Control Engineering Practice 74 (2018) 84–94

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

A hidden-Gamma model-based filtering and prediction approach for monotonic health factors in manufacturing

Gian Antonio Susto [a,*], Andrea Schirru [b], Simone Pampuri [b], Alessandro Beghi [a], Giuseppe De Nicolao [c]

[a] Department of Information Engineering, University of Padova, via G. Gradenigo 6/B, 35131 Padova, Italy
[b] Statwolf LTD, 38-39 Baggot Street Lower, Dublin, Ireland
[c] Department of Computer Engineering and Systems Science, University of Pavia, via Ferrata 1, 27100 Pavia, Italy

ARTICLE INFO

Keywords:
Decision support system
Gamma distribution
Industry 4.0
Monte Carlo methods
Particle filters
Predictive maintenance
Prognostic and health management
Semiconductor manufacturing

ABSTRACT

In the context of Smart Monitoring and Fault Detection and Isolation in industrial systems, the aim of Predictive Maintenance technologies is to predict the happening of process or equipment faults. In order for a Predictive Maintenance technology to be effective, its predictions have to be both accurate and timely for taking strategic decisions on maintenance scheduling, in a cost-minimization perspective. A number of Predictive Maintenance technologies are based on the use of "health factors", quantitative indicators associated with the equipment wear that exhibit a monotone evolution. In real industrial environment, such indicators are usually affected by measurement noise and non-uniform sampling time. In this work we present a methodology, formulated as a stochastic filtering problem, to optimally predict the evolution of the aforementioned health factors based on noisy and irregularly sampled observations. In particular, a hidden Gamma process model is proposed to capture the nonnegativity and nonnegativity of the derivative of the health factor. As such filtering problem is not amenable to a closed form solution, a numerical Monte Carlo approach based on particle filtering is here employed. An adaptive parameter identification procedure is proposed to achieve the best trade-off between promptness and low noise sensitivity. Furthermore, a methodology to identify the risk function associated to the observed equipment based on previous maintenance data is proposed. The present study is motivated and tested on a real industrial Predictive Maintenance problem in semiconductor manufacturing, with reference to a dry etching equipment.

1. Introduction

Advanced monitoring is a fundamental activity in the Industry 4.0 scenario to implement control, maintenance, quality, reliability, and safety policies (Arinton, Caraman, & Korbicz, 2012; Chioua, Bauer, Chen, Schlake, Sand, Schmidt, et al., 2016; Ma, Dong, Peng, & Zhang, 2017). In particular, Fault Detection and Isolation (FDI) (Ma et al., 2017) and Predictive Maintenance (PdM) (Nguyen, Do, & Grall, 2015) technologies have proliferated in the past recent years for diagnosis and prognosis of process/tool failures (Sikorska, Hodkiewicz, & Ma, 2011). While the aim of such technologies is similar and partly overlapped, PdM technologies are more focused on prognosis. Prognosis can be defined as the capability to provide early detection of the precursor and/or incipient fault condition of a component, and to design tools for managing and predicting the progression of such fault condition to component failure (Engel, Gilmartin, Bongort, & Hess, 2000). Given their goal, PdM technologies are typically applied to failures that are associated with wear and usage of the system/process (Susto, Schirru, Pampuri, McLoone, & Beghi, 2015), or, more generally, to failures that can be predicted in advance (Lewin, 1995; Susto, McLoone, Pagano, Schirru, Pampuri, & Beghi, 2013). Examples of such type of faults are the breaking of the source in ion-implantation processes in semiconductor manufacturing (Susto et al., 2015), the flute wear in cutting tool equipment (Benkedjouh, Medjaher, Zerhouni, & Rechak, 2015), and the lifespan of lithium-ion batteries (Liao & Köttig, 2016).

In this work we focus on the so-called 'Health Factors' (HFs), an important concept in prognostic.[1] HFs are quantitative indexes used to define the current status of a tool/process and to assess the future

[1] Health Factors are also indicated as 'Component Health' (Sikorska et al., 2011), 'State of Health'/'Health State' (Si, Wang, Hu, & Zhou, 2011; Zhou, Stein, & Ersal, 2017) or as 'Health Indicators' (Benkedjouh et al., 2015; Wang, Yu, Siegel, & Lee, 2008) by different authors and they are closely in relation with the concept of 'degradation data' (Chen, Lio, Ng, & Tsai, 2017).

* Corresponding author.
E-mail addresses: gianantonio.susto@dei.unipd.it (G.A. Susto), andrea.schirru@statwolf.com (A. Schirru), simone.pampuri@statwolf.com (S. Pampuri), beghi@dei.unipd.it (A. Beghi), giuseppe.denicolao@unipv.it (G. De Nicolao).

https://doi.org/10.1016/j.conengprac.2018.02.011
Received 9 July 2017; Received in revised form 3 January 2018; Accepted 23 February 2018
Available online 23 March 2018
0967-0661/

statuses of the system under exam (or of one of its components/subsystems), and its Remaining Useful Life (RUL) (Bressel, Hilairet, Hissel, & Bouamama, 2016; Butler & Ringwood, 2010; Wang et al., 2008), so that strategic decisions regarding maintenance scheduling and dynamic sampling plans can be taken (Nguyen et al., 2015). Being in direct relationship with wear, usage or stress of an equipment/component or system, HFs generally have a monotone evolution. A HF can be of very different nature: in its simplest form, HFs can be observable parameters that, thanks to specific domain expertise, can be associated with equipment/process health status. Example of health factors as quantities that are directly related to system health, such as the thermal index of a polymeric material (Xie, Jin, Hong, & Van Mullekom, 2017), the scar width in sliding metal wear (Hu, Li, & Hu, 2017), and the temperature difference in semiconductor manufacturing epitaxy processes (Susto, Beghi, & Luca, 2012). HFs can also be the output of Soft Sensor modules (Souza & Araujo, 2014; Wang, Liu, & Srinivasan, 2010), where the status health is impossible/costly to be monitored. Moreover, HFs can be the residual of first principle FDI models (Zhang & Canova, 2015). In fact, in many practical examples (Arinton et al., 2012; Butler & Ringwood, 2010; Hast, Findeisen, & Streif, 2015; Zhang & Canova, 2015), residuals have a monotonic behavior and threshold-based policies to maintenance management are implemented on such quantities. HFs are therefore relevant quantities in both model-based (Dey, Biron, Tatipamula, Das, Mohon, Ayalew, et al., 2016; Hast et al., 2015; Xu, Lee, Zhou, & Yang, 2015) and model-free (Arinton et al., 2012; Bakdi, Kouadri, & Bensmail, 2017; Ge, Song, & Gao, 2013; Ma et al., 2017) prognostic approaches.

In the present paper, the problem of designing a HF for Predictive Maintenance (PdM) purposes is considered (Ding, Yin, Peng, Hao, & Shen, 2013; Filev, Chinnam, Tseng, & Baruah, 2010; Susto et al., 2015). In particular, the issue of assessing the probability distribution of the HF future values given its past measurements is addressed, under the following assumptions: (i) the HF is monotonically increasing; (ii) its measurements are subject to random noise that may conceal its monotonic nature; (iii) measurements are non-uniformly sampled over time. The aforementioned features are typical traits of HF signals (Butler & Ringwood, 2010; Gorinevsky, 2004; Saha, Goebel, & Christophersen, 2009; Susto et al., 2012; You, Li, Meng, & Ni, 2010), but they are generally not simultaneously accounted for in the related literature. Non-stochastic models (see Si et al., 2011 for a broad review on RUL estimation) for HFs have been presented in literature, as well as inspection and intervention approaches for increasing maintenance actions effectiveness and decreasing the associated costs. However, such methodologies are well suited for noise-free scenarios and, given the aforementioned assumptions on the HF signals, it is here proposed to adopt a stochastic filtering paradigm (Wang, Hussin, & Jefferis, 2012). With the proposed approach, the HF is treated as a stochastic process, with the possibility to combine prior knowledge on the HF with statistical information regarding the observed noisy data. A simple approach to deal with the problem at hand is provided by the Wiener and Kalman predictors (Abdennadher, Venet, Rojat, Rétif, & Rosset, 2010; Lu, Tu, & Lu, 2007; Susto et al., 2012; Yang & Liu, 1999), which are statistically optimal for linear Gaussian models. However, such classical approaches may be considered suboptimal for signals with the characteristics given in assumptions (i)–(iii). As a matter of fact, far from being Gaussian, the HF derivative is in this work considered to be a nonnegative random variable.

Given such premises, a framework for HF filtering and prediction based on the Gamma distribution is here proposed. PdM applications employing Gamma distributions has been developed since the 1970s (Abdel-Hameed, 1975), especially in mechanical and civil engineering applications (Cinlar, Osman, & Bazoant, 1977; Lawless & Crowder, 2004; Lu, Pandey, & Xie, 2013) and, recently, in industrial environments (LeSon, Fouladirad, & Barros, 2016). Indeed, if the HF is modeled as the sum of Gamma distributed random variables, such sum is still Gamma distributed, with the advantage that convenient estimation and prediction algorithms can be derived. Given that in real-world industrial applications HFs are usually observed with noise, the approach proposed in this work considers the HF as a monotonic Gamma process (with time-varying shape parameter) corrupted by Gaussian noise (*hidden-Gamma* model). Such assumptions lead to the lack of closed-form solutions for the estimation of model parameters in the proposed approach. However, it will be shown that the prediction problem can be efficiently solved by resorting to particle filtering methods (Alrowaie, Gopaluni, & Kwok, 2012; Doucet, 1998; Doucet, DeFreitas, & Gordon, 2001), employing Monte Carlo (MC) simulations to derive the target posterior distributions. Finally, a recursive procedure to estimate the time-varying shape parameter is proposed. Such procedure allows to optimize a trade-off between the need for promptness and noise insensitivity/outlier rejection.

