Journal of Statistical Planning and Inference 140 (2010) 640 – 651

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Bayesian emulation of complex multi-output and dynamic computer models

Stefano Conti[a,*], Anthony O'Hagan[b]

[a] *Statistics Unit, Centre for Infections, Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK*
[b] *Department of Probability and Statistics, The Hicks Building, University of Sheffield, Sheffield S3 7RH, UK*

ARTICLE INFO

Article history:
Received 30 January 2007
Received in revised form
8 May 2009
Accepted 11 August 2009
Available online 19 August 2009

Keywords:
Bayesian inference
Computer experiments
Dynamic models
Hierarchical models

ABSTRACT

Computer models are widely used in scientific research to study and predict the behaviour of complex systems. The run times of computer-intensive simulators are often such that it is impractical to make the thousands of model runs that are conventionally required for sensitivity analysis, uncertainty analysis or calibration. In response to this problem, highly efficient techniques have recently been developed based on a statistical meta-model (the *emulator*) that is built to approximate the computer model. The approach, however, is less straightforward for dynamic simulators, designed to represent time-evolving systems. Generalisations of the established methodology to allow for dynamic emulation are here proposed and contrasted. Advantages and difficulties are discussed and illustrated with an application to the Sheffield Dynamic Global Vegetation Model, developed within the UK Centre for Terrestrial Carbon Dynamics.

1. Introduction

Large computer codes, implementing sophisticated mathematical models, are widely used in all fields of science and technology to describe and understand complex systems. We refer to any such program as a *simulator*. The size and complexity of a simulator can become a problem when it is necessary to make very many runs at different input values. For example, the model user may wish to study the sensitivity of model outputs to variations in its inputs, which entails many model evaluations when the number of inputs is large (as is very often the case). In particular, standard Monte Carlo-based methods of sensitivity analysis (extensively reviewed by Saltelli et al., 2000) typically require thousands of model runs. Another example is the practice of calibrating model parameters by varying them to fit a set of physical observations. Such explorations can become infeasible even for moderately large computer models requiring just a few seconds per run.

Following Sacks et al. (1989), a two-stage approach based on meta-modelling (*emulation*) of the simulator's response has been developed (see Haylock and O'Hagan, 1996; Kennedy and O'Hagan, 2001; Oakley and O'Hagan, 2002), offering substantial efficiency gains in terms of accuracy and computing time over standard Monte Carlo-based methods. These authors represent the simulator as a function $f(\cdot)$ which takes as input a vector $\boldsymbol{x}$ of parameters and produces an output $y = f(\boldsymbol{x})$. A Bayesian formulation assumes a Gaussian process prior distribution for the function $f(\cdot)$, conditional on various hyper-parameters. This prior distribution is updated using as data a preliminary *training sample* $\{y_1 = f(\boldsymbol{x}_1), \ldots, y_n = f(\boldsymbol{x}_n)\}$ of n selected simulator runs. Formally, the posterior distribution of $f(\cdot)$ is regarded as the emulator. This posterior distribution is also a Gaussian process conditional on the hyper-parameters; here conditioning upon the training set forces realisations from the emulator to interpolate the observed data points and induces posterior distributions for the hyper-parameters.

* Corresponding author. Tel.: +44 208 3277825; fax: +44 208 2007868.
E-mail address: stefano.conti@hpa.org.uk (S. Conti).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.08.006

The first stage of the two-stage approach is to build the emulator. Problems such as sensitivity analysis or calibration are then tackled in the second stage using the emulator. Since the emulator runs almost instantaneously, the computational cost of this approach for a large and computationally intensive simulator lies primarily in obtaining the training runs. Gains in efficiency arise through the emulation approach requiring far fewer simulator runs to achieve the same accuracy as Monte Carlo methods in tasks such as sensitivity analysis. Indeed, in practice the number of runs required is typically reduced by a factor of 100 or more, and it is usually possible to emulate the code output to a high degree of precision using only a few hundreds of training runs.

A number of research advances and applications dealing with statistical emulation of an ensemble of computer outputs were noted in recent years. Much of this work relies upon extensions of the univariate Gaussian process-based emulation framework, often in association with some dimension-reducing technique to ameliorate the complexity of the examined system. This was notably achieved through a principal component decomposition of the simulator's covariance structure (Higdon et al., 2008) or some basis function representation of its time-dependent, or otherwise functional, outputs, as attained via wavelets by Bayarri et al. (2007) and more widely discussed by Campbell et al. (2006). In these cases transformed and reduced outputs are treated as independent, effectively using what is referred to in Section 3 as the MS emulator. Although without explicitly referring to multi-output simulators, work developed by Qian et al. (2008) around the incorporation of qualitative, in addition to quantitative, factors into a Gaussian process emulator setting exhibits some similarity with the methodology herein proposed. Extensions to the conventionally employed Gaussian correlation function therein formulated can in principle be utilised to model dependence of the system of interest on time via an ordered categorical input, in a similar fashion to the TI emulator introduced in Section 3. Qian et al. (2008) also elaborate around an alternative 'independent analysis' approach, which basically coincides with the MS emulator discussed in Section 3. Outside the field of computer experimentation, Gelfand et al. (2004) formulate a non-stationary multivariate Gaussian point process in terms of a spatially varied linear model of coregionalisation to analyse multivariate data on commercial property transactions in three separate US real estate markets.

