

Article

# Online Subspace Tracking of Sensors Data for Damage Propagation Modeling and Predictive Analytics

Farhan Khan 

National University of Sciences and Technology, Islamabad; farhan.khan@seecs.edu.pk

**Abstract:** We analyze damage propagation modeling of turbo-engines in a data-driven approach. We investigate subspace tracking assuming a low dimensional manifold structure and a static behavior during the healthy state of the machines. Our damage propagation model is based on the deviation of the data from the static behavior and uses the notion of health index as a measure of the condition. Hence, we incorporate condition-based maintenance and estimate the remaining useful life based on the current and previous health indexes. This paper proposes an algorithm that adapts well to the dynamics of the data and underlying system, and reduces the computational complexity by utilizing the low dimensional manifold structure of the data. A significant performance improvement is demonstrated over existing methods by using the proposed algorithm on CMAPSS Turbo-engine datasets.

**Keywords:** online learning; machine prognostics; sensor systems; signal processing; damage propagation; predictive maintenance; intelligent sensing

## 0. Introduction

Machine prognostics and predictive analytics are widely investigated in control theory, industry applications and signal processing [1–4]. Condition based maintenance uses prognostics methods in a variety of applications such as manufacturing, automotive, heavy industry, consumer electronics and biomedical equipments. The correct estimate of future condition is helpful and add to the timely maintenance/replacement of the faulty component(s). In such manner predictive analytics helps save time, effort and cost and assure the smooth running of the required processes. In literature, various methods have been studied using physical modeling of the degradation. However, physical modeling in most modern applications are inadequate and at times extremely complex and stochastic. Therefore, more recently data-driven methods are investigated that use instantaneous sensors' as well as operational data collected from the machines [5]. To this end, we propose a novel data-driven algorithm that analyze time-series data for the degradation modeling and remaining useful life (RUL) estimation.

Various data-driven methods have been investigated in literature for damage propagation modeling and predictive analysis specifically in machine learning, signal processing, time-series analysis and deep learning [6,7]. Conventionality, the sensors' data from machines through their life cycle (till failure) is used to train the damage propagation model and estimate remaining useful life of new instances. In such scenario, the RUL is represented by a linear decaying function, i.e.,  $RUL_t = T - t$ , where  $T$  is time at failure. However it's more practical to assume a piece-wise linear function since, in the beginning when machine is operating in perfectly healthy condition, the health of machine cannot be taken as degrading. For that purpose, the RUL is approximated as [8]:

$$\begin{aligned} RUL_t &= T && \text{for } 0 \leq t \leq t_d \\ &= T - t && \text{for } t_d < t \leq T, \end{aligned} \quad (1)$$

where  $t_d$  is the point where the linear degradation starts. The piece-wise RUL is then used as target in supervised learning during the training phase of the algorithm. Several algorithms such as support vector regression (SVR), Convolutional neural networks (CNN) and more advanced deep learning involving Long Short-Term Memory (LSTM) and CNN have been used for degradation modeling and RUL estimation [8]. However, here we investigate a semi-supervised approach where no assumption is made with regards to RUL and the degradation modeling is learned from the input data entirely as in [2,4,9,10]. To this end, we assume that the time till the machine runs in healthy state is known that can be incorporated in a novelty detection model [9]. In this manner, the deviation in statistical distribution of the data is used as a measure of the degradation.

For damage propagation and performance degradation modeling, various methods are investigated in literature that mainly involve auto-encoders for input sequence reconstruction [9]. These methods rely on the reconstruction error as a measure of the damage (and hence) health of the machine at any given time. For instance, in [9], the authors use LSTM based encoder-decoder model that regenerates the input sequence, after training on the healthy samples, and use the error between estimated and true input sequence as measure of the degradation and health index. Furthermore, they employ linear regression along-with the encoder-decoder model for more robust modeling. However, we propose a subspace tracking approach to measure the variation in the distribution of input sequence by incorporating instantaneous manifold tracking [11,12]. In the proposed approach, instead of regenerating the whole input sequence, we estimate a low dimensional representation of the input and the subspace that it lies in, hence reducing the computational cost and overfitting. Furthermore, since the input and underlying submanifold subspace have different dimension, we incorporate approximate Mahalanobis distance for updating the model parameter during training and later as a measure of degradation [12–15]. We emphasize that our proposed algorithm suits well to the dynamics of the input data, reduces the computational complexity and achieves significantly higher accuracy than the state-of-the-art. We demonstrate the performance of the proposed algorithm by applying it to the well known CMAPSS datasets [5,16].

