Statistica Sinica **33** (2023), 1-26
doi:https://doi.org/10.5705/ss.202020.0456

A CLUSTERED GAUSSIAN PROCESS MODEL FOR COMPUTER EXPERIMENTS

Chih-Li Sung, Benjamin Haaland, Youngdeok Hwang and Siyuan Lu

Michigan State University, University of Utah,
City University of New York and IBM Thomas J. Watson Research Center

Abstract: The Gaussian process is one of the most important approaches for emulating computer simulations. However, the stationarity assumption common to Gaussian process emulation and the computational intractability for large-scale data sets limit accuracy and feasibility in practice. In this article, we propose a clustered Gaussian process model that *simultaneously* segments the input data into multiple clusters and fits a Gaussian process model in each cluster. The model parameters and the clusters are learned through the efficient stochastic expectation-maximization, which allows for emulations for large-scale computer simulations. Importantly, the proposed method provides valuable model interpretability by identifying clusters, which reveal hidden patterns in the input–output relationship. The number of clusters, which controls the bias–variance trade-off, is efficiently selected using cross-validation to ensure accurate predictions. In our simulations and a real application to solar irradiance emulation, our proposed method has smaller mean squared errors than its main competitors, with competitive computation time, and provides valuable insights from the data by discovering clusters. An R package for the proposed methodology is provided in an open repository.

Key words and phrases: Large-scale data, mixture models, nonstationarity, solar irradiance emulation, uncertainty quantification.

1. Introduction

Gaussian processes (GPs) are popular modeling tools in various research areas, including spatial statistics (Stein (2012)), computer experiments (Fang, Li and Sudjianto (2005); Santner, Williams and Notz (2018); Gramacy (2020)), machine learning (Rasmussen and Williams (2006)), and robot control (Nguyen-Tuong and Peters (2011)). GPs provide flexibility for a prior probability distribution over functions in Bayesian inference, and the posterior can be used both to estimate the unknown function at an unknown point, and to quantify the uncertainty in this estimate. This explicit probabilistic formulation for GPs has proved to be powerful for general function learning problems. However, its use is

Corresponding author: Chih-Li Sung, Department of Statistics and Probability, East Lansing, MI 48824-1312, USA. E-mail: sungchih@msu.edu.