Reliability Engineering and System Safety 172 (2018) 12–24

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Bayesian model updating with summarized statistical and reliability data

Eric VanDerHorn [a,b], Sankaran Mahadevan [a,*]

[a] Department of Civil & Environmental Engineering, Vanderbilt University, Nashville, TN 37235, USA
[b] American Bureau of Shipping, Houston, TX, USA

ARTICLE INFO

Keywords:
Reliability
Calibration
Summary statistics
Sufficient statistics
Bayesian network

ABSTRACT

The accuracy of model-based reliability analysis is affected by the uncertainty regarding the model parameters used to predict the behavior of the engineering system. The uncertainty in the model parameters can be reduced by combining prior knowledge about the parameters with observed data regarding system inputs and outputs. In some cases, the information about the observations is only available as abstracted data, where the original raw data have been reduced to a summarized representation. Common forms of abstracted data include summary statistics, such as the mean and variance for continuous variables and observed frequencies for discrete variables. In the context of reliability analysis, a common form of available information is summarized reliability data for various mechanical components (e.g., failure rates or failure probabilities) instead of detailed actual test data. This paper presents a methodology for updating the model parameters using these abstracted data forms through a Bayesian network. First, the concept of a statistics function is developed and linked to the abstracted data forms. The concept of arc reversal is then exploited to transform the Bayesian network to a form that can be used to incorporate the statistics function and thereby enable the updating of the model parameters. Several numerical examples are used to demonstrate the applicability and generality of the proposed method for several different forms of abstracted data.

1. Introduction

Decision making in engineering applications based on the results of reliability analysis often relies on the use of mathematical or computational models to predict the behavior of complex engineering systems. Reliability analysis is affected by both aleatory uncertainty (natural variability) and epistemic uncertainty lack of knowledge regarding the variables or the models). The epistemic uncertainty can further be classified into statistical uncertainty and model uncertainty to represent the lack of knowledge in variables and models respectively. The model uncertainty is related to model approximations as well as the uncertainty in the model parameters. It is important that the model parameters be calibrated based on the available information so that the model predictions accurately reflect the physical reality. This updating process is informed by data and requires that all available information be properly incorporated into the modeling and simulation.

The model calibration data may be available in many different forms, including but not limited to, experimental and operational data, inspection reports, health monitoring data, engineering plans, rules and standards, and expert opinion. These heterogeneous sources of information can lead to significant challenges for model calibration, as the data may often be imprecise, uncertain, ambiguous, and/or incomplete. Additional challenges may arise as the data may not be provided in a traditional format, such as point or interval data [1], but instead may be provided in abstracted formats such as sample statistics (e.g. mean, variance, median, max, etc.), probability or frequency data, or reliability data.

The term “abstracted data” in this paper refers to the case where raw data has been reduced to a simplified representation of portions or the entirety of the raw dataset. There are several sources of abstracted data in practical applications [2,3]. For example, instead of receiving the full data of all the outcomes of an experiment, sometimes the only information provided from testing may be in the form of summary statistics of the observed sample distribution (e.g., mean, variance etc.) or the observed frequencies for categorical data. In some cases, the performance of a population of components or system may be given as reliability data [4] or summarized results from acceptance testing [5], both of which can be considered as forms of abstracted data. Sometimes, experts may provide their point or interval estimates of moments, frequencies, or probability ranges. This calibration process can be further complicated if data is provided simultaneously in several of these heterogeneous abstracted forms.

The incorporation of abstracted data in inference is not a new concern. Early work focused on the use of abstracted data for distribution parameter estimation of random variables, particularly in cases where

* Corresponding author.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

https://doi.org/10.1016/j.ress.2017.11.023
Received 3 April 2017; Received in revised form 20 October 2017; Accepted 29 November 2017
Available online 2 December 2017
0951-8320/

the estimation based on the original raw data might be biased due to data inconsistencies, such as outliers [6–8]. Additionally, an initial formulation by Pratt [9] describes the incorporation of summary statistics into Bayesian inference. Pratt also conveyed the idea that in certain situations, the resulting posterior distribution of the parameter of interest closely approximates the posterior distribution that would be obtained using the full dataset. Summary statistics with this property are defined as Bayesian sufficient statistics [10]. Much of the previous work has considered the cases where the raw data was available, but abstracted data was used instead to ensure robust inference in the presence of data inconsistencies. The focus of these past studies was primarily on the selection of the most appropriate statistics for the given problem.