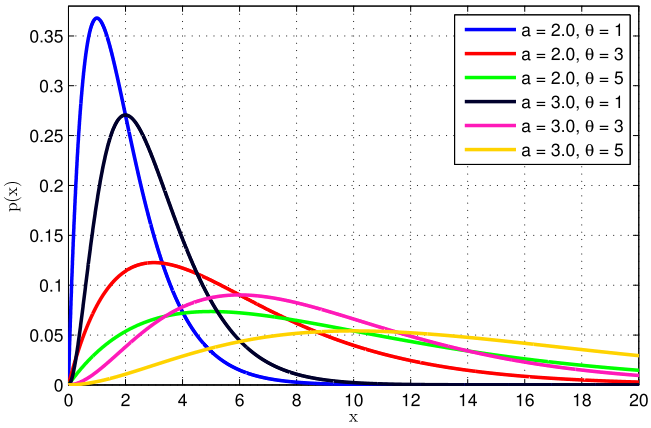

Fig. 1. Gamma probability distributions for different values of a and θ.

The paper is organized as follows. In Section 2 the hidden-Gamma model is presented. In Section 3.1 the principles of Particle Filtering (PF) are briefly summarized and adapted to Gamma processes. In Section 4 an adaptive recursive scheme for estimation of monotone HFs is presented. Section 5 is dedicated to the definition and estimation of an appropriate Risk Function for the proposed model. In Section 6 some experimental results on synthetic datasets are reported, whereas in Section 7 a real PdM semiconductor manufacturing problem related to dry etching is tackled.[2]

2. The hidden Gamma process

2.1. Gamma probability distribution

The most notable property of Gamma distributions is their non-negative support. We consider a random variable x with Gamma distribution $\Gamma(a,\theta)$, where a is the shape parameter and θ is the scale factor. The first two statistical moments of x are $E[x] = a\theta$ and $Var[x] = a\theta^2$ and the probability density function (PDF) is $p(x) = \frac{x^{a-1}e^{-\frac{x}{\theta}}}{\Gamma(a)\theta^a}$. Gamma distributed random variables enjoy the following property:

Property 1 (*Infinite Divisibility*). *If $x_1 \sim \Gamma(a_1,\theta)$ and $x_2 \sim \Gamma(a_2,\theta)$, then the sum $x = x_1 + x_2$ has a Gamma distribution with shape $a_1 + a_2$ and scale factor θ.*

The shape of the Gamma probability distribution for different values of a and θ is shown in Fig. 1.

[2] The present work is an extended version of Schirru, Pampuri, and DeNicolao (2010). Additional material concerns implementation details, the derivation of a risk function associated with the maintenance operation, and the use of synthetic data to better assess performance of the algorithms.

2.2. The hidden-Gamma model

In the following, the HF is denoted by $x(\cdot)$. Measurements $x(t_1), x(t_2), \dots, x(t_k)$ are available for time instants $t_1 \le t_2 \le \cdots \le t_k$. We indicate with $t_0 \le t_1$ the initial time instant, and with $t_{k+1} \ge t_k$ the instant points in which predictions of the HF are desired.

In the following, we adopt a Bayesian paradigm by assigning to $\{x_j, j = 1, \dots, k+1\}$ a joint prior distribution. As stated in Section 1, the HF is associated with equipment/process degradation, therefore the prior information on the HF can be formalized as follows:

- HF takes non negative values: $x_j \ge 0, \forall\ 0 \le j \le k+1$;
- HF has non negative increments: $\Delta x_j \ge 0,\ \forall\ k, 1 \le j \le k+1$, where $\Delta x_j := x_j - x_{j-1}$;
- Δx_j is positively correlated with the length of $t_j - t_{j-1}$.

The previous characteristics for the HF are captured by the following stochastic model:

Assumption 1. The HF evolution is governed by the following equation

$$x_{j+1} = x_j + w_{j+1}, \qquad j = 1, \dots, k \tag{1}$$

where $x_0 \sim \Gamma(a_0, \theta)$ and $w_j \sim \Gamma(\alpha(t_j - t_{j-1}), \theta)$. It is supposed that w_j are mutually independent random variables, also independent from x_0. □

It is straightforward to see from Eq. (1) that both the HF and its increments are non negative. Moreover, thanks to Property 1, it can be seen that $E[x(t_j)] = (a_0 + \alpha t_j)\theta$, which means that it is expected that the HF is linearly increasing with time.

Remark 1. the discrete-time model described in (1) can be obtained by sampling the continuous-time Gamma process $x(t)$ that satisfies

$$dx(t) = x(t)dt + dw(t), \qquad t \ge t_0, \tag{2}$$

where $x(t_0) \sim \Gamma(a_0, \theta)$ and $dw(t)$ is a Wiener process with $dw(t) \sim \Gamma(\alpha dt, \theta)$. Eq. (2) can be obtained thanks to Property 1. Such formulation allow us to estimate the HF for the generic time instant $t \ne t_j$. □

If $\{x_j\}$ were noiseless, the estimation of the unknown parameters $\{a_0, \alpha, \theta\}$, that specify the distribution of the future values of the HF, could be performed for example via maximum likelihood estimation (MLE). In such case, the posterior expectation can be employed as point predictor

$$\hat{x}_{k+1} := E[x_{k+1}|x_k] = x_k + E[w_{k+1}] = x_k + \alpha\theta(t_{k+1} - t_k).$$

Moreover, the knowledge of the distribution of w_{k+1} allows to define confidence intervals for the HF. Such confidence intervals could be exploited in a PdM perspective, by computing the probability of exceeding predefined thresholds associated with a maintenance action.

Unfortunately, in real world scenario, HFs are usually affected by measurement noise. For this reason, a *measurement equation* is added to the model (1):

Assumption 2. The HF is observed through the noisy measures

$$y_j = x_j + v_j, \qquad j = 1, \dots, k \tag{3}$$

where $v_j \sim \mathcal{N}(0, \sigma^2)$ (Gaussian distribution with 0 mean and standard deviation σ) is independent from the initial value x_0 and $\{w_j\}$.

The stochastic process y_k is named here a *hidden Gamma process* (HGP). Given the presence of the measurement noise v_j in (3), there is no guarantee that the sequence $\{y_j\}$ is monotonic. In the following, it is assumed that a_0, α, θ are known, as they can be estimated by MLE even in presence of noise. The formulation of the filtering problem is then the following:

Problem 1. Given the available measures $Y_k = \{y_j, j = 1, \dots, k\}$ and Assumptions 1–2, compute the posterior PDF $p(x_{k+1}|Y_k)$.

Since $p(x_{k+1}|x_k, x_{k-1}, \dots) = p(x_{k+1}|x_k)$, the HGP defined in (1) is a first-order Markov process. Then, Problem 1 can be approached by looking for a recursive solution where $p(x_{k+1}|Y_k)$ is computed by updating $p(x_k|Y_{k-1})$, once the measure y_k is available. In the noisy conditions we are considering in this work, such solution must be derived with numerical MC techniques like PF.

3. Particle filtering of Gamma processes

3.1. Basics of particle filtering

PFs or *Sequential Monte Carlo methods* are numerical approaches that allow to approximate intractable or complex distributions by employing discrete distributions whose statistical moments and confidence intervals can be easily calculated. PFs exploits the generations of N random variables, named *particles*, to approximate the unknown stochastic process posterior. A basic PF algorithm for a hidden state-space system as the one given by (1)–(3) is provided here[3] (a graphical representation of the PF procedure is depicted in Fig. 2):

Algorithm 1: Particle Filter (PF) algorithm

1. Set $k = 0$
2. Particles $x_0^{(j)}, j = 1, \dots, N$ are drawn from the initial distribution $p(x_0)$
3. Weights $w_0^{(j)} = p(x_0 = x_0^{(j)}),\ j = 1, \dots, N$ are computed
4. Update $k = k + 1$
5. From a suitable proposal distribution $G(x_k^{(j)}|x_{k-1}^{(j)})$, N particles $x_k^{(j)}, j = 1, \dots, N$ are sampled
6. The weights

$$w_k^{(j)} = w_{k-1}^{(j)} \frac{L(y_k; x_k^{(j)}) p(x_k^{(j)}; x_{k-1}^{(j)})}{G(x_k^{(j)}|x_{k-1}^{(j)})} \tag{4}$$

are adjusted, where L is a likelihood function defined by the measurement model (3) and the known statistics of v_j, while the state-transition probability $p(x_k^{(j)}; x_{k-1}^{(j)})$ is specified by the state-space model (1)

7. The weights are normalized in order to sum to 1
8. $p(x_k|Y_k)$ is approximated by

$$p(x_k|Y_k) \approx \sum_{j=1}^{N} w_k^{(j)} \delta(x_k - x_k^j), \tag{5}$$

a discrete distribution with support points $x_k^{(j)},\ j = 1, \dots, N$, where $\delta(\cdot)$ is the Dirac delta measure. 9. Go to Step 4.

Beside the selection of the number of particles N, the most important design choice of the PF procedure is the selection of G. The simplest choice is to set $G(x_k^{(j)}|x_{k-1}^{(j)}) = p(x_k^{(j)}; x_{k-1}^{(j)})$, so that only L is required in (4). Notably, with this choice, y_k has no influence on Step 5 of the procedure, reducing in general the robustness of the estimates. A possible alternative is to use a G that considers a preliminary approximation of $p(x_k|Y_{k-1})$; this can be achieved, for example, by means of a Kalman Filter (KF) approach (Douc & Cappé, 2005).

