A Unifying Approach to Data Fusion for Reliability Prediction

Rong Pan
Dept of Industrial, Systems and Operations Engineering
Arizona State University
Tempe, Arizona
United States
rong.pan@asu.edu

Abstract

In a data-rich environment, effectively and efficiently extracting information from heterogeneous data sources becomes critically important. Current data fusion methods are mostly geared to data organization and database management, instead of extracting information to build system models and to do model-based prediction. This paper focuses on the methodological development of an integrated, information-driven system modeling and analysis approach to reliability prediction. To be succinct, in this paper we only discuss how to establish a connection among various types of data from different sources. It generally requires that data structures can be explicitly established through either engineering knowledge or statistical modeling. We define three basic data structures - clustered data, functional structured data, and input-output data, and use them as the building blocks for analyzing complex systems. The hierarchical Bayesian data analysis is then the natural choice for model inference and model-based reliability prediction. Given its model integrity and adaptability, the hierarchical data modeling approach allows analysts to attack the reliability prediction problem through a meaningful data fusion.

1 Introduction

Reliability prediction can be improved by maximizing the use of all relevant data. For a complex system composed of multiple components with different development processes, integrating direct and indirect information to improve system performance assessment and to achieve the optimal use of resources is highly desirable. Indeed, most industrial products/processes are developed through an evolving, incrementally refining process. Using the reliability information obtained along the life cycle of previous products to design and optimize the next generation product stands out to be an extremely efficient and effective approach. This paper emphasizes data modeling techniques for problems with complex data structures. The idea of using hierarchical multilevel models to represent the internal structures of heterogeneous data from different sources will be presented first, followed by examples. Model estimation and model diagnosis using Bayesian methods are brief described.

2 Three basic structures

To establish a connection among various types of data from different sources, it generally requires that the data structure can be found through either engineering knowledge or some practical assumptions. Toward this goal, we categorize three basic data structures - cluster structure, functional structure, input-output structure - as the elementary components for analyzing multi-source, multi-level, heterogeneous reliability data. These basic structures are illustrated in the figure below.

Clustered data are the data collected from similar processes, but may exhibit extra variability due to the cluster-to-cluster variation. In reliability engineering, one may obtain failure or censoring time data of products from different manufacturing lines, different plants, or different user groups. Combining them can reduce the parameter estimation error of a statistical model and strengthen the statistical power of hypothesis testing, but the heterogeneity among clusters must be carefully accounted for. Hierarchical modeling becomes a natural choice for this cause. In general, observations from different sources are modeled by one distribution family; while some parameters in the family are assumed to be generated from a higher level distribution with hyperparameters, thus they are correlated. The hierarchical model refers to a hierarchical structure, from data to model parameter and hyperparameter. In fact, the

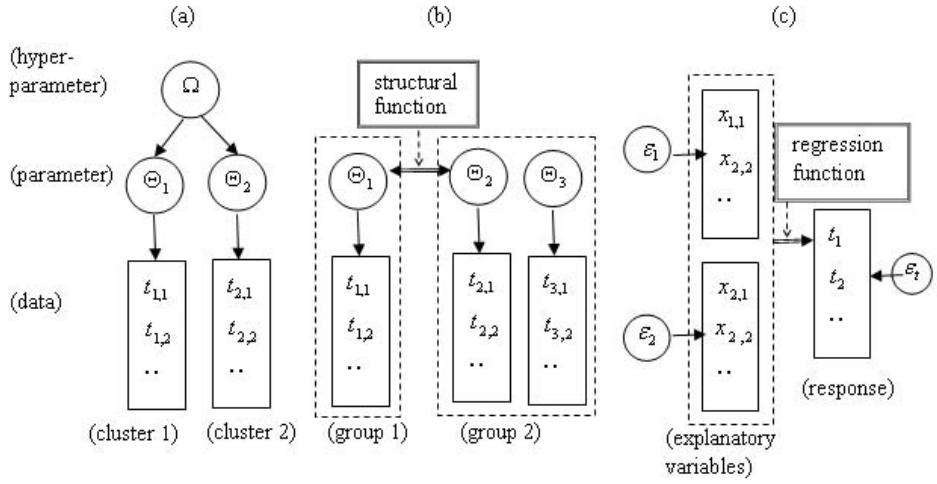

Figure 1: (a) Cluster structure; (b) Functional structure; (c) Input-output structure

