Computers & Geosciences 85 (2015) 38–48

Contents lists available at ScienceDirect

Computers & Geosciences

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

Case study

Accelerating Monte Carlo Markov chains with proxy and error models

Laureline Josset [a,*], Vasily Demyanov [b], Ahmed H. Elsheikh [b], Ivan Lunati [a]

[a] ISTE, University of Lausanne, Geopolis – UNIL Mouline, 1015 Lausanne, Switzerland
[b] IPE, Heriot-Watt University, Edinburgh, UK

ARTICLE INFO

Article history:
Received 26 November 2014
Received in revised form
17 April 2015
Accepted 7 July 2015
Available online 10 July 2015

Keywords:
Functional data analysis
Bayesian inference
Two-stage MCMC
Imperial College Fault test case

ABSTRACT

In groundwater modeling, Monte Carlo Markov Chain (MCMC) simulations are often used to calibrate aquifer parameters and propagate the uncertainty to the quantity of interest (e.g., pollutant concentration). However, this approach requires a large number of flow simulations and incurs high computational cost, which prevents a systematic evaluation of the uncertainty in the presence of complex physical processes. To avoid this computational bottleneck, we propose to use an approximate model (proxy) to predict the response of the exact model. Here, we use a proxy that entails a very simplified description of the physics with respect to the detailed physics described by the "exact" model. The error model accounts for the simplification of the physical process; and it is trained on a learning set of realizations, for which both the proxy and exact responses are computed. First, the key features of the set of curves are extracted using functional principal component analysis; then, a regression model is built to characterize the relationship between the curves. The performance of the proposed approach is evaluated on the Imperial College Fault model. We show that the joint use of the proxy and the error model to infer the model parameters in a two-stage MCMC set-up allows longer chains at a comparable computational cost. Unnecessary evaluations of the exact responses are avoided through a preliminary evaluation of the proposal made on the basis of the corrected proxy response. The error model trained on the learning set is crucial to provide a sufficiently accurate prediction of the exact response and guide the chains to the low misfit regions. The proposed methodology can be extended to multiple-chain algorithms or other Bayesian inference methods. Moreover, FPCA is not limited to the specific presented application and offers a general framework to build error models.

1. Introduction

Simulations of subsurface flow are important in many applications, such as groundwater protection and remediation, water prospection, exploration of hydrocarbon resources, and nuclear waste disposal. One of the main challenges is to estimate a continuous distribution of the underground model parameters from a sparse set of observational sites. This lack of information on model input propagates to the quantities of interest (for instance, the concentration of a pollutant in a drinking well), whose exact values remain uncertain. Model calibration using historical integrated data (for example, time series of concentration or pressure at observation wells) is often used to reduce the uncertainty on model parameters by relying on Bayes theorem. A widespread approach for numerical application of Bayes rule is to use Monte-Carlo Markov-Chain (MCMC) simulations (Robert and Casella, 2004) to sample the posterior probability density function. While MCMC is theoretically robust and ensures convergence to the true posterior distribution under mild constraints, in practice it is subject to several limitations due to the cost of the large number of required flow simulations, which can become prohibited in the presence of limited computational resources. Indeed, the finite length chains should be able to explore all areas of the prior space in order to provide samples from the posterior distribution. To achieve this goal, it is tempting to increase the step length of the chains, but this would result in a drastic reduction of the acceptance rate (which should ideally remain around 20–50% in multidimensional space) and subsequently in a high number of wasted simulations (Roberts et al., 1997).

To avoid these issues, Efendiev et al. (2005, 2006) and Christen and Fox (2005) have introduced a two-stage MCMC, which employs a less computationally expensive solver to obtain a first evaluation of the proposal and decide whether it is useful to run the exact solver. This allows them to reduce the number of exact simulations that will be rejected and thus increase the acceptance rate. This methodology has been first explored by Christen and Fox (2005) to recover resistor values of an electrical network from

* Corresponding author.
E-mail addresses: laureline.josset@unil.ch (L. Josset), vasily.demyanov@pet.hw.ac.edu.uk.co (V. Demyanov), ahmed.elsheikh@pet.hw.ac.edu.uk.co (A.H. Elsheikh), ivan.lunati@unil.ch (I. Lunati).

http://dx.doi.org/10.1016/j.cageo.2015.07.003
0098-3004/

measurements performed at the network boundary. They have obtained an increase in acceptance rate (the number of exact simulations accepted over the number of exact simulations run; first-stage simulations are not taken into account as their cost is assumed to be negligible). Both Efendiev et al. (2006) and Christen and Fox (2005) have shown that, under certain hypotheses, the solution converges to the posterior distribution. Efendiev et al. (2005, 2006) and Dostert et al. (2008) have applied this methodology in the context of flow in porous media. As first-stage solver they have used a multiscale method, which combines a global coarse solution with a number of local fine solutions. If the coarse solution is accepted, local solutions are employed to reconstruct a finer solution on the original grid, based on which the second-stage evaluation is performed. While this allows for the necessary convergence assumptions to be satisfied (namely, smoothness and strong correlation), the computational gain of the two-stage set-up is limited. Indeed, the reconstruction step (necessary for the second-stage evaluation) is cheap with respect to the cost of constructing and solving the coarse problem used at the first-stage. Other applications of two-stage MCMC have used polynomial chaos response surfaces (Zeng et al., 2012; Elsheikh et al., 2012; Laloy et al., 2013) as first-stage model. The computational gain is much higher, despite some additional cost required to set up the polynomial chaos model.

