

Physics-based prognostics-promises and challenges

Yiwei Wang¹, Nam H. Kim², and Raphael T. Haftka³

¹*Université de Toulouse, Toulouse, France*

wangyiwei_1988@hotmail.com

^{2,3}*University of Florida, Gainesville, Florida, 32611, USA*

nkim@ufl.edu

haftka@ufl.edu

ABSTRACT

In this paper, an interesting observation on the noise-dependent performance of prognostics algorithms is presented, as well as a method of evaluating the accuracy of prognostics algorithms without having the true degradation model is presented. We found that the randomness in the noise leads to very different ranking of the algorithms for different datasets. In particular, even for the algorithm that has the best performance on average, poor results can be obtained for some datasets. In absence of true damage information, we propose a metric, mean squared discrepancy (MSD), which measures the difference between the prediction and the data. It is shown that the ranking by MSD is strongly correlated with ranking with true degradation model. This may be particularly useful when information is available from multiple sites of damage for the same application.

1. INTRODUCTION

Model-based prognostic approaches can provide a better performance than data-driven approaches when a degradation model is available (An et al., 2015). However, we have not found studies of the effect of randomness in the data on the ranking of algorithms, which is the objective of the present paper.

Classical metrics such as the prognostic horizon, α - λ accuracy, (cumulative) relative accuracy and convergence (Saxena et al., 2010) require the knowledge of true damage degradation information, which in practice is not available. In this paper, we focus on four most commonly used model-based algorithms and verify their performance through a simple degradation model with multiple simulated measurement datasets.

The conventional metric, the mean squared error (MSE) measuring the difference between predicted and the true crack size, is firstly utilized to rank the four algorithms in terms of accuracy assuming the true information on crack

growth is available. We examine how much the ranking changes from one dataset to another due to randomness in the noise. We assume that difference in performance from one dataset to another is caused by specific realizations of the noise, which may be friendly to one algorithm and unfriendly to another. Then a new metric based on measurement data, called mean squared discrepancy (MSD), which measures the difference between predicted crack sizes and measured data, is proposed to be a performance indicator in the absence of true crack size. Based on our numerical tests, it shows that the performance of one algorithm varies from one dataset to another. No method can perform consistently well and always be the best for handling all datasets. The ranking based on the mean squared error (MSE) can be mostly preserved when the ranking based on the mean squared discrepancy (MSD) is used. The former requires the true model while the latter does not. This indicates that MSD can be considered to rank the algorithms when the true crack size is not available.

2. STRATEGY FOR COMPARISON AND METRICS FOR PERFORMANCE

When multiple predictions are available from different algorithms, it is important for the users to evaluate their performance. In this paper, we assess the performances of algorithms based on multiple randomly simulated datasets. In this section, the strategy for implementing performance comparison taking into account the randomness in data is introduced firstly, followed by the metrics used for performance evaluation.

2.1. Strategy for implementing performance comparison

We use a moving time window as an experiment strategy, as shown in Fig. 1, to examine how well a given algorithm predicts future crack propagation with increasing number of measurement data. In Fig. 1, the solid blue curve represents the true crack size, (denoted by a in the following text and equations), red dash line the median crack size predicted by

the algorithm (denoted by \hat{a}) based on the first N_f data points, black solid dots the fitting data, and red asterisks the validation data (denoted by a_v). Three different time windows for three different datasets are shown.

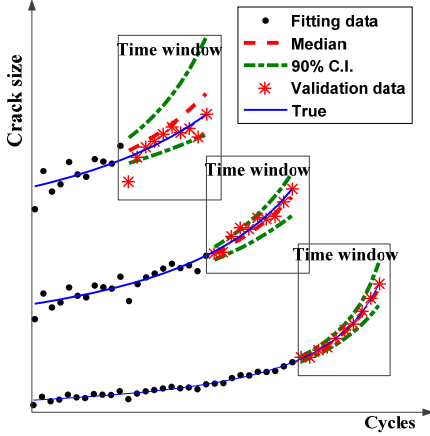


Figure 1 Schematic illustration of moving time window strategy.

