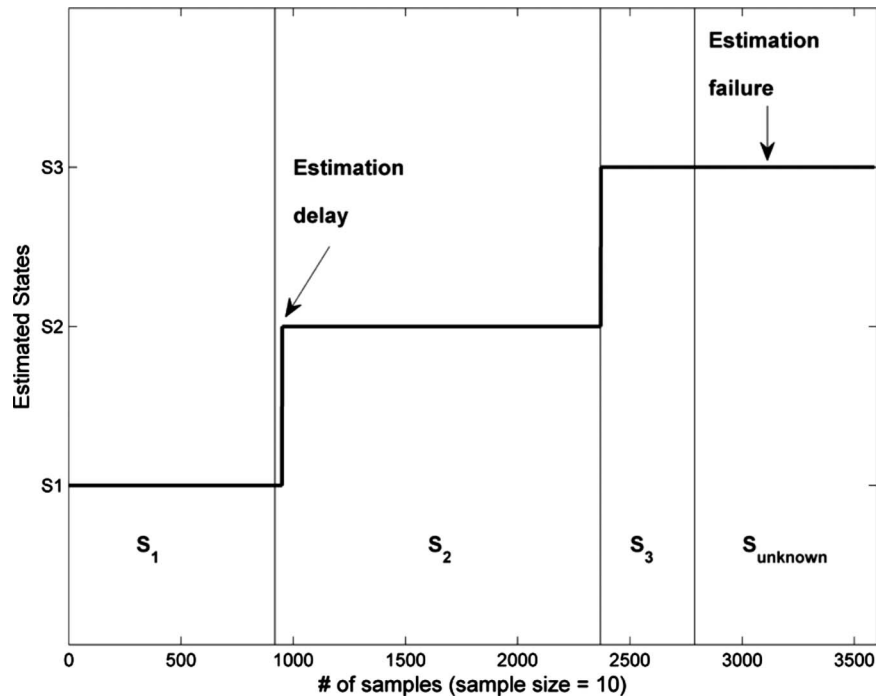


Fig. 11 Estimated states: vertical dashed lines indicate the true states, while the solid lines represent the estimated states

Fig. 7.

As shown in Fig. 8, we cannot see the wiggling in the posterior probabilities even with the presence of the unknown state in Fig. 7. In this case, the conventional HMM algorithm disguises the unknown state by calculating $P\{q(n)=S_1|O(1), \dots, O(n)\} = 0.9972$ after around the 650th sample. The conventional HMM misinterprets an unknown state S_5 as the first state S_1 with a high probability even though the unknown state is located far from the first state S_1 . The fourth row of the transition probability matrix $P(4, :) = [0.01 \ 0.00 \ 0.00 \ 0.99]$ is defined in such a way that a state will move to either state S_1 or state S_4 after being in state S_4 . The conventional HMM, however, excludes the chance of being in state S_4 after observing that an observation symbol is far away from state S_4 . Thus, the conventional HMM misjudges that $P\{q(n)=S_1|O(1), \dots, O(n)\} = 0.9972$.

These two examples lead us to conclude that considering only the posterior probability in the identification of the unknown state is not sufficient, based on the conventional HMM. This is why we propose the modified HMM algorithm to deal with challenges related to unknown states using the Hotelling multivariate control chart. We will illustrate how the MHMM operates in Sec. 3.2 with an example of the tool degradation process.

3.2 Case Study on the Tool Wear of the Turning Process.

The proposed MHMM has been tested with a turning process and is shown to be able to perform an adaptive diagnosis of the different failure modes as well as online degradation assessment. A ceramic tool is used to turn an Inconel718 workpiece with coolant supplied, as shown in Fig. 9. During the turning process, two orthogonal forces (the cutting and thrust forces) are measured by the dynamometer.

The first step is to train the MHMM using the training data sets associated with each state. The states are defined as the degree of the tool flank wear. Three different degrees of tool flank wears $S = \{S_1, S_2, S_3\}$ are used to train the MHMM. The cutting and thrust forces are measured under the same turning process conditions such as the depth of cut, feed rate, and cutting speed (see Table 1). Note that enough coolant was supplied during this train-

ing stage.

Observation symbol probability distributions for each state are then calculated from two force signals in the form of the joint Gaussian density functions. The resultant mean and covariance matrix with corresponding tool wears are displayed in Table 2.

We then restart the turning process with a new tool while measuring the cutting and thrust forces. As shown in Fig. 10, both cutting and thrust forces increase with the process duration as the cutting tool loses its sharpness.