In summary, we investigate online subspace tracking for the damage propagation assuming low dimensional manifolds. We further incorporate linear regression by using the low dimensional projection of the input data as new input and estimate the health index. We then estimate the RUL of a test case by comparing the health index curve to all available degradation models. We use an ensemble learning approach as in [9] to finally estimate the RUL. The organization of the paper is as follows: in Section 1, we formally describe problem setting and define various parameters in detail. In Section 2, we demonstrate the subspace tracking algorithm for damage propagation modeling and health index curves generation. In Section 3, we apply the proposed damage propagation modeling to real life datasets and demonstrate the performance evaluation in terms of root mean square error (RMSE) and the scores defined in [5]. We finally conclude the paper in Section 4.

## 1. Problem Description

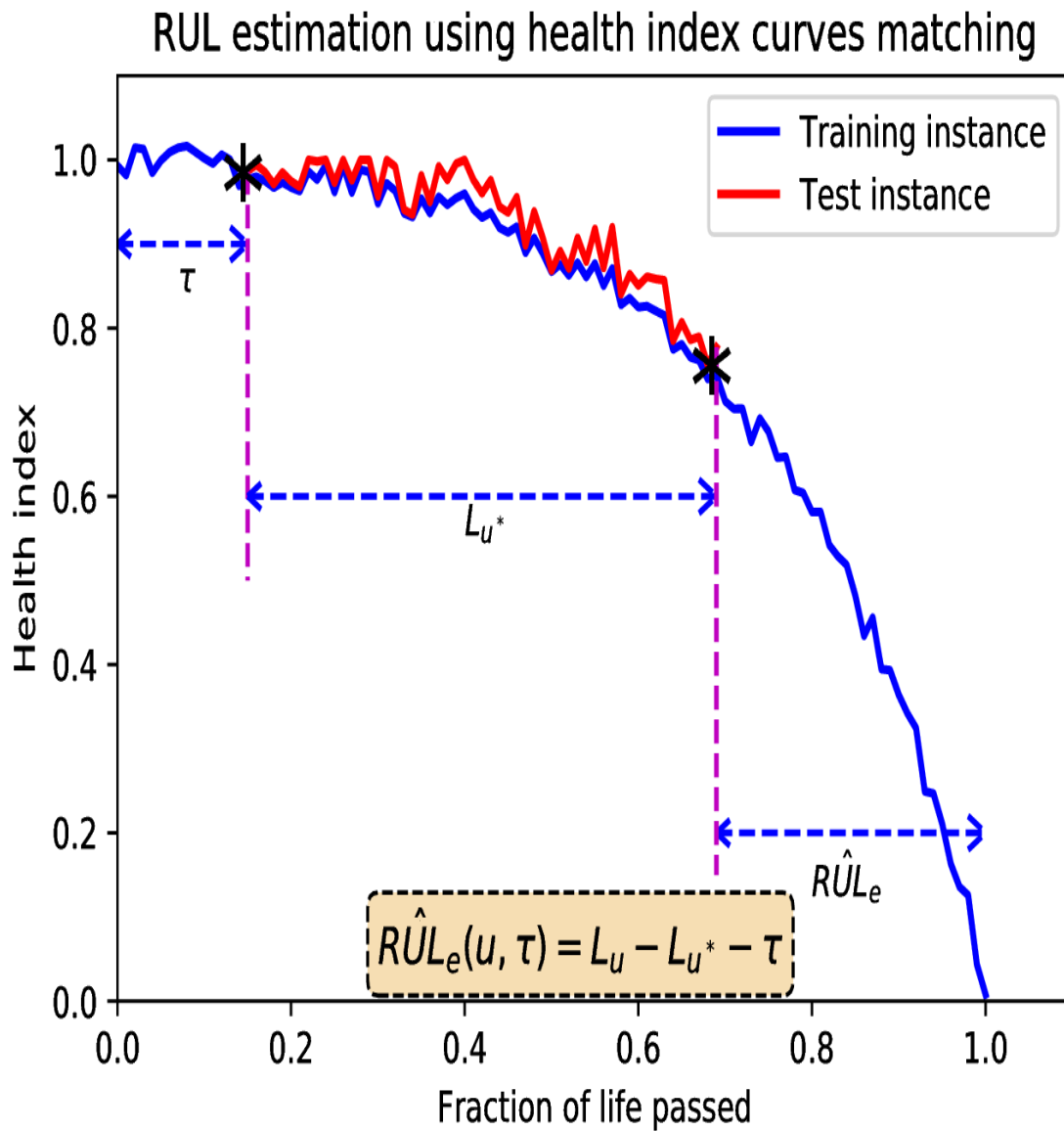
All vectors used in this paper are column vectors denoted by boldface lowercase letters. Matrices are denoted by boldface uppercase letters. For a vector  $\mathbf{x}$  (or a matrix  $\mathbf{U}$ ),  $\mathbf{x}^T$  ( $\mathbf{U}^T$ ) is the ordinary transpose.  $T$  is the total number of time-steps and an arbitrary time-step is denoted by  $t$  where  $0 \leq t < T - 1$ . The time index of a sequence vector is denoted by  $t$  in the subscript, as in  $\mathbf{x}_t$ .

