

THE LATENT STATE HAZARD MODEL, WITH APPLICATION TO WIND TURBINE RELIABILITY

BY RAMIN MOGHADDASS AND CYNTHIA RUDIN

Massachusetts Institute of Technology

We present a new model for reliability analysis that is able to distinguish the *latent* internal vulnerability state of the equipment from the vulnerability caused by temporary *external* sources. Consider a wind farm where each turbine is running under the external effects of temperature, wind speed and direction, etc. The turbine might fail because of the external effects of a spike in temperature. If it does not fail during the temperature spike, it could still fail due to internal degradation, and the spike could cause (or be an indication of) this degradation. The ability to identify the underlying latent state can help better understand the effects of external sources and thus lead to more robust decision-making. We present an experimental study using SCADA sensor measurements from wind turbines in Italy.

1. Introduction. One of the most important decisions that many companies face is when to turn off mechanical equipment in order to perform preventive maintenance. Considering wind farm maintenance, for instance, it is much more cost effective to shut a turbine off before it fails than to repair extensive damage caused by a failure. The goal then becomes one of prediction: if we stop the turbine too early before it would have failed, we lose valuable operating time. If we stop it too late, the turbine may have sustained a catastrophic failure that is expensive to repair. While the equipment is operating, its vulnerability to failure depends not only on external factors such as temperature, wind direction and speed, but also on latent degradation due to wear-and-tear. If it can be well estimated, this latent vulnerability state would be important to decision-makers because: (i) it would provide insight into the health state of the equipment without the influence of additional external factors, (ii) it would determine whether the turbine is likely to sustain extreme external conditions such as high temperatures, (iii) it would reveal how the various external factors influence the degradation levels of the equipment, and (iv) it would help provide maintenance decision-makers with a tool that can help prevent too many early or late warnings. Ideally, we would like to decouple the latent vulnerability state from the vulnerability due to external sources, making as few distributional assumptions as possible.

Like many other types of large mechanical equipment (e.g., oil drilling equipment, electrical feeders), wind turbines are usually equipped with supervisory control and data acquisition (SCADA) sensors that record various measurements of

Received September 2014; revised July 2015.

Key words and phrases. Performance monitoring, reliability, maintenance, decision-making, big data.

the dynamic environment every few minutes. For each of the 28 turbines in Italy that we are considering, there are over 30 different measurements collected every 10 minutes, including temperatures inside and outside the turbine, wind speed, measurements of the yaw system, and so on. These can be environmental measurements, measurements that reflect the degradation state, or summary statistics of these measurements. Wind farm operators make critical decisions on a regular basis that depend on condition monitoring; this is because failures occur fairly frequently, once every 2–6 weeks on average. The frequency of these failures makes statistically motivated decision-making strategies very relevant.

Our model is a generalization of the Cox proportional hazard model (PHM) in that there are two separate terms in the hazard function, the latent hazard μ and the transient hazard g . The main aspects of the model can be summarized as follows:

(1) *Latent State (Degradation) Term μ* : This term is monotonically nondecreasing over time, reflecting the fact that mechanical equipment like a turbine does not self-heal. This term takes into account the full history of the turbine as an integral of the degradation. In contrast, to include the full history in the regular Cox PHM, one would need to include a large number of terms and constrain their influence to prevent the appearance of the nonphysical self-healing.

(2) *Transient Hazard Term g* : The second element of the hazard function (g) reflects the instantaneous contribution to vulnerability due to current measurements. For instance, a spike in temperature would be reflected as a spike in the general g term and as a permanent rise in the μ term.

(3) *Form of the Model*: Our model generalizes the Cox proportional hazard model, which uses only the g term. Our model is a mixed hazard model [see, e.g., Lin and Ying (1995)] whose form permits data-driven parameter fitting using convex optimization techniques.

(4) *Use in Decision-Making for Maintenance*: Because we can isolate the latent state, we can better estimate the resilience of our equipment: the latent state would not be as sensitive to sudden but normal changes in the covariates. By triggering warnings using knowledge of the latent state, decision-makers may be able to issue better-timed warnings and alarms. We also provide an optimization procedure to assist with decision-making.

The paper is organized as follows: Section 2 includes the motivation for this work and reviews the literature on using analytical techniques for covariate-dependent degradation models, particularly for wind turbines using SCADA data. Section 3 describes the model and its derivation, along with an optimization procedure for making maintenance decisions. Section 4 discusses important properties and inference. Section 5 shows the result of applying the proposed model and the warning generation technique on turbines in Italy. Section 6 provides a set of numerical experiments, including a simulation study and comparisons with previous models. In the Supplementary Materials Moghaddass and Rudin (2015), we provide an interpretation of our model and motivation with respect to discrete multi-state degradation models.

2. Related work. The literature of degradation monitoring and failure analysis using time-varying covariates can be divided into three main categories, namely, (1) degradation-only modeling, (2) hazard modeling with diagnostic covariates, and (3) degradation modeling with partial information. Our work is related to all three categories.

Degradation-only models assume that the covariates are noisy versions of the degradation state itself [e.g., [Bian and Gebraeel \(2012, 2013\)](#), [Gebraeel and Pan \(2008\)](#), [Flory, Kharoufeh and Gebraeel \(2014\)](#), [Kharoufeh \(2003\)](#), [Kharoufeh and Cox \(2005\)](#), [Zhou, Serban and Gebraeel \(2011\)](#)]. In this way, one can assume, for instance, that each covariate could be generated separately from the degradation state plus random noise. Some of these studies have assumed that a failure occurs precisely when the degradation signal exceeds a predefined threshold [e.g., [Bian and Gebraeel \(2012, 2013\)](#), [Gebraeel and Pan \(2008\)](#)]. For example, [Bian and Gebraeel \(2013\)](#) presented a stochastic modeling framework for sensor-based degradation signals for systems operating under a time-varying environment. They assumed that the rate of degradation directly depends on the environment profile that is known, deterministic and evolves continuously. The overall degradation signal is defined as the sum of the effect of environmental factors and a time-dependent Brownian motion process.

For the case of wind turbines, it is not realistic to assume that we simply have noisy measurements of the degradation state. Our measurements all stem from (a possibly complicated) combination of external sources of vulnerability and the degradation state itself (e.g., temperature within the turbine); it is our job to separate these two sources, and we cannot assume that the signals are directly correlated with the underlying physical degradation processes. In our case, the degradation process is assumed to be unobservable, and there are no prior distributional assumptions on the parameters of the model. We also do not require a predefined failure threshold, which is often not available in real-world systems.

Works within the second group (hazard modeling with diagnostic covariates) assume that the hazard rate is influenced by internal and/or external time-varying covariates and aim to estimate the hazard rate. The Cox proportional hazard model (PHM) [[Cox \(1972\)](#)], its time-dependent version [[Fisher and Lin \(1999\)](#)] and its other extensions [see, e.g., [Gorjian et al. \(2009\)](#)] are examples of this second group. Works in the second group primarily aim to predict the hazard rate, and do not necessarily model degradation. Those that do model degradation generally take the perspective of the first category, where some of the signals are known to be noisy versions of the degradation state [e.g., [Banjevic and Jardine \(2006\)](#), [Banjevic et al. \(2001\)](#), [Jardine, Anderson and Mann \(1987\)](#), [Makis and Jardine \(1991\)](#), [Qian and Wu \(2014\)](#), [Wu and Ryan \(2011\)](#), [Zhao et al. \(2010\)](#)].

Some works consider Markov failure time processes [e.g., [Banjevic and Jardine \(2006\)](#), [Banjevic et al. \(2001\)](#)] where a predetermined subset of the covariates changes over time. Finite state space models are useful in that the vulnerability

states are finite and meaningful, but the assumption of a finite state space is restrictive and not particularly realistic. At the same time, relaxing the assumptions of finite state space and considering multiple covariates yield too many states and transitions, making it difficult to estimate all of the transition probabilities.

The model in this work, by contrast, does not require the number of states to be known a priori, and the unobservable vulnerability state is modeled as a function of past measurements. Furthermore, the model in this work does not prespecify which variables contribute to the degradation state, allowing this to be learned from data.

Our model is a particular time-dependent, additive-multiplicative mixed hazard model (AMMHM), containing both additive and multiplicative terms. There are examples of additive hazard models [e.g., Pijnenburg (1991)] and multiplicative hazard models [e.g., Kalbfleisch and Prentice (2002)] in different application domains. A review of hazard models with covariates, with an explanation of AMMHM, is given by Gorjian et al. (2009). Few studies have developed special-purpose mixed hazard models. Martinussen and Scheike (2002) proposed an additive-multiplicative model consisting of two components. The first component contains additive covariates through an additive Aalen model and the second component contains a multiplicative covariate effect through a Cox regression model. Two different feature sets were used to separately model baseline mortality and dose effects. Andersen and Vaeth (1989) considered a mixed model with additive and multiplicative terms, where one term is proportional to a known population mortality. They illustrated their model by predicting survival of medical patients after an operation for malignant melanoma. Another example of a mixed hazard model was proposed by Höhle (2009) for spatial and temporal infectious disease surveillance. For a theoretical analysis of AMMHM, interested readers may refer to the work of Lin and Ying (1995). Our model leverages the multiplicative and additive terms for a specific purpose: to separate the latent hazard state from the external risk factors. Our particular multiplicative term (μ) acts as a concise representation of the full history of covariates. It is unclear how one would encode the full history of covariates within, for instance, the Cox proportional hazards model, without introducing a large number of variables.

The third group of related models are partially observable degradation models. These models differ from the other two types in that the degradation process is assumed to be unobservable (hidden). Instead, covariates with an indirect relationship to the degradation process are monitored over time. These are also referred to as degradation processes with incomplete information [Hontelez, Burger and Wijnmalen (1996)] or partially-observed degradation processes [Ghasemi, Yacout and Ouali (2007)]. Most works in this category have used hidden Markov and hidden semi-Markov models with discrete unobservable degradation states [Moghaddass and Zuo (2012), Peng and Dong (2011)]. Our model is similar to those in this category in that it also provides insight into the latent degradation state. Our model is different from those in this category in that we do not need to specify the number of states, the transition probability distribution between states or the structure of

the stochastic relationship between the covariates and the degradation process. The covariates used in our work, by contrast, are not necessarily assumed to reflect the hidden degradation level.

There are other types of degradation models that make particular generative assumptions for specific types of environmental processes. For example, [Kharoufeh, Finkelstein and Mixon \(2006\)](#) considered a single-unit degrading system affected by its operating environment with a deterministic degradation threshold value. They considered a random shock process where the rate of wear is modulated by a discrete-space, continuous-time Markov chain, and additional damage is induced by a Poisson shock process. The total degradation is assumed to be the sum of the degradation due to wear and that due to shocks. For other types of degradation models used in reliability modeling, interested readers may refer to the works of [Gorjian et al. \(2009\)](#) and [Si et al. \(2011\)](#).

For wind farms, keeping maintenance costs low is essential; it is difficult to be competitive against the costs of other energy sources, such as oil and gas. There are several reviews on condition monitoring and fault detection at wind farms [e.g., [Hameed et al. \(2009\)](#), [Kusiak, Zhang and Verma \(2013\)](#), [Lu et al. \(2009\)](#), [Marquez et al. \(2012\)](#)]. In the wind industry, SCADA (supervisory control and data acquisitions) systems are the most commonly used mechanism for turbine health monitoring [[Marquez et al. \(2012\)](#)]. Although SCADA systems are relatively inexpensive to install and are used at almost all wind farms, relatively very little research effort has been devoted to analytics using wind turbine SCADA measurements. There are several potential reasons for this. For instance, one is that SCADA sampling frequency is too low to be used for spectral analysis, and another is that it does not collect all the information needed for full condition monitoring of any particular wind turbine component. On the other hand, since SCADA measurements do provide ample and cheap indirect information about the health state of the turbine, some research has begun to determine how to leverage these measurements for health monitoring. [Qiu et al. \(2012\)](#) proposed an alarm analysis and prioritization methodology using descriptive statistics of SCADA data. A method for processing SCADA data and a condition monitoring technique were developed by [Yang, Court and Jiang \(2013\)](#) using a regression approach to anomaly detection. Several papers [e.g., [Marvuglia and Messineo \(2012\)](#), [Zaher et al. \(2009\)](#)] proposed using machine learning techniques for condition monitoring using SCADA data. [Guo et al. \(2009\)](#) proposed a time-dependent reliability analysis based on the three-parameter Weibull distribution for wind turbine failure time data. A recent review of challenges for wind turbine maintenance was provided by [Yang et al. \(2014\)](#).

To summarize, the benefits of our model beyond those of previous works are that (i) it decouples the (unobserved) degradation state from the hazard due to transient sources, without having to specify anything about the relationship of the features to the degradation state, (ii) it takes the full history of measurements into account in a concise way, which cannot be done easily in a Cox proportional hazard model without including a large number of terms, (iii) it provides a new decision-making methodology through latent state inference.

3. The latent state hazard model. Although this work focuses on turbine modeling, our approach can be applied to any type of nonself-healing degrading system. Our notation is as follows:

N : Total number of units (turbines).

P : Number of features (SCADA measurements) used for monitoring.

T_i : The total lifetime of the i th unit.

Δ : Measurement interval.

$x_{i,k}(j)$: The value of the k th covariate at time $j\Delta$ for the i th unit.

$\mathbf{x}_i(j) = [x_{i,1}(j), x_{i,2}(j), \dots, x_{i,P}(j)]^\top$: Feature measurements at time $j\Delta$ for the i th unit.

$\mathbf{x}_i^h(j) = [\mathbf{x}_i(1), \mathbf{x}_i(2), \dots, \mathbf{x}_i(j)]$: History of features up to time $j\Delta$ for the i th unit.