randomness of a model parameter introduces a random effect into the overall reliability/failure function; therefore, this hierarchical model can be viewed as a type of random effect model.

Functional structured data are constructed by an explicit link function that bridges two statistical models, so it provides more flexibility than clustering techniques for modeling the internal relationship among multi-source data. For example, system reliability is in general a function of component reliability, so the failure observations of both system and components can be linked by this system reliability function. Hamada et al. (2004) discussed a full Bayesian approach of combining multi-level failure information for a simple system with three components. Calibration model is another type of functional structure. Oftentimes we can use a calibration function to link product failure rates from different product generations (Pan, 2009), combine computer simulation models with different model fidelity (Qian and Wu, 2008), or integrate empirical models obtained from experimental data with analytical models derived from physical/chemical principles (Reese et al., 2004).

Input-output data are responses with explanatory variables. Typically, a statistical regression model (in contrast to the causal model of functional structure) is built between inputs and outputs. In reliability engineering, the failure time regression model and the proportional hazard model are often used for regression. In survival data analysis, the frailty model is developed to extend the proportional hazard model to include random effects. It is common that not only output data, but also input data are measured with uncertainties. In a dynamic product use environment, for example, environmental stress variables may vary their values from time to time. Therefore, the regression model needs to be built with a careful consideration of variance components.

3 Examples and Bayesian inferences

Example 1: Analyzing recurrent data from multiple repairable systems with heterogeneous repair effectiveness. This problem considers multiple manufacturing lines that are maintained by different repair crews. A widely used parametric form of failure intensity function is the power law function, $\lambda(t) = \theta\beta t^{\beta-1}$, $\theta > 0$, $\beta > 0$ where θ is a scale parameter (also called the intrinsic failure rate) and β is a shape parameter. The intensity function is a monotone increasing function and it can be assumed that a repair action will cause an arithmetic reduction in intensity, i.e., $\lambda(t^+) = \lambda(t^-) - \rho\lambda(t^-)$, where t^- and t^+ are the times immediately before and after the repair, respectively, and ρ specifies the effectiveness of the repair (Doyen and Gaudoin, 2004). Both failure intensity and repair effectiveness cannot be directly

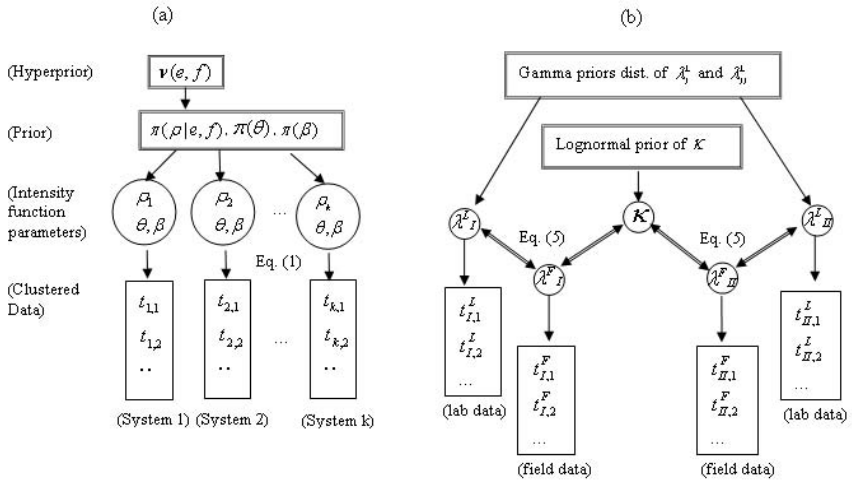

Figure 2: (a) Multiple repairable systems; (b) Lab-field data integration

observed. The failure density function of each individual system can be derived as