A second critical design issue in the PF procedure is the *resampling step* (Pitt & Shephard, 1999). If a large number of particles have their respective weights with very small values $w_k^{(j)} \ll \frac{1}{N}$, it is necessary that such particles are discarded and re-sampled from the distribution $p(x_k|Y_{k-1})$. This is done in order to allow the uninformative particles to contribute again to the estimation. Many resampling strategies have computational cost $\mathcal{O}(N)$, however less resource-demanding approaches for resampling can be implemented (Li, Bolic, & Djuric, 2015).

[3] We refer the interested readers to Arulampalam, Maskell, Gordon, and Clapp (2002) or Doucet et al. (2001) for more details on PF.

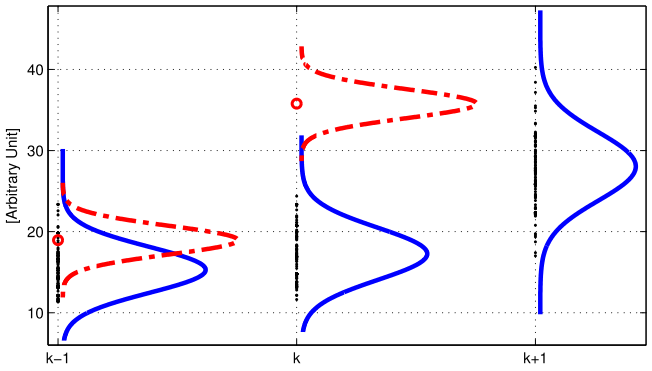

Fig. 2. A graphical representation of the PF procedure. At time instant k, the prior state distribution, represented by the blue solid curve, is approximated by the N particles (black dots). The posterior distribution arises from the interplay between the prior distribution and the observation likelihood, represented by the red dashed curve, generated by the noisy observation (the red circle).

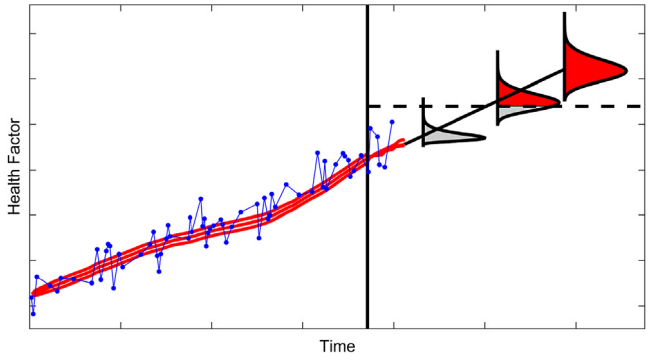

Fig. 3. HGP-based predictions (represented by the shaded areas) in three different future time instants for a synthetic HF. Thanks to the monotonic nature of the HGP model, the prediction uncertainty grows with time only in one direction. By computing the integral of the red shaded areas it is possible, for fixed thresholds, to estimate the threshold-crossing probability at different future time instants.

3.2. Adaptation to Gamma processes

A possible issue affecting the PF problem for system (1)–(3) is related to the nonnegativity of quantities w_j. Such issue is related to the fact that an overestimation of the lower limit of the distribution of x_k will propagate to all estimates x_i with $i > k$. In such case, lower limits of x_i will be overestimated as well, leading to the accumulation of one-sided errors. This issue is formally described in the following proposition.

Proposition 1. *If, for some k, the posterior distribution $p(x_k|Y_{k-1})$ is approximated by a representation with discrete support, the lower limit of the support of $p(x_{k+1}|Y_k)$ is greater or equal to that of $p(x_k|Y_{k-1})$.*

For a proof of Proposition 1, we refer the interested readers to Schirru et al. (2010). A possible way to mitigate the aforementioned propagation error is to employ a fixed-lag smoother (Arulampalam et al., 2002), where $p(x_{k+1}|Y_{k+W})$ is taken as the basis for future updates instead of $p(x_{k+1}|Y_k)$, $p(x_{k+1}|Y_{k+W})$ (the integer W denotes a fixed window size). With this approach, the smoothed $p(x_{k+1}|Y_{k+W})$ is generally more accurate and prone to overestimating the lower limit of the distribution, thanks to the increased availability of information.

The fixed-lag smoother can be derived as follows. The augmented state vector $\tilde{x}_j := [x_j \; x_{j+1} \; \dots \; x_{j+W}]'$ is introduced and, similarly, $\tilde{w}_j$, $\tilde{v}_j$, $\tilde{y}_j$. By exploiting (1)–(3) it is possible to obtain

$$\tilde{x}_{j+1} = \tilde{x}_j + \tilde{w}_j \tag{6}$$

$$\tilde{y}_j = \tilde{x}_j + \tilde{v}_j \tag{7}$$

The augmented-state model described by Eqs. (6)–(7) allows more robustness in the PF approach. In Fig. 4 an example with a synthetic HF is illustrated. It can be observed that larger lag sizes allow enhanced estimation stability (even in presence of outliers) and prevent systematic bias.

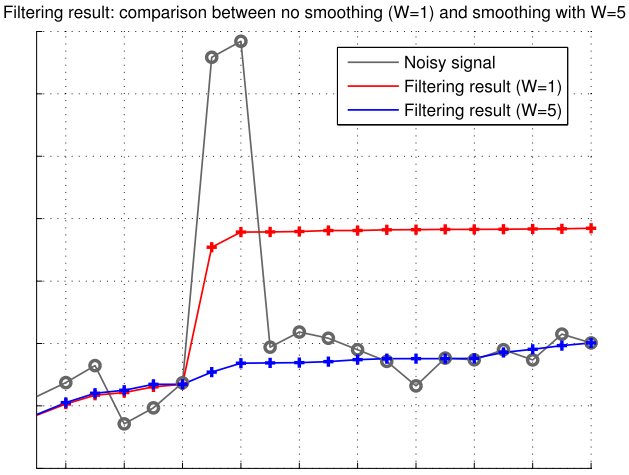

Fig. 4. State estimations with different lag values. It can be appreciated how larger values of lag are associated with higher stability in presence of outliers.

A final design guideline for the Hidden Gamma Process-PF regards resampling. *Conditional* resampling has been here implemented to make less likely for a particle to be sampled depending on its distance from the critical edge. While this procedure can lead to the creation of zones where particles are rarely resampled, a mitigated risk of error propagation is achieved.

4. Regularized adaptive filtering

The a-priori knowledge on the HF increment from t_{j-1} to t_j is expressed by the statistics $E[w_j] = \alpha(t_j - t_{j-1})\theta$ and $Var[w_j] = \alpha(t_j - t_{j-1})$: the higher the α, the higher the expected size of the increments. In real world industrial cases, variations in HF may happen quite suddenly with a steep rise of $x(t)$ after flat steady-state behavior (see the application case discussed in Section 7). For this reason, a time-varying shape factor $\alpha = \alpha(t)$ is used here (see Fig. 5), that must be estimated as well by the PF. Considering the discrete-time nature of (1)–(3), the shape factor can be denoted as $\alpha_j = \alpha(t_j)$, and the following hypothesis can be made:

Assumption 3. The shape factor α_j evolves according to

$$\alpha_{j+1} = \alpha_j + \delta_j, \qquad j = 1, \dots, k \tag{8}$$

where $\delta_j \sim \mathcal{N}(0, \lambda^2)$ is independent of x_0, $\{w_j\}$ and $\{v_j\}$.

In this perspective, the hyper-parameter λ^2 can be tuned to modify the variability of α_j. Indeed, large values of λ^2 lead to quickly varying α_j and promptly reactive adapting PF. On the other hand, small values of λ^2 can improve the noise sensitivity at the price of a less responsive PF.

A maximum a posteriori (MAP) estimate for α_k is adopted based on a moving window approach. If $\tilde{\alpha}_j := [\alpha_j \alpha_{j+1} \; \dots \; \alpha_{j+W}]'$, the MAP estimate $\hat{\tilde{\alpha}}_k$ is

$$\hat{\tilde{\alpha}}_k = \arg\max_{\tilde{\alpha}_k} p(\tilde{\alpha}_k | \tilde{y}_k, x_{k-W}, \alpha_{k-W+1}) \qquad \text{where}$$

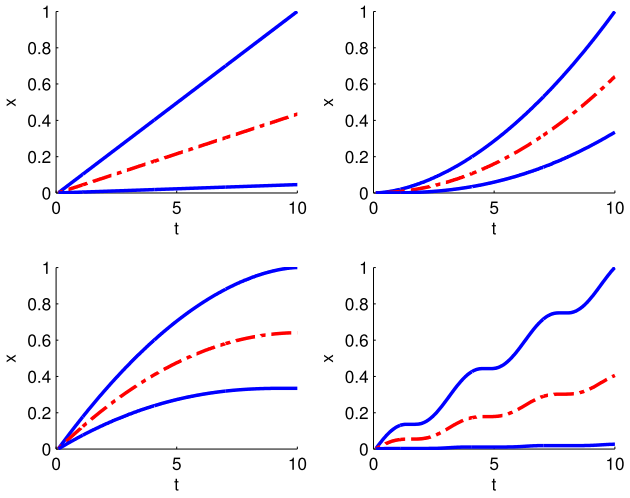

Fig. 5. Expected value (red line) and confidence limits (blue lines) corresponding with different choices of $\alpha(t)$ (from top-left, in clockwise order) constant, linear increasing, linear decreasing and periodic. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$p(\tilde{\alpha}_k|\tilde{y}_k, x_{k-W}, \alpha_{k-W}) \propto p(\tilde{y}_k|\tilde{\alpha}_k, x_{k-W})p(\tilde{\alpha}_k|\alpha_{k-W}) \tag{9}$$

with x_{k-W} and α_{k-W} set to be equivalent to their point estimates at previous iteration.