2.2. Dealing with randomness in data

The strategy in Section 2.1 is developed for one algorithm using one dataset. We randomly generate N_d datasets with the same noise level to assess the algorithms' performance in different realizations of noise. These N_d datasets act as a database shared by all algorithms. For each dataset, the moving time window strategy is used, i.e., the previous N_m data are used to identify the parameters while the following N_v data are used for validation. There could be multiple time windows when using one dataset to test one algorithm.

2.3. Metrics for performance evaluation

Mean squared error (MSE)

The mean squared error (MSE) integrated over the time window is used to measure the accuracy of an algorithm. MSE is defined as

$$MSE = \frac{1}{N_v} \sum_{i=1}^{N_v} (a_{(N_f+i)\Delta T} - \hat{a}_{(N_f+i)\Delta T})^2$$

where a is the true crack size, \hat{a} the predicted crack size, and subscript the time step (refer to Fig.1 for illustration).

Mean squared discrepancy (MSD)

In practice, the MSE is unavailable since the true crack size cannot be available. A straightforward way is to compare the predictions with data. The difference between prediction and data is referred to as *discrepancy*. We consider the mean squared discrepancy (MSD) as a possible candidate for performance metric, as

$$MSD = \frac{1}{N_v} \sum_{i=1}^{N_v} (a_{v,(N_f+i)\Delta T} - \hat{a}_{(N_f+i)\Delta T})^2$$

where a_v is the validation data, and \hat{a} the predicted crack size by the algorithm. In the numerical case study, we will show the feasibility of using MSD as the performance indicator to rank the algorithms for an individual dataset.

3. NUMERICAL CASE STUDY

In this section, we investigate the four algorithms by assessing their prognostic performance to $N_d=100$ randomly generated measurement datasets. For each dataset, we test the prognostic behavior of each algorithm and rank them in terms of metrics.

3.1. Ranking of prognostics algorithms is sensitive to noise in data

We first show that even for our simple degradation model, when dealing with multiple measurement datasets from different realizations of random noise, the performance of an algorithm varies from one dataset to another, and none of the methods performs best for all datasets. The accuracy metric MSE is used to assess the performance of the algorithms.

Table 1 compares the average MSE over 100 datasets for the three time windows (i.e., $N_f=10, 20,$ and $30,$ respectively). Table 1 shows that in terms of average performance, BM and PF outperform the other two while NLS yields the largest MSE, especially in the earlier stage when very few data are available. For BM, PF and EKF, the MSE is not monotonically reduced as more measurements are used but tend to be large at the steep section of the crack growth curve. This indicates the prediction error increases when the crack grows fast. For NLS, the MSE in the $N_f=10$ and $N_f=20$ are large. The reason is that MSE of NLS method has some abnormal large values. These outliers of large MSE contribute to the average value, which makes the average MSE much greater than other three algorithms.

Table 1 Average MSE (in m^2) over 100 datasets with different numbers of measurement data

Methods	$N_m=20$ ($N_f=10$)	$N_m=30$ ($N_f=20$)	$N_m=40$ ($N_f=30$)
BM	1.20e-6	6.54e-7	3.02e-6
PF	1.22e-6	7.85e-7	4.52e-6
NLS	0.27	0.034	1.04e-4
EKF	2.91e-6	2.54e-6	5.48e-5

Next, we rank the four algorithms in terms of their MSE for each dataset. To present the results we use letters B, P, N, and E to index the methods BM, PF, NLS, and EKF, respectively. There are 24 possible permutations of rank, and the number of times each permutation appears is presented in Table 2 for $N_f=10, 20,$ and $30.$ For example, for $N_f=10,$ 11 permutations out of the 24 occur. Among these, the ranking PBEN, which indicates $PF > BM > EKF > NLS,$

occurs the most frequently. It is interesting to note that even NLS, the worst performer on average, outperforms the others for nine out of the 100 datasets.

