Journal of Statistical Planning and Inference 157–158 (2015) 1–15

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

On the Bayesian treed multivariate Gaussian process with linear model of coregionalization

Bledar Konomi *, Georgios Karagiannis, Guang Lin

Pacific Northwest National Laboratory, WA, United States

A R T I C L E   I N F O

Article history:
Received 9 September 2013
Received in revised form 28 May 2014
Accepted 1 August 2014
Available online 1 October 2014

Keywords:
Multivariate Gaussian process
Linear model of coregionalization
Bayesian treed Gaussian process
Markov chain Monte Carlo

A B S T R A C T

The Bayesian treed multivariate Gaussian process (BTMGP) and Bayesian treed Gaussian process (BTGP) provide straightforward mechanisms for emulating non-stationary multivariate computer codes that alleviate computational demands by fitting models locally. Here, we show that the existing BTMGP performs acceptably when the output variables are dependent but unsatisfactory when they are independent while the BTGP performs contrariwise. We develop the BTMGP with linear model of coregionalization (LMC) cross-covariance, an extension of the BTMGP, that gives satisfactory fitting compared to the other two emulators regardless of whether the output variables are locally dependent. The proposed BTMGP is able to locally model more complex and realistic cross-covariance functions. The conditional representation of LMC in combination with the right choice of the prior distributions allow us to improve the MCMC mixing and invert smaller matrices in the Bayesian inference. We illustrate our empirical results and the performance of the proposed method through artificial examples, and one application to the multiphase flow in a full scale regenerator of a carbon capture unit.

1. Introduction

Computer codes have recently gained popularity because they can simulate physical systems which many times are too costly to be observed in practice. Despite the availability of faster and parallelized computational resources, it is often too expensive to run such models for all possible input conditions. To overcome this computational barrier, several methods based on Gaussian processes (Cressie, 1993) have been proposed to build up surrogate models which can be used to predict the response surface using only a few observations. To the best of our knowledge, the first attempt of the statistics community to build a computer surrogate starts with the seminal papers of Currin et al. (1988) and independently Sacks et al. (1989).

For multivariate output the modeling of the cross-covariance function in the Gaussian process is crucial for the best representation of the data; see (Gelfand et al., 2010; Cressie and Wickle, 2011) for recent reviews. The separable cross-covariance model (Mardia and Goodall, 1993; O'Hagan et al., 1999; Oakley and O'Hagan, 2002; Conti and O'Hagan, 2010) has been used as an easy and computationally fast model to deal with multivariate spatial data and computer experiments. Two main limitations of the separable model are the symmetric property and the assumption that the correlation parameters are the same over the input space for each distinct output. The linear model of coregionalization (LMC) (Grzebyk et al., 1994; Wackernagel, 2003) is a more general model of the cross-covariance function which is based on linear transformations of independent latent processes. Different variations of LMC have been proposed to deal with the computational difficulties

* Corresponding author.
E-mail address: alex.konomi@uc.edu (B. Konomi).

http://dx.doi.org/10.1016/j.jspi.2014.08.010

and non-stationarity in the variance (Schmidt and O'Hagan, 2003; Gelfand et al., 2004). Another approach, based on latent dimensions, to model the cross-covariance function has been proposed in Apanasovich and Genton (2010). However, most of the above literature focuses on stationary cross-covariance functions.

To address non-stationary cases Gelfand et al. (2004) proposed a method based on the idea that varying the coefficients of the latent variables results in varying variance matrix spatially. However, this can model only a special case of non-stationarity since it does not allow for the spatial correlation to vary on space. Moreover, the implementation comes with a huge computational cost. Konomi et al. (2014) developed a multivariate model based on the Bayesian treed multivariate Gaussian process (BTMGP) with separable cross-covariance function that extends the Bayesian tree models proposed by Gramacy and Lee (2008) to the multivariate output. The proposed BTMGP with separable cross-covariance leads to low computational cost Bayesian inference in a non-stationary environment. However, the separable cross-covariance is limited to only model some particular types of dependencies. On the other hand the multiple BTGP cannot model the dependency between outputs. Both of these methods may have problematic behavior depending on the application problem. In specific, the traditional univariate BTGP performs well in the independent scenario but not in the dependent scenario. For instance, the existing BTMGP performs well in the dependent scenario but not in the independent scenario.