Given that the conditional distributions of $\tilde{y}_k$ and $\tilde{\alpha}_k$ in (9) are Gaussian, the logposterior $\mathcal{L} = \log(p(\tilde{\alpha}_k|\tilde{y}_k, x_{k-W}, \alpha_{k-W}))$ is defined (up to a constant) as

$$\mathcal{L} = SSR + \frac{1}{\lambda^2}R \tag{10}$$

where SSR and R are respectively the sum of squared residuals and the sum of squares of δ_j, $j = k-W, \dots, k$. Eq. (10) is typical of regularization methods (Seeger, 2009) where a trade off choice between accuracy in fitting the training data and complexity of the estimated function has to be made. In regularization methods, the following family of penalty terms is usually considered:

$$R = \sum_{i=0}^{W-1} |\delta_{k-i}|^q \,. \tag{11}$$

For $q = 2$, the well-known Ridge Regression (RR) (James, Witten, Hastie, & Tibshirani, 2014) is obtain. The advantage of RR is that it admits a closed-form solution. For $q = 1$, a LASSO-type (Hastie, Tibshirani, & Friedman, 2009; Tibshirani, 1996) regularization is obtained instead. LASSO provides sparse solutions, an important property that makes LASSO the first choice in many applications over RR, even at the price of not admitting a closed-form solution. Values of q larger than 1 can also be adopted; for instance, $q \in]1, 2[$ leads to a penalization region similar to the Elastic Net (Zou & Hastie, 2005). The computational cost of this operation for $k > W$ is $\mathcal{O}(W^2k)$ for $q = 1$ and $\mathcal{O}(W^3 + W^2k)$ for $q = 2$; optimized approaches (Wahlberg, Boyd, Annergren, & Wang, 2012) are now available for the less frequent case $k < W$.

Remark 2. The discrete-time model with time-varying α_k can still be interpreted as the sampled-data version of the continuous time model with time-varying $\alpha(t)$. In fact, $p(x(t_{j+1})|x(t_j))$ in (2) depends on the mean value of $\alpha(t)$ in the interval $[t_j, t_{j+1}]$, and not on its evolution inside the interval. From Property 1, $w_j = x(t_{j+1}) - x(t_j)$ is Gamma distributed, that is, $w_j \sim \Gamma(\bar{\alpha}_j, \theta)$ where $\bar{\alpha}_j := \int_{t_j}^{t_{j+1}} \alpha(t)dt$. Then, by setting $\alpha_j := \frac{1}{t_{j+1}-t_j}\int_{t_j}^{t_{j+1}} \alpha(t)dt$ it follows that $w_j \sim \Gamma(\alpha_j(t_{j+1}-t_j), \theta)$ and the discrete-time increment model is obtained. □

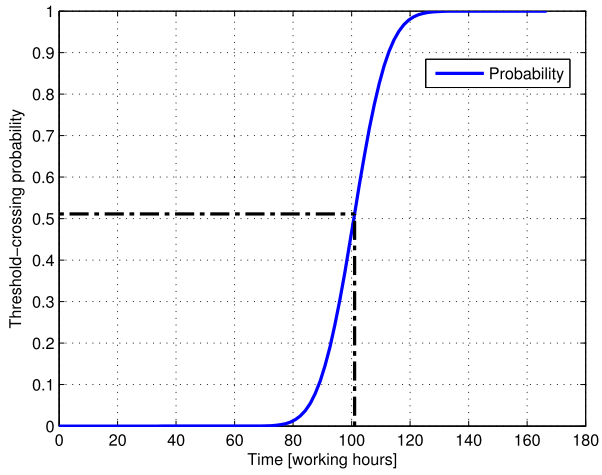

Fig. 6. Representation of a risk function: thanks to the properties of the proposed methodology, it is possible to assess the risk of crossing the threshold $\mathcal{H}(t)$ at a generic future time instant.

4.1. Implementation notes

The HGP-PF approach proposed in this work can be summarized (for the sequences $\{t_j, y_j\}$, $j = 1, \dots, k$) as:

1. Initial parameter are selected: the process noise-related quantities θ and α_0, initial state distribution $P(x_0)$, measurements noise variance σ^2, the PF design parameters N, W and λ^2.
2. For $j = 1, \dots, k$:
   (a) If $(j \geq W)$, $\tilde{\alpha}_j$ is updated by solving the regularization problem (Section 4);
   (b) $p(\tilde{x}_j|Y_j)$ is approximated;
3. Predictions and confidence intervals are computed using the newest estimation;

It can be shown, given that $p(x_{k+1}|x_k)$ is Gamma distributed (Schirru et al., 2010), that $p(x_{k+1}|Y_k)$ is a continuous mixture of Gamma distributions. Therefore $p(x_{k+1}|Y_k)$ will be approximated by a finite mixture of Gamma distributions since the PF provides an approximation of $p(x_k|Y_k)$ with discrete support (see Eq. (5)).

5. Risk function evaluation

Once a prediction of the future probability distribution of an observed HF is available, it can be compared with a *maintenance threshold* to compute and evaluate a risk function (RF) (Ding, Tian, & Yu, 2015) associated with the maintenance operation (Fig. 6). Such a maintenance threshold may be given from process/equipment operating conditions, or inferred from historical, noisy HF data, and it can be time/usage-dependent. In the following, a RF is formally defined and motivated in a maintenance optimization perspective. Furthermore, resorting to supervised classification theory, a method to estimate the maintenance threshold from historical maintenance data is also presented.

5.1. RF definition

Let $p_\tau^{t_k} := p(x(t_k + \tau)|Y_k)$, $\tau > 0$ be the predicted distribution of the observed health factor at time instant $t_k + \tau$. According to the paradigm established in the previous sections, p_τ is a mixture of Gamma distributions such that

$$p_\tau^{t_k}(x) = \sum_{i=1}^{N} w_i\omega_i(x, \tau) \tag{12}$$

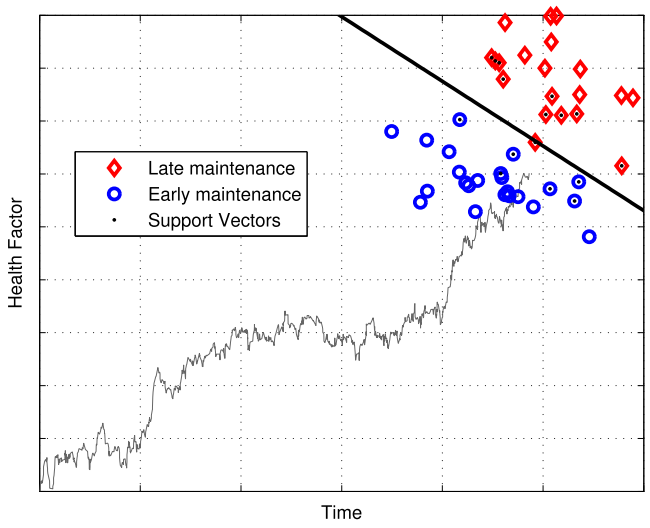

Fig. 7. A dataset D along with an example of HF reading (in gray). The optimal threshold $\mathcal{H}(t)$ is obtained using SVM ($p = 1$). It is to be noted that the optimal order of the class separation, p, can be determined by means of Generalized Cross Validation (GCV).

$$\omega_i(x,\tau) = \begin{cases} (x - x_i)^{\tau\alpha_k} \dfrac{e^{-(x-x_i)/\theta}}{\Gamma(\tau\alpha_k)\theta^{\tau\alpha_k}} & x \geq x_i \\ 0 & x < x_i \end{cases}$$

where $\sum_{i=1}^{N} w_i = 1$ and $\Gamma(\cdot)$ is the gamma function. Furthermore, let $\mathcal{H}(t)$ be a continuous real function with nonnegative codomain representing the time-dependent threshold for the analyzed HF. It follows that the probability of exceeding the threshold $\mathcal{H}$ at time instant $t_k + \tau$ is

$$\begin{aligned} p(x(t_k+\tau) > \mathcal{H}(t_k+\tau)|Y_k) &= 1 - \int_0^{\mathcal{H}(t_k+\tau)} p_\tau(x)dx \\ &= 1 - \sum_{i=1}^{N} w_i \int_0^{\mathcal{H}(t_k+\tau)} \omega_i(x,\tau)dx \end{aligned}$$

By observing that

$$\int_0^{\mathcal{H}(t_k+\tau)} \omega_i(x,\tau)dx = \frac{\gamma(\tau\alpha_k, (\mathcal{H}(t_k+\tau) - x_i)/\theta)}{\Gamma(\tau\alpha_k)}$$

where $\gamma(\cdot,\cdot)$ is the incomplete lower Gamma function, it is possible to compute the risk function

$$\mathcal{R}^{t_k}(\tau) := 1 - \sum_{i=1}^{N} w_i \frac{\gamma(\tau\alpha_k, (\mathcal{H}(t_k+\tau) - x_i)/\theta)}{\Gamma(\tau\alpha_k)} \tag{13}$$

The risk function $\mathcal{R}(t)$ depends on the choice of the threshold function $\mathcal{H}(t)$. Such choice can be done either by exploiting experts knowledge (for instance, a threshold beyond which the machine is known to malfunction) or by analyzing historical maintenance data. In the next subsection, classification theory is employed in order to estimate the optimal threshold when such data are available.

5.2. Threshold estimation

Let D be a set of N_m maintenance operations

$$D = \{T_i \in \mathbb{R}, y_i \in \mathbb{R}, s_i \in \{-1, 1\}\}_{i=1}^{N_m}$$

where T_i is the duration of the ith production cycle (maintenance to maintenance) and y_i is the last observed HF measurement. Furthermore, let s_i be an indicator of the effectiveness of the ith maintenance cycle. Since it is not possible to know what the status is of the maintained component before its replacement, two situations can occur. Let $s_i = -1$ conventionally represent an early replacement (the component is still functional when replaced) and let $s_i = 1$ represent a belated replacement (component replaced after it has broken).