We will assume, mainly for notational convenience, that hazard rates are constant over each small unit of time Δ and can be presented as piecewise constant functions of time. The hazard rate in each interval can thus be approximated by the hazard rate at the endpoint of that interval. Given the full history of covariate values $\mathbf{x}_i^h(t)$, the notation for the hazard rate is $\lambda(t|\mathbf{x}_i^h(t))$, which we model with two terms as follows:

$$(1) \quad \lambda(t|\mathbf{x}_i^h(t)) = \mu(t|\mathbf{x}_i^h(t)) + g(t|\mathbf{x}_i(t)),$$

where, assuming t is a multiple of Δ , we set

$$(2) \quad \begin{aligned} \mu(t|\mathbf{x}_i^h(t)) &= \int_0^t \mu_0(\tau) \exp(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i(\tau)) d\tau \\ &\approx \sum_{l \in \{\Delta, 2\Delta, 3\Delta, \dots, t\}} \mu_0(l) \exp(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i(l)) \Delta \\ &= \mu(t - \Delta|\mathbf{x}_i^h(t - \Delta)) + \mu_0(t) \exp(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i(t)) \Delta. \end{aligned}$$

The μ term is the latent state term. Because it is an integral of exponentials, it is monotonically nondecreasing in t . The approximation in the second line of equation (2) shows the discrete form of the integral (which we use in practice since our measurements are taken at discrete times), and the third line in equation (2) shows that it can be written as a recursion. The other term of λ is the g term, referred to in this paper as the transient hazard term, which is not necessarily monotonically increasing. The mathematical expression of the g term is

$$(3) \quad g(t|\mathbf{x}_i(t)) = g_0(t) \exp(\alpha_0 + \boldsymbol{\alpha}^\top \mathbf{x}_i(t)).$$

In our notation, the feature vector $\mathbf{x}_i(t)$ depends on time t , but $\mathbf{x}_i^h(t)$ could have components that are feature values from the previous times, or nonlinear transformations of measurements taken either at the current time or in the past. The coefficients β_0 and α_0 are intercept terms and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_P\}^\top$ and

$\alpha = \{\alpha_1, \dots, \alpha_P\}^\top$ are each the vector of regression coefficients associated with P features in μ and g , respectively. The functions $\mu_0(t)$ and $g_0(t)$ are, respectively, the baseline hazard functions associated with μ and g . We note here that β_0 and α_0 could be absorbed into the baseline hazard functions. Similarly, $\mu_0(t)$ and $g_0(t)$ could be absorbed into the exponential term. The function μ depends on the full history of covariates and encodes the latent hazard state. We conjecture that μ is often smooth, but will include jumps in the presence of fast-changing external factors to the degradation state. The g term should similarly fluctuate as a function of current external conditions (e.g., temperatures and wind speeds). It is possible to make the transient hazard term g depend on μ by adding terms within g related to μ . The main parameters of the model are the intercepts α_0 and β_0 , and the α and β vectors, each of size P , which encode the history of the degradation process and the temporary influences of the covariates.

In expressions (2) and (3) the covariates, such as temperature, pressure, etc., are not assumed to follow a certain distribution or well-structured time series. Figure 1 shows an example of the total hazard rate, latent vulnerability rate μ and transient vulnerability g , that we estimated from two of the wind turbines. For the first turbine shown in Figure 1(a)–(c), the total hazard rate is formed from a fairly balanced mix of both degradation and temporary sources. For the second turbine shown in Figure 1(d)–(f), the total hazard rate comes mainly from degradation, and there is minimal contribution from transient sources. The model parameters used for both turbines are the same, and were learned from a separate training set that did not include either of the turbines.

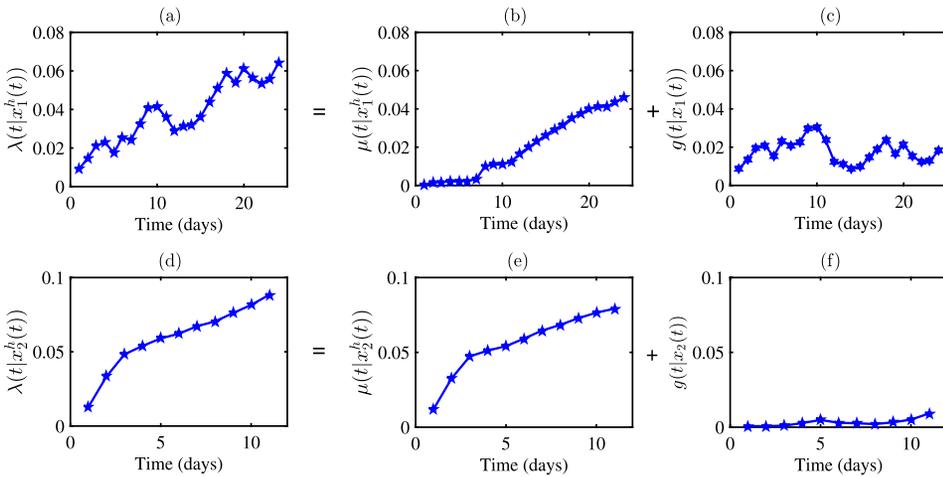


FIG. 1. Estimated hazard rates (λ , μ and g in columns 1–3, resp.) for turbine No. 1 (1st row) and turbine No. 2 (2nd row), decomposed into the latent degradation (μ) and the transient vulnerability (g), using the model presented in this study. (a) Turbine No. 1 (λ), (b) turbine No. 1 (μ), (c) turbine No. 1 (g), (d) turbine No. 2 (λ), (e) turbine No. 2 (μ), (f) turbine No. 2 (g).

3.1. *Cost-benefit analysis and decision-making.* We propose to use our model for generation of warnings. It is important that our warning generation method is accurate. For instance, alarm and warning rates that are too conservative can not only increase downtime and decrease productivity, but they can also reduce the operators’ sensitivity to failure, which can have catastrophic consequences. We can define “warning generation” as a decision process that depends on the estimated hazard of each unit over time. Let us define d as the ideal lead time between the warning point and the failure point determined by decision-makers. In other words, the warning generation system is considered efficient if it generates warnings when the actual time to failure (also called remaining useful life—RUL) is very close to d time units. To find the optimal warning policy, we can define γ_d as the threshold for warning generation, so that a warning is put in place as soon as the estimated hazard exceeds this threshold. To define the quality of the decision framework, we define a cost function $C_d(\xi)$, $\xi \geq 0$, to represent the cost of a warning at ξ units before the actual failure time. We should note that $C_d(d) = 0$, and there is a positive cost for warnings that are too early or too late. The cost of an actual failure without warning in advance, that is, when the warning time equals the failure time, is $C_d(0)$. We refer to this cost as the cost of warning at failure. Our model is general in the sense that any kind of cost function can be considered depending on the application (e.g., hinge, quadratic, logistic, exponential, etc.). A very simple form of this function is shown in Figure 2 and is typically called the “pinball loss” or “newsvendor cost” [see, e.g., Rudin and Vahn (2014)].

We define R_{i,γ_d} as the model’s warning time for the i th available lifetime based on the lead time d and threshold γ_d . For any d , one can determine the optimal policy γ_d^* such that the expected average cost of the warning generation process is

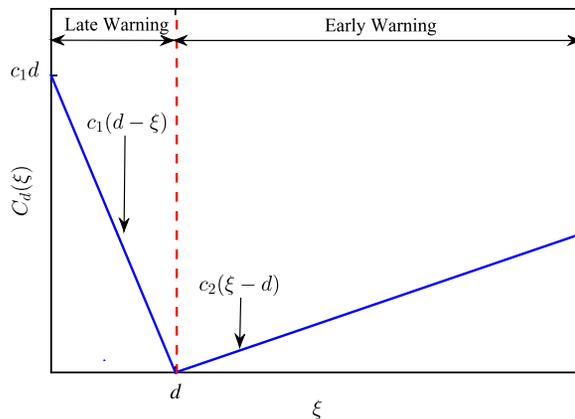


FIG. 2. An example of the function $C_d(\xi)$, as explained in the text. The cost of warning at exactly d units before failure is zero, and there is a positive cost for warnings that are too early (i.e., for $\xi > d$) or too late (i.e., for $\xi < d$). The unit costs of late warning and early warning are c_1 and c_2 , respectively.

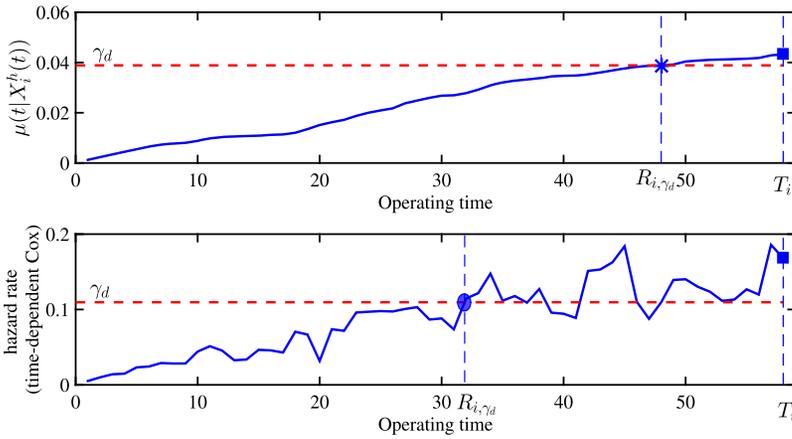


FIG. 3. Hazard rates (solid lines), decision thresholds (γ_d , dashed lines), warning times (R_{i,γ_d}) and failure times (T_i) for a single lifecycle of a turbine. The hazard rates have been estimated using (top) the μ term from the latent state hazard model and (bottom) the hazard rate from the Cox proportional hazard model. The decisions based on the latent state tend to be more robust to fluctuations earlier on in the timeline, possibly leading to more accurate lifecycle estimates.

minimized. We find the optimal solution γ_d^* using an empirical risk model (ERM). The optimization problem is

$$(4) \quad \underset{\gamma_d}{\operatorname{argmin}} J_d(\gamma_d) \quad \text{where } J_d(\gamma_d) = \frac{1}{|N_1|} \sum_{i \in N_1, T_i \geq d} C_d(T_i - R_{i,\gamma_d}),$$

where $R_{i,\gamma_d} = \min\{T_i, \inf\{j | \hat{\lambda}(j | \mathbf{x}_i^h(j)) \geq \gamma_d, j \geq 0\}\}, i \in N_1, \gamma_d \geq 0,$

where N_1 are the lifetimes used for training, $\inf\{j | \hat{\lambda}(j | \mathbf{x}_i^h(j)) \geq \gamma_d, j \geq 0\}$ is the time at which the estimated hazard rate ($\hat{\lambda}$) of the i th unit exceeds the threshold, and γ_d is the only decision variable. The warning time R_{i,γ_d} is either the time at which the unit fails (T_i) or the time at which its estimated hazard rate exceeds the threshold γ_d , whichever occurs first.

Figure 3 shows how decisions can be made using the latent state μ for our model, in contrast with how decisions would be made using the (more common) time-dependent proportional hazard model, where hazard rate is employed as the decision criteria. This figure also illustrates why decisions made using the latent degradation state μ tend to be more robust; they are resilient to fluctuations in the hazard rate earlier in the lifecycle of the turbine. This can lead to longer (more accurate) lifecycles and thus lower costs.

4. Properties and inference. In this section we discuss how to fit the model to data.

4.1. *Model training.* We use the method of maximum likelihood to infer the coefficients of the model. Given lifetimes with stopping times T_i for each turbine i , the continuous time version of the likelihood is

$$(5) \quad \mathcal{L}_N(\boldsymbol{\theta}|X) = \prod_{i=1}^N \left(\exp\left(-\int_0^{T_i} \lambda(t|\mathbf{x}_i^h(t)) dt\right) \right) \times \lambda(T_i|\mathbf{x}_i^h(T_i)),$$

where λ depends on $\boldsymbol{\theta} = \{\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta}\}$ through the definition of the model in equations (1), (2) and (3). If the lifetime associated with unit r ($r \in \{1, 2, \dots, N\}$) is censored at time S_r ($S_r < T_r$), then its contribution to the likelihood function is $\exp(-\int_0^{S_r} \lambda(t|\mathbf{x}_r^h(t)) dt)$. From now on, we assume all lifetimes are complete. For discretized time units, the likelihood has the following form:

$$(6) \quad \mathcal{L}_N(\boldsymbol{\theta}|X) = \underbrace{\prod_{i=1}^N \prod_{j=1}^{T_i-1} \exp(-\lambda(j|\mathbf{x}_i^h(j))\Delta)}_{\text{probability of survival}} \times \underbrace{(1 - \exp(-\lambda(T_i|\mathbf{x}_i^h(T_i))\Delta))}_{\text{probability of failure}},$$

where Δ is the length of each measurement interval. For mathematical convenience, we let $\Delta = 1$ and also define an indicator variable $y_{i,j}$ as

$$(7) \quad y_{i,j} = \begin{cases} 1, & j = T_i, \\ 0, & j \neq T_i. \end{cases}$$

Then the log-likelihood function may be written as follows:

$$(8) \quad \begin{aligned} \log \mathcal{L}_N(\boldsymbol{\theta}|X) &= \sum_{i=1}^N \sum_{j=1}^{T_i-1} -\lambda(j|\mathbf{x}_i^h(j)) + \sum_{i=1}^N \log(1 - \exp(-\lambda(T_i|\mathbf{x}_i^h(T_i)))) \\ &= \sum_{i=1}^N \sum_{j=1}^{T_i} \log(y_{i,j} + (1 - 2y_{i,j}) \exp(-\lambda(j|\mathbf{x}_i^h(j)))) \end{aligned}$$

where $\boldsymbol{\theta}$ is the set of parameters of the model. It is now clear that we have

$$\hat{\boldsymbol{\theta}} \in \arg \max_{\boldsymbol{\theta}} \log \mathcal{L}_N(\boldsymbol{\theta}|X),$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameters. We will show later that, under some regularity conditions, $\hat{\boldsymbol{\theta}}$ converges to the true parameter set $\boldsymbol{\theta}^0$ in probability when N is large. The procedure outlined above will produce a point estimate for $\boldsymbol{\theta}$. (If full Bayesian inference is desired, we could assume a normal prior on $\boldsymbol{\theta}$ and sample from the posterior distribution over $\boldsymbol{\theta}$. However, this would lose the interpretability of the single point estimate, be far less computationally tractable, and the mechanism for making decisions using the full posterior would likely require us to choose a point estimate anyway.)

4.2. *Regularization.* We use ℓ_2 regularization or, equivalently, a normal prior on model parameters α and β . Regularization helps prevent overfitting, and makes the log of the objective function strictly convex. In particular, we optimize

$$(9) \quad \begin{aligned} W_N(\theta|X) &= -\log \mathcal{L}_N(\theta|X) + \mathbf{C}\|\theta\|_2^2 \\ &= -\log \mathcal{L}_N(\theta|X) + C_1\|\alpha\|_2^2 + C_2\|\beta\|_2^2. \end{aligned}$$

Setting C_1 very large will cause the model to ignore the internal state. Similarly, setting C_2 very large will result in ignoring the transient hazard term g . In practice, we set C_1 and C_2 using cross-validation; however, they can be set manually, to force more weight to internal degradation or vice versa. As usual, the regularization constants should effectively vanish as N tends to infinity.

4.3. *Convexity of the loss function.*

PROPOSITION 1. *The loss function $-\log \mathcal{L}_N(\{\alpha, \beta\}|X) + C_1\|\alpha\|_2^2 + C_2\|\beta\|_2^2$ derived from equation (9) is strictly convex when $C_1 > 0, C_2 > 0$.*

The proof is given in Appendix A.

