Reliability Engineering and System Safety 261 (2025) 111145

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

A transfer learning approach for remaining useful life prediction subject to hard failure considering within and between population variations

Xinxing Guo [a], Song Huang [a], Jianguo Wu [a,*], Chao Wang [b,*]

[a] Department of Industrial Engineering and Management, Peking University, China
[b] Department of Industrial and Systems Engineering, University of Iowa, USA

ARTICLE INFO

Keywords:
Bayesian hierarchical model
Hard failure
Cox PH model
Transfer learning
Remaining useful life prediction

ABSTRACT

Accurate prediction of remaining useful life (RUL) of a unit plays a critical role in condition-based maintenance, especially for hard failure cases. In industrial practice, due to differences in units' types and working environments, there may exist multiple populations, and even within the same population, there are also variations among units. However, existing methods either assume that different units share the same population characteristics and ignore the between-population variations, or solely focus on between-population knowledge transfer while neglecting the within-population variations. To address this issue, this article proposes a transfer learning approach by integrating a Cox Proportional Hazards (PH) model with a Bayesian hierarchical model, which considers both within and between population variations. Specifically, a shared prior distribution is deployed to the parameters of the Cox model in each population, which builds the foundation for transfer learning across different populations. To model within-population variations, a linear mixed-effects model is utilized to represent heterogeneous degradation data of each unit. The effectiveness of the proposed method is demonstrated and compared with various benchmarks through a simulation study and a case study of turbine engines.

1. Introduction

Condition-based maintenance (CBM) has become increasingly popular across various sectors over recent years. CBM makes maintenance decisions based on the information collected by continuously or periodically monitoring the actual condition of in-operation units [1], thus ensuring the system's health before the failures of crucial components happen. With the advancements of sensor and information technologies, multiple sensors have been widely applied to perform conditional monitoring. The acquired sensor data reflects the degradation characteristics of units and creates a data-rich environment to employ data fusion techniques for degradation modeling and remaining useful life (RUL) prediction.

In literature, there are two types of failures when dealing with CM signals: soft failure and hard failure. Soft failure happens when a degradation signal hits its pre-determined failure threshold for the first time [2]. Once a machine encounters a soft failure, the machine's performance ceases to satisfy operational requirements, even though it may still be functioning. General path models [3] and stochastic process models [4–7] are commonly adopted to characterize the degradation process and predict when the failure will occur. Zhang et al. [8] and Kordestani et al. [9] provide a more recent review about data-driven prognostics of soft failure. On the other hand, hard failure usually results in immediate failure or inoperable condition of a unit. The occurrence of hard failure is probabilistically based on the risk level and does not assume a pre-specified and fixed failure threshold. Predicting the hard failure has received intensive attention since its stochastic occurrence adds difficulty and uncertainty for an accurate prediction. Such risk-based failure mechanism is usually captured by modeling the time-to-event data, which contains useful information about life-time distributions.

The joint prognostic model is a popular solution to deal with hard failure, which models both CM signals and survival data to achieve more reliable event prediction. Zhou et al. [10] proposed a two-stage prognostic framework for hard failure prediction by joint modeling of degradation signals and time-to-event data. Based on this work, Man and Zhou [11] further made improvements by employing the extended hazard (EH) model for the time-to-event data. Yue and Kontar [12] presented a non-parametric prognostic framework which exploits a multivariate Gaussian convolution process (MGCP) for characterizing time series signals. Hu and Chen [13] used a random-effects Wiener process to model the sensor signals and adopted a Weibull function for hazard modeling. To capture the highly non-linear relationship

* Corresponding authors.
E-mail addresses: j.wu@pku.edu.cn (J. Wu), chao-wang-2@uiowa.edu (C. Wang).

https://doi.org/10.1016/j.ress.2025.111145

Available online 21 April 2025
0951-8320/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

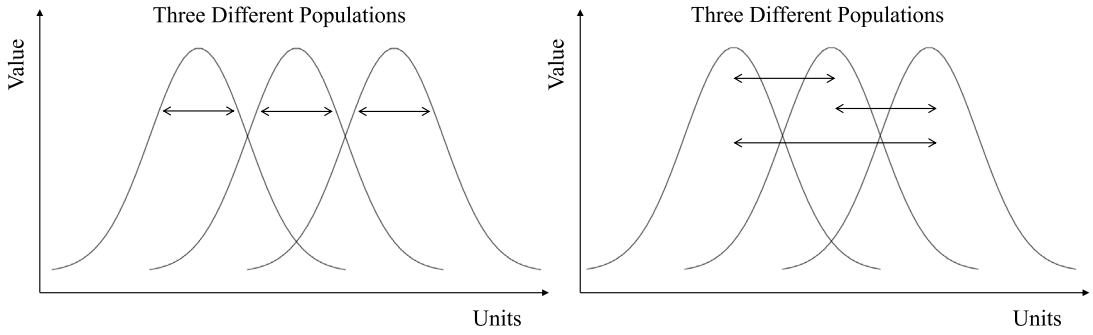

Fig. 1. Within-population (left) and between-population (right) variations. The x-axis represents different individual units, and the y-axis represents certain quantifiable features or characteristics associated with these units.

between the hazard function and covariates, Wen et al. [14] developed a neural network (NN) based proportional hazard model for the hard failure prediction. Specifically, in their joint prognostic modeling framework, the degradation signals were characterized through a mixed-effects model while the NN-based Cox model was employed for time-to-event data. All these existing models consider both the stochastic occurrence of hard failures and unit-to-unit variations, which contribute to the superior prognostic performance.

Nevertheless, most of existing methods assume that the stochastic behavior of each unit follows the same distribution. In other words, it is assumed that the units are from the same population, i.e., a group of units that share some similar characteristics. In practice, however, these units can be categorized into different populations based on their types, working environments, or other factors. Examples such as different types of batteries [15], bearings in rotating machines that suffer different failure modes [16], turbofan engines under multiple operating conditions [17] all represent different populations. In these cases, as shown in Fig. 1, there are variations not only among individuals within a population but also between different populations. An intuitive strategy to deal with such within-group and between-group variations is to ignore the between-group variation and model each group independently. However, it will reduce the modeling capability and simultaneously increase the computational complexity. More specifically, the data availability of each population may be different due to data acquisition cost and constraints, resulting in imbalanced data among different populations. In such case, while the data-rich populations can be well modeled, the data-scarce populations may suffer sub-optimal performance or over-fitting, leading to inferior RUL prediction [18,19].

A feasible solution to the data scarcity problem is to model the variation relationship among different populations so that the information in data-rich populations can be extracted and shared with the data-scarce populations. Such strategy is commonly known as transfer learning, and has been effectively applied to alleviate the data insufficiency in RUL prediction [20–26]. For example, transfer learning based approaches have been used in survival analysis problems such as predicting the death of patients in multiple cancer types [27,28]. Li, et al. [29] developed a transfer learning based Cox method, termed as Transfer-Cox, to distill valuable knowledge from the source domain and transfer it to the target domain. Its effectiveness in bolstering the predictive accuracy was demonstrated in the Cancer Genome Atlas (TCGA) dataset [30]. Wang, et al. [31] proposed a multi-task survival analysis approach and presented two models called Cox-TRACE and Cox-cCMTL to simultaneously learn multiple related survival prediction tasks and benefit from the tasks' relatedness. Based on these works, Wang, et al. [32] further introduced a novel cluster-boosted multitask learning framework and Liu, et al. [33] developed an asymmetric graph-guided multitask learning approach which incorporated self-paced learning. Despite the successful application of transfer learning in RUL prediction, existing methods mainly focus on between-population knowledge transfer and ignore the within-population variations. This is a significant limitation for practical use since units in any population may degrade heterogeneously due to the within-population variations. As a result, it is imperative to develop transfer learning methods for RUL prediction of hard failures by incorporating both within and between population variations.

In this article, we propose a novel transfer learning framework tailored for the Cox PH model to deal with the within and between population variations. Specifically, we integrate the Cox method with the Bayesian hierarchical model, denoted as Cox-BHM, to capture the within and between population variations embedded in degradation signals and time-to-event data. In this hierarchical model, we deploy a shared prior distribution to the parameters of the Cox model in each population, which builds the foundation for transfer learning across different populations. The within-population variations are modeled by a linear mixed-effects model to represent heterogeneous degradation data of each unit. The estimation of the model parameters in the framework has two stages: offline stage and online stage. In the offline stage, we first estimate the parameters that can reflect population-to-population characteristics such as the baseline hazard and the parameters in Cox model. In the online stage, the parameters of the population to which the in-process unit belongs will be updated first under the Bayesian hierarchical modeling framework, and then the parameters of the in-process unit will be updated as more online degradation signals are collected. Ultimately, the RUL prediction is achieved for the in-process unit by considering within and between population variations.

The rest of this article is organized as follows. Section 2 describes the details of the proposed Cox-BHM, including problem formulation and modeling, parameters estimation and online updating and RUL prediction. The effectiveness of proposed method is tested and verified in Sections 3 and 4 through a simulation study and a real case study. Section 5 concludes this article and discusses future research directions.