To derive the threshold function $\mathcal{H}(t)$ from D, a Support Vector Machine (SVM) approach is hereby proposed. SVM techniques allow to find an optimal nonlinear separation between two categories of data points (if such categories are separable) or, in the *soft-margin* version, to produce an optimal robust (with respect to data mislabeling) classification. In the following, the focus is set on the estimation of a $\mathcal{H}(t)$ represented by a pth degree polynomial function of t. Let $\tilde{t}_i = [t_i,\ t_i^2,\ \ldots,\ t_i^p]$ be the polynomial span of t_i up to the pth degree. Furthermore, let $\tilde{z}_i = [y_i\ \tilde{t}_i]' \in \mathbb{R}^{p+1}$.

The problem of estimating $\mathcal{H}$, according to SVM soft-margin theory, can then be seen as the research of a parameter vector $\tilde{c}$, defined as $\tilde{c} = [c^{(y)},\ c^{(t_1)},\ c^{(t_2)},\ \ldots,\ c^{(t_p)}]'$ that solves the following problem.

Problem 2. Find

$$\min_{\tilde{c},\xi,b}\ \max_{\mu,\beta}\ J$$

where the cost function J is defined as

$$J = \frac{1}{2}\|\tilde{c}\|^2 + \sum_{i=1}^{N_m} \left[(C - \beta_i)\xi_i - \mu_i(s_i(\tilde{c}'\tilde{z}_i - b) - 1 + \xi_i)\right]$$

under the constraints $\mu_i \geq 0, \beta_i \geq 0$. C is a tuning parameter that can be selected by means of Generalized Cross Validation (GCV).

Problem 2 is the standard representation of the SVM problem with soft margin. Its solution can be obtained by means of a Sequential Minimal Optimization (SMO) approach (Platt, 1998).

The optimal separating hyperplane resulting from the solution of Problem 2 (Fig. 7) satisfies the condition

$$b + c^{(y)}y + \sum_{i=1}^{p} c^{(t_i)}t^i = 0$$

The threshold function $\mathcal{H}(t)$ is then obtained as

$$\mathcal{H}(t) = -\sum_{i=1}^{p} \frac{c^{(t_i)}}{w^{(y)}}t^i - \frac{b}{c^{(y)}} \tag{14}$$

By combining (13) and (14), the risk function is then

$$\mathcal{R}^{t_k}(\tau) = 1 - \tag{15}$$
$$\sum_{i=1}^{N} w_i \frac{\gamma(\tau\alpha_k, -(\sum_{i=1}^{p} \frac{c^{(t_i)}}{c^{(y)}}(t_k+\tau)^i + \frac{b}{c^{(y)}} + x_i)/\theta)}{\Gamma(\tau\alpha_k)}$$

5.3. Use of RF for maintenance optimization

The main goal of a maintenance management system is the minimization of the costs associated with failures and maintenance operations. The cost of maintenance operations can be associated with the cost of several factors, such as spare parts, equipment/production downtime, staff performing the interventions, scrap products due to the *requalification* of the system (a post-maintenance phase in which the process needs to run before being 'stable' or within the production standards). At first glance, minimizing the number of interventions seems to be the natural approach to minimize the aforementioned costs. However, Run-to-Failure (R2F) policies, where maintenances are performed only after a failure has happened, are generally deprecated because the costs related to unexpected failures can be very high. The trade-off between early-stage maintenances (and associated *unexploited lifetime* UL of the system) and unexpected breaks UB can be optimized by using corresponding agglomerated costs (respectively c_{UL} and c_{UB}) and the proposed RF. Furthermore, in the proposed approach, reliable predictions can be obtained over given time frames, a relevant feature in applications where timeliness is a critical issue, such as when interventions have to be planned in advance or the monitoring of RF and re-scheduling can be guaranteed only with a fixed time delay.

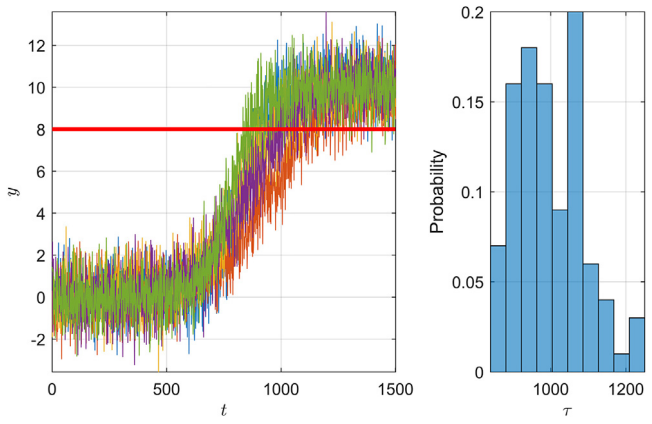

Fig. 8. [Sigmoid data] On the left panel a set of 5 time series from $D^{(1)}$ is shown. The red line (–) represents the maintenance threshold $\mathcal{H} = 8$. On the right panel an histogram of $\{\tau_i\}_{i=1}^{100}$ with $\mathcal{H} = 8$.

Table 1
Filtering performances of the considered approaches in terms of RMSE.

Experiment name	Kalman	HGP-PF
Sigmoid data	0.843	0.784
Sigmoid data with discontinuities	1.211	1.063
Industrial data	2.471	2.118

6. Simulation results

To test the proposed methodology, synthetic datasets representing HFs have been created. The generic dataset D has a total of N time series, the ith time series is defined as

$$S_i = \left\{ \{t_j, x_j\}_{j=1}^{n_i}, \{t_j, y_j\}_{j=1}^{n_i}, \tau_i \right\} = \left\{ [T_{n_i}\ X_{n_i}], [T_{n_i}\ Y_{n_i}], \tau_i \right\},$$

where τ_i indicates the time instant when the ith HF crosses a predefined threshold $\mathcal{H}$.

The accuracy of the prediction at a time instant $\tau_i - \Delta t$ is used to assess the performance of the proposed PdM algorithm. The initial filter parameters are chosen via likelihood maximization. A truncated multivariate Normal distribution is used as proposal distribution, that is then sampled through a Gibbs sampler (Casella & George, 1992). A Kalman Filter is used to obtain the non-truncated Normal distribution. Although the use of the Kalman filter is justified by model linearity, an Unscented Kalman Filter can also be used (VanDerMerwe, Doucet, DeFreitas, & Wan, 2004).

6.1. Sigmoid data

A first synthetic dataset $D^{(1)}$ of $N = 100$ time series has been generated as follows: for each time series the time domain is $T = [0, 1, \dots, 1500]$, while the ith HF at time t is

$$x_i(t) = \frac{10}{1 + e^{-a_i(t-b_i)}},$$

where $a_i \sim \mathcal{U}(0.005, 0.02)$, a uniform distribution of support $[0.005, 0.02]$, and $b_i \sim \mathcal{U}(750, 1000)$. Gaussian noise $v \sim \mathcal{N}(0, 1)$ is then added to the HF. A set of 5 time series belonging to $D^{(1)}$ is reported in Fig. 8. For this study a fixed maintenance threshold has been set to $\mathcal{H} = 8$.

The proposed HGP-PF is compared with a KF-based approach. The KF employed in this Section, and also in the experimental work detailed in Section 7 is formulated as follows. The following linear state-space model is considered:

$$z_{j+1} = Az_j + Gv'_j \quad (16)$$

$$y_j = Cz_j + w'_j, \quad (17)$$

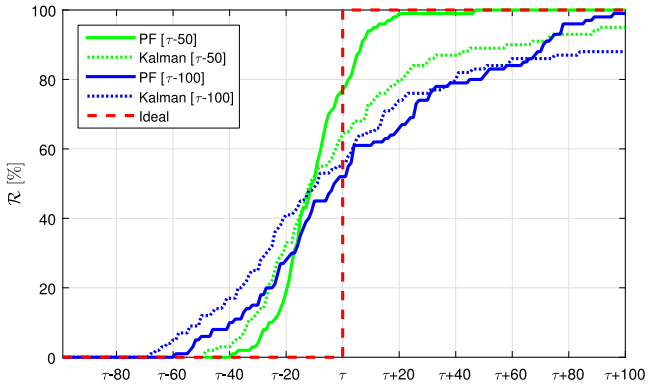

Fig. 9. [Sigmoid data] Averaged risk functions associated with different prediction approach and different prediction horizons.

where $z_j = [x_j \dots x_{j-n_O}]$ is a state vector that contains the estimated present and past values of the HF, $v'_j \sim \mathcal{N}(0, Q)$ is the model error, and $w'_j \sim \mathcal{N}(0, R)$ is the measurement error. v'_j and w'_j are supposed to be uncorrelated. Matrices A, C and G are estimated using the N4SID algorithm (Van Overschee & De Moor, 2012), while the model order n_O is chosen according to the Akaike Information Criterion (Bozdogan, 1987). The Kalman predictor for the 1-step ahead prediction has the classical formulation

$$\hat{z}_{j+1|j} = A\hat{z}_j + K_j \left[y_j - C\hat{z}_{j|j-1} \right], \quad (18)$$

$$K_j = AP_{j|j-1}C' \left[CP_{j|j-1}C' + R \right]^{-1}, \quad (19)$$

where $P_{j|j-1}$ is the variance matrix of the prediction error $\hat{z}_{j+1|j} - z_{j+1}$ and it is updated through the discrete Riccati equation. The tuning of Q and R has been done by computing a test on the residuals correlation

$$\mathbf{RE}_{Q,R}(\sigma) = \mathbb{E}\left[e_{Q,R}(j) e_{Q,R}(j+\sigma) \right], \quad (20)$$

with $e_{Q,R}(j) = y_j - C\hat{z}_{j|j}$ where the estimation $\hat{z}_{j|j}$ depends on the choice of Q and R. A grid search on different set of values of Q and R has been performed to minimize $\max_{\sigma>0} |\mathbf{RE}_{Q,R}(\sigma)|$. The multiple-step ahead prediction that can be easily derived by exploiting (16)–(18) (Susto et al., 2012).

