# Uncertainty-aware Remaining Useful Life predictors

**Luca Biggio**[1,2], **Manuel Arias Chao**[1], **Olga Fink**[1]
[1]ETH Zürich, Switzerland
[2]CSEM SA, Switzerland
lbiggio@inf.ethz.ch, manuel.arias@ethz.ch, ofink@ethz.ch

## Abstract

Remaining Useful Life (RUL) estimation is the problem of inferring how long a certain industrial asset is going to operate until a system failure occurs. Deploying successful RUL methods in real-life applications would result in a drastic change of perspective in the context of maintenance of industrial assets. In particular, the design of intelligent maintenance strategies capable of automatically establishing when interventions have to be performed has the potential of drastically reducing costs and machine downtimes. In light of their superior performances in a wide range of engineering fields, Machine Learning (ML) algorithms are natural candidates to tackle the challenges involved in the design of intelligent maintenance approaches. In particular, given the potentially catastrophic consequences associated with wrong maintenance decisions, it is desirable that ML algorithms provide uncertainty estimates alongside their predictions. In this work, we propose and compare a number of techniques based on Gaussian Processes (GPs) that can cope with this aspect. We apply these algorithms to the new C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset from NASA for aircraft engines. The results show that the proposed methods are able to provide very accurate RUL predictions along with sensible uncertainty estimates, resulting in more safely deployable solutions to real-life industrial applications.

## 1 Introduction

The current most widely employed maintenance strategy is based on scheduling interventions at fixed time intervals, regardless of the current state of the machine under consideration. The times at which maintenance operations are performed are set by human experts and often result in expensive and unnecessary machine downtimes. Predictive Maintenance (PM), on the other hand, represents a completely different paradigm since it aims at setting maintenance operations based on the information extracted from data describing the health state of the machine. Efficient RUL estimation is a key enabler of PM and the application of ML techniques to address this task is an active research area. However, very limited efforts have been made to equip data-driven techniques with tools for quantifying the level of uncertainty associated with their predictions. Uncertainty Quantification (UQ) is crucial in the context of PM due to the high risks deriving from insufficient or not timely maintenance interventions. The deployment of ML techniques in real-world engineering scenarios requires the design of reliable and transparent algorithms capable of estimating the level of confidence of their outputs.
The contribution of this paper is twofold: 1) We show that modern GP models can be successfully employed to predict RUL of complex industrial systems, thanks to a series of recent developments aimed at making standard GPs more scalable and expressive; 2) Our results highlight that, differently from standard Deep Learning (DL) approaches, the proposed techniques are able to take uncertainty into account without sacrificing RUL prediction accuracy.

Workshop on machine learning for engineering modeling, simulation and design @ NeurIPS 2020

## 2 Methods

### 2.1 Background

Standard DL techniques applied to RUL estimation range from relatively complex fully-connected networks to CNN or RNN based models. DL techniques have indeed the potential to bypass (or at least to limit) the process of manual feature extraction that is often performed by engineers. This aspect makes them particularly appealing in the context of PM, where the goal is to reduce the impact of human inductive biases and increase the level of automation of maintenance strategies. However, most of the DL approaches proposed in the literature do not take into account the level of uncertainty associated with their predictions, thus posing a strong limit on their deployment.

One of the most popular classes of ML models capable of addressing the UQ problem is that of Gaussian Processes (GPs) [Rasmussen and Williams, 2006]. Despite their desirable properties in terms of UQ and their elegant theoretical formulation, GPs are affected by two main limitations hindering their application to real world datasets. First, the analytical calculation of the marginal likelihood is computationally prohibitive when the number of data, $N$, is large. Second, their hypothesis space is completely determined by the choice of the kernel function, which might not be complex enough to model certain types of data. However, over the last 20 years, the ML community has found a number of workarounds to address the aforementioned limitations and to refine standard GP models. In this work, in particular, we investigate the performance of three of such methods, namely Stochastic Variational Gaussian Processes (SVGPs) [Hensman et al., 2015], Deep Gaussian Processes (DGPs) [Damianou and Lawrence, 2013, Salimbeni and Deisenroth, 2017] and Deep Sigma Point Processes (DSPPs) [Jankowiak et al., 2019, 2020]. In the following, we briefly review the basis features of each of these methods. A more comprehensive analysis of the above techniques can be found, for instance, in Jankowiak et al. [2020].

