SIViP (2013) 7:189–195
DOI 10.1007/s11760-011-0215-y

ORIGINAL PAPER

A variance-estimation-based stopping rule for symbolic dynamic filtering

Yicheng Wen · Asok Ray · Qiang Du

Received: 19 August 2010 / Revised: 10 December 2010 / Accepted: 25 January 2011 / Published online: 18 February 2011
© Springer-Verlag London Limited 2011

Abstract As an alternative to the batch means (BM) method in the stopping rule for symbolic dynamic filtering, this short paper presents an analytical procedure to estimate the variance parameter and to obtain a lower bound on the length of symbol blocks for constructing probabilistic finite state automata (PFSA). If the modulus of the second largest eigenvalue of the PFSA's state transition matrix is relatively small or if the symbol block length is not too large, then the performance of the proposed stopping rule is superior to that of the stopping rule based on BM method. The algorithm of the proposed stopping rule is validated on ultrasonic data collected from a fatigue test apparatus for damage detection in the polycrystalline alloy 7075-T6.

Keywords Probabilistic finite state automata · Markov chains · Variance estimation

Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

Y. Wen · A. Ray (✉)
Mechanical Engineering Department,
The Pennsylvania State University,
University Park, PA 16802, USA
e-mail: axr2@psu.edu

Y. Wen
e-mail: yxw167@psu.edu

Q. Du
Mathematics Department,
The Pennsylvania State University,
University Park, PA, 16802, USA
e-mail: qud2@psu.edu

1 Introduction

Recently Wen and Ray [1] have reported a stopping rule for symbolic dynamic filtering (SDF) [2], where a key issue is how to obtain a lower bound on the length of symbol blocks for computing the state probability vector of probabilistic finite state automata (PFSA). The stopping rule is formulated in the setting of Markov chain Monte Carlo (MCMC) and its key parameter to be estimated is the variance of the symbol sequence in the central limit theorem.

In the MCMC literature, there are many well-studied variance estimation techniques (e.g., batch means [3], regenerative simulation [4] and spectral variance estimators [5]). The Markov chain is known a priori in MCMC and its parameters contribute to the variance estimation. In the current context, the parameters of the Markov chain are unknown and need to be estimated. Hence, both regenerative simulation and spectral variance estimators, which rely on the knowledge of the chain parameters, are not suitable in the present formulation; and the BM method, used in an earlier stopping rule [1], is an asymptotic numerical tool that is purely based on the simulated data. However, the BM method requires partitioning the time series into several batches of data by heuristic rules, where an inappropriate choice of the batch size could adversely affect the convergence rate of the variance estimator. It has been observed in [1] that the BM method may require an unnecessarily long observed sequence, which is not conducive to the quasi-stationarity assumption in SDF.

Within a similar framework, the current paper formulates an alternative stopping rule, where a theoretical upper bound of the variance is derived from online estimates of the state transition matrix and the state probability vector of the Markov chain. It is shown by numerical simulations that the proposed variance estimation approach enhances the performance of the stopping rule, especially if the modulus of

the second largest eigenvalue of the PFSA's state transition matrix is relatively small or if the symbol block is not very long. The algorithm of the proposed stopping rule is validated on ultrasonic data collected from a fatigue test apparatus for damage detection in the polycrystalline alloy 7075-T6.

The paper is organized as follows. The stopping rule problem is formulated in Sect. 2. The algorithm of the stopping rule is developed in Sect. 3. Numerical results are presented in Sect. 4 to evaluate the performance of the proposed stopping rule algorithm relative to the algorithm [1] based on the BM method. The application for fatigue damage detection in polycrystalline alloys is discussed in Sect. 5. The paper is concluded in Sect. 6.

2 Statement of the problem

Let a statistically (quasi-)stationary dynamical system be modeled as a stationary Markov chain $\mathbb{S} = \{s_0, s_1, s_2, \ldots\}$ of finite order D, where D is a positive integer. Let the symbols $s_i \in \mathbb{S}$ belong to a (finite) alphabet Σ and let Σ^* be the set of all finite-length strings of symbols including the null string ϵ. Each state of $\mathbb{S}$ is labeled with a symbol block of length D belonging to Σ^* and let j be the unique label of the jth state. For example, if $\Sigma = 0, 1$ and $D = 2$, then the possible states are as follows: 00, 01, 10, and 11.