2. Bayesian hierarchical Cox PH model

2.1. Problem modeling and formulation

As the name suggests, within-population variations typically occur among individuals from the same population due to unit-to-unit heterogeneity [34]. For example, even if all units are of the same model and operate under similar conditions, each unit may degrade at a different rate due to unit-specific factors such as manufacturing differences. Between-population variations generally occur among different populations, caused by population heterogeneity. For instance, turbine engines operating under different conditions may exhibit different degradation behaviors due to varying environments. In this section, we will firstly employ the linear mixed-effects model to characterize the within-population variations. Subsequently, the between-population variations will be modeled by using the Cox model based on Bayesian hierarchical framework. Different from previous studies in survival analysis which

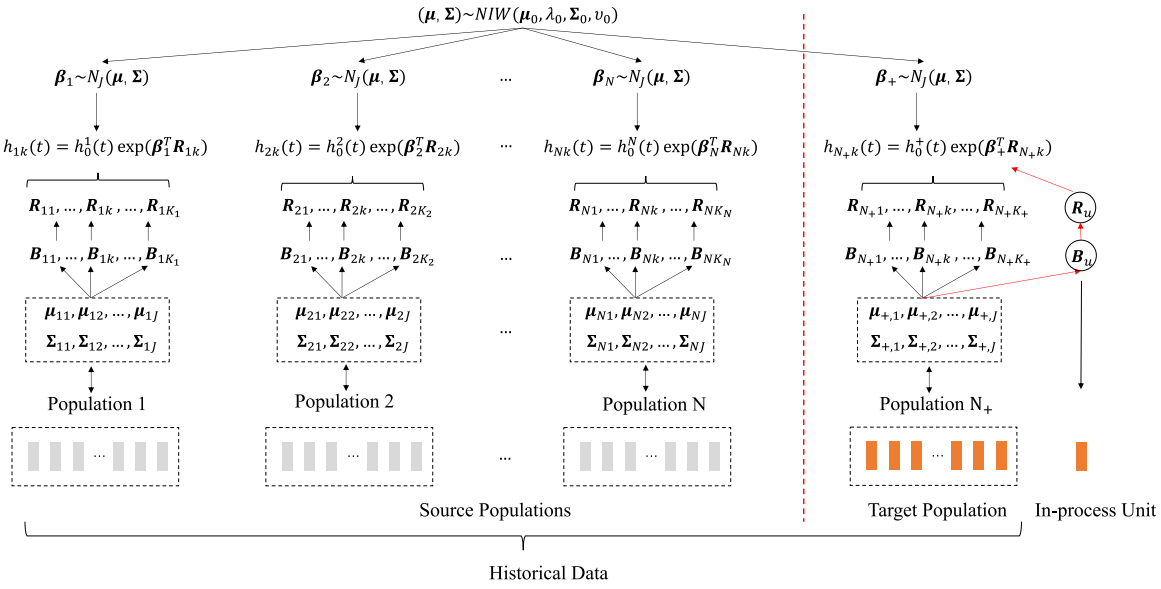

Fig. 2. The overall structure of the proposed Bayesian hierarchical model.

primarily focus on modeling risks at the population level, our approach aims to achieve the failure prediction for each individual unit.

Without loss of generality, we assume that there are N populations and each population has K_i units for $i = 1, \ldots, N$. For unit k in population i, suppose there are J observed sensor signals with each signal observed at time $t = t_1, t_2, \ldots, t_{\tau_{ik}}$, where τ_{ik} denotes the total number of observations and $t_{\tau_{ik}}$ represents the last observed time. Then the collected signal data $\boldsymbol{X}_{ik}$ can be denoted as:

$$\boldsymbol{X}_{ik} = \begin{bmatrix} X_{ik,1}(t_1) & X_{ik,2}(t_1) & \cdots & X_{ik,J}(t_1) \\ X_{ik,1}(t_2) & X_{ik,2}(t_2) & \cdots & X_{ik,J}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ X_{ik,1}\left(t_{\tau_{ik}}\right) & X_{ik,2}\left(t_{\tau_{ik}}\right) & \cdots & X_{ik,J}\left(t_{\tau_{ik}}\right) \end{bmatrix}, \tag{1}$$

where each element $X(t)$ represents the signal value at time t. Besides, the recorded event time is $F_{ik} = \min\{T_{ik}, C_{ik}\}$ (the unit failed at time T_{ik} or censored at time C_{ik}) and the event indicator is δ_{ik} ($\delta_{ik} = 0/1$ shows the unit has failed/censored). Then for unit k of population i, the associated data are denoted as $D_{ik} = \{\boldsymbol{X}_{ik}, F_{ik}, \delta_{ik}\}$ and for population i, the full observed data are $D_i = \left\{D_{i1}, D_{i2}, \ldots, D_{iK_i}\right\}$.

We first adopt a mixed-effects model to characterize the degradation signals of units since these signals reflect the health condition of the unit. It is documented that the unit-to-unit variations within a population can be well captured by the random coefficients in the mixed-effects model [10,35,36]. For sensor j of unit k from population i, the following model is employed to describe its degradation path:

$$X_{ik,j}(t) = r_{ik,j}(t) + \varepsilon_{ik,j} = \boldsymbol{Z}_j(t)\boldsymbol{B}_{ik,j} + \varepsilon_{ik,j}, \tag{2}$$

where $r_{ik,j}(t)$ represents the true but unobservable value of the degradation signal, $\varepsilon_{ik,j}$ is an independent and identically distributed white noise and follows a Normal distribution $N(0, \sigma_{ij}^2)$, $\boldsymbol{Z}_j(t) = [1, t, t^2, \ldots, t^{q_j}]$ and $\boldsymbol{B}_{ik,j}$ is a vector of random coefficients with $q_j + 1$ dimensions. $\boldsymbol{B}_{ik,j}$ is posited to follow a multivariate Normal distribution $N(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})$.

The within-population risk is usually described by a popular model named Cox PH model [37]. The hazard function of unit k from population i is assumed to have the following form:

$$h_{ik}(t) = h_0^i(t)\exp\left(\boldsymbol{\beta}_i^T \boldsymbol{R}_{ik}(t)\right), \tag{3}$$

where $h_0^i(t)$ is the baseline hazard function for population i, $\boldsymbol{R}_{ik}(t) = [r_{ik,1}(t), r_{ik,2}(t), \ldots, r_{ik,J}(t)]^T$ is time-dependent covariates which represents the unit's true degradation path, and $\boldsymbol{\beta}_i$ is a J dimensional vector associated with the covariates.

It is clear that Eq. (3) treats each population independently and does not leverage their inherent relationship. $\boldsymbol{\beta}_i$ directly characterize the impact of covariates on the hazard function and represents the main characteristics at the population level. To capture the variations cross populations and facilitate the transfer learning, we assume each $\boldsymbol{\beta}_i$ is sampled from the same J-dimensional Normal distribution $N_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the sample from a hyper-distribution, that is, the Normal-inverse-Wishart (NIW) distribution NIW $(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Sigma}_0, \upsilon_0)$.

As shown in Fig. 2, each population is composed of multiple units, where the gray bars represent units from the source population and the orange bars represent units from the target population. The differences between units within the same population are characterized by the parameter $\boldsymbol{B}$ of the mixed-effects model. And the between-population variation is leveraged by transfer learning through the shared hyper parameters, where the information from source populations can be transferred through the shared $\boldsymbol{\mu}_0$, λ_0, $\boldsymbol{\Sigma}_0$, υ_0 to the target population. The entire process of the proposed method can be divided into two parts. The first part is the offline modeling of multiple source populations, which involves estimating the parameters to capture the commonalities and differences between the risk models of different populations. The second part is updating the parameters and predicting the risks for in-process units of the target population. By using the estimated model parameters of the source populations as prior knowledge, and combining it with the existing data of the target population and the sensor signals of in-process units, the parameters are updated at both the population and individual levels.

2.2. Offline parameters estimation

All the model parameters can be denoted as $\psi = \left\{\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}, \sigma_{ij}^2, \boldsymbol{\beta}_i, h_0^i(t), \boldsymbol{\mu}, \boldsymbol{\Sigma}\right\}$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$. Due to the complexity of the posterior distribution, directly estimating the parameters proves to be challenging. Consequently, we adopt a two-stage approximation method [38]. In the first stage, the parameters $\left\{\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}, \sigma_{ij}^2\right\}$ in the mixed-effects model are estimated first. Subsequently, treating the mixed-effects model as given, we proceed to estimate the parameters $\left\{\boldsymbol{\beta}_i, h_0^i(t), \boldsymbol{\mu}, \boldsymbol{\Sigma}\right\}$ within the Bayesian hierarchical Cox model. The estimation of parameters in mixed-effects models can be achieved using the restricted maximum likelihood method.

As for the parameters in the second stage, we need to compute their posterior distributions. Denote $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N\}$, $D = \{D_1, D_2, \ldots,$

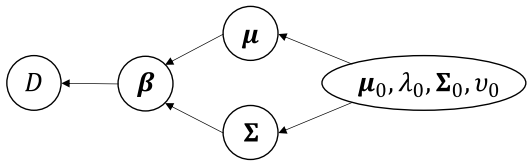

Fig. 3. A directed acyclic graph for variables.

$D_N\}$, our objective is to calculate $p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|D)$, which is expressed as

$$p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid D) \propto p(D \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(D \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (4)$$

where the three components in Eq. (4) represents the likelihood function, prior distribution of $\boldsymbol{\beta}$, and prior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, respectively. The expressions for each component are:

$$\begin{aligned} p(D \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} \prod_{k=1}^{K_i} h_{ik}\left(F_{ik}\right)^{\delta_{ik}} S_{ik}\left(F_{ik}\right) \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K_i} \left\{h_{ik}\left(F_{ik}\right) \exp\left[\boldsymbol{\beta}_i{}^T \boldsymbol{R}_{ik}\left(F_{ik}\right)\right]\right\}^{\delta_{ik}} \\ &\quad \times \exp\left\{-\int_0^{F_{ik}} h_0^i(q) \exp\left[\boldsymbol{\beta}_i{}^T \boldsymbol{R}_{ik}(q)\right] dq\right\}, \end{aligned} \quad (5)$$

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} p\left(\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-\frac{NJ}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T\right]\right\}, \end{aligned} \quad (6)$$

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{\left(\lambda_0\right)^{\frac{J}{2}} \left|\Sigma_0\right|^{\frac{\upsilon_0}{2}} |\boldsymbol{\Sigma}|^{-\frac{\upsilon_0+J+2}{2}}}{(2\pi)^{\frac{J}{2}} 2^{\frac{\upsilon_0 J}{2}} \Gamma_J\left(\frac{\upsilon_0}{2}\right)} \\ &\quad \times \exp\left[-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_0\right) - \frac{\lambda_0}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\right], \end{aligned} \quad (7)$$

where $S_{ik}(t)$ is the survival function of unit k in population i, $\Gamma_J(\cdot)$ is the multivariate gamma function and $\operatorname{tr}(\cdot)$ is the Trace of the given matrix.