4.4. *Coordinate descent method for model training.* Since the optimization problem is convex and differentiable, coordinate descent is a natural fit. The direction is provided by a Fréchet (directional) derivative. Denoting $C_1\|\alpha\|_2^2 + C_2\|\beta\|_2^2$ by $\mathbf{C}\|\theta\|_2^2$, we have

$$(10) \quad \begin{aligned} &\frac{\partial}{\partial \theta_k} [-\log \mathcal{L}_N(\theta|X) + \mathbf{C}\|\theta\|_2^2] \\ &= \sum_{i=1}^N \sum_{j=1}^{T_i} (1 - 2y_{i,j}) \frac{\exp(-\lambda(j|\mathbf{x}_i^h(j)))}{y_{i,j} + (1 - 2y_{i,j}) \exp(-\lambda(j|\mathbf{x}_i^h(j)))} \frac{\partial \lambda(j|\mathbf{x}_i^h(j))}{\partial \theta_k} \\ &\quad + \frac{\partial}{\partial \theta_k} \mathbf{C}\|\theta\|_2^2, \end{aligned}$$

where θ_k is the k th parameter of θ , and

$$(11) \quad \frac{\partial}{\partial \theta_k} \lambda(j|\mathbf{x}_i^h(j)) = \frac{\partial}{\partial \theta_k} \mu(j|\mathbf{x}_i^h(j)) + \frac{\partial}{\partial \theta_k} g(j|\mathbf{x}_i(j)).$$

Now, for the coefficients of β (denoting $\theta_k = \beta_{k_1}$), we have

$$(12) \quad \frac{\partial}{\partial \beta_{k_1}} \lambda(j|\mathbf{x}_i^h(j)) = \sum_{l=1}^j x_{i,k_1}(l) \exp(\beta_0 + \beta^\top \mathbf{x}_i(l)).$$

Similarly, for the coefficients of α (denoting $\theta_k = \alpha_{k_2}$), we have

$$(13) \quad \frac{\partial}{\partial \alpha_{k_2}} \lambda(j|\mathbf{x}_i^h(j)) = x_{i,k_2}(j) \exp(\alpha_0 + \alpha^\top \mathbf{x}_i(j)).$$

Algorithm 1 Coordinate Descent Algorithm

Let $W_N(\boldsymbol{\theta}|X) = -\log \mathcal{L}_N(\boldsymbol{\theta}|X) + \mathbf{C}\|\boldsymbol{\theta}\|_2^2$ as in equation (9), then:

1. Select the starting point $\boldsymbol{\theta}^{(1)}$ and the convergence parameter ε and let $k = 1$.
2. Compute $\frac{\partial}{\partial \theta_i}[W_N(\boldsymbol{\theta}^{(k)}|X)]$ for all i from equation (10).
3. Choose $i_k \in \arg \max_i |\frac{\partial}{\partial \theta_i}[W_N(\boldsymbol{\theta}^{(k)}|X)]|$.
4. Find the positive step size (φ_k) as

$$(14) \quad \varphi_k \in \arg \min_x W_N\left(\boldsymbol{\theta}^{(k)} - x \frac{\partial}{\partial \theta_{i_k}}[W_N(\boldsymbol{\theta}^{(k)}|X)]e_{i_k}|X\right).$$

Here, e_i is the i th coordinate vector in \mathbb{R}^{2P+2} .

5. Update the current point as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \varphi_k \frac{\partial}{\partial \theta_{i_k}}[W_N(\boldsymbol{\theta}^{(k)}|X)]e_{i_k}.$$

6. Evaluate $W_N(\boldsymbol{\theta}^{(k+1)}|X)$. If the condition $|W_N(\boldsymbol{\theta}^{(k+1)}|X) - W_N(\boldsymbol{\theta}^{(k)}|X)| < \varepsilon$ is satisfied, then terminate the algorithm and output $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(k+1)}$. Otherwise, set $k = k + 1$ and return to step 2.

In Algorithm 1, the steps for optimizing the loss function using the Coordinate Descent method are described.

4.5. *Asymptotic properties.* In this section we state a result (whose proof is in Appendix B) about the consistency properties of the maximum likelihood estimators of the parameters of our model.

THEOREM 1. *Let $X = (X_i, T_i), 1 \leq i \leq N$, be i.i.d. with the likelihood function $\mathcal{L}_N(\boldsymbol{\theta}|X)$ given in equation (6) with independent parameter set $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ where $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\theta} \in \Theta, |\boldsymbol{\alpha}| < M$, and $|\boldsymbol{\beta}| < M$, where M is a finite positive constant independent of N . Then with probability tending to 1 as N tends to infinity, there exist solutions $\hat{\boldsymbol{\theta}}_N = (\hat{\theta}_{N,1}, \dots, \hat{\theta}_{N,p})$ of the likelihood equations such that:*

- (i) $\hat{\theta}_{N,j}$ is consistent for estimating θ_j ,
- (ii) $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0)$ is asymptotically normal with (vector) mean zero and covariance matrix $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$, and
- (iii) $\hat{\theta}_{N,j}$ is asymptotically efficient in the sense that its variance attains the Cramér–Rao lower bound as N goes to infinity.

The proof is a consequence of Theorems 2 and 3, which are given in Appendix B.

5. Application to wind turbine data. Our collaboration is with Accenture (a consulting company) and ENEL (Italy's largest power company). Our data are from a wind farm in Europe that collects SCADA data from $N = 28$ (with N defined as in Section 3) pitch-regulated 2 mega-watt wind turbines. The original measurements used in this work were collected every 10 minutes over the course of a year, and values were averaged over the course of a day. Averaging smooths out small variations and is much more computationally efficient. However, too much averaging can potentially remove the effect of g . In order to make sure that we prevent this from happening, we repeated some of our experiments on smaller portions of our data, averaging over 12-hour intervals and 6-hour intervals. The estimates for μ and g for these experiments were similar to those for the 24-hour interval, implying that our choice of one day was reasonable.

The original data consist of $P = 49$ (with P defined as in Section 3) different signals including internal and external covariates. External covariates are those generated by an independent process that can influence (accelerate or decelerate) the degradation state, such as environmental covariates (humidity, wind speed, external temperature, time, etc.) and operational covariates (load). Also, external covariates include the nontime-varying covariates, such as turbine age in years, location and manufacturer. Internal covariates are those relevant to estimating the degradation process. Examples are oil temperature, gear box temperature and voltage. There were many missing observations in the data, resulting from the SCADA sensors being turned off or other SCADA malfunctions. We chose 95 lifetimes that had few missing points to minimize the bias of the SCADA malfunctions, though since we averaged data over each day, even a few hours of missing points would not change our results very much. A "lifetime" is a time interval between the time point at which the turbine is restarted after a work order ticket and when it fails. Thus, there are possibly multiple lifetimes for each turbine. This implicitly models the effect of maintenance as returning the turbine's condition to a similar "restored" condition at the beginning of each lifetime, though in reality we cannot always know the state of the turbine exactly after a particular type of maintenance was performed. We have assumed that these lifetimes are independent; if one has more information about the state of the system after a repair, our model can be modified to consider an initial degradation level [see [Moghaddass and Rudin \(2015\)](#)]. Figure 4 displays the histogram for 95 lifetimes associated with the 28 turbines (also shown are the mean and the standard deviation of the 95 lifetimes).

In addition to the modeling efforts discussed in this work, a lot of the effort for this project went into the derivation of appropriate features and labels. In particular, our SCADA data do not always include an automated failure "flag" that indicates what part of the turbine failed or whether the turbine had been shut off for reasons other than failure (e.g., inspection). We used a separate database of "work orders" written by the wind farm company to help us determine whether the turbine had been shut off due to failure; however, those data were not sufficient to differentiate reasons for the shut-offs or identify the particular parts of the turbines that

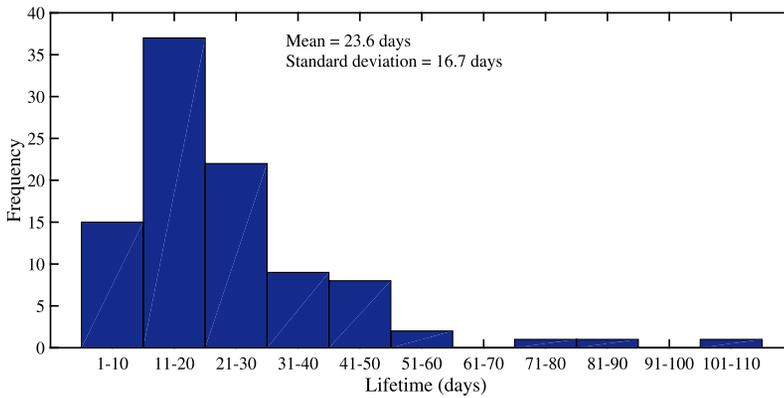


FIG. 4. Histogram for the 95 lifetimes associated with the 28 turbines.

failed. As a result, we chose to predict whether any failure mode will occur, and thus categorized the work orders into those that represented failures and those that represented nonfailures. Table 1 provides the list of covariates we used and which part of the turbine was measured to obtain each covariate. Since these covariates are completely typical of data collected by SCADA systems for wind turbines, we believe our approach could be widely applied by the wind industry.

In Figure 5, the SCADA measurements of 14 covariates for a single turbine are plotted for a period of three months. The dotted lines in each plot represent work order events (maintenance actions). No obvious trend can be discerned from any individual signal. We normalized each signal to be between zero and one, where the minimum and maximum used in the normalization were calculated over the training set and the same values were used for the test set. This helps with numerical stability and improves convergence speed for parameter estimation.

In addition to the features above, we added another class of features that compares the signal values of each turbine to the other turbines in the wind farm. For example, if the power output of one turbine is much lower than the average gener-

TABLE 1

Covariates selected for turbine health monitoring and the part of the turbine used to measure them

Feature name	Location	Feature name	Location
Pitch average	Rotor	Bus voltage	Generator
Hydraulic pressure	Hydraulic	Temperature transformer max	Generator
Nacelle temperature	Nacelle	Temperature radiator 2	Generator
Ambient temperature	Ambient	Temperature drive end bearing	Gearbox
Active power	Counter	Temperature nondrive end bearing	Gearbox
Power loss	Counter	Temperature gear oil	Gearbox
Temperature generator windings avg	Generator	Generator speed	Generator

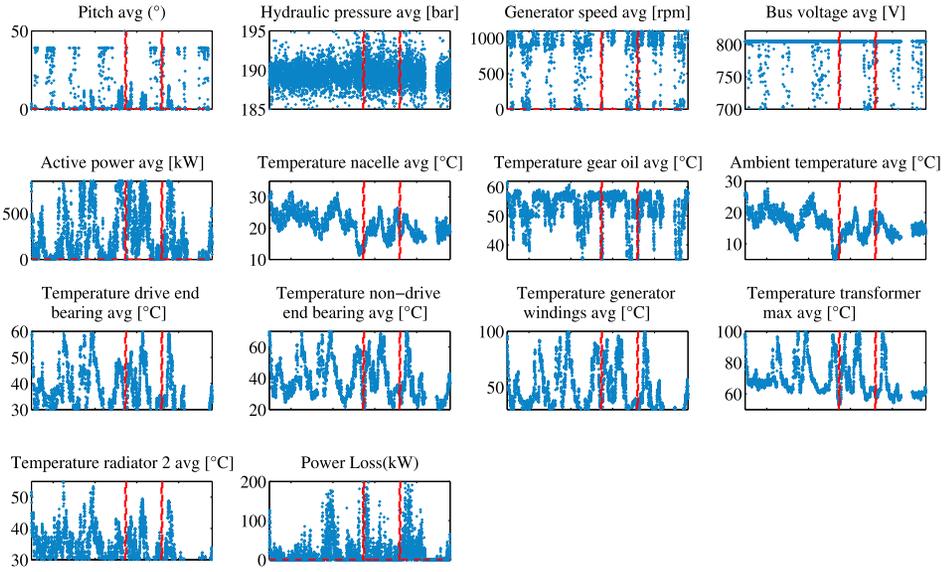


FIG. 5. Original feature values displayed for a period of 3 months for a sample turbine. The dotted vertical lines in each plot represent work order events (maintenance actions).

ated power of the other turbines, it could be an indication of a mechanical problem and a precursor to failure. One class of features (denoted by M_1) is the differences in percentile value of each signal from the median among other turbines within the wind farm. That is, we subtracted 0.5 from the normalized rank and took the absolute value to compute differences from the median. The second class of features (denoted by M_2) is the z -scores of the signal values. The formulas for these features are given below:

$$M_1(i, j, t) = \left| \frac{\sum_{j'} \mathbf{1}_{\{x_{i,j}(t) < x_{i,j'}(t)\}}}{n(t)} - 0.5 \right|,$$

$$M_2(i, j, t) = \left| \frac{x_{i,j}(t) - \bar{x}_{:,j}(t)}{\text{sd}(x_{:,j})(t)} \right|, \quad \forall(i, j, t),$$

where $n(t)$ is the total number of turbines with nonnull measurements at time t , $x_{i,j}(t)$ is the j th signal value associated with the i th lifetime at time t , and $\bar{x}_{:,j}(t)$ and $\text{sd}(x_{:,j})(t)$ are the average and the standard deviation of signal j collected from all available turbines at time t , respectively.

We used cross-validation to evaluate performance. We randomly divided our data set into five folds of equal size (19 lifetimes per fold). These sets are referred to as Test Set 1 (TS1) through Test Set 5 (TS5) hereafter. We trained the model with 4 folds and used the last fold for testing. This process was repeated 5 times so that all folds were used for testing. We used Cox–Snell residuals [see, for more details, Collett (2003)] to check if the estimated hazard functions model the set

of turbines’ lifetimes adequately. If the model fits the data well, the Cox–Snell residuals should approximately follow a unit exponential distribution. We used the Kolmogorov–Smirnov (K–S) test on each test fold to compare the estimated cumulative hazard function (Cox–Snell residuals) with the exponential distribution with mean 1. Since all calculated p -values of the associated K–S test for the 5 folds are large (i.e., 0.14, 0.26, 0.30, 0.31, 0.19), we do not reject the hypothesis that the model fits the data well. We should point out that the power of the test increases rapidly away from a unit exponential distribution with 19 lifetimes, which indicates that the K–S test results are reasonable.

5.1. *Interpretability.* In Figure 6, the estimated hazard function and its separation into μ and g are shown for eight lifetimes in the first test set. The model was trained on 76 lifetimes, using 0.1 for both regularization constants; the regularization constant was found via a validation set of 19 training lifetimes. The interesting feature of Figure 6 is that it shows a clear separation between internal and external effects, where some of the lifetimes are driven mainly by the μ term, others by the g term, and some by both terms. Some of these effects may be attributed to regularization, but not all. This means that some of the lifetimes may have been more robust than others to external factors. The increased robustness could be because the equipment is in a better mechanical state and/or it is possible

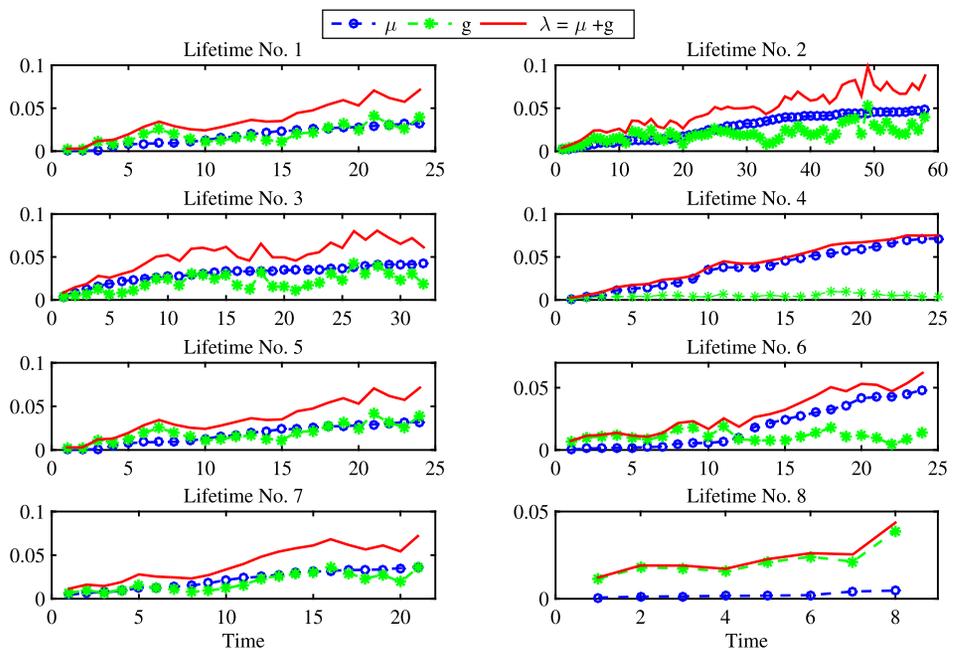


FIG. 6. Estimated λ (solid lines), μ (dashed-star lines) and g (dashed-circle lines) for eight lifetimes in Test Set No. 1.

that the external effects are mitigated due to the physical location of the turbine during that lifetime. These possibilities may be explored further by the wind turbine company, who may better be able to understand the cause of the failures of the turbines and to use this information for planning (locations of future turbines, maintenance policies, replacement policies, etc.).