We investigate the estimation of current health index (HI)  $\sigma_t$  and the prediction of remaining useful life  $RUL_t$  by analyzing the input data  $\mathbf{x}_t \in \mathbb{R}^D$  where  $D$  is the number of input features. The health index is modeled as a function of the input data as:

$$\sigma_t = f(\mathbf{x}_t), \quad (2)$$

where  $0 \leq \sigma \leq 1$ , with 1 and 0 corresponding to perfect health and failure respectively. Furthermore, we assume that the  $D$ -dimensional input data lies on a static or time varying submanifold,  $S_{k,t}$  with

reduced intrinsic dimension  $d$  such as  $d \ll D$ , where, for the multiscale modeling of a non-stationary submanifold,  $k$  is the index such as  $k \in \{1, 2, \dots, K\}$  [13,14].



**Figure 1.** Health index curve and RUL estimation using similarity

### 1.1. Health index and Remaining Useful Life Estimation

Given HI curves for the all training instance, we use curve matching for the estimation of RUL [2,4,10]. That is, as shown in Fig. 1, the HI trajectory for a test instance  $u^*$  is slided over training HI curve sets  $U$  for all times  $\tau_1 \leq \tau \leq \tau_2$  in order to minimize the euclidean distance. The similarity between test HI curve  $u^*$  and training HI curve  $u$  is defined by [2,4,10]:

$$s(u^*, u, \tau) = \exp(-d^2(u^*, u, \tau) / \beta), \quad (3)$$

where  $\tau \in \{\tau_1, \dots, \tau_2\}$  is the time-lag,  $\beta > 0$  is a small constant that controls the similarity and  $d^2(u^*, u, \tau)$  is the squared euclidean distance defined by,

$$d^2(u^*, u, \tau) = \frac{1}{L_{u^*}} \sum_{i=1}^{L_{u^*}} (\sigma_i^{u^*} - \sigma_{i+\tau}^u)^2. \quad (4)$$

For each training instance  $u$  and time-lag  $\tau$ , the RUL of test instance  $u^*$  is estimated as:  $RUL_{L_{u^*}}(u, \tau) = L_u - L_{u^*} - \tau$ . The final estimate of RUL is a linear combination of  $RUL_{L_{u^*}}(u, \tau)$  all  $u, \tau$  using the similarity measure  $s(u^*, u, \tau)$  as coefficients, i.e.,

$$R\hat{U}_{L_{u^*}} = \frac{\sum s(u^*, u, \tau) \sum RUL_{L_{u^*}}(u, \tau)}{\sum s(u^*, u, \tau)}. \quad (5)$$

The error  $e_{u^*} = R\hat{U}_{L_{u^*}} - RUL_{L_{u^*}}$  between estimated and true RUL on test sets is used as a measure of performance. We specifically use two performance metrics, i.e., the RMSE defined as,

$$RMSE = \sqrt{\sum_{u^*=1}^N (R\hat{U}_{L_{u^*}} - RUL_{L_{u^*}})^2}, \quad (6)$$

and score  $S$  defined as,

$$S = \sum_{u^*=1}^N (\exp(\frac{\gamma}{e_{u^*}}) - 1), \quad (7)$$

where  $\gamma = 1/13$  when  $e_{u^*} < 0$  and  $\gamma = 1/10$  when  $e_{u^*} \geq 0$ . This way a late detection of the failure (smaller estimated RUL) is penalized more as compared to early detection [5].