In Table 1 the filtering performances of the proposed HGP-PF are reported, and compared to that of a KF-based approach in terms of RMSE. As for prediction accuracy, averaged Risk Functions are reported in Fig. 9 for 2 cases, a 50-step and 100-step ahead predictions. Risk Functions are aligned with respect to the time instant of the fault τ and compared with a ideal function that provides

$$\mathcal{R}_{\text{ideal}}(t) = \begin{cases} 1 & \text{if } t > \tau \\ 0 & \text{otherwise} \end{cases}$$

The HGP-PF beats the KF both in terms of RMSE and in PdM accuracy. In particular, it can be appreciated in Fig. 9 how the PF of the HGP-PF is qualitatively closer to $\mathcal{R}_{\text{ideal}}(\cdot)$ than that of the KF.

6.2. Sigmoid data with discontinuities

A sigmoid dataset with discontinuities $D^{(2)}$ (with $N = 100$ time series and support $T = [0, 1, \dots, 1500]$, as before) has been generated as follows: The ith HF at time t is

$$x_i(t) = \frac{10}{1 + e^{-a_i(t-b_i)}} + 0.6c_i,$$

where $a_i \sim \mathcal{U}(0.005, 0.02)$, $b_i \sim \mathcal{U}(750, 1000)$ and $c_i \sim \mathcal{B}(0.01, 1)$ is a binomial distribution with 1 trial and 0.01 success probability. Gaussian noise $v \sim \mathcal{N}(0, 1)$ is then added to the HF. A set of 5 time series in $D^{(2)}$ is reported in Fig. 10. A fixed maintenance threshold has been set to $\mathcal{H} = 14$. All of the N generated HFs exceed $\mathcal{H}$ in T. In Fig. 11 the

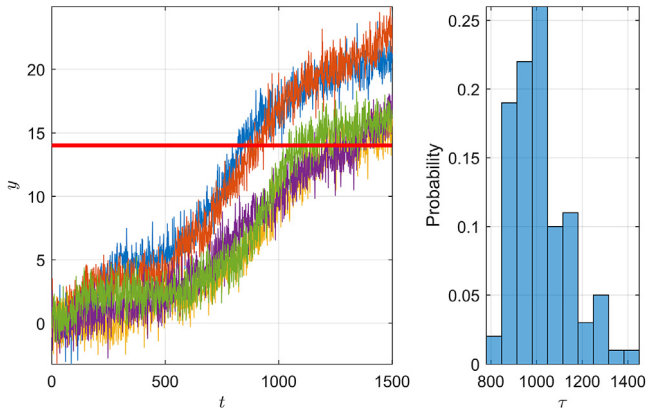

Fig. 10. [Sigmoid data with discontinuities] On the left panel a set of 5 time series from $\mathcal{D}^{(1)}$ is depicted. The red line (–) represents the maintenance threshold $\mathcal{H} = 14$. On the right panel an histogram of $\{\tau_i\}_{i=1}^{100}$ with $\mathcal{H} = 14$.

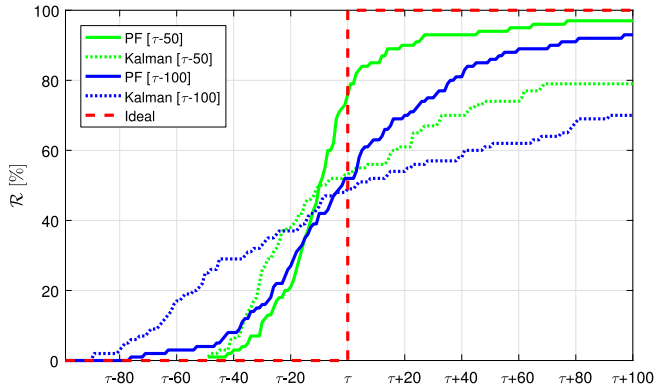

Fig. 11. [Sigmoid data with discontinuities] Averaged risk functions associated with different prediction approach and different prediction horizons.

averaged Risk Functions of PF and Kalman Filter fot the 50-step and 100-step ahead predictions are reported. RMSE performance is given in Table 1. In this case, too, the HGP-PF outperforms the KF both in terms of RMSE and in PdM accuracy.

7. Application to predictive maintenance of a dry etching equipment for semiconductor manufacturing

Maintenance optimization strategies based on HFs are receiving increasing attention in the field of semiconductor manufacturing (Butler & Ringwood, 2010; Susto et al., 2012). In this Section, the HPG-PF approach developed in the present paper is tested on a PdM problem related to a dry etching equipment. Etching is a fundamental step in semiconductor fabrication that is employed to chemically remove layers from the wafer surface. In some etching tools, the wafer is held on a Electrostatic Chuck (ESC) thanks to electrostatic charge, while a backside helium flow cools down the wafer and prevents problems during the unloading of the product (Wang, Cheng, Wang, Yang, Sun, Cao, et al., 2014). During this operation, the quartz parts around the ESC (and the ESC itself) undergo the action of aggressive plasma, causing a wearing out that affects both wafer quality and process stability. The intensity of the helium flow intensity in the etching equipment, represented in Fig. 12, reflects the wear of the ESC and can be considered a HF for the degradation problem at hand. The helium flow exhibits a monotonically increasing trend (masked by noise) and it is common practice that, in a Condition-based Maintenance fashion, when a given threshold is exceeded, maintenance operations take place, including the (expensive) replacement of several components. Predicting future

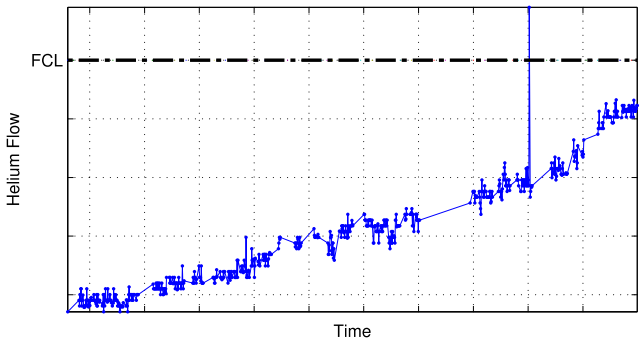

Fig. 12. Anonymized helium flow data and an example of fixed control limit (FCL). It can be noticed the observation noise (possibly resulting in outliers), as well as the time-varying behavior of the signal.

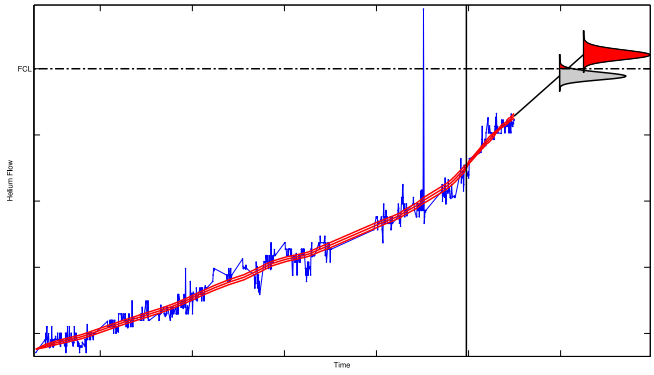

Fig. 13. [Industrial data] Filtering and prediction of the signal presented in Fig. 12.

values of the HF is fundamental to timely schedule maintenance actions and minimize the trade-off between unnecessary replacements and unexpected breaks. For such reasons, stability and reliability of the predictions are primary needs in the problem at hand, as in general for HF signal employed in PdM.

To test the proposed methodology, a dataset consisting of 17 complete helium flow readings (from maintenance to maintenance) is employed as benchmark. The ith test series is defined as

$$S_i = \left\{ \{t_j, y_j\}_{j=1}^{n_i}, \bar{t}_i \right\}, \; i = 1, \ldots, 17$$

where τ_i is the actual time instant at which a maintenance operation has taken place, and $\{t_j, y_j\}_{j=1}^{n_i}$ are the related helium flow readings. Specifically, the accuracy of the prediction at a time $\bar{t}_i - \Delta t$ is used to assess the performance of the proposed algorithm.

In Table 1 the filtering performances of the proposed HGP-PF and of a KF in terms of RMSE are reported. As in the simulation studies of previous section, the HGP-PF outperforms the KF. Fig. 13 reports filtering and prediction results based on the data of Fig. 12. In Fig. 3, the predictions distributions are represented with gray and red areas if they are below or above the FCL, respectively. Fig. 14 shows the estimated risk (i.e., the threshold-crossing probability) at time instant $\bar{t}_i - \Delta t$ for the actual failure time $\bar{t}_i$ (i.e., $\mathcal{R}(\bar{t}_i)$ computed with all the available information at time instant $\bar{t}_i - \Delta t$). Rather unsurprisingly, the best performances are obtained with the smallest Δt, namely, 100 working hours (left panel of Fig. 14). In this case, only 3 out of 17 estimated risks are below 50%, and only one below 40%. When Δt is increased to 200 working hours (central panel of Fig. 14), the increased uncertainty does not allow to obtain optimal results. For $\Delta t = 250$ working hours (right panel of Fig. 14) the phenomenon is even more evident, as the distribution of the risk function assessment is almost flat.