The above discussion illustrates that even with a simple crack growth model, the performance rank of the methods varies from dataset to dataset, even if the difference among datasets comes only from random noise with the same noise level. In addition to the rank, random noise can also lead to a large different in algorithm performance, that is to say, the best algorithm on average can have a very poor accuracy for some datasets. Specifically, from Table 2, we see that PF is the best algorithm on average; i.e., PF is the best 127 times out of 300 (51, 34, and 42 times in $N_f=10, 20,$ and 30 cases, respectively). However, Table 2 also shows that for nine cases out of 300 datasets PF is the worst algorithm.

In summary, the performance of prognostics algorithms strongly depends on a specific dataset. Therefore, it may not have much sense to say one algorithm is better than the other. It would be better to choose the best algorithm based on a given specific dataset. We also see that the number of measurements can make a big difference in the ranking. For example, particle filter is the best 51% of the time for $N_f=10$, 34% of the time for $N_f=20$, and 42% for $N_f=30$.

Table 2 Statistics of MSE rank, MSE-Means Square Error.

$N_m=20$ ($N_f=10$)		$N_m=30$ ($N_f=20$)		$N_m=40$ ($N_f=30$)	
Cases	Times	Cases	Times	Cases	Times
EBPN	6	ENBP	1	EBPN	1
NEPB	1	EPBN	2	EBPN	3
NEBP	1	EBPN	4	NPBE	2
NPBE	5	ENBP	1	NBPE	5
NBPE	4	NEPB	1	PNEB	2
PNBE	2	NPBE	3	PNBE	3
PBEN	39	NBPE	5	PEBN	2
PBNE	10	NBEP	5	PBEN	9
BPNE	10	PNBE	1	PBNE	26
BPEN	20	PEBN	5	BNPE	8
BEPN	2	PBEN	21	BPNE	25
		PBNE	7	BPEN	14
		BNPE	4		
		BNPE	1		
		BPNE	9		
		BPEN	27		
		BEPN	3		

3.2. Discrepancy can rank the algorithms in absence of true damage information

In practice, the MSE is unavailable since the true model parameters are unknown and thus the true crack size. We attempt to use another metric to assess the performance of algorithms. A straightforward way of evaluating the performance is to compare the predictions with validation data, which is the MSD presented in Section 2.3. To verify our hypothesis, we consider the correlation between MSD and MSE. It was found that in general, MSD is highly correlated with MSE for all algorithms in all time-window

scenarios, meaning that MSD could be considered as a performance indicator to assess the algorithm performance.

When MSE is unknown, we naturally ask whether MSD can be used to rank the algorithms in terms of accuracy. This can be achieved by studying how consistent the rank based on MSE is with that based on MSD. Specifically, for each dataset, we rank the four algorithms based on their MSE. This is the *actual rank* in terms of accuracy. Then we re-rank the algorithms based on their MSD, which is referred to as *predicted rank*, and compare these two ranks to see to what extent they match.

The standard approach for comparing ranks would be to use the Spearman correlation coefficient. We opt instead for a weighted measure D_R of the discrepancy in ranking, which assigns a weight of 4 to a discrepancy in the first place and 1 to a discrepancy in the last place. The weight is introduced to take into account the importance of mistaking different positions. We assume that the importance of mismatching the i -th position diminishes with increasing i . Higher weight is assigned for smaller i -th position for accounting this importance. Therefore, the smaller D_R is, the better the two ranks match each other.

$$D_R = \sum_{i=1}^4 (5-i)d_i^2$$

The statistics of the D_R for the 100 datasets are given in Table 3 where $D_R=3$ corresponds to switching 3rd and 4th positions, $D_R = 5$ to switching 2nd and 3rd, and $D_R = 7$ to switching 1st and 2nd. It is seen from the table that out of all 300 datasets, 232 have a perfect agreement in rank and 49 others have a single adjacent permutation so that only 19 have a substantial difference in rank.

