Journal of Computational Physics 348 (2017) 694–714

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Discovering variable fractional orders of advection–dispersion equations from field data using multi-fidelity Bayesian optimization

Guofei Pang [a], Paris Perdikaris [b], Wei Cai [c], George Em Karniadakis [d,*]

[a] Algorithms Division, Beijing Computational Science Research Center, Beijing 100193, China
[b] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Department of Mathematics, Southern Methodist University, Dallas, TX 75275, USA
[d] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 7 May 2017
Received in revised form 1 July 2017
Accepted 27 July 2017
Available online 2 August 2017

Keywords:
Fractional modeling
Porous media
Machine learning
Inverse problem
Gaussian process regression
Uncertainty quantification
Model uncertainty

ABSTRACT

The fractional advection–dispersion equation (FADE) can describe accurately the solute transport in groundwater but its fractional order has to be determined *a priori*. Here, we employ multi-fidelity Bayesian optimization to obtain the fractional order under various conditions, and we obtain more accurate results compared to previously published data. Moreover, the present method is very efficient as we use different levels of resolution to construct a stochastic surrogate model and quantify its uncertainty. We consider two different problem set ups. In the first set up, we obtain variable fractional orders of one-dimensional FADE, considering both synthetic and field data. In the second set up, we identify constant fractional orders of two-dimensional FADE using synthetic data. We employ multi-resolution simulations using two-level and three-level Gaussian process regression models to construct the surrogates.

1. Introduction

The fractional advection–dispersion equation (FADE) is an effective mathematical model in describing the solute transport in porous and fractured media, such as contaminant transport in aquifers [1–3]. However, being a phenomenological model, the FADE requires parameter calibration before being applied to predict concentration profiles, namely, the phenomenological parameters, fractional derivative orders, have to be determined by fitting the experimental data. As a result, a parameter identification inverse problem needs to be solved. Due to the high computational cost of solving forward problems using fractional discretization multiple times, sometimes hundreds of times, reaching a converged inverse solution presents a serious computational challenge. Recent efforts for fractional order identification have mainly focused on time-fractional advection-dispersion equations (ADEs) [4–9] and, to a lesser extent, on parameter identification of the space-fractional ADEs but only limited to constant fractional order [10–12]. The FADEs for describing the real-world solute transport, however, are often complex [1] and involve 2-D or 3-D spatial fields as well as variable fractional orders. Thus, the parameter identification of these FADEs is more challenging than those previously considered in the literature.

* Corresponding author.
E-mail address: george_karniadakis@brown.edu (G.E. Karniadakis).

http://dx.doi.org/10.1016/j.jcp.2017.07.052
0021-9991/

To the best of our knowledge, parameter identification of high-dimensional and variable-order FADEs has not yet been investigated using a systematic method. Ref. [1] identifies the fractional orders by trial and error approach, which is feasible for identification of a small number of parameters, but becomes intractable for a large number of parameters as it is extremely time-consuming to transverse the whole parameter space. In this paper, to reduce the computational time, we apply the multi-fidelity model inversion method [13] to identify the fractional orders in such challenging FADEs. Numerical results show that not only the computational overhead is considerably reduced but also the parameters identified by our method lead to more accurate results than those estimated by the trial and error approach in ref. [1]. Specifically, the model inversion method we adopt is based on multi-fidelity Gaussian process regression (multi-fidelity GPR) and Bayesian optimization [13]. The method targets the accurate construction of response surfaces in parameter space, and an efficient pursuit to identify global optima while keeping the number of expensive function evaluations at a minimum. The method has been successfully applied to the calibration of outflow boundary conditions in blood flow simulations, as well as the parameter estimation in elliptic PDEs [13]. In the current paper, we extend the method to parameter identification of FADEs and we also use available field data in addition to synthetic data.

