Advances in Water Resources 76 (2015) 1–10

Contents lists available at ScienceDirect

Advances in Water Resources

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

Inversion and uncertainty of highly parameterized models in a Bayesian framework by sampling the maximal conditional posterior distribution of parameters

Thierry A. Mara [a,*], Noura Fajraoui [b], Anis Younes [b], Frederick Delay [b]

[a] PIMENT, EA 4518, Université de La Réunion, FST, 15 Avenue René Cassin, 97715 Saint-Denis, Reunion
[b] LHyGeS, UMR-CNRS 7517, Université de Strasbourg/EOST, 1 rue Blessig, 67084 Strasbourg, France

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 26 November 2014
Accepted 27 November 2014
Available online 8 December 2014

Keywords:
Inverse modeling
Bayesian parameter estimation
Model parameter identification
Highly parameterized model
Heterogeneous transmissivity field
Karhunen–Loève expansion

ABSTRACT

We introduce the concept of maximal conditional posterior distribution (MCPD) to assess the uncertainty of model parameters in a Bayesian framework. Although, Markov Chains Monte Carlo (MCMC) methods are particularly suited for this task, they become challenging with highly parameterized nonlinear models. The MCPD represents the conditional probability distribution function of a given parameter knowing that the other parameters maximize the conditional posterior density function. Unlike MCMC which accepts or rejects solutions sampled in the parameter space, MCPD is calculated through several optimization processes. Model inversion using MCPD algorithm is particularly useful for highly parameterized problems because calculations are independent. Consequently, they can be evaluated simultaneously with a multi-core computer. In the present work, the MCPD approach is applied to invert a 2D stochastic groundwater flow problem where the log-transmissivity field of the medium is inferred from scarce and noisy data. For this purpose, the stochastic field is expanded onto a set of orthogonal functions using a Karhunen–Loève (KL) transformation. Though the prior guess on the stochastic structure (covariance) of the transmissivity field is erroneous, the MCPD inference of the KL coefficients is able to extract relevant inverse solutions.

1. Introduction

Models are tools on which environmental risk-assessment and decision-making strategies can rely, provided it is proved that the models are relevant to the problem under investigation. This relevance can be addressed by facing a model prediction to observation data knowing that the whole procedure also requires assigning model parameter values. Some parameters can be directly measured while some others ought to be indirectly estimated by comparing model predictions with observations. The present work addresses the issue of parameter identification for highly parameterized models. The notion of identification encompasses seeking the parameter values and assessing the uncertainty on parameters and on model predictions.

During the past two decades, the increasing power of computers was conducive to emphasize and promote the so-called Bayesian parameter estimation techniques. In essence, the Bayesian framework leads to the definition of the parameter joint posterior probability density function (pdf), for instance inferred by means of Markov Chain Monte Carlo (MCMC) samplings [1–4]. The notion of posterior pdf is associated with the fact that the parameter's pdf is conditioned both on plausible (prior) parameter values and on observation data. MCMC provides draws directly sampled from the posterior pdf which leads to exploration of the plausible areas in the parameter space. The Bayesian estimation using MCMC has been subject to many developments and improvements during the last decade (e.g. [5–8] among others). However, MCMC samplers remain computationally expensive because many draws are rejected by the statistical test embedded in the sampler. Furthermore, with MCMC, the parameters marginal posterior distributions cannot be investigated independently. Recently, several strategies have been proposed to increase MCMC efficiency (see [9–13]).

In the present work we propose a new method, partly grounded in optimization techniques, to cope with the identification of model parameters. The first step of this approach is to seek all the probable local optima of the joint posterior pdf of the whole set of parameters (including the maximum a posteriori estimate). Next, several maximizations of the conditional pdf are performed for different prescribed values of one selected parameter. The

* Corresponding author.
E-mail address: mara@univ-reunion.fr (T.A. Mara).

http://dx.doi.org/10.1016/j.advwatres.2014.11.013
0309-1708/

values assigned to this parameter are picked from a range around its optimal value(s). The value of the other parameters is investigated by maximizing the conditional pdf. This provides what we call the maximal conditional posterior distribution (MCPD) of the selected parameter. It actually corresponds to a discrete approximation of the pdf of a single parameter conditioned on data such that the conditional pdf is maximized.