$$f_{\mathbf{t}} = \prod_{i=1}^{n} p(t_{k,i} \mid t_{k,i-1}, ..., t_{k,1}) = \theta^n \beta^n \prod_{i=1}^{n} (t_{k,i}^{\beta-1} - \rho_k t_{k,i-1}^{\beta-1}) e^{-\theta[t_{k,n}^{\beta} - \rho_k \beta \sum_{i=1}^{n} t_{k,i-1}^{\beta-1}(t_{k,i} - t_{k,i-1})]}$$

where $p(t_{k,i} \mid t_{k,i-1}, ..., t_{k,1})$ is the conditional probability function of recurrent failure time, and $(t_{k,1}, ..., t_{k,n})$ are failure time data clustered by systems. The data variability between systems may reflect the difference of maintenance practices on these systems; therefore, one can treat the parameter ρ_k of each system as a random value from a distribution, which characterizes the variation of repair effectiveness from one maintenance team to another. Essentially, repair actions introduce a random effect into the general repairable system model. Figure 2(a) illustrates the data structure and the hierarchical Bayesian (HB) model of this example. A study of synthesizing repairable system failure information using HB has been conducted in Pan and Rigdon (2009). The result highlights the advantage of combining information from similar systems to improve model prediction on individual systems.

Example 2: Lifetime data from two product generations. Consider a new product is developed based on a technological modification of an old product. The question is: How to use the lifetime information from the old product to help predict the reliability of the new product? This problem was studied from Bayesian perspective in Whitmore et al. (1994) and Young (1994). We had extended the study to include the accelerated life testing and field testing data of both old and new products into consideration (Pan, 2009). Assume product failure times from both generations follow exponential distribution. Because of the variation of environmental stresses which are uncontrollable under the product use condition, the failure times from laboratory testing, where stress variables are controllable, may not fully agree with field observations. In general, field lifetime data exhibit larger variation than the lab data. To show the functional relationship between lab and field lifetime distributions, a calibration factor, κ, is used to account for the effect of the variation of environmental stresses, i.e., $\lambda^F = \kappa\lambda^L/AF$, where AF is an acceleration factor, assumed to be known. Expert opinions can be used to generate the prior distributions of the lab testing failure rates and the calibration factor. Figure 2(b) depicts the functional structure of these parameters and data in a hierarchical model. In the observation space we have $(t_{I,1}^L, ...)$, $(t_{I,1}^F, ...)$, $(t_{II,1}^L, ...)$ and $(t_{II,1}^F, ...)$; while in the parameter space, we have $(\lambda_I^F, \lambda_I^L, \lambda_{II}^F, \lambda_{II}^L, \kappa)$. Using the hierarchical model, the estimation of field failure rate will depend on experts' opinion, accelerated life testing data, as well as a calibration factor, which can be inferred from the life cycle data of the old product. Therefore, the effect of stress variation between field and lab is addressed.

Example 3: Step-stress accelerated life testing (SSALT) data analysis. SSALT is in general difficult for data analyst because the stress condition (covariate) is changing over time. Lee and Pan (2009)

