Abstracts from the book "Analysis of scientific computer models. Methodology in computer simulator data analysis"

Ksenia N. Kyzyurova*

September 9, 2019

This monograph provides methodological foundation for analysis of data from scientific computationally challenging mathematical models. Mathematical topics include statistical modeling, Bayesian inference, stochastic processes, and decision-theoretic model assessment.

Prerequisites: an emulator of a computer model

Abstract: Science attempts to describe complex natural or engineering phenomena by use of mathematical processes. These computer models are usually either computationally slow and/or too resources demanding for operation of the model and sometimes even for storage of the data produced by the model. This makes the computer model output at *any* desirable input not available. However, fast approximations to such an output may be obtained, once an *emulator* of the model (its *statistical approximation*), using only a handful of input-output data points, is constructed. Construction of the 'default' emulator within its objective implementation is outlined in this chapter.

Chapter 1: Assessment of a statistical model

Abstract: Protagoras argued that a man is a measure of all things. Within the decision-theoretic approach this is mathematically shown to be indeed so: *scoring rules* calculated for predictive model evaluation and comparison are SUBJECTIVE. This means that the choice of a scoring rule for model comparison affects the results of the comparison and, therefore, the decision on a model choice. What's more, the scoring rules are hardly interpretable. Recommendation is to, instead, employ three independent *frequency estimates* of the quality of model predictions: (1) empirical frequency coverage, (2) predictive bias, and (3) uncertainty (variability) in predictions.

Chapter 2: Computer model with multivariate output

Abstract: Computer models often produce multivariate output for every single run of the model. There have been attempts to account for correlation among outputs in the construction of a Gaussian process emulator of such a model with the goal of achieving a more accurate emulator. Both, theoretical evidence and simulations are presented here which demonstrate that multivariate emulator does not lead to "better" (that is, more accurate, precise or less uncertain) emulation results compared to independent modeling of each component of the output.

*Current affiliation: The University of Sheffield, School of Mathematics and Statistics, the United Kingdom; web: kseniak.ucoz.net

Chapter 3: Approximation to a system of computer models

Abstract: Direct approximation of a system which consists of several computer models is difficult for computational and logistical reasons. The methodology of the linked Gaussian approximation has been demonstrated to be a successful alternative. This is outlined in this chapter.

Chapter 4: Calibration of a computer model

Abstract: If observations corresponding to the output of the model are collected, a theoretical model may be assessed on the agreement to the collected data. Moreover, given a computer model and collected data, one might inquire which values of inputs to the model could have generated the collected data; thus, performing *calibration* of a model.

Calibration is, therefore, analogous to finding an inverse image of function $f : Z \to Y$, given a set of n values $\mathbf{y} = (y_1, \ldots, y_n) \in Y$, that is $f^{-1}(\mathbf{y}) = \{\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n) \in Z : f(\zeta_i) = y_i \ \ \forall\ i = 1, \ldots, n\}$. However, this is indeed only an analogy, since in practice calibration involves (a) a computationally intensive computer model or a system of such models, and (b) *noisy* collected data which does not arise from the computer model itself but is obtained in either an experiment or observed in nature. Calibration framework which accounts for these two levels of complexity is presented in this study.

Chapter 5: Censoring for a computer model with zero-inflated output

Abstract: Computer model of a volcano pyroclastic flow, given a set of initial conditions, produces an output, maximum height of the flow at thousands of geographical locations. The output is non-negative and often results in exact zero, thus, indicating the absence of a flow, and resulting in that the zero-height value of an output has a *non-zero* probability to occur, as opposed to all other simulator output values. In order to account for these features of the output, the methodology of a censored GASP approximation to such a computer model is employed. Customarily employed, usual GASP, does not allow to construct an approximation which would incorporate these features of the simulator.

Chapter 6: Design of experiments for large-scale simulators

Abstract: A simulator is defined as LARGE-SCALE if the number of inputs is such that construction of its emulator (which involves optimization over its parameters) is prohibitively time-consuming. In order to facilitate the exploration of such a simulator, useful is to divide inputs $\mathbf{x}$ into two groups, that is $\mathbf{x} = (\boldsymbol{\kappa}, \boldsymbol{\lambda})$. First, design over the range of $\boldsymbol{\kappa}$, choosing, say, m points $\{\boldsymbol{\kappa}_i\}_{i=1}^m$. For each $\boldsymbol{\kappa}_i$ develop a Gaussian process emulator over the rest of inputs $\boldsymbol{\lambda}$. Second, for each set of *fixed* inputs $\boldsymbol{\lambda}$, an emulator over $\boldsymbol{\kappa}$ conditional on fixed $\boldsymbol{\lambda}$ is constructed.

This methodology may be used for facilitating parameter estimation and fast emulation of a model with many inputs. Depending on the purpose and implementation of this methodology in practice, but the Gaussian process over the entire input space may be lost, although useful approximations are constructed.