## 2. Multi-scale Subspace Tracking for Predictive Analytics

In the most basic form, we propose to use a single subspace tracking algorithm for the damage propagation modeling. In this sense, the input data  $\mathbf{x}_t \in \mathbb{R}^D$  in healthy state is assumed to be lying on a static submanifold with intrinsic dimension  $d$ . We project the input data on the subspace  $S_t$  of the submanifold and then determine the approximate Mahalanobis distance  $d(x, S)$  as in [14]. The square root of distance is used as error to update the parameters of estimated subspace  $\{\mathbf{U}, \mathbf{c}, \mathbf{\Lambda}\}$ . Here matrix  $\mathbf{U}$  is the eigen-vector matrix of the covariance matrix representing the orientation of the subspace, vector  $\mathbf{c}$  is the mean of input space and  $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_d\}$  represents the spread.

The subspace tracking is run on the healthy data in epochs till the error is minimized and converged while using the instantaneous error to update the parameters in a stochastic gradient descent manner. Once the subspace parameters are finalized, the trained model is used on the remaining cycles of data till failure and the approximate Mahalanobis distance is recorded for each time instant as [12,13]:

$$d_t(\mathbf{x}, S) \triangleq \delta(\mathbf{x} - \mathbf{c})^T \mathbf{U}_1 \mathbf{\Lambda}_1^{-1} \mathbf{U}_1^T (\mathbf{x} - \mathbf{c}) + \|\mathbf{U}_2^T (\mathbf{x} - \mathbf{c})\|^2, \quad (8)$$

where  $\delta > 0$  is a small constant that depends on the distribution of data beyond the submanifold. We then use this distance to generate health index  $\sigma_t$  as:

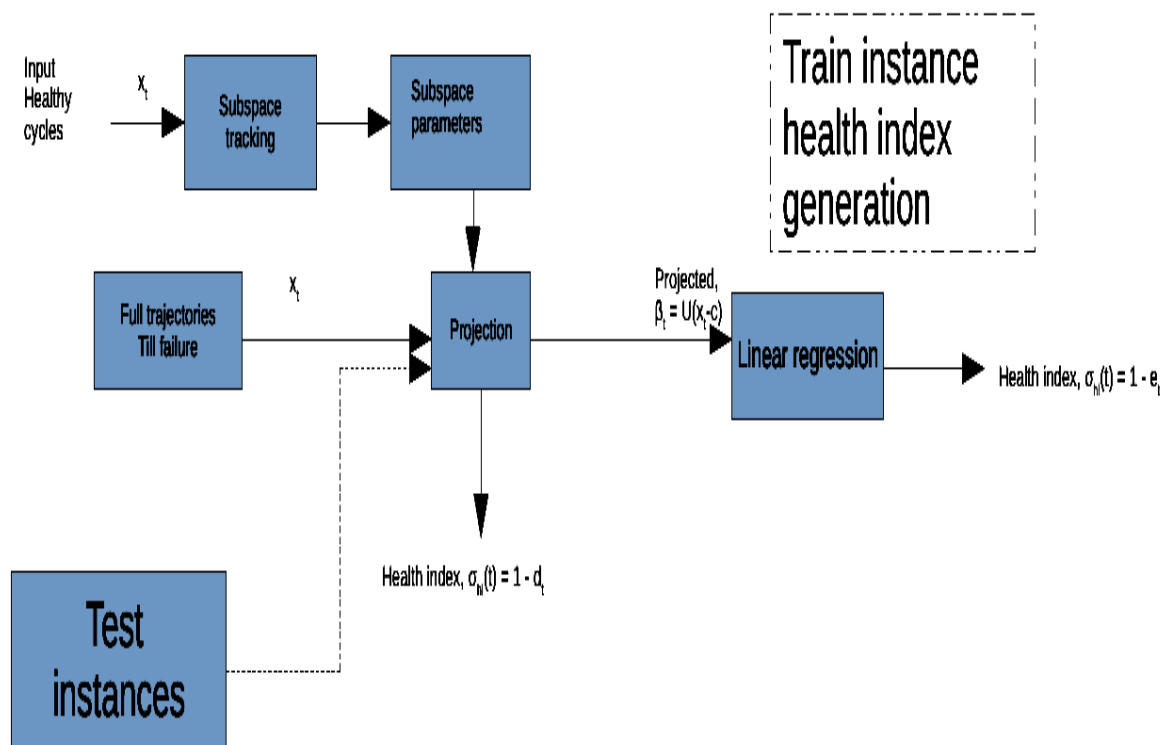
$$\sigma_t = 1 - \sqrt{\hat{d}_t(\mathbf{x}, S)}, \quad (9)$$

where  $0 \leq \hat{d}_t(x, S) \leq 1$  is the scaled version of  $d(x, S)$ . We get an exponentially decaying curve of the health index values that reaches 0 at failure as shown in Fig. 1.