The use of inexact solvers requires designing error models to account for the discrepancy between approximate and exact responses. In the context of multiscale approaches, Kennedy and O'Hagan (2001) used a Gaussian-process method to represent model inadequacy. O'Sullivan and Christie (2005, 2006) employed error modeling to reduce the bias in history matching resulting from the use of upscaled reservoir models. Efendiev et al. (2009) proposed non-linear error models in the context of ensemble-level upscaling. Scheidt et al. (2010), for instance, used a distance metric to account for upscaling errors in ensemble history matching. More specifically to two-stage MCMC, Cui et al. (2011) proposed to adapt the error model at each iteration: they used information on the discrepancy between the exact and approximate models at the previous iteration to correct the result of the successive iteration. However, this approach provides a good correction only for problems that are smooth enough.

Here, we propose a different strategy that combines a two-stage MCMC set-up with a methodology recently presented by Josset et al. (2015). We use an approximate model (proxy) that assumes a very simplified physics with respect to the problem under consideration, and we construct an error model to account for the approximation errors. The error model is purpose oriented as it is tailored directly for the quantities of interest following an approach typical of machine learning. For a subset of realizations, the responses of both the proxy and the exact models are evaluated and the mapping between the two is learned by means of tools from functional data analysis (Ramsay, 2006; Ramsay et al., 2009). Josset et al. (2015) applied this methodology to propagate the uncertainty on the permeability field to the concentration of a pollutant in the observational well. Here, the methodology is tested on a complex problem of Bayesian inference, the Imperial College Fault (ICF) test case, which is a benchmark problem first published by Tavassoli et al. (2004) and repeatedly explored in many studies (e.g., Demyanov et al., 2010; Mohamed et al., 2011, 2012).

The paper is structured as follows: we first describe the ICF test case and review the literature about the calibration of this model (Section 2). Next, we present the novel methodology, which uses a purpose-oriented error model within a two-stage MCMC set-up (Section 3). Then, we specifically construct and evaluate the error-model approach for the ICF problem (Section 4.1). Finally, we compare and discuss the results of the two-stage MCMC with the classic Metropolis–Hastings algorithm (Section 4.2).

2. The Imperial College Fault (ICF) test case

The ICF test case was first published by Tavassoli et al. (2004, 2005) as a simple yet challenging example of history matching in petroleum engineering applications. Since then, ICF has proved a difficult test for optimization techniques due to numerous local minima. The ICF model consists of a layered reservoir disrupted by a fault (Fig. 1), in which water is injected at the left-hand boundary while the displaced fluids are recovered at the right-hand boundary. The layer-cake model of the reservoir permeability is described by three parameters: the conductivity of the high permeability facies, K_{high}, the conductivity of the low permeability facies, K_{low}, and the fault throw, h. The true parameters are K_{high} = 131.6 md, K_{low} = 1.3 md and h = 10.4 ft. A uniform distribution $\mathcal{U}_{[a,b]}$ (where a and b are the bounds of the distribution) is attributed to each parameter as prior.

The calibration of the parameters to the observational data (oil and water production rates) appeared to be a challenging history matching problem. Due to the nature of the permeability field, several parameter combinations, corresponding to narrow regions of the parameter space, can reproduce the observational data with satisfactory accuracy. Between these regions of good quality, the misfit is very high due to the very irregular response surface that results from the strong fluctuations of the connectivity across the fault when h is varied. We refer to Fig. 9 for a 1D cross-section cut of the complex misfit surface that characterizes this problem.

Many optimizations and inference techniques have been applied to the ICF problem over the years. The first studies of this test case (Tavassoli et al., 2004, 2005; Carter et al., 2006) have employed a pure Monte Carlo approach, which required nearly 160,000 samples of the parameter space. Christie et al. (2006) demonstrated that a good representation of the uncertainty can be inferred from a few thousand samples using Genetic Algorithm Important Sampling with artificial neural network proxy. More recently, Demyanov et al. (2010) have used Support Vector Machines (SVM) with a small number of flow simulations (about 700); and Mohamed et al. (2011) have employed Particle Swarm Optimization (PSO) using 2050 flow simulations. A Bayesian inference approach close to two-stage MCMC has been presented by

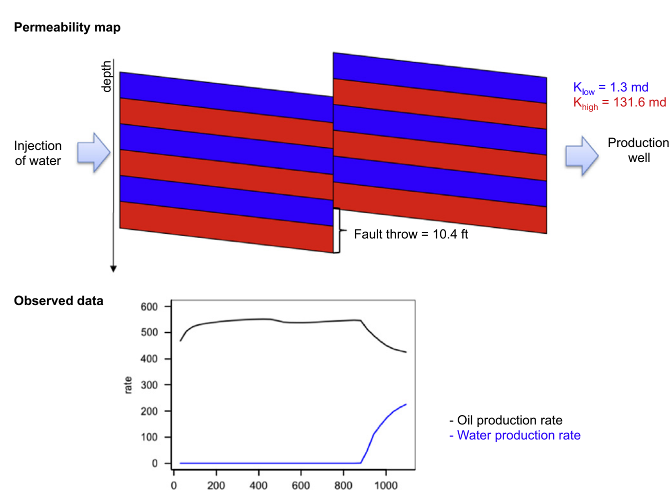

Fig. 1. The permeability map of the ICF test case and the observed data used for the history matching. As prior, a uniform distribution is attributed to each parameter, i.e., $P(h) = \mathcal{U}_{[0,60]}$ for the fault throw h, $P(K_{high}) = \mathcal{U}_{[100,200]}$ for the permeability of the most permeable facies K_{high}, and $P(K_{low}) = \mathcal{U}_{[0,50]}$ for the permeability of the least permeable facies K_{low}.