Model Construction of the Basic Theories of Probability Theory in the Quantification of Uncertainty

Zirui Wang
Shanghai Experimental Foreign Language School, Shanghai, China
**Corresponding Author*

Abstract: Probability theory, as a core mathematical tool for quantifying uncertainty, provides a scientific basis for decision-making by constructing probability models. This paper systematically reviews the model construction path of the basic theories of probability theory in the quantification of uncertainty. Starting from the theoretical framework of probability space and random variables, it analyzes the influence of parameter estimation methods on model uncertainty, explores the advantages of Bayesian inference in the quantification of cognitive uncertainty, and further elaborates on the application of probability models in the propagation of uncertainty in complex systems. Research shows that the fundamental theories of probability theory provide a full-chain methodology from data generation to decision support for the quantification of uncertainty, which is of great value in enhancing the reliability of models and the scientific nature of decisions.

Keywords: Probability Theory; Quantification of Uncertainty; Model Construction; Parameter Estimation; Bayesian Inference

1. Introduction

Uncertainty is widely present in various fields such as science, engineering, economy and society. Its essence has a dual nature: on the one hand, it stems from the randomness of the data itself, that is, accidental uncertainty, such as sensor measurement errors, market price fluctuations, etc. On the other hand, it stems from the limitations of model cognition, that is, cognitive uncertainty, such as the unknown physical mechanisms of complex systems or the simplification of model assumptions [1]. In autonomous driving scenarios, sensor noise (accidental uncertainty) and the algorithm's insufficient recognition of rare road conditions (cognitive uncertainty) may jointly lead to decision-making errors. In medical diagnosis, individual differences among patients (accidental uncertainty) and incomplete coverage of disease mechanisms by diagnostic models (cognitive uncertainty) may affect diagnostic accuracy. Therefore, the comprehensive quantification of uncertainty is the key to enhancing the safety and effectiveness of decision-making.

Traditional deterministic models describe system behavior through fixed parameters or a single predicted value, ignoring the potential impact of uncertainty on the results. For instance, deep neural networks may give overly confident and incorrect predictions on samples outside the training data distribution [2], while probabilistic models can clearly reveal the credibility of the prediction results by outputting the prediction distribution (such as mean and variance) rather than a single value. This characteristic makes the fundamental theories of probability theory the core support for constructing quantitative models of uncertainty. Probability theory transforms uncertainty into measurable probability distributions by constructing mathematical tools such as probability spaces, random variables, and conditional probabilities, providing decision-makers with a basis for risk assessment (such as the probability of extreme events) and confidence level judgment (such as prediction intervals) [3].

In recent years, with the rapid development of data science and artificial intelligence, the application of probability theory in uncertainty quantification has shown new trends. On the one hand, the combination of deep learning and probabilistic models (such as Bayesian neural networks) significantly enhances the adaptability of models to scenarios with scarce data [4]; On the other hand, causal discovery under high-dimensional data and dynamic causal system modeling have become research hotspots, promoting the in-depth application of probability theory in the analysis of complex systems [5].

This paper starts from the core concepts of probability theory and systematically explores its model construction path in uncertainty quantification, aiming to provide theoretical references for uncertainty management in complex systems.

2. Basic Theory of Probability: The Theoretical Cornerstone of Quantifying Uncertainty

2.1 Probability Space and Random Variables: A Mathematical Framework for Quantifying Uncertainty

Probability theory constructs the mathematical basis for quantifying uncertainty through probability spaces (sample spaces, event sets, probability measures) and random variables. The sample space describes the set of all possible outcomes. For example, in medical diagnosis, the sample space can be defined as two states: "the patient is ill" and "the patient is healthy". The event set is a subset of the sample space, such as "The patient is over 50 years old and has the disease"; The probability measure assigns a numerical value to the possibility of each event occurring through the probability axioms (non-negativity, normativity, and columnability and additivity) [6]. For instance, if the prevalence rate of a certain disease is 5%, then the probability measure of the event "the patient is ill" is 0.05.

Random variables, as the core tool of probability theory, map the results of the sample space to real numbers, thereby achieving a numerical description of uncertainty. Discrete random variables (such as binomial distribution and Poisson distribution) are suitable for describing counting uncertainties, for example, the fluctuation in the number of customers arriving per unit time (following a Poisson distribution). Continuous random variables (such as normal distribution and uniform distribution) are used to describe the uncertainty of continuous changes. For example, the distribution of sensor measurement errors (usually assumed to be normal distribution) [7]. Selecting an appropriate random variable distribution model is a key step in building an effective probability model, and its rationality directly affects the accuracy of uncertainty quantification. For instance, in financial risk assessment, if it is wrongly assumed that the return on assets follows a normal distribution, the probability of extreme events (such as financial crises) may be underestimated, leading to distorted risk measurement [8].