To this end, we use subspace tracking of the  $d$ -dimensional submanifold and use the tracking error as measure of degradation for all training instances. However, after using a certain amount of data for generating HI curves, we next utilize linear regression for the remaining instances and cycles

to estimate the health index. In other words, we use the known HI values (estimated through subspace tracking) as target and learn the linear regression model by least squares method, then estimate the HI values for the remaining data. Here, instead of using the original  $D$ -dimensional input, we use the  $d$ -dimensional projections of the data on the submanifold subspace as the new input. In this manner, we further reduce the computational cost of the overall algorithm.

For non-stationary setting, i.e., when the input data does not follow a static distribution in the healthy state, we assume the data lies on a time varying submanifold and use a multi-model learning of the underlying subspace. We use the notion of multi-scale tracking and MOUSSE algorithm as in [12–14]. In multi-scale subspace tracking, the input space is partitioned into  $K$  regions where there is a different subspace representing each individual submanifold. The input data  $x_t$  at time  $t$  is projected on each subspace and the one with minimum distance is used and updated for the next cycle. Also, for the health index calculation, the minimum distance is used as a measure of degradation.



**Figure 2.** The proposed learning model for health index generation

### 2.1. Algorithm Description

Here, we briefly describe the algorithm shown in Fig. 2 step by step. Initially, we use data from the healthy state and train the subspace tracking algorithm till convergence, i.e., when the health index,  $\sigma_t = 1$ . Then during the second stage (inference), we use the estimated parameters of the underlying subspace to estimate health index for the remaining cycles till the failure point. In this manner, we get an exponentially decaying curve. Furthermore, we use the trained model to estimate  $\sigma_t$  for the test

