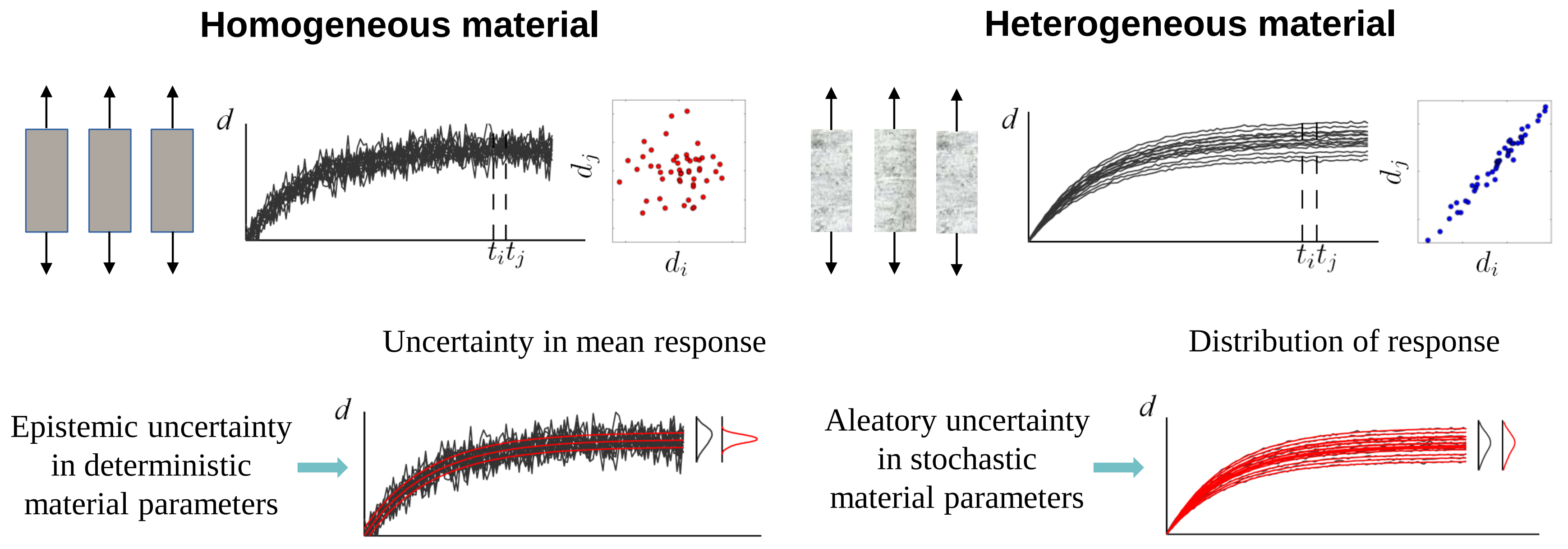
Uncertainty quantification through Bayesian nonparametric modelling

E. Kočková, A. Kučerová, J. Sýkora

CTU in Prague, Faculty of Civil Engineering, Thákurova 7, 166 29 Prague 6, Czech Republic

1 Introduction

For appropriate uncertainty quantification one has to distinguish between two principal types of uncertainties, specifically, they are epistemic and aleatory uncertainties [1], see Fig. 1. The first uncertainty type is connected to a lack of knowledge, e.g. measurement errors or a small number of measurements. This epistemic uncertainty can be reduced by any additional information. On the other side, there is aleatory uncertainty or variability which is irreducible. The aleatory uncertainty represents natural variability or randomness of a considered quantity, which arises from neglecting some problem dimension. In other words, this variability originates from data collection, when the data are singled out e.g. from different locations or times and modelled as a random variable.


Homogeneous material

Heterogeneous material

Uncertainty in mean response

Distribution of response

Epistemic uncertainty in deterministic material parameters

Aleatory uncertainty in stochastic material parameters

Fig. 1: Uncertainties in properties of a homogeneous and heterogeneous material.

An estimation of uncertain factors influencing behaviour of an investigated system is a crucial task in predicting of future events. Inferring a probability distribution which is an infinite-dimensional object is a very complex problem. Commonly applied approaches are based on low-dimensional parameterisations of the unknown density function, traditionally they consist in prescribing some specific parameterised family of probability density functions [2]. The corresponding unknown statistical moments can be considered as uncertain random variables and inferred in the Bayesian way. This approach is based on the Bayesian parametric models whose basic feature is a fixed number of unknown parameters. The significant disadvantage of this method is the necessity of making the strong assumption about the density function structure. An inappropriate guess can lead to a totally misleading result, especially in the regions of low probability which are important e.g. in reliability analysis of building structures where the design is based on a very low failure probability.