5.2. Examining the hazard rate at the point of failure. We wanted to know whether the hazard rates were high at the times when the turbines actually failed. We used rank statistics to do this. In particular, for each lifetime, we evaluated the hazard rank percentile when it failed. That is, we took each lifetime of length (T_i) and considered it a part of a cohort of all other lifetimes equal to or exceeding T_i . Then at time T_i , we calculated the rank percentile of the turbine that failed, which is the fraction of turbines whose hazard rate was lower. The higher the percentiles, the better our prediction method performed in terms of distinguishing failures from nonfailures.

Figure 7(a) presents the hazard rank percentile for all turbines in the five test sets (TS1–TS5) for one day and one week before failure. Figure 7(b) shows the same information, but in the form of box plots. It can be observed from these figures that the hazard percentile of the failed turbine is generally higher than those of the other operating turbines with the same age. This is particularly true one day before the turbine fails, but even one week before failure, the hazard rank percentile of many turbines is still high, with the median percentile rank well above 50. This indicates that our model is performing well.

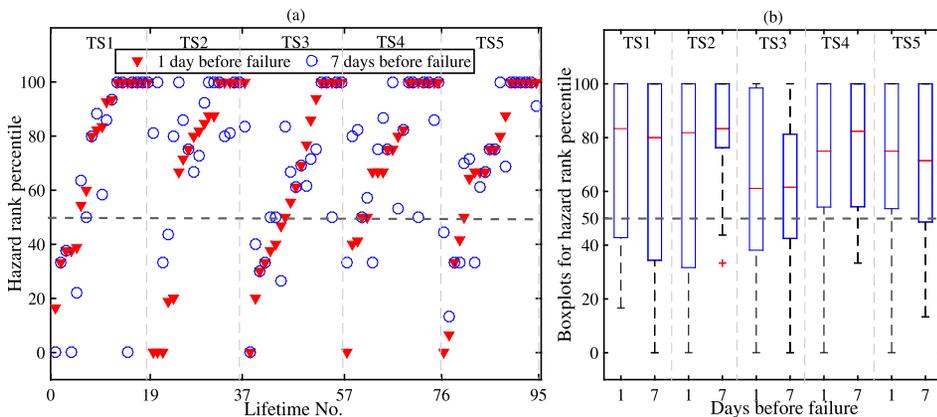


FIG. 7. (a) Hazard rank percentile for the five test sets, with each point representing the percentile rank of one turbine relative to all working turbines with the values sorted by the percentile rank one day before failure and (b) the box plots for the rank percentile for the five test sets; the first box plot in each test set is for 1 day before failure. The dashed horizontal lines are located at the 50th percentiles. In all test sets, most of the rank percentile distribution is above the 50th percentile.

5.3. *Comparison between our model and the Cox proportional hazard model.* We compared our model with the feature-based time-dependent Cox proportional hazard model (Cox PHM). We expect that the hazard rate from our model and from the Cox PHM should be of similar accuracy; however, if our assumption is true that the hazard rate can be split into an external and an internal vulnerability state, the Cox PHM should not be able to predict failures in advance as accurately as our model, which uses the internal state to make decisions rather than the total hazard. The Cox PHM tends to be very sensitive to the covariate values at the previous time step, making its hazard rate fluctuate and leading to possible problems with decision-making. We used a time-dependent Cox PHM with a Weibull-based baseline hazard function, trained on the same training sets as our model. The Weibull distribution is the most commonly used distribution in reliability and degradation analysis, and is often used with the Cox PHM [Boutros and Liang (2011)]. We repeated the experiments on all 5 splits of data, tuning the regularization constant to 0.1 through cross-validation. To compare our model with the Cox PHM, we calculated the hazard rank percentiles at the failure point for both our model and for the Cox PHM for all five test sets. We then counted the number of times in each test set that our model gives a higher hazard rank, and then performed the sign test to assess whether our model significantly outperformed the Cox PHM, which it did (p -value = 0.0480). The individual hazard ranks of each lifetime using both models are shown in Figure 8(a). In Figure 8(b), we show the box plots of rank percentile for our model and the Cox PHM. These figures show that our method outperformed the Cox PHM for most lifetimes and, in particular, that the vulnerability levels of failed turbines were higher in our model than for the Cox PHM.

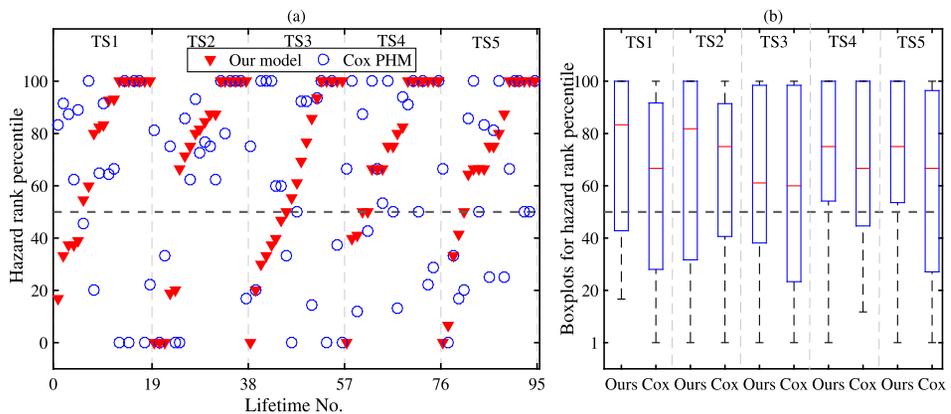


FIG. 8. Comparison between our model and Cox on the wind turbine data set, (a) hazard rank percentile for the five test sets, with each point representing the percentile rank of one turbine relative to all working turbines with the values sorted by our model and (b) the box plots for the rank percentile for the five test sets; the first box plot in each test set is for our model. The dashed horizontal lines are located at the 50th percentiles.

5.4. *Cost-benefit analysis and decision-making.* As discussed in Section 3.1, the value of these techniques lies in their power for maintenance decisions. We considered cost function $C_d(\xi)$, shown in Figure 2. We let c_1 and c_2 denote the unit cost of a late and early warning, respectively. We assumed that early warnings are preferred over late warnings, that is, $c_1 \geq c_2$. We applied the above model on all five training sets to find the optimal threshold of $d = 5$ days and then calculated the associated cost on the test sets. The summary of results is given in Table 2. We have also reported the total cost associated with warning at the failure point, the cost associated with the Cox PHM, and the mean and standard deviation (sd) of each model over the five folds. We repeated this experiment on three different combinations of c_1 and c_2 . As c_1 was increased, the costs for all models also increased. Table 2 indicates that our model performs slightly better (across these combinations of c_1 and c_2) than the Cox PHM with respect to cost; this is in addition to its distinct benefit of being more interpretable. As expected, both models perform substantially better than warning at failure.

Figure 9 illustrates another mechanism for making decisions. It shows the trade-off between the percentage of missing operating time and the percentage of unexpected failures. Here, the percentage of missing operating time is the fraction of total potential operating time when the turbine does not operate due to early warnings. The percentage of unexpected failures is the fraction of replacements that happen as a result of late warnings. That is, it is the number of failures that happen while the turbine is operating divided by the total number of replacements. This trade-off is shown for the 19 lifetimes in one test set. A figure like this can be used to determine the desired cost-benefit trade-off between early and late warnings. Using the training data, we can then find the corresponding threshold for the hazard rate to generate warnings, which depends on the cost of failure replacements

TABLE 2
Summary of results (total cost) for the warning generation process given our model and the Cox PHM for three combinations of c_1 and c_2

Test set No.	$c_1 = c_2$ cost of failure warning = 95		$c_1 = 5c_2$ cost of failure warning = 475		$c_1 = 10c_2$ cost of failure warning = 950	
	Our model	Cox PHM	Our model	Cox PHM	Our model	Cox PHM
1	89	99	291	306	421	425
2	95	91	180	210	255	198
3	92	87	193	223	215	198
4	92	95	364	393	404	484
5	95	98	197	209	223	221
Mean (sd)	92.6 (2.51)	94.00 (5.0)	245.0 (79.8)	268.2 (80.5)	303.60 (100.7)	305.20 (138.2)

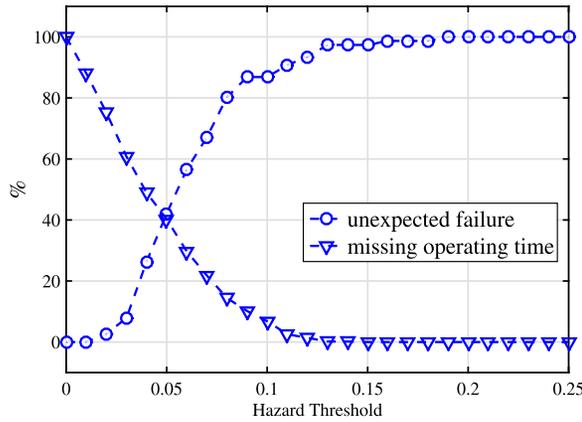


FIG. 9. Trade-off between the percentage of missing operating time and the percentage of unexpected failures. This is shown over various hazard thresholds using data from one of the test sets. Here, the percentage of missing operating time is the fraction of total potential operating time when the turbine does not operate due to early warnings. The percentage of unexpected failures is the fraction of replacements that happen as a result of late warnings. That is, it is the number of failures that happen while the turbine is operating divided by the total number of replacements. This trade-off is shown for the 19 lifetimes in one test set.

and the cost of nonfailure replacements. One might choose the threshold on the hazard rate for which the long-run average unit cost of the system is minimized.

6. Numerical experiments. In this section we provide a set of numerical experiments, including a simulation study and comparisons with previous models.

6.1. *Simulation study.* In this section we demonstrate through simulation experiments (1) the motivation of this work and (2) the empirical consistency of the parameter estimation method for recovering true parameters. In our simulation, a single feature is used as the observable signal over time (we might consider this variable as representing the external temperature near a wind farm). The rest of the parameters associated with μ and g used in this example are as follows: $\beta_0 = -7$, $\beta_1 = 0.5$, $\alpha_0 = -14$, and $\alpha_1 = 5$. These values were chosen so that the parameters μ and g were different, but both would be on approximately the same scale. These four parameters specify the internal and external effects. Therefore, the overall hazard rate for the i th lifetime at time t is

$$\lambda(t|x_i(1), \dots, x_i(t)) = \left[\sum_{\ell \in \{1,2,3,\dots,t\}} \exp(-7 + 0.5x_i(\ell)) \right] + \exp(-14 + 5x_i(t)),$$

where each $x_i(t) \sim \mathcal{N}(0, 1)$. To generate a survival time T_i for each lifetime i given the covariate measurements, we first draw a random number v from $\mathcal{U}(0, 1)$. The failure time is the time point at which the conditional survival function [which can

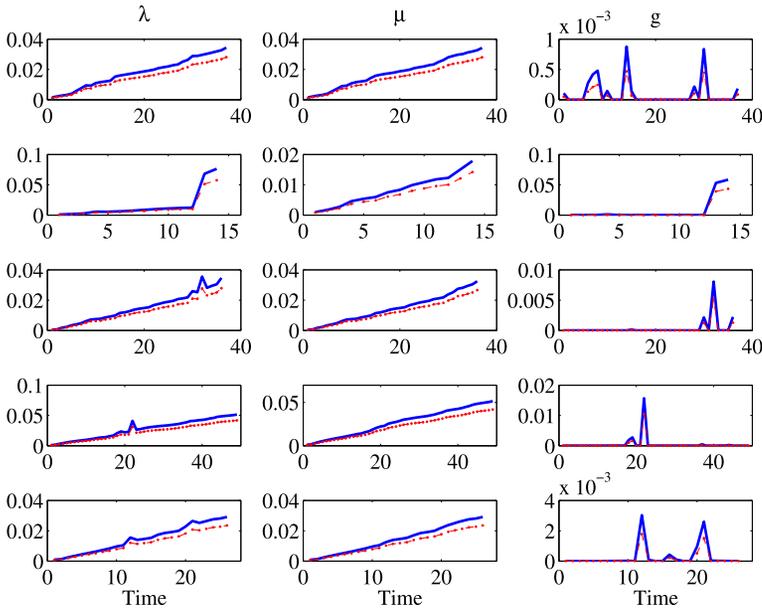


FIG. 10. True (solid lines) overall hazard rates (λ), internal (μ) and external (g) hazard curves for 5 lifetimes (rows 1–5) and their estimates (dashed lines) based on parameter estimation using $N = 50$ sample lifetimes.

be computed at time t as $\exp(-\int_0^t \lambda(u|x_i(1), \dots, x_i(u)) du]$ equals v . We then used N (50, 100, 200, 400, 800) simulated trajectories to estimate the parameters of the model. In Figure 10, the true internal and external terms, and the overall hazard rates are shown as solid lines. Their estimated values based on parameter estimation with $N = 50$ sample lifetimes are shown as dashed lines. Figure 10 shows that the model was able to approximately capture both μ and g terms. In this particular experiment, the model slightly underestimated β_0 and β_1 , leading to a small bias in estimates for μ and λ in all of the lifetimes. The discrepancy between the actual and the estimated rate increases as a function of time (due to the cumulative nature of the hazard rate).