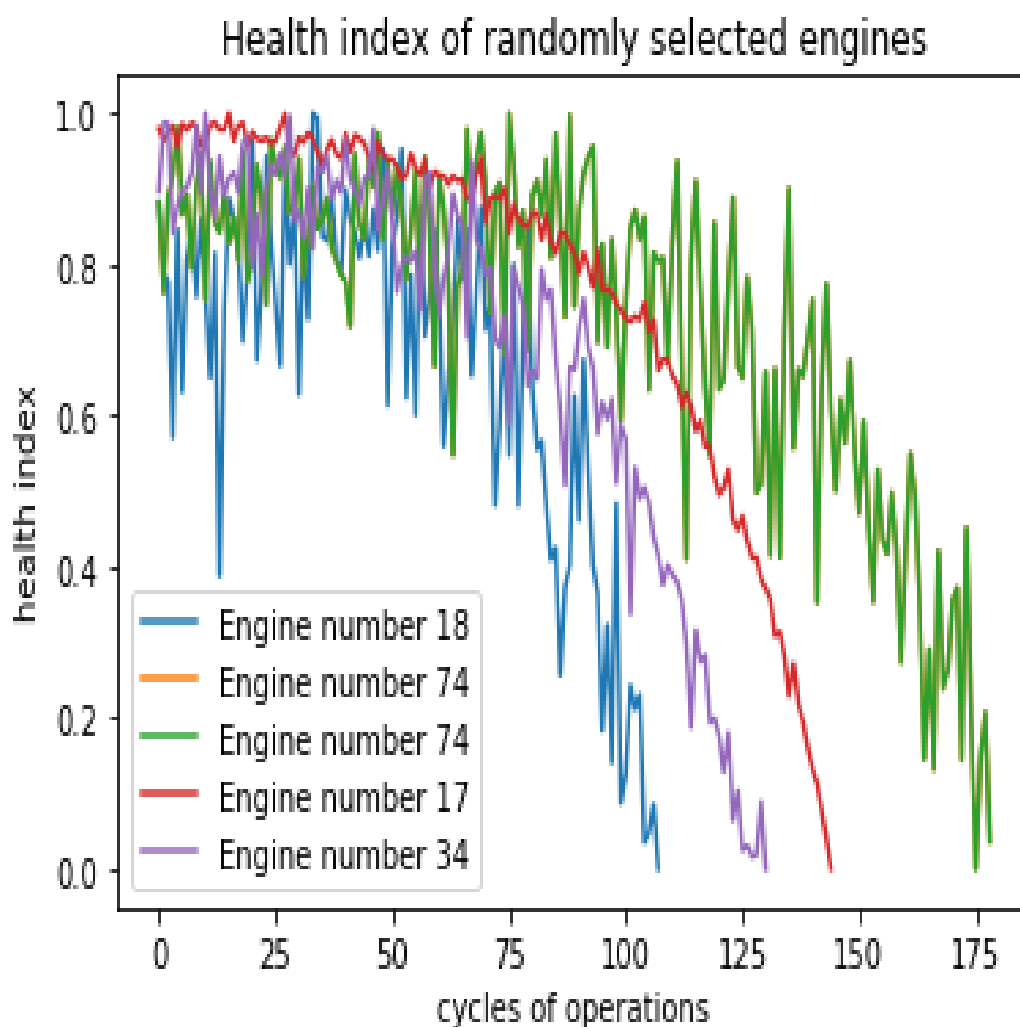


Figure 3. Health index curves of FD001 training dataset

instances that are truncated before the failure. Finally, we use similarity measures for curve matching for the estimation of RUL as described in subsection 1.1.

### 3. Results and Performance Analysis

To validate the proposed algorithm, we use the widely investigated CMAPSS turboengine datasets as benchmark [5,16]. In principal, the CMAPSS datasets consist of four independent datasets for various number of engines and time instances, where each one is a pair of training (for a complete cycle till failure) and test (where the data is truncated at a point before the failure point). The target is to estimate the remaining useful life of the test instances based on the behavior of the training data. The four datasets (in pair) are named as FD001, FD002, FD003 and FD004 in literature and each input consists of 24 features that include 3 operational setting features and 21 sensor values. In all of the experiments, we assume the first 20 cycles of each engine as healthy and employ subspace tracking. We use grid search cross-validation to choose the hyper-parameters as  $\alpha = 0.87$ ,  $\tau_1 = 1$ ,  $\tau_2 = 40$ , and  $\beta = 0.0235$ . The intrinsic dimension of submanifold is set as  $d = 3$ . We use three variations of the algorithm, i.e., Single subspace tracking (SST) for all HI curves generation, Single subspace tracking with linear regression (SST-LR) and multiple subspace tracking where the input data is non-stationary during the healthy state.