Since the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|D)$ is a highly complex function of the model parameters, sampling directly from it presents a challenge. Therefore, we employ a Gibbs sampling scheme, a popular Markov chain Monte Carlo (MCMC) method to address this issue. To facilitate the sampling procedure, given the collected data and the other parameters, the conditional posterior distribution of each parameter is needed. As shown in Fig. 3, the conditional independencies among these variables are depicted by the graphical representation called directed acyclic graphs (DAGs). Leveraging the graph's conditional independence structure, we derive the following conditional distributions:

$$p(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, D) = \frac{p(D|\boldsymbol{\beta})\, p(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{p(D)} \propto p(D|\boldsymbol{\beta})\, p(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (8)$$

$$p(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, D) \propto p(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad (9)$$

$$p(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\mu}, D) \propto p(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\mu}). \quad (10)$$

The conditional distributions can be computed based on Lemma 1:

Lemma 1. *$p(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, D), p(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, D)$ and $p(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\mu}, D)$ can be sampled through MCMC. $(\boldsymbol{\mu}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, D)$ follows a multivariate Normal distribution $N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ and $(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\mu}, D)$ follows an inverse Wishart distribution $W^{-1}(\boldsymbol{\Sigma}_0{}^*, \nu^*)$, where*

$$\boldsymbol{\mu}^* = \frac{\lambda_0}{\lambda_0 + N} \boldsymbol{\mu}_0 + \frac{1}{\lambda_0 + N} \sum_{i=1}^{N} \boldsymbol{\beta}_i, \quad (11)$$

$$\boldsymbol{\Sigma}^* = \frac{\boldsymbol{\Sigma}}{\left(N + \lambda_0\right)}, \quad (12)$$

$$\boldsymbol{\Sigma}_0{}^* = \boldsymbol{\Sigma}_0 + \sum_{i=1}^{N} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T + \lambda_0 \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T, \quad (13)$$

$$\nu^* = N + \upsilon_0 + 1. \quad (14)$$

The proof is given in Appendix A. Based on Lemma 1, the overall Gibbs sampling algorithm can be summarized in Algorithm 1:

Algorithm 1 Gibbs Sampling for Cox-BHM

1: Initialize $(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\beta}^{(\mathbf{0})}, \boldsymbol{\mu}^{(\mathbf{0})}, \boldsymbol{\Sigma}^{(\mathbf{0})}$
2: **for** iteration $l = 1, 2, \cdots, L$, **do**
3: Sample $\boldsymbol{\beta}^{(l+1)} \sim p\left(\boldsymbol{\beta} \middle| \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}, D\right)$
4: Sample $\boldsymbol{\mu}^{(l+1)} \sim p\left(\boldsymbol{\mu} \middle| \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\Sigma}^{(l)}, D\right)$
5: Sample $\boldsymbol{\Sigma}^{(l+1)} \sim p\left(\boldsymbol{\Sigma} \middle| \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\mu}^{(l+1)}, D\right)$
6: **end for**
7: **return** $\left\{\boldsymbol{\beta}^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}\right\}_{l=1}^{L}$

To compute the hazard rate for the units in each population, we use the sample mean as the point estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}$ be the mean value of $\boldsymbol{\beta}$, then the baseline hazard function $h_0^i(t)$ can be estimated by the Breslow estimator [39]:

$$\hat{h}_0^i\left(t_m\right) = \frac{d_m}{\sum_{g \in \mathbb{R}(t_m)} \exp\left(\hat{\boldsymbol{\beta}}_i^T \boldsymbol{R}_{ig}\left(t_m\right)\right)}, \quad (15)$$

where $\mathbb{R}(t_m)$ denotes the set of all units at risk at time t_m in population i, $\mathbb{D}(t_m)$ is the set of units that fail at time t_m and d_m is the size of $\mathbb{D}(t_m)$.

2.3. Online parameters update and RUL prediction

During the online stage, prognostics for an in-process unit u from a new population N_+ can be made by utilizing the previously proposed model and the estimated parameters. Compared with the existing populations $1, 2, \ldots, N$, population N_+ exhibits similarities in certain aspects, such as the distribution of parameters in the risk model. As mentioned before, populations $1, 2, \ldots, N$ are referred to as the source populations while population N_+ is the target population. Suppose the new population consists of K_+ units, and similar to these K_+ units, unit u can be viewed as an individual sampled from the overall population. For a more intuitive representation, Fig. 2 illustrates the relationship among them.

Given the collected data of the new population as well as the degradation signals of unit u gathered until time t^*, our primary objective is to predict the RUL of unit u. Denote the collected sensor signals of unit u as $\boldsymbol{X}_u|_{t=1:t^*}$, and to achieve a precise RUL prediction, the parameters of the mixed-effects model are required to be updated first by integrating both historical data and information specific to unit u. Since $X_{u,j}(t) = \boldsymbol{Z}_{u,j}(t)\, \boldsymbol{B}_{u,j} + \varepsilon_{u,j}$, we can prove that the posterior distribution $p\left(\boldsymbol{B}_{u,j}|\boldsymbol{X}_u|_{t=1:t^*}\right)$ of sensor j is also multivariate Normal under the Bayesian framework [10,14]. The corresponding mean $\hat{\boldsymbol{\mu}}_{u,j}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{u,j}$ can be updated as follows:

$$\hat{\boldsymbol{\mu}}_{u,j} = \hat{\boldsymbol{\Sigma}}_{u,j} \left[\left(\hat{\boldsymbol{\Sigma}}_{+,j}\right)^{-1} \hat{\boldsymbol{\mu}}_{+,j} + \frac{\left(\boldsymbol{Z}_{u,j}|_{t=1:t^*}\right)^T \boldsymbol{X}_u|_{t=1:t^*}}{\hat{\sigma}_{+,j}^2}\right], \quad (16)$$

$$\hat{\boldsymbol{\Sigma}}_{u,j} = \left[\frac{\left(\boldsymbol{Z}_{u,j}|_{t=1:t^*}\right)^T \boldsymbol{Z}_{u,j}|_{t=1:t^*}}{\hat{\sigma}_{+,j}^2} + \left(\hat{\boldsymbol{\Sigma}}_{+,j}\right)^{-1}\right]^{-1}, \quad (17)$$

where $\hat{\boldsymbol{\mu}}_{+,j}, \hat{\boldsymbol{\Sigma}}_{+,j}, \hat{\sigma}_{+,j}^2$ are the estimated parameters of the mixed-effects model for the new population N_+.

Thus, we have obtained the parameters to depict the characteristics of unit u at the individual level. Assume the hazard function of unit u

is

$$h_u(t) = h_0^+(t)\exp\left(\boldsymbol{\beta}_+^T \hat{\boldsymbol{R}}_u(t)\right), \tag{18}$$

where $\boldsymbol{R}_u(t)$ can be fully characterized by the parameters estimated in Eqs. (16) and (17), $h_0^+(t)$ and $\boldsymbol{\beta}_+$ are parameters in the risk model to characterize the overall population. Obviously, for $h_0^+(t)$ and $\boldsymbol{\beta}_+$, once we get the estimated value of $\boldsymbol{\beta}_+$, $h_0^+(t)$ can be calculated through Eq. (15). Since the prior distribution of $\boldsymbol{\beta}_+$ is known, the next work is to estimate its posterior distribution. To leverage the knowledge learned from the source populations, the updating framework for $\boldsymbol{\beta}_+$ is shown as follows:

$$p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N, D_+\right) \propto p\left(D_+|\boldsymbol{\beta}_+\right) p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right), \tag{19}$$

where $p\left(D_+|\boldsymbol{\beta}_+\right)$ is the partial likelihood of the new population and $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right)$ represents the prior knowledge transferring from the source populations, which is expected to benefit the estimation of $\boldsymbol{\beta}_+$. In Eq. (19), $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right)$ can be obtained based on Lemma 2.

Lemma 2.

$$p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right) = \int p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right) \prod_{i=1}^{N} p\left(\boldsymbol{\beta}_i|D_i\right) d\boldsymbol{\beta}, \tag{20}$$

where $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N\right\}$, $p\left(\boldsymbol{\beta}_i|D_i\right)$ *has been calculated by Gibbs Sampling in the offline stage, and* $p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right)$ *follows a t-distribution:*

$$p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right) \sim t_{\upsilon_N - p + 1}\left(\boldsymbol{\mu}_N, \frac{\boldsymbol{\Sigma}_N\left(\lambda_N + 1\right)}{\lambda_N\left(\upsilon_N - p + 1\right)}\right). \tag{21}$$

The proof is given in Appendix B. As can be seen from Eq. (20), the N source populations influence the distribution of $\boldsymbol{\beta}_+$ through $\prod_{i=1}^{N} p\left(\boldsymbol{\beta}_i|D_i\right)$, which is the product of the posterior distributions of their parameters. It functions by leveraging information from all source populations to infer the underlying population parameters, and then use them to derive the prior distribution for the new population. Hence, by virtue of this mechanism, the information from the source populations is effectively "transferred" into the modeling process for the target population.

To get samples from Eq. (19), which has no closed form, we resort to the importance resampling method, where the importance distribution is simply $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right)$ and the weight function is set as $p\left(D_+|\boldsymbol{\beta}_+\right)$. According to Eq. (20), sampling from the distribution $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right)$ can be performed in two steps. The first step is to get samples $\boldsymbol{\beta}$ from $p\left(\boldsymbol{\beta}|D_1, D_2, \ldots, D_N\right)$, which has already been done by the Gibbs sampling in the offline parameters estimation stage. We use S to denote the number of samples and $\left\{\boldsymbol{\beta}^{(s)}\right\}_{s=1}^{S}$ denote the sample set. The second step is to sample $\boldsymbol{\beta}_+^{(s)}$ according to $p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}^{(s)}\right)$ shown in Eq. (21) based on $\boldsymbol{\beta}^{(s)}$ for each s. Then the sample set $\left\{\boldsymbol{\beta}_+^{(s)}\right\}_{s=1}^{S}$ follows the distribution $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N\right)$. Furthermore, we assign each sample $\boldsymbol{\beta}_+^{(s)}$ the weight $p\left(D_+|\boldsymbol{\beta}_+^{(s)}\right)$, which can be computed by Eq. (A.3) in Appendix A. Denote the normalized weight $W_s = \frac{p\left(D_+|\boldsymbol{\beta}_+^{(s)}\right)}{\sum_{l=1}^{S} p\left(D_+|\boldsymbol{\beta}_+^{(l)}\right)}$, for $s = 1, 2, \ldots, S$. Finally, we will resample S samples from the sample set $\left\{\boldsymbol{\beta}_+^{(s)}\right\}_{s=1}^{S}$ where for each s the sample $\boldsymbol{\beta}_+^{(s)}$ will be selected with probability W_s to make the samples $\left\{\tilde{\boldsymbol{\beta}}_+^{(s)}\right\}_{s=1}^{S}$ follow the distribution $p\left(\boldsymbol{\beta}_+|D_1, D_2, \ldots, D_N, D_+\right)$.