To evaluate the efficiency of the parameter estimation method, we sampled lifetimes and used our method to recover the true parameters. The simulation was repeated for $N = 50, 100, 200, 400$ and 800 lifetimes to assess the convergence rate to the underlying true values. To assess variance, the experiment was performed for 100 simulation runs for each choice of N . We used the squared error between the simulated and true parameter values to evaluate the estimation results. In Figure 11, the mean estimate from the 100 runs and its 95% prediction interval (the upper and lower bounds are denoted by UB and LB, resp.) are shown for different values of N and the parameters of the model ($\alpha_0, \alpha_1, \beta_0, \beta_1$). Figure 11 illustrates that the prediction intervals narrow and the estimated values converge to the true values as N increases. Table 5 in Appendix C presents numerical values for the

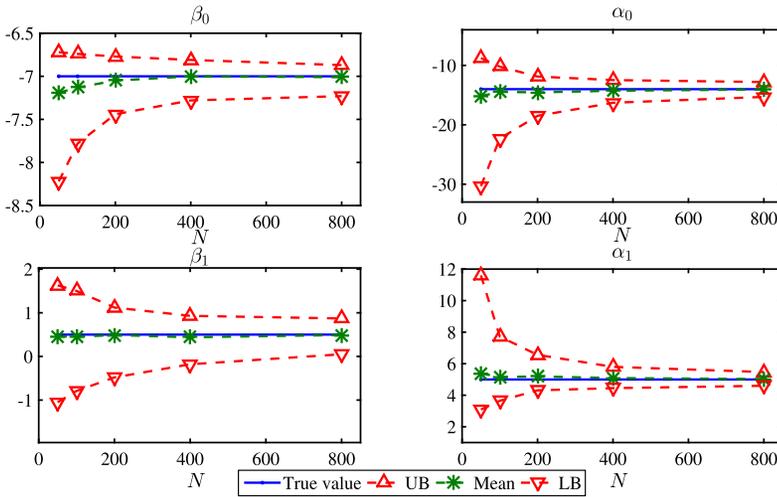


FIG. 11. The true value of each of the 4 model parameters α_0 , α_1 , β_0 and β_1 (solid blue lines in the middle), the corresponding mean estimate based on 100 runs (middle dashed lines in blue) and the 95% prediction interval (the outer two dashed lines in red), as a function of the number of sample lifetimes N used for estimation.

mean, the standard deviation and the mean squared error of estimation over the 100 simulation runs for the various choices of N .

To demonstrate the quality of the estimation for the overall hazard rate, we show the result of estimating λ (overall hazard rate) using N simulated lifetimes for one randomly chosen lifetime and various N in Figure 12. We can observe

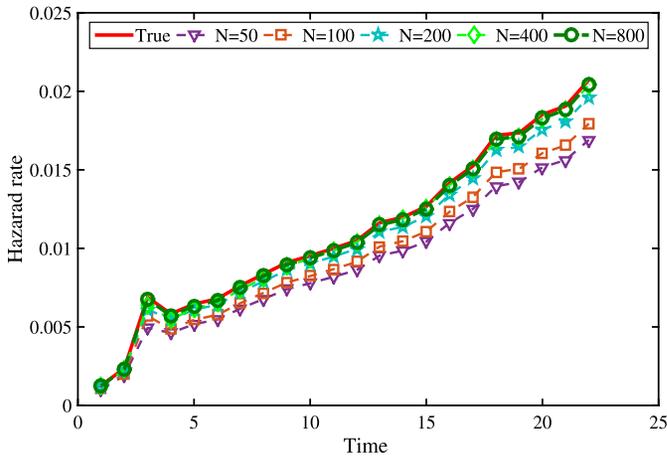


FIG. 12. Estimated results for $\lambda = \mu + g$ for one simulated lifetime (dashed lines). The average percent errors for $N = 50, 100, 200, 400$ and 800 are 19.06%, 13.97%, 5.59%, 2.42% and 1.40%, respectively. This figure shows that the estimation error decreases as N increases.

that the estimated hazard rates converge to the actual hazard rate as the number of training data used for estimation increases. However, the discrepancy between the actual and the estimated rate goes up as a function of time (due to the cumulative nature of the hazard rate), and this is particularly pronounced for the smaller values of N .

6.2. Comparison with existing models. In this subsection we illustrate possible benefits of our model as compared to the models in each of the three categories described in Section 2. The purpose is not to show that our model outperforms previous methods with respect to the estimation of the hazard rate; here, we aim to show that our model possesses the complexity to reproduce behavior generated by existing approaches. We will also show that our model has some additional benefits, such as fewer assumptions and parameters.

6.2.1. Comparison with models with degradation signals. We first show that our model can be easily made to reproduce behaviors of other models that employ explicit degradation signals for degradation modeling. We compare our model with one of the recent papers in this category [Bian and Gebraeel (2013)], which employs historical and real-time signals related to environmental conditions, as well as an observable degradation signal representing the underlying degradation process. The model of the degradation signal denoted by $s_i(t)$ corresponding to the i th unit is expressed as

$$(15) \quad s_i(t) = \int_0^t [\alpha_i + \beta_i \omega(v)] dv + \gamma_i B(t),$$

where α , β and γ are parameters of the model with normal prior distributions, $\omega(t)$ is a deterministic environmental condition that evolves according to a sine function: $\omega(t) = 2 + \sin(\frac{\pi t}{12})$, and $B(t)$ is a standard Brownian Motion process. Unlike the assumptions for our model, Bian and Gebraeel (2013) assumed (i) a predefined formula for the degradation signal itself, (ii) a predefined threshold for failure (predict failure when s exceeds a predefined value), (iii) a known time-dependent distribution with a sine function for the environmental condition, and (iv) a Brownian noise distribution, and normal priors on all other parameters. We simulated 1000 signals (500 for training and 500 for testing) from their model, and then used the simulated degradation signals $[s_i(t), 1 \leq i \leq 500]$ and the environmental observations to train our model. In Figure 13, three sample degradation signals and their corresponding estimates of $\log(\hat{\mu})$ and \hat{g} using our model are presented. Our model managed to decompose the hazard rate perfectly into a monotonic latent degradation function and a transient vulnerability rate—without needing to make the assumptions of Bian and Gebraeel (2013) for the form of signal s_i . To assess whether the trained model is useful for decision-making, we used five different levels of leading time d (assuming $c_1 = c_2$), and calculated the warning generation

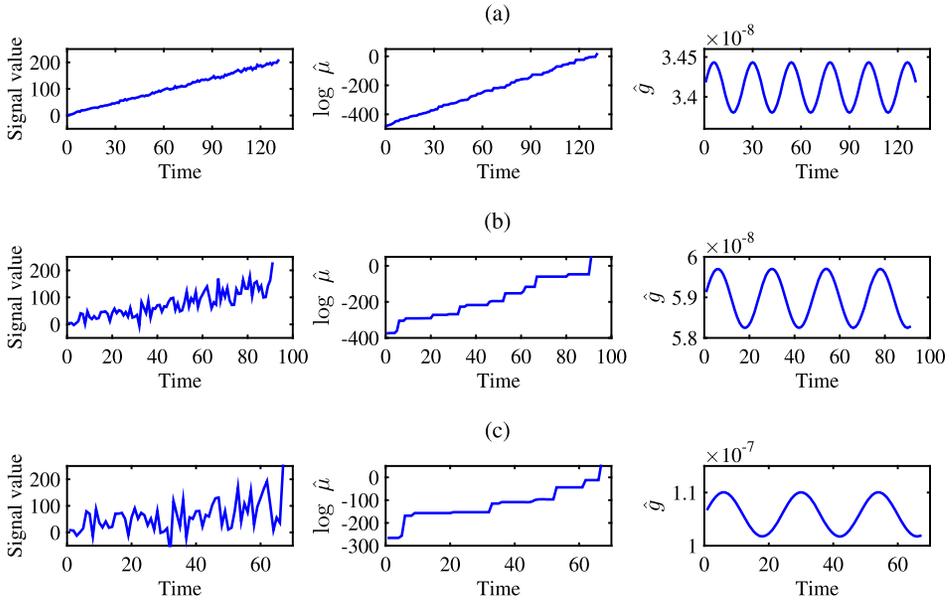


FIG. 13. Three lifetimes of degradation signals (column 1) generated from the model of *Bian and Gebraeel (2013)*, and estimated $\log \hat{\mu}$ (column 2) and \hat{g} from our model (column 3). (a) Degradation signal, (1) $\gamma = 1$; (b) degradation signal, (2) $\gamma = 5$; (c) degradation signal, (3) $\gamma = 10$.

cost reduction percentage using our model with respect to warning at failure. From Figure 14 we can observe that the cost reduction distribution mostly (though not always) takes positive values; in particular, the medians of the cost reductions are

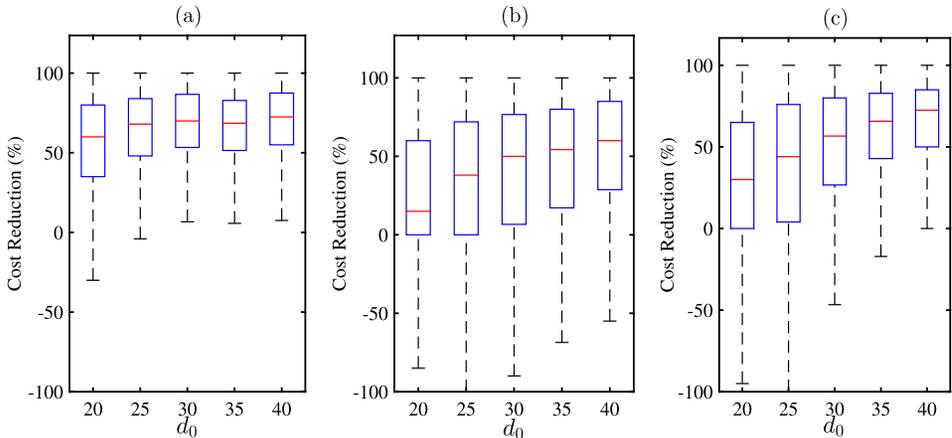


FIG. 14. Box plots for the cost reduction (%) using our model for three levels of noise (γ) and five levels of d_0 . (a) $\gamma = 1$, (b) $\gamma = 5$, (c) $\gamma = 10$.

always positive and become larger as γ decreases (i.e., when there is less noise). This figure shows that, on average, using the latent degradation state of our model to choose warning times actually achieves reasonable performance with respect to the warning generation process, despite the data being generated from the model of [Bian and Gebraeel \(2013\)](#).

[Bian and Gebraeel \(2013\)](#) aimed to predict the remaining useful life, so we used our model to do the same. We evaluated predictions at the time of the 75th percentile of the true lifetime. We first calculated d as the remaining life at the 75th percentile of the lifetime (we calculated this for each simulated signal). Then, using the corresponding threshold value of d (called γ_d), we calculated the warning time (R_{i,γ_d}) as explained in Section 3.1. This means our estimated lifetime is $R_{i,\gamma_d} + d$. Then, we computed the prediction error using the relative percentage difference, in the same way as [Bian and Gebraeel \(2013\)](#), which is

$$\text{Prediction error} = 100 \times \frac{|\text{Actual lifetime} - \text{Estimated lifetime}|}{\text{Actual lifetime}}.$$

The box plots in Figure 15 show that our model provides remaining useful life predictions that are approximately as accurate as those of [Bian and Gebraeel's](#) [see Figure 3 in [Bian and Gebraeel \(2013\)](#)]. These results indicate that our model could potentially represent a degradation model as complex as that given by [Bian and Gebraeel \(2013\)](#), using fewer parameters and without making heavy distributional assumptions on the parameters or the structure of the degradation signal or environmental conditions. Our model also has the advantage that it does not require a predetermined failure threshold, making the implementation of our model easier in real-world problems.

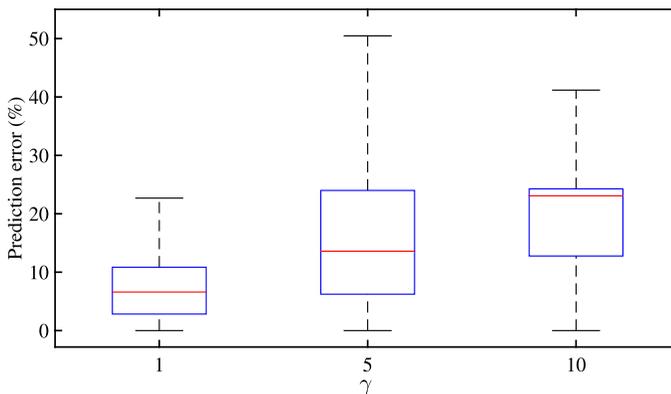


FIG. 15. Results (prediction errors) of remaining useful life prediction as a function of the level of noise γ . These were evaluated at the 75th percentile of the true lifetime (given that the true remaining useful life is known).

6.2.2. *Comparison with hazard models with time-varying covariates.* We performed a set of experiments on prognostic turbofan engine data from the publicly available NASA Prognostic Data Repository [DeCastro, Litt and Frederick (2008), Saxena et al. (2008)]. The data set, FD001, includes run-to-failure time series signals that were collected from a dynamic simulation process for 100 engines. These represent data from a modern dual-spool, high-bypass ratio turbofan engine, which has been the focus of many controls and diagnostics/health management studies over the past few years [see Saxena et al. (2008) and references therein]. To generate condition monitoring features, NASA developed a comprehensive logical order of events that is similar to that of real engines. The engine operates normally at the start of each time series and the fault grows in magnitude until the system fails. Each record is a 24-element vector that is a run-to-failure lifetime corresponding to a given operation cycle. The vector consists of three values for the operational settings and 21 values for engine performance measurements, which are averaged over three cycles. We chose 3 cycles because it was computationally more efficient to perform repeated experiments. We did a small-scale study with averaging removed, and we also considered averaging over 2 cycles. In both cases, the results were the same as averaging over 3 cycles. All failures are caused by HPC (High-Pressure Compressor) degradation. We randomly divided our data set into five folds (subsets) of equal size (20 engines per fold). We trained the model with 4 folds and used the last fold for testing, where 20 engines from the training set were used as a validation set to fit the regularization constant. This process was repeated 5 times so that all folds were used for testing.

An interesting observation made from the trained models is that the latent state hazard rate (μ) dominates the transient hazard term (g), which means that failures are mainly the result of soft degradation. This is consistent with the procedure by which these data were generated (i.e., all failures are due to soft degradation) as described by Saxena et al. (2008). The trained models suggest that the hazard rate is very low during the early life of the engines but increases significantly during the last 10% of the lifetimes. We have also observed that the hazard rate of each engine has some significant jumps during its lifetime, which can potentially be an indication of different damage levels. We present an assessment of the failure prediction ability of our model. We calculated the hazard rank percentile at the failure points and 10 cycles before failure for all engines in the 5 test sets. Results reported in Table 3 show that our method performed well on all test sets, that is, at almost all cases the hazard rank at the failure point and the hazard rank at 10 cycles before the failure point were higher than those of other engines with longer lifetimes. Our model and the Cox PHM performed comparably on all test sets.