If $\mathbb{S}$ is an ergodic Markov chain (i.e., the associated state transition matrix $\mathbf{\Pi}$ is irreducible), then it follows from the Perron–Frobenius theorem [6,7] that there exists a unique probability vector $\boldsymbol{p} = [p_1, p_2, \ldots, p_n]$, where $n \leq |\Sigma|^D$, such that $\boldsymbol{p}\mathbf{\Pi} = \boldsymbol{p}$ with the constraints: $\sum_i p_i = 1$ and $p_i > 0 \,\forall\, i$. Then, $\boldsymbol{p}$ is called the state probability vector of the Markov chain $\mathbb{S}$. In the setting of SDF, the quasi-stationary state probability vector of the PFSA is commonly selected as a feature vector [8], which captures the statistical property of the dynamical system. Let N_i^r be the number of times the block $\mathbf{s}^r$ visits the state i and N_{ij}^r be the number of times the block $\mathbf{s}^r$ encounters a transition to the state j from the state i.

For a particular symbol block $\mathbf{s}^r = \{s_i\}_{i=1}^r$ generated by $\mathbb{S}$, let $\hat{\boldsymbol{p}}(r) \triangleq [\hat{p}_1(r)\hat{p}_2(r)\ldots\hat{p}_n(r)]$ be the estimated state probability vector. Each element of $\hat{\boldsymbol{p}}(r)$ is defined as

$$\hat{p}_i(r) \triangleq \frac{1}{r}\sum_{j=1}^{r} \mathbb{J}_i \circ T^j(\mathbf{s}^r) = \frac{N_i^r}{r} \tag{1}$$

where $\mathbb{J}_i(x)$ is an indicator function, i.e.,

$$\mathbb{J}_i(x) = \begin{cases} 1 & \text{if the state } i \text{ is a prefix of the string } x \\ 0 & \text{otherwise} \end{cases}$$

and T is the left shift operator [i.e., $T(s_1 s_2 s_3 \ldots) = s_2 s_3 \ldots$].

By the ergodic theorem for Markov chain [6], it is guaranteed that $\hat{\boldsymbol{p}}(r) \to \boldsymbol{p}$ as $r \to \infty$. The problem is to find a *minimal* stopping point r_{stop} such that $\hat{\boldsymbol{p}}(r_{\text{stop}})$ computed from a symbol block of length r_{stop} satisfies the following condition:

$$\|\hat{\boldsymbol{p}}(r_{\text{stop}}) - \boldsymbol{p}\|_\infty < \varepsilon \quad \text{with a confidence level } (1-\alpha) \tag{2}$$

where $\| \bullet \|_\infty$ is the max norm of the finite-dimensional vector $\bullet$ and ε is the absolute error bound of estimation.

3 Algorithm development

For the finite-order Markov chain $\mathbb{S} = \{s_0, s_1, s_2, \ldots\}$, let $g(\bullet)$ be a real-valued, μ-integrable function on the signal space Ω such that the second moment $E_\mu\, g^2 \triangleq \int_\Omega g^2(x)\mu(\mathrm{d}x) < \infty$. Defining the time average $\bar{g}_r \triangleq \frac{1}{r}\sum_{i=0}^{r-1} g(s_i)$, the estimate $E_\mu\, g \triangleq \int_\Omega g(x)\mu(\mathrm{d}x)$ is obtained by Markov chain Monte Carlo (MCMC) tools and an application of the central limit theorem [4] as

$$\sqrt{r}\left(\bar{g}_r - E_\mu\, g\right) \xrightarrow{\text{distribution}} N(0, \sigma_g^2) \quad \text{as} \quad r \to \infty \tag{3}$$

where the variance $\sigma_g^2 \triangleq \mathrm{Var}_\mu\{g(s_0)\} + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\mu\{g(s_0), g(s_i)\}$. The ergodic theorem [4] yields

$$\bar{g}_r \xrightarrow{\text{almost surely}} E_\mu\, g \quad \text{as} \quad r \to \infty \tag{4}$$

Taking g as the indicator function of the state i, the limit of the time average $\bar{g}_r$ becomes p_i, the ith element of the state probability vector $\boldsymbol{p}$.

Next an estimate for the variance σ_g^2 is analytically derived by restricting the irreducible state transition matrix $\mathbf{\Pi}$ to be aperiodic. Then, for each state i, $(\sigma_g)_i^2$ is defined as

$$(\sigma_g)_i^2 \triangleq \mathrm{Var}_\mu\{g(s_0)\} + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\mu\{g(s_0), g(s_k)\} \tag{5}$$

$$= \mathrm{Var}_\mu\{\mathbb{I}(s_0 = i)\} + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\mu\{\mathbb{I}(s_0 = i), \mathbb{I}(s_k = i)\} \tag{6}$$

Assuming that the initial condition s_0 has a stationary distribution $\boldsymbol{p}$, it follows that

$$\Pr(s_k = i) = p_i \quad \text{and} \quad \Pr(s_k \neq i) = 1 - p_i \tag{7}$$

Therefore, we obtain

$$\mathbb{E}\{\mathbb{I}(s_k = i)\} = p_i \quad \text{and} \quad \mathrm{Var}\{\mathbb{I}(s_k = i)\} = p_i(1 - p_i) \tag{8}$$