Table 3 Statistics of match extent of MSE rank and MSD rank

$N_m=20$ ($N_f=10$)		$N_m=30$ ($N_f=20$)		$N_m=40$ ($N_f=30$)	
D_R	times	D_R	times	D	times
				R	
0	81	0	65	0	86
3	1	3	4	3	1
5	2	5	6	7	11
7	11	7	13	9	1
15	1	10	1	15	1
21	4	15	2		
		21	6		
		25	3		

We further probe the correlation between MSD and MSE with different magnitude of the noise level. We test other two noise levels, i.e., $\sigma=0.3\text{mm}$ and $\sigma=2.85\text{mm}$, which are equivalent to 3.75% and 35% coefficient of variation (COV) with respect to the initial crack size (8mm). In fact, given initial crack size 8mm, 35% is almost the largest COV we could try, because for larger COVs negative values of crack length appear. It is found that the correlation between MSE and MSD in both two noise levels are high, except the case

of EKF with 35% COV noise level in the case of $N_f=10$, which is relatively low (0.64). In addition, in both two noise levels, the match extent between MSE rank and MSD rank are satisfactory. The results indicate that the MSD rank is tolerant to a relatively large measurement noise.

4. CONCLUSION

In this paper, the four most commonly used algorithms, Bayesian method, particle filter, nonlinear least squares, and Extended Kalman filter, are applied on a simple crack growth model with simulated random measurement noise. We investigate their performance statistically by testing their performance using 100 randomly generated measurement datasets with the same noise level. The mean squared error (MSE) is used as a metric to rank the four algorithms in terms of accuracy for each dataset. It is found that the performance of prognostics algorithms strongly depends on the realization of random noise, and none of the algorithms can be the best one in all realizations. It was found that on average the two algorithms based on Bayesian inference substantially outperformed the other two. However, the statistics of MSE rank showed that the performance of the algorithms varies substantially from one dataset to another even if the data are generated with the same noise level. As a result, for some data sets the worst on-average algorithm can substantially outperform the best one. Since the exact solution is not available in practice, the discrepancy between predictions and measurements (MSD) has to stand for the actual error (MSE). We found a very good correlation between MSE and MSD and that ranking the algorithms based on discrepancy with measurements is a good stand in for their true rank in terms of accuracy.

REFERENCES

- An D, Kim N-H, Choi J-H. Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. *Reliability Engineering & System Safety* 2015;133:223-36.
- Saxena A, Jose C, Bhaskar S, Sankalita S, Kai G. Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management* 2010;1: 4-23.

BIOGRAPHIES

Yiwei WANG received her B.Sc. degree (2010) in Mechanical Engineering from Beijing Jiaotong University and M.Sc. (2013) degree in Mechanical Engineering from Beihang University, Beijing, China. She is currently a Ph.D. candidate in Institut National des Sciences Appliquées (INSA), Toulouse, France and at the same time studying in Lab Institut Clément Ader (CNRS), France. Her research interests include uncertainty modeling, prognostic methods, reliability and probabilistic approaches, Bayesian methods.

Nam-Ho Kim is presently Professor of Mechanical and Aerospace Engineering at the University of Florida. He graduated with a Ph.D. in the Department of Mechanical Engineering from the University of Iowa in 1999 and worked at the Center for Computer-Aided Design as a postdoctoral associate until 2001. His research area is structural design optimization, design sensitivity analysis, design under uncertainty, structural health monitoring, nonlinear structural mechanics, and structural-acoustics. He has published five books and more than hundred fifty refereed journal and conference papers in the above areas.

Raphael T. Haftka is a Distinguished Professor of Mechanical and Aerospace Engineering at the University of Florida. Before coming to the University of Florida in 1995, he was the Chris Kraft Professor of Aerospace and Ocean Engineering at Virginia Tech. His areas of research include structural and multidisciplinary optimization, design under uncertainty, and surrogate based global optimization. In the last decade, his research has focused on the contribution of structural tests and structural health monitoring to reducing uncertainty and improve safety. He is a fellow of the AIAA, and a recipient of the AIAA MDO award and the AIAA/ASC James H. Starnes award. He was president of the International Society of Structural and Multidisciplinary Optimization 1995-1999, and chaired their world congress in 2013. He has directed more than 50 PhD students, and has more than 24,000 Google Scholar citations.