Reliability Engineering and System Safety 145 (2016) 183–189

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Bayesian hazard modeling based on lifetime data with latent heterogeneity

Mingyang Li [a], Jian Liu [b,*]

[a] Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA
[b] Department of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 2 January 2015
Received in revised form
31 August 2015
Accepted 12 September 2015
Available online 25 September 2015

Keywords:
Bayesian inference
Mixture model
Hazard regression
Gibbs sampler
Model selection

ABSTRACT

Lifetime data collected from reliability tests or field operations often exhibit significant heterogeneity patterns caused by latent factors. Such latent heterogeneity indicates that lifetime observations may belong to different sub-populations with different distribution parameters. As a result, the assumption on data homogeneity adopted by conventional reliability modeling techniques becomes inappropriate. Effective identification and quantification of such heterogeneity is crucial for more reliable model estimation and subsequent optimal decision making in a variety of reliability assurance activities. This research proposes a full Bayesian modeling framework for statistical hazard modeling of latent heterogeneity in lifetime data. The proposed framework is generic and comprehensive by systematically addressing different modeling aspects, which include modeling sub-populations with different hazard rates changing over time and different responses to the same stress factors, determining the number of sub-populations, identifying the most appropriate sub-population model structures, estimating model parameters and performing predictive inference. A numerical case study demonstrates the validity and effectiveness of the proposed approach.

1. Introduction

Statistical analysis of lifetime data is important in reliability assessment, prediction and improvement. Lifetime data obtained from laboratory tests in the design and development phase often provide the basis for reliability assessment, verification and prediction. Lifetime data collected from manufacturers' follow-up actions in the field operational phase, such as warranty claims and maintenance records, are also the important feedback information. Such feedback information helps manufacturers to improve the product design and manufacturing and to establish various optimal service policies.

Conventional statistical modeling of lifetime data assumes that the underlying product population is homogeneous. However, in real-world practice, lifetime data are often heterogeneous and the homogeneity assumption does not hold. A typical example is that, in the semiconductor industry, some units of integrated circuits belong to a weak sub-population and fail much earlier than other units under the same usage conditions. This difference in lifetime distribution is due to defective internal bonds or contaminator corrosions resulting from manufacturing defects [1]. Such phenomenon becomes even more obvious for maturing manufacturing processes where evolving and new technologies are applied [2]. Root cause analysis of all failures is often expensive and time-consuming. Therefore, there is limited ways to classify the product units into different sub-populations and specify their sub-population memberships prior to the statistical lifetime modeling. In this paper, the phenomenon of product units with heterogeneous lifetime distribution but unknown sub-population membership is defined as latent heterogeneity.

To account for the heterogeneity, different modeling approaches have been developed and studied. One popular approach is based on the hazard modeling. Since the hazard function of the heterogeneous population often exhibits patterns that can be broken down into segments, change-point models were employed to capture the hazard with piecewise functions separated by change points [6,7]. These change-point based methods cannot be applied in cases where a single hazard function is defined over the entire lifetime domain. To overcome this limitation, frailty models were introduced to construct a single hazard model with a multiplicative random variable (i.e., so-called frailty term) to represent unobserved heterogeneity [8,9]. However, frailty models often require known sub-population membership of each product unit. Such assumption is reasonable in some medical studies [10], since patients' background information may be used for classification. In some engineering applications, a unit's membership is often unknown and thus, latent heterogeneity is

* Corresponding author. Tel.: +1 520 621 6548; fax: +1 520 621 6555.
E-mail address: jianliu@email.arizona.edu (J. Liu).

http://dx.doi.org/10.1016/j.ress.2015.09.007
0951-8320/

inevitable. Instead of modeling lifetime data with a hazard function, another popular approach is to utilize mixture of lifetime distributions, which represent a heterogeneous population with a finite number of homogeneous sub-populations. Distributions such as Weibull [3,4] and Lognormal [5] are employed to model the lifetimes of each sub-population. Such formulation allows both mathematical simplicity of a single model defined on the entire domain and practical concerns of latent heterogeneity.

Both the aforementioned approaches have their own advantages. Mixture distribution approaches quantify the proportion of each sub-population in the overall population. Such proportion information is useful in the evaluations of product design and production performance [1]. Hazard modeling approaches, on the other hand, are more informative in capturing the underlying failure mechanism [11]. Thus, it is desirable to combine both advantages in heterogeneity modeling. To achieve this, this paper proposes a hazard modeling approach with the formulation of mixture distributions. Specifically, the heterogeneous population is represented by a mixture of a finite number of homogeneous sub-populations while hazard modeling is realized within each sub-population. Existing researches along this direction are limited and mainly focused on model estimation. Attardi et al. [12] analyzed interval-censored data through maximum likelihood estimation of the mixture of Weibull regression. Rosen and Tanner [13] considered the mixture of Cox proportional hazards model under the Frequentist framework. For model estimation comparison among different non-Bayesian estimation methods, see [14] and references therein.