After the tool wear status reaches state S_3 , the coolant supply is removed to introduce a different tool wear mode. The cutting force seems to increase when the coolant is not supplied. This dry machining condition generates nonexperienced forces from an unknown state S_{unknown} that has not been seen during the training stage. Figures 11 and 12 demonstrate the problem or drawback of the conventional HMM, showing that a conventional HMM fails to estimate S_{unknown} with high Mahalanobis distances. The distance statistic D^2 becomes larger than the UCL after around the 2800th sample, which corresponds to the moment when the coolant is shut off. The estimation failure in Fig. 11 causes higher Mahalanobis distance in Fig. 12. On the other hand, the MHMM is able to update its structure to add new states successfully by calculating a statistical distance between the current forces and known states. Since the MHMM has a new state to represent an unknown condition, the Mahalanobis distance between the incoming data and the new state is less than the UCL, as shown in Fig. 13. Although the estimation delay from state S_1 to state S_2 causes some nonconsecutive data points to be out of control, these points are not statistically significant to add another state in the MHMM. However, the appearance of unknown states after the 2800th sample does trigger the MHMM to add another state, resulting in Fig. 13. It is critical to diagnose coolant shortage as early as possible to avoid excessive tool wear, as shown in Fig. 10(b). The appearance of unknown states can be identified through the emergence of a new state in the MHMM. Figures 12 and 13 illustrate that the MHMM is not only a stochastic modeling technique but also an adaptive fault detector.