The MCPD returns information about the model parameter values supported by the data and any correlations between parameters. The MCPD sample also allows uncertainty bounds to be assigned to the model predictions. The main advantage of the approach is that MCPD inferences for different parameters are independent and can be evaluated simultaneously by easily distributing the calculations over a multi-core computer (or several computers). This feature drastically decreases the computation time and makes the inversion of highly parameterized problems feasible.

The main topics addressed in the present paper are organized as follows. A short outline on inverse modeling within a Bayesian framework is proposed in Section 2, and then followed by the details on the MCPD sampling in Section 3. The first exercise testing the MCPD approach is proposed in Section 4 and addresses the ability of the sampler to retrieve a multimodal probability density function. The second test in Section 5 applies the MPCD approach to identify the Karhunen–Loève expansion [14] of a stochastic transmissivity field for a two-dimensional steady-state groundwater flow problem.

2. Bayesian inference

In inverse modeling, the parameter set (of size s) $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_s\}$ of a given model is estimated from a set of observation data $\mathbf{d}$. In the following, we assume that the model does not suffer from misconceptions. The model is therefore supposed to be exact regarding the processes and the system that it mimics. However, observation data remain uncertain (random variables) making the model parameters to be also random and characterized by a joint probability density function $p(\boldsymbol{\theta})$. We denote by Ω_i the probable prior uncertainty range of θ_i. In a Bayesian framework, the parameter joint posterior pdf is defined by

$$p(\boldsymbol{\theta} \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{d})} \tag{1}$$

where $p(\mathbf{d})$ is a scaling factor called evidence, $p(\boldsymbol{\theta})$ is the prior density corresponding to a first guess on parameters before collecting the observations, while $p(\mathbf{d} \mid \boldsymbol{\theta})$ is called the likelihood function measuring how well the model describes the data.

The parameter set that maximizes Eq. (1)

$$\boldsymbol{\theta}^{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbf{d}) \tag{2}$$

is called the maximum a posteriori estimate. It is the most probable parameter set given our knowledge about the system (i.e. the data $\mathbf{d}$ and the prior pdf of the parameters $p(\boldsymbol{\theta})$) and it is sought by appropriate optimization algorithms (e.g., descent methods, evolutionary algorithms, etc ...). Unfortunately, finding $\boldsymbol{\theta}^{MAP}$ does not allow to (fully) characterize the posterior uncertainty of the parameters (except for linear models, see [15]). This uncertainty should be assessed by calculating the marginal posterior density for each parameter, defined as follows

$$p(\theta_i \mid \mathbf{d}) = \int p(\theta_i, \boldsymbol{\theta}_{-i} \mid \mathbf{d}) d\boldsymbol{\theta}_{-i}, \quad \forall i = 1, \ldots, s \tag{3}$$

where $\boldsymbol{\theta}_{-i}$ represents the vector of parameters $\boldsymbol{\theta}$ without θ_i. The integral in (3) can be approximated by a multidimensional quadrature method or by a sampling-based method such as the Markov Chain Monte Carlo (MCMC). Nevertheless, the computational effort can be prohibitive and sometimes unaffordable for problems with a large number of parameters.

In the present work, we propose an optimization-based method in order to assess the parameter uncertainty for models post-conditioned on available observation data. For this purpose, we introduce the concept of maximal conditional posterior distribution. One could raise that relying on an optimization-based method will require solving many problems, as is classical with standard inversion techniques when obtaining a large set of solutions is contemplated. As shown hereafter, the maximal conditional posterior distribution has some specific features diminishing the calculation loads.