For the second term in Eq. (6), we have

$$\begin{aligned} &\mathrm{Cov}\{\mathbb{I}(s_0 = i), \mathbb{I}(s_k = i)\} \\ &= \mathbb{E}(\mathbb{I}(s_0 = i) \cdot \mathbb{I}(s_k = i)) - \mathbb{E}(\mathbb{I}(s_0 = i))\mathbb{E}(\mathbb{I}(s_k = i)) \\ &= \Pr(\mathbb{I}(s_0 = i) = 1, \mathbb{I}(s_k = i) = 1) - p_i^2 \\ &= \Pr(s_0 = i)\Pr(s_k = i | s_0 = i) - p_i^2 \\ &= p_i \Pr(s_k = i | s_0 = i) - p_i^2 \\ &= p_i\left(\mathbf{\Pi}_{ii}^{(k)} - p_i\right) \end{aligned} \tag{9}$$

where $\mathbf{\Pi}_{ii}^{(k)}$ is the ith row and ith column element of the kth power of the state transition matrix $\mathbf{\Pi}$. By the ergodicity property, Eq. (9) approaches zero when k becomes large, which implies that s_0 and s_k are uncorrelated in the limit. Then, Eq. (6) becomes

$$(\sigma_g)_i^2 = p_i(1-p_i) + 2p_i \sum_{k=1}^{\infty} \left(\mathbf{\Pi}_{ii}^{(k)} - p_i\right) \tag{10}$$

The "plug-in" rule is directly applied here to obtain an estimate of $(\sigma_g)_i^2$ based on the consistent estimates of p_i and Π online, namely,

$$\hat{\Pi}_{ij} = \frac{N_{ij}^r}{N_i^r} \quad \text{and} \quad \hat{p}_i = \frac{N_i^r}{r} \tag{11}$$

Since there is an infinite sum in Eq. (10), we need to choose where to truncate the infinite tail. To bound the infinite tail, we introduce the following classical convergence theorem with a proposition for finite Markov chains.

Theorem 3.1 ([9]) *Let $\mathbb{S}$ be a finite-order Markov chain. If the state transition matrix $\mathbf{\Pi}$ is irreducible and aperiodic, then there exists $0 < q < 1$ and $b > 0$ such that*

$$\left\| \mathbf{\Pi}_{i\cdot}^{(k)} - \boldsymbol{p} \right\|_{\infty} < bq^k \tag{12}$$

where $\boldsymbol{p}$ is the unique state probability vector and $\mathbf{\Pi}_{i\cdot}^{(k)}$ is the ith row of of the kth power of the state transition matrix $\mathbf{\Pi}$.

Theorem 3.1 does not specify the values of b and q. Classical results show that the rate of convergence in Eq. (12) is asymptotic as dictated by the modulus of the second largest eigenvalue of the state transition matrix $\mathbf{\Pi}$, but the constant b is not provided for this exact bound. Moreover, the second largest eigenvalue itself is difficult to compute online [10]. For these reasons, it is preferable to use the bound, stated in the following Proposition 3.1, which is easy to compute and is shown to be a good approximation in many cases.

Proposition 3.1 ([11]) *If the conditions in Theorem 3.1 hold, then Eq. (12) is valid with the following choice of b and q:*

$$q = 1 - \beta \quad \textit{and} \quad b = 1 \tag{13}$$

where $\beta = \sum_i \min_j \mathbf{\Pi}_{ij}$, which is simply the sum of the minimum values of the entries in each column of the state transition matrix $\mathbf{\Pi}$.

Let us define the finite sum of the second term in Eq. (10) as follows

$$\left(\sigma_L^2\right)(i) \triangleq p_i(1-p_i) + 2p_i \sum_{k=1}^{L} \left(\mathbf{\Pi}_{ii}^{(k)} - p_i\right) \tag{14}$$

where $L \in \mathbb{N}$. Then, by applying Theorem 3.1 and Proposition 3.1 in the last two steps, the error between $(\sigma_L^2)(i)$ and $(\sigma_g^2)(i)$ is obtained as

$$\begin{aligned} e_L(i) &\triangleq \left| \left(\sigma_g^2\right)(i) - (\sigma_L^2)(i) \right| \le 2p_i \sum_{k=L+1}^{\infty} \left| \mathbf{\Pi}_{ii}^{(k)} - p_i \right| \\ &\le 2p_i \sum_{k=L+1}^{\infty} \left\| \mathbf{\Pi}_{i\cdot}^{(k)} - \boldsymbol{p} \right\|_{\infty} \\ &\le \frac{2p_i b q^{L+1}}{1-q} = \frac{2p_i(1-\beta)^{L+1}}{\beta} \end{aligned} \tag{15}$$