The outline of this paper is as follows. Section 2 introduces the basic ideas behind the multi-fidelity GPR as well as the Bayesian optimization which constitute the model inversion framework. For more details we refer the readers to ref. [13]. Section 3 gives a description of the parameter identification problem to be solved, followed by numerical results on parameter identification of 1-D variable-order and 2-D multiscaling constant-order FADEs. Discussion and conclusion are given in the last two sections. In the first three Appendices we include some more details of the convergence of the model inversion method, and in the last Appendix, we briefly introduce the implementation of the random walk method for solving the forward problem of 2-D FADEs.

2. Methodology

The section introduces the basic building blocks of the model inversion method. The multi-fidelity Gaussian process regression constructs a fast-evaluation response surface for the original expensive system function (or input-output mapping) using limited computational resources, and the Bayesian optimization adaptively searches the resulting response surface with the help of acquisition function and rapidly finds the optimum of the original system function.

2.1. Multi-fidelity GPR

A multi-variate Gaussian random vector $\mathbf{v}_N$ of N components corresponds to the probability density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{K}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\}, \tag{1}$$

where $\mathbf{m}$ and $\mathbf{K}$ are the mean vector and the covariance matrix, respectively. The index set for the random vector $\mathbf{v}_N = [v_1, v_2, \cdots, v_N]^T$ is a finite number set $\{1, 2, \cdots, N\}$. By extending the finite index set to an infinite continuous set, say $\Omega \subset \mathbf{R}^d$, we have the Gaussian process (GP),

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{2}$$

where $\mathbf{x} \in \Omega$ is the index and for the fixed $\mathbf{x}$, $f(\mathbf{x})$ is a Gaussian random variable; $m(\cdot)$ and $k(\cdot,\cdot)$ are the mean and the covariance functions, respectively. On any finite subset of the index set, $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, the corresponding random variables form a multi-variate Gaussian random vector, namely, $\mathbf{f}_N = [f(\mathbf{x}_1), f(\mathbf{x}_2), \cdots, f(\mathbf{x}_N)]^T$.

GP regression constructs the response surface based on the known input-output pairs (also known as training data). Suppose we have obtained N training data points, and the training data set $\mathbf{D}$ is given by

$$\mathbf{D} = \{(\mathbf{x}_1, f(\mathbf{x}_1)), (\mathbf{x}_2, f(\mathbf{x}_2)), \cdots, (\mathbf{x}_N, f(\mathbf{x}_N))\}, \tag{3}$$

where $f(\mathbf{x})$ for the fixed index $\mathbf{x}$ is the output of the system of interest when the system input is the very $\mathbf{x}$. Generally, the system output is contaminated by noise, and the final observation is actually

$$y(\mathbf{x}_i) = f(\mathbf{x}_i) + \epsilon_i, i = 1, 2, \cdots, N. \tag{4}$$

To simplify the problem, Gaussian white noise is usually considered, namely $\boldsymbol{\epsilon}_N \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I})$ where $\boldsymbol{\epsilon}_N$ is the vector of noise on all training data, $\mathcal{N}(\cdot,\cdot)$ is the normal distribution, σ_n^2 is the noise variance, and $\mathbf{I}$ is the identity matrix.

The goal of GPR is to predict the output f at arbitrary test input $\mathbf{x}_t$. Actually, this can be easily done by using the conditional distribution property for a multi-variate Gaussian random vector. We write the joint distribution of the test output $f(\mathbf{x}_t)$ and the noisy training data as a $(N+1)$-variate Gaussian random vector

$$\begin{pmatrix} f(\mathbf{x}_t) \\ y_(\mathbf{x}_1) \\ \vdots \\ y_(\mathbf{x}_N) \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_t) \\ m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_t,\mathbf{x}_t) & k(\mathbf{x}_t,\mathbf{x}_1) & \dots & k(\mathbf{x}_t,\mathbf{x}_N) \\ k(\mathbf{x}_1,\mathbf{x}_t) & k(\mathbf{x}_1,\mathbf{x}_1)+\sigma_n^2 & \dots & k(\mathbf{x}_1,\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N,\mathbf{x}_t) & k(\mathbf{x}_N,\mathbf{x}_1) & \dots & k(\mathbf{x}_N,\mathbf{x}_N)+\sigma_n^2 \end{bmatrix} \right), \tag{5}$$