MascotNum Annual Conference, April 23-25, 2014, ETH Zürich (Switzerland)

Bayesian Multilevel Model Calibration for Inversion of "Perfect" Data in the Presence of Uncertainty

Joseph B. Nagel & Bruno Sudret
ETH Zürich

Affiliation: *Chair of Risk, Safety and Uncertainty Quantification*, ETH Zürich, D-BAUG, Institute of Structural Engineering, Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland

Email: nagel@ibk.baug.ethz.ch – **URL**: http://www.ibk.ethz.ch/su/people/nagelj

Master: Rheinische Friedrich-Wilhelms-Universität Bonn

Ph.D. (2012-2015): ETH Zürich

Supervisor(s): Prof. Dr. Bruno Sudret (ETH)

Abstract: Hierarchical or multilevel modeling establishes a convenient framework for solving complex inverse problems [1, 2] in the presence of uncertainty. In the last two decades it has been studied from a frequentist [3] and a Bayesian perspective [4]. We will adopt a Bayesian point of view to statistical inversion and uncertainty quantification and present a Bayesian multilevel framework that allows for inversion and optimal analysis of "perfect" or noise-free data in the presence of aleatory and epistemic types of uncertainty and in experimental situations when data is scarce or expensive to acquire. In this contribution to the annual MascotNum workshop we will discuss the abovementioned framework on the basis of an application example within the domain of aerospace engineering [5]. We will not only illustrate the very potential of Bayesian multilevel modeling as well as ways to overcome its immanent major challenges, but more importantly we will discuss the main observations, considerations and key questions that the practical problem solution [6] has given rise to.

A forward model $\mathcal{M}\colon (\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \mapsto \tilde{\boldsymbol{y}}$ describes a system or phenomenon under consideration. Throughout a number of $i = 1, \ldots, n$ experiments forward model inputs may be represented corresponding to a certain model of epistemic and aleatory uncertainty. There are fixed albeit insufficiently well-known model parameters $\boldsymbol{m}$, model inputs $\boldsymbol{x}$ and $\boldsymbol{\zeta}$ that are subject to imperfectly or perfectly known variability, respectively, and experimental conditions $\boldsymbol{d}$ that are known with certainty. Constant yet unknown model parameters are represented as random variables $\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m})$ where $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ is a Bayesian prior belief about their true values. Model inputs with perfectly known variability are modeled as experiment-specific realizations $\boldsymbol{\zeta}_i$ of random variables $(\boldsymbol{Z}_i|\boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})$ with known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}}$ that prescribe the variability. Model inputs with imperfectly known variability are modeled as experiment-specific realizations $\boldsymbol{x}_i$ of exchangeable random variables $(\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$ with hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ about which only Bayesian prior knowledge $\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ is available. Experimental conditions $\boldsymbol{d}_i$ possibly differ throughout the experiments yet they are (deterministic) perfectly known values.

A "complex" inverse problem is posed when model responses $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ are measured in n experiments, forward model inputs comply with the aforementioned uncertainty model and inference focuses on the unknowns $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. While classical Bayesian multilevel modeling deals with the analysis of "imperfect" data $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$, i.e. model-measurement discrepancy is accounted for by residual terms that are modeled as outcomes $\boldsymbol{\varepsilon}_i$ of a random variables $\boldsymbol{E}_i \sim f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i)$ with distributions $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i)$, the problem formulation at hand deals with "perfect" data $\tilde{\boldsymbol{y}}_i$. Interestingly, in the context present the analysis of "perfect" data is more involved than the analysis of "imperfect" data in mathematical and numerical terms. Thus firstly we will devise a Bayesian multilevel model involving "perfect" data. Subsequently we will show how Bayesian calibration of the formulated multilevel model can be accomplished by analyzing the entirety of collected data $\langle \tilde{\boldsymbol{y}}_i \rangle = (\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n)$. The inferential prior distribution $\pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$, that represents the knowledge about the quantities of interest $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ prior to analyzing the data, will be updated in order to obtain the posterior distribution $\pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \tilde{\boldsymbol{y}}_i \rangle)$. To that end a likelihood function $\mathcal{L}(\langle \tilde{\boldsymbol{y}}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \boldsymbol{\theta}_{\boldsymbol{Z}})$ has to be formulated as well as a means for its efficient evaluation.