### 2.2 Applied Models

**SVGP.** SVGP is a popular inducing point method [Snelson and Ghahramani, 2006, Titsias, 2009] based on variational inference [Blei et al., 2017] that makes the application of the GP framework to big datasets possible. On one hand, the introduction of inducing points alleviates the computational cost associated with the operations involving the $N \times N$ kernel matrix. On the other hand, SVGP is based on the optimization of the ELBO (evidence lower bound), which can be compactly written as a sum over data points. This important aspect allows the application of stochastic gradient methods and data sub-sampling, resulting in a more flexible and scalable optimization process.

**Deep Gaussain Processes.** The classes of functions modelled by standard GP models, including SVGP, are limited by the expressiveness of the chosen kernel. One way to tackle this shortcoming is to use a deep neural network to automatically learn the kernel from data Wilson et al. [2016]. However, these approaches often require problem-specific architectures and are prone to overfitting. DGPs consist of hierarchical compositions of GPs and offer a powerful alternative solution to increase the representational power of "single-layer"-GPs. They retain much of the advantages of shallow GPs and introduce a relatively small number of parameters to optimize compared to standard neural network models. In this work, we employ a variant of DGPs recently proposed by Salimbeni and Deisenroth [2017] in order to address some drawbacks of the original DGP formulation [Damianou and Lawrence, 2013]. This improved model enjoys the same advantages of SVGPs, i.e. it reduces computational complexity by introducing inducing variables for each GP in the hierarchy and supports mini-batch training[1].

**Deep Sigma Point Processes.** Despite their many practical successes, variational inference methods often tend to provide overly confident uncertainty estimates [Turner and Sahani, 2011, Bauer et al., 2017]. A recent series of works [Jankowiak et al., 2019, 2020] try to address this limitation by reformulating the variational inference scheme at the basis of SVGP and DGPs. In particular, the authors note an inconsistency between the ELBO (the objective function to be optimized) and the predictive distribution to be used a test time. More specifically, both quantities are written as functions of two variance terms, one input-dependent, $\sigma_f(x)^2$, and one input-independent $\sigma^2_{obs}$. By opportunely modifying the ELBO to fix the aforementioned asymmetry between objective and

[1]The ELBO can again be written as a sum over data points

predictive posterior, the authors introduce a new loss function where the two variance terms, $\sigma_f(x)^2$ and $\sigma_{obs}^2$, are treated consistently. In [Jankowiak et al., 2019], the authors show that equipping SVGP with this new objective results in significant improvement in terms of UQ. Finally, in [Jankowiak et al., 2020], the DGP framework proposed in [Salimbeni and Deisenroth, 2017] is combined with the new loss function introduced in [Jankowiak et al., 2019]. DSPPs arise from the necessity of overcoming one last theoretical obstacle: the direct application of the objective introduced in [Jankowiak et al., 2019] to the DGP setting would result in the computation of the logarithm of a continuous mixture of Normal distributions. The approximation of such expectation via Monte-Carlo sampling would yield a biased estimator. To cope with this issue, the authors propose to replace the continuous mixture of Gaussians with a parametric (finite) mixture of Gaussians. This procedure is practically implemented by applying an opportune quadrature rule (e.g. the Gauss-Hermite quadrature rule).

### 2.3 Dataset

We evaluate and compare the UQ capabilities of the selected ML techniques on a new version of the popular C-MAPSS dataset for benchmark of prognostics models Saxena et al. [2008]. The new C-MAPSS dataset is a synthetic dataset providing the full degradation trajectories of a fleet comprising nine large turbofan engines under real flight conditions. Concretely, the flight data cover climb, cruise and descend flight conditions corresponding with different commercial flight routes. The degradation trajectories are given in the form of multivariate time-series of sensor readings and physics-inferred process features derived following the method in Arias Chao et al. [2020a] (i.e., virtual sensors and system health parameters). The dataset was generated with the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dynamical model Frederick et al. [2007]. Two distinctive failure modes are present in the development dataset: a high pressure turbine (HPT) efficiency degradation and a more complex failure mode that affects the low pressure turbine (LPT) efficiency and flow in combination with the high pressure turbine (HPT) efficiency degradation. Test units (i.e., Units 11, 14, & 15) are subjected to the latter complex failure mode. The sampling rate of the data is 0.1 Hz resulting in a total size of the dataset of 0.53M samples for model development and 0.12M samples for testing. More details about the generation process can be found in Arias Chao et al. [2020b].