It is still necessary to select the parameter L that should be sufficiently large to make the error $e_L(i)$ small. However, the computational load increases as L becomes larger. In this setting, a ratio $\gamma_L \in (0, 1)$ is introduced.

$$\gamma_L \triangleq \frac{2p_i(1-\beta)^{L+1}}{\beta \sigma_L^2(i)} \tag{16}$$

where the parameter γ_L is the ratio of the tail portion of the sum to the finite sum. The user is free to choose an upper bound γ for γ_L to make a trade-off between accuracy and efficiency of the finite approximation. In most cases, $\gamma_L = 0.1$ or 0.05 should suffice. The infinite sum is truncated after L terms and a bound is added on the tail if $\gamma_L < \gamma$. Defining the estimated variance as $\hat{\sigma}_g^2(i) \triangleq (\sigma_L^2)(i) + \frac{2p_i(1-\beta)^{L+1}}{\beta}$, it follows that

$$\left(\sigma_g^2\right)(i) \le \left(\sigma_L^2\right)(i) + e_L(i) \le \hat{\sigma}_g^2(i) \tag{17}$$

Algorithm 1 Variance Estimation Algorithm

Input: symbol block $\mathbf{s}^r$ of length r, number of states n, parameter $0 < \gamma < 1$;
Output: variance estimation $\hat{\sigma}_g^2$;
Compute N_i^r and N_{ij}^r;
$\gamma_L \leftarrow 1$ and $L \leftarrow 0$;
$\hat{\mathbf{\Pi}}_{ij} \leftarrow \frac{N_{ij}^r}{N_i^r}$;
$\left(\hat{\sigma}_L^2\right)(i) \leftarrow \hat{p}_i(1-\hat{p}_i)$, $\hat{p}_i \leftarrow \frac{N_i^r}{r}$, and $\hat{\beta} \leftarrow \sum_i \min_j \frac{N_{ij}^r}{N_i^r}$;
while $\gamma_L > \gamma$ **do**
  $L \leftarrow L + 1$;
  $\left(\hat{\sigma}_L^2\right)(i) \leftarrow \left(\hat{\sigma}_L^2\right)(i) + 2\hat{p}_i\left(\hat{\mathbf{\Pi}}_{ii}^{(L)} - \hat{p}_i\right)$;
  $\gamma_L = \max_{1 \le i \le n} \left\{ \frac{2\hat{p}_i(1-\hat{\beta})^{L+1}}{\hat{\beta}\hat{\sigma}_L^2(i)} \right\}$;
end while
$\hat{\sigma}_g^2(i) = \left(\hat{\sigma}_L^2\right)(i) + \frac{2\hat{p}_i(1-\hat{\beta})^{L+1}}{\hat{\beta}}$ for all $1 \le i \le n$

4 Numerical results

The following three state transition matrices for 1-D Markov chains [2] are analyzed to compare the performance of the

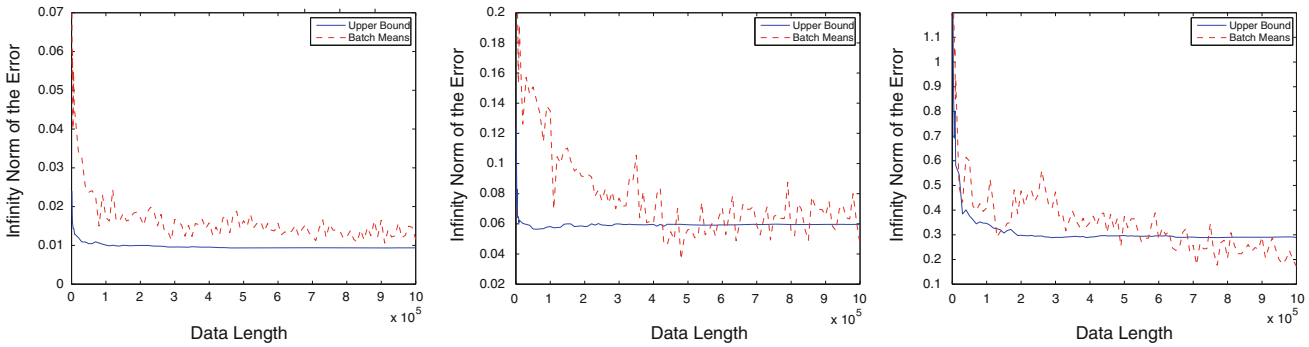

Fig. 1 Comparison of the variance (σ_g^2) estimation errors computed by Algorithm 1 and by the BM method. **a** $\mathbf{\Pi}_1(|\lambda_2| = 0.20)$, **b** $\mathbf{\Pi}_2(|\lambda_2| = 0.70)$, **c** $\mathbf{\Pi}_3(|\lambda_2| = 0.93)$

proposed variance estimation algorithm with that of the BM algorithm [1].