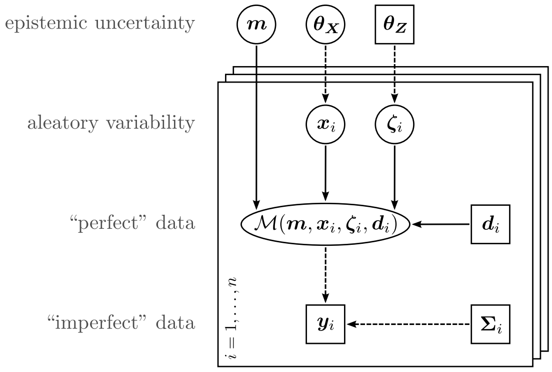

Figure 1: **DAG of the generic multilevel model.** Vertices symbolize unknown (○) or known (□) quantities and directed edges represent their deterministic (→) or probabilistic (⇢) relations. Quantities are shown in a way that reflects their uncertainty.

Since for the specific problem at hand such a likelihood function is not available in closed-form, we will propose a statistical simulator of the likelihood which is based on Monte Carlo (MC) sampling and kernel density estimation (KDE). Moreover, in order to explore the posterior of the quantities of interest, we will devise dedicated Markov chain Monte Carlo (MCMC) algorithms. The very principle of MCMC is to construct a Markov chain whose long-run distribution approaches the desired posterior. By virtue of Bayes' law closed-form approximations of the likelihood directly induce approximations on the level of the posterior. However, if calls to the likelihood function $\mathcal{L}$, over the course of the Markov chain, are replaced by calls to a statistical estimator $\hat{\mathcal{L}}$, an approximation is introduced on the level of the Markov chain transition kernel. This raises the distinctly important question as to which degree the induced equilibrium distribution is in conformity with the true posterior, i.e. the issue of *posterior fidelity*. In turn the practical question becomes how to "optimally" tune free algorithmic parameters, e.g. the number of MC samples, the bandwidth of the KDE and parameters of the MCMC simulation. We will present a heuristic way of approaching those delicate issues. Beyond that, we will demonstrate how data augmentation [7] can be utilized in the outlined multilevel context. Data augmentation is a powerful technique from the vast MCMC toolkit that traditionally aims at enhancing MCMC efficiency by introducing hidden data as auxiliary variables. Instead we will herein introduce latent data as auxiliary variables in order to enhance the adequacy of likelihood estimations and the fidelity of the posterior densities that are eventually obtained.

Short biography – Joseph Benjamin Nagel studied theoretical physics and wrote his diploma thesis about quantum field theories, MCMC simulations and GPU computations. After working on geophysical variational data assimilation he joined the Chair of Risk, Safety and Uncertainty Quantification, where his research focuses on "Bayesian techniques for model calibration and stochastic inverse problems".

References

[1] E. de Rocquigny and S. Cambier. Inverse probabilistic modelling of the sources of uncertainty: A non-parametric simulated-likelihood method with application to an industrial turbine vibration assessment. *Inverse Prob. Sci. Eng.*, 17(7):937–959, 2009.

[2] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. Nonlinear methods for inverse statistical problems. *Comput. Stat. Data Anal.*, 55(1):132–142, 2011.

[3] H. T. Banks, Z. R. Kenz, and W. C. Thompson. A review of selected techniques in inverse problem nonparametric probability distribution estimation. *J. Inverse Ill-posed Prob.*, 20(4):429–460, 2012.

[4] D. J. Lunn. Bayesian analysis of population pharmacokinetic/pharmacodynamic models. In D. Husmeier, R. Dybowski, and S. Roberts, editors, *Probabilistic Modeling in Bioinformatics and Medical Informatics*, Advanced Information and Knowledge Processing, pages 351–370. Springer, London, 2005.

[5] L. G. Crespo, S. P. Kenny, and D. P. Giesy. The NASA Langley multidisciplinary uncertainty quantification challenge, 2013. NASA Langley Research Center.

[6] J. B. Nagel and B. Sudret. A Bayesian multilevel framework for uncertainty characterization and the NASA Langley multidisciplinary UQ challenge. In *Proc. 16th AIAA Non-Deterministic Approaches Conference (SciTech 2014), National Harbor, Maryland*, January 13-17, 2014.

[7] D. A. Dyk and X.-L. Meng. The art of data augmentation. *J. Comp. Graph. Stat.*, 10(1):1–50, 2001.