To demonstrate the potential of using our model for warning generation, we applied the model to all 5 splits of training data to find the optimal threshold when $d = 5$ cycles and then calculated the associated cost on the test sets. The summary

TABLE 3

Summary of the hazard rank percentile on the turbofan data set (TS1–TS5) given by our model and the Cox PHM calculated at 1 and 10 cycles before failure

Test set No.	1 cycle before failure		10 cycles before failure	
	Our model	Cox PHM	Our model	Cox PHM
TS1	99.4	100.00	91.0	86.4
TS2	97.9	100.0	79.8	86.1
TS3	99.3	99.7	74.0	78.9
TS4	98.5	99.3	84.8	85.9
TS5	99.5	100.0	90.0	92.8
Mean (sd) over 5 folds	98.9 (0.6)	99.8 (0.3)	84 (7.1)	86 (4.9)

of results given in Table 4 and Figure 16 indicates that our model, as expected, performs better than warning at failure. Compared to the Cox PHM, although the differences between the two models may not be significantly different, our model generally tends to result in a lower average cost and standard deviation than the Cox PHM.

The remaining cycles to failure calculated at the suggested warning times given by our model and the Cox PHM for the 20 engines in the first training set are shown in Figure 17. It can be seen from this figure that the remaining cycles to failure at the warning point given by our model are closer to the desired one, which is 5 cycles. The Cox PHM had some predictions that were very poor, where the warning time was much too early. This is due to problems with robustness of the

TABLE 4

Summary of results (total cost for warning generation on turbofan data set) given by our model and the Cox PHM for three combinations of c_1 and c_2

Set No.	$c_1 = c_2$ cost of failure warning = 100		$c_1 = 5c_2$ cost of failure warning = 500		$c_1 = 10c_2$ cost of failure warning = 1000	
	Our model	Cox PHM	Our model	Cox PHM	Our model	Cox PHM
TS1	31	33	73	76	102	223
TS2	27	30	64	70	88	77
TS3	37	72	51	174	76	254
TS4	44	41	81	87	103	87
TS5	32	32	80	72	100	97
Mean (sd) over 5 folds	34.20 (6.5)	41.60 (17.5)	69.80 (12.5)	95.80 (44.2)	93.80 (11.6)	147.60 (84.0)

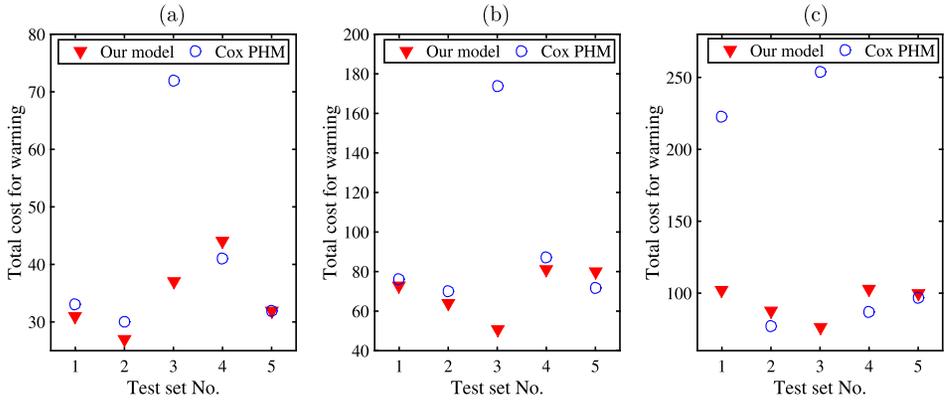


FIG. 16. Total cost for warning generation on the turbofan data set (TS1–TS5) given by our model and the Cox PHM for three combinations of c_1 and c_2 . The mean and standard deviation (sd) of the total cost for our model for the three cases of $c_1 = c_2$, $c_1 = 5c_2$ and $c_1 = 10c_2$ are 34.20 (6.5), 69.80 (12.5) and 93.8 (11.6), respectively. The mean and standard deviation (sd) of the total cost in the Cox PHM model for the three cases of $c_1 = c_2$, $c_1 = 5c_2$ and $c_1 = 10c_2$ are 41.6 (17.5), 95.80 (44.2) and 147.6 (84), respectively. (a) $c_1 = c_2$, (b) $c_1 = 5c_2$, (c) $c_1 = 10c_2$.

Cox PHM that our model does not generally have due to its natural regularization resulting from the use of the internal state. As shown in Figure 18, our model is more robust with respect to changes in covariates, which leads to better decision-making.

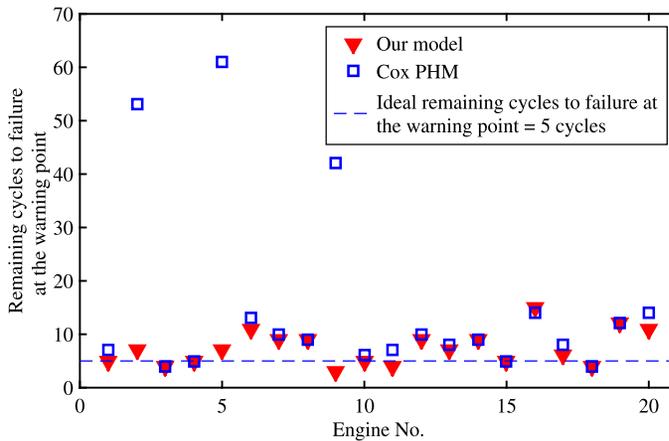


FIG. 17. The remaining cycles to failure calculated at warning points from our model and the Cox PHM. Ideally, the warning should be issued as close as possible to 5 cycles before failure (dashed line). The suggested warning times for Engines 2, 5 and 9 are too early using the Cox PHM. For instance, for Engine 2, the Cox PHM model issued a warning over 53 cycles too early, whereas our model issued a warning that was near the ideal of 5 cycles before failure.

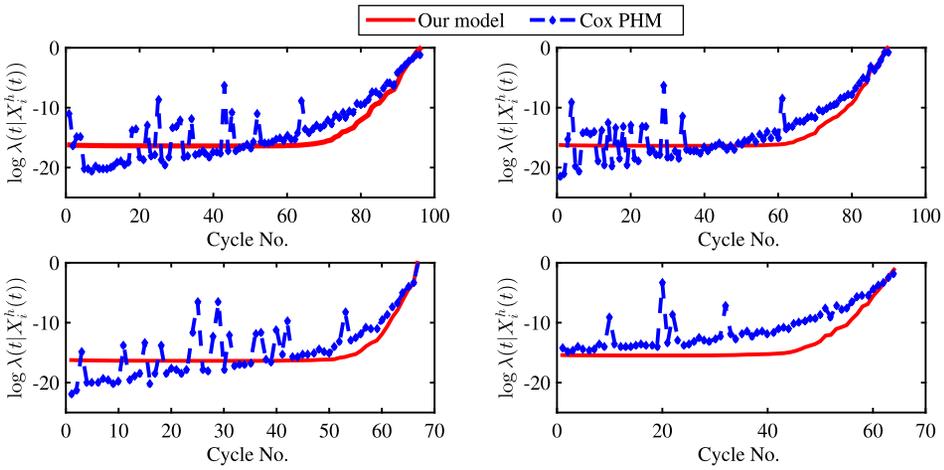


FIG. 18. Sample estimated log of the hazard rates from our model (solid lines) and the Cox PHM (dashed lines).

6.2.3. Comparison with partially-observed degradation models. We compared our model with one of the recent models in the literature [Ghasemi, Yacout and Ouali (2010)], which used a discrete, hidden multistate stochastic process for degradation modeling. The degradation process $Z(t)$ is assumed to be a three-state Markov process with a transition matrix \mathbf{P} . The states are only indirectly observable through condition monitoring. The output of condition monitoring at time t , denoted by y_t , is one of five possible values, and y_t is stochastically related to the actual level of degradation. This stochastic relationship is represented by matrix $\mathbf{Q} = [q_{j,i}]$, where $q_{j,i}$ is the probability of getting the i th output ($i \in \{1, 2, \dots, 5\}$) while the system is in degradation state j , $j \in \{1, 2, 3\}$. From the degradation state, failures are generated according to a time-dependent proportional hazard model. The hazard function is

$$\lambda(t, Z(t)) = \frac{\zeta}{\eta} \left(\frac{t}{\eta}\right)^{\zeta-1} \exp(\gamma Z(t)),$$

where ζ and η are the scale and shape parameters associated with a Weibull baseline and γ is the regression coefficient. The parameters used in our analysis are given in Appendix D. We simulated 1000 lifetimes based on this model, 500 for training and 500 for testing. We used $\log(t)$ and y_t as our covariates. We compared the cost reduction of using our model with that of the model of Ghasemi, Yacout and Ouali (2010) over multiple combinations of d and c_1/c_2 . Figure 19 shows that (i) as c_1/c_2 increases, there is more cost reduction for both models and (ii) our model has very similar performance to that of Ghasemi, Yacout and Ouali (2010). Our model has the benefits that it has fewer parameters, and training our model is computationally less expensive. Our model has only 5 parameters

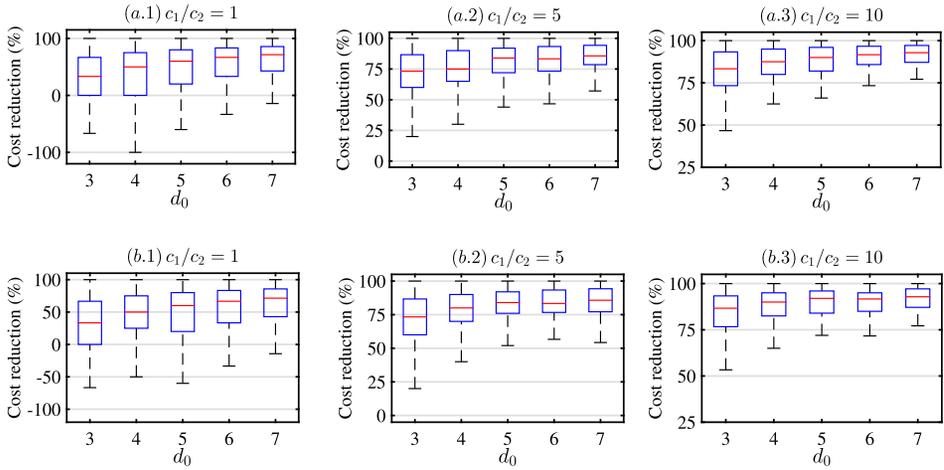


FIG. 19. Cost reduction comparison between (a.1)–(a.3) our model and (b.1)–(b.3) hidden multi-state models for three values of c_1/c_2 . The upper plots are similar to their corresponding lower plots. The same trends are visible within each upper plot and lower plot.

$(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2)$, but the hidden Markov process described here has 27 parameters (9 for $\mathbf{P}_{3 \times 3}$, 15 for $\mathbf{Q}_{3 \times 5}$ and 3 for η, ζ and γ).

7. Concluding remarks. We presented a method for separating the latent internal hazard rate from the temporary hazard due to external sources. We showed that the method has some major advantages over the Cox proportional hazard model, in that it can encode the full history of the turbine within the estimated degradation state in a natural way that the Cox PHM cannot. Further, because the latent degradation state is estimated, it can be used for making maintenance decisions. There are many possible extensions and uses for this model. Although we designed the model for predictive maintenance at wind farms, it can be used for any type of equipment failure prediction, health condition maintenance and in many other application domains (e.g., healthcare). It is also possible that additional prior knowledge is available about the influence of the external factors on the latent state, which can be incorporated in an extended version of the model. In cases where self-healing is possible due to external factors, a third term (in addition to μ and g) that reduces the hazard rate could be introduced. The latent state hazard model has the benefit that it requires very few distributional assumptions and can be trained in a computationally efficient way through convex optimization.

APPENDIX A: PROOF OF PROPOSITION 1

The proof relies on the fact that the composition of convex functions is convex, as follows.

LEMMA 1. *Composition rule for convex functions. Suppose*

$$f(x) = h(\omega_1(x), \omega_2(x), \dots, \omega_k(x)),$$

where $h : \mathbb{R}^k \rightarrow \mathbb{R}$ is convex, and $\omega : \mathbb{R}^n \rightarrow \mathbb{R}$. The function f is convex if one of the following holds:

- h is nondecreasing in the i th argument, and ω_i is convex.
- h is nonincreasing in the i th argument, and ω_i is concave.

PROOF OF PROPOSITION 1. It is sufficient to prove that

$$H_{i,j}(\boldsymbol{\theta}) = -\log(y_{i,j} + (1 - 2y_{i,j}) \exp(-\lambda(j|\mathbf{x}_i^h(j))))$$

is convex in $\boldsymbol{\theta}$ for all (i, j) , based on the fact that the sum of convex functions is also convex. Now, if $y_{i,j} = 0$, then

$$H_{i,j}(\boldsymbol{\theta}) = \lambda(j|\mathbf{x}_i^h(j)),$$

which is convex in $\boldsymbol{\theta}$ due to the fact that the sum of two convex functions $\mu(t|\mathbf{x}_i^h(t))$ and $g(t|\mathbf{x}_i(t))$ is convex. Recall that $g(t|\mathbf{x}_i(t))$ is the exponential of an affine function and $\mu(t|\mathbf{x}_i^h(t))$ is a sum of exponentials of affine functions, which are both convex in $\boldsymbol{\theta}$. If $y_{i,j} = 1$, then

$$H_{i,j}(\boldsymbol{\theta}) = -\log(1 - \exp(-\lambda(j|\mathbf{x}_i^h(j)))).$$

We will invoke Lemma 1. Since $-\log(\cdot)$ is nonincreasing, we have that $-\log f(\cdot)$ is convex if f is concave and positive. That is, $H_{i,j}(\boldsymbol{\theta})$ for $y_{i,j} = 1$ is convex only if $(1 - \exp(-\lambda(j|\mathbf{x}_i^h(j))))$ is concave and positive. Now, since the hazard rate is always nonnegative by its definition, then $\lambda(j|\mathbf{x}_i^h(j)) > 0$ and, thus, $\exp(-\lambda(j|\mathbf{x}_i^h(j))) < 1$, so

$$(1 - \exp(-\lambda(j|\mathbf{x}_i^h(j)))) > 0.$$

Also, since $\lambda(j|\mathbf{x}_i^h(j))$ is convex in $\boldsymbol{\theta}$ (i.e., the sum of two convex functions of μ and g is also convex), then $(1 - \exp(-\lambda(j|\mathbf{x}_i^h(j))))$ is concave in $\boldsymbol{\theta}$. We can conclude that for $y_{i,j} = 1$, $H_{i,j}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. We now have that $H_{i,j}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$ for all (i, j) . Also, note that the ℓ_2 regularization term is strictly convex and, therefore,

$$W_N(\boldsymbol{\theta}|X) = -\log \mathcal{L}_N(\boldsymbol{\theta}|X) + C_1 \|\boldsymbol{\alpha}\|_2^2 + C_2 \|\boldsymbol{\beta}\|_2^2$$

is strictly convex. \square

APPENDIX B: ASYMPTOTIC PROPERTIES

To our knowledge, no work on AMMHM has studied asymptotic properties for a class of models that includes ours. Thus, we prove basic asymptotic properties of the maximum likelihood estimators of the parameters of our model here. There are regularity conditions for the maximum likelihood estimate that can guarantee consistency and asymptotic normality [Lehmann and Casella (1998)]. We verify these regularity conditions for our model. Similar analysis was done in Rashid and Shifa (2009) for the logistic regression model. Let $X_i, 1 \leq i \leq N$ be i.i.d. random variables with a p.d.f. $f(\theta, X_i)$ that depends on parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. The regularity conditions are as follows:

(C0) The distributions $f(\theta, X)$ of the observations are distinct (otherwise, θ cannot be estimated consistently).