## RUL estimation --health index curves generated by subspace tracking

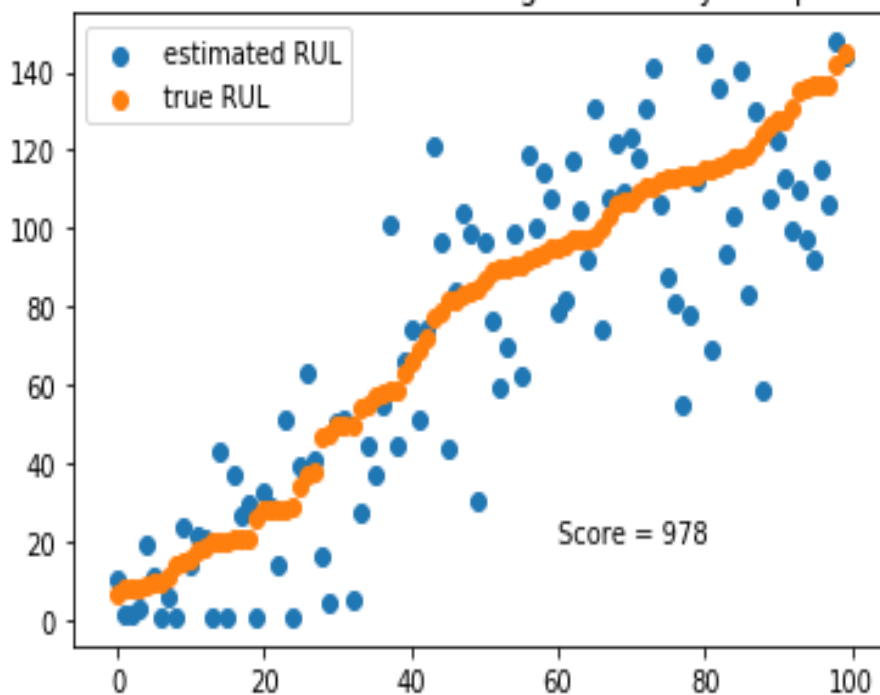


Figure 4. Subspace tracking (SST) without linear regression

## RUL estimation --health index curves generated by linear regression

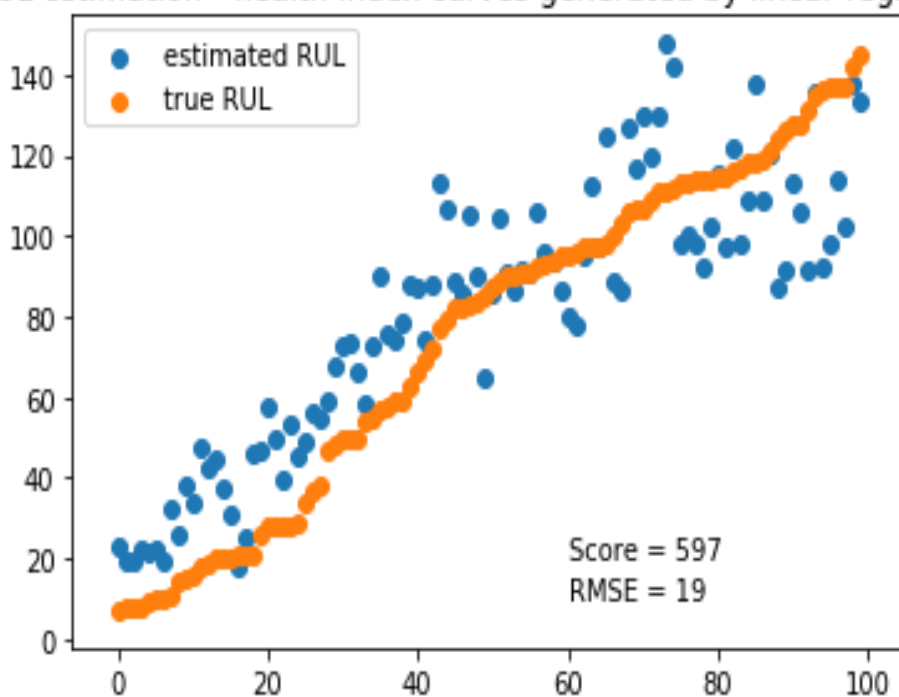


Figure 5. SST with linear regression (SST-LR)

As second stage of the algorithm (inference), we plot the health index curves for all training and test instances. Fig. 3 shows health index curves for randomly selected five engines where  $\sigma_t$  reaches zero at the end-of-life. Similar to [7], we match the HI curve for a certain test instance with all training instance curves and use (5) to estimate RUL. In Fig. 4 and Fig. 5, the estimated RUL for each test

**Table 1.** Turbofan Engine: Performance comparison w.r.t scores

| <i>Algorithm</i> | <i>FD001</i> | <i>FD002</i>       | <i>FD003</i> | <i>FD004</i>       |
|------------------|--------------|--------------------|--------------|--------------------|
| <i>SVR</i>       | 1380         | $5.90 \times 10^5$ | 1603         | $3.71 \times 10^5$ |
| <i>CNN</i>       | 1290         | $1.36 \times 10^4$ | 1602         | 7892               |
| <i>DeepLSTM</i>  | 338          | 4452               | 852          | 5554               |
| <i>LSTM – ED</i> | 1260         | —                  | —            | —                  |
| <b>SST</b>       | 978          | <b>4230</b>        | <b>822</b>   | <b>4401</b>        |
| <b>SST-LR</b>    | 597          | <b>3351</b>        | <b>634</b>   | <b>3381</b>        |

**Table 2.** Turbofan Engine: Performance comparison w.r.t RMSE values

| <i>Algorithm</i> | <i>FD001</i> | <i>FD002</i> | <i>FD003</i> | <i>FD004</i> |
|------------------|--------------|--------------|--------------|--------------|
| <i>SVR</i>       | 20.96        | 42.00        | 21.05        | 45.35        |
| <i>CNN</i>       | 18.45        | 30.29        | 19.82        | 29.16        |
| <i>DeepLSTM</i>  | 16.14        | 24.49        | 16.18        | 28.17        |
| <i>LSTM – ED</i> | 23.36        | —            | —            | —            |
| <b>SST</b>       | <b>16.22</b> | 30.21        | 17.02        | 28.21        |
| <b>SST-LR</b>    | <b>15.02</b> | 29.12        | 16.95        | <b>26.03</b> |

instance of dataset FD001 are plotted in the ascending order, using HI curves generated by SST and SST-LR respectively. The results show a good match between the true and estimated RUL, specifically with SST-LR.