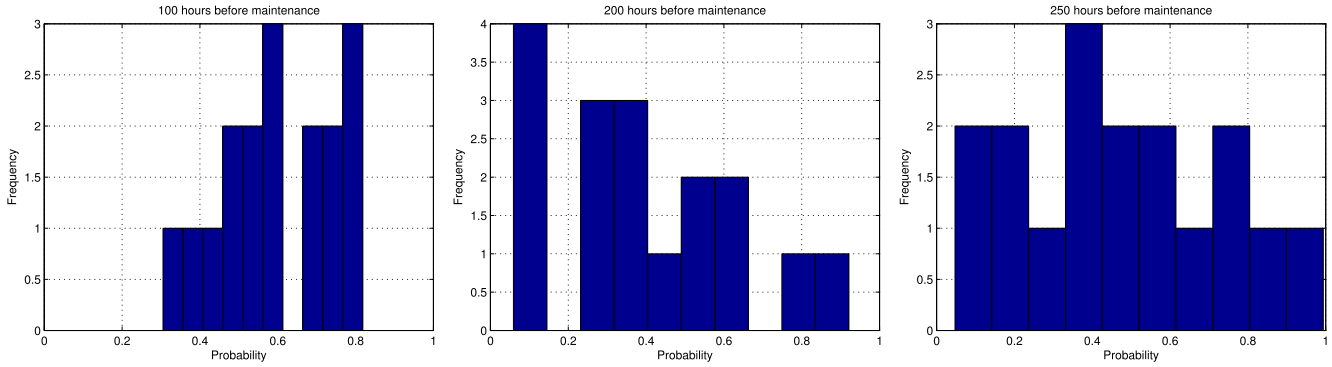

Fig. 14. [Industrial] Risk function assessments for the experimental dataset.

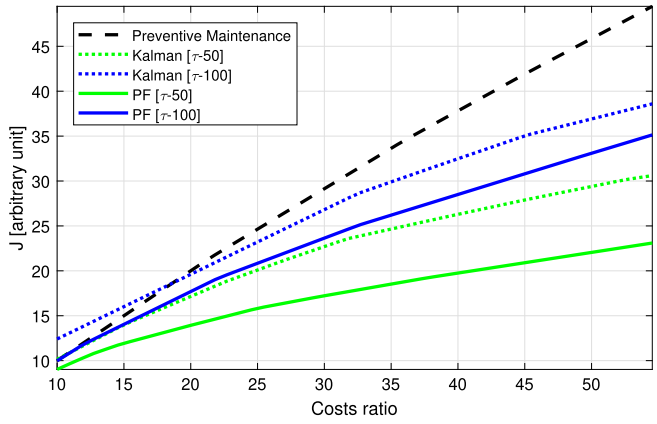

Fig. 15. [Industrial data] Optimal value of J for the various maintenance strategies as a function of the ratio of the costs $\frac{c_{UL}}{c_{UB}}$ (averaged results over 100 Monte Carlo simulations).

From an operational point of view, the most important feature of the proposed methodology is its capability to precisely assess the fault event probability when the life-cycle of the equipment is coming to an end, thus allowing to schedule an early maintenance operation. With respect to this requirement, the performances presented in Fig. 14 are satisfactory. The experimental results presented in this section show that the choice of a proper prediction range is crucial: Indeed, a long range would result in almost uninformative predictions, while a short range would yield extremely precise predictions when the optimal time to adjust the maintenance schedule is already passed. Such trade-off between precision and timeliness is consistent with the characteristics of proposed prediction paradigm.

Finally, the HGP-PF based and KF based PdM policies have been tested versus a simulated Preventive Maintenance (PvM) tool. PvM is a really popular approach to maintenance that triggers actions based on the amount of time passed from previous maintenance. Here the PvM tool has been simulated by computing a risk factor $\mathcal{R}_{\text{PvM}}$ on a different set D_0 of N_0 time series from the same distributions of D as follows

$$\mathcal{R}_{\text{PvM}}(t) = \frac{\sum_{i=i}^{N_0} \Theta(\tau_i - t)}{N_0}\%, \tag{21}$$

where $\Theta(\cdot)$ is the Heaviside step function. Observe that the Risk Function (21) is computed on the training data and does not depend on current sensor readings. Let ρ_{UB} be the percentage of unexpected breaks and ρ_{Ul} the number of unexploited runs. In Fig. 15 the overall costs $J = c_{UB}\rho_{UB} + c_{UL}\rho_{UL}$ are reported, over a range of different values of the ratio $\frac{c_{UL}}{c_{UB}}$, for the different policies. It can be appreciated how PdM approaches are generally superior to PvM and PF outperforms the Kalman Filtering-based PdM policy.

Table 2
[Industrial Data] HGP-PF-based PdM and PvM performances in terms of overall cost J for fixed values of the ratio $\frac{c_{UL}}{c_{UB}}$.

Maintenance policy	$c_{UL}/c_{UB} = 25$	$c_{UL}/c_{UB} = 40$
PdM HGP-PF $\Delta t = 25$	11.13	13.25
PdM HGP-PF $\Delta t = 50$	15.81	19.75
PdM HGP-PF $\Delta t = 75$	17.34	22.43
PdM HGP-PF $\Delta t = 100$	20.14	26.29
PdM HGP-PF $\Delta t = 150$	23.49	30.71
PdM HGP-PF $\Delta t = 200$	26.56	38.31
PdM HGP-PF $\Delta t = 250$	27.12	39.12
PvM	24.84	37.81

To better highlight the trade-off between prediction accuracy and timeliness of the proposed HGP-PF approach, performances for fixed values of the ratio $\frac{c_{UL}}{c_{UB}}$ and different values of Δt for the HGP-PF PdM-based and for the PvM-based maintenance management policy. are reported in Table 2. It can be noticed that, for $\Delta t = 200$ and $\Delta t = 250$, the PvM-based policy is more effective than the HGP-PF PdM-based one. It can also be appreciated that, as expected, the more timely a prediction is, the lower the performance is in terms of costs associate with unexpected breaks and unexploited lifetime. However, in a cost-minimization perspective such lower performance could be justified in some real world example by the cost savings associated with timely maintenance planning.

8. Conclusions and discussion

In this work, a hidden Gamma process particle-filter approach for health factor has been presented. The proposed approach is well suited for real-world industrial health factors, characterized by monotonic behavior and observed through irregularly sampled and noisy measurements. The proposed approach provides predictions based on a Particle Filter that employs Monte Carlo simulation to approximate the health factor posterior distribution from the aforementioned data. To account for changes in variability of the health factor, an adaptive filtering scheme, based on a regularization approach, has also been proposed. Furthermore, the definition and generation of a proper risk function associated with the model has been discussed. The proposed approach has been tested on both synthetic data and experimental data coming from a semiconductor manufacturing application. In both cases, the new approach has proved to ensure better performance with respect to those based on Kalman Filtering when applied to the definition of Predictive Maintenance policies. Furthermore, the possibility of calculating the future distribution of the health factor can be used to obtain a quantitative assessment of failure risks.

In Fault Detection and Isolation, model robustness and reliability are crucial issues (Chen & Patton, 2012; Ding, 2008). For this reason, many data-driven approaches, thanks to their simple forms and limited requirements in terms of design and engineering efforts, have become

more and more popular both in industry and academia (Beghi, Brignoli, Cecchinato, Menegazzo, Rampazzo, & Simmini, 2016). One of the main advantages of the approach proposed in this work is that it only relies on the simple model assumption that the Health Factor and its increments are nonnegative. As shown in Section 1, such assumption is common and realistic in most industrial/real-world scenarios.

Beside robustness and reliability, current research in Fault Detection and Isolation and Predictive Maintenance is dedicated to the derivation of multi-component, incipient faults (Ji, He, Shang, & Zhou, 2017), and model-free solutions. In the present work the latter 2 aspects have been addressed, while multi-component problems have not been explored. In fact, in complex, real-world industrial scenarios processes are described by a large number of variables that can be related to a given fault. For this reason, many works on multi-dimensional diagnosis have been presented in the past recent years (Beghi et al., 2016; Li, Qin, Ji, & Zhou, 2010; Ma et al., 2017). However, in terms of prognosis, the focus of the proposed approach, Health Factors can be traced back to a single variable/quantity, identified through experience/domain expertise, the output of multi-dimensional Soft Sensor, or the residual of a Fault Detection & Identification multi-dimensional procedure. One aspect that has not been tackled in this work and may be subject of future research activities is the case of multiple fault problems that may be associated with multiple Health Factors. While the Hidden-Gamma Process-Particle Filter procedure can still be applied in such cases (prognostic for each Health Factor can be run in parallel by different instances of the proposed approach), a jointly risk function has to be considered to implement Predictive Maintenance policies considering multiple faults.

References

Abdel-Hameed, M. (1975). A gamma wear process. *IEEE Transactions on Reliability*, *24*(2), 152–153.

Abdennadher, K., Venet, P., Rojat, G., Rétif, J.-M., & Rosset, C. (2010). A real-time predictive-maintenance system of aluminum electrolytic capacitors used in uninterrupted power supplies. *IEEE Transactions on Industry Applications*, *46*(4), 1644–1652.

Alrowaie, F., Gopaluni, R., & Kwok, K. (2012). Fault detection and isolation in stochastic non-linear state-space models using particle filters. *Control Engineering Practice*, *20*(10), 1016–1032.

Arinton, E., Caraman, S., & Korbicz, J. (2012). Neural networks for modelling and fault detection of the inter-stand strip tension of a cold tandem mill. *Control Engineering Practice*, *20*(7), 684–694.

Arulampalam, M., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.

Bakdi, A., Kouadri, A., & Bensmail, A. (2017). Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme. *Control Engineering Practice*, *66*, 64–75.

Beghi, A., Brignoli, R., Cecchinato, L., Menegazzo, G., Rampazzo, M., & Simmini, F. (2016). Data-driven fault detection and diagnosis for HVAC water chillers. *Control Engineering Practice*, *53*, 79–91.