Once the posterior distribution of $\boldsymbol{\beta}_+$ is obtained, the baseline hazard function of the new population $h_0^+(t)$ can be estimated through Eq. (15). Then with the estimated population-level parameters $\left\{\hat{\boldsymbol{\beta}}_+, \hat{h}_0^+(t)\right\}$ and the individual-level parameters $\left\{\hat{\boldsymbol{\mu}}_{u,j}, \hat{\boldsymbol{\Sigma}}_{u,j}\right\}$ for $j = 1, 2, \ldots, J$, we can calculate the marginal survival function of unit u as follows:

$$S\left(t|t^*, \boldsymbol{X}_u|_{t=1:t^*}\right) = \int S\left(t|t^*, \boldsymbol{B}_u\right) p\left(\boldsymbol{B}_u|\boldsymbol{X}_u|_{t=1:t^*}\right) d\boldsymbol{B}_u, \tag{22}$$

where $\boldsymbol{B}_u = \boldsymbol{B}_{u,1}, \boldsymbol{B}_{u,2}, \ldots, \boldsymbol{B}_{u,p}$ and $S\left(t|t^*, \boldsymbol{B}_u\right)$ is the survival function:

$$S\left(t|t^*, \boldsymbol{B}_u\right) = \exp\left\{-\int_{t^*}^{t} \hat{h}_0^+(q)\exp\left(\hat{\boldsymbol{\beta}}_+^T \boldsymbol{Z}(q)\,\boldsymbol{B}_u\right) dq\right\}. \tag{23}$$

Eq. (23) can be computed by Monte Carlo simulation and the integration can be done by Gauss–Legendre quadrature method [40] as shown in Wen, et al. [14]. After obtaining the estimated marginal survival function $\hat{S}\left(t|t^*, \boldsymbol{X}_{u,1:t^*}\right)$, the remaining useful life of unit u can be calculated by the following expression:

$$RUL\left(t^*\right) = \int_{t^*}^{\infty} \hat{S}\left(t|t^*, \boldsymbol{X}_u|_{t=1:t^*}\right) dt. \tag{24}$$

3. Simulation study

To demonstrate the effectiveness of the proposed method, we conduct a simulation study in this section. Four benchmarks Cox PH model [37], Cox-L_{21} [29], Cox-Trace [31] and Cox–Zhou [10] are chosen to compare the performance with our method. The first benchmark is the traditional Cox PH model, which is applied to characterize the risk for one single population. The second and third benchmark methods are transfer learning approaches applied in survival analysis. They are both designed to address the risk assessment issue for the patients in multiple cancer types, and they can be regarded as only concentrating on modeling the between-population variations and ignoring the within-population variations. We refer these two models as Cox-L_{21} and Cox-Trace, respectively, where the former employs the $l_{2,1}$-norm to encourage multiple coefficient vectors to share similar sparsity patterns, while the latter assumes the estimated coefficients from different tasks sharing a low-dimensional subspace. The final benchmark is the method proposed by Zhou [10], which considers the within-population variations but ignore the between-population variations, i.e., only focus on one population and do not apply transfer learning, and we denote it as Cox–Zhou.

The simulation experiment is divided into two parts. The first part focuses on the offline stage, primarily comparing the characterization of the source populations by Cox, Cox-L_{21}, Cox-Trace, and Cox-BHM, including the estimation errors of the parameters and the mean concordance-index (C-index) values. It is worth noting that the concordance index [41], or concordance probability, is one of the most commonly applied evaluation metrics in survival analysis. It measures the proportion of all usable patient pairs whose predicted survival times are correctly ordered. Specifically, a higher C-Index indicates better model performance in terms of predicting the order of events. The second part concentrates on the online stage and compares the performance of Cox–Zhou, Cox-L_{21}, Cox-Trace and Cox-BHM in risk modeling and remaining useful life prediction for in-process units of the new population.

The simulation settings are as follows. We set $N = 10$ and $K_i = 100$ for $i = 1, 2, \ldots, N$, which means there are 10 source populations, and each population has 100 units. By generating multiple source populations, we aim to capture a diverse range of degradation behaviors and failure patterns, which helps to improve the prediction accuracy for the target population and enhance the robustness and generalizability of our transfer learning model. For each unit, we suppose that the degradation signals are collected from two sensors. And for unit k of population i, signals are assumed as:

$$X_{ik,j}(t) = \boldsymbol{Z}_j(t)\,\boldsymbol{B}_{ik,j} + \varepsilon_{ik,j}, \quad j = 1, 2. \tag{25}$$

The parameters used in Eq. (25) are listed in Table 1.

Fig. 4 shows 16 randomly generated degradation signals based on Eq. (25).

To generate the hazard rate function, we generate the covariates-associated parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{10}$ from Normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Given the degradation signals of unit k in population i, the related hazard rate function for generating time-to-event data is

$$h_{ik}(t) = h_0(t)\exp\left[w * \left(\beta_{ik,1}\boldsymbol{Z}_1(t)\boldsymbol{B}_{ik,1} + \beta_{ik,2}\boldsymbol{Z}_2(t)\boldsymbol{B}_{ik,2}\right)\right], \tag{26}$$

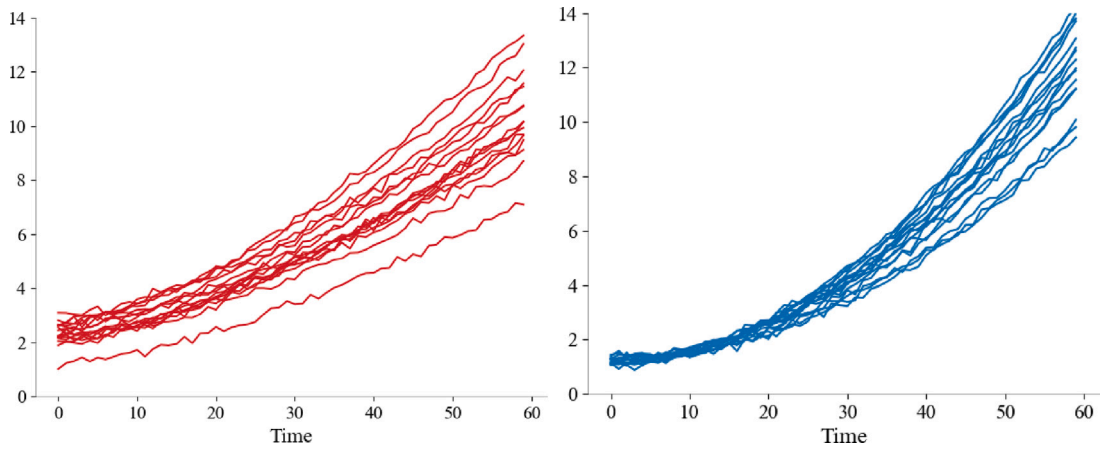

Fig. 4. Generated two degradation signals for each unit.

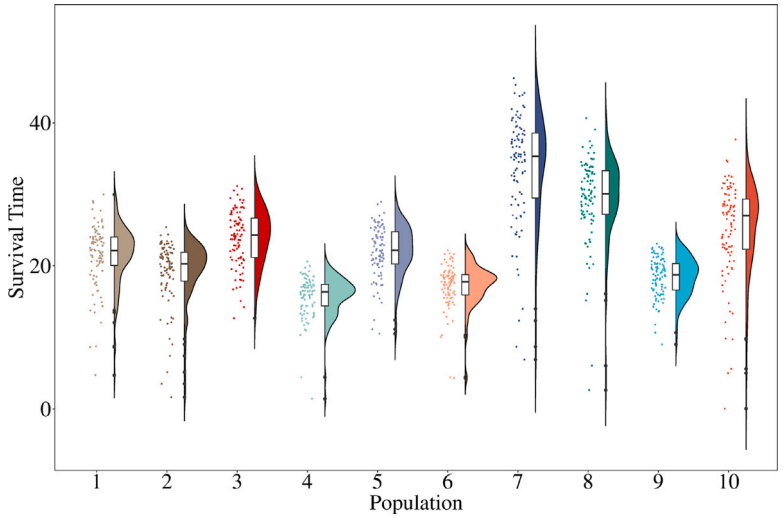

Fig. 5. Units' failure time distribution in 10 populations.

Table 1
Parameter settings for signals generation.

Parameters	$j=1$	$j=2$
$\boldsymbol{Z}^j(t)$	$[1, t^{1.2}, t^{1.6}]^T$	$[1, t, t^2]^T$
$\boldsymbol{\mu}_{i,j}$	[2.4, 0.01, 0.01]	[1.2, 0.005, 0.003]
$\boldsymbol{\Sigma}_{i,j}$	$\begin{bmatrix} 0.2 & -4e{-4} & 7e{-5} \\ -4e{-4} & 3e{-6} & 1e{-7} \\ 7e{-5} & 1e{-7} & 3e{-6} \end{bmatrix}$	$\begin{bmatrix} 0.01 & -3e{-4} & 4e{-7} \\ -3e{-4} & 2e{-7} & 2e{-7} \\ 4e{-7} & 2e{-7} & 2e{-7} \end{bmatrix}$
σ_j^2	0.01	0.01

where $h_0(t)$ is the Weibull baseline hazard rate function and

$$h_0(t) = \lambda\alpha t^{\alpha-1} = 0.001 \times 1.05 t^{1.05-1}. \tag{27}$$

w is a tuning parameter and here $w = 0.1$.

With the hazard function in (18), the probability density function of the failure time of unit k in population i can be computed as follows:

$$f_{ik}(t) = h_{ik}(t) S_{ik}(t) = h_{ik}(t) \exp\left(\int_0^t -h_{ik}(s)\, ds\right), \tag{28}$$

where $S_{ik}(t)$ is the survival function. Fig. 5 displays the overall distribution of units' failure time in 10 different populations. The percentage of censored units for each population is set to 5%. Table 2 presents a detailed description of data generation process.

Once obtaining the generated data, four models Cox-BHM, Cox, Cox-$L_{2,1}$, Cox-Trace can be applied to estimate the population-level parameters for each population. After conducting 100 experiments, we first compared the mean absolute errors (MAE) between the estimated parameters and the true values, which is defined as

$$MAE = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{i}^{N}\left|\hat{\boldsymbol{\beta}}_{il} - \boldsymbol{\beta}_{il}\right|, \tag{29}$$

where N is the number of generated populations, L denotes the number of experiments and $\hat{\boldsymbol{\beta}}_{il}$ represents the estimated parameter for population i in lth experiment while $\boldsymbol{\beta}_{il}$ is the true value.

Fig. 6 presents the mean estimation error of parameters using four different methods in 100 experiments, while Fig. 7 demonstrates a scatter plot and a box plot of the errors. From these two figures, it is clear that the performance of the method we proposed surpasses that of the other three benchmark methods, exhibiting lower errors in parameter estimation. In addition, we also compared the mean C-index values of ten populations in 100 experiments by using different models. And the calculation method of the mean C-index we used is shown as Eq. (30):

$$Mean\ CI = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{i}^{N} CI_{il}, \tag{30}$$

where CI_{il} represents the C-index value for population i in lth experiment.