We apply the proposed algorithms on the remaining three datasets while using the multiple subspace tracking as described in subsection 2. Specifically, by clustering the operational setting data, we observe that there are six different scenarios in case of datasets FD002 and FD004. This makes these datasets ideal candidate for multiple subspace tracking as in each operational case, the data lies on a one of the six submanifolds. Similarly, while using linear regression for HI values, we train a different model for each case. We compare the performance of proposed algorithms with that of support vector regression (SVR), Convolutional neural networks (CNN), LSTM based deep learning (Deep LSTM) [6,8] and LSTM based encoder-decoder model (LSTM-ED) [9] as shown in Table I and Table II. The use of multi-model analysis makes the algorithm more robust and achieves significant performance improvement over SVR and CNN, and competes well against Deep LSTM while using a reduced computational complexity.

#### 4. Conclusion

We investigate online manifold learning and subspace tracking for the damage propagation modeling. We propose a novel algorithm that generates health index values for each input based on the distribution of the data. We then use the HI curves for the estimation of RUL. We specifically investigate a damage propagation model of Turbo-engine, however, the proposed algorithm can be extensively applied to other applications for predictive analysis. The proposed algorithm is computationally efficient and adapts to the dynamics of the data both in static and non-stationary scenarios. We implement the proposed algorithm for the RUL estimation of CMAPSS Turbo-engines and achieves a significant performance improvement over the state-of-the-art in terms of RMSE and timely detection of the damage. **Conflicts of Interest:** The authors declare no conflict of interest.



## References

1. Abbas, M.; Vachtsevanos, G.J. A hierarchical framework for fault propagation analysis in complex systems. AUTOTESTCON, 2009 IEEE. IEEE, 2009, pp. 353–358.
2. Wang, T.; Yu, J.; Siegel, D.; Lee, J. A similarity-based prognostics approach for remaining useful life estimation of engineered systems. Prognostics and Health Management, 2008. PHM. International Conference on. IEEE, 2008, pp. 1–6.
3. Javed, K.; Gouriveau, R.; Zerhouni, N. A New Multivariate Approach for Prognostics Based on Extreme Learning Machine and Fuzzy Clustering. *IEEE Transactions on Cybernetics* **2015**, *45*, 2626–2639.
4. Eker, Ö.F.; Camci, F.; Jennions, I.K. A similarity-based prognostics approach for remaining useful life prediction. Prognostics and Health Management Society. 2nd European Conference of the. PHM Society, 2014, pp. 1–5.
5. Ramasso, E.; Saxena, A. Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets. *International Journal of Prognostics and Health Management* **2014**, *5*, 1–15.
6. Babu, G.S.; Zhao, P.; Li, X.L. Deep convolutional neural network based regression approach for estimation of remaining useful life. International conference on database systems for advanced applications. Springer, 2016, pp. 214–228.
7. Heimes, F.O. Recurrent neural networks for remaining useful life estimation. Prognostics and Health Management, 2008. PHM 2008. International Conference on. IEEE, 2008, pp. 1–6.
8. Zheng, S.; Ristovski, K.; Farahat, A.; Gupta, C. Long short-term memory network for remaining useful life estimation. Prognostics and Health Management (ICPHM), 2017 IEEE International Conference on. IEEE, 2017, pp. 88–95.
9. Malhotra, P.; TV, V.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; Shroff, G. Multi-sensor prognostics using an unsupervised health index based on lstm encoder-decoder. *arXiv preprint arXiv:1608.06154* **2016**.
10. Wang, T. Trajectory similarity based prediction for remaining useful life estimation. PhD thesis, University of Cincinnati, 2010.
11. Lin, T.; Zha, H. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2008**, *30*, 796–809.
12. Khan, F.; Kari, D.; Karatepe, I.A.; Kozat, S.S. Universal nonlinear regression on high dimensional data using adaptive hierarchical trees. *IEEE Transactions on Big Data* **2016**, *2*, 175–188.
13. Xie, Y.; Willett, R. Online logistic regression on manifolds. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3367–3371.
14. Xie, Y.; Huang, J.; Willett, R. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing* **2013**, *7*, 12–27.
15. Zhang, Z.; Wang, J.; Zha, H. Adaptive Manifold Learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2012**, *34*, 253–265.
16. Liu, Y.; Frederick, D.K.; DeCastro, J.A.; Litt, J.S.; Chan, W.W. User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS): Version 2. *NASA/TM-pp.2012-217432[R]* **2012**.