Benkedjouh, T., Medjaher, K., Zerhouni, N., & Rechak, S. (2015). Health assessment and life prediction of cutting tools based on support vector regression. *Journal of Intelligent Manufacturing*, *26*(2), 213–223.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Bressel, M., Hilairet, M., Hissel, D., & Bouamama, B. O. (2016). Remaining useful life prediction and uncertainty quantification of proton exchange membrane fuel cell under variable load. *IEEE Transactions on Industrial Electronics*, *63*(4), 2569–2577.

Butler, S., & Ringwood, J. (2010). Particle filters for remaining useful life estimation of abatement equipment used in semiconductor manufacturing. In *Control and fault-tolerant systems, 2010 conference on* (pp. 436–441).

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Chen, D.-G. D., Lio, Y., Ng, H. K. T., & Tsai, T.-R. (2017). *Statistical modeling for degradation data*. Springer.

Chen, J., & Patton, R. J. (2012). *Robust model-based fault diagnosis for dynamic systems, Vol. 3*. Springer Science & Business Media.

Chioua, M., Bauer, M., Chen, S.-L., Schlake, J. C., Sand, G., Schmidt, W., et al. (2016). Plant-wide root cause identification using plant key performance indicators (KPIs) with application to a paper machine. *Control Engineering Practice*, *49*, 149–158.

Cinlar, E., Osman, E., & Bazoant, Z. (1977). Stochastic process for extrapolating concrete creep. *Journal of the Engineering Mechanics Division*, *103*(6), 1069–1088.

Dey, S., Biron, Z. A., Tatipamula, S., Das, N., Mohon, S., Ayalew, B., et al. (2016). Model-based real-time thermal fault diagnosis of Lithium-ion batteries. *Control Engineering Practice*, *56*, 37–48.

Ding, S. X. (2008). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media.

Ding, X., Tian, Y., & Yu, Y. (2015). A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations. *IEEE Transactions on Industrial Informatics*, *PP*(99), 1–1. http://dx.doi.org/10.1109/TII.2015.2436337.

Ding, S., Yin, S., Peng, K., Hao, H., & Shen, B. (2013). A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill. *IEEE Transactions on Industrial Informatics*, *9*(4), 2239–2247.

Douc, R., & Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and signal processing and analysis* (pp. 64–69). IEEE.

Doucet, A. (1998). *On sequential simulation-based methods for Bayesian filtering, Tech. rep*. Citeseer.

Doucet, A., DeFreitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer Verlag.

Engel, S. J., Gilmartin, B. J., Bongort, K., & Hess, A. (2000). Prognostics, the real issues involved with predicting life remaining. In *Aerospace conference proceedings, 2000 IEEE, Vol. 6* (pp. 457–469). IEEE.

Filev, D., Chinnam, R., Tseng, F., & Baruah, P. (2010). An industrial strength novelty detection framework for autonomous equipment monitoring and diagnostics. *IEEE Transactions on Industrial Informatics*, *6*(4), 767–779.

Ge, Z., Song, Z., & Gao, F. (2013). Review of recent research on data-based process monitoring. *Industrial and Engineering Chemistry Research*, *52*(10), 3543–3562.

Gorinevsky, D. (2004). Monotonic regression filters for trending deterioration faults. In *American control conference, Vol. 6* (pp. 5394–5399). IEEE.

Hast, D., Findeisen, R., & Streif, S. (2015). Detection and isolation of parametric faults in hydraulic pumps using a set-based approach and quantitative–qualitative fault specifications. *Control Engineering Practice*, *40*, 61–70.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Hu, L., Li, L., & Hu, Q. (2017). Degradation modeling, analysis, and applications on lifetime prediction. In *Statistical modeling for degradation data* (pp. 43–66). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning*. Springer.

Ji, H., He, X., Shang, J., & Zhou, D. (2017). Incipient fault detection with smoothing techniques in statistical process monitoring. *Control Engineering Practice*, *62*, 11–21.

Lawless, J., & Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, *10*(3), 213–227.

LeSon, K., Fouladirad, M., & Barros, A. (2016). Remaining useful lifetime estimation and noisy gamma deterioration process. *Reliability Engineering & System Safety*, *149*, 76–87.

Lewin, D. R. (1995). Predictive maintenance using PCA. *Control Engineering Practice*, *3*(3), 415–421.

Li, T., Bolic, M., & Djuric, P. M. (2015). Resampling methods for particle filtering: classification, implementation, and strategies. *IEEE Signal Processing Magazine*, *32*(3), 70–86.

Li, G., Qin, S. J., Ji, Y., & Zhou, D. (2010). Reconstruction based fault prognosis for continuous processes. *Control Engineering Practice*, *18*(10), 1211–1219.

Liao, L., & Köttig, F. (2016). A hybrid framework combining data-driven and model-based methods for system remaining useful life prediction. *Applied Soft Computing*, *44*, 191–199.

Lu, D., Pandey, M. D., & Xie, W.-C. (2013). An efficient method for the estimation of parameters of stochastic gamma process from noisy degradation measurements. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *227*(4), 425–433.

Lu, S., Tu, Y.-C., & Lu, H. (2007). Predictive condition-based maintenance for continuously deteriorating systems. *Quality and Reliability Engineering International*, *23*(1), 71–81.

Ma, L., Dong, J., Peng, K., & Zhang, K. (2017). A novel data-based quality-related fault diagnosis scheme for fault detection and root cause diagnosis with application to hot strip mill process. *Control Engineering Practice*, *67*, 43–51.

Nguyen, K.-A., Do, P., & Grall, A. (2015). Multi-level predictive maintenance for multi-component systems. *Reliability Engineering & System Safety*, *144*, 83–94.

Pitt, M., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, *94*(446), 590–599.

Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines, Tech. rep*. (p. 21). Microsoft.

Saha, B., Goebel, K., & Christophersen, J. (2009). Comparison of prognostic algorithms for estimating remaining useful life of batteries. *Transactions of the Institute of Measurement and Control*, *31*, 293–308.

Schirru, A., Pampuri, S., & DeNicolao, G. (2010). Particle filtering of hidden gamma processes for robust predictive maintenance in semiconductor manufacturing. In *IEEE conference on automation science and engineering* (pp. 51–56).

Seeger, M. (2009). *Bayesian modelling in machine learning: A tutorial review, Tech. rep*. EPFL.

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation–A review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14.

Sikorska, J., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, *25*(5), 1803–1836.

Souza, F., & Araujo, R. (2014). Online mixture of univariate linear regression models for adaptive soft sensors. *IEEE Transactions on Industrial Informatics*, (ISSN: 1551-3203) *10*(2), 937–945.

Susto, G., Beghi, A., & Luca, C. D. (2012). A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. *Semiconductor Manufacturing, IEEE Transactions on*, *25*(4), 638–649.

Susto, G. A., McLoone, S., Pagano, D., Schirru, A., Pampuri, S., & Beghi, A. (2013). Prediction of integral type failures in semiconductor manufacturing through classification methods. In *Emerging technologies & factory automation, 2013 IEEE 18th conference on* (pp. 1–4). IEEE.

Susto, G., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: a multiple classifiers approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812–820.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

VanDerMerwe, R., Doucet, A., DeFreitas, N., & Wan, E. (2004). The unscented particle filter. In *Computer vision-ECCV 2004* (pp. 28–39). Springer.

Van Overschee, P., & De Moor, B. (2012). *Subspace identification for linear systems: Theory, implementation, applications*. Springer Science & Business Media.

Wahlberg, B., Boyd, S., Annergren, M., & Wang, Y. (2012). An ADMM algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, *45*(16), 83–88.

Wang, X., Cheng, J., Wang, K., Yang, Y., Sun, Y., Cao, M., et al. (2014). Modeling of Electrostatic chuck and simulation of electrostatic force. *Applied Mechanics and Materials*, *511*, 588–594.

Wang, W., Hussin, B., & Jefferis, T. (2012). A case study of condition based maintenance modelling based upon the oil analysis data of marine diesel engines using Stochastic Filtering. *International Journal of Production Economics*, *136*(1), 84–92.

Wang, D., Liu, J., & Srinivasan, R. (2010). Data-driven soft sensor approach for quality prediction in a refining process. *IEEE Transactions on Industrial Informatics*, *6*(1), 11–17.

Wang, T., Yu, J., Siegel, D., & Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *Prognostics and health management, 2008. PHM 2008. International conference on* (pp. 1–6). IEEE.

Xie, Y., Jin, Z., Hong, Y., & Van Mullekom, J. H. (2017). Statistical methods for thermal index estimation based on accelerated destructive degradation test data. In *Statistical Modeling for Degradation Data* (pp. 231–251). Springer.

Xu, Q.-N., Lee, K.-M., Zhou, H., & Yang, H.-Y. (2015). Model-based fault detection and isolation scheme for a rudder servo system. *IEEE Transactions on Industrial Electronics*, *62*(4), 2384–2396.

Yang, S., & Liu, T. (1999). State estimation for predictive maintenance using kalman filter. *Reliability Engineering & System Safety*, *66*(1), 29–39.

You, M.-Y., Li, L., Meng, G., & Ni, J. (2010). Cost-effective updated sequential predictive maintenance policy for continuously monitored degrading systems. *Automation Science and Engineering, IEEE Transactions on*, *7*(2), 257–265.

Zhang, Q., & Canova, M. (2015). Fault detection and isolation of automotive air conditioning systems using first principle models. *Control Engineering Practice*, *43*, 49–58.

Zhou, X., Stein, J. L., & Ersal, T. (2017). Battery state of health monitoring by estimation of the number of cyclable Li-ions. *Control Engineering Practice*, *66*(Suppl. C), 51–63. http://dx.doi.org/10.1016/j.conengprac.2017.05.009.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.