Table 3 displays the performance results of the mean C-index values of different models. The result indicates that three transfer learning

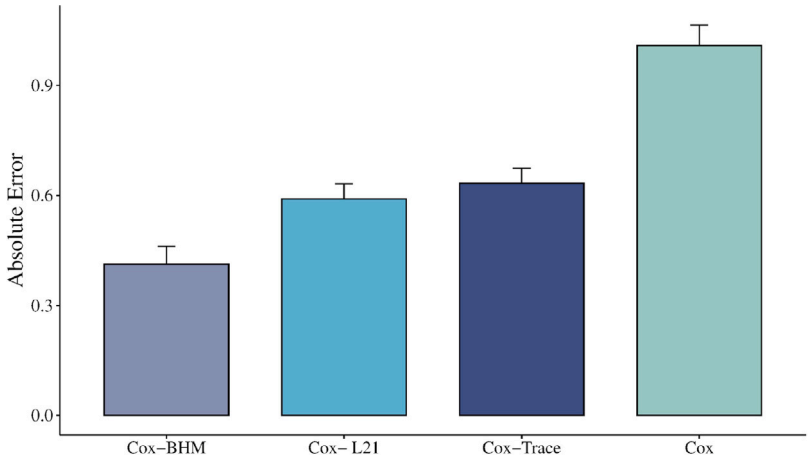

Fig. 6. The mean errors of parameters for four models.

Table 2
The detailed procedure of data generation.

(1)	Generate $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N$ for N populations from Normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are sampled according to $NIW(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Sigma}_0, \upsilon_0)$.
(2)	Generate signals' parameters of 100 units in each population. Specifically, for sensor j of unit k in population i, generate $\boldsymbol{B}_{ik,j}$ from $N(\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j})$.
(3)	Generate the failure time T_{ik} for each unit by drawing random samples using reject sampling from $f_{ik}(t)$ as described in Eq. (28). Select 5% of the units in each population as censored, and the censoring time C_{ik} is sampled from uniform distribution $U(1, T_{ik})$.
(4)	Degradation signals of unit k in population i are generated with noise according to Eq. (25) until its time of failure or censoring.

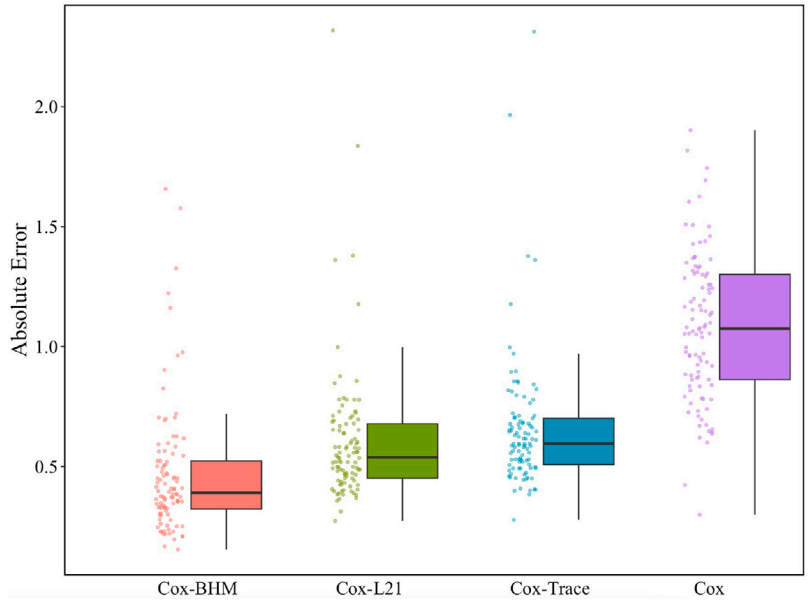

Fig. 7. Boxplots for the mean errors of parameters in 100 experiments.

Table 3
Performance comparison of different models using mean C-index values.

Models	Cox-BHM	Cox	Cox-$L_{2,1}$	Cox-Trace
Mean C-index	**0.8046**	0.6442	0.7408	0.7393

models perform better than traditional Cox model, and the performance of our model is superior to Cox-$L_{2,1}$ and Cox-Trace.

To further compare the performance of Cox-BHM, Cox-$L_{2,1}$, Cox-Trace and Cox–Zhou, during the online stage, we first generate 10 source populations and one target population consisting of K_+ units. For Cox-BHM, by employing the parameter updating steps described in the second section, we can obtain the estimation of relevant parameters for the target population. And for a new in-process unit u from the target population, we can then estimate its risk at a specific moment t^* as well as its remaining useful life. The detailed procedure for online stage is shown in Table 4.

During the online stage, we set two variables, K_+ and t^*, with different values. The K_+ represents the size of the target population, which affects the accuracy of $p\left(D_+|\boldsymbol{\beta}_+\right)$ in Eq. (19). The t^* denotes the observed length of degradation signals for the in-process unit u. Typically, if K_+ is very small, the information about the target population will be inadequate, leading to less accurate estimation of population-level parameters. Similarly, if t^* is very small, there will be less information about unit u, resulting in larger errors when predicting the remaining useful life. Due to the different underlying mechanisms of Cox-BHM,

Table 4
The simulation procedure for online stage.

(1)	Generate ten source populations and one target population consisting of K_+ units and K_+ = 10, 40, 80 respectively.
(2)	Update the posterior distribution of β_+ for the target population based on Eq. (19), and take the average as the estimated value of Cox-BHM.
(3)	Calculate the estimated parameters of the target population by using Cox–Zhou, Cox-$L_{2,1}$, Cox-Trace.
(4)	Generate degradation signals for an operating unit u similar to units in the target population until the time instant of prediction t^* and t^* = 5, 15, 25, 35.
(5)	Use four models to calculate the RUL of unit u and then compare the prediction result with its true value.
(6)	Repeat (4) and (5) for 1000 times to calculate the mean prediction errors.

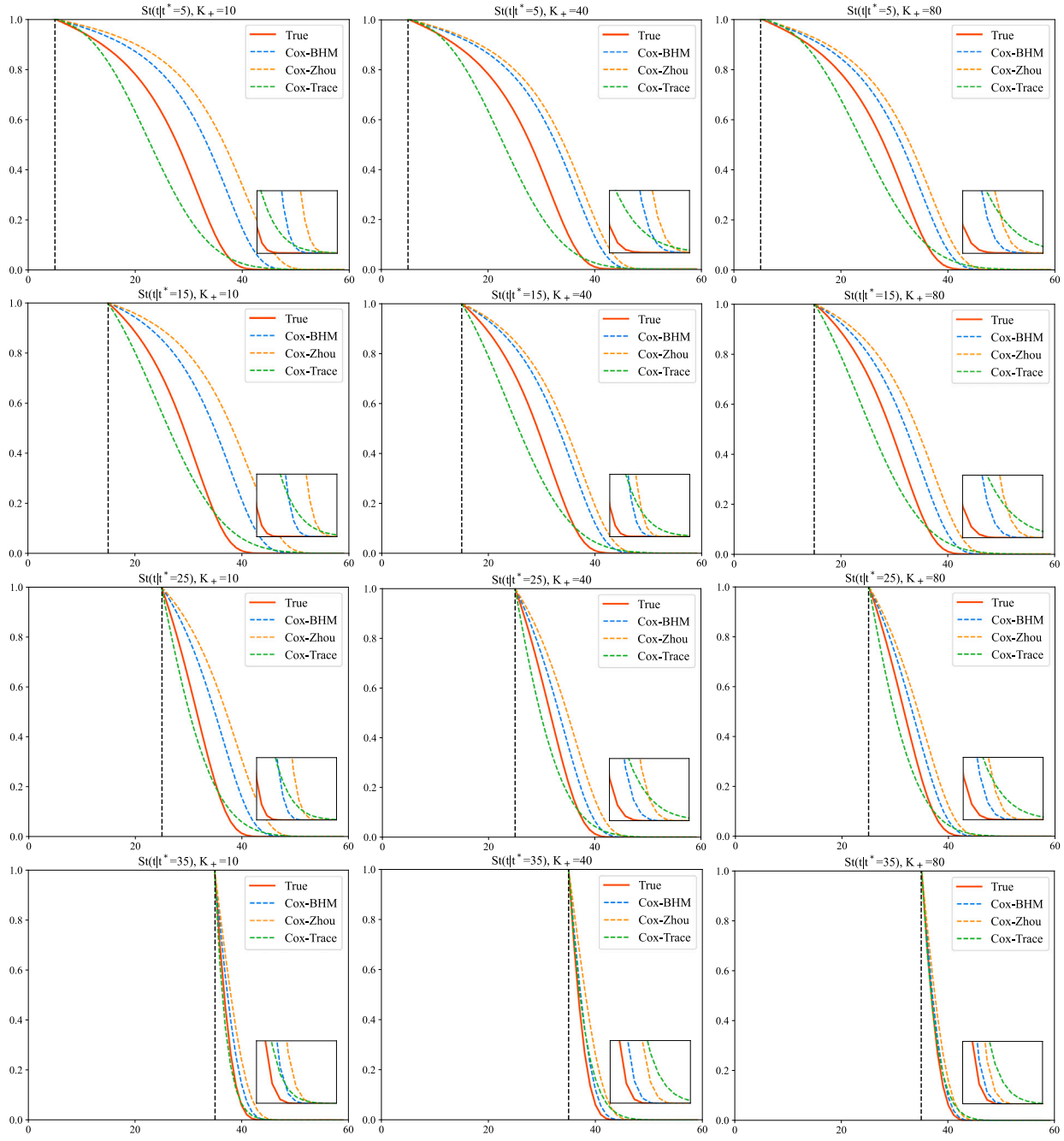

Fig. 8. Comparison between true and estimated conditional survival curves for in-process units using different models at t^* = 5, 15, 25, 35 and K_+ = 10, 40, 80.

Cox–Zhou, Cox-$L_{2,1}$ and Cox-Trace, their prediction performance varies for different K_+ and t^*.