Relaxing the density structure assumption is allowed by Bayesian nonparametric modelling which serves to model selection and adaptation according to the available data. In order to ensure consistency of the estimation, in other words to obtain undistorted inference results, some prior distribution with enough large support is necessary. In the case of density estimation, it is reasonable to use an infinite-dimensional nonparametric prior on the space of density functions, i.e. to construct a probability model for the unknown probability distribution itself [3]. Commonly used nonparametric priors include stochastic processes or their mixtures, the specific setting is problem-dependent. The Gaussian processes are mostly applied in nonlinear regression problems, the mixtures of Dirichlet processes are suitable for density estimations [4]. Practically, despite the infinite dimensionality of the assumed prior, a finite dimensional formulation is employed in the computations. The model complexity is determined on a basis of the available data, it means that the dimensionality of the Bayesian nonparametric model can change with a growing data set [5].

In this contribution, we focus on estimating a probability density function of random factors from a countable number of observations with a help of the Bayesian nonparametrics allowing to capture distribution properties such as multimodality, asymmetry or heavy-tailedness. Specifically, the unknown but fixed probability density function is expressed by a hierarchical model based on the Dirichlet process mixture, which enables to model a continuous density function [6].

2 Density estimation via Bayesian nonparametrics

The most popular nonparametric method for estimating a probability distribution is a histogram, more sophisticated is a kernel density estimation widely used by frequentists. In the Bayesian nonparametrics, the Dirichlet process is well-known tool introduced as a suitable class of prior distributions with available analytical formulations of posterior distributions given a sample of observations [7]. Particularly, the Dirichlet process is a probability distribution over the set of probability distributions, i.e. every realization of the process is a probability distribution. Nevertheless, the samples of the Dirichlet process are of a discrete nature, which makes it unsuitable for the density estimation of a continuous random variable. To overcome this obstacle, a hierarchical model based on the Dirichlet process is utilized producing a mixture of Dirichlet processes also called a Dirichlet process mixture (DPM) model [8].

Assuming a set of statistically exchangeable i.i.d. samples

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim F, \tag{1}$$

where $\sim$ stands for “distributed according to” and $x_i \in \boldsymbol{R}$, the goal is to infer the unknown probability density function f as a DPM model, where

$$f(\boldsymbol{x}) = \sum_{j=1}^{\infty} w_j \, g_\theta\big(\boldsymbol{x}|\boldsymbol{\theta}_j\big), \tag{2}$$

which is an infinite weighted mixture of smooth probability densities from a parametric family $\boldsymbol{G} = \{g_\theta | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with latent variables $\boldsymbol{\theta}$. Weights w_j represent a Dirichlet process and their sum is equal to one. Considering P_0 as a probability measure on the parameter space $\boldsymbol{\Theta}$, the DPM has the following hierarchical structure:

$$\begin{aligned} P &\sim \mathrm{DP}(\alpha, P_0) \\ \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | P &\sim P \\ \boldsymbol{x}_i | \boldsymbol{\theta}_i &\sim g_\theta(\boldsymbol{x}, \boldsymbol{\theta}_i), \quad i = 1, \dots, n. \end{aligned} \tag{3}$$

A random probability distribution P is generated by a Dirichlet process with a positive scalar α called a concentration (or precision) parameter because it defines a spread of the prior probability distribution P around the base (or center) distribution P_0, which is the prior expectation of P. A higher value of α means a higher level of the centralization.

3 Illustrative example

We present a simple example of observations from a mixture of normally distributed random variables where the observed data coincide with the random effect whose unknown probability distribution is the object of the Bayesian inference. In this case, the densities g_θ are assumed to be Gaussians with unknown mean values $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The base distribution P_0 is assumed to be the normal-inverse-Wishart distribution which is conjugate prior distribution for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and has its own four parameters. Multiplying this prior density by the normal likelihood gives a posterior density of the same family, which fundamentally simplifies the actual computations [4]. The inference is focused on the marginalized posterior distribution $p(\boldsymbol{\theta}_{1:n}|\boldsymbol{x}_{1:n})$ since the infinite-dimensional P is integrated out with a help of Polya urn representation of the Dirichlet process [9]. The posterior samples can be obtained almost directly by Gibbs sampling. The estimated density function in a comparison with the true density and observations is depicted in Fig. 2.