### 2.4 Problem Formulation

Given are multivariate time-series of condition monitoring sensors readings and physics-inferred process features $X_i = [x_i^{(1)}, \ldots, x_i^{(n_i)}]^T \in R^m$ and their corresponding RUL i.e., $Y_i = [y_i^{(1)}, \ldots, y_i^{(n_i)})]^T$ from a fleet of six units (i.e., $N_{train}$= 6). The length of the input feature vector for the *i-th* unit is given by $n_i$; which differs from unit to unit. The total combined length of the available data set is $N = \sum_{i=1}^{N_{train}} n_i$ and the dimension of the input features is 41 (i.e., $m = 41$). More compactly, we denote the available dataset as $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N_{train}}$. Given this set-up, the task is to obtain a predictive model that provides a reliable RUL estimate ($\hat{\mathbf{Y}}$) with UQ on a test dataset of $M = 3$ units $\mathcal{D}_{T*} = \{X_{s_j*}\}_{j=1}^{M}$.

## 3 Results

In this section, we apply the methods described above to the CMAPPS dataset in order to predict the RUL of the three test units (i.e., units 11, 14 and 15) and quantify the uncertainty of the predictions. We report the results provided by the SVGP [Hensman et al., 2015], DGPs [Salimbeni and Deisenroth, 2017] and DSPPs [Jankowiak et al., 2020] models, all equipped with the new objective introduced in [Jankowiak et al., 2019][2]. Indeed, we empirically find that, in agreement with the results of Jankowiak et al. [2019], using the ELBO has a negative impact on UQ for both SVGP and DGPs. All our models are implemented using the open-source library GpyTorch [Gardner et al., 2018].

**Visualization.** In this section, we provide visualizations to demonstrate the effectiveness of our GP-based techniques in the UQ task. Since all our models provide very similar confidence bounds,

[2]In particular, in the case of DGPs, we compute a biased estimator of the continuous mixture of Gaussians obtained by applying Monte-Carlo sampling.

we report only the results obtained by the DSPP model. Fig. 1 shows the true and predicted RUL provided by a standard fully-connected neural network (FFNN) model (left) and the DSPP model (right) for each of the test units. The full lines correspond to the true RUL, the dots are the average RUL estimates at each cycle and the shaded surface shows the $\pm 3\sigma$ confidence bounds of the RUL predictions within each cycle. In both cases, the models' predictions tend to align with the true values as the RUL decreases. However, while the FFNN does not provide an estimation of the uncertainty, DSPP does. The confidence bounds show an important desirable characteristic for RUL models. The values of the predictive variance drop over time. This is physically meaningful since predictions when the machine is far away from its breaking point are much more uncertain. As a result, the confidence bounds associated with early operating times are significantly larger that those corresponding to the machine's end of life. Such a property has very important practical implications since it enables to design risk-aware maintenance strategies.

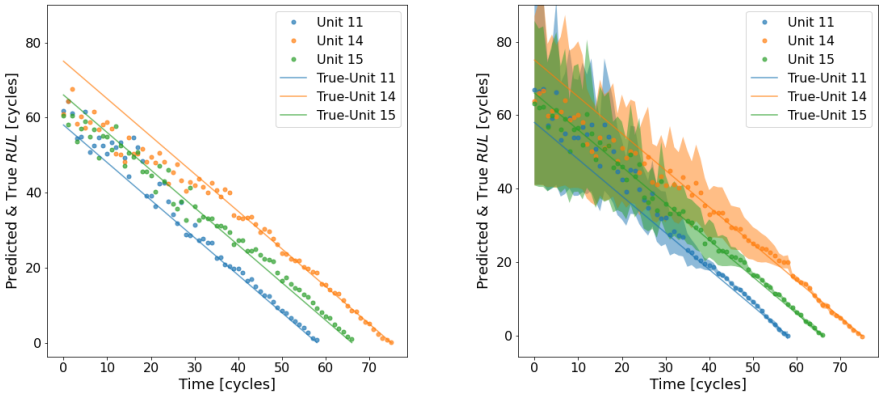


Figure 1: Predicted vs. ground-truth RUL values for units 11,14,15 for a standard feed-forward deep neural network (left) and a our DSPP model (right). $\pm 3\sigma$ confidence intervals are represented.