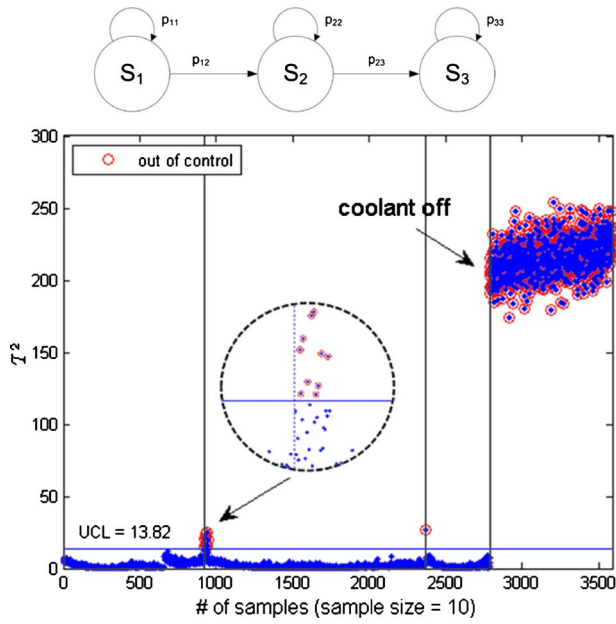


Fig. 12 Control chart with conventional HMM

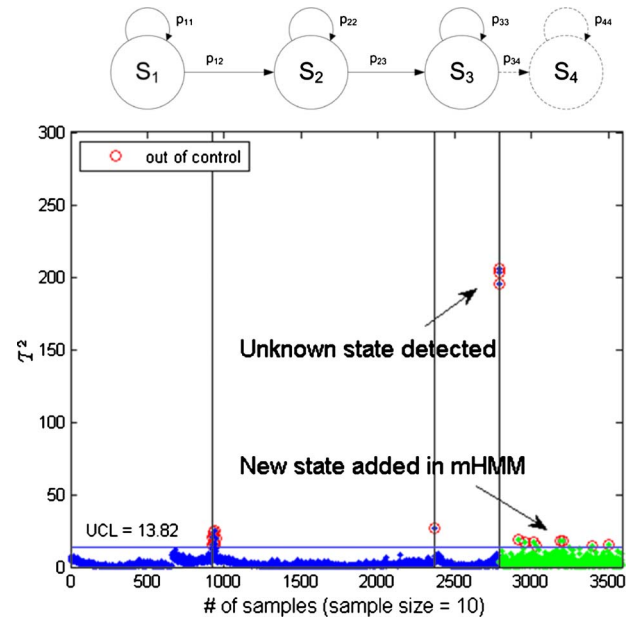
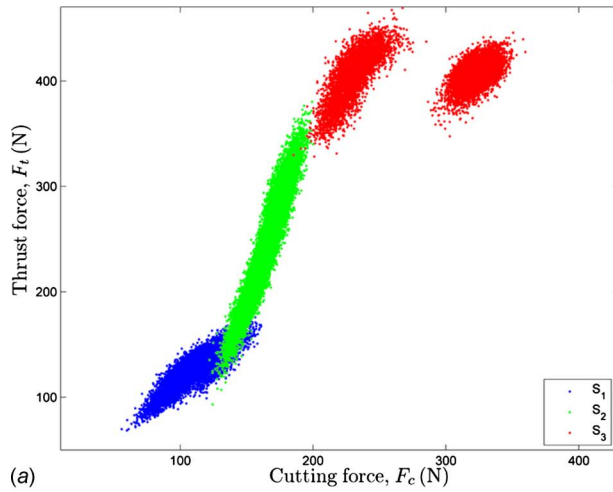
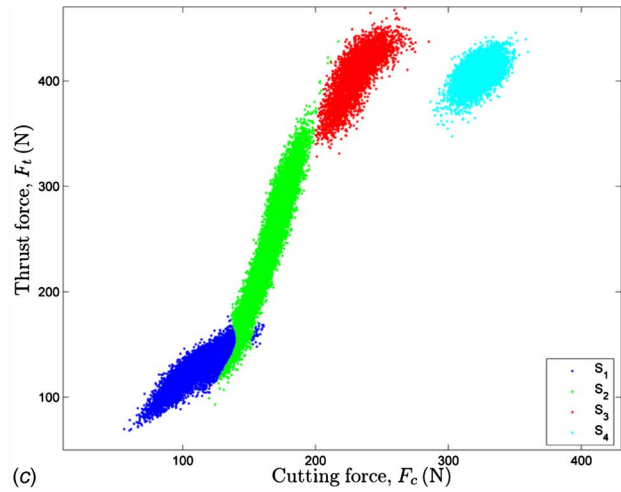


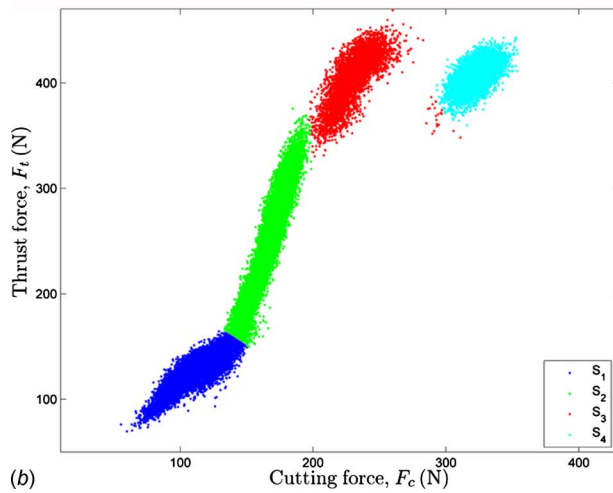
Fig. 13 Control chart with MHMM



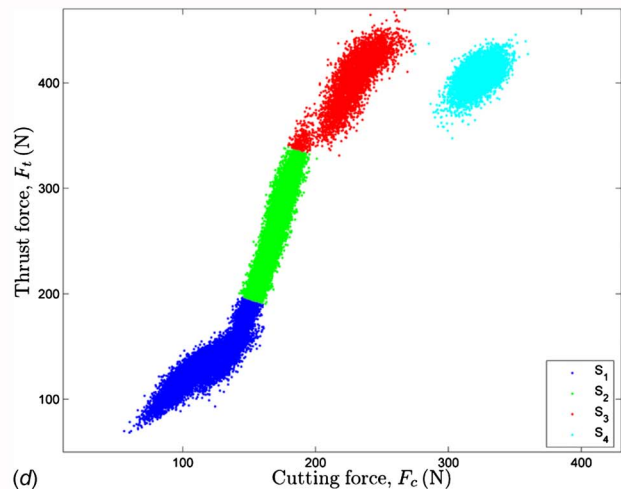
(a)



(c)



(b)



(d)

Fig. 14 The estimated states via various algorithms: (a) HMM, (b) neural network, (c) GMM, and (d) K-means

Table 3 Correct estimation rate comparison

Estimation methods	Accuracy (%)
MHMM	99.06
HMM	77.69
Neural networks	97.71
GMM	96.49
K-means	92.81

We have also compared the proposed MHMM with other typical clustering algorithms such as neural networks, Gaussian mixture model (GMM), and K-means clustering [21,25]. Artificial neural networks are motivated by biological neural networks and have been used extensively over the past three decades for both classification and clustering [28]. GMM is based on the idea that the data can be clustered using a mixture of multivariate Gaussian distributions. On the other hand, K-means is the simplest and most commonly used algorithm. K-means starts with a random initial partition and keeps re-assigning the patterns to clusters based on the similarity between the patterns and cluster centers until a convergence criterion is met [29].