2.2 Conditional Probability and Independence: Revealing the Interrelationships among Uncertainties

Conditional probability provides an analytical tool for the propagation of uncertainty in complex systems by quantifying the influence of known information on uncertainty. For instance, in financial risk assessment, under the condition of known market volatility, the conditional probability of asset price decline can help investors adjust their holding strategies. Independence, as a special case of conditional probability, simplifies the complexity of quantifying multivariable uncertainty. If two random variables are independent, the joint probability is equal to the product of the edge probabilities. This feature significantly reduces the computational complexity when constructing high-dimensional probability models.

However, the dependencies among variables in real systems are often nonlinear and complex. For instance, in climate models, variables such as temperature, humidity, and air pressure have significant interactions. At such times, it is necessary to describe the dependency structure among these variables through conditional probability distributions or Bayesian networks. This ability to quantify the correlation enables probability theory to more truly reflect the uncertain characteristics in complex systems.

3. Parameter Estimation Method: A Bridge from Data to Probability Distribution

3.1 Maximum Likelihood Estimation: Parameter Inference Based on Data Frequency

Maximum likelihood Estimation (MLE) seeks the parameter value that maximizes the probability of data occurrence by maximizing the likelihood function of the observed data. The core idea lies in "letting the data speak", that is, assuming that the data generation process conforms to a specific probability distribution, the distribution parameters are inferred from the sample data. For instance, in market demand forecasting, if it is assumed that demand follows a normal distribution, MLE can estimate the mean and variance parameters, thereby quantifying the uncertainty of demand fluctuations.

The excellent properties of MLE include asymptotic normality (the estimator follows a normal distribution as the sample size increases) and consistency (the estimator converges to the true value as the sample size increases), which make it highly efficient in big data scenarios. However, the sensitivity of MLE to model assumptions may lead to estimation bias. When the data distribution does not conform to the assumed model (such as when the data has thick-tail features but is assumed to be a normal distribution), MLE may provide unreliable parameter estimates. In addition, the stability of MLE is poor in small sample scenarios and it is easily affected by outliers.

3.2 Bayesian Estimation: Parameter Inference Integrating Prior Knowledge

Bayesian estimation combines the prior distribution with the likelihood function through Bayes' theorem to obtain the posterior distribution of the parameters. Unlike MLE, the Bayesian method regards parameters as random variables rather than fixed values, thereby being able to quantify the uncertainty of the parameters themselves. For instance, in the evaluation of drug efficacy, if historical research shows that the efficacy rate of a certain drug follows a Beta distribution (prior distribution), combined with current experimental data (likelihood function), Bayesian estimation can update the posterior distribution of the drug's efficacy rate, providing more comprehensive uncertainty information for decision-making.

The advantage of Bayesian estimation lies in its ability to utilize prior knowledge to make up for the insufficiency of data, and it is particularly suitable for scenarios with small samples or high-dimensional parameters. For instance, in the diagnosis of rare diseases, due to the scarcity of patient data, MLE may not be able to provide reliable estimates. However, Bayesian methods can enhance the robustness of diagnostic models by integrating prior information from medical literature. However, Bayesian estimation has a relatively high computational complexity and requires approximate methods such as Markov Chain Monte Carlo (MCMC) or variational inference to solve the posterior distribution, which limits its application in large-scale data.

4. Bayesian Inference: A Quantitative Paradigm for Cognitive Uncertainty

4.1 Distinction between Cognitive Uncertainty and Model Uncertainty

Uncertainty can be divided into accidental uncertainty and cognitive uncertainty. Accidental uncertainty stems from the randomness of the data itself (such as sensor noise) and cannot be eliminated by increasing the data. Cognitive uncertainty stems from the model's insufficient understanding of the data (such as incomplete coverage of training data), which can be reduced by supplementing the data or improving the model. Bayesian inference provides a theoretical framework for the quantification of cognitive uncertainty by quantifying the posterior distribution of model parameters.

For instance, in autonomous driving systems, the model's insufficient understanding of rare road conditions (such as extreme weather) may lead to incorrect predictions. Bayesian Neural networks (BNN) quantify the uncertainty of the model with regard to the input by treating the weights as random variables and outputting the mean and variance of the predictions. When the input data is far from the training distribution, the prediction variance of BNN will increase, indicating that the system needs to make decisions with caution.