More recent applications of abstracted data in the inference process include the identification and use of summary statistics for improving computational efficiency as commonly seen in methodologies such as the Approximate Bayesian Computation (ABC). ABC methods have been explored for a wide variety of applications in Bayesian inference [11–14] and were developed to circumvent the need to evaluate the likelihood function which might be analytically intractable or computationally expensive to evaluate. ABC is a simulation-based approach, where model parameters are randomly selected from a prior distribution and used to generate a sample dataset. The discrepancy between the simulated data set and the observed dataset is examined to determine if the model parameter values used to generate the simulated dataset should be accepted or rejected. The accepted model parameter values are used to generate the posterior distribution. For high-dimensional data, the probability of selecting appropriate model parameter values that generate a simulated dataset matching the observed dataset within a prescribed tolerance significantly decreases. To improve computational efficiency, low-dimensional summary statistics are used in ABC instead of the raw data. The main concern in using summary statistics is the loss of information associated with condensing the data, which may bias the discrimination between two models [15]. A sizeable amount of work in ABC has gone into determining the optimal summary statistics to minimize the information loss [16,17]. An area of research that has not currently been considered is the application of ABC when only sample summary statistics are provided instead of the raw data. However, previous efforts have shown that the computations can be particularly problematic when the considered statistic is not relevant to the current inference problem [18].

Another recent approach that has incorporated summary statistics into the inference process is maximum relative entropy (MrE). The MrE method has been proposed as a generalized framework to unify classical Bayesian inference with the concept of Maximum Entropy (MaxEnt) [19]. This has been found to be particularly useful in cases where both point data and moment data[1] are given [20]. Based on the axioms of maximum entropy [21], the optimal posterior distribution is the one that maximizes the relative entropy between the prior and posterior distributions. Observation data is incorporated into the inference process through the placement of constraints on the posterior distribution. This formulation is beneficial since it allows for the incorporation of any form of data that can be written as a constraint on the posterior distribution. While the advantage of this approach for engineering problems has been demonstrated [22], there is still some debate concerning its performance. Particularly, some studies argue that in certain situations, the MrE method conflicts with Bayes' rule [23,24] and that it can often lead to counterintuitive consequences [25]. This is perhaps most apparent when considering the sequence effects of processing different types of information, where the processing of constraints simultaneously vs. in different sequential orders could result in different posteriors. Another limitation of the MrE method is that the formulated constraints placed on the posterior are hard constraints, meaning that all posterior distributions which violate the constraint are ruled out. This can lead to difficulties if there is uncertainty in the abstracted data; the incorporation of data uncertainty is a challenge that is yet to be fully addressed in the MrE method.

The use of summary statistics in the Bayesian inference process has typically been examined as an alternative to using the raw data in order to ensure robust inference in the presence of data inconsistencies or to reduce the computational effort in inference, and the focus in both these cases has been on identifying the appropriate form of the statistics to achieve accurate inference results. This paper considers the case where only the abstracted data is provided and the raw data is unavailable, and we seek to incorporate this form of data in Bayesian inference.

There are several major challenges when considering the inference problem if only abstracted data is provided. These include the potential loss of information resulting from the data abstraction resulting in an insufficient statistic and the incorporation of the uncertainty associated with the probability distribution of the statistic resulting from a random sample of limited and potentially unknown sample size.

Bayesian networks provide a convenient and well-established approach for facilitating the inference of unknown or unobservable parameters, utilizing observations of random variables conditional on these unknown parameters. Therefore, this paper proposes a novel idea to incorporate "observations" of abstracted data through a Bayesian network representation to enable the Bayesian inference process in the presence of abstracted data. The relevant theory is developed from first principles and is suitable for practical problems. Since the proposed approach is based on a Bayesian network representation, this approach can be easily extended beyond the simple inference of distribution parameters of random variables (based on observations of that random variable) to the calibration of model parameters in physics models.

The remainder of the paper is organized as follows. Section 2 reviews Bayesian networks and inference. Section 3 proposes the methodology to incorporate abstracted data into the Bayesian network and presents a generalized form of Bayesian inference with abstracted data. Section 4 illustrates the proposed approach for inferring distribution parameters using material yield strength data. Section 5 then demonstrates the use of the proposed approach for physics model calibration using manufacturing acceptance testing/component reliability data. Section 6 provides concluding remarks.

2. Bayesian networks and inference

In this section, we first discuss stochastic and deterministic nodes in Bayesian networks and discuss how Bayesian methods provide a convenient framework for combining prior beliefs about parameters with current evidence gained from data.

2.1. Bayesian networks

Bayesian networks provide a convenient framework for graphically representing probabilistic relationships among multiple variables. More specifically, a Bayesian network is a directed, acyclic graph (DAG) representation of a multivariate distribution, expressing its decomposition into a combination of marginal and conditional probabilities. An example of a DAG model is given in Fig. 1.

Each node in a Bayesian network denotes a random variable and the directed edges between nodes (arcs) are associated with conditional probabilities. If there exists a directed edge between two nodes, the upstream node is designated the parent node and the downstream node is designated the child node. The dependence between these nodes can be described mathematically by a conditional probability distribution. Based on the *directed Markov condition*, a node is independent of its nondescendant nodes when conditioned on its parent nodes, therefore the Bayesian network can be decomposed into a product of conditional and marginal probabilities using the graphical structure and the chain rule of probability [26,27]. If the random variables in a Bayesian network are

[1] Moment data in this case refers to information about the expected values of moments of the distribution, and is a form of abstracted data.