The MHMM algorithm is based on online data streaming, which is more applicable in equipment condition diagnosis, while GMM and K-means clustering approaches are based on off-line but unsupervised machine learning. The classification results of the three different algorithms are illustrated in Fig. 14 and summarized in Table 3. The accuracies are calculated by counting the errors between the true state and the state estimated via the clustering algorithms. The MHMM clearly outperforms the others in terms of the estimation accuracy because the MHMM makes use of the information regarding the transition probability as well as the observation symbol distributions.

The MHMM enables the identification of anomalous behavior of a system by measuring Mahalanobis distance. We have shown that the proposed MHMM algorithm is successfully able to modify its structure by increasing the number of states and estimate the state of a system even in the existence of an unknown state.

4 Conclusion and Future Work

In this paper, the MHMM algorithm is developed to deal with variable state space. A method in the SPC has been combined into MHMM for unknown state detection and diagnosis. The results illustrate that the proposed MHMM can (1) estimate current tool conditions more effectively than other classification algorithms such as GMM, K-means, and neural network; (2) detect anomalous behavior or an unknown state at an early stage by using the Hotelling multivariate control chart; and (3) change its structure to represent degradation processes more accurately in the presence of unknown faults.

Future work will involve further experimental validations of the MHMM algorithm. The assumption that the monitoring signals follow the Gaussian density distribution is strict in the case of having different observation symbol probability distributions. The MHMM needs to be modified to handle general distributions. Furthermore, the false detection rate and the average time required to detect faults should be theoretically examined in order to understand the limits of MHMM.

Nomenclature

- $S = \{S_1, \dots, S_M\}$ = state space for the discrete degradation process
- $P = \{p_{ij}\}_{M \times M}$ = state transition probability matrix
- $b_i(O(n))$ = observation symbol probability distribution, $P\{O(n) | q(n) = S_i\}$

- $\pi = \{\pi_i\}_M$ = initial state probability distribution, $\pi_i = P\{q(1) = S_i\}$
- $O_n = \{O(1), \dots, O(n)\}$ = sequence of all observation symbols up to time n
- $q_n = \{q(1), \dots, q(n)\}$ = actual state sequence up to time n
- $\hat{q}_n = \{\hat{q}(1), \dots, \hat{q}(n)\}$ = maximum likelihood state sequence up to time n
- $\lambda = (P, b, \pi)$ = HMM model parameters
- $\alpha_n(i) = \alpha_n(i) = P\{O(1), \dots, O(n) \wedge q(n) = S_i | \lambda\}$
- $\xi_n(i, j) = \xi_n(i, j) = P\{q(n) = S_i \wedge q(n+1) = S_j | O_N \wedge \lambda\}$
- $\gamma_n(i) = \gamma_n(i) = P\{q(n) = S_i | O_N \wedge \lambda\}$
- μ = mean value
- Σ = covariance matrix
- $\bar{O}(n)$ = vector of the observation symbol mean of the n th observation
- $D^2(\bar{O}(n), \mu_i)$ = weighted distance of $\bar{O}(n)$ from the μ_i