In this paper, a full Bayesian modeling framework is proposed by comprehensively addressing modeling issues, such as model estimation, model selection, and model prediction, in a systematic manner. Bayesian inference is considered due to its advantage of incorporating possible prior knowledge and its practical convenience in different aspects of modeling (see details in Section 3). Specifically, mixture model formulation is considered for modeling heterogeneity and each sub-population is modeled by a specific hazard regression. Hazard regression is more generic in a sense that Weibull regression [12] and Cox proportional hazards model [13] can be treated as its special case. Challenges and difficulties involved in model estimation and model selection are also addressed comprehensively. The proposed modeling approach features in (i) modeling each sub-population with a hazard regression to quantify possible influence of reliability impact factors; (ii) considering different inherent hazard rates and different responses to the same reliability impact factors among sub-populations; (iii) considering Bayesian modeling approach by incorporating possible prior knowledge into the model formulation and estimation; (iv) providing a coherent framework by comprehensively addressing model construction, estimation, selection and prediction. For practitioners, the proposed work provides a complete procedure (e.g., model construction, estimation, selection, prediction) in analyzing heterogeneous lifetime data. It also allows practitioners to specify their domain knowledge as priors and incorporate them into the data analysis procedure. The rest of paper is organized as follows. Section 2 introduces the proposed model. Section 3 discusses a variety of modeling issues. Section 4 provides further illustration with a numerical case study and Section 5 draws the conclusion.

2. Model formulation

Consider a population consisting of m homogeneous sub-populations. Units from a certain sub-population, j, exhibit similar failure characteristics and thus their lifetimes can be assumed as independent and identically distributed random variables with probability density function, $f_j(t)$, reliability function, $R_j(t)$, and hazard function, $h_j(t), j=1,\ldots,m$. Hazard regression can be used to model reliability of units from a homogeneous sub-population by explicitly taking into account the influence of possible reliability impact factors. Specifically, a hazard function of the jth sub-population can be expressed as

$$h_j(t|\boldsymbol{\beta}_j) = h_j^{\mathrm{b}}(t)\exp(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}), \quad j=1,\ldots,m, \tag{1}$$

where $h_j^{\mathrm{b}}(t)$ is the baseline hazard function, $\mathbf{x}$ and $\boldsymbol{\beta}_j$ are $p\times 1$ vectors of covariates (i.e., reliability impact factors) and the corresponding covariate coefficients. Baseline hazard function corresponds to the situation where all covariates are equal to zero, i.e., $h_j^{\mathrm{b}}(t) = h_j(t|\mathbf{x}=\mathbf{0})$. Covariate coefficients explicitly quantify the influence of covariates on the hazard rate. $h_j^{\mathrm{b}}(t)$ and $\boldsymbol{\beta}_j$ are identical for units within a sub-population but may be different for units across sub-populations. A typical example is a population consisting of a small proportion of weak units due to the imperfections in manufacturing processes. As opposed to the remaining quality products with increasing failure rate (IFR, e.g., $\partial h_1^{\mathrm{b}}(t)/\partial t > 0$), such small proportion of defective units may have significantly higher hazard rates in their early period of usage and exhibit decreasing failure rate (DFR, e.g., $\partial h_2^{\mathrm{b}}(t)/\partial t < 0$). The impacts of stress factors on their hazards will also be different from those on quality products, i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$.

Hazard regression is employed to model reliability of each individual homogeneous sub-population due to its flexibility in quantifying influences of covariates, $\mathbf{x}$, and its great flexibility in representing many widely used models in reliability engineering and survival analysis. If $h_j^{\mathrm{b}}(t)$ is modelled parametrically, it can represent lifetime distributions such as exponential, Weibull, and extreme value. If $h_j^{\mathrm{b}}(t)$ is modelled non-parametrically, it becomes a Cox proportional hazard model proposed by Cox [15]. With each sub-population represented by a hazard regression model, the lifetime distribution, $f(t)$, of the overall population is given by

$$f(t|\mathbf{x},\Theta) = \sum_{j=1}^{m} w_j f_j(t|\mathbf{x},\boldsymbol{\beta}_j), \tag{2}$$

where w_j is the mixing proportion of the jth sub-population and Θ denotes a collection of all unknown parameters, i.e., $\Theta = \{w_j, h_j^{\mathrm{b}}, \boldsymbol{\beta}_j, j=1,\ldots,m\}$. $f_j(t|\mathbf{x},\boldsymbol{\beta}_j)$ is the conditional probability density of the jth sub-population and can be uniquely determined by model (1) through $f_j(t|\mathbf{x},\boldsymbol{\beta}_j) = h_j(t|\mathbf{x},\boldsymbol{\beta}_j)\,\exp(-\int_0^t h_j(s|\mathbf{x},\boldsymbol{\beta}_j)\,\mathrm{d}s)$. Eqs. (1) and (2) constitute the proposed model of modeling lifetime heterogeneity. It is noted that the proposed model has some connection with the discrete frailty model in [16]. In particular, it can be reduced into the discrete frailty model if the membership of each unit is given. The proposition below specifies their relationships followed with some interpretations.

Proposition. *Hazard function $h(t)$ based on Eqs. (1) and (2) can be reduced into the discrete frailty model if the following two conditions hold (see proofs in Appendix A):*

(i) $h_j^{\mathrm{b}}(t)/h_m^{\mathrm{b}}(t) = C_j,\ j=1,\ldots,m-1$, *where C_j's are constants and* $C_m = 1$;
(ii) $\boldsymbol{\beta}_j = \boldsymbol{\beta},\ \forall j=1,\ldots,m.$

And the resulting discrete frailty model is given by

$$h(t|V) = V\bar{h}^{\mathrm{b}}(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \tag{3}$$

where $\bar{h}^{\mathrm{b}}(t) = \sum_{j=1}^{m} w_j h_j^{\mathrm{b}}(t)$, V is a discrete random variable following the categorical distribution, i.e., $\Pr(V = \frac{C_j}{\sum_{j=1}^{m} w_j C_j}) = w_j, j=1,\ldots,m$, and satisfying $\mathbf{E}(V)=1$.