Fig. 8 shows a comparison between true and estimated conditional survival curves for in-process units using different models at t^* = 5,15,25,35 and K_+ = 10, 40, 80 (The survival curve of Cox-$L_{2,1}$ is not included in Fig. 8 as its performance is not significantly different from that of Cox-Trace). Table V presents the mean RUL prediction errors of 1000 Monte Carlo samples, which are calculated as Eq. (31), where $\hat{R}_m(t^*)$ and $R_m(t^*)$ are the predicted and true RUL of sample m at time t^*, respectively. From Fig. 8 and Table 5, it is obvious that

Table 5
The mean RUL prediction errors of different models with different K_+ and t^*.

t^*	$K_+ = 10$				$K_+ = 40$				$K_+ = 80$			
	Cox-BHM	Cox-$L_{2,1}$	Cox-Trace	Cox-Zhou	Cox-BHM	Cox-$L_{2,1}$	Cox-Trace	Cox-Zhou	Cox-BHM	Cox-$L_{2,1}$	Cox-Trace	Cox-Zhou
$t^* = 5$	**6.7461**	7.5351	7.6322	9.7214	**5.0021**	7.4402	7.3920	6.1359	**4.1333**	6.3013	6.3142	4.5240
$t^* = 15$	**5.6420**	7.7391	7.9831	8.3308	**4.1895**	7.8103	7.9234	5.3340	3.5498	6.0249	5.8459	**3.4934**
$t^* = 25$	**4.5766**	7.4063	7.3405	7.8120	**2.2310**	7.0142	6.9734	4.3593	**1.6430**	5.9402	6.3345	2.0205
$t^* = 35$	**3.0521**	6.8964	6.7241	6.9912	**1.0935**	6.9013	7.0132	2.7714	**0.3745**	5.6839	5.5580	0.8920

Table 6
Settings of four health condition indices.

Population	Form	a	b
Source Population 1	$\Gamma_1(t) = a_1 + b_1 t$	$a_1 \sim N(0.25, 0.015^2)$	$b_1 \sim N(0.15, 0.01^2)$
Source Population 2	$\Gamma_2(t) = a_2 t + b_2 t^2$	$a_2 \sim N(0.01, 0.001^2)$	$b_2 \sim N(0.003, 0.0003^2)$
Source Population 3	$\Gamma_3(t) = e^{a_3 + b_3 t}$	$a_3 \sim N(0.008, 0.004^2)$	$b_3 \sim N(0.04, 0.001^2)$
Target Population	$\Gamma_4(t) = 0.6\Gamma_2(t) + 0.3\Gamma_3(t)$	/	/

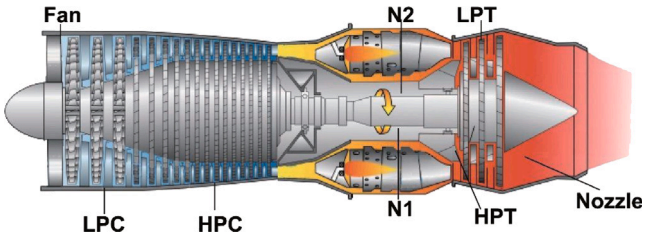

Fig. 9. Simplified engine diagram simulated in the software.

when the sample size of the target population is limited, e.g. $K_+ = 10$, the transfer learning methods, Cox-BHM, Cox-$L_{2,1}$ and Cox-Trace, outperform Cox–Zhou in predicting the RUL of in-process units. This superior performance can be attributed to the knowledge transferred from the source populations. However, as the sample size of the target population increases, the advantages of Cox-BHM, Cox-$L_{2,1}$ and Cox-Trace over Cox–Zhou begin to diminish. Notably, at $K_+ = 40$ and $K_+ = 80$, the predictive performance of Cox–Zhou surpasses that of Cox-Trace. Since both Cox-BHM and Cox–Zhou account for within-population variations and dynamically update the parameters as more observations for individuals become available, their prediction errors tend to decrease over time. In contrast, Cox-Trace and Cox-$L_{2,1}$ does not exhibit this behavior. In summary, Cox-BHM, by updating parameters within and between populations, effectively combines the advantages of both Cox-Trace and Cox–Zhou.

$$Mean\ Error = \frac{1}{L}\sum_{l=1}^{L}\left|\hat{R}_l\left(t^*\right) - R_l\left(t^*\right)\right|. \tag{31}$$

4. Case study on gas turbine engine dataset

In this section, we use the dataset of aircraft gas turbine engines generated from a high-fidelity turbine engine simulation software called TEACHES [42] to evaluate the performance of the proposed method. In this software, users are able to input indicative parameters of health condition of different degradation modes, thus obtaining turbine engines which experience different degradation processes. Fig. 9 shows a schematic diagram of a commercial aircraft gas turbine engine. Here, we choose fan outer flow drop, fan inner efficiency drop, and HP turbine flow drop as the degradation modes of three source populations, respectively. And HP compressor flow drop is selected as the degradation mode of the target population. The health condition indices of four failure modes are assumed to follow different forms as shown in Table 6.

With the input health indices, the software will simulate the degradation process of each unit accordingly. And for each generated turbine engine, 16 sensor signals are collected to monitor its health condition. Detailed information of these sensors is given in Table 7.

Table 7
Description of 16 TEACHES outputs.

Symbol	*Description*	*Units*
NL	LP shaft speed	rpm
NH	HP shaft speed	rpm
P13	Fan outer outlet pressure	bar
P26	HP compressor inlet pressure	bar
T26	HP compressor inlet temperature	C
P3	HP compressor outlet pressure	bar
T3	HP compressor outlet temperature	C
T6	Exhaust gas temperature	C
EPR	Engine pressure ratio	–
T13	Fan outer outlet temperature	C
P42	HP turbine outlet pressure	bar
T42	HP turbine outlet temperature	C
P5	LP turbine outlet pressure	bar
T41	HP turbine inlet temperature	C
Thrust	Thrust	Nt
Wf	Fuel flow rate	kg/s

To simulate hard failure cases, we generate the failure time for each unit by randomly sampling from its probability density function as defined in Eq. (28). And for each population, we employ the same baseline hazard function, which is a Weibull distribution with the shape and scale parameter $\lambda = 0.001$ and $\gamma = 1.05$, respectively. Fig. 10 shows the true probability density function of the failure time for one randomly selected unit from every generated population. In addition, for each source population, we generated 100 units, of which 5% were censored. For the target population, we generated 200 units, with 100 units utilized for estimating population-level parameters, and the remaining 100 units serving as in-process individuals for prognostic purposes.

Once we obtained the generated data, following the steps in simulation study, we primarily evaluated the average performance of different methods in predicting the remaining useful life of 100 in-process individuals under varying values of K_+ and t^*. We first normalized the sensor data and 11 sensor signals including NL, NH, T3, P26, P42, P3, T6, T26, T41, T42 and Thrust are screened out since they show an obvious change for all units. The mean absolute prediction error, as defined in Eq. (31), was employed as the index for assessing the model's efficacy.

From Fig. 11, it can be seen that when the sample size of the target population is relatively small, especially when $K_+ = 20$, the two transfer learning methods, Cox-BHM and Cox-Trace, have smaller prediction errors for the RUL of in-process individuals compared to Cox–Zhou. This indicates that Cox-BHM and Cox-Trace have stronger predictive power and adaptability in cases of data scarcity. This advantage mainly stems from the transfer learning mechanisms of these two models,

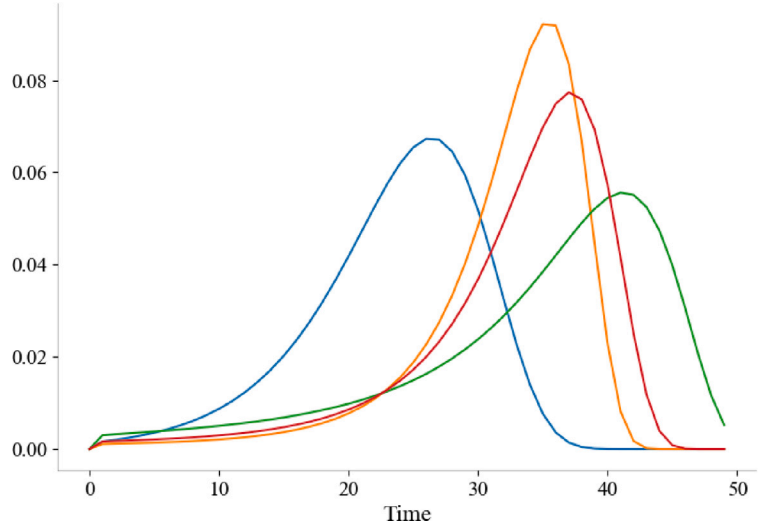

Fig. 10. True failure time distribution of four populations.

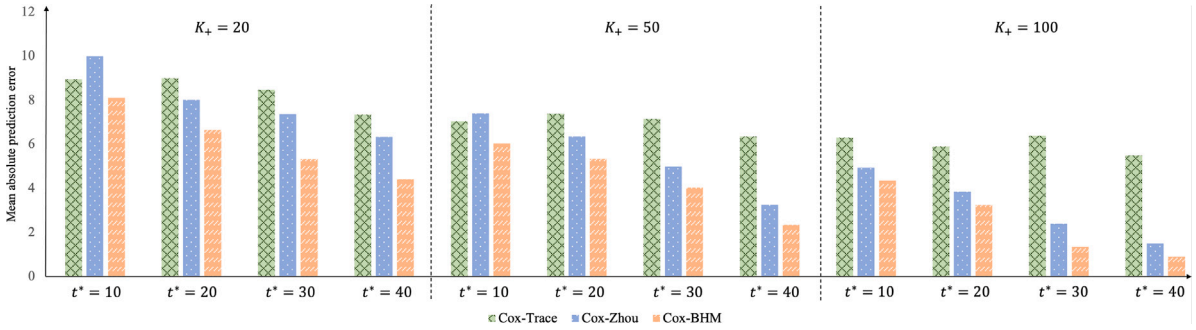

Fig. 11. The mean prediction errors of different models with different K_+ and t^*.

enabling them to more effectively utilize information from the source population to improve the accuracy of estimating the target population parameter β_+. However, as K_+ increases, the performance gap between Cox-BHM, Cox-Trace, and Cox–Zhou starts to narrow. When $K_+ = 100$, the predictive performance of Cox–Zhou even surpasses that of Cox-Trace. This is because as the size of target population data increases, Cox–Zhou can also learn more about the characteristics of the target population, leading to more accurate estimation of β_+. Additionally, as t^* increases, the prediction errors of Cox-BHM and Cox–Zhou gradually decrease, while Cox-Trace does not show the same trend. This indicates that Cox-Trace cannot dynamically update individual parameters, which is related to its modeling strategy that only considers between-population variations. Consistent with the conclusions of the simulation study, our proposed Cox-BHM possesses the capability to dynamically update both population-level and individual-level parameters, thereby exhibiting certain advantages over Cox-Trace and Cox–Zhou.