3. Maximal conditional posterior distribution

3.1. The concept

We define the maximal conditional posterior distribution (MCPD) of θ_i as

$$\mathcal{P}_i(\theta_i) = \max_{\boldsymbol{\theta}_{-i}}(p(\boldsymbol{\theta}_{-i} \mid \mathbf{d}, \theta_i)) \times p(\theta_i \mid \mathbf{d}) \tag{4}$$

$\mathcal{P}_i(\theta_i)$ is interpreted as the posterior probability function that maximizes the conditional posterior distribution $p(\boldsymbol{\theta}_{-i} \mid \mathbf{d}, \theta_i)$ and encompasses the MAP probability (i.e. $\mathcal{P}_i(\theta_i^{MAP}) = p(\boldsymbol{\theta}^{MAP} \mid \mathbf{d})$). By using the Bayes theorem, one can write, $\max(p(\boldsymbol{\theta}_{-i} \mid \mathbf{d}, \theta_i)) \times p(\theta_i \mid \mathbf{d}) = \max(p(\theta_i \mid \boldsymbol{\theta}_{-i}, \mathbf{d}) \times p(\boldsymbol{\theta}_{-i} \mid \mathbf{d}))$. Therefore, the MCPD in (4) can also be viewed as the distribution of the parameter θ_i, knowing that the other parameters $\boldsymbol{\theta}_{-i}$ are at their optimal values. The MCPD of θ_i is assessed in a discrete form by sampling Eq. (4). A parameter θ_i is frozen at a prescribed value and the other parameters $\boldsymbol{\theta}_{-i}$ are optimized to find (according to the Bayesian definition) the maximal probability of these parameters. Changing the prescribed value of θ_i allows scanning the distribution of θ_i. In practice, the sampled values of θ_i (denoted below θ_i^*) are picked around the MAP estimate θ_i^{MAP} (estimated beforehand) within its prior uncertainty range Ω_i (see Fig. 1). This gives,

$$\boldsymbol{\theta}_{-i}^* = \arg\max_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta}_{-i} \mid \mathbf{d}, \theta_i = \theta_i^*) \tag{5}$$

$$\mathcal{P}_i(\theta_i^*) = p\left(\boldsymbol{\theta}_{-i}^* \mid \mathbf{d}, \theta_i^*\right) \times p(\theta_i^* \mid \mathbf{d}) = p(\boldsymbol{\theta}^* \mid \mathbf{d}) \tag{6}$$

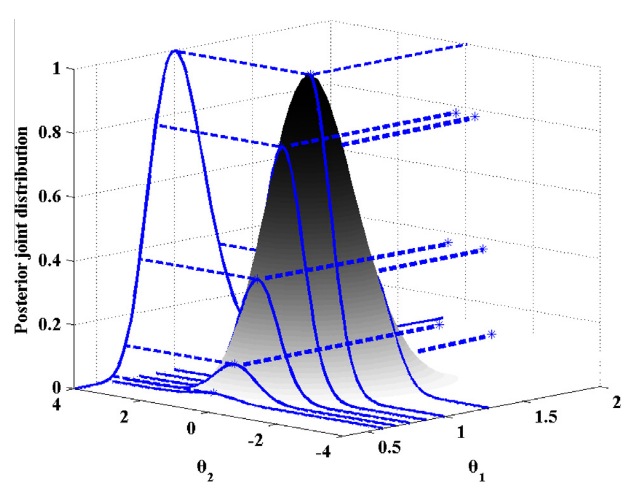

Fig. 1. A two-dimensional illustration of the MCPD assessment. The MCPD of θ_1 is built by maximizing the conditional joint posterior distribution. During this maximization, the most probable value of θ_2 is derived while the value of θ_1 is fixed. This operation is repeated by fixing θ_1 farther and farther from its maximum a posteriori estimate.