Journal of Computational Physics 295 (2015) 189–208

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems

Kathryn Farrell, J. Tinsley Oden, Danial Faghihi

Institute for Computational Engineering and Sciences, The University of Texas at Austin, United States

ARTICLE INFO

Article history:
Received 26 September 2014
Received in revised form 18 February 2015
Accepted 30 March 2015
Available online 16 April 2015

Keywords:
Coarse graining models
Bayesian inference
Output sensitivities
Model plausibility
Model validation

ABSTRACT

A general adaptive modeling algorithm for selection and validation of coarse-grained models of atomistic systems is presented. A Bayesian framework is developed to address uncertainties in parameters, data, and model selection. Algorithms for computing output sensitivities to parameter variances, model evidence and posterior model plausibilities for given data, and for computing what are referred to as Occam Categories in reference to a rough measure of model simplicity, make up components of the overall approach. Computational results are provided for representative applications.

1. Introduction

The subject of model validation in the presence of uncertainties – in observational data, in model parameters, and uncertainties in the generally subjective process of selecting the model itself – lies at the very foundations of the scientific method. Scientific knowledge is acquired through observations of physical events and through the development of scientific hypotheses on the causes of the events. The former requires the acquisition of relevant observational data and the latter are the consequences of inductive logic leading to mathematical and computational models, the validity of which must be tested against experiments or observations.

Here we consider issues of model selection and validation in connection with coarse graining of atomic systems: the creation of models of atomistic systems by aggregating clusters of atoms into “beads” or “super atoms” so as to dramatically reduce the number of degrees of freedom and also to extend time scales in which physical quantities of interest can be observed. The overriding issues in developing coarse-grained (CG) models are how accurately they approximate key quantities of interest captured by the all-atom (AA) system; that is, is the CG model valid in some sense, and, more fundamentally, how does one select the CG model itself to faithfully represent relevant properties of the AA model? It is important that inherent uncertainties encountered in each step of the model validation and selection process must also be taken into account. In this study, a Bayesian framework is developed to address these issues.

A large literature on CG models exists, going back over a half century as early versions of CG approximations appeared in the 1940s, with more general approaches appearing decades later in computer implementations (*e.g.* [22,25]). The recent survey of Noid [42] of CG models of bio-molecular system cites almost 600 works and the survey of Li et al. [36] with over 400 references attests to the great interest in CG methodologies. Among methods proposed for constructing CG models

E-mail addresses: kfarrell@ices.utexas.edu (K. Farrell), oden@ices.utexas.edu (J.T. Oden), danial@ices.utexas.edu (D. Faghihi).

http://dx.doi.org/10.1016/j.jcp.2015.03.071
0021-9991/

using AA-model data, we mention force-matching methods [18,28–31], the so-called "multiscale coarse-graining" methods of Izvekov and Voth et al. [26,27,31,61,66] and Das and Andersen [16,17] which were shown to be "physically consistent" in [41,43,44], the iterative Boltzmann inversion methods of [54] in which CG parameters are chosen to fit specific probability distributions in AA systems, the Reverse Monte Carlo (RMC) method [38–40], the Conditional Reverse Work (CRW) method [9,10], and the methods of Shell et al. based on parameter selections that minimize the relative entropics of CG and AA systems (*e.g.* [12,13,19,60]).

The present work, which builds on the general framework for model validation and selection advanced in [21] and [48], follows somewhat different arguments than earlier approaches, as it attempts to make specific the processes of model selection and validation in the presence of uncertainties.

Our general approach toward selecting and validating CG models can be described as an adaptive modeling paradigm that employs Bayesian and information theoretics and consists of the following five components:

I *Parametric Model Classes*: A family of possible parametric model classes with unknown parameters $\boldsymbol{\theta}$ defining a particular CG interatomic potential.

II *Parameter Sensitivity Analysis*: A parameter sensitivity analysis performed over the entire set of $\mathcal{M}$ models. Here we employ a variance-based sensitivity analysis that computes the sensitivity of a representative output function (here the total potential energy) due to variations in parameters using Monte Carlo samples of parameter probability densities, retaining only parameters that significantly influence the output.

III *Occam Categories*: We partition the surviving models into categories numbered according to the number of unknown parameters in each model. Here we equate the number of parameters to the simplicity of the model in the spirit of Occam's Razor, knowing that this is not always a precise designation.

IV *Model Plausibilities and Parameters Calibration*: For models in a given set, we compute the Bayesian posterior plausibility of each model for given AA data in a calibration scenario. The use of model plausibilities in selecting CG models is discussed in [20,21,49]. The calibration scenario is viewed as an initial characterization of probability densities of parameters for models of representative molecules that are to be tested in one or more validation tests. Upon calculating plausibilities, the most plausible is selected.

V *Validation Tests*: One or more validation tests are designed to test the accuracy with which the most plausible model from IV predicts CG observables. A validation criterion must be specified to judge if the prediction is adequate and that, therefore, the model is not invalid ("valid"). If the model is valid, the process is terminated. The validated model is then used to solve the forward problem for the target quantities of interest (QoI) and the uncertainty in the QoI is quantified using standard measures.

Other proposals for Bayesian calibration methods for molecular systems can be found in the papers of Liu et al. [37], Español and Zúñiga [19], and Wang et al. [64], and more recently in the work of Angelikopoulos et al. [2,3]. Also relevant are the works of Das and Anderson [16,17] on their multiscale coarse-graining methods.

Following this introduction, we present an overview of modeling atomistic systems and creating a coarse-grained model. In Section 3, a Bayesian framework for building predictive models is detailed. Section 4 presents the Occam-Plausibility Algorithm, which pairs the concept of Occam's Razor with the Bayesian framework laid down in Section 3. An example application of this algorithm to coarse-grained polyethylene is given in Section 5. Concluding remarks can be found in Section 6.

2. The problem setting

2.1. The AA model

We begin with the general process of creating CG models designed to deliver approximations of properties of a "ground-truth" all-atom (AA) model, the properties of which are assumed to be known (the force field, parameters, *etc.*). The AA model thus provides synthetic observational data for calibrating and validating the CG model for specific observables and quantities of interest (QoI). For specificity, the behavior of the atomic system is assumed to be captured through molecular dynamics (MD) simulations, implemented using a hardened and generally accepted MD code. Also for simplicity, and without loss in generality, we restrict attentions to configurational energies of canonical ensembles of AA systems. In the present investigation, we employ LAMMPS [52]. A force field (the functional form of the potential energy) of the general form calibrated in [34,35] is employed,

$$u(\mathbf{r}) = V_{\text{bond}}(\mathbf{r}) + V_{\text{angle}}(\mathbf{r}) + V_{\text{dihedral}}(\mathbf{r}) + V_{\text{non-bonded}}(\mathbf{r}) + V_{\text{coulomb}}(\mathbf{r}), \tag{1}$$

where $\mathbf{r} = \mathbf{r}^n = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ is the vector of atomic coordinates in an n-atom system, and, as typical examples of energy characterizations,

$$V_{\text{bond}}(\mathbf{r}) = \sum_{i=1}^{N_b} \frac{1}{2} k_{ri}(r_i - r_{0i})^2, \tag{2a}$$