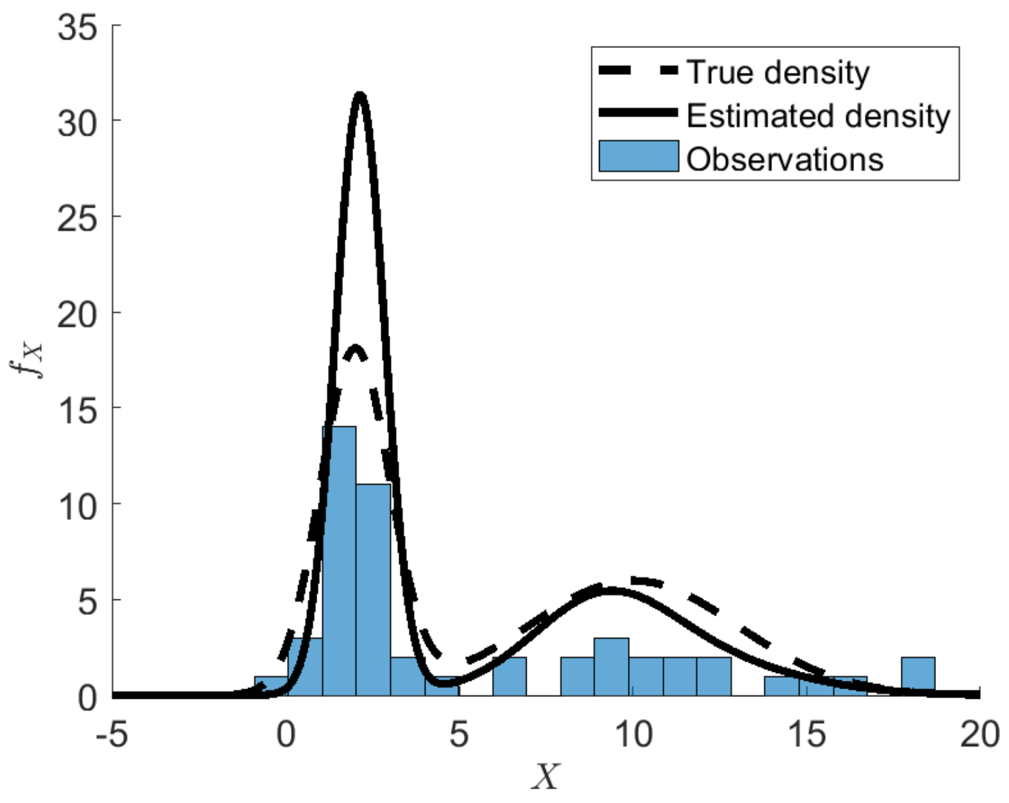


Fig. 2: Example of density estimation for mixture of two Gaussians $0.5\mathrm{N}(2{,}1){+}0.5\mathrm{N}(10{,}3)$. *Comparison of true and estimated probability density function based on Dirichlet process mixture of Gaussians considering set of 50 observations.*

4 Conclusion

The Bayesian nonparametric methods enable to quantify uncertainties more precisely without making restrictive assumptions about their probability distributions as it is done in the parametric approaches, where the structure and a number of parameters of the estimated probability density function are prescribed a priori. Specifically, properties such as multimodality or asymmetry of the density function are usually omitted which can lead to unrealistic predictions and then to a wrong evaluation of risks connected to the modelled system.

Usually, a limited number of observations of the uncertain effect is available and the hierarchical model based on the Dirichlet process mixture allows to share information among these samples. The nonparametric inference results in the density estimation of aleatory uncertainty formulated as a weighted finite-dimensional mixture of densities with random parameters. The number of components is determined on a basis of clustering the processed data so the parameterisation is not fixed.

This paper gives a very brief view into the world of the Bayesian nonparametrics with a simple illustrative example, however modelling density estimation especially in higher dimensions is not trivial. This topic is very actual and different effective methods have been developed in this area. Besides using the Dirichlet process mixtures, some researchers are focused on constructing hierarchical models based on the Pólya tree [10]. Another method is based on separating marginal and joint distribution by using copula transform [11].

References

[1] W. L. Oberkampf et al.: Error and uncertainty in modeling and simulation. *Reliab. Eng. Syst. Safe*. 75 (2002), 333-357.

[2] J. B. Nagel, B. Sudret: A unified framework for multilevel uncertainty quantification in Bayesian inverse problems. *Probabilistic Eng. Mech.* 43 (2016), 68-84.

[3] S. N. MacEachern: Nonparametric Bayesian methods: a gentle introduction and overview. *Commun. Stat. Appl. Methods.* 23 (2016), 6, 445-466.

[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin: *Bayesian data analysis.* 3rd ed., Boca Raton: CRC Press, 2014.

[5] S. J. Gershman, D. M. Blei: A tutorial on Bayesian nonparametric models*. J. Math. Psychol*. 56 (2012), 1, 1-12.

[6] S. Ghosal, A. Van der Vaart: *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press. Cambridge series in statistical and probabilistic mathematics, 2017.

[7] T. S. Ferguson: A Bayesian analysis of some nonparametric problems. *Ann. Stat*. 1 (1973), 2, 209-230.

[8] C. E. Antoniak: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat*. 2 (1974), 6, 1152-1174.

[9] D. Blackwell, J. B. MacQueen: (1973) Ferguson Distributions Via Polya Urn Schemes. *Ann. Stat*. 1 (1973), 2, 353-355.

[10] J. Christensen, L. Ma: A Bayesian hierarchical model for related densities by using Pólya trees. *J. R. Stat. Soc. Ser. B Methodol.* 82 (2019), 1, 127-153.

[11] A. Majdara, S. Nooshabadi: Nonparametric Density Estimation Using Copula Transform, Bayesian Sequential Partitioning and Diffusion-Based Kernel Estimator. *IEEE Trans. Knowl. Data Eng.* 32 (2020), no. 4, 821-826.

Acknowledgement

The authors gratefully acknowledge the financial support from the Czech Science Foundation (Project No. 18-04262S) and the Grant Agency of the Czech Technical University in Prague (Project No. SGS20/039/OHK1/1T/11).

GAČR
CZECH SCIENCE FOUNDATION

CZECH
TECHNICAL
UNIVERSITY
IN PRAGUE