$$\mathbf{\Pi}_1 = \begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.5 & 0.2 \end{pmatrix}, \mathbf{\Pi}_2 = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.7 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \text{ and}$$

$$\mathbf{\Pi}_3 = \begin{pmatrix} 0.4 & 0.5 & 0.1 \\ 0.03 & 0.95 & 0.02 \\ 0.03 & 0.02 & 0.95 \end{pmatrix}$$

Each of the above three (irreducible and aperiodic) stochastic matrices has exactly one unity eigenvalue by the Perron–Frobenius theorem [6,7] and the moduluss of their second largest eigenvalues ($|\lambda_2|$) are 0.20, 0.70, and 0.93, respectively. We conducted 10 independent simulation runs of 1-D Markov chains for each of these state transition matrices. Both Algorithm 1 and the BM method [1] were applied to estimate the variance σ_g^2 for different values of the data length r ranging from 10^3 to 10^6; the parameter γ_L was set to be 0.05 in Algorithm 1 and the batch size was taken as $\sqrt{r}$ in the BM method.

Figure 1 exhibits the averaged errors between σ_g^2 and $\hat{\sigma}_g^2$ computed by the L_∞ norm as the data length r varies for both algorithms in all three cases. The solid (blue) lines indicate the error curves generated by Algorithm 1 and the dashed (red) curves for the BM method. It is concluded from Fig. 1a that the proposed algorithm outperforms the BM method for the entire range of the data length when $|\lambda_2|$ is relatively small. As $|\lambda_2|$ is increased, the BM method progressively becomes more accurate and outperforms Algorithm 1 for large r as seen in Fig. 1b, c. It is observed that the convergence of the upper bound $\hat{\sigma}_g^2$ in Algorithm 1 is fast in the sense that a more accurate estimate of variance is generated even if the data length r is not very large.

5 Algorithm validation on experimental data

This section presents validation of the stopping rule on ultrasonic data collected from a test apparatus for fatigue damage detection in polycrystalline alloys [12]. A description of the test apparatus, the experimental procedure, and the analysis of experimental data are presented in the next three subsections.

5.1 Test apparatus and experimental procedure

The experimental apparatus is a special-purpose uniaxial fatigue damage testing machine that is instrumented with ultrasonic flaw detectors and an optical traveling microscope, as shown in Fig. 2. The apparatus is operated under load control or strain control at speeds up to 12.5 Hz. The tests have been conducted using center notched 7075-T6 aluminum specimens at a constant amplitude sinusoidal load, where the maximum and minimum loads were kept constant at 87 and 4.85 MPa. The test specimens are 3 mm thick and 50 mm wide and have a slot of 1.58 × 4.5 mm at the center. The central notch is made to increase the stress concentration factor that

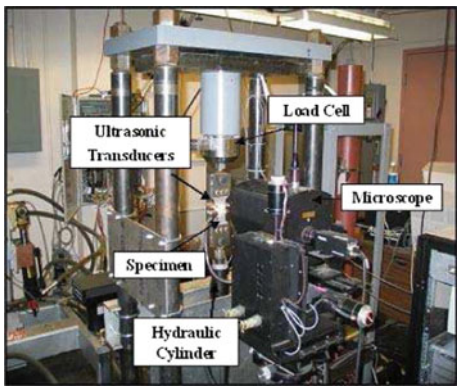

Fig. 2 Computer-instrumented and computer-controlled fatigue test apparatus

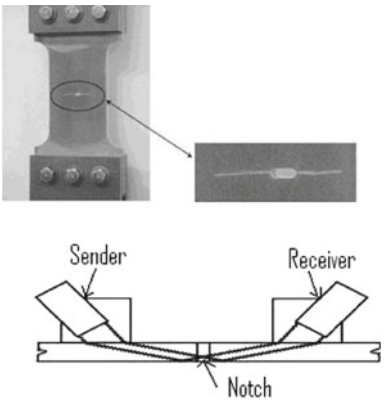

Fig. 3 Schematic of ultrasonic sensing on a test specimen

ensures crack initiation and propagation at the notch ends. The test apparatus is equipped with two types of sensors that have been primarily used for damage detection: traveling optical microscope and ultrasonic flaw detector. The top plate in Fig. 3 shows a cracked specimen and the bottom plate in Fig. 3 illustrates a schematic of the ultrasonic flaw detector mounted on the specimen; the sampling frequency of the ultrasonic sensing device is 20 MHz.