(C1) The distributions $f(\theta, X)$ have common support.

(C2) The random variables are $X = (X_1, \dots, X_N)$, where the X_i 's are i.i.d. with probability density $f(\theta, X_i)$ with respect to probability measure μ .

(C3) There exists an open subset ω of Ω containing the true parameter point θ^0 such that for almost all x the density $f(\theta, X)$ admits all third derivatives $(\frac{\partial^3}{\partial \theta_k \partial \theta_j \partial \theta_z} f(\theta, X))$ for all $\theta \in \omega$.

(C4) The first and second logarithmic derivatives of $f(\theta, X)$ satisfy the equations

$$(16) \quad \mathbb{E} \left[\frac{\partial}{\partial \theta_k} \log f(\theta, X) \right] = 0 \quad \forall k,$$

$$(17) \quad \begin{aligned} I_{jk}(\theta) &= \mathbb{E} \left[\frac{\partial}{\partial \theta_j} \log f(\theta, X) \frac{\partial}{\partial \theta_k} \log f(\theta, X) \right] \\ &= \mathbb{E} \left[-\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(\theta, X) \right] \quad \forall j, k. \end{aligned}$$

(C5) Since the $p \times p$ matrix $\mathbf{I}(\theta)$ is a covariance matrix, it is positive semidefinite. We shall assume that the $I_{j,k}(\theta)$'s are finite and that the matrix $\mathbf{I}(\theta)$ is positive definite for all θ in ω , and hence that the statistics

$$\frac{\partial}{\partial \theta_1} \log f(\theta, X), \dots, \frac{\partial}{\partial \theta_p} \log f(\theta, X)$$

are affinely independent with probability 1.

(C6) Finally, we will assume that there exists function $M_{k,m,z}$ such that

$$\left| \frac{\partial \partial^3}{\partial \theta_k \partial \theta_m \partial \theta_z} \log f(\theta, X) \right| \leq M_{k,m,z}(X) \quad \text{for all } \theta \in \omega,$$

where $m_{k,m,z} = \mathbb{E}[M_{k,m,z}(X)] < \infty$ for all k, m, z .

If the above assumptions are satisfied, the following theorem obtained from Lehmann and Casella (1998) can be used for the asymptotic properties of the maximum likelihood estimator.

THEOREM 2. *Let X_1, \dots, X_N be i.i.d. each with a density $f(\boldsymbol{\theta}, X)$, with bounded $\boldsymbol{\theta}$ and X , which satisfies (C0)–(C6). Then with probability tending to 1 as N tends to infinity, there exist solutions $\hat{\boldsymbol{\theta}}_N = (\hat{\theta}_{N,1}, \dots, \hat{\theta}_{N,p})$ of the likelihood equations such that:*

- (i) $\hat{\theta}_{N,j}$ is consistent for estimating θ_j ,
- (ii) $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0)$ is asymptotically normal with (vector) mean zero and covariance matrix $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$, and
- (iii) $\hat{\theta}_{N,j}$ is asymptotically efficient in the sense that its variance attains the Cramér–Rao lower bound as N goes to infinity.

The following theorem states that the regularity conditions are satisfied in our model. Here, $f(\boldsymbol{\theta}, X)$ should be replaced with the likelihood probability given in equation (6).

THEOREM 3. *Let $X = (X_i, T_i)$, $1 \leq i \leq N$, be i.i.d. with the likelihood function $\mathcal{L}_N(\boldsymbol{\theta}|X)$ given in equation (6) with independent parameter set $\boldsymbol{\theta} \in \Theta$, $|\boldsymbol{\alpha}| < M$, and $|\boldsymbol{\beta}| < M$, where M is a finite positive constant independent of N . Then, the regularity conditions (C0), ..., (C6) are satisfied for this model.*

PROOF. Each regularity condition is verified separately as follows:

(C0) The condition (C0), also called identifiability, refers to the fact that the true but unknown parameters of the model ($\boldsymbol{\theta}^0$) should be identified (estimable). The parameters $\boldsymbol{\theta}$ are identified if for any parameter vector $\boldsymbol{\theta}'$ ($\boldsymbol{\theta}' \neq \boldsymbol{\theta}^0$), for some X , $\mathcal{L}_N(\boldsymbol{\theta}'|X) \neq \mathcal{L}_N(\boldsymbol{\theta}^0, X)$. The log-likelihood of our model is an additive multi-index model, which is a linear combination of nonlinear transformations of a linear combination of explanatory variables. Since the parameter space is convex and the regularized log-likelihood function is strictly convex, then the solution of the maximum likelihood problem is unique and, therefore, the true parameter set $\boldsymbol{\theta}^0$ is identified.

(C1) To show that all distributions $P_{\boldsymbol{\theta}}$ have common support, we can prove without loss of generality that the set $A = \{x | \mathcal{L}_N(\boldsymbol{\theta}|X) > 0\}$ is independent of $\boldsymbol{\theta}$. For the probability distribution of our model, since $|\alpha_i| < M$, $|\beta_i| < M$, and feature values are also bounded, then $\mathcal{L}_N(\boldsymbol{\theta}|X)$ is always greater than zero [i.e., the two elements of the likelihood function given in equation (6) are greater than zero regardless of $\boldsymbol{\theta}$]. The only exception is when all feature values are 0, which would make the likelihood 0 regardless of $\boldsymbol{\theta}$.

(C2) This condition is one of the assumptions of the model whereby unit lifetimes are assumed to be i.i.d. with probability distribution generated from the hazard rate function given in equations (1). Thus, condition (C2) is satisfied.

(C3) As the log-likelihood function is a linear combination of other nonlinear functions for each (i, j) , it is sufficient to prove that the third derivative exists for each choice of (i, j) with respect to θ . It is easy to show that the third derivative of each observation in $\mathcal{L}_N(\theta|X)$ exists if the first, second and the third derivatives of $\lambda(j|\mathbf{x}_i^h(j))$ exist for all (i, j) . Since $\lambda = \mu + g$, then we can show that condition (C3) is met if the first, second, and the third derivative of μ and g exists for all $\theta \in \omega$. Now, we can show that

$$(18) \quad \begin{cases} \frac{\partial \lambda(j|\mathbf{x}_i^h(j))}{\partial \alpha_k} = x_{i,k}(j) \exp(\alpha^\top \mathbf{x}_i(j)), \\ \frac{\partial^2 \lambda(j|\mathbf{x}_i^h(j))}{\partial \alpha_k \partial \alpha_m} = x_{i,k}(j)x_{i,m}(j) \exp(\alpha^\top \mathbf{x}_i(j)), \\ \frac{\partial^3 \lambda(j|\mathbf{x}_i^h(j))}{\partial \alpha_k \partial \alpha_m \partial \alpha_z} = x_{i,k}(j)x_{i,m}(j)x_{i,z}(j) \exp(\alpha^\top \mathbf{x}_i(j)), \end{cases}$$

$$(19) \quad \begin{cases} \frac{\partial \lambda(j|\mathbf{x}_i^h(j))}{\partial \beta_k} = \sum_{l=1}^j x_{i,k}(l) \exp(\beta^\top \mathbf{x}_i(l)), \\ \frac{\partial^2 \lambda(j|\mathbf{x}_i^h(j))}{\partial \beta_k \partial \beta_m} = \sum_{l=1}^j x_{i,k}(l)x_{i,m}(l) \exp(\beta^\top \mathbf{x}_i(l)), \\ \frac{\partial^3 \lambda(j|\mathbf{x}_i^h(j))}{\partial \beta_k \partial \beta_m \partial \beta_z} = \sum_{l=1}^j x_{i,k}(l)x_{i,m}(l)x_{i,z}(l) \exp(\beta^\top \mathbf{x}_i(l)). \end{cases}$$

Since all of the above derivatives exist for all θ (given that θ and X are bounded), then (C3) is satisfied for our model.

(C4) This condition shows two properties of the score function and the Hessian matrix of the log-likelihood function. To verify these relationships, we first rewrite the first derivative of the log-likelihood function in terms of a sum over time $u = 1, \dots, \max\{T_1, \dots, T_N\}$ using equation (10) as

$$(20) \quad \begin{aligned} & \frac{\partial}{\partial \theta_k} (\log \mathcal{L}_N(\theta|X)) \\ &= \sum_u \sum_i -Y_i(u-1)(1-2y_{i,u}) \frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{y_{i,u} + (1-2y_{i,u}) \exp(-\lambda(u|\mathbf{x}_i^h(u)))} \\ & \quad \times \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_k}, \end{aligned}$$

where u 's are the time points and $Y_i(u-1)$ is a binary indicator function being one only if unit i has survived until time point $u-1$. This allows the sum over u to

range over all time instead of only until the failure times as in equation (10). The term $\mathbb{E}[\frac{\partial}{\partial\theta_k}(\log \mathcal{L}_N(\boldsymbol{\theta}|X))]$ can now be calculated as

$$\begin{aligned}
 & \mathbb{E}\left[\frac{\partial}{\partial\theta_k}(\log \mathcal{L}_N(\boldsymbol{\theta}|X))\right] \\
 (21) \quad & = \sum_u \sum_i \mathbb{E} \\
 & \quad \times \underbrace{\left[-Y_i(u-1) \frac{(1-2y_{i,u}) \exp(-\lambda(u|\mathbf{x}_i^h(u)))}{y_{i,u} + (1-2y_{i,u}) \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \frac{\partial\lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k}\right]}_{F(i,u,k)}.
 \end{aligned}$$

So, if $Y_i(u-1)$ is zero, the internal quantity is zero. Otherwise, we can expand $y_{i,u}$ based on the two possibilities of (1) failure within the next interval $(u-1, u)$ with the probability of $(1 - \exp(-\lambda(u|\mathbf{x}_i^h(u))))$, and (2) survival within the next interval with the probability of $(\exp(-\lambda(u|\mathbf{x}_i^h(u))))$. Considering the interior of equation (21), we have

$$\begin{aligned}
 & \mathbb{E}(F(i, u, k)) \\
 (22) \quad & = -Y_i(u-1) \mathbb{E}\left[\frac{\partial\lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k} \times (\exp(-\lambda(u|\mathbf{x}_i^h(u))))\right] \\
 & \quad - \frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{1 - \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \frac{\partial\lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k} \times (1 - \exp(-\lambda(u|\mathbf{x}_i^h(u)))) \\
 & = -Y_i(u-1) \times 0 = 0,
 \end{aligned}$$

which proves equation (16) in (C4). We use a similar approach to prove equation (17). The left-hand side of equation (17) can be calculated as

$$\begin{aligned}
 & \mathbb{E}\left[\sum_u \sum_{u'} \sum_i \sum_{i'} F(i, u, k) \times F(i', u', m)\right] \\
 & = \mathbb{E}\left[\sum_u \sum_i F(i, u, k) \times F(i, u, m)\right] \\
 (23) \quad & + \mathbb{E}\left[\sum_u \sum_i \sum_{i' \neq i} F(i, u, k) \times F(i', u, m)\right] \\
 & + \mathbb{E}\left[\sum_u \sum_{u' \neq u} \sum_i F(i, u, k) \times F(i, u', m)\right] \\
 & + \mathbb{E}\left[\sum_u \sum_{u' \neq u} \sum_i \sum_{i' \neq i} F(i, u, k) \times F(i, u', m)\right].
 \end{aligned}$$

The first part of the above equation can be rewritten using equation (21) as

$$\begin{aligned}
 & \sum_u \sum_i \mathbb{E}[F(i, u, k) \times F(i, u, m)] \\
 &= \sum_u \sum_i \mathbb{E} \left[Y_i(u-1)^2 \left(\frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{y_{i,u} + (1 - 2y_{i,u}) \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \right)^2 \right. \\
 & \quad \left. \times \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_k} \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_m} \right] \\
 &= \sum_u \sum_i Y_i(u-1)^2 \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_k} \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_m} \\
 (24) \quad & \times \left[1 \times \exp(-\lambda(u|\mathbf{x}_i^h(u))) + \left(\frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{1 - \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \right)^2 \right. \\
 & \quad \left. \times (1 - \exp(-\lambda(u|\mathbf{x}_i^h(u)))) \right] \\
 &= \sum_u \sum_i Y_i(u-1)^2 \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_k} \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial \theta_m} \\
 & \quad \times \left(\frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{1 - \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \right).
 \end{aligned}$$

In an analogous way, we can prove that the second element of equation (23) equals zero. The third element of equation (23) is the sum of two zero terms as follows:

$$\begin{aligned}
 & \mathbb{E} \left[\sum_u \sum_{u' \neq u} \sum_i F(i, u, k) \times F(i, u', m) \right] \\
 &= \mathbb{E} \left[\sum_u \sum_{u' < u} \sum_i F(i, u, k) \times F(i, u', m) \right] \\
 (25) \quad & + \mathbb{E} \left[\sum_u \sum_{u' > u} \sum_i F(i, u, k) \times F(i, u', m) \right] \\
 &= \left[\sum_{u' < u} \sum_u \sum_i F(i, u', m) \times \underbrace{\mathbb{E}[F(i, u, k)]}_0 \right] \\
 & + \left[\sum_{u < u'} \sum_u \sum_i F(i, u, k) \times \underbrace{\mathbb{E}[F(i, u', m)]}_0 \right] = 0,
 \end{aligned}$$

where we used equation (22). In an analogous way, we can prove that the last element of equation (23) equals zero. Now, from equations (23), (24) and (25), we

can conclude that

$$\mathbb{E}\left[\sum_u \sum_{u'} \sum_i \sum_{i'} F(i, u, k) \times F(i', u', m)\right] = \mathbb{E}\left[\sum_u \sum_i F(i, u, k) \times F(i, u, m)\right].$$

We have finished simplifying the left-hand side of equation (17). The right-hand side of equation (17), which equals the expected value of the negative of the Hessian matrix of $\log \mathcal{L}_N(\boldsymbol{\theta}|X)$, can be rewritten as follows:

$$\begin{aligned} & \mathbb{E}\left[-\frac{\partial^2}{\partial\theta_k \partial\theta_m} \log \mathcal{L}_N(\boldsymbol{\theta}|X)\right] \\ (26) \quad & = \sum_u \sum_i \mathbb{E}\left[Y_i(u-1) \left((1-2y_{i,u}) \frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{y_{i,u} + (1-2y_{i,u}) \exp(-\lambda(u|\mathbf{x}_i^h(u)))} \right. \right. \\ & \quad \left. \left. \times \left(\frac{\partial^2 \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k \partial\theta_m} - \frac{\frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k} \times \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_m} \times y_{i,u}}{y_{i,u} + (1-2y_{i,u}) \exp(-\lambda(u|\mathbf{x}_i^h(u)))} \right) \right) \right]. \end{aligned}$$

Pulling the expectation into the sum and separating cases where $y_{i,u} = 1$ from $y_{i,u} = 0$, we can simplify equation (26) as

$$\begin{aligned} & \mathbb{E}\left[-\frac{\partial^2}{\partial\theta_k \partial\theta_m} \log \mathcal{L}_N(\boldsymbol{\theta}|X)\right] \\ & = \sum_u \sum_i Y_i(u-1) \left[\left(\frac{\partial^2 \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k \partial\theta_m} \right) \exp(-\lambda(u|\mathbf{x}_i^h(u))) \right. \\ (27) \quad & \left. - \frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{1 - \exp(-\lambda(u|\mathbf{x}_i^h(u)))} \right. \\ & \quad \left. \times \left(\frac{\partial^2 \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k \partial\theta_m} - \frac{\frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k} \times \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_m}}{1 - \exp(-\lambda(u|\mathbf{x}_i^h(u)))} \right) (1 - \exp(-\lambda(u|\mathbf{x}_i^h(u)))) \right] \\ & = \sum_u \sum_i Y_i(u-1) \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_k} \frac{\partial \lambda(u|\mathbf{x}_i^h(u))}{\partial\theta_m} \times \left(\frac{\exp(-\lambda(u|\mathbf{x}_i^h(u)))}{1 - \exp(-\lambda(j|\mathbf{x}_i^h(u)))} \right). \end{aligned}$$

Since the last term in equation (27) is equivalent to the last term in equation (24), and equations (27) and (24) are, respectively, the right- and left-hand sides of equation (17), we can conclude that equation (17) is satisfied.

