Computational Statistics and Data Analysis 83 (2015) 168–181

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A Monte Carlo approach to quantifying model error in Bayesian parameter estimation

Staci A. White *, Radu Herbei

Department of Statistics, The Ohio State University, Columbus, OH 43210, United States

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 26 August 2014
Accepted 6 October 2014
Available online 15 October 2014

Keywords:
ϕ-divergence
Kullback–Leibler
Hellinger distance
Model error

ABSTRACT

Quantifying the discrepancy between two distributions is considered, using the concept of ϕ-divergence. The motivation is a Bayesian inference scenario where one is interested in comparing different posterior distributions. Strongly consistent estimators for the ϕ-divergence between two posterior distributions are developed. The proposed estimators alleviate known computational difficulties with estimating normalizing constants. This approach can be used to study the impact that using an approximate likelihood has on the resulting posterior distribution and also to compare the effectiveness of different model approximations. The methodology is applied to two first-order emulator models and an oceanographic application where evaluation of the likelihood function involves the solution to a partial differential equation.

1. Introduction

In modern science we use models with ever increasing complexity, motivated by large, multivariate data sets exhibiting rich covariance structures. Unfortunately, in many such cases we are forced to use approximations of various kinds, given that analytical calculations are often impossible. As a classical example, Bayesian inference is typically done via Markov chain Monte Carlo (MCMC) methods and one would approximate samples from the posterior distribution with draws from a simulated Markov chain. Appropriate asymptotic theoretical results are making this approach acceptable to the scientific community. However, it is rare that one would quantify the magnitude of such an approximation and describe its consequences on the subsequent inference steps.

The current work is motivated by the following typical Bayesian approach to the inverse problem of estimating the coefficients of a partial differential equation (PDE), given noisy observations of its solution. We refer the reader to Stuart (2010), Wunsch (1996) and Abramovich and Silverman (1998) for some theoretical aspects of this problem. Briefly, one is interested in estimating a multi-dimensional parameter $\boldsymbol{\theta}$ representing the coefficients and boundary values of a PDE, based on observations $\mathbf{Y}$ representing discrete and noisy measurements of its solution, represented as *the forward map*: $\boldsymbol{\theta} \mapsto \mathsf{F}(\boldsymbol{\theta})$. Typically, this is an ill-posed inverse problem as there usually does not exist an inverse operator $\mathbf{Y} \mapsto \mathsf{F}^{-1}(\mathbf{Y})$. Moreover, as the data are contaminated by noise, it may be possible that $\mathbf{Y}$ is not even in the range of $\mathsf{F}(\cdot)$. In a Bayesian framework, one explores the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathsf{F})$, which is typically proper under very mild assumptions. However, in practice the forward map $\mathsf{F}(\cdot)$ is unavailable and is consequently replaced by an approximation $\tilde{\mathsf{F}}$, and one reports samples from the approximate posterior $p(\boldsymbol{\theta} \mid \mathbf{Y}, \tilde{\mathsf{F}})$. Our goal is to quantify the *model error*, which we define as the discrepancy between two distributions of interest. In this setting, model error is the difference between $p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathsf{F})$ and $p(\boldsymbol{\theta} \mid \mathbf{Y}, \tilde{\mathsf{F}})$.

* Corresponding author. Tel.: +1 614 292 5375.
E-mail addresses: white.1715@osu.edu (S.A. White), herbei@stat.osu.edu (R. Herbei).

http://dx.doi.org/10.1016/j.csda.2014.10.008
0167-9473/

The issue of model error resulting from using various approximations has been well documented in the literature. In the majority of cases, scientists account for model bias by introducing a new field which can be estimated through a hierarchical Bayesian framework. For example, Kennedy and O'Hagan (2001) account for model error in the predictions from computer model output by including a model inadequacy function to correct for inaccuracies in the computer model. The authors assign Gaussian process prior distributions to both the computer model output and model inadequacy functions. Their work has been extended to accommodate high dimensional computer model output by using the singular value decomposition to reduce the dimensionality (Higdon et al., 2008). In this case, prior distributions which are linear combinations of Gaussian processes on both the computer model output and the model inadequacy function are selected.