5.2 Experimental procedure

The ultrasonic sensing device is triggered at a frequency of 5 MHz at each peak of the sinusoidally fluctuating load at ∼12.5 Hz. The time epochs, at which data are collected, have been chosen to be 1,000 load cycles (i.e., ∼80 s) apart. At the beginning of each time epoch, the ultrasonic data points are collected for 50 load cycles (i.e., ∼4 s) which produced a string of 15,000 data points. It is assumed that no significant changes occur in the fatigue crack behavior of the test specimen during the tenure of data acquisition at a given time epoch. The nominal condition at the slow-time epoch τ_0 is chosen to be 1.0 kilocycles to ensure that the electro-hydraulic system of the test apparatus comes to a steady state and that no significant damage occurs till that point. The anomalies at subsequent slow-time epochs, $\tau_1, \tau_2, \ldots, \tau_k \ldots$, are then calculated with respect to the nominal condition at τ_0. There are in total 46 epochs, which are taken every 1 kilocycles in the fatigue testing.

5.3 Results of experimental data analysis

Under the nominal condition, the time-series data set is converted to a symbol sequence based on the maximum entropy partition (MEP) [13] on the alphabet of size $|\Sigma| = 6$. Each partition segment is associated with a unique symbol in the alphabet. The symbol sequence characterizes the evolving fatigue damage and is modeled via a first-order Markov chain that has six states. By the property of the MEP, the stationary state probability vector $\boldsymbol{p}_0$ of the resulting probabilistic finite state automaton (PFSA) model is uniformly distributed, that is, $\boldsymbol{p}_0 = \frac{1}{6}\boldsymbol{e}$, where $\boldsymbol{e}$ is the 6-dimensional row vector of all ones.

A pre-specified number ($r_{\min}$) of data points are collected from the test apparatus. A symbol block $\mathbf{s}^{r_{\min}}$ is obtained for each of the time series at individual epochs by using the data partitioning constructed under the nominal condition. Then, Algorithm 1 is used to estimate the variance and r_{stop}. After the stationary probability vector $\hat{p}$ is estimated at an epoch, the damage increment is quantified in terms of the divergence of the probability distribution from the nominal condition. To that end, a statistic test is conducted to obtain the corresponding p-value [14] as the damage measure. It follows from Eq. (3) that $\hat{p}_i \sim N(\boldsymbol{p}_0, \hat{\sigma}_g^2(i))$. Hence, $Z = \sum_{i=1}^{n} (\hat{p}_i(r) - \boldsymbol{p}_0)^2 / \hat{\sigma}_g^2$ follows χ^2-distribution with $n-1$ degrees of freedom and thus the corresponding p-value can be found. This online anomaly detection procedure is described in Algorithm 2.

Algorithm 2 Online Anomaly Detection Algorithm

Input: Observed symbol block $\mathbf{s}^{r_{\min}}$ of length $r_{\min}$ at a slow-time epoch τ, state probability vector under the nominal condition $\boldsymbol{p}_0$, absolute error bound ε, number of states n, significance level α, and variance residual bound γ;
Output: p-value p_{value};
$r_{\text{stop}} = r \leftarrow r_{\min}$;
while 1 **do**
  Compute $\hat{\sigma}_g^2$ for the symbol block $\mathbf{s}^r$ based on Algorithm 1;
  $r_{\text{stop}} = \max_i \frac{(\hat{\sigma}_g)_i^2}{\varepsilon^2} \left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]^2$;
  if $r < r_{\text{stop}}$ **then**
    Observe more data (update $\mathbf{s}^r$);
  else
    break;
  end if
end while{Obtain p-value at the slow-time epoch τ}
Compute $\hat{p}_i(r)$ for all state i by Eq. (1);
Compute test statistics $Z = \sum_{i=1}^{n} (\hat{p}_i(r) - \boldsymbol{p}_0)^2 / \hat{\sigma}_g^2$;
$p_{\text{value}} = 1 - F_n(Z)$;
{where F_n is the cumulative distribution function of χ^2-distribution with $n-1$ degrees of freedom}

Algorithm 2 is implemented using the fatigue data with both proposed variance estimation method (Algorithm 1) and the BM method. Figure 4a shows the computed p-values versus the increasing cycles of the specimen. We notice that both methods give similar p-values. Figure 4b gives the r_{step} for specimen at different cycles computed by two methods. It is noted that the BM method results in much larger r_{step} than our proposed method. In other words, the BM method requires much more data points to compute the stationary probability vector than our method. This illustrates that our proposed method is more efficient than the BM method under certain circumstance.