4.2 Bayesian Model Averaging: Uncertainty Quantification Integrating Multiple Models

Bayesian Model Averaging (BMA) quantifies the uncertainty of the model structure itself by weighting and integrating the posterior probabilities of multiple candidate models. For instance, in climate prediction, climate models with different physical assumptions may yield differentiated prediction results. BMA allocates weights based on the fit between the model and the data (posterior probability), and the final prediction result is the weighted average of the predictions of each model. At the same time, the uncertainty of the model structure is quantified through the variance of the weights.

The advantage of BMA lies in its ability to automatically select the optimal model combination, avoiding the bias of a single model. However, its computational complexity increases exponentially with the number of candidate models, and the computational cost needs to be reduced through approximate methods (such as variational inference). In addition, BMA is sensitive to the selection of the model's prior distribution, and an unreasonable prior may lead to a deviation in weight distribution.

5. Application of Probabilistic Models in Uncertainty Propagation in Complex Systems

5.1 Polynomial Chaotic Expansion: Uncertainty Propagation Based on Orthogonal Polynomials

Polynomial chaotic expansion (PCE) quantifies the impact of input uncertainty on the output by representing the system output as an orthogonal polynomial series of input random variables. For instance, in structural mechanics, fluctuations in material parameters may lead to uncertainties in structural stress. PCE assesses the risk of structural failure by constructing a polynomial expansion of stress with regard to material parameters and calculating the statistical characteristics of stress (such as mean and variance).

The advantage of PCE lies in its ability to decompose high-dimensional input uncertainties into lower-order terms, significantly reducing computational complexity. However, its applicability depends on the known probability distribution of the input variables, and the polynomial order needs to balance accuracy and computational cost.

5.2 Monte Carlo Simulation: Uncertainty Propagation Based on Random Sampling

Monte Carlo simulation generates a large number of system output samples by randomly sampling from the probability distribution of the input variables, and then statistically analyzes the distribution characteristics of the output variables. For instance, in the optimization of financial investment portfolios, fluctuations in asset returns may lead to uncertainties in the portfolio's value. Monte Carlo simulation provides a basis for risk assessment by generating thousands of yield scenarios and calculating the distribution of portfolio value.

The advantage of Monte Carlo simulation lies in that it does not rely on analytical solutions and is applicable to any complex system model. However, its computational efficiency is affected by the number of samples. In high-dimensional input scenarios, the efficiency needs to be improved through variance reduction techniques (such as importance sampling). In addition, the results of the Monte Carlo simulation are random and its stability needs to be verified through multiple runs.

6. Conclusion

The fundamental theory of probability provides a full-chain methodology for the quantification of uncertainty, ranging from the construction of probability space to parameter estimation, and from Bayesian inference to uncertainty propagation. By choosing random variables and probability distributions, probability models can quantify the accidental uncertainty of the data itself. Through parameter estimation and Bayesian inference, the model can further reveal cognitive uncertainty. Uncertainty propagation in complex systems is achieved through polynomial chaos expansion and Monte Carlo simulation. These theories and methods have demonstrated significant value in fields such as autonomous driving, medical diagnosis, and financial risk control, significantly enhancing the scientificity and reliability of decision-making.

Future research needs to be deepened in the following directions: First, develop efficient approximate inference methods to reduce the computational costs of Bayesian inference and high-dimensional PCE; Secondly, build a hybrid uncertainty quantification framework that integrates physical models and data-driven approaches to enhance the adaptability of the model in scenarios where data is scarce. Thirdly, explore the combination of causal reasoning and uncertainty quantification, distinguish false correlations from true causal effects, and provide a more accurate basis for decision-making. With the development of data science and artificial intelligence, the fundamental theories of probability theory will play a more crucial role in the quantification of uncertainty, promoting the analysis of complex systems to shift from "experience-driven" to "evidence-based".

References

[1] McDonald, S., & Campbell, D. (2021). A review of uncertainty quantification for density estimation.

[2] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?. Advances in neural information processing systems, 30.

[3] Ross, S. M. (2020). Introduction to probability and statistics for engineers and scientists. Academic press.

[4] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep

learning. In international conference on machine learning (pp. 1050-1059). PMLR.

[5] Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.

[6] Durrett, R. (2019). Probability: theory and examples (Vol. 49). Cambridge University Press.

[7] Ross, S. M. (2020). A first course in probability. Harlow, UK: Pearson.

[8] Taleb, N. N. (2020). Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. arXiv preprint arXiv:2001.10488.