Often, the collection of outputs has a spatial and/or temporal structure. For instance, the oilfield simulator studied by Craig et al. (1996) outputs its predictions of the pressure at a given well over time, so that we can view these outputs as a time series. Similarly, the atmospheric dispersion model used by Kennedy et al. (2002) predicts deposition of radioactive particles at points on a spatial grid. Dynamic variation in underlying trends and stochastic volatility in a simulator output are also addressed by Liu and West (2009), whose strategy revolves around a time-varying auto-regressive model with stochastic innovations linked across the input space via a Gaussian process. Although existing theory for single-output emulation may be used to emulate each output individually, this can be a laborious process and may lose important information about correlations between outputs. The purpose of the present article is to propose a *multi-output* emulator, and to compare it with two other approaches based on single-output emulation.

We will base our analysis particularly on approaches to emulating *dynamic* simulators that model a system evolving over time, thereby producing a time series of outputs. One such model is the Sheffield Dynamic Global Vegetation Model (henceforth SDGVM), which is used to simulate the carbon dynamics of forests and other kinds of vegetation. The SDGVM will be used as a practical illustration of the performance of alternative emulation approaches. However, much of our discussion is relevant to emulating simulators which produce multiple outputs in other structures, for instance on a spatial grid or at different frequencies in a power spectrum.

The single-output Bayesian methodology elaborated by O'Hagan (1992), Oakley and O'Hagan (2002) is extended in Section 2 to enable the simultaneous emulation of a vector of outputs. In Section 3 we present three alternative approaches to modelling the output of a dynamic simulator based on single-output emulation, and contrast the assumptions of these methods with those of the multi-output emulator. The methods are contrasted in a practical example using SDGVM in Section 4. Section 5 discusses the benefits and limitations of the multi-output emulator, and contrasts it with other approaches to multiple outputs in the literature.

2. Emulating multiple outputs

We consider a deterministic simulator returning outputs $\boldsymbol{y} \in \mathbb{R}^q$ from inputs $\boldsymbol{x}$ lying in some (often high-dimensional) input space $\mathscr{X} \subseteq \mathbb{R}^p$. The simulator is essentially a function $\boldsymbol{f}: \mathscr{X} \longmapsto \mathbb{R}^q$, and due to its deterministic nature it returns the same output if repeatedly executed on the same set of inputs. Despite $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ being in principle known for any $\boldsymbol{x}$, in practice the complexity of the simulator requires the computer code to be executed in order to determine $\boldsymbol{y}$. From a Bayesian perspective, we thus regard $\boldsymbol{f}(\cdot)$ as an unknown function, and in line with e.g. O'Hagan et al. (1999) we represent the uncertainty surrounding it by means of the q-dimensional Gaussian process

$$\boldsymbol{f}(\cdot) \mid B, \Sigma, \boldsymbol{r} \sim GP(\boldsymbol{m}(\cdot), c(\cdot,\cdot)\Sigma) \tag{1}$$

conditional on hyper-parameters B, Σ and $\boldsymbol{r}$.

The multivariate Gaussian process is a straightforward extension of the univariate Gaussian process: in line with established results from probability theory, any q-variate Gaussian process $\boldsymbol{Y}(\cdot) \sim GP(\boldsymbol{\mu}(\cdot), \Gamma(\cdot,\cdot))$ can be expressed as a linear combination $\boldsymbol{Y}(\cdot) = \boldsymbol{\mu}(\cdot) + \mathrm{T}(\cdot,\cdot)\boldsymbol{z}(\cdot)$ of q i.i.d. standard univariate Gaussian processes $z_i(\cdot) \sim GP(0,1)$ for any symmetric positive-definite matrix $\mathrm{T}(\cdot,\cdot) \in \mathbb{R}_{q,q}$ such that $\Gamma(\cdot,\cdot) = \mathrm{T}(\cdot,\cdot)\mathrm{T}^{\mathrm{T}}(\cdot,\cdot)$ (the Choleski decomposition of $\Gamma(\cdot,\cdot)$ fulfils such requirement). Such decomposition,