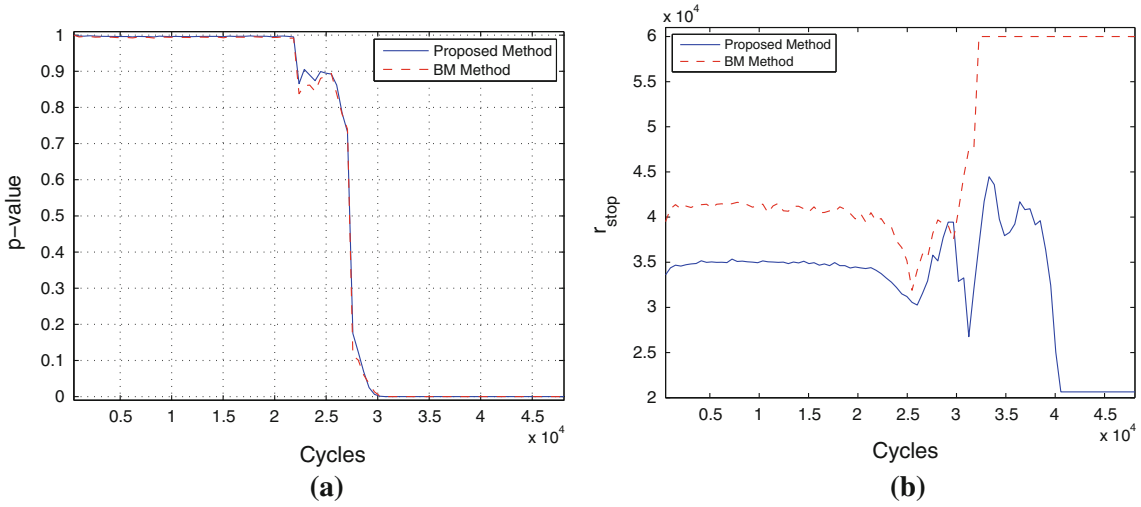

Fig. 4 Comparison of the results. **a** p-Value. **b** Comparison of r_{stop}

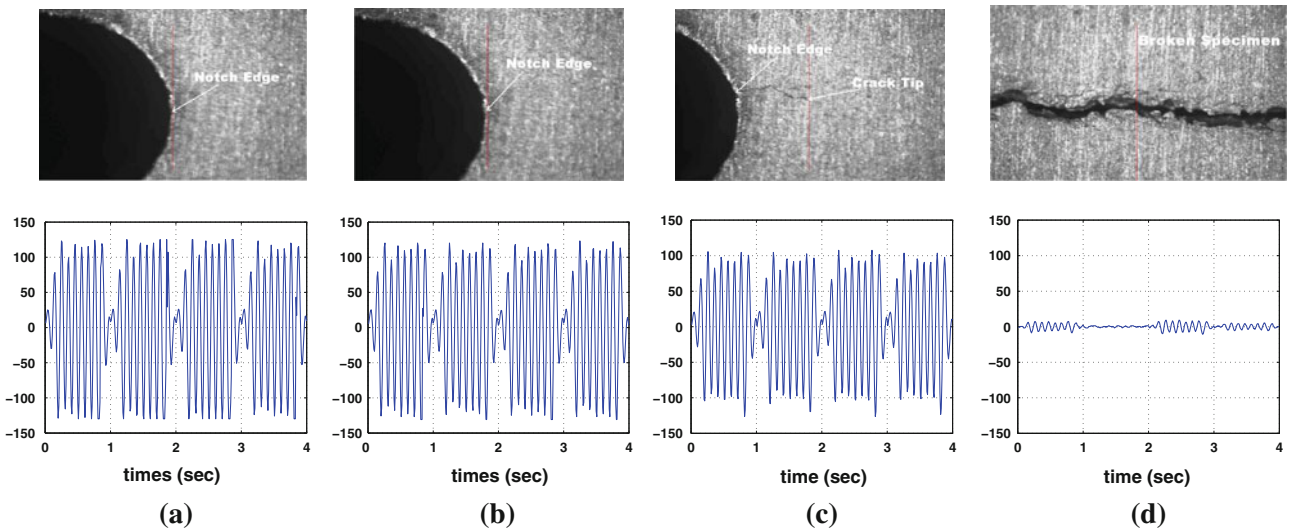

Fig. 5 Optical microscope images of surface crack evolution. **a** Nominal condition at 5 kilocycles. **b** Internal damage at 15 kilocycles. **c** Surface at 23 kilocycles. **d** Fully developed crack at 34 kilocycles

To better understand the physical meaning of the p-values in Fig. 4a, the two-dimensional (optical microscope) images of the specimen surface and the corresponding profiles of ultrasonic sensor outputs are, respectively, presented in the top row and bottom row of Fig. 5 at different slow-time epochs, approximately at 5, 15, 23, and 34 kilocycles, to exhibit the gradual evolution of fatigue damage from the reference condition. As seen in Fig. 4a, the p-values remain essentially unchanged prior to 23 kilocycles. The fluctuations in p-values and r_{stop} around 23 kilocycles, as seen in Fig. 4a, b, result possibly due to the uncertainties in the phase transition from crack initiation to crack propagation. Figures 5a, b show that there is no visible surface crack until ∼23 kilocycles. Beyond 23 kilocycles, Fig. 5c shows the appearance of a crack on the image of the specimen surface and the corresponding p-values are also seen to decrease to ∼0.9 at around 23 kilocycles. Afterward, the p-values dramatically decrease to close to zero in agreement with Fig. 5d when the surface crack is fully developed at 34 kilocycles.