5. Conclusion

In this article, we proposed a novel transfer learning framework for the Cox model which takes into account both within and between population variations. The framework integrates the Cox method with Bayesian hierarchical modeling, aiming to achieve more accurate remaining useful life prediction for in-process units from a new population. Existing methods either consider only between-population variations or within-population variations. In comparison, our proposed approach addresses both cross-population knowledge transfer and the characterization of individual heterogeneity, leading to more accurate and individualized remaining useful life predictions. This framework presents two main advantages: on one hand, it can extract valuable information from existing source populations to better model the new target population; On the other hand, for in-process units, it can dynamically update the unit-related parameters as more degradation signals are collected over time. That is to say, it can simultaneously achieve parameters updates at both the population and individual levels, which has not been accomplished in the existing works. The proposed approach is validated through a simulation study and a case study of turbine engines. The results showed that our method can successfully capture the within and between population variations embedded in degradation signals and time-to-event data, consistently outperforming several benchmark models.

There are still several topics worthy of further investigation. First, the proposed method only considers the transfer learning of parameters associated with covariates in the Cox PH model. How to characterize the potential knowledge sharing mechanism in the baseline hazard function among different populations still remains challenging. Moreover, for units from different populations, the dimensions of collected sensor signals might not be the same in practice. In such instance, our proposed method might not be suitable. Therefore, it will be one future research direction to develop a methodology capable of addressing the problem of dimension inconsistency while facilitating knowledge transfer.

CRediT authorship contribution statement

Xinxing Guo: Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Song Huang:** Validation, Methodology, Investigation, Formal

analysis, Conceptualization. **Jianguo Wu:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Formal analysis, Data curation. **Chao Wang:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant No 72171003 and 12288101).

Appendix A. Proof of Lemma 1

For Eq. (8), given that $\boldsymbol{\mu}=\boldsymbol{\mu}^0, \boldsymbol{\Sigma}=\boldsymbol{\Sigma}^0$, the conditional distribution $p(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Sigma},D)$ can be written as:

$$p\left(\boldsymbol{\beta}|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0,D\right)\propto p(D|\boldsymbol{\beta})\,p\left(\boldsymbol{\beta}|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right)=\prod_{i=1}^{N}p\left(D_i|\boldsymbol{\beta}_i\right)p\left(\boldsymbol{\beta}_i|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right), \tag{A.1}$$

where $p\left(D_i|\boldsymbol{\beta}_i\right)$ represents the likelihood of population i and $p\left(\boldsymbol{\beta}_i|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0\right)$ is a J-dimension Normal distribution $N_J(\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0)$. By using the Cox partial likelihood, $p\left(D_i|\boldsymbol{\beta}_i\right)$ is given by

$$p\left(D_i|\boldsymbol{\beta}_i\right)=\prod_{k=1}^{K_i}\left[\frac{\exp\left(\boldsymbol{\beta_i}^T\boldsymbol{R}_{ik}\left(F_{ik}\right)\right)}{\sum_{g\in\mathbb{R}(F_{ik})}\exp\left(\boldsymbol{\beta_i}^T\boldsymbol{R}_{ig}\left(F_{ik}\right)\right)}\right]^{\delta_{ik}}, \tag{A.2}$$

where $\mathbb{R}(F_{ik})$ denotes the set of all units at risk at time F_{ik} in population i. However, Eq. (15) is not able to handle with the tied failures, i.e., two or more failure events that occur at the same time. Therefore, Efron method [43] is applied to address this issue. Let $t_1<t_2<\cdots<t_{M_i}$ be the ordered and distinct failure times of population i, $\mathbb{D}(t_m)$ be the set of individuals that fail at time t_m and d_m be the size of $\mathbb{D}(t_m)$, then the partial likelihood function can be rewritten as:

$$p\left(D_i|\boldsymbol{\beta}_i\right)=\prod_{m=1}^{M_i}\left(\frac{\prod_{g\in\mathbb{D}(t_m)}\exp\left(\boldsymbol{\beta_i}^T\boldsymbol{R}_{ig}\left(t_m\right)\right)}{\prod_{j=1}^{d_m}\left(\sum_{k\in\mathbb{R}(t_m)}\exp\left(\boldsymbol{\beta_i}^T\boldsymbol{R}_{ik}\left(t_m\right)\right)-\frac{j-1}{d_m}\sum_{g\in\mathbb{D}(t_m)}\exp\left(\boldsymbol{\beta_i}^T\boldsymbol{R}_{ig}\left(t_m\right)\right)\right)}\right). \tag{A.3}$$

Obviously, $p\left(\boldsymbol{\beta}|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0,D\right)$ has an explicit formulation and it can be sampled through MCMC sampling.

For Eq. (9), given that $\boldsymbol{\beta}=\boldsymbol{\beta}^0, \boldsymbol{\Sigma}=\boldsymbol{\Sigma}^0$, the conditional distribution $p(\boldsymbol{\mu}|\boldsymbol{\beta},\boldsymbol{\Sigma},D)$ is expressed as:

$$p\left(\boldsymbol{\mu}|\boldsymbol{\beta}^0,\boldsymbol{\Sigma}^0,D\right)\propto p\left(\boldsymbol{\mu}|\boldsymbol{\beta}^0,\boldsymbol{\Sigma}^0\right)\propto p\left(\boldsymbol{\mu},\boldsymbol{\beta}^0,\boldsymbol{\Sigma}^0\right)\propto p\left(\boldsymbol{\beta}^0|\boldsymbol{\mu},\boldsymbol{\Sigma}^0\right)p\left(\boldsymbol{\mu}|\boldsymbol{\Sigma}^0\right). \tag{A.4}$$

Obviously, $p\left(\boldsymbol{\beta}^0|\boldsymbol{\mu},\boldsymbol{\Sigma}^0\right)$ is composed of N independent and identically distributed Normal distributions $N_J(\boldsymbol{\mu},\boldsymbol{\Sigma}^0)$ and $p\left(\boldsymbol{\mu}|\boldsymbol{\Sigma}^0\right)$ is also a Normal distribution with mean $\boldsymbol{\mu}_0$ and covariance $\frac{1}{\lambda_0}\boldsymbol{\Sigma}^0$. Then we have the following equation:

$$\begin{aligned}p\left(\boldsymbol{\beta}^0|\boldsymbol{\mu},\boldsymbol{\Sigma}^0\right)p\left(\boldsymbol{\mu}|\boldsymbol{\Sigma}^0\right)&=(2\pi)^{-\frac{Np}{2}}\left|\boldsymbol{\Sigma}^0\right|^{-\frac{N}{2}}\\&\quad\times\exp\left[-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{0^{-1}}\sum_{i=1}^{N}\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}\right)\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}\right)^T\right]\right]\\&\quad\times(2\pi)^{-\frac{p}{2}}\left|\boldsymbol{\Sigma}^0\right|^{-\frac{1}{2}}\\&\quad\times\exp\left[-\frac{\lambda_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}^{0^{-1}}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right],\end{aligned} \tag{A.5}$$

which is a new Normal distribution: $\boldsymbol{\mu}|\boldsymbol{\beta}^0,\boldsymbol{\Sigma}^0,D\sim N(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$, and

$$\boldsymbol{\mu}^*=\frac{\lambda_0}{\lambda_0+N}\boldsymbol{\mu}_0+\frac{1}{\lambda_0+N}\sum_{i=1}^{N}\boldsymbol{\beta}_i^0=\frac{\lambda_0}{\lambda_0+N}\boldsymbol{\mu}_0+\frac{N}{\lambda_0+N}\bar{\boldsymbol{\beta}}^0, \tag{A.6}$$

$$\boldsymbol{\Sigma}^*=\frac{\boldsymbol{\Sigma}^0}{(N+\lambda_0)}. \tag{A.7}$$

For Eq. (10), given that $\boldsymbol{\mu}=\boldsymbol{\mu}^0, \boldsymbol{\beta}=\boldsymbol{\beta}^0$, the conditional distribution $p(\boldsymbol{\Sigma}|\boldsymbol{\beta},\boldsymbol{\mu},D)$ can be written as:

$$p\left(\boldsymbol{\Sigma}|\boldsymbol{\beta}^0,\boldsymbol{\mu}^0,D\right)\propto p\left(\boldsymbol{\Sigma}|\boldsymbol{\beta}^0,\boldsymbol{\mu}^0\right)\propto p\left(\boldsymbol{\beta}^0|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\mu}^0|\boldsymbol{\Sigma}\right)p(\boldsymbol{\Sigma}). \tag{A.8}$$

And

$$\begin{aligned}&p\left(\boldsymbol{\beta}^0|\boldsymbol{\mu}^0,\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\mu}^0|\boldsymbol{\Sigma}\right)p(\boldsymbol{\Sigma})\\&=(2\pi)^{-\frac{Np}{2}}|\boldsymbol{\Sigma}|^{-\frac{N}{2}}\exp\left[-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}^0\right)\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}^0\right)^T\right]\right]\\&\quad\times\frac{1}{Z_{NIW}}|\boldsymbol{\Sigma}|^{-\frac{\upsilon_0+p+2}{2}}\\&\quad\times\exp\left[-\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0\right)-\frac{\lambda_0}{2}\left(\boldsymbol{\mu}^0-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^0-\boldsymbol{\mu}_0\right)\right],\end{aligned} \tag{A.9}$$

where Z_{NIW} is a constant:

$$Z_{NIW}=2^{\frac{\upsilon_0 p}{2}}\boldsymbol{\Gamma}_p\left(\frac{\upsilon_0}{2}\right)\left(\frac{2\pi}{\lambda_0}\right)^{\frac{p}{2}}|\boldsymbol{\Sigma}_0|^{-\frac{\upsilon_0}{2}}. \tag{A.10}$$

From Eq. (A.9), we can conclude that $\boldsymbol{\Sigma}|\boldsymbol{\beta}^0,\boldsymbol{\mu}^0,D$ follows an Inverse Wishart distribution $W^{-1}(\boldsymbol{\Sigma}_0{}^*,\nu^*)$, and

$$\boldsymbol{\Sigma}_0{}^*=\boldsymbol{\Sigma}_0+\sum_{i=1}^{N}\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}^0\right)\left(\boldsymbol{\beta}_i^0-\boldsymbol{\mu}^0\right)^T+\lambda_0\left(\boldsymbol{\mu}^0-\boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}^0-\boldsymbol{\mu}_0\right)^T, \tag{A.11}$$

$$\nu^*=N+\upsilon_0+1. \tag{A.12}$$

Appendix B. Proof of Lemma 2

Denote $\boldsymbol{\beta}=\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_N$ and conditioning on $\boldsymbol{\beta}$, then $p\left(\boldsymbol{\beta}_+|D_1,D_2,\ldots,D_N\right)$ can be written as the following equation:

$$\begin{aligned}P\left(\boldsymbol{\beta}_+|D_1,D_2,\ldots,D_N\right)&=\int p\left(\boldsymbol{\beta}_+,\boldsymbol{\beta}|D_1,D_2,\ldots,D_N\right)d\boldsymbol{\beta}\\&=\int p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right)p\left(\boldsymbol{\beta}|D_1,D_2,\ldots,D_N\right)\boldsymbol{d}\boldsymbol{\beta}.\end{aligned} \tag{B.1}$$

Since $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_N$ are independent and identically distributed, $p\left(\boldsymbol{\beta}|D_1,D_2,\ldots,D_N\right)$ can be expressed as the product of N independent terms, that is, $p\left(\boldsymbol{\beta}|D_1,D_2,\ldots,D_N\right)=\prod_{i=1}^{N}p\left(\boldsymbol{\beta}_i|D_i\right)$. $p\left(\boldsymbol{\beta}_i|D_i\right)$ is the posterior distribution, which has been calculated by Gibbs Sampling in the offline stage.