In a different framework, Majda and Gershgorin (2010) use empirical information theory to quantify the uncertainty in atmospheric/oceanic models. Empirical information theory builds the least biased probability measure that is consistent with observed measurements of functions of the variables of interest. The discrepancy between the least biased probability measure and a second probability measure is quantified using the relative entropy, or the Kullback–Leibler distance. The authors show that minimizing the Kullback–Leibler distance to the true model is equivalent to minimizing the distance to the least biased probability measure. An information theoretic approach has also been used in model selection to determine the effective number of parameters in a Bayesian hierarchical model, as in Spiegelhalter et al. (2002). This relies on the deviance which quantifies the difference between a model with a subset of parameters versus the saturated model. The authors develop a Deviance Information Criterion which takes the difference between the posterior mean of the deviance and the deviance at the posterior estimates of the parameters of interest.

The above methods have the common goal of improving accuracy in predictions. However, the aforementioned do not offer any indication on the impact of model bias when the goal is parameter estimation (rather than state prediction). Parameter estimation requires exploring the distribution of interest. When approximations are used a different distribution is explored, and the resulting inference will be affected. A first step in studying the impact that approximations have on inference procedures is to understand how different the two distributions under study are. Throughout the paper we use the term *model* to refer to a probability model, i.e., a probability distribution. In a more general framework, let $\boldsymbol{\theta} \in \mathbb{R}^d$ be a parameter of interest, and let $P(d\boldsymbol{\theta})$ and $G(d\boldsymbol{\theta})$ be two probability models defined over $\mathbb{R}^d$. In a Bayesian analysis, these would be two posterior distributions for $\boldsymbol{\theta}$ constructed under different scenarios. We understand that P is the ideal or *target* model which is computationally intractable. For instance, one can think of P as the model that would be used if there were unlimited computational resources. The *working* model G is tractable at least from a computational stand-point, and it represents an approximation of P. We aim to calculate or estimate the model error, i.e., a quantity $D(P, G)$ which describes how different P and G are. Estimating $D(P, G)$ allows one to assess the quality of model approximations with regard to their influence of parameter estimation.

Approximating target models is common in current practice and occurs for a variety of reasons. For example, Approximate Bayesian Computation is a large class of methods which rely heavily on approximations. In this case the likelihood function is analytically and computationally intractable, and one would approximate the posterior distribution using a rejection simulation approach. A recent overview of this approach is given in Marin et al. (2011). Another instance occurs when likelihood evaluation involves intractable integrals over very large spaces, see Beaumont et al. (2002) for an example in phylogenetics. In these situations, one integrates out nuisance parameters. However, since such integrals can rarely be evaluated exactly, the solution is to estimate nuisance parameters off-line and use these estimates in the likelihood evaluation, thus introducing model error. In Bayesian environmental applications, scientists use numerical solvers based on fine grids, discrete-time approximations, etc., without attempting to quantify the bias introduced by this practice (Wikle and Hooten, 2006; Nychka et al., 2002).

In our work, we select $D(P, G)$ to be a *ϕ-divergence* (defined below), and we suggest a general Monte Carlo method for estimating it. We work under the following general assumptions. The approximate model, i.e. $G(d\boldsymbol{\theta})$, can be explored efficiently and one could sample this distribution either directly or via a MCMC approach. We assume that the target model $P(d\boldsymbol{\theta})$ is very computationally demanding, and one would never consider running a MCMC sampler to explore it. Note that our methodology only requires $G(d\boldsymbol{\theta})$ and $P(d\boldsymbol{\theta})$ to have densities which are known up to a normalizing constant. In Section 2, we provide an overview of discrepancy metrics between probability measures. In Section 3, we introduce the proposed estimator for $D(P, G)$ and discuss its asymptotic properties. Section 4 illustrates our methodology in the context of emulator models, and Section 5 focuses on a more complex application from oceanography. Concluding remarks are given in Section 6.

2. Preliminaries

Let P and G be two probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We assume that P and G have densities p and g with respect to the Lebesgue measure μ, so that $p, g : \mathbb{R}^d \to [0, \infty)$. Let $\phi : [0, \infty) \to \bar{\mathbb{R}}$ be a convex function such that $\phi(1) = 0$. The ϕ-divergence between P and G is defined as

$$D_\phi(P, G) = \int_{\mathbb{R}^d} g(x) \cdot \phi\left(\frac{p(x)}{g(x)}\right) \mu(dx). \tag{1}$$

We note that $D_\phi(P, G)$ may not be a metric in the proper sense. However, (1) can still be useful for defining a notion of distance between probability measures. Using Jensen's inequality, it follows that $D_\phi(P, G) \geq 0$, and $D_\phi(P, G) = 0$ if and