5.4 Discussion

This SDF-based anomaly detection algorithm is formulated in [2] and can be applied to detect slowly evolving anomalies in any dynamical systems that satisfy the following two assumptions:

1. The system behavior is stationary at the fast time scale of the process dynamics;
2. An observable non-stationary behavior of the dynamical system can be associated with anomaly(ies) evolving at a slow-time scale.

The statistical patterns captured by the state probability vectors for the symbol sequences at the different slow-time epochs are compared to the reference pattern under the nominal condition for making decisions on detection of anomaly behavior. The proposed stopping rule allows to circumvent the effects of uncertainties in the state probability vectors due to the restriction of finite-horizon observations by computing the p-values. The proposed stopping rule generalizes the SDF-based anomaly detection algorithm, described in [2], by adaptation of decision making to the observed data length.

6 Summary and conclusions

This short paper presents a stopping rule for symbolic dynamic filtering (SDF) [2] by variance estimation, where a theoretical upper bound of the variance is derived from online estimates of the state transition matrix and the state probability vector of the underlying Markov chain. The proposed stopping rule is an alternative to another stopping rule [1] that is based on the batch means (BM) method [4]. The proposed algorithm uses online consistent estimates for both the state transition matrix and state probability vector of the Markov chain to calculate an upper bound on the variance. The performance of the proposed variance-estimation-based stopping rule is shown to be superior to that of the BM method [1] if the modulus of the second largest eigenvalue is small, or if the data length is not too large. Due to the quasi-stationarity assumption in SDF, the proposed algorithm is more suitable to compute the variance than the BM method in the stopping rule in many applications; this is demonstrated by numerical simulations. Efficacy of the proposed stopping rule is demonstrated on experimental data for detection of fatigue damage in the 7075-T6 polycrystalline alloy.

Acknowledgments This work has been supported in part by the US Office of Naval Research under Grant No. N00014-09-1-0688, by the US Army Research Laboratory and the US Army Research Office (ARO) under Grant No. W911NF-07-1-0376, and by National Science Foundation (NSF) under Grant No. DMS-016073.

References

1. Wen, Y., Ray, A.: A stopping rule for symbolic dynamic filtering. Appl. Math. Lett. **23**, 1125–1128 (2010)
2. Ray, A.: Symbolic dynamic analysis of complex systems for anomaly detection. Signal Process. **84**(7), 1115–1130 (2004)
3. Flegal, J., Haran, M.: Markov chain Monte Carlo: can we trust the third significant figure. Stat. Sci. **23**(2), 250–260 (2008)
4. Jones, G., Haran, M., Caffo, B., Neath, R.: Fixed-width output analysis for Markov chain Monte Carlo. J. Am. Stat. Assoc. **101**, 1537–1547 (2006)
5. Flegal, J., Jones, G.: Batch means and spectral variance estimators. In Markov chain Monte Carlo. technical report, University of Minnesota, Department of Statistics (2008)
6. Berman, A., Plemmons, R.: Nonnegative Matrices in the Mathematical Sciences. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1997)
7. Bapat, R., Raghavan, T.: Nonnegative Matrices and Applications. Cambridge University Press, Cambridge (1997)
8. Gupta, S., Ray, A.: Statistical mechanics of complex systems for pattern identification. J. Stat. Phys. **134**(2), 337–364 (2009)
9. Bhattacharya, R., Waymire, E.: Stochastic Processes with Applications. Wiley-Interscience, New York (1990)
10. Garren, S.T., Smith, R.L. : Estimating the second largest eigenvalue of a Markov transition matrix. Bernoulli **6**, 215–242 (2000)
11. Rosenthal, J.: Convergence rates for Markov chains. Soc. Ind. Appl. Math. J. **37**(3), 387–445 (1995)
12. Gupta, S., Ray, A., Keller, E.: Symbolic time series analysis of ultrasonic data for early detection of fatigue damage. Mech. Syst. Signal Process. **21**(2), 866–884 (2007)
13. Rajagopalan, V., Ray, A.: Symbolic time series analysis via wavelet-based partitioning. Signal Process. **86**(11), 3309–3320 (2006)
14. Schervish, M.: p-values: what they are and what they are not. Am Stat. **50**, 203–206 (1996)