As for $p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right)$, it is a posterior predictive for the multi-variate Normal with NIW as the conjugate prior [44], then $p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right)$ follows a t-distribution:

$$p\left(\boldsymbol{\beta}_+|\boldsymbol{\beta}\right)\sim t_{\upsilon_N-p+1}\left(\boldsymbol{\mu}_N,\frac{\boldsymbol{\Sigma}_N\left(\lambda_N+1\right)}{\lambda_N\left(\upsilon_N-p+1\right)}\right). \tag{B.2}$$

where

$$\boldsymbol{\mu}_N = \frac{\lambda_0}{\lambda_0 + N}\boldsymbol{\mu}_0 + \frac{1}{\lambda_0 + N}\sum_{i=1}^{N}\boldsymbol{\beta}_i = \frac{\lambda_0}{\lambda_0 + N}\boldsymbol{\mu}_0 + \frac{N}{\lambda_0 + N}\bar{\boldsymbol{\beta}}, \tag{B.3}$$

$$\lambda_N = \lambda_0 + N, \tag{B.4}$$

$$\upsilon_N = \upsilon_0 + N, \tag{B.5}$$

$$\boldsymbol{\Sigma}_N = \boldsymbol{\Sigma}_0 + \sum_{i=1}^{N}\left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}\right)^T + \frac{\lambda_0 N}{\lambda_0 + N}\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\mu}_0\right)\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\mu}_0\right)^T. \tag{B.6}$$

In addition, λ_0, υ_0, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are hyper-parameters of the NIW distribution. We define these parameters to make the hyper NIW distribution noninformative so that the integration results can be mainly dominated by the data.

References

[1] Jardine AK, Lin D, Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech Syst Signal Process 2006;20(7):1483–510.

[2] Lu CJ, Meeker WO. Using degradation measures to estimate a time-to-failure distribution. Technometrics 1993;35(2):161–74.

[3] Gebraeel NZ, Lawley MA, Li R, Ryan JK. Residual-life distributions from component degradation signals: A Bayesian approach. IiE Trans 2005;37(6):543–57.

[4] Peng W, Li Y, Yang Y, Huang H, Zuo MJ. Inverse Gaussian process models for degradation analysis: A Bayesian perspective. Reliab Eng Syst Saf 2014;130:175–89.

[5] Wen Y, Wu J, Das D, Tseng TB. Degradation modeling and RUL prediction using Wiener process subject to multiple change points and unit heterogeneity. Reliab Eng Syst Saf 2018;176:113–24.

[6] Xu X, Tang S, Yu C, Xie J, Han X, Ouyang M. Remaining useful life prediction of lithium-ion batteries based on wiener process under time-varying temperature condition. Reliab Eng Syst Saf 2021;214:107675.

[7] He R, König F, Wang Y, Wirsing F, Tian Z, Zuo M, Ye Z. Wear and life predictions for bearings considering simulation-to-reality variability. Mech Syst Signal Process 2025;229:112498.

[8] Zhang Z, Si X, Hu C, Lei Y. Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods. European J Oper Res 2018;271(3):775–96.

[9] Kordestani M, Saif M, Orchard ME, RazaviFar R, Khorasani K. Failure prognosis and applications—A survey of recent literature. IEEE Trans Reliab 2019;70(2):728–48.

[10] Zhou Q, Son J, Zhou S, Mao X, Salman M. Remaining useful life prediction of individual units subject to hard failure. IIE Trans 2014;46(10):1017–30.

[11] Man J, Zhou Q. Remaining useful life prediction for hard failures using joint model with extended hazard. Qual Reliab Eng Int 2018;34(5):748–58.

[12] Yue X, Kontar RA. Joint models for event prediction from time series and survival data. Technometrics 2021;63(4):477–86.

[13] Hu J, Chen P. Predictive maintenance of systems subject to hard failure based on proportional hazards model. Reliab Eng Syst Saf 2020;196:106707.

[14] Wen Y, Guo X, Son J, Wu J. A neural-network-based proportional hazard model for IoT signal fusion and failure prediction. IISE Trans 2023;55(4):377–91.

[15] Li J, Zhou S, Han Y. Advances in battery manufacturing, service, and management systems. John Wiley & Sons; 2016.

[16] Ragab A, Yacout S, Ouali M-S, Osman H. Prognostics of multiple failure modes in rotating machinery using a pattern-based classifier and cumulative incidence functions. J Intell Manuf 2019;30:255–74.

[17] Xu D, Xiao X, Liu J, Sui S. Spatio-temporal degradation modeling and remaining useful life prediction under multiple operating conditions based on attention mechanism and deep learning. Reliab Eng Syst Saf 2023;229:108886.

[18] Ferreira C, Gonçalves G. Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. J Manuf Syst 2022;63:550–62.

[19] Ding Y, Jia M, Zhuang J, Cao Y, Zhao X, Lee C-G. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. Reliab Eng Syst Saf 2023;230:108890.

[20] Cheng H, Kong X, Wang Q, Ma H, Yang S, Xu K. Remaining useful life prediction combined dynamic model with transfer learning under insufficient degradation data. Reliab Eng Syst Saf 2023;236:109292.

[21] Zhang Z, Chen X, Zio E, Li L. Multi-task learning boosted predictions of the remaining useful life of aero-engines under scenarios of working-condition shift. Reliab Eng Syst Saf 2023;237:109350.

[22] Dong S, Xiao J, Hu X, Fang N, Liu L, Yao J. Deep transfer learning based on Bi-LSTM and attention for remaining useful life prediction of rolling bearing. Reliab Eng Syst Saf 2023;230:108914.

[23] Fan Y, Nowaczyk S, Rögnvaldsson T. Transfer learning for remaining useful life prediction based on consensus self-organizing models. Reliab Eng Syst Saf 2020;203:107098.

[24] Le Xuan Q, Munderloh M, Ostermann J. Self-supervised domain adaptation for machinery remaining useful life prediction. Reliab Eng Syst Saf 2024;250:110296.

[25] Huang C-G, Li H, Peng W, Tang LC, Ye Z-S. Personalized federated transfer learning for cycle-life prediction of lithium-ion batteries in heterogeneous clients with data privacy protection. IEEE Internet Things J 2024.

[26] Pan T, Chen J, Ye Z, Li A. A multi-head attention network with adaptive meta-transfer learning for RUL prediction of rocket engines. Reliab Eng Syst Saf 2022;225:108610.

[27] Saha B, Gupta S, Phung D, Venkatesh S. Multiple task transfer learning with small sample sizes. Knowl Inf Syst 2016;46:315–42.

[28] Liao Q, Ding Y, Jiang ZL, Wang X, Zhang C, Zhang Q. Multi-task deep convolutional neural network for cancer diagnosis. Neurocomputing 2019;348:66–73.

[29] Li Y, Wang L, Wang J, Ye J, Reddy CK. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In: 2016 IEEE 16th international conference on data mining. ICDM, IEEE; 2016, p. 231–40.

[30] Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. Mutational landscape and significance across 12 major cancer types. Nat 2013;502(7471):333–9.

[31] Wang L, Li Y, Zhou J, Zhu D, Ye J. Multi-task survival analysis. In: 2017 IEEE international conference on data mining. ICDM, IEEE; 2017, p. 485–94.

[32] Wang L, Chignell M, Jiang H, Charoenkitkarn N. Cluster-boosted multi-task learning framework for survival analysis. In: 2020 IEEE 20th international conference on bioinformatics and bioengineering. BIBE, IEEE; 2020, p. 255–62.

[33] Liu C, Cao W, Wu S, Shen W, Jiang D, Yu Z, Wong H. Asymmetric graph-guided multitask survival analysis with self-paced learning. IEEE Trans Neural Netw Learn Syst 2020;33(2):654–66.

[34] Fallahdizcheh A, Wang C. Data-level transfer learning for degradation modeling and prognosis. J Qual Technol 2023;55(2):140–62.

[35] Wen Y, Wu J, Yuan Y. Multiple-phase modeling of degradation signal for condition monitoring and remaining useful life prediction. IEEE Trans Reliab 2017;66(3):924–38.

[36] Chen N, Tsui KL. Condition monitoring and remaining useful life prediction using degradation signals: Revisited. IiE Trans 2013;45(9):939–52.

[37] Cox DR. Regression models and life-tables. J R Stat Soc Ser B Stat Methodol 1972;34(2):187–202.

[38] Tsiatis AA, Degruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. J Amer Statist Assoc 1995;90(429):27–37.

[39] Breslow NE. Contribution to discussion of paper by DR Cox. J R Stat Soc Ser B 1972;34:216–7.

[40] Hildebrand F. Introduction to numerical analysis: Courier corporation. Chelmsford, MA, USA: Courier Corporation; 1987.

[41] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama 1982;247(18):2543–6.

[42] Mathioudakis K, Stamatis A, Tsalavoutas A, Aretakis N. Computer models for education on performance monitoring and diagnostics of gas turbines. Int J Mech Eng Educ 2002;30(3):204–18.

[43] Efron B. The efficiency of Cox's likelihood function for censored data. J Amer Statist Assoc 1977;72(359):557–65.

[44] Murphy KP. Conjugate Bayesian analysis of the Gaussian distribution. Def 2007;1:16.