In this paper we extend the BTMGP with separable cross-covariance to that of BTMGP with LMC cross-covariance. The Bayesian tree can overcome most of the stationary LMC cross-covariance limitations. Moreover, the use of the Bayesian tree reduces the computational cost by fitting the multivariate Gaussian process independently in every MCMC iteration. In spite of these nice features of the Bayesian tree, the trans-dimensional reversible jump pair of moves in the Bayesian inference become cumbersome since a lot of parameters have to be proposed. In addition, sampling from the full posterior distribution of the joint LMC is a challenging task and the MCMC sampler may result in a very slow convergence (Gelfand et al., 2004). To solve these issues we utilize the conditional representation of LMC and assign a particular set of prior distributions. We manage to integrate out the parameters associated with the mean and the variance and reduce the number of parameters proposed in the trans-dimensional reversible jump moves. Moreover, inference based on the conditional representation of LMC becomes computationally easier since our method inverts smaller covariance matrices inside each external node of the Bayesian tree.

Given that the independent and the separable model are special cases of the LMC, one can expect the performance of the BTMGP with conditional representation of LMC cross-covariance to give better results in terms of prediction. To show this in practice, the proposed BTMGP model is compared, in several case studies, to the BTMGP with independent cross-covariance model, the multiple BTGP proposed by Gramacy and Lee (2008) and the BTMGP with separable cross-covariance model proposed by Konomi et al. (2014). We perform the comparison in two artificial examples, and one application in the multiple flow in a full scale regenerator of carbon capture unit.

Compared to BTMGP with separable cross-covariance model, significant improvements are shown when the spatial variation of the multivariate computer experiment is different for different outputs. This is shown mainly in the first illustration study. The simulation study shows that the proposed BTMGP is more robust than BTMGP with separable cross-covariance on possible deviations from the assumption of dependent output. Moreover, it maintains the good features of BTMGP with separable cross-covariance when the outputs are dependent. Compared to the multiple BTGP, in the proposed model significant improvements are shown when there is a dependency between outputs and similar results when the dependence assumption is violated. Significant differences are shown mostly in the second illustration study. Moreover, in the application we show improvement in the prediction intervals of the multivariate output.

The rest of the paper is organized as follows: In Section 2 we review the LMC, its variations, and the Bayesian tree. In Section 3 we describe the Bayesian inference and prediction for the Bayesian tree with coregionalization. In Section 4 we illustrate the BTMGP with LMC cross-covariance and compare it with BTMGP and multiple BTGP in artificial examples and real application of multiple flow in a full scale regenerator of carbon capture unit. Conclusions are presented in Section 5.

2. Model

Let us consider a physical problem with input (or spatial) domain $\mathcal{X} \subset \mathbb{R}^{k_x}$, where k_x is the dimension of the input (spatial) space. Let $\boldsymbol{\eta}(\boldsymbol{x}_i) \in \mathbb{R}^q$ denote the $q \times 1$ vector observed output at input $\boldsymbol{x}_i$, n denote the number of input (spatial) observations, $\tilde{\boldsymbol{Y}} = (\boldsymbol{\eta}^T(\boldsymbol{x}_1), \ldots, \boldsymbol{\eta}^T(\boldsymbol{x}_n))^T$ denote the $(nq) \times 1$ observed output vector and $\boldsymbol{Y} = (\boldsymbol{\eta}(\boldsymbol{x}_1), \ldots, \boldsymbol{\eta}(\boldsymbol{x}_n))^T$ denote the $N = n \times q$ observed output matrix.

2.1. Bayesian tree

The Bayesian tree provides a straightforward mechanism for modeling nonstationary data and can reduce the computational demands by fitting simple models locally. A Bayesian model averaging (BMA) approach allows for explicit estimation of predictive uncertainty, which can vary over space. In many applications, fitting a stationary multivariate GP may not be appropriate since the mean, the variance, and the spatial dependency may differ from one input subregion to the other.

The Bayesian tree (Chipman et al., 1998) partitions the input space in a tree form. Chipman et al. (1998) uses a linear model, and Gramacy and Lee (2008) uses a Gaussian process inside of each external node. In the multivariate case Konomi