proposed a GLM formulation which simplifies the data analysis by looking into the inner data structure of SSALT failure and censored data. By assuming exponential failure time distribution and log-linear function of failure rate on covariates, the likelihood function can be expressed in two terms, while the first term is the kernel of Poisson distribution and the second term does not depend on the covariate. As a simple example, suppose we have two step-stress levels (l_1, l_2) and the termination times on these levels are τ_1 and τ_2 ($\tau_2 > \tau_1$). Let three specimens be tested. The failure or censoring times are t_1 ($t_1 < \tau_1$), t_2 ($\tau_1 < t_2 < \tau_2$) and τ_2 (right censored). Then, the data can be transformed to triplets (y_k, c_k, x_k), where y_k is the time observed at each specimen-stress combination, c_k is failure/censoring indicator and x_k is the stress level. So, the new dataset of this example are $(t_1, 1, 1)$, $(\tau_1, 0, 1)$, $(t_1 - \tau_1, 1, 2)$, $(\tau_1, 0, 1)$ and $(\tau_2 - \tau_1, 0, 2)$. Connecting to the log-linear life-stress function, the GLM formulation can be written as

- Distribution: $c_k \sim Poisson(\mu_k)$
- Linear predictor: $\eta_k = \beta_0 + \beta_1 x_k$
- Link function: $\eta_k = \log \mu_k - \log y_k$ or $\log \mu_k = \beta_0 + \beta_1 x_k + \log y_k$

After clarifying this data structure, the inference on model parameters, β_0 and β_1, can be carried out through the conventional GLM method or Bayesian method. The benefit of Bayesian method lies on that any prior knowledge of failure rate or life-stress function can be easily incorporated into data analysis.

4 Conclusions

As is evident, hierarchical modeling is an extremely adaptable tool for information integration. It is able to mix fixed and random effects into statistical models based on the problem context and/or engineering knowledge. It is easy to interpret and communicate these models to reliability engineers and managers. The afore-described three basic data structures are the building blocks for modeling more complicated data scenarios. Although this research is still in its infancy; it has shown an enormous potential toward meaningful data fusion. Unlike some other machine learning techniques, which treat the system understudied as a black box, this research intends to elicit and test the embedded relationship of available information and decompose it to basic data structures. Typically, a link function is built upon the unobserved model parameters and conventional distribution functions are applied on observations. This hierarchical model is similar to the latent variable model and the structural equation model studied in psychology and social sciences, but it is more interpretable and implementable given the meaning of model and model parameters in engineering applications.

References

[1] Doyen, L. and Gaudoin, O. (2004) Classes of imperfect repair models based on reduction of failure intensity or virtual age, *Reliability Engineering and System Safety*, vol. 84, pp. 45-56.

[2] Hamada, M., Martz, H. F., Reese, C. S., Graves, T., Johnson, V. and Wilson, A. G. (2004) A fully Bayesian approach for combining multilevel failure information in fault tree quantification and optimal follow-on resource allocation, *Reliability Engineering and System Safety*, vol. 86, pp. 297-305.

[3] Lee, J. and Pan, R., (2009) Statistical inference methods for step-stress accelerated life testing using generalized linear models, Technical Report.

[4] Pan, R. (2009) A Bayes approach to reliability prediction utilizing data from accelerated life tests and field failure observations, *Quality and Reliability Engineering International*, vol. 25, no. 2, pp. 229-240.

[5] Pan, R. and Rigdon, S. (2009) Bayes Inference for General Repairable Systems, *Journal of Quality Technology*, vol. 41, no. 1, pp. 82-94.

[6] Qian, Z.G. and Wu, C.F. (2008) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments, *Technometrics*, vol. 50, no. 2, pp. 192-204.

[7] Reese, C. S., Wilson, A. G., Hamada, M., Martz, H. F. and Ryan, K. J. (2004) Integrated analysis of computer and physical experiments, *Technometrics*, vol. 46, no. 2, pp. 153-165.

[8] Whitmore, G. A., Young, K. D. S. and Kimber, A. C. (1994) Two-stage reliability tests with technological evolutions: a Bayesian analysis, *Applied Statistics*, vol. 43, no. 2, pp. 295-307.

[9] Young, K. D. S. (1994) A Bayesian analysis of undated component data, *The Statistician*, vol. 43, no. 1, pp. 129-137.