**Quantitative analysis.** In addition to the qualitative evaluation of the predictive uncertainties, we also compare the prediction accuracy obtained by our three techniques in terms of RMSE and negative log-likelihood (NLL). Both the DGP and DSPP models have one hidden-layer. We select the number of inducing points $I \in \{100, 500, 1000\}$, the width of the hidden layer $W \in \{2, 3\}$ and the number of quadrature sites $Q \in \{5, 8, 10\}$ by performing a grid-search over the hyperparameter space for each model. Additionally, we provide the results achieved by two DL models: a standard deep feed-forward neural network (FFNN) and a one-dimensional Convolutional Neural Network (1-d CNN). The FFNN model is characterized by $L \in \{2, 3, 4, 5\}$ hidden linear layers, with $H_f \in \{50, 100, 200\}$ hidden units each. The search space of the 1-d CNN model consists of $C \in \{2, 3, 4\}$ convolutional modules, each with $F \in \{10, 20, 30\}$ filters of size $K \in \{10, 20\}$. A fully connected linear layer with $H_c \in \{50, 100\}$ units is used after the convolutional blocks. ReLU activation functions are used in both DL models. The results are shown in Tab. 1.

| **Gaussian Processes** | | |
|---|---|---|
| **Models** | **RMSE** | **NLL** |
| SVGP [Hensman et al., 2015, Jankowiak et al., 2019] | 4.90 | 2.72 |
| DGP [Salimbeni and Deisenroth, 2017, Jankowiak et al., 2019] | 4.74 | 2.57 |
| DSPP [Jankowiak et al., 2020] | **3.97** | **2.46** |
| **Deep Neural Networks** | | |
| **Models** | **RMSE** | **NLL** |
| FFNN | 4.11 | - |
| 1d CNN [Arias Chao et al., 2020a] | 4.18 | - |

Table 1: Comparison of SVGP, DeepGP, DSPP, FFNN and 1d CNN in terms of RMSE and negative log-likelihood (NLL) on the test data.

The above results clearly show that the RMSE values provided by our three GP-models are compatible with the result given by the neural network model. In the case of DSPP, the obtained RMSE is slightly below the neural network one.

## 4 Conclusion

In this work, we provide the first evidence that modern GP models can be successfully applied to the domain of PM of industrial assets. This is made possible by a number of relatively recent advances in the ML literature, which have helped to make GPs more scalable and expressive. Our experiments show that the application of such techniques to the C-MAPSS dataset results in predictive performances close to or superior than those obtained by two strong DL baselines. Nevertheless, the proposed GP-models are able to provide physically meaningful uncertainty estimates alongside their RUL estimates. As shown in our visualizations, this aspect is in stark contrast to the behaviour of a standard deep neural network model which does not take uncertainty into account in its predictions. In light of the results obtained, we believe that the models analyzed in this work could be successfully employed in several real-world scenarios, especially in those cases where overly confident predictions might result in catastrophic consequences.

## References


M. Arias Chao, C. Kulkarni, K. Goebel, and O. Fink. Fusing physics-based and deep learning models for prognostics, 2020a.

M. Arias Chao, C. S. Kulkarni, K. Goebel, and O. Fink. Damage propagation modeling for aircraft engine run-to-failure simulation under real flight conditions. *Under review*, 2020b.

M. Bauer, M. van der Wilk, and C. E. Rasmussen. Understanding probabilistic sparse gaussian process approximations, 2017.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

A. Damianou and N. Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

D. K. Frederick, J. A. Decastro, and J. S. Litt. User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS). Technical report, 2007. URL `http://www.sti.nasa.gov`.

J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7587–7597, Red Hook, NY, USA, 2018. Curran Associates Inc.

J. Hensman, A. Matthews, and Z. Ghahramani. Scalable variational gaussian process classification. 2015.

M. Jankowiak, G. Pleiss, and J. R. Gardner. Parametric gaussian process regressors, 2019.

M. Jankowiak, G. Pleiss, and J. R. Gardner. Deep sigma point processes. *arXiv*, pages arXiv–2002, 2020.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, Jan. 2006.

H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

A. Saxena, K. Goebel, D. Simon, and N. Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management*, pages 1–9. IEEE, oct 2008. ISBN 978-1-4244-1935-7. doi: 10.1109/PHM.2008.4711414. URL `http://ieeexplore.ieee.org/document/4711414/`.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, T. Cemgil, and S. Chiappa, editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.