References

- [1] Jardine, A. K. S., Lin, D., and Banjevic, D., 2006, "A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance," *Mech. Syst. Signal Process.*, **20**(7), pp. 1483–1510.
- [2] Heng, A., Zhang, S., Tan, A. C. C., and Mathew, J., 2009, "Rotating Machinery Prognostics: State of the Art, Challenges and Opportunities," *Mech. Syst. Signal Process.*, **23**(3), pp. 724–739.
- [3] Wang, W., and Christer, A. H., 2000, "Towards a General Condition Based Maintenance Model for a Stochastic Dynamic System," *J. Oper. Res. Soc.*, **51**(2), pp. 145–155.
- [4] Wang, G., Luo, Z., Qin, X., Leng, Y., and Wang, T., 2008, "Fault Identification and Classification of Rolling Element Bearing Based on Time-Varying Autoregressive Spectrum," *Mech. Syst. Signal Process.*, **22**(4), pp. 934–947.
- [5] Randall, R. B., Antoni, J., and Chobsaard, S., 2001, "The Relationship Between Spectral Correlation and Envelope Analysis in the Diagnostics of Bearing Faults and Other Cyclostationary Machine Signals," *Mech. Syst. Signal Process.*, **15**(5), pp. 945–962.
- [6] Bonato, P., Ceravolo, R., De Stefano, A., and Knaflitz, M., 1997, "Bilinear Time-Frequency Transformations in the Analysis of Damaged Structures," *Mech. Syst. Signal Process.*, **11**(4), pp. 509–527.
- [7] Paya, B. A., Esat, I. I., and Badi, M. N. M., 1997, "Artificial Neural Network Based Fault Diagnostics of Rotating Machinery Using Wavelet Transforms as a Preprocessor," *Mech. Syst. Signal Process.*, **11**(5), pp. 751–765.
- [8] Jin, J., and Shi, J., 2000, "Diagnostic Feature Extraction From Stamping Tonnage Signals Based on Design of Experiments," *ASME J. Manuf. Sci. Eng.*, **122**(2), pp. 360–369.
- [9] Rabiner, L. R., 1989, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, **77**(2), pp. 257–286.
- [10] Ertunc, H. M., Loparo, K. A., and Ocak, H., 2001, "Tool Wear Condition Monitoring in Drilling Operations Using Hidden Markov Models (HMMs)," *Int. J. Mach. Tools Manuf.*, **41**(9), pp. 1363–1384.
- [11] Wang, L., Mehrabi, M. G., and Kannatey-Asibu, J. E., 2002, "Hidden Markov Model-Based Tool Wear Monitoring in Turning," *ASME J. Manuf. Sci. Eng.*, **124**(3), pp. 651–658.
- [12] Li, Z., Wu, Z., He, Y., and Fulei, C., 2005, "Hidden Markov Model-Based Fault Diagnostics Method in Speed-Up and Speed-Down Process for Rotating Machinery," *Mech. Syst. Signal Process.*, **19**(2), pp. 329–339.
- [13] Smyth, P., 1994, "Markov Monitoring With Unknown States," *Selected Areas in Communications, IEEE Journal on*, **12**(9), pp. 1600–1612.
- [14] Smyth, P., 1994, "Hidden Markov Models for Fault Detection in Dynamic Systems," *Pattern Recognit.*, **27**(1), pp. 149–164.
- [15] Tang, K., Williams, W. W., Jwo, W., and Gong, L. G., 1999, "Performance Comparison Between On-Line Sensors and Control Charts in Manufacturing Process Monitoring," *IIE Trans.*, **31**(12), pp. 1181–1190.
- [16] Ross, S. M., 1996, *Stochastic Processes, Probability and Statistics*, Wiley, New York.
- [17] Viterbi, A., 1967, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inf. Theory*, **13**(2), pp. 260–269.
- [18] Forney, G. D., 1973, "The Viterbi Algorithm," *Proc. IEEE*, **61**(3), pp. 268–278.
- [19] Baum, L. E., and Petrie, T., 1966, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Ann. Math. Stat.*, **37**(6), pp. 1554–1563.
- [20] Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, "Maximum Likelihood From Incomplete Data via the EM Algorithm," *J. R. Stat. Soc. Ser. B (Methodol.)*, **39**(1), pp. 1–38.
- [21] Duda, R. O., Hart, P. E., and Stork, D. G., 2001, *Pattern Classification*, Wiley, New York.

- [22] Montgomery, D. C., 2004, *Introduction to Statistical Quality Control*, Wiley, New York.
- [23] Lowry, C. A., and Montgomery, D. C., 1995, "A Review of Multivariate Control Charts," *IIE Trans.*, **27**(6), pp. 800–810.
- [24] Jackson, J. E., 1985, "Multivariate Quality Control," *Commun. Stat: Theory Meth.*, **14**(11), pp. 2657–2688.
- [25] Bishop, C. M., 2006, *Pattern Recognition and Machine Learning*, Springer, New York.
- [26] Ryan, T. P., 2000, *Statistical Methods for Quality Improvement*, Wiley, New York.
- [27] Nelson, L. S., 1984, "The Shewhart Control Chart—Test for Special Cause," *J. Quality Technol.*, **16**(4), pp. 237–239.
- [28] Jain, A. K., Mao, J., and Mohiuddin, K. M., 1996, "Artificial Neural Networks: A Tutorial," *IEEE Computer*, **29**(3), pp. 31–44.
- [29] Jain, A. K., Murty, M. N., and Flynn, P. J., 1999, "Data Clustering: A Review," *ACM Comput. Surv.*, **31**(3), pp. 264–323.