(C5) This condition is satisfied since we proved earlier that the log-likelihood is strictly convex and, therefore, $\mathbf{I}(\boldsymbol{\theta})$ (which is the negative of the Hessian matrix) is positive definite and $I_{k,m}(\boldsymbol{\theta})$ is finite for all k and m .

(C6) To show that (C6) holds true, we need to show that the third derivatives are absolutely bounded. We first show that there exists a positive real number \mathcal{M}

such that

$$\left| \frac{\partial^3}{\partial \theta_k \partial \theta_m \partial \theta_z} \log \mathcal{L}_N(\boldsymbol{\theta} | X) \right| \leq \sum_i \sum_j |x_{i,k}(j)| |x_{i,m}(j)| |x_{i,z}(j)| \mathcal{M}$$

for all $\boldsymbol{\theta} \in \omega$.

Since $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and X are all bounded, we can assume that there exists a positive real number \mathcal{M}_1 such that $\exp(\boldsymbol{\alpha}^\top X) < \mathcal{M}_1$ and $\exp(\boldsymbol{\beta}^\top X) < \mathcal{M}_1$ for all X . As a result, we can conclude using equations (18)–(19) that

$$\frac{\partial^3 \lambda(j | \mathbf{x}_i^h(j))}{\partial \theta_k \partial \theta_m \partial \theta_z} \leq \sum_{l=1}^j |x_{i,k}(l)| |x_{i,m}(l)| |x_{i,z}(l)| \mathcal{M}_1.$$

Also, as $\frac{\partial^3 \log \mathcal{L}_N(\boldsymbol{\theta} | X)}{\partial \theta_k \partial \theta_m \partial \theta_z}$ is a linear combination of some nonlinear functions all including $x_{i,k}(j)$, $x_{i,m}(j)$ and $x_{i,z}(j)$, and as the sum and the product of bounded functions are also bounded, there exists a real number \mathcal{M} for which

$$\left| \frac{\partial^3}{\partial \theta_k \partial \theta_m \partial \theta_z} \log \mathcal{L}_N(\boldsymbol{\theta} | X) \right| \leq \sum_i \sum_j |x_{i,k}(j)| |x_{i,m}(j)| |x_{i,z}(j)| \mathcal{M} \quad \forall \boldsymbol{\theta} \in \omega.$$

As feature values are all bounded, then

$$\begin{aligned} \mathbb{E} \left[\sum_i \sum_j |x_{i,k}(j)| |x_{i,m}(j)| |x_{i,z}(j)| \mathcal{M} \right] &= \sum_i \sum_j \mathbb{E} |x_{i,k}(j) x_{i,m}(j) x_{i,z}(j)| \mathcal{M} \\ &< \infty, \end{aligned}$$

which can be used to prove that (C6) is satisfied. These theorems together provide an immediate proof for Theorem 1. \square

APPENDIX C: TABLE OF RESULTS FOR SIMULATION STUDY

TABLE 5

Table of results for simulation study—[mean, standard deviation (sd) and mean squared error (MSE)] for 5 values of N

Parameters	β_0	α_0	β_1	α_1
True values	-7	-14	0.5	5
N	(Mean, sd, MSE)			
50	(-7.19, 0.39, 0.19)	(-15.17, 4.68, 23.27)	(0.46, 0.69, 0.48)	(5.39, 1.75, 3.22)
100	(-7.13, 0.26, 0.08)	(-14.45, 2.52, 6.54)	(0.45, 0.59, 0.36)	(5.15, 0.87, 0.77)
200	(-7.05, 0.18, 0.03)	(-14.54, 1.53, 2.64)	(0.49, 0.36, 0.13)	(5.19, 0.54, 0.32)
400	(-7.00, 0.12, 0.02)	(-14.26, 0.94, 0.95)	(0.44, 0.27, 0.08)	(5.09, 0.32, 0.11)
800	(-7.01, 0.09, 0.01)	(-14.05, 0.66, 0.44)	(0.49, 0.19, 0.04)	(5.02, 0.23, 0.05)

APPENDIX D: INPUT PARAMETERS OF HMM USED IN SECTION 6.2.3

$$\mathbf{P} = \begin{bmatrix} 0.9 & 0.09 & 0.01 \\ 0 & 0.87 & 0.13 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0.6 & 0.3 & 0.05 & 0.05 & 0 \\ 0.1 & 0.2 & 0.4 & 0.2 & 0.1 \\ 0 & 0.05 & 0.05 & 0.3 & 0.6 \end{bmatrix},$$

$$\zeta = 20, \eta = 4.5, \text{ and } \gamma = 1.4.$$

Acknowledgments. We would like to thank the Accenture—MIT Alliance for funding this work and ENEL for providing the data. We would specifically like to thank Giuseppe Panunzio, Cristian Corbetti, Andrew Fano and Thania Villatoro from Accenture. We would also like to thank Şeyda Ertekin and Ken Cohn for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplementary material for “The latent state hazard model, with application to wind turbine reliability” (DOI: [10.1214/15-AOAS859SUPP](https://doi.org/10.1214/15-AOAS859SUPP); .pdf). The supplementary material includes three sections: A: Interpretation of the model; B: Notes on the reliability assumption, and Supplement; and C: Making the coefficients interpretable.

REFERENCES

- ANDERSEN, P. K. and VAETH, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45** 523–535.
- BANJEVIC, D. and JARDINE, A. K. S. (2006). Calculation of reliability function and remaining useful life for a Markov failure time process. *IMA J. Manag. Math.* **17** 115–130. [MR2216398](#)
- BANJEVIC, D., JARDINE, A. K. S., MAKIS, V. and ENNIS, M. (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR* **39** 32–50.
- BIAN, L. and GEBRAEEL, N. (2012). Computing and updating the first-passage time distribution for randomly evolving degradation signals. *IIE Transactions (Institute of Industrial Engineers)* **44** 974–987.
- BIAN, L. and GEBRAEEL, N. (2013). Stochastic methodology for prognostics under continuously varying environmental profiles. *Stat. Anal. Data Min.* **6** 260–270. [MR3062269](#)
- BOUTROS, T. and LIANG, M. (2011). Detection and diagnosis of bearing and cutting tool faults using hidden Markov models. *Mech. Syst. Signal Process.* **25** 2102–2124.
- COLLETT, D. (2003). *Modelling Binary Data*, 2nd ed. *Chapman & Hall/CRC Texts in Statistical Science Series*. Chapman & Hall/CRC, Boca Raton, FL. [MR1999899](#)
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- DECASTRO, J. A., LITT, J. S. and FREDERICK, D. K. (2008). A modular aero-propulsion system simulation of a large commercial aircraft engine. In *AIAA-2008-4579, 44th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*. Hartford, USA.
- FISHER, L. D. and LIN, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* **20** 145–157.

- FLORY, J. A., KHAROUFEH, J. P. and GEBRAEEL, N. Z. (2014). A switching diffusion model for lifetime estimation in randomly varying environments. *IIE Transactions (Institute of Industrial Engineers)* **46** 1227–1241.
- GEBRAEEL, N. and PAN, J. (2008). Prognostic degradation models for computing and updating residual life distributions in a time-varying environment. *IEEE Trans. Reliab.* **57** 539–550.
- GHASEMI, A., YACOUT, S. and OUALI, M. S. (2007). Optimal condition based maintenance with imperfect information and the proportional hazards model. *Int. J. Prod. Res.* **45** 989–1012.
- GHASEMI, A., YACOUT, S. and OUALI, M. S. (2010). Evaluating the reliability function and the mean residual life for equipment with unobservable states. *IEEE Trans. Reliab.* **59** 45–54.
- GORJIAN, N., MA, L., MITTINTY, M., YARLAGADDA, P. and SUN, Y. (2009). A review on reliability models with covariates. In *4th World Congress on Engineering Asset Management, WCEAM 2009* 385–397. Athens, Greece.
- GUO, H., WATSON, S., TAVNER, P. and XIANG, J. (2009). Reliability analysis for wind turbines with incomplete failure data collected from after the date of initial installation. *Reliab. Eng. Syst. Saf.* **94** 1057–1063.
- HAMEED, Z., HONG, Y. S., CHO, Y. M., AHN, S. H. and SONG, C. K. (2009). Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renew. Sustain. Energy Rev.* **13** 1–39.
- HÖHLE, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biom. J.* **51** 961–978. [MR2744450](#)
- HONTELEZ, J. A. M., BURGER, H. H. and WIJNMALEN, D. J. D. (1996). Optimum condition-based maintenance policies for deteriorating systems with partial information. *Reliab. Eng. Syst. Saf.* **51** 267–274.
- JARDINE, A. K. S., ANDERSON, P. M. and MANN, D. S. (1987). Application of the Weibull proportional hazards model to aircraft and marine engine failure data. *Qual. Reliab. Eng. Int.* **3** 77–82.
- KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, Hoboken, NJ. [MR1924807](#)
- KHAROUFEH, J. P. (2003). Explicit results for wear processes in a Markovian environment. *Oper. Res. Lett.* **31** 237–244. [MR1967296](#)
- KHAROUFEH, J. P. and COX, S. M. (2005). Stochastic models for degradation-based reliability. *IIE Transactions (Institute of Industrial Engineers)* **37** 533–542.
- KHAROUFEH, J. P., FINKELSTEIN, D. E. and MIXON, D. G. (2006). Availability of periodically inspected systems with Markovian wear and shocks. *J. Appl. Probab.* **43** 303–317. [MR2248566](#)
- KUSIAK, A., ZHANG, Z. and VERMA, A. (2013). Prediction, operations, and condition monitoring in wind energy. *Energy* **60** 1–12.
- LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. [MR1639875](#)
- LIN, D. Y. and YING, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann. Statist.* **23** 1712–1734. [MR1370304](#)
- LU, B., LI, Y., WU, X. and YANG, Z. (2009). A review of recent advances in wind turbine condition monitoring and fault diagnosis. In *2009 IEEE Power Electronics and Machines in Wind Applications*. Lincoln, NE.
- MAKIS, V. and JARDINE, A. K. S. (1991). Computation of optimal policies in replacement models. *IMA J. Manag. Math.* **3** 169–175.
- MARQUEZ, F. P. G., TOBIAS, A. M., PREZ, J. M. P. and PAPAELIAS, M. (2012). Condition monitoring of wind turbines: Techniques and methods. *Renew. Energy* **46** 169–178.
- MARTINUSSEN, T. and SCHEIKE, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89** 283–298. [MR1913959](#)
- MARVUGLIA, A. and MESSINEO, A. (2012). Monitoring of wind farms' power curves using machine learning techniques. *Appl. Energy* **98** 574–583.

- MOGHADDASS, R. and RUDIN, C. (2015). Supplement to “The latent state hazard model, with application to wind turbine reliability.” DOI:10.1214/15-AOAS859SUPP.
- MOGHADDASS, R. and ZUO, M. J. (2012). A parameter estimation method for a condition-monitored device under multi-state deterioration. *Reliab. Eng. Syst. Saf.* **106** 94–103.
- PENG, Y. and DONG, M. (2011). A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction. *Mech. Syst. Signal Process.* **25** 237–252.
- PIJNENBURG, M. (1991). Additive hazards models in repairable systems reliability. *Reliab. Eng. Syst. Saf.* **31** 369–390.
- QIAN, X. and WU, Y. (2014). Condition based maintenance optimization for the hydro generating unit with dynamic economic dependence. *International Journal of Control and Automation* **7** 317–326.
- QIU, Y. N., FENG, Y. H., TAVNER, P. J., RICHARDSON, P., ERDOS, G. and CHEN, B. (2012). Wind turbine SCADA alarm analysis for improving reliability. *Wind Energy* **15** 951–966.
- RASHID, M. and SHIFA, N. (2009). Consistency of the maximum likelihood estimator in logistic regression model: A different approach. *Journal of Statistics* **16** 1–11.
- RUDIN, C. and VAHN, G.-Y. (2014). The big data newsvendor: Practical insights from machine learning. Working paper.
- SAXENA, A., GOEBEL, K., SIMON, D. and EKLUND, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In 2008 *International Conference on Prognostics and Health Management, PHM 2008*. Denver, USA.
- SI, X.-S., WANG, W., HU, C.-H. and ZHOU, D.-H. (2011). Remaining useful life estimation—A review on the statistical data driven approaches. *European J. Oper. Res.* **213** 1–14. MR2795805
- WU, X. and RYAN, S. M. (2011). Optimal replacement in the proportional hazards model with semi-Markovian covariate process and continuous monitoring. *IEEE Trans. Reliab.* **60** 580–589.
- YANG, W., COURT, R. and JIANG, J. (2013). Wind turbine condition monitoring by the approach of SCADA data analysis. *Renewable Energy* **53** 365–376.
- YANG, W., TAVNER, P. J., CRABTREE, C. J., FENG, Y. and QIU, Y. (2014). Wind turbine condition monitoring: Technical and commercial challenges. *Wind Energy* **17** 673–693.
- ZAHER, A., MCARTHUR, S. D. J., INFELD, D. G. and PATEL, Y. (2009). Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy* **12** 574–593.
- ZHAO, X., FOULADIRAD, M., BÉRENGUER, C. and BORDES, L. (2010). Condition-based inspection/replacement policies for non-monotone deteriorating systems with environmental covariates. *Reliab. Eng. Syst. Saf.* **95** 921–934.
- ZHOU, R. R., SERBAN, N. and GEBRAEEL, N. (2011). Degradation modeling applied to residual lifetime prediction using functional data analysis. *Ann. Appl. Stat.* **5** 1586–1610. MR2849787

SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: raminm@mit.edu

COMPUTER SCIENCE AND
ARTIFICIAL INTELLIGENCE LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: cynthia@mit.edu